IRTG 1792 Discussion Paper 2018-063

Causal Inference using Machine Learning An Evaluation of recent Methods through Simulations

Daniel Jacob*
Stefan Lessmann*
Wolfgang Karl Härdle*

* Humboldt-Universität zu Berlin, Germany

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

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

International Research Training Group 1792

Causal Inference using Machine Learning*

An Evaluation of recent Methods through Simulations

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

November, 2018

Abstract

The estimation of a causal parameter in a high-dimensional setting where the functions are potentially complex is a challenging task. Parametric and linear modelling is not sufficient to generate unbiased and consistent estimators. Modern approaches, therefore, use machine learning (ML) algorithms to learn these nuisance functions. However, this leads to new problems like the regularization bias or overfitting that are common when using ML models.

This paper considers different novel methods that overcome these problems or at least address them. These methods differ in terms of the target parameter, namely the average treatment effect of the population, group heterogeneity or the conditional average treatment effect for each individual. Each method is first investigated and tested separately and second, they are compared among each other. To do this in a disciplined manner, simulations with synthetic data are used. This ensures that all distributions of the generated treatment effect parameters are known. The findings are that each method has its limits in terms of unbiased estimation, the detection of heterogeneity and also the determination of which covariates are responsible for different causal effects.

JEL classification: C01, C14, C31, C63

Keywords: *causal inference, machine learning, simulation study, sample-splitting double machine learning, sorted group ATE (GATES), causal tree*

*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 authors 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

This paper aims at investigating some novel methods that can be used to estimate the structural or causal effect of a treatment on some outcome. To highlight the difference between marginal and causal effects the Neyman-Rubin Causal Model is described by using an example. This framework, whit all its assumptions, has gained substantial acceptance for analyzing this sort of problems.

It is assumed that some data from a newsletter campaign are given and the task is to evaluate the effect that this campaign has in terms of the outcome e.g. the purchase amount. The first assumption is that every customer has two potential outcomes: Y_i^1 and Y_i^0. These outcomes have fixed values and only one of them can be observed. The first, if a customer has received the newsletter or the latter if not. These two states can be seen as either being in the treatment group ($D = 1$) or the control group ($D = 0$). The fixed outcomes require a necessary assumption in the potential outcome framework, namely the **stable unit treatment value assumption** (SUTVA) (see Rubin (1980)). It guarantees that the potential outcome of a customer is unaffected by changes in the treatment assignment of all other customers. This assumption might be violated if customers can interact with each other or if some customers are connected. Of course, the outcome state is not only depending on the newsletter but also on some observed features or covariates (X) like the e-mail address that belongs either to a company or a private person. Maybe companies buy more expensive products. However, it could be the case that they get a lot of newsletters and are therefore not that sensitive for a respond. In this scenario, the treatment assignment would also depend on the covariates. If the assignment of the newsletter is completely random; the covariates are equally balanced among the two groups. Given this setting, the so-called unconfoundedness assumption holds: $(Y_i^1, Y_i^0, X_i) \perp\!\!\!\perp D_i$.

This assumption, often also called exogeneity or strong ignorability, states that the treatment assignment is independent of all pre-treatment characteristics. Especially from the potential outcome. This means that whether or not a customer received the treatment is independent on how she would respond to the newsletter. In practice however, it is often the case that the treatment assignment depends at least on some covariates. In this case, the **definition of ignorability** or **unconfoundedness** needs to be expanded to a weaker but more realistic case:

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

As defined by Rubin (1978), *"ignorability of treatment assignment holds when all variation in D is completely random within strata defined by all combinations of values on all variables, X, that systematically determine all treatment assignment patterns."* Since both states cannot be observed at the same time, the treatment effect $(Y_i^1 - Y_i^0)$ for every customer cannot be estimated directly. What can be done in a random assignment setting is to estimate the average treatment effect (ATE) and report if the campaign had an effect at all. Likewise, if there would have been two different newsletters the question which one had the bigger impact on the outcome could be answered. The ATE (θ_0) for one campaign would then be calculated as follows:

$$
\begin{aligned}
\theta_0 &= E[\theta_i] \\
&= E[Y_i^1 - Y_i^0] \\
&= E[Y_i^1] - E[Y_i^0] && \text{due to linearity in expectations} \\
&= E[Y_i^1 | D_i = 1] - E[Y_i^0 | D_i = 0] && \text{due to the independence assumption.}
\end{aligned}
$$

If the setting is expanded to observational studies where the treatment D depends on X, a calculation of the ATE conditional on the covariates, namely the conditional average treatment effect (CATE) is possible:

$$
\begin{aligned}
\theta_0(x) &= E[Y_i^1 | D_i = 1, X_i = x] - E[Y_i^0 | D_i = 0, X_i = x], && (2) \\
&= \mu_1(x) - \mu_0(x). && (3)
\end{aligned}
$$

Given the covariates, subgroups can be defined where customers are identical in terms of observables but different in their treatment assignment. Nearby effects would then be observed and they could be interpreted as group effects or even individual effects. This is a big advantage over the ATE. However, it is not that easy. For the average treatment effect, it is easy to see that under the specific assumption this estimate will be unbiased and consistent. To calculate the CATE it needs to be guaranteed that no subpopulation defined by $X = x$ is entirely located in the treatment or control group. Both outcomes need to be identified with respect to the covariates, otherwise the differential could not be built. This constraint, which is implicitly used in equation 2, is called the **overlap assumption** (Morgan and Winship, 2015) which formally states:

$$
\forall x \in supp(X), \quad 0 < P(D = 1 | X = x) < 1. \tag{4}
$$

In marketing campaigns, it is not unusual that this assumption is violated. Two possible solutions would then be to either drop these observations or limit the identification in terms of the observed covariates to ensure that both groups are represented. The treatment assignment probability as a function of the covariates $P(D = 1|X = x)$ can further be defined as the **propensity score**:

$$
\pi(x) \equiv P(D = 1 | X = x). \tag{5}
$$

Especially for the estimation of heterogeneous treatment effects, propensity score weighting reduces the effect of confounding covariates. The idea is to assign a weight to each treated customer equal to the inverse of the propensity score and to each untreated customer the weight one minus the propensity score. This is called **inverse probability weighting** by Rosenbaum and Rubin (1983), whereby similar baseline characteristics are obtained. The weights w for a linear regression can be represented as

$$
w = \frac{D}{\pi(x)} + \frac{(1 - D)}{1 - \pi(x)}. \tag{6}
$$

Athey and Imbens (2017) have shown that using propensity score methods lead to semi parametrically efficient estimators for the average treatment effect if the number of

covariates is low and fixed (e.g. Austin (2011)). In high dimension, the use of machine learning methods, such as boosting or random forests to estimate the propensity score, works quite well as McCaffrey et al. (2004) and Wyss et al. (2014) show. In fact, the estimation of probabilities given a large set of covariates is nothing less than a prediction problem in where ML methods are superior. A concern with this method is the high variance given the variability on the weights which depends heavily on the used estimation method. This especially applies in a setting with many covariates (see Low et al. (2016)).

In the presence of high-dimensional nuisance functions one different way to estimate a low dimensional average treatment effect is to use regularized regression. The idea is to overcome the regularization bias through partialing out the effect that the covariates have on the treatment assignment as well as on the outcome. This approach is known as **orthogonalization** and is similar to the Frisch-Waugh-Lovell theorem (1930). There is a variety of literature that focuses on obtaining such estimates that are $\sqrt{N}$-consistent and asymptotically normal distributed. While, among others, Levit (1976) and Robinson (1988) use kernels or series for the estimation of the non-parametric functions, Belloni et al. (2014) and Chernozhukov et al. (2015) extend this approach in a way that allows the use of any machine learning method. Using the same observations for both, the non-parametric and the semi-parametric estimation, does however lead to a bias which is why Chernozhukov et al. (2018a) introduced **sample-splitting** as an extension to the orthogonalization approach. The resulting method is called double machine learning (DML) since it relies on estimating primary and auxiliary predictive models. They show that the low-dimensional estimator is $n^{\frac{1}{2}}$ consistent as long as the non-parametric regression converges at the rate of $n^{-\frac{1}{4}}$. This is the first method which is presented and investigated in this paper. Recently, Mackey et al. (2017) show that the rate of convergence can be improved given a more complex or higher-dimensional setting. They restore robustness by employing a k-th order of orthogonality that results in an $n^{-1/(2k+2)}$ rate of convergence. The orthogonal approach is very promising for cross-section data but can also be extended to panel data as Chernozhukov et al. (2017b) show. In their paper they further allow for a number of treatments that grow with the sample size.

Of increasing interest is heterogeneous treatment effect estimation. An already well-established approach is to build a causal tree (Athey and Imbens, 2016). The idea is to build leaves in terms of covariates to get conditional average treatment effects. Since a tree structure is easy to interpret this method is used to investigate heterogeneous treatment effects. This approach can be extended towards the use of many trees (a forest) instead of just one (see Wager and Athey (2017)). One important assumption for these methods to work is that a randomized treatment assignment exists. An exception can be found by Athey et al. (2016) where they use gradient boosing in a flexible high-dimensional and observational environment. While the estimation of individual treatment effects is still lacking consistency under fewer assumptions, one way to address this problem comes from Chernozhukov et al. (2018b). Instead of trying to report treatment effects for every possible subgroup, which can be a large number and hence might lead to overfitting, they only concentrate on estimating sorted group average treatment effects (GATES) and try to detect specific characteristics, like the heterogeneity, about these groups. Their approach is agnostic and does not rely on unrealistic assumptions. They adopt the two main ingredients from Chernozhukov et al. (2018a) namely the sample-splitting and the orthogonal approach. Based on profound research for this paper, this approach has not been investigated independently in any other paper which is the reason to include this into the analysis within this paper. Furthermore, the use of orthogonalization seems to become

more popular. Oprescu et al. (2018) combine this technique with the generalized random forests while using a partially linear model. They show that their algorithm consistently outperforms baseline approaches.

An overview of the intersection of machine learning and causal inference in where several methods are described is given by Athey (2017). For an extended overview of recent methods in heterogeneous treatment effect models see Powers et al. (2017). The authors compare and evaluate the performance of different methods such as pollinated transformed outcome (PTO) forests, causal boosting and causal multivariate adaptive regression splines (MARS) and others. However, they do not use any of the methods that are investigated in this paper.

While many papers apply their methods directly on real data to estimate causal parameters (for example the DML method in Chernozhukov et al. (2018a) and the GATES algorithm in Chernozhukov et al. (2018b)) the use of controlled simulations comes rather short. If the purpose is to evaluate ML methods in terms of prediction it is easy to do this on an independent validation set. For the estimation of causal parameters, however, the true value is not observed which makes it hard to empirically test these methods. In the above-mentioned paper from Powers et al. (2017) they use synthetic data to evaluate the performance of different methods and find that causal boosting performs best on average but not on all data generating processes. Schuler et al. (2018) use simulations for randomized and observational data and compare different approaches. They not only use the same scenarios based on Powers et al. (2017) but also come to the same conclusion: Researches should not rely on a single method but evaluate multiple models using an objective function learned from the validation set. This out-of-sample set again only evaluates the performance of the ML algorithm in terms of prediction of the nuisance functions. This paper suggests, to carefully use the hypothesis that the best ML model is also the best method for estimating the treatment parameter. Only observed variables like the outcome can be evaluated but not the true parameter of interest. Model complexity, dimensionality, heterogeneity and the form of treatment assignment needs to be considered to choose the right model.

The generation of such variables for simulations comes with certain obstacles which is why Wendling et al. (2018) use real covariate and treatment assignment data and only simulate the outcome based on non-parametric models of the real outcome. They find that, among different evaluated methods, Bayesian additive regression trees as well as again causal boosting provide the lowest bias.

In this paper, some novel methods for the estimation of causal parameters are investigated. To do this in a disciplined fashion, simulations with synthetic data is used. This data is altered, according to the different objectives of the chosen models. Therefore, every model is analyzed in a single section on independent datasets that differ in terms of the above-mentioned parameters - especially the relationship between covariates and treatment assignment as well as the effect. The methods are also compared among each other under different model and data specifications for the one common parameter, namely the average treatment effect. This paper concludes with a discussion about the findings and recommendations on how an empiricist should use these methods. Among the methods, a perspective on further research and some notable problems are discussed.

