Erasmus University Rotterdam
Erasmus School of Economics

Master Thesis Quantitative Marketing & Business Analytics

The Added Value of Machine Learning to Economics: Evidence from Revisited Studies

Menno de Jong
413841

Supervisor: dr. A.A. Naghi

Second Assessor: dr. M. Zhelonkin

September 21, 2018

Abstract

Recently, promising new methods based on Machine Learning (ML) have been introduced to conduct causal inference, see for an overview Section 4 of Athey (2017). These methods can pick up complex and high-dimensional nuisance relationships such that they improve plausibility of the widely used unconfoundedness and IV assumptions that identify causal effects. Accordingly, causal inference might become more credible than with established methods like difference-in-difference, OLS, fixed effects or two stage least squares estimation. However, the merits of these new methods in empirical applications have not been studied yet. The purpose of this paper is therefore to employ ML-based methods to conduct causal inference on the Average Treatment Effect (ATE) and on Heterogeneous Treatment Effects (HTEs) by revisiting two well-known and well-cited applied papers. We compare to the original results the new results from different ML-based causal inference methods for the ATE (Double Machine Learning and Approximate Residual Balancing). We find that our ML-based methods for the ATE give estimates that deviate from original ones. This implies that these methods might improve causal inference from established methods to a great extent. Then, we extend the original papers by applying different ML-based methods for HTEs (heterogeneous DML and Causal Forests). This gives us additional relevant findings which could not have been obtained with established causal inference techniques.

Keywords: Causal Inference, Machine Learning, Average Treatment Effect, Heterogeneous Treatment Effects, Unconfoundedness, IV

1 Introduction

Causal inference has long been a challenging task in the economic literature. This stems from the limited possibilities to conduct randomized controlled experiments, partly due to political or ethical reasons. Hence, the applied economist has to rely mostly on observational data. Causal effect estimates can still be obtained from observational data if we are willing to make identifying assumptions. For the relationship between education and earnings, this is illustrated nicely by the famous study of Angrist and Krueger (1991). They argue that variation in school start age due to season of birth together with fixed compulsory schooling age by law generates a natural experiment. This natural experiment can in turn be exploited to estimate the causal effect of education on earnings. Two widely used identifying assumptions for causal inference are unconfoundedness and valid instrumental variables (IVs). Both assumptions involve the relationships between other variables on the one hand and the causal variable of interest and outcome variable on the other hand, which make causal effect estimates invalid if they are not taken into account. These relationships are called nuisance or confounding relationships and the invalidating effects on the causal effect estimate nuisance or confounding effects. Unconfoundedness states that, when taking into account observed covariates by using a control specification of some functional form, all confounding effects are captured. This basically means that there are only confounding effects from observed variables. IVs capture possible confounding effects from unobserved variables as well when they are valid, which mainly requires that they satisfy the exclusion assumption. This assumption states that the IVs cannot determine the outcome other than through the causal variable, again after controlling for observed covariates in some functional form.

Since the identifying assumptions cannot be tested empirically, their credibility in practice often remains debatable. Recently, however, there has been considerable interest in applying Machine Learning (ML) methods to causal inference, see Section 4 of Athey (2017). ML methods excel at using data to select functional forms flexibly, by trading off regularization against overfitting. This enables them to pick up arbitrary complex, possibly nonlinear nuisance relationships. ML methods can also handle large amounts of covariates in an efficient manner, allowing the researcher to correct efficiently for high-dimensional nuisance relationships. Therefore, in a setting where these complex or high-dimensional nuisance relationships might be present, ML methods could be employed to improve credibility of the unconfoundedness and exclusion assumptions, because they allow for fully flexible control specifications instead of simple, linear control specifications.

However, directly applying ML in this complex or high-dimensional setting leads to biased causal effect estimates with a slower than $n^{-1/2}$ rate of convergence, where n denotes the sample size, as a result of regularization and overfitting. For that reason, it is necessary to adapt ML methods to the particular goal of causal inference. Different methods have been proposed recently depending on the target causal parameter and identifying assumption. In this study, we focus on the average treatment effect (ATE) and heterogeneous treatment effects (HTEs). The ATE gives the average causal effect of a treatment over all observations. In contrast, HTEs indicate how the treatment effect varies across the covariates. Hence, it provides more detailed information on the efficacy of the treatment for specific observations.

When estimating the ATE in the complex or high-dimensional setting, some ML-based methods seem very promising (Athey et al., 2017). One of these methods is Double Machine Learning (DML; Chernozhukov et al., 2016). DML allows the use of any consistent ML method with $n^{-1/4}$ rate of convergence[1] to estimate nuisance relationships and subsequently causal effects, under both unconfoundedness and IV-identification. DML estimators have good statistical properties: they are asymptotically unbiased and normally distributed with $n^{-1/2}$ rate of convergence, enabling valid ATE inference. Another method is Approximate Residual Balancing (ARB; Athey et al., 2016a), which offers a competitive alternative. Contrary to DML, while showing similar statistical properties, ARB does not impose the need to estimate conditional treatment probabilities consistently, implying that it performs well in situations with complex treatment assignment relationships. This comes however at the cost of additional linearity assumptions of which the most important is a linear, sparse relationship between covariates and the outcome. Primarily regularized regression methods are suitable to estimate such an outcome model. Moreover, ARB is customized to ATE estimation of a binary treatment under unconfoundedness, indicating that it cannot be applied as widely as DML.

[1] Examples of ML methods that can be used are random forests, neural networks and the LASSO

When estimating HTEs in the complex or high-dimensional setting, Chernozhukov et al. (2017) advocate to adjust DML by replacing OLS with the LASSO for final causal effect estimation. This enables estimation of a larger amount of heterogeneous treatment variables, while the use of any well-performing ML method to estimate the nuisance relationships remains possible. They derive new statistical properties that enable valid HTE inference, under the assumption of sparsity of the HTEs. The heterogeneous DML method is suitable for HTE inference under unconfoundedness, excluding identification via IVs. As opposed to heterogeneous DML, the Causal Forest (Wager and Athey, 2017) does not rely on any form of sparsity of the HTEs. Generic statistical properties from the generalized random forest framework (Athey et al., 2016b) permit valid HTE inference, under both unconfoundedness and IV-identification. Causal Forests build on the traditional random forests, but interpret them as nearest neighbor estimator instead and use different trees that are specifically designed for causal effect heterogeneity.

By applying these new methods instead of established methods such as difference-in-difference, OLS, fixed effect or two stage least squares estimation, we expect to improve causal estimates. Nevertheless, to the best of our knowledge, there are not many economic studies examining the merits of these new approaches in empirical applications. For the ATE, we found only the short estimation exercise on heart catherization data from Athey et al. (2017) that involves both DML and ARB. Separate applications of DML and ARB are presented in the original papers, where Chernozhukov et al. (2016) revisit three empirical applications, for instance IV-estimation of the causal effect of institutions on economic performance, following Acemoglu et al. (2001). Athey et al. (2016a) reexamine the efficacy of a welfare-to-work program, following Hotz et al. (2006). For HTEs, there do not exist any joint applications of heterogeneous DML and the Causal Forest at all. Again, we do find some separate applications, with Chernozhukov et al. (2017) applying heterogeneous DML to estimate demand elasticities for a major wholesale food distributor. Athey et al. (2016b) apply Causal Forests to see how the treatment effect of women's child-rearing on labor market participation varies with selected covariates, including the mother's age and the father's income.

Thus, in this study, we examine what modified and extra insights we can get from applying the ML-based causal inference methods instead of established methods in empirical applications. In order to achieve that, we revisit two applied papers from the literature that match the complex or high-dimensional setting. The first paper from DellaVigna and Kaplan (2007) estimates the causal effect of the introduction of the Fox News channel on the Republican vote share in the 2000 U.S. presidential elections by assuming unconfoundedness. The second paper from Nunn (2007) estimates the causal effect of contract enforcement quality on trade flows by using IVs. Both papers focus primarily on estimation of the ATE. In the context of these papers, we implement the ML-based causal inference methods from Chernozhukov et al. (2016; DML), Athey et al. (2016a; ARB), Chernozhukov et al. (2017; heterogeneous DML) and Athey et al. (2016b; Causal Forests) using the original ML methods proposed in these four papers, but also an additional established ML method: Support Vector Machines. Our purpose is fourfold: firstly we assess whether DML and/or ARB change substantive conclusions from the original papers and if yes what the implications of these changes are. Secondly, we inspect if DML and ARB agree and if not what factors explain the difference. Thirdly, we investigate what additional substantive insights we get by applying heterogeneous DML and Causal Forests to find HTEs. Fourthly, we examine if the heterogeneous DML and Causal Forest produce similar estimates and if not what the source of differences is.

For ATE estimation under unconfoundedness, we find that DML estimates exceed the original OLS estimates, with both our main and dynamic specification controls. Further inspection suggests that this increase could be due to nonlinear confounding relationships that happen to be present and are captured with DML. The DML estimates are in line with original estimates that use controls and employ fixed effects and/or data weighting in the main specification, but they are larger than most of these estimates in the dynamic specification. DML seems to strengthen the original results for a significant positive causal effect here. Under IV-identification, we find that DML estimates correspond to a positive reverse causal effect for the relationship between contract enforcement and trade flows. This agrees with intuition, in contrast to the negative reverse causal effect from original IV estimates. DML possibly makes the exclusion assumption more plausible here. Next, for ATE estimation under unconfoundedness with ARB, we also find larger estimates than original OLS in some cases, suggesting that taking into account nonlinear confounding indeed increases the estimates. DML and ARB agree in the main specification but disagree in the dynamic specification. The latter might reflect the difference between finite sample and asymptotic

optimality when controlling for confounding. For HTE estimation under unconfoundedness, we find some similarities between heterogeneous DML and the Causal Forest. Extra substantive findings agreed upon by both methods are that the largest effects of Fox News channel availability on the Republican vote share in 2000 occur in New York ($0.8 - 1.1\%$), Michigan ($0.7 - 0.8\%$) and Wyoming, while in Wisconsin there is a small effect. For HTE estimation with IVs, we conclude from Causal Forest estimates that the influence of vertical integration on the relationship between contract enforcement and trade flows might not be as strong as originally thought.

This paper is organized as follows. In Section 2, we give an overview of the literature on causal inference on the ATE and HTEs, under unconfoundedness and IV-identification. Readers already familiar with this well-known literature could skip this section. In Section 3, we start with discussing the DML and ARB methods, after which we elaborate on heterogeneous DML and the Causal Forest. In Section 4, we present the revisited papers and their data. We also motivate why specifically for these papers ML-based methods might be useful and could give additional insights for causal inference. In Section 5, we present our results for ATE estimation with DML and ARB. We also compare our results across DML and ARB as well as to the original results here. In Section 6, we discuss our additional findings from HTE estimation with heterogeneous DML and Causal Forests and compare them. We also motivate the relevance of the HTE estimates there. Finally, in Section 7, we conclude the research.

2 Literature review

For ATE estimation under unconfoundedness, the generally accepted framework is the potential outcomes model (Neyman, 1923; Rubin, 1974) postulating the potential outcomes $Y_i(1)$ and $Y_i(0)$ which represent the outcomes for observation i if treated or not treated, respectively. Only one of these outcomes is observed in practice. An additional important concept is the conditional probability of treatment given the covariates, or propensity score $e(X_i)$. Furthermore, the traditional literature not involving ML focuses on semiparametric or even nonparametric models for these quantities, although parametric models were considered in earlier work. This is partly due to the groundbreaking paper of LaLonde (1986), who concludes that parametric econometric estimates do not replicate experimental findings, implying substantial specification error in more complex settings.

Traditional estimators for the ATE under unconfoundedness can be divided into several classes (Imbens and Wooldridge, 2009), each controlling differently for covariates. Regression methods estimate regression functions for the conditional expectations of the potential outcomes $\mathbb{E}[Y(1)|X_i = x]$ and $\mathbb{E}[Y(0)|X_i = x]$, see for instance the global smoothing method of Hahn (1998). Next, propensity score methods build on a result of Rosenbaum and Rubin (1983) suggesting that under unconfoundedness, it is sufficient to correct solely for differences in propensity scores instead of differences in all the covariates. An example is Inverse Probability Weighting (IPW; Hirano et al., 2003), where propensity scores serve as weights when estimating $\mathbb{E}[Y(1)|X_i = x]$ and $\mathbb{E}[Y(0)|X_i = x]$ as weighted average of the treatment and control group observations. Further, matching methods estimate the missing potential outcomes using observed outcomes of a few nearest neighbors of the opposite treatment group, where distance is measured as similarity in terms of the covariates. See for instance Abadie and Imbens (2006), who propose a matching estimator and derive its large sample properties. From all previously discussed classes of methods, we obtain estimators that are asymptotically efficient. Therefore, if we have a large sample size relative to the number of covariates and we apply suitable nonparametric estimators, they theoretically perform well. However, a consistent finding from the literature is that the best methods in practice involve both conditional expectation and propensity score estimation (Athey et al., 2017). Such doubly robust methods rely only on consistent estimation of either the conditional expectations or the propensity scores, hence removing sensitivity to misspecification. A fundamental doubly robust method is Augmented Inverse Propensity Weighting (AIPW; Robins et al., 1994).

The previous methods all assume that we have a small, fixed number of covariates. Conversely, the targeted maximum likelihood methods as described by Van Der Laan and Rubin (2006) estimate a low-dimensional parameter, in our case the ATE, when nuisance parameters that are used to control for covariates are high-dimensional. The methods (iteratively) perform maximum likelihood estimation in a least favorable direction, leading to a locally efficient estimator of the parameter of interest. One of the advantages is that likelihood cross-validation can be used to estimate the nuisance parameters. This

allows the use of flexible methods with less well established properties for nuisance parameter estimation.

Recent advances show that ML methods also yield great improvement compared to traditional methods, when covariates are high-dimensional (Athey, 2017). Early cases used ML in a sparse setting: a lot of covariates are included but only a small number of them are relevant. An often used ML method in this setting is the LASSO (a form of regularized regression). However, if the LASSO is used directly in a regression of the outcome on both the treatment variable and covariates, it produces a biased estimate of the treatment effect (regularization bias, see Belloni et al., 2014). The reason is that the LASSO tends to drop covariates with a small effect on the outcome but a large effect on the treatment, because it focuses completely on prediction. Modifications are necessary here, such as the use of Neyman orthogonality as proposed by Chernozhukov et al. (2015).

Modified methods along these lines do however all limit nuisance parameter complexity (as measured by entropy growth), thereby placing strong restrictions on our models. Now, DML pragmatically avoids these strong complexity restrictions by means of cross-fitting, following among others Belloni et al. (2012), who use sample splitting to weaken sparsity conditions. Moreover, DML also uses Neyman orthogonality to remove regularization bias. DML generalizes the method of Robinson (1988) to be able to apply any consistent ML method with $n^{-\frac{1}{4}}$ rate of converge as estimator of the nuisance parameter. The reason for this is that Robinson's kernel regression breaks down in high-dimensional settings. Finally, DML is doubly robust, hence it builds on ideas from the traditional literature as well.

The previously discussed ML-based methods all require estimation of the propensity score to eliminate regularization bias. However, in situations where we are willing to rely on linearity and sparsity of the outcome model, we can circumvent this step. Accurately balancing the moments of the covariate distributions across the treatment and control group is namely more important for finite sample performance than modeling the propensity score well. For example, methods like IPW and AIPW can perform poorly in finite samples if the propensity score comes very close to the boundaries of 0 and 1. Zubizarreta (2015) points out that exact balance of the covariates is in general not possible in high dimensions, but extends the idea to approximate balance. Next, the ARB method goes a step further by allowing for valid inference on the ATE in high dimensions as well. ARB resembles AIPW computationally, but the used weights and motivation for them are very different.

Next, for HTE estimation under unconfoundedness, only since recently data sets have been informative enough for exploring heterogeneity in treatment effects. Early approaches employ for example matching (see Lee, 2009), but analogous to traditional methods for inference on the ATE, these break down in settings with high-dimensional covariates. As a consequence, ML based methods are proposed increasingly for HTE estimation, too[2]. In the linear, sparse setting, Imai et al. (2013) and Tian et al. (2014) propose LASSO-based methods. The heterogeneous DML method is derived in a general framework of high-dimensional treatment variable estimation in a setting with high-dimensional nuisance parameters. The framework lends itself perfectly to HTE estimation under sparsity, applying a version of the LASSO, too.

Other ML methods also give promising results in the context of HTEs. A simple approach is to transform the data in order to apply standard ML methods, a path taken by many researchers in this area (e.g. Dudík et al., 2011). There are however also adjustments to standard ML methods in order to accommodate HTE estimation. Tree-based methods are for instance a natural means to divide treatment effects in heterogeneous subgroups. Zeileis et al. (2008) develop a generic framework for estimating models at the leaves of a tree. It incorporates HTE estimation, which uses a linear model with an intercept and treatment indicator as leaf model. They also suggest to apply statistical tests instead of cross-validation to decide when to prune the tree. In line with this, Su et al. (2009) propose to use a squared t-statistic to test whether the ATE is equal in two potential leaves. However, Athey and Imbens (2016) argue that both methods do not effectively combine goodness-of-fit improvement and the particular purpose of finding heterogeneous treatment subgroups, hence they introduce causal trees. Causal trees can data-adaptively determine heterogeneous subgroups without giving up on valid confidence intervals. This is achieved via an honest approach that implements sample splitting. Causal trees compare favorably to other tree-based methods; the only drawback is a loss of efficiency.

While causal trees provide subgroups with different treatment effects, they do not offer full non-parametric specification of treatment heterogeneity. For this, forest-based algorithms seem to be most suitable. Foster et al. (2011) apply standard regression forests to predict unobserved potential outcomes

[2]There is also research on targeted learning in this area, see for instance Rosenblum and van der Laan (2011)

for HTE estimation. Additionally, Bayesian additive regression trees have been applied successfully to obtain HTE estimates and posterior credible intervals via MCMC sampling (e.g. Green and Kern, 2012). Nonetheless, although vital in practice, valid statistical results for inference are not given in these papers, as indicated by Wager and Athey (2017). This motivates them to develop Causal Forests and their asymptotic theory, finding that Causal Forests dominate other forest-based methods in terms of bias and variance. Athey et al. (2016b) reconsider Causal Forests in the generalized random forest framework, thereby finding improved HTE estimation accuracy compared to the standard Causal Forest when applying a preliminary orthogonalization step in the spirit of Robinson (1988).

