Psychological Test and Assessment Modeling, Volume 61, 2019 (4), 363-364

Editorial

Tree-based methods for regression and classification – Statistical methods at the interface of graphics and statistics

Mark Stemmler[1] *& Alexander von Eye*[2]

The editors of this Special Issue invited scholars to contribute new formal and statistical developments as well as interesting applications in the field of tree-based methods in psychological research. Among the many interesting properties of tree-based methods is that they enable researchers to investigate effects of multiple independent variables on one or several dependent variables without any restriction on level of measurement.

The contributions included in this issue contain theoretical considerations, new methods for tree-based analysis, simulation studies, and sample analyses of interesting psychological data sets. Also included are software developments and applications (R packages such as confreq, rpart, dHISC and partykit). The reader will find articles that encompass classification and regression tree methods (CART), CHAID, and bootstrap regression tree methods. This special issue first presents a selection of new developments and theoretical contributions, and, second, a selection of applications of tree-based methods.

Section I: Theory and Methods

In the first manuscript of the theory and methods section, Michael P. van Wie, Xintong Li, and Wolfgang Wiedermann propose and discuss methods for the *Identification of confounded subgroups using linear model-based recursive partitioning*. The authors aim at detecting confounded subgroups in linear regression models by way of combining a confounder detection approach, based on kernel-based independence testing, with model-based recursive partitioning.

[1] *Correspondence concerning this article should be addressed to:* Prof. Dr. Mark Stemmler, Institut für Psychologie, Universität Erlangen-Nürnberg, Nägelsbachstraße 49c, 91052 Erlangen, Germany; email: mark.stemmler@fau.de

[2] Michigan State University, USA

The second article, contributed by Markus Fritsch, Harry Haupt, Friedrich Lösel, and Mark Stemmler, discusses data science alternatives to ordinary least squares: Investigating the risk factors for corporal punishment with decision trees and random forests. The authors demonstrate that decision trees and random forests enable data-driven modeling. These methods are valuable tools in psychological research to gain new insights and to validate existing results. In an application, Fritsch and colleagues examine the behavior of father's aggressiveness, dysfunctional parent-child interactions, and other risk factors for corporal punishment of children by their fathers.

This article is followed by *Analyzing Tree Structures with Configural Frequency Analysis and the R-package confreq* by Mark Stemmler, Jörg Henrik-Heine and Susanne Wallner. This article describes the use of Configural Frequency Analysis (CFA) for detecting a tree structure in data that were analyzed with the SPSS module Answer Tree. In addition, the application of the R package *confreq* is demonstrated. The data example is taken from a longitudinal study on deviant and delinquent behavior in juveniles.

In the last article of the theory and methods section, which is entitled *Log-linear and Configural Analysis of Tree Structures*, Alexander von Eye, Wolfgang Wiedermann and Stefan von Weber propose two methods for the analysis of existing tree structures. These methods are log-linear modeling and Configural Frequency Analysis (CFA). In a data example, students' decisions and life satisfaction are examined.

Section II: Applications

The first article of the application section is contributed by Cody Ding and Yuyang Zhao. The authors present on *Using tree-based regression to examine factors related to math ability among 15-year old students* to predict students' mathematics ability, particularly subgroups of students who share characteristics that are associated with different levels of math ability. Based on PISA 2012 data from the United States and China, the authors used regression tree analysis to select the most salient predictors of math ability, and identify subgroups of 15-year-old students who were likely to be proficient in math ability.

The second article of the application section, co-authored by Eun-Young Mun and Feng Geng is entitled *Predicting post-experiment fatigue among healthy young adults: Random forest regression analysis*. In this paper, a random forest regression analysis is used to predict post-experiment fatigue in a sample of 212 participants between the ages of 18 and 30 following a mildly stressful experiment. The random forest regression analysis is used together with a built-in cross-validation function.

We hope that the readers will enjoy this Special Issue.

Mark Stemmler & Alexander von Eye

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