The Foerder Institute for Economic Research
at Tel Aviv University

The Eitan Berglas School of Economics

מכון למחקר כלכלי על שם ד"ר ישעיהו פורדר
על יד אוניברסיטת תל אביב

בית-הספר לכלכלה ע"ש איתן ברגלס

Machine Learning for Treatment Effect Heterogeneity: Recovering Partial Effects

Elad Guttman, Dor Leventer, Itay Saporta-Eksten, Analia Schlosser

Discussion Paper No. 5-2023

The Foerder Institute for Economic Research
and
The Sackler Institute of Economic Studies

Machine Learning for Treatment Effect Heterogeneity: Recovering Partial Effects *

Elad Guttman
Tel Aviv University eladguttman@mail.tau.ac.il

Dor Leventer
Tel Aviv University dorleventer@mail.tau.ac.il

Itay Saporta-Eksten
Tel Aviv University , *CEPR*, and *IZA* itaysap@tauex.tau.ac.il

Analia Schlosser
Tel Aviv University , *CEPR*, *CESifo*, and *IZA* analias@tauex.tau.ac.il

Abstract

Recent developments in the causal inference literature introduced Machine Learning (ML) algorithms to the analysis of heterogeneous treatment effects. Relying on these methods, various studies examine how treatment effects vary as a function of covariates. We highlight the potential interpretation challenges when one analyzes treatment effect heterogeneity without taking into account correlated covariates, and propose to examine the partial effect of a covariate on the estimated conditional average treatment effect. Our approach introduces the application of Partial Dependence Plots (PDP) and Accumulated Local Effects (ALE) used in the prediction literature, to the analysis of heterogeneous treatment effects.

JEL classification codes: C18, C21

Keywords: Machine Learning, Treatment Effects, Heterogeneous Treatment Effects, CATE, Partial Dependence Plots, Accumulated Local Effects.

*For valuable comments we thank Oren Danieli, Dan Zeltzer, Uri Shalit, and Daniel Nevo. Saporta-Eksten and Schlosser aknowledge financial support from the Pinhas Sapir Center for Development and from the Foerder Institute for Economic Research.

1. Introduction

The causal inference literature has expanded beyond the estimation of average treatment effects in an attempt to uncover the distributional impacts of treatment and to examine treatment effect heterogeneity. New recent developments include the application of Machine Learning (ML) methods to estimate heterogeneous treatment effects. Several recent studies include the estimation of Conditional Average Treatment Effects (CATE) for the purpose of optimizing treatment assignment and highlighting potential economic mechanisms underlying the response to treatment. The discussion of mechanisms based on ML estimation of treatment effect heterogeneity typically includes a comparison of covariates between those with high versus low predicted treatment effects, as well as the examination of how average treatment effects non-parametrically vary when the value of a covariate changes. The former is typically referred to as classification analysis (CLAN; see Chernozhukov et al. (2020)), while the latter is sometimes referred to as marginal plots in the ML literature (see, e.g., Molnar (2020), and Apley and Zhu (2020)). These types of analyses are informative; however, they cannot reveal whether a treatment effect varies due to a direct interaction of the treatment with a specific covariate, or rather through the covariance with other variables associated with the treatment effect.

In this paper, we highlight the potential interpretation challenges in applying marginal plots and CLAN to heterogeneity analysis and propose an approach that goes one step further to examine the *partial effect* of a covariate on the estimated CATE, while holding all other observed covariates constant. Specifically, we apply ML methods used to examine partial effects of a data feature on the response variable to the analysis of treatment effect heterogeneity. We show how to use Partial Dependence Plots (PDP) introduced by Friedman (2001), as well as Accumulated Local Effects (ALE) plots developed by Apley and Zhu (2020), to non-parametrically describe how the estimated CATE varies as a function of a specific covariate, while holding all other covariates constant.[1] Like any estimation of CATE, our approach has the limitation that it can explore only variation in treatment effects that arises from variation in observed covariates, and not omitted confounders. Yet, despite its limitations, our approach will help to highlight important patterns of predicted treatment effect heterogeneity to explore in further research. Moreover, in cases where the main drivers of heterogeneity are observed, our proposed methods provide important insights into why a specific treatment matters and for whom.

To motivate our approach, consider the following example. Suppose that there is an intervention where the treatment effect varies as a function of household income. In particular, assume that low-income families benefit more from treatment. Assume also that income is negatively associated with the number of children in the household. By contrast, a treatment effect does not vary with the number of children. In this setup, a simple analysis of treatment effect heterogeneity that does not account for correlations between covariates will show that the treatment effect is larger for families with more children even though there is no direct link between the treatment effect and family size. This might not matter if the objective of the analysis is to provide predictions in order to target treatment or create assignment rules. However, if the objective is to learn about why the treatment works and about possible mechanisms, we might arrive at wrong conclusions if we simply examine how a treatment effect varies as a function of a specific

[1] To keep the discussion simple, we always refer to a change in one variable while holding all other variables constant. The approach can be generalized to changing a set of variables while holding all other variables constant.

covariate without holding other covariates fixed (i.e., without accounting for correlations between covariates and their interactions with the treatment effect).

In Section 2 we illustrate these ideas using simulations, where we show that the heterogeneity analysis based on ML estimation using marginal plots and CLAN cannot capture the true heterogeneity in treatment effects with respect to covariates, even if all sources of heterogeneity are observed by the researcher.

We proceed in Section 3 by presenting two approaches to examining how the estimated CATE varies with respect to a specific covariate, while holding other covariates constant. We start by applying PDP to the analysis of heterogeneity in treatment effects. PDP are used in machine-learning applications to visualize how predictions of an outcome variable vary as a function of a specific feature of the data while all other features are kept constant. We propose an application of this method to examine how treatment effect predictions (instead of outcome predictions) vary as a function of a specific covariate while all other covariates are kept fixed. As discussed in the literature (see, e.g., Apley and Zhu (2020)), in cases of high correlation between covariates, PDP might sometimes fail to reflect the partial effect of covariates as they involve extrapolation over points beyond the envelope of the training data. Hence, we proceed by proposing the application of ALE to the analysis of treatment effect heterogeneity. We implement both methods within the Chernozhukov et al. (2020) framework, using their derivation of the Best Linear Predictor (BLP) of CATE.

We evaluate the PDP and ALE approaches in terms of performance and, using a simulation study, compare them to the heterogeneity analysis used in most recent papers in economics (i.e., marginal plots and CLAN). The simulations are based on different DGPs, where we also vary the degree of correlation between covariates and the functional form of the heterogeneous treatment effects. These simulations illustrate the stark differences between applying marginal plots and CLAN, and the two proposed approaches (PDP and ALE), where PDP and ALE succeed in capturing direct links between treatment and covariates, while holding other covariates fixed.

In Section 4 we reanalyze data from a field experiment in order to compare the alternative approaches. We estimate CATE by applying several ML methods and show that across all methods ALE and PDP deliver different results compared to marginal plots. These differences are consistent with the idea that marginal plots do not account for correlated covariates.

Our paper is directly related to the growing literature on the heterogeneity in treatment effects. Traditionally, the examination of treatment effect heterogeneity has been done in a classical regression framework by including interactions between the treatment indicator and covariates, while controlling for the main effects. Following recent advances in ML methods for causal inference (see, e.g., Athey and Imbens (2016); Wager and Athey (2018); Künzel et al. (2019); Athey, Tibshirani and Wager (2019); Chernozhukov et al. (2020); Hahn, Murray and Carvalho (2020); Nie and Wager (2021); Fan et al. (2022)), a growing number of papers include an analysis of treatment effect heterogeneity using a ML approach (see, e.g., O'Neill and Weeks (2018); Hoffman and Mast (2019); Davis and Heller (2020); Deryugina et al. (2019); Lechner, Strittmatter and Knaus (2020); Haaland and Roth (2020); Bertrand et al. (2021); Breda et al. (2021); Carlana and La Ferrara (2021), Farbmacher, Kögel and Spindler (2021); Sylvia et al. (2021)).[2] The use of ML methods to explore treatment effect heterogeneity has several appealing features as it requires

[2]For a review of recent methodological papers and applications see Knaus, Lechner and Strittmatter (2021).

fewer assumptions on the functional form or sources of heterogeneity.

Our paper contributes to the literature on the estimation of treatment effect heterogeneity by pointing out a potential pitfall in the interpretation of treatment effect heterogeneity with respect to observed covariates in setups where these covariates are correlated. We propose to apply two approaches that come from the ML literature to evaluate how the estimated CATE varies as a function of a specific covariate, while holding all other variables fixed.

2. Treatment Effect Heterogeneity Analysis in the Presence of Correlated Covariates

We start by showing a simple simulated example that illustrates the potential interpretation issues in the analysis of treatment effect heterogeneity based on CLAN and marginal plots in the presence of correlated covariates. Consider a Data Generating Process (DGP) with three variables $(z_1, z_2, z_3) \sim N(0, \Sigma)$ such that:

$$\Sigma = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & 0 \\ \rho & 0 & 1 \end{pmatrix}, \tag{1}$$

and an outcome function defined by

$$Y = \alpha_1 (W \times I(z_1 > 0)) + \gamma_1 z_1 + \alpha_2 (W \times I(z_2 > 0)) + \gamma_2 z_2 + \alpha_3 (W \times I(z_3 > 0)) + \gamma_3 z_3 + u, \tag{2}$$

where W is an indicator for participation in the treatment group and I is the indicator function.

This DGP is designed such that the heterogeneity in treatment effect is a step function over z_1, z_2, and z_3, allowing for correlation between z_1 and z_2 and between z_1 and z_3. We examine a particular case where $\alpha_1 = -\alpha_2 = 0.15$, $\gamma_1 = -\gamma_2 = 0.06$, and $\alpha_3 = \gamma_3 = 0$. This implies that while z_1 and z_2 are correlated (for $\rho \neq 0$) they have opposite effects on Y and their interaction with the treatment effect is also of the opposite sign. At the same time, z_3 is correlated with z_1, but it does not affect Y by itself, and naturally does not induce any treatment effect heterogeneity.

To obtain CATE estimates for this DGP, we use three popular ML methods, Generalized Random Forests (GRF), Gradient Boosting (GBM), and Neural Networks (NNET), which are non-parametric algorithms that differ in their smoothness level, allowing us to achieve a good fit to various DGPs.[3,4] The estimates were obtained by applying all algorithms to estimate the proxy for CATE. Following the approach in Chernozhukov et al. (2020), we choose between the proxies using the performance measure (Λ), which is informative on the fit in the regression of the estimated proxy on the true CATE. We then obtain a measure of the CATE by calculating the BLP of CATE (see more details about the implementation in Appendix D).[5]

[3] Whenever we apply GRF, we refer to the implementation of causal forests as discussed in Athey, Tibshirani and Wager (2019).

[4] For the GBM and NNET algorithms, we estimate the proxy for CATE using a T-learner methodology (see, e.g., Künzel et al. (2019)).

[5] As opposed to classical ML algorithms, GRF (Athey, Tibshirani and Wager (2019)) predicts treatment effects rather than outcomes and produces consistent estimates for CATE. This fact renders estimation of BLP unnecessary. Nevertheless, for comparability with the other methods, we apply the same procedure using the BLP of CATE throughout the paper.

Figure 1 shows *marginal plots* for different values of z_1 and z_3. Marginal plots offer a popular way to summarize the estimated CATE function obtained from the ML procedure. They are calculated as an average of the CATE estimates over a single dimension. The plots in Figure 1 were obtained using the GRF algorithm, which had the higher performance measure (Λ) in this case. The displayed estimates were obtained averaging over 500 simulated data sets from this DGP. Each data set was compiled of 10,000 observations with a treatment assignment probability of 0.5 and $u \sim N(0, 0.1^2)$.

Starting from the left panel of the figure, as we move along the z_1 dimension, the distribution of z_2 and z_3 changes as well, in accordance with equation (1). Hence, the marginal plot over z_1 captures (1) a direct relation between z_1 and the treatment effect (the partial effect of z_1), and (2) the changes in CATE due to the correlation between z_1 and the other two covariates, through their relation with the treatment effect.

The "direct effect" line in the figure represents changes in the treatment effect due to changes in z_1, holding constant the other variables, centered around the average treatment effect (ATE).[6] The red diamonds show the estimated marginal plot when there is no correlation between the zs (i.e., $\rho = 0$). In this scenario, there is no indirect effect. Hence the marginal plot provides a good representation of the direct relationship between z_1 and the treatment effect. The black circles show the estimated marginal plot for the case where $\rho = 0.6$. Unlike the no-correlation case, the marginal plot over z_1 incorporates also the indirect effect driven by the relation between the other covariates and the treatment effect. The deviation of the estimates from the "direct effect" line is larger as z_1 grows (in absolute terms). Specifically, z_2 has the opposite effect on the treatment effect, and is positively correlated with z_1, which explains the growing difference between the marginal plot and the "direct effect" for high values of z_1.

The right panel in Figure 1 shows the marginal plot over the z_3 dimension, a covariate that does not interact with the treatment effect in the DGP (and, in fact, does not affect Y at all). In this case, the marginal plot captures only the indirect effect of z_3 through its correlation with z_1. For the case of $\rho = 0.6$, this plot demonstrates how marginal plots in the presence of correlated covariates may erroneously lead one to conclude that an irrelevant variable (z_3) is an important driver of treatment effect heterogeneity.

