Testability of Reverse Causality Without Exogenous Variation

Christoph Breunig* Patrick Burauel†

April 29, 2024

Abstract

This paper shows that testability of reverse causality is possible even in the absence of exogenous variation, such as in the form of instrumental variables. Instead of relying on exogenous variation, we achieve testability by imposing relatively weak model restrictions and exploiting that a dependence of residual and purported cause is informative about the causal direction. Our main assumption is that the true functional relationship is nonlinear and that error terms are additively separable. We extend previous results by incorporating control variables and allowing heteroskedastic errors. We build on reproducing kernel Hilbert space (RKHS) embeddings of probability distributions to test conditional independence and demonstrate the efficacy in detecting the causal direction in both Monte Carlo simulations and an application to German survey data.

Keywords: Endogeneity; Reverse causality; Conditional Independence; Causal Discovery; Reproducing kernel Hilbert spaces.

1 Introduction

Endogeneity is a central problem in econometric models which potentially invalidates estimates of causal effects. Reverse causality, where the dependent variable causes the independent variable, is one source of such endogeneity. The aim of this paper is to show that reverse causality is testable under mild assumptions and without relying on exogenous variation. We build on Hoyer et al. (2009) who provide a link between nonlinear model structure and causality, namely, that a nonlinear relation between cause and effect leads to observable signals about causal direction using observational data. While their theoretical results are striking, their results assume homoskedastic errors and do not generalize to settings with additional control variables. We show that this assumption can be relaxed to allow for heteroskedasticty with respect to additional control variables, as is commonly the case in econometric applications. Our primary conditions for achieving testability are twofold: First, we necessitate a nonlinear relationship between the dependent variable

*University of Bonn, 53113 Bonn, Germany. Email: cbreunig@uni-bonn.de
†California Institute of Technology, Pasadena, CA 91125, USA. Email: pburauel@caltech.edu

and the regressor (with no restrictions on how controls are incorporated into the model). Second, we impose additively separable errors.

Following Hoyer et al., 2009, we also demonstrate how our testability result can be applied in empirical practice. Specifically, we show that identification of the causal direction is equivalent to a conditional independence test of covariates and error terms given control variables. We make use of conditional independence tests based on kernel mean embeddings, i.e., maps of probability distributions into reproducing kernel Hilbert spaces (RKHS) (see Muandet et al., 2017, for a survey). Intuitively, this corresponds to approximating conditional distributions with unconditional ones by weighting with an appropriate kernel, and evaluating their covariance in an RKHS. The method can detect nonlinear dependencies.

We consider two formal applications of our testability result. First, we explore testing for causal direction based on conditional independence. As already indicated in the related literature, achieving exact size control can be challenging within this framework, and we provide a detailed discussion of this issue below. Second, we conduct causal discovery, where we remain agnostic about the causal direction and compare test statistics for two rivaling models to gain insight into which one represents the true causal structure (see Peters et al., 2014).

In Monte Carlo simulations, we investigate the power of our approach to detect reverse causality. We see that the degree of nonlinearity increases the power to detect reverse causation. The procedure has surprisingly high accuracy in detecting the true causal direction even under moderate form of nonlinearities and can be powerful even in linear models under restrictions on the distribution of errors terms, which is a result described by (Shimizu et al., 2006). Furthermore, we provide an empirical illustration using data from the German Survey of Income and Expenditure. We show that our algorithm can infer from purely observational data that work experience is a causal driver for income, not vice versa. Substantively, this is not a surprising result; however, the fact that it can be inferred without exogenous variation is.

Related literature Our test rests on the idea that X causing Y implies an independence between the error of a regression of Y on X, and X. This idea goes back to Engle et al. (1983), who propose a definition of an exogenous relation in terms of conditional densities. 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 rely on analyzing to what extent parameter values are sensitive to exogenous interventions on the purported cause. Thus, their results specifically rely on exogenous variation (e.g. in the form of instrumental variables) whereas the approach at hand does not require such variation.

The problem of identifying causal structure from non-experimental data is receiving considerable attention in the causal machine learning literature (for comprehensive overviews see Mooij et al., 2016; Peters et al., 2017; Schölkopf et al., 2021). 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 identification is possible.[1] Shimizu et al. (2006) show that non-Gaussianity of observed variables leads to identifiability. Subsequently, 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. 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.

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 W serve as controls.

In this paper, we rely on RKHSs and kernel mean embeddings, which are not widely used in the econometrics literature, albeit with notable exceptions. Carrasco and Florens (2000) and Carrasco et al. (2007) discuss the usefulness of RKHS theory given infinite number of moment conditions. Singh et al. (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. Zhang et al. (2020) propose kernelized moment restrictions to estimate nonlinear IV estimators. 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.

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, Fève et al. (2018) and Breunig (2020) for models with nonseparable errors.

The remainder of the paper is organised as follows. Section 2 establishes testability of reverse causality under nonlinear regression functions and introduces the RKHS test for independence. In Section 3, we analyze finite sample power of our RKHS procedure in a Monte Carlo simulation study. Section 4 provides an application of our method to empirical data. Appendix A provides a proof of our main testability result. Appendix B gives a review of the construction of RKHS. Appendix C contains additional Monte Carlo simulation results.

[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.

2 Testing Reverse Causality

We show how to test for reverse causality between two variables X and Y in the presence of additional covariates W. 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 test. Third, we present the implementation of the test.

2.1 Model and Assumptions

Consider a model where observable continuous scalar variable X causes observable scalar variable Y in the presence of the vector of covariates W (which we refer to in the following as the *model*):

$$Y = h(X, W) + U \quad \text{where} \quad U = \sigma(W)\,\varepsilon \;\text{ and }\; \varepsilon \perp\!\!\!\perp (X, W) \tag{1}$$

where ε are unobservable variables and $\sigma(\cdot)$ some strictly positive function. In addition, we assume $E[\varepsilon] = 0$ without loss of generality.

Note that the error U is additively separable. The additive separability of U precludes the dependence of marginal effects on unobservables except through a dependence via W. The model allows for heteroskedasticity of the error term U with respect to the control variables W. In particular, model equation (1) implies $U \perp\!\!\!\perp X|W$, which is also known as conditional exogeneity (see White and Chalak, 2010). It corresponds to the unconfoundedness assumption in the treatment effects literature (Imbens and Rubin, 2015) and is also closely related to the special regressor assumption (see Lewbel, 2014, for an overview).

The main idea of this paper is to study the conditions under which this model is distinguishable form reversed analog without relying on exogenous information. The *reverse model*, where Y is causing X, again in the presence of the vector of covariates W, is defined as

$$X = \widetilde{h}(Y, W) + \widetilde{U} \quad \text{where} \quad \widetilde{U} = \widetilde{\varepsilon}\,\widetilde{\sigma}(W) \text{ and } \widetilde{\varepsilon} \perp\!\!\!\perp (Y, W) \tag{2}$$

where $\widetilde{\varepsilon}$ are unobservable variables and $\widetilde{\sigma}(\cdot)$ some strictly positive function.

We denote the probability density function of a random vector V by f_V and make the following assumptions.

Assumption 1 (Regularity)**.** *The functions h, $\widetilde{h}$, $f_{X|W}$, $f_{Y|W}$, f_ε, and $f_{\widetilde{\varepsilon}}$ are three times differentiable.*

Assumption 2 (Nonlinearity)**.** *The functions h and $\tilde{h}$ are nonlinear in their first arguments.*

The nonlinearity of regression functions (Assumption 2) can be used to make inference on the causal structure of a model is obtained first by Hoyer et al. (2009). We extend their work by allowing for additional control variables W and also considering heteroskedasticity of the error term with respect to these covariates. Intuitively, nonlinearity of h ensures that the error terms in the reverse model are not independent of the regressor, which provides power of the test. While linear models are used in many economic applications, they are typically seen as approximations of nonlinear relationships between dependent variable and regressors.

2.2 Testability

We are now in a position to formulate the main theorem of this paper.

Theorem 1. *Let Assumptions 1 and 2 be satisfied. Both model* (1) *and the reverse model* (2) *exist only if* $\xi(x,w) := \log f_{X|W}(x|w)$ *satisfies the linear inhomogeneous differential equation*

$$\frac{\partial^3 \xi(x,\bar{w})}{\partial x^3} = \frac{\partial^2 \xi(x,\bar{w})}{\partial x^2} G_1(x,\bar{y},\bar{w}) + G_2(x,\bar{y},\bar{w}) \tag{3}$$

for all x *and some fixed* $(\bar{y},\bar{w})$*, where* $G_1(x,\bar{y},\bar{w})$ *and* $G_2(x,\bar{y},\bar{w})$ *are defined in Appendix A.*

The proof of this statement can be found in Appendix A. For the proof of the result, we build on Hoyer et al. (2009) and extend their result to allow for control variables W with additional form of heteroskedasticity. Intuitively, it is shown that causal and anti-causal models can only exist simultaneously under very specific circumstances: if the joint distribution of (Y,X,W) satisfies both a causal and a anticausal model, we can show that densities $\log f_{X|W}$ and $\log f_\varepsilon$ of the causal model have to satisfy the linear inhomogeneous differential equation (3). The solutions of this differential equation restrict the log density of X given W to lie in a (specific) three-dimensional space, although a priori the (generic) space of possible log marginal densities of X given W is infinite-dimensional (this argument follows Hoyer et al., 2009, closely). To achieve testability of reverse causality, we need to exclude those distributions that satisfy the differential equation, i.e., we need to exclude specific combinations of $\log f_{X|W}$, $\log f_\varepsilon$, and $h(\cdot)$.