Finally, for estimation of the ATE or HTEs with IVs, there are some semiparametric and nonparametric methods that extend the traditional two-stage least squares (2SLS) estimator for more complex settings. For example, Abadie (2003) consider semiparametric IV-estimation using kernel estimation. Darolles et al. (2011) derive a nonparametric IV-estimator based on ridge regression. However, approaches based on ML often handle high-dimensional settings better, where dimensionality refers to the control functions[3]; in contrast to the number of instruments. Gautier and Tsybakov (2011) consider IV-estimation of the outcome equation in very high-dimensional sparse settings, where their motivational examples of rich heterogeneity, many exogenous covariates due to nonlinearities and many control variables to justify the use of an instrument particularly resemble our setup. They propose the self tuning instrumental variables estimator, which is related to versions of the LASSO. Valid confidence intervals on endogenous variables, in our case the causal variable, can be obtained in this way if we assume some upper bound on the number of controls with nonzero effect. Furthermore, in a similar setting but without assuming sparsity, Hartford et al. (2017) propose DeepIV, applying deep neural networks for IV-estimation. They decompose the analysis into a first step of modeling the conditional distribution of the causal variable given the instruments and covariates, followed by minimization of a loss function that is based on the first step conditional distribution. They also present a method for out of sample causal validation of hyperparameters. DeepIV performs well relative to 2SLS and standard ML methods. Obtaining valid confidence intervals is however more challenging with this method. Next, due to their generality, i.e. they are specified in a GMM framework, we can apply DML and Causal Forests as well for IV-estimation in this setting, without relying on sparsity.

3 Methodology

3.1 Average Treatment Effect

3.1.1 Double Machine Learning

Consider the setting with complex, that is nonlinear, and high-dimensional nuisance relationships. Formally, these relationships are modeled using so-called highly-complex nuisance parameters/functions, which are functions of the covariates having an entropy that increases with the sample size. Here, ML methods often do well in practice by striking a balance between regularization and overfitting. Regularization reduces estimation variance in order to learn from the data in this setting, but it necessarily induces a bias. Furthermore, data-driven model and parameter selection brings about overfitting with respect to the particular data sample, resulting in bias as well. In a pure prediction context, these biases do not matter because the one and only priority is out-of-sample predictive quality; which is optimized by ML methods.

However, in the causal inference context, one does care about these biases because target treatment parameters cannot be observed directly. In this context, one relies crucially on statistical properties of the used estimator. Thus, the estimator obtained by directly plugging ML estimates for the nuisance parameters into the estimating equations does not suffice here, because according to Chernozhukov et al. (2016) it is biased and fails to be $n^{-1/2}$ consistent.

DML removes the negative impact of ML on statistical properties of the causal effect estimator by means of two ingredients: Neyman orthogonality and cross fitting. Firstly, Neyman orthogonality is a property that is imposed on the moment conditions that are used for estimation, in order to reduce their sensitivity to noisy nuisance parameter estimates. This in turn leads to removal of regularization

[3]Control functions are used here to correct for factors that relate to the instrument, treatment and outcome.

bias. Secondly, sample splitting eliminates overfitting bias and ensures a favorable rate of convergence, without requiring strong entropy growth restrictions. Cross fitting adds a follow up step in which the roles of the splitted samples are swapped and the results averaged in order to regain efficiency. In turn, we discuss both ingredients followed by the implied statistical properties, adopting the original notation of Chernozhukov et al. (2016).

Before elaborating on Neyman orthogonality, we explain the DML estimation strategy and introduce necessary concepts. The causal parameter of interest is $\theta_0 \in \Theta$, which is assumed to satisfy the following population moment conditions:

$$\mathbb{E}[\phi(W_i; \theta_0, \eta_0)] = 0, \tag{1}$$

where W_i is a random variable vector, η_0 the true value of the highly-complex nuisance parameter $\eta \in T$ and $\phi = (\phi_1, \ldots, \phi_{d_\theta})'$ a vector of known score functions $\phi_j : \mathcal{W} \times \Theta \times T \to \mathbb{R}$, to be specified later in this section. d_θ denotes the fixed, low dimension of the causal variable of interest. A random data sample $\{W_i\}_{i=1}^n$ is available. In the GMM framework, the sample analog of (1) is used for estimation. For $\tilde{T} = \{\eta - \eta_0, \eta \in T\}$, the Gateaux derivative map $D_r : \tilde{T} \to \mathbb{R}^{d_\theta}$ is defined as follows:

$$D_r[\eta - \eta_0] = \partial_r \Big\{ \mathbb{E}\big[\phi(W_i; \theta_0, \eta_0 + r(\eta - \eta_0)\big] \Big\}, \quad \eta \in T,$$

for all $r \in [0, 1)$. The Gateaux derivative indicates the change in value of the moment conditions due to a size r deviation from the true nuisance function towards η. Furthermore, a nuisance realization set $\mathcal{T}_n \subseteq T$ holds the nuisance parameters that are taken with high probability by estimators $\hat{\eta}_0$ of η_0. It is used to model that ML nuisance parameter estimates are approximately correct, but noisy due to regularization.

Neyman orthogonality is defined as follows. The score ϕ is Neyman orthogonal at (θ_0, η_0) with respect to $\mathcal{T}_n$ if (1) holds, the Gateaux derivative $D_r[\eta - \eta_0]$ exists for all $r \in [0, 1)$ and $\eta \in \mathcal{T}_n$ and

$$D_0[\eta - \eta_0] = 0, \quad \text{for } r = 0 \text{ and all } \eta \in \mathcal{T}_n. \tag{2}$$

Intuitively, Neyman orthogonality means that the moment conditions for estimating θ_0 are insensitive to the value of the nuisance parameters in a small neighborhood around η_0. This ensures that the moment conditions continue to be valid if noisy ML estimates for the nuisance parameters are plugged in. Validity of the moment conditions is key for obtaining desirable statistical properties for the GMM estimator.

Next, to reduce sensitivity to the particular estimation sample used for ML and overcome overfitting bias, cross fitting is implemented with the following algorithm. The first step is to split the full sample by taking a K-fold random partition $\{I_k\}_{k=1}^K$ of the observation indices. For each fold $k \in [K] = \{1, \ldots, K\}$, let $I_k^c = \{1, \ldots, n\} \setminus I_k$ indicate the corresponding ML estimation sample. Then, in the second step, for each $k \in [K]$ a ML estimator of η_0 is constructed based on different estimation samples:

$$\hat{\eta}_{0,k} = \hat{\eta}_0((W_i)_{i \in I_k^c}). \tag{3}$$

In the last step, the target parameter estimate $\hat{\theta}_0$ is computed by solving the equation that follows from the Neyman orthogonal score:

$$\hat{\theta}_0 = \arg\min_{\theta_0} \left\{ \left\| \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in I_k} \phi(W_i; \theta_0, \hat{\eta}_{0,k}) \right\|_2 \right\}.$$

We directly pool the sample moment conditions over all folds here. There is also a version of DML where the moment conditions are solved for each fold and aggregation is performed afterwards. The current version is however more stable according to Chernozhukov et al. (2016). Finally, note that for larger values of K, more observations are put in I_k^c and less in I_k such that there is more data to estimate the nuisance parameters with ML but less data to estimate the causal effect with GMM. Since the first part appears to be most difficult, we partly follow the recommendation of Chernozhukov et al. (2016) to use a moderate value of $K = 5$. However, because higher K increases computation time and because the estimates differ only slightly for different K in the empirical applications of Chernozhukov et al. (2016), we mostly stick with the simple case of $K = 2$. We examine the impact of K in a sensitivity check.

Chernozhukov et al. (2016) assume that certain regularity conditions regarding identification, the score function and the quality of nuisance parameter estimates hold, including those requiring a crude rate of convergence of $n^{-1/4}$ for the ML methods. They then derive that:

$$\sqrt{n}\Sigma^{-1/2}(\hat{\theta}_0 - \theta_0) \to N(0, I_{d_\theta}),$$

where the approximate variance $\Sigma = (J_0^{-1})\mathbb{E}[\phi(W_i;\theta_0,\eta_0)\phi(W_i;\theta_0,\eta_0)'](J_0^{-1})'$. Hence, the desired $n^{-1/2}$ rate of convergence is achieved and valid confidence intervals can be obtained if we have a consistent variance estimate.

Under the previously discussed regularity conditions, there are variance estimates available that concentrate around the true variance. For that, we write the score functions as a linear function in θ:

$$\phi(w;\theta,\eta) = \phi(w;\eta)^a\theta + \phi(w;\eta)^b$$

The expression for the variance estimates is in that case:

$$\widehat{\Sigma} = (\widehat{J}_0^{-1})\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\phi(W_i;\hat{\theta}_0,\hat{\eta}_{0,k})\phi(W_i;\hat{\theta}_0,\hat{\eta}_{0,k})'(\widehat{J}_0^{-1})', \tag{4}$$

with

$$\widehat{J}_0 = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\phi(W_i;\hat{\eta}_{0,k})^a.$$

To reduce the impact of random sample splits, the estimation procedure is repeated S times, obtaining $\hat{\theta}_0^s$ and $\widehat{\Sigma}_s$ for $s = 1,\dots,S$. These are then aggregated to get point and variance estimates. In addition, an estimate of the variance which includes variation induced by sample splitting can be computed. To be more robust against outliers, we use the median for aggregation, which leads to the following expressions for the point and variance estimates, respectively:

$$\hat{\theta}_0^{med} = \text{median}\{\hat{\theta}_0^s\}_{s=1}^S, \qquad \widehat{\Sigma}^{conv} = \text{median}\{\widehat{\Sigma}_s\}, \qquad \widehat{\Sigma}^{med} = \text{median}\{\widehat{\Sigma}_s + (\hat{\theta}_0^s - \hat{\theta}_0^{med})(\hat{\theta}_0^s - \hat{\theta}_0^{med})'\}_{s=1}^S,$$

where the median is applied coordinatewise for $\hat{\theta}_0^{med}$ and where median corresponds to the matrix with median operator norm for $\widehat{\Sigma}^{conv}$ and $\widehat{\Sigma}^{med}$, which ensures non-negative definiteness. We refer to $\widehat{\Sigma}^{conv}$ as conventional variance estimates and to $\widehat{\Sigma}^{med}$ as median variance estimates, where only the latter incorporate variability induced by sample splitting. Due to computational limitations, we take half the amount of random data splits from Chernozhukov et al. (2016), i.e. $S = 50$.

DML can be used in conjunction with several causal effect models. The implementation in each model is characterized by a specific Neyman orthogonal score function. Firstly, we consider the Partially Linear Regression (PLR) model (Robinson, 1988):

$$Y_i = D_i\beta_0 + g_0(X_i) + U_i, \qquad \mathbb{E}[U_i|D_i,X_i] = 0 \tag{5}$$

$$D_i = m_0(X_i) + V_i, \qquad \mathbb{E}[V_i|X_i] = 0 \tag{6}$$

where Y_i is the scalar outcome variable, D_i the scalar causal variable of interest, $X_i = (X_{i1},\dots,X_{ip})'$ a vector of covariates and U_i, V_i disturbances. The outcome equation (5) is the main equation and the ATE equals β_0 in this model. The treatment assignment equation (6) keeps track of confounding, with the highly-complex nuisance function $m_0 : \text{supp}(X_i) \to (0,1)$ denoting the influence of the covariates on D_i. The highly-complex nuisance function $g_0 : \text{supp}(X_i) \to \mathbb{R}$ denotes the influence of the covariates on Y_i. For the PLR model, Chernozhukov et al. (2016) propose a Robinson (1988) style score function:

$$\phi(W_i;\theta_0,\eta_0) = \{Y_i - l_0(X_i) - \beta_0(D_i - m_0(X_i))\}\{D_i - m_0(X_i)\},$$

where $W_i = (Y_i, D_i, X_i)$, $\theta_0 = \beta_0$ and $\eta_0 = (l_0, m_0)$. In addition, $l_0 = \mathbb{E}[Y|X]$ and $g_0 = l_0 - \beta_0 m_0$, parametrizing g_0 differently because l_0 can be learned directly with ML, in contrast to g_0. Solving the implied moment conditions boils down to OLS of $\hat{\tilde{Y}}_i = Y_i - \hat{l}_0(X_i)$ on $\hat{\tilde{D}}_i = D_i - \hat{m}_0(X_i)$, i.e. OLS of residualized Y_i on residualized D_i, to get the estimate $\hat{\beta}_0^{DML}$ in the PLR model. Calculating standard

errors $\hat{\sigma}^{DML}$ via (4) is in this model equivalent to calculating heteroscedasticity-consistent standard errors with the residualized variables. Our derivations for these simpler forms are given in Section A.4 of the Appendix.

Secondly, we examine the interactive regression model with fully heterogeneous causal effects:

$$\begin{aligned} Y_i &= g_0(D_i, X_i) + U_i, & \mathbb{E}[U_i|D_i, X_i] &= 0 \\ D_i &= m_0(X_i) + V_i, & \mathbb{E}[V_i|X_i] &= 0 \end{aligned}$$

where $D_i \in \{0, 1\}$ now and where the highly-complex function $g_0 : (0, 1) \times \text{supp}(X_i) \to \mathbb{R}$ denotes the joint influence of the causal variable and the covariates on Y_i. The ATE is in this model defined by:

$$\beta_0 = \mathbb{E}[g_0(1, X_i) - g_0(0, X_i)].$$

For the interactive regression model, Chernozhukov et al. (2016) follow Robins and Rotnitzky (1995) by suggesting the score function:

$$\phi(W_i; \theta_0, \eta_0) = \psi(W_i; \eta_0) - \beta_0$$

with

$$\psi(W_i; \eta_0) = g_0(1, X_i) - g_0(0, X_i) + \frac{D_i\{Y_i - g_0(1, X_i)\}}{m_0(X_i)} - \frac{(1 - D_i)\{Y_i - g_0(0, X_i)\}}{1 - m_0(X_i)},$$

and where $W_i = (Y_i, D_i, X_i)$, $\theta_0 = \beta_0$ and $\eta_0 = (g_0, m_0)$. Solving the associated moment conditions yields $\hat{\beta}_0^{DML}$ equal to the mean of $\psi(W_i; \eta_0)$ over all observations, after trimming the propensity score weights $1/m_0(X_i)$ and $1/(1 - m_0(X_i))$ at $(0.01, 0.99)$ to reduce their extreme impact on the estimate (Chernozhukov et al., 2016). Computing the standard errors $\hat{\sigma}^{DML}$ via (4) is in this model equivalent to taking the standard deviation of $\psi(W_i; \eta_0)$ over all observations and dividing by $\sqrt{n}$. Our derivations for these simpler forms are given in Section A.4 of the Appendix. In both the PLR and the interactive regression model, DML relies on the unconfoundedness assumption to identify the ATE. This assumption corresponds to the combination of $\mathbb{E}[U_i|D_i, X_i] = 0$ and $\mathbb{E}[V_i|X_i] = 0$ in these models.

Thirdly, we consider the Partially Linear Instrumental Variable (PLIV) model:

$$\begin{aligned} Y_i &= D_i\beta_0 + g_0(X_i) + U_i, & \mathbb{E}[U_i|Z_i, X_i] &= 0 \\ Z_i &= m_0(X_i) + V_i, & \mathbb{E}[V_i|X_i] &= 0 \end{aligned}$$

with D_i the scalar causal variable, Z_i a scalar instrumental variable and β_0 the structural parameter of interest. For the PLIV model, again a Robinson (1988) style score function is proposed:

$$\phi(W_i; \theta_0, \eta_0) = \{Y_i - l_0(X_i) - \beta_0(D_i - r_0(X_i))\}\{Z_i - m_0(X_i)\},$$

where $W_i = (Y_i, D_i, X_i)$, $\theta_0 = \beta_0$ and $\eta_0 = (l_0, r_0, m_0)$. Additionally, $l_0 = \mathbb{E}[Y|X]$, $r_0 = \mathbb{E}[D|X]$ and $g_0 = l_0 - \beta_0 r_0$, parametrizing g_0 differently because l_0 and r_0 can be learned directly with ML. Solving the corresponding moment conditions to compute $\hat{\beta}_0^{DML}$ boils down in this model to IV estimation of $Y_i - \hat{l}_0(X_i)$ on $D_i - \hat{r}_0(X_i)$ using as instruments $Z_i - \hat{m}_0(X_i)$. That is, two stage least squares estimation of residualized Y_i on residualized D_i using as instruments residualized Z_i. Computing standard errors $\hat{\sigma}^{DML}$ via (4) is equivalent to computing heteroscedasticity-consistent IV standard errors with the residualized variables. Our derivations for these simple forms are given in Section A.4 of the Appendix. DML relies on IV to identify the causal parameter in the PLIV model. The exclusion restriction is imposed via the combination $\mathbb{E}[U_i|Z_i, X_i] = 0$ and $\mathbb{E}[V_i|X_i] = 0$ in this model. Finally, it can be verified for each of the score functions that the moment conditions (1) and the Neyman orthogonality conditions (2) hold, see Chernozhukov et al. (2016).

Regarding the ML submethods that are used to obtain the ML estimator $\hat{\eta}_{0,k}$ in (3), we apply the same methods as Chernozhukov et al. (2016): a regression tree, random forest, boosted regression tree, neural network, the LASSO and the hybrid Best method. We do not consider the ensemble method from Chernozhukov et al. (2016) because initial runs show that it greatly increases computation time. Moreover, Chernozhukov et al. (2016) choose to use the LASSO, boosting, random forest and neural network for the ensemble, but without clear motivation for these. Furthermore, we opt for the support vector machine (SVM) as alternative ML submethod, since Smola and Schölkopf (2004) argue that

excellent performance has been obtained in several empirical applications with SVMs, including some in the context of regression and time-series analysis. By adding another ML submethod, we are able to test the theory of Chernozhukov et al. (2016) stating that DML estimates from every sensible ML method of estimating the nuisance function should be similar. In Section A.3 of the Appendix, we shortly discuss each of these ML methods and the associated choice of tuning parameters.