While so far we have focused on marginal plots, it is important to highlight that other popular approaches to analyzing heterogeneity in treatment effects, such as CLAN, raise similar interpretation issues. CLAN aims to characterize individuals with high versus low treatment effects in terms of their observed characteristics. This is done by first splitting the sample according to the level of the predicted treatment effect obtained using the ML procedure (for example, by quartiles of treatment effect), and then showing means of observed characteristics for individuals in the different categories.[7] Table 1 shows the results for a CLAN analysis of the simulation study discussed above, for the cases of $\rho = 0$ and $\rho = 0.6$. Each row in Table 1 reports the mean of a specific covariate for observations with predicted treatment effect in the lower quartile (columns (1) and (4)) and in the upper quartile (columns (2) and (5)) of the proxy predictor $S(z)$. Columns (3) and (6) show the difference between the means of the lower and the upper quartiles and report in parentheses rejection rates at the 10% significance level for a test of the hypothesis that the

[6] In Section 3, we discuss two different concepts that correspond to the direct effect. However, for the specific DGP discussed in the current section, the two concepts coincide, due to the lack of interaction between the covariates in the treatment effect function.

[7] See Appendix D for more details about CLAN analysis.

difference equals zero, over the 500 replications. The table illustrates that when covariates are correlated, using the difference in mean between variables among those with high and low predicted treatment effects could give rise to erroneous conclusions regarding the true drivers of treatment effect heterogeneity. For example, when $\rho = 0.6$, the difference in the mean of z_3 between observations with treatment effects in the lower quartile and the upper quartile (column (6)), is the largest, even though z_3 is not even in the DGP for Y and has no interaction with treatment.

In the next section, we show how to apply two methods using machine-learning predictions of treatment effects to examine treatment effect heterogeneity with respect to specific covariates, while holding other covariates constant.

3. Heterogeneous Treatment Effects, Partial Dependence Plots, and Accumulated Local Effects

Motivated by the above examples, we introduce two methods that allow researchers to look into the black box of ML predictions, and explore the direct relation between a covariate and the estimated CATE, while accounting for other correlated features of the data. We start by discussing PDP, first introduced by Friedman (2001), and then present ALE plots, as discussed in Apley and Zhu (2020). We briefly describe the general ideas behind PDP and ALE, as well as their interpretation in the context of treatment effect heterogeneity. We close this section by showing how PDP and ALE perform in a simulation study.

3.1. Partial Dependence Plots

A Partial Dependence Plot (PDP) is a model-agnostic tool aimed at illustrating the partial effect of one variable (or a set of variables) on the predicted outcome of a machine learning model, which is usually trained using many (correlated) variables. This is done by fixing the variable of interest at specific values and then, for each value, averaging the predicted outcome over the *marginal* distribution in the sample (Friedman (2001)). In this section we apply this approach to CATE estimation.

Denote by $Y(1)$ and $Y(0)$ treated and untreated potential outcomes respectively and $s_0(\mathbf{z}) = \mathbb{E}[Y(1) - Y(0)|z]$ be the conditional average treatment effect (CATE), where $\mathbf{z}$ is a vector of covariates that potentially interact with the treatment effect, z_l is a single covariate of interest, and z_{-l} denotes all variables except z_l. Lastly, we denote by $\hat{s}$ some estimator of s_0.

One way to illustrate the dependency of $\hat{s}$ on z_l is to use marginal plots, calculated by $\mathbb{E}[\hat{s}(\mathbf{z}) \mid z_l]$. However, as discussed above, by using the conditional expectation, we also capture the indirect effects of z_l on $\hat{s}(\mathbf{z})$ over zs that are correlated with z_l (analogous to omitted variable bias in a regression context). Our first proposal to overcome this problem is to adapt the PDP approach by Friedman (2001) to the heterogeneous treatment environment. We define the PDP estimator of the CATE as

$$\hat{s}_l^{PDP}(z_l^*) = \mathbb{E}_{z_{-l}}\left[\hat{s}(z_l^*, z_{-l})\right], \tag{3}$$

which traces how the estimator changes when we change z_l, while keeping constant all other z_{-l}. The concept in (3) can thus be interpreted as a counterfactual prediction of the CATE estimator, when all z_{-l} are kept constant and only z_l varies. By keeping all other covariates constant, we do not allow z_{-l} to change as we vary z_l, and hence we observe only the direct effect of z_l. To estimate (3), one can replace the expectation operator with sample averages.

It is worth noting that when the data consists of strongly correlated covariates, PDP may lead to extrapolation of the treatment effect outside the joint support of the covariates (see the discussion in Apley and Zhu (2020)). In this case, the computation of equation (3) for a covariate, which is highly correlated with other covariates, involves averaging predictions of data points that rarely exist in reality, making the PDP estimates sensitive to model extrapolation. ALE, discussed in the next section, is not sensitive to such extrapolation issues.

3.2. Accumulated Local Effects

The second approach, Accumulated Local Effects (ALE), relies on using the partial derivatives of the estimator with respect to a specific variable z_l. The aim of ALE, described by Apley and Zhu (2020), is to average the predicted derivative of a function with respect to z_l over the *conditional* distribution of z_{-l}, and then to integrate these local derivatives to obtain a local averaged prediction. As above, we adapt this idea to the heterogeneous treatment environment.

To define the estimand of ALE, consider first the mean partial derivative of $\hat{s}(\mathbf{z})$ with respect to z_l at a specific point t:[8]

$$PD_l^{ALE}(t) = \mathbb{E}_{z_{-l}}\left[\frac{\partial \hat{s}(z_l, z_{-l})}{\partial z_l} \mid z_l = t\right]. \tag{4}$$

This expression has a clear interpretation: it captures how the estimated CATE changes locally with z_l around a specific point (t) when all other variables are held constant. ALE accumulates these local changes by integrating over the partial derivatives. Integrating over (4) we get the ALE function for $\hat{s}(\mathbf{z})$:

$$\bar{s}_l^{ALE}(z_l^*) = \int_{z_{l,min}}^{z_l^*} PD_l^{ALE}(t)dt - c. \tag{5}$$

To estimate (5), one needs to estimate (4) using $\hat{s}$, as well as to numerically calculate the integral in (5). In Appendix E we outline the implementation of the ALE estimator following Apley and Zhu (2020).

3.3. Comparing the Interpretation of PDP and ALE

Before proceeding with the implementation of PDP and ALE, it is useful to compare them. These two approaches capture different statistical concepts, and hence generally lead to different estimates. Specifically, predicting how the estimated CATE changes over the z_l dimension, PDP integrates over the *marginal* distribution of other covariates z_{-l}, regardless of the conditional distribution around specific values of z_l. By contrast, ALE takes into account the fact

[8] For simplicity we present here the definition in which $\hat{s}$ is assumed to be differentiable, but there is another version of this definition in which this assumption is relaxed. See, e.g., equation (6) in Apley and Zhu (2020), and application of ALE to discrete covariates, discussed in Appendix F.

that partial derivatives of the estimated CATE with respect to z_l may depend on other covariates. To do so, it uses the *conditional* distribution of z_{-l} when recovering local partial derivatives around specific values of z_l.

To further illustrate the difference between PDP and ALE, suppose that $\hat{s}$ and its partial derivative w.r.t. z_l are continuous and consider the partial derivative of the PDP estimand (3) with respect to z_l:

$$\frac{\partial E_{z_{-l}}\left[\hat{s}\left(z_l^*, z_{i,-l}\right)\right]}{\partial z_l} = E_{z_{-l}}\left[\frac{\partial \hat{s}\left(z_l^*, z_{i,-l}\right)}{\partial z_l}\right]. \tag{6}$$

This derivative is the *unconditional* expectation of a partial derivative of the CATE estimator with respect to a single covariate. This is in contrast to the derivative of ALE as shown in (4), which is a *conditional* expectation of a partial derivative of the CATE estimator with respect to a single covariate.

In Appendix C we show sufficient conditions for the estimands for PDP and ALE to coincide (though the estimators are not generally numerically identical).[9] In the simulation study below, we introduce a specific DGP, which demonstrates where the two approaches diverge.

3.4. Applying PDP and ALE to Study Heterogeneous Treatment Effects

We turn now to the application of PDP and ALE to the analysis of heterogeneous treatment effects. We implement PDP and ALE within the Chernozhukov et al. (2020) framework. This general framework can be applied to study heterogeneous treatment effects, using any machine learning approach (for more details see Appendix D.1). A central idea in Chernozhukov et al. (2020) is to provide consistent estimates and confidence intervals of features of the CATE instead of focusing on the CATE itself. This approach overcomes the challenges of getting a valid estimation framework for the CATE obtained from generic ML algorithms.

Let $s_0(\mathbf{z})$ be the true CATE, and $S(\mathbf{z})$ be a proxy predictor of $s_0(\mathbf{z})$. There is no requirement for the proxy to be a consistent estimate of CATE, and hence we may use a wide variety of ML methods to estimate this proxy. To obtain consistent estimates and confidence intervals for a specific feature of CATE, the estimation approach involves randomly splitting the data into main and auxiliary samples where the model is trained using the auxiliary sample and inference is performed on the main sample. To account for the uncertainty induced by the random splitting of the data, Chernozhukov et al. (2020) propose using Variational Estimation and Inference (VEIN), i.e., repeating the estimation process many times and reporting the median over the splits for point estimates and confidence intervals.

One of the features derived by Chernozhukov et al. (2020) is the Best Linear Predictor (BLP) of CATE, $s_0(\mathbf{z})$. Given the proxy $S(\mathbf{z})$, the BLP of CATE is defined as

$$BLP(s_0(\mathbf{z}) \mid S(\mathbf{z})) = \beta_1 + \beta_2 \left(S(\mathbf{z}) - \mathbb{E}[S(\mathbf{z})]\right). \tag{7}$$

In this equation, β_1 is interpreted as the ATE, while β_2 measures both the presence of heterogeneity and whether the

[9] By construction, as an integral, ALE is defined up to a scale and therefore, to show that PDP and ALE coincide, we show that their partial derivatives w.r.t. z_l coincide at every point. As mentioned, we suggest centering ALE around the ATE. However, the mean of the PDP function is not necessarily the ATE, and therefore if one wishes to have the two functions coincide in levels, ALE needs to be centered around the PDP mean.

proxy $S(\mathbf{z})$ is a relevant predictor of CATE. Chernozhukov et al. (2020) develop two approaches to estimate (7).[10]

We apply the PDP and ALE approaches to the estimated $BLP(s_0(\mathbf{z}) \mid S(\mathbf{z}))$ and use the VEIN method to calculate confidence bands (see details in Appendix D). Note that Chernozhukov et al. (2020) prove that this method provides consistent estimates and valid confidence bands for the BLP of CATE. While PDP and ALE are informative of the properties of the estimated CATE function, there is no formal proof that they recover the properties of the true CATE since they rely on the *derivative* of the BLP of CATE. As a result, the confidence bands that we provide are mainly meant to illustrate the amount of uncertainty of our PDP and ALE estimates. Nevertheless, in the simulation study described below we show that both PDP and ALE estimates successfully capture the direct effects of covariates on the treatment effect for various DGPs with different types of heterogeneity structures and correlation between covariates.

3.5. A Simulation Study

We illustrate the two proposed methods, PDP and ALE, for the analysis of treatment effect heterogeneity using simulations. We conduct simulations for three DGPs. First, the DGP defined in equations (1) and (2), which we term the "step" DGP. The second DGP has the same correlation matrix between the covariates as defined in equation (1), but the heterogeneity in the treatment effects is linear, defined in equation (8) below. In what follows we term this the "linear" DGP. In the third DGP we introduce interactions between two covariates and the treatment effects in order to highlight the differences between PDP and ALE.

We first focus on the step DGP defined in equations (1) and (2), and a linear DGP of the following form:

$$Y = \alpha_1 z_1 W + \gamma_1 z_1 + \alpha_2 z_2 W + \gamma_2 z_2 + \alpha_3 z_3 W + \gamma_3 z_3 + u. \tag{8}$$

Each simulated data set consists of 10,000 observations, the treatment assignment probability is 0.5, and $u \sim N(0, 0.1^2)$. We set $\alpha_1 = -\alpha_2 = 0.09$, $\gamma_1 = -\gamma_2 = 0.04$, and $\alpha_3 = \gamma_3 = 0$.[11] We conduct 500 replications, and use the same estimation procedure described in Section 2 for our main results. Figure 2 summarizes the results from the simulation study for the step and the linear DGPs (we introduce the third DGP below). The solid black line denotes the direct effect of z_i on CATE, when z_{-i} is held constant. Because the treatment effect is additive, in this case the direct effect is equivalent under both PDP and ALE (see the discussion in Appendix C). The triangles show the average marginal plot estimates over the simulated data sets, the diamonds show the average PDP estimates, and the circles show the average ALE estimates. The colors show the share of data sets for which we rejected the test that the estimate is different from the true direct effect (indicated by the solid black line) at the 10% significance level. Green denotes low rejection values, implying that the estimator performs well. To represent the variance over the simulated samples, we also report the 0.005 and 0.995 percentiles of the different estimators over the 500 replications for each point in the figure.

Panels A and B of Figure 2 plot the results for the step DGP obtained by applying GRF as the proxy predictor ML algorithm (given that it provided a higher performance measure than NNET and GBM). Panel A plots the results

[10] See further discussion in Chernozhukov et al. (2020) and in Appendix D.
[11] For $\rho = 0$ these values maintain the same R^2 as that of the step DGP.

for a relatively low correlation case ($\rho = 0.3$) and Panel B plots the results when there is higher correlation between covariates ($\rho = 0.6$). As shown in Panel A, PDP and ALE plots deliver highly similar results. However, they both differ from the marginal plot, which erroneously attributes part of the effect of z_1 to the heterogeneity in treatment effect with respect of z_3. Similarly, the marginal plot differs from the direct effect of z_1 as it attributes part of the (negative) treatment effect of z_2 to z_1, while ALE and PDP capture direct effect correctly. This is even more salient in Panel B ($\rho = 0.6$), where differences between the marginal plot and the direct effect are larger due to the higher correlation between the covariates.