2 Simulated Data

Since the true treatment effect is not known beforehand, a simulation model is necessary to evaluate different approaches in terms of biasedness for parameter estimation. The generation of synthetic data also allows for the control of all the amount of observations, the dimensionality and the distributions of the variables. The possibility to specify datasets for different simulations and scenarios helps to investigate the methods used in this paper.

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

$$Y = \theta_0 D + g_0(X) + U, \qquad E[U|X, D] = 0, \tag{7}$$

$$D = m_0(X) + V, \qquad E[V|X] = 0, \tag{8}$$

$$\theta_0 = t_0(Z) + W \qquad E[W|Z] = 0, Z \subset X \tag{9}$$

Y is the outcome variable (e.g. checkout amount from an online shop (continuous) or the status that a customer purchased something i.e. the conversion rate (binary)). θ_0 is the true treatment effect or population uplift, while D is the treatment status. The vector $X = (X_1, ..., X_k)$ consists of k 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 8 is the propensity score. In the case of completely random treatment assignment the propensity score $\hat{m}_0(X_i) = 0.5$ for all units $(i = 1, ..., N)$ with N being the number of observations. The covariates X are generated from a random multivariate normal distribution $(N(0, 1))$. Note that all values are continuous. In business applications, discrete values (categorical variables) are very common. For the data generation process as well as for the evaluation it would make no difference if such variables are present or not. This is due to the fact that the used machine learning methods can handle categorical variables quite well. The generation is done as follows:

Algorithm for Covariates (X)

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

The number of covariates is flexible and within the simulations different amounts $k = \{20, 40, 60, 200\}$ are used, depending on the purpose of evaluation. An illustration of the distribution for $k = 10$ and $N = 5000$ observations is given in Figure 2.1.

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 ML algorithm, especially the regularization bias, in a more realistic manner.

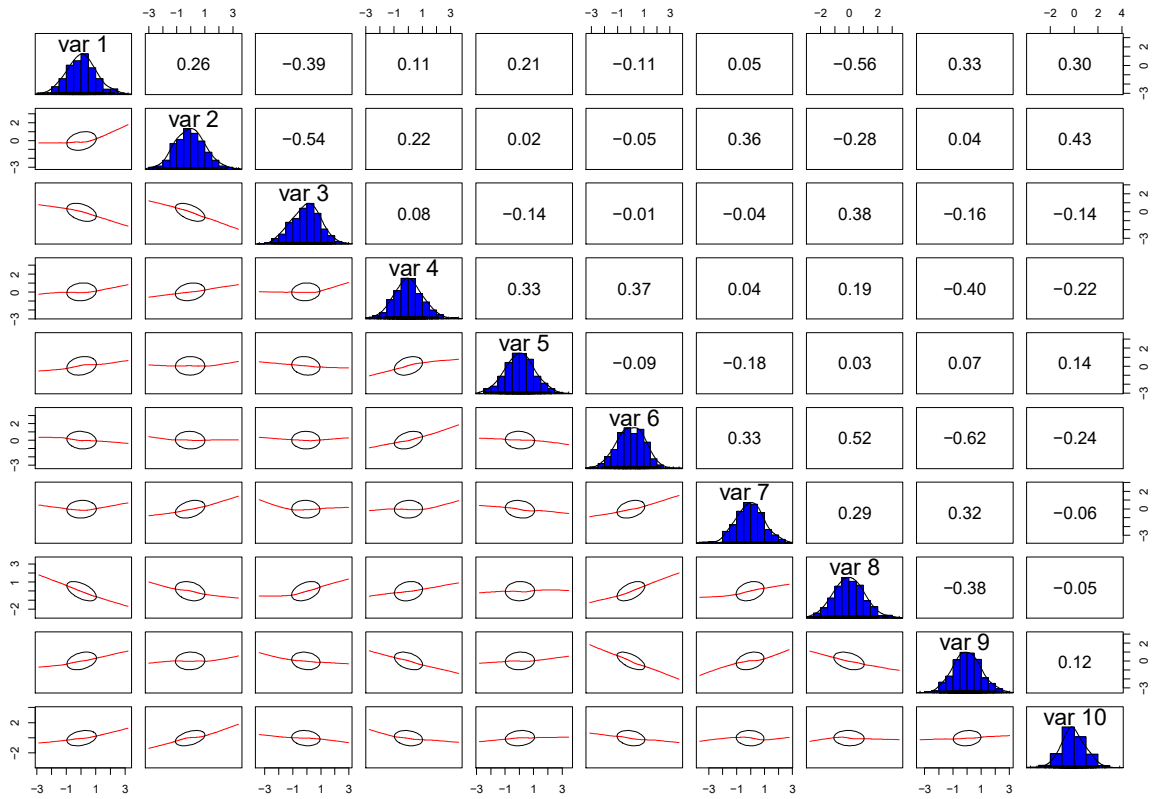


Figure 2.1: Correlation Matrix of Covariates. Correlation metric is bravais-pearson.

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

$$g_0(X) = \cos(X \times b)^2 \tag{10}$$

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 binary treatment assignment is generated through a Bernoulli function. This implies per default a sort of uncertainty or random error. Even if the probability from the propensity score is at 90% for $D = 1$ there is still a 10% chance that it is generated to be zero. This generation seems to differ from the above equation 8. Since the random error V does not matter in its specific values it can be seen as equivalent. The functions are generated as follows:

Algorithm for Treatment Assignment $(m_0(X, \theta))$

Option 1: m_0

$$D \overset{ind.}{\sim} \text{Bernoulli}(m_0), \qquad \text{with} \quad m_0 = 0.5 \tag{11}$$

Option 2: $m_0(X)$

1. Multiply the matrix X by vector $b = \frac{1}{l}$ with $l \in \{1, 2, ..., k\}$ to get vector a.
2. Calculate the probability distribution for the vector a from the normal distribution function:

$$m_0(X) = \Phi\left(\frac{a - \mu(a)}{\sigma(a)}\right) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{a - \mu(a)}{\sigma(a)\sqrt{2}}\right)\right] \tag{12}$$

3. Apply a 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 $D \overset{ind.}{\sim} \text{Bernoulli}(m_0(X))$.

To show how well a matching is done, the mean of each covariate against the estimated propensity score, separately by treatment status is plotted. The better the matching, the more equal is the mean of each covariate for the treatment and control group at each value of the propensity score. Figure A.1 presents the randomized assignment (option 1) whereas Fig A.2 shows the conditioned assignment (option 2). It can be seen, that the propensity score for the former case only takes values in the interval $(0.3, 0.67)$. The mean is clearly concentrated around 0.5 which is to be assumed given a random assignment of the treatment status.

Regarding the treatment effect, three different options are considered. First, θ_0 is a constant for every unit. Second, θ_0 depends on all covariates and is continuous. Third, 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 where the treatment effect only depends on two covariates and is binary.

Algorithm for Treatment Effect (θ_0)

Option 1: θ_0 = constant (c), with c = 0.2

Option 2: $\theta_0(X) = [0.1, 0.3]$

1. Apply trigonometric function:

$$t_0(X) = \sin(X \times b)^2 + W = \gamma, \tag{13}$$

$$W = \text{random normal distributed values} : (N(0, 0.25)) \tag{14}$$

2. Standardize the treatment effect within the interval [0.1,0.3]. This ensures θ_0 to be at most 30% of the baseline outcome ($g_0(X)$):

$$\theta_0(X) = \frac{\gamma - min(\gamma)}{max(\gamma) - min(\gamma)}(0.3 - 0.1) + 0.1 \tag{15}$$

Option 3: $\theta_0(Z) = \{0.1, 0.3\}$

1. Define Z as some feature space of X and apply CDF as in 12 and run Bernoulli trials:

$$Z = (X_6 \circ (X_1 \times X_5) \circ X_2)^2 \tag{16}$$

$$t_0(Z) = \Phi\left(\frac{Z - \mu(Z)}{\sigma(Z)}\right) \tag{17}$$

$$\theta_0(Z) \overset{ind.}{\sim} \text{Bernoulli}(t_0(Z)) \tag{18}$$

2. Standardize the treatment effect within the set {0.1,0.3} as above.

3 Experiments and Model Evaluation

In this section, the chosen methods that use machine learning algorithms to estimate treatment effects are presented. These methods are semi-parametric as they usually try to estimate a low-dimensional parameter of interest with parametric assumptions while estimating non-parametric nuisance functions with the help of machine learning methods.

Three methods are being focused here that are promising or already established in the literature for causal parameter estimation. The motivation for the selection of these methods is based on the parameter of interest namely the treatment effect (TE). One might be interested only in the average treatment effect for which the double machine learning (DML) method by Chernozhukov et al. (2018a) is used. If there is reason to believe that there is heterogeneity in the TE the second approach by Chernozhukov et al. (2018b) focuses on key features like average effects sorted by impact groups and also most and least impacted units. To get scores about individuals the last method introduced by Athey and Imbens (2016) creates partitions of the feature space with the use of a tree method and estimates a treatment effect for each leaf. However, the last method needs strong assumptions on the low-dimensional parameter and can't handle high-dimensional covariates very well. Every method is tested on different simulated data and the behaviour

and accuracy are investigated. It is shown under which assumptions the models are potentially biased or lack in interpretable results. For every model a detailed guide on how they could be implemented and what kind of considerations might need to be considered when applying them to real datasets is given.

3.1 Double (Debiased) Machine Learning (DML)

The method proposed by Chernozhukov et al. (2018a) and likewise Chernozhukov et al. (2017a) efficiently estimate an average treatment effect parameter within a partial linear model. It orthogonalizes out the effect from the covariates on both, the outcome and the treatment assignment. Therefore, they run non-parametric regressions using sophisticated machine learning algorithms. In a second step, they run a parametric linear regression model on the residuals of outcome and treatment to get the low-dimensional treatment effect parameter. A notable extension to the idea already introduced by Robinson (1988) is the use of sample-splitting. Using the whole sample to learn the nuisance functions as well as to estimate the target parameter would lead to overfitting. However, through sample-splitting, the subsamples are smaller and specific and could be biased. To account for this introduced uncertainty Chernozhukov et al. (2018a) suggest to repeat the main estimation procedure for a large number M and rearrange the data in each replication. They then report estimates from a distribution rather than point estimates.

The algorithm can be implemented as follows:

Algorithm for DML

1. Split Data in $k = 2$ samples: I^a and I with $I^a \cup I$
2. Train $Y_i = \hat{g_0}(X_i) + \hat{U}_i$, with $i \in I^a$
3. Train $D_i = \hat{m_0}(X_i) + \hat{V}_i$, with $i \in I^a$
4. Estimate $\hat{Y}_i = \hat{g_0}(X_i)$, with $i \in I$
5. Estimate $\hat{D}_i = \hat{m_0}(X_i)$, with $i \in I$
6. Residualize $\hat{W}_i = Y_i - \hat{Y}_i \quad \text{and} \quad \hat{V}_i = D_i - \hat{D}_i$, with $i \in I$
7. Estimate $\hat{\theta}_0(I, I^a) = \left(\sum_{i \in I} \hat{V}_i D_i \right)^{-1} \sum_{i \in I} \hat{V}_i (Y_i - \hat{g_0}(X_i))$
8. Switch sample I^a and I
9. Repeat steps 2 to 6
10. Estimate DML:

$$\hat{\theta}_0(I^a, I) = \left(\sum_{i \in I^a} \hat{V}_i D_i \right)^{-1} \sum_{i \in I^a} \hat{V}_i (Y_i - \hat{g_0}(X_i)) \tag{19}$$

11. Average the two parameters:

$$\tilde{\theta}_0 = \frac{1}{2} \left(\hat{\theta}_0(I, I^a) + \hat{\theta}_0(I^a, I) \right) \tag{20}$$

12. Repeat steps 1 to 11 M times and average the resulting $\tilde{\theta}_0$.

The authors show that the second stage is $n^{-\frac{1}{2}}$ consistent as long as the estimator of the first stage converges at a rate faster than $n^{-\frac{1}{4}}$. Furthermore, not only will the bias of the target parameter disappear but also that it is asymptotically normal distributed. The orthogonalization removes the bias which is introduced through regularization from the machine learning algorithms and guarantees a robust estimate of the average treatment effect.