3.1.2 Approximate Residual Balancing

Assuming unconfoundedness requires that we balance covariate distributions between the treatment and control group in order to obtain valid causal effect estimates. Two distinct balancing approaches are regression adjustments and weighting with weights based on the covariates. Either of the two gives estimators with favorable properties in a low-dimensional setting. However, a consistent finding from the literature is that both are required for $n^{-1/2}$ consistency of the estimator in the high-dimensional setting. The recurring issue is that ML, or more specifically regularized regression methods in this case, produce regularization bias. Weighting then acts in the process of removing this bias; in other words, debiasing.

Doubly robust methods like DML implement the combination of regression adjustments and weighting by using inverse propensity scores as weights. This builds on the early result of Rosenbaum and Rubin (1983), who find that under unconfoundedness, controlling for the propensity score is asymptotically sufficient to remove biases of any functional form related to observed covariates. Hence, these methods always require a consistent propensity score model that converges fast enough to the truth. One might wonder whether it is always possible to construct an accurate enough propensity score model and if yes whether it is worth the effort, given that there is no guarantee that an accurate propensity score model also yields good debiasing weights in finite samples.

Athey et al. (2016a) therefore intend to circumvent any propensity score modeling at all with ARB. This comes at the cost of two additional linearity assumptions: a sparse, linear outcome model and linear debiasing via weights. The key implication of using the linear outcome model is that we only have to remove linear biases. Propensity score based weighting methods solve an unnecessary difficult problem here by trying to remove biases of any functional form. Instead, ARB fully exploits the linearity assumptions in order to derive weights that optimally trade off the implied balance and variance, achieving at least approximate balance of the covariate distributions between the treatment and control group. This turns out to be exactly sufficient to remove only the linear biases in the high-dimensional setting. All in all, ARB consists of three steps. Firstly, a regularized regression model is fitted in the treatment and control group separately, to capture strong causal effects. Secondly, the approximate balancing weights are computed. Thirdly, the regularized regression results are debiased by adding their reweighted residuals, using the approximate balancing weights. This ensures that remaining, small causal effects are captured. Next, we discuss the linear outcome model assumption, the three steps of ARB and finally the statistical properties.

Consider the outcome equation of the fully heterogeneous interactive regression model:

$$Y_i = g_0(D_i, X_i) + U_i, \qquad \mathbb{E}[U_i|D_i, X_i] = 0$$

with ATE target parameter:

$$\beta_0 = \mathbb{E}[g_0(1, X_i) - g_0(0, X_i)].$$

In this model, the linear outcome equation assumption can be stated as follows:

$$\mu_t(X_i) = g_0(1, X_i) = X_i'\beta_t$$
$$\mu_c(X_i) = g_0(0, X_i) = X_i'\beta_c.$$

The ARB estimation strategy is to use an estimator of the form:

$$\hat{\beta}_0 = \hat{\mu}_t - \hat{\mu}_c = \Big(\overline{X}'\hat{\beta}_t + \sum_{\{i:D_i=1\}} \gamma_{t,i}(Y_i - X_i'\hat{\beta}_t)\Big) - \Big(\overline{X}'\hat{\beta}_c + \sum_{\{i:D_i=0\}} \gamma_{c,i}(Y_i - X_i'\hat{\beta}_c)\Big), \tag{7}$$

where $\overline{X}$ denotes the mean of X_i over all observations and where $\hat{\beta}_t, \hat{\beta}_c$ denote coefficient estimates for the linear regression of Y_i on X_i for the treatment and control group separately, respectively. Furthermore,

$\gamma_{t,i}, \gamma_{c,i}$ correspond to the balancing weights for observation i for the treatment and control group, respectively. In the first step of ARB, ML or more precisely regularized regression methods are used to compute $\hat{\beta}_t$ and $\hat{\beta}_c$, obtaining $\hat{\beta}_t^{Reg}$ and $\hat{\beta}_c^{Reg}$. Regularized regression methods are suitable in the high-dimensional setting due to their optimal tradeoff of regularization and overfitting.

In the second step, the vector of approximate balancing weights is calculated by using:

$$\gamma_t^{Aprox} = \arg\min_{\tilde{\gamma}} \Big\{ (1-\zeta)||\tilde{\gamma}||_2^2 + \zeta||\overline{X} - \boldsymbol{X}_t'\tilde{\gamma}||_\infty^2, \text{ s.t. } \sum_{\{i:D_i=1\}} \tilde{\gamma}_i = 1 \text{ and } 0 \leq \tilde{\gamma}_i \leq n_t^{-2/3} \Big\},$$

with $\boldsymbol{X}_t$ the covariate matrix for the n_t treatment group observations and $\zeta \in (0,1)$ a tuning parameter stating the importance of bias and variance. The expression for the control group weights γ_c^{Approx} is similar but with $\boldsymbol{X}_c'$, n_c substituted for $\boldsymbol{X}_t'$, n_t and summing over observations with $D_i = 0$ instead of 1. Here, $\boldsymbol{X}_c$ is the covariate matrix for the n_c control group observations. The first term of the objective function reflects the implied variance of estimators of the form (7), whereas the second term reflects the implied covariate balance achieved by such estimators. Intuitively, the weights are constructed in such a way that the mean of the reweighted treatment or control group observations $\boldsymbol{X}_t'\tilde{\gamma}$,$\boldsymbol{X}_c'\tilde{\gamma}$ match the overall sample mean $\overline{X}$ as closely as possible, implying balanced covariate distributions between the treatment and control group. The difference with the estimator of Zubizarreta (2015) is that ARB does not constrain the implied balance term to be practically small, because this leads in general to infeasibility in the high-dimensional setting. In the third step, the ARB estimate $\hat{\beta}_0^{ARB}$ is computed by filling in the abstract form (7), setting $\hat{\beta}_t = \hat{\beta}_t^{Reg}$, $\hat{\beta}_c = \hat{\beta}_c^{Reg}$ and $\gamma_{t,i} = \gamma_{t,i}^{Aprox}$, $\gamma_{c,i} = \gamma_{c,i}^{Aprox}$.

For the statistical properties of the ARB estimator, Athey et al. (2016a) make the often used assumptions of sparsity of the parameter vectors β_t, β_c and overlap, i.e. the treatment and control group cannot be too dissimilar with respect to the covariate distribution. Under these assumptions, they derive asymptotic distributions for $\hat{\mu}_t$ and $\hat{\mu}_c$:

$$(\hat{\mu}_t - \mu_t) \Big/ \sqrt{\hat{\sigma}_t^2} \to N(0,1), \quad \text{with } \hat{\sigma}_t^2 = \sum_{\{i:D_i=1\}} (\gamma_{t,i}^{Aprox})^2 (Y_i - X_i'\hat{\beta}_t^{Reg})^2$$

$$(\hat{\mu}_c - \mu_c) \Big/ \sqrt{\hat{\sigma}_c^2} \to N(0,1), \quad \text{with } \hat{\sigma}_c^2 = \sum_{\{i:D_i=0\}} (\gamma_{c,i}^{Aprox})^2 (Y_i - X_i'\hat{\beta}_c^{Reg})^2,$$

where $\hat{\sigma}_t^2, \hat{\sigma}_c^2$ are consistent variance estimates for $\hat{\mu}_t$ and $\hat{\mu}_c$, respectively. In addition, conditional on X_i and D_i, $\hat{\mu}_t$ and $\hat{\mu}_c$ are independent. This suggests that $(\hat{\beta}_0^{ARB} - \beta_0)/\hat{\sigma}^{ARB} \to N(0,1)$, where $(\hat{\sigma}^{ARB})^2 = \hat{\sigma}_c^2 + \hat{\sigma}_t^2$. In the end, this result can be used to conduct inference on the ATE in the interactive regression model using ARB. The achieved rate of convergence is $n^{-1/2}$.

We adopt the suggestion of Athey et al. (2016a) to apply the LASSO and the elastic net as regularized regression methods in step one. The elastic net offers additional stability compared to the LASSO with respect to small changes in the data and is therefore suitable here. To be able to examine the impact of using the elastic net instead of the LASSO well, we assign equal importance to ridge regression and the LASSO within the elastic net. In Section A.3 of the Appendix, we describe our implementation of the LASSO and elastic net for ARB in more detail. Further, we follow the recommendation of Athey et al. (2016a) to set $\zeta = 0.5$. We however also investigate the impact of a larger and smaller ζ by setting $\zeta = 0.3$ and $\zeta = 0.7$.

A final note is that in practice, high-dimensionality of the covariate vector could occur for two reasons: there is either a large number of covariates from the start or a small number of covariates but due to including first, second and higher order interactions of the original covariates the dimensionality gets high. In the latter case, the linear outcome model assumption is substituted for the assumption of a quadratic, cubic, and higher order smooth outcome model[4], as indicated by Athey et al. (2016a). This implies that we can make the ARB estimates less sensitive to linearity by including higher order interactions. Due to the sharp increase in computation time when including higher-order terms, we only add interactions up to the first order (including squares) in our applications. We also discuss the difference between ARB estimates that do not add any extra higher order terms (fully linear) and ARB estimates that do add those terms (quadratic).

[4] Hirshberg and Wager (2017) show that the approximate balancing weights converge to the true inverse propensity weights in that situation, indicating that there is an explicit connection with doubly robust methods like DML.

3.2 Heterogeneous Treatment Effects

3.2.1 Heterogeneous Double Machine Learning

When estimating HTEs, the dimension of the treatment variable vector typically grows very large. Due to the assumption of fixed treatment dimensionality d_θ, original DML does not suffice here. Hence, Chernozhukov et al. (2017) modify it to be applicable in situations with high-dimensional treatments, assuming sparsity of the HTEs. Firstly, instead of residualizing each of the heterogeneous treatment variables separately, they build on affine transformations of a single residualized base treatment variable, to improve precision and save computation time. Secondly, where original DML applies OLS, they apply the LASSO or a debiased version of the LASSO on the residualized variables to estimate the causal parameters in the second stage. The LASSO is suitable for estimation, while the debiased LASSO also enables asymptotically valid inference. We discuss the modifications in turn, followed by statistical properties of the resulting estimators.

Consider the HTE model with modeled heterogeneity, which adjusts the PLR model in order to allow for a vector causal variable of interest:

$$\begin{aligned} Y_i &= F_i'\beta_0 + g_0(X_i, H_i) + U_i, & \mathbb{E}[U_i|D_i, X_i, H_i] = 0 \\ F_i &= D_i H_i \\ D_i &= m_0(X_i, H_i) + V_i, & \mathbb{E}[V_i|X_i, H_i] = 0 \end{aligned}$$

where β_0 is a $d \times 1$ parameter vector holding HTEs and $H_i \in \{h_1, \ldots, h_k\}$ a $d \times 1$ vector of characteristics that induces treatment heterogeneity by taking one of k prespecified values. The highly-complex nuisance functions $g_0 : \text{supp}(H_i,X_i) \to \mathbb{R}$ and $m_0 : \text{supp}(H_i,X_i) \to (0,1)$ specify the joint influence of the covariates and heterogeneity variables on Y_i and D_i, respectively.

There is a base treatment variable D_i, but the final treatment is equal to $F_i = D_i H_i$. Chernozhukov et al. (2017) show that in this case, the estimated base treatment variable can be used directly to construct estimated final treatment. More specifically, $\widehat{F}_i = \mathbb{E}[F_i|X_i, H_i] = H_i \widehat{m}_0(X_i, H_i)$, so that ML methods only need to be applied once to estimate the nuisance parameter, for D_i, and not for all heterogeneous treatments individually.

Given the estimated residuals $\widehat{\tilde{Y}}_i = Y_i - \widehat{g}_0(X_i)$ and $\widehat{\tilde{F}}_i = F_i - \widehat{F}_i$, original DML applies OLS of $\hat{\tilde{Y}}_i$ on $\hat{\tilde{F}}_i$ in the second stage to estimate the causal parameter. However, when the dimension of the treatment variable vector increases due to heterogeneity, one may improve on this orthogonal least squares by using ML. Here, the same logic as for estimation of high-dimensional nuisance parameters is followed. Chernozhukov et al. (2017) opt for the LASSO and refer to it as the orthogonal LASSO:

$$\hat{\beta}_0^{OL} = \arg\min_{\theta_0} \Big\{ \sum_{i=1}^{n} (\widehat{\tilde{Y}}_i - \widehat{\tilde{F}}_i'\theta_0)^2 + \lambda ||\theta_0||_1 \Big\},$$

where $\lambda > 0$ is a penalty parameter. The value of λ cannot be determined with cross-validation, because true causal effects are not observed. Chernozhukov et al. (2017) suggest however to choose $\lambda = c \cdot \max\{l_n m_n, s m_n^2, \lambda_n\}$. Here, $c > 1$, s is the sparsity level which Chernozhukov et al. (2017) assume to be equal to $\sqrt{n/\log(\#\text{Controls})}$. l_n, m_n are the rates of convergence for ML estimation of the outcome and treatment assignment equation, respectively, which lie between $n^{-1/4}$ and $n^{-3/4}$ according to Chernozhukov et al. (2017). λ_n is another rate of convergence which is smaller than $n^{-1/2}$ according to Chernozhukov et al. (2017). All in all, we deduce that this leads to $\lambda \geq n^{-3/2}\sqrt{n/\log(\#\text{Controls})}$. We take a shrinkage level λ around this lower bound because d is not very large in our applications, suggesting that we do not need very much regularization. Along these lines, we also use a higher shrinkage level for larger d.

Since the LASSO produces ML bias, it has to be debiased in order to be suitable for inference. For simplicity, we explain the debiasing strategy of Chernozhukov et al. (2017) for the case where there is no first stage ML estimation error. The true residuals are denoted $\widetilde{Y}_i = Y_i - g_0(X_i)$, $\widetilde{F}_i = F_i - \mathbb{E}[F_i|X_i, H_i]$ and the treatment variable covariance matrix is denoted $Q = \mathbb{E}[\widetilde{F}_i\widetilde{F}_i'] \approx \frac{1}{n}\sum_{i=1}^{n} \widetilde{F}_i\widetilde{F}_i'$. A sparsity assumption on Q^{-1} is further made, which we satisfy because Q is diagonal in our case[5]. Then, Q^{-1} can be estimated

[5] see Assumption 3.8 of Chernozhukov et al. (2017)

reliably by using the Constrained Linear Inverse Matrix Estimation (CLIME) matrix $M = [m_1, \ldots, m_d]$:

$$m_j = \arg\min_m \left\{ ||m||_1 \text{ s.t. } \left|\left| Qm - e_j \right|\right|_\infty < a\sqrt{\frac{\log(d)}{n}} \right\}, \qquad j = 1, \ldots, d$$

with a a suitably large constant for which we use a grid of values and e_j the unit vector in dimension j.

Let $\widetilde{U}_i = \widetilde{Y}_i - \widetilde{F}_i \hat{\beta}_0^{OL}$ further be the LASSO residual. This residual can be decomposed into a true disturbance and estimation error: $\widetilde{U}_i = U_i + \widetilde{F}_i'(\beta_0 - \hat{\beta}_0^{OL})$. Next, consider the following correction term and its decomposition implied by the decomposition of the LASSO residual:

$$\sqrt{n} m_j' \frac{1}{n} \sum_{i=1}^n \widetilde{F}_i \widetilde{U}_i = m_j' \frac{1}{\sqrt{n}} \sum_{i=1}^n \widetilde{F}_i U_i + \sqrt{n} m_j' Q(\beta_0 - \hat{\beta}_0^{OL}) \qquad j = 1, \ldots, d \tag{8}$$

According to the central limit theorem, the first term is normally distributed with mean zero and certain variance $(\sigma_j^{DOL})^2$. The second term is approximately equal to $\sqrt{n}(\beta_{0,j} - \hat{\beta}_{0,j}^{OL})$ by definition of the CLIME matrix M (that is $m_j' Q \approx e_j'$). Therefore, adding the correction term to the jth LASSO estimate gives the asymptotic distribution:

$$\sqrt{n}(\hat{\beta}_{0,j}^{OL} + m_j' \frac{1}{n} \sum_{i=1}^n \widetilde{F}_i \widetilde{U}_i - \beta_{0,j}) \approx m_j' \frac{1}{\sqrt{n}} \sum_{i=1}^n \widetilde{F}_i U_i \approx N(0, (\sigma_j^{DOL})^2) \qquad j = 1, \ldots, d \tag{9}$$

Result (9) indicates that adding the correction term (8) to $\hat{\beta}_0^{OL}$ leads to successful debiasing. This continues to be the case if one takes into account first stage ML estimation error, see Chernozhukov et al. (2017). Thus, the debiased LASSO estimator is defined as follows:

$$\hat{\beta}_0^{DOL} = \widehat{M}' \frac{1}{n} \sum_{i=1}^n \widehat{\widetilde{F}}_i (\widehat{\widetilde{Y}}_i - \widehat{\widetilde{F}}_i' \hat{\beta}_0^{OL}) + \hat{\beta}_0^{OL},$$

where the CLIME matrix $\widehat{M}$ is now computed using $\widehat{Q} = \frac{1}{n} \sum_{i=1}^n \widehat{\widetilde{F}}_i \widehat{\widetilde{F}}_i'$ instead of Q. Under certain regularity conditions, $\hat{\beta}_0^{DOL}$ is asymptotically normally distributed with covariance matrix Σ^{DOL} that can be estimated consistently by $\widehat{\Sigma}^{DOL} = \widehat{M}'(\frac{1}{n} \sum_{i=1}^n \widehat{\widetilde{F}}_i \widehat{U}_i \widehat{U}_i' \widehat{\widetilde{F}}_i') \widehat{M}$, where $\widehat{U}_i = \widehat{\widetilde{Y}}_i - \widehat{\widetilde{F}}_i' \hat{\beta}_0^{OL}$. Thus, asymptotic standard errors are available and we can conduct inference.