Panels C and D of Figure 2 show the results for the linear DGP obtained by applying NNET as the proxy predictor ML algorithm (due to a higher performance measure than GRF and GBM). Again, we see that the marginal plot estimates attribute part of the treatment effect heterogeneity to z_3 even though this covariate does not interact with the treatment. In addition, the effect of z_1 in determining heterogeneity in treatment effects is much smaller (in absolute terms) than the direct effect, especially in the high correlation case (Panel B). PDP and ALE plots almost overlap in this case and are again very close to the direct effect.

The simulation has shown no evident difference between the PDP and ALE estimates. As discussed in Appendix C, this result is not surprising when the treatment effect does not include interactions between the covariates, as was the case in the two DGPs discussed so far. The third DGP will highlight the difference between PDP and ALE when the treatment effect is not additive. Specifically, we add to the linear DGP in (8) an interaction between the treatment and both z_1 and z_2 as follows:

$$Y = \alpha_1 z_1 W + \gamma_1 z_1 + \alpha_2 z_2 W + \gamma_2 z_2 + \alpha_3 z_3 W + \gamma_3 z_3 + \alpha_4 z_1 z_2 I(z_1 > 0) W + u, \tag{9}$$

where we use the exact same specifications as before, and set $\alpha_4 = \alpha_1 = 0.09$.

We focus on the case of $\rho = 0.6$ and present the results in Figure 3 for the NNET as it achieved the highest performance measure. In the left panel, triangles represent the average marginal plot estimates, diamonds represent average PDP estimates, and the black line represents the direct effect defined in to equation (3). PDP estimates fit this line well, while the marginal plot estimates do not. Similarly, in the right panel triangles represent average marginal plot estimates, circles represent average ALE estimates, and the black line represents the direct effect line defined in to equation (5). It is evident that the marginal plot estimates do not fit the direct effect line well, while the ALE estimates approximate this line with high precision.

We also conduct several robustness tests to demonstrate that our results hold when we change the stylized DGPs in the above simulation. Specifically, we replicate Figures 2 and 3 three times with the following changes: we reduce the number of observations by half from 10,000 to 5,000 (Figures A1 and A2); we increase the noise from $u \sim N(0, 0.1^2)$ to $u \sim N(0, 0.2^2)$ (Figures A3 and A4); and we decrease the heterogeneity by half from $\beta_1 = \beta_2 = 0.15$ to $\beta_1 = \beta_2 = 0.075$ in equation (2) and from $\alpha_1 = \alpha_2 = 0.09$ to $\alpha_1 = \alpha_2 = 0.045$ in equations (8) and (9) (Figures A5 and A6). In all cases the results remain very similar to the ones presented in this section, even though the best algorithm according to the Λ performance measure varies across some specifications.[12]

[12] Due to the long computational time, we lowered the number of MC replications in the robustness tests from 500 to 100 when using GRF as the

4. Application

In this section we show the results from the different approaches discussed above using data from a field experiment. We use data from a General Social Survey (GSS) wording experiment, illustrating that marginal plots might lead to different conclusions than PDP and ALE estimates regarding the sources of treatment effect heterogeneity.

For about five decades, the GSS has been collecting Americans' perspectives and views on national spending priorities, as well as on many other social issues. Starting in the mid-1980s, respondents randomly received different versions of the question regarding national spending. While some were asked about their view on the level of spending on "welfare," others were asked about the level of spending on "assistance to the poor."[13] It has been long documented that the framing of the question has an impact on the response. Specifically, Americans tend to hold the opinion that national spending on welfare is too high, but that assistance to the poor is too low.[14]

The sample we use includes 32,814 observations, of which 17,567 are assigned to treatment and 15,247 to control.[15] The outcome variable is the response to the national spending question coded "1" for "too much" and "0" otherwise. Treatment is defined as receiving the "welfare" version of the question and is randomly assigned in each survey year. A comparison of the characteristics of the treatment and control groups (Appendix Table B1) suggests that within a survey year, the assignment protocol produced well-balanced groups. Our set of covariates include age, years of education, party identification coded on a scale of 0 to 6 (higher values denote affiliation with the Republican party), political views coded on a scale of 1 to 7 (higher values denote more conservative views), gender, race, and a vector of year dummy indicators. Appendix Figure A7 shows a high correlation between some covariates. For example, the race covariate "Black" is negatively correlated to conservative parties.

We use GRF, gradient boosting, and neural network algorithms to compute treatment effects.[16] Appendix Table B2 reports the results from estimating the BLP (Chernozhukov et al. (2020)) for the three algorithms. The GBM algorithm (column (2)) does a slightly better job than GRF in capturing heterogeneity, and a much better job than neural network. All the algorithms report a β_2 that is significantly different from zero, suggesting that there is heterogeneity in the treatment effects.

In Figure 4, we illustrate how PDP and ALE may lead to a different interpretation from that of marginal plots, by focusing on the race and party identification covariates. The figure reports results from the three algorithms (GBM, GRF, and NNET) along with empirical confidence bands. The solid line indicates the average treatment effect and the dotted lines indicate its confidence band. We also report at the top of each column the sample distribution of the variable of interest. Marginal plot estimates show large differences in the treatment effect for Blacks versus non-Blacks, while PDP and ALE show much smaller differences along this dimension. This finding may be explained by the negative correlation between Black and conservative party identification, which is found to be an important driver

proxy predictor, after verifying that the main results were very similar.

[13] See Appendix G for the full wording of the question

[14] See, e.g., Smith (1987), Rasinski (1989), and Green and Kern (2012) for some studies that analyzed this experiment.

[15] We use the GSS data for the years 1986–2010 taken from Green and Kern (2012). The data is available from Susan Athey and Guido Imbens' course "Machine Learning and Econometrics" (AEA continuing Education, 2018), see: https://github.com/gsbDBI/ExperimentData.

[16] We account for the randomization process by including survey year dummies as covariates in the ML model fit, as well as fixed effects in the BLP estimation, along the lines of the application in Chernozhukov et al. (2020).

of heterogeneity in the PDP and ALE analysis. Relatedly, PDP and ALE show smaller variation in treatment effects with respect to party identification relative to the pattern described by the marginal plots, which does not account for the correlation between this variable and other variables. Importantly, the three algorithms produce very similar results when we compare marginal plots with PDP and ALE.[17]

5. Conclusion

The treatment effect literature has evolved beyond analyzing average treatment effects to analyzing heterogeneous treatment effects. To this end, recent studies apply ML methods to estimate Conditional Average Treatment Effects (CATE). These estimates are sometimes used to discuss the economic mechanisms underlying the response to treatment. The discussion of economic mechanisms in this context typically includes a comparison of covariates between those with high versus low predicted treatment effects, and the examination of how average treatment effects vary with the values of the covariates.

In this paper, we proposed an approach that goes one step further than the classical heterogeneity analysis by examining the *partial effect* of a covariate on the estimated CATE function. Our approach brings ML tools that are used to uncover the role of specific variables in black-box predictions, specifically PDP and ALE, to the domain of heterogeneous treatment effects. We show in simulations that they perform well in uncovering partial effects. Finally, we demonstrate the differences between marginal plot estimates and the PDP and ALE approaches using data from a field experiment.

[17] In Appendix Figure A8 we report the results for the rest of the covariates, focusing on the GBM method that had the highest performance measure (Λ).

References

Apley, Daniel W., and Jingyu Zhu. 2020. "Visualizing the effects of predictor variables in black box supervised learning models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4): 1059–1086.

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

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized random forests." *The Annals of Statistics*, 47(2): 1148 – 1178.

Bertrand, Marianne, Bruno Crepon, Alicia Marguerie, and Patrick Premand. 2021. "Do Workfare Programs Live Up to Their Promises? Experimental Evidence from Cote DâIvoire." National Bureau of Economic Research.

Breda, Thomas, Julien Grenet, Marion Monnet, and Clémentine Van Effenterre. 2021. "Do female role models reduce the gender gap in science? Evidence from French high schools."

Carlana, Michela, and Eliana La Ferrara. 2021. "Apart but connected: Online tutoring and student outcomes during the COVID-19 pandemic."

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val. 2020. "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments."

Davis, Jonathan MV, and Sara B Heller. 2020. "Rethinking the benefits of youth employment programs: The heterogeneous effects of summer jobs." *Review of Economics and Statistics*, 102(4): 664–677.

Deryugina, Tatyana, Garth Heutel, Nolan H Miller, David Molitor, and Julian Reif. 2019. "The mortality and medical costs of air pollution: Evidence from changes in wind direction." *American Economic Review*, 109(12): 4178–4219.

Fan, Qingliang, Yu-Chin Hsu, Robert P Lieli, and Yichong Zhang. 2022. "Estimation of conditional average treatment effects with high-dimensional data." *Journal of Business & Economic Statistics*, 40(1): 313–327.

Farbmacher, Helmut, Heinrich Kögel, and Martin Spindler. 2021. "Heterogeneous effects of poverty on attention." *Labour Economics*, 71: 102028.

Friedman, Jerome H. 2001. "Greedy function approximation: a gradient boosting machine." *Annals of statistics*, 1189–1232.

Green, Donald P, and Holger L Kern. 2012. "Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees." *Public opinion quarterly*, 76(3): 491–511.

Haaland, Ingar, and Christopher Roth. 2020. "Labor market concerns and support for immigration." *Journal of Public Economics*, 191: 104256.

Hahn, P Richard, Jared S Murray, and Carlos M Carvalho. 2020. "Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion)." *Bayesian Analysis*, 15(3): 965–1056.

Hoffman, Ian, and Evan Mast. 2019. "Heterogeneity in the effect of federal spending on local crime: Evidence from causal forests." *Regional Science and Urban Economics*, 78: 103463.

Knaus, Michael C, Michael Lechner, and Anthony Strittmatter. 2021. "Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence." *The Econometrics Journal*, 24(1): 134–161.

Künzel, Sören R, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. 2019. "Metalearners for estimating heterogeneous treatment effects using machine learning." *Proceedings of the national academy of sciences*, 116(10): 4156–4165.

Lechner, Michael, Anthony Strittmatter, and Michael Knaus. 2020. "Heterogeneous Employment Effects of Job Search Programmes: A Machine Learning Approach." *Journal of Human Resources*.

Molnar, Christoph. 2020. *Interpretable machine learning.* Lulu. com.

Nie, Xinkun, and Stefan Wager. 2021. "Quasi-oracle estimation of heterogeneous treatment effects." *Biometrika*, 108(2): 299–319.

O'Neill, Eoghan, and Melvyn Weeks. 2018. "Causal tree estimation of heterogeneous household response to time-of-use electricity pricing schemes." *arXiv preprint arXiv:1810.09179*.

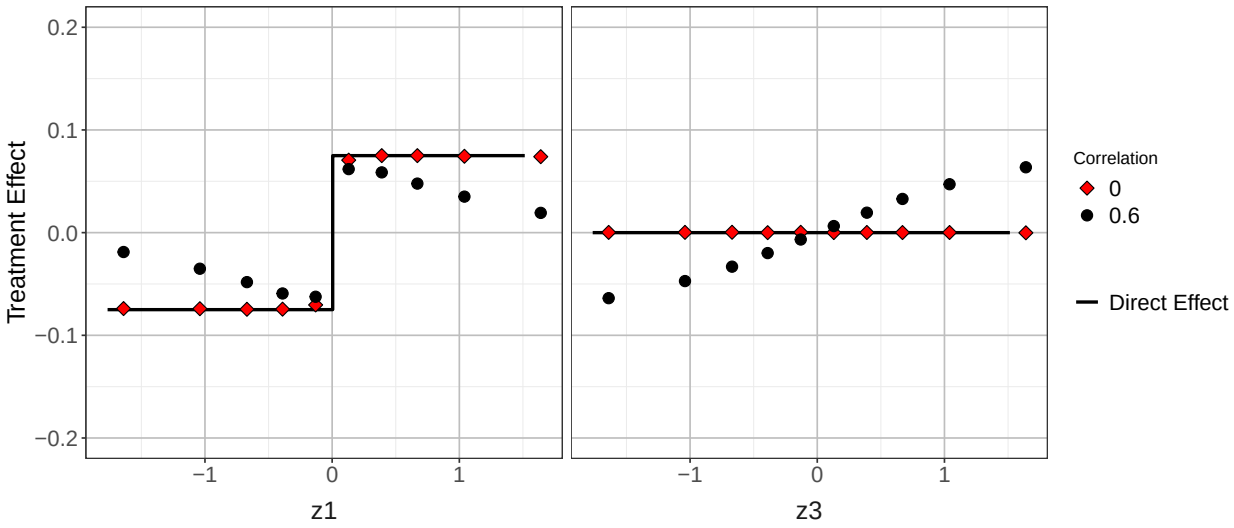
Rasinski, Kenneth A. 1989. "The effect of question wording on public support for government spending." *Public Opinion Quarterly*, 53(3): 388–394.

Smith, Tom W. 1987. "That which we call welfare by any other name would smell sweeter an analysis of the impact of question wording on response patterns." *Public opinion quarterly*, 51(1): 75–83.

