Psychological Test and Assessment Modeling, Volume 61, 2019 (4), 365-387

Identification of confounded subgroups using linear model-based recursive partitioning

Michael P. van Wie[1], *Xintong Li*[2] & *Wolfgang Wiedermann*[3]

Abstract

The absence of confounding is the fundamental assumption to endow parameters of a statistical model with causal meaning. Causal inference is prone to biases due to confounding when data are purely observational. Often the assumption of unconfoundedness may be too rigid for the entire population under study, but may be plausible for subpopulations. The present article introduces an approach to detect confounded subgroups in linear regression models through combining a recently proposed confounder detection approach based on kernel-based independence testing with model-based recursive partitioning. Results of a simulation study indicate that Bonferroni-corrected independence tests are able protect the (family-wise) Type I error rate of multiple independence testing across recursively partitioned local models. We discuss data scenarios under which the proposed approach can be expected to show adequate statistical power to detect confounded subgroups. Data requirements to ensure best practice for applications and strategies to further improve the statistical power of the approach are discussed.

Keywords: causal inference, recursive partitioning, confounding, Hilbert-Schmidt independence criterion, non-normality

[1] University of Missouri

[2] University of Missouri

[3] *Correspondence concerning this article should be addressed to:* Wolfgang Wiedermann, PhD, Statistics, Measurement, and Evaluation in Education, Department of Educational, School, and Counseling Psychology, College of Education, University of Missouri, 13B Hill Hall, Columbia, MO, 65211, USA; email: wiedermannw@missouri.edu

This article discusses model-based regression tree methods from the perspective of causal inference (Wiedermann & von Eye, 2016). While the presence of confounding is well known to bias causal inference, the effect of hidden confounders on the performance of regression tree algorithms (well-suited to study causal effect heterogeneity) is less known. The present study evaluates the robustness of a model-based regression tree algorithm against hidden confounding and introduces a confounder detection approach that makes use of non-normality of variables to test the independence assumption of linear models. Because violations of the independence assumption are characteristic for confounded (sub)samples, the approach presented in this article can be used to detect (un)confounded subgroups in a purely data-driven way.

Randomized designs are the gold-standard to estimate the causal effect of an explanatory variable (the predictor or regressor) on a dependent variable (the outcome or regressand). In many experimental settings, randomization enables researchers to evaluate a treatment effect without the need to consider all possibly relevant covariates (i.e., additional variables which may have an effect on the outcome variable). However, in practical applications, covariates are routinely included in the analysis of data obtained from randomized controlled trials (RCTs) to increase precision of causal effect estimates and statistical power to detect treatment effects. In observational studies (i.e., when randomization is not feasible, e.g., due to ethical or financial constraints), researchers are commonly advised to collect and consider (a potentially large number of) covariates to statistically control for potential confounding factors. While statistical adjustment alone can never be a sufficient replacement for randomization, the hope here is that it is "good enough" to make valid statements about the causal effect under study. Several statistical approaches are available for causal inference in observational data. For example, regression discontinuity designs (Thistlethwaite & Campbell, 1960) can be used to quantify causal effects by assigning subjects to "control" and "treatment" groups according to a pre-determined cut-off value of a pre-program measure. Propensity score techniques (Rosenbaum & Rubin, 1983) are available to reduce the effect of confounders through accounting for covariates that predict treatment status. As a third option, instrumental variable (Imbens & Angrist, 1994) approaches exist to estimate causal effects in the presence of confounding.

A question that is closely tied to the identification of causal effects, is whether the causal effect is constant for all subjects under study or whether effect heterogeneity is present. The latter case describes situations in which the causal effect systematically differs across subpopulations. Such subpopulations are usually defined by additional subject characteristics (so-called moderators; eligibility criteria for moderators are given in Kraemer, Kiernan, Essex, and Kupfer, 2008). Research on moderated causal effects helps to inform theories about the exact conditions under which causal effects can be expected to be large (in experimental settings, such follow-up analyses help to identify "for whom" the intervention works best). However, standard moderation/subgroup analysis can give misleading results when testing purely exploratory (data-driven) hypotheses without accounting for multiple testing (Wang & Ware, 2013). Without proper adjustment, the probability of a false positive test result increases with the number of subgroup/interaction tests performed. For example, when the causal effect is constant for all

subjects and one performs 10 independent subgroup/interaction tests, the probability of finding at least one significant interaction effect is about 40% (Lagakos, 2006).

The machine learning literature has developed a variety of statistical methods to maximize predictive accuracy of outcomes as a function of covariates, one of them being regression trees. Regression tree techniques have also been discussed in the context of testing causal effect heterogeneity. For example, Dusseldorp and Van Mecherlen (2014) suggested so-called qualitative interaction trees (QUINT) to evaluate whether the effectiveness of two treatments is equal for all subgroups of subjects (see also Doove et al., 2016). Athey and Imbens (2016) used "honest estimation" (a modified classification and regression tree [CART] algorithm) to identify subpopulations that differ in the magnitude of their treatment effects while preserving validity of confidence intervals of causal effects.

The present study focuses on another extension of the conventional CART algorithm, model-based recursive partitioning (MOB; Zeileis, Hothorn & Hornik, 2008). Recently, Fokkema et al. (2018) discussed MOB to identify treatment-subgroup interactions in the context of nested (multilevel) data. MOB is commonly described as a method that seeks to find "better fitting" local (subgroup-specific) statistical models compared to a global model based on the total sample.

Recursive partitioning techniques are well-suited to 1) increase the predictive performance and 2) capture interaction effects and complex nonlinear relations. While it is well-known that minimal changes in the data can change either the variables and/or the cutpoints selected for building a regression tree (Li & Belford, 2002; Philipp, Zeileis, & Strobl, 2016), violations of statistical model assumptions impose additional challenges on finding stable tree structures. The common characterization of regression tree methods as tools to find "better fitting models" might be misleading with respect to the assumptions made for the statistical model of interest. It is important to realize that submodels resulting from recursive partitioning rest on exactly the same statistical assumptions as the global model. In other words, any application of MOB needs to be complemented by a critical evaluation of model assumptions using regression diagnostics. The present study focuses on the absence of confounding assumption (i.e., independence of the counterfactual outcomes and the exposure[4]; cf. VanderWeele & Shpitser, 2013) in the context of recursively partitioned linear models using observational (non-experimental) data. Absence of confounding is the fundamental assumption to interpret model parameter estimates as causal (Pearl, 2009). In practical applications, the absence of confounding assumption may often be too rigid for the total sample, but might hold for certain subgroups.

The aims of the present article are two-fold: 1) to evaluate the impact of confounding on the performance of MOB in the context of the ordinary least square (OLS) regression model and 2) to combine MOB with a recently proposed confounder detection approach

[4] Absence of confounding is also referred to as "ignorability" (Rubin, 1978), "exchangeability" (Greenland & Robins, 1986), "selection on observables" (Barnow, Cain, & Goldberger, 1980) or "exogeneity" (Imbens, 2004).

for non-normal variables (Wiedermann & Li, 2018, 2019). The latter enables researchers to test the crucial assumption of unconfoundedness in local (subgroup-specific) models. The remainder of the article is structured as follows: We start with introducing the theoretical foundations of model-based recursive partitioning. We then briefly review the assumption of unconfoundedness in the standard OLS regression model and consequences of assumption violations and show that, in the presence of confounders, stochastic non-independence of regressors and model errors becomes testable when variables deviate from the Gaussian distribution. Then, we introduce a kernel-based measure of independence that can be used to detect dependence patterns of linearly uncorrelated variables and propose a simple stepwise procedure to detect (un)confounded subgroups of a sample. Results from a Monte-Carlo simulation study are presented which (1) quantify the impact of unobserved confounders on the accuracy of MOB regression trees and (2) evaluate the performance of kernel-based tests of independence to detect (un)confounded subgroups. The article closes with a discussion of data requirements and analytic strategies to guarantee best practice application of the proposed approach.

