Essays on Methods for Causal Inference

Inaugural-Dissertation zur Erlangung des akademischen Grades

eines Doktors der Wirtschaftswissenschaft

Fachbereich Wirtschaftswissenschaft

Freie Universität Berlin

vorgelegt von

Patrick F. Burauel

Berlin, 29. Januar 2020

Erstgutachter

Prof. Dr. Carsten Schröder, DIW Berlin & Freie Universität Berlin

Zweitgutachter

Assistant Prof. Christoph Breunig, PhD, Emory University

Datum der Disputation: 3. Juli 2020

Essays on Methods for Causal Inference

Patrick F. Burauel

January 2020

Acknowledgements

I thank my advisors Carsten Schröder and Christoph Breunig for their continued encouragement and support throughout my time at the DIW Berlin while working on this dissertation. Collaborating with each has been enriching in my development as a researcher and has been invaluable in moulding intitially rather amourphous research ideas into well-formed research papers. Their support of my projects ranging from rather unorthodox choices for summer schools to rather non-economicsy research stays abroad has been invaluable to my development.

I am particularly grateful to Carsten Schröder for giving me ample room to read widely, to dive into topics that drew my curiosity, to develop my ideas freely without the stifling aversion to new perspectives. Beyond the exciting research hustle and bustle, I appreciate him knowing that sometimes there are more important things in life than research, and his acting accordingly.

I thank Uri Shalit for welcoming me at the Technion in Haifa to do research on causality. The collaboration with him and the time in Israel have a long-lasting buoying impact that I do not want to miss.

I thank my colleagues at the DIW Berlin, especially the 2015 GC cohort which I am happy to be part of; we walked through sunflower fields to Café Einstein together, and there's a lot more to come. I thank my fellow PhD students and colleagues at the SOEP for uncountable encounters, for discussions about common trends and free will. The red carpet is for you.

Collaboration with Coauthors and Publications

Kapitel 2: The German Minimum Wage and Wage Growth: Heterogeneous Treatment Effects using Causal Forests

- Kapitel 2 basiert auf einem bisher unveröffentlichten Artikel, der zu gleichen Teilen mit Carsten Schröder verfasst wurde.
- Burauel and Schröder (2019)

Kapitel 3: A Reverse Causality Test Without Instruments

- Kapitel 3 basiert auf einem bisher unveröffentlichten Artikel, der zu gleichen Teilen mit Christoph Breunig verfasst wurde.
- Breunig and Burauel (2020)

Kapitel 4: Structural Autonomy and Instrument Validity

- Kapitel 4 basiert auf einem bisher unveröffentlichten Artikel, der in Alleinarbeit angefertigt wurde.
- Burauel (2020)

Contents

1 Introduction

This dissertation consists of three papers sharing the objective to analyze how machine learning (ML) methods can be useful to economists and econometricians in their pursuit to understand causal mechanisms operating in the economy.[1] Such causal knowledge is essential when designing policies that help achieve societal goals. ML techniques are increasingly applied in and adapted to practical policy settings. These settings share the types of endogeneity problems that make actionable inference from data difficult with the domains that economics is occupied with. Thus, there are many potential synergies between ML and economics that are surfacing on both the academic and policy-making agendas. In the context of this dissertation, it is useful to make a distinction between two points of interchange between the two fields. First, ML can be used to improve or extend widely-used identification techniques in economics and, second, insights into causal modeling from the ML community can be introduced as novel routes to identification in economics. The first paper of this dissertation falls in the former, the second and third paper in the latter category. I briefly introduce each in turn.

As machine learning methods excel at prediction tasks, much of the existing work leverages this comparative advantage by either focusing on problems where a superior prediction in itself is the objective (see Kleinberg et al., 2015) or where it serves as one brick within a broader methodology. The latter is addressed here. Many techniques for causal identification in observational studies have an element of pure prediction. For example, the superior predictive power of ML techniques can be employed in the first stage of an instrumental variable regression (Hartford et al., 2016) or to estimate propensity scores (Cannas and Arpino, 2019). Furthermore, regularization and systematic model selection are gaining prominence in e.g. models for demand estimation and, more generally, structural econometrics (Bajari et al., 2015). Systematic model selection is particularly

[1] The exact scope of the term ‘machine learning’ is contentious. I use the term liberally denoting all work on modeling data originating in the ML community, while acknowledging that many ‘machine learning algorithms’ are well-known and widely-used in economics; after all, ML textbooks often start with a description of ordinary least squares regression. Whether the reader wants to categorize the literature that I draw upon in this work as ‘machine learning,’ ‘computer science,’ ‘statistics,’ or ‘philosophy’ should not distract from the contributions that go beyond these semantics.

important in machine learning because overfitting is a common pitfall due to the flexibility of the employed models, most evidently seen in deep neural networks (Goodfellow et al., 2016). Though *a priori* unrelated to causal inference, Athey and Imbens (2016) combine one such flexible modeling technique, the random forest (Breiman, 2001), with the dominant conceptualization of causality in economics, namely the potential outcomes framework (Rubin, 2005). They show how random forests are used for estimating heterogeneity of causal treatment effects in a data-driven manner. In the first paper of this dissertation, we adapt the causal forest methodology proposed by Athey et al. (2019) to estimate heterogeneous treatment effects in difference-in-differences studies and analyze heterogeneous effects on wage growth of the 2015 introduction of the statutory minimum wage in Germany.

The starting point for the second and third paper of this dissertation is the second point of interchange. There is a tendency to argue that ML techniques are about “prediction and prediction only" (Agrawal et al., 2017). See also Mullainathan and Spiess (2017) who state that “machine learning revolves around the problem of prediction" (p. 88). However, above and beyond the idea that superior prediction can be useful in causal inference problems, developments in the ML community question this dictum: Techniques to model causal relations and to identify them from observational data are emerging (for a survey see Peters et al., 2017).

The literature on causality in the computer science community is pioneered by Pearl (2009, the first edition of which was published in 2000), whose conceptualization is commonly referred to as the ‘graphical approach’ to causality. I follow this nomenclature. Pearl’s work, along with the literature it spawned, is only paid scant attention by economics. However, with the increased interest in applications of machine learning in economics that Athey and Imbens (2016) epitomizes, interest in the work on causality originating in the computer science community seems to be surging as well. Guido Imbens, who widely contributed to (and argues for) the rivaling potential outcomes framework, states that Pearl’s graphical approach to causality “has not had as much impact in economics as it should have" (Imbens, 2019, pp. 1). This constitutes a change in attitude relative to earlier claims that “economists have not felt that graphical models have much to offer them" (Imbens, 2014b, p. 376). An insightful overview of how the two approaches can benefit from an appreciation of each other’s strengths is given in Bareinboim and Paul (2019). In the graphical approach to causality, causal relations are represented in structural equation models, which are accompanied by graphical representations of the causal links. Such representations are closely linked to the notion of structural invariance or the autonomous nature of causal relations that is due to Frisch, Haavelmo and fellow Cowles Commission members (Haavelmo, 1944, see Appendix 4.10.6 for more details).

A central tenet of causal machine learning is that the observed joint distribution of a number of random variables contains causal information in the form of such invariance properties. This causal information can be exploited by appropriate statistical techniques. In that sense, the causal machine learning literature offers novel pathways to causal understanding that are not yet exploited in economics. The originality of the second and third paper lies in exploring the potential of these novel pathways: In the second paper, a test for reverse causality that relies on the insight that imposing functional form assumptions can help identify the causal direction between two observed variables is suggested. In the third paper, a test for instrument validity is proposed, relying both on the notion that causal relations are autonomous and on a method to quantify to which extent an observed statistical relation describes autonomous, i.e. causal, mechanisms, which is introduced by Janzing and Schölkopf (2018).

2 The German Minimum Wage and Wage Growth: Heterogeneous Treatment Effects using Causal Forests

This chapter is based on joint work with Carsten Schröder, see Burauel and Schröder (2019). We thank participants at the SOEP User Conference 2018, the 34th Annual Congress of the European Economic Association 2019, the 6th PhD Workshop in Empirical Economics at Potsdam University, and the SOEP Applied Panel Analysis Brownbag Seminar for their comments.

2.1 Introduction

A broad economic literature seeks to understand how public policies change socio-economic outcomes. Standard micro-econometric workhorses to analyze such policy changes are difference-in-differences and regression discontinuity designs. To better understand effect heterogeneities, i.e. differences in policy-induced changes by population subgroups, either models are estimated by subgroup, or incorporate interactions between treatment and subgroup dummies. Both approaches incur problems of multiple hypothesis testing: as the number of subgroups increases, the likelihood of erroneous inferences increases. While statistical approaches to address multiple testing problems exist (e.g. Bonferroni or Benjamini-Hochberg adjustments) and pre-analysis plans may help narrow the number of potentially relevant subgroups, such plans are selective. Usually, they focus on a relatively small number of groups, such that unexpected heterogeneities across groups that are determined by more complex interactions of covariates remain unobserved. Specifically, we show how previously found heterogeneities can turn out to be spurious when interactions of covariates with the treatment indicator are taken into account.

The causal forest approach provides an alternative statistical framework. It allows an evaluation of heterogeneous treatment effects for randomized control trials without the need to specify pre-analysis plans (Athey and Imbens, 2016). A regression tree is a popular machine learning algorithm that systematically splits the covariate space into recursively smaller subsets and estimates the value of an individual's outcome Y_i as the mean outcome of those Y_j with similar covariates. The estimation involves a parameter that penalizes model complexity. Since this parameter and the structure of the tree are estimated on independent subsamples, overfitting is avoided. Athey and Imbens (2016) modify such regression trees to optimize for differences in treatment effects rather than to maximize the mean squared predictive error. This paper relies upon the flexible moment-based implementation provided by Athey et al. (2019). This paper relies upon Athey et al. (2019), which is an extended and flexible moment-based implementation of the original idea in Athey and Imbens (2016).

A fruitful area of application is the evaluation of the effects of minimum wages. Effect heterogeneities – for instance with respect to employment, working hours, and wage changes – are expected in this context because of productivity differentials across groups of employees and between regions or the ease with which an employer can track hourly productivity. For instance, Ahlfeldt et al. (2018) analyze spatial heterogeneity of wage convergence in a difference-in-differences setting using the 2015 German minimum wage introduction as the institutional background. Bonin et al. (2018) use variation in regional treatment intensity to identify a decline in marginal employment that is larger in regions with high treatment intensity. In contrast, a decline in regular employment is not found. Burauel et al. (2020) (which we henceforth refer to as B20) analyze heterogeneity in effects on hourly wage growth for marginal and regularly employed as well as for East and West Germany. Overall, the credibility of approaches that rely on *ex ante* definitions of specific subgroups is challenged by the multitude of possible sub-populations and interaction effects of the grouping variables. We complement such approaches by adapting causal forests to infer heterogeneity in treatment effects without an *ex ante* group categorization in difference-in-difference settings. This allows an assessment of whether previously observed heterogeneities are spurious and instead result from more complex interactions of covariates.

Germany's 2015 minimum wage introduction at EUR 8.50 serves as our institutional background. The case of Germany is particularly interesting because the introduction was a high-impact labor-market intervention: almost all employees are eligible and the minimum wage has a considerable bite (more than 10% of all eligible employees earned less than it in the year prior to its introduction). By using the same data and treatment-control-group design as a previous study, B20, we assess to what extent the effect heterogeneities reported therein are driven by more complex underlying interactions of covariates. We analyze heterogeneity in treatment effects associated with pre-reform characteristics and employment situation of the eligible workers.

B20 implement a differential trend-adjusted difference-in-differences design (DTADD). In their setting, the inter-temporal changes in wages of a treatment group are compared with the changes for a control group. Employees with wages below the minimum wage in 2014 form the treatment group. Defining a control group is not straightforward since the minimum wage applies to all employees in principle. The authors address this issue by defining employees with wages slightly above the minimum wage as control group. Spillover effects are a challenge to this route to identification since employees earning slightly above the minimum wage cannot serve as control if their wages are indirectly affected by the introduction of the minimum wage. That both employees and employers might have a motivation to keep the wage structure constant thereby leading employers to increase wages also for employees earning above EUR 8.50 is a potential reason for such spillover effects. Though there is empirical evidence for such effects in the US (Brochu et al., 2015; Neumark and Wascher, 2004), no such effects can be discerned in the context of the 2015 minimum wage introduction in Germany (Caliendo et al., 2019). To avoid lagged responses to the reform in the first months of 2015, the authors use wage growth over two years, namely between 2014 and 2016, as the main outcome variable. The authors find an intention-to-treat (ITT) effect of 6.5 percentage points of additional growth in hourly wages that can be causally attributed to the minimum wage introduction.[1] Regression by subgroups suggest larger treatment effect for marginally relative to full-time employed as well as for residents in East relative to West Germany. It is subject of the study to check whether these results are spurious in the sense that they vanish after taking interactions of other covariates into account.

We first replicate these estimations and then adapt the forest methodology proposed by Athey et al. (2019) to study the extent of effect heterogeneities, paying particular attention to the above reported heterogeneities across types of employment and regions of residence. In sum, the forest methodology reveals substantial effect heterogeneities: subgroup-specific conditional intention-to-treat effects (CITEs) range from about 1.5 percentage points to about 13 percentage points. The nature of this heterogeneity is determined by complex interdependencies of employer-employee characteristics, including firm size, nature of the employment contract, skill degree of the occupation etc. Our analysis reveals that previously reported higher treatment effects in East Germany turn out to be spurious after interactions of employer-employee characteristics are taken into account, while higher ITT for marginally employed do not.

Our fine-grained information on CITEs is most interesting to policy makers as it can,

[1] They estimate an ITT, not an average treatment effect (ATE), since there is non-compliance; a considerable part of the eligible population does not earn the minimum wage even after its official introduction (Burauel et al., 2018).

for instance, be used to investigate whether those groups of employees whose wages were lowest before the reform experienced the highest wage increases. Our analyses show that this goal was only partially achieved and that not all eligible groups receive a lawful wage following the minimum wage introduction.

The article is structured as follows. Section 2.2 provides some background on the minimum wage reform in Germany and the data we use. Section 2.3 explains the causal forest methodology and how we adapt it to the study at hand. Section 2.4 provides the results. Section 4.9 concludes.

2.2 Application to evaluation of the minimum wage reform

2.2.1 Data and descriptive statistics

We use data from the German Socio-economic Panel (SOEP). The SOEP is a panel study surveying about fifteen thousand households every year. It provides information on a wide variety of socio-economic variables such as household composition, income, job characteristics, education, life satisfaction etc., see Goebel et al. (2018). Most importantly for our purposes, it provides detailed information on agreed working hours (weekly) and gross earnings (monthly), thus enabling us to derive the core variable of our analysis: agreed hourly wages. This wage concepts differs from actual hourly wages, which is not subject of this study. Thus, we refer to agreed hourly wages as hourly wages.

The SOEP consists of several subsamples that, together and weighted, represent the entire population of Germany. In a typical year of our observation period, SOEP includes about 16,000 employed individuals. To ensure comparability, we replicate the sample restriction employed in B20.[2] We exclude employees from the sample who are either not eligible for the minimum wage or work in sectors where sector-specific minimum wages existed prior to the reform. In this paper, we primarily utilize the longitudinal sample, focusing on the period around the reform (2010-2016). Because the SOEP field work mostly takes place in the first half of a year and previous studies report delays in implementation (see, for example, Caliendo et al., 2017), we study wage changes between two consecutive years, e.g. between 2014 and 2016. Therefore, individuals that are not observed in $t+2$ are dropped from the sample in survey year t. Individuals that lose their job between t and $t+2$ are dropped from the sample since their hourly wages are undefined despite the fact that such job losses might be attributable to the minimum wage introduction. As

[2] See B20 for further details on the sample selection.

a natural consequence, our results pertain to those workers who stay employed. In any case, such effects will be small, since there is only weak evidence for employment effects in the short-run (Caliendo et al., 2018; Bonin et al., 2018; Bossler and Gerner, 2016) and the number of individuals dropped due to job loss are small. Furthermore, due to item non-response, we do not have access to critical information such as job characteristics in $t + 2$ for some individuals. Consequently, these are dropped from the sample. The sample restrictions applied throughout the paper are summarized in Table 2.1. In sum, only part of the cross-sectional sample fulfills these requirements and we lose roughly a third of observations by moving from cross-section to panel setting. This raises concerns about the representativeness of the panel sample, which we refute in Table 2.7 in Appendix 2.6.3 by showing that descriptive statistics of important variables remain largely unchanged when moving from the cross-section to the panel sample.

Table 2.1: *Working Sample Size*

	2012	2013	2014	2015	2016	Total
Employed	16,155	18,199	16,066	15,822	14,895	81,137
Hourly wage undefined	-3,734	-4,236	-3,392	-3,553	-3,445	-18,360
Exempt from minimum wage or has sector-specific minimum wage	-2,522	-2,904	-2,458	-2,727	-2,447	-13,058
Cross-Sectional Sample	9,899	11,059	10,216	9,542	9,003	49,719
Not observed in $t + 2$	-3,341	-4,026	-3,336	-/-	-/-	-29,248
Job loss	-62	-51	-75	-/-	-/-	-188
Missing information	-363	-279	-330	-/-	-/-	-972
2-Year Panel Sample	6,133	6,703	6,475	-/-	-/-	19,311

Source: SOEP v33 2012-2016, own calculations.

2.2.2 Definition of objective variables and descriptive statistics

SOEP respondents are asked about individual monthly gross earnings (w_{gross}), agreed (h_c) and actual weekly working hours. Following B20, We focus on agreed hourly wages, defined as gross monthly income divided by agreed working hours per month (which is calculated as 4.33 (weeks in a month) times observed weekly working hours).

Monthly earnings include payments for overtime work. Therefore, if overtime work is not compensated through work time adjustment and paid out, agreed hourly wages exceed the effective wages of the employed. That earnings do not include special payments such as holiday payments or profit bonuses works in the opposite direction, although it should be noted that such bonuses are typically paid out at the higher end of the earnings distribution. Against this background, our wage concept should be viewed as conservative.

Table 2.2 shows average monthly gross wages, average agreed weekly working hours, as well as the average of hourly wages. In the pre-reform period, annual growth of monthly earnings is about 1%. In 2015, when the minimum wage is introduced, earnings grow by about 4%, between 2015 and 2016 again by about 1%. Annual growth of hourly wages is below 1% prior to the reform, almost 4% between 2014 and 2015 and about 1% between 2015 and 2016. Average working hours hardly differ over time.

Since our aim is to analyze the extent of treatment effect heterogeneity across employee groups, Table 2.3 lists for several groups the shares below and above the minimum wage. The groups are distinguished by gender, geographic location (East and West Germany), migration background, and work arrangement (full-, part-time, marginally employed).

Altogether, about 14.5% of the eligible population received less than the minimum wage prior to the reform. This share declines but does not vanish with the introduction of the minimum wage, suggesting considerable non compliance issues (for a more detailed discussion see B20). In 2016, the share of non-compliance amounts to about 10%.

There is considerable heterogeneity in group composition: In the low-wage group, there is a strong over-representation of employees which are female (about 71%, compared to 30% male in 2014), are resident in East Germany (about 32%, relative to 68% in West Germany in 2014), have a migration background (about 16%, relative to 85% withouth migrational background in 2014), and work as part-timers (about 16% in 2014) or marginally employed (about 38% in 2014). We define all individuals who do not indicate 'German' as their nationality as having a migrational background. Part-time employed are those individuals whose weekly working hours reach a maximum of thirty hours. Marginally employed are those individuals whose monthly gross income is a maximum of EUR 450. Full- and part-time employed are also socially-insured.

Another way of visualizing this heterogeneity are Pen's Parades (Pen, 1971). These are constructed by ranking all individuals according to their agreed hourly wage and plotting their agreed hourly wage as a function of their wage percentiles. We show such Pen's

Table 2.2: *Descriptive Statistics of the Working Sample*

	2012	2013	2014	2015	2016
Monthly gross earnings [EUR]	2651.22	2674.80	2703.05	2818.06	2846.49
	(1539.18)	(1582.59)	(1639.14)	(1684.00)	(1685.27)
Weekly working hours	34.38	34.29	33.75	34.03	33.98
	(9.71)	(9.56)	(9.96)	(9.78)	(9.77)
Agreed hourly wages [EUR]	17.32	17.48	17.88	18.54	18.74
	(8.54)	(8.76)	(9.06)	(9.25)	(9.24)
Observations	9899	11059	10216	9542	9003

Note: Own calculations based on cross-sectional sample. Weighted by cross-sectional weights. Standard errors in parentheses. *Source:* SOEP v33 2012-2016.

Table 2.3: *Composition of the sample below and above minimum wage.*

	2014		2015		2016	
	< MW	≥ MW	< MW	≥ MW	< MW	≥ MW
Female	70.5	46.5	70.0	47.2	69.2	47.7
Male	29.5	53.5	30.0	52.8	30.8	52.3
East	32.2	17.7	27.0	18.0	26.4	17.8
West	67.8	82.3	73.0	82.0	73.6	82.2
Migration background	16.2	8.6	15.5	9.2	19.4	9.6
No migration background	83.8	91.4	84.5	90.8	80.6	90.4
Full-time	46.5	80.4	47.5	80.6	45.5	80.7
Part-time	15.6	15.9	17.2	15.7	19.0	15.3
Marginally employed	37.8	3.7	35.4	3.7	35.6	4.0
Observations	1184	9032	846	8696	710	8293

Note: MW denotes minimum wage. All estimates based on agreed working hours, cross-sectional sample. Weighted by cross-sectional weights *Source:* SOEP v33 2014-2016, own calculations.

Parades for regularly employed and marginally employed employees in Figure 2.1. Several observations are noteworthy. The Pen's Parade for regular employees is much steeper than it is for the marginally employed. Among the marginally employed, a much larger fraction was remunerated below the EUR 8.50 before (roughly 60%) as well as after (roughly 40%) the introduction of the minimum wage. Importantly, the wage increase experienced by the marginally employed from 2014 to 2016 (i.e. from the dotted to the black line) is much larger than that by the regularly employed. To what extent such a startling difference

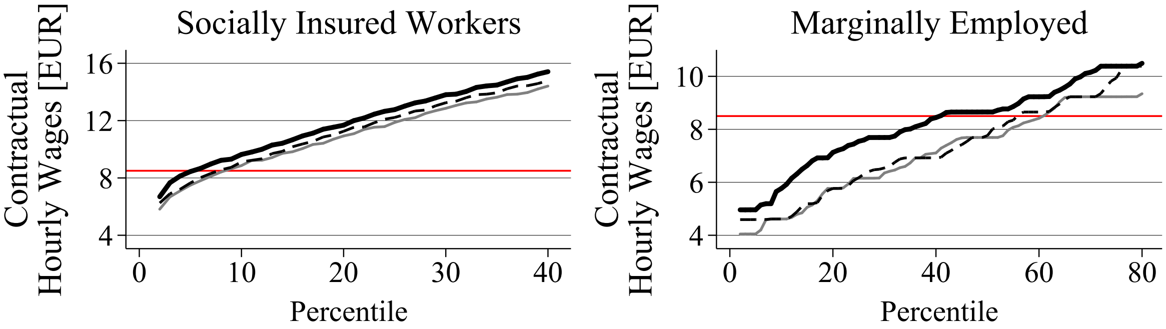


Figure 2.1: Pen's Parades. *These Pen's Parades show the level of agreed hourly wages as a function of the percentile in the agreed wage distribution. The left panel shows Pen's Parades for socially-insured or regularly employed workers, the right panel for marginally employed workers. Red horizontal line indicates the level of the minimum wage of EUR 8.50. Grey line: 2012, dotted line: 2014, black line: 2016. Source: own calculations, SOEPv33, 2012-2016*

is causally explained by the minimum wage introduction and whether there are similar differences across other subgroups and combinations of subgroups is the subject of the causal tree methodology.

2.3 Methodology

To understand effect heterogeneities of the minimum-wage introduction, we first estimate the overall effect of the reform on hourly wages. The estimation relies on a standard difference-in-differences (DiD) setting, with the treatment group being employees whose wages were below the minimum wage in 2014 and the control group being employees whose wages were slightly above. Next, we derive effects by different subgroups using the forest methodology.

2.3.1 Step 1: Estimating the average intention-to-treat effect

Difference-in-differences approaches rely on a common trend assumption that demands the dependent variable to have similar trends in treated and control groups absent the treatment. As shown in B20, the validity of this assumption cannot be taken for granted in the present context since hourly wage trends of treatment and control groups differ already prior to reform. More, specifically, people at the lower end of the wage distribution tend to have larger yearly wage increases than those at the higher end of the wage distribution regardless of the introduction of the minimum wage. Hence, following Stewart (2004) and B20, we employ a differential trend adjusted DiD estimator (DTADD), assuming that the *differences* in wage growth dynamics between treated and control groups remain constant over time. Thus, the treatment effect δ on wage growth Δy_{it} is identified by

$$\underbrace{[\mathbf{E}(\Delta y^t_{it=2014}) - \mathbf{E}(\Delta y^t_{it=2012})]}_{\text{factual}} - \underbrace{[\mathbf{E}(\Delta y^c_{it=2014}) - \mathbf{E}(\Delta y^c_{it=2012})]}_{\text{counterfactual}}$$

where superscripts t and c denote treated and control groups respectively and $\Delta y_{it} = \log\left(\frac{y_{it+2}}{y_{it}}\right) \times 100$ is wage growth between t and $t+2$.[3]

The DTADD approach relies on three differencing steps. First, we calculate the log difference in wages in each group (treated or control) in pre- and post-treatment periods; i.e. we calculate Δy_{it} for each group. Second, we take differences between group-specific changes in wage dynamics in the period before (2012 to 2014) and after the minimum-wage introduction (2014-2016) ($\mathbf{E}(\Delta y^t_{it=2014}) - \mathbf{E}(\Delta y^t_{it=2012})$ and $\mathbf{E}(\Delta y^c_{it=2014}) - \mathbf{E}(\Delta y^c_{it=2012})$

[3] As indicated in Section 2.2, we study wage growth over two years to abstract from possible delays in implementation.

respectively). Third, we take differences between treatment ('factual') and control group ('counterfactual').

The pooled OLS regression model that we use to implement this identification strategy takes the form,

$$\Delta y_{it} = \beta_0 + \delta(W_{it}\mathbf{1}_{t=2014}) + \delta_0(W_{it}\mathbf{1}_{t=2012}) + \beta_1 W_{it} + \beta_2 \mathbf{1}_t + \beta_\mathbf{3} \mathbf{Z_{it}} + \varepsilon_{it}, \tag{2.1}$$

with wage growth between t and $t+2$ as dependent variable, $t \in \{2010, 2012, 2014\}$ denoting the time periods, and ε_{it} an individual error term.

The term W_{it},

$$W_{it} = \begin{cases} 0, & \text{if } y_{it} \in [8.50, 10] \\ 1, & \text{if } y_{it} < 8.50, \end{cases} \tag{2.2}$$

distinguishes observations of the treatment and control group, while $\mathbf{1}_t$ denotes time dummies for 2012 and 2014. The vector $\mathbf{Z}_{it}$ contains a list of socio-economic characteristics: age, gender, marital status, migration status, level of education, presence of kids below age 16 in the household, East/West, as well as several job characteristics (type of contract (full-time, part-time, marginally employed), dummy for a temporary contract, size and sector of the firm; dummies indicating whether the employee changed jobs, sectors or firm size, moved into a job that is not eligible for the minimum wage, changed to or from a temporary contract). The coefficient δ is the treatment effect of interest, while δ_0 is the placebo treatment effect, which should not be significantly different from zero.

B20 estimate the model in eq. (2.1) for the full sample and then re-estimate the model for employees distinguished along two dimensions: employment status and region of residence. Their evidence suggests larger treatment effects for marginally employed and residents in East Germany. In Section 2.4.2, we assess whether these differences are attributable to the above-mentioned dimensions or whether there are more complex patterns, i.e. interactions of several covariates, that actually drive the differences.

The DTADD strategy has higher data requirements compared to a regular DiD approach as we need to compute the difference in wage dynamics between control and treated groups prior to the reform. The pooled OLS framework effectively treats individuals observed in different time periods as independent observations, which might create bias as it implies neglecting individual-specific effects. While poolability can be tested, the time-varying definition of the treatment dummy W_{it} in combination with the identification design precludes the use of a regression design with individual-specific fixed effects in any case. We relegate the detailed discussion of the reasons to Appendix 2.6.6 since the paper at hand takes the identification strategy proposed in B20 as given. As our main aim is to study treatment heterogeneity in the setting of B20, we stick to the pooled OLS framework.

2.3.2 Step 2: Estimating effect heterogeneities

Equation (2.1) does not account for potential heterogeneity of the treatment effect since interactions between the treatment dummy and covariates as well as high-order interactions of covariates, which would capture the heterogeneity, are not included as additional terms in the model. As discussed, the number of interaction terms would become prohibitively large due to the large number of potential subgroups. Subsequently, we are interested in heterogeneity of treatment effects across groups that are implicitly defined by a set of variables $\mathbf{X}$ that is only partly overlapping with $\mathbf{W}$. Under the assumption that the high-order interactions of $\mathbf{X}$ with the treatment indicator are uncorrelated with any of the control variables in (2.1), their effect, i.e. the heterogeneity of the treatment effect, is captured by the sum of residuals, $\hat{\varepsilon}_{it}$, and intention-to-treat effect.

Thus, a suitable outcome for the random forest is composed of two parts: i) the estimated average intention-to-treat effect; and ii) the residuals of (2.1), which represent variation in hourly wage growth rates purged of a) the time trend between pre- and post-treatment periods; b) the differences between the treatment and control groups prior to intervention; as well as c) those effects that can be attributed to the control variables $\mathbf{Z}_{it}$:

$$\text{forest outcome variable: } \widetilde{\Delta y}_{it} := \hat{\varepsilon}_{it} + \hat{\delta}(WT_{it}). \tag{2.3}$$

A discussion of the main assumption follows. If the omitted interactions are correlated with any of the control covariates in (2.1), the coefficients on these control covariates will capture some of the effect of the interactions; in other words, the interaction effect will then not be captured by the sum of error term and intention-to-treat effect. Therefore, to the extent that the interactions of $\mathbf{X}$ with the treatment indicator are correlated with any of the included independent variables in (2.1), we will get a biased estimate of the amount of heterogeneity. Two important observations follow. First, except for the East/West, marginal employment contract and firm size dummies, the variables we subject to the heterogeneity analysis are different from the control variables (see Section 2.4). Therefore, for the vast majority of variables we do not run into the mechanical problem that the control variables capture some effect of the omitted interaction variables simply by their virtue of being composed of the same control variables. Second, the intention-to-treat effect δ will capture some of the effect of the interactions with the treatment dummy due to them being correlated by definition. For this reason, we include the intention-to-treat effect in the calculation of the forest outcome variable.

Subsequently, we feed $\widetilde{\Delta y}_{it}$ to the causal forest algorithm to understand treatment effect heterogeneity by the groups implicitly defined by covariates $\mathbf{X}$. We regard $\widetilde{\Delta y}_{it}$ 'as if observed' although it is based on estimated quantities. This should lead to an understatement of standard errors for the results that follow. Properly accounting for the uncertainty about $\widetilde{\Delta y}_{it}$ in the causal forest step needs to be addressed in future work. There is a countervailing effect at play also. The treatment effect δ in eq. (2.1) is estimated on a sample that includes all eligible workers. The fact that this is a heterogeneous group is the motivation of the study at hand. A direct consequence of such heterogeneity is that δ will be estimated imprecisely relative to estimates that are based on more homogeneous samples that include only individuals enjoying a large treatment effect. Since this is effectively what the causal forest methodology does, the standard errors of group-specific treatment effects tend to be lower than the standard error for δ in eq. (2.1).

Athey and Imbens (2016) first proposed to use regression trees to study heterogeneity in randomized controlled trials (RCTs). Athey et al. (2019) recast this initial idea in the framework of generalized random forests (GRFs), the backbone of our analysis. The switch from 'trees' to 'forests' is not merely semantic. A forest consists of many trees; each characterized a different order and subset of the variables based on which the successive splits are made. Consequently, the group each given individual is allocated to will not be the same across all trees. The calculation of consistent variance estimates is based on the variability of these tree-specific CITE estimates, which is similar to well-known bootstrap procedures to estimate the variance of a given statistical quantity. The ability to estimate consistent estimates for the variance is a main contribution of Athey et al. (2019). We provide a short introduction to the machinery of causal forests in Appendix 2.6.1.

We begin by making some general remarks about the moment-based formulation of the GRF methodology before describing how we are using it in the case at hand. The flexible GRF method can estimate any quantity of interest, $\tau(x)$, identified by the local moment condition,

$$\mathbb{E}[\psi_{\tau(x),\nu(x)}(O_i)|X_i = x] = 0. \tag{2.4}$$

ψ denotes some scoring function, $\nu(x)$ an optional nuisance parameter, O_i, X_i are both observed data, O_i being relevant to estimate τ, and X_i contains auxiliary variables. In a standard regression problem O_i is the outcome variable Y_i. In treatment effect estimation O_i contains both the outcome variable and the treatment indicator. One way to approach such an estimation is to define some similarity weights ω_i, which must be positive and sum to one, measuring the importance of observation i to estimate ψ at x,

$$\Big(\hat{\tau}(x), \hat{\nu}(x)\Big) = \text{argmin}_{\tau,\nu} \left\{ \left\| \sum_{i=1}^{n} \omega_i(x) \psi_{\tau,\nu}(O_i) \right\|_2 \right\}. \tag{2.5}$$

In a generalized random forest procedure, the estimates ω_i are defined implicitly by a set of $b = 1, \ldots, B$ subsampled trees: intuitively, the more often i ends up in the same final leaf as the observation defined by x, the more important it is to estimate $\Big(\hat{\tau}(x), \hat{\nu}(x)\Big)$. In other words, GRF estimates ω_i by optimizing the moment conditions given in (2.4).

Therefore, reformulating the model of interest in this paper in the form of a moment condition as (2.4) opens the door to using GRF to estimate heterogeneous treatment effects. We follow Athey et al. (2019) and posit the random coefficients model,

$$\widetilde{\Delta y}_{it} = \alpha_i(x_i) + \tau_i(x_i) WT_{it} + u_i, \tag{2.6}$$

with $\tau(x) = \mathbb{E}[\tau_i | X = x_i]$. Under the assumption that WT_{it} is independent of unobservables conditionally on X_i, $\{\tau_i, u_i\} \perp\!\!\!\perp WT_{it} | X_i$, $\tau(x)$ identifies the conditional intention-to-treat effect (CITE). $\tau(x)$ can be estimated via the generalized random forest methodology by defining,

$$\psi_{\tau(x),\alpha(x)}(\widetilde{\Delta y}_{it}, WT_{it}) := \big(\widetilde{\Delta y}_{it} - \tau_i(x_i) WT_{it} - \alpha_i(x_i)\big) \begin{pmatrix} 1 \\ WT_{it}, \end{pmatrix} \tag{2.7}$$

which results in two moment conditions, one for the intercept and another for the random coefficient of interest, i.e. the conditional intention-to-treat effect $\tau_i(x_i)$. To corroborate that the conditional independence assumption is reasonable in the case at hand, it is important to keep in mind that the set of control variables in the base model (2.1) is extensive and, therefore, can be credibly believed to have rendered $\hat{\varepsilon}_{it}$ independent of WT_{it}. Since $\hat{\varepsilon}_{it}$ is the essential component of $\widetilde{\Delta y}_{it}$, this, in turn, lends credibility to the conditional independence assumption underlying the identification of $\tau_i(x_i)$ in (2.6). We employ the GRF methodology of Athey et al. (2019) since it also provides variance estimates corresponding to each final CITE estimate, which we will rely on in the next subsection to test for statistical significance of the estimated CITEs. The identifying assumption in the DTADD approach is that existing differences in wage growth dynamics between treated and control group remain constant absent the treatment. Equally strong business cycles over the years 2012-2014 and 2014-2016 indicate that such differences are likely to have remained constant.[4] Another threat to this identification are spillover effects, which cannot be discerned in the case at hand (see above). In the estimation of heterogeneous treatment effects, we assume that assumptions of equally strong business cycles or absence of spillover effects (see above) hold within each subgroup identified by the causal forest. Moreover, the fact that our results are robust to the inclusion of high-order

[4] B20 point out that GDP growth amounted to 5.8% in 2012-2014 and 5.7% in 2014-2016.

interactions of X in eq. (2.1) provides evidence that differences in wage growth dynamics in groups defined by such interactions have been sufficiently controlled for, see Section 2.4.4.

It is worth noting that the conditioning set X_i in the conditional independence assumption only includes contemporaneous realizations of the covariates. Therefore, issues that were to arise when conditioning on post-treatment variables can be neglected.

Other approaches to evaluate treatment effect heterogeneity include k-nearest neighbour or kernel estimation procedures (Crump et al., 2008). Such methods work well with a small number of covariates, but performance is reduced when number of covariates increases or interactions among covariates becomes important (Wager and Athey, 2015). Since the objective of this study is to find heterogeneous effects among groups characterized by a combination of covariates we can leverage the comparative strengths of causal trees. A study on heterogeneous employment effects by Wang et al. (2019) is close in spirit to our approach since it uses a C-lasso technique, another data-driven approach, to reveal the number of groups with different effects.

2.3.3 Empirical illustration

In this subsection, we illustrate how the causal forest methodology splits the intention-to-treat effect into ever finer-grained subgroups. For the purpose of illustration, we focus on a small subset of three control variables: a dummy for marginal employment (x_1), firm size (x_2), and skill degree (x_3). We proceed iteratively by first estimating the causal forest as described in Section 2.3.2 with $\mathbf{x} = (x_1)$, then with $\mathbf{x} = (x_1, x_2)$, and finally with $\mathbf{x} = (x_1, x_2, x_3)$. Thus, through these consecutive steps, we distinguish between, first, those employees in marginal and those in regular employment (full- or part-time); second, between those in marginal employment at a large firm, those in marginal employment at a small firm, those in regular employment at a large firm, and those regularly employed at a small firm; and third, between those in marginal employment at a large firm with a low skill degree, those in marginal employment at a large firm with a high skill degree, etc.

Figure 2.2 illustrates the iterative process by means of a Sankey diagram. Moving down to the bottom of the figure, which represents the estimate of the intention-to-treat effect of 6.5 percentage points, we illustrate the successive splits along x_1, x_2, x_3. The horizontal position of each node (i.e. group defined by the corresponding covariates) represents the level of the CITE, more right-ward nodes representing larger CITEs. The width of each edge is proportional to the number of observations of the sending node that are assigned to the receiving node. The color of each edge represents the level of the CITE in the receiving node. Therefore, the more diverse the color spectrum leaving a given node, the

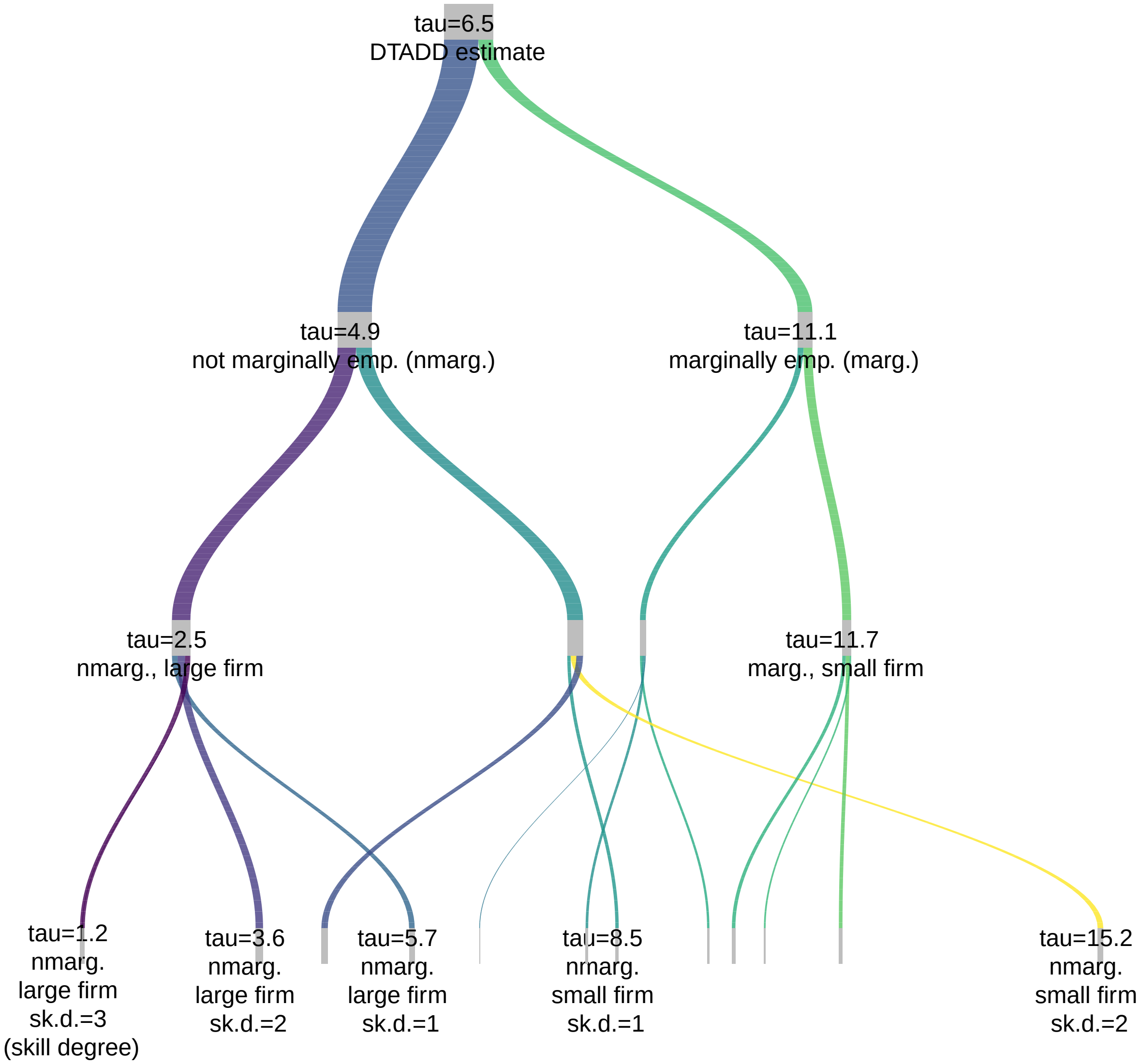


Figure 2.2: Structure of a Causal Tree. *This Figure illustrates how the causal forest splits the intention-to-treat effect (on the top of the graph) into treatment effects specific to ever more fine-grained groups. The edges' color is calibrated based on treatment effect levels in the receiving layer (from yellow = high to violet = low). Therefore, the diversity of colors leaving a particular node indicates how heterogeneous the group actually is. Since this graph serves to illustrate what type of interactions the causal forest methodology helps uncover, we present labels only for selected groups. See main text for additional information.*

more heterogeneous that node is. In order to keep the illustration as simple, we only label nodes relevant for the ensuing discussion.