What is worth mentioning is the fact that even if the data generating process follows strict rules about distribution with first and second moment about the covariance matrix and therefore the covariates, the created data can differ in each generation. This affects also the treatment assignment as well as the treatment effect and hence the level of outcome. This variation in the generation of the data could be minimized by setting treatment assignment to random (option 1 for $m_0(X)$) and also make the treatment effect independent of the covariates by assigning a fixed scalar (option 1 for θ_0). However, it seems to be the case that even with this little variation the algorithm does not capture the true parameter every time.

To see this result, 20 independent datasets with all parameters being constant (i.e. 5000 observations, 20 covariates, treatment assignment dependent on X and theta = 0.2) are generated. Then the DML with M = 500 repetitions for each data is applied. In each of the 500 runs the one dataset is repartitioned into two samples which are then averaged

to estimate the parameter of interest $\tilde{\theta}$. Hence, 500 estimates are produced per dataset which are then averaged to obtain the final estimator. Then the dataset is completely newly generated and the process starts at the beginning. Since the treatment effect is held constant among all sets and within for all individuals at a level of 0.2 the datasets are comparable in terms of estimation error. Figure 3.1 shows the performance of the 20 datasets. The comparison value is the median within the boxplots.The dashed line at 0.2 indicates the true median ATE and the other two dashed lines above and below show the highest and lowest estimated value among the 20 datasets. In only 25% the DML estimates the true parameter. The absolute (estimation) error (AE) for this test is between -0.035 and +0.052. However, if the medians over the 20 datasets would be averaged again, the final parameter value would be 0.197.

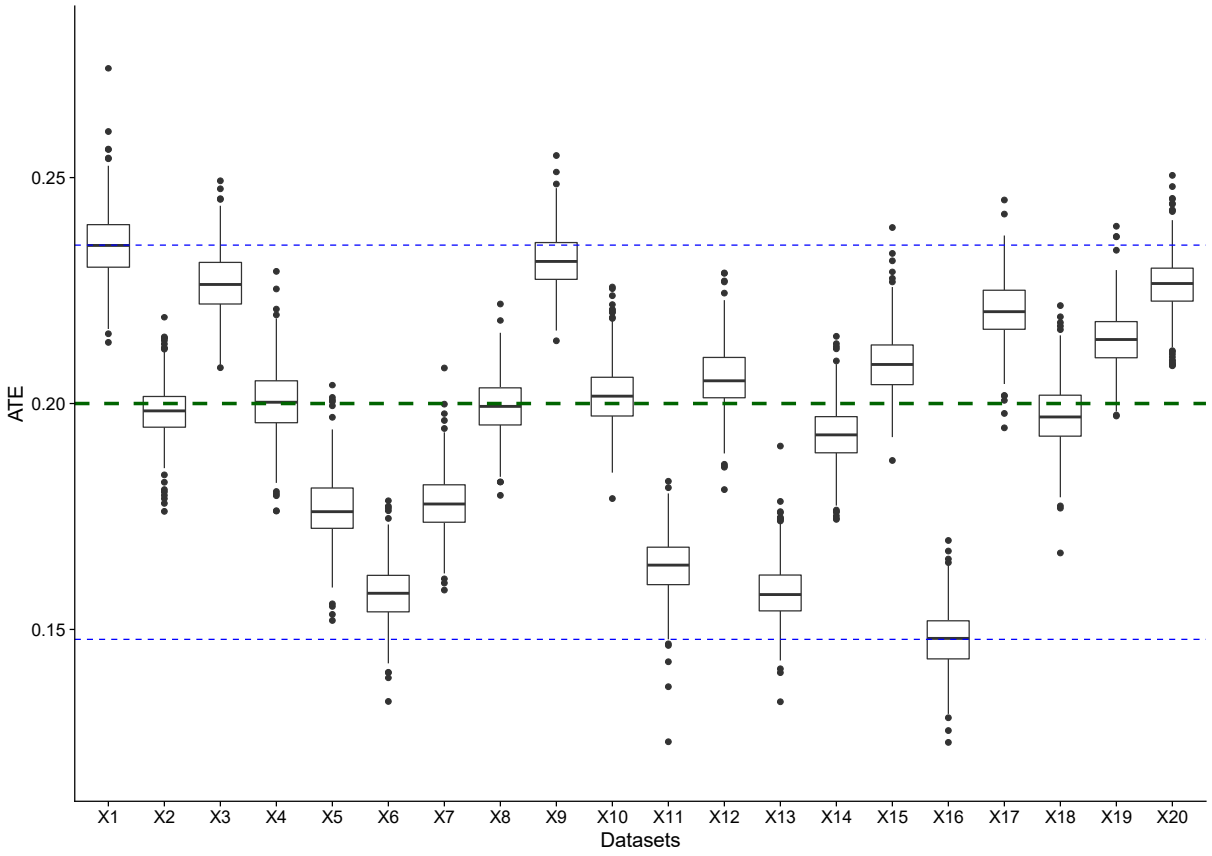


Figure 3.1: Boxplots for 20 independent datasets with 500 repetitions each. Dashed lines show the maximum and minimum estimated parameter value, respectively. True ATE at 0.2 (dashed line in the middle). N = 5000.

This is similar to what Chernozhukov et al. (2018a) do in their simulation. They generate new data for every number of the M repetitions. The result is a point estimate for M datasets. Even if the extreme points are taken, given in this example of 0.125 and 0.276 shown in A.4, their mean would be close to the true parameter of 0.2. However, in reality, there is only one dataset available. It should not be the case that the structure could not be identified and that the ML methods lack in estimating the nuisance parameter such that the estimation of the parameter of interest is biased. It needs to be stated that the idea of using different datasets and average their estimator is similar to Monte Carlo

simulations to learn the distribution of the functions. This reduces the bias from small samples and is computational faster than using only one dataset with an increased amount of observations. To demonstrate this behaviour, the simulation is rerun under the same conditions but instead of N = 5000, 50.000 observations are generated. Now the AE has a range from -0.011 to +0.026. This is a decrease of approx. 58%. Figure 3.2 shows the decrease in deviation from the true parameter. Due to computational reasons only 10 datasets are used. The assumption is that even if the generating process is repeated there would not be an increase in variance hence that this selection is unbiased. Not only is the between bias decreasing but also the within variance of the distribution as they are more concentrated. For completeness a test for group heterogeneity in means by applying an ANOVA analysis is done. The assumption of homoscedasticity based on a Lavene test with median centers to control for outliers is however partially violated. The results confirm that for both simulations the groups still differ in their means.

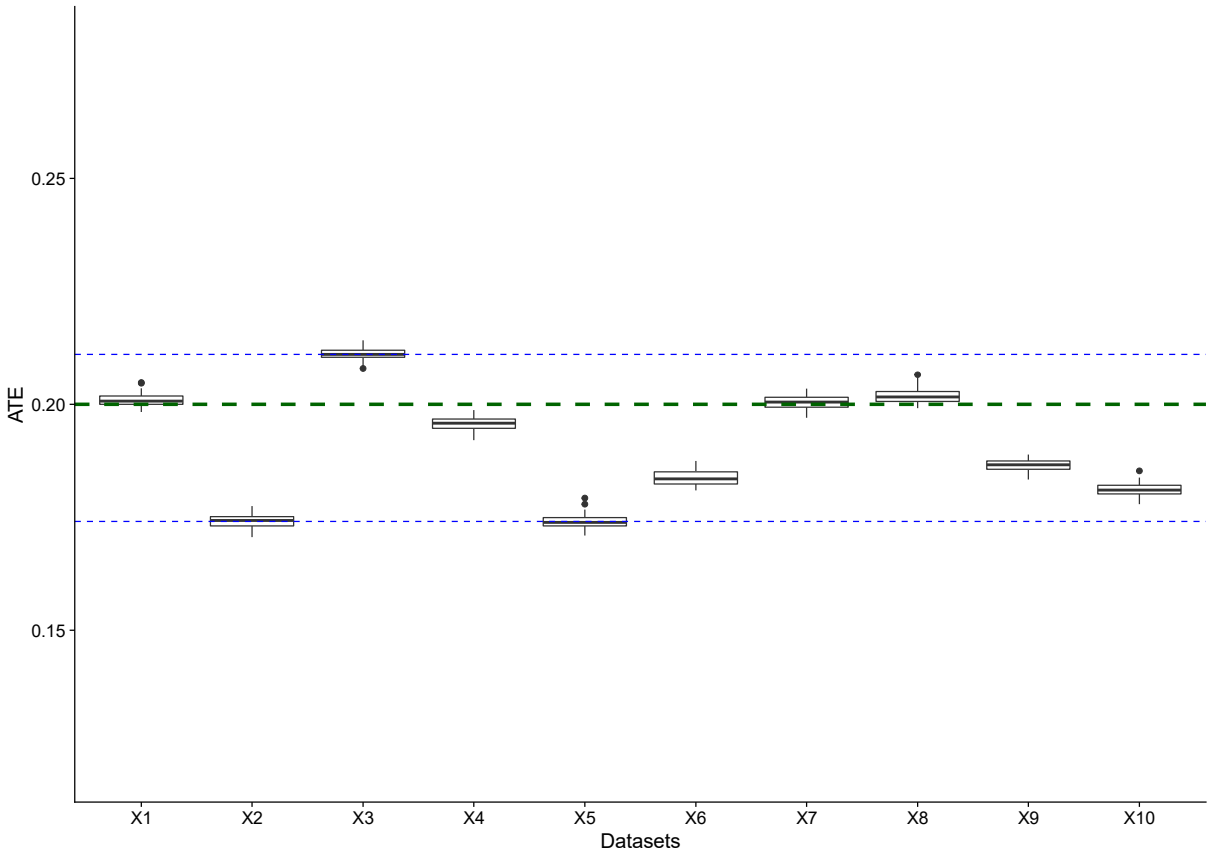


Figure 3.2: Boxplots for 10 independent datasets with 500 repetitions each. Dashed lines show the maximum and minimum estimated parameter value, respectively. True ATE at 0.2 (dashed line in the middle). N = 50.000.

Their paper refers to estimate a consistent "low-dimensional parameter of interest" (Chernozhukov et al., 2018a). To understand what low-dimensional means, the algorithm is applied to three datasets that differ in terms of the treatment effect parameter. The distribution is either constant (θ_0 = 0.2), binary ($\theta_0 \in \{0.1, 0.3\}$) or of continuous level ($\theta_0 \in [0.1, 0.3]$). The options are the same as described in section 2.

To investigate the performance of the DML algorithm it is assumed that there is a fixed dataset where the covariates, the functions $g(X)$ and $m(X)$ are being held constant. This ensures that there is no uncertainty in getting a biased dataset for the different estimations of the parameter. Only the treatment effect and hence the outcome (Y) differs. The ATE varies between being constant ($\theta_0 = 0.2$), binary, with $E[\theta_0^b] = 0.22$ and continuous, with $E[\theta_0^c] = 0.197$. The motivation in this simulation is to identify if and when the DML algorithm cannot handle the nonlinearity of the parameter. The standard error ($\hat{\sigma}_{se}$) for the average treatment effect is given by $\hat{\sigma}/\sqrt{N}$, with

$$\hat{\sigma}^2 = \left(\frac{1}{N}\sum_{i=1}^{N}\hat{V}_i^2\right)^{-2}\frac{1}{N}\sum_{i=1}^{N}\hat{V}_i^2\left(Y_i - D_i\tilde{\theta}_0 - g_0(X)\right). \tag{21}$$

Due to the repetition of the algorithm again a number of M estimators and standard errors is produced. Two measurements to evaluate the validity are provided. First, the calculation of equivalent root median squared errors (22) and second, median standard errors (23) for each partition $j = 1...M$. The median adjustment in 22 incorporates the potential bias due to the sample-splitting for each repetition. Using the adjusted standard error results in more conservative inference as it takes the uncertainty into account. Moreover, it is estimated using the median $\hat{\theta}$ which, as Chernozhukov et al. (2018a) describe, accounts for outliers and is therefore more robust.

$$\hat{\sigma}_{se}^{adj} = median\left\{\sqrt{\hat{\sigma}_{se,j}^2 + \left(\tilde{\theta}_j - \tilde{\theta}_{median}\right)^2}\right\}_{j=1}^{S}, \tag{22}$$

$$\hat{\sigma}_{se}^{median} = median\left\{\hat{\sigma}_{se,j}\right\}_{j=1}^{S}. \tag{23}$$

The corresponding 95% confidence intervals are calculated using the sample split adjusted standard error from 22 and the assumption that the estimator is approximately normally distributed. It could be shown that, given a specific dataset, the accuracy of the estimator heavily depends on the number of observations and also on the complexity of the model. To make sure that the statements about validity regarding dimensionality of the parameter do not appear given a specific distribution of the covariates and the outcome, a Monte Carlo simulation with 1000 iterations is applied. Hence, new data is generated for every run, to learn the distribution of the baseline effect (g_X). The results shown in Table 3.1 are median averages of the 1000 runs. For the constant and binary case, the estimator is unbiased. In the continuous case the DML method underestimates the true treatment effect. However, all CI's incorporate the true parameter at a 95% level. These results are interesting since the authors state that they estimate a low-dimensional parameter but do not comment how they define this exactly nor do they explain to what extent the DML method can capture the distribution of the treatment effect. The results in these simulations show that a binary case does work in the same way as a constant parameter value work. If the heterogeneity is however continuous, the estimation differs.

