IRTG 1792 Discussion Paper 2019-028

Group Average Treatment Effects for Observational Studies

Daniel Jacob [*]
Stefan Lessmann [*]
Wolfgang Karl Härdle [* *2 *3]

[*] Humboldt-Universität zu Berlin, Germany
[*2] Singapore Management University, Singapore
[*3] Xiamen University, China

This research was supported by the Deutsche Forschungsgesellschaft through the International Research Training Group 1792 "High Dimensional Nonstationary Time Series".

http://irtg1792.hu-berlin.de
ISSN 2568-5619

Group Average Treatment Effects for Observational Studies*

Daniel Jacob† Stefan Lessmann† Wolfgang Karl Härdle†‡§

This Draft: November 6, 2019

Abstract

The paper proposes an estimator to make inference on key features of heterogeneous treatment effects sorted by impact groups (GATES) for non-randomised experiments. Observational studies are standard in policy evaluation from labour markets, educational surveys and other empirical studies. To control for a potential selection-bias we implement a doubly-robust estimator in the first stage. Keeping the flexibility to use any machine learning method to learn the conditional mean functions as well as the propensity score we also use machine learning methods to learn a function for the conditional average treatment effect. The group average treatment effect is then estimated via a parametric linear model to provide p-values and confidence intervals. The result is a best linear predictor for effect heterogeneity based on impact groups. Cross-splitting and averaging for each observation is a further extension to avoid biases introduced through sample splitting. The advantage of the proposed method is a robust estimation of heterogeneous group treatment effects under mild assumptions, which is comparable with other models and thus keeps its flexibility in the choice of machine learning methods. At the same time, its ability to deliver interpretable results is ensured.

JEL classification: C01, C14, C31, C63

Keywords: *causal inference, machine learning, simulation study, confidence intervals, multiple splitting, sorted group ATE (GATES), doubly-robust estimator*

*Financial support from the Deutsche Forschungsgemeinschaft via the IRTG 1792 "High Dimensional Non Stationary Time Series", Humboldt-Universität zu Berlin, is gratefully acknowledged. All correspondence may be addressed to the author by e-mail at daniel.jacob@hu-berlin.de.

†School of Business and Economics, Humboldt-Universität zu Berlin, Spandauer Straße 1, 10178 Berlin, Germany.

‡Sim Kee Boon Institute for Financial Economics, Singapore Management University, 50 Stamford Road, 178899 Singapore, Singapore

§W.I.S.E. - Wang Yanan Institute for Studies in Economics, Xiamen University, 422 Siming S Rd, 361005 Fujian, China

1 Introduction

When evaluating a causal effect of some policy, marketing action or another treatment indicator, it might not be sufficient to only report the average treatment effect (ATE). The estimation of heterogeneous effects, e.g. the conditional (on covariates) average treatment effect (CATE), provides further insight into causal mechanisms and helps researchers and practitioners to actively adjust the treatment assignment towards an efficient allocation. The more information in terms of characteristics i.e. covariates we are provided with, the better can heterogeneity be observed. If we have little deterministic information it might be that heterogeneity effects are overlooked. The trade-off here is that the more covariates datasets have, the more complex they get. This is why parametric models are often insufficient when applied on high-dimensional, non-linear datasets (Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey & Robins, 2018). Therefore, recent methods for treatment effect estimation use machine learning models that have shown to be superior in high-dimensional prediction problems (Hastie, Tibshirani & Friedman, 2009). The idea is to learn nuisance functions and regularize the parameter space while making as little assumptions as possible. This is especially helpful when the data does not come from randomised experiments where treatment is randomly assigned to the individuals. In observational studies, self-selection into treatment can arise which introduces a bias that has to be corrected for (i.e. self-selection bias) (Heckman, Ichimura, Smith & Todd, 1998). For the ATE one would use the nuisance parameter to orthogonalize the effect that covariates have on both, the treatment assignment and the outcome variable. See Chernozhukov et al. (2018) for a recent approach which they call double machine learning.

Recent papers that study and evaluate different models that are designed for the estimation of heterogeneous treatment effects are (Knaus, Lechner & Strittmatter, 2018; Künzel, Sekhon, Bickel & Yu, 2019; Powers, Qian, Jung, Schuler, Shah, Hastie & Tibshirani, 2018). The two most prominent methods used to estimate the CATE may be the general random forest, which builds on the idea of controlling for observed confounders through a tree structure and then estimates the CATE within each final leaf (Athey, Wager & Tibshirani, 2019). The second one is the causal boosting, which uses boosted trees to increase performance (Powers et al., 2018). The conditional average treatment effects can then be interpreted as individualised treatment effects. What the aforementioned methods lack, however, is that they are built on tree algorithms and therefore do not allow a flexible estimation of heterogeneous treatment effects in terms of the model choice. A recent method called R-learner does provide such flexibility and shows competitive performance in the estimation of the CATE to other existing proposals (Nie & Wager, 2017). Other models, known as meta-learners, decompose the modelling procedure into sub-regression functions, which can be solved using any supervised learning method. This can e.g. be done by a two-model approach (TMA) where we train a response function (conditional mean) on the treated and another one on the non-treated observations. In randomised experiments, the difference of the two functions can thus be interpreted as the individualised treatment effect (Künzel et al., 2019). Applying the two-model approach on data from non-randomized experiments would incorporate a potential bias that needs to be corrected for. One way to address the problem is to use double-robust estimator as proposed by (Robins & Rotnitzky, 1995). Using the results from the two-model approach and, in a second step, use inverse probability weighting (IPW) decreases the variance of the estimator (see e.g. Lunceford & Davidian (2004)). Additional orthogonalization using the two conditional mean functions produced by the TMA also decreases the bias of

the parameter of interest (Lee, Okui & Whang, 2017). The doubly-robust estimator can even be used in high-dimensional settings to estimate a reduced dimensional conditional average treatment effect function. Using machine learning methods to learn the nuisance functions and a kernel regression for the low-dimensional covariates of interest, functional limit theory can be derived (Fan, Hsu, Lieli & Zhang, 2019).