Remark 1 (Characterization of differential equation (3))**.** *In Table 1, we reproduce an exhaustive list of all model specifications that satisfy the differential equation* (3) *by Zhang and Hyvärinen (2009) under the additional assumption that the error* ε *has large support. The list of* $(f_X, f_\varepsilon, h(\cdot))$ *tuples that satisfy the differential equation is even smaller than in Zhang and Hyvärinen (2009) because we constrain the model space by assuming a nonlinear* h*. It is remarkable that in each specification I, II, and III, the density of* ε *is not even integrable. Even more,* $E[\varepsilon]$ *does not exist. This illustrates that even though solutions to the differential equation* (3) *can be computed, they are not relevant in most (if not all) empirical applications.*

	f_ε	f_X
I	$f_\varepsilon(u) = c_1 \exp(c_2 u) + c_3 u + c_4$	$(\log f_X(x))' \to c_1 \neq 0$ as $x \to +\infty$ or $x \to -\infty$
II	$f_\varepsilon(u) = c_1 \exp(c_2 u) + c_3 u + c_4$	$f_X(x) = \{c_1 \exp(c_2 x) + c_3 \exp(c_4 x)\}^{c_5}$
III	$f_\varepsilon(u) = \{c_1 \exp(c_2 u) + c_3 \exp(c_4 u)\}^{c_5}$	$\lim_{x\to-\infty}(\log f_X(x))' = c_1 \neq 0$, $\lim_{x\to+\infty}(\log f_X(x))' = c_2 \neq 0$

Table 1: *All situations in which reverse causality is not testable. Constants* c_1, c_2, c_3, c_4, c_5 *might be different in different cases. In all situations,* h *strictly monotonic, and* $h'(x) \to 0$*, as* $x \to +\infty$ *or as* $x \to -\infty$ *under the maintained assumption of large support of* ε*.*

The next result provides a more concrete formulation of Theorem 1 and how it can be applied to model specification testing. This corollary follows immediately from Theorem 1 and its proof is thus omitted.

Corollary 1. *Let Assumptions 1 and 2 be satisfied. Then model* (1) *rules out the reverse model* (2) *if the joint distribution of* (X, W) *does not satisfy the differential equation* (3).

Corollary 1 allows identification of 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.

Note that even when h is linear ($Y = X + U$, simplifying by abstracting from W), it is only possible to rewrite this as a reverse model, $X = Y + \widetilde{U}$, with Y independent of $\widetilde{U}$, if all variables are Gaussian. If at most one of the exogenous variables is non-Gaussian, residual and purported cause are dependent in the reverse model (Shimizu et al., 2006).

2.3 Implementation the Reverse Causality Test

Algorithm 1 shows detailed steps of the implementation of the test. In words, after some pre-processing (Step 1 and 2), we propose estimating a nonlinear model in both directions, i.e. with Y and X as dependent variables respectively (Step 3), calculating residuals for both models (Step 4) and testing for conditional independence using a kernel conditional independence test (KCI) introduced by Zhang et al. (2011) (Step 5). We discuss this test statistic, see eq. (7), and its development in Section 2.5; see the detailed discussion in Section 2.6.

Data: $\mathcal{D} = \{Y_i, X_i, W_i\}_{i=1}^n$
Input: hyperparameters for KCI test: heuristic kernel bandwidth λ, set at $\lambda = 0.8$ if sample size $n \leq 200$, $\lambda = 0.3$ if sample size $n > 1200$ and $\lambda = 0.5$ otherwise; heuristic regularization parameter λ_R is set to 10^{-3}
Output: Decision whether to reject the model in (1)
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 $\mathcal{D}^{tr} = \{Y_i, X_i, W_i\}_{i=1}^{n/2}$ and test set $\mathcal{D}^{te} = \{Y_i', X_i', W_i'\}_{i=(n/2)+1}^{n}$
Step 3: Estimate generalized additive models (GAMs) based on $\mathcal{D}^{tr}$
GAM1: $Y = h(X, W) + U$, call resulting estimate $\widehat{h}$
GAM2: $X = \widetilde{h}(Y, W) + \widetilde{U}$, call resulting estimate $\widehat{\widetilde{h}}$
Step 4: calculate residuals based on $\mathcal{D}^{te}$
$\widehat{U} := Y' - \widehat{h}(X', W')$, and
$\widehat{\widetilde{U}} := X' - \widehat{\widetilde{h}}(Y', W')$
Step 5: Test conditional independence with KCI test (based on residuals from Step 4)
use $\widehat{U}$, X' and W' to test $U \perp\!\!\!\perp X|W$ with KCI test whose null hypothesis is conditional independence; call resulting p-value p_{model}
use $\widehat{\widetilde{U}}$, Y' and W' to test $\widetilde{U} \perp\!\!\!\perp Y|W$ with KCI test whose null hypothesis is conditional independence; call resulting p-value p_{reverse}
Step 6: Decide to reject model (1) if $p_{\text{model}} < \alpha$.

Algorithm 1: Reverse causality test at nominal level $\alpha \in (0, 1)$

2.4 Causal Discovery

Alternatively, instead of imposing model (1) as a maintained hypothesis, we can also remain agnostic about the true causal relationship and infer which of the models is the correct one. Formally, this requires assuming that either model (1) or (2) hold. This is formalized in the following corollary which follows immediately from Theorem 1 and its proof is thus omitted.

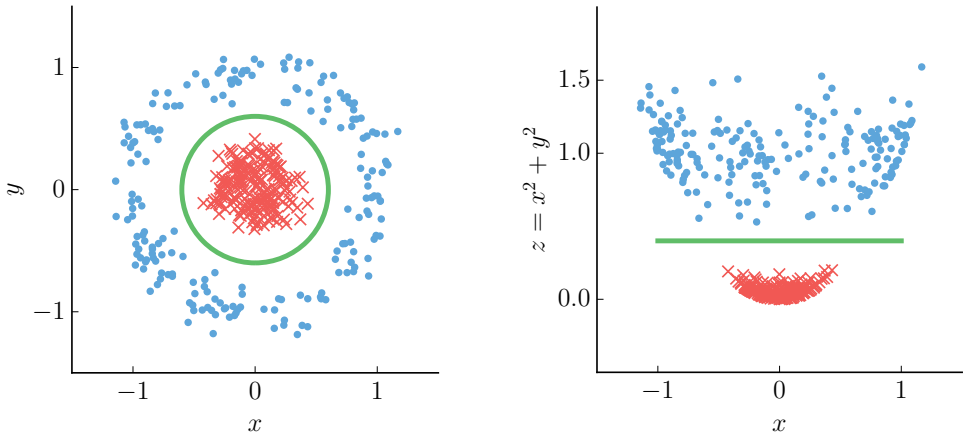
Corollary 2. *Let Assumptions 1 and 2 be satisfied. Suppose either the model* (1) *or the reverse model* (2) *holds. Then, if the joint distribution of* (X, W) *does not satisfy the differential equation* (3)*, the true model is identified.*

Data: $\mathcal{D} = \{Y_i, X_i, W_i\}_{i=1}^n$
Input: hyperparameters for KCI test: heuristic kernel bandwidth λ, set at $\lambda = 0.8$ if sample size $n \leq 200$, $\lambda = 0.3$ if sample size $n > 1200$ and $\lambda = 0.5$ otherwise; heuristic regularization parameter λ_R is set to 10^{-3}.
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 $\mathcal{D}^{tr} = \{Y_i, X_i, W_i\}_{i=1}^{n/2}$ and test set $\mathcal{D}^{te} = \{Y_i', X_i', W_i'\}_{i=(n/2)+1}^{n}$
Step 3: Estimate generalized additive models (GAMs) based on $\mathcal{D}^{tr}$
GAM1: $Y = h(X, W) + U$, call resulting estimate $\widehat{h}$
GAM2: $X = \widetilde{h}(Y, W) + \widetilde{U}$, call resulting estimate $\widehat{\widetilde{h}}$
Step 4: calculate residuals based on $\mathcal{D}^{te}$
$\widehat{U} := Y' - \widehat{h}(X', W')$, and
$\widehat{\widetilde{U}} := X' - \widehat{\widetilde{h}}(Y', W')$
Step 5: Test conditional independence with KCI test (based on residuals from Step 4)
use $\widehat{U}$, X' and W' to test $U \perp\!\!\!\perp X|W$ with KCI test; call resulting test statistic $\text{KCI}_{\text{causal}}$
use $\widehat{\widetilde{U}}$, Y' and W' to test $\widetilde{U} \perp\!\!\!\perp Y|W$ with KCI test; call resulting test statistic $\text{KCI}_{\text{anticausal}}$
Step 6: Decide on causal direction
if $\text{KCI}_{causal} < \text{KCI}_{anticausal}$ **then**
  accept $X \to Y$ as correct model
else if $\text{KCI}_{causal} > \text{KCI}_{anticausal}$ **then**
  accept $Y \to X$ as correct model
else if $\text{KCI}_{causal} = \text{KCI}_{anticausal}$ **then**
  inconclusive result
end

Algorithm 2: Bivariate causal discovery with control covariates

Algorithm 2 shows detailed steps of the implementation of the bivariate causal discovery which is motivated by 2. Steps 1 to 4 are the same as in 1. In step 5, we compute the test statistics corresponding to two conditional independence tests: one for model (1) and one

for (2). The relative size of the resulting test statistics is informative about which model is the correct causal model (Step 6).