Model-based recursive partitioning

Tree-based methods are valuable alternatives to standard parametric methods and have extensively been studied in the past (see, e.g., Breiman et al., 1984; Hothorn, Hornik & Zeileis, 2006; Quinlan, 1993; Morgan & Sonquist, 1963; Strobl et al., 2009; Zhang & Singer, 2010). The ability to automatically detect interactions and nonlinearities paired with straight-forward interpretation and visualization makes them useful statistical tools for applied researchers. In conventional CART algorithms, the covariate space is recursively partitioned to identify subgroups with different values of an outcome variable. In contrast, MOB uses parameters of a model (instead of values of an outcome) as the basis for recursive partitioning. In other words, the MOB algorithm partitions a set of covariates by evaluating parameter instabilities of a model (e.g., the linear regression model). Identifying a significant instability with respect to a partitioning covariate implies that subgroup-specific (conditional) effects exist in the dataset. MOB can be used to estimate such conditional effects and identify the corresponding subgroups. More specifically, a parametric model is formulated to represent a theory-driven empirical question (in the present study, the parametric model of interest is the standard linear regression model and the parameters of interest are the regression slopes). Following the formulation of the research question, the corresponding parametric model is fed into the MOB algorithm that tests whether relevant covariates exist which would alter the parameters of the model. Regression trees proceed through a search of all possible splits (algorithmic details are given below). A large tree is constructed and then pruned back with a cross-validation scheme, to avoid over-fitting. The MOB algorithm terminates in end nodes, each of which consists of a local parametric model.

The MOB algorithm consists of four steps: Step one, *parameter estimation*, starts by fitting the model $M(Y,\theta)$ (with Y being the dataset and θ describing the model parame-

ters), to all observations in a node by estimating θ via minimization of the objective function Ψ (usually the negative log-likelihood)

$$\hat{\theta} = \text{argmax}_{\theta} \sum_{i=1}^{n} \Psi\ (y_i, \theta)$$

with

$$\log L\left(\theta | y_1, \ldots y_n\right) = \sum_{i=1}^{n} \Psi(y_i, \theta)$$

and $\Psi\left(y_i, \theta\right)$ being the likelihood contribution of the *i*-th subject (i = 1, ..., n). The second step, *testing parameter instability*, assesses parameter estimates with respect to every ordering of the partitioning variables, $Z_1, \ldots, Z_t (j = 1, \ldots, t)$. Under the null hypothesis of parameter stability, we do not expect systematic structural changes. In contrast, parameter instabilities are present when one or more of the model parameters change significantly due to the ordering caused by a partitioning variable Z_j. Here, the subject-wise score/estimating function (i.e., the derivative of the log-likelihood distribution with respect to θ)

$$\psi\left(y_i, \theta\right) = \frac{\partial \Psi\left(y_i, \theta\right)}{\partial \theta}$$

is a general measure of deviations in log-likelihood based models. For OLS regression models, the score function is given by the product of the OLS residuals and the model matrix. These deviations are cumulatively aggregated along the (ordered) covariates and generalized *M*-fluctuation tests (Zeileis & Hornik, 2007) are used to test stability of the score function. Because the number of partitioning variables can be large, fluctuation tests should be corrected for multiple testing (e.g., using Bonferroni adjustment). If parameter instability is identified, the variable with the smallest *p*-value is selected. The third step, *splitting*, computes the split points that locally optimize the partitioned likelihood (i.e., the sum of the likelihoods before and after a split point τ)

$$\sum_{i \in L(\tau)} \Psi(y_i, \hat{\theta}^{(L)}) + \sum_{i \in R(\tau)} \Psi(y_i, \hat{\theta}^{(R)})$$

with $\hat{\theta}^{(L)}$ and $\hat{\theta}^{(R)}$ being the model parameters based on the two subsamples before (i.e., $L\left(\tau\right) = \{i | Z_{ij} \leq \tau\}$) and after the split point (i.e., $R\left(\tau\right) = \{i | Z_{ij} > \tau\}$). The entire procedure (Steps 1 – 3) is repeated until no parameter instabilities are detected or the number of subjects in a subsample is smaller than an a priori selected minimum node size (e.g., $n \geq$ 10). Each terminal node consists of a subgroup-specific local (parametric) model $M_k\left(Y, \theta_k\right)$ (k = 1, …, K) with subgroup-specific model parameters θ_k.

Confounders in linear regression models

While researchers in the psychological, educational, and behavioral sciences often loosely define a confounder (u) as a variable that is simultaneously associated with the focal predictor (x) and the outcome (y) without linking x and y in the sense of a mediational causal chain ($x \rightarrow u \rightarrow y$), this definition is inadequate from the perspective of "confounder control" to eliminate biases in causal effect estimates: There exist covariates that are associated with the predictor and the outcome, the control of which introduces (rather than eliminates) biases in causal effect estimates (VanderWeele & Shpitser, 2013). Consider, for example, a causal mechanism of the form $x \rightarrow u \leftarrow y$, i.e., u is a common effect of x and y. While u is in line with the somewhat loose definition of "simultaneous association with x and y", controlling for (or conditioning on) u induces a bias in the causal effect estimate while ignoring u in the analysis would lead to an unbiased estimate of the causal effect. This phenomenon is known as a collider-bias (Elwert & Winship, 2014). A more rigorous definition of a confounder has been proposed by VanderWeele and Shpitser (2013)[5].

The adverse effects of the presence of a confounder can be illustrated as follows. Suppose that the "true" underlying data-generating mechanism in the k-th subgroup can be written as (without loss of generality, we assume that model intercepts are zero)

$$x^{[k]} = b_{xu}^{[k]} u^{[k]} + e_x^{[k]}$$
$$y^{[k]} = b_{yx}^{[k]} x^{[k]} + b_{yu}^{[k]} u^{[k]} + e_y^{[k]}$$

with $b_{yx}^{[k]}$ being the causal effect of interest, $b_{xu}^{[k]}$ and $b_{yu}^{[k]}$ being the regression slopes of the confounder u of the k-th local regression model where $b_{xu}^{[k]} b_{yu}^{[k]} \neq 0$ for at least one of the K local models. Let $\hat{b}_{yx}^{[k]}$ be the estimated causal effect of the k-th local model. Further, $e_x^{[k]}$ and $e_y^{[k]}$ denote error terms which are assumed to be independent of the corresponding regressors and of each other and $\hat{e}_x^{[k]}$ and $\hat{e}_y^{[k]}$ are the observed model residuals in the k-th local model. When controlling for u, the regression coefficient $\hat{b}_{yx}^{[k]}$ is an unbiased estimate of the "true" causal effect of x on y, that is, $E\left[\hat{b}_{yx}^{[k]}\right] = b_{yx}^{[k]}$ (with E being the expected value operation).

---

[5] These authors focused on two fundamental properties that need to be fulfilled for an adequate definition of a confounder, 1) whether control for all confounders is sufficient to control for confounding and 2) whether each confounder can be used to eliminate or reduce the confounding bias. Based on these two necessary properties, the authors defined a confounder "… as a pre-exposure covariate C for which there exists a set of other covariates X such that effect of the exposure on the outcome is unconfounded conditional on (X, C) but such that for no proper subset of (X, C) is the effect of the exposure on the outcome unconfounded given the subset" (p. 196).

Next, suppose that u has not been observed (or, equivalently, u is erroneously omitted from the model). In this case, the model can be written as

$$y^{[k]} = b_{yx}^{'[k]} x^{[k]} + e_y^{'[k]}$$

and $\hat{b}_{yx}^{'[k]}$ will now be a biased estimate for $b_{yx}^{[k]}$. Specifically, one obtains

$$E\left[\hat{b}_{yx}^{'[k]}\right] = b_{yx}^{[k]} + b_{yu}^{[k]} \frac{cov\left(x^{[k]}, u^{[k]}\right)}{\sigma_{x^{[k]}}^2}$$

with $cov\left(x^{[k]}, u^{[k]}\right)$ being the covariance of $x^{[k]}$ and $u^{[k]}$. From the above equation, it follows that $b_{yx}^{'[k]} \neq b_{yx}^{[k]}$ when $b_{yu}^{[k]} \neq 0$ and $cov\left(x^{[k]}, u^{[k]}\right) \neq 0$ (or, equivalently, $b_{xu}^{[k]} \neq 0$).