Regarding further evaluation whether there are some observable indicators why the estimation differs the nuisance functions are investigated. When the machine learning method is a random forest the mean of squared residuals (MSE) and the pseudo R^2 (explained variance) can be calculated in terms of the output Y. A result is that the MSE does not differ between datasets that differ in the estimation of the parameter. The explained variance however does. Since the latter is calculated as $1 - (MSE/Var(Y))$ it has to be the variance of Y that might be too small in some cases. Comparing all

Table 3.1: M=500 repetitions. Treatment effect differs between constant ($\theta_0 = 0.2$), binary = $E[\theta_0^b] = 0.22$ and continuous = $E[\theta_0^c] = 0.197$. Median ATE and standard errors, with SE(median) as in equation 22 and SE as in 23 are reported. The CI is given for a 95% interval.

	constant	binary	continuous
Median ATE	0.195	0.217	0.144
SE(median)	0.041	0.042	0.039
SE	0.035	0.035	0.035
CI upper	0.276	0.298	0.268
CI lower	0.114	0.136	0.019

20 datasets the assumption that the R-squared is an indicator for estimation bias does not hold. There are datasets with a small explained variance that perform well and vice versa. One point, however, can be made for the classification part i.e. the propensity score estimation. The prediction accuracy stays almost constant for all datasets.

Given a dataset with random treatment assignment, N = 5000 observations and a constant treatment effect (with ATE = 0.20) it can be shown that the bias can be reduced by selecting a best ML method used to estimate the nuisance functions. Therefore, five different methods are chosen for both estimations (regression and classification) as well as a method called "best", in which different methods are used to estimate each function as suggested by Chernozhukov et al. (2018a). Table 3.2 shows the comparison in terms of ATE and the two standard errors. The highest AE of 0.0202 is produced by the neural net and the lowest with -0.0057 by the gradient boosting. The best method has an AE of -0,0068 which is close enough to say that the estimator is unbiased. Even if the parameter differs among the ML methods, the standard errors are almost equal. In fact, the neural net has, besides the tree method, the smallest error.

Table 3.2: Median ATE is estimated over 500 random splits. True theta = 0.2 for all 5000 observations. Default tuning parameter used for each algorithm. Method best consists of gradient boosting for regression and random forest for classification.

	RLasso	Boosting	Neural Net	Trees	Random Forest	best
Median ATE	0.2173	0.2057	0.1798	0.1885	0.2106	0.2068
se(median)	(0.0392)	(0.0392)	(0.0387)	(0.0368)	(0.0383)	(0.0391)
se	(0.0391)	(0.0389)	(0.0365)	(0.0361)	(0.038)	(0.0388)

3.2 Generic Machine Learning for Group ATE (GATES)

Chernozhukov et al. (2018b) notice that, *"in high dimensional settings, absent strong assumptions, generic ML tools may not even produce consistent estimates of the conditional average treatment effect (CATE)."* To provide valid estimation and inference they, therefore, focus on features of the CATE. One of these features is the **Sorted Group Average Treatment Effect (GATES)**. The idea is to find groups of observations depending on

the estimated treatment effect heterogeneity. This section, first of all, describes how this approach can be implemented. Secondly, different ML algorithms are evaluated to choose the best one for further calculations. Thirdly, besides the ATE, different standard errors are calculated while applying this method on a binary and continuous theta distribution. In a final step, this method is investigated on a constant treatment effect to evaluate if it can detect homogeneity.

$$E[\theta(X)|G_k], \qquad G_k : k^{th} \quad \text{n-tile of estimated} \quad \hat{\theta}(X) \tag{24}$$

The groups are ex-post defined by the predicted treatment effect in the first stage. These functions are trained by ML algorithms and might be biased. The authors state that it might not be true that these groups contain the observations with the true CATE. If, however the treatment effect for the groups would be consistent, it asymptotically holds that

$$E[\theta(X)|G_1] \leqslant E[\theta(X)|G_2] \leqslant ... \leqslant E[\theta(X)|G_k], \tag{25}$$

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

$$Y = \alpha' X_1 + \sum_{k=1}^{K} \gamma_k \times (D - p(X)) \times 1(G_k) + \nu, \qquad E[w(X)\nu W] = 0, \tag{26}$$

where $X_1 = [1, B(X)]$ and $W = (X_1', W_2')$ with $W_2 = (\{(D - p(X))1(G_k)\}_{k=1}^{K})'$. $B(X) = E[Y|D = 0, X]$ is the baseline function without treatment and the projection $S_0(X) = E[Y|D = 1, X] - E[Y|D = 0, X]$ is the treatment effect (Chernozhukov et al., 2018b). The parameter γ is the sorted group average treatment effect for each group $1...K$. This is the main identification result:

$$\gamma = (\gamma)_{k=1}^{K} = (E[s_0(X)|G_k])_{k=1}^{K} \tag{27}$$

The sorted group average treatment effect can be implemented with the following steps:

Algorithm for GATES

1. Split Data in $k = 2$ samples: I^a and I with $I^a \cup I$
2. Train $Y_i = \hat{g}_0(X_i, D) + \hat{U}_i$, with $i \in I^a$
3. Train $D_i = \hat{m}_0(X_i) + \hat{V}_i$, with $i \in I^a$
4. Estimate conditional baseline function (a) and treatment function (b)
   (a) $\hat{Y}_i = \hat{g}_0(X_i, D = 0)$, with $i \in I$
   (b) $\hat{Y}_i = \hat{g}_0(X_i, D = 1)$, with $i \in I$
5. Divide observations into groups according to the difference from (a) and (b)
6. Estimate the propensity score: $\hat{D}_i = \hat{m}_0(X_i)$, with $i \in I$
7. Calculate propensity score offset: $\hat{V}_i = D_i - \hat{D}_i$, with $i \in I$
8. Estimate the GATES parameter by weighted OLS using I
9. Repeat Steps 1 to 9 for a large number M (e.g. M = 100) and report medians at the end

Since the ML methods used to predict (a) and (b) can be biased, the authors suggest to use a variety of methods and choose the best one. They propose three different evaluation criteria. The first is to choose the best ML methods based on the auxiliary sample. This is the one which minimizes the errors in the weighted prediction of Y on B and $(D - p(Z))(S - E[S]))$. The second and third option uses the main sample to either maximize the R^2 in the regression for the baseline functions or, which is the third option, maximize the variation between the groups of the GATES method. Since these options are independent, option two and three are used in this paper and the best ML method is the one which maximizes both conditions. Similar to the paper four different algorithms are considered: elastic net, gradient boosting, neural network and a random forest.

Furthermore, different tuning parameters are used. Some that are independent of the methods like repeated cross-validation and specific ones to the algorithm like e.g. maximum iterations and the number of trees. Table 3.3 shows the ML methods based on the best linear prediction (BLP) and the GATES variation. The elastic net and the neural net outperform the other two methods in terms of the BLP. However, focusing on the variation between groups the choice would be to use the random forest instead of the elastic net. The authors do not describe such a case in the paper but a recommendation would be to average both options. In that case, it would still be the last two ML methods to use for the second stage. At a recent workshop from NBER called "SI 2018 Development Economics" Esther Duflo, one of the authors of (Chernozhukov et al., 2018b) published the R-code used in this paper. Investigating the code, it seems that they choose the best ML method based on the best linear predictor.

As in the DML approach, the use of sample-splitting is crucial to overcome the regularization bias as well as to avoid overfitting. However, it comes with a price. Due to the introduced uncertainty from data splitting the authors prove that under repeated partition the confidence interval will cover the true parameter in $(1 - 2\alpha)\%$ of the time. If

Table 3.3: Best linear predictor option and best maximizing variation for GATES over 500 splits.

	Elastic Net	Boosting	Neural Net	Random Forest
Best BLP	0.096	0.077	0.089	0.087
Best GATES	0.114	0.082	0.179	0.142

α is set to 0.05, which is the default in this paper, the CI would be at a 90% probability instead of the 95%.

Table 3.4 shows the conditional average treatment effect and the heterogeneity loading (HET) for the chosen methods. In parentheses the median adjusted 90% confidence intervals are reported. Since the last step in the GATES algorithm is a weighted linear regression, adjusted p-values are also reported in brackets. The true ATE for this dataset with N = 5000 observations and binary theta is 0.212. The estimated ATE for all methods is biased in the same way as for the DML approach.

Table 3.4: Medians over 500 splits. 90% confidence intervals in parenthesis. P-values for the hypothesis that the parameter is equal to zero in brackets. Table adapted from (Chernozhukov et al., 2018b).

Neural Net		Random Forest	
ATE	HET	ATE	HET
0.249	0.464	0.250	0.503
(0.165,0.333)	(0.236,0.687)	(0.164,0.336)	(0.194,0.805)
[0.000]	[0.000]	[0.000]	[0.003]

A more detailed look provides Figure 3.3 in where all four methods are plotted for the five quintiles. Since there are only two levels of heterogeneity by design, 0.1 and 0.3, it can be shown that the predicted groups are not perfectly separated. Since it is not known ex-ante how many groups there would be, quintiles are considered where in the best case at least three of them would have the same ATE. They all predict the least affected group near zero and slightly overestimate the most affected group with values over 0.4. In line with the paper is that there is heterogeneity in the treatment effect parameter and also that the groups monotone increase. The authors show that this results asymptotically into consistency. The ATE (blue dashed line in the middle) is calculated as an average over all groups.

As a next step, some results for the continuous case are presented. The theta distribution is shown in Figure A.5. This is where the treatment effect takes any values between 0.1 and 0.3. The distribution shows a higher density in the lower region while less observations take values near 0.3. This is consistent with the output from the GATES method which is shown in Table 3.5. The average treatment effect is calculated using the median method to average over all repetitions. The first three groups show a median treatment effect between 0.092 and 0.124. Group four has a higher effect and group five estimates that there are some observations at 0.38. The 90% confidence intervals for each group are reported using estimated standard errors. They also confirm that there is heterogeneity in the data even if the CI for each group is quite large. For the first three they even take negative values.