The difficulty, however, is that machine learning methods are often a black box that is not easy to interpret and hence hinder the information on drivers for effect heterogeneity. In this paper, we, therefore, build on the ideas of Chernozhukov, Demirer, Duflo & Fernandez-Val (2018) who concentrate to estimate a sorted group average treatment effect (GATE) in randomised experiments. The heterogeneity between these groups can then be interpreted through covariates which shed some light on the question of what characteristics determine the differences between groups. We extend the approach to estimate the GATE parameter towards the use in observational studies and also towards the possibility to estimate the CATE based on the group heterogeneity. The advantage of the proposed method is a robust estimation of heterogeneous treatment effect that is comparable with other models thus keeping its flexibility in the choice of machine learning methods and at the same time its ability to interpret the results. The latter is especially useful in all areas of empirical economics like policy or labour markets. It also has the advantage to control for potential self-selection bias. The idea of going beyond the average, but not as deep as to estimate conditional average treatment effects for many covariates, is first considered in Chernozhukov, Fernández-Val & Luo (2018). They provide standard errors and confidence bands for the estimated sorted group effects and related classification analysis and provide confidence sets for the most and least affected groups. While they only use parametric estimators, a nonparametric attempt to estimate group average treatment effects and also provide insights from the heterogeneity in terms of observed covariates comes from Zimmert & Lechner (2019). They use a two-step estimator of which the second step consists of a kernel estimator. Our contribution is to keep machine learning methods to learn the nuisance parameter in the first step but use a parametric model in the last step. We also include a second step which uses the idea of a doubly-robust estimator to make inference about group average treatment effect when a randomized control trial is not given. This paper consists of three parts. First, we state the methodology for randomized experiments and second, the extensions to deliver robust results in observational studies. Third, we simulate data that include selection bias and are high-dimensional and non-linear. We compare the results for the GATE obtained with the two-model approach and the our extended doubly-robust method. Through averaging of the results for each observation we report the mean absolute error from the true heterogeneous treatment effects for both methods.

2 Generic Machine Learning for Group ATE

2.1 Potential Outcome Assumptions

Throughout this paper, we make use of the potential outcome theorem (Rosenbaum & Rubin, 1983) and state three necessary assumptions. The first assumption is the ignorability of treatment, conditional on observed covariates (X), from the two potential outcomes. It is also known as unconfoundedness or simply conditional independence:

$$\left(Y_i^1, Y_i^0\right) \perp\!\!\!\perp D_i | X_i. \tag{1}$$

With Y^1 denoting the potential outcome under treatment and Y^0 if not being treated. D is the treatment assignment variable.

The second assumption, the Stable Unit Treatment Value Assumption (SUTVA), guarantees that the potential outcome of an individual is unaffected by changes in the treatment assignment of others. This assumption might be violated if individuals can interact with each other (peer and social effects). In randomised controlled experiments, the first two assumptions are fulfilled by design or, at least, cancel out.

The third assumption, called overlap, guarantees that for all $x \in supp(X)$, the probability of being in the treatment group (propensity score $e(x)$), is bounded away from 0 and 1:

$$\begin{aligned} 0 < \mathrm{P}(D = 1|X = x) < 1. \\ e(x) = \mathrm{P}(D = 1|X = x). \end{aligned} \tag{2}$$

We control for the common support by estimating the propensity score and balance the treatment and control group based on the distribution. We hence exclude all observations that have a propensity score lower 0.02 or higher than 0.98. The fundamental problem of causal inference is that we only observe one of the two potential outcomes at the same time. The counterfactual for a nontreated (treated) person, namely, what would have happened if this person were (not) treated, is always missing. We can represented this statement through a switching regression where the observed outcome (Y_i) depends on the two potential outcomes and the treatment assignment:

$$Y_i = Y_i^0 + D(Y_i^1 - Y_i^0). \tag{3}$$

We further assume that, for the estimation of standard errors, the following moments exist: $\mathrm{E}\left[|Y^j|^q\right] < \infty$ for $q \geq 4$ and $j = 0, 1$.

2.2 Randomized Control Trial

To provide valid estimation and inference for a causal interpretation of parameters, Chernozhukov et al. (2018) focus on features of the CATE. One of main features is the **Sorted Group Average Treatment Effect**. The idea is to find groups of observations depending on the estimated treatment effect heterogeneity. Their proposed method relies on a two-model approach in the first step. Here, two response functions are trained separately for the treated and non-treated observations. This approach can be biased if the data sample is from an observational study. Consider e.g. self-selection into a treatment. In randomized control trials, difference of the two functions provides an estimate of the treatment effect for every observation. To denote that this function might not be consistent or unbiased it is further called score-function:

$$\begin{aligned} \tau(X) &= \mathrm{E}[Y|D = 1, X] - \mathrm{E}[Y|D = 0, X], \\ \hat{S}(X) &= \hat{g}_1\left(X_i, \widehat{\alpha}_1\right) - \hat{g}_0\left(X_i, \widehat{\alpha}_0\right). \end{aligned} \tag{4}$$