(a) Raw data (b) Data representation

Figure 1: A nonlinear classifier. *Panel (a): data can only be separated by a nonlinear decision boundary (green circle), 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 decision boundary (green line) to separate the data. Note that the y dimension is ommited in (b); it extends perpendicular to the 2D sheet of paper/screen. In a 3D graph, the data looks like an inverted cone with a rounded bottom. Figure credit: Lopez-Paz (2016, Figure 3.1)*

2.5 Testing Conditional Independence

This section introduces the concept of Hilbert Space embeddings of probability 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, we discuss the procedure in detail. We first intuitively introduce important underlying concepts such as feature maps, reproducing kernel Hilbert spaces, etc. keeping technical details to a minimum before turning to how these constructs can help to formulate a conditional independence test. See Appendix B for the formal statements.

2.5.1 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 each characterized by two dimensions, say x = weight and y = height, by using a linear classifier (i.e. a linear regression that serves as a boundary between the two classes). If the data looks like those in Figure 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 practice the feature space will have many more dimensions). In this higher-dimensional space, there is a linear boundary that separates the two classes, see Figure 1(b). This example is adopted from Lopez-Paz (2016).

Similarly to the linear classifier in Figure 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 nonlinear 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.

2.5.2 Kernels as inner product of implicit feature map

In practice, instead of manually defining a set of appropriate features (such as $z = x^2+y^2$ in the previous example), flexible functions can be used to define the feature map. Formally, we define a feature map Φ from input space $\mathcal{X}$ to the space of functions $\mathbb{R}^{\mathcal{X}}$:

$$\begin{aligned} \Phi: \quad & \mathcal{X} \to \mathbb{R}^{\mathcal{X}} \\ & x \mapsto k(\cdot, x) \end{aligned}$$

where k is a positive-definite kernel[2], such as the Gaussian kernel, which is defined as

$$k(v, v') := \exp\Big(-\frac{\|v - v'\|_{\ell_2}^2}{\lambda}\Big) \tag{4}$$

for arbitrary vectors v and v' and a bandwidth parameter $\lambda > 0$, where $\|\cdot\|_{\ell_2}$ denotes the ℓ_2 norm. Each data point can thus be richly represented by its similarity (defined by the kernel) to all other data points. It can be shown that the inner product of two such feature maps in an RKHS reduces to an evaluation of the kernel itself (see Schölkopf and Smola, 2001, and Appendix B):

$$\langle k(\cdot, x), k(\cdot, x')\rangle = \langle \Phi(x), \Phi(x')\rangle = k(x, x').$$

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

[2]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.

the data into a higher-dimensional Reproducing Kernel Hilbert Space. 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 can be populated by moments of increasing order when embedding a probability distribution in the RKHS, which gives rise to a unique representation of the probability distribution.

2.5.3 Partial cross-covariance operators and conditional independence

The conditional independendece test we use relies on a characterization of conditional independence as a vanishing partial cross-covariance operator between two RKHSs. To get an intuition, consider an analogy to the characterization of conditional independence for jointly Gaussian variables in terms of vanishing partial correlation. 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. 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 because of 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 into higher-dimensional RHKS enables the use of linear algorithms to study nonlinear relationships. This reasoning also underlies the following characterization of conditional independence for arbitrarily distributed random variables.

Daudin (1980) establishes the equivalence

$$\sup_{f \in \mathcal{F}_{\widetilde{X}}, g \in \mathcal{F}_U} \mathbb{E}[f(\widetilde{X}) g(U)] = 0 \Leftrightarrow X \perp\!\!\!\perp U | W, \tag{5}$$

for properly chosen function spaces, based on function spaces $\mathcal{F}_{\widetilde{X}} := \left\{ f \in L^2_{\widetilde{X}} : \ \mathbb{E}[f(\widetilde{X})|W] = 0 \right\}$ and $\mathcal{F}_{\widetilde{U}} := \left\{ g : \ g(\widetilde{U}) = \check{g}(U) - \mathbb{E}[\check{g}(U)|W] \text{ where } \check{g} \in L^2_U \right\}$ where for any random variable Z the Hilbert space $L^2_Z = \{ f : \mathbb{E}[f^2(Z)] < \infty \}$.

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. Further, we make use of the notation $\widetilde{X} = (X, W)$ and $\widetilde{U} = (U, W)$ and define $k_{\widetilde{\mathcal{X}}} = k_{\mathcal{X}} \times k_{\mathcal{W}}$ and corresponding RKHS $\mathcal{H}_{\widetilde{\mathcal{X}}}$. Denote the feature maps corresponding to RKHS $\mathcal{H}$ with $\phi_{\mathcal{H}}$. To reduce the complexity of Daudin's equivalence result, Zhang et al. (2011) show that restricting function spaces of f and g to lie in RKHSs $\mathcal{H}_{\widetilde{\mathcal{X}}}$ and $\mathcal{H}_{\mathcal{U}}$. This restrictions are sufficient to derive an estimable statistic in terms of the Hilbert Schmidt norm of a partial cross-covariance operator:

$$\Sigma_{\widetilde{X}U|W} := \Sigma_{\widetilde{X}U} - \Sigma_{\widetilde{X}W} \Sigma_{WW}^{-1} C_{WU}$$

where $\Sigma_{\widetilde{X}U}$ is the covariance operator $\Sigma_{\widetilde{X}U} : \mathcal{H}_{\widetilde{\mathcal{X}}} \to \mathcal{H}_{\mathcal{U}}$ defined as

$$\langle \Sigma_{\widetilde{X}U} \phi_{\widetilde{\mathcal{X}}}, \phi_{\mathcal{U}} \rangle_{\mathcal{H}_{\mathcal{U}}} = \mathrm{Cov}(\phi_{\widetilde{\mathcal{X}}}(\widetilde{X}), \phi_U(U)). \tag{6}$$

Following Zhang et al. (2011), a vanishing Hilbert Schmidt norm of the partial cross-covariance operator characterizes conditional independence yielding the equivalence result:

$$\left\|\Sigma_{\widetilde{X}U|W}\right\|_{HS} = 0 \Leftrightarrow X \perp\!\!\!\perp U|W \tag{7}$$

where the Hilbert Schmidt norm of an operator $A : \mathcal{H}_{\widetilde{\mathcal{X}}} \to \mathcal{H}_{\mathcal{U}}$ is defined as $\|A\|_{HS} = \sqrt{\sum_{j,l\geq 1}\langle f_j, Ae_l\rangle^2_{\mathcal{H}_{\mathcal{U}}}}$ where $\{e_j\}_{j\geq 1}$ and $\{f_j\}_{j\geq 1}$ are an orthonormal basis in $\mathcal{H}_{\widetilde{\mathcal{X}}}$ and $\mathcal{H}_{\mathcal{U}}$, respectively.

2.5.4 The KCI test statistic

Zhang et al. (2011) build on the conditional independence characterization in (7) and define the KCI test statistic:

$$\text{KCI} = \frac{1}{n}\text{tr}(\widetilde{K}_{\widetilde{X}|W}\ \widetilde{K}_{U|W}) \tag{8}$$

where $\widetilde{K}_{\widetilde{X}|W}$ and $\widetilde{K}_{U|W}$ are centralized kernel matrices defined as follows. The centralized kernel matrix for any variable Z is given by $\widetilde{K}_Z = HK_ZH$ where K_Z is the uncentralized kernel matrix, i.e. a matrix whose (i, j) element is given by $k(x_i, x_j)$ where k is the Gaussian kernel in eq. (4), and $H = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top$ where $\mathbf{I}_n$ denotes the identity matrix of size n and $\mathbf{1}_n$ a vector of ones of length n. These centralized kernel matrices need to be adjusted to reflect the conditioning on W. This is achieved using kernel ridge regression to derive a matrix R_W

$$R_W = \mathbf{I}_n - \widetilde{K}_W(\widetilde{K}_W + \lambda_R\mathbf{I}_n)^{-1}$$

where λ_R is a regularization parameter. Finally, the kernel matrix $\widetilde{K}_{\widetilde{X}|W}$ can be expressed as

$$\widetilde{K}_{\widetilde{X}|W} = R_W\widetilde{K}_{\widetilde{X}}R_W.$$

Similarly, the construction for $\widetilde{K}_{U|W}$ is analogous and yields

$$\widetilde{K}_{U|W} = R_W\widetilde{K}_UR_W.$$

Following Zhang et al. (2011), we normalize the data and choose the hyperparameters heuristically as follows. The bandwidth parameter λ is set at $\lambda = 0.8$ if sample size $n \leq 200$, $\lambda = 0.3$ if sample size $n > 1200$, and $\lambda = 0.5$ otherwise for the construction of $\widetilde{K}_{\widetilde{X}}$ and $\widetilde{K}_U$, and at half that value for $\widetilde{K}_W$. The regularization parameter λ_R is set to 10^{-3}. These parameters are deemed appropriate when the dimenstionality of W is small, say less than three. For higher-dimensional W, Zhang et al. (2011) recommend choosing different λ and λ_R for $\widetilde{K}_{\widetilde{X}}$ and $\widetilde{K}_U$, which can be achieved using cross-validation.