3.2.2 Causal Forest

Tree-based ML methods partition the data in heterogeneous subgroups, after which each subgroup gets a different target parameter estimate. They are therefore a natural starting point to derive methods that discover heterogeneity in causal effects, which in practice translates into minimizing the mean squared error of causal effects. However, since one does not observe the ground truth for causal effects, applying standard regression tree methodology directly with this new target quantity is not possible. Athey and Imbens (2016) overcome this with their causal trees by deriving an unbiased estimate of the mean squared error of causal effects that can be used instead, enabling estimation of HTEs. Furthermore, to ensure valid causal inference on HTEs as well, they suggest to impose honesty on the trees. Honesty states that the same observation cannot be used for both selecting a model and estimating the model afterwards. For trees, this boils down to not using the same data for determining the tree splits and estimating the HTEs for each subgroup identified by the tree. Honesty comes at the cost of inefficiency for causal trees, because it requires sample splitting.

HTE estimates for each of the subgroups of observations that the causal tree identifies can thus be obtained. Yet, causal trees do not produce a causal effect estimate for each specific observation, as identified by its particular covariate values. It follows that we do not necessarily find treatment effect heterogeneity for the covariates that we are interested in. To be able to at least examine treatment effect heterogeneity for our covariates of interest, we resort to Causal Forests (Wager and Athey, 2017). Causal Forests produce ATE estimates conditional on the covariate values of an individual, called Conditional Average Treatment Effects (CATE), which allow us to estimates HTEs across our covariates of interest.

Causal Forests also enable valid causal inference on these HTEs. Moreover, considering that Causal Forests aggregate estimates of multiple causal trees, they reduce the loss of efficiency from honesty, because different trees use different data splits such that all observations typically get involved in both determining the tree splits and estimating final HTEs.

To apply Causal Forests, as originally introduced by Wager and Athey (2017), the researcher must choose between two procedures to realize honesty. Both procedures have their strengths: either doing well in correcting for confounding or being able to pick up heterogeneity in causal effects. Having to make this choice is therefore rather unfortunate, given that we ideally enjoy both of the strengths. The generalized random forests framework (Athey et al., 2016b) offers the solution by deriving Causal Forest estimation anew but now via moment conditions. Both strengths can then be combined by using an orthogonalized version of these moment conditions, analogous to being able to combine biased ML estimation with valid causal inference for DML by means of orthogonalization.

Thus, Causal Forests in the generalized random forest framework seem to suit the purpose of estimating HTEs in our applications best. The generalized random forest framework implements a couple of key ideas. Firstly, to remove regularization bias, it departs from the perspective that random forests aggregate estimates from regression trees. Instead, the framework casts random forests as weighted nearest neighbor estimator, producing forest weights that can be used for valid causal inference with moment conditions. Secondly, the framework speeds up the process of discovering heterogeneity in causal effects by using a linear, gradient-based approximation of the moment conditions rather than the nonlinear moment conditions themselves. Next, we discuss the generalized random forest algorithm, afterwards the gradient approximation of our moment conditions and finally some statistical properties.

Similar to DML, the estimation strategy of Causal Forests is to find a solution to moment conditions, which in this case have the form:

$$\mathbb{E}[\psi(O_i; \theta_0(x^*), \eta_0(x^*))|X_i^* = x^*] = 0,$$

where O_i is random variable vector, $X_i^* \in \mathcal{X}$ an auxiliary covariate vector, $x^* \in \mathcal{X}$ an input test point specifying covariate values for which we intend to compute the causal effect, $\theta_0(x^*)$ the target parameter and $\eta_0(x^*)$ a nuisance parameter. Note that the parameters specifically belong to the input test point. However, to simplify notation, we do not explicitly write this if unnecessary and use θ_0 and η_0 instead. θ_0 is estimated by solving the weighted sample analog:

$$(\hat{\theta}_0, \hat{\eta}_0) = \arg\min_{\theta_0, \eta_0} \{ \big|\big|\Psi(\theta_0, \eta_0)\big|\big|_2 \}, \qquad \Psi(\theta_0, \eta_0) = \sum_{i=1}^{n} \alpha_i(x^*)\psi(O_i; \theta_0, \eta_0). \tag{10}$$

The forest weight $\alpha_i(x^*)$ equals the frequency with which X_i^* falls into the same tree subgroup (or leaf) as x^*, across all grown trees $b = 1, \ldots, B$. For tree b, let $L_b(x^*)$ denote the subset of covariate values of observations in the leaf corresponding to x^*. Then, formally the weights are calculated via:

$$\alpha_i(x^*) = \frac{1}{B} \sum_{b=1}^{B} \alpha_{ib}(x^*), \qquad \alpha_{ib}(x^*) = \frac{I[X_i^* \in L_b(x^*)]}{|L_b(x^*)|}. \tag{11}$$

To guarantee plenty of tree variation in order to control overfitting, tree construction is preceded by sampling a fraction s from the full sample without replacement, obtaining an estimation sample $\mathcal{I}_b$. Next, to ensure honesty, $\mathcal{I}_b$ is divided into two evenly sized samples $\mathcal{J}_{1b}$ and $\mathcal{J}_{2b}$. $\mathcal{J}_{1b}$ is used to build the tree, whereas $\mathcal{J}_{2b}$ is employed to determine $\alpha_{ib}(x^*)$ with (11) after the tree has been built. After the complete forest has been constructed, the forest weights $\alpha_i(x^*)$ are obtained from (11) for i in $\mathcal{J}_{2b}$ after which θ_0 is estimated using (10).

Next, for tree construction, note that every tree split starts with a parent node $P \subseteq \mathcal{X}$. Causal Forests split P into two nonoverlapping children $C_1, C_2 \subseteq \mathcal{X}$ such as to maximize heterogeneity of the target parameter, using only the randomly chosen variables tried at that split. For that, the solution of the estimating equation at a node $C \subseteq \mathcal{X}$ is defined by:

$$(\hat{\theta}_C, \hat{\eta}_C) = \arg\min_{\theta, \eta} \Big\{ \Big|\Big| \sum_{\{i \in \mathcal{J}_{1b}: X_i^* \in N\}} \psi(O_i; \theta, \eta) \Big|\Big|_2 \Big\}. \tag{12}$$

We denote n_C the number of observations from $\mathcal{J}_{1b}$ in node C. Then, heterogeneity is maximized with the criterion:

$$\hat{\Delta}(C_1, C_2) = \frac{n_{C_1} n_{C_2}}{n_P^2} (\hat{\theta}_{C_1} - \hat{\theta}_{C_2})^2, \tag{13}$$

where $\hat{\theta}_{C_1}, \hat{\theta}_{C_2}$ follow from (12), and n_{C_1}, n_{C_2}, n_P denote the number of observations from $\mathcal{J}_{1b}$ in C_1, C_2 and P.

Since directly optimizing (13) turns out to be computationally intensive, an approximate criterion is used instead. For a node $C \subseteq \mathcal{X}$, the gradient approximation of $\hat{\theta}_C$ in the parent node is given by:

$$\tilde{\theta}_C = \hat{\theta}_P - \frac{1}{n_C} \sum_{\{i \in \mathcal{J}_{1b} : X_i^* \in C\}} \xi' A_P^{-1} \psi(O_i; \hat{\theta}_P, \hat{\eta}_P), \tag{14}$$

where $\hat{\theta}_P$, $\hat{\eta}_P$ follow from solving (12) once in the parent node, ξ denotes a vector picking out the θ coordinates of (θ, η), and A_P denotes a consistent estimate of the gradient of the expectation of the moment function, i.e. $\nabla \mathbb{E}[\psi(O_i; \hat{\theta}_P, \hat{\eta}_P) | X_i^* \in P]$. We follow Athey et al. (2016b) by using

$$A_P = \frac{1}{n_P} \sum_{\{i \in \mathcal{J}_{1b} : X_i^* \in P\}} \nabla \psi(O_i; \hat{\theta}_P, \hat{\eta}_P). \tag{15}$$

Inserting the gradient approximation (14) for C_1 and C_2 into the criterion (13) and simplifying yields the following steps to compute an approximate criterion. Firstly, the parent node estimates $\hat{\theta}_P, \hat{\eta}_P$ and A_P are calculated by using (12) and (15), after which the pseudo outcomes are determined:

$$\rho_i = -\xi' A_P^{-1} \psi(O_i; \hat{\theta}_P, \hat{\eta}_P) \in \mathbb{R}.$$

Secondly, the following expression for the approximate criterion is filled in:

$$\tilde{\Delta}(C_1, C_2) = \sum_{j=1}^{2} \frac{1}{n_{C_j}} \left(\sum_{\{i \in \mathcal{J}_{1b} : X_i^* \in C_j\}} \rho_i \right)^2 \tag{16}$$

Hence, the optimal split into children C_1, C_2 is determined by maximizing (16). This procedure is recursively repeated by relabeling the obtained children as parent node. We stop when a minimum number of observations in the node has been reached. In contrast to the exact criterion (13), it is possible with the approximate criterion (16) to evaluate all split points for a given covariate in a single pass through the data. This highlights the advantage of Causal Forests in the generalized random forest framework over standard Causal Forests, given that the former uses (13), whereas the latter uses (16).

Under weak assumptions, Athey et al. (2016b) derive that $(\hat{\theta}_0 - \theta_0)/(\sigma^{CF})^2 \to N(0,1)$, where $(\sigma^{CF})^2$ is a scale parameter related to the variance of $\hat{\theta}_0$. Moreover, $(\sigma^{CF})^2$ can be estimated consistently via the Delta method, which states that $(\sigma^{CF})^2 \approx \xi'(V^{-1})H(V^{-1})'\xi$. V is a problem specific curvature parameter that can be estimated using separate standard random forests, see Athey et al. (2016b). In order to estimate H, the subsampling scheme needs to be altered slightly. Instead of random sampling B times a fraction s of the full sample, only $g = 1, \ldots, B/\ell$ random half samples $\mathcal{H}_b$ are drawn, where ℓ denotes little bag size. Next, the estimation samples $\mathcal{I}_b$ used to build the forest are generated such that $\mathcal{I}_b \subseteq \mathcal{H}_{\lceil b/\ell \rceil}$, i.e. the forest is constructed using little bags of ℓ trees, where trees in the same bag use the same half sample. Then, by using the forest score $\widehat{\Psi} = \Psi(O_i; \hat{\theta}, \hat{\eta})$ from (10) and a similar forest score $\widehat{\Psi}(\mathcal{H}_g)_b$ which only uses tree b from little bag g, $\widehat{H}$ is computed via:

$$\widehat{H} = \sum_{g=1}^{B/\ell} \Big(\frac{1}{\ell} \sum_{b=1}^{\ell} \widehat{\Psi}(\mathcal{H}_g)_b - \widehat{\Psi} \Big)^2 - \frac{1}{\ell - 1} \sum_{g=1}^{B/\ell} \Big(\frac{1}{\ell} \sum_{b=1}^{\ell} \big(\widehat{\Psi}(\mathcal{H}_g)_b - \frac{1}{\ell} \sum_{b=1}^{\ell} \widehat{\Psi}(\mathcal{H}_g)_b \big)^2 \Big),$$

such that asymptotic confidence intervals can be constructed.

The previously discussed generalized random forest framework enables HTE estimation in several models. The implementation in each model is characterized by the score function. Firstly, we consider

the Conditional Average Treatment Effect (CATE) model:

$$Y_i = D_i\beta(X_i, H_i) + g_0(X_i, H_i) + U_i, \qquad \mathbb{E}[U_i|D_i, X_i, H_i] = 0$$
$$\beta_0 = \Big[\beta(\text{median}(X_i), H_i = h_1), \ldots, \beta(\text{median}(X_i), H_i = h_k)\Big]'$$

where $\beta(X_i, H_i)$ is an estimable function giving the covariate specific treatment effect and β_0 a $k \times 1$ parameter vector holding HTEs. We follow (Athey et al., 2016b) by setting covariates X_i that we do not inspect for heterogeneity to their median values. To estimate $\beta(X_i, H_i)$ for the CATE model, the following score function is used:

$$\psi(O_i; \theta_0(x^*), \eta_0(x^*)) = (Y_i - \theta_0(x^*)D_i - \eta_0(x^*))(1 \; D_i)',$$

where $O_i = (Y_i, D_i)$, $x^* = (x, h) \in \{(\text{median}(X_i), h_j)\}_{j=1}^k$, $\theta_0(x^*) = \beta_0(x, h)$ and $\eta_0(x^*) = g_0(x, h)$. This yields the closed form expression $\hat{\beta}(x, h)^{CF} = \Big(\sum_i \alpha_i(x)(D_i - \overline{D}_\alpha)^2\Big)^{-1}\Big(\sum_i \alpha_i(x)(D_i - \overline{D}_\alpha)(Y_i - \overline{Y}_\alpha)\Big)$ by solving the moment conditions, where $\overline{D}_\alpha = \sum_i \alpha_i(x)D_i$ and $\overline{Y}_\alpha = \sum_i \alpha_i(x)Y_i$. A derivation for this closed form is given in Section A.4 of the Appendix. After having constructed the forest for the first test point, it is sufficient to keep the forest and compute $\alpha_i(x^*)$ anew for each next test point $j = 2, \ldots, k$.

Obtaining HTEs via the CATE is also possible for IV-identification. We use the CATE IV model then:

$$Y_i = D_i\beta(X_i, H_i) + g_0(X_i, H_i) + U_i, \qquad \mathbb{E}[U_i|Z_i, X_i, H_i] = 0$$
$$\beta_0 = \Big[\beta(\text{median}(X_i), H_i = h_1), \ldots, \beta(\text{median}(X_i), H_i = h_k)\Big]'$$

To estimate $\beta(X_i, H_i)$ in the CATE IV model, a slightly different score function is used:

$$\psi(O_i; \theta_0(x^*), \eta_0(x^*)) = (Y_i - \theta_0(x^*)D_i - \eta_0(x^*))(1 \; Z_i)',$$

where $O_i = (Y_i, D_i, Z_i)$ and from which we calculate $\hat{\beta}(x, h)^{CF}$ using (10).

Although asymptotically valid inference on β_0 is possible, finite sample performance of generalized random forests can still be improved by using an initial orthogonalization step. For that, consider the conditional expectations $\mathbb{E}[Q_i|X_i^* = x^*]$ for each $Q_i \in O_i$. The idea is then to apply preliminary steps $\widetilde{Q}_i = Q_i - \hat{q}^{(-i)}(X_i^*)$ before estimating the forest, where $\hat{q}^{(-i)}(X_i^*)$ correspond to leave-one-out-estimates of the conditional expectations, computed with separate standard random forests. In fact, any residualization scheme that does not depend on the data would work, including cross-fitting. However, leave-one-out-estimation is much more practical in the forest context (Athey et al., 2016b). Finally, we use standard choices of $B = 10,000$ and $\ell = 2$, whereas the sampling fraction s, the minimum node size, the number of variables tried at each split and other parameters are tuned using cross-validation.

4 Revisited Papers and Data

In this section, we introduce the two revisited papers: their goal, data, data preparation, specifications and causal effect identification strategy. Furthermore, we motivate the use of ML, explain our modifications to the original methodology and describe how and why we introduce heterogeneity. A comprehensive overview is given in Table 1.

4.1 Fox News

The revisited paper from DellaVigna and Kaplan (2007) has a twofold purpose. The first is to estimate the causal effect of Fox News channel availability on the Republican vote share in the 2000 U.S. presidential elections, called the Fox News effect. The second is to quantify the media persuasion rate of watching the Fox News channel, since it is a commonly targeted quantity in media studies such that it can be used to compare the persuasive effect of Fox News to other media forms. For the first purpose, a combined data set with information about cable companies, elections and demographics is available. For the second

purpose, there is another data set with household viewing behavior data including Fox News watching, in addition to the first data set. Next, we discuss both data sets in more detail.

The first data set combines data from three main sources. The first source is the Television and Cable Factbook (Warren, 2001), giving information about local cable companies including channel availability. It contains information as of November 2000, i.e. right before the 2000 presidential elections. The second source is the Election Division of the Secretary of State of each state, offering local election data for the years 2000, 1996, 1992 and 1988. The third source is Census, making available demographics from 1990 and 2000 which have been aggregated to the town level. This gives 10,126 towns.

The second data set is from Scarborough research. Scarborough collects demographics and calculates the so-called diary and recall audience measure for TV channels using household panels. The recall audience measure indicates the share of panel respondents who confirm that they watched a channel the past week, while the diary audience measure indicates the share of panel respondents who watched a channel for at least a full half hour block the past week according to their TV diary. The available sample of 11,388 respondents includes the diary audience measure and was recorded between February 2000 and August 2001. It has already been aggregated to the town level via zip codes and matched to the first data set afterwards. This gives 568 towns.

Regarding data preparation for the first data set we follow DellaVigna and Kaplan (2007). We drop 289 towns for which we do not know Fox News availability. Also, we exclude 324 towns that do not have CNN as part of the cable package, because the cable offerings only rebroadcast local cable channels and hence they differ too much from other offerings in terms of content. Finally, we drop 257 towns where the pattern of voting data is problematic, i.e. unrealistic changes between 1996 and 2000 such as a voter turnout difference of more than 100%. A final sample of 9,256 towns for 28 states remains. The second data set is used without preparation, giving us a final sample of 568 towns for 7 states.