Common confounder detection approaches rely on so-called instrumental variables (IVs). IVs are used to isolate that part of the predictor variation that is not influenced by the confounder. In general, two conditions need to be met to ensure that an IV is reliable (see, e.g., Pearl, 2009): First, the IV must be independent of all exogenous factors that affect the outcome when the predictor of interest is held constant (known as exclusion restriction). Second, the IV is assumed to be correlated with the predictor of interest (known as the strength of an IV). While a "weak" IV is likely to produce a biased effect estimate (Bound, Jaeger, & Baker, 1995), the exclusion restriction assumption cannot be tested using standard methods of correlation and regression in just-identified models (i.e., models with as many predictors as IVs). Therefore, strong substantial rationale is needed to justify the role of a variable as an IV. When an IV is available, a Hausman-type specification test (Hausman, 1978) can be used to test the equality of an IV-based two-stage least square effect estimate (b_{IV}) and the standard OLS estimate (b_{OLS}) where $b_{IV} = b_{OLS}$ holds under unconfoundedness. Because IVs may be hard to come by in practical applications, we focus on testing the assumption of unconfoundedness without requiring IVs. Instead of making use of additional external data information, the present approach assumes that variables under study are non-normally distributed. Under non-normality, asymmetry patterns of the independence assumption inherent to the linear regression model (i.e., regressands are assumed to be independent of the error term) emerge. Such independence properties have been used in the past in the development of causal learning algorithms (Shimizu et al., 2011), confirmatory methods of testing causal effect directionality (Wiedermann & von Eye, 2015, Wiedermann & Li, 2018), and automated covariate selection algorithms (Entner, Hoyer, & Spirtes, 2012). Further, Wiedermann and Li (2019) used these independence properties to detect confounding in linear models.

Confounder detection under non-normality

The confounder detection approach proposed by Wiedermann and Li (2019) assumes that 1) the relation of the two focal variables (x and y) can be described by the linear regression model (the issue of nonlinear relations is addressed in the Discussion section), 2) the predictor is exogenous, continuous, and non-normally distributed (i.e., the cause of x lies outside the model, x is at least interval-scaled, and x deviates from the perfect Gaussian distribution), and 3) the error term of the unconfounded model is non-normally distributed and independent of all regressors. The theoretical foundations for detecting confounders under non-normality are summarized in the so-called Darmois-Skitovich (DS) theorem (Darmois, 1953; Skitovich, 1953). The DS theorem states that if two stochastically independent variables v_1 and v_2 are linear functions of the same independent random variables w_i ($i = 1, \ldots, l$ with $i \geq 2$),

$$v_1 = \sum_i^l \alpha_i w_i \quad \text{and} \quad v_2 = \sum_i^l \beta_i w_i$$

(with α_i and β_i being constants), then all component variables w_i where $\alpha_i \beta_i \neq 0$ follow a normal distribution. The reverse corollary, therefore, implies that if a common variable w_i exists that is *non-normal*, then v_1 and v_2 must be *non-independent*. It is easy to show that this reverse corollary applies in the context of the linear regression model whenever a confounder u is present and the variables under study deviate from the normal distribution (for notational simplicity, in the following, we drop the subgroup index k). The regression model is then given by

$$\begin{aligned} x &= b_{xu} u + e_x \\ y &= b_{yx} x + b_{yu} u + e_y \\ &= \left(b_{yx} b_{xu} + b_{yu} \right) u + b_{yx} e_x + e_y . \end{aligned}$$

Thus, the error term of the mis-specified model $y = b'_{yx} x + e'_y$ can be written as

$$\begin{aligned} e'_y &= y - b'_{yx} x \\ &= \left(b_{yx} b_{xu} + b_{yu} \right) u + b_{yx} e_x + e_y - b'_{yx} \left(b_{xu} u + e_x \right) \\ &= \left[b_{yu} + \left(b_{yx} - b'_{yx} \right) b_{xu} \right] u + \left(b_{yx} - b'_{yx} \right) e_x + e_y \end{aligned}$$

from which follows that x and e'_y consist of the same common component variables u and e_x whenever $\left(b_{yx} - b'_{yx} \right) \neq 0$ (which holds by definition when confounding is present). According to the DS theorem, x and e'_y will be non-independent when at least one of the two component variables (u and e_x) are non-normal. Because x is a convolution of u and e_x, it follows that non-normality of x implies that at least one of the two

components is non-normal. In contrast, under $b_{yu} = 0$ one obtains $\left(b_{yx} - b_{yx}^{'}\right) = 0$ and the above equation reduces to $e_y^{'} = e_y$. As a consequence, *x* will be independent of $e_y^{'}$ because of its independence of e_y. When *x* and *u* are independent, it follows that $b_{xu} = 0$ which, again, implies $\left(b_{yx} - b_{yx}^{'}\right) = 0$ and the above equation reduces to $e_y^{'} = b_{yu}u + e_y$. Again, $e_y^{'}$ and *x* will be independent due to the independence of *u, x*, and e_y. Irrespective whether confounding is present or not, estimated residuals $\hat{e}_y^{'}$ will always be uncorrelated with *x*. Therefore, to test the presence of confounders, statistical inference methods beyond linear uncorrelatedness are needed. In the following section, we introduce such methods, so-called kernel-based independence tests.

Testing non-independence in uncorrelated variables

As shown in the previous section, under non-normality of variables, hidden confounders can be detected through evaluating stochastic independence of a (linearly uncorrelated) regressor (*x*) and the residuals ($\hat{e}_y^{'}$). Stochastic independence of two variables (v_1 and v_2) is defined as $E\left[f\left(v_1\right)g\left(v_2\right)\right] - E\left[f\left(v_1\right)\right]E\left[g\left(v_2\right)\right] = 0$ for any absolutely integrable functions $f(\bullet)$ and $g(\bullet)$. Thus, stochastic independence implies uncorrelatedness, however, the reverse statement does not hold, i.e., a zero first-order correlation does not imply stochastic independence. In principle, tests of stochastic independence can be constructed by inserting functions $f(\bullet)$ and $g(\bullet)$ and evaluate whether $cov\left(f\left(v_1\right)g\left(v_2\right)\right) = 0$ holds (so-called non-linear correlation tests; see, e.g., Hyvärinen, Karhunen & Oje, 2001). However, inserting all possible functions is not feasible in practical applications which implies that such tests show inflated Type II errors (Wiedermann, Artner, & von Eye, 2017). Therefore, we focus on a kernel-based measure of stochastic independence – the Hilbert-Schmidt Independence Criterion (HSIC; cf. Gretton et al., 2008) – which can be shown to be an omnibus measure for detecting any form of dependence in the large sample limit.

The HSIC is defined as follows: Let v_1 and v_2 be two continuous variables with sample size *n* and define $H = I - n^{-1}11^T$ with I being an *n*-th order identity matrix, and 1 being a $n \times 1$ vector of ones (1^T is the transpose of 1). Further, K and L are $n \times n$ matrices with cell entries $k_{ij} = k\left(v_{1(i)}, v_{1(j)}\right)$ and $l_{ij} = l\left(v_{2(i)}, v_{2(j)}\right)$ where k and l define Gaussian kernels of the form $k\left(v_1, v_1^T\right) = \exp\left(-\sigma^{-2} \| v_1 - v_1^T \|^2\right)$ and $l\left(v_2, v_2^T\right) = \exp\left(-\sigma^{-2} \| v_2 - v_2^T \|^2\right)$ with $\| v_1 - v_1^T \|^2$ and $\| v_2 - v_2^T \|^2$ being squared Euclidean distances of v_1 and v_1^T and v_2 and v_2^T. The parameter σ represents a bandwidth

parameter which can be determined using the so-called median heuristic (i.e., the median of all pairwise Euclidian distances; Sriperumbudur et al., 2009). The HSIC statistic can be obtained through $\text{HSIC} = n \bullet \hat{T}_n$ where $\hat{T}_n$ is based on the trace of the matrix product $KHLH$, $\hat{T}_n = 1/n^2\, trace\left(KHLH\right)$. When v_1 and v_2 are stochastically independent, $\hat{T}_n$ approximates zero. In contrast, if the HSIC significantly deviates from zero, the null hypothesis of independence of v_1 and v_2 can be rejected. To test the null hypothesis of stochastic independence, Gretton et al. (2008) suggested to approximate the null distribution of T_n as a two parameter gamma-distribution.