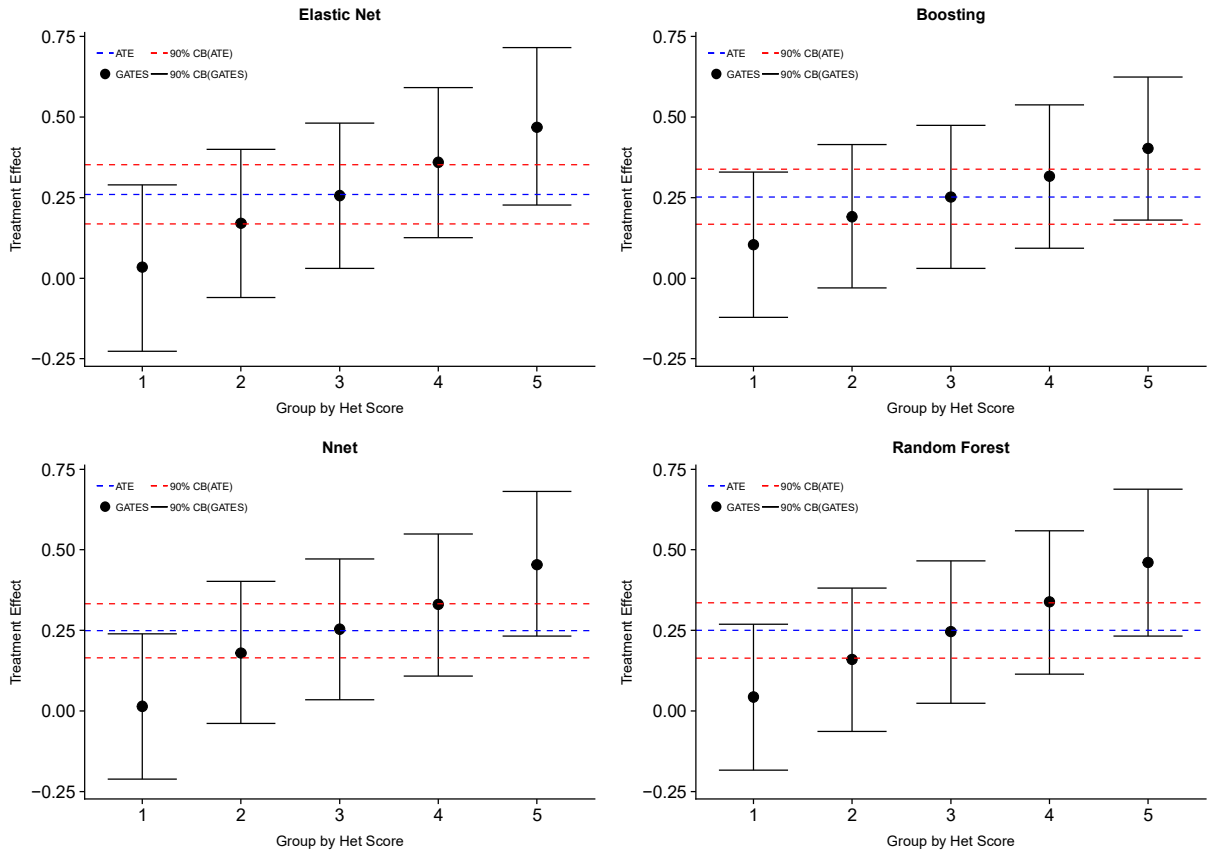


Figure 3.3: GATES for binary theta. K=5 groups defined by quintiles of $S(X)$ over 500 random splits. N=5000. Figure style adapted from (Chernozhukov et al., 2018b).

Since the samples are always different (in observations) this approach cannot assign individuals to a specific group. The groups always change with each repetition (splitting). However, the aim is not to make inference on the individual level rather to find valid estimators for features, namely groups, based on the size of the difference in ATE for each group (HET score). The effect is therefore always an average over all splits. It is still a promising approach since it is valid in high dimensional settings. There are no heavy assumptions and it is applicable to a broad bench of ML methods.

Regarding testing for a homogeneous treatment effect, all groups are supposed to have the same effect. Therefore, a dataset with constant theta is used to estimate a parameter for five groups. The two best methods, namely the elastic net and the random forest, predict the ATE being 0.238 and 0.245, respectively. Table B.1 shows the results together with the CI's and standard errors. They are chosen based on the BLP and HET score in which they both outperform the other two considered models. Figure 3.4 shows the ATE for every group and model. They show the same heterogeneity as in the binary and continuous case. Especially the elastic net estimates a huge difference between the first and the last group. These results would conclude that there is heterogeneity, at least for two groups. Further research would be needed to investigate why this method does not detect the homogeneity.

Table 3.5: Numeric GATES output. 500 random splits used. Actual treatment effect is of form continuous with $\theta \in (0.1, 0.3)$

Quintiles	ATE	lower 90% CI	upper 90% CI
Group1	0.092	-0.101	0.285
Group2	0.102	-0.081	0.285
Group3	0.124	-0.055	0.305
Group4	0.187	0.000	0.376
Group5	0.384	0.159	0.610

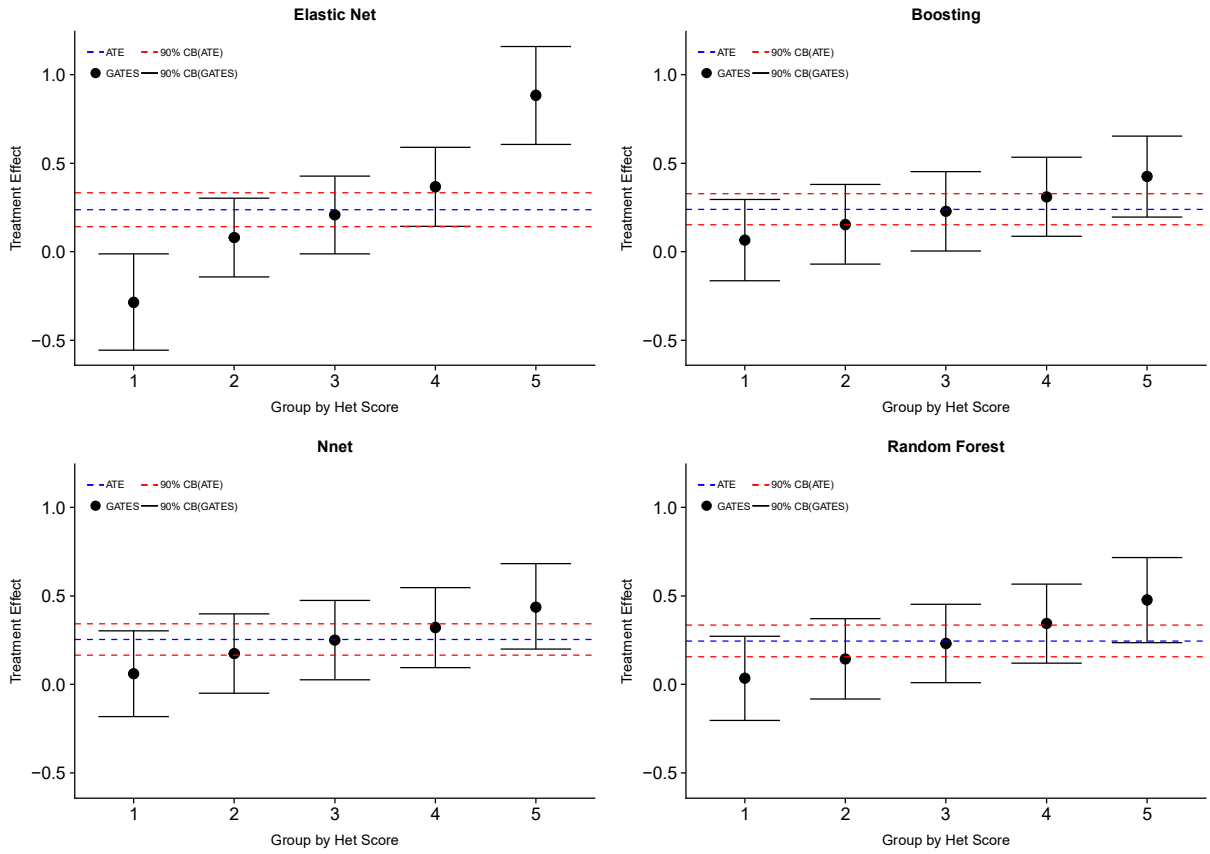


Figure 3.4: GATES for constant theta. K=5 groups defined by quintiles of $S(X)$ over 500 random splits. N=5000. Figure style adapted from (Chernozhukov et al., 2018b).

3.3 Causal Trees (CT)

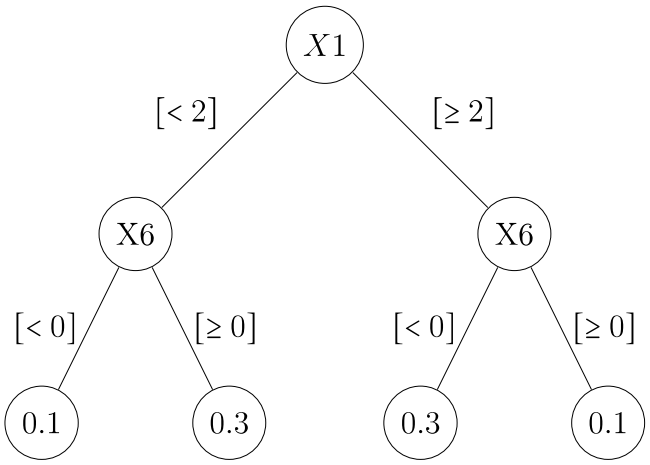
Moving one step further, from group heterogeneity towards heterogeneity with respect to observed covariates, the **causal tree** method by Athey and Imbens (2016) is an easily applicable approach. It is especially useful in repeated campaigns with the same treatment and the same covariates. The treatment effect mapped from the observed characteristics would then be comparable. Ceteris paribus, the results would answer questions like whom to treat and how heterogeneous the treatment effect is. The trees are built as follows:

Algorithm for Causal Trees

1. Split Data in $k = 2$ samples: I^a and I with $I^a \cup I$
2. Use I^a and I to grow a tree via recursive partitioning with (X_i, D_i, Y_i) where $i \in N$ but $Y_i \notin I$
3. Choose splits by maximizing the variance of $\hat{\theta}(X_i)$ for $i \in I^a$
4. Estimate the treatment effect for each leaf using $Y(X_i|D = 1) - Y(X_i|D = 0)$ for $i \in I$.

Step 4 refers to their so-called **honest tree** splitting, where the idea is to use different samples. One sample to construct the partition and another for estimating the treatment effect. As a result, the confidence intervals for the estimated treatment effects have nominal coverage.

In a simple scenario, a new option for theta is generated where it is only dependent on two covariates, namely:



As in the binary option, theta takes values of either 0.1 or 0.3. The distribution has almost the same density for the two values which results in balanced data (true ATE = 0.201). The dependence and the strength are verified in Table B.3. An easy linear (or probit) model does not find a significant connection between $X1$ and theta but for $X6$ and the treatment effect. Now, the causal tree algorithm is used to find the structure above. Since the CT model does not take a propensity weighting into account the treatment assignment for the used data is completely random. Therefore, all leaves have equal weights of 0.5. The result of the causal tree model is a tree with 187 splits based on pruning with respect to the lowest error function. To detect the structure and since the error does not decrease significantly, a plot of the tree with only 15 splits is shown in Figure 3.5. The second split on the left is for $X6$ if it is smaller than -0.044, and accounts for 22% of the observations. The treatment effect is then either -0.22 or +0.23. For the remaining 78%, there is no split for $X6$ but one for $X1$ which accounts for 46%. This is interesting since first, this node includes the highest percentage of observations in one single leaf (41%) whit a predicted treatment effect is 0.26. Second, the dependence of $X1$ was not found from a linear model but here accounts for almost half of the sample. However, the true

correlation includes both variables which the model fails to detect. The ATE over all splits is biased with a value of 0.158.

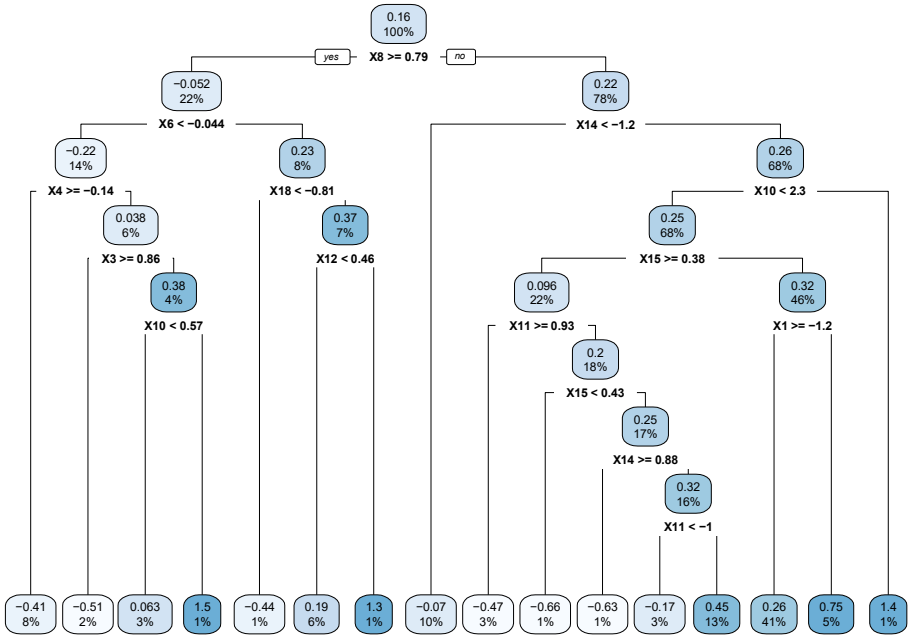


Figure 3.5: Binary theta depending on X6 and X1. Treatment assignment is random. N = 5000 observations.