To achieve the two purposes of the paper, two specifications are used. The outcome variable in the first specification is the Republican vote share for the 2000 U.S. presidential election on the town level, $v^{Rep}_{k,2000}$. The binary treatment variable is $d^{FOX}_{k,2000}$, valued one if all cable systems in a town offer Fox News in 2000 and zero if no cable system offers Fox News. In addition, there is a large set of control variables available. For the main specification, we firstly have cable system characteristics $C_{k,2000}$ for the year 2000. These are nine dummy variables indicating decile 2 until 10 in the number of channels across towns[6] and another nine indicating decile 2 until 10 in the potential subscribers (i.e. the total voting-age population covered by a cable system) across towns. Secondly, we have the lagged vote share for the 1996 U.S. presidential election, denoted $v^{Rep}_{k,1996}$. Thirdly, we possess twelve town-level demographic summary statistics for both 1990 and 2000, including gender, race, education, income and marital status. We include the demographics for 2000, $X_{k,2000}$, and the difference between 2000 and 1990, $X_{k,00-90}$. The number of controls equals 43 in the main specification of the first model (898 including squares and first order interactions, for methods relying on sparsity). For the dynamic specification, extra lagged vote shares for the 1992 and 1988 U.S. presidential elections are added, which are denoted $v^{Rep}_{k,1992}$ and $v^{Rep}_{k,1988}$. The number of controls rises to 45 in that case (985 for methods relying on sparsity) and the number of observations decreases to $3,722$. Finally, in the second specification the diary audience measure e^{FOX}_{k} is used as outcome variable, while the control set consists of $C_{k,2000}$, $X_{k,2000}$ and $X_{k,00-90}$. The number of controls in the main specification of the second model is 42 (590 for methods relying on sparsity).

We rely on the unconfoundedness assumption to identify causal effects. In a preliminary selection regression, DellaVigna and Kaplan (2007) investigate the plausibility of this assumption. It becomes clear that conditional on the demographic controls $X_{k,2000}, X_{k,00-90}$, cable controls $C_{k,2000}$ and geographical differences, the treatment variable $d^{FOX}_{k,2000}$ is uncorrelated with political outcomes, including the outcome variable $v^{Rep}_{k,2000}$. Hence, they argue that unconfoundedness does not seem to be unrealistic. However, considering that a fairly large number of controls are included, we might improve credibility of the unconfoundedness assumption even further by applying ML in order to handle the high-dimensional controls adequately. An additional advantage is that we are able to control for nonlinear confounding relationships as well with ML. We thus argue that DML and ARB match perfectly with the particular purpose and situation of this paper and we apply both to estimate the ATE.

Methodologically, we follow the same estimation procedure for the ATE as DellaVigna and Kaplan (2007), with a few exceptions. Originally, the observations are weighted with turnout data and the model

[6]Decile 1 is omitted in order to circumvent multicollinearity problems with constant terms.

includes U.S. district or county fixed effects, with county being the smaller geographical level. Weighting is done to improve accuracy of the causal effect estimate. In particular, it ensures that we control for differences in town populations. Moreover, the vote share variables are more accurately measured for larger turnout values, such that weighting with turnout incorporates data quality information. Weighting is however not possible when adopting the interactive model. Apart from weighting the outcome variable and controls, it namely also requires that we weight the treatment variable; that is we multiply the binary treatment values with the turnout for each town. Hence, we obtain a nonbinary treatment variable after weighting, which cannot be used with the interactive model. Thus, for consistency, we decide to leave weights out in all of our models. Taking into account fixed effects is relatively a lot more complicated with our (partially) nonlinear models than in the original linear models. In the linear models, we simply apply within estimation to control for fixed effects. However, this does not work in (partially) nonlinear models and a general method to incorporate fixed effects in these models has not been devised (yet). Therefore, we are forced to leave out fixed effects.

Furthermore, the standard errors are originally clustered because the Fox News availability treatment is assigned per local cable company, which can supply multiple towns, and not per individual town. However, incorporating standard errors that vary across clusters of towns instead of individual towns requires modifications in each of the ML-based causal inference methods. Moreover, we do not expect clustering to lead to very different standard errors because there are only roughly three towns per local cable company. Preliminary checks with the original methods support this. We thus decide to leave out clustering of the standard errors. Finally, the original specification uses as outcome the differenced variable $v^{Rep}_{k,2000} - v^{Rep}_{k,1996}$ to control for the vote share in 1996. Instead, we argue that including $v^{Rep}_{k,1996}$ as an additional control variable is preferred here, since it allows to capture possible nonlinearities involving this lagged vote share for 1996 if we apply ML. To include the vote shares in a consistent manner into the model, we also use the extra lagged control variables $v^{Rep}_{k,1992}$ and $v^{Rep}_{k,1988}$ separately instead of differenced.

Beside using the original procedure to estimate the ATE, we introduce HTEs across geographical Census regions that are organized in a hierarchy, see Figure 13 in Section A.2 of the Appendix. We use three levels with increasing degree of heterogeneity, named level 1, level 2 and level 3. This setup extends the research of DellaVigna and Kaplan (2007), since they only examined broad differences of the Fox News effect between Democratic and Republican states. In practice, one way in which these regional estimates of the Fox News effect are relevant is to make a valid comparison with the findings from other media studies. Many media studies are based on field and laboratory experiments, but these are most often conducted in only a small geographical region due to cost constraints, for example. Then, it would be fairer to compare the findings from such an experiment in a particular region to the regional heterogeneous Fox News effect there, instead of the average Fox News effect. Accordingly, HTEs allow us to compare accurately the persuasive effect of watching the Fox News channel to among others door-to-door canvassing, direct phone calls and watching political ads. We demonstrate the consequences of using HTEs instead of the ATE in the comparison of media persuasion rates when we discuss our results for HTEs.

4.2 Contracts and Trade

In the revisited paper from Nunn (2007), the goal is to estimate the causal effect of contract enforcement on trade flows. The following channel is considered. Bad contract enforcement leads to under-investment, when investments are relationship-specific. Under-investment then induces a cost disadvantage, which in turn leads to less export. The point of departure from other studies that examine the relationship between contract enforcement and trade is that the effect of contract enforcement on the levels of trade is excluded. A data set with several production factor variables for countries and industries is available here, which we discuss more thoroughly in the following.

Firstly, there is export data on the country industry pair level from Feenstra (1996). The primary measure of contract enforcement quality at the country level is the rule of law from Kaufmann et al. (2004), computed as weighted average of several variables indicating perceptions of the effectiveness and predictability of contract enforcement. Nunn (2007) construct a variable indicating the importance of relationship-specific investments across industries based on U.S. IO Tables and data from Rauch (1999) on the need of relationship-specific investments to produce goods. The final variable is given by $z^{rs}_i = \sum_j \theta_{ij} R_j$, where $\theta_{ij} = u_{ij}/u_i$ with u_{ij} the value of input j in industry i and $u_i = \sum_j u_{ij}$ the total

value of all inputs in industry i. R_j indicates what proportion of input j is sold neither on an organized exchange nor reference priced, indicating relationship-specificity of the input. The previously discussed variables give 32,426 country variable pairs. Secondly, there is industry level data on the skill and capital intensities of production from Bartlesman and Gray (1996), but only for manufacturing industries. Skill and capital stock data is also available from Antweiler and Trefler (2002). They measure capital stock as the log of the average capital stock per worker and labor stock as the log of the ratio of workers completing high school to those not completing it. Thirdly, we possess country level log income per capita and log private credit to GDP ratio variables and industry level share of value added in shipments, amount of intra-industry trade, TFP growth in the previous twenty years and Herfindahl index of input concentration variables. Finally, we have countries' legal origins from La Porta et al. (1999), being one of British common law, French civil law, German civil law, Scandinavian civil law or Socialist.

Again, we follow the original paper for data preprocessing. Potentially, there are 32,426 observations available. However, 1,396 observations have missing values and 8,418 observations are valued zero, hence we drop them. Next, for factor endowments and intensities, we only have 12,740 observations. Everything combined, we arrive at a sample of 10,976 observations for $n_c = 70$ countries and $n_i = 128$ industries.

The outcome variable is the log export per industry country pair, $ln(x_{ic})$. The continuous treatment variable is the interaction of the contract enforcement quality variable Q_c with the importance of relationship-specific investments z_i^{rs}. Furthermore, in the main specification, we use two variables to control for the possibility that labor or capital rich countries export more in labor or capital intensive industries. These are the interaction of the skill intensity h_i and skill stock H_c and the interaction of the capital intensity k_i and capital stock K_c. The number of controls equals five in the main specification (20 for methods based on sparsity). For the extended specification, another five variables are constructed to control for a possible higher export of high-income or financially developed countries in certain industries. Firstly, interactions of the log income per capita $ln(y_c)$ with the share of value added va_i, intra-industry trade iit_i, TFP technology growth Δtfp_i, and one minus the Herfindahl index of input concentration $1 - hf_i$ and secondly the interaction of the log of private bank credit to GDP ratio CR_c with the capital intensity. The number of controls rises to 9 in the extended specification (54 for methods based on sparsity) and the number of observations decreases to $10,816$.

We turn to instrumental variables to identify the causal effect since Nunn (2007) argues that there might be reverse causality between trade flows and contract enforcement. This essentially means that high exports in a contract intensive industry improve the quality of contract enforcement in a country. The causal effect of interest can then be isolated by using legal origins as instrumental variable. This is motivated by noticing that legal origins do affect contract enforcement quality and hence trade flows indirectly, but possibly do not affect trade flows directly. The latter becomes invalid if there are other factors that are associated to legal origins, contract enforcement and trade flows. The previously introduced controls may be such factors, if for example certain legal origins also induced high labor or capital endowments. Hence, it is vital to control adequately for the confounding effects from the controls.

The counterintuitive increase of estimates from original IV compared to OLS suggests however that the original linear control function might not capture all confounding effects. Nunn (2007) point out that this might be because it is difficult to capture the effect of country level variables on trade flows since it requires to identify the correct industry level variable for specialization. Given the complexity of relationships here, we might be able to execute this task much better by using flexible controlling with ML. We therefore argue that ML could improve the results here. It is however not possible to adopt the interactive model because this model is restricted to binary treatment variables. Thus, we can only apply DML in the partially linear model, but not ARB and DML in the interactive model.

For estimating the ATE, we remain with the original methodology from Nunn (2007) but with a single modification. As discussed before, allowing for fixed effects in our (partially) nonlinear models is much less straightforward than in the original linear models. Nonetheless, contrary to the Fox News application, we do not think it is a good idea to simply drop the industry and country fixed effects here. To be specific, these fixed effects are primarily used to control for the effect of contract enforcement on the levels of trade. This ensures that we focus on trade flows instead of levels, which is a vital part of the study of Nunn (2007). Consequently, we want to keep the focus on trade flows but we cannot apply within estimation in our nonlinear models to take into account fixed effects, as explained before. Hence, we apply an ad hoc procedure to still incorporate them. Consider a variant of the original linear model

in this application:

$$Y_{ic} = \mu + \alpha_i + \alpha_c + X_{ic}\beta_{lin} + U_{ic}, \tag{17}$$

where $Y_{ic} = ln(x_{ic})$ is the outcome, $X_{ic} = (D_{ic}\ C_{ic})$ a vector consisting of the treatment $D_{ic} = z_i^{rs}Q_c$ and control variable vector C_{ic} and U_{ic} an unobserved error term for industry i in country c. μ and β_{lin} are unknown parameters and α_i, α_c correspond to industry and country fixed effects. Then, we can estimate the fixed effects as follows (Greene, 2002, Section 13.3.3):

$$\hat{\alpha}_i = (\overline{Y}_{i.} - \overline{Y}) - (\overline{X}_{i.} - \overline{X})\hat{\beta}_{lin}^W \qquad i = 1, \ldots, n_i$$
$$\hat{\alpha}_c = (\overline{Y}_{.c} - \overline{Y}) - (\overline{X}_{.c} - \overline{X})\hat{\beta}_{lin}^W \qquad c = 1, \ldots, n_c$$

where $Y_{i.}$ is the average of Y_{ic} over countries, $Y_{.c}$ the average of Y_{ic} over industries, $\overline{Y}$ the overall average and similarly so for X_{ic}. $\hat{\beta}_{lin}^W$ is the two way fixed effects within estimator in the linear model (17). Hence, we can still take into account fixed effects by adding these estimated fixed effects from the linear model as control variables in our nonlinear models. We call them the industry and country level variables. The 4 and 9 controls in the main and extended specification include these country and industry level variables.

Next, we introduce HTEs across the number of inputs in the production process, because it serves as measure of the difficulty of vertical integration, assuming that there are fixed costs associated with the production of each input. As such, we build on the insight that under-investment due to bad contract enforcement might be reduced by means of vertical integration. It enables us to investigate the degree to which vertical integration affects the relationship between contract enforcement and trade flows. This setup extends the results of Nunn (2007), who only investigates broad differences in the causal effect between industries with a number of inputs above or below the median. Since heterogeneous DML has been designed for causal effect identification via unconfoundedness, we cannot apply it here without rigorous modifications. Thus, we only apply the Causal Forest in the generalized random forest framework, which does include IV-identification.

Table 1: Summary of the revisited applied papers' specifications

	Fox News			**Contracts and Trade**	
	Vote share		*Audience*	*Export*	
	Main	Dynamic	Main	Main	Extended
# Obs.	9,256	3,722	568	10,976	10,816
# Controls	(43, 898)	(45, 985)	(42, 590)	(4, 20)	(9, 54)
Y_i	$v_{k,2000}^{Rep}$	$v_{k,2000}^{Rep}$	e_k^{FOX}	$ln(x_{ic})$	$ln(x_{ic})$
D_i	$d_{k,2000}^{FOX}$	$d_{k,2000}^{FOX}$	$d_{k,2000}^{FOX}$	$z_i^{rs}Q_c$	$z_i^{rs}Q_c$
X_i	$v_{k,1996}^{Rep}$	$v_{k,1996}^{Rep}$	$X_{k,2000}$	Country lvl	Country lvl
	$X_{k,2000}$	$X_{k,2000}$	$X_{k,90-00}$	Industry lvl	Industry lvl
	$X_{k,90-00}$	$X_{k,90-00}$	$C_{k,2000}$	k_iK_c	k_iK_c
	$C_{k,2000}$	$C_{k,2000}$	-	h_iH_c	h_iH_c
	-	$v_{k,1992}^{Rep}$	-	-	$va_iln(y_c)$
	-	$v_{k,1988}^{Rep}$	-	-	$iit_iln(y_c)$
	-	-	-	-	$\Delta tfp_iln(y_c)$
	-	-	-	-	$(1 - hf_i)ln(y_c)$
	-	-	-	-	k_iCR_c
Small H_i	Regions (4)	-	-	# Inputs (c)	-
Medium H_i	Divisions (9)	-	-	-	-
Large H_i	States (28)	-	-	-	-
Z_i	-	-	-	Legal origins	Legal origins

Note: the number of controls is denoted (s, l), with s the actual amount and l the amount including all first order interactions and squares, used for methods relying on sparsity. H_i is accompanied by the amount of variables in parentheses, which needs to be added to # Controls to get the total number of controls for HTE estimation. For example, the total number of controls in the main specification with a small degree of heterogeneity is $43 + 4 = 47$. (c) for contracts and trade heterogeneity denotes that H_i is a count variable instead of a binary variable.

5 Results Average Treatment Effect

5.1 Double Machine Learning: Fox News

5.1.1 Vote share regression

We apply DML to estimate the ATE of Fox News availability on the Republican vote share in 2000. We start with the main specification which includes the demographic controls, cable controls and the lagged Republican vote share for 1996. The results are presented in Table 2, where double asterisks indicate significance of the DML estimate at the 5% level, assessed with median standard errors. A description of each of the ML submethods can be found in Section A.3 of the Appendix. Fraction best ML method denotes the fraction of nuisance function estimation problems in which the ML submethod outperforms all the other ML submethods in terms of out-of-sample MSE, over the folds and S replications. It reflects relative performance of the ML submethods and provides more information on which ML submethods are combined into the hybrid Best ML submethod.

In the PLR model, we obtain DML estimates of roughly 0.2 to 0.6%. With the median standard errors, the DML estimates are significant at the 5% level[7] for the tree-based ML submethods regression tree, boosting and random forest. With the conventional standard errors, we find significant DML estimates for the LASSO and neural network as well, in addition to those from the tree-based ML submethods. The qualitative difference is due to the fact that the two kinds of standard errors differ noticeably sometimes, which is in line with the results from several empirical applications of Chernozhukov et al. (2016). It shows the enlarging effect on the standard errors of incorporating variation from random sample splitting. Overall, the DML estimates in the PLR model suggest a positive Fox News effect, that is Fox News availability led to a higher Republican vote share in 2000.

Table 2: DML estimates for the Fox News vote share ATE using the main specification, $S = 50$

	SVM	**LASSO**	**Reg. Tree**	**Boost-ing**	**Random Forest**	**Neural Net.**	**Best**
A. Partially Linear Regression Model							
Fraction best ML submethod	0.50	0.00	0.00	0.00	0.50	0.00	–
Median ATE (2 fold)	0.0018	0.0034*	0.0059**	0.0054**	0.0044**	0.0037*	0.0020
	[0.0022]	[0.0019]	[0.0028]	[0.0023]	[0.0019]	[0.0020]	[0.0022]
	(0.0013)	(0.0013)	(0.0018)	(0.0016)	(0.0014)	(0.0013)	(0.0013)
B. Interactive Regression Model							
Fraction best ML submethod	0.67	0.00	0.00	0.00	0.33	0.00	–
Median ATE (2 fold)	0.0014	0.0031	0.0073	0.0054**	0.0074*	0.0040	0.0015
	[0.0027]	[0.0024]	[0.0046]	[0.0025]	[0.0039]	[0.0038]	[0.0033]
	(0.0015)	(0.0014)	(0.0023)	(0.0015)	(0.0020)	(0.0032)	(0.0014)

Note: Fraction best ML method denotes the fraction of nuisance function estimation problems in which the ML submethod outperforms all the others in terms of out-of-sample MSE, over the folds and S replications. We present conventional standard errors in parentheses and median standard errors that incorporate the variation of sample splitting in square brackets. The symbols $*$, $**$ and $***$ denote significance at the 10, 5 and 1% level, respectively, assessed with median standard errors.