Sylvia, Sean, Nele Warrinnier, Renfu Luo, Ai Yue, Orazio Attanasio, Alexis Medina, and Scott Rozelle. 2021. "From quantity to quality: Delivering a home-based parenting intervention through Chinaâs family planning cadres." *The Economic Journal*, 131(635): 1365–1400.

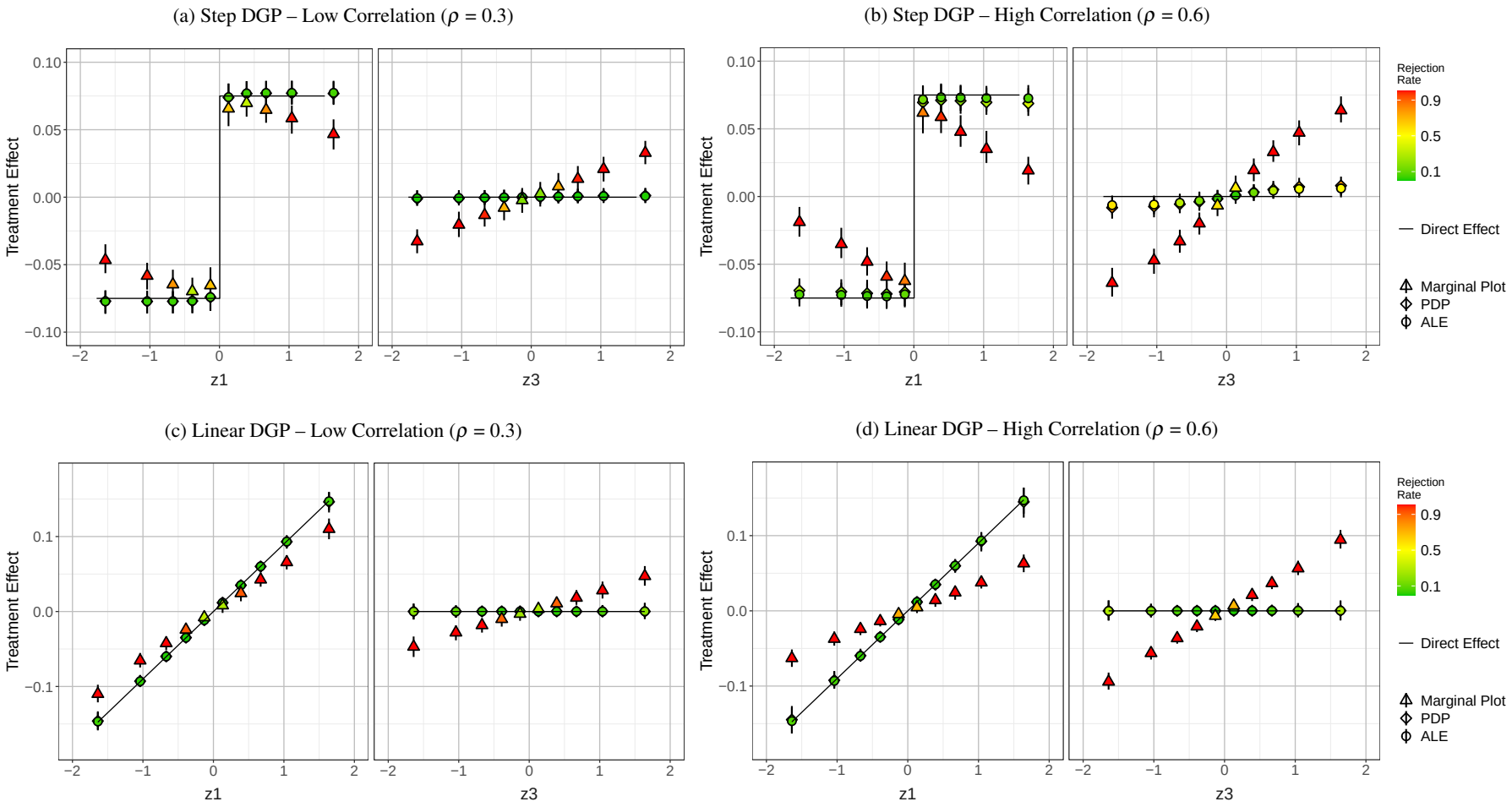
Wager, Stefan, and Susan Athey. 2018. "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association*, 113(523): 1228–1242.

Figure 1: Simulated Estimated Marginal Plots with Correlated Covariates



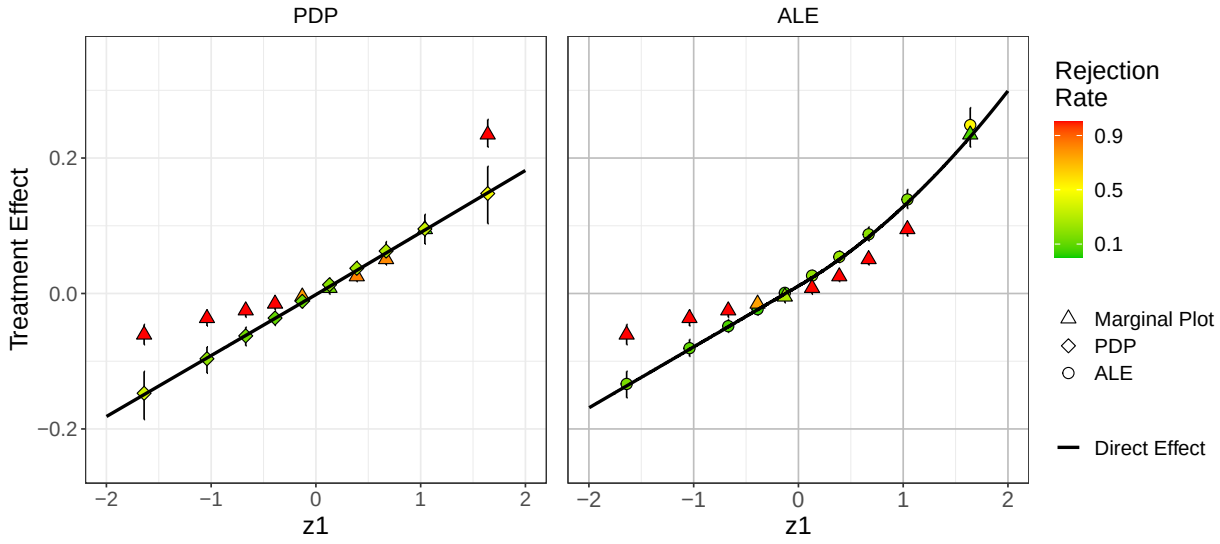
Notes: The figure shows results obtained by applying the GRF algorithm as the ML proxy predictor, averaged over 500 replications. GRF, GBM, and NNET algorithms were applied. The algorithm with the highest performance measure was chosen. The "direct effect" line in the figure represents changes in the treatment effect due to changes in z_1 (left panel) or z_3 (right panel), when the other variables are held constant. Step DGP is defined in equations (1) and (2).

Figure 2: Simulation Study of Step and Linear Heterogeneous Treatment Effect



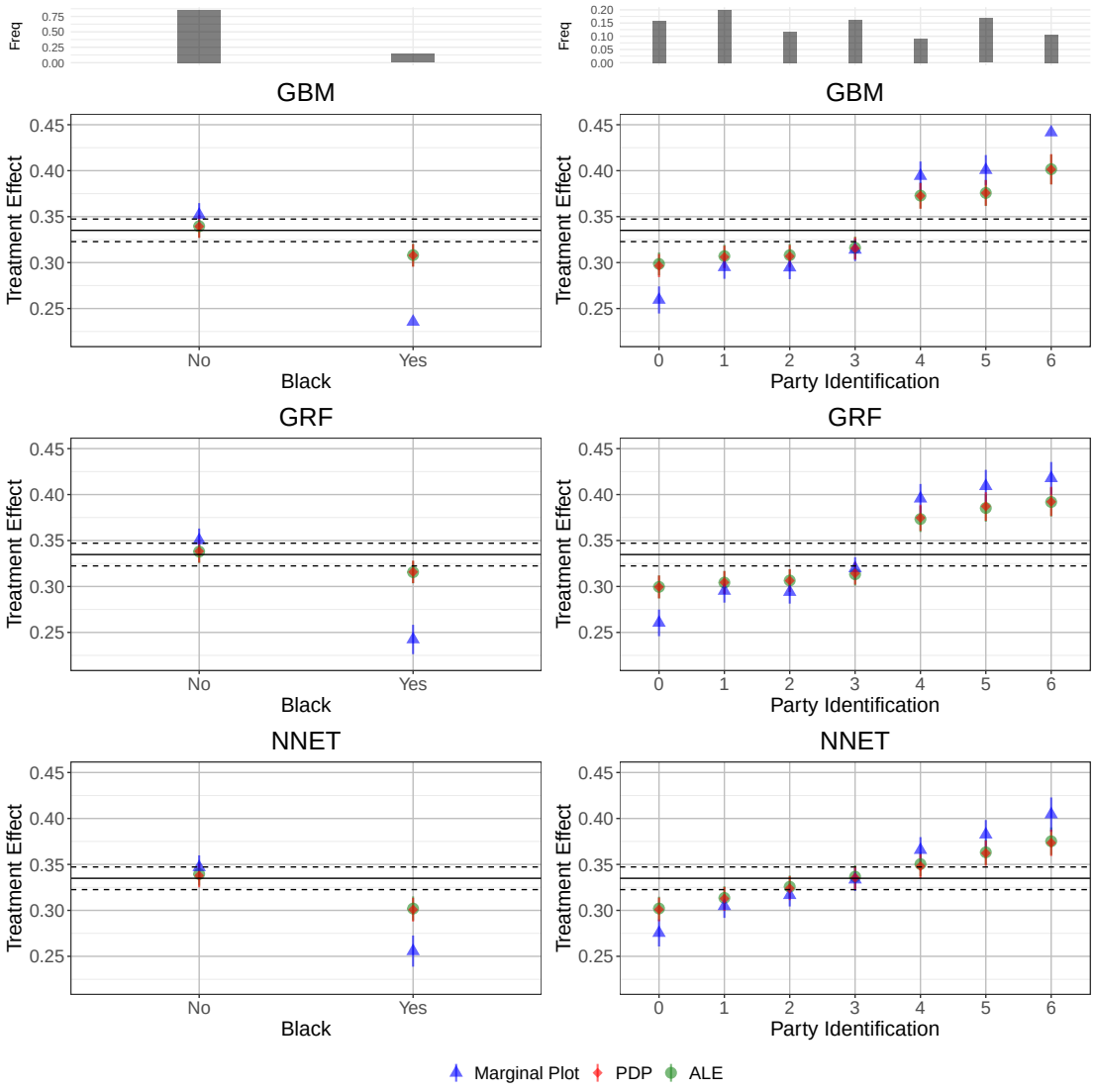
Notes: The figure shows results obtained by applying the GRF for the Step DGP and NNET for the Linear DGP as the ML proxy predictor, averaged over 500 replications. GRF, GBM, and NNET algorithms were applied. The algorithm with the highest performance measure was chosen. Step DGP is defined in equations (1) and (2), and Linear DGP is defined in equations (1) and (8).

Figure 3: Simulation Study of Treatment Effect with Interactions



Notes: The figure shows results obtained by applying the NNET algorithm as the ML proxy predictor, averaged over 500 replications. GRF, GBM, and NNET algorithms were applied, with NNET obtaining a higher performance measure. Interactions DGP defined in equations (1) and (9) for $\rho = 0.6$.

Figure 4: Application – Heterogeneity Along Black and Party Identification Variables



Notes: The ML algorithm is reported at the top of each panel and the covariate at the bottom. The top of each column shows the covariate's distribution. The black horizontal lines represent the estimated ATE $\hat{\beta}_1$ (solid) and its 90% confidence interval (dashed).