Zhang et al. (2011) derive the asymptotic distribution and corresponding p-values of the KCI test statistic under H_0 : conditional independence and show that it achieves pointwise asymptotic level (see also 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.

2.6 On Critical Values

The theory implies an independence of errors and the covariate in the causal model and a dependence between the errors and the covariate in the anticausal model. The algorithm involves testing the independence between errors and covariate conditional on 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 $U := Y - h(X, W)$ in eq. (1) at their disposal and, therefore, must rely on estimated errors $\widehat{U} := Y - \widehat{h}(X, W)$, and the respective estimated errors of the model in eq. (2), to investigate which model is correct. That these residuals are estimated and, in particular, that they depend on the estimated $\widehat{h}$, poses a challenge that we discuss now.

Mooij et al. (2016) and Hoyer et al. (2009) propose randomly splitting the available data $\mathcal{D} = \{Y_i, X_i, W_i\}_{i=1}^{n}$ in training and test sets, denoted $\mathcal{D}^{tr} = \{Y_i, X_i, W_i\}_{i=1}^{n/2}$ and $\mathcal{D}^{te} = \{Y_i', X_i', W_i'\}_{i=(n/2)+1}^{n}$, respectively. $\mathcal{D}^{tr}$ is used to get an estimate $\widehat{h}$ of the true regression function h. $\mathcal{D}^{te}$ is then used to get estimates $\widehat{\varepsilon}' := Y' - \widehat{h}(X', W')$ of the true errors ε. An error in the estimated $\widehat{h}$ induces a dependence of $\widehat{\varepsilon}'$ and X' (conditional on W') even though ε and X are truly independent (conditional on W). Consequently, conventional thresholds for the independence test tend to be too loose and would ideally incorporate the fact that $\widehat{h}$ 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 $U \perp\!\!\!\perp X|W$, 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 $\widehat{\varepsilon}'$ and X'.[3] 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 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 imposed in Corollary 2, which states that either causal or anticausal model exist, allows us to infer the directionality with Algorithm 2.

In particular, under the conditions of Corollary 2, 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

[3]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 exercises, to indicate the level of λ_α as a function of observable or estimable quantities.

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.

3 Monte Carlo Simulations

We investigate finite sample performance of our heteroskedasticity robust reverse causality test and its use in the causal discovery context with Monte Carlo experiments. The results are based on 500 Monte Carlo replications in each experiment and the sample size is varied with $n \in \{250, 500, 1000\}$.

We simulate data for the following model:

$$\begin{pmatrix} X_i \\ W_i \\ U_i^* \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right)$$

and

$$U_i = c_{\rho,q} \operatorname{sgn}(U_i^*) \left| (1 + \phi(W_i))^{\rho/2} U_i^* \right|^q, \tag{9}$$

where ϕ denotes the standard normal probability density function, $\operatorname{sgn}(\cdot)$ is the sign function, and $q, \rho, c_{\rho,q}$ are constants which vary in the experiments below. The dependent variable is generated by

$$Y_i = \kappa_j(X_i, W_i, \tau) + U_i$$

where $j \in \{1, 2\}$ and the functions κ_j are given by

$$\kappa_1(x, w, \tau) = x + \tau x^2 + w,$$
$$\kappa_2(x, w, \tau) = x + \tau \sin(x + \pi/2) + w + w^2.$$

Here, τ controls the degree of nonlinearity between Y and X. The linear case corresponds to $\tau = 0$. The parameter ρ captures degree of the heteroskedasticity w.r.t. the control variable W. That is, $\operatorname{Var}(U|W = w) = (1 + \phi(w))^\rho$ and hence, $\rho = 0$ corresponds to the homoskedastic case. We simulate data for $\tau \in \{0, 0.25, 0.5, 0.75, 1\}$ and $\rho \in \{0, 1\}$. We rely on the R package `CondIndTests` (Heinze-Deml et al., 2019) for the implementation of the HSIC independence test.

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 distributions. We choose the constant $c_{\rho,q}$ (via numerical approximation) such that the variance of U is normalized to one under each choice of ρ and q. We estimate both causal and reverse models with a Generalized Additive Model using smoothing splines.

3.1 Testing for Reverse Causality

We implement the test as described in Algorithm 1 for the nominal level $\alpha = 0.05$. Figure 2 reports the empirical rejection probabilities of testing conditional independence of covariates and estimated residuals under the reverse model (2). Overall, the empirical rejection

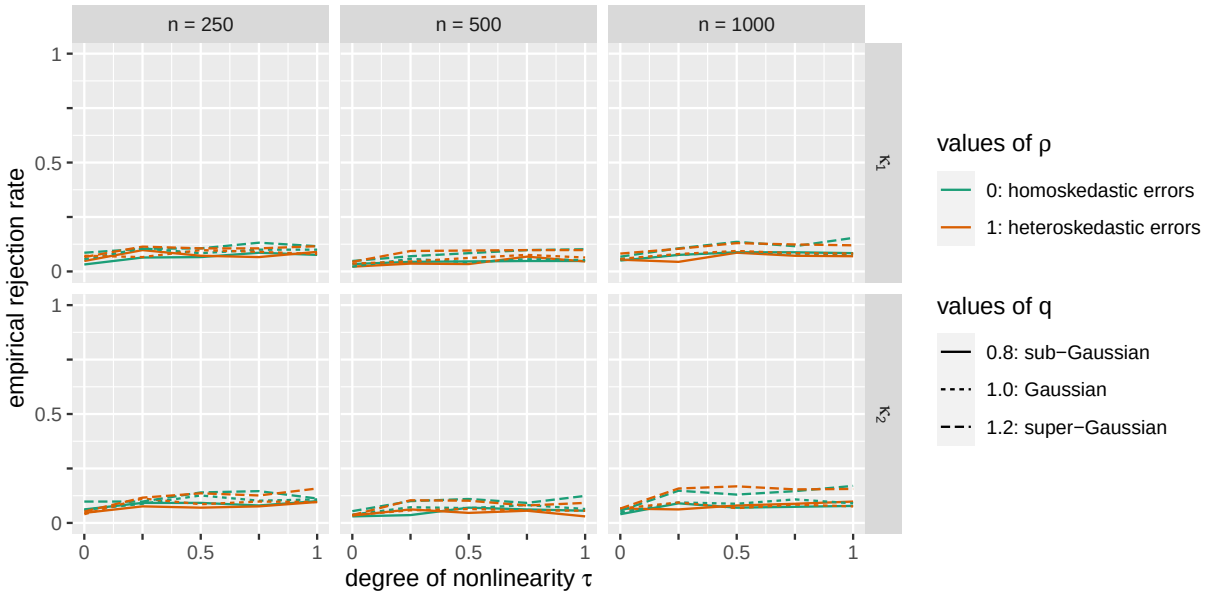


Figure 2: *Empirical rejection rates of testing conditional independence of covariates and estimated residuals under the model* (1) *when* $c_{\rho,q}$ *is chosen such that* $\mathrm{Var}(U) = 1$.

rates are close to the nominal level for different choices of ρ, q, and τ. This is remarkable as the testing problem is complex and builds on nonparametric procedures. As such, we cannot expect exact control over the significance level; see also the discussion in Section 2.6. The test shows some degree of oversizing under super-Gaussian error distributions, indicating the complexity of the testing problem.

We illustrate the empirical power of the reverse causality test in Figure 3. The power of the test, i.e., the probability of rejecting the independence of estimated residual and candidate cause in model (2), increases sharply as the degree of nonlinearity τ increases. While our identifiability results rely on a nonlinear h, see Assumption 2, it can be seen that non-Gaussianity can be a source of power similar to nonlinearity: with sufficiently many samples (rightmost column), the empirical rejection rates lie above 0.4 even in the linear case ($\tau = 0$) as long as $q \neq 1$. This is a well-known result in the causal discovery community, see e.g. Shimizu et al., 2006.

3.2 Causal Discovery

In Figure 4 we report empirical probabilities of correct classification of the causal direction. When the relationship between X and Y is linear, i.e. $\tau = 0$, and the error Gaussian, i.e. $q = 1$, the algorithm performs at about chance level. This is consistent with the theory since the causal direction is not identifiable in the linear case.[4] As soon as the relation between cause and effect becomes nonlinear, i.e. $\tau \neq 0$, the accuracy of the reverse causal discovery algorithm increases. For instance, when $n = 500$ and the relation between cause and $\tau = 1$ the algorithm arrives at the correct conclusion in more than 95% of the Monte Carlo runs,

[4]Note that Shimizu et al. (2006) show that the direction *is* identifiable in the linear case with a non-Gaussian, homoskedastic error distribution, i.e. when $q \neq 1$ and $\rho = 0$. Though our simulation results suggest that the result also holds with heteroskedastic errors w.r.t. W, we do not extended it formally.