In the present study, we propose to apply the HSIC test to detect potential confounding in recursively partitioned linear regression models by making use of the following stepwise procedure:

1. Use the MOB algorithm to obtain *K* local (subgroup-specific) linear regression models $y^{[k]} = b_{yx}^{[k]} x^{[k]} + e_y^{[k]}$ (k = 1, …, K) with $y^{[k]}$ and $x^{[k]}$ being the outcome and predictor scores for subgroup *k*.
2. Extract the OLS residuals of the *k*-th local model, i.e., $\hat{e}_y^{[k]} = y^{[k]} - \hat{b}_{yx}^{[k]} x^{[k]}$.
3. Evaluate whether stochastic independence holds for $x^{[k]}$ and $\hat{e}_y^{[k]}$ using the HSIC test and reject the null hypothesis of independence if the HSIC significantly differs from zero.

To adjust for multiple testing across the *K* subgroups, a simple Bonferroni correction can be applied. In other words, given *K* local models, the significance of the HSIC statistic is evaluated using a nominal significance level of $\alpha^* = \alpha / K$.

Monte-Carlo simulation

To quantify the impact of unobserved confounders on the accuracy of MOB regression trees and to assess the Type I error and power properties of the HSIC test in the context of detecting local confounding, a simulation experiment was performed using the R statistical programing environment (R Core Team, 2019). The data generating mechanism is given in Figure 1. Each simulated dataset (n = 1000) consisted of four different subgroups with Z_1, Z_2, and Z_5 representing the relevant partitioning variables among a set of t covariates. The four subgroups differed in the magnitude of the causal effect ($b_{yx}^{[k]}$; intercepts were fixed at zero throughout the study) and in the presence/absence of an unobserved confounder. In subgroup 1 (i.e., cases where $Z_1 \leq 17$ and $Z_2 \leq 30$), the outcome scores were generated using the unconfounded model $y^{[1]} = b_{yx}^{[1]} x^{[1]} + e_y^{[1]}$ with $b_{yx}^{[1]} = 0.75$. In subgroups 2 and 3 (i.e., when $Z_1 > 17$ and $Z_2 \leq 30$, or when $Z_5 \leq 63$ and $Z_2 > 30$) no confounders were present and the causal effect was set to zero ($b_{yx}^{[2]} = b_{yx}^{[3]} = 0$). In subgroup 4 (when $Z_5 > 63$ and $Z_2 > 30$), an unobserved confounder u was present, i.e., $x^{[4]} = b_{xu}^{[4]} u^{[4]} + e_x^{[4]}$ and

$y^{[4]} = b_{yx}^{[4]} x^{[4]} + b_{yu}^{[4]} u^{[4]} + e_y^{[4]}$, the causal effect was fixed at $b_{yx}^{[4]} = -0.75$, and the confounding effects ($b_{xu}^{[4]}$ and $b_{yu}^{[4]}$) where 0, 0.5, or 1. The number of partitioning variables (Z_1, …, Z_t) was either $t = 5$ or 15. Following Dusseldorp and Van Mecherlen (2014) and Fokkema et al. (2018), all partitioning variables were randomly drawn from a multivariate normal distribution with means for Z_1, Z_2, Z_4, and Z_5 of 10, 30, –40, and 70. The means of the remaining covariates (Z_3 and depending on the value of t, Z_6, … Z_{15}) were randomly generated from the uniform distribution on the interval [–70, 70]. The standard deviation of all covariates were set to 10, the correlations among pairwise partitioning covariates were fixed at 0.3. Because confounder detection using kernel-based independence tests requires non-normality of variables, the predictor (x), the confounder (u), and the error terms e_x and e_y were independently sampled from the gamma distribution with pre-specified skewnesses of 0, 1, and 2. In case of zero skewnesses, data were generated from the normal distribution. The simulation factors were fully crossed and 1000 samples were generated for each of the 3 (magnitude of $b_{xu}^{[4]}$) × 3 (magnitude of $b_{yu}^{[4]}$) × 2 (number of covariates t) × 3 (distribution of x) × 3 (distribution of u) × 3 (distribution of e_x) × 3 (distribution of e_y) = 1458 simulation conditions.

For each dataset, the MOB algorithm was used to identify local linear regression models of the form $y^{[k]} = b_{yx}^{[k]} x + e_y^{[k]}$ (k = 1, …, K) with at least 20 observations per terminal

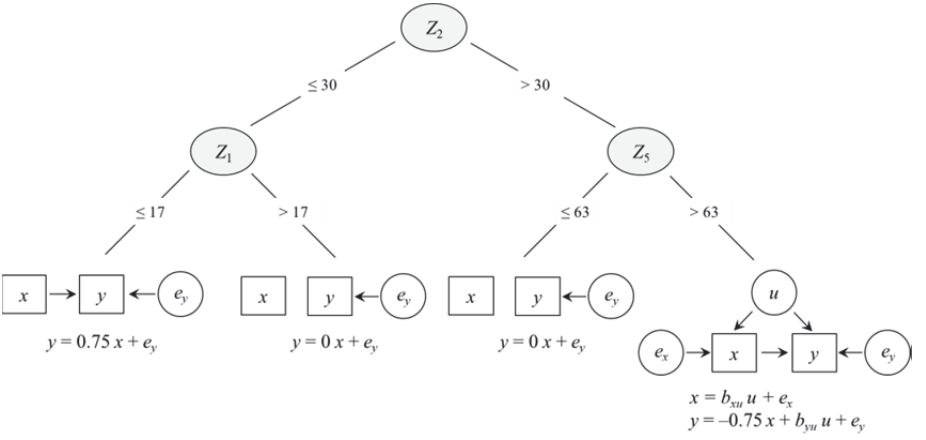

Figure 1:
Data-generating mechanism. Z_1, Z_2, and Z_5 are the partitioning variables, x is the focal predictor, y is the focal outcome, and u is an unobserved confounder

node. Linear models were estimated as generalized linear models[6] (GLM; McCullagh & Nelder, 1989; for a discussion of MOB in the context of the GLM see Rusch and Zeileis, 2013) using the identity link and a Gaussian error. Bonferroni-corrected parameter stability tests were performed using a nominal significance level 0.05. Because regression tree methods can be prone to overfitting in case of large samples, post-pruning based on the Bayes Information Criterion (BIC) was used for each estimated regression tree. Next, for each of the *K* subgroups, we computed OLS residuals and used the Bonferroni-corrected HSIC test (with median heuristic-based bandwidth parameters) to evaluate the independence assumption. To quantify the Type I error robustness of the HSIC test, we used Bradley's (1978) liberal robustness criterion, i.e., a test is considered robust if the empirical Type I error rates fall within the interval 2.5% – 7.5%. Regression trees were estimated using the `glmtree` function of the `partykit` package (Hothorn & Zeileis, 2015). Independence testing was performed using the `dhsic.test` function of the R package `dHSIC` (Pfister & Peters, 2017).

Results

In the following section, we summarize the results of the simulation study. First, we focus on the performance of the MOB algorithm to detect the correct tree structure in the presence of confounding. Specifically, we focus on tree size and overall tree accuracy. Following Fokkema et al. (2018), we defined an accurately recovered tree as 1) having seven nodes in total (i.e., three splitting and four terminal nodes), 2) the first split involving variable Z_2 with a value of 30 $\pm$ 5, 3) the next split on the left involving the variable Z_1 with a value of 17 $\pm$ 5, and 4) the next split on the right involving the variable Z_5 with a value of 63 $\pm$ 5 (here, $\pm$ 5 corresponds to $\pm$ half the population standard deviations of the partitioning covariates). Second, we focus on the Type I error and power rates of the HSIC test. Finally, we focus on the performance of HSIC tests to identify the confounded subgroup.

Tree size and accuracy

Figure 2 gives the average tree size as a function of regressor skewness (γ_x), number of covariates, and magnitude of confounding effects (b_{yu} and b_{xu}). The distributional shape of the confounder (*u*) and the two error terms (e_x and e_y) did not have an effect on the average tree size. Therefore, averages in Figure 2 were computed across all skewness levels of *u*, e_x , and e_y . In the absence of confounding, we observed an average tree size of 6.68 (*SD* = 0.77). The number of covariates did not affect the average tree size. However, tree size systematically decreased with the magnitude of the confounding effects.