The topmost layer of the Sankey diagram provides the estimated intention-to-treat effect of 6.5 percentage points. On the first layor, the causal forest splits treated individuals in regular employed (flow to the left, i.e. smaller CITE) and marginally employed (flow to the right, i.e. larger CITE). The second layer splits along the firm size dummy. Being employed in a small firm is associated with a larger CITE. Interestingly, the positive effect of being marginally employed is almost completely off-set by the negative effect of being employed at a large firm. Vice versa, the negative effect of being regularly employed is off-set by being employed at a small firm. As a result, these two groups (marginally employed at a large firm, and regularly employed at a small firm) have almost indistinguishable CITEs. The third layer splits along skill degree.[5] Low skill levels tend to be associated with larger CITEs. This is seen, for instance, in the nodes emanating from the regularly employed at a large firm (leftmost node on the third layer); the resulting nodes on the fourth, bottom layer represent individuals with skill degree of three, two, and one respectively moving from low (left) to high CITE (right). We can also observe an unexpected break in this pattern in the nodes emanating from the node representing the regularly-employed at a large firm on the third layer. When this subgroup is further split by skill degree, it is not the lowest skill degree individuals who are associated with the largest CITE but those with an intermediate skill degree level, which move all the way to the right at an estimated CITE of 15.4. The subgroup with the lowest skill level is associated with an estimated CITE of merely 8.5. In other words, although low skill degree is associated with the largest CITEs among individuals employed regularly at a large firm, this is not the case for individuals employed regularly at small firms. It is these complex interactions that the causal forest methodology enables us to detect.

A natural objection to these types of calculation is that the number of individuals represented in each group decreases substantially down the tree. Therefore, it is important to keep in mind that the GRF methodology produces consistent variance estimates for each CITE estimate that capture the statistical uncertainty due to low numbers of observations. With this note, we conclude the illustrative example and move to the presentation of the heterogeneity results when using the whole list of covariates $\mathbf{X}$.

2.4 Results

Guided by previous labor-market research, we consider the following covariate set, $\mathbf{X}$, for constructing subgroups:

[5] It seems as if there are only three edges emanating at the level of the third split, which seems to be at odds with the skill degree variable having five levels. Do note that the higher skill levels are relatively rare (especially among the marginally employed) and some of the resulting narrow flows happen to lie just below wider edges and, therefore, are not visible.

- Firm size: small vs. non-small firms
- Educational attainment according to the 8-level ISCED code

  1 Primary education

  2 Lower secondary education

  3 Upper secondary education

  4 Post-secondary non-tertiary education

  5 Short-cycle tertiary education

  6 Bachelor's or equivalent level

  7 Master's or equivalent level

  8 Doctoral or equivalent level
- Residence: East and West Germany
- Skill degree of occupation:

  1 Untrained blue-collar worker / untrained white-collar worker

  2 Semi-trained blue-collar worker / trained white-collar

  3 Trained blue-collar worker / qualified professional

  4 Foreman / highly-qualified professional

  5 Master Craftsman / managerial position
- Blue- or white-collar worker
- Degree of autonomy in job ([0] none to [5] high autonomy)
- Marginal vs. non-marginal (regular) employment.

These covariates define 4,320 potential subgroups. The number of groups actually represented in the data is 248. First, we give an overview of the results (Section 2.4.1), then we go into detail as to how our results compare to a traditional heterogeneity analysis (Section 2.4.2), relate them to pre-reform gaps to the minimum wage prior to the reform (Section 2.4.3), and describe the robustness check we implement (Section 2.4.4).

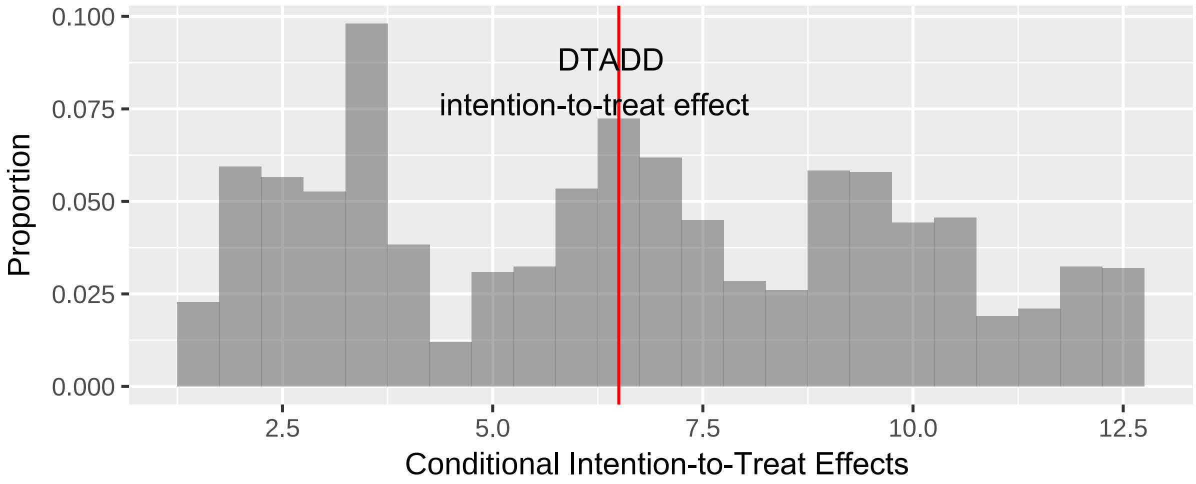


Figure 2.3: Histogram of CITEs. *This figure shows a histogram of the estimated conditional intention-to-treat effects. The red vertical line indicates the level of the intention-to-treat effect from the DTADD model.*

2.4.1 Treatment effect heterogeneity: An overview

Table 2.4 reproduces the results from the DTADD model proposed in B20 with an intention-to-treat effect of 6.5 percentage points which we now set out to decompose into subgroup-specific effects. The causal forest implements 208 splits that lead to sufficiently different CITEs in the resulting subgroups. The arithmetic mean of these 208 CITE estimates weighted by the number of observations in each group amounts to an intention-to-treat effect of 6.6 percentage points, which is close to the average DTADD effect. Note that the common trend assumption, which underlies the identification strategy in the DTADD model (as discussed in B20), is more likely to hold in the various subgroups under study here since these are more homogeneous, i.e. more likely to show similar wage growth dynamics, than the whole sample.

Table 2.5 provides the treatment effects for the 16 groups with at least 10 observations.[6] Groups are provided in decreasing order of estimated effect sizes. The treatment effects of the 16 groups vary considerably – from 1.2 percentage points to 12.7 percentage points – suggesting remarkable heterogeneity hidden behind the intention-to-treat effect. Employees with the following set of characteristics experience the largest CITE of 12.7 percentage points: trained white-collar marginally-employed workers in a small firm in West Germany with short-cycle tertiary education undertaking semi-autonomous work. On the lower end of the spectrum, with a CITE of 1.2 percentage points, you find trained

[6] Table 2.8 in the Appendix contains the treatment effect estimates for groups with at least 3 observations.

Table 2.4: *Regression results for the DTADD base model.*

	DTADD
Treatment indicator	13.002***
	(1.966)
Causal effect	6.493**
	(2.720)
Placebo effect	1.591
	(2.762)
Observations	2,874
R^2	0.105
Adjusted R^2	0.096
Residual Std. Error	29.607 (df = 2844)
F Statistic	11.507*** (df = 29; 2844)

*p<0.1; **p<0.05; ***p<0.01

blue-collar workers employed regularly in a large firm in West Germany with short-cycle tertiary education undertaking semi-autonomous work.

Table 2.5: *This table shows the conditional intention-to-treat effects ($\hat{\tau}(x)$) for all terminal leaves the causal forest, subject to the restriction that the group contains at least 10 observations. Underlying model for Step 1 of our methodology is the basic DTADD model, eq.* (2.1).

running no.	N	$\hat{\tau}(x)$	s.e.	Marginally employed dummy	Small Firm Size dummy	Type of work	Degree of autonomy	ISCED	East dummy	White collar job dummy
1	11	12.7	1.01	1	1	2	3	5	0	1
2	15	12.2	0.88	0	1	2	4	5	1	0
3	14	10.9	0.89	1	1	2	4	5	0	0
4	12	10.8	1.01	1	1	1	4	5	0	0
5	11	10.0	1.20	1	1	3	5	5	0	0
6	13	9.9	0.87	0	1	3	4	5	1	1
7	10	9.4	1.06	1	1	1	4	4	0	0
8	12	7.1	0.98	0	0	2	3	4	0	1
9	10	6.6	1.07	0	0	2	3	5	1	1

Table 2.5: *This table shows the conditional average treatment effec (continued)*

running no.	N	$\hat{\tau}(x)$	s.e.	Marginally employed dummy	Small Firm Size dummy	Type of work	Degree of autonomy	ISCED	East dummy	White collar job dummy
10	12	6.5	1.11	0	1	3	5	5	1	0
11	16	6.5	0.96	0	0	2	4	5	1	0
12	10	6.0	1.15	0	1	2	4	5	0	0
13	14	3.4	0.74	0	0	3	4	5	1	1
14	15	2.3	1.08	0	1	3	5	5	0	0
15	16	2.2	0.70	0	0	1	4	5	0	0
16	12	1.2	0.95	0	0	3	5	5	0	0

Note: Own calculations, based on SOEP v33 2010-2016.

In the absence of any interaction effects, one would expect individuals who share a given characteristic to cluster around a specific CITE and not to spread across the whole spectrum of CITEs. For example, if white-collar work were to have a purely uniform positive effect (i.e. without any interaction effects) on the CITE, then all groups of individuals in white-collar work would cluster at the upper end of the CITE distribution (i.e. in the upper lines of Table 2.5). This, however, is not the case. On the contrary, the CITE for white-collar workers ranges from 12.7 percentage points (line 1) to 3.5 percentage points (line 13) depending on other characteristics. Figure 2.3 shows a histogram of the estimated CITEs. Such a wide range is suggestive that interactions between the covariates are important determinants for the size of the CITE.

An important question pertains to the statistical significance of the observed treatment heterogeneity. To investigate this matter, we conduct pairwise t-tests for mean equality (with the conservative Bonferroni correction). The results of these tests are in Table 2.6, which shows that the pairwise test of the bottom four groups with the top six groups are all statistically significant. We discuss these comparisons in detail now.

Table 2.6: *This table corresponds to the results in Table 2.5. It shows results for t-tests for mean differences in the treatment effects of the groups specified in column and row. A '1' indicates that the null hypothesis of mean equality is rejected at Bonferroni-corrected level of 0.05. Underlying model for Step 1 of our methodology is the basic DTADD model, eq.* (2.1).

White collar job dummy (WC)							1	0	0	0	0	1	0	1	1	0	0	0	1	0	0	0
East dummy (E)							0.	1.	0.	0.	0.	1.	0.	0.	1.	1.	1.	0.	1.	0.	0.	0.
ISCED (I)							5.	5.	5.	5.	5.	5.	4.	4.	5.	5.	5.	5.	5.	5.	5.	5.
Degree of autonomy (DA)							3.	4.	4.	4.	5.	4.	4.	3.	3.	5.	4.	4.	4.	5.	4.	5.
Type of work (TW)							2.	2.	2.	1.	3.	3.	1.	2.	2.	3.	2.	2.	3.	3.	1.	3.
Small Firm Size dummy (SF)							1.	1.	1.	1.	1.	1.	1.	0.	0.	1.	0.	1.	0.	1.	0.	0.
Marginally employed dummy (ME)							1.	0.	1.	1.	1.	0.	1.	0.	0.	0.	0.	0.	0.	0.	0.	0.
ME	SF	TW	DA	I	E	WC																
1.	1.	2.	3.	5.	0.	1	0															
0.	1.	2.	4.	5.	1.	0	0	0														
1.	1.	2.	4.	5.	0.	0	0	0	0													
1.	1.	1.	4.	5.	0.	0	0	0	0	0												
1.	1.	3.	5.	5.	0.	0	0	0	0	0	0											
0.	1.	3.	4.	5.	1.	1	0	0	0	0	0	0										
1.	1.	1.	4.	4.	0.	0	0	0	0	0	0	0	0									
0.	0.	2.	3.	4.	0.	1	0	0	0	0	0	0	0	0								
0.	0.	2.	3.	5.	1.	1	0	0	0	0	0	0	0	0	0							
0.	1.	3.	5.	5.	1.	0	0	0	0	0	0	0	0	0	0	0						
0.	0.	2.	4.	5.	1.	0	1	1	0	0	0	0	0	0	0	0	0					
0.	1.	2.	4.	5.	0.	0	1	0	0	0	0	0	0	0	0	0	0	0				
0.	0.	3.	4.	5.	1.	1	1	1	1	1	1	1	1	0	0	0	0	0	0			
0.	1.	3.	5.	5.	0.	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0		
0.	0.	1.	4.	5.	0.	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	
0.	0.	3.	5.	5.	0.	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0

Note:
ME: Marginally employed dummy, SF: Small Firm Size dummy, TW: Type of work, DA: Degree of autonomy, I: ISCED, E: East dummy, WC: White collar job dummy; Own calculations, based on SOEP v33 2012-2016.

2.4.2 Unfolding uni-dimensional effect heterogeneities

B20 estimate equation (2.1) for the full sample and by two subgroups. They find larger treatment effect for marginally employed (15.5 percentage points, $p < 0.05$ for the null hypothesis of a zero effect), relative to full-time employed (7.8 percentage points, $p < 0.05$); and East Germans (8.1 percentage points, $p < 0.05$) relative to West Germans (4.9 percentage points, $p > 0.1$). Subsequently, we pay specific attention to these comparisons and check to what extent these observed differences are robust to adding interactions between the marginal employment and the East Germany dummy, respectively. For instance, although it seems that there is a large gap in treatment effect between workers in East and West Germany, those employed in either the East or the West with a certain skill degree might actually have very similar treatment effects.

We find confirming evidence for the previously reported larger treatment effects for marginally employed. Depending on other covariates, the treatment effect for marginally-employed varies between 9.4 and 12.7 percentage points. The CITE for regularly-employed typically lies below 7.1 percentage points. Therefore, marginal employment is associated with higher treatment effects regardless of the value of other covariates. The larger effect for marginally employed reported by B20 does not hide important interaction effects.

We find little evidence for the existence of a particularly strong treatment effect for employees resident in East Germany. Instead, we document large CITEs for certain groups of employees in both East and West. For instance, among the regularly employed qualified professionals (this title is determined jointly by the white-collar and skill degree dummies) with a relatively high educational level and degree of autonomy who are living in the East, those working at a small firm have a CITE of 9.9 percentage points, whereas those working at a large firm only of 3.4 percentage points (lines 6 and 13 respectively). The location of residence, thus, interacts with, e.g., firm size in a way that produces differences in the CITE of roughly six percentage points. This observation is part of a more general pattern: employees working in small firms do not generally enjoy a larger CITE since those working in small firms *with a regular position* have a below-average CITE. These observations show that the intention-to-treat effect estimated with traditional methods is hiding complex interaction effects that, once accounted for, reveal a large spectrum of conditional intention-to-treat effects.

This is important information to a policy-maker who might be interested to complement the minimum wage with other measures for those groups of workers who have benefited least from the introduction. E.g. active labour market policies might then be made available contingent on individual characteristics that identify those groups at the lower end of the CITE spectrum such as those individuals working in small firms with a regular position.

Moreover, a policy-maker interested in increasing the compliance with the minimum wage law where it shows the lowest effects might mistakenly focus their resources on West Germany if basing their decisions on a uni-dimensional heterogeneity analysis. If such a policy-maker were to rely on the heterogeneity analysis presented here, they could target non-compliance investigations more precisely, thereby maximizing impact and minimizing resources to monitor non-compliance.

2.4.3 Contrasting effect heterogeneities with pre-reform wage gaps

Ex ante, one would expect that, with the introduction of the minimum wage, the treatment effect of those employees would be largest whose wages were previously furthest below the threshold. Figure 2.4 shows a scatter plot of group-specific CITEs and group-specific gaps to the minimum wage level (defined as the negative relative distance to EUR 8.50) prior to reform. The figure conveys several interesting results: First, the pre-reform wage gap is only a few percentage points for some groups, for others almost 50 percent. Second, for groups with relatively small gaps, the effects triggered by the minimum wage frequently exceed the increase that would have been necessary to raise them above the wage threshold (even when disregarding the regular wage growth that takes place independently of the minimum wage introduction).

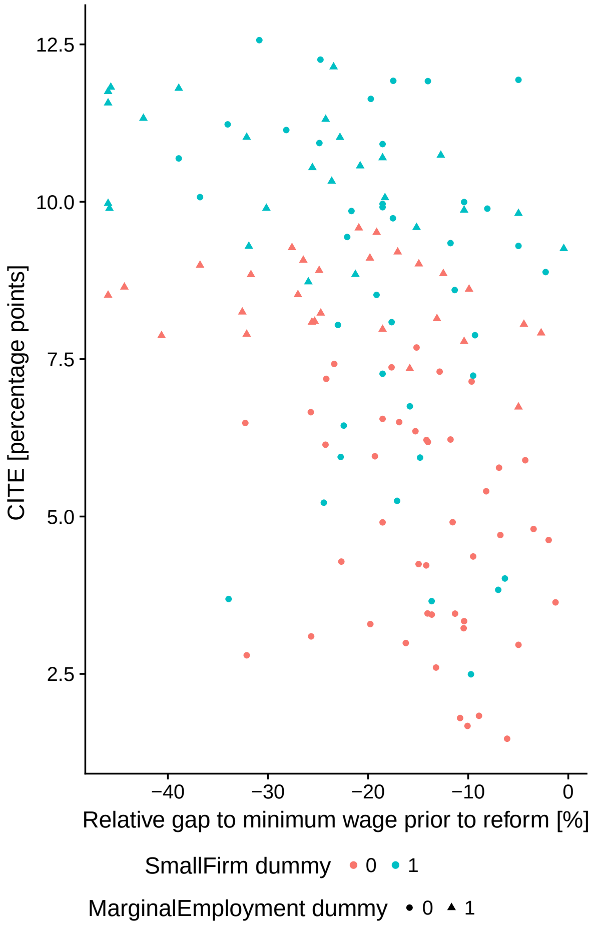


Figure 2.4: CITEs as a function of gap to minimum wage. *This Figure shows CITE as a function of the relative gap to the minimum wage prior to reform (i.e. '-10' translates to 10% below minimum wage prior to reform). A large gap is generally associated with a large CITE. The group with both the largest CITE and largest gap is marginally-employed at small firms (denoted with green triangles).*

The opposite result is observed for those groups where the gap was particularly large, say less than -40%. In fact, the effects for these groups vary between about 8 and 12 percentage points, far below the level needed to push these groups above the threshold (again, in the absence of regular wage growth). This is an indication that there are problems in enforcing the minimum wage, especially for those with particularly low wages before the reform. On the upside, one can discern a generally negative correlation between the wage gap and the level of CITE: those groups that are furthest away from the threshold prior to reform experience the largest CITEs. Such information can also help the policy-maker envisioned in the previous paragraph to direct resources to monitor non-compliance to those subgroups who are furthest away from the minimum wage prior to reform yet do not hava a large CITE.

Note that it is possible to include the distance to the minimum wage prior to reform in $\mathbf{X}$ and check directly for heterogeneous effects w.r.t. that distance. We opt not do that in the study at hand to keep comparability with previously reported results in B20. However, including that distance directly is certainly worthwhile analyzing in future research since it would also address the issue of varying treatment intensities for workers at varying distances from the minimum wage prior to the reform.

2.4.4 Robustness check

In the main body, the forest outcome variable is composed of an average and an individual-level treatment effect derived from the OLS regression (DTADD model in eq. (2.1)) (see Section 2.3.2). We implement the following procedure to check the robustness of our findings. We follow the DTADD specification, as detailed above, but include all possible four-way interactions of the variables in $\mathbf{X}$ as additional explanatory variables. This precludes the objection that instead of implicitly capturing heterogeneous treatment effects, the ε_{it} in eq. (2.3) in fact measures differences in average outcomes regardless of treatment status. The results remain constant upon introducing the four-way interactions in the DTADD model.

2.5 Conclusion

Analyzing the heterogeneity of treatment effects in both observational or randomized controlled studies is difficult as the number of potential subgroups increases, since either one quickly runs into multiple testing problems or is liable to the criticism of hand-picking groups that *ex post* show differences in treatment effects. The method proposed here, an adaptation of Athey et al. (2019) applied to a difference-in-differences setting, distills heterogeneity in a data-driven manner, thus obviating these problems.

Applying the proposed method to heterogeneities in treatment effects induced by the introduction of the statutory minimum wage in Germany in 2015 reveals that the structure of heterogeneity is determined by complex interactions of covariates. An *ex ante* specification of subgroups precludes controlling for complex interactions due to the multitude of possible subgroups and consequent multiple testing problems. Using a data-driven approach enables the researcher to detect spurious heterogeneities that vanish as soon as complex interactions are controlled for. For instance, residing in East Germany interacts with other employer-employee characteristics such that deducing a larger treatment effect of the minimum wage introduction in the East, though in a narrow sense true, is a misleading representation of reality.

Obtaining such fine-grained estimates of treatment effects for a given reform are relevant to the policy-maker for a number of reasons. First, to understand the efficacy of a policy it is important to assess if those groups that were intended to benefit most actually benefit most. Secondly, and building on the first reason, the results indicate which groups of workers, respectively their employers, should be subject to stricter control mechanisms to increase compliance.

2.6 Appendix

2.6.1 From regression trees to causal forests

As a courtesy to the reader, we provide a short introduction to the causal forests. We start by reviewing regression trees before moving to their adaptation to estimate causal effects. This subsection does not contain new results.

Regression trees

Following Athey and Imbens (2016), denote with $\Pi \in \mathbb{P}$ a partitioning of the covariate space, with $\#(\Pi)$ the number of elements in the partition each of which is called $\ell \in \Pi$ and $\pi : \mathbb{S} \to \mathbb{P}$ a function that maps a sample $\mathcal{S} \in \mathbb{S}$ to a partition $\Pi \in \mathbb{P}$. In the upper panel of Figure 1 we illustrate a partition of a two-dimensional covariate space: $\Pi = \{X_1 \leq t_1 \wedge X_2 \leq t_2\}, \{X_1 \leq t_1 \wedge X_2 > t_2\}, \{X_1 \leq t_3\}, \{X_1 > t_3 \wedge X_2 \leq t_4\}, \{X_1 > t_3 \wedge X_2 > t_4\}$.

A basic tree algorithm estimates the individual outcome variable Y_i as the mean Y of observations that are similar with respect to their covariates. Its objective is to maximize the negative expectation of the mean squared error (MSE_μ):

$$MSE_\mu(\mathcal{S}^{te}, \mathcal{S}^{tr}, \pi(\mathcal{S}^{tr})) := \frac{1}{\#(\mathcal{S}^{te})} \sum_{i \in \mathcal{S}^{te}} \Big((Y_i - \hat{\mu}(X_i, \mathcal{S}^{tr}, \pi(\mathcal{S}^{tr})))^2 - Y_i^2 \Big) \tag{2.8}$$

and the estimate for the conditional mean in each leaf is given by

$$\hat{\mu}(x, \mathcal{S}, \Pi) := \frac{1}{\#(i \in \mathcal{S} : X_i \in \ell(x; \Pi))} \sum_{i \in \mathcal{S} : X_i \in \ell(x;\Pi)} Y_i. \tag{2.9}$$

Note that (1) can be written as

$$\begin{aligned}
&MSE_\mu(\mathcal{S}^{te}, \mathcal{S}^{tr}, \pi(\mathcal{S}^{tr})) \\
&\quad = \frac{1}{\#(\mathcal{S}^{te})} \sum_{i \in \mathcal{S}^{te}} \Big(\hat{\mu}^2(X_i, \mathcal{S}^{tr}, \pi(\mathcal{S}^{tr})) - 2Y_i \hat{\mu}(X_i, \mathcal{S}^{tr}, \pi(\mathcal{S}^{tr})) \Big) \\
&\quad = \frac{1}{\#(\mathcal{S}^{te})} \sum_{i \in \mathcal{S}^{te}} \Big(\hat{\mu}^2(X_i, \mathcal{S}^{tr}, \pi(\mathcal{S}^{tr})) \Big) \\
&\qquad - \frac{2}{\#(\mathcal{S}^{te})} \sum_{i \in \mathcal{S}^{te}} \Big(\hat{\mu}(X_i, \mathcal{S}^{te}, \pi(\mathcal{S}^{tr})) \hat{\mu}(X_i, \mathcal{S}^{tr}, \pi(\mathcal{S}^{tr})) \Big)
\end{aligned} \tag{2.10}$$

since $\hat{\mu}$ is the same for all individuals within one leaf.

In order to avoid overfitting, the overall sample $\mathcal{S}$ is split into a training sample $\mathcal{S}^{tr}$, which is used to estimate the tree and leaf means, a cross-validation sample $\mathcal{S}^{cv}$ which is used to choose a complexity penalty, and a test sample $\mathcal{S}^{te}$, which is used to evaluate out-of-sample performance. The CART algorithm is implemented in a two-step process: first, it recursively partitions the covariate space of the training sample by maximizing $-MSE_{\mu}(\mathcal{S}^{tr}, \mathcal{S}^{tr}, \Pi)$ for each splitting decision until a minimum number of observations in each leaf or a specified tree depth is reached. That means that the algorithm has decided that by first splitting the covariate space into $\{X_1 \leq t_1\}$ and $\{X_1 > t_1\}$, it achieves the lowest MSE at that level. Second, to avoid overfitting it then chooses a complexity penalty parameter based on the cross-validation sample (Breiman et al., 1984). We do not go further into detail here and rather focus on the adjustments made by Athey and Imbens (2016) to estimate heterogeneity of treatment effects.[7]

Causal trees

The central modification of Athey and Imbens (2016) is to replace the leaf means by treatment effect estimates. Thereby they leverage the power of the underlying algorithm to estimate heterogeneous treatment effects.

We follow the potential outcomes model or Rubin causal model (Rubin, 1974) and postulate two potential outcomes for each individual i: $(Y_i(W_i = 0), Y_i(W_i = 1))$ where $W_i \in \{0, 1\}$ is a binary treatment indicator. We assume unconfoundedness $W_i \perp\!\!\!\perp (Y_i(0), Y_i(1))|X$. Define the conditional intention-to-treat effect (CITE) as $\tau(X_i) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i]$. The problem is that we can only observe either $Y_i(0)$ or $Y_i(1)$ for any individual. The key to overcoming this problem is re-defining the outcome variable such that it is equal to the CITE in expectation:

$$\begin{aligned} \tau(y, \Pi) &= \mathbb{E}[Y_i(1) - Y_i(0)|X_i \in \ell(x, \Pi)] \\ &= \mathbb{E}[Y_i(1)|X_i \in \ell(x, \Pi)] - \mathbb{E}[Y_i(0)|X_i \in \ell(x, \Pi)] \\ : &= \mu(W = 1, x, \mathcal{S}, \Pi) - \mu(W = 0, x, \mathcal{S}, \Pi) \end{aligned} \tag{2.11}$$

We can now replace these population quantities by sample estimates. The subsample $\mathcal{S}_{W=1}$ refers to the treated observations for which $W_i = 1$ and $\mathcal{S}_{W=0}$ refer to the untreated

[7] Athey and Imbens (2016) propose an ‘honest’ splitting rule in which they use one sample to estimate the splits and another to estimate leaf means. The honest criterion penalizes small leaf size only if it results in a higher within-leaf MSE. This enables a more granular estimation of heterogeneous treatment effects; for illustrative purposes, however, we stick to the conventional splitting rule here. However, in the actual analysis we also implement such an honest splitting rule.

observations for which $W_i = 0$. Redefine the average leaf outcome depending on W as

$$\hat{\mu}(w, x, \mathcal{S}_{W=w}, \Pi) := \frac{1}{\#(i \in \mathcal{S}_{W=w} : X_i \in \ell(x; \Pi))} \sum_{i \in \mathcal{S}_{W=w} : X_i \in \ell(x;\Pi)} Y_i. \tag{2.12}$$

Consequently, we have

$$\hat{\tau}(x, \mathcal{S}, \Pi) := \hat{\mu}(W = 1, x, \mathcal{S}_{W=1}, \Pi) - \hat{\mu}(W = 0, x, \mathcal{S}_{W=0}, \Pi). \tag{2.13}$$

Now we can transform the tree objective from leaf means to leaf treatment effects by replacing the $\hat{\mu}$ in (3) with $\hat{\tau}$ from (6):

$$\begin{aligned} &MSE_{\tau}(\mathcal{S}^{te}, \mathcal{S}^{tr}, \pi(\mathcal{S}^{tr})) \\ &\quad := \frac{1}{\#(\mathcal{S}^{te})} \sum_{i \in \mathcal{S}^{te}} \Big(\hat{\tau}^2(X_i, \mathcal{S}^{tr}, \pi(\mathcal{S}^{tr})) \Big) \\ &\quad - \frac{2}{\#(\mathcal{S}^{te})} \sum_{i \in \mathcal{S}^{te}} \Big(\hat{\tau}(X_i, \mathcal{S}^{te}, \pi(\mathcal{S}^{tr})) \hat{\tau}(X_i, \mathcal{S}^{tr}, \pi(\mathcal{S}^{tr})) \Big) \end{aligned} \tag{2.14}$$

Wager and Athey (2015) provide inferential theory for causal trees, which rests on the re-estimation of a causal tree on a number of random subsamples and averaging their predictions. The result of this multitude of trees resulting from such subsampling is called a random forest, in the causal setting they are termed causal forests. How do they proceed? Random forests combine regression trees with bootstrap aggregation. Let us briefly address each in turn. Regression trees recursively partition the covariate space by maximizing some criterion (e.g. mean squared prediction error) until some stopping criterion is met (Hastie et al., 2009). Trees typically show low bias yet high variance properties. This shortcoming is addressed by exploiting the fact that the variance of the average of B correlated random variables (think, prediction of a regression tree) is given by $\sigma^2_{overall} = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$ meaning one can decrease $\sigma^2_{overall}$ by increasing B while ideally keeping the correlation ρ as small as possible. This is achieved by re-estimating the tree B times but restricting the set of variables considered at each split to be a finite subset of the complete set of covariates in order to decrease ρ.

Athey et al. (2019) have further developed the theory laid out in Wager and Athey (2015), reformulate it in terms of moment conditions and provide an efficient implementation, which we use in our paper.

2.6.2 *t*-tests for the difference in mean treatment effects

The t-test used to investigate whether there are statistically significant differences between the group-specific treatment effects are implemented as follows.

Let $\hat{\tau}_g$ be a group-specific estimate of the treatment effect, n_g the number of observations in that group, σ_g^2 the variance (as estimated by the GRF). Using these ingredients, we calculate a t-statistic based on a Welch-Satterthwaite approximation of the degrees of freedom. The null hypothesis of the test is $H_0 : \tau_j = \tau_g$. The Bonferroni correction is implemented by requiring a p-value of $\frac{0.05}{\#oftests}$ for a test to be declared rejected at the 5% level.

2.6.3 Further descriptive statistics

Table 2.7: *Comparing Descriptive Statistics of Cross-section and Panel sample*

	2012		2014	
	cross-section	panel	cross-section	panel
Contractual hourly wages	17.32	17.90	17.88	18.56
	(8.54)	(8.41)	(9.06)	(8.93)
Size Of Company	5.27	5.34	5.10	5.18
	(1.84)	(1.79)	(1.84)	(1.78)
ISCED	4.01	4.14	4.04	4.16
	(1.73)	(1.70)	(1.77)	(1.68)
East dummy	0.18	0.20	0.19	0.20
	(0.39)	(0.40)	(0.39)	(0.40)
Skill degree of occupation	2.62	2.71	2.59	2.73
	(1.20)	(1.17)	(1.20)	(1.14)
Blue collar worker dummy	0.28	0.25	0.26	0.24
	(0.45)	(0.44)	(0.44)	(0.42)
Degree of autonomy	2.79	2.86	2.79	2.87
	(1.05)	(1.01)	(1.05)	(1.00)
Marginally employed	0.06	0.04	0.07	0.04
	(0.24)	(0.19)	(0.26)	(0.20)
Observations	9899	6133	10216	6475

Note: Averages of respective variables. Standard errors in parentheses. Based on cross-sectional and panel samples. Weighted by cross-sectional weights. *Source:* SOEP v33 2012-2016.

By moving from cross-section to panel samples we lose roughly a third of all observations in a given survey year, see Table 2.1. To show that this does not lead to unrepresentative samples, we compare descriptive statistics for covariates $\mathbf{X}$ and agreed hourly wages in Table 2.7 for 2012 and 2014 and for both cross-sectional and panel samples. One can see that the descriptive statistics change only slightly indicating that a threat to the representativeness of the sample is not warranted. Note that we use cross-sectional weights for both the cross-section as well as panel samples since we do not account for the panel structure of the data in the main specification.

2.6.4 Further results

Table 2.8: *This table shows the conditional intention-to-treat effects ($\hat{\tau}(x)$) for all terminal leaves the causal forest, subject to the the restriction that the group contains at least 3 observations. Underlying model for Step 1 of our methodology is the basic DTADD model, eq.* (2.1).

running no.	N	$\hat{\tau}(x)$	s.e.	Marginally employed dummy	Small Firm Size dummy	Type of work	Degree of autonomy	ISCED	East dummy	White collar job dummy
1	4	12.9	1.28	0	1	2	3	5	1	1
2	11	12.7	1.01	1	1	2	3	5	0	1
3	8	12.2	1.11	0	1	2	3	4	0	1
4	15	12.2	0.88	0	1	2	4	5	1	0
5	7	12.1	1.27	0	1	2	3	5	0	1
6	5	11.6	1.54	1	1	2	3	4	0	1
7	5	11.0	1.45	1	1	1	3	4	0	1
8	6	11.0	1.13	0	1	1	4	5	1	0
9	14	10.9	0.89	1	1	2	4	5	0	0
10	12	10.8	1.01	1	1	1	4	5	0	0
11	3	10.4	1.92	1	1	2	4	6	0	0
12	3	10.3	1.90	0	1	1	3	4	0	1
13	5	10.0	1.51	0	1	1	3	5	0	1
14	11	10.0	1.20	1	1	3	5	5	0	0
15	13	9.9	0.87	0	1	3	4	5	1	1
16	4	9.8	1.66	1	0	1	3	5	0	1
17	4	9.8	1.76	0	1	2	4	8	1	0
18	10	9.4	1.06	1	1	1	4	4	0	0
19	3	9.1	2.18	1	0	2	3	5	0	1
20	4	9.0	1.81	1	0	1	3	4	0	1
21	7	9.0	1.53	1	0	2	4	5	0	0
22	4	8.9	1.97	1	0	3	5	5	0	0
23	3	8.7	2.33	1	1	3	5	8	0	0
24	6	8.5	2.05	0	1	3	4	5	0	1
25	7	8.5	1.57	1	0	2	3	4	0	1
26	3	8.3	2.35	1	0	1	4	6	0	0
27	4	8.1	1.83	1	0	1	4	4	0	0
28	3	8.1	1.99	0	1	2	4	8	0	0
29	3	7.4	1.97	0	0	2	3	4	1	1
30	6	7.2	1.24	0	0	1	3	5	1	1
31	3	7.1	1.96	0	0	2	3	3	0	1

Table 2.8: *This table shows the conditional average treatment effec (continued)*

running no.	N	$\hat{\tau}(x)$	s.e.	Marginally employed dummy	Small Firm Size dummy	Type of work	Degree of autonomy	ISCED	East dummy	White collar job dummy
32	12	7.1	0.98	0	0	2	3	4	0	1
33	8	7.0	1.13	0	0	1	4	5	1	0
34	5	6.9	1.30	0	0	1	3	4	0	1
35	10	6.6	1.07	0	0	2	3	5	1	1
36	12	6.5	1.11	0	1	3	5	5	1	0
37	16	6.5	0.96	0	0	2	4	5	1	0
38	3	6.5	1.57	0	0	1	4	4	1	0
39	10	6.0	1.15	0	1	2	4	5	0	0
40	4	5.9	1.55	0	1	2	4	4	0	0
41	6	5.8	1.16	0	0	0	4	7	0	
42	7	5.8	1.15	0	0	1	3	5	0	1
43	8	5.5	1.18	0	0	2	3	5	0	1
44	3	5.1	1.91	0	0	2	4	6	1	0
45	7	5.0	1.27	0	1	1	4	4	0	0
46	6	5.0	1.48	0	1	1	4	5	0	0
47	4	4.2	1.92	0	0	2	4	6	0	0
48	3	4.0	1.71	0	0	1	4	8	0	0
49	5	3.8	1.54	0	1	3	5	6	1	0
50	9	3.6	1.15	0	1	3	5	8	1	0
51	14	3.4	0.74	0	0	3	4	5	1	1
52	4	3.2	1.27	0	0	1	4	4	0	0
53	5	2.9	1.33	0	0	3	5	5	1	0
54	15	2.3	1.08	0	1	3	5	5	0	0
55	3	2.3	2.41	0	1	4	6	5	0	0
56	3	2.3	2.18	0	0	3	5	6	0	0
57	16	2.2	0.70	0	0	1	4	5	0	0
58	3	2.1	2.33	0	0	3	5	8	0	0
59	8	1.8	1.23	0	0	2	4	5	0	0
60	12	1.2	0.95	0	0	3	5	5	0	0
61	3	1.0	1.61	0	0	3	5	8	1	0

Note:
Own calculations, based on SOEP v33 2012-2016.

2.6.5 Robustness check

As a robustness check, we modify the DTADD specification to include all possible four-way interactions of the variables in **X** as additional explanatory variables. Table 2.9 shows the regression results of these two models and the DTADD specification of the main text.

Table 2.9: *Regression results of DTADD model and DTADD model with interactions*

	DTADD	interaction DTADD
	(1)	(2)
Treatment indicator	13.002***	12.466***
	(1.966)	(1.962)
Causal effect estimate	6.493**	6.952***
	(2.720)	(2.693)
Placebo	1.591	3.004
	(2.762)	(2.742)
Observations	2,874	2,848
R^2	0.105	0.156
Adjusted R^2	0.096	0.118
Residual Std. Error	29.607 (df = 2844)	28.694 (df = 2723)
F Statistic	11.507*** (df = 29; 2844)	4.070*** (df = 124; 2723)

Note: *p<0.1; **p<0.05; ***p<0.01
Own calculations, based on SOEP v33 2012-2016.

DTADD model with interactions of X

Ideally, we want ε_{it} in eq. (2.1) to contain only the sum of average and individual-specific treatment effects. In order to show that ε_{it} is not prohibitively contaminated by level changes in the dependent variable that can be explained purely by covariates $\mathbf{X}$, we add all possible four-way interactions of all variables in $\mathbf{X}$ to the model in eq. (2.1), then calculate the forest outcome variable as before and implement the methodology. The results, analogous to the main text, are reproduced here.

Table 2.10: *This table shows the conditional average treatment effects ($\hat{\tau}(x)$) for all terminal leaves the causal forest, subject to the restriction that the group contains at least 10 observations. Underlying model for Step 1 of our methodology is the basic DTADD model complemented with interactions of X.*

running no.	N	$\hat{\tau}(x)$	s.e.	Marginally employed dummy	Small Firm Size dummy	Type of work	Degree of autonomy	ISCED	East dummy	White collar job dummy
1	11	12.7	0.96	1	1	2	3	5	0	1
2	15	11.4	0.86	0	1	2	4	5	1	0
3	14	11.4	0.89	1	1	2	4	5	0	0
4	12	11.3	1.07	1	1	1	4	5	0	0
5	10	11.1	1.15	1	1	1	4	4	0	0
6	11	11.1	1.15	1	1	3	5	5	0	0
7	13	9.3	0.89	0	1	3	4	5	1	1
8	10	6.7	0.94	0	0	2	3	5	1	1
9	12	6.6	0.93	0	0	2	3	4	0	1
10	12	6.5	0.98	0	1	3	5	5	1	0
11	16	6.2	0.88	0	0	2	4	5	1	0
12	10	5.4	1.13	0	1	2	4	5	0	0
13	14	3.7	0.70	0	0	3	4	5	1	1
14	15	2.6	0.95	0	1	3	5	5	0	0
15	16	2.3	0.71	0	0	1	4	5	0	0
16	12	1.3	1.02	0	0	3	5	5	0	0

Note:

Own calculations, based on SOEP v33 2012-2016.

Discussion

Figure 2.5 shows the CITE estimates for two models (DTADD, as well as DTADD with interactions) in a scatter plot. The x-axis denotes the DTADD estimates, the y-axis describes CITE based on DTADD with interaction. Table 2.9 shows estimation results of the two models. The CITE estimates with the underlying DTADD with interactions model align very closely at the 45-degree line. This precludes the objection that instead of implicitly capturing heterogeneous treatment effects, the ε_{it} in (2.3) in fact measures differences in averages outcomes regardless of treatment status.

Table 2.11: *This table corresponds to the results in Table 2.10. It shows results for t-tests for mean differences in the treatment effects of the groups specified in column and row. A '1' indicates that the null hypothesis of mean equality is rejected at Bonferroni-corrected level of 0.05. Underlying model for Step 1 of our methodology is the basic DTADD model complemented with interactions of* **X**.