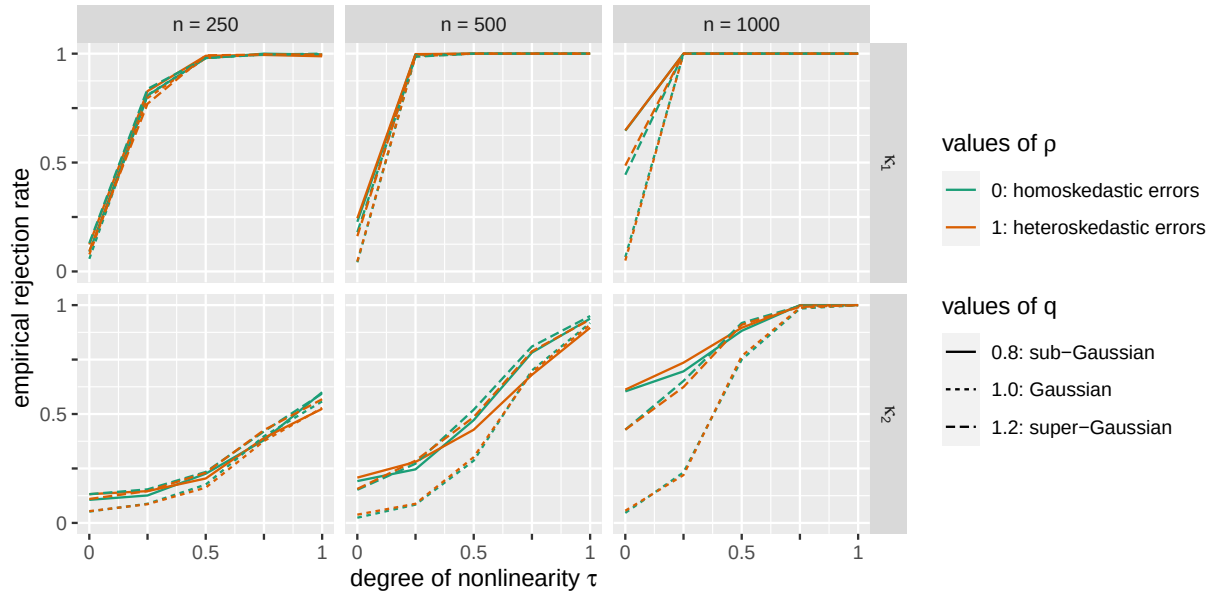


Figure 3: *Empirical rejection rates of testing conditional independence of covariates and estimated residuals under the reverse model* (2) *when* $c_{\rho,q}$ *is chosen such that* $\mathrm{Var}(U) = 1$.

regardless of the level of heteroskedasticity (parameterized by ρ) and shape of the error distribution (parameterized by q). We also see that our procedure has power to detect the correct causal directions in cases where the nonlinearity is less pronounced, i.e., $0 < \tau < 1$. Moreover, the performance of the algorithm is robust to changes in the specification of the functional form, which can be seen by comparing the κ_1 and κ_2 rows in Figure 4. For a given τ, the test has more power for κ_1 because the nonlinear relation between X and Y is more pronounced than for κ_2. Overall we see that the accuracy of causal discovery increases with sample size, where this change is stronger when the regression function is given by κ_2. We show that the results remain robust to different error variances in Appendix C.Specifically, we consider a low noise regime where $c_{\rho,q}$ is chosen such that $\mathrm{Var}(U) = 0.8$ and a high noise regime where $\mathrm{Var}(U) = 1.2$.

In sum, two observations are worth stressing. First, the results show that the algorithm has surprisingly high power when cause and effect are related nonlinearly. Second, the performance of the algorithm does not suffer from heteroskedastic errors w.r.t. W.

4 Empirical Illustration

We use data from the 2013 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

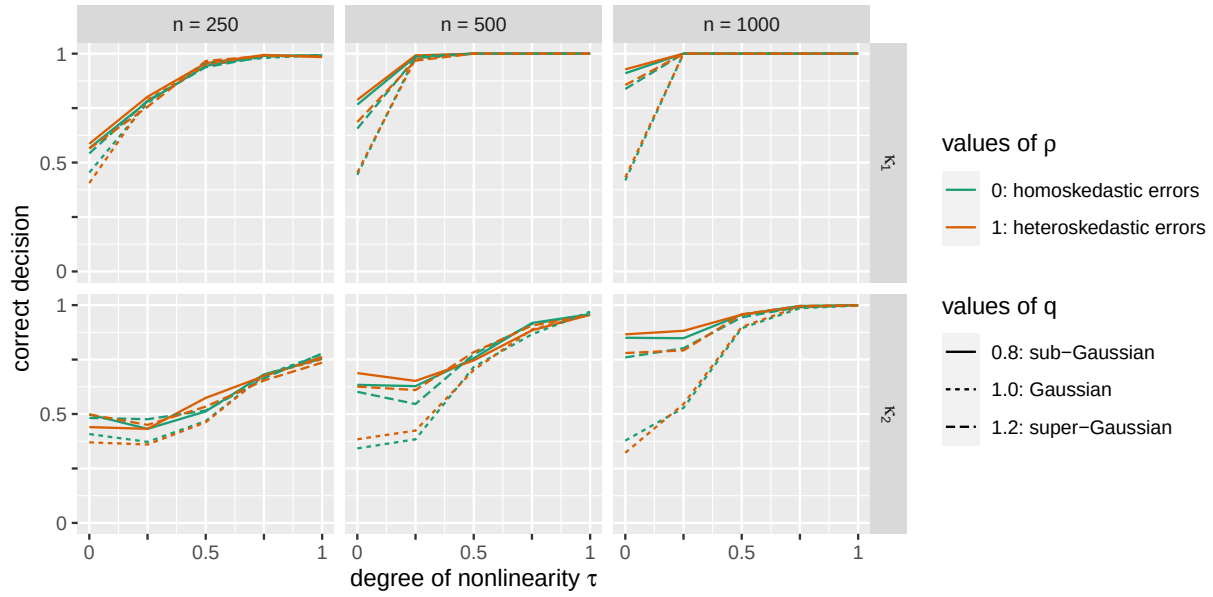


Figure 4: *Empirical probabilities of correct recovery of the causal direction when* $c_{\rho,q}$ *is chosen such that* $\mathrm{Var}(U) = 1$.

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 (Inc) and age where age can be interpreted as proxy for work experience (Exp). We posit the correct causal model to be

$$\text{Inc} = h(\text{Exp}, W) + \varepsilon_y, \tag{10}$$

where experience Exp causes income Inc. Vice versa, the anticausal model in which income causes work experience is given as

$$\text{Exp} = \widetilde{h}(\text{Inc}, W) + \varepsilon_e \tag{11}$$

where in each model W contains all remaining covariates as control (expenditure, highest educational attainment of the main earner, highest professional training of the main earner).

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 expenditure 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. This argument applies even more strongly as the expenditure distribution is split into a larger number of quantiles. 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 identify the correct causal direction. Therefore, we show results for a set of $n_q = \{4, \ldots, 20\}$ quantiles.[5] For each number of quantiles n_q, we run the causal discovery algorithm 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 5 refers to the

[5]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 $n_q = \{1, 2, 3\}$.

number of quantiles the income distribution is split in, not the quantiles as such). For example, the bar above $n_q = 5$ in Figure 5 denotes that in 4 of the 5 quantiles, i.e. 80%, the algorithm concludes that experience causes income. Regardless of n_q, the test always favours the model where work experience causes income in at least 50% of quantiles. The algorithm 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.

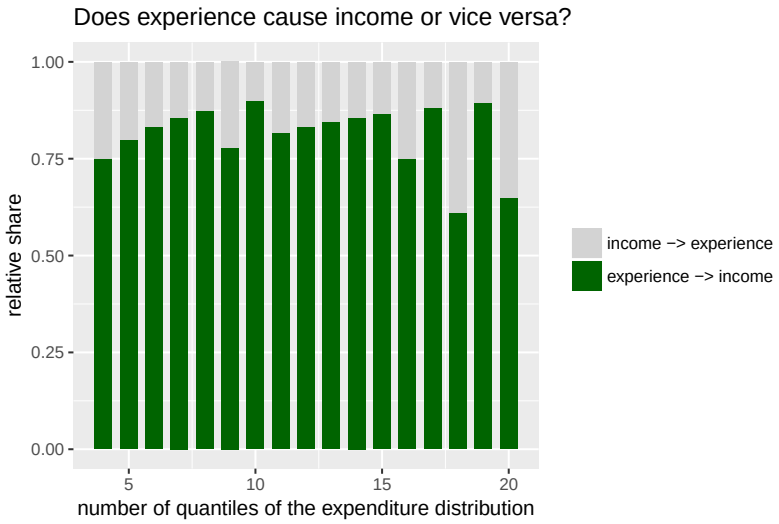


Figure 5: *This Figure shows the results of the empirical application of Algorithm* **??** *to the question whether work experience causes income or vice versa. The x-axis shows the number of quantiles that the expenditure distribution is split in (not to be mistaken with the quantiles as such). The stacked bars show the shares of the respective number of quantiles the algorithm decides the causal or anticausal model is the correct model.*

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 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 W and heteroskedasticity w.r.t. them and, thus, provide a heteroskedasticity-robust method to test for reverse causality. In addition, we show how this test can be extended to a bivariate causal discovery algorithm by comparing the test statistics of residual and purported cause of two candidate models. We extend known results on causal identification and causal discovery to settings with heteroskedasticity with respect to additional control covariates.