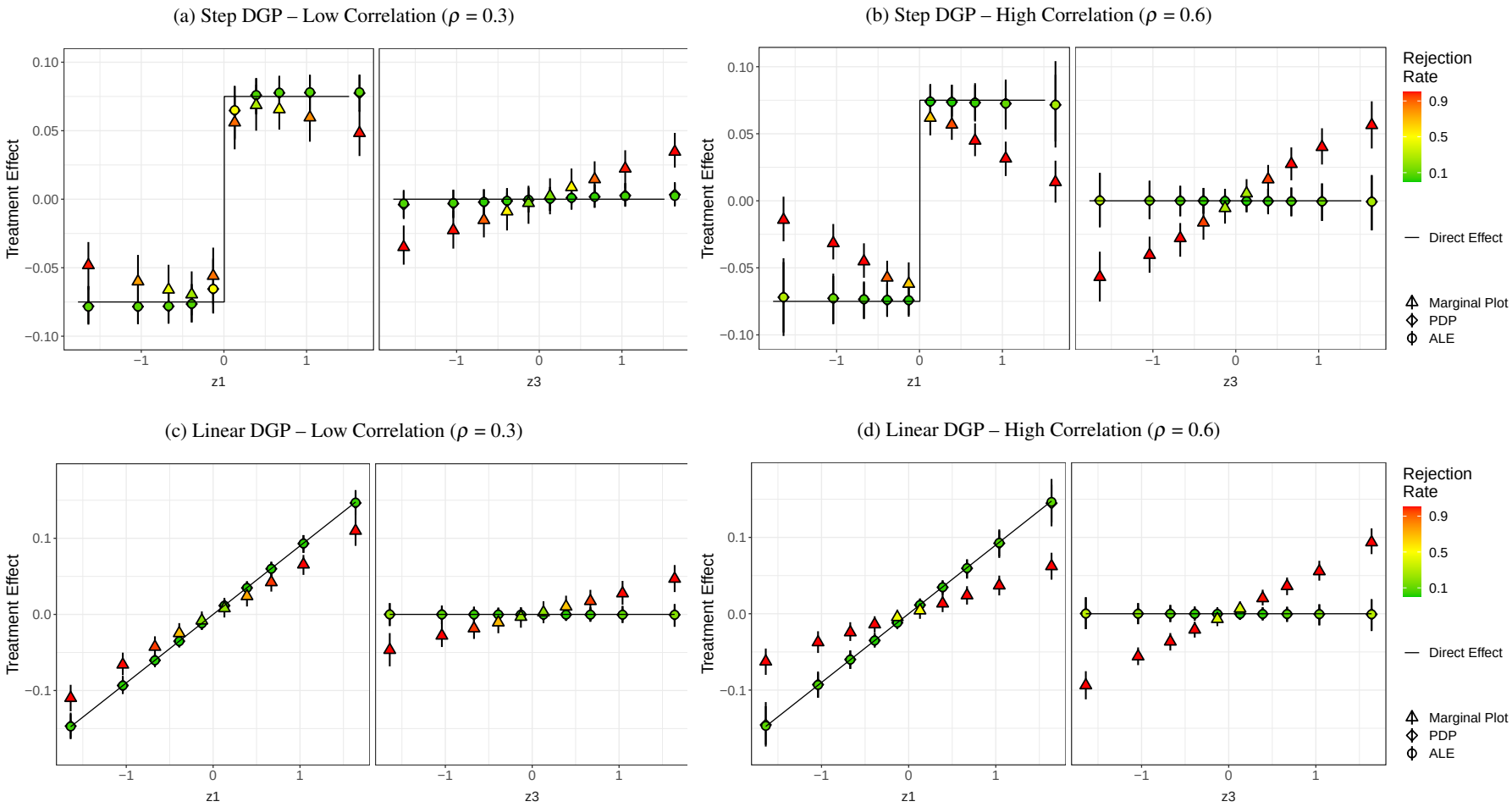
Table 1: Simulated CLAN – Relevant and Irrelevant Variables

	$\rho = 0$			$\rho = 0.6$		
	Lower Quartile (1)	Upper Quartile (2)	Difference (3)	Lower Quartile (4)	Upper Quartile (5)	Difference (6)
z_1	-0.801	0.798	1.600	-0.509	0.525	1.044
			(1.000)			(1.000)
z_2	0.797	-0.801	-1.598	0.367	-0.359	-0.716
			(1.000)			(1.000)
z_3	0.000	-0.001	-0.001	-0.816	0.833	1.643
			(0.027)			(1.000)

Notes: 500 replications. Rejection rates are in parentheses. z_3 is not in the DGP for Y, but is correlated with z_1. The DGP is defined in equations (1) and (2).

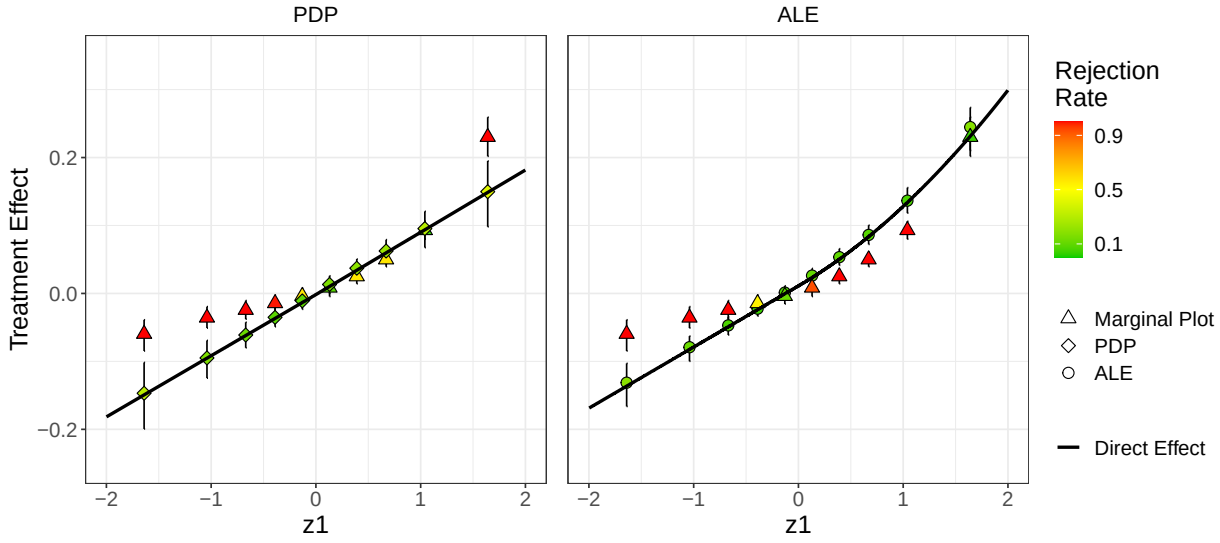
Appendix A Appendix Figures

Figure A1: Robustness – Simulation Study of Step and Linear Heterogeneous Treatment Effect Using Half the Observations



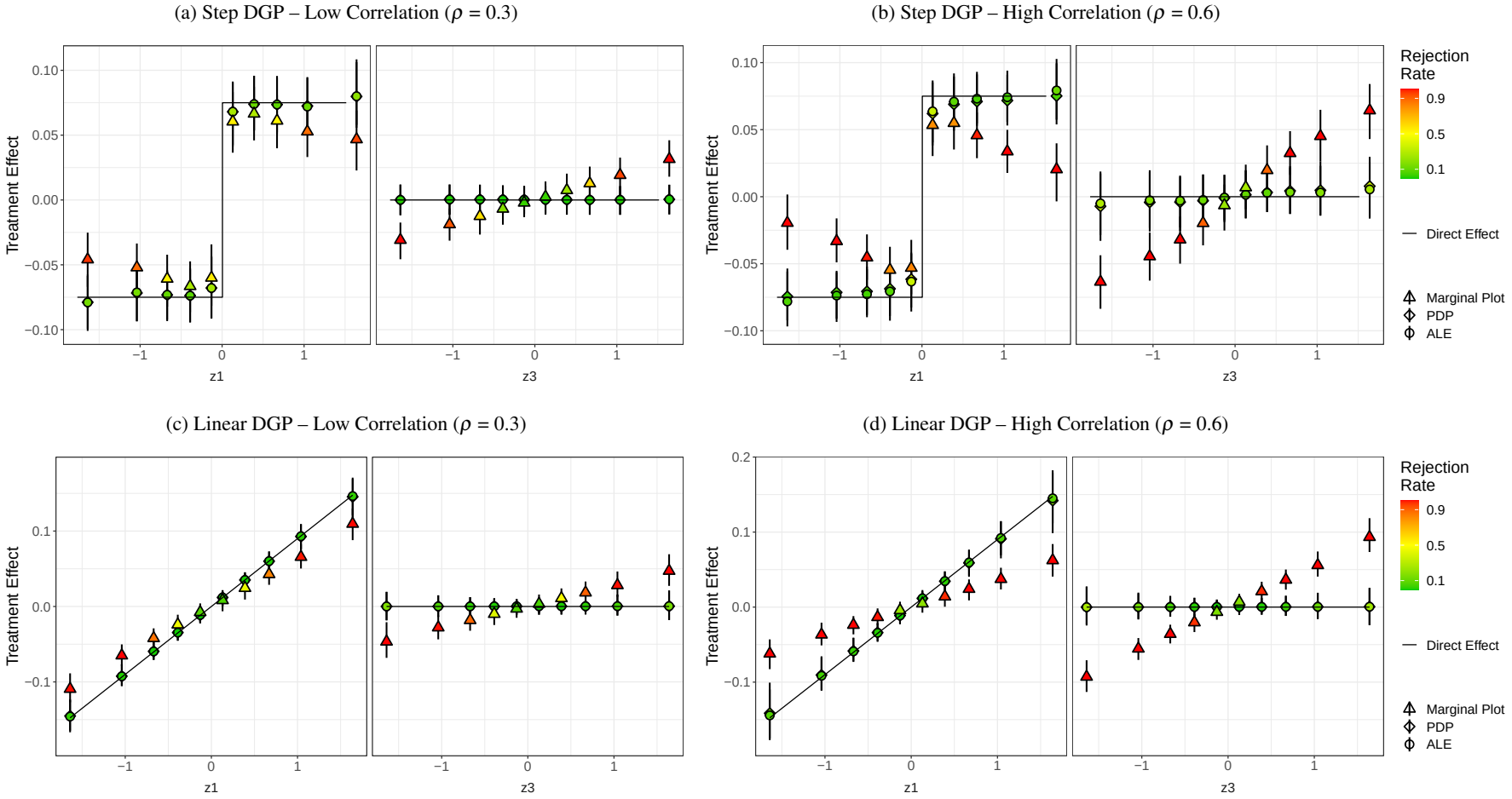
Notes: The figure shows results obtained by applying the GRF for the Step DGP with low correlation ($\rho = 0.3$), NNET for the Step DGP with high correlation ($\rho = 0.6$) and NNET for the Linear DGP as the ML proxy predictor. GRF, GBM, and NNET algorithms were applied (100 replications for GRF and 500 replications for GBM and NNET). The algorithm with the highest performance measure was chosen. Step DGP is defined in equations (1) and (2), and Linear DGP is defined in equations (1) and (8).

Figure A2: Robustness – Simulation Study of Treatment Effect with Interactions Using Half the Observations



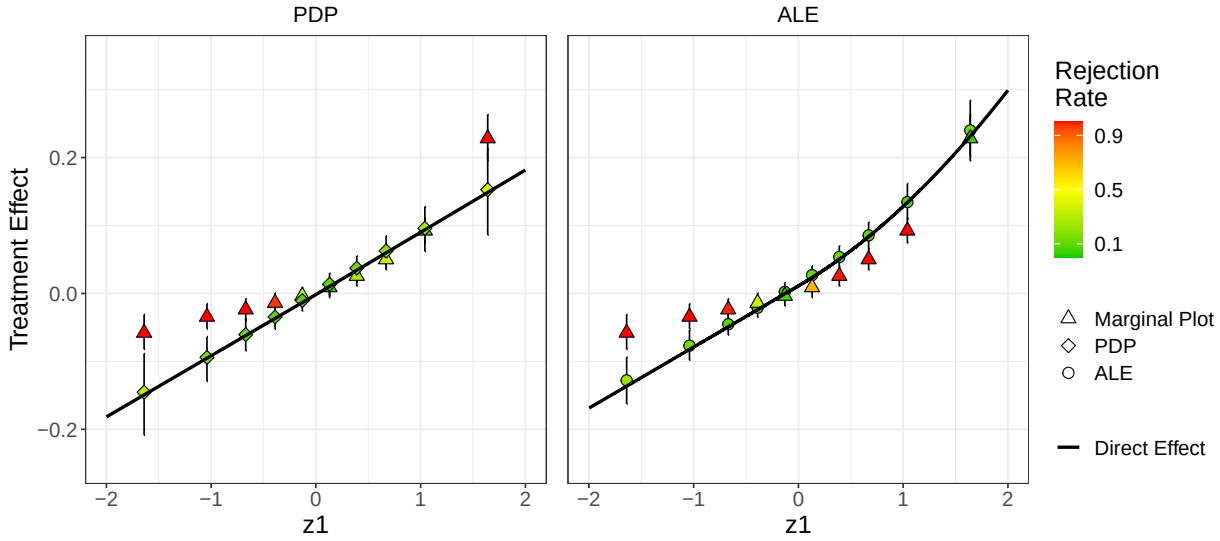
Notes: The figure shows results obtained by applying the NNET as the ML proxy predictor. The GRF, GBM, and NNET were applied (100 replications for GRF and 500 replications for GBM and NNET). The algorithm with the highest performance measure was chosen. Interactions DGP is defined in equations (1) and (9) for $\rho = 0.6$.