Here $\hat{g}_D(X_i, \widehat{\alpha}_D) = \mathrm{E}(Y|D, X)$ is the regression model of the outcome variable on X separately for $D \in \{0, 1\}$ and $\hat{\alpha}_D$ represents the parameters for treatment and control group. These two functions can be estimates with a broad range of supervised machine learning methods. The target parameters are

$$\mathrm{E}[\tau(X)|G_k] \qquad G_k : k^{th} \quad \text{n-tile of estimated} \quad \hat{S}(X), \tag{5}$$

where G is an indicator of group membership. The groups are ex-post defined by the predicted scores ($\hat{S}(X)$) in the first stage. If the treatment effect for the groups are consistent, it asymptotically holds that

$$\mathrm{E}[\tau(X)|G_1] \leqslant \mathrm{E}[\tau(X)|G_2] \leqslant ... \leqslant \mathrm{E}[\tau(X)|G_k], \tag{6}$$

which is the monotonicity restriction. Furthermore, it can be tested whether there is a homogeneous effect if $\mathrm{E}[\tau(X)|G_k]$ would be equal for all k groups. The weighted linear projection equation to recover the GATES parameter is:

$$YH = \hat{\beta}^{\intercal} A_1 H + \hat{\gamma} \times (D - \hat{e}(X)) \times \mathbf{I}(\hat{S}(X) \in I_k) + \nu, \tag{7}$$

with $A_1 = (1, B(X))$ and $B(X) = \mathrm{E}[Y|D = 0, X]$ being the baseline function without treatment. $\hat{S}_i(X) = \mathrm{E}[Y|D = 1, X] - \mathrm{E}[Y|D = 0, X]$ is the treatment effect projection. $I_k = [\ell_{k-1}, \ell_k)$ and ℓ_k is the k/K-quantile of $\{\hat{S}_i\}_{i \in M}$. Subscript M denotes that these are all out-of-sample predictions. This becomes clearer in the pseudo-code of Algorithm 1, which describes the implementation of this method. The weights H represent the Horvitz-Thompson transformation (Horvitz & Thompson, 1952):

$$H = H(D, Z) = \frac{D - \hat{e}(X)}{\hat{e}(X)(1 - \hat{e}(X))}. \tag{8}$$

This estimator, which is applied to account for different proportions of observations within strata in a target population, is equivalent to the simple inverse probability weighting estimator. These estimators, however, might exhibit a high variance if the identification (the precision) of the propensity scores is lacking Lunceford & Davidian (2004).

The main identification result is that the projection coefficients γ_k can be represented in the following way:

$$\gamma = (\gamma)_{k=1}^{K} = (\mathrm{E}[\tau(X)|G_k])_{k=1}^{K}. \tag{9}$$

Algorithm 1: GATES

for *b=1 to B* **do**
  Split Data in $k = 2$ samples: I^a and M with $I^a \cup M$
  Train $Y_i^0 = g_0(X_i, D = 0) + U_{0i}$, with $i \in I^a$
  Train $Y_i^1 = g_1(X_i, D = 1) + U_{1i}$, with $i \in I^a$
    Predict $\hat{Y}_i^0 = \hat{g}_0(X_i)$, with $i \in M$
    Predict $\hat{Y}_i^1 = \hat{g}_1(X_i)$, with $i \in M$
      Calculate $S_b(X|i) = \hat{Y}_i^1 - \hat{Y}_i^0$
  Train $D_i = e_0(X_i) + V$, with $i \in I^a$
    Predict $\hat{D}_i = \hat{e}(X_i)$, with $i \in M$
      Calculate $\hat{V}_i = D_i - \hat{e}(X_i)$, with $i \in M$
  Estimate GATES parameters (γ) with weighted OLS using M (see equation 7)
end
Average γ over B iterations: $\tilde{\gamma} = median\{\gamma\}$

2.3 Observational Studies

To use the best linear predictor for group heterogeneity in observational studies, we need to change and extend the first and second stage. First, we replace the two-model approach by a doubly-robust estimator. This means we not only weight by the inverse of the propensity score but also orthogonalize the outcome variable by subtracting the mean. We also use the sample splitting as a form of cross-fitting by using the auxiliary sample to estimate the score function via the doubly- robust estimator and then use the main sample to predict the final score function, which is used in the parametric step. In this way, we limit the danger of over-fitting. The parametric second stage simplifies by plugging in the robust score function without the use of inverse probability weighting. The resulting function is a more robust version of the CATE for each individual as well as for the GATE function. The two steps are described in more detail in the following.

The separate estimation of the outcome conditioning on the treatment assignment only works for randomised experiments. Assume that in observational studies individual's self-select themselves into the treatment. If this is the case, then the distribution of the covariates is different given treatment status. As a consequence, the estimated score-function, which is the difference between the estimated outcomes from $\hat{g}_1\left(X_i, \widehat{\alpha}_1\right) - \hat{g}_0\left(X_i, \widehat{\alpha}_0\right)$ might not reflect the treatment effect rather than observed differences based on the covariates. To account for a selection bias, we replace the simple two-model approach by a doubly-robust estimator, which accounts for this potential bias via an extension of inverse probability weighting and orthogonalization of the outcome variable Y_i via the conditional expectation functions $\hat{g}_D\left(X_i, \widehat{\alpha}_D\right)$ for $D \in \{0, 1\}$.

We can also think of this estimator as a transformed outcome estimator, which is why we denote the outcome as Y_i^*. This new outcome is calculated on the training data (the I^a sample). In a second step, a new supervised model is trained on the transformed outcome using I^a while predictions are made on the test set M. Algorithm 2 describes this process.

$$\hat{Y}_{i,DR}^* = \hat{g}_1\left(X_i, \widehat{\alpha}_1\right) - \hat{g}_0\left(X_i, \widehat{\alpha}_0\right) + \frac{D_i\left(Y_i - \hat{g}_1\left(X_i, \widehat{\alpha}_1\right)\right)}{\hat{e}\left(X_i\right)} - \frac{\left(1 - D_i\right)\left(Y_i - \hat{g}_0\left(X_i, \widehat{\alpha}_0\right)\right)}{\left(1 - \hat{e}\left(X_i\right)\right)} \tag{10}$$

$$\tau(X) = \hat{Y}_{i,DR}^* = m(X_i) + \omega \tag{11}$$

In equation 10, $\hat{g}_1(X_i, \widehat{\alpha}_1) - \hat{g}_0(X_i, \widehat{\alpha}_0)$ is equivalent to the the score-function from the two-model approach. Simulation evidence of Knaus et al. (2018) suggests that estimators based on $Y^*_{i,DR}$ might be more stable because of the doubly-robust property and that the performance is competitive for the estimation of heterogeneous treatment effects in observational studies. The doubly-robust property states that, at least for the ATE, the estimator is consistent and unbiased if only one of the models, the regression or the propensity score, is correctly specified (Robins, Rotnitzky & Zhao, 1994; Robins & Rotnitzky, 1995). Lunceford & Davidian (2004); Williamson, Forbes & White (2014); Belloni, Chernozhukov & Hansen (2014) study the theoretical properties and highlight implications for practice. One of the findings is that the variance can be decreased when using the doubly-robust estimator instead of a simple inverse probability estimator (Lunceford & Davidian, 2004). Chernozhukov & Semenova (2018) show that equation 10 is conditionally locally robust to the estimation error of the nuisance parameter.