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 evidence that the true causal direction is from work experience to income. Though this result is not substantively surprising because income 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. This underscores the value of our proposed methodology to shed light on the causal structure of economic phenomena without resorting to exogenous variation.

6 Bibliography

Blundell, Richard and Joel Horowitz (2007). "A Non Parametric Test of Exogeneity". *Review of Economic Studies* 74.4, pp. 1035–1058.

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". *Econometric Theory* 36.4, pp. 583–625.

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.

Daudin, JJ (1980). "Partial association measures and an application to qualitative regression". *Biometrika* 67.3, pp. 581–590.

Engle, Robert, David Hendry, and Jean-Francois Richard (1983). "Exogeneity". *Econometrica* 51.2.

Engle, Robert F. and David F. Hendry (1993). "Testing superexogeneity and invariance in regression models". *Journal of Econometrics* 56.1-2, pp. 119–139.

Favero, Carlo and David F. Hendry (1992). "Testing the Lucas critique: A review". *Econometric Reviews* 11.3, pp. 265–306.

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*.

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.

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

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.

Heinze-Deml, Christina, Jonas Peters, and Asbjoern Marco Sinius Munk (2019). *CondIndTests: Nonlinear Conditional Independence Tests*. R package version 0.1.5.

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.

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.

Imbens, Guido and Donald Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press. ISBN: 9780521885881.

Lewbel, Arthur (2014). "An Overview of the Special Regressor Method". *Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*. Ed. by J. Racine A. Ullah and L. Su. Oxford: Oxford University Press.

Lopez-Paz, David (2016). "From Dependence to Causation". PhD thesis. Max Planck Institute for Intelligent Systems and University of Cambridge.

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.

Muandet, Krikamol, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf (2017). "Kernel Mean Embedding of Distributions: A Review and Beyond". *Foundations and Trends® in Machine Learning* 10.1-2, pp. 1–141.

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

Schölkopf, Bernhard and Alexander Smola (2001). *Learning with Kernels*.

Schölkopf, Bernhard, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio (2021). "Toward causal representation learning". *Proceedings of the IEEE* 109.5, pp. 612–634.

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.

Singh, Rahul, Maneesh Sahani, and Arthur Gretton (2019). "Kernel Instrumental Variable Regression". *ArXiv e-prints*, pp. 1–31. arXiv: `arXiv:1906.00232v1`.

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.

White, Halbert and Karim Chalak (2010). "Testing a conditional form of exogeneity". *Economics Letters* 109.2, pp. 88–90.

Zhang, Kun and Aapo Hyvärinen (2009). "On the Identifiability of the Post-Nonlinear Causal Model". *Uncertainty in Artificial Intelligence* 2013, pp. 1–8.

Zhang, Kun, 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.

Zhang, Rui, Masaaki Imaizumi, Bernhard Schölkopf, and Krikamol Muandet (2020). "Maximum Moment Restriction for Instrumental Variable Regression". *arXiv preprint arXiv:2010.07684*.

A Irreversibility Proof

Proof of Theorem 1. For the proof of the result, we proceed in three steps. First, for the reverse model (2), we establish restrictions on the third derivative of the conditional density of the model. Second, we perform this calculation explicitly for the model (1). Third, from the previous steps, we derive distributional restrictions when both models are satisfied.

Step 1. For the reverse model (2), we first derive an expression for the conditional density of X given (Y, W), and calculate a third-order derivative of that expression. Under the reverse model (2), the cumulative distribution function of X given (Y, W) satisfies

$$\begin{aligned} P(X \leq x|Y = y, W = w) &= P(\widetilde{h}(Y, W) + \widetilde{\varepsilon}\widetilde{\sigma}(W) \leq x|Y = y, W = w) \\ &= P\left(\widetilde{\varepsilon} \leq \frac{x - \widetilde{h}(Y, W)}{\widetilde{\sigma}(W)}|Y = y, W = w\right) \\ &= P\left(\widetilde{\varepsilon} \leq \frac{x - \widetilde{h}(y, w)}{\widetilde{\sigma}(w)}\right) \end{aligned} \tag{12}$$

where the last step uses the independence assumption $\widetilde{\varepsilon} \perp\!\!\!\perp (Y, W)$. Thus, we conclude for the conditional probability density functions that

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

and, in particular,

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

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

$$\pi(x, y, w) : = \log f(x, y|w) = \eta(y, w) + \widetilde{\nu}\left(\frac{x - \widetilde{h}(y, w)}{\widetilde{\sigma}(w)}\right)$$

for all (x, y, w) such that $f_{\widetilde{\varepsilon}}\big((x - \widetilde{h}(y, w))/\widetilde{\sigma}(w)\big) > 0$ and $f_{Y|W}(y, |w) > 0$. Taking partial derivatives yields

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

and

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

which, in turn, results in

$$\frac{\frac{\partial^2 \pi(x,y,w)}{\partial x^2}}{\frac{\partial^2 \pi(x,y,w)}{\partial x \partial y}} = -\frac{1}{\frac{\partial \widetilde{h}(y,w)}{\partial y}}. \tag{13}$$

Therefore, taking the derivative of the ratio in eq. (13) w.r.t. x, we conclude

$$\frac{\partial}{\partial x}\left(\frac{\frac{\partial^2\pi}{\partial x^2}}{\frac{\partial^2\pi}{\partial x \partial y}}\right) = 0. \tag{14}$$

Step 2. We derive restrictions similar to eq. (14) for the causal model in eq. (1). First, we derive an expression for the conditional density of $Y|X, W$. Similar to (12), we can write

$$P(Y \leq y|X = x, W = w) = P(h(X, W) + \varepsilon\sigma(W) < y|X = x, W = w) = P\left(\varepsilon \leq \frac{y - h(x, w)}{\sigma(w)}\right)$$

which uses the independence assumption $\varepsilon \perp\!\!\!\perp (X, W)$. This lets us conclude for the probability density functions

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

Therefore, the conditional density of $X, Y|W$ can be expressed as

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

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

$$\pi(x, y, w) : = \log f_{X,Y|W}(x, y|w) = \xi(x, w) + \nu\Big(\frac{y - h(x, w)}{\sigma(w)}\Big)$$

for all (x, y, w) such that $f_\varepsilon\big((x - h(y, w))\sigma(w)\big) > 0$ and $f_{Y|W}(y, |w) > 0$. Taking partial derivatives, we conclude

$$\begin{aligned}\frac{\partial^2\pi(x, y, w)}{\partial x^2} &= \frac{1}{\sigma^2(w)}\left\{\left(\frac{\partial h(x, w)}{\partial x}\right)^2 \nu''\left(\frac{y - h(x, w)}{\sigma(w)}\right) - \frac{\partial^2 h(x, w)}{\partial x^2}\nu'\left(\frac{y - h(x, w)}{\sigma(w)}\right)\right\} \\ &\quad + \frac{\partial^2\xi(x, w)}{\partial x^2} \\ &=: \phi_1(x, y, w) + \frac{\partial^2\xi(x, w)}{\partial x^2}\end{aligned} \tag{15}$$

and

$$\frac{\partial^2\pi(x, y, w)}{\partial x \partial y} = -\nu''\left(\frac{y - h(x, w)}{\sigma(w)}\right)\frac{\partial h(x, w)}{\partial x}\frac{1}{\sigma(w)^2} =: \phi_2(x, y, w). \tag{16}$$

In the following derivations we omit the arguments (x, w) for ξ, and (x, y, w) for ϕ_1 and ϕ_2. The ratio of eqs. (15) and (16) is given by

$$\frac{\frac{\partial^2\pi}{\partial x^2}}{\frac{\partial^2\pi}{\partial x \partial y}} = \frac{\phi_1(x, y, w) + \partial^2\xi(x, w)/\partial x^2}{\phi_2(x, y, w)} \tag{17}$$

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

$$\frac{\partial}{\partial x}\left(\frac{\frac{\partial^2\pi}{\partial x^2}}{\frac{\partial^2\pi}{\partial x \partial y}}\right) = \frac{\partial^3\xi/\partial x^3}{\phi_2} - \frac{(\partial^2\xi/\partial x^2)(\partial\phi_2/\partial x)}{\phi_2^2} + \frac{(\partial\phi_1/\partial x)\phi_2 - (\partial\phi_2/\partial x)\phi_1}{\phi_2^2}. \tag{18}$$

Step 3. If the reverse model (2) holds, we know from (14) that (18) must equal zero. By setting (18) equal to zero and given h, ν, we obtain for each fixed y and w, which we denote $\bar{y}$ and $\bar{w}$, respectively, a linear inhomogenous differential equation for ξ:

$$\frac{\partial^3\xi(x,\bar{w})}{\partial x^3} = \frac{\partial^2\xi(x,\bar{w})}{\partial x^2}G_1(x,\bar{y},\bar{w}) + G_2(x,\bar{y},\bar{w}). \tag{19}$$

where $G_1 = \frac{\partial\phi_2/\partial x}{\phi_2}$ and $G_2 = \frac{(\partial\phi_2/\partial x)\phi_1 - (\partial\phi_1/\partial x)\phi_2}{\phi_2}$. Making use of the notation $\chi(x,w) := \partial^2\xi(x,w)/\partial x^2$, we may write

$$\frac{\partial\chi(x,\bar{w})}{\partial x} = \chi(x,\bar{w})G_1(x,\bar{y},\bar{w}) + G_2(x,\bar{y},\bar{w}),$$

which completes the proof. □

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

$$\chi(x,\bar{w}) = \chi(x_0,\bar{w})e^{\int_{x_0}^{x} G_1(\widetilde{x},\bar{y},\bar{w})d\widetilde{x}} + \int_{x_0}^{x} e^{\int_{\widehat{x}}^{x} G_1(\widetilde{x},\bar{y},\bar{w})d\widetilde{x}} G_2(\widehat{x},\bar{y},\bar{w})d\widehat{x}.$$

Following Hoyer et al. (2009), we can see that the set of all functions satisfying the condition in eq. (19) is a 3-dimensional affine space. We can fix $\xi(x_0,w)$, $\xi'(x_0,w)$, $\xi''(x_0,w)$ for some arbitrary x_0, and w at $\bar{w}$, which determines ξ. Given fixed f and ν, and arbitrary $\bar{w}$, the set of all ξ admitting an anticausal model is, thus, contained in a three-dimensional subspace, and therefore not generic.