It is worth noting that one should not fall into the trap of interpreting the output of the causal tree as causal effects. In this case, it can be seen that only $X6$ and $X1$ are the driving force for a different treatment effect. In reality, it could be the case, that one of these two variables are dependent on a third one which is not plotted in the tree. Hence the output of causal trees is not really a representation of the true CATE function. It only finds some heterogeneity on average which is useful for relevant heterogeneity investigation but not to state which covariate affects the treatment effect in which way.

3.4 Comparison of Models

In this section, the performance of the models among each other is evaluated. The idea is to check if there is one model that performs best and if there is some structure if the estimation is biased. The one common target parameter that can be calculated with every model is the average treatment effect. The DML estimates this per default. For the GATES method, the estimated ATE from the groups are simply averaged. The purpose of the CT model is to estimate conditional average treatment scores. The data can be used to predict a treatment effect for every single individual (i.e. the leaf CATE) which is then averaged to get the final ATE. These assumptions guarantee that the models can be compared among each other.

Table 3.6 presents the structure and parameters for different datasets. Scenario 1-3 equals 4-6, respectively. The difference is the treatment assignment which is dependent on the covariates in the former and random in the latter cases. The same holds for scenarios 7-8 and 10-12. All different theta distributions are considered as well as different

amounts of covariates. Since the true ATE differs between the options the resulting estimates are centered for every distribution with the true mean. This makes the scenarios comparable. Each dataset is held constant among the repetitions which are set to 100 due to computational reasons. The same evaluation is done with 60 covariates for three datasets. The results do not differ from the setting with 200 covariates which is why they are not shown here.

Table 3.6: Dataset variation for simulation.

Scenarios	1,4	2,5	3,6	7,10	8,11	9,12
n	5000	5000	5000	5000	5000	5000
k	20	20	20	200	200	200
t_0	constant	binary	continuous	constant	binary	continuous

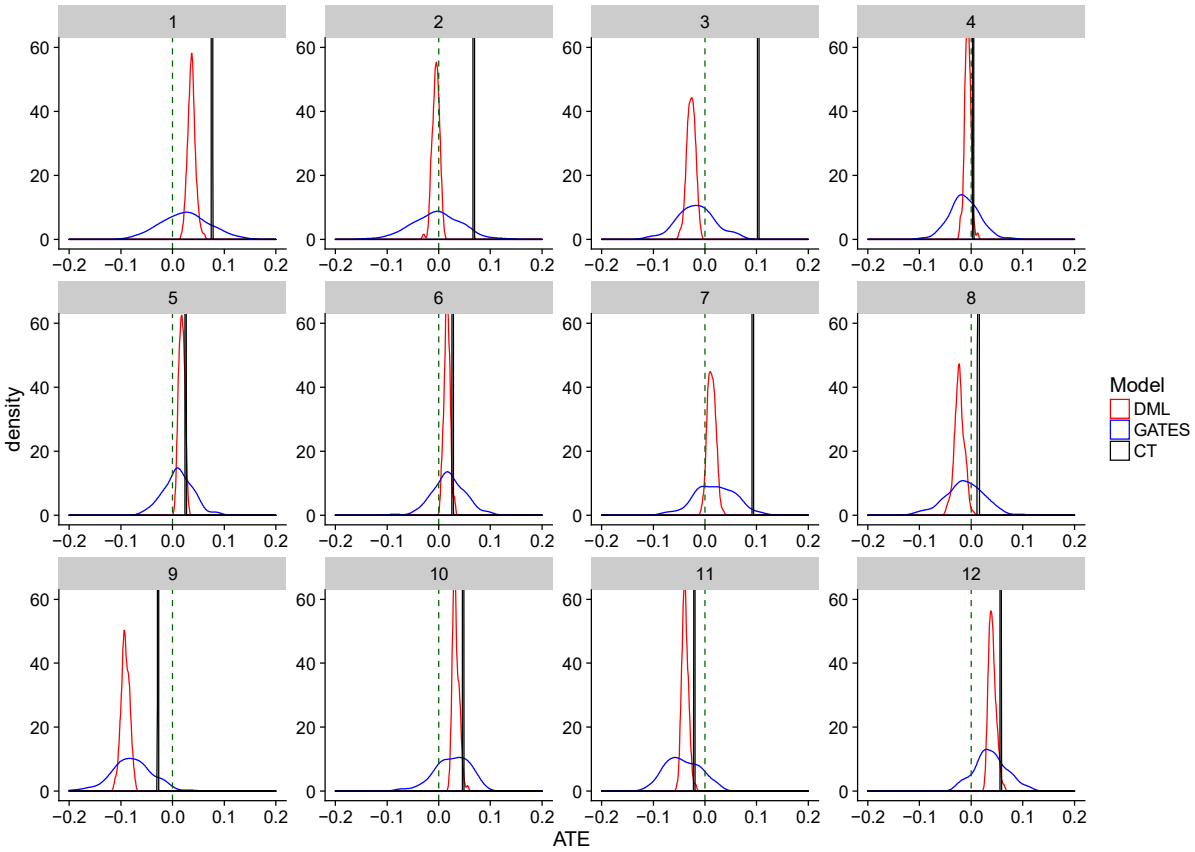


Figure 3.6: Model comparison between DML, GATES and CT for different datasets. True treatment effect (dashed green line) centered at 0.0 for all theta distributions. Nearly point distribution indicates CL. Lowest density, highest variance is the GATES method. DML nearly normal distributed.

Comparing all scenarios, the GATES model outperforms all others in terms of estimation accuracy for the ATE. The second-best model is the DML. For both models it holds that they always under- or overestimate in the same direction. This can be seen in scenario 1 or 9. This differs from the estimation for the CT model. While in scenario 3 DML and

GATES slightly underestimate the ATE, the CT overestimates the true mean with an error of 0.1. This is also the highest MAE in this simulation. Scenario 1, 2 and 3 have a non-random treatment assignment which is probably the reason why the CT model overestimates the true parameter. This is in line with scenario 4 to 6 in which D is random and the error from the CT model decreases. Looking at the simulations with 200 covariates this rule does not hold anymore. In scenario 8 the parameter estimated from CT is closer to the true one than in scenario 12. Looking at the distributions it can be shown that the CT is nearly a point estimate over all repetitions. The DML and the GATES method are quite normal distributed whereas the latter shows a higher variance. Interesting is also the distribution from the DLM which has a decreasing variance if the treatment assignment is random. While this does not affect the biasedness, it is in line with consistency since it could be assumed that the nuisance function shows a higher accuracy for the random assignment option. While in the double machine learning section the prediction accuracy for the propensity score is found to be constant among the datasets it is not investigated if this would change given different treatment assignment.

4 Discussion

In this paper, three novel methods for the estimation of a causal parameter are revisited. All methods use ML algorithms to increase the accuracy. While the DML and the GATES approach directly allow the use of different ML methods which can be compared to each other, the causal trees only use some sort of k-nearest neighbor for the estimation. Extensions to the latter one would be to use a random forest or a gradient forest. A recommendation on what method to use cannot be given in terms of accuracy. However, a suggestion would be that researchers start with the generic machine learning method to detect if there is heterogeneity in the treatment effect. If this is not the case then the DML would be a good choice to estimate only the average treatment effect. It is the only approach that estimates a consistent and asymptotically unbiased as well as normal distributed parameter. The consistent estimation of standard error and hence confidence intervals at a $1 - \alpha$ probability is an advantage over the GATES method that, due to the doubling factor, only produces a CI at $1 - 2\alpha$. The same nominal coverage of only 90% given alpha = 0.05 can be achieved by the causal tree approach.

Using the GATES method to prove if the treatment effect is homogeneous does however not perform appropriately. It can be shown that in simulations with a constant theta of 0.2 for all individuals this method does still find heterogeneous effects among the defined groups. Ideally, it should be the other way around. If the heterogeneity is rather small in its range, the method should not detect different effects. In this scenario, it could be argued that only big differences between groups would be detected. Since it can be shown that for heterogeneous effects the DML method underestimates the true parameter, it is more challenging to determine if the DML can be applied in a second stage to estimate the ATE without bias. This is however only problematic if the distribution of the treatment effect is high-dimensional. In the case of a binary theta the ATE estimated via the DML is still unbiased.

The recommendation to use the DML when the parameter of interest is the ATE is based on the fact that for the GATES approach several tuning parameters have been considered in the above-mentioned simulations to find the best ML method. While for the DML it can be shown that tuning leads to a decrease in absolute estimation error it was not done in such an expended way. For further research the same tuning parameters as

well as an algorithm selection needs to be applied on the DML to guarantee an honest model comparison.

Furthermore, it would be interesting to determine if there are differences in the estimation if the amount of groups changes. Another approach regarding heterogeneity in groups, that would however leave the framework of this paper, could be to only use the observations that are "stable" in terms of group affiliation, within the repetitions. These observations are the ones that have a probability for group x higher than some predefined threshold. Even if this would result in only using a subsample for the estimation, it could be the case that the heterogeneity is more accurate.

Regarding the selection of the ML algorithms within a model there is no clear answer for practitioners. For the DML approach no observed characteristic that would ensure a best-chosen method is found. However, in this paper, the evaluation of the pseudo r-squared and the MSE is only based on different datasets that conclude that there is no indicator on when an algorithm performs better in terms of the parametric regression. Furthermore, only the random forest algorithm is used for this evaluation. Different ML methods estimate different parameters which concludes that there are some differences in estimation accuracy. A next step would be to find other observable evaluation criteria that are correlated with the estimation accuracy of the parameter of interest. Especially the GATES model shows that these metrics not necessarily imply the latter. In the simulation with a constant parameter, the best ML method is the elastic net based on two different evaluation criteria. However, the model also estimates a heterogeneity between least and most affected group of 1.18 while the true difference is zero since the effect is 0.2 for all treated individuals. In the comparison of the presented model on the same data, the GATES method as well as the DML outperform the CT in terms of prediction accuracy for the ATE. The error however does depend on the specification of the data.

All the examples conclude that further research is necessary when trying to consistently estimate causal parameters given complex and high-dimensional data.

References

Athey, S. (2017). The impact of machine learning on economics. In *Economics of Artificial Intelligence*. University of Chicago Press.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

Athey, S. and Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32.

Athey, S., Tibshirani, J., Wager, S., et al. (2016). Solving heterogeneous estimating equations with gradient forests. https://arxiv.org/abs/1610.01271.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2):29–50.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2017a). Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65.

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

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

Chernozhukov, V., Goldman, M., Semenova, V., and Taddy, M. (2017b). Orthogonal machine learning for demand estimation: High dimensional causal inference in dynamic panels. https://arxiv.org/abs/1712.09988.

Chernozhukov, V., Hansen, C., and Spindler, M. (2015). Valid post-selection and post-regularization inference: An elementary, general approach. *Annu. Rev. Econ.*, 7(1):649–688.

Levit, B. Y. (1976). On the efficiency of a class of non-parametric estimates. *Theory of Probability & Its Applications*, 20(4):723–740.

Low, Y. S., Gallego, B., and Shah, N. H. (2016). Comparing high-dimensional confounder control methods for rapid cohort studies from electronic health records. *Journal of comparative effectiveness research*, 5(2):179–192.

Mackey, L., Syrgkanis, V., and Zadik, I. (2017). Orthogonal machine learning: Power and limitations. https://arxiv.org/abs/1711.00342.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4):403.

Morgan, S. L. and Winship, C. (2015). *Counterfactuals and causal inference*. Cambridge University Press.

Oprescu, M., Syrgkanis, V., and Wu, Z. S. (2018). Orthogonal random forest for heterogeneous treatment effect estimation. https://arxiv.org/abs/1806.03467.

Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T., and Tibshirani, R. (2017). Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in Medicine*, 37(11).

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

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58.

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

Schuler, A., Baiocchi, M., Tibshirani, R., and Shah, N. (2018). A comparison of methods for model selection when estimating individual treatment effects. https://arxiv.org/abs/1804.05146.

Wager, S. and Athey, S. (2017). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, (just-accepted).

Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N., and Gallego, B. (2018). Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Statistics in medicine*.