[6] The reason for this is that BIC-based post-pruning strategies assume that the objective function used corresponds to the negative log-likelihood.

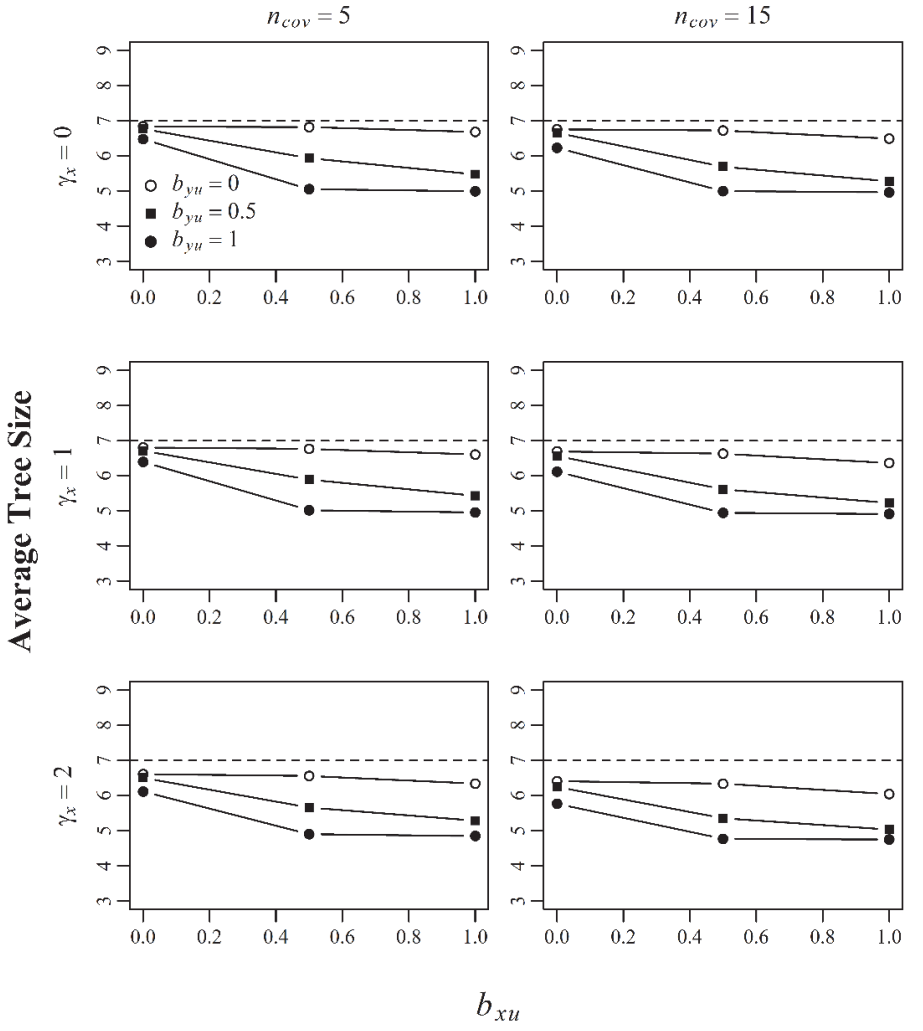

Figure 2:
Average tree size as a function of regressor skewness (γ_x), number of covariates, and magnitude of confounding effects (b_{yu} and b_{xu})

For example, for $b_{yu} = b_{xu} = 0.5$, we obtained an average tree size of 5.69 (*SD* = 1.13) nodes, for $b_{yu} = b_{xu} = 1$, the average tree size further declined to 4.90 (*SD* = 0.69) nodes. Similarly, the average tree size decreased with the skewness of *x*, however, to a far lesser extent. Here, the average tree size decreased from 6.04 (*SD* = 1.08) nodes (for $\gamma_x = 0$) to 5.98 (*SD* = 1.11; for $\gamma_x = 1$) and 5.75 (*SD* = 1.21; for $\gamma_x = 2$).

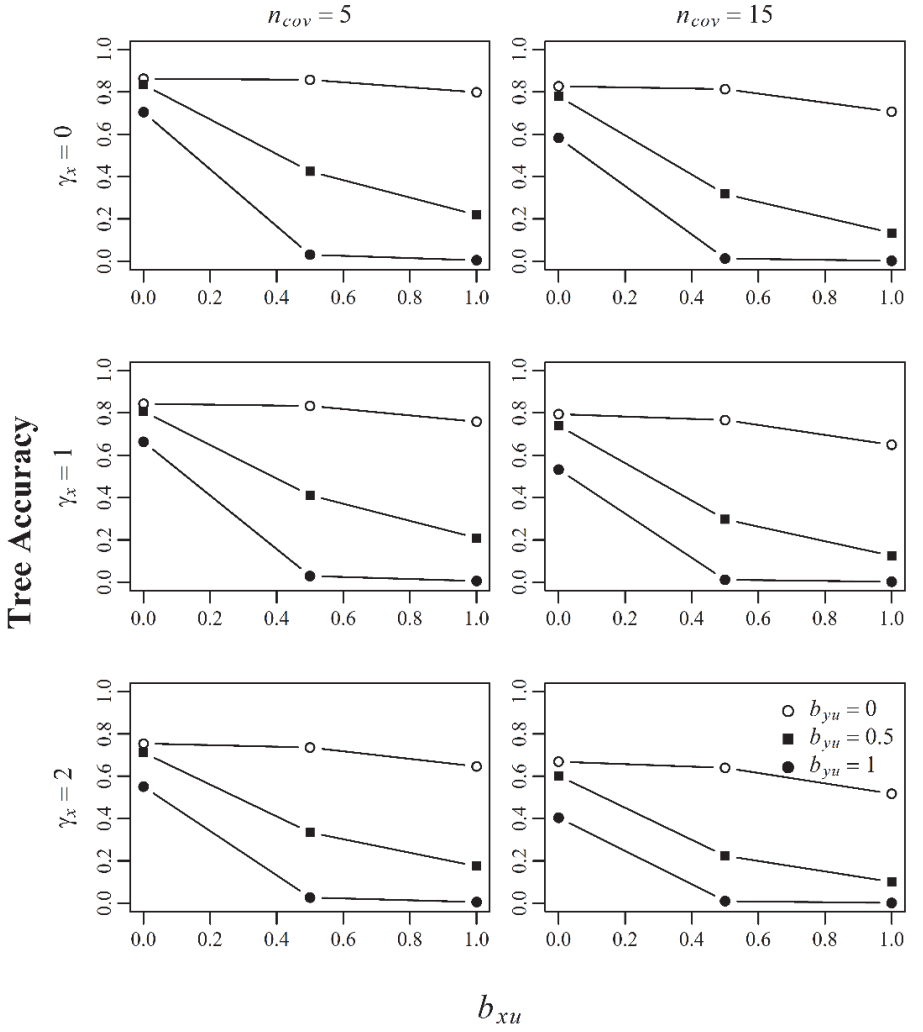

Figure 3:
Tree accuracy as a function of regressor skewness (γ_x) , number of covariates, and magnitude of confounding effects (b_{yu} and b_{xu}).

Figure 3 summarizes the tree accuracy as a function of predictor skewness, number of covariates, and magnitudes of confounding effects. When no confounders were present and all variables followed a normal distribution, the accuracy to recover the entire regression tree was 89.3%. Again, the distributional shapes of u, e_x , and e_y had virtually no effect on tree accuracy. However, tree accuracy slightly decreased with the number of

covariates and the skewness of x. As expected, the presence of confounding heavily influenced the probability to accurately recover the regression tree. For example, the probability to recover the "true" regression tree was close to zero when $b_{yu} = 1$ and $b_{yu} \geq 0.5$. Thus, we can conclude that unobserved confounders are not only likely to bias causal effect estimates, they also heavily distort the structure of estimated regression trees.