White collar job dummy (WC)							1	0	0	0	0	0	1	1	1	0	0	0	1	0	0	0
East dummy (E)							0.	1.	0.	0.	0.	0.	1.	1.	0.	1.	1.	0.	1.	0.	0.	0.
ISCED (I)							5.	5.	5.	5.	4.	5.	5.	5.	4.	5.	5.	5.	5.	5.	5.	5.
Degree of autonomy (DA)							3.	4.	4.	4.	4.	5.	4.	3.	3.	5.	4.	4.	4.	5.	4.	5.
Type of work (TW)							2.	2.	2.	1.	1.	3.	3.	2.	2.	3.	2.	2.	3.	3.	1.	3.
Small Firm Size dummy (SF)							1.	1.	1.	1.	1.	1.	1.	0.	0.	1.	0.	1.	0.	1.	0.	0.
Marginally employed dummy (ME)							1.	0.	1.	1.	1.	1.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
ME	SF	TW	DA	I	E	WC																
1.	1.	2.	3.	5.	0.	1	0															
0.	1.	2.	4.	5.	1.	0	0	0														
1.	1.	2.	4.	5.	0.	0	0	0	0													
1.	1.	1.	4.	5.	0.	0	0	0	0	0												
1.	1.	1.	4.	4.	0.	0	0	0	0	0	0											
1.	1.	3.	5.	5.	0.	0	0	0	0	0	0	0										
0.	1.	3.	4.	5.	1.	1	0	0	0	0	0	0	0									
0.	0.	2.	3.	5.	1.	1	1	0	0	0	0	0	0	0								
0.	0.	2.	3.	4.	0.	1	1	0	0	0	0	0	0	0	0							
0.	1.	3.	5.	5.	1.	0	1	0	0	0	0	0	0	0	0	0						
0.	0.	2.	4.	5.	1.	0	1	1	1	0	0	0	0	0	0	0	0					
0.	1.	2.	4.	5.	0.	0	1	0	0	0	0	0	0	0	0	0	0	0				
0.	0.	3.	4.	5.	1.	1	1	1	1	1	1	1	1	0	0	0	0	0	0			
0.	1.	3.	5.	5.	0.	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0		
0.	0.	1.	4.	5.	0.	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	
0.	0.	3.	5.	5.	0.	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0

Note:
ME: Marginally employed dummy, SF: Small Firm Size dummy, TW: Type of work, DA: Degree of autonomy, I: ISCED, E: East dummy, WC: White collar job dummy; Own calculations, based on SOEP v33 2012-2016.

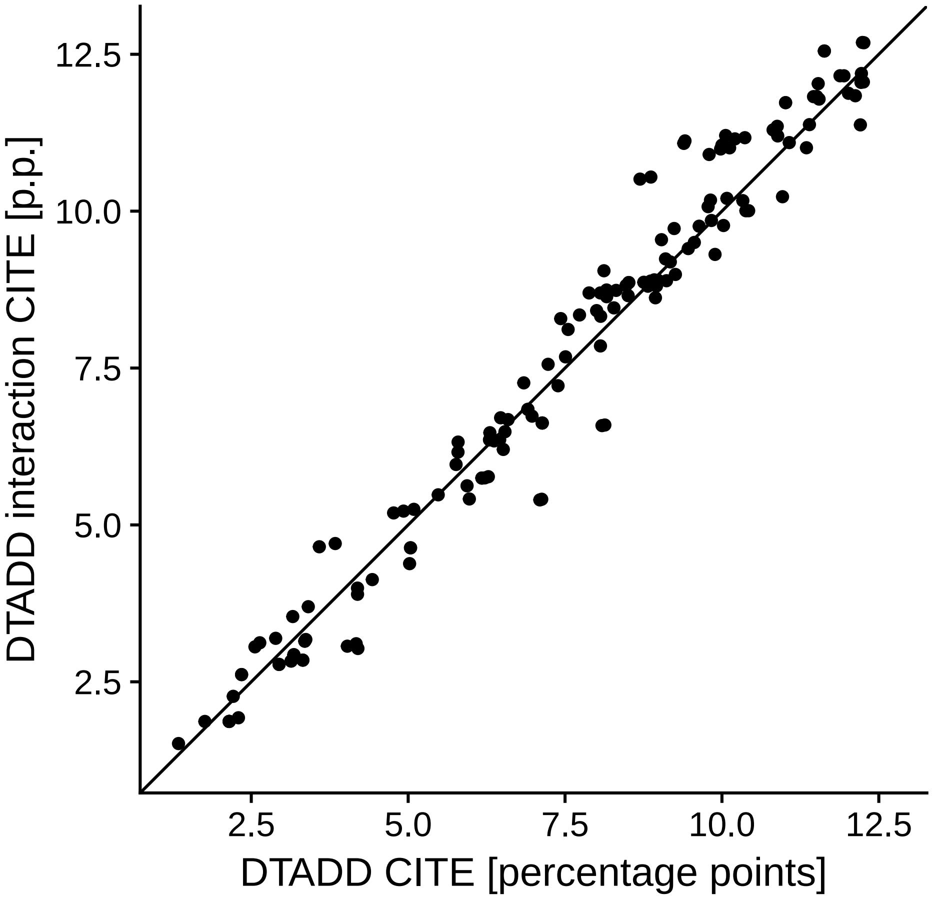


Figure 2.5: DTADD interaction CITE vs. DTADD CITE *This Figure shows CITE estimates based on DTADD with interactions against the corresponding estimates estimated based on the DTADD model.*

2.6.6 Why individual-specific fixed effects cannot be used

We replicate eq. (2.4), that shows our identification strategy, as a starting point of the discussion:

$$\underbrace{[\overbrace{\mathbf{E}(\Delta y^t_{it=2014})}^{G1} - \overbrace{\mathbf{E}(\Delta y^t_{it=2012})}^{G2}]}_{\text{factual}} - \underbrace{[\overbrace{\mathbf{E}(\Delta y^c_{it=2014})}^{G3} - \overbrace{\mathbf{E}(\Delta y^c_{it=2012})}^{G4}]}_{\text{counterfactual}},$$

where superscripts t and c denote treated and control groups respectively.

The employed DTADD approach relies on three differencing steps. First, we take (relative) difference in wages in each group G1, G2, G3, and G4; i.e. we calculate Δy_{it} for each group. Second, we take differences between group-specific changes in wage dynamics in the period before (2012 to 2014) and after the minimum-wage introduction (2014-2016) ($G1-G2$ and $G3-G4$ respectively). Third, we take differences between treatment ('factual') and control group ('counterfactual'). The second difference is the additional step that renders the underlying common trend assumption credible.

The OLS regression model we use to estimate δ takes the form,

$$\Delta y_{it} = \delta(W_{it}\mathbf{1}_{t=2014}) + \delta_0(W_{it}\mathbf{1}_{t=2012}) + \beta_1 W_{it} + \beta_2\mathbf{1}_t + \beta_\mathbf{3}\mathbf{Z_{it}} + \varepsilon_{it}, \tag{2.15}$$

with wage growth, $\Delta y_{it} = \log\left(\frac{y_{it+2}}{y_{it}}\right) \times 100$, as dependent variable, $t \in \{2010, 2012, 2014\}$ denoting the differenced time periods, and ε_{it} an individual error term.

To show that employing a fixed effects regression in this identification strategy leads to bias, let us consider the simplest version of the model, with only two time periods (2012 and 2014) and no control variables. We include fixed effects η_i to illustrate the problem.

$$\Delta y_{it} = \delta(W_{it}\mathbf{1}_{t=2014}) + \beta_1 W_{it} + \beta_2\mathbf{1}_{t=2012} + \beta_3\mathbf{1}_{t=2014} + \eta_i + \varepsilon_{it}$$

First, let us consider the hypothetical, since unrealistic in the application at hand, case where each individual is categorized as either treated (subscript j) or control (subscript k) for both periods. This implies the following four terms.

$$\begin{aligned}
\Delta y^t_{jt=2012} &= \beta_1 W_{jt} + \beta_2\mathbf{1}_{t=2012} + \eta_j + \varepsilon_{jt}\\
\Delta y^t_{jt=2014} &= \delta(W_{jt}\mathbf{1}_{t=2014}) + \beta_1 W_{jt} + \beta_3\mathbf{1}_{t=2014} + \eta_j + \varepsilon_{jt}\\
\Delta y^c_{kt=2012} &= \beta_2\mathbf{1}_{t=2012} + \eta_k + \varepsilon_{kt}\\
\Delta y^c_{kt=2014} &= \beta_3\mathbf{1}_{t=2014} + \eta_k + \varepsilon_{kt}
\end{aligned} \tag{2.16}$$

Then we have the following.

$$\delta = \underbrace{[\overbrace{\mathbf{E}(\Delta y^t_{it=2014})}^{G1} - \overbrace{\mathbf{E}(\Delta y^t_{it=2012})}^{G2}]}_{\text{factual}} - \underbrace{[\overbrace{\mathbf{E}(\Delta y^c_{it=2014})}^{G3} - \overbrace{\mathbf{E}(\Delta y^c_{it=2012})}^{G4}]}_{\text{counterfactual}} \tag{2.17}$$

$$= \Big[[\delta(W_{jt}\mathbf{1}_{t=2014}) + \beta_1 W_{jt} + \beta_3\mathbf{1}_{t=2014} + \eta_j] - [\beta_1 W_{jt} + \beta_2\mathbf{1}_{t=2012} + \eta_j]\Big] - \tag{2.18}$$

$$\Big[[\beta_3\mathbf{1}_{t=2014} + \eta_k] - [\beta_2\mathbf{1}_{t=2012} + \eta_k]\Big] \tag{2.19}$$

$$= \delta \tag{2.20}$$

The treatment effect is identified through δ.

The DTADD approach rests on the assumption that, for both control and treatment groups, the wage growth dynamics between 2014 and 2016 are the same as the wage growth dynamics between 2012 and 2014 absent the reform. If this is so, we can compare the difference in wage growth dynamics in 2012 and 2014 between treatment and control with the same difference in 2014-2016 to estimate the treatment effect. It is possible, even likely given that wage growth is more dynamic at the lower end of the wage distribution, that an individual who is earning below the minimum wage in 2012 (i.e. classified as treated) will earn above the minimum wage in 2014 (i.e. classified as control).

More realistically, let's say we observe individuals in each of the following groups: 1) group j: treatment group in 2012, treatment group in 2014, 2) group k: control group in 2012, control group in 2014, 3) group l: treatment group in 2012, control group in 2014

In other words, it might occur that an individual appears in group G2 for year $t = 2012$ and in group G3 in year $t = 2014$. We now illustrate that this will lead to a biased treatment effect estimate.

To simplify the argument, assume that we observe one individual in each group. This implies the following terms.

$$\Delta y^t_{jt=2012} = \beta_1 W_{jt} + \beta_2 \mathbf{1}_{t=2012} + \eta_j + \varepsilon_{jt} \tag{2.21}$$

$$\Delta y^t_{jt=2014} = \delta(W_{jt}\mathbf{1}_{t=2014}) + \beta_1 W_{jt} + \beta_3 \mathbf{1}_{t=2014} + \eta_j + \varepsilon_{jt} \tag{2.22}$$

$$\Delta y^c_{kt=2012} = \beta_2 \mathbf{1}_{t=2012} + \eta_k + \varepsilon_{kt} \tag{2.23}$$

$$\Delta y^c_{kt=2014} = \beta_3 \mathbf{1}_{t=2014} + \eta_k + \varepsilon_{kt} \tag{2.24}$$

$$\Delta y^t_{lt=2012} = \beta_1 W_{lt} + \beta_2 \mathbf{1}_{t=2012} + \eta_l + \varepsilon_{lt} \tag{2.25}$$

$$\Delta y^c_{lt=2014} = \beta_3 \mathbf{1}_{t=2014} + \eta_l + \varepsilon_{lt} \tag{2.26}$$

Then we have the following.

$$\begin{aligned}
\delta =& \underbrace{[\overbrace{\mathbf{E}(\Delta y^t_{it=2014})}^{G1} - \overbrace{\mathbf{E}(\Delta y^t_{it=2012})]}^{G2}}_{\text{factual}} - \underbrace{[\overbrace{\mathbf{E}(\Delta y^c_{it=2014})}^{G3} - \overbrace{\mathbf{E}(\Delta y^c_{it=2012})]}^{G4}}_{\text{counterfactual}} \\
=& \Big[[\delta(W_{jt}\mathbf{1}_{t=2014}) + \beta_1 W_{jt} + \beta_3 \mathbf{1}_{t=2014} + \eta_j] - [\beta_1 W_{jt} + \beta_2 \mathbf{1}_{t=2012} + \frac{\eta_j + \eta_l}{2}]\Big] \\
&- \Big[[\beta_3 \mathbf{1}_{t=2014} + \frac{\eta_k + \eta_l}{2}] - [\beta_2 \mathbf{1}_{t=2012} + \eta_k]\Big] \\
=& \delta + \eta_j - \frac{\eta_j + \eta_l}{2} - \frac{\eta_k + \eta_l}{2} + \eta_k \\
=& \delta + \frac{\eta_j}{2} - \eta_l + \frac{\eta_k}{2} \\
\neq& \delta
\end{aligned} \tag{2.27}$$

Thus, if we were to include individual-specific fixed effects, part of the difference between the ‘factual’ and ‘counterfactual’ wage dynamics would be soaked up by that individual fixed effect, and would therefore bias our estimate of the treatment effect.

3 A Reverse Causality Test Without Instruments

3.1 Introduction

Endogeneity is a central problem in econometric models which potentially invalidates estimates of causal effects. Existing tests of endogeneity often require that a potential solution in the form of instruments is available. Moreover, even when instruments are available, imposed exclusion restrictions on such instruments are often controversial on their own. We provide a test for one source of endogeneity, namely reverse causality of a single regressor, that does not require instruments. To detect this type of endogeneity we show that it is sufficient to impose a nonlinear model structure. We require the errors to be additively separable but allow for heteroscedasticity w.r.t. additional control variables.

Testability of reverse causality relies on restrictions of the model under consideration. In particular, we show that additively separable errors and nonlinearity of the true underlying functional relationship implies testable restrictions. We build on Hoyer et al. (2009), who establish that the causal direction between two variables is identifiable, while maintaining independence between covariate and the additively separable error term. The main contribution of this paper is to extend the framework of Hoyer et al. (2009) to the heteroscedastic case with additional covariates and clarify the usefulness of this approach in economic applications.

We define a causal model in which Y is generated as a function of covariate X and control variables $\mathbf{W}$ and a anticausal model in which X is generated as a function of covariate Y and control variables $\mathbf{W}$. The testable restriction implies that the independence between the error and covariate can only hold in *either* the causal *or* the anticausal model, not both.

This testable restriction involves the unobserved true errors. In the practical algorithmic implementation, we need to rely on estimates of these true errors, which has implications for the asymptotic distribution of the test statistic. More specifically, the estimated errors show a particular dependence with the covariate even if the true errors are independent of the covariate. This dependence originates from the fact that the residuals are estimated with a model that is itself estimated based on the covariate which we want to test independence with. This requires us to make an additional assumption, namely that either the causal or the anticausal model represents the true data generating process. Given this assumption, we leverage advances on testing conditional independence based on kernel mean embeddings, i.e. maps of probability distributions into reproducing kernel Hilbert spaces (RKHS) (Muandet et al., 2016). We show that our procedure has high accuracy in detecting the true causal direction in simulated data. Furthermore, the provide an empirical application, which show that our test can provide suggestive evidence about the causal link between income and work experience, which we proxy by age.

Related literature The idea that a causal link between two variables implies an independence between the error and the cause variable, which our test ultimately rests on, has precursors in the literature. In particular, Robert Engle et al. (1983) propose a definition of an exogeneous relation in terms of conditional densities that is close in spirit to the test idea in the paper at hand. In particular, they argue that, if a joint probability density of two random variables Y and X factorizes as $f(Y, X) = f(Y|X)f(X)$ and the conditional density $f(Y|X)$ is invariant to changes in the marginal density $f(X)$, then X is called "super exogenous" (p. 278). Statistical tests for the notion of "super exogeneity" are proposed by Favero and Hendry (1992), Engle and Hendry (1993), and Hendry and Santos (2010). These tests specifically rely on analyzing to what extent parameter values are sensitive to interventions in the economy. Such interventions can be either natural or experimental. The approach at hand neither relies on experimental or natural interventions nor on the stability of parameter values. We interpret the invoked invariance to changes in terms of independence between the true error and the covariate and derive testable implications.

This paper is also related to a strand of the literature, which make use of exogenous variations to detect endogeneity of regressors. The idea to make use of instrumental variables to detect endogeneity was originally proposed by Hausman (1978). More recently, Blundell and Horowitz (2007) and Breunig (2015) provide exogeneity tests using instrumental variables for nonparametric models with additively separable errors and Fève et al. (2018) and Breunig (2020) for models with nonseparable errors.

3.2 Reverse causality test

We show how to test for reverse causality between two variables X and Y in the presence of additional covariates $\mathbf{W}$ where $\mathbf{W}$ need not be independent of the regression error. First, we introduce the model, discuss how the model specification relates to the existing causal discovery literature, and derive testable implications. Second, we present the conditional independence test that is a central component of the algorithm. Third, we present the algorithmic implementation.

3.2.1 Model and testability

The problem of identifying causal structure from non-experimental data is receiving considerable attention in the nascent causal machine learning literature (Mooij et al., 2016; Peters et al., 2017). In its bivariate form, the problem is concerned with deciding whether a variable X is causing Y or vice versa solely based on a non-experimental joint probability distribution of the two variables. Without making any assumptions regarding the true underlying data-generating process, no headway is possible.[1] In the following, we discuss the assumptions we make that enable us to identify the causal direction.

Consider the following model where X is causing Y:

$$Y = h(X, \mathbf{W}) + U \tag{3.1}$$

where X, Y and U are scalars and $\mathbf{W}$ is a vector of covariates. This model is called the 'causal model' in the following. We make the following assumptions.

Note that the error U in eq. (3.1) is additively separable. We stress that, with this assumption, we place our paper in that strand of the literature exploring identifiability of the causal direction by restricting the model class under consideration. This strand can broadly be sub-divided into two approaches. First, Shimizu et al. (2006) show that identification of the causal direction is possible in a model with linear h when the error term is non-Gaussian. See Appendix 3.6.3 for a proof of the central idea. We show that this identification result holds in our Monte Carlo simulations, yet we do not extend it formally (see Appendix 3.6.2). Zhang and Hyvärinen (2009) weaken that assumption. They consider a 'post-nonlinear' model of the form $Y = h_2(h_1(X) + \varepsilon)$ and show that the causal direction is identifiable when h_2 is invertible. Thus, nonlinear rescaling of the data e.g. through typical log transformations of income data in economics can be taken into account. Moreover, Mooij et al. (2011) extend the results in Hoyer et al. (2009) to cyclic models under the assumption of Gaussian error terms.

The additive separability of the error term U precludes the dependence of marginal

[1] Previous work shows that the causal direction cannot be identified without making further assumptions. Peters (2012, Proposition 2.6) proves that for every joint distribution of two variables, X and Y, there is a model $Y = h(X, \varepsilon)$, with $X \perp\!\!\!\perp \varepsilon$ with h a measurable function and ε a real-valued noise variable. The roles of X and Y can be easily interchanged showing that the joint distribution itself does not identify the causal direction in this most general form.

effects on unobservables. Therefore, we interpret the error term in a traditional sense as measurement error of the variable of interest.

Assumption 3.2.1 (Nonlinearity)**.** *The function $h(.,\mathbf{w})$ is nonlinear for each $\mathbf{w}$ in the support of $\mathbf{W}$.*

Hoyer et al. (2009) show that nonlinearity of h can play a similar role as regards the identifiability of the causal direction as non-Gaussianity of U, which Shimizu et al. (2006) rely on. They show that, if the true model is of a nonlinear form, one can infer the causal direction without making any assumptions about the distribution of the error. Our work is most closely related to this route of identification, which we complement by considering heteroskedasticity of the error term with respect to additional covariates $\mathbf{W}$.

Assumption 3.2.2 (Heteroskedasticity)**.** *Assume*

$$U = \sigma(\mathbf{W})\,\varepsilon \;\; \textit{with} \;\; \varepsilon \perp\!\!\!\perp (X, \mathbf{W}) \tag{3.2}$$

for some strictly positive function $\sigma(\cdot)$.

Existing causal discovery algorithms do not take into account heteroskedastic error structures. Assumption 3.2.2 explicitly introduces such heteroskedasticity of the error term with respect to the control variables $\mathbf{W}$. Note that Assumption 3.2.2 implies $U \perp\!\!\!\perp X|\mathbf{W}$.

Given Assumptions 3.2.1 and 3.2.2 we now formulate the main theorem of this paper.

Theorem 3.2.1 (Identifiability)**.** *Let Assumptions 3.2.1 and 3.2.2 about causal model* (3.1) *be satisfied. Then there cannot be a anticausal model,*

$$X = \tilde{h}(Y, \mathbf{W}) + \tilde{U}, \tag{3.3}$$

where $\tilde{U} = \tilde{\varepsilon}\,\tilde{\sigma}(\mathbf{W})$ and $\tilde{\varepsilon} \perp\!\!\!\perp (Y, \mathbf{W})$ is fulfilled. The proof of this statement can be found in Appendix 3.6.1.

Theorem 3.2.1 implies that if $U \perp\!\!\!\perp X|\mathbf{W}$ then $\tilde{U} \perp\!\!\!\perp Y|\mathbf{W}$ cannot simultaneously be true. This enables inferring the causal direction from observational data by analyzing to what extent the independence of errors and covariates holds. The nonlinearity of h and the additive separability of the error term, U, give the proposed test power.

In the remainder, we make an additional assumption that we require for the algorithmic implementation, namely we require the existence of causal *or* anticausal model:

Assumption 3.2.3 (Existence)**.** *Assume that the data generating process satisfies the causal model in* (3.1) *or the anticausal model in* (3.3)*.*

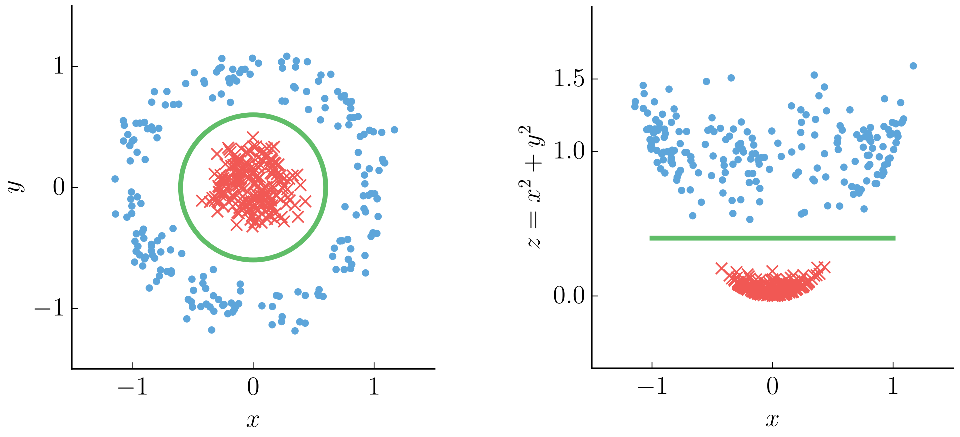
We conclude this section with some further observations about the causal discovery literature. Next to the *a priori* restriction of the model class, which we follow in this paper, there are more proposals to identify causal directionality. First, there is work relying on information-geometric arguments: the essential idea is that the conditional distribution of the effect given its cause does not contain information about the marginal distribution of the cause (Janzing and Schölkopf, 2010). The information content is formalized using the notion of Kolmogorov complexity, which in turn is approximated by the entropy of underlying probability distributions. Second, there are constraint-based causal discovery algorithms. These methods construct a causal model based on an exhaustive list of statistical independencies of any two observed variables conditional on sets of the other observed variables (Peters et al., 2014). One needs at least three observed variables to apply such methods. Thus, the bivariate nature of the problem we are addressing precludes the application of constraint-based causal discovery algorithms. Furthermore, there are score-based methods that compare, e.g., penalized likelihoods across models and base inference of causal direction thereon (see Nowzohour and Bühlmann, 2016, for an example).

3.2.2 Testing conditional independence

This section introduces the concept of Hilbert Space embeddings of distributions and their use for (un)conditional independence testing of random variables. Since this notion is not common in the econometrics literature and conditional independence testing forms a central part of the proposed algorithm, it is pertinent to discuss the procedure in detail. We proceed step-wise and first introduce important underlying concepts such as feature maps, reproducing kernel Hilbert spaces, etc. before turning to how these constructs can help to formulate a conditional independence test.

Feature maps

To introduce the usefulness of a feature map, consider the following problem. Terms used loosely in this paragraph are precisely defined below. Imagine you want to distinguish between two groups of subjects that you are given data about by using a linear classifier, i.e. a linear regression line that serves as a boundary between the two classes. If the data looks like those in Figure 3.1(a), a linear classifier will perform poorly since there is no linear decision boundary that it could uncover. A solution to the problem lies in mapping the data from two-dimensional input space to a higher-dimensional feature space by introducing an additional feature $z = x^2 + y^2$ that complements existing features x and y (here the map is from a two-dimensional to a three-dimensional space). In this higher-dimensional space, there is a linear boundary that separates the two classes, see



(a) Raw data (b) Data representation

Figure 3.1: A non-linear classifier. *Panel (a): data can only be separated by a nonlinear decision boundary, a linear algorithm fails. Panel (b) mapping the data to a higher-dimensional space by introducing an additional feature $z = x^2 + y^2$ enables a linear algorithm to separate the data, source: Lopez-Paz (2016, Figure 3.1)*

Figure 3.1(b). This example is adopted from Lopez-Paz (2016).

Similarly to the linear classifier in Figure 3.1(a) that does not succeed in distinguishing between two classes that are separated by a nonlinear decision boundary in input space, the (linear) covariance between two random variables does not succeed in detecting non-linear statistical dependencies. Mapping the data from input to feature space enables the exemplary classifier to linearly describe the decision boundary in feature space despite it being nonlinear in input space. Similarly, one can use the theory on reproducing kernel Hilbert spaces (RKHS) to construct a representation of marginal and conditional probability distributions in higher-dimensional feature space. The covariance operator between two random variables in that feature space is then informative about nonlinear dependencies in input space. In sum, *any linear algorithm in high-dimensional feature space corresponds to a nonlinear algorithm in input space.* Crucially, inner products between feature space representations can be estimated without knowing the exact feature representation itself (the so-called ‘kernel trick’). We now turn to a formal definition of a RKHS and kernel mean embedding of probability distributions.

Construction of the RKHS

To further elucidate the usefulness of RKHS representations, we proceed step-wise and follow Schölkopf and Smola (2001) and Muandet et al. (2016) in their expositions. Since a RKHS is a kind of Hilbert space and a Hilbert space is a vector space that possesses an inner product, we start by spanning a vector space and define an inner product on

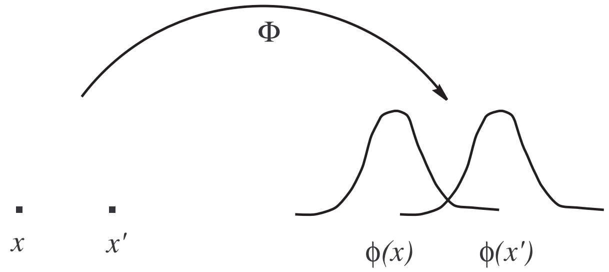


Figure 3.2: Illustration of feature map* Φ*. *Each data point x in input space is mapped to a function $\phi(x)$ in feature space, which represents x in terms of its similarity to all other data points. Figure credit: Schölkopf and Smola (2001, p. 32)*

this space. In particular, we span the vector space as convex combination of a given positive definite kernel function. This subsequently implies the 'reproducing property' of the RKHS.

1. Generalizing the example illustrated in Figure 3.1, we consider higher-dimensional feature representations formalized as kernel functions. The Gaussian kernel, for instance, defined as

$$k(v, v') := \exp\Big(- \frac{\|v - v'\|^2}{\lambda}\Big). \tag{3.4}$$

   for arbitrary vectors v and v' and parameter λ, can serve as a higher-dimensional feature representation. In particular, each data point x is mapped from input space to higher-dimensional feature space where it is represented by its distance to all other data points, i.e. $k(\cdot, x)$. This step is illustrated in Figure 3.2: each data point is richly represented by its similarity (defined by the kernel) to all other data points.

   Formally, we define a feature map Φ from input space $\mathcal{X}$ to the space of functions $\mathbb{R}^{\mathcal{X}}$:

$$\Phi : \ \mathcal{X} \to \mathbb{R}^{\mathcal{X}} \tag{3.5}$$

$$x \mapsto k(\cdot, x) \tag{3.6}$$

   where k is a positive-definite kernel. A positive-definite kernel is a kernel with an associated kernel matrix K, which has entries $K_{ij} := k(x_i, x_j)$, that is positive-definite. Thus, each data point x is represented by a theoretically infinite-dimensional vector or, in other words, *a function* $k(\cdot, x)$. In practice, a data point x is represented by an n-dimensional vector where n is the number of data points in the sample.

2. The next step in constructing an RKHS is opening the vector space. Consider linear

combinations of the feature representations of the form

$$f(\cdot) = \sum_{i=1}^{m} \alpha_i k(\cdot, x_i) \tag{3.7}$$

for $\alpha_i \in \mathbb{R}$ and samples $x_1, \ldots, x_m$ of input space $\mathcal{X}$ where m is an integer index.

3. Given a similarly constructed function

$$g(\cdot) = \sum_{j=1}^{m'} \beta_j k(\cdot, x'_j)$$

with $\beta_i \in \mathbb{R}$ and samples $x'_1, \ldots, x'_{m'}$ of input space $\mathcal{X}$ where m' is an integer index, we can define an inner product between f and g as

$$\langle f, g \rangle := \sum_{i=1}^{m} \sum_{j=1}^{m'} \alpha_i \beta_j k(x_i, x'_j) \tag{3.8}$$

4. Then complete the space spanned by (3.8) by adding the limit points of sequences in the norm defined by $||f|| := \sqrt{\langle f, f \rangle}$, the resulting space $\mathcal{H}$ is called reproducing kernel Hilbert space (RKHS).

This construction implies the 'reproducing property' of the positive-definitive kernel that gives rise to $\mathcal{H}$:

$$\langle k(\cdot, x), f \rangle = f(x). \tag{3.9}$$

In particular,

$$\langle k(\cdot, x), k(\cdot, x') \rangle = \langle \Phi(x), \Phi(x') \rangle = k(x, x'). \tag{3.10}$$

This result shows that the inner product of possibly infinite-dimensional feature representations, $\langle \Phi(x), \Phi(x') \rangle$, can be evaluated through the kernel k without making the feature representation explicit (the so-called 'kernel trick' in machine learning). Any algorithm or other data processing technique that relies on calculating inner products between data representations can be 'kernelized,' i.e. transformed into a nonlinear algorithm by mapping the data into a higher-dimensional space $\mathcal{H}$. The covariance, which can be defined as a dot product, falls into this category.

Instead of representing a specific data point by means of a feature vector, we subsequently intend to represent a whole probability distribution in terms of a higher-dimensional vector. One way to think about this procedure intuitively is to note that probability distributions can be characterized uniquely by an infinite sequence of their moments. Thus, the elements of the infinite-dimensional feature vector are populated by moments of increasing order when embedding such a distribution in the RKHS, which gives rise to a unique representation of the probability distribution.

Kernel mean embedding

The kernel mean embedding extends the concept of a feature map Φ to the space of probability distributions. The map of a probability distribution to a RKHS is defined as

$$\mu : M(\mathcal{X}) \to \mathcal{H} \tag{3.11}$$

$$\mathbb{P} \mapsto \int_{\mathcal{X}} k(x, \cdot) d\mathbb{P}(x) \tag{3.12}$$

where $M(\mathcal{X})$ consists of all probability measures on a measurable space $\mathcal{X}$ and the integral is a Bochner integral. Such a map of probability distribution $\mathbb{P}$ is denoted $\mu_{\mathbb{P}}$ and contains information on all moments of the random variable $\mathbb{P}$ if $k(\cdot, \cdot)$ fulfills some mild conditions.

For a specific class of kernels, called characteristic kernels, the map is injective meaning that the distance of two distributions $\mathbb{P}$ and $\mathbb{Q}$ in $\mathcal{H}$ is zero if and only if the distributions are the same: $||\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}|| = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$. Many widely-used kernels such the Gaussian, Laplacian, Exponential, Poisson etc. are shown to be characteristic (Muandet et al., 2016, Section 3.3.1).

So far, the theory on RKHSs and kernel mean embeddings receives scant attention in the econometrics literature, albeit with notable exceptions. Carrasco et al. (2007) discuss the usefulness of RKHS theory in cases where the researcher has an infinite number of moment conditions that they want to use efficiently. This seemingly rare situation might occur when the moment conditions can be expressed as a *function*, i.e. a vector of infinite length. For instance, Carrasco and Florens (2000) further generalize (already) generalized method of moments estimators to account for infinitely many moment conditions. To analyze the infinitely many moment conditions requires inverting a covariance operator. Akin to the procedure described here, they show that the generalized inverse of such operator only exists in the RKHS. Singh, Rahul; Sahani and Gretton (2019) study the use of kernel methods in the context of instrumental variable (IV) methods. They use kernel mean embeddings of the conditional distribution of the covariates given the instrument to propose a nonlinear extension of linear IV implementations. Grünewälder et al. (2012) analyze connections between kernel mean embeddings and vector-valued functions to analyze Markov decision processes. Flaxman et al. (2015) use kernel mean embeddings to analyze who cast their votes for Obama in the 2012 US presidential election.

Cross-covariance operators and unconditional independence

In addition to mean embeddings, covariance and cross-covariance operators can be defined on the RKHS (Baker, 1973; Fukumizu et al., 2004). These are essential for the formulation of conditional independence tests. To see the connection between the expressive power of the RKHS and the difficult task of (nonlinear) independence testing, consider the result of Rényi (1959), who shows that the two random variables U and X are independent if and only if the maximal covariance is zero:

$$\sup_{f,g} Cov(f(X), g(U)) = 0 \Leftrightarrow X \perp\!\!\!\perp U. \tag{3.13}$$

However, the space of functions that one needs to search over is too large for the result to be of practical use. Gretton et al. (2005) address this problem and show that a RKHS generated by a universal kernel is sufficiently large for the result to hold and sufficiently small for search to be possible (the universality of the kernel relates to the denseness of the RKHS into the space of bounded continuous functions).

Consider an RKHS $\mathcal{H}_x$ with positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ where $\mathcal{X}$ is the domain of X and corresponding feature representation $\phi(x) \in \mathcal{H}_x$. Analogously, we define a second RKHS, $\mathcal{H}_u$, with kernel $l : \mathcal{U} \times \mathcal{U} \to \mathbb{R}$ where $\mathcal{U}$ is the domain of U and feature map ψ on $\mathcal{H}_u$. Now, we can define the cross covariance operator $C_{XU} : \mathcal{H}_u \to \mathcal{H}_x$ as

$$\begin{aligned} C_{XU} :&= \mathbb{E}_{XU}[(\phi(x) - \mu_{\mathbb{P}_X}) \otimes (\psi(u) - \mu_{\mathbb{P}_U})] \\ &= \mathbb{E}_{XU}[\phi(x) \otimes \psi(u)] - \mu_{\mathbb{P}_X} \otimes \mu_{\mathbb{P}_U} \\ &= \mu_{\mathbb{P}_{XU}} - \mu_{\mathbb{P}_X} \otimes \mu_{\mathbb{P}_U} \end{aligned} \tag{3.14}$$

where the expectation is taken over the joint distribution of (X, U), *cf.* eq. (3.12).

Define the Hilbert-Schmidt Independence Criterion (HSIC) as the squared Hilbert-Schmidt norm of the cross-covariance operator C_{XU} (Gretton et al., 2007):

$$HSIC(\mathbb{P}_{XU}, \mathcal{H}_x, \mathcal{H}_u) := \|C_{XU}\|^2_{HS} = \sum_j \lambda_j^2 \tag{3.15}$$

where λ_j are all j eigenvalues of C_{XU}.

If the product kernel $k(\cdot,\cdot) \times l(\cdot,\cdot)$ is a characteristic kernel on $\mathcal{X} \times \mathcal{U}$, i.e. the map μ is injective, one can show the following central result:

$$HSIC(\mathbb{P}_{XU}, \mathcal{H}_x, \mathcal{H}_u) = 0 \Leftrightarrow X \perp\!\!\!\perp U. \tag{3.16}$$

For the estimation of (3.15) and its asymptotic distribution see Pfister et al. (2018).

For the implementation of the proposed method, a test for *conditional* independence is needed. We turn to a formalization of such a test in the following. The discussion of an unconditional independence test serves to fix ideas that continue to be relevant.

Partial cross-covariance operators and conditional independence

A natural next step to build a test for conditional independence is to rely on a characterization of conditional independence in terms of a conditional cross covariance operator $C_{(X,U)|W}$ of (X, U) given W, i.e. an extension of the cross covariance operator in eq. (3.14) to the case with conditioning variables. Indeed, Fukumizu et al. (2008) show $C_{(X,U)|W} = 0 \Leftrightarrow X \perp\!\!\!\perp U|W$. However, the distribution of that conditional cross covariance operator under the null of conditional independence is difficult to derive, which – so far – precludes its use for hypothesis testing.

However, an analogy to the characterization of conditional independence for jointly Gaussian variables in terms of vanishing partial correlation shows how to make progress. To illustrate, first note that, for jointly Gaussian variables (Z_1, Z_2, Z_W), the conditional independence, $Z_1 \perp\!\!\!\perp Z_2|Z_W$ can be characterized as the correlation between $Z_1|Z_W$ and $Z_2|Z_W$ being zero. In other words, since independence and correlation coincide for jointly Gaussian variables, one can conclude from the partial correlation between Z_1 and Z_2 given Z_W going to zero that $Z_1 \perp\!\!\!\perp Z_2|Z_W$. Partial correlation is a *linear* concept defined by the orthogonality of *linear* maps of Z_1 and Z_2 on the space orthogonal to Z_W. It can only characterize conditional independence for *jointly* Gaussian variables. This limited applicability to jointly Gaussian variables lies in the linearity of the underlying maps. Intuitively, one can extend the results to apply to *nonlinear* dependence of *arbitrarily* distributed random variables if such maps can be described more flexibly. We have seen how maps of data and whole probability distributions into higher-dimensional RHKS enables the use of linear algorithms to study non-linear relationships. This reasoning also underlies the following characterization of, and test for, conditional independence for arbitrarily distributed random variables.

Throughout the remainder of this section, we consider continuous random variables X, U and W with domains $\mathcal{X}$, $\mathcal{U}$ and $\mathcal{W}$, and with positive definite kernels $k_{\mathcal{X}}$, $k_{\mathcal{U}}$, and $k_{\mathcal{W}}$ defined on these domains. These give rise to RKHSs $\mathcal{H}_{\mathcal{X}}$, $\mathcal{H}_{\mathcal{U}}$ and $\mathcal{H}_{\mathcal{W}}$ respectively. We make use of the notation $\tilde{X} = (X, W)$ and $\tilde{U} = (U, W)$. Define $k_{\tilde{\mathcal{X}}} = k_{\mathcal{X}} \times k_{\mathcal{W}}$ and corresponding RKHS $\mathcal{H}_{\tilde{\mathcal{X}}}$.

First, consider a result due to work by Daudin (1980), who characterizes conditional independence as partial correlation of appropriate functions in appropriate function spaces being zero. Consider, L^2 function spaces

$$\mathcal{F}_{\tilde{X}} := \{f \in L^2_{\tilde{X}} \mid \mathbb{E}[f(\tilde{X})|W] = 0\}, \tag{3.17}$$
$$\mathcal{F}_U := \{g|g(U, W) = g'(U) - h_{g'}(W), g' \in L^2_U\} \tag{3.18}$$

where, for an arbitrary Z, L^2_Z denotes the space of square integrable functions of Z, $h_{g'}(W)$ is a nonlinear regression function of $g'(U)$ on W.

For any function $\tilde{f} \in L^2_{\tilde{X}}$ in $\mathcal{F}_{\tilde{X}}$ define

$$f(\tilde{X}) := \tilde{f}(\tilde{X}) - h_{\tilde{f}}(W), \tag{3.19}$$

where $h_{\tilde{f}}$ is a nonlinear regression function of $\tilde{f}(\tilde{X})$ on W.

With these definition, Daudin (1980) shows that

$$\sup_{f,g} \mathbb{E}[f(\tilde{X})g(\tilde{U})] = 0 \Leftrightarrow X \perp\!\!\!\perp U|W. \tag{3.20}$$

The applicability of Daudin's result, similar to the one by Rényi (1959), is limited in practice because the considered L^2 space of functions does not admit a concise expression for $\mathbb{E}[f(\tilde{X})g(\tilde{U})]$. Kun Zhang et al. (2011) show, similar to Gretton et al. (2005) w.r.t. eq. (3.13) above, that restricting function classes of f and g to lie in RKHSs $\mathcal{H}_{\tilde{X}}$ and $\mathcal{H}_U$ is sufficient to make Daudin's result in eq. (3.20) operational in practice. Specifically, they define a partial cross-covariance operator as

$$C_{\tilde{X}U\cdot W} := C_{\tilde{X}U} - C_{\tilde{X}W} C^{-1}_{WW} C_{WU} \tag{3.21}$$

and show

$$KCI(\mathbb{P}_{XUW}, \mathcal{H}_{\tilde{X}}, \mathcal{H}_U) := \|C_{\tilde{X}U\cdot W}\|^2_{HS} = 0 \Leftrightarrow X \perp\!\!\!\perp U|W \tag{3.22}$$

which is the KCI test statistic we use in the subsequent algorithmic implementation. For the derivation of the asymptotic distribution see Kun Zhang et al. (2011) and Strobl et al. (2019).