Wyss, R., Ellis, A. R., Brookhart, M. A., Girman, C. J., Jonsson Funk, M., LoCasale, R., and Stürmer, T. (2014). The role of prediction modeling in propensity score estimation: an evaluation of logistic regression, bcart, and the covariate-balancing propensity score. *American journal of epidemiology*, 180(6):645–655.

A Figures

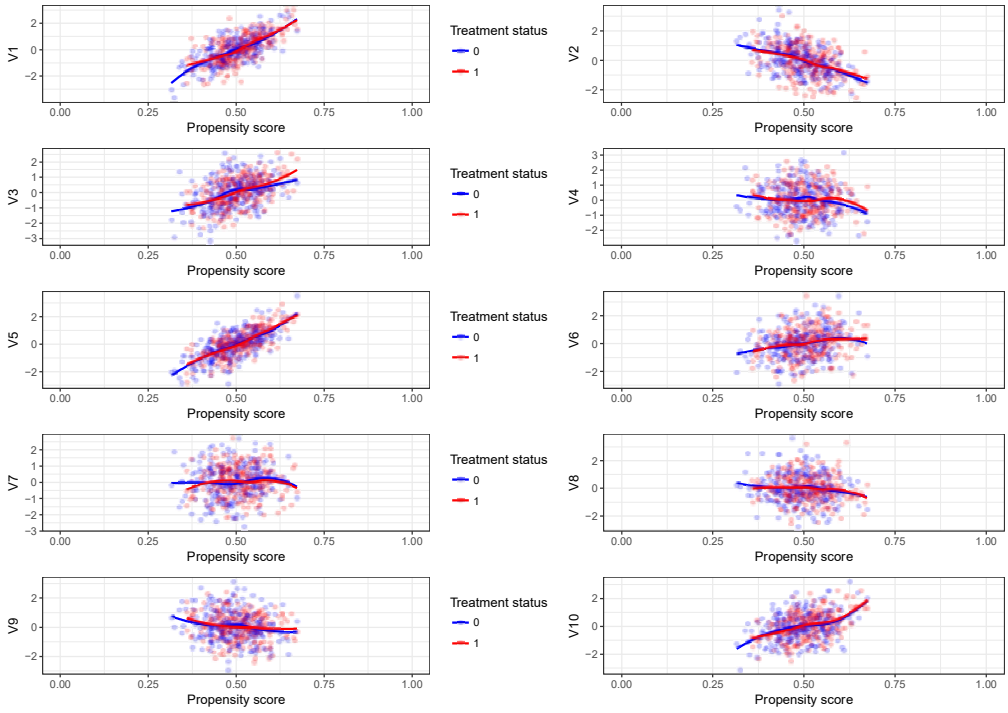


Figure A.1: Propensity score independent of covariates.

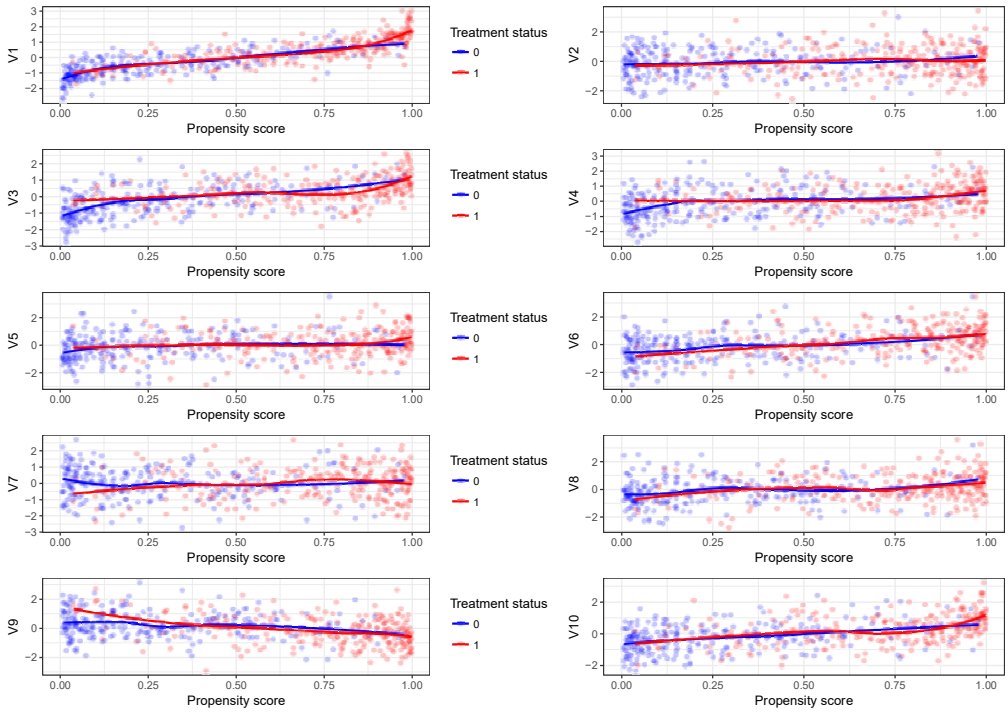


Figure A.2: Propensity score depending on covariates.

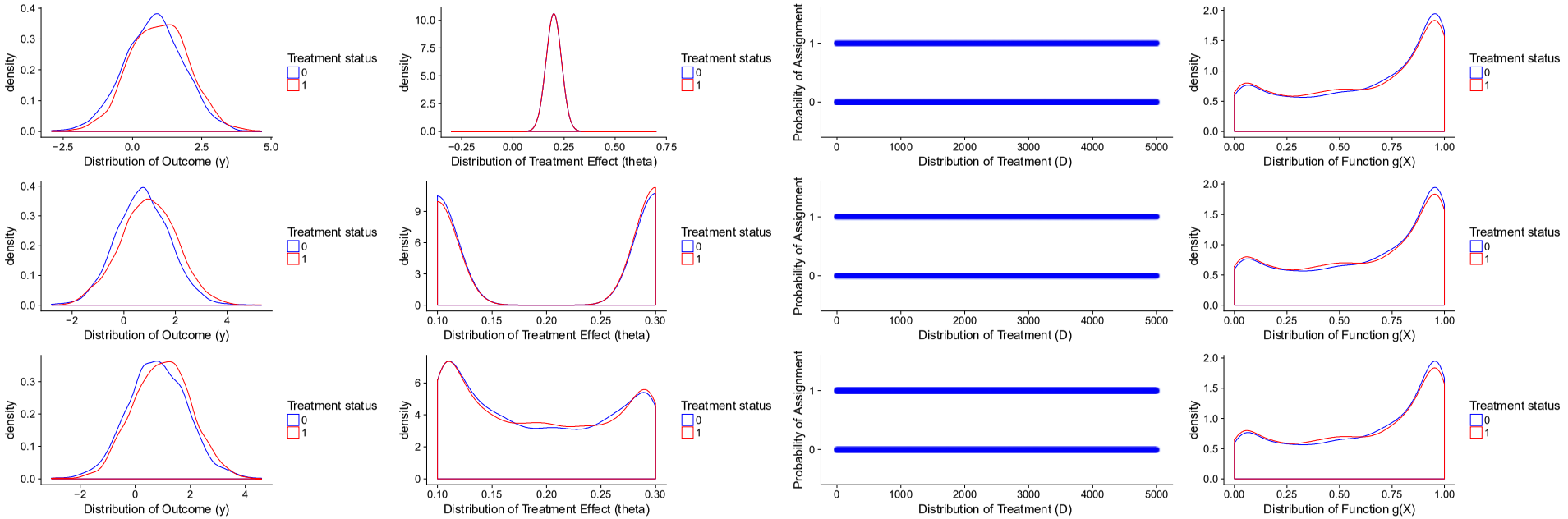


Figure A.3: Shows the distribution of the simulated data. The baseline effect $g(X)$ is held constant. Treatment effect differs between a scalar ($\theta = 0.2$), binary ($\theta \in \{0.1, 0.3\}$) and continuous ($\theta \in [0.1, 0.3]$). The treatment assignment is random in this figure.

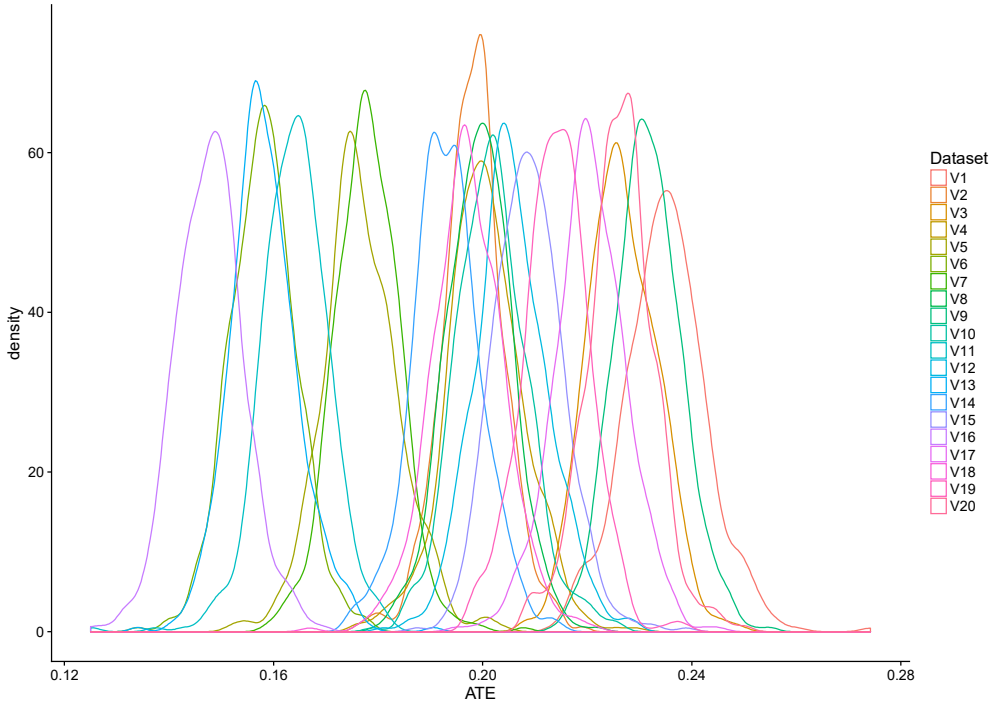


Figure A.4: Distribution of the 20 datasets with 500 repetitions. N = 5000 observations. True ATE at 0.2.

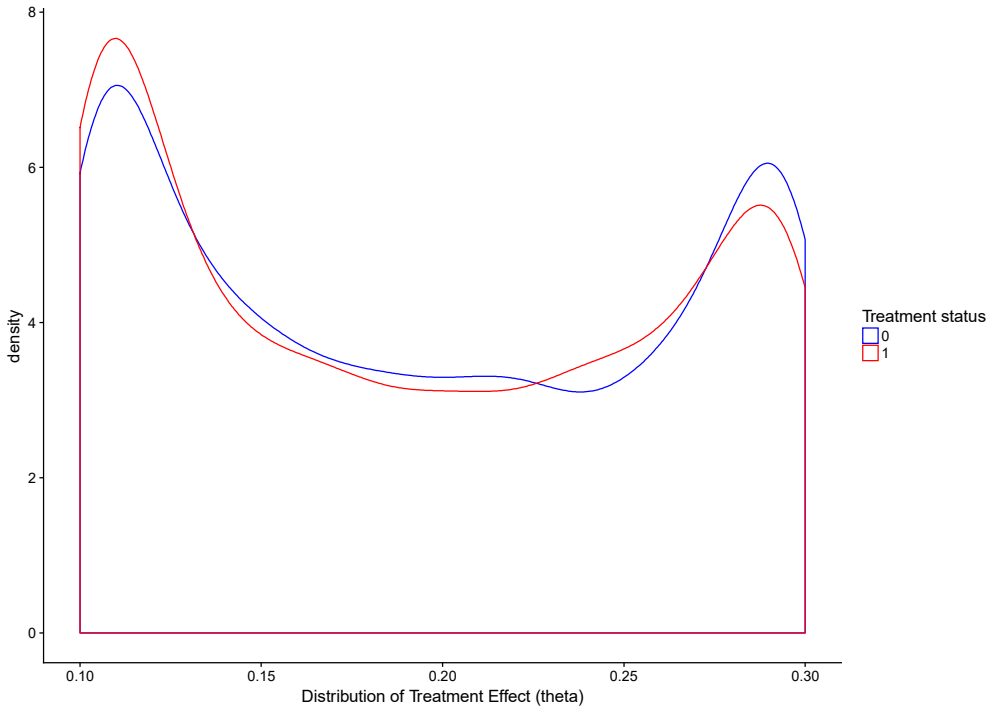


Figure A.5: Distribution of treatment effect for the continuous case. Theta takes any values between 0.1 and 0.3. N = 5000 observations.