B Construction of the RKHS

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

   $$k(v,v') := \exp\Big(-\frac{\|v-v'\|_{\ell_2}^2}{\lambda}\Big).$$

   for arbitrary vectors v and v' and parameter $\lambda > 0$. This kernel 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. $k(\cdot,x)$. 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 $\mathbb{R}^{\mathcal{X}}$:

   $$\begin{aligned}\Phi:\ \ &\mathcal{X} \to \mathbb{R}^{\mathcal{X}}\\ &x \mapsto k(\cdot,x)\end{aligned}$$

where k is a positive-definite kernel. A positive-definite kernel is a kernel whose associated kernel matrix K, which has entries $K_{ij} := k(x_i, x_j)$, 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

$$f(\cdot) = \sum_{j=1}^{m} \alpha_j k(\cdot, x_j)$$

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

3. Given a similarly constructed function

$$g(\cdot) = \sum_{l=1}^{m'} \beta_l k(\cdot, x'_l)$$

with $\beta_l \in \mathbb{R}$ 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

$$\langle f, g \rangle := \sum_{j=1}^{m} \sum_{l=1}^{m'} \alpha_j \beta_l k(x_j, x'_l) \tag{20}$$

4. Then complete the space spanned by (20) 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 $\mathcal{H}$:

$$\langle k(\cdot, x), f \rangle = f(x),$$

see also Schölkopf and Smola (2001, Section 2.2., p.33) for more details. In particular, we obtain

$$\langle k(\cdot, x), k(\cdot, x') \rangle = \langle \Phi(x), \Phi(x') \rangle = k(x, x').$$

C Further Simulation Results

In this section, we present further simulations results analogous to Figures 2, 3, 4 but with $\mathrm{Var}(U) = 0.8$ and $\mathrm{Var}(U) = 1.2$.

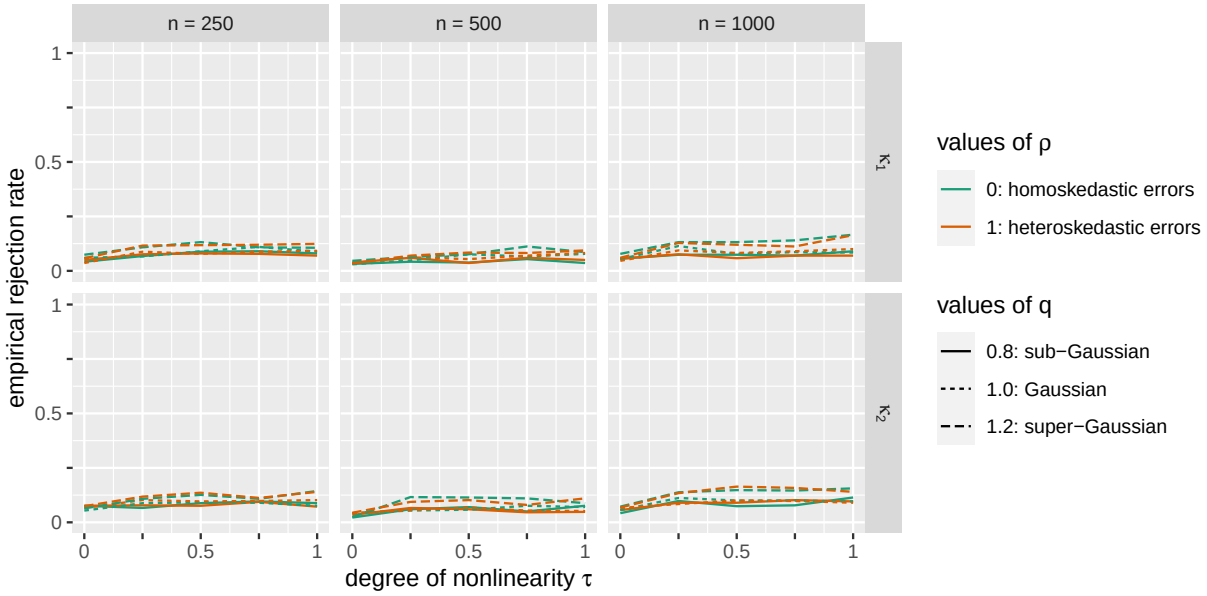


Figure 6: *Empirical rejection rates of independence of estimated residual and* X *of model* (1) *when* $c_{\rho,q}$ *is chosen such that* $\mathrm{Var}(U) = .8$.

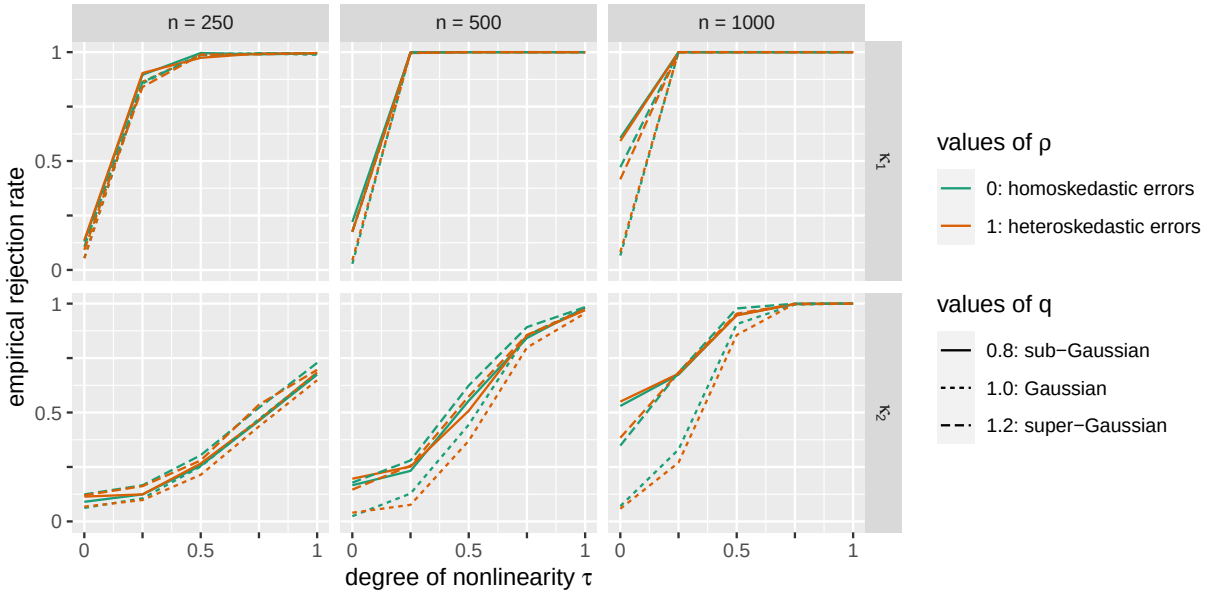


Figure 7: *Empirical rejection rates of independence of estimated residual and* Y *of model* (2) *when* $c_{\rho,q}$ *is chosen such that* $\mathrm{Var}(U) = .8$.