In sum, the idea of feature representations motivates the map of the distributions of X and U conditional on W into higher-dimensional spaces where linear correlations correspond to nonlinear dependencies in original space. The KCI test statistic is a central component of the algorithm to infer causal direction presented in the following.

3.2.3 Algorithmic implementation

Hoyer et al. (2009) and Mooij et al. (2016) discuss inference of the causal direction between *two* random variables (cause and effect) from observational data. Peters et al. (2014) constitutes a theoretical extension of these methods to more than two variables. The paper at hand falls between these two strands as it accounts for more than two variables, yet its primary concern is the causal directionality between a subset of just two of them. The remaining variables $\mathbf{W}$ serve as controls.

The theory implies an independence of errors and the covariate in the causal model; whereas, an independence between the errors and the covariate does not hold in the anticausal model. The algorithm involves testing the independence between errors and covariate conditional on $\mathbf{W}$ in both causal and anticausal model. Ideally, the test would conclude with the following decisions: i) if independence can be rejected at a pre-specified significance level in one model but not in the other, one would conclude that the latter model represents the correct causal relation, ii) if independence is rejected in both models, one would conclude that the relation between X and Y is confounded, and iii) if independence cannot be rejected in either model, one would conclude that the test does not have sufficient power to decide on the causal direction. It is not possible to implement such a strategy in practice because the true errors are unobserved and the practitioner has to rely on estimated errors. Specifically, the practitioner does not have a sample of $U := Y - h(X, \mathbf{W})$ in eq. (3.1) at their disposal and, therefore, must rely on estimated errors $\hat{U} := Y - \hat{h}(X, \mathbf{W})$, and the respective estimated errors of the model in eq. (3.3), to investigate which model is correct. That these residuals are estimated and, in particular, that they depend on the estimated $\hat{h}$, poses a challenge that we discuss now.

Mooij et al. (2016) and Hoyer et al. (2009) propose randomly splitting the available data $\mathcal{D} = \{Y_i, X_i, \mathbf{W}_i\}_{i=1}^{n}$ in training and test sets, denoted $\mathcal{D}^{tr} = \{Y_i, X_i, \mathbf{W}_i\}_{i=1}^{n/2}$ and $\mathcal{D}^{te} = \{Y_i', X_i', \mathbf{W}_i'\}_{i=(n/2)+1}^{n}$, respectively. $\mathcal{D}^{tr}$ is used to get an estimate $\hat{h}$ of the true regression function h. $\mathcal{D}^{te}$ is then used to get estimates $\hat{\varepsilon}' := Y' - \hat{h}(X', \mathbf{W}')$ of the true errors ε. An error in the estimated $\hat{h}$ induces a dependence of $\hat{\varepsilon}'$ and X' (conditional on $\mathbf{W}'$) even though ε and X are truly independent (conditional on $\mathbf{W}$). Consequently, conventional thresholds for the independence test tend to be too loose and would ideally incorporate the fact that $\hat{h}$ is estimated. Specifically, for a conventional threshold of, say, $\alpha^* = 0.05$ the empirical rejection rate will be larger than α^* in the causal model even though under H_0 we have that $U \perp\!\!\!\perp X|\mathbf{W}$, which should lead to an empirical rejection rate roughly equal to α^*. To achieve an empirical size of α^*, one needs to use a threshold $\alpha = \alpha^* \times \lambda_\alpha$ with $0 < \lambda_\alpha < 1$. There are no theoretical results on how to choose λ_α to account for the dependence of $\hat{\varepsilon}'$ and X'.[2] However, Mooij et al. (2016) show that one can infer the correct directionality under additional assumption that the causal *or* the anticausal model exist. Therefore, the identifiability result in Theorem 3.2.1, which states that either causal or anticausal model, but not both, can satisfy the independence of the error with the covariate, in combination with the existence assumption in Assumption 3.2.3, which states that either causal or anticausal model exist, allows us to infer the directionality with Algorithm 1.

In particular, under Assumption 3.2.3, one can infer that the model with the lower KCI test statistic (i.e. a larger p-value of the conditional independence test) is the correct causal model, thereby circumventing the lack of theoretical guidance about an appropriate threshold. Making this assumption comes at a cost; namely, a procedure that relies on comparing two test statistics can never conclude that there is not enough information in the data to decide on the causal direction. In other words, such a procedure will never conclude that there is a lack of power to make a decision. However, we show in subsequent Monte Carlo simulations that the procedure almost always picks the correct causal model. Assumption 3.2.3 is strong; yet, if one is willing to make it, the probability of inferring the wrong direction is very low.

A comment on the sample splitting procedure follows. We compare the sample splitting procedure proposed above to an alternative sample splitting procedure. For simplicity, we neglect the conditional nature of the applied tests in this paragraph. In this alternative sample splitting procedure $\mathcal{D}^{tr}$ is used to estimate both h and the residuals $\hat{\varepsilon} := Y - \hat{h}(X, \mathbf{W})$. The independence test is then implemented by using $\hat{\varepsilon}$ and $\mathcal{D}^{te} = \{X_i', \mathbf{W}_i'\}_{i=(n/2)+1}^{n}$. Here, $\hat{\varepsilon}$ and X_i' are draws of two independent random variables, *by construction.* Moreover, in this alternative approach paired sample tests cannot be used. However, in order to find evidence that favours either the causal or anticausal model, we need to rely on *paired* samples of the estimated residual and the covariate, in both the causal and anticausal models. If the alternative sample splitting were used, the estimated errors were independent of the covariate in *both* causal and anticausal model *by construction* yielding no insight into the causal direction.

[2] Simulation studies, which are not replicated here, show that λ_α depends on the type of distribution that the true error follows. Since there is no way for a practitioner to get a hold on that error distribution, it is impossible to propose rules of thumb, substantiated by simulation excercises, to indicate the level of λ_α as a function of observable or estimable quantities.

```
Data: $\mathcal{D} = \{Y_i, X_i, \mathbf{W}_i\}_{i=1}^{n}$
Output: Decision whether true causal model is $X \to Y$ or $Y \to X$
Step 1: Normalize data to have mean equal to zero and variance equal to one.
Step 2: Randomly split data in half to form training $\mathcal{D}^{tr} = \{Y_i, X_i, \mathbf{W}_i\}_{i=1}^{n/2}$ and
  test set $\mathcal{D}^{te} = \{Y_i', X_i', \mathbf{W}_i'\}_{i=(n/2)+1}^{n}$
Step 3: Estimate generalized additive models (GAMs) based on $\mathcal{D}^{tr}$
GAM1: $Y = h(X, \mathbf{W}) + U$, call resulting estimate $\hat{h}$
GAM2: $X = \tilde{h}(Y, \mathbf{W}) + \tilde{U}$, call resulting estimate $\hat{\tilde{h}}$
Step 4: calculate residuals based on $\mathcal{D}^{te}$
$\hat{U} := Y' - \hat{h}(X', \mathbf{W}')$, and
$\hat{\tilde{U}} := X' - \hat{\tilde{h}}(Y', \mathbf{W}')$
Step 5: Test conditional independence with KCI test (based on residuals from
  Step 4)
use $\hat{U}$, $X'$ and $\mathbf{W}'$ to test $U \perp\!\!\!\perp X|\mathbf{W}$ with $KCI$ test; call resulting test statistic
  $KCI_{\text{causal}}$
use $\hat{\tilde{U}}$, $Y'$ and $\mathbf{W}'$ to test $\tilde{U} \perp\!\!\!\perp Y|\mathbf{W}$ with $KCI$ test; call resulting test statistic
  $KCI_{\text{anticausal}}$
Step 6: Decide on causal direction
if $KCI_{causal} < KCI_{anticausal}$ then
    accept $X \to Y$ as correct model
else if $KCI_{causal} > KCI_{anticausal}$ then
    accept $Y \to X$ as correct model
else if $KCI_{causal} = KCI_{anticausal}$ then
    inconclusive test
end
```

Algorithm 1: Reverse causality test

3.3 Monte Carlo simulations

In order to understand the sensitivity of the algorithm with respect to the nonlinearity of the function relating the cause to its effect we introduce a parameter, τ, that controls the strength of this nonlinearity. Furthermore, we want to understand the robustness of the algorithm to heteroskedasticity w.r.t. a third variable W. For this purpose, we introduce a parameter ρ, which controls the strength of the heteroskedasticity.

We simulate the following data for $\tau \in \{0, 1\}$, $\rho \in \{0, 1\}$, $n \in \{500, 1000\}$, and 500

Monte Carlo draws:

$$(X, W) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2) \tag{3.23}$$

$$Y = \kappa(X, W, \tau) + U \tag{3.24}$$

where U is defined as

$$U \sim \mathcal{N}(0, (1 + f(W))^{\rho}) \tag{3.25}$$

and f is the density function of W.

Each ρ-τ-combination is implemented with the following specifications for $\kappa(\cdot, \tau)$ for each $i = 1, \ldots, n$:

$$\kappa_1(X_i, W_i, \tau) = X_i + \tau X_i^3 + W_i \tag{3.26}$$

$$\kappa_2(X_i, W_i, \tau) = X_i + \tau \sin(X_i + \pi/2) + W + W_i^2 / \max\{W_1^2, \ldots, W_n^2\} \tag{3.27}$$

To explore the robustness of our results with respect to the distribution of the error U, we run the simulation with errors drawn from sub- and super-Gaussian distribution by raising the draws in (3.25) to q while keeping sign and variance. We estimate both causal and anticausal models with a Generalized Additive Model using smoothing splines. See Algorithm 1. The results of these Monte Carlo studies are found in Tables 3.1 and 3.2.

When the relationship between X and Y is linear, i.e. $\tau = 0$, the algorithm randomly chooses between the two directions. This is consistent with the theory since the causal direction is not identifiable in the linear case. It also shows the implication of the lack of theoretical guidance for choosing λ_α (see discussion above). Note that the direction is identifiable in the linear case with a non-Gaussian error distribution, see also Appendix 3.6.2 and 3.6.3. As soon as the relation between cause and effect becomes non-linear, i.e. $\tau \neq 0$, the algorithm makes the correct decision in more than 95% of the cases. For instance, when $n = 1000$, the relation between cause and effect is nonlinear ($\tau = 1$) the algorithm arrives at the correct conclusion in 98% of the Monte Carlo runs, regardless of the level of heteroskedasticity and q. We show that the results remain robust to different error variances and q in Appendix 3.6.2.

In sum, two observations are worth stressing. First, the results show that the algorithm has power when cause and effect are related non-linearly. Second, the performance of the algorithm does not suffer from heterskedastic errors w.r.t. W. Next, we turn to an empirical illustration of the algorithm.

Table 3.1: *This Table shows Monte Carlo results of the procedure to infer the causal direction between two variables described in Algorithm 1. Underlying data is simulated as* $Y = \kappa_1(X, W, \tau) + U \times 1$ *where* $U \sim \mathcal{N}(0, (1 + f(W))^\rho)$*, where* f *is the probability density function of W. U raised to q while keeping its sign and variance. The model where X is causing Y is the correct model.* 500 *Monte Carlo runs.*

				share of decision	
n	τ	ρ	q	correct	false
500	0	0	0.8	0.774	0.226
			1.0	0.428	0.572
			1.2	0.652	0.348
		1	0.8	0.702	0.298
			1.0	0.456	0.544
			1.2	0.700	0.300
	1	0	0.8	0.890	0.110
			1.0	0.918	0.082
			1.2	0.900	0.100
		1	0.8	0.916	0.084
			1.0	0.930	0.070
			1.2	0.932	0.068
1000	0	0	0.8	0.888	0.112
			1.0	0.468	0.532
			1.2	0.840	0.160
		1	0.8	0.918	0.082
			1.0	0.466	0.534
			1.2	0.816	0.184
	1	0	0.8	0.988	0.012
			1.0	0.996	0.004
			1.2	0.976	0.024
		1	0.8	0.990	0.010
			1.0	0.988	0.012
			1.2	0.990	0.010

Table 3.2: *This Table shows Monte Carlo results of the procedure to infer the causal direction between two variables described in Algorithm 1. Underlying data is simulated as* $Y = \kappa_2(X, W, \tau) + U \times 1$ *where* $U \sim \mathcal{N}(0, (1 + f(W))^\rho)$*, where* f *is the probability density function of W. U raised to* q *while keeping its sign and variance. The model where* X *is causing* Y *is the correct model.* 500 *Monte Carlo runs.*

				share of decision	
n	τ	ρ	q	correct	false
500	0	0	0.8	0.680	0.320
			1.0	0.450	0.550
			1.2	0.634	0.366
		1	0.8	0.734	0.266
			1.0	0.450	0.550
			1.2	0.606	0.394
	1	0	0.8	0.982	0.018
			1.0	0.984	0.016
			1.2	0.974	0.026
		1	0.8	0.986	0.014
			1.0	0.974	0.026
			1.2	0.984	0.016
1000	0	0	0.8	0.928	0.072
			1.0	0.432	0.568
			1.2	0.834	0.166
		1	0.8	0.922	0.078
			1.0	0.486	0.514
			1.2	0.834	0.166
	1	0	0.8	1.000	0.000
			1.0	0.998	0.002
			1.2	1.000	0.000
		1	0.8	1.000	0.000
			1.0	1.000	0.000
			1.2	0.998	0.002

3.4 Empirical illustration

We use data from the Survey of Income and Expenditure (Einkommens- und Verbrauchsstichprobe, EVS), which is a voluntary survey of roughly 60,000 households in Germany, to test the proposed algorithm. We consider the following variables: income, expenditure, highest educational attainment of the main earner, highest professional training of the main earner, and age group of the main earner. We analyze the causal direction between income and work experience, which we proxy by age group.

Hump-shaped income profiles over the life-cycle are well-documented in labor economics (Heckman et al., 2006). It is interesting to test the algorithm for a cause-effect pair where the causal direction is *a priori* clear. Since work experience mechanically increases over the life-cycle, it can be credibly assumed not to be caused by income changes. Therefore, we analyze the directionality between income and age where age can be interpreted as proxy for work experience. We posit the correct causal model to be

$$Y = h(E, \mathbf{Z}) + \varepsilon_y, \tag{3.28}$$

where experience E is causing income Y. Vice versa, the anticausal model in which income is causing work experience is given as

$$E = \tilde{h}(Y, \mathbf{Z}) + \varepsilon_e \tag{3.29}$$

where in each model $\mathbf{Z}$ contains all remaining covariates as control.

We aim to alleviate the problem that we are likely to omit many crucial confounding variables by splitting the data in n_q quantiles of the income distribution. At least part of the omitted confounding factors can be assumed to be fixed within given quantiles as they collect individuals with roughly similar life-styles etc. This argument applies more strongly the larger the number of quantiles the income distribution is split in. On the other hand, the larger n_q the smaller the number of observations within each quantile and the lower the power of the test to prefer the correct causal direction. Therefore, we show results for a set of $n_q = \{4, \ldots, 20\}$ quantiles.[3] For each number of quantiles n_q, we run the test in each of these n_q quantiles and plot the share of quantiles in which the algorithm prefers either model (note that the x-axis in Figure 3.3 refers to the *number* of quantiles the income distribution is split in, not the quantiles as such). For example, the bar above $n_q = 5$ in Figure 3.3 denotes that in 3 of the 5 quantiles, i.e. 60%, the algorithm concludes

[3] The KCI test, which forms an important part of the algorithm, requires the inversion of $n \times n$ matrices where n is the number of observations. Constraints on local computing power preclude running the test on the whole sample with roughly 60,000 observations or with $n_q = \{1, 2, 3\}$.

that experience is causing income. Regardless of n_q, the test always favours the model where work experience, proxied by age, is causing income in at least 50% of quantiles.

Since income is one of the potential causes in this application, splitting the distributions into quantiles of income has the drawback that the variation in income that the algorithm can use within each quantile is mechanically reduced as n_q increases. Therefore, we replicate the analysis by splitting the data into quantiles of the expenditure distribution and controlling for income as one of the control variables in $\mathbf{Z}$. The results can be seen in Figure 3.4. The algorithm now favours the causal model in more than 75% of the quantiles for most n_q.

In sum, this application documents that our algorithm gives economically meaningful results in empirical applications.

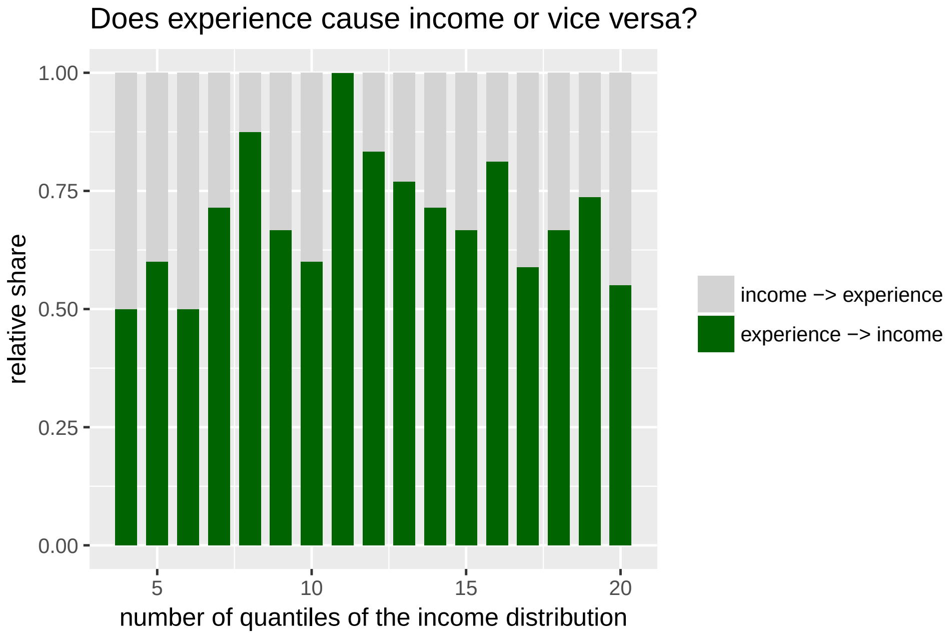


Figure 3.3: *This Figure shows the results of the empirical application of Algorithm 1 to the question whether work experience is causing income or vice versa. The x-axis shows the number of quantiles that the income distribution is split in (not to be mistaken with the quantiles as such). The stacked bars show the shares of the respective number of quantiles the algorithm decides the causal or anticausal model is the correct model.*

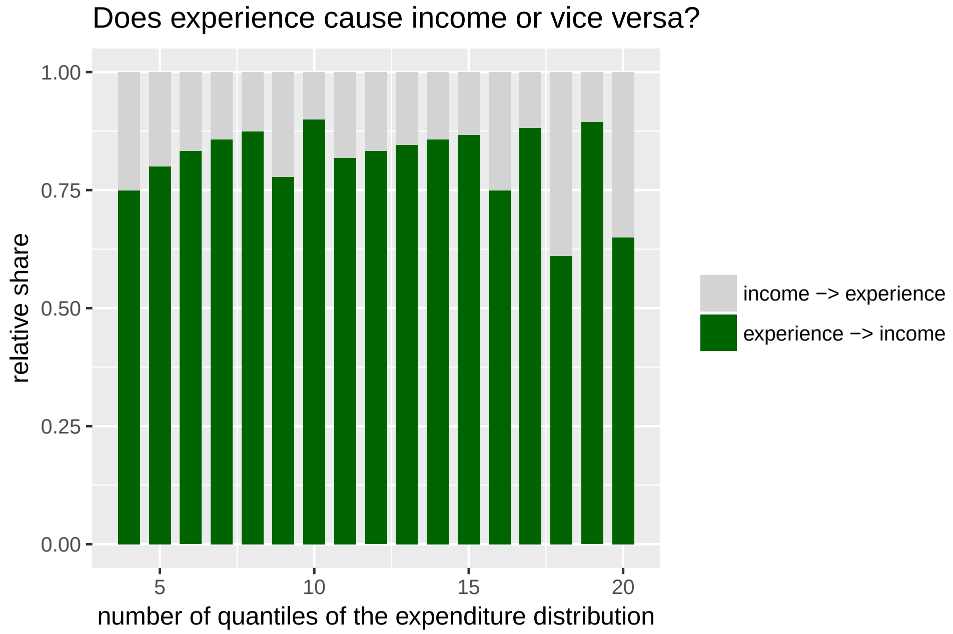


Figure 3.4: *This Figure shows the results of the empirical application of Algorithm 1 to the question whether work experience is causing income or vice versa. The x-axis shows the number of quantiles that the expenditure distribution is split in (not to be mistaken with the quantiles as such). The stacked bars show the shares of the respective number of quantiles the algorithm decides the causal or anticausal model is the correct model.*

3.5 Conclusion

Endogeneity is a common threat to causal identification in econometric models. Reverse causality is one source of such endogeneity. We build on work done by Hoyer et al. (2009) and Mooij et al. (2016) who have shown that the causal direction between two variables X and Y is identifiable in models with additively separable error terms and nonlinear function forms. We extend their results by allowing for additional control covariates $\mathbf{W}$ and heteroskedasticity w.r.t. them.

An empirical application underscores the feasibility of the proposed algorithm. We analyze the causal link between income and work experience, as proxied by age, and show that our procedure provides suggestive evidence that the true causal direction is from work experience to income. Though substantively not surprising precisely because income mechanically cannot causally influence work experience, it is encouraging that our algorithm can distinguish between the causal directions without resorting to instruments or other sources of exogenous variation.

A central problem that must be addressed in future research is how to adjust the critical values of the involved conditional independence tests, i.e. how to choose λ_α. Achieving progress in this direction will make Assumption 3.2.3 unnecessary and would increase the usefulness of the provided test.

3.6 Appendix

3.6.1 Irreversibility proof

The goal is to infer from observational data whether the model in eq. (3.1) or its anticausal version,

$$X = \tilde{h}(Y, \mathbf{W}) + \underbrace{\tilde{\varepsilon}\tilde{\sigma}(\mathbf{W})}_{\tilde{U}} \quad \text{with} \quad \tilde{\varepsilon} \perp\!\!\!\perp (Y, \mathbf{W}), \tag{3.30}$$

is the correct causal model.

Proof of Theorem 3.2.1. We prove that the anticausal model with independence assumption, $\tilde{\varepsilon} \perp\!\!\!\perp (Y, \mathbf{W})$, does not exist in general, if the causal model fulfills the corresponding independence assumption $\varepsilon \perp\!\!\!\perp (Y, \mathbf{W})$. We complement Hoyer et al. (2009) by showing that irreversibility can be proven in the presence of errors that are heteroskedastic w.r.t. a set of additional covariates $\mathbf{W}$.

Step 1. Referring to model (3.30), we derive an expression for the conditional density of $X|Y, \mathbf{W}$:

$$\begin{aligned} P(X \le x | Y = y, \mathbf{W} = \mathbf{w}) &= P(\tilde{h}(Y, \mathbf{W}) + \tilde{\varepsilon}\tilde{\sigma}(\mathbf{W}) \le x | Y = y, \mathbf{W} = \mathbf{w}) \\ &= P\left(\tilde{\varepsilon} \le \frac{x - \tilde{h}(Y, \mathbf{W})}{\tilde{\sigma}(\mathbf{W})} | Y = y, \mathbf{W} = \mathbf{w}\right) \\ &= P\left(\tilde{\varepsilon} \le \frac{x - \tilde{h}(y, \mathbf{w})}{\tilde{\sigma}(\mathbf{w})}\right) \end{aligned} \tag{3.31}$$

where the last step uses the independence assumption $\tilde{\varepsilon} \perp\!\!\!\perp (Y, \mathbf{W})$. Thus, we conclude

$$f_{X|Y,\mathbf{W}}(x|y, \mathbf{w}) = f_{\tilde{\varepsilon}}\left(\frac{x - \tilde{h}(y, \mathbf{w})}{\tilde{\sigma}(\mathbf{w})}\right). \tag{3.32}$$

Assuming that the anticausal model (3.30) does indeed exist, we can express the joint density of x and y conditional on $\mathbf{w}$ as

$$f_{X,Y|\mathbf{W}}(x, y|\mathbf{w}) = f_{\tilde{\varepsilon}}\left(\frac{x - \tilde{h}(y, \mathbf{w})}{\tilde{\sigma}(\mathbf{w})}\right) f_{Y|\mathbf{W}}(y|\mathbf{w}). \tag{3.33}$$

We define $\tilde{\nu} := \log f_{\tilde{\varepsilon}}$, $\eta := \log f_{Y|\mathbf{W}}$ and

$$\begin{aligned}\pi(x, y, \mathbf{w}) : &= \log f(x, y|\mathbf{w}) \\ &= \eta(y, \mathbf{w}) + \tilde{\nu}\left(\frac{x - \tilde{h}(y, \mathbf{w})}{\tilde{\sigma}(\mathbf{w})}\right).\end{aligned} \tag{3.34}$$

Taking partial derivatives yields

$$\frac{\partial^2 \pi(x, y, \mathbf{w})}{\partial x \partial y} = -\tilde{\nu}''\left(\frac{x - \tilde{h}(y, \mathbf{w})}{\tilde{\sigma}(\mathbf{w})}\right)\frac{\tilde{h}'(y, \mathbf{w})}{\tilde{\sigma}(\mathbf{w})^2} \tag{3.35}$$

and

$$\frac{\partial^2 \pi(x, y, \mathbf{w})}{\partial x^2} = \tilde{\nu}''\left(\frac{x - \tilde{h}(y, \mathbf{w})}{\tilde{\sigma}(\mathbf{w})}\right)\frac{1}{\tilde{\sigma}(\mathbf{w})^2}. \tag{3.36}$$

which, in turn, results in

$$\frac{\frac{\partial^2 \pi(x,y,\mathbf{w})}{\partial x^2}}{\frac{\partial^2 \pi(x,y,\mathbf{w})}{\partial x \partial y}} = -\frac{1}{\tilde{h}'(y, \mathbf{w})}. \tag{3.37}$$

Therefore, taking the derivative of the ratio (3.37) w.r.t. x, we conclude

$$\frac{\partial}{\partial x}\left(\frac{\frac{\partial^2 \pi}{\partial x^2}}{\frac{\partial^2 \pi}{\partial x \partial y}}\right) = 0. \tag{3.38}$$

Step 2. Now, we derive similar restrictions for the causal model, (3.1).

First, we derive an expression for the conditional density of $Y|X, \mathbf{W}$. Similar to (3.31), we can write

$$\begin{aligned}P(Y \leq y|X = x, \mathbf{W} = \mathbf{w}) &= P(h(X, W) + \varepsilon\sigma(\mathbf{W}) < y|X = x, \mathbf{W} = \mathbf{w}) \\ &= P\left(\varepsilon \leq \frac{y - h(x, \mathbf{w})}{\sigma(\mathbf{w})}\right)\end{aligned} \tag{3.39}$$

which uses the independence assumption $\varepsilon \perp\!\!\!\perp (X, \mathbf{W})$. This lets us conclude

$$f_{Y|X,\mathbf{W}}(y|x, \mathbf{w}) = f_\varepsilon\left(\frac{y - h(x, \mathbf{w})}{\sigma(\mathbf{w})}\right). \tag{3.40}$$

Therefore, the conditional density of $X, Y|\mathbf{W}$ can be expressed as

$$f_{X,Y|\mathbf{W}}(x,y|\mathbf{w}) = f_\varepsilon\left(\frac{y-h(x,\mathbf{w})}{\sigma(\mathbf{w})}\right) f_{X|\mathbf{W}}(x|\mathbf{w}) \tag{3.41}$$

with $f_{X|\mathbf{W}}$ and f_ε probability densities on $\mathbb{R}$.

We define $\nu := \log\ f_\varepsilon$, $\xi := \log\ f_{X|\mathbf{W}}$ and

$$\begin{aligned}\pi(x,y,\mathbf{w}) :&= \log f_{X,Y|\mathbf{W}}(x,y|\mathbf{w}) \\ &= \xi(x,\mathbf{w}) + \nu\Big(\frac{y-h(x,\mathbf{w})}{\sigma(\mathbf{w})}\Big).\end{aligned} \tag{3.42}$$

Taking partial derivatives, we conclude

$$\begin{aligned}\frac{\partial^2\pi(x,y,\mathbf{w})}{\partial x^2} &= \frac{h'^2(x,\mathbf{w})}{\sigma(\mathbf{w})^2}\nu''\left(\frac{y-h(x,\mathbf{w})}{\sigma(\mathbf{w})}\right) - \frac{h''(x,\mathbf{w})}{\sigma(\mathbf{w})}\nu'\left(\frac{y-h(x,\mathbf{w})}{\sigma(\mathbf{w})}\right) + \xi''(x,\mathbf{w}) \\ &=: \phi_1(x,y,\mathbf{w}) + \xi''(x,\mathbf{w})\end{aligned} \tag{3.43}$$

and

$$\begin{aligned}\frac{\partial^2\pi(x,y,\mathbf{w})}{\partial x\partial y} &= -\nu''\left(\frac{y-h(x,\mathbf{w})}{\sigma(\mathbf{w})}\right)\frac{h'(x,\mathbf{w})}{\sigma(\mathbf{w})^2} \\ &=: \phi_2(x,y,\mathbf{w}).\end{aligned} \tag{3.44}$$

In the following derivations we omit the arguments $(x,\mathbf{w})$ for ξ, and $(x,y,\mathbf{w})$ for ϕ_1 and ϕ_2. The ratio of eqs. (3.43) and (3.44) is given by

$$\frac{\frac{\partial^2\pi}{\partial x^2}}{\frac{\partial^2\pi}{\partial x\partial y}} = \frac{\phi_1(x,y,\mathbf{w}) + \xi''(x,\mathbf{w})}{\phi_2(x,y,w)} \tag{3.45}$$

which we derive w.r.t. x to conclude

$$\frac{\partial}{\partial x}\left(\frac{\frac{\partial^2\pi}{\partial x^2}}{\frac{\partial^2\pi}{\partial x\partial y}}\right) = \frac{\xi'''}{\phi_2} - \frac{\xi''\phi_2'}{\phi_2^2} + \frac{\phi_1'\phi_2 - \phi_1\phi_2'}{\phi_2^2}. \tag{3.46}$$

If a anticausal model exists, we know that (3.46) must equal zero (from (3.38), which is derived from the anticausal model). By setting (3.46) equal to zero and given h, ν, we obtain for each fixed y and $\mathbf{w}$, which we denote $\bar{y}$ and $\bar{\mathbf{w}}$, respectively, a linear inhomogenous differential equation for ξ:

$$\xi'''(x, \bar{\mathbf{w}}) = \xi''(x, \bar{\mathbf{w}})G(x, \bar{y}, \bar{\mathbf{w}}) + H(x, \bar{y}, \bar{\mathbf{w}}). \tag{3.47}$$

where $G(x, y, \mathbf{w}) = \frac{\phi_2'}{\phi_2}$ and $H(x, y, \mathbf{w}) = \frac{\phi_1\phi_2' - \phi_1'\phi_2}{\phi_2}$. Defining $z := \xi''$, we have

$$\frac{\partial z(x, \bar{\mathbf{w}})}{\partial x} = z(x, \bar{\mathbf{w}})G(x, \bar{y}, \bar{\mathbf{w}}) + H(x, \bar{y}, \bar{\mathbf{w}}). \tag{3.48}$$

Finally, given such a solution for $z(x, \bar{\mathbf{w}})$ exists, it is given by

$$z(x, \bar{\mathbf{w}}) = z(x_0, \bar{\mathbf{w}})e^{\int_{x_0}^{x} G(\tilde{x}, \bar{y}, \bar{\mathbf{w}})d\tilde{x}} + \int_{x_0}^{x} e^{\int_{\hat{x}}^{x} G(\tilde{x}, \bar{y}, \bar{\mathbf{w}})d\tilde{x}} H(\hat{x}, \bar{y}, \bar{\mathbf{w}})d\hat{x} \tag{3.49}$$

Thus, "the set of all functions satisfying linear inhomogenous differential [eq. (3.47)] is a 3-dimensional affine space: Once we have fixed $\xi(x_0)$, $\xi'(x_0)$, $\xi''(x_0)$ for some arbitrary x_0, ξ is completely determined. Given fixed f and ν, the set of all ξ admitting a anticausal model is contained in this subspace." (Hoyer et al., 2009, after Theorem 1) □

More intuitively, it is shown that causal and anticausal models can only exist simultaneously under specific circumstances. In fact, if the joint distribution of X and Y is to allow for both a causal and a anticausal model, we show that the causal model has to satisfy differential equation (3.47). This is a requirement that a generic causal model does not fulfill. What exactly do *specific* and *generic* refer to? The solutions of the differential equation restrict the log density of X to lie in a (specific) *three-dimensional* space, although *a priori* the (generic) space of possible log marginal densities of X is *infinite-dimensional.*

3.6.2 Further simulation results

In this section, we provide further simulation results. In particular, we replicate the simulations discussed in the main part for different variances of U. These results are found in Tables 3.3 to 3.6.

The fact that the algorithm rejects the anticausal model despite the functional relationship between cause and effect being linear ($\tau = 0$) when the error distribution is either sub- or super-Gaussian ($q \in \{0.8, 1.2\}$) reflects the identifiability results of Shimizu et al. (2006, see also Appendix 3.6.3).

Table 3.3: *This Table shows Monte Carlo results of the procedure to infer the causal direction between two variables described in Algorithm 1. Underlying data is simulated as $Y = \kappa_1(X, W, \tau) + U \times 1.1$ where $U \sim \mathcal{N}(0, (1 + f(W))^\rho)$, where f is the probability density function of W. U raised to q while keeping its sign and variance. The model where X is causing Y is the correct model.* 500 *Monte Carlo runs.*

				share of decision	
n	τ	ρ	q	correct	false
500	0	0	0.8	0.800	0.200
			1.0	0.474	0.526
			1.2	0.698	0.302
		1	0.8	0.792	0.208
			1.0	0.468	0.532
			1.2	0.708	0.292
	1	0	0.8	0.948	0.052
			1.0	0.930	0.070
			1.2	0.958	0.042
		1	0.8	0.940	0.060
			1.0	0.930	0.070
			1.2	0.938	0.062
1000	0	0	0.8	0.936	0.064
			1.0	0.434	0.566
			1.2	0.846	0.154
		1	0.8	0.934	0.066
			1.0	0.458	0.542
			1.2	0.866	0.134
	1	0	0.8	0.996	0.004
			1.0	0.990	0.010
			1.2	0.988	0.012
		1	0.8	0.990	0.010
			1.0	0.996	0.004
			1.2	0.990	0.010

Table 3.4: *This Table shows Monte Carlo results of the procedure to infer the causal direction between two variables described in Algorithm 1. Underlying data is simulated as* $Y = \kappa_2(X, W, \tau) + U \times 1.1$ *where* $U \sim \mathcal{N}(0, (1 + f(W))^\rho)$*, where* f *is the probability density function of W. U raised to* q *while keeping its sign and variance. The model where* X *is causing* Y *is the correct model.* 500 *Monte Carlo runs.*

				share of decision	
n	τ	ρ	q	correct	false
500	0	0	0.8	0.756	0.244
			1.0	0.428	0.572
			1.2	0.672	0.328
		1	0.8	0.750	0.250
			1.0	0.476	0.524
			1.2	0.688	0.312
	1	0	0.8	0.972	0.028
			1.0	0.954	0.046
			1.2	0.962	0.038
		1	0.8	0.960	0.040
			1.0	0.968	0.032
			1.2	0.948	0.052
1000	0	0	0.8	0.942	0.058
			1.0	0.424	0.576
			1.2	0.838	0.162
		1	0.8	0.936	0.064
			1.0	0.426	0.574
			1.2	0.872	0.128
	1	0	0.8	0.998	0.002
			1.0	0.998	0.002
			1.2	1.000	0.000
		1	0.8	1.000	0.000
			1.0	0.998	0.002
			1.2	1.000	0.000

Table 3.5: *This Table shows Monte Carlo results of the procedure to infer the causal direction between two variables described in Algorithm 1. Underlying data is simulated as* $Y = \kappa_1(X, W, \tau) + U \times 0.89$ *where* $U \sim \mathcal{N}(0, (1 + f(W))^{\rho})$*, where* f *is the probability density function of W. U raised to q while keeping its sign and variance. The model where* X *is causing* Y *is the correct model.* 500 *Monte Carlo runs.*

				share of decision	
n	τ	ρ	q	correct	false
500	0	0	0.8	0.700	0.300
			1.0	0.398	0.602
			1.2	0.654	0.346
		1	0.8	0.748	0.252
			1.0	0.426	0.574
			1.2	0.680	0.320
	1	0	0.8	0.910	0.090
			1.0	0.916	0.084
			1.2	0.886	0.114
		1	0.8	0.934	0.066
			1.0	0.912	0.088
			1.2	0.866	0.134
1000	0	0	0.8	0.908	0.092
			1.0	0.406	0.594
			1.2	0.814	0.186
		1	0.8	0.884	0.116
			1.0	0.472	0.528
			1.2	0.818	0.182
	1	0	0.8	0.980	0.020
			1.0	0.984	0.016
			1.2	0.964	0.036
		1	0.8	0.998	0.002
			1.0	0.988	0.012
			1.2	0.984	0.016

Table 3.6: *This Table shows Monte Carlo results of the procedure to infer the causal direction between two variables described in Algorithm 1. Underlying data is simulated as* $Y = \kappa_2(X, W, \tau) + U \times 0.89$ *where* $U \sim \mathcal{N}(0, (1 + f(W))^\rho)$*, where* f *is the probability density function of W. U raised to q while keeping its sign and variance. The model where* X *is causing* Y *is the correct model.* 500 *Monte Carlo runs.*

				share of decision	
n	τ	ρ	q	correct	false
500	0	0	0.8	0.720	0.280
			1.0	0.444	0.556
			1.2	0.644	0.356
		1	0.8	0.740	0.260
			1.0	0.414	0.586
			1.2	0.642	0.358
	1	0	0.8	0.986	0.014
			1.0	0.996	0.004
			1.2	0.994	0.006
		1	0.8	0.990	0.010
			1.0	0.990	0.010
			1.2	0.996	0.004
1000	0	0	0.8	0.876	0.124
			1.0	0.394	0.606
			1.2	0.824	0.176
		1	0.8	0.904	0.096
			1.0	0.406	0.594
			1.2	0.824	0.176
	1	0	0.8	1.000	0.000
			1.0	1.000	0.000
			1.2	1.000	0.000
		1	0.8	1.000	0.000
			1.0	1.000	0.000
			1.2	1.000	0.000

3.6.3 Linear Models with Additive Non-Gaussian Noise

Shimizu et al. (2006) prove the following theorem (in its multivariate form).

Theorem 3.6.1. *Assume joint distribution $f(X, Y)$ admits the linear model*

$$Y = \alpha X + \varepsilon \text{ with } \varepsilon \perp\!\!\!\perp X, \tag{3.50}$$

then there exists β and a random variable $\tilde{\varepsilon}$ such that

$$X = \beta Y + \tilde{\varepsilon} \text{ with } \tilde{\varepsilon} \perp\!\!\!\perp Y \tag{3.51}$$

if and only if ε and X are Gaussian.

The proof of Theorem 3.6.1 relies on the following results.

Lemma 3.6.2. *Take A and B two independent variables, assume B to be nondeterministic. Then $B \not\perp\!\!\!\perp B + A$ (Peters, 2008).*

Furthermore, we rely on a characterization of Gaussian distributions due to Darmois (1953), Skitovich (1954), and Skitovich (1962) who independently show the following result.

Theorem 3.6.3. *Let $X_1, \ldots, X_d$ be independent, non-degenerate random variables. If there are nonvanishing coefficients ($\forall i, a_i \neq 0 \neq b_i$), $a_1, \ldots, a_d$, and $b_1, \ldots, b_d$, such that the two linear combinations*

$$\begin{aligned} l_1 &= a_1 X_1 + \cdots + a_d X_d \\ l_2 &= b_1 X_1 + \cdots + b_d X_d \end{aligned} \tag{3.52}$$

are independent, then each X_i is normally distributed.

We proceed with the proof of Theorem 3.6.1 (cf. Peters et al., 2017, Appendix C.1).

Proof. ($\Rightarrow$) If X and ε are normally distributed

$$\beta := \frac{Cov(X, Y)}{Cov(Y, Y)} = \frac{\alpha Var(X)}{\alpha^2 Var(X) + Var(\varepsilon)}. \tag{3.53}$$

Define $\tilde{\varepsilon} := X - \beta Y$. $\tilde{\varepsilon}$ and Y are uncorrelated by construction and, since they are jointly Gaussian, it follows that they are independent.

($\Leftarrow$) Assume that

$$\begin{aligned} Y &= \alpha X + \varepsilon, \text{ and} \\ \tilde{\varepsilon} &= (1 - \alpha\beta) X - \beta \varepsilon \end{aligned} \tag{3.54}$$

are independent. Make a case distinction.

1. $(1 - \alpha\beta) \neq 0$, and $\beta \neq 0$

   Since $Y \perp\!\!\!\perp \tilde{\varepsilon}$ by assumption, Theorem 3.6.3 implies that $X \perp\!\!\!\perp \varepsilon$ in the case at hand, cf. eq (3.54). Therefore, $f_{X,Y}(x, y)$ is bivariate Gaussian.

2. $\beta = 0$

   This implies $X \perp\!\!\!\perp \alpha X + \varepsilon$, cf. eq (3.54), which contradicts Lemma 3.6.2.

3. $1 - \alpha\beta = 0$

   This implies $-\beta\varepsilon \perp\!\!\!\perp \alpha X + \varepsilon \perp\!\!\!\perp$, and thus $\varepsilon \perp\!\!\!\perp \alpha X + \varepsilon$ which contradicts Lemma 3.6.2.

This completes the proof. □

Therefore, it is sufficient that ε or X is non-Gaussian for the causal direction to be identifiable (Shimizu et al., 2006).