Type I error and statistical power of the HSIC test

Type I error scenarios (i.e., rejecting the true null hypothesis of independence of predictors and residuals) only exist when all variables are normally distributed. The reason for this is that uncorrelatedness implies independence only in the normal case (Hyvärinen et al., 2001). Table 1 provides an overview of the Type I error rates of Bonferroni-adjusted and unadjusted HSIC tests under normality of variables and errors. Type I error rates of unadjusted HSIC tests are inflated irrespective of the number of covariates and the magnitude of the confounding effects. In other words, multiple independence testing across the K subgroups increases the risk of false positive results. In contrast, adjusted HSIC tests are better suited to protect the nominal significance level. While application of a simple Bonferroni adjustment renders the HSIC test conservative in statistical decisions (i.e., Type I error rates tend to be close to the lower robustness bound), it is important to note that conservative Type I error rates do not invalidate significance tests per se. Instead, conservative Type I error rates simply imply that, compared to an ideal significance test with Type I error rates close to the nominal significance level, power rates of the Bonferroni-adjusted test can be expected to be lower. However, overall, the results in Table 1 confirm the importance of adjustment for multiple independence testing.

Table 1:
Type I error rates of unadjusted and α-adjusted HSIC tests

		5 Covariates		15 Covariates	
b_{xu}	b_{yu}	HSIC	adj. HSIC	HSIC	adj. HSIC
0.0	0.0	**0.099**	**0.023**	**0.109**	0.028
0.5	0.0	**0.088**	0.027	**0.093**	**0.024**
1.0	0.0	**0.100**	0.027	**0.107**	0.032
0.0	0.5	**0.097**	**0.024**	**0.096**	0.025
0.5	0.5	**0.091**	**0.024**	**0.078**	**0.014**
1.0	0.5	**0.089**	0.029	**0.117**	0.041
0.0	1.0	**0.118**	0.034	**0.101**	0.029
0.5	1.0	**0.080**	0.026	**0.101**	0.031
1.0	1.0	**0.084**	0.028	**0.085**	0.028

Note: Type I error rates outside Bradley's (1976) robustness interval .025 - .075 are bold.

Next, we focus on the statistical power of the Bonferroni-adjusted HSIC test to detect predictor-residual non-independence due to confounding. Because of non-robust Type I error rates under independence, we do not focus on the power of unadjusted HSIC tests. Figures 4 and 5 detail the statistical power of the Bonferroni-adjusted HSIC test to detect non-independence in at least one of the K subgroups. Because statistical power was not affected by the distributional shape of x and the number of covariates, we present power curves as a function of error skewnesses (γ_{e_x} and γ_{e_y}), magnitudes of confounding

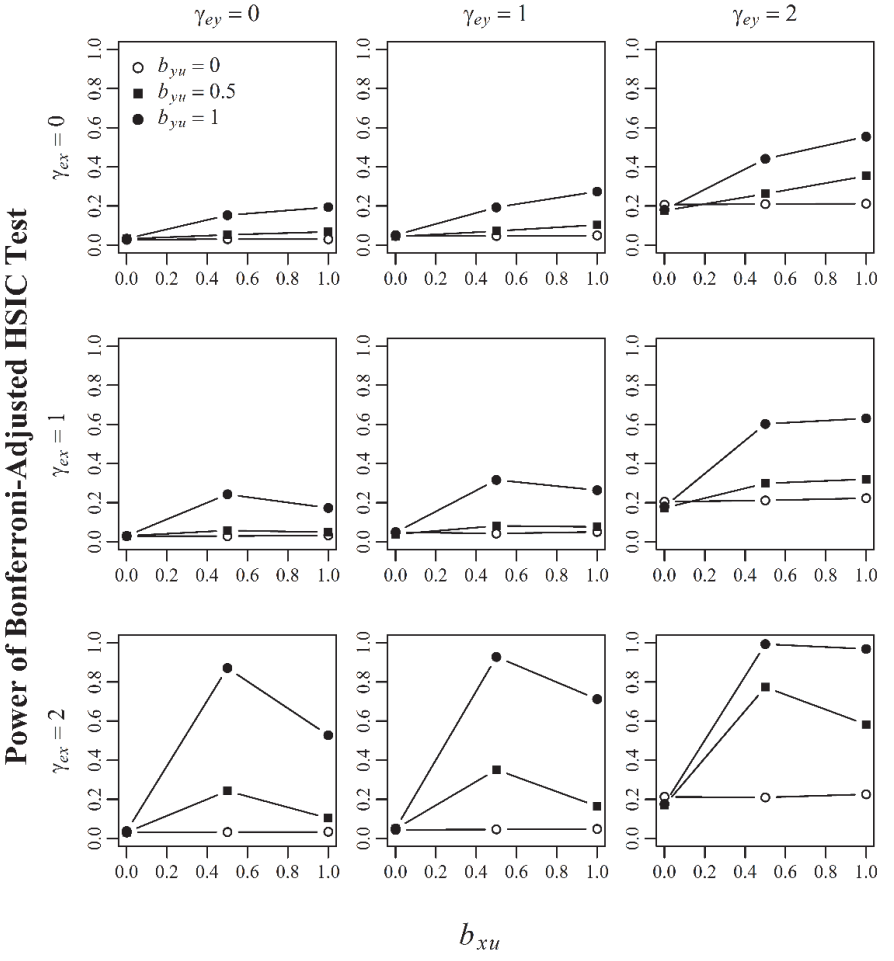

Figure 4:
Statistical power of α-adjusted HSIC tests as a function of error skewnesses (γ_{e_x} and γ_{e_y}) and magnitudes of confounding effects (b_{yu} and b_{xu}) for $\gamma_u = 1$

effects (b_{xu} and b_{yu}), and the distributional shape of the confounder (Figures 4 and 5 give the power curves for γ_u = 1 and 2, respectively). Overall, statistical power increases with error skewnesses and magnitudes of confounding effects which can be expected from the theoretical results presented above. Power curves followed an inverse U-shaped pattern when e_x was highly skewed and $\gamma_{e_y} \leq 1$. The skewness of *u* affects the power of the HSIC test only for cases where error distributions are symmetric or moderately skewed.

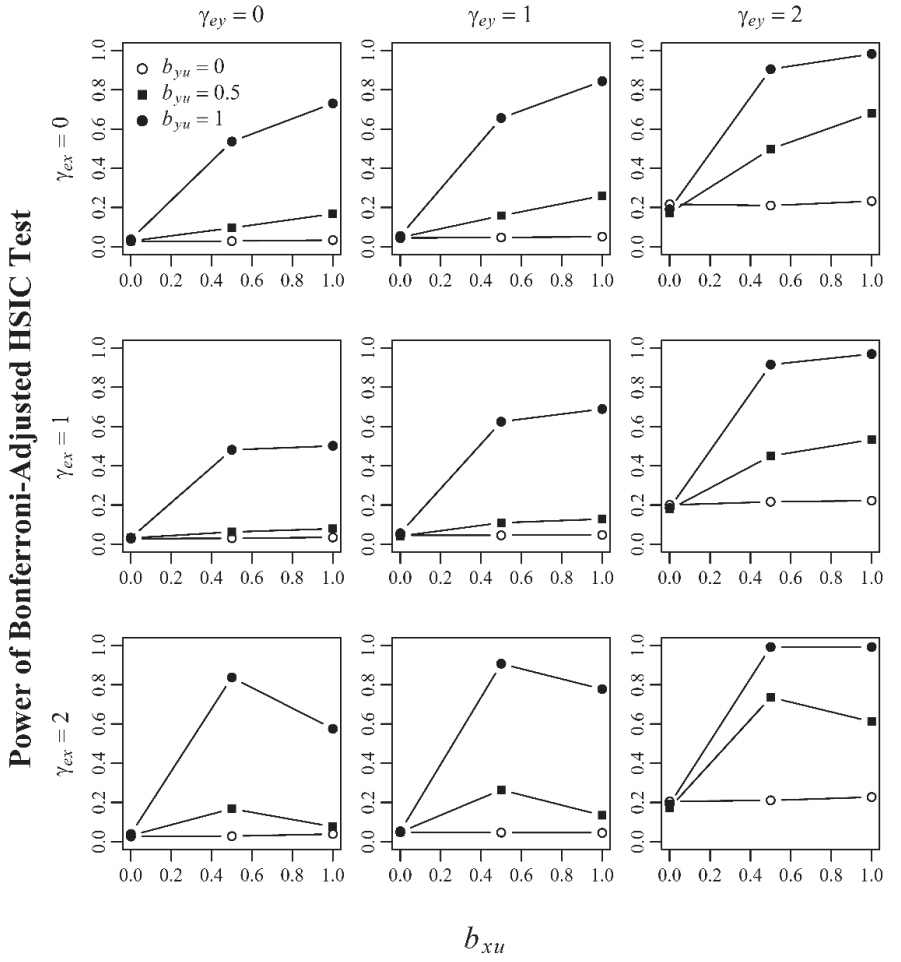

Figure 5:
Statistical power of α-adjusted HSIC tests as a function of error skewnesses (γ_{e_x} and γ_{e_y}) and magnitudes of confounding effects (b_{yu} and b_{xu}) for γ_u = 2