Next we state some asymptotic results to recover the the CATE. From equation 5 it follows that

$$\tau(X) = \mathrm{E}\left\{\mathrm{E}[Y|D=1, X] - \mathrm{E}[Y|D=0, X]|X = x_i\right\} \tag{12}$$

Let $\eta(X) \coloneqq (e(X), g_1(X_i, \alpha_1), g_0(X_i, \alpha_0))$ be the true high dimensional nuisance parameters. Following Fan et al. (2019) we can define

$$\psi(D, Y, X, \eta(X)) = g_1(X_i, \alpha_1) - g_0(X_i, \alpha_0) + \frac{D_i(Y_i - g_1(X_i, \alpha_1))}{e(X_i)} - \frac{(1 - D_i)(Y_i - g_0(X_i, \alpha_0))}{(1 - e(X_i))} \tag{13}$$

.

Theorem 1.1
(i) under Assumption 1,2,3,4

$$\mathrm{E}\left[g_1(X_i, \alpha_1) + \frac{D_i(Y_i - g_1(X_i, \alpha_1))}{e(X_i)}|X = x_i\right] = \mathrm{E}\left(Y^1|X = x_i\right),$$
$$\mathrm{E}\left[g_0(X_i, \alpha_0) + \frac{(1 - D_i)(Y_i - g_0(X_i, \alpha_0))}{1 - e(X_i)}|X = x_i\right] = \mathrm{E}\left(Y^0|X = x_i\right)$$

(ii)$\mathrm{E}[\psi(D, Y, X, \eta(X)) - \tau(X)|X = x_i] = 0$ given (i). This moment condition satisfies the Neyman-orthogonality condition. Neyman-orthogonality is a key component in ensuring that the CATE estimators are robust to the regularization bias inherent for the nuisance functions which are learned via machine learning models.

Through the doubly-robust estimator, $\tau(X)$, the weighted linear projection equation changes to

$$Y = \hat{\gamma} \times (D - \hat{e}(x)) \times \mathbf{I}(\hat{S}(X) \in I_k) + \nu, \tag{14}$$

with $\hat{S}(X) = \hat{m}(X_i)$. The interaction $(D - \hat{e}(x))$ is an orthogonalization of the treatment variable to all other covariates and used to increase precision. The Horvitz-Thompson transformation is excluded since the inverse probability weighting is already included in the doubly-robust estimator.

The second extension to the method is to weight each individual based on the group inclusion probability. Instead of taking the median over B repetitions for the K groups we store the information about the group estimate for each individual i over the B repetitions. The median is then taken over B repetitions for each individual rather than the groups. This allows us to get an estimate for each individual which can be used for comparison with other methods and to make predictions. Naturally, we can do the same in the first step and apply this weighting procedure on the score-function. The result is a robust estimate for the conditional average treatment effect.

Algorithm 2: Extended GATES

for *b=1 to B* **do**
  Split Data in $k = 2$ samples: I^a and M with $I^a \cup M$
  Train $Y_i^0 = g_0(X_i, D = 0) + U_{0i}$, with $i \in I^a$
  Train $Y_i^1 = g_1(X_i, D = 1) + U_{1i}$, with $i \in I^a$
  Train $D_i = e_0(X_i) + V$, with $i \in I^a$
    Predict $\hat{Y}_i^0 = \hat{g}_0(X_i)$, with $i \in I^a$
    Predict $\hat{Y}_i^1 = \hat{g}_1(X_i)$, with $i \in I^a$
    Predict $\hat{D}_i = \hat{e}(X_i)$, with $i \in I^a$
      Train $Y^*_{i,DR}$ on $X_i = l(X_i) + W$ with $i \in I^a$
      Predict $\hat{Y}^*_{i,DR} = \hat{l}(X_i)$ with $i \in M$
        Calculate $\hat{V}_i = D_i - \hat{e}(X_i)$, with $i \in M$
        Calculate $S_b(X|i) = \hat{Y}^*_{i,DR}$
  Estimate GATES parameters (γ) with OLS using M (see equation 14)
end
Average γ over B iterations: $\tilde{\gamma} = median\{\gamma\}$
Calculate Density for every i: $S_i(X)$ given $S_b(X|i)$ over all b
Calculate Final score-function ($\tilde{S}_i(X)$) given density of medians for i = 1 to N

3 Simulation Study

3.1 Data Generating Process

To evaluate the advantage of the proposed extensions i) doubly-robust first stage and ii) simplified parametric second stage, we use simulated data where the true treatment effects are known. In the following we describe the data generating process (DGP) in detail and show the variations that we consider. We generate the covariates X in a way that they are partially correlated among each other. The process is described in Algorithm 3.

Algorithm 3: Correlation Matrix

1 **Generate** random positive definite covariance matrix Σ based on a uniform distribution over the space $p \times p$ of the correlation matrix
2 **Scale** covariance matrix. This equals the correlation matrix and can be seen as the covariance matrix of the standardised random variables $\Sigma = \frac{X}{\sigma(X)}$.
3 **Generate** random normal distributed variables $X_{N \times p}$ with mean $= 0$ and variance $= \Sigma$

An illustration of the distribution for $p = 10$ and $N = 5000$ observations is given in Figure 3.1.