Figure A3: Robustness – Simulation Study of Step and Linear Heterogeneous Treatment Effect Doubling the Noise



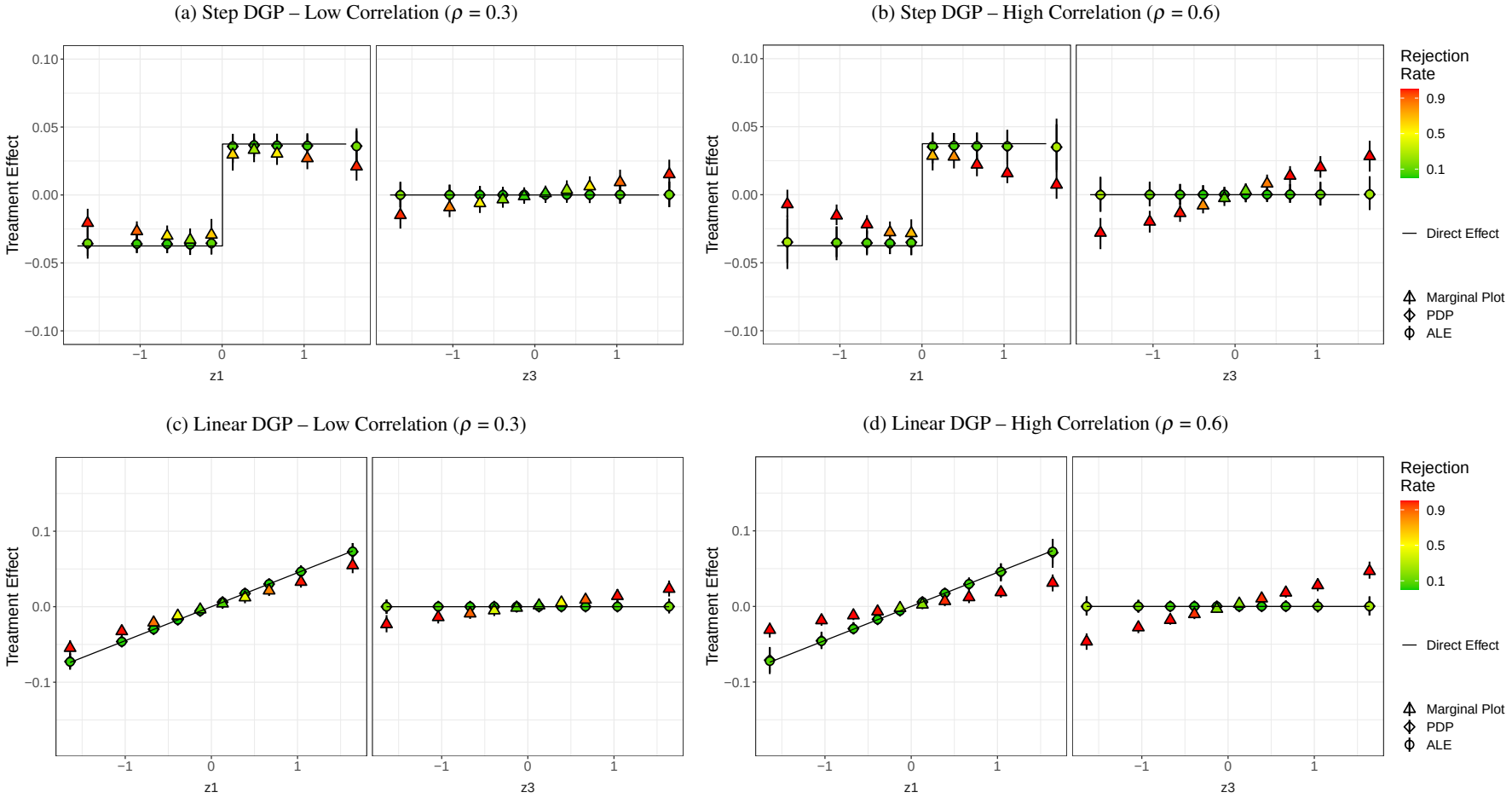
Notes: The figure shows results obtained by applying the GBM for the Step DGP and NNET for the Linear DGP as the ML proxy predictor. GRF, GBM, and NNET algorithms were applied (100 replications for GRF and 500 replications for GBM and NNET). The algorithm with the highest performance measure was chosen. Step DGP is defined in equations (1) and (2), and Linear DGP is defined in equations (1) and (8).

Figure A4: Robustness – Simulation Study of Treatment Effect with Interactions Doubling the Noise



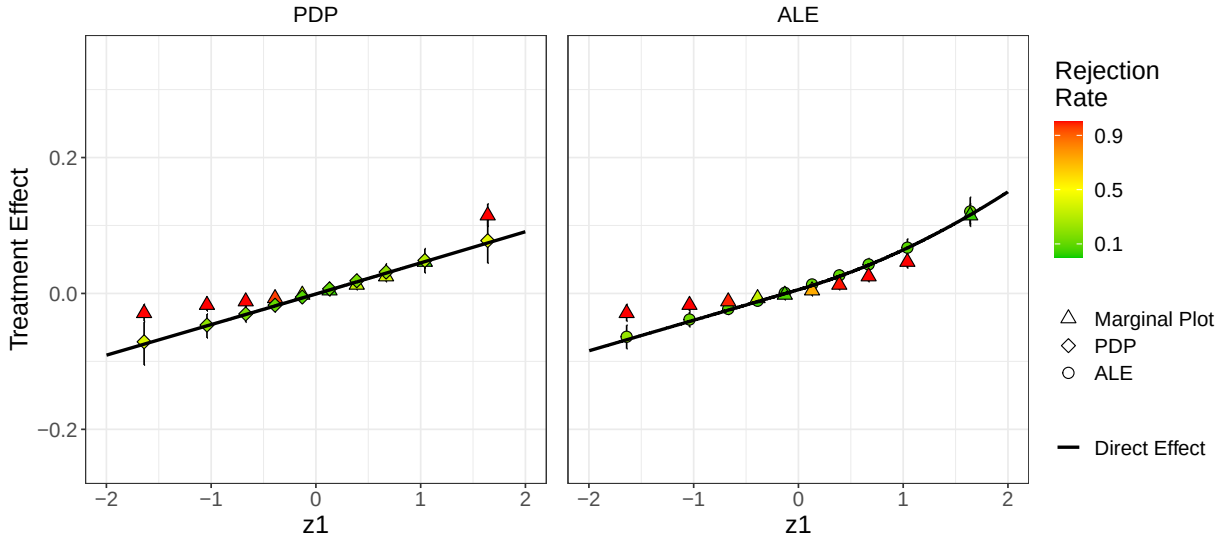
Notes: The figure shows results obtained by applying NNET as the ML proxy predictor. The GRF, GBM, and NNET were applied (100 replications for GRF and 500 replications for GBM and NNET). The algorithm with the highest performance measure was chosen. Interactions DGP is defined in equations (1) and (9) for $\rho = 0.6$.

Figure A5: Robustness – Simulation Study of Step and Linear Heterogeneous Treatment Effect Halving the Heterogeneity



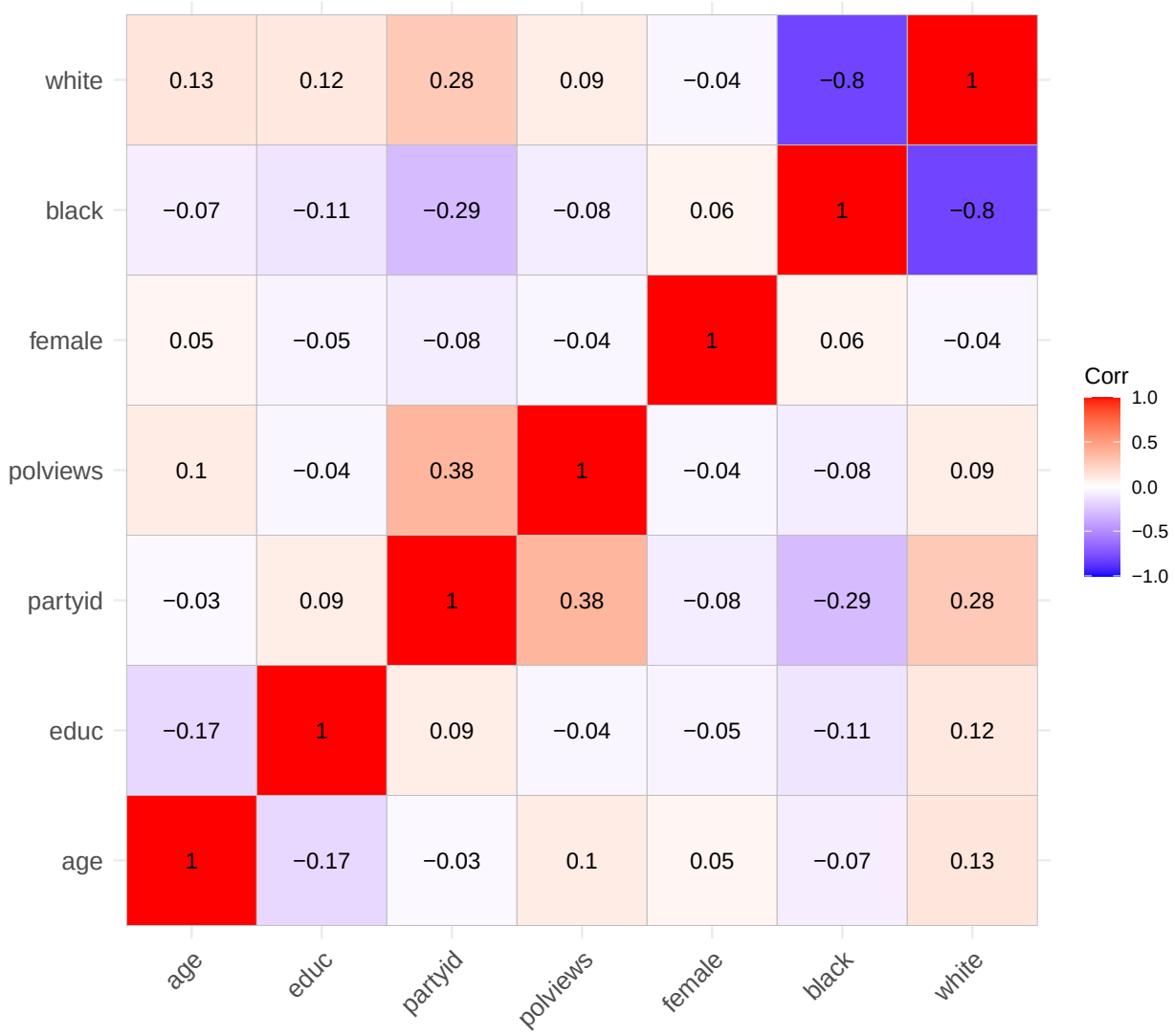
Notes: The figure shows results obtained by applying NNET as the ML proxy predictor. GRF, GBM, and NNET algorithms were applied (100 replications for GRF and 500 replications for GBM and NNET). The algorithm with the highest performance measure was chosen. Step DGP is defined in equations (1) and (2), and Linear DGP is defined in equations (1) and (8).

Figure A6: Robustness – Simulation Study of Treatment Effect with Interactions Halving the Heterogeneity



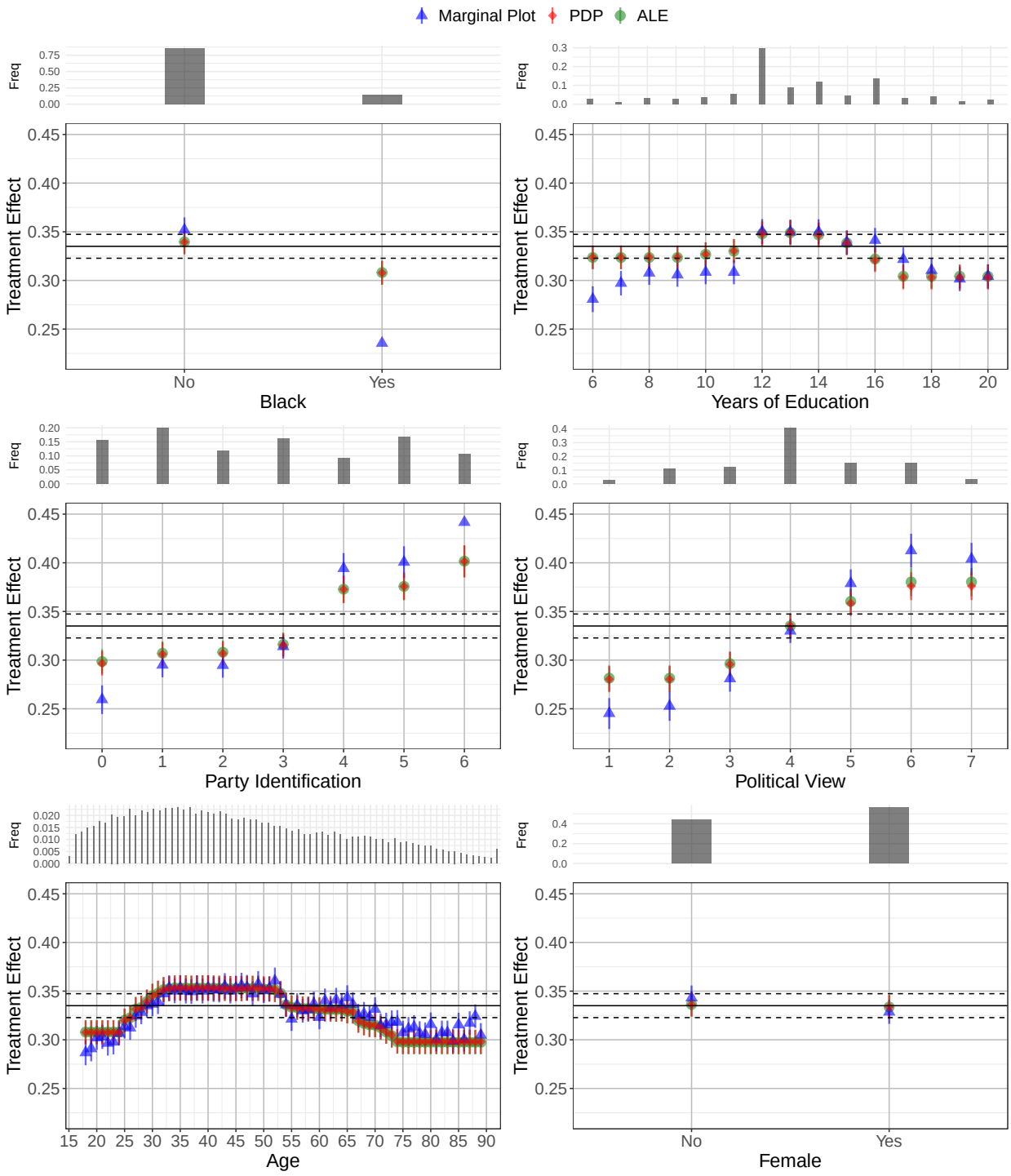
Notes: The figure shows results obtained by applying NNET as the ML proxy predictor. The GRF, GBM, and NNET were applied (100 replications for GRF and 500 replications for GBM and NNET). The algorithm with the highest performance measure was chosen. Interactions DGP is defined in equations (1) and (9) for $\rho = 0.6$.