In the interactive regression model, we do not make any functional form assumption with respect to the joint effect of Fox News availability and the controls on the Republican vote share in 2000. The DML estimates range from approximately 0.1 to 0.7% here. With median standard errors, we find a significant DML estimate only for boosting. With conventional standard errors, we obtain a significant DML estimate for all tree-based ML submethods and the LASSO. Again, we there is some discrepancy

[7] In the following, we use the 5% level to assess statistical significance.

between the two standard errors. The DML estimates in the interactive model also indicate a positive Fox News effect, but the insignificance of many estimates suggests that we cannot rule out that there is no Fox News effect. The DML estimates in the interactive model are not consistently smaller or larger than in the PLR. However, we observe a wider range of ATE estimates across the ML submethods in the interactive model compared to the PLR model. Moreover, the standard errors in the interactive model are larger than in the PLR model in 13 out of 14 cases. The wider range and larger standard errors reflect a larger uncertainty around the estimates when allowing for fully heterogeneous treatment effects. Preciseness of the DML estimates depends thus highly on the model assumptions.

Table 3: Original ATE estimates with corresponding table numbers and columns from DellaVigna and Kaplan (2007)

	Vote share			
	Unweighted		Weighted	
No controls				
Diff-in-diff	-0.0124^{***}	Earlier results	-0.0025	Table IV column (1)
Main specification				
No f.e. (LS)	0.0027^{**}	Earlier results	0.0080^{***}	Table IV column (3)
District f.e.	0.0014	Table A.III column (1)	0.0042^{***}	Table IV column (4)
County f.e.	0.0040^{***}	Table A.III column (2)	0.0069^{***}	Table IV column (5)

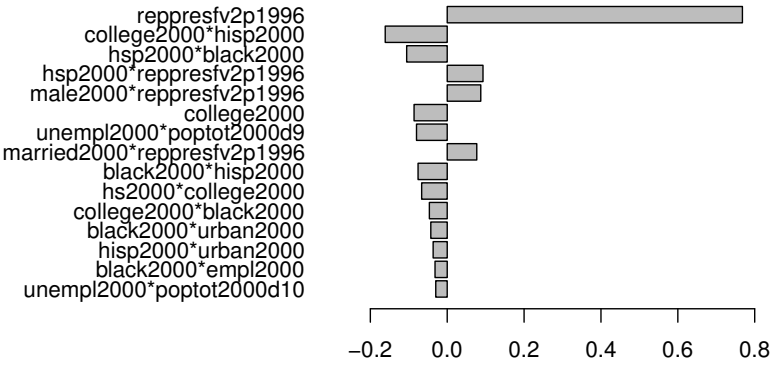
Note: Unweighted and weighted estimation corresponds to using the observations directly and after weighting them by the amount of votes cast in 1996, respectively. Diff-in-diff denotes the simple difference-in-difference estimator computed with only the 2000 and 1996 vote shares. No f.e. (LS), district f.e. and county f.e. add controls from the main specification, see Table 1. No f.e. (LS) denotes the least squares estimator without fixed effects, while district f.e. and county f.e. denote different fixed effect estimators. The symbols $*$, $**$ and $***$ denote significance at the 10, 5 and 1% level, respectively. Earlier results indicates an estimate that belongs to earlier original methodology, but where originally a different estimation sample was used. We computed new values by using the final instead of the earlier estimation sample to exclude sample effects from comparisons.

For comparison, we present the original results in Table 3. In contrast to the difference-in-difference estimators, we do not find any counterintuitive negative Fox News effect estimates with DML, neither in the PLR nor in the interactive model. This indicates that DML can have an advantage over prevailing simple causal effect estimation methods. Furthermore, the DML estimates in either of the models are larger than the original unweighted least squares (OLS) estimate[8]. The estimate may increase here due to adequate controlling for high-dimensional and nonlinear confounding relationships via ML.

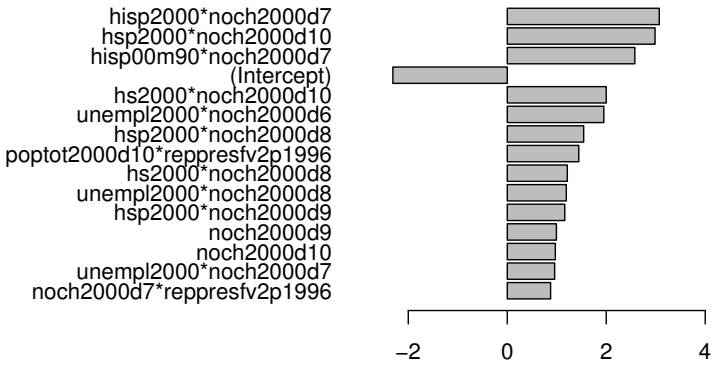
To examine the importance of nonlinear confounding further, we use the LASSO since it is the only ML submethod that produces readily comparable estimates of the nuisance effects of different nonlinear transformations of the controls. For the LASSO, these nonlinear transformations are interactions and squared terms. The LASSO thus estimates coefficient sizes for the controls, their interaction terms and their squared terms. The mean LASSO coefficient sizes over DML folds and replications then indicate which of these terms have the largest nuisance effect. The mean LASSO coefficient estimates for the main specification are plotted in Figure 1. We observe many interaction terms with a relatively large coefficient size in the flexible control functions, for the Republican vote share in 2000 (Y) and for Fox News availability (D). This suggests that nonlinear confounding indeed plays a role here, possibly explaining part of the increased coefficient estimate compared to original OLS, which employs a fully linear control specification. To illustrate, we consider the dynamic effect of the lagged Republican vote share of 1996 (reppresfv2p1996) on that of 2000 (Y). The interactions suggest that this dynamic effect depends on the percentage of town inhabitants in 2000 that are college educated (hsp2000), male (male2000) and married (married2000). In other words, nonlinear political trends exist. Furthermore, the probability that Fox News is available in a town (D) when there are relatively many channels (noch2000d8, noch2000d9, noch2000d10) is higher for a greater percentage of inhabitants in 2000 with lower education (only high

[8]This does not hold for DML with the SVM and Best ML submethods. Given the high fraction of prediction problems in which the SVM tops the ML submethods in terms of out-of-sample MSE (fraction best ML method), it follows that the Best ML submethod depends highly on the SVM and thus shows very similar estimates. The smaller size of the DML estimate due to SVM could relate to improper tuning, since SVMs need to be tuned in order to work well.

school, hs2000; some college, hsp2000). This possibly reflects the (complex) business strategy of Fox News to introduce their channel sooner to lower educated people in competitive markets. Both of these nonlinear nuisance effects cannot be fully captured with a linear control specification.



(a) $\mathbb{E}[Y|X]$



(b) $\mathbb{E}[D|X]$

Figure 1: Fox News LASSO terms with the largest absolute mean coefficient size over the folds and S replications, in the PLR using the main specification. * in the name of the term denotes an interaction. The abbreviations for the control variables are explained in Table 27 in Section A.2 of the Appendix.

Next, we return to our comparison with the original estimates in Table 3, but concentrate on the more advanced original estimates, i.e. those that use controls and apply weighting and/or fixed effects. We see that the DML estimates in both the PLR and interactive model are slightly attenuated relative to the weighted least squares (WLS) estimate. However, they are similar to the unweighted original estimate with either of the fixed effects and the weighted original estimate with district fixed effects. Although being insignificant, some of the DML estimates in the interactive model are also in line with the weighted original estimate with county fixed effects. We observe fairly much similarity overall, indicating that DML and these more advanced original methods capture roughly the same causal effects here. Both types of methods correct for confounding in a very different way, however, such that DML strengthens the original results. Thus, DML might be a valuable alternative to established advanced causal effect estimation methods.

Credibility of the unconfoundedness assumption might be improved further by using more historical vote shares, since it increases the variety of political trends that we are able to capture. Hence, we add the lagged Republican vote shares for 1992 and 1988 to the controls, leading to the dynamic specification. The downside of this specification is that the sample size reduces to 3,722 towns. We present DML estimates of the ATE of Fox News availability on the Republican vote share in 2000 in Table 4. In the PLR model, we get estimates of roughly 0.7 to 1.2%. All of them are significant regardless of whether we use median or conventional standard errors. In the interactive model, we find values ranging from approximately 0.6 to 1.2%. All estimates except the one for the neural network are significant with both median and conventional standard errors. DML estimates are insignificant for the neural network due to

very large standard errors. Further inspection suggests that this is caused by propensity score weights close to 0, even after trimming them at 0.01. This motivates substituting the propensity score weights with other weights using ARB empirically as well. We also observe once more a wider range of estimates and larger standard errors (in 10 out of 14 cases) in the interactive model compared to the PLR model. All in all, the DML estimates for the dynamic specification strongly indicate a fairly large positive Fox News effect, in both models.

Table 4: DML estimates for the Fox News vote share ATE using the dynamic specification, $S = 50$

	SVM	**LASSO**	**Reg. Tree**	**Boost-ing**	**Random Forest**	**Neural Net.**	**Best**
A. Partially Linear Regression Model							
Fraction best ML method	0.41	0.00	0.00	0.00	0.54	0.05	–
Median ATE (2 fold)	0.0073***	0.0083***	0.0115***	0.0109***	0.0082***	0.0077***	0.0075***
	[0.0022]	[0.0019]	[0.0035]	[0.0035]	[0.0021]	[0.0022]	[0.0020]
	(0.0019)	(0.0018)	(0.0024)	(0.0021)	(0.0019)	(0.0019)	(0.0018)
B. Interactive Regression Model							
Fraction best ML method	0.60	0.00	0.00	0.00	0.37	0.03	–
Median ATE (2 fold)	0.0075**	0.0064**	0.0092***	0.0102***	0.0117***	0.0083	0.0071**
	[0.0036]	[0.0025]	[0.0035]	[0.0028]	[0.0037]	[0.0079]	[0.0036]
	(0.0026)	(0.0017)	(0.0029)	(0.0020)	(0.0022)	(0.0059)	(0.0026)

Note: Fraction best ML method denotes the fraction of nuisance function estimation problems in which the ML submethod outperforms all the others in terms of out-of-sample MSE, over the folds and S replications. We present conventional standard errors in parentheses and median standard errors that incorporate the variation of sample splitting in square brackets. The symbols $*$, $**$ and $***$ denote significance at the 10, 5 and 1% level, respectively, assessed with median standard errors.

A higher number of folds in the DML procedure ensures a larger amount of observations to be used for learning the control functions with ML. Considering that the latter seems to be the most difficult task within the DML procedure, we might improve DML estimates by using more folds. We therefore report the DML estimates in the dynamic specification with 5 instead of 2 folds in Table 20 in Section A.1 of the Appendix. In the PLR and interactive model, we find values of $0.7 - 1.2\%$ and $0.6 - 1.1\%$, respectively. With median standard errors, all DML estimates are significant in the PLR model and all except the ones for the neural network and Best ML submethod are significant in the interactive model. Furthermore, we point out that the standard errors from DML with 5 folds are greater or equal than those from DML with 2 folds in the PLR. This was also found by Chernozhukov et al. (2016), when estimating the effect of institutions on economic growth using DML under IV-identification. All in all, we conclude that the DML estimates with 5 folds resemble those with 2 folds. The amount of folds used within DML does not seem to be very important.

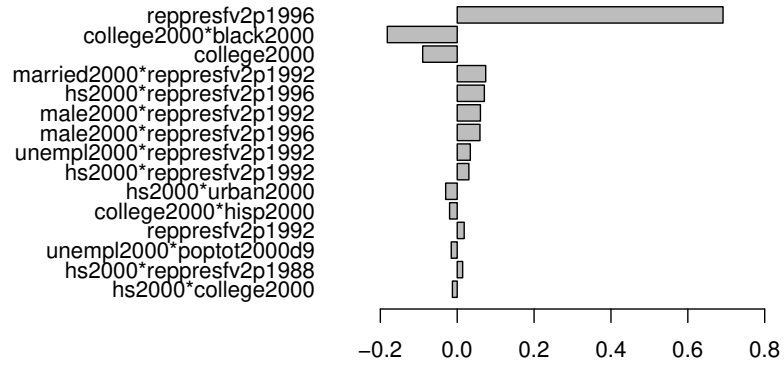
We provide the original results for the dynamic specification in Table 5. The DML estimates from Table 4 in either of the models are all greater than the original unweighted least squares (OLS) estimate. Capturing high-dimensional, possibly nonlinear confounding relationships via ML could have increased the estimates once again. To gain further understanding in these relationships, we plot the mean coefficient estimates of the LASSO terms for the dynamic specification in Figure 2. We find a lot of interactions among the terms with relatively large coefficient size in the flexible control functions, for the Republican vote share in 2000 (Y) and Fox News availability (D). This indicates the importance of controlling for nonlinear nuisance effects. To give an example, we consider the dynamic effect of the lagged Republican vote share controls for 1996 (reppresfv2p1996), 1992 (reppresfv2p1992) and 1988 (reppresfv2p1988) on the Republican vote share in 2000 (Y). There are many interactions involving these lagged controls and for 1992 the interactions rank even higher than the lagged variable itself. This indicates again that nonlinearities happen to be important for political trends, also over a longer time period. Additionally,

the probability that Fox News is available (D) when the cable company in a town reaches a percentage of voting age inhabitants below average in 2000 (poptot2000d4) is higher in towns with more unemployment (unempl2000). This possibly reflects the business strategy of the Fox News channel to introduce sooner in rural areas with more unemployment. We highlight once again that trends like these cannot be captured with a fully linear control specification.

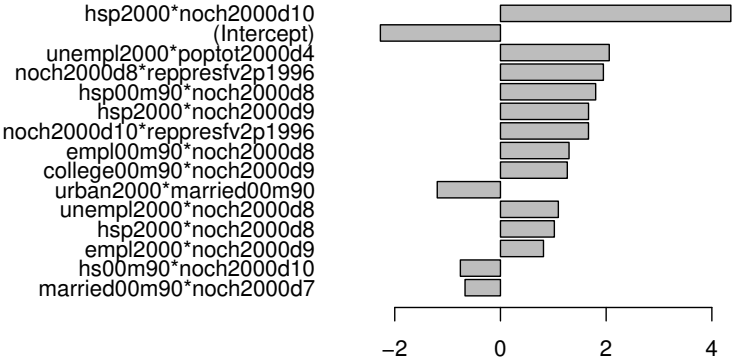
Table 5: Original ATE estimates with corresponding table numbers and columns from DellaVigna and Kaplan (2007)

	Vote share			
	Unweighted		Weighted	
Dynamic specification				
No f.e. (LS)	0.0055^{***}	Earlier results	0.0090^{***}	Earlier results
District f.e.	0.0024	Earlier results	0.0037^{*}	Table IV column (6)
County f.e.	0.0055^{**}	Earlier results	0.0048^{**}	Table IV column (7)

Note: Unweighted and weighted estimation correspond to using the observations directly and after weighting them by the amount of votes cast in 1996, respectively. Diff-in-diff denotes the simple difference-in-difference estimator computed with only the 2000 and 1996 vote shares. No f.e. (LS), district f.e. and county f.e. add controls from the dynamic specification, see Table 1. No f.e. (LS) denotes the least squares estimator without fixed effects, while district f.e. and county f.e. denote different fixed effect estimators. The symbols $*$, $**$ and $***$ denote significance at the 10, 5 and 1% level, respectively. Earlier results indicates an estimate that belongs to earlier original methodology, but where originally a different estimation sample was used. We computed new values by using the final instead of the earlier estimation sample to exclude sample effects from comparisons.



(a) $\mathbb{E}[Y|X]$



(b) $\mathbb{E}[D|X]$

Figure 2: Fox News LASSO terms with the largest absolute mean coefficient size over the folds and S replications, in the PLR using the main specification. * in the name of the term denotes an interaction. The abbreviations for the control variables are explained in Table 27 in Section A.2 of the Appendix.

We move back to Table 5 now to compare DML to the more advanced original estimates in the dynamic specification. In both the PLR and the interactive model, the DML estimates are comparable to the original weighted least squares estimate, but they exceed the unweighted and weighted original estimates with either of the fixed effects. Moreover, the DML estimates are generally significant whereas the original unweighted and weighted district fixed effect estimates are not. DML and the more advanced original methods shows overall a lot of difference; they capture other causal effects. A possible explanation is that DML modifies the advanced original estimates in order to strengthen them, since DML is able to correct adequately for nonlinear nuisance effects such as nonlinear political trends. The difference between DML and the advanced original estimates in the dynamic specification contrasts with findings in the main specification. The extra dynamic controls may play a role in this, but we should keep in mind that a different sample is used in both specifications, which could also have an influence.

5.1.2 Audience regression and persuasion rates

Next, we compute what fraction of viewers actually got persuaded by the Fox News channel to vote Republican in 2000, i.e. the Fox News persuasion rate. DellaVigna and Kaplan (2007) determine the persuasion rate for the in Section 4.2 described recall audience and compute it as follows:

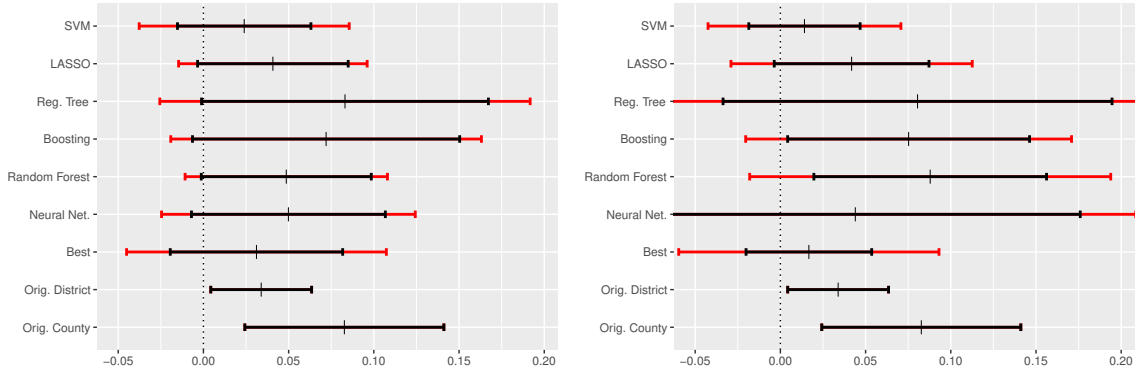
$$f = \frac{\hat{\beta}_{0,VoteShare}}{\hat{\beta}_{0,Audience}} \frac{t_C t_T}{c_{Recall} d}, \tag{18}$$

where $t_C = t_T = 0.560$ corresponds to the turnout fraction for control and treatment group, $d = 0.306$ to the share of Democratic voters and $c_{Recall} = 3.43$ to a constant converting the diary audience to the recall audience. t_C and t_T are used here to correct for turnout effects, whereas d is used to differentiate between convincing a Democrat or a nonvoter to vote Republican. In addition, $\hat{\beta}_{0,VoteShare}$ is an estimate for the effect of Fox News availability on the Republican vote share in 2000, while $\hat{\beta}_{0,Audience}$ is an estimate of the effect of Fox News availability on Fox News watching. Since we obtain main specification ATE estimates $\hat{\beta}_{0,VoteShare}$ from Section 5.1.1, we only need to run a second regression of Fox News watching on Fox News availability in order to obtain $\hat{\beta}_{0,Audience}$ and compute the persuasion rate. We refer to this regression as the audience regression.