Detection of confounded subgroups

Finally, we focus on the accuracy of detecting confounded subgroups based on the results of Bonferroni-adjusted HSIC tests. Figure 6 gives the percentages of subjects which were correctly identified as members of the confounded subgroup. Across all simulation conditions, the true number of confounded subjects ranged from 154 to 493 (M = 312.5, SD = 105.1). The distributional shape of the confounder (γ_u) and the predictor (γ_x) did

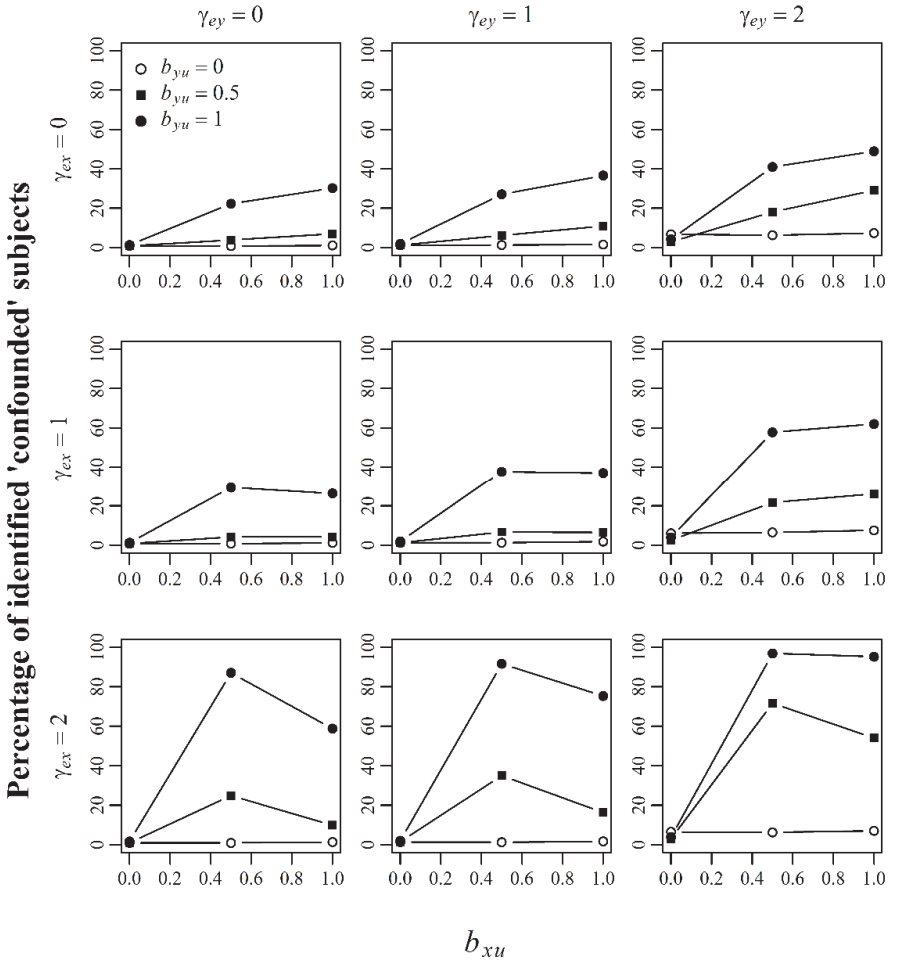

Figure 6:
Percentage of subjects correctly identified as being members of the confounded subgroup as a function of error skewnesses (γ_{e_x} and γ_{e_y}) and magnitudes of confounding effects (b_{yu} and b_{xu})

not affect the estimated percentage of confounded subjects. The patterns presented in Figure 6 are quite similar to the power curves presented in Figures 4 and 5. That is, the percentage of identified confounded subjects increases with error skewnesses and magnitudes of confounding effects. However, a nonlinear pattern is observed when $\gamma_{e_y} \leq 1$ and $\gamma_{e_x} \geq 1$. The presented approach identified 100% of the confounded subjects when errors were highly skewed and confounding effects were large, as expected.

Discussion

The present study combined asymmetry principles of the linear regression model which emerge under non-normality of variables with a model-based recursive partitioning algorithm to detect confounded subgroups in linear models. While the presence of confounding biases the structure of the data generating regression tree, kernel-based tests of independence can still be used to identify members of a confounded subpopulation.

The proposed approach rests on a number of assumptions that need to be evaluated critically to ensure valid results. First, the present approach builds on asymmetry properties of the linear model that emerge under non-normality of variables. In other words, in the present context, non-normality is the crucial element to detect confounded subgroups. This also implies that, in practical applications, non-normality needs to be an inherent distributional feature of the variable under study and not an artificial by-product of outlying observations or ceiling/floor effects. Data visualizations (such as histograms or QQ-plots) and normality tests are readily available to evaluate the distributional requirement of non-normality.

Second, it is important to reiterate that MOB-based local regression models rest on the same model assumptions as a standard (global) linear model. This implies that biases caused by influential observations, heteroscedasticity, nonlinearity, or structural misspecifications (e.g., simultaneity and reverse causation biases) also jeopardize local regression models. Thus, regression diagnostics are a necessary adjunct to the use of recursively partitioned linear models. Further, it is important to realize that some biasing factors also affect the performance of the HSIC test. For example, the presence of simultaneity (Wooldridge, 2010) and reverse causation biases (Wiedermann & von Eye, 2015b) can introduce dependencies between predictors and errors which can also be detected by the HSIC test. Therefore, a significant HSIC test may point at the presence of reciprocal causation (which can be considered a special case of confounding) or directional model-misspecifications where the "true" predictor is erroneously used as the outcome (the latter bias can be detected using so-called Direction Dependence Analysis; Wiedermann & von Eye, 2015a; Wiedermann & von Eye, 2015b; Wiedermann & Sebastian, 2019). Further, predictor-residual dependencies may also emerge from unconsidered nonlinear effects (for a discussion of the HSIC as a global goodness-of-fit test see Sen and Sen, 2014). To distinguish such a functional model-misspecification from the presence of confounding, one can start with recursively partitioning a regression model that includes higher order polynomials followed by an examination of local independence using local model residuals and predicted scores (instead of raw values) of *x*.

Simulation evidence suggests that multiple independence testing across subgroups leads to inflated Type I error rates. The present study focused on a simple Bonferroni correction to adjust the family-wise error rate (i.e., the probability of erroneously rejecting independence in one or more subgroups). The Bonferroni adjustment is able to protect the family-wise error rate regardless of the correlation among the HSIC tests. However, it tends to be conservative because only hypotheses with p-values $\leq \alpha / K$ are rejected and may, under certain conditions, lack sufficient statistical power. As an alternative, more powerful sequential adjustment procedures have been proposed by Holm (1979), Simes (1986), Hochberg (1988), Hommel (1988), and Benjamini and Hochberg (1995). For example, Holm's (1979) procedure is based on the ordering of the p-values (from smallest to largest) for the K null hypotheses H_i (i = 1, …, K), i.e., $p_1 \leq p_2 \leq p_i \leq \ldots \leq p_K$, and the corresponding adjustment terms are computed sequentially using $(K - i + 1)$. Since Holm's method is known to be more powerful than the conventional Bonferroni correction, we can also expect that the power of multiple HSIC tests to detect confounded subgroups increases when using Holm's adjustment. Similarly, one can further improve the power of the HSIC test through replacing the gamma-approximated HSIC test by a resampling-based procedure proposed by Sen and Sen (2014). These performance optimizations constitute material for future research.

Acknowledgement

We thank Edgar C. Merkle and the guest editors Mark Stemmler and Alexander von Eye for their constructive comments on an earlier version of the article. Further, we are indebted to Marjolein Fokkema for sharing the R code accompanying Fokkema et al. (2018).