Figure A7: Covariates' Correlation Matrix



Notes: The matrix reports the covariance between the variables used in the analysis: age, years of education, party identification coded on a scale of 0 to 6 (higher values denote affiliation with the Republican party), political views coded on a scale of 1 to 7 (higher values denote more conservative views), gender, and race (White, Black, and other).

Figure A8: Application – Marginal Plots, PDP, and ALE



Notes: The black horizontal lines represent the estimated ATE $\hat{\beta}_1$ (solid) and its 90% confidence interval (dashed). The figure shows results for GBM due to highest performance measure.

Appendix B Appendix Tables

Table B1: Treatment – Control Comparison

	Treatment Mean (1)	Control Mean (2)	Difference (3)
Age	46.01	46.21	-0.25
			(0.27)
Years of Education	13.16	13.15	0.02
			(0.05)
Party Identification	2.74	2.77	-0.02
			(0.02)
Political Views	4.12	4.12	0.00
			(0.01)
Female	0.57	0.56	0.01
			(0.01)
Black	0.14	0.14	0.00
			(0.00)
White	0.79	0.8	0.00
			(0.00)
N	17,567	15,247	

Notes: The table reports means for the treatment and control groups and the mean difference between the groups controlling for year of survey. Clustered standard errors (at the survey-year level) are reported in parentheses.

Table B2: Application – BLP Results

	Generalized Random Forest (1)	Gradient Boosting (2)	Neural Network (3)
ATE (β_1)	0.335	0.335	0.335
	[0.322, 0.347]	[0.323, 0.347]	[0.323, 0.347]
Heterogeneity (β_2)	1.313	0.740	0.631
	[1.132, 1.494]	[0.639, 0.839]	[0.496, 0.774]
Performance (Λ)	0.0074	0.0080	0.0060

Notes: Point estimates are medians over 100 splits. 90% confidence intervals in brackets. Each algorithm is trained with a sample consisting of 32,814 observations and includes the following variables: survey year dummies, age, years of education, party identification, political view, gender, and race.

Appendix C Proofs

In this appendix we prove that mean ALE and PDP slopes w.r.t. a single covariate are equivalent when either (i) covariates are independent, or (ii) when there is no interactions of covariates within the treatment effect function. This implies that they trace out the same function, up to a level-shift constant. We define

$$PD^{PDP}\left(z_l^*\right) = E_{z_{-l}}\left[\frac{\partial \hat{s}\left(z_l^*, z_{-l}\right)}{\partial z_l}\right],$$
$$PD_l^{ALE}\left(z_l^*\right) = E_{z_{-l}}\left[\frac{\partial \hat{s}\left(z_l, z_{-l}\right)}{\partial z_l} \mid z_l = z_l^*\right],$$

where $\mathbf{z}$ is a covariate vector, z_l is a single covariate, and z_{-l} are all the covariates in $\mathbf{z}$ exept z_l, and $\hat{s}$ is the estimated CATE function. The integral over each of these definitions corresponds to equations (3) and (5), and so the equivalence of $PD^{PDP}\left(z_l^*\right)$ and $PD_l^{ALE}\left(z_l^*\right)$ implies that (3) and (5) are equivalent, up to a constant.

Proposition C.1. *If z_l and z_{-l} are independent then $PD_l^{PDP}\left(z_l^*\right) = PD_l^{ALE}\left(z_l^*\right)$.*

Proof. Assume that z_l and z_{-l} are independent. Then

$$\begin{aligned}
PD_l^{ALE}\left(z_l^*\right) &= E_{z_{-l}}\left[\frac{\partial \hat{s}\left(z_l, z_{-l}\right)}{\partial z_l} \mid z_l = z_l^*\right] \\
&= \int_{z_{-l}} \frac{\partial \hat{s}\left(z_l^*, z_{-l}\right)}{\partial z_l} f_{z_{-l}|z_l=z_l^*} dz_{-l} \\
&= \int_{z_{-l}} \frac{\partial \hat{s}\left(z_l^*, z_{-l}\right)}{\partial z_l} f_{z_{-l}} dz_{-l} \\
&= E_{z_{-l}}\left[\frac{\partial \hat{s}\left(z_l^*, z_{-l}\right)}{\partial z_l}\right] \\
&= PD^{PDP}\left(z_l^*\right).
\end{aligned}$$

□

Proposition C.2. *If $\hat{s}$ is an additive function of z_l and z_{-l} without interactions of the form $\hat{s}\left(z_l, z_{-l}\right) = h\left(z_l\right) + g\left(z_{-l}\right)$ then $PE^{PDP}\left(z_l^*\right) = PE_l^{ALE}\left(z_l^*\right)$.*

Proof. Assume that $\hat{s}(z_l, z_{-l}) = h(z_l) + g(z_{-l})$. Starting from $PD_l^{ALE}\left(z_l^*\right)$, we get

$$\begin{aligned} PD_l^{ALE}(z_l^*) &= E_{z_{-l}}\left[\frac{\partial\left\{h(z_l) + g(z_{-l})\right\}}{\partial z_l} \mid z_l = z_l^*\right] \\ &= E_{z_{-l}}\left[\frac{\partial h(z_l)}{\partial z_l} + \frac{\partial g(z_{-l})}{\partial z_l} \mid z_l = z_l^*\right] \\ &= E_{z_{-l}}\left[\frac{\partial h(z_l)}{\partial z_l} \mid z_l = z_l^*\right] \\ &= \frac{\partial h\left(z_l^*\right)}{\partial z_l}. \end{aligned}$$

And starting from $PD^{PDP}\left(z_l^*\right)$ we get the same expression:

$$\begin{aligned} PD^{PDP}(z_l^*) &= E_{z_{-l}}\left[\frac{\partial\left\{h\left(z_l^*\right) + g(z_{-l})\right\}}{\partial z_l}\right] \\ &= E_{z_{-l}}\left[\frac{\partial h\left(z_l^*\right)}{\partial z_l}\right] \\ &= \frac{\partial h\left(z_l^*\right)}{\partial z_l}. \end{aligned}$$

□

Appendix D Short review of Chernozhukov et al. (2020) Algorithm

D.1 The Best Linear Predictor (BLP) of the CATE

We start by providing some notation. Let Y, W, and $\mathbf{z}$ denote the outcome variable, the treatment variable, and the covariate vector, respectively. Under the assumption of conditional random assignment, the CATE is identified by

$$s_0(\mathbf{z}) = \mathbb{E}(Y|W=1,\mathbf{z}) - \mathbb{E}(Y|W=0,\mathbf{z}), \tag{D.1}$$

and the propensity score is given by[18]

$$p(\mathbf{z}) = P(W=1|\mathbf{z}). \tag{D.2}$$

In addition, we denote the ML estimator of the CATE by $S(\mathbf{z})$. This ML estimator can be obtained by any ML method (see the discussion below about our implementation). Chernozhukov et al. (2020) refer to this estimator as a proxy predictor of CATE, and derive the statistical properties of the coefficients of the Best Linear Predictor (BLP) of the CATE given the ML proxy. The BLP of the CATE is defined as

$$BLP(s_0(\mathbf{z}) \mid S(\mathbf{z})) = \beta_1 + \beta_2 \left(S(\mathbf{z}) - \mathbb{E}[S(\mathbf{z})]\right). \tag{D.3}$$

The idea in Chernozhukov et al. (2020) is that instead of focusing on obtaining a consistent estimator for CATE, we can consistently estimate the BLP of CATE given a proxy predictor. This is appealing for two reasons. First, BLP coefficients have an important interpretation: β_1 is the Average Treatment Effect (ATE), and β_2 is interpreted as a measure of both the presence of heterogeneity and the relevance of the proxy $S(\mathbf{z})$ as a predictor of $s_0(\mathbf{z})$, in the sense that $\beta_2 = 0$ suggests that the correlation between $S(\mathbf{z})$ and $s_0(\mathbf{z})$ is 0. Second, the BLP is the personalized prediction of $s_0(\mathbf{z})$, which, as described below, can be used to explore treatment effect heterogeneity with a wide variety of LM algorithms and with no requirement for the proxy to be a consistent estimate for CATE.

Estimation of equation (D.3) involves a three-step procedure: (1) split the data randomly into a main sample and an auxiliary sample; (2) fit the proxy predictor of CATE using the auxiliary sample; (3) proceed to the main sample and exploit the proxy obtained in the previous step to estimate the following weighted linear projection, which identifies the coefficients of the BLP:[19]

$$\begin{aligned} Y &= \alpha_0 + \alpha_1 B(\mathbf{z}) + \beta_1 \big(W - p(\mathbf{z})\big) + \beta_2 \big(W - p(\mathbf{z})\big)\big(S(\mathbf{z}) - \mathbb{E}[S(\mathbf{z})]\big) + \varepsilon, \\ q(z) &= \frac{1}{p(\mathbf{z})(1-p(\mathbf{z}))}, \end{aligned} \tag{D.4}$$

where $q(z)$ are the weights in the linear regression, and $B(\mathbf{z})$ is the proxy predictor estimate for $b_0(\mathbf{z})$, the baseline

[18] The propensity score depends on the entire set of variables $\mathbf{z}$, or on a subset of thereof. In some applications the propensity score is a known function. In cases where the propensity score is unknown to the experiment designer, it can be computed as a preliminary stage. See, e.g., Section 6.3 in Chernozhukov et al. (2020).

[19] Chernozhukov et al. (2020) provide two different specifications for estimation of equation (D.3). Here we describe a short version of the first specification (equation (2.1) in their paper).

value of the outcome, defined as $b_0(\mathbf{z}) = E[Y|W=0,\mathbf{z}]$. Chernozhukov et al. (2020) show that $\hat{\beta}_1$ and $\hat{\beta}_2$ obtained from estimating equation (D.4) are consistent estimates for β_1 and β_2 from equation (D.3), and can be used to (1) calculate the estimated ATE (i.e., $\hat{\beta}_1$), (2) test for the presence of heterogeneity and whether $S(\mathbf{z})$ is its relevant predictor by assessing the null hypothesis $\beta_2 = 0$, and (3) calculate the estimated BLP of CATE.

As described above, the identification of BLP relies on random sample splitting, which introduces an additional source of uncertainty as different splits lead to different estimates. To account for that uncertainty, Chernozhukov et al. (2020) suggest the following method, which they refer to as Variational Estimation and Inference (VEIN): repeat the estimation process many times. For each data partition calculate an estimate for the parameter of interest θ, a p-value for the null hypothesis $\theta = \theta_0$, and a confidence interval (CI) for θ. Then report the median estimate over the splits as a point estimate, and the median upper and lower bounds of a $1-\alpha$ CI as a CI with a confidence level of $1-2\alpha$. Reject the null hypothesis $\theta = \theta_0$ at a significance level of α if the median p-value is lower than $\alpha/2$.

D.2 Implementation Algorithms

We describe in this subsection our implementation algorithm for producing CATE estimates over one dimension.

1. Calculating the propensity score. In the simulations (Sections 2 and 3.5) we assign the known treatment probability for each observation while in the application we compute the propensity score by using a logistic regression of W on $\mathbf{z}$ and survey year dummies.

2. Training the proxy predictor. We split the data randomly into main and auxiliary samples. Using only the auxiliary samples, we train three proxy predictors with Generalized Random Forests (GRF), Gradient Boosting (GBM), and Neural Network (NN).[20] For the GBM and NN algorithms we take a T-learner approach, where we fit two ML models to predict Y based on $\mathbf{z}$, first using only the treated observations, and second using the untreated observations. We then predict for observations in the main sample the proxy predictor $S(z)$ using the difference in predictions between the model trained on the treated observations and the model trained on the untreated observations. We also estimate the baseline predictor $B(z)$ using the prediction of the model trained on the untreated observations. For the GRF algorithm, we fit a causal forest model to produce estimates for $S(z)$. For this case, we estimate the baseline predictor by estimating the outcome and treatment (using a regression forest) and subtracting the outcome from the estimated treatment propensity and the CATE proxy predictor: $B(z) := \hat{Y}(z) - \hat{W}(z) \times S(z)$.[21]

3. Classification Analysis (CLAN). Using the proxy predictor $S(z)$, we split the observations in the main sample into K equally sized groups where group 1 consists of observations with the lowest CATE and group K consists of observations with the highest CATE. Using the subsample of observations belonging to the lowest and highest groups, i.e., observations in groups 1 and K, we calculate the mean of each covariate in $\mathbf{z}$ in each group and their difference,

[20] All algorithms were implemented in R. For GRF we use the `grf` package and the default tuning option `tune.paramters = "all,"` which tunes the following hyper-parameters using 100 cross-validated repitions: `min.node.size` (minimum node size), `sample.fraction` (fraction of sample used in each tree), `mtry` (number of variables considered in each split), `honesty.fraction` (fraction of sample used to grow vs. populate leaves), `honesty.prune.leaves` (whether to prune empty leaves in honest splitting), `alpha` (the maximum imbalance of a split), and `imbalance.penalty` (how harshly imbalanced splits are penalized). For gradient boosting we use the `gbm` package and tune the following hyper-parameters: `shrinkage` (regularization parameter), `interaction.depth` (tree depth), `n.minobsinnode` (minimum node size), and `n.trees` (number of trees). For neural network we use the `nnet` package and tune the following hyper-parameters: `size` (hidden layer size), `decay` (regularization parameter), and `linout` (activation function).