We apply DML in the audience regression to estimate the ATE of Fox News availability on Fox News watching. Our specification includes the demographic and cable controls that were used before. The results are given in Table 21 in Section A.1 of the Appendix. In the PLR model, we get DML estimates ranging approximately from 1.9 to 2.7% and being significant for 5 out of 7 ML submethods with median standard errors. In the interactive model, we find a range of 2.1 to 2.9% with significant effects for only 3 out of 7 ML submethods with median standard errors. Thus, Fox News availability led to Fox News watching for $2 - 3\%$ of the households, but preciseness of the estimates depends again highly on the model assumptions. Comparing with the original estimates in Table 19 in Section A.1 of the Appendix, we observe that most DML estimates in the PLR model are broadly consistent with the original difference-in-difference and unweighted least squares estimates. Firstly, this indicates that prevailing simple causal effect estimation methods still suffice in this case. Secondly, it follows that controlling for high-dimensional and nonlinear confounding via ML has a smaller impact than in the vote share regression. The lack of dynamics in the audience regression possibly explains this, given our previous finding that nonlinearities are particularly important for dynamic trends. The smaller sample size could also have an influence, since ML methods often require more observations than linear methods to perform optimally, due to their flexibility.

After having estimated the audience regression, we compute Fox News persuasion rates from DML using (18) and plot them in Figure 3 together with the original persuasion rate estimates. We also present corresponding confidence intervals for median and conventional standard errors. Median standard errors always produce wider confidence intervals since they incorporate an extra source of uncertainty: random sample splitting. In the PLR, the DML estimates for the SVM, LASSO and Best ML submethods lie close to the original district fixed effect estimate, while those for the regression tree and boosting ML submethods lie close to the original county fixed effect estimate. Note however that the confidence intervals are a lot wider in both cases for median and conventional standard errors. Not clustering the standard errors for towns with the same local cable company with DML could partly explain this result.

In the interactive model, the DML estimates for the tree-based ML submethods (regression tree, boosting, random forest) approach the original county fixed effect estimate, whereas those for the LASSO and neural network approach the original district fixed estimate. Yet again, we see much wider confidence intervals, but this is partly the consequence of adopting the interactive model as well now. The extreme widths for the regression tree and neural network[9] ML submethods follow from the large audience regression standard errors and relate possibly to the small sample size there. Hence, DML produces persuasion rate estimates similar to the original estimates, but we obtain larger standard errors such that we cannot exclude that Fox News watching did not actually persuade people to vote Republican.



(a) PLR Model (b) Interactive Model

Figure 3: DML estimates and 95% confidence intervals for persuasion rates. Standard errors are computed via the Delta method[10], following DellaVigna and Kaplan (2007). Confidence intervals computed with conventional standard errors are presented in black, while those computed with median standard errors are given in red. The original weighted district and county fixed effect persuasion rate estimates are denoted Orig. District and Orig. County, respectively.

5.2 Double Machine Learning: Contracts and Trade

Next, we employ DML to estimate the effect of contract enforcement quality on trade flows. We begin with the main specification including the labor and capital interactions as controls, in addition to the country and industry level variables. The results are presented in Table 6. Firstly, without turning to instrumental variables yet, we observe estimates of roughly $0.29 - 0.38$ in the PLR model, all of them significant with either of the standard errors. The two standard errors differ however still noticeably, with even a different order of magnitude here. Without incorporating potential reverse causality, we find that countries with better contract enforcement export relatively more in relationship important industries.

When using instrumental variables by adopting the PLIV model, we obtain estimates of approximately $0.17 - 0.37$. All of them except those for the LASSO and regression tree are significant with median standard errors. The DML estimate for the LASSO becomes significant with conventional standard errors. The large difference between the two standard errors from the PLR is maintained. After correcting for potential reverse causality, we continue to find a positive effect of contract enforcement quality on comparative advantage in contract intensive industries. However, most importantly, we notice that for each of the ML submethods, the estimate is reduced compared to the PLR. This decrease from IV is largest for the tree-based ML submethods[11] and only small for the LASSO and neural network. The decrease suggests possibly that reverse causality exists. In this case, reverse causality would be a positive effect of comparative advantage in contract intensive industries on contract enforcement quality, which is in line with our intuition.

[9]The median standard error confidence interval is too wide for the neural network to represent it completely in the Figure.

[10]The Delta method computes first order approximate standard errors for any function of available estimates, which in our case given by (18), by using the corresponding estimated covariance matrix.

[11]We also count the Best ML submethod as part of the tree-based ML submethods in this subsection because it is essentially equal to the random forest, as follows from the fact that fraction best ML method for the latter is equal to 1.

Table 6: DML estimates for the contracts and trade ATE using the main specification, $S = 50$

	SVM	**LASSO**	**Reg. Tree**	**Boost-ing**	**Random Forest**	**Neural Net.**	**Best**
A. Partially Linear Regression Model							
Fraction best ML method	0.00	0.00	0.00	0.00	1.00	0.00	–
Median ATE (2 fold)	0.3683***	0.3795***	0.2870***	0.3245***	0.3194***	0.3640***	0.3194***
	[0.0444]	[0.0550]	[0.0405]	[0.0381]	[0.0318]	[0.0393]	[0.0313]
	(0.0070)	(0.0071)	(0.0084)	(0.0080)	(0.0087)	(0.0071)	(0.0087)
B. Partially Linear IV Model							
Fraction best ML method	0.00	0.00	0.00	0.00	1.00	0.00	–
Median ATE (2 fold)	0.2472***	0.3720*	0.1653	0.1661***	0.1791**	0.3408***	0.1791**
	[0.0839]	[0.1938]	[0.1258]	[0.0476]	[0.0908]	[0.1314]	[0.0904]
	(0.0271)	(0.0269)	(0.0881)	(0.0437)	(0.0772)	(0.0342)	(0.0772)

Note: Fraction best ML method denotes the fraction of nuisance function estimation problems in which the ML submethod outperforms all the others in terms of out-of-sample MSE, over the folds and S replications. We present conventional standard errors in parentheses and median standard errors that incorporate the variation of sample splitting in square brackets. The symbols $*$, $**$ and $***$ denote significance at the 10, 5 and 1% level, respectively, assessed with median standard errors.

Table 7: Original ATE estimates with corresponding table numbers and columns from Nunn (2007)

	Export			
	No Instrumental Variables		Instrumental Variables	
Main specification	0.3260***	Table VII column (3)	0.5390***	Table VII column (4)
Extended specification	0.2960***	Table VII column (5)	0.5200***	Table VII column (6)

Note: The symbols $*$, $**$ and $***$ denote significance at the 10, 5 and 1% level, respectively. Components of the main and extended specification are given in Table 1.

Comparing the DML estimates to the original results in Table 7, it becomes clear that DML in the PLR model produces estimates similar to the original estimate using the main specification. Especially DML with tree-based ML submethods seems to come very close to this estimate, while DML with the other ML submethods gives slightly larger estimates. Hence, there does not seem to be much nonlinear confounding. Further, we find that the DML estimates in the PLIV model are attenuated compared to the original IV estimate using the main specification. The original IV estimate is larger than the original estimate without IVs and hence indicates a counterintuitive negative reverse causal effect, stating that more export in relationship important industries would lead to reduced contract enforcement quality. In contrast, our DML estimates in the PLIV model are smaller than the DML estimates in the PLR model and for most ML submethods also smaller than the original estimate without IVs. Therefore, DML suggests a more plausible positive reverse causal effect, which agrees with the positive reverse causal effect that was found by Nunn (2007) by using propensity score matching. A possible explanation is that we more successfully control for alternative ways in which the legal origin instruments affect trade flows by using ML, instead of the original simple linear method. The key advantage of ML might be that it enables to pick up complex patterns of industry specialization for country controls that could not be identified before.

To inspect the importance of nonlinearities further, we plot the mean LASSO coefficient sizes in Figure 10 in Section A.1 of the Appendix. We find a few interactions and a squared term with relatively large LASSO coefficient, suggesting that nonlinear nuisance effects may play a role here. This could possibly explain the positive reverse causal effects from DML. For example, we find a squared term in the relationship between the labor variable (skill1_times_at_hk) and legal origins (Z). This suggest decreasing marginal returns to better legal origins here, which cannot be incorporated with fully linear

controlling. Note however that the LASSO might not be the best method to reveal important nonlinear terms here, since the found positive reverse causal effect is only small for the LASSO relative to the other ML submethods. This perhaps implies that nonlinearities other than first order interactions and squared terms are important here.

There are economic consequences from the positive reverse causal effect belonging to the attenuation of the DML estimate with IVs compared to the original or DML estimate without IVs. Nunn (2007) concludes that the effect of contract enforcement quality on trade flows is much larger than the effect of capital and labor interactions combined (the latter effect being equal to 0.19, see Table IV column (3)). This follows indeed from the original estimates with or without IVs in the main specification. Conversely, our DML estimates in the PLIV model range from $0.17 - 0.18$ for 4 out of 7 ML submethods, possibly suggesting that the effect of contract enforcement quality on trade flows is more or less equal to the combined effect of the capital and labor variables. Thus, the effect of contract enforcement might have been overstated originally due to inadequate removal of reverse causal effects.

Subsequently, we try to make the IV exclusion restriction of no alternative channel through which legal origins affect trade flows other than contract enforcement even more plausible by including additional control variables. More specifically, we control for a possible comparative advantage of high income countries in certain industries[12] and a possible effect of financial development on trade flows. This gives us the extended specification, with a decreased sample size of $10,816$ observations. We present DML estimates for the effect of contract enforcement quality on trade flows in Table 8. To start with, we inspect DML in the PLR model, which does not employ the instrumental variables. The estimates range from $0.10 - 0.30$ and are significant except for the regression tree. It seems that the tree-based ML submethods produce substantially smaller DML estimates than the other ML submethods. Overall, without taking into account reverse causality, we still find that countries with better contract enforcement export relatively more in relationship important industries.

Table 8: DML estimates for the contracts and trade ATE using the extended specification, $S = 50$

	SVM	**LASSO**	**Reg. Tree**	**Boost-ing**	**Random Forest**	**Neural Net.**	**Best**
A. Partially Linear Regression Model							
Fraction best ML method	0.00	0.00	0.00	0.00	1.00	0.00	–
Median ATE (2 fold)	0.2993***	0.2950***	0.1021	0.1470***	0.1993**	0.2959***	0.1993**
	[0.0999]	[0.0955]	[0.1017]	[0.0538]	[0.0916]	[0.0937]	[0.0929]
	(0.0075)	(0.0076)	(0.0102)	(0.0099)	(0.0142)	(0.0074)	(0.0142)
B. Partially Linear IV Model							
Fraction best ML method	0.00	0.00	0.00	0.00	1.00	0.00	–
Median ATE (2 fold)	0.2165***	0.4975**	0.2907	0.3061***	0.2305	0.2687***	0.2305*
	[0.0667]	[0.2316]	[0.3860]	[0.0985]	[0.1406]	[0.0570]	[0.1282]
	(0.0415)	(0.0477)	(0.2820)	(0.0723)	(0.1203)	(0.0471)	(0.1203)

Note: Fraction best ML method denotes the fraction of nuisance function estimation problems in which the ML submethod outperforms all the others in terms of out-of-sample MSE, over the folds and S replications. We present conventional standard errors in parentheses and median standard errors that incorporate the variation of sample splitting in square brackets. The symbols $*$, $**$ and $***$ denote significance at the 10, 5 and 1% level, respectively, assessed with median standard errors.

Next, for the PLIV model, we observe DML estimates of $0.22 - 0.50$, being significant with either of the two standard errors except for the regression tree, random forest and Best ML submethods. After correcting for potential reverse causality, we still have a significant positive effect of contract enforcement quality on trade flows for DML with the SVM, LASSO, neural network and boosting. Furthermore,

[12] These are industries with high value added, fragmentation of the production, rapid technological development and high product complexity.

like in the main specification, the estimates decrease compared to the PLR model for the SVM and neural network. They become even smaller than the DML estimate in the PLIV model using the main specification. This might indicate that we capture an even larger positive reverse causal effect now. Evidence for a positive reverse causal effect is however much less convincing in this specification, given that we only find it for two ML submethods.

Comparing the DML estimates to the original results in Table 7, we notice that DML in the PLR model generates estimates very close to the original estimate in the extended specification for the SVM, LASSO and neural network, but relatively smaller estimates for the tree-based ML submethods. This might suggest that nonlinear confounding exists but that it can only be picked up by DML with the tree-based ML submethods. Next, we find again that the DML estimates in the PLIV model are attenuated compared to the original IV estimate, except for the LASSO. The LASSO does not seem to give much difference compared to a simple linear control specification, possibly again indicating that the LASSO might not be the best method to pick up nonlinearities here. The original results in the extended specification suggest again a negative reverse causal effect. Conversely, for the SVM and neural network, the DML estimates in the PLIV model are smaller than the DML estimates in the PLR model and the original estimate without IVs. This indicates a positive reverse causal effect. For the tree-based ML submethods, the DML estimate in the PLIV model is either larger than in the PLR model or insignificant. When significant, it is broadly in line with the original estimate without IVs. Evidence for a reverse causal effect is mixed here: either it is negative or there is none. All in all, it appears that controlling for alternative effects of legal origins on trade flows not via contract enforcement is still more effective by using ML. However, evidence is weaker given that we only find positive reverse causal effects for 2 out of 7 ML submethods.

Terms with the largest mean LASSO coefficient in Figure 11 in Section A.1 of the Appendix include only a single nonlinear term here; an interaction. This probably explains why DML with the LASSO does not deviate much from original IV with linear controlling. Unfortunately, the LASSO coefficients do not very much improve our understanding of possibly important nonlinearities here.

Finally, we increase the number of folds from 2 to 5 in order to check sensitivity of the results to the number of folds. The DML estimates with 5 folds are presented in Table 22 in Section A.1 of the Appendix. We find estimates of $0.11 - 0.31$ in the PLR, being significant except for the regression tree. The DML with 5 folds estimates resemble the DML with 2 folds estimates closely, both qualitatively and quantitatively. In the PLIV model, we obtain estimates in the range of $0.10 - 0.47$, being significant for all ML submethods excluding the regression tree; including the random forest and Best method, however, in contrast to DML estimates with 2 folds for these ML submethods. The DML estimates with the two amounts of folds are again very similar and previous interpretations do not change. Moreover, standard errors for DML do not consistently grow or shrink due to using more folds. Hence, it becomes clear that the number of folds does not have a large influence on any estimation results in this application.

5.3 Approximate Residual Balancing: Fox News

5.3.1 Vote share regression

We also apply ARB to estimate the ATE of Fox News availability on the Republican vote share in 2000. We adopt the main specification with the demographic controls, cable controls and the lagged Republican vote share for 1996. Firstly, we consider a fully linear version of ARB where we do not take into account nonlinearities. We present the results in Table 9. The ARB estimates show little dispersion and concentrate around 0.3%, but none of them are significant. This implies again a positive Fox News effect, although we cannot exclude that the Fox News effect does not exist. Secondly, we consider a quadratic version by adding to the controls all first order interactions and squares of the census controls, cable controls and the lagged Republican vote share for 1996. The results are presented in Table 10. The ARB estimates range from 0.2 to 0.5% now but we still do not find any significant effects. Most estimates appear to increase by adding nonlinearities, but the standard errors grow as well. Even when taking into account some nonlinear confounding, we cannot rule out that that there is no Fox News effect. Again but now for ARB instead of DML, we find that there is a lot of uncertainty around the estimates when allowing for fully heterogeneous causal effects by using the interactive model. This supports the previous claim that preciseness of the Fox News effect estimates depends highly on the model assumptions.

Comparing both versions of the ARB estimates to the original estimates in Table 3, we notice that ARB produces, even though insignificant, positive Fox News effects that are in line with intuition, in contrast to the difference-in-difference estimators. This suggests that ARB can have an advantage over prevailing simple causal effect estimation methods. The fully linear ARB estimates are similar to the original unweighted least squares (OLS) estimate, whereas the quadratic ARB estimates are slightly larger, except for $\zeta = 0.7$. This increase is consistent with the findings from DML and is possibly the consequence of adequately taking into account nonlinear confounding. The ARB estimates are smaller than the original weighted least squares (WLS) estimates. However, they fall within the range of values from the unweighted original estimate with either of the fixed effects and the weighted original estimate with district fixed effects. ARB gives estimates similar to the more advanced original estimates, but it addresses confounding very differently. ARB might thus be another valuable alternative to established advanced causal effect estimation methods. Given the larger uncertainty around the estimates from having to use the interactive model, ARB seems to be a more conservative method than DML.