B Tables

Table B.1: Medians over 500 splits. 90% confidence intervals in parenthesis. P-values for the hypothesis that the parameter is equal to zero in brackets. Table adapted from (Chernozhukov et al., 2018b).

Elastic Net		Random Forest	
ATE	HET	ATE	HET
0.238	1.434	0.245	0.454
(0.142,0.334)	(1.077,1.807)	(0.156,0.335)	(0.173,0.736)
[0.000]	[0.000]	[0.000]	[0.003]

Table B.2: Most and least affected groups for the two best ML algorithms. Note that the standard error for the least affected group estimated by random forest is too high and therefore this estimate is not significant.

Elastic Net			Random Forest		
Most Affected	Least Affected	Difference	Most Affected	Least Affected	Difference
0.884	-0.286	1.177	0.471	0.047	0.429
(0.650,1.115)	(-0.514,-0.055)	(0.882,1.471)	(0.268,0.678)	(-0.160,0.253)	(0.149,0.708)
[0.000]	[0.031]	[0.000]	[0.000]	[1.000]	[0.005]

Table B.3: Linear model for theta on dependent covariates (X1 and X6).

	Dependent variable:
	theta
z[, 6]	0.077^{***}
	(0.001)
z[, 1]	0.0002
	(0.001)
z[, 6]:z[, 1]	-0.010^{***}
	(0.001)
Constant	0.200^{***}
	(0.001)
Observations	5,000
R^2	0.592
Adjusted R^2	0.592
Residual Std. Error	0.064 (df = 4996)
F Statistic	$2{,}416.195^{***}$ (df = 3; 4996)
Note:	*p<0.1; **p<0.05; ***p<0.01

IRTG 1792 Discussion Paper Series 2018

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

001 "Data Driven Value-at-Risk Forecasting using a SVR-GARCH-KDE Hybrid" by Marius Lux, Wolfgang Karl Härdle and Stefan Lessmann, January 2018.

002 "Nonparametric Variable Selection and Its Application to Additive Models" by Zheng-Hui Feng, Lu Lin, Ruo-Qing Zhu asnd Li-Xing Zhu, January 2018.

003 "Systemic Risk in Global Volatility Spillover Networks: Evidence from Option-implied Volatility Indices " by Zihui Yang and Yinggang Zhou, January 2018.

004 "Pricing Cryptocurrency options: the case of CRIX and Bitcoin" by Cathy YH Chen, Wolfgang Karl Härdle, Ai Jun Hou and Weining Wang, January 2018.

005 "Testing for bubbles in cryptocurrencies with time-varying volatility" by Christian M. Hafner, January 2018.

006 "A Note on Cryptocurrencies and Currency Competition" by Anna Almosova, January 2018.

007 "Knowing me, knowing you: inventor mobility and the formation of technology-oriented alliances" by Stefan Wagner and Martin C. Goossen, February 2018.

008 "A Monetary Model of Blockchain" by Anna Almosova, February 2018.

009 "Deregulated day-ahead electricity markets in Southeast Europe: Price forecasting and comparative structural analysis" by Antanina Hryshchuk, Stefan Lessmann, February 2018.

010 "How Sensitive are Tail-related Risk Measures in a Contamination Neighbourhood?" by Wolfgang Karl Härdle, Chengxiu Ling, February 2018.

011 "How to Measure a Performance of a Collaborative Research Centre" by Alona Zharova, Janine Tellinger-Rice, Wolfgang Karl Härdle, February 2018.

012 "Targeting customers for profit: An ensemble learning framework to support marketing decision making" by Stefan Lessmann, Kristof Coussement, Koen W. De Bock, Johannes Haupt, February 2018.

013 "Improving Crime Count Forecasts Using Twitter and Taxi Data" by Lara Vomfell, Wolfgang Karl Härdle, Stefan Lessmann, February 2018.

014 "Price Discovery on Bitcoin Markets" by Paolo Pagnottoni, Dirk G. Baur, Thomas Dimpfl, March 2018.

015 "Bitcoin is not the New Gold - A Comparison of Volatility, Correlation, and Portfolio Performance" by Tony Klein, Hien Pham Thu, Thomas Walther, March 2018.

016 "Time-varying Limit Order Book Networks" by Wolfgang Karl Härdle, Shi Chen, Chong Liang, Melanie Schienle, April 2018.

017 "Regularization Approach for NetworkModeling of German EnergyMarket" by Shi Chen, Wolfgang Karl Härdle, Brenda López Cabrera, May 2018.

018 "Adaptive Nonparametric Clustering" by Kirill Efimov, Larisa Adamyan, Vladimir Spokoiny, May 2018.

019 "Lasso, knockoff and Gaussian covariates: a comparison" by Laurie Davies, May 2018.

This research was supported by the Deutsche

IRTG 1792 Discussion Paper Series 2018

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

020 "A Regime Shift Model with Nonparametric Switching Mechanism" by Haiqiang Chen, Yingxing Li, Ming Lin and Yanli Zhu, May 2018.

021 "LASSO-Driven Inference in Time and Space" by Victor Chernozhukov, Wolfgang K. Härdle, Chen Huang, Weining Wang, June 2018.

022 "Learning from Errors: The case of monetary and fiscal policy regimes" by Andreas Tryphonides, June 2018.

023 "Textual Sentiment, Option Characteristics, and Stock Return Predictability" by Cathy Yi-Hsuan Chen, Matthias R. Fengler, Wolfgang Karl Härdle, Yanchu Liu, June 2018.

024 "Bootstrap Confidence Sets For Spectral Projectors Of Sample Covariance" by A. Naumov, V. Spokoiny, V. Ulyanov, June 2018.

025 "Construction of Non-asymptotic Confidence Sets in 2 -Wasserstein Space" by Johannes Ebert, Vladimir Spokoiny, Alexandra Suvorikova, June 2018.

026 "Large ball probabilities, Gaussian comparison and anti-concentration" by Friedrich Götze, Alexey Naumov, Vladimir Spokoiny, Vladimir Ulyanov, June 2018.

027 "Bayesian inference for spectral projectors of covariance matrix" by Igor Silin, Vladimir Spokoiny, June 2018.

028 "Toolbox: Gaussian comparison on Eucledian balls" by Andzhey Koziuk, Vladimir Spokoiny, June 2018.

029 "Pointwise adaptation via stagewise aggregation of local estimates for multiclass classification" by Nikita Puchkin, Vladimir Spokoiny, June 2018.

030 "Gaussian Process Forecast with multidimensional distributional entries" by Francois Bachoc, Alexandra Suvorikova, Jean-Michel Loubes, Vladimir Spokoiny, June 2018.

031 "Instrumental variables regression" by Andzhey Koziuk, Vladimir Spokoiny, June 2018.

032 "Understanding Latent Group Structure of Cryptocurrencies Market: A Dynamic Network Perspective" by Li Guo, Yubo Tao and Wolfgang Karl Härdle, July 2018.

033 "Optimal contracts under competition when uncertainty from adverse selection and moral hazard are present" by Natalie Packham, August 2018.

034 "A factor-model approach for correlation scenarios and correlation stress-testing" by Natalie Packham and Fabian Woebbeking, August 2018.

035 "Correlation Under Stress In Normal Variance Mixture Models" by Michael Kalkbrener and Natalie Packham, August 2018.

036 "Model risk of contingent claims" by Nils Detering and Natalie Packham, August 2018.

037 "Default probabilities and default correlations under stress" by Natalie Packham, Michael Kalkbrener and Ludger Overbeck, August 2018.

038 "Tail-Risk Protection Trading Strategies" by Natalie Packham, Jochen Papenbrock, Peter Schwendner and Fabian Woebbeking, August 2018.

This research was supported by the Deutsche

IRTG 1792 Discussion Paper Series 2018

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

039 "Penalized Adaptive Forecasting with Large Information Sets and Structural Changes" by Lenka Zbonakova, Xinjue Li and Wolfgang Karl Härdle, August 2018.

040 "Complete Convergence and Complete Moment Convergence for Maximal Weighted Sums of Extended Negatively Dependent Random Variables" by Ji Gao YAN, August 2018.

041 "On complete convergence in Marcinkiewicz-Zygmund type SLLN for random variables" by Anna Kuczmaszewska and Ji Gao YAN, August 2018.

042 "On Complete Convergence in Marcinkiewicz-Zygmund Type SLLN for END Random Variables and its Applications" by Ji Gao YAN, August 2018.

043 "Textual Sentiment and Sector specific reaction" by Elisabeth Bommes, Cathy Yi-Hsuan Chen and Wolfgang Karl Härdle, September 2018.

044 "Understanding Cryptocurrencies" by Wolfgang Karl Härdle, Campbell R. Harvey, Raphael C. G. Reule, September 2018.

045 "Predicative Ability of Similarity-based Futures Trading Strategies" by Hsin-Yu Chiu, Mi-Hsiu Chiang, Wei-Yu Kuo, September 2018.

046 "Forecasting the Term Structure of Option Implied Volatility: The Power of an Adaptive Method" by Ying Chen, Qian Han, Linlin Niu, September 2018.

047 "Inferences for a Partially Varying Coefficient Model With Endogenous Regressors" by Zongwu Cai, Ying Fang, Ming Lin, Jia Su, October 2018.

048 "A Regime Shift Model with Nonparametric Switching Mechanism" by Haiqiang Chen, Yingxing Li, Ming Lin, Yanli Zhu, October 2018.

049 "Strict Stationarity Testing and GLAD Estimation of Double Autoregressive Models" by Shaojun Guo, Dong Li, Muyi Li, October 2018.

050 "Variable selection and direction estimation for single-index models via DC-TGDR method" by Wei Zhong, Xi Liu, Shuangge Ma, October 2018.

051 "Property Investment and Rental Rate under Housing Price Uncertainty: A Real Options Approach" by Honglin Wang, Fan Yu, Yinggang Zhou, October 2018.

052 "Nonparametric Additive Instrumental Variable Estimator: A Group Shrinkage Estimation Perspective" by Qingliang Fan, Wei Zhong, October 2018.

053 "The impact of temperature on gaming productivity: evidence from online games" by Xiaojia Bao, Qingliang Fan, October 2018.

054 "Topic Modeling for Analyzing Open-Ended Survey Responses" by Andra-Selina Pietsch, Stefan Lessmann, October 2018.

055 "Estimation of the discontinuous leverage effect: Evidence from the NASDAQ order book" by Markus Bibinger, Christopher Neely, Lars Winkelmann, October 2018.

056 "Cryptocurrencies, Metcalfe's law and LPPL models" by Daniel Traian Pele, Miruna Mazurencu-Marinescu-Pele, October 2018.

057 "Trending Mixture Copula Models with Copula Selection" by Bingduo Yang, Zongwu Cai, Christian M. Hafner, Guannan Liu, October 2018.

This research was supported by the Deutsche

IRTG 1792 Discussion Paper Series 2018

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

058 "Investing with cryptocurrencies – evaluating the potential of portfolio allocation strategies" by Alla Petukhina, Simon Trimborn, Wolfgang Karl Härdle, Hermann Elendner, October 2018.

059 "Towards the interpretation of time-varying regularization parameters in streaming penalized regression models" by Lenka Zbonakova, Ricardo Pio Monti, Wolfgang Karl Härdle, October 2018.

060 "Residual's Influence Index (Rinfin), Bad Leverage And Unmasking In High Dimensional L2-Regression" by Yannis G. Yatracos, October 2018.

061 "Plug-In L2-Upper Error Bounds In Deconvolution, For A Mixing Density Estimate In R^d And For Its Derivatives" by Yannis G. Yatracos, October 2018.

062 "Conversion uplift in e-commerce: A systematic benchmark of modeling strategies" by Robin Gubela, Artem Bequé, Fabian Gebert, Stefan Lessmann, November 2018.

063 "Causal Inference using Machine Learning. An Evaluation of recent Methods through Simulations" by Daniel Jacob, Stefan Lessmann, Wolfgang Karl Härdle, November 2018.

This research was supported by the Deutsche