4 Structural Autonomy and Instrument Validity

This chapter is single-authored, see Burauel (2020). I thank Uri Shalit, Max Schäfer, Johannes König, and participants at the Association for the Advancement of Artificial Intelligence 2019 Spring Symposium 'Beyond Curve Fitting – Causation, Counterfactuals and Imagination-Based AI', the 2019 Barcelona GSE Summer Forum 'Machine Learning for Economics' workshop, and the DIW SOEP Applied Panel Analysis Brownbag Seminar for their comments.

4.1 Introduction

Concerns about unobserved confounding, which can invalidate estimates of causal effects, are widespread in non-experimental studies in economics and beyond. To estimate a causal effect in spite of such problems, a common solution is to resort to instrumental variable (IV) approaches. Researchers often rely on institutional knowledge or a policy change to justify the strong assumptions that an IV must fulfill to identify the sought causal effect. Such justifications are seldom rigorously data-driven and often controversial. For instance, in a study on the causal effect of economic development on democracy, Acemoglu et al. (2008) use changes in past savings rates to instrument for income. They argue that, "it seems plausible to expect that changes in the savings rate over periods of five to ten years should have no direct effect on the culture of democracy, the structure of political institutions, or the nature of political conflict within society" (p. 822). That *plausability* argument is augmented by controlling for a number of additional covariates and checking whether the coefficient of interest changes. Yet, this is shown to be an uninformative procedure in observational studies (Oster, 2019).[1] This underscores the necessity to develop and make accessible statistical tests that can falsify critical IV assumptions.

Such assumptions are difficult to evaluate since they involve unobservable quantities. Nevertheless, the assumed causal structure in IV models implies testable constraints on

[1] The authors employ an overidentification test, which *assumes* validity of at least one instrument and, therefore, is rather mute.

the outcome distributions of four groups of individuals defined by two observed quantities, treatment status and instrument assignment. These are described by Balke and Pearl (1997) and leveraged first by Kitagawa (2015) to propose a test for instrument validity. Kitagawa (2015) writes that these testable implications are "optimal," in the sense that "any other feature of the data distribution cannot contribute to screening out invalid instruments" (p. 2048). This paper provides a test that can detect invalid instruments that does not rely on Balke and Pearl's testable implications and, therefore, shows that other features of the data distribution do contain evidence to identify invalid instruments.

I argue that an instrumental variable that violates either the exclusion restriction or the exchangeability assumption (such instruments are, henceforth, called invalid; vice versa, an instrument fulfilling both these assumptions is called valid) implies a biased treatment effect estimate. My approach builds on testing whether the instrument induces a biased treatment effect estimate to infer whether the instrument is invalid. In particular, I build on work by Janzing and Schölkopf (2018, JS henceforth), who show how to measure the extent to which an observed statistical relationship in multivariate linear models is due to confounding or genuine causation. Whereas Kitagawa (2015) relies on restrictions of the outcome distribution of groups implied by the interaction of treatment and instrument variables to test IV validity, my test relies on the genericity of the estimated parameter vector w.r.t. the covariance matrix of independent variables. By applying the methodology laid out in JS to the problem of testing IV validity, this paper constitutes a bridge between the surging literature on causal modeling in the machine learning community and traditional econometric problems (see Peters et al., 2017, for an overview of the former). I apply the proposed test to data by Card (1995) to show its feasiblity in practice.

The paper is structured as follows. In Section 4.2, I provide an overview of the literature. In Section 4.3, I discuss the Principle of Independent Mechanisms that underlies the methodology to measure degree of confounding in multivariate linear models proposed by Janzing and Schölkopf (2018), which I also introduce on an intuitive level. This paper, in turn, uses that methodology in the construction of an instrument validity test. In Section 4.4, I describe the IV model to be analyzed, discuss assumptions and links to the potential outcomes framework. In Section 4.5, I present the test for instrument validity. In Section 4.6, I present results of Monte Carlo simulations. In Section 4.7, I provide an empirical application. In Section 4.8, I discuss limitations and future extensions. Finally, in Section 4.9, I conclude. The Appendix provides, *inter alia*, a detailed discussion of Janzing and Schölkopf (2018) (Appendix 4.10.1) and a discussion of the historic origins of the Principle of Independent Mechanisms (Appendix 4.10.6).

4.2 The literature

The Sargan (1958)-Hansen (1982) J-test for overidentifying restrictions arguably spawned the substantial literature on specification testing in instrumental variable (IV) models. The J-test can be used to test instrument validity when there are more instruments than endogenous regressors. Conditional on the assumption that at least one instrument is valid, the test can help decide whether *all* instruments are valid. However, the test is not able to detect a situation in which all instruments are invalid.

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 that does not rely on non-parametric IV estimation (which often suffers from slow convergence that, in turn, results in low power of such tests). Two related papers that both study nonparametric IV models are Breunig (2015) and Gagliardini and Scaillet (2017). The former uses series estimators to propose a test for instrument exogeneity and the latter employ a Tikhonov Regularized estimator of the functional parameter to minimize the distance criterion corresponding to the moment conditions. Breunig (2020) extends these results to nonparametric quantile regression with nonseparable structural disturbances. In broad terms, what unites many of these papers is their reliance on testing whether the moment conditions implied by the instrumental variable model are fulfilled. By analyzing higher-order moments as well, these models can resort to overidentifying restrictions even when there is only one instrument per endogenous variable.

Although diverse methods to test the exclusion restriction in *overidentified* IV models are 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 instrument validity 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 (1997) and Heckman and Vytlacil (2005). These imply constraints on the outcome distributions of groups defined by the interaction of their observed treatment and instrument status (denoted T_i and Z_i respectively). In particular, if the outcome distributions of individuals with $Z_i = 1,\ T_i = 0$ and $Z_i = 0,\ T_i = 0$ or those of individuals with $Z_i = 1,\ T_i = 1$ and $Z_i = 0,\ T_i = 1$ intersect, instrument validity is violated. More specifically, among treated individuals, the outcome density of those who have received the instrument ($Z_i = 1$) should lie above that of those who have not ($Z_i = 0$). Conversely, among control individuals, the outcome density of those who have received the instrument ($Z_i = 1$) should lie below that of those who have not ($Z_i = 0$). Huber and Mellace (2015) and Mourifié and Wan (2017) provide closely related extensions to Kitagawa (2015). The former propose a similar test that relies on mean potential outcomes rather than their distributions. Mourifié and Wan (2017) build on Kitagawa (2015) by representing his test in terms of moment inequalities conditional on additional covariates.

The test proposed in this paper builds on a strand of the literature on the identification of causal signals in non-experimental data. The underlying idea, which goes back to Haavelmo (1944), is that invariance structures in observed data justify statements about the underlying causal structure of the system under study. The idea that invariant structures are informative about causal structure is formalized from an information-geometric perspective as the Principle of Independent Mechanisms (PIM) (Janzing et al., 2012; Peters et al., 2017). Janzing and Schölkopf (2018) show that traces of violations of PIM can be discerned in the spectral measures of variance-covariance matrices in the presence of unobserved confounding. In this paper, I show how this reasoning can be employed to analyze instrument validity.

Thus, the main contribution of this paper is to develop a novel testing approach that relies neither on moment restrictions nor on the potential outcomes framework. Instead, my approach is based on the decomposition of the spectral measure of the covariates' covariance matrix induced by the corresponding parameter vector. Since the present work is based on the Principle of Independent Mechanisms, it adds to the growing literature using this principle as a powerful concept to guide causal identification (Peters et al., 2016; Besserve et al., 2017; Besserve et al., 2018).

4.3 The Janzing-Schölkopf Methodology

4.3.1 The Principle of Independent Mechanisms and Generic Orientation

The Principle of Independent Mechanisms (PIM) underlies many contributions to causal inference from the machine learning community (for an overview see Schölkopf, 2019).[2] It also serves as the basis for the test proposed in this paper. The notion goes back to Haavelmo and Frisch, who identified the search for and analysis of 'autonomous relations' as the ultimate goal of econometrics (see Appendix 4.10.6 for a brief historical overview). Despite considering it an important guiding principle, they did not employ the notion of autonomy as an empirical identification technique as such. In fact, Frisch and Haavelmo

[2] In its bivariate incarnation, the principle is referred to as Independence between Cause and Mechanism (ICM), see Appendix 4.10.6 for an illustration.

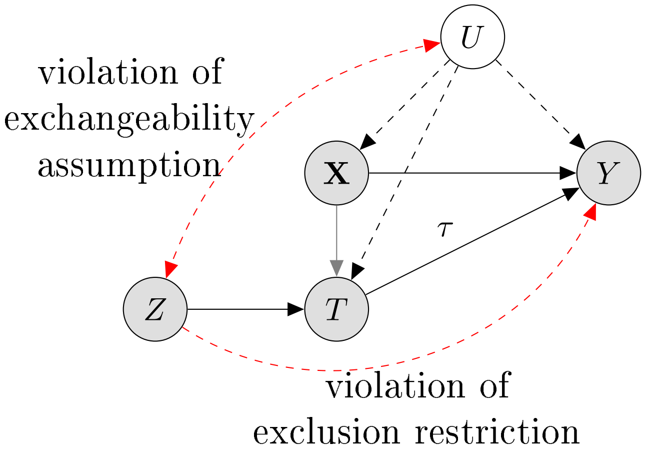


Figure 4.1: *Graphical representation of the IV model under study. T represents the binary treatment variable of interest. The red arrow from Z to Y indicates how the exclusion restriction can be violated by Z's direct effect on Y. The double-edged arrow between U and Z indicates how the exchangeability assumption can be violated when there is an unobserved confounder influencing both Z and Y. 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.*

argued that the autonomous nature of mechanisms cannot be identified empirically but must be motivated by (economic) theory. Although not all issues that preclude its use as an identification tool have been resolved, some important advances have been made in the first two decades of the twenty-first century. The proposal by Janzing and Schölkopf (2018) to estimate the degree of confounding in multivariate linear models, which is motivated by the notion of autonomy, is an example of this progress.

To illustrate the idea, consider a set of random variables $\{V_1, \ldots, V_n\}$ whose causal relations can be represented in a directed acyclic graph (DAG) and an accompanying structural equation model (Pearl, 2009). The joint probability distribution that is consistent with the causal structure given in the DAG can be factorized as

$$P(V_1, \ldots, V_n) = \prod_{j=1}^{n} P(V_j | Pa(V_j)) \tag{4.1}$$

where $Pa(V_j)$, the *parents* of V_j, denotes the set of random variables that causally influence V_j. Naturally, there are many other types of factorizations of the joint distribution:

$$P(V_1, \ldots, V_n) = \prod_{j=1}^{n} P(V_j | V_{j+1}, \ldots, V_n). \tag{4.2}$$

However, only the conditionals in eq. (4.1) are independent of each other (in a sense made precise below) and, therefore, represent causal mechanisms that translate causes (or parents, $Pa(V_j)$) into their effects (children, V_j). Causes that do not have parents in the model under investigation appear as marginal distributions in this formulation. Using algorithmic information theory, Janzing and Schölkopf (2010) and Lemeire and Janzing (2013) show that the conditionals on the right-hand-side are algorithmically independent of each other if the DAG represents the causal structure. Intuitively, knowing about one of the mechanisms does not provide any information about other mechanisms. In this sense each of the mechanisms operates independently of the others. Since the formal deduction of the mechanism's algorithmic independence relies on the theoretical notion of Kolmogorov complexity that is uncomputable, it is not obvious how to conceive of the independence of mechanisms in practice. Thus, the algorithmic independence of mechanisms amounts less to a precise recipe for uncovering autonomous relations in observational data than to a rigorous guiding principle to design algorithms that do.

To make the notion of 'independent mechanisms' practically relevant, what precisely is meant by 'independence' must be defined in a way that allows data-driven quantification. Janzing and Schölkopf (2018) propose such a feasible interpretation of the Principle of Independent Mechanisms. Moreover, they show a way to measure the degree of violation of PIM in observational data. This degree of violation is a measure of confounding in multivariate linear models. I spend the rest of this section discussing their notion of independence and how they can infer a measure of confounding. Though this is not an exhaustive discussion, the technical details are provided in Appendix 4.10.1, which reproduces the arguments in JS as a courtesy to the reader.

To illustrate the proposal by Janzing and Schölkopf (2018), consider an illustrative multivariate linear model $Y = \mathbf{X}\beta + \varepsilon$ and suppose that there is no unobserved confounding such that multidimensional $\mathbf{X}$ is causing Y and the least-squares estimate of β is unbiased. β is the crucial parameter representing the 'mechanism' that translates the causes $\mathbf{X}$ into effect Y. The causes, in turn, are represented by the covariance matrix of the right-hand-side variables $\Sigma_{\mathbf{XX}}$. Independence between β and $\Sigma_{\mathbf{XX}}$ amounts to β lying in a generic orientation with respect to $\Sigma_{\mathbf{XX}}$. To give a counterexample, a vector aligning with the first eigenvector of $\Sigma_{\mathbf{XX}}$ would not lie in a generic orientation with respect to $\Sigma_{\mathbf{XX}}$.

To recap, what the PIM implies on an intuitive level is that the *mechanism* translating cause into effect, represented by the true parameter vector, and the *input to the mechanism* or *cause*, represented by $\Sigma_{\mathbf{XX}}$, should be 'independent'. JS make the concept of 'independence' operational by arguing that, if PIM 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, $\Sigma_{\mathbf{XX}}$. In technical terms, 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 equal to the (unweighted) tracial spectral measure of $\Sigma_{\mathbf{XX}}$. Technical details are presented in Appendix 4.10.1.

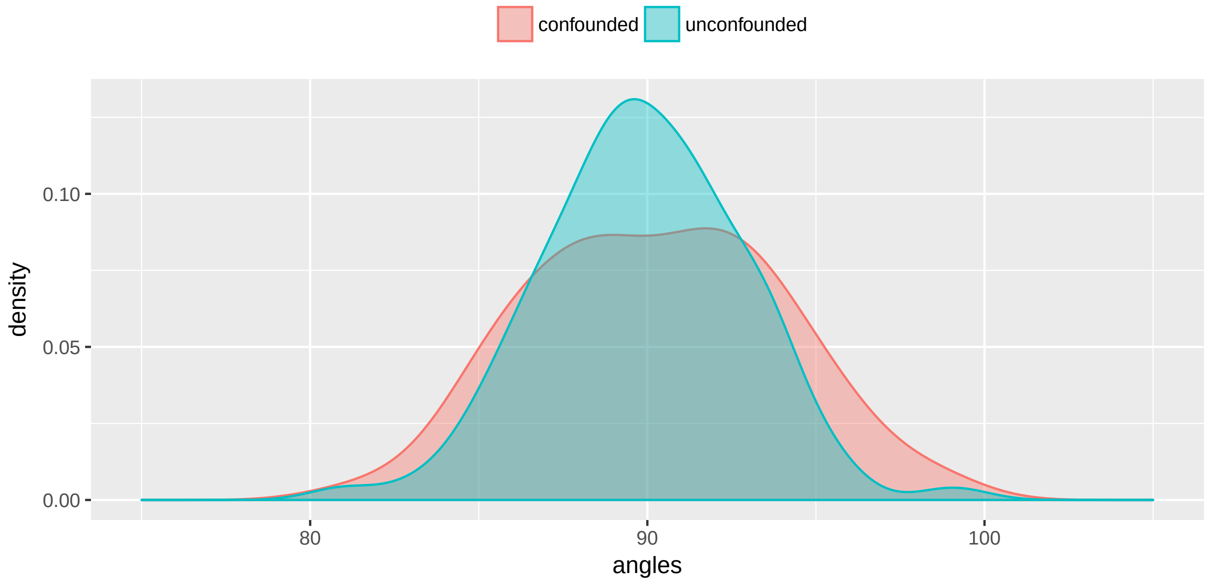


Figure 4.2: Illustration of genericity of vectors. *This Figure shows density plots of the angles between the least-squares parameter vector of both confounded and unconfounded models with each of the d eigenvectors of the covariance matrix of the covariates. In the unconfounded model, the least-squares parameter vector should lie in generic orientation with respect to (the eigenspace spanned by the) eigenvectors of the covariance matrix of the covariates. Genericity of two vectors can be understood as their dot product being zero or their angle being 90 degrees. Thus, as expected, the distribution of angles in the unconfounded case clusters around 90 degrees. Crucially, in the confounded case, the distribution of angles is considerably wider. Thus, a trace of confounding is reflected in the less generic angles of the confounded parameter vector w.r.t. the eigenvectors; their distribution is characterized by a more frequent divergence from the generic angle of 90 degrees. This illustrates the type of confounding signal that Janzing and Schölkopf (2018) leverage in their methodology. Details on the simulation setting is found in Appendix 4.10.7; here I set* $d = 100$*, and* $n = 50000$.

I now provide a graphical illustration of the traces that a violation of PIM leaves in purely observational data. I simulate data from a confounded and an unconfounded model, then compute the estimated parameter vector in each case (see Appendix 4.10.7 for details on the simulation). In the unconfounded case, the estimated parameter vector represents genuine causes and is not biased due to unobserved confounding. Following JS, that true parameter vector should lie in generic orientation w.r.t. the eigenvectors of the covariance matrix. Two vectors lie in generic orientation w.r.t. each other if their dot product is zero (or the angle they span is ninety degrees). At first glance orthogonality seems like a specific, not generic, relation between any two vectors. However, it is important to note that such genericity is a high-dimensional phenomenon: the angle between two randomly drawn vectors approaches ninety degrees as the their dimensionality increases (see e.g. Gorban and Tyukin, 2018). This is also why the asymptotic results in JS rely on the dimensionality of the covariate space going to infinity. Intuitively, two generic vectors do not share any information since they are pointing in two orthogonal directions.

Therefore, I compute the angle between the estimated parameter vector and each of the eigenvectors of the normalized covariance matrix of the covariates for both the confounded and unconfounded setting.[3] For both settings, I simulate data for $d = 100$ dimensions and $n = 50000$ observations. Then, I plot the resulting distribution of angles between d eigenvectors and the least-squares estimate $\hat{\beta}$. Figure 4.2 plots these distributions for one draw of the data; Figure 4.9 in Appendix 4.10.7 plots the same information for 100 draws of the data. The distribution of angles for both settings centers around ninety degrees, which is not surprising given our simulation setting. Crucially, one can see that the distribution of angles is more widespread for the confounded setting. Consequently, in the presence of confounding the estimated parameter vector lies in a less generic direction w.r.t. the eigenvectors of the covariance matrix. This deviation from genericity is what Janzing and Schölkopf (2018) exploit to measure the degree of confounding. Although the modeling assumptions underlying both this graphical depiction as well as the JS methodology in general are idealized, they do provide useful insight into the elusive nature of unobserved confounding (see also Section 4.8).

4.3.2 Estimating the degree of confounding

I have presented an intuitive understanding that the orientation of a parameter vector w.r.t. the eigenspaces of the corresponding $\Sigma_{\mathbf{XX}}$ contains a confounding signal. JS propose a method to measure deviations from the generic orientation to estimate the *degree of confounding* in multivariate linear models. Technically, generic orientation is instantiated as the equivalence of two spectral measures of $\Sigma_{\mathbf{XX}}$: first, the unweighted spectral measure (called tracial spectral measure and denoted $\mu^{Tr}_{\Sigma_{\mathbf{XX}}}$), and second, the spectral measure weighted by a vector such as a parameter vector β (called vector-induced spectral measure

[3] For the purpose of illustration, I depart slightly from JS here. I compute the genericity of the estimated parameter vector for every eigenvector *in isolation*. However, JS postulate a generic orientation w.r.t. the eigenspace spanned by the collection of eigenvectors. In other words, they jointly consider the whole set of eigenvector-eigenvalue pairs. Technically, they consider the distribution of eigenvalues weighted by the estimated parameter vector. See Appendix 4.10.1.

denoted $\mu_{\Sigma_{\mathbf{XX}},\beta})$[4]:

$$\text{generic orientation of } \beta \text{ w.r.t. } \Sigma_{\mathbf{XX}} \Leftrightarrow \mu^{Tr}_{\Sigma_{\mathbf{XX}}} \simeq \mu_{\Sigma_{\mathbf{XX}},\beta}. \tag{4.3}$$

Ideally, one would check whether the spectral measure induced by the estimated parameter vector is equivalent to that induced by the true parameter vector. However, the latter is not estimable from observed data. Nevertheless, the equivalence of the tracial spectral measure and that induced by the true parameter vector makes it possible to compare the spectral measure induced by the estimated, and possibly biased, vector to the estimable tracial spectral measure to infer a degree of confounding.

The crucial result in JS is that the computable spectral measure induced by the estimated (and possibly biased) parameter vector $\mu_{\Sigma_{\mathbf{XX}},\hat{\beta}}$ can be decomposed into one part that is due to confounding and a second part that represents genuine causation. More specifically, $\mu_{\Sigma_{\mathbf{XX}},\hat{\beta}}$ can be decomposed into the spectral measure induced by the true parameter vector and that induced by the bias of the estimated parameter vector from the true parameter vector. The relative sizes of these two components define the degree of confounding κ:

$$\mu_{\Sigma_{\mathbf{XX}},\hat{\beta}} \simeq (1-\kappa)\,\mu_{\Sigma_{\mathbf{XX}},\beta} + \kappa\,\mu_{\Sigma_{\mathbf{XX}},(\hat{\beta}-\beta)}. \tag{4.4}$$

κ ranges from 0 (no confounding) to 1 (observed statistical relation is fully due to confounding). Without confounding,

$$\mu_{\Sigma_{\mathbf{XX}},\hat{\beta}} \simeq \mu_{\Sigma_{\mathbf{XX}},\beta} \simeq \mu^{Tr}_{\Sigma_{\mathbf{XX}}}, \tag{4.5}$$

i.e. $\hat{\beta}$ is generically oriented.

Still, $\mu_{\Sigma_{\mathbf{XX}},\beta}$ and $\mu_{\Sigma_{\mathbf{XX}},(\hat{\beta}-\beta)}$ are uncomputable since they involve the unknown true β. However, the computable $\mu_{\Sigma_{\mathbf{XX}},\hat{\beta}}$ can be parameterized by a two-parametric family of probability measures. The algorithm proposed by JS finds those two parameter values that minimize the distance between the two-parametric estimate and the observed spectral measure induced by the estimated (and possibly) biased parameter vector. One of the parameters is κ.

In the remainder of this paper, I take their method as given and show how it can be employed as a workhorse in testing instrument validity. A detailed description of the procedure to estimate the confounding strength κ is available in Appendix 4.10.1.

[4] I use $\simeq$ in this and the following expressions in this subsection to indicate that the following statements are not precise in the sense that I do not explicitly state the types of and rates of convergence as well as conditions for convergence. See Appendix 4.10.1 for details.

4.4 Model and assumptions

Building on Angrist et al. (1996), I consider the following latent index model,

$$Y = \mathbf{X}\beta + \tau T + \varepsilon_Y \tag{4.6}$$

$$T^* = \mathbf{X}\beta_T + \alpha Z + \varepsilon_T \tag{4.7}$$

$$\text{with} \quad T = \begin{cases} 1 & \text{if} \quad T^* > 0 \\ 0 & \text{if} \quad T^* \leq 0. \end{cases} \tag{4.8}$$

$\mathbf{X}$ represents a set of d covariates, T the binary treatment indicator, and Z a binary instrument. β and β_T are d-dimensional vectors and α a one-dimensional vector of coefficients. τ is the causal effect of interest. ε_Y and ε_T are unobserved structural errors. Y is the outcome variable of interest. The difference between the model considered here and Angrist et al. (1996) is the presence of covariates $\mathbf{X}$.

Though I am considering a binary instrument in the paper at hand, I do not require it in the theoretical development of the test, i.e. it can also be applied to in settings with continuous instruments.

In this structural model, the object of interest is the true causal parameter τ. In structural models, such parameters are also referred to as deep parameters, i.e. those that are policy-invariant. Since the JS methodology effectively measures the degree of autonomy of observed statistical relations, it is natural to theoretically embed the proposed instrument validity test in a structural model framework such as eqs. (4.6)-(4.8).

The endogeneity problem is caused by a potential dependence of the structural error terms. If $Cov(\varepsilon_Y, \varepsilon_T) \neq 0$, then $Cov(T, \varepsilon_Y) \neq 0$ and, consequently, T is not exogenous and a naive estimate of τ will be biased. An instrumental variable Z that fulfills the following assumptions identifies τ and can be used to estimate it consistently.

Assumption 4.4.1. *The instrument Z correlates neither with ε_Y nor with ε_T:*

$$Cov(Z, \varepsilon_Y) = 0 \textit{ and } Cov(Z, \varepsilon_T) = 0. \tag{4.9}$$

Assumption 4.4.2. *The instrument Z correlates with T:*

$$Cov(Z, T) \neq 0. \tag{4.10}$$

With these assumptions I can provide a definition of IV validity:

Definition 4.4.1. A variable Z is called a valid instrumental variable if if fulfills Assumptions 4.4.1 and 4.4.2. Vice versa, an invalid instrumental variable does not fulfill either Assumption 4.4.1 or 4.4.2.

Under Assumptions 4.4.1 and 4.4.2, the instrumental variable can be used to estimate τ consistently, for instance by two-stage least squares (see Wooldridge, 2002). The underlying idea of the method detailed in Section 4.5 is to evaluate instrument validity by checking whether, after instrumenting T with Z, the estimated τ is still biased. If it is, the instrument is invalid. Before turning to a description of the test idea, I compare the structural model approach to the potential outcomes framework.

Comparison to the potential outcomes framework. I will now briefly discuss standard IV assumptions in the potential outcomes framework (PO), which is popularized in its modern form by Rubin (1974) and Holland (1986) and widely-used in empirical practice. I draw comparisons to the structural framework. Ultimately, the goal is to compare my test to the one proposed by Kitagawa (2015), who relies on the PO framework. Therefore, it is important to show that the two frameworks can be shown to estimate the same object of interest, namely τ, under certain assumptions.

Unlike the structural approach, the potential outcomes framework relies on positing a set of individual-specific potential outcomes as a function of instrument and treatment assignment: each individual i has potential outcomes $Y_i(\mathbf{Z}, \mathbf{T})$ where $\mathbf{Z}$ and $\mathbf{T}$ are vectors of potential instrument and treatment assignments. Some of these are by definition not observable.

A major argument in favour of the PO framework is the intuitive interpretation of crucial IV assumptions since they are not formulated in terms of unobserved structural error terms. Specifically, Assumption 4.4.1 can be disentangled into the exclusion restriction and the random assignment of treatment or exchangeability assumption (Angrist et al., 1996). The exclusion restriction,

$$Y(\mathbf{Z}, \mathbf{T}) = Y(\mathbf{Z}', \mathbf{T}) \quad \forall \quad \mathbf{Z},\ \mathbf{Z}',\ \mathbf{T}, \tag{4.11}$$

states that the instrument Z must only have an influence on Y through its effect on T and not directly. This assumption is reflected in the absence of Z in eq. (4.6) and $Cov(Z, \varepsilon_Y) = 0$ in Assumption 4.4.1.

The exchangeability assumption

$$\Pr(\mathbf{Z} = \mathbf{c}) = \Pr(\mathbf{Z} = \mathbf{c}'), \tag{4.12}$$

where $\mathbf{c}$ and $\mathbf{c}'$ are vectors of instrument assignments, states that instrument assignment is random. It is reflected in $Cov(Z, \varepsilon_T) = 0$ in Assumption 4.4.1.

The PO assumption corresponding closely to Assumption 4.4.2 is that Z needs to have some effect on the probability of treatment,

$$\mathbb{E}(D_i(T_i = 1) - D_i(T_i = 0)) \neq 0. \tag{4.13}$$

Furthermore, monotonicity must be assumed. The monotonicity assumption

$$D_i(Z_i = 1) \geq D_i(Z_i = 0) \tag{4.14}$$

states that there are no individuals who would opt to take the treatment ($T_i = 1$) if they are not induced to do so ($Z_i = 0$) but would choose not to take the treatment ($T_i = 0$) if induced to do so ($Z_i = 1$); there are no so-called defiers. There is no direct equivalent of this assumption in the structural framework. However, the assumption is implicitly fulfilled because α is not individual-specific in eq. (4.7).[5]

In the PO framework, the causal effect is defined as the difference in potential outcomes:

$$\tau^{PO} := \mathbb{E}(Y_i(T_i = 1) - Y_i(T_i = 0)). \tag{4.15}$$

In an IV setting, this treatment effect is identified for the subgroup of compliers, i.e. those individuals who can be induced ($Z_i = 1$) to take the treatment ($T_i = 1$) and who would not take the treatment ($T_i = 0$) if not induced ($Z_i = 0$), and is called the local average treatment effect (LATE).

The potential outcomes $Y_i(T_i)$ are implicitly defined in a structural model such as in eqs. (4.6)-(4.8) if one supposes that the potential outcome is a linear function of the treatment and control covariates (see also Imbens, 2014a). I make this assumption to compare the structural and the PO framework.

Assumption 4.4.3. *The potential outcome $Y_i(T_i)$ is a linear function of treatment and control variables:*

$$\mathbb{E}(Y_i(T_i)|\mathbf{X}_i) = \mathbf{X}_i\beta + \tau T_i. \tag{4.16}$$

Furthermore, I make explicit the following assumption that is implicit in eq. (4.6):

Assumption 4.4.4. *The treatment effect τ is constant.*

Under these additional assumptions, the treatment effect as it is typically defined in

[5] In addition, the stable unit treatment value assumption (SUTVA) must be fulfilled. I can take this for granted for the purposes at hand and do not discuss it further.

the potential outcomes framework coincides with the deep parameter τ:

$$\begin{aligned}\tau^{\text{PO}} &= \mathbb{E}(Y_i(T_i = 1) - Y_i(T_i = 0)) \\ &= (\mathbf{X}\beta + \tau) - (\mathbf{X}\beta) \\ &= \tau.\end{aligned}$$

Making Assumptions 4.4.3 and 4.4.4 enables me to compare my approach to existing IV validity tests that are based on the PO framework. Moreover, Angrist et al. (1996) state that "pooling [the exclusion restriction and exchangeability assumption] into the single assumption of zero correlation between instruments and disturbances [Assumption 4.4.1] has led to confusion about the essence of the identifying assumptions and hinders assessment and communication of the plausibility of the underlying model" (p. 450). Following this argument, I introduce violations of IV validity in the form of violations of exclusion restriction and exchangeability assumption in the Monte Carlo studies that follow. This facilitates the interpretability of these violations.

4.5 Test for Instrument Validity

In this section, I describe the test for instrument validity in a step-by-step manner.

4.5.1 Reduced form model and connection to Janzing Schölkopf

It is useful to reformulate the model in eqs. (4.6)-(4.8) in its reduced form to explicitly show how Janzing and Schölkopf (2018) can be applied to the problem of instrument validity. Following Wooldridge (2002), the treatment variable T is instrumented by two-stage least squares with the instrument Z. I call the instrumented treatment variable $\hat{T}$. Then, the resulting reduced form is

$$Y = \{\mathbf{X}, T_{\text{instrumented}}\} \begin{pmatrix} \beta \\ \tau \end{pmatrix} + cu + \varepsilon \tag{4.17}$$

$$\{\mathbf{X}, T_{\text{instrumented}}\} = \mathbf{E} + u \begin{pmatrix} \mathbf{b} & b_\tau \end{pmatrix} \tag{4.18}$$

where $\{\mathbf{X}, T_{\text{instrumented}}\}$ denotes a matrix of control variables $\mathbf{X}$ and the instrumented treatment variable, which we will also refer to as $\hat{T}$ in the following. β is a d-dimensional parameter vector, τ the true causal effect of interest. ε is a reduced form error. u is an unobserved confounder, which influences Y when $c \neq 0$ and $\{\mathbf{X}, \hat{T}\}$ when $\begin{pmatrix} \mathbf{b} & b_\tau \end{pmatrix} \neq \mathbf{0}$. Like Janzing and Schölkopf (2018), I use u to parameterize confounding in this model. $\mathbf{E} = \{\mathbf{X}^*, \hat{T}^*\}$ represents hypothetical unconfounded versions of $\mathbf{X}$ and $\hat{T}$. Confounding is introduced by adding $u\begin{pmatrix}\mathbf{b} & b_\tau\end{pmatrix}$. Specifically, each element of the vector $\begin{pmatrix}\mathbf{b} & b_\tau\end{pmatrix} = \begin{pmatrix}b_1 & \dots & b_d & b_\tau\end{pmatrix}$ parameterizes the confounding of the corresponding dimension of $\{\mathbf{X}, \hat{T}\}$, e.g. $X_1 = E_1 + ub_1$.

The level of confounding that JS make it possible to estimate is defined as

$$\kappa := \frac{\left\| c\Sigma^{-1}_{\mathbf{X}\hat{T}} \begin{pmatrix}\mathbf{b} \\ b_\tau\end{pmatrix} \right\|^2}{\left\| \begin{pmatrix}\beta \\ \tau\end{pmatrix} \right\|^2 + \left\| c\Sigma^{-1}_{\mathbf{X}\hat{T}} \begin{pmatrix}\mathbf{b} \\ b_\tau\end{pmatrix} \right\|^2} \tag{4.19}$$

where $\Sigma_{\mathbf{X}\hat{T}}$ is the covariance matrix of $\{\mathbf{X}, \hat{T}\}$ and $c\Sigma^{-1}_{\mathbf{X}\hat{T}} \begin{pmatrix}\mathbf{b} \\ b_\tau\end{pmatrix} = \begin{pmatrix}\hat{\beta} - \beta \\ \hat{\tau} - \tau\end{pmatrix}$, i.e. the deviation of the least-squares parameter vector $\begin{pmatrix}\hat{\beta} \\ \hat{\tau}\end{pmatrix}$ from its true values $\begin{pmatrix}\beta \\ \tau\end{pmatrix}$. Therefore, κ is the deviation of $\begin{pmatrix}\hat{\beta} \\ \hat{\tau}\end{pmatrix}$ from $\begin{pmatrix}\beta \\ \tau\end{pmatrix}$ relative to the sum of squared length of $\begin{pmatrix}\hat{\beta} \\ \hat{\tau}\end{pmatrix}$. See Appendix 4.10.1 for a discussion of this methodology, and in particular eq. (4.53) for a motivation of that interpretation of κ.

By quantifying the average bias of the estimated parameter vectors from their true values, κ gives an overall degree of confounding of the whole model. However, I am interested in the bias of $\hat{\tau}$. Therefore, building on JS, I propose a way to estimate confounding of a single parameter. I do this by estimating a counterfactual degree of confounding κ_s that would be obtained if $\hat{T}$ were unconfounded. Then, I compare this counterfactual to the actual degree of confounding observed. The following description of the algorithm I propose shows how I leverage JS to achieve that.

4.5.2 Test Procedure

The test procedure is succinctly described in Algorithm 2. In the following main text I provide a description that focuses on the intuition behind the procedure.

1. Normalize the data such that all variables have the same mean and variance as the treatment indicator T.

2. Instrument T with Z using two-stage least squares. Call the instrumented treatment variable $\hat{T}$.

3. Generate a synthetic variable T_s that has the same covariance structure to $\mathbf{X}$ as does $\hat{T}$, i.e. T_s satisfies
$$Cov(\mathbf{X}, T_s) = Cov(\mathbf{X}, \hat{T}).$$
See Algorithm 3 for details on how to construct T_s.[6]

4. Estimate $\kappa_s := \kappa(\{\mathbf{X}, T_s\}; Y)$ following JS.[7]

T_s is a synthetically generated variable that does not have a causal effect on Y and is, conditionally on $\mathbf{X}$, uncorrelated with the unobserved error while having the same covariance structure to $\mathbf{X}$ as $\hat{T}$. Intuitively, κ_s measures the counterfactual overall degree of confounding of the model that would be obtained if the instrument were valid and $\hat{T}$ unconfounded. Thus, κ_s is an important component to which the actual degree of confounding that is estimated in the following step can be compared to evaluate instrument validity.

5. Estimate $\kappa_i := \kappa(\{\mathbf{X}, \hat{T}\}; Y)$ following JS.

6. Calculate $\delta := \kappa_i - \kappa_s$.

Intuitively, if κ_i is larger than κ_s, i.e. $\delta > 0$, instrumenting leads to a level of confounding of the model that is larger than would be obtained if the instrument were valid. Therefore, $\delta > 0$ is evidence for an invalid instrument.

7. To incorporate uncertainty about these metrics in the subsequent decision, bootstrap over steps 3-6 above. For each bootstrap sample $b \in \{1, \ldots, B\}$ calculate
$$\delta_b = \kappa_i - \kappa_s \tag{4.20}$$

8. Calculate the share of samples with $\delta_b \leq 0$,
$$\delta_B^{\mathbb{1}} = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}(\kappa_{i,b} \leq \kappa_{s,b}), \tag{4.21}$$
$\delta_B^{\mathbb{1}}$ can be interpreted as a pseudo-p-value for the hypothesis
$$H_0 : (\text{the instrument is valid}) \Rightarrow \kappa_i \leq \kappa_s \tag{4.22}$$

[6] I thank participants at CrossValidated for insightful discussions on how to generate such a variable T_s. Algorithm 3 is based on the answer by `whuber` to be found at `https://tinyurl.com/syntheticT`.

[7] I use the code provided by Janzing and Schölkopf to estimate κ.

against the alternative

$$H_1 : \kappa_i > \kappa_s \Rightarrow \text{(the instrument is invalid)}. \tag{4.23}$$

For a proof of the statements (the instrument is valid) $\Rightarrow \kappa_i \leq \kappa_s$ and $\kappa_i > \kappa_s \Rightarrow$ (the instrument is invalid) see proof of Theorem 4.10.1 and Corollary 4.10.1.1 in Appendix 4.10.2

9. Finally, I propose the following decision rule:

$$\psi_\delta(\alpha) = \mathbb{1}(\delta_B^{\mathbb{1}} \leq \alpha) = \begin{cases} 1 & \Longrightarrow \text{reject } H_0 \\ 0 & \Longrightarrow \text{do not reject } H_0 \end{cases} \tag{4.24}$$

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

There is a caveat to the way the synthetic T_s is generated. T_s is uncorrelated with the structural error conditional on $\mathbf{X}$ and not causally related to Y. In other words, the true causal effect of T_s is equal to zero. Therefore, κ_s measures the degree of confounding that would be obtained if the instrument were valid *and* $\tau = 0$. However, this caveat does not affect the validity of Theorem 4.10.1.

```
Data: sample of the outcome variable, control covariates, treatment indicator,
      and instrumental variable $\mathcal{D} = \{Y_i, \mathbf{X}_i, T_i, Z_i\}_{i=1}^{n}$
Input: data $\mathcal{D}$, threshold value $\alpha$, number of bootstraps $B$
Output: pseudo-$p$-value and rejection decision $\psi(\alpha)$ for the hypothesis
        $H_0 :$ $Z$ is a valid instrument
Normalize data such that all variables have mean zero and variance equal to
  $Var(T)$
Implement two-stage least squares IV approach: regress $Z$ on $\{\mathbf{X}, T\}$ and call
  resulting parameter vector $\beta_{IV}$, calculate prediction $\hat{T} = \{\mathbf{X}, T\}\beta_{IV}$
for $b = 1$ to $B$ do
    Draw a bootstrap sample $\mathcal{D}_b$ of size $n$ with replacement
    Generate synthetic variable $T_s$ based on $\mathcal{D}_b$ by following Algorithm 3
    Estimate $\kappa_s := \kappa(\{\mathbf{X}, T_s\}; Y)$ based on $\mathcal{D}_b$ following JS
    Estimate $\kappa_i := \kappa(\{\mathbf{X}, \hat{T}\}; Y)$ based on $\mathcal{D}_b$ following JS
    Calculate $\delta_b = \kappa_i - \kappa_s$
end
Calculate the pseudo-$p$-value

    $$p = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}(\delta_b \leq 0)$$

Decide whether to reject $H_0$: $\psi(\alpha) = \mathbb{1}(p \leq \alpha)$
```

Algorithm 2: Test for instrument validity

```
Data: $n \times d$ matrix of covariates $\mathbf{X}$,
$n \times 1$ vector of instrumented treatment $\hat{T}$
Output: a random variable $T_s$ with $Cov(\mathbf{X}, T_s) = Cov(\mathbf{X}, \hat{T})$
Define $\rho := \begin{pmatrix} Cov(X_1, \hat{T}) & \dots & Cov(X_d, \hat{T}) \end{pmatrix}^\top$
Draw $W \sim \mathcal{N}(0, 1)$.
Regress $W$ on $\mathbf{X}$ and compute residuals: $\hat{\eta} := W - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'W$
Compute the singular value decomposition of $\mathbf{X}$:
    $$\mathbf{X} = \mathbf{U}\, \Sigma\, \mathbf{V}^\top$$
    where the diagonal elements $\{\sigma_j\}_{j=1}^d$ of $\Sigma$ are the singular values of $\mathbf{X}$, $\mathbf{U}$ contains the left-singular vectors, $\mathbf{V}$ contains the right-singular vectors
Compute $\mathbf{X}_{\text{dual}} := (n-1) \times \mathbf{U} \times diag(1/\sigma_j) \times \mathbf{V}^\top$ where $1/\sigma_j$ is replaced with zero if $\sigma_j = 0$
Compute $s := \sqrt{\frac{1 - \rho^\top \times Cov(\mathbf{X}_{\text{dual}}, \mathbf{X}_{\text{dual}}) \times \rho}{Var(\hat{\eta})}}$
Compute $T_s := \mathbf{X}_{\text{dual}} \times \rho + s \times \hat{\eta}$
```

Algorithm 3: Generate synthetic T_s

4.6 Monte Carlo Simulation

I consider the model in eqs. (4.6)-(4.8). To analyze the effectiveness of the instrument validity 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, I present the simulation to study violations of the exclusion restriction, followed by the simulation to study violations of the exchangeability assumption.