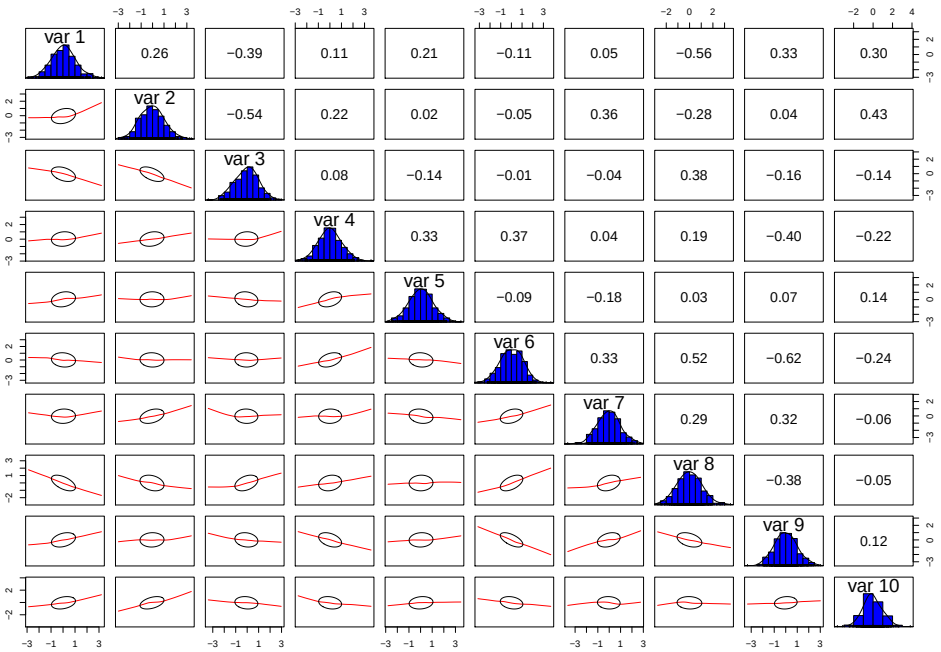

Figure 3.1: Correlation Matrix of Covariates. Correlation metric is bravais-pearson.

It shows that the covariates are correlated among each other. This is guaranteed through the uniform distribution of the covariance matrix which is then transformed to a correlation matrix. This assumption is more common in real datasets and helps to investigate the performance of machine learning algorithms, especially the regularization bias, in a more realistic manner.

The basic model used in this paper is a partially linear regression model based on Robinson (1988) with extensions:

$$Y = \tau(X)D + g_0(X) + U, \qquad E[U|X, D] = 0, \tag{15}$$

$$D = m_0(X) + V, \qquad E[V|X] = 0, \tag{16}$$

$$\tau(X) = t_0(Z) + W \qquad E[W|Z] = 0, Z \subset X \tag{17}$$

with Y being a continuous outcome variable. $\tau(X)$ is the true treatment effect or population uplift, while D is the treatment status. The vector $X = (X_1, ..., X_p)$ consists

of p different features, covariates or confounders, while the vector Z is a subspace of X and represents the variables on which the treatment effect is dependent. U, V and W are unobserved covariates which follow a random normal distribution = $N(0,1)$.

Equation 16 is the propensity score. In the case of completely random treatment assignment the propensity score $\hat{m_0}(X_i) = \theta_0$ for all units ($i = 1, ..., N$) with N being the number of observations. The scalar θ_0 can take any value between (0,1). Here we use 0.5 (balanced assignment). The covariates X are generated from a random multivariate normal distribution ($N(0,1)$) as follows:

The function $g_0(X)$ is calculated via a trigonometric function to make the covariates non-linear and potentially complicated for estimation.

$$g_0(X) = \cos(X \times b)^2 \tag{18}$$

The vector $b = \frac{1}{l}$ with $l \in \{1, 2, ..., k\}$ represents weights for every covariate. Next, a description of how to build the function $m_0(X)$ as well as how to create a heterogeneous treatment effect is given. A varying treatment effect implies that its strength differs among the observations and is therefore conditioned on some covariates Z. Regarding the treatment assignment (D) two options are considered. Option 1 assumes D to be completely random assigned among the observations. In this case, D is just a vector of random numbers with values 0 or 1. In the second option, the treatment assignment is dependent on the covariates. The functions are generated as follows:

Algorithm 4: Treatment Assignment

if *random assignment* **then**
  Generate $D \overset{ind.}{\sim} \text{Bernoulli}(c)$, with $c \in [0,1]$;
else
  Create Vector Multiply the matrix X by vector $b = \frac{1}{l}$ with $l \in \{1, 2, ..., p\}$ to get vector a.
  Make nonlinear $a = a + X_4 * X_8 + sin(X_5) + X_2$
  Calculate probability distribution for the vector a from the normal distribution function:

$$m_0(X) = \Phi\left(\frac{a - \mu(a)}{\sigma(a)}\right) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{a - \mu(a)}{\sigma(a)\sqrt{2}}\right)\right] \tag{19}$$

  Apply random number generator from a Binomial function $B(N, k, p)$ with probability (p) for success equals $m_0(X)$. This creates a vector $D \in \{0; 1\}$ such that $D \overset{ind.}{\sim} \text{Bernoulli}(m_0(X))$.
end

Regarding the treatment effect, three different options are considered. First, $\tau(X)$ is a constant for every unit. Second, $\tau(X)$ depends on all covariates and is continuous. Third, $\tau(X)$ only depends on some space Z of the covariates and further takes only two different values. The latter two options are especially useful when examining heterogeneous treatment effects. In the causal tree section, there will also be a fourth option in which

the treatment effect only depends on two covariates and is binary.