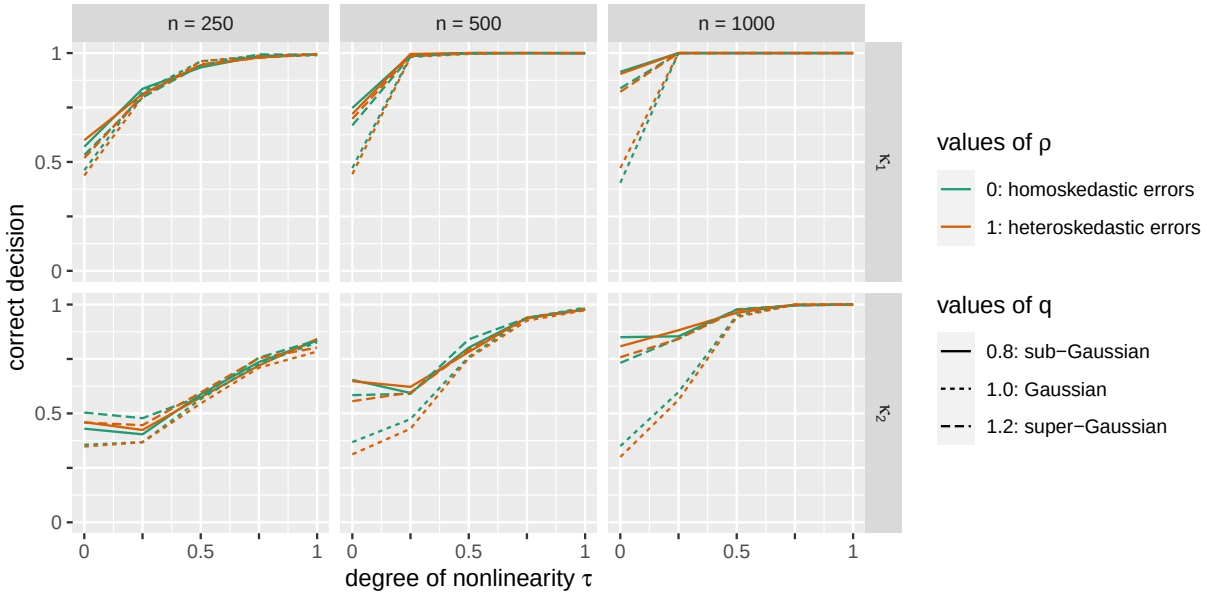


Figure 8: *Empirical probabilities of correct recovery of the causal direction when* $c_{\rho,q}$ *is chosen such that* $\mathrm{Var}(U) = .8$.

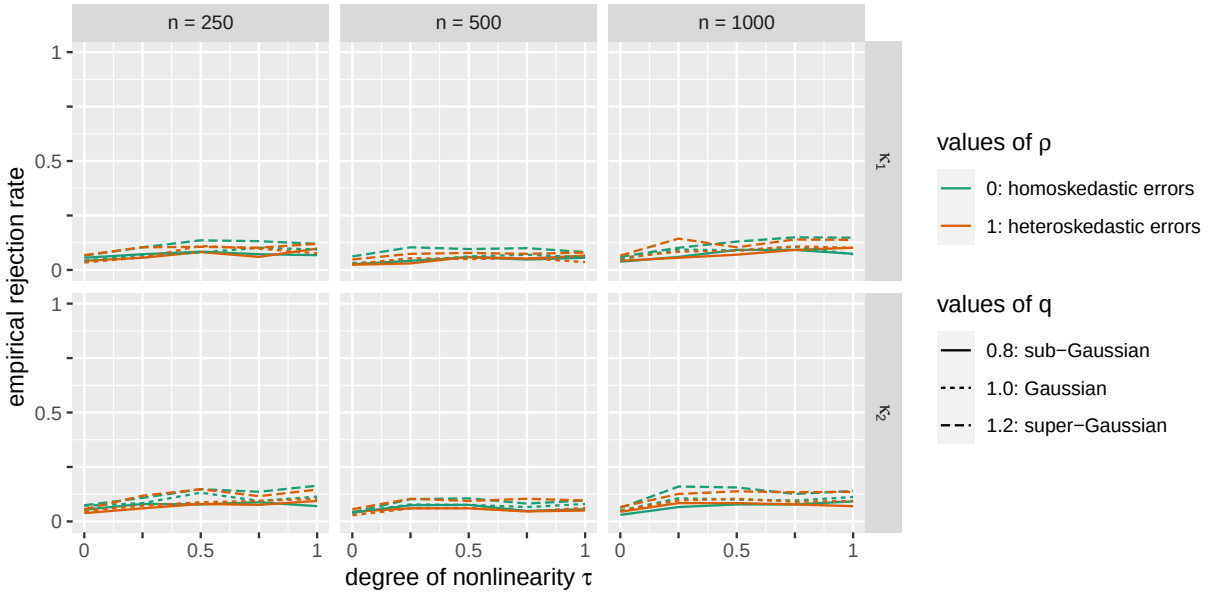


Figure 9: *Empirical rejection rates of independence of estimated residual and* X *of model* (1) *when* $c_{\rho,q}$ *is chosen such that* $\mathrm{Var}(U) = 1.2$.

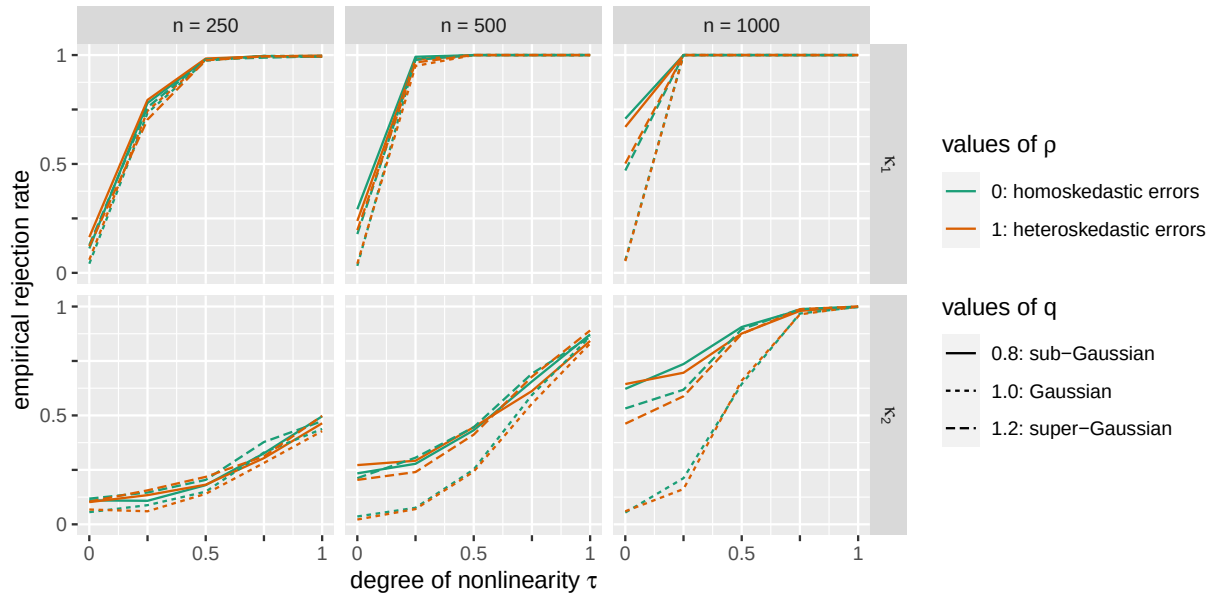


Figure 10: *Empirical rejection rates of independence of estimated residual and* Y *of model* (2) *when* $c_{\rho,q}$ *is chosen such that* $\mathrm{Var}(U) = 1.2$.

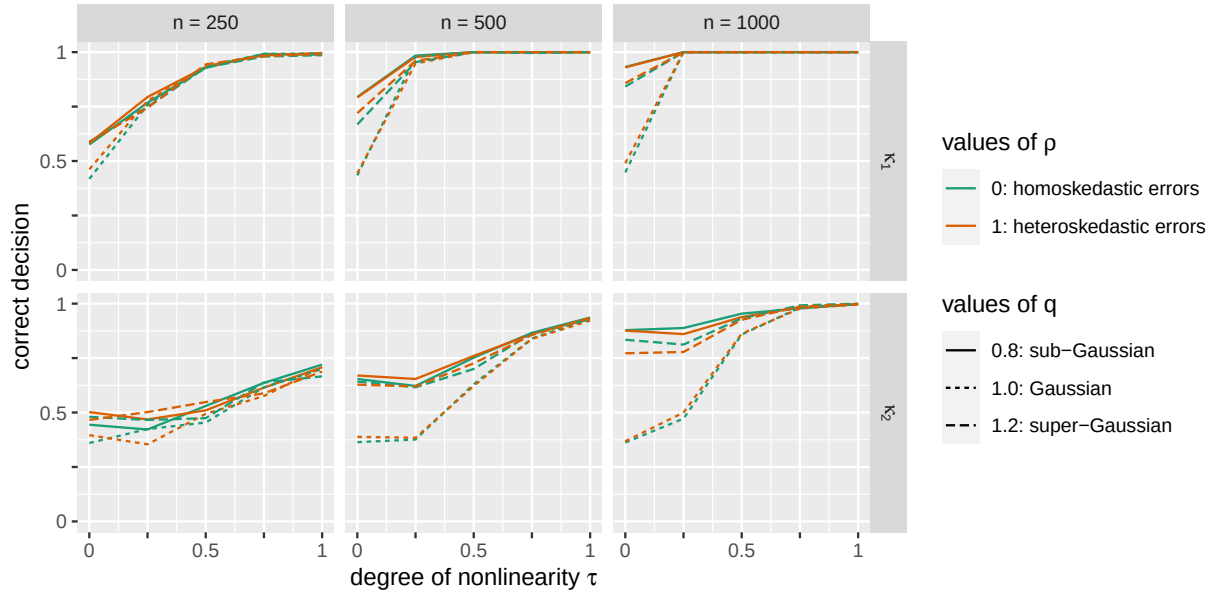


Figure 11: *Empirical probabilities of correct recovery of the causal direction when* $c_{\rho,q}$ *is chosen such that* $\mathrm{Var}(U) = 1.2$.