Throughout, $\|a\|$ denotes the L_2 norm of the d-dimensional vector $a = \begin{pmatrix} a_1 \\ \vdots \\ a_d \end{pmatrix}$:

$$\|a\| := \Big(\sum_{i=1}^{d} a_i^2\Big)^{1/2}.$$

4.6.1 Simulation Regime 1: Violation of exclusion restriction

Let n-dimensional vectors of disturbances, ε_Y and ε_T, be drawn from

$$\begin{pmatrix} \varepsilon_Y \\ \varepsilon_T \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \omega_3 \\ \omega_3 & 1 \end{pmatrix} \right), \tag{4.25}$$

and the instrument, Z, be generated by

$$Z \sim \text{Bernoulli}(0.5). \tag{4.26}$$

The set of covariates is generated by first drawing $d-1$ eigenvalues from a uniform distribution:

$$\lambda_i \sim \mathcal{U}(0.5, 1.5) \tag{4.27}$$

which then populate the diagonal of a $(d-1) \times (d-1)$ matrix Λ. Then draw a random orthonormal matrix $\mathbf{O}$ of dimension $(d-1)$ and set

$$\Sigma = \mathbf{O}\Lambda\mathbf{O}^\top \tag{4.28}$$

and draw $\mathbf{X}_{\text{temp}}$ from a multivariate normal distribution

$$\mathbf{X}_{\text{temp}} \sim \mathcal{N}(\mathbf{0}, \Sigma). \tag{4.29}$$

Draw a random d-dimensional vector

$$\beta_{c,temp} \sim \mathcal{N}(0, 1) \tag{4.30}$$

and, to keep the variance of Y for various dimensions d comparable, normalize $\beta_c = \beta_{c,temp}/\left\|\beta_{c,temp}\right\|$. With these ingredients set

$$\mathbf{X} = \mathbf{X}_{\text{temp}} + \varepsilon_Y \beta_c'. \tag{4.31}$$

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 $\mathcal{N}(0,1)$,

$$\beta_{T,temp} \sim \mathcal{N}(0, 1) \tag{4.32}$$

and set $\beta_T = \left(\beta_{T,temp}\right) / \left\|\left(\beta_{T,temp}\right)\right\|$ to keep the relative influence of $\mathbf{X}$ on T constant regardless of the number of covariates d.

Further, generate treatment, T, as

$$T = \mathbb{1}\Big(\mathbf{X}\beta_T' + \omega_2 Z + \varepsilon_T > T'\Big). \tag{4.33}$$

where T' is the mean of $\mathbf{X}\beta_T' + \varepsilon_T$ and $\mathbb{1}$ is the indicator function.

To simulate the outcome variable, first generate a random d-dimensional vector

$$\beta_{\text{temp}} \sim \mathcal{N}(0, 1). \tag{4.34}$$

To keep the variance of Y constant regardless of the number of covariates d, set $\beta = \Big(\beta_{\text{temp}}\Big) / \Big\|\Big(\beta_{\text{temp}}\Big)\Big\|$.

The true coefficient of the treatment variable is set to

$$\tau = 1.$$

Finally, generate outcome Y as

$$Y = \mathbf{X}\beta' + \omega_1 Z + \tau T + \varepsilon_Y. \tag{4.35}$$

4.6.2 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, I replace (4.26) with

$$Z = \mathbb{1}(\varepsilon_Z + \omega_1 \varepsilon_Y > 0) \tag{4.36}$$

where ε_Z is drawn from a standard Gaussian. Thus, ω_1 controls the degree of violation of the exchangeability assumption. Finally, I replace (4.35) with

$$Y = \mathbf{X}\frac{\beta_{\text{temp}}'}{\|\beta_{\text{temp}}\|} + \tau T + \varepsilon_Y. \tag{4.37}$$

4.6.3 Parameter constellations

An overview of the interpretation of the parameters is provided:

- ω_1: endogeneity of the instrument, Z
- ω_2: relevance of the instrument, Z

- ω_3: endogeneity of treatment, T

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, I implement Monte Carlo simulations for each combination of the following parameters: number of observations: $n \in \{500, 1000\}$, number of covariates: $d \in \{10, 20\}$ (one endogenous treatment variable: T, along with $d - 1$ exogenous variables: $X_1, \ldots, X_{d-1}$), degree of the endogeneity of T: $\omega_3 = 0.5$, degree of the relevance of the instrument: $\omega_2 \in \{0.3, 0.6\}$ degree of the endogeneity of the instrument, Z: $\omega_1 \in \{0, 0.1, 0.2, 0.4, 0.5\}$. Moreover, the following parameters are fixed: number of bootstrap samples $B = 200$, number of Monte Carlo draws $M = 500$. In Appendix 4.10.4 I show simulation results for $\omega_3 = \{0.25, 0.75\}$, which are not discussed in the main body of the paper.

I also report the average difference between the κs over all boostrap draws:

$$\delta_B = \frac{1}{B}\sum_{b=1}^{B}(\kappa_{i,b} - \kappa_{s,b}). \tag{4.38}$$

4.6.4 Results of Monte Carlo Study

I begin the discussion with Simulation Regime 1, i.e. simulated violations of the exclusion restriction. Figure 4.3 shows the evolution of the average over 300 Monte Carlo runs of pseudo-p-value and δ_B as a function of the degree of endogeneity of the instrument (ω_1). Both measures are increasing with endogeneity, which shows that they are picking up the confoundedness signal in the data. The empirical rejection rate based on the pseudo-p-value ($\psi(\alpha)$) with $\alpha = 0.05$ increases as a function of the instrument endogeneity. The null hypothesis of instrument validity is rejected more and more often as the level of endogeneity is increasing. For combinations of large d and large n, the empirical rejection probability moves up from close to 0 to effectively 1 as endogeneity is introduced. Generally, both a larger d and a larger n improve the performance of the test; however, given d, increasing n improves performance by more than increasing d given n. Considering that the asymptotic results in JS require $d \to \infty$, I show Monte Carlo results for relatively small d and still achieve good performance.

In order to evaluate the trade-off between making type I and type II errors I calculate the area under the ROC curve (AUC) and plot it as a function of the endogeneity of the instrument, Figure 4.4 (see Appendix 4.10.5 for details on the calculation). It is noteworthy

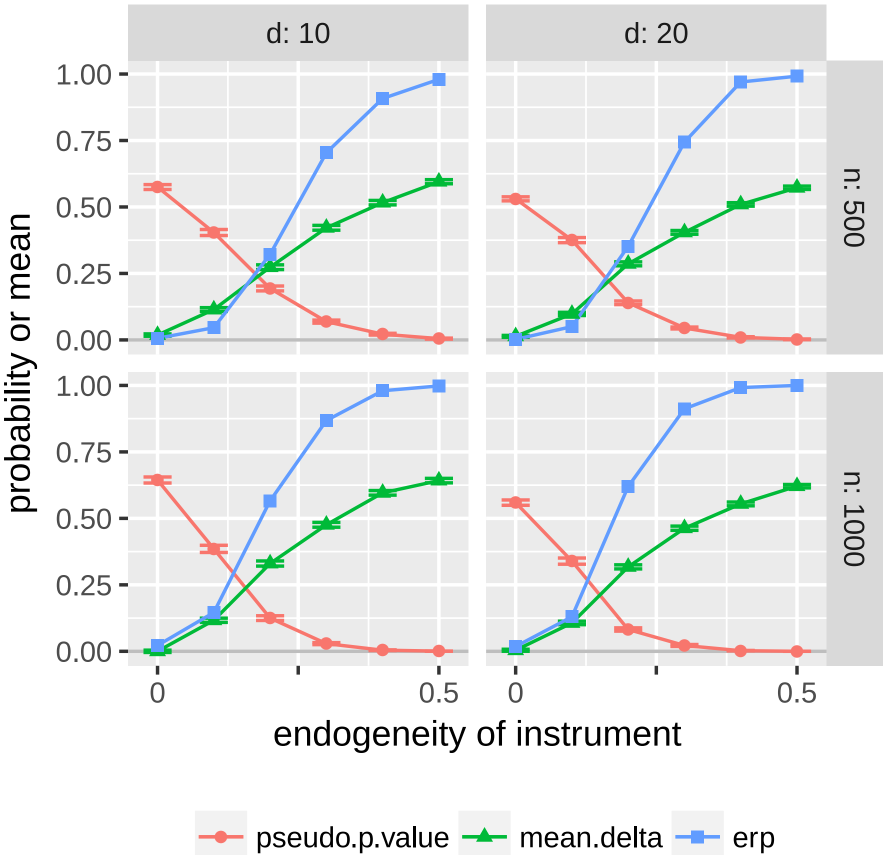


Figure 4.3: Simulation results: pseudo-p-values, δ_B, and empirical rejection rate as a function of ω_1. *This figure shows averages over all M Monte Carlo draws of the p-value, δ_B, and the empirical rejection probability (based on the p-value with threshold parameter $\alpha = 0.05$) as a function of the degree of instrument endogeneity where the source of confounding is a* ***violation of the exclusion restriction****, by number of covariates d and number of observations n. δ_B rises sharply with the degree of confounding, the p-value goes down as the degree of confounding increases. Consequently, the empirical rejection probabilities increase as the degree of confounding increases indicating that, if the degree of condounding is sufficiently high, the test rejects the null of instrument validity in all Monte Carlo draws.*

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 is implicitly accompanied by 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); the superior performance of the algorithm as ω_2 decreases mirrors this result. I assume a constant treatment effect in the present setting and, thus, speaking of compliers, always-takers, etc. is not precise. Nevertheless, as ω_2 increases Z contains less and less additional variation that can be leveraged in the IV implementation or in the validity test. In the extreme, Z and T collapse to one variable

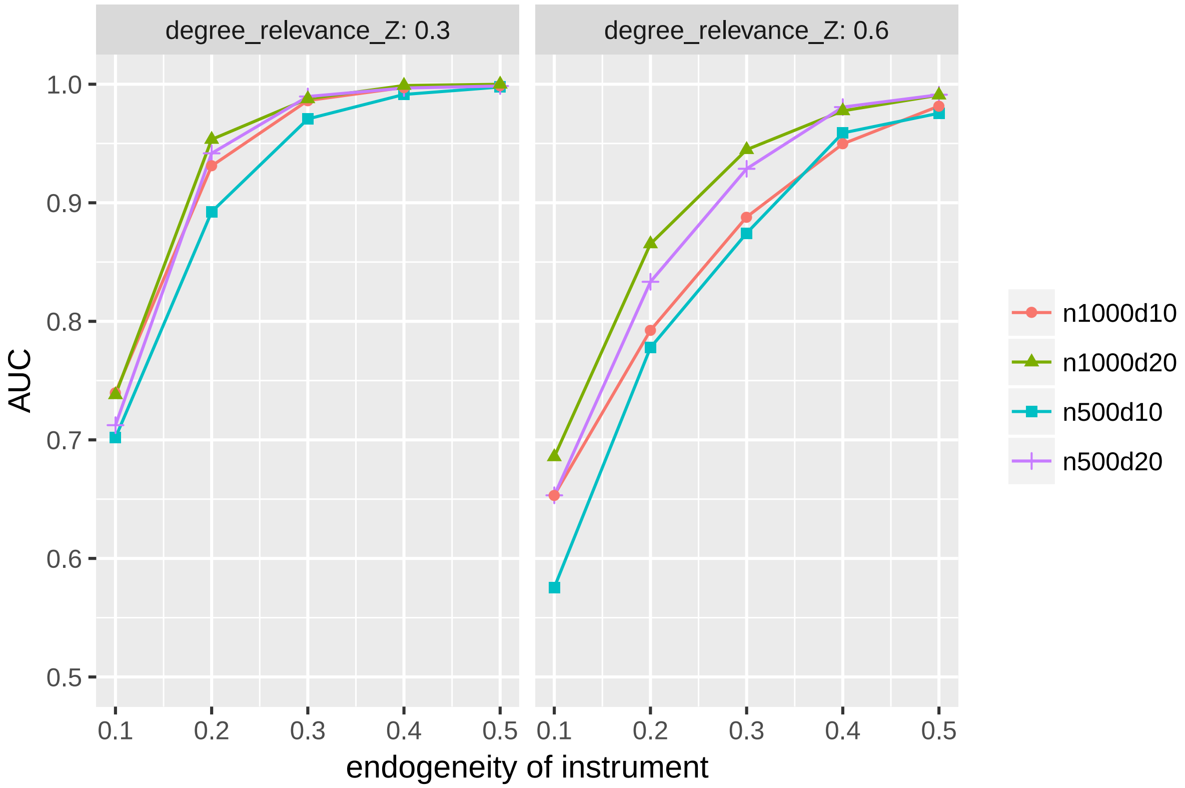


Figure 4.4: Simulation results: AUC curves. *This Figure shows the area under the ROC curve (AUC) as a function of the degree of instrument endogeneity where the source of confounding is a **violation of the exclusion restriction**, for various combinations of number of covariates, d, and number of observations, n, by instrument relevance degree (ω_2, horizontal). The 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 the test performance increases.*

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 the unobserved error) 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 instrument) to 0.9 (high degree of endogeneity).

In Simulation Regime 2 I analyze whether the algorithm can also detect an invalid instrument when its invalidity stems from the fact that the exchangeability assumption is violated. The results are presented in Appendix 4.10.4. Figures 4.6 and 4.7 report results for Simulation Regime 2 in the same form as previous figures for Simulation Regime 1. The performance of the test is similar.

Table 4.1: *For combinations of number of observations n, number of covariates d, confounding degree of instrument ω_1 this Table shows the area under the ROC curve (AUC), the average of $\delta_B^{\mathbb{1}}$ over all $M = 500$ Monte Carlo draws $\bar{\delta}_B^{\mathbb{1}}$, the empirical rejection probability for $\alpha = 0.05$ as well as the empirical rejection rate under H_0, i.e. when $\omega_1 = 0$. Simulation Regime 1: violation of exclusion restriction. $\omega_2 = 0.3$, $\omega_3 = 0.5$, $Var(\varepsilon_Y) = Var(\varepsilon_T) = 1$, $B = 200$.*

n	d	ω_1	AUC	$\bar{\delta}_B^{\mathbb{1}}$	*erp*	*erp* under H_0
500	10	0.1	0.702	0.114	0.046	0.006
		0.2	0.892	0.273	0.322	0.006
		0.3	0.971	0.422	0.704	0.006
		0.4	0.991	0.516	0.908	0.006
		0.5	0.998	0.595	0.980	0.006
	20	0.1	0.712	0.098	0.050	0.002
		0.2	0.942	0.286	0.352	0.002
		0.3	0.990	0.404	0.744	0.002
		0.4	0.997	0.510	0.970	0.002
		0.5	0.998	0.572	0.992	0.002
1000	10	0.1	0.740	0.117	0.146	0.022
		0.2	0.931	0.330	0.566	0.022
		0.3	0.986	0.476	0.868	0.022
		0.4	0.997	0.596	0.980	0.022
		0.5	0.999	0.642	0.998	0.022
	20	0.1	0.738	0.107	0.130	0.018
		0.2	0.953	0.318	0.620	0.018
		0.3	0.988	0.463	0.912	0.018
		0.4	0.999	0.554	0.992	0.018
		0.5	1.000	0.621	1.000	0.018

Table 4.2: *For combinations of number of observations n, number of covariates d, confounding degree of instrument ω_1 this Table shows the area under the ROC curve (AUC), the average of $\delta^{\mathbb{1}}_B$ over all $M = 500$ Monte Carlo draws $\bar{\delta}^{\mathbb{1}}_B$, the empirical rejection probability for $\alpha = 0.05$ as well as the empirical rejection rate under H_0, i.e. when $\omega_1 = 0$. Simulation Regime 1: violation of exclusion restriction. $\omega_2 = 0.6$, $\omega_3 = 0.5$, $Var(\varepsilon_Y) = Var(\varepsilon_T) = 1$, $B = 200$.*

n	d	ω_1	AUC	$\bar{\delta}^{\mathbb{1}}_B$	*erp*	*erp* under H_0
500	10	0.1	0.575	0.026	0.080	0.018
		0.2	0.778	0.128	0.264	0.018
		0.3	0.874	0.221	0.458	0.018
		0.4	0.959	0.353	0.722	0.018
		0.5	0.976	0.436	0.832	0.018
	20	0.1	0.653	0.034	0.072	0.012
		0.2	0.833	0.135	0.274	0.012
		0.3	0.929	0.234	0.546	0.012
		0.4	0.981	0.340	0.790	0.012
		0.5	0.991	0.430	0.920	0.012
1000	10	0.1	0.653	0.035	0.156	0.042
		0.2	0.792	0.127	0.374	0.042
		0.3	0.888	0.224	0.556	0.042
		0.4	0.950	0.354	0.784	0.042
		0.5	0.982	0.446	0.888	0.042
	20	0.1	0.686	0.040	0.190	0.042
		0.2	0.865	0.136	0.430	0.042
		0.3	0.945	0.250	0.706	0.042
		0.4	0.978	0.345	0.878	0.042
		0.5	0.991	0.440	0.970	0.042

4.6.5 Comparison of Performance to Kitagawa (2015)

I compare the performance of my instrument validity test to the one proposed by Kitagawa (2015). Note that these two tests rely on two entirely different approaches: the former on the genericity of estimated parameter vectors w.r.t. the covariance matrix of independent variables, the latter on restrictions on the outcome distribution of subsets of the data implied by the interaction of instrument and treatment assignment.

Tables 4.3 and 4.4, for $\omega_2 = 0.3$ and $\omega_2 = 0.6$ respectively, show comparisons of AUC levels for the test proposed in this paper ('mine') and the one proposed by Kitagawa (2015). The AUC levels for my approach always lie above those corresponding to Kitagawa's approach. Especially for low levels of ω_1 my approach outperforms Kitagawa's. Theoretically, I diverge from Kitagawa (2015) by assuming constant treatment effects. Nevertheless, my simulation approach implicitly generates compliers, always-takers, and never-takers, whose respective outcome distributions are essential for Kitagawa (2015).

I do not compare my approach to the one proposed by Huber and Mellace (2015) since their implementation requires the sample mean of T given $Z = 1$ to be larger than the sample mean of T given $Z = 0$. If this is not the case, the bounds for the quantile function they use lie outside the interval [0,1]. In my simulation, this need not always be the case, especially when $\omega_2 = 0$. Therefore, I compare my approach with Kitagawa's work.

4.7 Empirical application

I follow Kitagawa (2015) and Huber and Mellace (2015) and apply the proposed test to the IV study by Card (1995). Card proposes the proximity to a four-year college as an instrument of educational attainment to estimate returns to schooling, measured by log of weekly earnings. Card himself casts doubt on the validity of college proximity as an instrument as there might be factors such as family preferences or local labor market conditions that are be related to both the proximity to a college and the outcome variable. However, the instrument is likely to be valid, so his argument, in subsamples defined by the following set of covariates $\{S\} :=$ {ethnicity dummy, father's educational level, living in South dummy for 1966 and 1976, urban residence dummy for 1966 and 1976}. Unlike, the tests proposed by Kitagawa (2015) and Huber and Mellace (2015), my test requires the inclusion of covariates by construction. Therefore, to evaluate Card's argument on the validity of his instrument, I run my test three times: first, I include the full set of covariates which includes, beyond $\{S\}$, information on IQ levels, the knowledge of the world score, availability of a library card in the household head's childhood home, marital status, labor market experience, etc. Call this set of additional covariates $\{R\}$. Second, I include only

Table 4.3: *For combinations of number of observations (n), number of covariates (d), confounding degree of instrument ω_1 this table shows the area under the ROC curve (AUC) for my approach and Kitagawa (2015).* $\omega_2 = 0.3$, $\omega_3 = 0.5$, $Var(\varepsilon_Y) = Var(\varepsilon_T) = 1$

			AUC	
n	d	ω_1	mine	Kitagawa
500	10	0.1	0.702	0.495
		0.2	0.892	0.530
		0.3	0.971	0.580
		0.4	0.991	0.613
		0.5	0.998	0.733
	20	0.1	0.712	0.501
		0.2	0.942	0.520
		0.3	0.990	0.570
		0.4	0.997	0.628
		0.5	0.998	0.689
1000	10	0.1	0.740	0.481
		0.2	0.931	0.540
		0.3	0.986	0.586
		0.4	0.997	0.711
		0.5	0.999	0.792
	20	0.1	0.738	0.512
		0.2	0.953	0.526
		0.3	0.988	0.614
		0.4	0.999	0.712
		0.5	1.000	0.795

Table 4.4: *For combinations of number of observations (n), number of covariates (d), confounding degree of instrument* ω_1 *this table shows the area under the ROC curve (AUC) for my approach and Kitagawa (2015).* $\omega_2 = 0.6$, $\omega_3 = 0.5$, $Var(\varepsilon_Y) = Var(\varepsilon_T) = 1$

			AUC	
n	d	ω_1	mine	Kitagawa
500	10	0.1	0.575	0.499
		0.2	0.778	0.511
		0.3	0.874	0.548
		0.4	0.959	0.592
		0.5	0.976	0.651
	20	0.1	0.653	0.498
		0.2	0.833	0.510
		0.3	0.929	0.526
		0.4	0.981	0.565
		0.5	0.991	0.631
1000	10	0.1	0.653	0.517
		0.2	0.792	0.519
		0.3	0.888	0.569
		0.4	0.950	0.625
		0.5	0.982	0.698
	20	0.1	0.686	0.497
		0.2	0.865	0.537
		0.3	0.945	0.576
		0.4	0.978	0.605
		0.5	0.991	0.715

Table 4.5: Results of empirical application to Card (1995). *This Table shows results of the empirical application, based on Card (1995). $\{S\}$ denotes the set of covariates implicitly defining the subgroups in which the instrument is valid according to Card. $\{R\}$ contains all remaining covariates (for details see main text). Consistent with Card's argument, the null hypothesis of instrument validity cannot be rejected when all covariates are included; see column $\{R, S\}$. Similarly, when only the six covariates $\{S\}$ are included the instrument validity can also not be rejected; see column $\{S\}$. Dropping all variables $\{S\}$ and keeping only those in $\{R\}$, the test rejects instrument validity.*

	test results for different sets of covariates		
	$\{R, S\}$	$\{S\}$	$\{R\}$
pseudo-p-value	0.18	0.29	0.00
no. of covariates	29	6	23
no. of observations	3612	3612	3612

variables $\{S\}$. In these first two cases, I expect the test not to reject instrument validity since I am controlling for those variables $\{S\}$ that render the instrument valid according to Card. Third, I include only variables $\{R\}$ and exclude variables $\{S\}$. In the third run, I expect the test to reject the null hypothesis of instrument validity if Card's argument holds.

The results in Table 4.5 show that the test does not reject the null of instrument validity if I control for variables $\{S\}$. On the contrary, once $\{S\}$ is left out of the set of covariates, the test rejects instrument validity. This is consistent with Card's reasoning and the results of Kitagawa (2015) and Huber and Mellace (2015). The latter approaches are based on an IV model with heterogeneous effects whereas my approach relies on the assumption of a constant τ.

These results show that the proposed test is able to detect information on the validity of the instrument solely based on the spectra of the covariates induced by the estimated parameter vectors.

4.8 Discussion

In this section, I discuss limitations and future extensions of the proposed method.

Distribution of pseudo-p-values under H_0 and type I error control. I call the quantity in eq. (4.21) a *pseudo-p*-value and not a p-value because it does not have a uniform distribution under H_0. Rather, the following is the case. As $d \to \infty$, $\kappa_i \approx \kappa_s$ under H_0 because both T_s and $\hat{T}$ are unconfounded and have the same covariance structure with $\mathbf{X}$ (see also Appendix 4.10.2). Therefore, it is subject to chance whether $\kappa_i > \kappa_s$

(i.e. $\delta > 0$) or $\kappa_i < \kappa_s$ (i.e. $\delta < 0$). Thus, the pseudo-p-value, which is the share of $\delta \leq 0$ across B bootstrap draws, will converge to 0.5; in expectation $\delta_B^{\mathbb{1}} = 0.5$ under H_0. One can see this behavior of $\delta_B^{\mathbb{1}}$ in Figure 4.3 where the pseudo-p-value curve converges to 0.5 for a valid instrument ('endogeneity of instrument = 0') for large d and n. As d and n increase, and κ_i as well as κ_s are more and more precisely estimated, more and more mass of the pseudo-p-value distribution will lie above 0.5. Therefore, although I cannot guarantee that the empirical size of the test converges to its nominal size, this behavior of the pseudo-p-values implies that the empirical size of the test will not exceed the nominal size.

This argument remains informal. It will be subject of further research to investigate whether an alternative specification of $\delta_B^{\mathbb{1}}$ or its transformation guided by the insight that it converges to 0.5 under H_0 can guarantee a uniform distribution under H_0.

Robustness of κ to rescaling of the data. An important limitation of the algorithm proposed by JS is that the estimated κ is, in theory, not robust to rescaling of the data as this introduces a dependence between the covariance matrix of the covariates and the parameter vector. For instance, consider income as one of many independent variables. Its rescaling to logarithms changes both the covariance structure of independent variables and the parameter vector, whose independence drives the method proposed by JS. The authors acknowledge this, yet claim and show in simulations that the estimated κ is robust to rescaling of the data in practice.[8] However, the proposed test relies on a comparison of *two* κs, which is useful beyond the fact that such a comparison allows focusing on the bias of *one* covariate: Both κs are influenced by rescaling in the same fashion, which one can therefore expect to leave the sign of their differences, i.e. δ, unaffected. In Appendix 4.10.3 I document the robustness of the proposed algorithm to typical data transformations: the observed AUC levels are insensitive to rescaling of the data and the pseudo-p-values of the validity test on untransformed and transformed data show a correlation coefficient that exceeds 0.95.

[8] An interesting insight in this context is due to Holmes and Caiola (2018). A given regression techniques should fulfill certain properties to be useful. Two such properties are scale invariance (it should not matter whether data is measured in centimeters or inches) and rotational invariance (it should not matter 'from which angle you are looking at the data'). As an example, ordinary least-squares is scale-invariant but not rotationally invariant; Principal Component Analysis is rotationally invariant but not scale-invariant. Holmes and Caiola derive the incompatibility of these two criteria. For this reason, it might not seem surprising that the JS methodology, which relies on some limited type of rotational invariance, is not scale-invariant. Note that JS assume rotational invariance of the prior on the structural parameter vectors; they do not assume rotational invariance of the model itself.

Idealized modeling assumptions in Janzing and Schölkopf (2018). The JS methodology relies on a number of results on the limiting distribution of spectral measures in high dimensions. In particular, they consider a sequence of symmetric matrices $\Sigma_{d\times d}$ whose distribution of eigenvalues (spectral measure) converges weakly to some probability measure μ^∞ as d increases. Wishart matrices provide a common theoretical starting point to analyze such asymptotic properties. A Wishart matrix M is defined as $M = n^{-1}X^\top X$ where X is a $n \times p$ matrix with each column containing n independent samples from a real-valued random variable. Marchenko and Pastur (1967) show that the spectral measure of a Wishart matrix converges to an asymptotic distribution, which is a crucial result underlying the JS methodology (see also Götze and Tikhomirov, 2004).

The modeling assumptions that JS impose on the multivariate linear model to show the decomposability of the spectral measure induced by the potentially biased parameter vector are strong. Yet, it can be useful to impose constraints on a given model to obtain information about an unobserved quantity of interest. In the case at hand the unobserved quantity of interest is relation between the unobserved structural errors influencing both outcome and treatment – a quantity that is both notoriously difficult to characterize and of paramount importance in observational causal effect studies. The attempt to quantify this crucial quantity, albeit at the cost of idealized modeling assumptions, may be informative. In the paper at hand, the informational content gained is the evidence on instrument validity.

To give another example in which idealized assumptions enable the researcher to characterize some unobserved quantity, consider the work done by Oster (2019) to assess the robustness of an estimated treatment effect to unobserved confounding. To assess the severity of an unobserved confounder problem, researchers commonly check how sensitive the treatment parameter of interest is to the inclusion of the observed controls (e.g. Lacetera et al., 2012; Acemoglu et al., 2008). That observed confounders and their relationship to the treatment are informative about the relation between unobserved confounders and the treatment is the underlying, sometimes implicit, assumption of this procedure. Oster (2019) shows that evaluating robustness to unobserved confounders by observing coefficient movements alone is insufficient. Rather, it is important to take into account reactions by both the coefficient and the coefficient of determination, R^2, to the selective inclusion of observed confounders. Under the assumption that the relationship between unobserved confounders and treatment can be fully recovered from the relationship between observed confounders and the treatment, Oster (2019) shows how to bound the true treatment effect. Full recoverability in this work and the modelling assumptions in JS are both idealized assumptions that nevertheless yield important insights.

The crucial modeling assumption in JS is that the structural model parameters are drawn from a rotation-invariant prior distribution, namely from a sphere with fixed radius. In contrast, the uninformative prior in Bayesian linear regression analysis is uniform. Despite the fact that JS require only a rotationally invariant prior on the structural model parameter and not a rotationally invariant *model* itself, it is worth highlighting that even the latter are invoked in econometrics (see e.g. Andrews et al., 2006).

Accounting for heterogeneous coefficients. Heterogeneity in effect sizes across individuals is a common notion in economics: for instance, the causal effect of a policy intervention, such as the introduction of a statutory minimum wage on wage growth, might differ among individuals with different educational levels or employers. In other words, the effect size might depend on covariates. At first glance, such heterogeneity seems incompatible with the PIM or the Independence between Cause and Mechanism, which, after all, postulates an independence between the true causal parameter vector and the covariates representing the causes. In this respect, it is worthwhile to analyze random coefficient models as the workhorse of heterogeneity analysis in economics.

Typically, heterogeneity is modeled with random coefficient models such as $Y_i = \beta_i x_i + \varepsilon_i$ with individual-specific slope parameters β_i. These usually come with constant mean and variance assumptions on the distribution of β_i: $\beta_i = \beta + \alpha_i$ where $\mathbb{E}(\alpha_i) = 0$, $\mathbb{E}(\alpha_i \alpha_i') = \Lambda$, some covariance matrix (Hsiao and Pesaran, 2008; Swamy, 1970). β is the average effect.[9] Thus, the notion of independence underlying PIM and ICM can be understood as the independence between the covariates and the average effect β. Future work will analyze to what extent the JS methodology can be adapted to such cases. In any case, there is no inherent contradiction between PIM and heterogeneous effects as modeled by random coefficient models.

Parameter η. As indicated in Section (4.3) and Appendix 4.10.1, the JS methodology estimates a degree of confounding by minimizing the distance between an empirical vector-induced spectral measure and a two-parametric probability measure. One of the parameters is κ, the degree of confounding. The second parameter, η, measures the explanatory power of the unobserved error u for the covariates $\{\mathbf{X}, \hat{T}\}$ in eq. (4.18). η is introduced to distinguish between cases where a rescaling of $\begin{pmatrix} \mathbf{b} \\ b_\tau \end{pmatrix}$ and c in eqs. (4.17) and (4.18) leads to the same κ, see the discussion above eq. (4.54) in Appendix 4.10.1. A given κ can be consistent with a range of ηs. In the paper at hand, I do not take this ambiguity into account. This implies that I abstract from whether a given level of bias

[9] See also Hoderlein et al. (2010), who assume that β is independent of x to analyze a non-parametric random coefficient model.

in the estimated parameter vector originates from, on the one hand, a high explanatory power of u for Y and a low explanatory power of u for $\{\mathbf{X}, \hat{T}\}$ or, on the other hand, a low explanatory power of u for Y and a high explanatory power of u for $\{\mathbf{X}, \hat{T}\}$.

Alternatives to bootstrapping to assess uncertainty. Algorithm 2 assesses the statistical uncertainty about κ by calculating it for B bootstrap values. This procedure does not take into account the statistical uncertainty that underlies the estimation of the parameter vector *per se*. The standard errors associated with each parameter estimate might be used in an alternative strategy to assess the uncertainty about κ. More specifically, one might calculate a series of κs for different draws from the distribution of the estimated parameter vector. Given the centrality of the estimated parameter vector to estimate the degree of confounding in the JS methodology, this seems a worthwhile strategy to pursue. It will be subject of future research.

Measuring relative validity of the instrumented vs. original treatment variable. While it is straightforward to build a test on a comparison of $\kappa_1 = \kappa(\mathbf{X}, T; Y)$ and $\kappa_2 = \kappa(\mathbf{X}, \hat{T}; Y)$, it is not desirable for the objective of this paper. Such a comparison would merely yield information about whether instrumenting makes $\hat{T}$ *less* confounded than T. However, the question of whether a comparison of κ_1 and κ_2 in combination with the corresponding observed, yet biased, treatment effect estimates enables an extrapolation to the unbiased, yet unobserved, τ is subject of ongoing research.

4.9 Conclusion

Since the justification of IV assumptions is in practice seldom statistically-grounded and often relies on controversial context-specific arguments, it is pertinent to provide statistically-grounded methods to evaluate IV validity empirically. The proposed method leverages statistical traces of confounding, measured with the methodology laid out in Janzing and Schölkopf (2018), to test whether a potential instrument is valid. As such, it provides a novel way to test IV validity. It relies on Schölkopf and Janzing's insight that, under idealized assumptions, the spectral measure of the covariance matrix of the independent variables in a multivariate linear model that is induced by the estimated parameter vector can be decomposed into a causal part and a confounded part, which then yields information on the degree of confounding.

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 and covariates increases. In addition, I document the feasibility of the proposed test in an empirical application. I show that the test can reproduce the argument on instrument validity made by Card (1995) in spite of the likely violation of idealized modeling assumptions that underlie the estimation of the degree of confounding, which, in turn, underlies the test procedure.

In contrast to the few existing methods to test for instrument validity, my test relies neither on the Potential Outcomes framework nor on higher-order moment restrictions. Therefore, it constitutes a novel approach to evaluating IV validity that can be applied to IV applications in structural equation models. Despite different theoretical approaches, I compare the performance of my test to the one proposed by Kitagawa (2015). My test performs favorably.

4.10 Appendix

4.10.1 Quantifying the degree of confounding

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 causal influence of $\mathbf{X}$ on Y or due to an unobserved confounder influencing both $\mathbf{X}$ and Y. In multivariate linear models, they point out that the spectral measure of the covariance matrix of the 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 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.

The set-up

Consider the following linear structural equation model:

$$\mathbf{X} = \mathbf{b}u + \mathbf{E} \tag{4.39}$$

$$Y = \mathbf{X}^{\top}\mathbf{a} + cu^{\top} + \varepsilon \tag{4.40}$$

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 zero-mean errors drawn independently from u. ε is a $n \times 1$ vector of errors. c is a scalar. Without loss of generality, u is assumed to have unit variance.

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

$$\hat{\mathbf{a}} := \Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{X}Y}, \tag{4.41}$$

where Σ denotes covariance matrices. Generally, I am 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

$$\begin{aligned}\Sigma_{\mathbf{X}Y} &= Cov(\mathbf{X}, Y) = Cov(\mathbf{b}u + \mathbf{E}, \mathbf{X}^\top \mathbf{a} + cu^\top + \varepsilon)\\ &= (\Sigma_{\mathbf{EE}} + \mathbf{b}\mathbf{b}^\top)\mathbf{a} + c\mathbf{b}\\ \Sigma_{\mathbf{XX}} &= Cov(\mathbf{X}, \mathbf{X}) = Cov(\mathbf{b}u + \mathbf{E}, \mathbf{b}u + \mathbf{E})\\ &= \Sigma_{\mathbf{EE}} + \mathbf{b}\mathbf{b}^\top,\end{aligned}$$

and therefore

$$\hat{\mathbf{a}} = \mathbf{a} + (\Sigma_{\mathbf{EE}} + \mathbf{b}\mathbf{b}^\top)^{-1} c\mathbf{b} = \mathbf{a} + c\Sigma_{\mathbf{XX}}^{-1}\mathbf{b}. \tag{4.42}$$

Genericity assumptions

The idea underlying this method is the Independence between Cause and Mechanism (ICM) postulate (Peters et al., 2017), which states that the causal mechanism, represented by the conditional distribution of effect, Y, given cause, $\mathbf{X}$, $f(Y|\mathbf{X})$, is independent of the marginal distribution of the cause, $f(\mathbf{X})$. The ICM postulate is discussed in Appendix 4.10.6.

To understand what the ICM amounts to in the case at hand, note that the crucial determinant for $f(\mathbf{X})$ is $\Sigma_{\mathbf{XX}}$, likewise the crucial determinant for $f(Y|\mathbf{X})$ is $\mathbf{a}$. 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 X, and, thus, also the covariance matrix $\Sigma_{\mathbf{XX}}$, $\mathbf{a}$ is not likely to be aligned with its first principal component.[10] I next discuss what the concept of 'generic orientation' amounts to.

In order to make the notion of 'generic orientation' precise, some definitions are needed. First of all, assuming that all eigenvalues of a matrix are different from each other (i.e. the matrix is non-degenerate), each such symmetric $d \times d$ matrix A has a unique decomposition

$$A = \sum_{j=1}^{d} \lambda_j \phi_j \phi_j^\top \tag{4.43}$$

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

The renormalized trace is defined to be

$$\tau(A) := \frac{1}{d} tr(A) \tag{4.44}$$

(note that the τ in this notation is unrelated to the treatment effect that it denotes in the main body of the paper).

[10] To be precise, for the structural model in (4.39), the argument involves a generic orientation of $\mathbf{a}$ and *the eigenspaces* of $\Sigma_{\mathbf{XX}}$.

Definition 4.10.1. (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 $\lambda_1, \ldots, \lambda_d$:

$$\mu_A^{\mathrm{Tr}} := \frac{1}{d} \sum_{j=1}^{d} \delta_{\lambda_j} \tag{4.45}$$

where δ_{λ_j} denotes the point measure on λ_j.

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

Definition 4.10.2 (vector-induced spectral measure)**.** Given a symmetric $d \times d$ matrix A with associated eigenvalues λ_j and corresponding eigenvectors ϕ_j, the spectral measure induced by an arbitrary vector $v \in \mathbb{R}^d$ is given by

$$\mu_{A,v} = \sum_{j=1}^{d} \left(v^\top \phi_j\right)^2 \delta_{\lambda_j} \tag{4.46}$$

where δ_{λ_j} denotes the point measure on λ_j.

Intuitively, $\mu_{A,v}$ describes the squared length of components of a vector projected onto the eigenspace of $\Sigma_{\mathbf{XX}}$. 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, i.e. 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 $w = (1/d, \ldots, 1/d)$ representing the uniform weight of 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. (4.39) 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

$$\mu_{\Sigma_{\mathbf{XX}},\mathbf{a}} \approx \mu_{\Sigma_{\mathbf{XX}}}^{\mathrm{Tr}} ||\mathbf{a}||^2 \tag{4.47}$$

2. Vector $\mathbf{b}$ has generic orientation relative to $\Sigma_{\mathbf{EE}}$ in the sense that

$$\mu_{\Sigma_{\mathbf{EE}},\mathbf{b}} \approx \mu_{\Sigma_{\mathbf{EE}}}^{\mathrm{Tr}} ||\mathbf{b}||^2. \tag{4.48}$$

3. Vector $\mathbf{a}$ is generic relative to $\mathbf{b}$ and $\Sigma_{\mathbf{EE}}$ in the sense that

$$\mu_{\Sigma_{\mathbf{XX}},\mathbf{a}+c\Sigma_{\mathbf{XX}}^{-1}\mathbf{b}} \approx \mu_{\Sigma_{\mathbf{XX}},\mathbf{a}} + \mu_{\Sigma_{\mathbf{XX}},c\Sigma_{\mathbf{XX}}^{-1}\mathbf{b}}. \tag{4.49}$$

Intuitively, (4.47) states that 'decomposing $\mathbf{a}$ into eigenvectors of $\Sigma_{\mathbf{XX}}$ yields weights that are close to being uniformly spread over the spectrum.' (4.48) captures a similar statement for $\mathbf{b}$ and $\Sigma_{\mathbf{EE}}$: the weights of $\mathbf{b}$ are uniformly distributed across the spectrum of $\Sigma_{\mathbf{EE}}$.

Eq. (4.49) contains a crucial ingredient for the ability to detect confounding: the $\hat{\mathbf{a}}$-induced spectral measure (left-hand-side of (4.49), recall $\hat{\mathbf{a}} = \mathbf{a} + c\Sigma_{\mathbf{XX}}^{-1}\mathbf{b}$) 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).

Quantifying confounding

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

Definition 4.10.3 (correlative strength of confounding)**.** The correlative strength of confounding gives the degree to which the confounder contributes to the covariance between $\mathbf{X}$ and T.

$$\gamma := \frac{\|\Sigma_{\mathbf{X}Z}\|^2}{\|\Sigma_{\mathbf{X}T}\|^2 + \|\Sigma_{\mathbf{X}Z}\|^2} \tag{4.50}$$

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

Definition 4.10.4. (structural strength of confounding)

$$\kappa_{\mathrm{JS}} := \frac{\left\|\Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{X}u}\right\|^2}{\left\|\Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{X}Y}\right\|^2 + \left\|\Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{X}u}\right\|^2} = \frac{\left\|c\Sigma_{\mathbf{XX}}^{-1}\mathbf{b}\right\|^2}{\|\mathbf{a}\|^2 + \left\|c\Sigma_{\mathbf{XX}}^{-1}\mathbf{b}\right\|^2}, \tag{4.51}$$

$$\kappa_{\mathrm{JS}} \in [0,1]. \tag{4.52}$$

Note that from (4.49) and a normalizing condition

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

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

$$\kappa \approx \frac{\left\|c\Sigma_{\mathbf{XX}}^{-1}\mathbf{b}\right\|^2}{\|\hat{\mathbf{a}}\|^2} = \frac{\|\hat{\mathbf{a}} - \mathbf{a}\|^2}{\|\hat{\mathbf{a}}\|^2}. \tag{4.53}$$

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 $\hat{\mathbf{a}}$.

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 $\mathbf{X}$,

$$\eta := tr(\Sigma_{\mathbf{XX}} - tr(\Sigma_{\mathbf{XX}|u})) = tr(\Sigma_{\mathbf{XX}}) - tr(\Sigma_{\mathbf{EE}}) = \|\mathbf{b}\|^2 . \tag{4.54}$$

Estimating confounding

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

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

$$\frac{1}{\|\hat{\mathbf{a}}\|^2}\mu_{\Sigma_{\mathbf{XX}},\hat{\mathbf{a}}} - \nu_{\kappa,\eta} \to 0 \text{ (weakly in probability)} \tag{4.55}$$

where

$$\nu_{\kappa,\eta} := (1-\kappa)\,\nu^{\text{causal}} + \kappa\,\nu_{\eta}^{\text{confounded}} . \tag{4.56}$$

I inspect each part in turn.