Algorithm 5: Treatment Effect

if *constant effect* **then**
  $\tau(X) = c$ with $c \in (-2, 5)$;
else if *simple heterogeneous effect* **then**
  Generate $\tau(X) \sim N(\mu, \sigma)$
  $\tau(X) = X_1 + (X_2 > 0) + N(0, 0.1)$;
else if *non-linear heterogeneous effect* **then**
  Apply trigonometric function:

$$\tau(X) = \sin(X \times b)^2 + W, \tag{20}$$
$$W \sim (N(0, 0.1)) \tag{21}$$

else
  Define Z as some feature space of X and apply CDF as in 19 and run Bernoulli trials:

$$Z = (X_6 \circ (X_1 \times X_5) \circ X_2)^2 \tag{22}$$
$$t_0(Z) = \Phi\left(\frac{Z - \mu(Z)}{\sigma(Z)}\right) \tag{23}$$
$$\tau(Z) \overset{ind.}{\sim} \text{Bernoulli}(t_0(Z)) \tag{24}$$

  Standardise the treatment effect within the set {-2,+5}.

$$\tau(X) = \frac{\tau(Z) - min(\tau(Z))}{max(\tau(Z)) - min(\tau(Z))}(5 + 2) - 2 \tag{25}$$

end

Treatment assignment (D) depends on some covariates, which includes a selection bias. The treatment effect ($\tau(X)$) is heterogeneous among the observations (continuous within the interval of approximately $(-2, +5)$). The outcome variable Y is created through a partially linear model in the form of $Y = \tau(X)D + g_0(X) + U$ with $g_0(X)$ being a non-linear function (e.g. $cos(x)$) and $U \sim N(0, 0.1)$ is the error-term.

3.2 Simulation Results

Figure 3.2 shows the densities for 49 randomly selected observations. The simulated data has the following properties. $N = 1000$, $X = \mathbb{R}^{20}$, $\mathrm{P}(D = 1) = 0.5$ and $\tau(X) \in (0.1, 0.3)$. We show that even in randomised experiments, the point estimates differ due to the sample-splitting in the first step. Averaging them by taking the mean leads to a more stable conditional treatment effect function over all observations.

Figure 3.3 shows the results from a simulation, which compares the ATE estimated with the two-model and the doubly-robust estimator, respectively. The average treatment effect is used to build one estimator over the K groups. The groups are divided by a quantile function and hence have equal length. This leads to unequal frequency binning

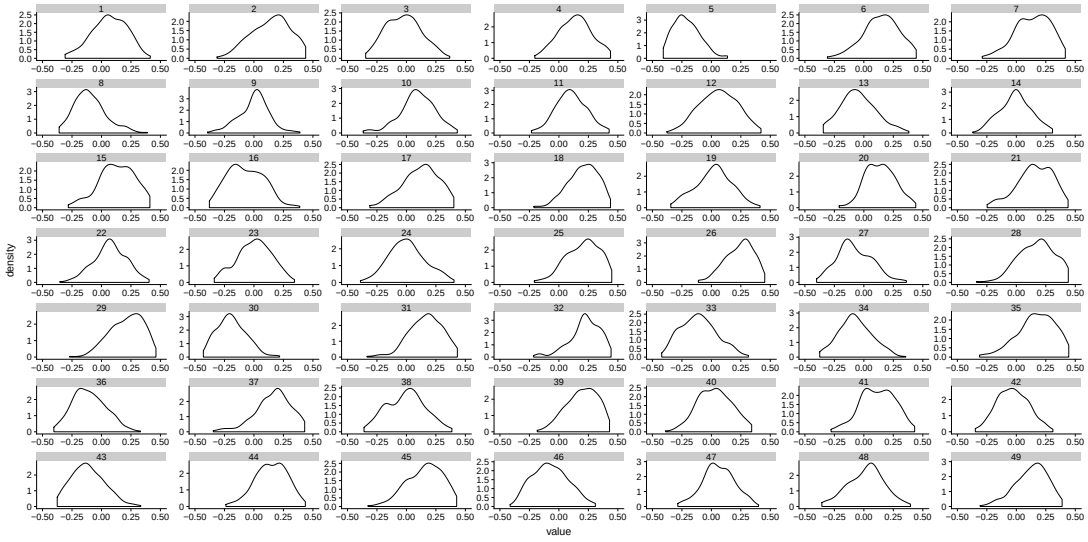

Figure 3.2: Distribution of scores ($\tau(X)$) for 49 randomly selected individuals.

within the groups. We, therefore, assign every observation the value of the group average treatment effect in where they belong to. After M iterations we take the median to get one estimate for every observation. To get the estimated ATE we simply take the mean over each observation in the whole dataset. We show the absolute deviation from the true ATE for different settings. We use Monte Carlo resampling 10 times for each setting and show the single results in Figure 3.3. We also state the average result (error) for each data generating process in Table 3.1. A two-sample Welch t-test confirms that the hypothesis of equal means can be rejected based on a 1% significance level for each setting. Algorithm 6 describes the estimation for the ATE based on the k groups from the GATES as well as the MAE estimation over S datasets. Naturally, this imposes a new estimator for the ATE.

Table 3.1: Settings and Monte Carlo averages

Scenarios	A	B	C	D	E	F
N	1000	1000	5000	5000	5000	5000
$\mathbb{R}^p$	100	200	100	200	200	500
$P(D = 1)$	$m(X)$	$m(X)$	$m(X)$	$m(X)$	0.5	0.5
$\tau(X)$	constant	continuous	continuous	binary	continuous	binary
Average error Two-Model	0.19	0.20	0.16	0.16	0.18	0.20
Average error Doubly-Robust	0.15	0.16	0.13	0.12	0.13	0.15

Algorithm 6: ATE estimation from GATES and error estimation

```
for s=1 to S do
    for b=1 to B do
        Assign group average treatment effect from group k to observations in
         group k
        Store results in some matrix R_nxB
    end
        Average Take median for each observation over B bootstraps
        Estimate mean absolute error (MAE) from true treatment effect
                  for both estimators
        Store results in some matrix Q_Sx2
    Resample keeping specifications constant (monte carlo study)
end
Average errors over S iterations
```