[21] Baseline prediction taken from GRF tutorials: https://grf-labs.github.io/grf/articles/muhats.html.

which allows the comparison of the mean of each covariate between those with the smallest and largest predicted treatment effects.

4. Estimating the performance of the proxy predictor. For each data split we estimate equation (D.4) in the main sample and obtain $\hat{\beta}_1$ and $\hat{\beta}_2$ for each proxy predictor separately. We also obtain a point estimate for the performance measure $\Lambda = |\beta_2|^2 var(S(\mathbf{z}))$, which is informative on the fit in the regression of the proxy predictor of the true CATE. To see this, note that

$$\beta_2 = \frac{cov(S(z), s_0(z))}{var(S(z))} = cor(S(z), s_0(z)) \times \frac{\sqrt{var(s_0(z))}}{\sqrt{var(S(z))}}$$

and so $\Lambda = [cor(S(z), s_0(z))]^2 \times var(s_0(z))$. Since $var(s_0(z))$ is constant, choosing the proxy predictor that maximizes Λ is equivalent to choosing the proxy predictor with the highest correlation with the true CATE. We estimate $\hat{\Lambda}$ using $\hat{\beta}_2$ and the known variance $\hat{\Lambda} = (\hat{\beta}_2)^2 \times var(S(\mathbf{z}))$.

5. Calculating marginal plots. For each data partition we use $\hat{\beta}_1$ and $\hat{\beta}_2$ to calculate the personalized BLP prediction (equation (D.3)). In order to draw the marginal plot over z_l, we first group observations into bins in the following way: if z_l is a continuous variable, we divide the support of z_l into deciles and group together observations whose z_l falls in the same decile. Otherwise, in cases where z_l is a discrete variable, we group together observations that have the same value of z_l. We denote by z_l^* a specific bin of z_l and calculate for each z_l^* the average BLP prediction and its CI, using the fact that standard errors can be calculated using $V(\hat{\beta})$ as follows:

$$\begin{aligned} &var\left(\frac{1}{N(z_l^*)}\sum_{i:\, z_{l,i}\in z_l^*}\left[\hat{\beta}_1 + \hat{\beta}_2\big(S(\mathbf{z}_i) - \mathbb{E}[S(\mathbf{z})]\big)\right] \mid S(\cdot), \mathbf{z}\right) = \\ &var(\hat{\beta}_1) + var(\hat{\beta}_2)\Big(\frac{1}{N(z_l^*)}\sum_{i:\, z_{l,i}\in z_l^*} S(\mathbf{z}_i) - \mathbb{E}[S(\mathbf{z})]\Big)^2 + 2cov(\hat{\beta}_1, \hat{\beta}_2)\Big(\frac{1}{N(z_l^*)}\sum_{i:\, z_{l,i}\in z_l^*} S(\mathbf{z}_i) - \mathbb{E}[S(\mathbf{z})]\Big), \end{aligned} \tag{D.5}$$

where $N(z_l^*)$ is the number of observations that fall into z_l^*.

6. Using the VEIN procedure. We iterate steps 2–5 100 times, and draw the median estimates of β_1, β_2, Λ and the personalized predictions and their respective median CI (with an adjusted confidence level).

Appendix E ALE Implementation

In this appendix we present for completeness the ALE estimator for a generic function f as proposed in Apley and Zhu (2020). Using the notation in the main text, for the generic function f the ALE estimand is

$$\bar{f}_{l,ALE}(z_l^*) = \int_{z_{l,min}}^{z_l^*} \mathbb{E}\Big[\frac{\partial f(z_l, z_{-l})}{\partial z_l} \mid z_l = t\Big] dt - c. \tag{E.6}$$

Let $\hat{f}$ denote an estimator of f. Let $P_l^M = \{x_l^m : m = 0, 1, ..., M\}$ be a fixed partition of the support of z_l into M intervals such that $x_l^0 = z_{l,min}$ and $x_l^M = z_{l,max}$. Let $N_l(m)$ and $n_l(m)$ denote the mth interval and the number of observations that fall into it, i.e., $N_l(m) = (x_l^{m-1}, x_l^m)$ and $\sum_{m=0}^{M} n_l(m) = N$. Finally, let $m_l(z_l^*)$ denote the index of the interval into which z_l^* falls, i.e., $z_l^* \in N_l(m_l(z_l^*))$. Then Apley and Zhu (2020) suggest estimating the uncentered component of equation (E.6) using

$$\hat{g}_{l,ALE}(z_l^*) = \sum_{m=1}^{m_l(z_l^*)} \frac{1}{n_l(m)} \sum_{\{i:\, z_{i,l} \in N_l(m)\}} \Big(\hat{f}(x_l^m, z_{i,-l}) - \hat{f}(x_l^{m-1}, z_{i,-l})\Big) \tag{E.7}$$

and to choose the constant c to be an average of $\hat{g}_{l,ALE}(z_l^*)$ over z_l, i.e.,

$$\hat{\bar{f}}_{l,ALE}(z_l^*) = \hat{g}_{l,ALE}(z_l^*) - \frac{1}{N}\sum_{i=1}^{N} \hat{g}_{l,ALE}(z_{i,l}) \tag{E.8}$$

so that the plot of $\hat{\bar{f}}_{l,ALE}(z_l^*)$ will be centered around zero.

Appendix F Implementing PDP and ALE to BLP and VEIN

Here we follow again steps 1–3 from Appendix D.2 and estimate $\hat{\beta}_1$ and $\hat{\beta}_2$. Then, we use the BLP function to study how the estimated CATE changes when changing the value of a single variable z_l, while holding all other variables constant, according to the implemented method.

F.1 Applying PDP to the BLP of CATE

To study the direct relation between the estimated treatment effect and z_l using PDP and ALE, we apply the PDP and ALE approaches to the estimated $\widehat{BLP}(\mathbf{z})$.[22] Starting with PDP, we implement this approach by substituting the estimated BLP into the PDP estimator. Formally,

$$\widehat{\overline{BLP}}_{l,PDP}(z_l^*) = \frac{1}{N}\sum_{i=1}^{N}\widehat{BLP}(z_l^*, z_{i,-l}) = \hat{\beta}_1 + \hat{\beta}_2\Big(\frac{1}{N}\sum_{i=1}^{N} S(z_l^*, z_{i,-l}) - \mathbb{E}[S(\mathbf{z})]\Big). \tag{F.9}$$

To keep the PDP estimation comparable to the marginal plots, we obtain point estimates and confidence intervals following the same VEIN approach of Chernozhukov et al. (2020) as described in Appendix D.1. The computation of confidence intervals *within* a split requires the estimation of standard errors (within a split), which we derive as follows:

$$\begin{aligned} &var\Big(\widehat{\overline{BLP}}_{l,PDP}(z_l^*) \mid S(\cdot), \mathbf{z}\Big) = var\Big(\hat{\beta}_1 + \hat{\beta}_2\big(\frac{1}{N}\sum_{i=1}^{N} S(z_l^*, z_{i,-l}) - \mathbb{E}[S(\mathbf{z})]\big) \mid S(\cdot), \mathbf{z}\Big) \\ &= var(\hat{\beta}_1) + var(\hat{\beta}_2)\Big(\frac{1}{N}\sum_{i=1}^{N} S(z_l^*, z_{i,-l}) - \mathbb{E}[S(\mathbf{z})]\Big)^2 + 2cov(\hat{\beta}_1, \hat{\beta}_2)\Big(\frac{1}{N}\sum_{i=1}^{N} S(z_l^*, z_{i,-l}) - \mathbb{E}[S(\mathbf{z})]\Big), \end{aligned} \tag{F.10}$$

where the covariance matrix of $\hat{\beta}$ is obtained in the estimation of the BLP. The equality in the second line of (F.10) derives from the fact that $S(\mathbf{z})$ is calculated using the auxiliary sample and, thus, is uncorrelated with $\hat{\beta}$.

The computation itself is conducted as follows. We first start by fixing a set of values from the support of z_l. In the application, where all covariates are discrete, we essentially use the entire support. In the simulations, where all covariates are continuous, we use the median value of each decile, so that the results will be comparable to the marginal plots. Then, for each value z_l^* in the set, we use the main sample to calculate $\widehat{\overline{BLP}}_{l,PDP}(z_l^*)$ using equation (F.9) and its CI using $var(\widehat{\overline{BLP}}_{l,PDP}(z_l^*))$ defined in equation (F.10). Finally, we aggregate the results over the splits using the VEIN method and draw the median estimate and the median CI (with an adjusted confidence level).

[22] As a result, we use the estimated $\widehat{BLP}(\mathbf{z})$ as our estimator for $s_0(\mathbf{z})$. Having said that, recall that this is a consistent estimator for the BLP of $s_0(\mathbf{z})$, rather than directly for $s_0(\mathbf{z})$.

F.2 Applying ALE to the BLP of CATE

As with PDP, we suggest applying ALE to the BLP of CATE. First, we define the population ALE of BLP by substituting the BLP function (7) into (5):

$$\begin{aligned}\overline{BLP}_{l,ALE}(z_l^*) &= \int_{z_{l,min}}^{z_l^*} \mathbb{E}\Big[\frac{\partial\left(\beta_1+\beta_2\left(S(z_l,z_{-l})-\mathbb{E}[\mathbf{z}]\right)\right)}{\partial z_l} \mid z_l=t\Big]dt-c \\ &= \beta_2\int_{z_{l,min}}^{z_l^*} \mathbb{E}\Big[\frac{\partial S(z_l,z_{-l})}{\partial z_l} \mid z_l=t\Big]dt-c=\beta_2\bar{S}_{l,ALE}(z_l^*)-c. \end{aligned} \tag{F.11}$$

As can be seen, applying ALE to the BLP of CATE is equivalent to first applying ALE to the proxy function S and then multiplying the result by β_2. Hence, to estimate the ALE of BLP we apply the estimation procedure described in Appendix E for the proxy function S and then multiply the results by $\hat{\beta}_2$. In addition, we suggest centralizing the results around the estimated ATE rather than zero, and so the final estimate of equation (F.11) is

$$\widehat{\overline{BLP}}_{l,ALE}(z_l^*)=\hat{\beta}_1+\hat{\beta}_2\hat{\bar{S}}_{l,ALE}(z_l^*). \tag{F.12}$$

As before, to obtain the point estimates and confidence intervals of the ALE of the estimated CATE we apply VEIN. That is, we use the within-split covariance matrix of $\hat{\beta}$ to calculate within-split standard errors, using again the fact that $S(\mathbf{z})$ is uncorrelated with $\hat{\beta}$.

While the focus of this subsection is on the derivation of PDP and ALE within the Chernozhukov et al. (2020) framework, it is worth noting that in the case where treatment effect heterogeneity is estimated using GRF, and because GRF provides consistent estimates directly to the CATE, PDP and ALE can be applied directly to the prediction function $\hat{s}_0$ estimated by GRF to explore treatment effect heterogeneity rather than to the BLP.

We compute the ALE estimates as follows. In order for ALE to be computationally feasible and comparable to previous methods, we always apply ALE to discrete covariates. In the case where z_l is continuous, we first discretize it by replacing each value in the support of z_l with the median value of the decile to which this value belongs. For the sake of clarity, we describe here a discrete version of the algorithm described in Section 3.2. Let $P_l^M=\{x_l^m : m=0,1,...,M\}$ be the ordered support of the discrete covariate z_l. We initialize $\hat{g}_{l,ALE}(x_l^0)=0$ and calculate for each $m>0$ the discrete version of equation (E.7) for the function $S(\mathbf{z})$ using the main sample:

$$\hat{g}_{l,ALE}(x_l^m)=\sum_{k=1}^{m}\frac{1}{n_l(k)}\sum_{\{i:\ z_{i,l}\in[x_l^k,x_l^{k-1}]\}}\Big(S(x_l^k,z_{i,-l})-S(x_l^{k-1},z_{i,-l})\Big), \tag{F.13}$$

where $n_l(k)$ is the number of observations that fall into the interval $[x_l^k,x_l^{k-1}]$. In the next step we centralize $\hat{g}_{l,ALE}(\cdot)$ by calculating the discrete version of equation (E.8) as follows:[23]

[23] In the R package `ALEPlot`, which implements Apley and Zhu (2020), this is done a bit differently as the package always assumes that z_l is a continuous variable.

$$\hat{\tilde{S}}_{l,ALE}(x_l^m) = \hat{g}_{l,ALE}(x_l^m) - \frac{1}{N}\sum_{i=1}^{N} \hat{g}_{l,ALE}(z_{i,l}), \tag{F.14}$$

where $\hat{g}_{l,ALE}(z_{i,l})$ is well defined as z_l is a discrete variable.

At this stage we finish applying ALE to the proxy predictor $S(\mathbf{z})$. The final step is to calculate $\widehat{\overline{BLP}}_{l,ALE}(x_l^m)$ using equation (F.12), which is the ALE of BLP centered around the ATE, and its CI. As before, we aggregate the results over the splits using the VEIN method and draw the median estimate and the median CI (with an adjusted confidence level).

Appendix G General Social Survey Questions

Below we provide the full wording of the questions used in the GSS wording experiment.

The question reads:

> "We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount."

This was followed by:

> "are we spending too much, too little, or about the right amount on (ITEM)?"

(ITEM) was replaced with different spending categories. Specifically, in the experiment, for some respondents, which were randomly chosen, (ITEM) was replaced with "welfare" (defined in our analysis as the treatment group), and for others with "assistance to the poor" (the control group).