1. ν^{causal} is the hypothetical spectral measure that would be obtained in the absence of confounding. Following (4.47), it is defined as

$$\nu^{\text{causal}} := \mu_{\Sigma_{\mathbf{XX}}}^{\text{Tr}} \tag{4.57}$$

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).

2. To define the corresponding confounding part, JS propose an approximation to the spectral measure of $\Sigma_{\mathbf{XX}}$ induced by the vector $\Sigma_{\mathbf{XX}}^{-1}\mathbf{b}$. Recall that $\mathbf{b}$ has generic orientation relative to $\Sigma_{\mathbf{EE}}$, see eq. (4.48). 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.

   (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 $\Sigma_{\mathbf{EE}}$. Although $\mathbf{b}$ is unknown, one does know that it is generic relative to $\Sigma_{\mathbf{EE}}$. Therefore, I can replace $\mathbf{b}$ with a vector that is 'particularly generic', namely $\mathbf{g} := (1, \ldots, 1)^\top/\sqrt{d}$, which satisfies

   $$\mu_{M_E,\mathbf{g}} = \mu_{M_E}^{\text{Tr}}.$$

   Therefore, one can approximate the spectral measure of $\Sigma_{\mathbf{XX}}$ induced by the vector $\Sigma_{\mathbf{XX}}^{-1}\mathbf{b}$ 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:

   $$\mu_{\Sigma_{\mathbf{XX}}}^{\text{Tr}} \approx \mu_{\Sigma_{\mathbf{EE}}}^{\text{Tr}},$$

   see 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

$$\nu_\eta^{\text{confounded}} := \frac{1}{\|T^{-1}\mathbf{g}\|^2}\mu_{T,T^{-1}\mathbf{g}}. \tag{4.58}$$

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 (4.55) 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

$$D(w, w') := \|K(w - w')\|_1 \tag{4.59}$$

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 $\nu_{\kappa,\eta}$.

4.10.2 Proofs: Relation between δ and IV validity

First, I recall the definition of a valid IV.

Definition 4.10.5. A variable Z is called a valid instrumental variable if if fulfills Assumptions 4.4.1 and 4.4.2. Vice versa, an invalid instrumental variable does not fulfill either Assumption 4.4.1 or 4.4.2.

For convenience, I reproduce the reduced form model that forms the starting ground for the test in Section 4.5:

$$Y = \{\mathbf{X}, \hat{T}\} \begin{pmatrix} \beta \\ \tau \end{pmatrix} + cu + \varepsilon \tag{4.60}$$

$$\{\mathbf{X}, \hat{T}\} = \mathbf{E} + u \begin{pmatrix} \mathbf{b} & b_\tau \end{pmatrix} \tag{4.61}$$

Each element of the vector $\mathbf{b} = \begin{pmatrix} b_1 & \dots & b_d & b_\tau \end{pmatrix}$ parameterizes the confounding of the corresponding dimension of $\{\mathbf{X}, \hat{T}\}$, e.g. $X_1 = E_1 + ub_1$. If Z is a valid IV, the instrumented treatment variable $\hat{T}$ is unconfounded, and $b_\tau = 0$.

Note that

$$\delta = \kappa_i - \kappa_s \leq 0 \Leftrightarrow \frac{\kappa_i}{\kappa_s} \leq 1,$$

which will simplify the proof. For convenience, I reproduce the definition of κ_i and κ_s here and introduce some placeholders.

$$\kappa_s = \frac{\overbrace{\left\| c_s \Sigma^{-1}_{\mathbf{X}T_s} \begin{pmatrix} \mathbf{b} \\ b_{T_s} \end{pmatrix} \right\|^2}^{\bar{c}_s}}{\underbrace{\left\| \begin{pmatrix} \mathbf{a} \\ a_{T_s} \end{pmatrix} \right\|^2}_{\bar{a}_s} + \left\| c_s \Sigma^{-1}_{\mathbf{X}T_s} \begin{pmatrix} \mathbf{b} \\ b_{T_s} \end{pmatrix} \right\|^2} \tag{4.62}$$

$$\kappa_i = \frac{\overbrace{\left\| c \Sigma^{-1}_{\mathbf{X}\hat{T}} \begin{pmatrix} \mathbf{b} \\ b_{\hat{T}} \end{pmatrix} \right\|^2}^{\bar{c}_\tau}}{\underbrace{\left\| \begin{pmatrix} \mathbf{a} \\ \tau \end{pmatrix} \right\|^2}_{\bar{a}_\tau} + \left\| c \Sigma^{-1}_{\mathbf{X}\hat{T}} \begin{pmatrix} \mathbf{b} \\ b_{\hat{T}} \end{pmatrix} \right\|^2} \tag{4.63}$$

Note that $a_{T_s} = b_{T_s} = 0$ since I draw T_s independently of Y and the structural error ε_Y. By virtue of how T_s is generated, $\Sigma_{\mathbf{X}T_s} = \Sigma_{\mathbf{X}\hat{T}}$. By replacing $\hat{T}$ with T_s the relation between Y and u does not change and, therefore, $c_s = c$.

Theorem 4.10.1. *If the instrumental variable is valid, $\delta \leq 0$.*

Proof. If the instrumental variable is valid, $b_{\hat{T}} = 0$. Then,

$$\frac{\kappa_i}{\kappa_s} = \frac{\left\| \begin{pmatrix} \mathbf{a} \\ a_{T_s} \end{pmatrix} \right\|^2 + \|\bar{c}\|^2}{\left\| \begin{pmatrix} \mathbf{a} \\ \tau \end{pmatrix} \right\|^2 + \|\bar{c}\|^2} \leq 1 \tag{4.64}$$

where $\bar{c} = c\Sigma^{-1}_{\mathbf{X}\hat{T}} \begin{pmatrix} \mathbf{b} \\ b_{\hat{T}} \end{pmatrix} = c\Sigma^{-1}_{\mathbf{X}T_s} \begin{pmatrix} \mathbf{b} \\ b_{T_s} \end{pmatrix}$ because neither $\bar{c}_\tau$ nor $\bar{c}_s$ contain τ or a_{T_s}, which are the only quantities that differ between κ_s and κ_i if the IV is valid. The last inequality is due to the fact that $a_{T_s} = 0$ by construction. Therefore, it follows

$$\text{IV valid} \Rightarrow \delta \leq 0.$$

□

Corollary 4.10.1.1. *If $\delta > 0$, instrumental variable is invalid.*

Proof. From Theorem 4.10.1 I have

$$\text{IV valid} \Rightarrow \delta \leq 0.$$

Therefore, by contrapositive,

$$\delta > 0 \Rightarrow \text{IV invalid}.$$

□

Thus, the proposed test evaluates the null hypothesis H_0 : IV valid

4.10.3 Robustness to rescaling

As discussed, a drawback of the JS methodology to estimate a degree of confounding is that it is theoretically not robust to rescaling of the data as this introduces a dependence of the parameter vector and the covariance matrix of the covariates.

I adjust the data generating process slightly in order to be able to use logarithmic transfromations. In particular, I transform $\mathbf{X}$, as defined in (4.31), by

$$\mathbf{X} := \mathbf{X} - \min(\min(\mathbf{X}), 0) + \mathbb{1}(\min(\mathbf{X}) < 0) \tag{4.65}$$

and I replace Y as defined in eq. (4.35) by

$$Y := Y - \min(\min(Y), 0) + \mathbb{1}(\min(Y) < 0) \tag{4.66}$$

where $\mathbb{1}$ is the indicator function, which equals 1 if the condition in brackets is fulfilled. I then implement the following three data transformations:

1. $X_1 := \log(X_1)$, and $X_2 := X_2^2$
2. $Y := \log(Y)$, $X_1 := \log(X_1)$, and $X_2 := X_2^2$
3. $Y := \log(Y)$, $X_1 := \log(X_1)$, and $X_2 := \log(X_2)$

For the original data and for each of the three transformations, I implement the algorithm described in the main text and compare the pseudo-p-values that result. First, I show scatter plots of pseudo-p-values for each data transformation against those of the original data. For each data transformation, the pseudo-p-values correlate almost perfectly with those from the original data. Second, I show AUC levels for the original as well as the three data transformations in Table 4.6. The AUC levels are not sensitive to data transformations.

Table 4.6: *For combinations of number of observations (n), number of covariates (d), confounding degree of instrument (ω_1) this table shows the area under the ROC curve (AUC) for the original model and three transformations specified in the main text.* $\omega_2 = 0.3$, $\omega_3 = 0.5$, $Var(\varepsilon_Y) = Var(\varepsilon_T) = 1$.

			AUC			
				transformed models		
n	d	ω_1	original	1	2	3
500	10	0.1	0.691	0.691	0.657	0.661
		0.2	0.889	0.883	0.867	0.866
		0.3	0.972	0.971	0.959	0.957
		0.4	0.994	0.993	0.988	0.989
		0.5	0.999	0.999	0.998	0.998
	20	0.1	0.733	0.733	0.697	0.697
		0.2	0.925	0.926	0.902	0.900
		0.3	0.986	0.986	0.983	0.979
		0.4	0.999	0.998	0.997	0.997
		0.5	1.000	1.000	1.000	1.000
1000	10	0.1	0.729	0.722	0.719	0.712
		0.2	0.929	0.926	0.912	0.912
		0.3	0.981	0.979	0.976	0.976
		0.4	0.996	0.995	0.993	0.993
		0.5	1.000	0.999	0.998	0.998
	20	0.1	0.811	0.808	0.801	0.802
		0.2	0.967	0.966	0.956	0.957
		0.3	0.996	0.996	0.995	0.995
		0.4	1.000	1.000	1.000	1.000
		0.5	1.000	1.000	1.000	1.000

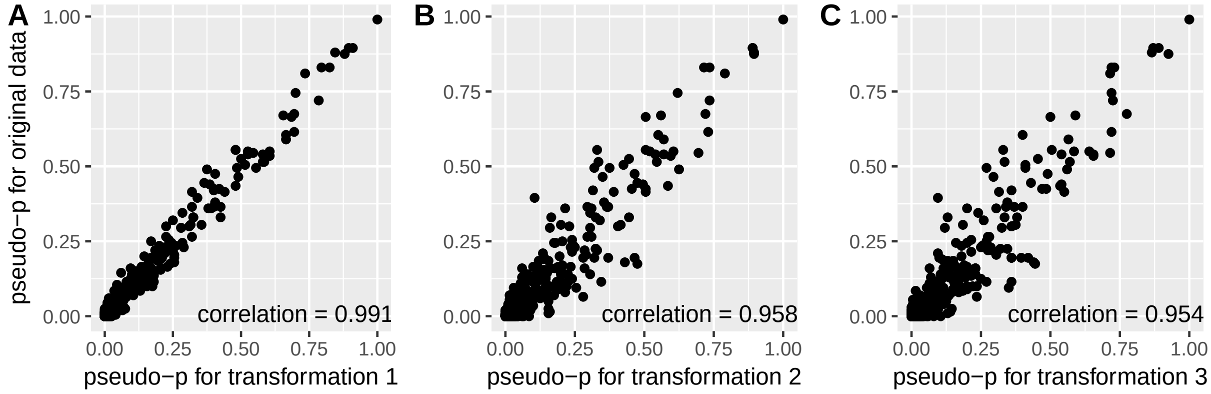


Figure 4.5: Scatter plot of pseudo-p-values. *This Figure shows scatter plots of pseudo-p-values estimated based on transformed data against pseudo-p-values estimated based on the original data. Each panel corresponds to one transformation of the data. The p-values remain largely invariant with each scatter plot displaying a correlation larger than 0.95. This is evidence for the robustness of the proposed test for instrument validity with respect to rescaling of the data.* $n = 1000$, $d = 20$, $\omega_1 = 0.3$, $\omega_2 = 0.3$, $\omega_3 = 0.5$, $Var(\varepsilon_Y) = Var(\varepsilon_T) = 1$

4.10.4 Further results

Further results for Simulation Regime 1

In this section I provide further simulation results for Simulation Regime 1: Violation of the exclusion restriction to show robustness of the results for different variances of ε_T and ε_Y, see Tables 4.7 to 4.11. In addition, I provide results for different levels of treatment endogeneity ω_3, see Tables 4.12 and 4.13.

Simulation Regime 2: Violation of Exchangeability Assumption

Figures 4.6 and 4.7 show results for the simulations for the violation of the exchangeability assumption, see Section 4.6.2. The test performs well also for this violation. Note that the degree of endogeneity of the instrument is not directly comparable to Simulation Regime 1 since ω_1 enters the simulation inside an indicator function for Simulation Regime 2.

Table 4.7: *For combinations of number of observations n, number of covariates d, confounding degree of instrument ω_1 this Table shows the area under the ROC curve (AUC), the average of $\delta_B^{\mathbb{1}}$ over all $M = 500$ Monte Carlo draws $\bar{\delta}_B^{\mathbb{1}}$, the empirical rejection probability for $\alpha = 0.05$ as well as the empirical rejection rate under H_0, i.e. when $\omega_1 = 0$. Simulation Regime 1: violation of exclusion restriction. $\omega_2 = 0.6$, $\omega_3 = 0.5$, $Var(\varepsilon_Y) = Var(\varepsilon_T) = 0.5$, $B = 200$.*

n	d	ω_1	AUC	$\bar{\delta}_B^{\mathbb{1}}$	erp	erp under H_0
500	10	0.1	0.674	0.056	0.144	0.036
		0.2	0.817	0.137	0.278	0.036
		0.3	0.916	0.270	0.530	0.036
		0.4	0.969	0.399	0.758	0.036
		0.5	0.980	0.472	0.864	0.036
	20	0.1	0.721	0.047	0.104	0.014
		0.2	0.889	0.164	0.338	0.014
		0.3	0.964	0.289	0.634	0.014
		0.4	0.988	0.392	0.824	0.014
		0.5	0.997	0.493	0.950	0.014
1000	10	0.1	0.679	0.048	0.200	0.074
		0.2	0.813	0.146	0.426	0.074
		0.3	0.914	0.270	0.664	0.074
		0.4	0.961	0.406	0.842	0.074
		0.5	0.981	0.496	0.952	0.074
	20	0.1	0.740	0.044	0.182	0.020
		0.2	0.913	0.180	0.558	0.020
		0.3	0.974	0.316	0.812	0.020
		0.4	0.989	0.406	0.928	0.020
		0.5	0.994	0.498	0.968	0.020

Table 4.8: *For combinations of number of observations n, number of covariates d, confounding degree of instrument ω_1 this Table shows the area under the ROC curve (AUC), the average of $\delta_B^{\mathbb{1}}$ over all $M = 500$ Monte Carlo draws $\bar{\delta}_B^{\mathbb{1}}$, the empirical rejection probability for $\alpha = 0.05$ as well as the empirical rejection rate under H_0, i.e. when $\omega_1 = 0$. Simulation Regime 1: violation of exclusion restriction. $\omega_2 = 0.6$, $\omega_3 = 0.5$, $Var(\varepsilon_Y) = Var(\varepsilon_T) = 1$, $B = 200$.*

n	d	ω_1	AUC	$\bar{\delta}_B^{\mathbb{1}}$	*erp*	*erp* under H_0
500	10	0.1	0.575	0.026	0.080	0.018
		0.2	0.778	0.128	0.264	0.018
		0.3	0.874	0.221	0.458	0.018
		0.4	0.959	0.353	0.722	0.018
		0.5	0.976	0.436	0.832	0.018
	20	0.1	0.653	0.034	0.072	0.012
		0.2	0.833	0.135	0.274	0.012
		0.3	0.929	0.234	0.546	0.012
		0.4	0.981	0.340	0.790	0.012
		0.5	0.991	0.430	0.920	0.012
1000	10	0.1	0.653	0.035	0.156	0.042
		0.2	0.792	0.127	0.374	0.042
		0.3	0.888	0.224	0.556	0.042
		0.4	0.950	0.354	0.784	0.042
		0.5	0.982	0.446	0.888	0.042
	20	0.1	0.686	0.040	0.190	0.042
		0.2	0.865	0.136	0.430	0.042
		0.3	0.945	0.250	0.706	0.042
		0.4	0.978	0.345	0.878	0.042
		0.5	0.991	0.440	0.970	0.042

Table 4.9: *For combinations of number of observations n, number of covariates d, confounding degree of instrument ω_1 this Table shows the area under the ROC curve (AUC), the average of $\delta_B^{\mathbb{1}}$ over all $M = 500$ Monte Carlo draws $\bar{\delta}_B^{\mathbb{1}}$, the empirical rejection probability for $\alpha = 0.05$ as well as the empirical rejection rate under H_0, i.e. when $\omega_1 = 0$. Simulation Regime 1: violation of exclusion restriction. $\omega_2 = 0.6$, $\omega_3 = 0.5$, $Var(\varepsilon_Y) = Var(\varepsilon_T) = 1.5$, $B = 200$.*

n	d	ω_1	AUC	$\bar{\delta}_B^{\mathbb{1}}$	*erp*	*erp* under H_0
500	10	0.1	0.610	0.046	0.104	0.016
		0.2	0.766	0.129	0.274	0.016
		0.3	0.866	0.223	0.488	0.016
		0.4	0.934	0.326	0.700	0.016
		0.5	0.960	0.393	0.806	0.016
	20	0.1	0.643	0.038	0.072	0.012
		0.2	0.818	0.128	0.302	0.012
		0.3	0.915	0.217	0.550	0.012
		0.4	0.974	0.310	0.770	0.012
		0.5	0.983	0.372	0.888	0.012
1000	10	0.1	0.610	0.039	0.182	0.064
		0.2	0.774	0.134	0.408	0.064
		0.3	0.873	0.233	0.594	0.064
		0.4	0.938	0.322	0.768	0.064
		0.5	0.962	0.395	0.850	0.064
	20	0.1	0.691	0.039	0.226	0.024
		0.2	0.832	0.111	0.424	0.024
		0.3	0.933	0.218	0.702	0.024
		0.4	0.981	0.302	0.876	0.024
		0.5	0.994	0.380	0.966	0.024

Table 4.10: *For combinations of number of observations n, number of covariates d, confounding degree of instrument ω_1 this Table shows the area under the ROC curve (AUC), the average of $\delta_B^{\mathbb{1}}$ over all $M = 500$ Monte Carlo draws $\bar{\delta}_B^{\mathbb{1}}$, the empirical rejection probability for $\alpha = 0.05$ as well as the empirical rejection rate under H_0, i.e. when $\omega_1 = 0$. Simulation Regime 1: violation of exclusion restriction. $\omega_2 = 0.3$, $\omega_3 = 0.5$, $Var(\varepsilon_Y) = Var(\varepsilon_T) = 0.5$, $B = 200$.*

n	d	ω_1	AUC	$\bar{\delta}_B^{\mathbb{1}}$	*erp*	*erp* under H_0
500	10	0.1	0.731	0.113	0.076	0.008
		0.2	0.925	0.324	0.400	0.008
		0.3	0.987	0.488	0.802	0.008
		0.4	0.998	0.597	0.960	0.008
		0.5	1.000	0.656	0.996	0.008
	20	0.1	0.802	0.118	0.048	0.006
		0.2	0.958	0.312	0.358	0.006
		0.3	0.996	0.488	0.844	0.006
		0.4	0.999	0.584	0.974	0.006
		0.5	1.000	0.640	1.000	0.006
1000	10	0.1	0.782	0.148	0.210	0.018
		0.2	0.952	0.357	0.624	0.018
		0.3	0.995	0.547	0.942	0.018
		0.4	1.000	0.649	0.996	0.018
		0.5	1.000	0.699	0.996	0.018
	20	0.1	0.835	0.152	0.200	0.010
		0.2	0.981	0.370	0.682	0.010
		0.3	0.998	0.546	0.966	0.010
		0.4	0.999	0.628	0.998	0.010
		0.5	0.999	0.684	1.000	0.010

Table 4.11: *For combinations of number of observations n, number of covariates d, confounding degree of instrument ω_1 this Table shows the area under the ROC curve (AUC), the average of $\delta_B^{\mathbb{1}}$ over all $M = 500$ Monte Carlo draws $\bar{\delta}_B^{\mathbb{1}}$, the empirical rejection probability for $\alpha = 0.05$ as well as the empirical rejection rate under H_0, i.e. when $\omega_1 = 0$. Simulation Regime 1: violation of exclusion restriction. $\omega_2 = 0.3$, $\omega_3 = 0.5$, $Var(\varepsilon_Y) = Var(\varepsilon_T) = 1.5$, $B = 200$.*

n	d	ω_1	AUC	$\bar{\delta}_B^{\mathbb{1}}$	*erp*	*erp* under H_0
500	10	0.1	0.670	0.100	0.058	0.014
		0.2	0.870	0.250	0.312	0.014
		0.3	0.956	0.367	0.666	0.014
		0.4	0.978	0.454	0.840	0.014
		0.5	0.996	0.529	0.952	0.014
	20	0.1	0.699	0.100	0.062	0.010
		0.2	0.900	0.217	0.266	0.010
		0.3	0.977	0.357	0.704	0.010
		0.4	0.993	0.420	0.928	0.010
		0.5	0.995	0.486	0.970	0.010
1000	10	0.1	0.700	0.104	0.168	0.006
		0.2	0.919	0.293	0.540	0.006
		0.3	0.967	0.419	0.818	0.006
		0.4	0.995	0.515	0.942	0.006
		0.5	0.998	0.577	0.988	0.006
	20	0.1	0.745	0.110	0.148	0.008
		0.2	0.951	0.280	0.578	0.008
		0.3	0.994	0.411	0.924	0.008
		0.4	0.998	0.476	0.986	0.008
		0.5	0.999	0.550	0.998	0.008

Table 4.12: *For combinations of number of observations n, number of covariates d, confounding degree of instrument ω_1 this Table shows the area under the ROC curve (AUC), the average of $\delta_B^{\mathbb{1}}$ over all $M = 500$ Monte Carlo draws $\bar{\delta}_B^{\mathbb{1}}$, the empirical rejection probability for $\alpha = 0.05$ as well as the empirical rejection rate under H_0, i.e. when $\omega_1 = 0$. Simulation Regime 1: violation of exclusion restriction. $\omega_2 = 0.6$, $\omega_3 = 0.25$, $Var(\varepsilon_Y) = Var(\varepsilon_T) = 1$, $B = 200$.*

n	d	ω_1	AUC	$\bar{\delta}_B^{\mathbb{1}}$	erp	erp under H_0
500	10	0.25	0.839	0.206	0.426	0.024
		0.50	0.979	0.443	0.856	0.024
	20	0.25	0.912	0.217	0.524	0.026
		0.50	0.995	0.464	0.956	0.026
1000	10	0.25	0.844	0.201	0.562	0.062
		0.50	0.982	0.467	0.910	0.062
	20	0.25	0.903	0.215	0.664	0.054
		0.50	0.993	0.465	0.988	0.054

Table 4.13: *For combinations of number of observations n, number of covariates d, confounding degree of instrument ω_1 this Table shows the area under the ROC curve (AUC), the average of $\delta_B^{\mathbb{1}}$ over all $M = 500$ Monte Carlo draws $\bar{\delta}_B^{\mathbb{1}}$, the empirical rejection probability for $\alpha = 0.05$ as well as the empirical rejection rate under H_0, i.e. when $\omega_1 = 0$. Simulation Regime 1: violation of exclusion restriction. $\omega_2 = 0.6$, $\omega_3 = 0.75$, $Var(\varepsilon_Y) = Var(\varepsilon_T) = 1$, $B = 200$.*

n	d	ω_1	AUC	$\bar{\delta}_B^{\mathbb{1}}$	erp	erp under H_0
500	10	0.25	0.853	0.177	0.342	0.010
		0.50	0.986	0.414	0.814	0.010
	20	0.25	0.894	0.162	0.316	0.010
		0.50	0.996	0.411	0.932	0.010
1000	10	0.25	0.870	0.186	0.482	0.032
		0.50	0.986	0.420	0.880	0.032
	20	0.25	0.919	0.174	0.522	0.014
		0.50	0.997	0.427	0.966	0.014

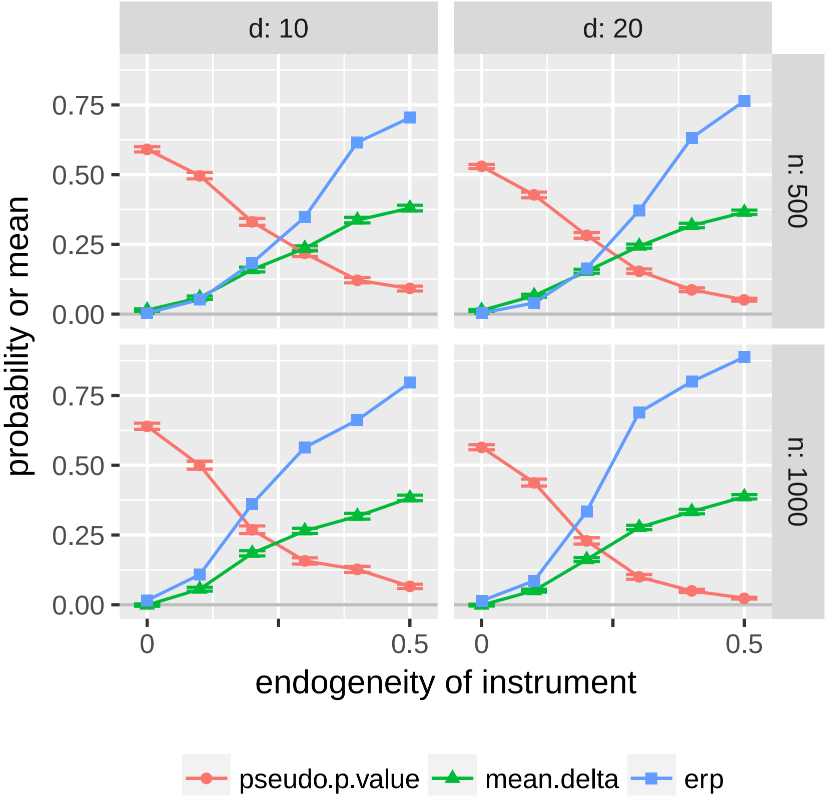


Figure 4.6: *This Figure shows the pseudo-p-value, δ_B, and the empirical rejection probability (based on the pseudo-p-value with threshold parameter $\alpha = 0.05$) as a function of the degree of instrument endogeneity where the source of confounding is a* ***violation of the exchangeability assumption****, 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.*

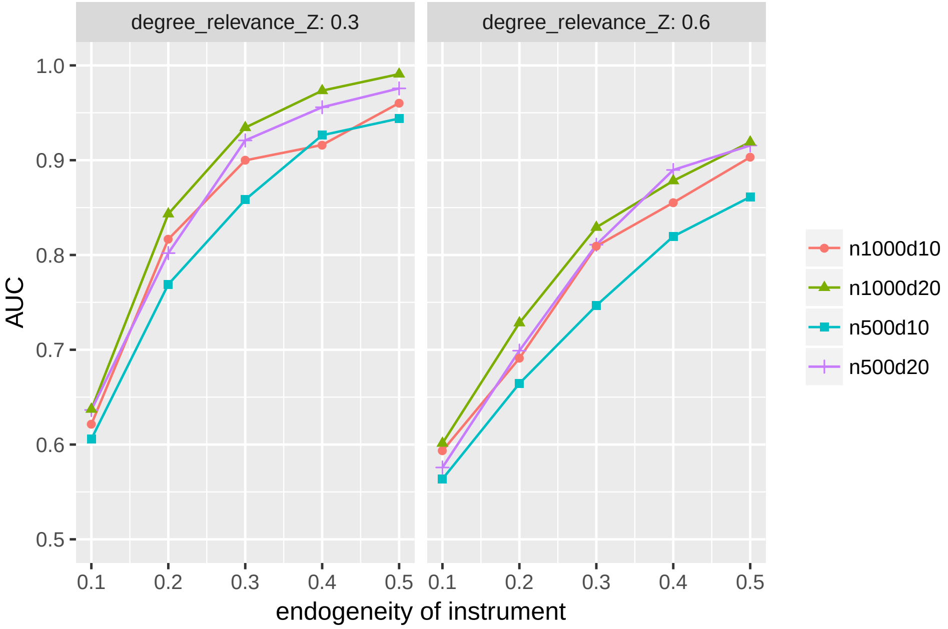


Figure 4.7: AUC curves for violations of the exchangeability assumption. *This Figure shows the area under the ROC curve (AUC) as a function of the degree of instrument endogeneity where the source of confounding is a* ***violation of the exchangeability assumption****, for various combinations of number 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.*

4.10.5 ROC curves

ROC curves are an insightful way to evaluate the performance of a binary classifier (valid vs. invalid 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, i.e. 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, i.e. rejections of H_0, across M Monte Carlo draws in which H_0 is in fact true (i.e. the instrument valid). 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 (i.e. the

Table 4.14: *For combinations of number of observations n, number of covariates d, confounding degree of instrument ω_1 this Table shows the area under the ROC curve (AUC), the average of $\delta_B^{\mathbb{1}}$ over all $M = 500$ Monte Carlo draws $\bar{\delta}_B^{\mathbb{1}}$, the empirical rejection probability for $\alpha = 0.05$ as well as the empirical rejection rate under H_0, i.e. when $\omega_1 = 0$. Simulation Regime 2: violation of exchangeability assumption. $\omega_2 = 0.3$, $\omega_3 = 0.5$, $Var(\varepsilon_Y) = Var(\varepsilon_T) = 1$, $B = 200$.*

n	d	ω_1	AUC	$\bar{\delta}_B^{\mathbb{1}}$	*erp*	*erp* under H_0
500	10	0.1	0.606	0.059	0.052	0.004
		0.2	0.769	0.160	0.184	0.004
		0.3	0.858	0.236	0.348	0.004
		0.4	0.926	0.337	0.616	0.004
		0.5	0.944	0.380	0.704	0.004
	20	0.1	0.636	0.066	0.040	0.004
		0.2	0.802	0.154	0.164	0.004
		0.3	0.921	0.244	0.372	0.004
		0.4	0.956	0.318	0.632	0.004
		0.5	0.976	0.365	0.764	0.004
1000	10	0.1	0.621	0.056	0.108	0.016
		0.2	0.817	0.184	0.362	0.016
		0.3	0.900	0.265	0.564	0.016
		0.4	0.916	0.317	0.662	0.016
		0.5	0.960	0.383	0.796	0.016
	20	0.1	0.637	0.051	0.086	0.014
		0.2	0.843	0.162	0.334	0.014
		0.3	0.934	0.277	0.690	0.014
		0.4	0.973	0.334	0.800	0.014
		0.5	0.991	0.387	0.888	0.014

Table 4.15: *For combinations of number of observations n, number of covariates d, confounding degree of instrument ω_1 this Table shows the area under the ROC curve (AUC), the average of $\delta_B^{\mathbb{1}}$ over all $M = 500$ Monte Carlo draws $\bar{\delta}_B^{\mathbb{1}}$, the empirical rejection probability for $\alpha = 0.05$ as well as the empirical rejection rate under H_0, i.e. when $\omega_1 = 0$. Simulation Regime 2: violation of exchangeability assumption. $\omega_2 = 0.6$, $\omega_3 = 0.5$, $Var(\varepsilon_Y) = Var(\varepsilon_T) = 1$, $B = 200$.*

n	d	ω_1	AUC	$\bar{\delta}_B^{\mathbb{1}}$	*erp*	*erp* under H_0
500	10	0.1	0.564	0.020	0.074	0.030
		0.2	0.664	0.054	0.134	0.030
		0.3	0.747	0.111	0.274	0.030
		0.4	0.820	0.173	0.406	0.030
		0.5	0.861	0.212	0.504	0.030
	20	0.1	0.576	0.011	0.052	0.018
		0.2	0.699	0.060	0.174	0.018
		0.3	0.811	0.128	0.340	0.018
		0.4	0.890	0.183	0.476	0.018
		0.5	0.916	0.220	0.592	0.018
1000	10	0.1	0.593	0.013	0.126	0.060
		0.2	0.691	0.054	0.232	0.060
		0.3	0.809	0.127	0.400	0.060
		0.4	0.855	0.177	0.516	0.060
		0.5	0.903	0.219	0.588	0.060
	20	0.1	0.601	0.012	0.136	0.048
		0.2	0.728	0.059	0.262	0.048
		0.3	0.829	0.121	0.460	0.048
		0.4	0.878	0.180	0.600	0.048
		0.5	0.919	0.216	0.714	0.048

instrument invalid):

$$\begin{aligned} \mathrm{FP}(\alpha) &= \frac{1}{M}\sum_{m=1}^{M}\psi_{\delta,m}(\alpha) \text{ when } \omega_1 \neq 0 \\ \mathrm{TP}(\alpha) &= \frac{1}{M}\sum_{m=1}^{M}\psi_{\delta,m}(\alpha) \text{ when } \omega_1 = 0. \end{aligned} \tag{4.67}$$

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 well as chance) and 1 (perfect accuracy).

4.10.6 Historical antecedents of the Principle of Independent Mechanisms

Since the algorithm proposed in Janzing and Schölkopf (2018) is justified by the Principle of Independent Mechanisms, which is called Independence between Cause and Mechanism in its bivariate version, it is instructive to have a brief look at the historical origins of this concept in econometrics.

A central problem in econometrics lies in identifying underlying economic relationships from observable data that are generated by these relationships. Pioneers of econometrics such as Frisch and Haavelmo worked on this problem in the mid-twentieth century and proposed important concepts of 'autonomy' and 'confluent relationships' in this context. 'Confluent' relations describe regularities that can be passively observed from the data. Autonomous relations, on the other hand, are those that are invariant to changes elsewhere in the system under study. Frisch et al. (1938) preface a memorandum introducing this work by stating that, "for any economic relation[] [. . .] I may ask: How autonomous is it? This question is extremely important. In one sense it is the most basic question one may raise in all sorts of econometric work" (p. 1). Haavelmo resorts to a mechanical analogy to illustrate the concept: if a man did not know anything about automobiles, and he wanted to understand how they work, we should not advise him to spend time and effort in measuring [the relationship between the pressure on the gas pedal and the corresponding speed]. Why? Because (1) such a relation leaves the whole inner mechanism of a car in complete mystery, and (2) such a relation might break down at any time, as soon as there is some disorder or change in any working part of the car. We say that such a relation has very little *autonomy*, because its existence depends upon the simultaneous fulfilment of a great many other relations, some of which are of a transitory nature. On the other hand, the general laws of thermodynamics, the dynamics of friction, etc., etc., are highly autonomous relations with respect to the automobile mechanism, because these relations describe the functioning of some parts of the mechanism irrespective of what happens in some other parts" (Haavelmo, 1944, pp. 27).[11] In addition to being more stable and more comprehensible, autonomous relations are essential for devising policy recommendations that rely on pinpointing those structures that remain invariant after a policy has changed.

It is instructive to look at the problem in a bivariate, linear setting to understand both why it is difficult to establish causal relations from observational data and how progress may nevertheless be achieved. Consider a system of two variables X and Y:

$$\begin{aligned} Y &= \theta X + \varepsilon_y \\ X &= \varepsilon_x \end{aligned} \tag{4.68}$$

where θ is a parameter, ε_y, ε_x are independent Gaussian noise variables. Simon (1953) argues that there is an implicit causal order in such a system because one needs to know X in order to know Y, yet one does not need to know Y to know about X. However, one can write a statistically equivalent model with completely reversed order as

$$\begin{aligned} X &= \delta Y + \omega_x \\ Y &= \omega_y \end{aligned} \tag{4.69}$$

where the parameters in the second system are calibrated such that the errors, ω_3 and ω_2, in system (4.69) are also independent: $\delta = \frac{\theta Var(\varepsilon_x)}{\theta^2 Var(\varepsilon_x) + Var(\varepsilon_y)}$, $\omega_y = \varepsilon_y + \theta\varepsilon_x$, $\omega_x = (1-\delta\theta)\varepsilon_x - \delta\varepsilon_y$ (Hoover, 2008). In such a bivariate linear Gaussian setting, systems (4.68) and (4.69) cannot be distinguished based on observational data alone without making further assumptions.

This observational equivalence of the system where X is causing Y and the system where Y is causing X is the root of the identification problem. The structural approach to causality represented by Frisch and Haavelmo at the Cowles Commission argues that this undecidability can only be resolved by means of substantive (economic) theory. Though generally sympathetic to the Cowles Commission's conceptualization of causality, Simon (1953) proposes a different approach to resolving the issue. He argues that one can deduce the causal direction by detecting invariance of conditional distributions analyzing either controlled or natural experiments without relying on economic theory.[12] Namely, imagine

[11] Haavelmo credits Frisch (1938) for coining the term 'autonomy.'

[12] Since instrumental variables are basically packaged natural experiments, much of the progress made under the umbrella of the credibility revolution in empirical economics is, in spirit, wedded to Simon's approach sketched out here (see Hoover, 2008).

an experiment that would alter the marginal distribution of X without altering the conditional distribution of $Y|X$. These alterations are only possible in a system defined by eqs. (4.68): altering the marginal distribution of X amounts to changing $Var(\varepsilon_x)$. This does not translate into a change in the conditional distribution of $Y|X$. If, on the other hand, the observations were guided by the system in eqs. (4.69), the change in $Var(\varepsilon_x)$ would influence ω_y and the distribution of $Y|X$ in turn. Therefore, one can deduce that X is causing Y.[13]

Thus, the idea of autonomous relations relates to the invariability of causal relations upon intervening on the cause. To give another example, the underlying causal structure between 'smoking' and 'lung cancer' does not change upon varying the number of cigarettes smoked. This is precisely why we expect to be able to change lung cancer incidence after intervening on smoking habits.[14]

Still, it seems that observational equivalence cannot be resolved by analyzing just a single data set. For Simon's approach to work, one would need two samples of the joint distribution of X and Y: one before and one after a (natural) experiment to analyze invariant structures. However, recent work that has been done largely outside of econometrics shows that progress is possible.[15]

Causal thinking is receiving increasing interest in the machine learning community (Peters et al., 2017). Early work on how to estimate causal parameters includes Pearl (2009) and Spirtes et al. (2000). Building on this seminal work, there are a number of proposals for how to resolve the aforementioned observational equivalence problem (Mooij et al., 2016; Shimizu et al., 2006; Hoyer et al., 2009). Building on the invariance principle expounded on by early Cowles Commission researchers, the overarching idea in this literature is that the causal structure of a system is composed of invariant mechanisms that do not inform or influence one another, and that are, therefore, mutually independent. Such a collection of invariant structural relations give rise to the independence of the conditional distribution of the effect given the cause and the marginal distribution of

[13] Simon's approach is also reflected in later work by Leamer, who defines exogeneity as follows: "If the observed conditional distribution of the variable y given a set of variables x is invariant under any modification of the system selected from a specified family of modifications that alter the process generating x, then the variables x are said to be *exogenous* to y." Leamer (1985, p. 262)

[14] This example is taken from Illari and Russo (2014, Section 10.3.2).

[15] It is worth mentioning a relevant argument due to Simon (1962) in this context. Simon discusses the "architecture of complexity" by first observing that there are two broad types of complexity: those that are characterized by a hierarchical structure and those that are not. He goes on to argue that hierarchical systems evolve more quickly than non-hierarchical ones and, therefore, are more common. The hierarchical structure of a complex system implies the possibility to characterize it as a collection of (nearly-)decomposable sub-structures. This (near-)decomposability, in turn, implies that intra-sub-structure mechanisms can be analyzed independently from other sub-structures – which is reminiscent of the Principle of Independent Mechanisms. Even in the presence of feedback or cyclical behavior, independence of sub-structures can be maintained by under mild conditions.

the cause: $f(\text{effect}|\text{cause}) \perp\!\!\!\perp f(\text{cause})$. Since the conditional distribution $f(\text{effect}|\text{cause})$ represents the causal mechanism, this independence is generally referred to as the Independence between Cause and Mechanism (ICM) (Peters et al., 2017). In other words, the ICM introduces a 'causal asymmetry' in the statistically symmetric factorization of the joint distribution of cause and effect:

$$\begin{aligned} f(\text{cause}, \text{effect}) &= f(\text{cause}|\text{effect}) f(\text{effect}) && (4.70)\\ &= \underbrace{f(\text{effect}|\text{cause})}_{\text{`mechanism'}} f(\text{cause}). && (4.71) \end{aligned}$$

Although statistically symmetric, the invariance of the causal process induces an independence between the mechanism and cause distributions in eqs. (4.71), which there is no reason to expect as well in the 'anticausal' direction, eqs. (4.70).[16]

This claim can be visualized with an example from Hoyer et al. (2009). Underlying

[16] An application of this reasoning in economics is found in Hoover (1990). He considers a setting in which money supply M causes price level P:

$$\begin{aligned} P &= aM + \varepsilon \quad \text{with } \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)\\ M &= b + \rho \qquad \text{with } \rho \sim \mathcal{N}(0, \sigma_\rho^2) \end{aligned}$$

with reduced form equations

$$P = ab + a\rho + \varepsilon \text{ and } M = b + \rho.$$

The joint distribution can be partitioned in two ways: $f(M,P) = f(M|P)f(P) = f(P|M)f(M)$. These distributions can be calculated more explicitly:

$$\begin{aligned} f(P|M) &= \mathcal{N}(aM, \sigma_\varepsilon^2)\\ f(M) &= \mathcal{N}(b, \sigma_\rho^2)\\ f(M|P) &= \mathcal{N}\left(\frac{a\sigma_\rho^2 P + b\sigma_\varepsilon^2}{a^2\sigma_\rho^2 + \sigma_\varepsilon^2}, \frac{\sigma_\varepsilon^2\sigma_\rho^2}{a^2\sigma_\rho^2 + \sigma_\varepsilon^2}\right)\\ f(P) &= \mathcal{N}(ab, a^2\sigma_\rho^2 + \sigma_\varepsilon^2) \end{aligned}$$

Now consider two changes to the system. First, the conduct of monetary policy changes, which means that either b or σ_ρ^2 changes. As a consequence, notice that $f(M|P)$ and $f(M)$, as well as $f(P)$, change, but crucially that $f(P|M)$ stays invariant, i.e. in the ('causally directed') factorization $f(P|M)f(M)$ only the latter component is affected, whereas in the ('anticausal') factorization $f(M|P)f(P)$ both components are affected. Second, the price setting procedure changes, which means that either a or σ_ε^2 changes. As a consequence, $f(M)$, $f(M|P)$, and $f(P)$ change, but $f(P|M)$ remains invariant. In the ('causally directed') factorization $f(P|M)f(M)$ only the first component is affected, whereas in the ('anticausal') factorization $f(M|P)f(P)$, again, both components are affected. As Hoover concludes, the "[causal] partition $f(P|M)f(M)$ is clearly more stable to Ill-defined interventions than the [anticausal] partition $f(M|P)f(P)$." The idea of a more stable partition in the causal direction links nicely with causal discovery methods based on complexity of the conditional distributions put forward by a.o. Peters and Bühlmann (2014) (for more references consult Mooij et al., 2016, p. 3). Those methods are based on the observation that the factorization of the joint distribution in the causal direction, $p(\text{effect}|\text{cause})p(\text{cause})$, yields models of lower total complexity (as for instance defined as Kolmogorov complexity). Note that the juxtaposition of causal and anticausal directions is described in the work of Schölkopf et al. (2012).