References

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360. https://doi.org/10.1073/pnas.1510489113

Barnow, B. S., Cain, G. G., & Goldberger, A. S. (1980). Issues in the analysis of selectivity bias. In *Evaluation Studies* (Vol. 5). San Francisco: Sage.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak. *Journal of the American Statistical Association*, *90*(430), 443–450. https://doi.org/10.2307/2291055

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144–152. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. L. (1984). *Classification and Regression Trees*. California: Wadsworth.

Darmois, G. (1953). Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle linéaire [General analysis of stochastic links]. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, *21*(1/2), 2. https://doi.org/10.2307/1401511

Doove, L. L., Van Deun, K., Dusseldorp, E., & Van Mechelen, I. (2016). QUINT: A tool to detect qualitative treatment–subgroup interactions in randomized controlled trials. *Psychotherapy Research*, *26*(5), 612–622. https://doi.org/10.1080/10503307.2015.1062934

Dusseldorp, E., & Van Mechelen, I. (2014). Qualitative interaction trees: a tool to identify qualitative treatment-subgroup interactions. *Statistics in Medicine*, *33*(2), 219–237. https://doi.org/10.1002/sim.5933

Elwert, F., & Winship, C. (2014). Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. *Annual Review of Sociology*, *40*(1), 31–53. https://doi.org/10.1146/annurev-soc-071913-043455

Entner, D., Hoyer, P. O., & Spirtes, P. (2012). Statistical test for consistent estimation of causal effects in linear non-Gaussian models. *The Journal of Machine Learning Research*, *22*, 364–372.

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., & Kelderman, H. (2018). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behavior Research Methods*, *50*(5), 2016–2034. https://doi.org/10.3758/s13428-017-0971-x

Greenland, S., & Robins, J. M. (1986). Identifiability, Exchangeability, and Epidemiological Confounding. *International Journal of Epidemiology*, *15*(3), 413–419. https://doi.org/10.1093/ije/15.3.413

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., & Smola, A. J. (2008). A kernel statistical test of independence. *Advances in Neural Information Processing Systems*, *20*, 585–592.

Hausman, J. A. (1978). Specification Tests in Econometrics. *Econometrica*, *46*(6), 1251–1271. https://doi.org/10.2307/1913827

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, *75*(4), 800–802. https://doi.org/10.1093/biomet/75.4.800

Holm, S. (1979). A simple sequential rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, *75*(2), 383–386. https://doi.org/10.1093/biomet/75.2.383

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674. https://doi.org/10.1198/106186006X133933

Hothorn, T., & Zeileis, A. (2015). partykit: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research*, *16*, 3905–3909.

Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis.* New York, NY: Wiley & Sons.

Imbens, G. W. (2004). Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *Review of Economics and Statistics*, *86*(1), 4–29. https://doi.org/10.1162/003465304323023651

Imbens, G. W., & Angrist, J. D. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica*, *62*(2), 467–475. https://doi.org/10.2307/2951620

Kraemer, H. C., Kiernan, M., Essex, M., & Kupfer, D. J. (2008). How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. *Health Psychology*, *27*(2S), S101–S108. https://doi.org/10.1037/0278-6133.27.2(Suppl.).S101

Lagakos, S. W. (2006). The Challenge of Subgroup Analyses – Reporting without Distorting. *New England Journal of Medicine*, *354*(16), 1667–1669. https://doi.org/10.1056/NEJMp068070

Li, R. H., & Belford, G. G. (2002). *Instability of decision tree classification algorithms*. Presented at the Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London, UK: Chapman & Hall.

Morgan, J. N., & Sonquist, J. A. (1963). Problems in the Analysis of Survey Data, and a Proposal. *Journal of the American Statistical Association*, *58*(302), 415–434. https://doi.org/10.1080/01621459.1963.10500855

Pearl, J. (2009). *Causality: models, reasoning, and inference (2nd ed.)*. Cambridge: Cambridge University Press.

Pfister, N., & Peters, J. (2017). dHSIC: Independence testing via Hilbert Schmidt independence criterion. *R Package Version 2.0. Https://CRAN.R-Project.Org/Package= dHSIC*.

Philipp, M., Zeileis, A., & Strobl, C. (2016). A toolkit for stability assessment of tree-based learners. *Proceedings of Compstat 2016 – 22nd International Conference on Computational Statistics*, 315–325. The International Statistical Institute/International Association for Statistical Computing.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Mateo: Morgan Kaufmann.

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. Retrieved from http://www.R-project.org/

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rubin, D. B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics*, *6*(1), 34–58. https://doi.org/10.1214/aos/1176344064

Rusch, T., & Zeileis, A. (2013). Gaining insight with recursive partitioning of generalized linear models. *Journal of Statistical Computation and Simulation*, *83*(7), 1301–1315. https://doi.org/10.1080/00949655.2012.658804

Sen, A., & Sen, B. (2014). Testing independence and goodness-of-fit in linear models. *Biometrika*, *101*(4), 927–942. https://doi.org/10.1093/biomet/asu026

Shimizu, S. et al. (2011). DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, *12*, 1225–1248.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, *73*(3), 751–754. https://doi.org/10.1093/biomet/73.3.751

Skitovich, W. P. (1953). On a property of the normal distribution. *Doklady Akademii Nauk SSSR [Reports of the Academy of Sciences USSR]*, *89*, 217–219.

Sriperumbudur, B., Fukumizu, K., Gretton, A., Lanckriet, G. R. G., & Schölkopf, B. (2009). Kernel choice and classifiability for RKHS embeddings of probability distributions. In *Advances in neural information processing systems* (Vol. 22, pp. 1750–1758). La Jolla, CA: Neural Information Processing Systems Foundation.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), 323–348. https://doi.org/10.1037/a0016973

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, *51*(6), 309–317. https://doi.org/10.1037/h0044319

VanderWeele, T. J., & Shpitser, I. (2013). On the definition of a confounder. *The Annals of Statistics*, *41*(1), 196–220. https://doi.org/10.1214/12-AOS1058

Wang, R., & Ware, J. H. (2013). Detecting Moderator Effects Using Subgroup Analyses. *Prevention Science*, *14*(2), 111–120. https://doi.org/10.1007/s11121-011-0221-x

Wiedermann, W., Artner, R., & von Eye, A. (2017). Heteroscedasticity as a Basis of Direction Dependence in Reversible Linear Regression Models. *Multivariate Behavioral Research*, *52*(2), 222–241. https://doi.org/10.1080/00273171.2016.1275498

Wiedermann, W., & Li, X. (2018). Direction dependence analysis: A framework to test the direction of effects in linear models with an implementation in SPSS. *Behavior Research Methods*, *50*(4), 1581–1601. https://doi.org/10.3758/s13428-018-1031-x

Wiedermann, W., & Li, X. (2019). Confounder detection in linear mediation models: Performance of kernel-based tests of independence. *Behavior Research Methods*. https://doi.org/10.3758/s13428-019-01230-4

Wiedermann, W., & Sebastian, J. (2019). Direction dependence analysis in the presence of confounders: Applications to linear mediation models. *Multivariate Behavioral Research*, *(in press)*.

Wiedermann, W., & von Eye, A. (2015a). Direction of effects in multiple linear regression models. *Multivariate Behavioral Research*, *50*(1), 23–40. https://doi.org/10.1080/00273171.2014.958429

Wiedermann, W., & von Eye, A. (2015b). Direction-dependence analysis: A confirmatory approach for testing directional theories. *International Journal of Behavioral Development*, *39*(6), 570–580. https://doi.org/10.1177/0165025415582056

Wiedermann, W., & von Eye, A. (2016). *Statistics and Causality: Methods for Applied Empirical Research*. Hoboken, NJ: Wiley and Sons.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data.* (2nd ed.). Cambridge, MA: MIT Press.

Zeileis, A., & Hornik, K. (2007). Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica*, *61*(4), 488–508. https://doi.org/10.1111/j.1467-9574.2007.00371.x

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-Based Recursive Partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492–514. https://doi.org/10.1198/106186008X319331

Zhang, H., & Singer, B. (2010). *Recursive partitioning and applications* (2nd ed.). New York, NY: Springer.