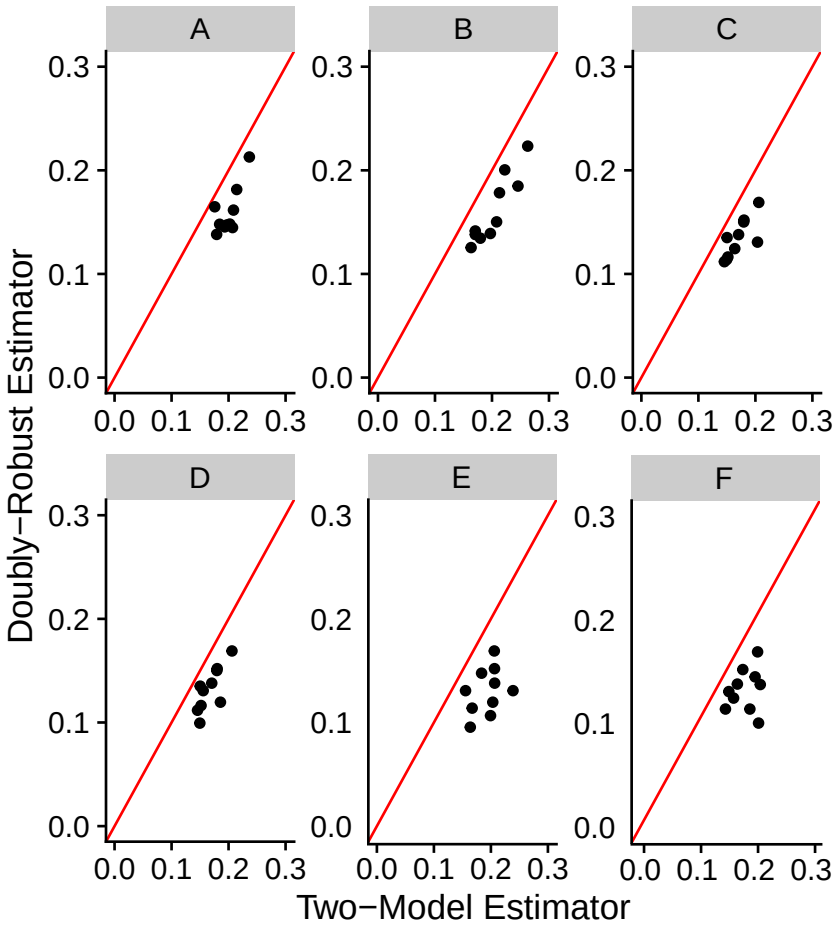

Figure 3.3: Comparison of two-model approach and doubly-robust. Axes show absolute error between estimated ATE and true ATE. 45-degree line indicates the equality of both methods.

Our simulation shows that, for all the considered data settings, our method decreases the error of the true individual treatment effect. Setting A:D show results for non-randomized

settings with different parameters. We even find that the proposed extensions produce a smaller MAE in randomized control trials (see Figure 3.3: E, F). This is true for every resampling of the DGP and each setting. Surprisingly there is no difference between the treatment assignment mechanism. The doubly-robust method is always better but even in randomized settings the error is in the range of 0.1 to 0.3. Looking at the MAE we find the highest difference between the two methods for random assignment (probability = 0.5).

4 Conclusion

In this paper, we extend the idea of reporting group average treatment effects towards the combination of machine learning methods and parametric estimation for non-randomized control trials. Since flexibility in terms of the model choice, as well as interpretability of the results, is of main interest we extend the idea of the GATES approach towards the use of a doubly-robust estimator. This ensures to control for self-selection into treatment which is a realistic assumption in observational studies.

We find that using a doubly-robust estimator with cross-fitting, in combination with a simplified parametric model, decreases the error compared to a two-model approach significantly. A disadvantage when estimating the CATE from the GATE is, that we can only assign k different values to the individuals. In our setting, we considered only five different groups. This amount could be increased to e.g. 10 or even more groups. In empirical settings, it would depend on the sample size. If we want to have at least 30 observations within a group we could have $\frac{N}{30 \times \Lambda}$ groups, with Λ-splits or folds of the dataset in the first stage. Here we considered only two-folds. However, there is no general relationship between the number of folds in cross-fitting and the precision of the estimator (see Chernozhukov et al. (2018) for an example with different folds). Due to computational reasons we only use $B = 10$ iterations within the same sample and $S = 10$ Monte Carlo re-samplings of the same data generating process. This amount needs to be increased to e.g. 50 and 100, respectively. At this stage, we only consider a boosting-trees algorithm (with parameter tuning via 10 fold cross-validation) as a machine learning method. In a further draft, we will extend this to the use of boosted gradient descent (XGBoost), random forest algorithm, neural networks and some linear methods like variants of the Elastic Net. We can even consider different methods for each nuisance function.

In a further draft, we would also compare the ATE as well as the CATE, all resulting from the group average treatment effect, and compare them with recent methods that estimate the former parameter or function.

References

Athey, S., Wager, S., & Tibshirani, J. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148–1178.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, *81*(2), 608–650.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68.

Chernozhukov, V., Demirer, M., Duflo, E., & Fernandez-Val, I. (2018). Generic machine learning inference on heterogenous treatment effects in randomized experiments. Preprint at https://arxiv.org/abs/1712.04802.

Chernozhukov, V., Fernández-Val, I., & Luo, Y. (2018). The sorted effects method: Discovering heterogeneous effects beyond their averages. *Econometrica*, *86*(6), 1911–1938.

Chernozhukov, V. & Semenova, V. (2018). Simultaneous inference for Best Linear Predictor of the Conditional Average Treatment Effect and other structural functions. CeMMAP working papers CWP40/18, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Fan, Q., Hsu, Y.-C., Lieli, R. P., & Zhang, Y. (2019). Estimation of conditional average treatment effects with high-dimensional data. Preprint at https://arxiv.org/abs/1908.02399.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction, Springer Series in Statistics*. Springer New York.

Heckman, J., Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing selection bias using experimental data. Technical report, National bureau of economic research.

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*(260), 663–685.

Knaus, M., Lechner, M., & Strittmatter, A. (2018). Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. *CEPR Discussion Paper No. DP13402*.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, *116*(10), 4156–4165.