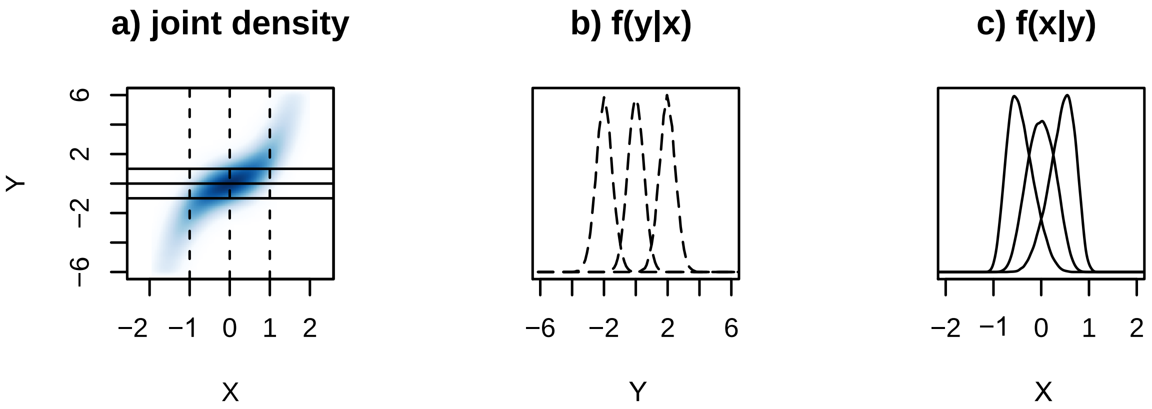


Figure 4.8: Illustration of the Independence between Cause and Mechanism postulate. *Panel a) shows the joint density $p(x, y)$. The horizontal and vertical lines indicate the levels of the conditioning variables. Panels b) and c) show the conditional densities $p(y|x)$ and $p(x|y)$, respectively. One can see by visual inspection that the shape of the conditional distribution $p(y|x)$ is invariant to the choice of x. However, the conditional distribution $p(x|y)$ does have a different shape depending on y. Following the modularity reasoning, this provides evidence for x causing y. Figure adapted from Hoyer et al. (2009).*

this example is the generative model

$$y = x + x^3 + \varepsilon \tag{4.72}$$

where x causes y, x and the error ε are independent following a Gaussian $\mathcal{N}(0; 0.5^2)$. This exercise allows one to obtain a visual intuition as to how causal identification by invariance works. As seen in the previous paragraph, the invariance argument rests on the independence between conditional distributions and purported causes. In particular, we expect an independence between the conditional distribution $p(\text{effect}|\text{cause})$ and the marginal $p(\text{cause})$. Since we know the causal model that has generated the data for this example ($x \to y$), we can check whether we see this independence relation exists in the simulated data. For this purpose, we visualize the joint distribution $p(x, y)$, as well as conditional distributions $p(x|y)$ and $p(y|x)$ in Figure 4.8. The shape of the conditional distribution $p(x|y)$ varies with y, which casts doubt on the hypothesis that $y \to x$. The shape of $p(y|x)$ is invariant to changes in the conditioning variable x, a feature that qualifies x as cause of y.

This concept closely resembles the definition of ‘super-exogeneity’ of Robert Engle et al. (1983). They define a variable z to be super-exogenous if its joint density with y factorizes as $f(y, z|\lambda) = f(y|z, \lambda_1)f(z|\lambda_2)$ and the conditional density $f(y|z, \lambda_1)$ is invariant to changes in the marginal density $f(z|\lambda_2)$ where λ, λ_1, and λ_2 are parameters.

4.10.7 Simulation for the illustration of PIM

The illustration in Figures 4.2 and 4.9 is based on the following simulation.

First, construct a covariance matrix Σ as follows. Draw $d+1$ eigenvalues

$$\lambda \sim \mathcal{U}(0.5, 1.5)$$

which populate the diagonal of a matrix V. Then I draw a random orthogonal matrix L and set $\Sigma = VLV^{\top}$. I multiply each element in the last row and last column of Σ by 5 to induce more unexplained variation in Y. For the unconfounded case, I replace the last row and last column of Σ with zeroes but leave the $(d+1, d+1)$ entry untouched:

$$\begin{aligned} \text{confounded: } S_c &= \Sigma_{d+1\times d+1} \\ \text{unconfounded: } S_u &= \begin{pmatrix} \Sigma_{(1:d)\times(1:d)} & \mathbf{0} \\ \mathbf{0} & \sigma_{d+1\times d+1} \end{pmatrix} \end{aligned} \tag{4.73}$$

I simulate data by drawing the structural error term ε_Y and $\mathbf{X}$ from a jointly normal distribution

$$\begin{pmatrix} \mathbf{X} \\ \varepsilon_Y \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, S_i) \tag{4.74}$$

where $i \in \{c, u\}$.

Next, draw the d-dimensional true parameter vector

$$\beta \sim \mathcal{N}(\mathbf{0}, diag(1))$$

and divide each element of β by $d^{0.5}$ (to keep the variance of Y comparable for different d).

Finally, set

$$Y = \mathbf{X}\beta + \varepsilon_Y. \tag{4.75}$$

We estimate $\hat{\beta}$ by OLS.

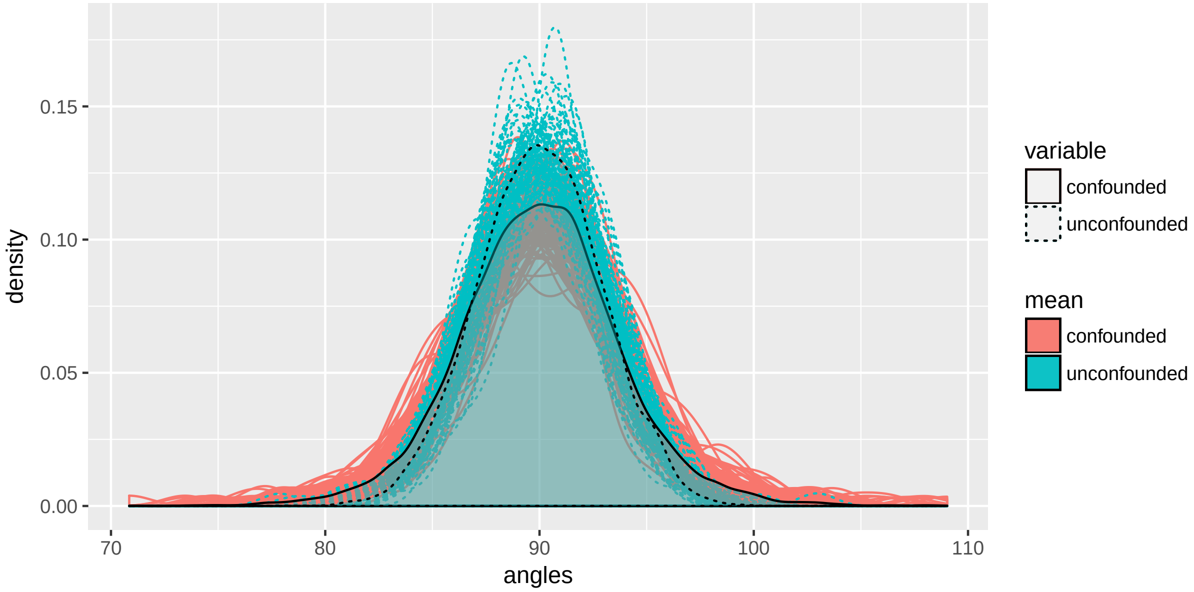


Figure 4.9: Illustration of genericity of causal parameter vectors. *This Figure shows density plots of the angles between the least-squares parameter vector of both confounded and unconfounded models with each of the d eigenvectors of the covariance matrix of the covariates. In the unconfounded model, the least-squares parameter vector should lie in generic orientation with respect to (the eigenspace spanned by the) eigenvectors of the covariance matrix of the covariates. Genericity of two vectors can be understood as their dot product being zero, or their angle being 90 degrees. As expected, therefore, the distribution of angles in the unconfounded case clusters around 90 degrees. Crucially, in the confounded case, the distribution of angles is considerable wider. A trace of confounding is thus reflected in the less generic angles of the confounded parameter vector w.r.t. the eigenvectors; their distribution is characterized by a more frequent divergence from the generic angle of 90 degrees. This illustrates the type of confounding signal that JS leverage in their methodology. The Figure shows angle distributions for 100 simulation runs with* $d = 100$*, and* $n = 50000$*, the respective means are depicted with black lines, solid for the confounded and dashed for the unconfounded case.*

5 Conclusion

The synergies between machine learning and, more generally, computer science and economics in the field of causal inference are wide-ranging. This dissertation focuses on two points of contact between the two fields. First, traditional machine learning algorithms can be used to complement econometric techniques for causal identification. Second, insights into causal modeling from the computer science community can be employed in economics.

The first paper falls into the former category. Carsten Schröder and I adapt the causal forest methodology proposed by Athey et al. (2019) to a difference-in-differences setting and analyze to what extent effect heterogeneities of the 2015 introduction of the minimum wage in Germany can be discerned in a data-driven manner. The Socio-economic panel (SOEP) serves as empirical basis. Two contributions are made. First, we show how the causal forest methodology can be applied in difference-in-differences settings. Second, we show that previously documented effect heterogeneities can be explained by interactions of other covariates. These interactions define subgroups of the population according to the level of treatment effect in a fine-grained manner. Such information is useful for policy-makers who wish to target measures complementary to the minimum wage or direct controls for non-compliance to least-benefiting groups.

Ultimately, the second and third paper are applications of Haavelmo's concept of the autonomy of causal relations (Haavelmo, 1944). Though useful as an intuitive guide as to what characterizes a causal relation and implicitly used to motivate existing techniques for causal inference, it has hitherto not been employed as an empirical tool for identification of causal relations *eo ipso*.

The second paper relies on a strand in the computer science literature that uses functional form restrictions to infer the causal direction between two random variables (Peters et al., 2014). Existing work shows that in models characterized by additively separable error terms and a nonlinear relation between cause and effect, the mechanism linking cause and effect is 'independent' of the cause. This implies an independence between the error term and the covariate. Christoph Breunig and I show how such reasoning can be used to address problems of reverse causality as one source of endogeneity, which is a central problem in econometric models that potentially invalidates estimates of causal effects. Existing tests of endogeneity often require that a potential solution in the form of instruments is available (see extensive body of work initiated by Hausman, 1978). In situations in which instruments are not available, a test that does not require them is needed. This paper presents a test for reverse causality of a single regressor without requiring instruments. The mean independence assumption is moved beyond and the error is required to be independent of the regressor, thus allowing us to infer the causal direction between the variables at hand. Advances on testing independence of random variables based on kernel-based procedures are leveraged. The contribution is twofold. First, we extend existing research from the computer science community on the identifiability of the causal direction by addressing heteroskedastic error structures and the presence of additional control variables. Second, we provide a test for reverse causality that does not rely on instruments.

While the second paper leverages the implications of structural autonomy in the bivariate case, the third paper does so in the multivariate case. Specifically, it builds on Janzing and Schölkopf (2018), who propose a method to estimate an overall degree of confounding in multivariate linear models. Given the often controversial identifying assumptions in instrumental variable models, whose justification is rarely statistically-grounded, such a method is a valuable addition to the empirical economics toolkit. I make two contributions. First, I address the limitation of Janzing and Schölkopf (2018) of providing an *overall* degree of confounding for the whole model and provide a way to use their method to estimate a degree of confounding of a *single* covariate in multivariate linear models. Second, I show how this method can be employed to test for instrument validity in instrumental variable models and provide an empirical application.

Thus, this dissertation contributes to the nascent literature on using ML techniques to answer questions in economics. A particular focus is on the potential of methods for causal identification developing in the ML community that have hitherto received scant attention in economics.

6 Bibliography

Acemoglu, Daron, Simon Johnson, James A Robinson, and Pierre Yared (2008). "Income and democracy". *American Economic Review* 98.3, pp. 808–42.

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.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb (2017). "Prediction, Judgment and Complexity". *NBER Conference: Economics of Artificial Intelligence (slides available at https://www.economicsofai.com/nber-conference-toronto-2017/).*

Ahlfeldt, Gabriel M, Duncan Roth, and Tobias Seidel (2018). "The regional effects of Germany's national minimum wage". *Economics Letters* 172, pp. 127–130.

Andrews, Donald W K, Marcelo J Moreira, James H Stock, and James H Stock (2006). "Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression". *Econometrica* 74.3, pp. 715–752.

Angrist, Joshua D, Guido W Imbens, and Donald B Rubin (1996). "Identification of Causal Effects Using Instrumental Variables". *Journal of the American Statistical Association* 91.434, pp. 444–455.

Athey, Susan and Guido Imbens (2016). "Recursive partitioning for heterogeneous causal effects". *Proceedings of the National Academy of Sciences of the United States of America* 113.27, pp. 7353–60. ISSN: 1091-6490. DOI: 10.1073/pnas.1510489113. arXiv: 1504.01132.

Athey, Susan, Julie Tibshirani, and Stefan Wager (2019). "Generalized Random Forests". *Annals of Statistics* 47.2, pp. 1148–1178. arXiv: 1610.01271. URL: http://arxiv.org/abs/1610.01271.

Bajari, Patrick, Denis Nekipelov, Stephen P Ryan, and Miaoyu Yang (2015). "Machine learning methods for demand estimation". *The American Economic Review* 105.5, pp. 481–485.

Baker, Charles R (1973). "Joint measures and cross-covariance operators". *Transactions of the American Mathematical Society* 186, pp. 273–289.

Balke, Alexander and Judea Pearl (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.

Bareinboim, Elias and H Paul (2019). “Causal Inference and Data-Fusion in Econometrics”. *arXiv*, pp. 1–46. URL: https://arxiv.org/abs/1912.09104.

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

Besserve, Michel, Rémy Sun, and Bernhard Schoelkopf (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.

Bonin, Holger et al. (2018). “Auswirkungen des gesetzlichen Mindestlohns auf Beschäftigung, Arbeitszeit und Arbeitslosigkeit, Studie im Auftrag der Mindestlohnkommission”. *Forschungsinstitut zur Zukunft der Arbeit, Evaluation Office Caliendo, Deutsches Institut für Wirtschaftforschung and others.*

Bossler, Mario and Hans-Dieter Gerner (2016). *Employment effects of the new German minimum wage: Evidence from establishment-level micro data.* Tech. rep. IAB-Discussion paper.

Breiman, Leo (2001). “Random forests”. *Machine Learning* 45.1, pp. 5–32. ISSN: 08856125. DOI: 10.1023/A:1010933404324. arXiv: /dx.doi.org/10.1023{\%}2FA{\%}3A1010933404324 [http:].

Breiman, Leo, Jerome Friedman, Richard Olshen, and Charles Stone (1984). *Classification and regression trees.* Wadsworth Brooks.

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

— (2020). “Specification testing in nonparametric instrumental quantile regression”. *forthcoming in Econometric Theory.*

Breunig, Christoph and Patrick Burauel (2020). “A Reverse Causality Test Without Instruments”. *Mimeo.*

Brochu, Pierre, David A Green, Thomas Lemieux, and James Townsend (2015). *The minimum wage, turnover, and the shape of the wage distribution.* Tech. rep. Unpublished.

Burauel, Patrick (2020). “What the Degree of Structural Autonomy Can Say about Instrument Validity”. URL: https://dx.doi.org/10.2139/ssrn.3344981.

Burauel, Patrick and Carsten Schröder (2019). “The German Minimum Wage and Wage Growth: Heterogeneous Treatment Effects Using Causal Forests”. *Available at SSRN 3415479*. URL: https://dx.doi.org/10.2139/ssrn.3415479.

Burauel, Patrick, Marco Caliendo, Markus Grabka, Cosima Obst, Malte Preuss, and Carsten Schröder (2018). “Auswirkungen des gesetzlichen Mindestlohns auf die Lohnstruktur, Studie im Auftrag der Mindestlohnkommission”. *Mimeo.*

Burauel, Patrick, Marco Caliendo, Markus Grabka, Cosima Obst, Malte Preuss, Carsten Schröder, and Cortnie Shupe (2020). “The Impact of the German Minimum Wage on Individual Wages and Monthly Earnings”. *Jahrbücher fur Nationalökonomie und Statistik* 240.2-3, pp. 1–31. ISSN: 00214027. DOI: 10.1515/jbnst-2018-0077.

Caliendo, Marco, Alexandra Fedorets, Malte Preuss, Carsten Schröder, and Linda Wittbrodt (2017). “The short-term distributional effects of the German minimum wage reform”. *SOEPpaper No. 948.*

— (2018). “The short-run employment effects of the German minimum wage reform”. *Labour Economics* 53, pp. 46–62.

Caliendo, Marco, Carsten Schröder, and Linda Wittbrodt (2019). “The Causal Effects of the Minimum Wage Introduction in Germany–An Overview”. *German Economic Review.*

Cannas, Massimo and Bruno Arpino (2019). “A comparison of machine learning algorithms and covariate balance measures for propensity score matching and weighting”. *Biometrical Journal* 61.4, pp. 1049–1072. ISSN: 15214036. DOI: 10.1002/bimj.201800132.

Card, David (1995). “Using Geographic Variation in College Proximity to Estimate the Return to Schooling”. *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp*. Ed. by Loizos Christofides, Kenneth Grant, and Robert Swidinsky. Toronto: University of Toronto Press, pp. 201–222. URL: http://www.nber.org/papers/w4483.

Carrasco, Marine and Jean-Pierre Florens (2000). “Generalization of GMM to a Continuum of Moment Conditions”. *Econometric Theory* 16.6, pp. 797–834.

Carrasco, Marine, Jean-Pierre Florens, and Eric Renault (2007). “Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization”. *Handbook of Econometrics*. Vol. 6. Elsevier. Chap. 7.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik (2008). “Nonparametric tests for treatment effect heterogeneity”. *The Review of Economics and Statistics* 90.3, pp. 389–405.

Darmois, George (1953). “Analyse générale des liaisons stochastiques: etude particulière de l’analyse factorielle linéaire”. *Revue de l’Institut international de statistique*, pp. 2–8.

Daudin, B Y J J (1980). “Partial association measures and an application to qualitative regression”. *Biometrika* 67.3, pp. 581–590.

Engle, Robert F. and David F. Hendry (1993). “Testing superexogeneity and invariance in regression models”. *Journal of Econometrics* 56.1-2, pp. 119–139. ISSN: 03044076. DOI: 10.1016/0304-4076(93)90103-C. arXiv: arXiv:1011.1669v3.

Favero, Carlo and David F. Hendry (1992). “Testing the lucas critique: A review”. *Econometric Reviews* 11.3, pp. 265–306. ISSN: 15324168. DOI: 10.1080/07474939208800238.

Fève, Frédérique, Jean-Pierre Florens, and Ingrid Van Keilegom (2018). “Estimation of conditional ranks and tests of exogeneity in nonparametric nonseparable models”. *Journal of Business & Economic Statistics* 36.2, pp. 334–345.

Flaxman, Seth R, Yu-xiang Wang, and Alexander J Smola (2015). “Who Supported Obama in 2012 ? Ecological Inference through Distribution Regression”. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Frisch, Ragnar (1938). “Statistical versus Theoretical Relations in Economic Macro-Dynamics”. *Mimeographed memorandum prepared for the Business Cycle Conferecne at Cambridge, England, July 18-20, 1938.*

Frisch, Ragnar, Trygve Haavelmo, T.C. Koopmans, and J. Tinbergen (1938). “Autonomy of Economic Relations”. *League of Nations Memorandum.*

Fukumizu, Kenji, Francis R Bach, and Michael I Jordan (2004). “Dimensionality Reduction for Supervised Learning with Reproducing Kernel Hilbert Spaces”. *Journal of Machine Learning Research* 5, pp. 73–99.

Fukumizu, Kenji, Arthur Gretton, and Bernhard Schölkopf (2008). “Kernel Measures of Conditional Dependence”. *Advances in Neural Information Processing Systems*, pp. 1–8.

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.

Goebel, Jan, Markus M. Grabka, Stefan Liebig, Martin Kroh, David Richter, Carsten Schröder, and Jürgen Schupp (2018). “The German Socio-Economic Panel (SOEP)”. *Journal of Economics and Statistics.*

Goodfellow, Ian; Yoshua Bengio; and Aaron Courville (2016). *Deep learning.* ISBN: 9780521835688. DOI: 10.1038/nmeth.3707. arXiv: arXiv:1312.6184v5. URL: http://goodfeli.github.io/dlbook/{\%}0Ahttp://dx.doi.org/10.1038/nature14539.

Gorban, A. N. and I. Y. Tyukin (2018). “Blessing of dimensionality: Mathematical foundations of the statistical physics of data”. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376.2118. ISSN: 1364503X. DOI: 10.1098/rsta.2017.0237. arXiv: 1801.03421.

Götze, Friedrich and Alexander Tikhomirov (2004). “Rate of convergence in probability to the Marchenko – Pastur law”. *Bernoulli* 10.3, pp. 503–548.

Gretton, Arthur, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf (2005). "Measuring statistical dependence with Hilbert-Schmidt norms". *International conference on algorithmic learning theory*. Springer, pp. 63–77.

Gretton, Arthur, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, and Alexander J. Smola (2007). "A kernel statistical test of independence". *Neural Information Processing Systems*, pp. 585–592. URL: http://eprints.pascal-network.org/archive/00004335/.

Grünewälder, Steffen, Guy Lever, Luca Baldassarre, Sam Patterson, Arthur Gretton, and Massimiliano Pontil (2012). "Conditional mean embeddings as regressors". *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pp. 1823–1830. arXiv: 1205.4656.

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.

Hartford, Jason, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy (2016). "Counterfactual Prediction with Deep Instrumental Variables Networks". *ArXiv e-prints*. arXiv: 1612.09596. URL: http://arxiv.org/abs/1612.09596.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning*. 2nd editio. New York: Springer.

Hausman, Jerry A (1978). "Specification tests in econometrics". *Econometrica*, pp. 1251–1271.

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.

Heckman, James J, Lance J Lochner, and Petra E Todd (2006). "Earnings functions, rates of return and treatment effects: The Mincer equation and beyond". *Handbook of the Economics of Education* 1, pp. 307–458.

Hendry, David F and Carlos Santos (2010). "Automatic Tests of Super Exogeneity". *Volatility and Time Series Econometrics, Essays in Honor of Robert Engle*, pp. 1–44.

Hoderlein, Stefan, Jussi Klemelä, and Enno Mammen (2010). "Analyzing the Random Coefficient Model Nonparametrically". *Econometric Theory* 26.3, pp. 804–837. DOI: 10.1017/S0266466609990107.

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

Holmes, Michael and Mark Caiola (2018). "Invariance properties for the error function used for multilinear regression". *PLoS ONE*, pp. 1–25.

Hoover, Kevin D. (1990). “The Logic of Causal Inference: Econometrics and the Conditional Analysis of Causation”. *Economics and Philosophy* 6.02, p. 207. ISSN: 0266-2671. DOI: 10.1017/S026626710000122X.

Hoover, Kevin D (2008). “causality in economics and econometrics”. *The New Palgrave Dictionary of Economics*. Ed. by Steven N Durlauf and Lawrence E Blume. Basingstoke: Palgrave Macmillan.

Hoyer, Patrik, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf (2009). “Nonlinear causal discovery with additive noise models”. *Advances in Neural Information Processing Systems*, pp. 689–696. ISSN: 15337928. DOI: 10.1.1.144.4921. arXiv: arXiv:1309.6779v1. URL: http://eprints.pascal-network.org/archive/00005377/.

Hsiao, Cheng and M Hashem Pesaran (2008). “Random Coefficient Models”. *The econometrics of panel data*. Chap. Chapter 6, pp. 185–213.

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.

Illari, Phyllis and Federica Russo (2014). *Causality: Philosophical theory meets scientific practice*. Oxford University Press.

Imbens, Guido (2014a). “Instrumental Variables : An Econometrician’s Perspective”. *Statistical Science* 29.3, pp. 323–358. DOI: 10.1214/14-STS480.

— (2014b). “Rejoinder”. *Statistical Science* 29.3, pp. 375–379. ISSN: 08834237, 21688745. URL: http://www.jstor.org/stable/43288516.

— (2019). “Potential Outcome and Directed Acyclic Graph Approaches to Causality : Relevance for Empirical Practice in Economics”. *arXiv*. arXiv: arXiv:1907.07271v1.

Janzing, Dominik and Bernhard Schölkopf (2010). “Causal inference using the algorithmic Markov condition”. *IEEE Transactions on Information Theory* 56.10, pp. 5168–5194. ISSN: 00189448. DOI: 10.1109/TIT.2010.2060095. arXiv: 0804.3678.

— (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.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer (2015). "Prediction Policy Problems". *American Economic Review: Papers & Proceedings* 105.5, pp. 491–495. ISSN: 0002-8282. DOI: `10.1257/aer.p20151023`. arXiv: `15334406`. URL: `http://pubs.aeaweb.org/doi/10.1257/aer.p20151023`.

Kun Zhang, J. Peters, D. Janzing, and B. Schölkopf (2011). "Kernel-based Conditional Independence Test and Application in Causal Discovery". *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pp. 804–813. arXiv: `1202.3775`. URL: `http://www.is.tuebingen.mpg.defileadmin/user{\_}upload/files/publications/2011/UAI-2011-Zhang.pdf`.

Lacetera, Nicola, Devin G Pope, and Justin R Sydnor (2012). "Heuristic thinking and limited attention in the car market". *American Economic Review* 102.5, pp. 2206–36.

Leamer, Edward E. (1985). "Vector Autoregressions for Causal Inference?" *Carnegie-Rochester Conference Series on Public Policy* 22, pp. 255–304. ISSN: 0167-2231. DOI: `DOI:10.1016/0167-2231(85)90035-1`.

Lemeire, Jan and Dominik Janzing (2013). "Replacing causal faithfulness with algorithmic independence of conditionals". *Minds and Machines* 23.2, pp. 227–249. ISSN: 09246495. DOI: `10.1007/s11023-012-9283-1`.

Lopez-Paz, David (2016). "From Dependence to Causation". PhD thesis. Max Planck Institute for Intelligent Systems and University of Cambridge. URL: `https://arxiv.org/pdf/1607.03300v1`.

Marchenko, V. A. and L. A Pastur (1967). "Distribution of eigenvalues for some sets of random matrices". *Matematicheskii Sbornik* 114.4, pp. 507–536.

Mooij, Joris, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf (2016). "Distinguishing cause from effect using observational data: methods and benchmarks". *Journal of Machine Learning Research* 17.April, pp. 1–102. arXiv: `arXiv:1412.3773v1`.

Mooij, Joris M., Dominik Janzing, Tom Heskes, and Bernhard Schölkopf (2011). "On causal discovery with cyclic additive noise models". *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, NIPS 2011*, pp. 1–9.

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`.

Muandet, Krikamol, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf (2016). "Kernel Mean Embedding of Distributions: A Review and Beyonds". *arXiv preprint arXiv:1605.09522.*

Mullainathan, Sendhil and Jann Spiess (2017). “Machine Learning: An Applied Econometric Approach”. *Journal of Economic Perspectives* 31.2, pp. 87–106. ISSN: 0895-3309. DOI: `10.1257/jep.31.2.87`.

Neumark, David and William Wascher (2004). “The Influence of Labour Market Institutions on the Disemployment Effects of the Minimum Wage”. *CESifo DICE Report* 2.2, pp. 40–47.

Nowzohour, Christopher and Peter Bühlmann (2016). “Score-based causal learning in additive noise models”. *Statistics* 50.3, pp. 471–485.

Oster, Emily (2019). “Unobservable Selection and Coefficient Stability : Theory and Evidence Unobservable Selection and Coefficient Stability : Theory and Evidence”. *Journal of Business and Economic Statistics* 37.2, pp. 187–204. ISSN: 0735-0015. DOI: `10.1080/07350015.2016.1227711`. URL: `https://doi.org/10.1080/07350015.2016.1227711`.

Pearl, Judea (2009). *Causality: Models, Resoning, and Inference*. Cambridge University Press.

Pen, Jan (1971). *Income Distribution*. London: Allen Lane: The Penguin Press.

Peters, J. and P. Bühlmann (2014). “Identifiability of Gaussian structural equation models with equal error variances”. *Biometrika* 101.1, pp. 219–228. ISSN: 00063444. DOI: `10.1093/biomet/ast043`. arXiv: `1205.2536`.

Peters, Jonas (2008). “Asymmetries of time series under inverting their direction”. *Diploma thesis* 9, p. 29.

— (2012). “Restricted structural equation models for causal inference”. PhD thesis. ETH Zürich.

Peters, Jonas, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf (2014). “Causal discovery with continuous additive noise models.” *Journal of Machine Learning Research* 15.1, pp. 2009–2053.

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`.

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

Pfister, Niklas, Peter Bühlmann, Bernhard Schölkopf, and Jonas Peters (2018). “Kernel-based tests for joint independence”. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 80.1, pp. 5–31. ISSN: 14679868. DOI: `10.1111/rssb.12235`. arXiv: `1603.00285`.

Rényi, Alfréd (1959). "On measures of dependence". *Acta mathematica hungarica* 10.3-4, pp. 441–451.

Robert Engle, David Hendry, and Jean-Francois Richard (1983). "Exogeneity". *Econometrica* 51.2. ISSN: 00130133. DOI: 10.2307/2223855. URL: http://www.jstor.org/stable/2223855?origin=crossref.

Rubin, Donald B. (1974). "Estimating causal effects of treatment in randomized and non-randomized studies". *Journal of Educational Psychology* 66.5, pp. 688–701. URL: http://www.fsb.muohio.edu/lij14/420{_}paper{_}Rubin74.pdf.

Rubin, Donald B (2005). "Causal inference using potential outcomes: Design, modeling, decisions". *Journal of the American Statistical Association* 100.469, pp. 322–331.

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

Schölkopf, Bernhard (2019). "Causality for Machine Learning". *arXiv*, pp. 1–20. arXiv: 1911.10500. URL: http://arxiv.org/abs/1911.10500.

Schölkopf, Bernhard and Alexander Smola (2001). *Learning with Kernels.* ISBN: 0262194759. DOI: 10.1198/jasa.2003.s269. arXiv: arXiv:1011.1669v3. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.167.5140{\&}rep=rep1{\&}type=pdf.

Schölkopf, Bernhard, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij (2012). "On Causal and Anticausal Learning". *arXiv.* arXiv: arXiv:1206.6471. URL: http://arxiv.org/abs/1206.6471{\%}5Cnhttp://www.arxiv.org/pdf/1206.6471.pdf.

Shimizu, Shohei, Patrik Hoyer, Aapo Hyvärinen, and Antti Kerminen (2006). "A linear non-Gaussian acyclic model for causal discovery". *Journal of Machine Learning Research* 7, pp. 2003–2030. ISSN: 10985522. URL: http://dl.acm.org/citation.cfm?id=1248619.

Simon, Herbert A. (1953). "Causal Ordering and Identifiability". *Studies in Econometric Method.* Ed. by T Hood, W. and Koopmans. New Haven: Yale University Press, pp. 1–2.

Simon, Herbert A (1962). "The Architecture of Complexity". *Proceedings of the American Philosophical Society* 106.6, pp. 467–482.

Singh, Rahul; Sahani, Maneesh and Arthur Gretton (2019). "Kernel Instrumental Variable Regression". *ArXiv e-prints*, pp. 1–31. arXiv: arXiv:1906.00232v1.

Skitovich, Viktor Pavlovich (1954). "Linear forms of independent random variables and the normal distribution law". *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya* 18.2, pp. 185–200.

Skitovich, Viktor Pavlovich (1962). “Linear combinations of independent random variables and the normal distribution”. *Selec. Transl, in Math., Stat. and Prob.* 2, pp. 211–229.

Spirtes, Peter, Clark Glymour, and Richard Scheines (2000). *Causation, prediction, and search*. Cambridge, MA: MIT Press.

Stewart, Mark B (2004). “The impact of the introduction of the UK minimum wage on the employment probabilities of low-wage workers”. *Journal of the European Economic Association* 2.1, pp. 67–97.

Strobl, Eric V, Kun Zhang, and Shyam Visweswaran (2019). “Approximate Kernel-Based Conditional Independence Tests for Fast Non-Parametric Causal Discovery”. *Journal of Causal Inference* 7.1.

Swamy, P A V B (1970). “Efficient Inference in a Random Coefficient Regression Model”. *Econometrica* 38.2, pp. 311–323.

Wager, Stefan and Susan Athey (2015). “Estimation and inference of heterogeneous treatment effects using random forests”. *arXiv preprint arXiv:1510.04342*, pp. 1–43. arXiv: 1510.04342. URL: http://arxiv.org/abs/1510.04342.

Wang, Wuyi, Peter C. B. Phillips, and Liangjun Su (2019). “The heterogeneous effects of the minimum wage on employment across states”. *Economics Letters* 174, pp. 179–185.

Wooldridge, Jeffrey M (2002). *Econometric analysis of cross section and panel data*. MIT press.

Zhang, Kun and Aapo Hyvärinen (2009). “On the Identifiability of the Post-Nonlinear Causal Model”. *Uncertainty in Artificial Intelligence* 2013, pp. 1–8. arXiv: 1309.2178. URL: http://arxiv.org/abs/1309.2178.

7 Summary

This dissertation consists of three papers sharing the objective to analyze how machine learning methods can be useful to economists and econometricians in their pursuit to understand causal mechanisms operating in the economy. Such causal knowledge is essential when designing policies that help achieve societal goals. ML techniques are increasingly applied in and adapted to practical policy settings. These are characterized by the same type of endogeneity problems that make actionable inference from data difficult and that economists are occupied with. Thus, there are many potential synergies between ML and economics that are surfacing on both the academic and policy-making agendas. Contributions to two points of interchange between the two fields are made. First, ML can be used to improve or extend widely-used identification techniques in economics and, second, insights into causal modeling from the ML community can be introduced as novel routes to identification in economics. The first paper of this dissertation falls in the former, the second and third paper in the latter category.

In the first paper of this dissertation, we adapt the causal forest methodology proposed by Athey et al. (2019) to estimate heterogeneous treatment effects in difference-in-differences studies and analyze heterogeneous effects on wage growth of the 2015 introduction of the statutory minimum wage in Germany. Two contributions are made. First, we show how the causal forest methodology can be applied in difference-in-differences settings. Second, we show that previously documented effect heterogeneities can be explained by interactions of other covariates.

The starting point for the second and third paper of this dissertation is the second point of interchange. There is a tendency to argue ML techniques' strength is their superior predictive capacity. However, above and beyond the idea that superior prediction can be useful in causal inference problems, developments in the ML community question this dictum: Techniques to model causal relations and to identify them from observational data are emerging (for a survey see Peters et al., 2017).

A central tenet of causal machine learning is that the observed joint distribution of a number of random variables contains causal information in the form of invariance properties. This causal information can be exploited by appropriate statistical techniques, even

in the absence of quasi-experimental techniques. In that sense, the causal machine learning literature offers novel pathways to causal understanding that are not yet exploited in economics. The originality of the second and third paper lies in exploring the potential of these novel pathways.

In the second paper, we propose a test for reverse causality that relies on the insight that making functional form assumptions can help identify the causal direction between two observed variables. Two contributions are made. First, we extend existing research from the computer science community on the identifiability of the causal direction by addressing heteroskedastic error structures and the presence of additional control variables. Second, we provide a test for reverse causality that does not rely on instruments.

In the third paper, I propose a test for instrument validity, which relies on a method proposed by Janzing and Schölkopf (2018) to quantify confounding in multivariate linear models. Given the often controversial identifying assumptions in instrumental variable models, whose justification is rarely statistically-grounded, such a method is a valuable addition to the empirical economics toolkit. Two contributions are made. First, I address the limitation of Janzing and Schölkopf (2018) of providing an *overall* degree of confounding for the whole model and provide a way to use their method to estimate a degree of confounding of a *single* covariate in multivariate linear models. Second, I show how this method can be employed to test for instrument validity in instrumental variable models and provide an empirical application.

8 Zusammenfassung

Die Synergien zwischen auf der einen Seite den Feldern des Machine Learning (ML) und der Informatik und auf der anderen Seite der Ökonomie sind weitreichend. Diese Dissertation befasst sich mit zwei Überschneidungspunkten zwischen diesen Feldern. Erstens können ML Methoden traditionelle ökonometrische Verfahren zur Kausalinferenz komplementieren. Zweitens können Erkenntnisgewinne in die Modellierung und Identifikation kausaler Zusammenhänge aus der ML Literatur in die Ökonomie eingeführt werden.

Das erste Papier fällt in erstere Kategorie. Carsten Schröder und ich adaptieren die Kausalwälder Methodik von Athey et al. (2019) für ein Differenzen-in-Differenzen Ansatz und analysieren, inwieweit Effektheterogenitäten der Mindestlohneinführung in Deutschland 2015 auf Lohnwachstum in einer datengetriebenen Weise identifiziert werden können. Mit dieser Studie tragen wir in zweierlei Hinsicht zur Literatur bei. Erstens zeigen wir wie die Kausalwälder Methodik in einem Differenz-in-Differenzen Ansatz verwendet werden kann. Zweitens zeigen wir, wie bereits dokumentierte Effektheterogenitäten sich als nicht echt herausstellen, sobald komplexe Interaktionen aus zusätzlichen Variablen in das Modell aufgenommen werden.

Der Ausgangspunkt für das zweite und dritte Papier dieser Dissertation ist der zweite Überschneidungspunkt. Es gibt eine Tendenz ML Methoden als lediglich mächtige Vorhersagemodelle zu verstehen, die niemals ein Verständnis kausaler Zusammenhänge erreichen können. Neue Entwicklungen in der ML Gemeinschaft stellen diese Aussage jedoch in Frage. Es gibt Fortschritte bezüglich Methoden, die einem erlauben kausale Zusammenhänge in rein observierten, nicht experimentellen, Daten zu erkennen. Die Originalität des zweiten und dritten Projektes besteht darin, diese neuen Entwicklungen in die Ökonomie einzuführen.

Im zweiten Papier stellen wir einen Test für umgekehrte Kausalität als eine Quelle von Endogenität in ökonometrischen Modellen vor. Es basiert auf der Einsicht, dass relativ schwache Annahmen zum funktionalen Zusammenhang zweier Variablen ausreichen, um deren kausale Richtung zu identifizieren. Mit dieser Studie leisten wir einen zweifachen Beitrag zur Literatur. Erstens erweitern wir theoretische Resultate aus der ML Literatur zur Identifizierbarkeit der kausalen Richtung zwischen zwei Variablen, indem wir zusätzliche Kontrollvariablen in das Modell aufnehmen und Heteroskedastizität hinsichtlich dieser zusätzlichen Variablen erlauben. Zweitens, zeigen wir wie die Methodik für einen Test für umgekehrte Kausalität verwendet werden kann, der keine Instrumente braucht.

Im dritten Papier, stelle ich einen Test für Instrumentenvalidität vor, der auf einer Methodik von Janzing and Schölkopf (2018) basiert. Diese Methodik erlaubt es den Grad, zu welchem eine observierte statistische Korrelation auf unbeobachtete Störvariablen zurückzuführen ist, zu quantifizieren. Gegeben der oft höchst umstrittenen Annahmen, die für eine kausale Identifizierung mit Hilfe von Instrumentalvariablen nötig sind, und deren Rechtfertigung selten statistisch fundiert ist, ist ein solcher Test ein sinnvoller Beitrag zum Instrumentarium in der empirischen Ökonomie. Diese Studie trägt entscheidend zur Literatur bei, denn erstens zeige ich wie die Methodik von Janzing und Schölkopf, die ein Störgrad für das gesamte Modell quantifiziert, genutzt werden kann um den Störgrad von einer einzelnen Variable zu quantifizieren. Zweitens zeige ich wie die Methodik für einen Instrumentalvariablentest genutzt werden kann und diskutiere eine empirische Anwendung.

List of Figures

List of Tables

Ehrenwörtliche Erklärung

Erklärung gemäß §4 Abs. 2

Hiermit erkläre ich, dass ich mich noch keinem Promotionsverfahren unterzogen oder um Zulassung zu einem solchen beworben habe, und die Dissertation in der gleichen oder einer anderen Fassung bzw. Überarbeitung einer anderen Fakultät, einem Prüfungsausschuss oder einem Fachvertreter an einer anderen Hochschule nicht bereits zur Überprüfung vorgelegen hat.

(Unterschrift, Ort, Datum)

Erklärung gemäß §10 Abs. 3

Ich habe meine Dissertation soweit nicht anders vermerkt selbständig verfasst.

Folgende Hilfsmittel wurden benutzt

- Statistik und Mathematik: Stata, R
- Schriftsatz und Formatierung: LaTeX

(Unterschrift, Ort, Datum)