Lee, S., Okui, R., & Whang, Y.-J. (2017). Doubly robust uniform confidence band for the conditional average treatment effect function. *Journal of Applied Econometrics*, *32*(7), 1207–1225.

Lunceford, J. K. & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, *23*(19), 2937–2960.

Nie, X. & Wager, S. (2017). Quasi-oracle estimation of heterogeneous treatment effects. Preprint at https://arxiv.org/abs/1712.04912.

Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T., & Tibshirani, R. (2018). Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in Medicine*, *37*(11), 1767–1787.

Robins, J. M. & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, *90*(429), 122–129.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, *89*(427), 846–866.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 931–954.

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.

Williamson, E. J., Forbes, A., & White, I. R. (2014). Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in Medicine*, *33*(5), 721–737.

Zimmert, M. & Lechner, M. (2019). Nonparametric estimation of causal heterogeneity under high-dimensional confounding. Preprint at https://arxiv.org/abs/1908.08779.

IRTG 1792 Discussion Paper Series 2019

For a complete list of Discussion Papers published, please visit http://irtg1792.hu-berlin.de.

001 "Cooling Measures and Housing Wealth: Evidence from Singapore" by Wolfgang Karl Härdle, Rainer Schulz, Taojun Xie, January 2019.

002 "Information Arrival, News Sentiment, Volatilities and Jumps of Intraday Returns" by Ya Qian, Jun Tu, Wolfgang Karl Härdle, January 2019.

003 "Estimating low sampling frequency risk measure by high-frequency data" by Niels Wesselhöfft, Wolfgang K. Härdle, January 2019.

004 "Constrained Kelly portfolios under alpha-stable laws" by Niels Wesselhöfft, Wolfgang K. Härdle, January 2019.

005 "Usage Continuance in Software-as-a-Service" by Elias Baumann, Jana Kern, Stefan Lessmann, February 2019.

006 "Adaptive Nonparametric Community Detection" by Larisa Adamyan, Kirill Efimov, Vladimir Spokoiny, February 2019.

007 "Localizing Multivariate CAViaR" by Yegor Klochkov, Wolfgang K. Härdle, Xiu Xu, March 2019.

008 "Forex Exchange Rate Forecasting Using Deep Recurrent Neural Networks" by Alexander J. Dautel, Wolfgang K. Härdle, Stefan Lessmann, Hsin-Vonn Seow, March 2019.

009 "Dynamic Network Perspective of Cryptocurrencies" by Li Guo, Yubo Tao, Wolfgang K. Härdle, April 2019.

010 "Understanding the Role of Housing in Inequality and Social Mobility" by Yang Tang, Xinwen Ni, April 2019.

011 "The role of medical expenses in the saving decision of elderly: a life cycle model" by Xinwen Ni, April 2019.

012 "Voting for Health Insurance Policy: the U.S. versus Europe" by Xinwen Ni, April 2019.

013 "Inference of Break-Points in High-Dimensional Time Series" by Likai Chen, Weining Wang, Wei Biao Wu, May 2019.

014 "Forecasting in Blockchain-based Local Energy Markets" by Michael Kostmann, Wolfgang K. Härdle, June 2019.

015 "Media-expressed tone, Option Characteristics, and Stock Return Predictability" by Cathy Yi-Hsuan Chen, Matthias R. Fengler, Wolfgang K. Härdle, Yanchu Liu, June 2019.

016 "What makes cryptocurrencies special? Investor sentiment and return predictability during the bubble" by Cathy Yi-Hsuan Chen, Roméo Després, Li Guo, Thomas Renault, June 2019.

017 "Portmanteau Test and Simultaneous Inference for Serial Covariances" by Han Xiao, Wei Biao Wu, July 2019.

018 "Phenotypic convergence of cryptocurrencies" by Daniel Traian Pele, Niels Wesselhöfft, Wolfgang K. Härdle, Michalis Kolossiatis, Yannis Yatracos, July 2019.

019 "Modelling Systemic Risk Using Neural Network Quantile Regression" by Georg Keilbar, Weining Wang, July 2019.

IRTG 1792, Spandauer Strasse 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.

IRTG 1792 Discussion Paper Series 2019

For a complete list of Discussion Papers published, please visit http://irtg1792.hu-berlin.de.

020 "Rise of the Machines? Intraday High-Frequency Trading Patterns of Cryptocurrencies" by Alla A. Petukhina, Raphael C. G. Reule, Wolfgang Karl Härdle, July 2019.

021 "FRM Financial Risk Meter" by Andrija Mihoci, Michael Althof, Cathy Yi-Hsuan Chen, Wolfgang Karl Härdle, July 2019.

022 "A Machine Learning Approach Towards Startup Success Prediction" by Cemre Ünal, Ioana Ceasu, September 2019.

023 "Can Deep Learning Predict Risky Retail Investors? A Case Study in Financial Risk Behavior Forecasting" by A. Kolesnikova, Y. Yang, S. Lessmann, T. Ma, M.-C. Sung, J.E.V. Johnson, September 2019.

024 "Risk of Bitcoin Market: Volatility, Jumps, and Forecasts" by Junjie Hu, Weiyu Kuo, Wolfgang Karl Härdle, October 2019.

025 "SONIC: SOcial Network with Influencers and Communities" by Cathy Yi-Hsuan Chen, Wolfgang Karl Härdle, Yegor Klochkov, October 2019.

026 "Affordable Uplift: Supervised Randomization in Controlled Exprtiments" by Johannes Haupt, Daniel Jacob, Robin M. Gubela, Stefan Lessmann, October 2019.

027 "VCRIX - a volatility index for crypto-currencies" by Alisa Kim, Simon Trimborn, Wolfgang Karl Härdle, October 2019.

028 "Group Average Treatment Effects for Observational Studies" by Daniel Jacob, Stefan Lessmann, Wolfgang Karl Härdle, November 2019.

IRTG 1792, Spandauer Strasse 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.