Table 9: Fully linear ARB estimates for the Fox News vote share ATE using the main specification

	LASSO			**Elastic Net**		
	$\zeta = 0.3$	$\zeta = 0.5$	$\zeta = 0.7$	$\zeta = 0.3$	$\zeta = 0.5$	$\zeta = 0.7$
Interactive Regression Model						
ATE	0.0029	0.0025	0.0026	0.0033	0.0026	0.0025
	(0.0020)	(0.0021)	(0.0023)	(0.0021)	(0.0021)	(0.0023)

Note: The symbols $*$, $**$ and $***$ denote significance at the 10, 5 and 1% level, respectively. ζ is a tuning parameter that trades off bias against variance, with lower values corresponding to less variance but more bias.

Table 10: Quadratic ARB estimates for the Fox News vote share ATE using the main specification

	LASSO			**Elastic Net**		
	$\zeta = 0.3$	$\zeta = 0.5$	$\zeta = 0.7$	$\zeta = 0.3$	$\zeta = 0.5$	$\zeta = 0.7$
Interactive Regression Model						
ATE	0.0034	0.0031	0.0025	0.0045*	0.0032	0.0022
	(0.0021)	(0.0023)	(0.0027)	(0.0023)	(0.0024)	(0.0029)

Note: The symbols $*$, $**$ and $***$ denote significance at the 10, 5 and 1% level, respectively. ζ is a tuning parameter that trades off bias against variance, with lower values corresponding to less variance but more bias.

Next, we turn to the dynamic specification by adding the lagged Republican vote share for 1992 and 1988 to the controls. We begin with a fully linear version of ARB, with the results given in Table 11. The ARB estimates range from approximately 0.5 to 0.7%, being significant in all cases. This strongly indicates a positive Fox News effect. Subsequently, we examine the quadratic version of ARB with all squares and first order interactions, including those involving the lagged vote shares for 1992 and 1988. We present the results in Table 12. The ARB estimates go from roughly 0.4 to 0.8% and half of them are significant. The majority of ARB estimates increase by adding nonlinearities, but the standard errors grow even more. Hence, we still obtain a positive Fox News effect when we account for some nonlinear confounding, but we cannot exclude anymore that there is no Fox News effect. ARB produces mostly significant positive estimates when using the dynamic specification, even though we adopt the interactive model. This is in line with DML in the interactive model for the dynamic specification. However, both versions of the ARB estimates lie on the low side of the DML estimates in the interactive model (compare to Table 4). The ARB estimates only resemble these DML estimates for the LASSO, SVM or Best ML submethods. The difference in size between the DML and ARB estimates possibly reflects two factors: the ability to handle nonlinear confounding and properties of the balancing weights. ARB might not be able to deal with nonlinear confounding as efficiently as DML with ML submethods other than those based on linear combinations, which could lead to smaller estimates. Simultaneously, using finite sample optimal ARB balancing weights instead of asymptotically optimal DML propensity score weights might also adjust estimates downwards.

Table 11: Fully linear ARB estimates for the Fox News vote share ATE using the dynamic specification

	LASSO			**Elastic Net**		
	$\zeta = 0.3$	$\zeta = 0.5$	$\zeta = 0.7$	$\zeta = 0.3$	$\zeta = 0.5$	$\zeta = 0.7$
Interactive Regression Model						
ATE	0.0061^{***}	0.0054^{**}	0.0052^{**}	0.0067^{***}	0.0060^{***}	0.0056^{**}
	(0.0020)	(0.0022)	(0.0024)	(0.0021)	(0.0023)	(0.0025)

Note: The symbols $*$, $**$ and $***$ denote significance at the 10, 5 and 1% level, respectively. ζ is a tuning parameter that trades off bias against variance, with lower values corresponding to less variance but more bias.

Table 12: Quadratic ARB estimates for the Fox News vote share ATE using the dynamic specification

	LASSO			**Elastic Net**		
	$\zeta = 0.3$	$\zeta = 0.5$	$\zeta = 0.7$	$\zeta = 0.3$	$\zeta = 0.5$	$\zeta = 0.7$
Interactive Regression Model						
ATE	0.0057^{**}	0.0056^{*}	0.0042	0.0076^{**}	0.0065^{**}	0.0062^{*}
	(0.0028)	(0.0032)	(0.0030)	(0.0031)	(0.0030)	(0.0033)

Note: The symbols $*$, $**$ and $***$ denote significance at the 10, 5 and 1% level, respectively. ζ is a tuning parameter that trades off bias against variance, with lower values corresponding to less variance but more bias.

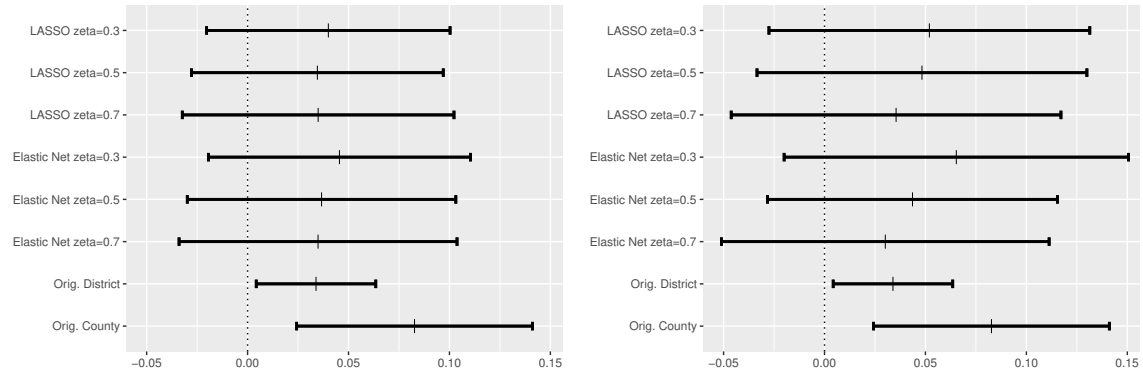
Comparing the ARB estimates to the original estimates for the dynamic specification in Table 5, it follows that there is a lot of similarity between the fully linear ARB estimates and the original unweighted least squares (OLS) estimate. The quadratic ARB estimates are also broadly consistent with this original estimate for the LASSO, but seem slightly larger for the elastic net. Quadratic ARB with the elastic net thus adjust the estimates from OLS upwards, which is consistent with the results from DML. It could be due to controlling for nonlinear confounding again. Further, most of the ARB estimates resemble the unweighted and weighted original county fixed effect estimates and some of them come close to the original weighted least squares or weighted district fixed effect estimates. In contrast to DML, ARB seems to give estimates in the neighborhood of the more advanced original estimates, even though ARB and these estimators capture confounding very differently. Hence, ARB might be another useful alternative to established advanced causal effect estimation methods, although it may not be as effective as DML when it comes to handling nonlinearity.

Finally, we point out that a higher value for the tuning parameter ζ almost every time leads to smaller ATE estimates and larger standard errors, regardless of the specification or whether we incorporate nonlinearities or not. This means that less bias and more variance leads to smaller estimates, often moving away from more advanced original or DML estimates towards the original OLS estimates. Although this does not provide direct guidance on choosing ζ (an important open issue with respect to ARB; Athey et al., 2016a), it helps to further understand the influence of this tuning parameter in practice.

5.3.2 Audience regression and persuasion rates

In the following, we compute Fox News persuasion rates for ARB by using (18). For that, we use the main specification ATE estimates $\hat{\beta}_{0,VoteShare}$ from Section 5.3.1. Next, we need to run the audience regression of Fox News watching on Fox News availability to get $\hat{\beta}_{0,Audience}$ and compute the persuasion rate. We present fully linear and quadratic audience regression ARB estimates in Table 23 and 24 in Section A.1 of the Appendix, respectively. The fully linear ARB estimates lie around $2.1 - 2.2\%$ and are all significant. The quadratic ARB estimates range from $1.9 - 2.2\%$ and are all significant as well. Thus, ARB strongly suggests that Fox News availability led to Fox News watching, for approximately 2% of the households. We can reject that the entry of Fox News did not induce Fox News watching. This stands in contrast with the audience regression DML estimates in the interactive model. Next, we notice that the LASSO and elastic net ARB estimates become exactly identical, which could indicate that ML does not perform optimally here due to the smaller sample size. Comparing to the original estimates in Table 19 in Section A.1 of the Appendix, it becomes clear that the fully linear and quadratic ARB estimates

resemble the original difference-in-difference and unweighted least squares (OLS) estimates. We conclude once again that in this regression, prevailing simple causal effect estimation methods cannot be rejected and that nonlinear controlling does not have a very large impact.



(a) Fully Linear (b) Quadratic

Figure 4: ARB estimates and 95% confidence intervals for persuasion rates. Standard errors are computed via the Delta method, following DellaVigna and Kaplan (2007). The original weighted district and county fixed effect persuasion rate estimates are denoted Orig. District and Orig. County, respectively.

Afterwards, we calculate the ARB persuasion rates and plot them in Figure 4. The fully linear ARB rate estimates show little difference across the LASSO, elastic net and their tuning parameter values and lie around the original district fixed effect rate estimate. However, the confidence intervals become a lot wider. The quadratic ARB rate estimates show more dispersion, while those using $\zeta = 0.7$ come very close to the original district fixed effect rate. The confidence intervals become even wider than for the fully linear version. The increase in confidence interval width for ARB compared to the original methods is partly the consequence of using the interactive model, which brings more uncertainty. It could be due to omitting the clustering of standard errors for towns with the same local cable company with ARB as well. All in all, the ARB persuasion rate estimates are in line with the original estimates, but the larger standard errors render them insignificant such that we cannot rule out that Fox News watching did not actually persuade viewers to vote Republican.

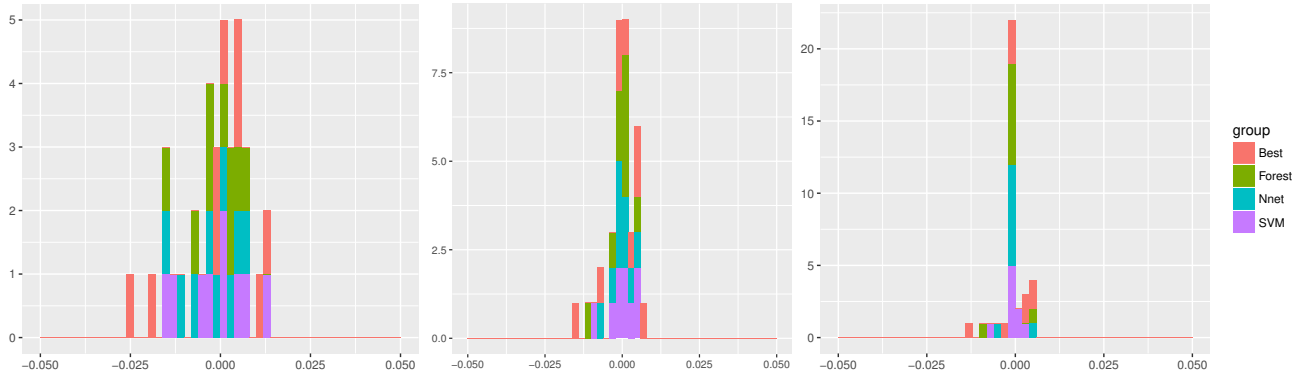
6 Results Heterogeneous Treatment Effects

6.1 Heterogeneous Double Machine Learning: Fox News

6.1.1 Vote share regression

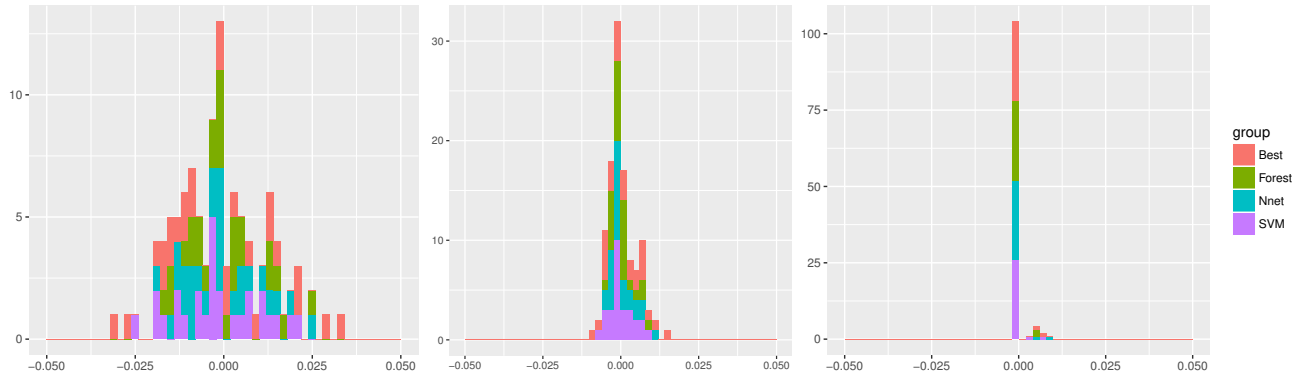
Now that we have established that ML-based causal inference methods can be used successfully to obtain the Fox News ATE, we extend the original analysis by estimating HTEs. We look at geographical heterogeneity, as described schematically in Figure 13 in Section A.2 of the Appendix. We use three levels of heterogeneity: level 1 corresponds to Census regions, level 2 splits these regions to obtain Census divisions and level 3 indicates the highest degree of heterogeneity by using individual states. We apply the main specification because this gives us the larger sample, which leads to better estimation accuracy. Due to limited computational resources, we choose to restrict the set of ML submethods for heterogeneous DML in this section to the random forest, SVM, neural net and Best. We also restrict ourselves to conventional standard errors due to the large impact of sample splitting on the median standard errors for HTEs, as a result of smaller sample sizes. In particular, the median standard errors for all HTEs grow very large, rendering all estimates insignificant. Before we discuss heterogeneous DML estimates, we have to make a choice between orthogonal least squares (LS), orthogonal debiased LASSO and orthogonal LASSO for each level of heterogeneity. Since the dimension of the treatment vector remains low ($d = 4$) for level 1 HTEs, we expect that estimation variance of OLS does not get too large. Hence, we do not have to apply regularization yet such that we can use orthogonal LS for level 1.

In order to choose between methods for level 2 and 3, we plot their estimates in Figure 5 and 6, respectively. From the left histograms, we obtain already much dispersion of the orthogonal LS estimates for level 2 HTEs ($d = 9$), but even more for level 3 HTEs ($d = 28$). This pattern reflects the explosion of OLS estimation variance when the treatment dimension d rises. The variance explosion results in increasingly more negative Fox News effect estimates, which seems to be intuitively implausible as it corresponds to a decrease in the Republican vote share due to Fox News availability. The orthogonal LASSO applies shrinkage to restrict estimation variance, at the cost of estimation bias. From the right figures, we observe that this leads to a smaller, more credible HTE range. For level 3 HTEs, we do not even get any negative estimates anymore. The introduced bias does however not allow for valid inference.



(a) Orthogonal LS (b) Orthogonal debiased LASSO (c) Orthogonal LASSO

Figure 5: Histograms of estimates from different versions of heterogeneous DML for level 2 Fox News vote share HTEs (see Figure 13). We use a level of shrinkage of $\lambda = 0.00004 \approx n^{-3/2}\sqrt{n/log(\#\text{Controls})}$ and set $a = 2^5$ based on a grid search. We represent groups of estimates for 4 ML submethods: random forest, neural net, SVM and Best.



(a) Orthogonal LS (b) Orthogonal debiased LASSO (c) Orthogonal LASSO

Figure 6: Histograms of estimates from different versions of heterogeneous DML for level 3 Fox News vote share HTEs (see Figure 13). We use a level of shrinkage of $\lambda = 0.00008$, such that it is roughly twice as large as for level 2 heterogeneity, following Chernozhukov et al. (2017). Also, we set $a = 2^5$ based on a grid search. We represent groups of estimates for 4 ML submethods: random forest, neural net, SVM and Best.

From the middle figures we see that the orthogonal debiased LASSO strikes a middle ground: it produces a slightly wider but still credible range of HTEs with the majority being positive. Meanwhile, it limits the amount of shrinkage and therefore estimation bias, such that it can still be used to conduct inference on the HTEs. The difference in pattern of the HTE estimates across the methods resembles that of Chernozhukov et al. (2017) for heterogeneous DML, but the differences across methods are a bit larger. The latter could relate to the fact that our sample size is only small compared to theirs. Moreover, it depends on our choice for the shrinkage and mixing parameters λ and a. It is not fully clear which particular value should be used in practice. Altogether, we conclude that the debiased LASSO is the preferred option to conduct inference on level 2 and 3 HTEs.

Heterogeneous DML orthogonal LS estimates for level 1 HTEs of Fox News availability on the Republican vote share in 2000 are presented in Table 13. For the Northeast region, we obtain a significant estimate for all ML submethods. The size of the Fox News effect ranges from $0.4 - 0.6\%$ here. This corresponds roughly with the size of the ATE from the more advanced original estimates and from DML for several ML submethods, using the main specification. For the South, Midwest and West regions, we get insignificant estimates with much variation across ML submethods. For the West, the estimates are positive and for the Midwest and South they are close to zero or even negative. Accordingly, we find that the significant positive Fox News ATE is driven in particular by the Northeast region and perhaps the West region, but not the South and Midwest regions. This indicates that geographical heterogeneity of the Fox News effect indeed exists.

Next, heterogeneous DML orthogonal debiased LASSO estimates for level 2 HTEs are given in Table 14. The Middle Atlantic division gets for all ML submethods a significant Fox News effect estimate of around $0.4 - 0.6\%$. Simultaneously, we notice from Figure 12 and 13 in Section A.2 of the Appendix that all member states of this division are Democratic states. Furthermore, we find for some ML submethods significant positive Fox News effect estimates for the divisions New England ($0.1 - 0.7\%$), Pacific ($0.2 - 0.6\%$) and East North Central ($0 - 0.4\%$). Again, it follows that these generally consist of Democratic towns, considering that 2/3, 5/6 and 2/3 of their member states had a Democratic victory in 2000, respectively. The divisions West North Central and South Atlantic get (partly) significant negative Fox News effect estimates, while East South Central, Mountain and West South Central show insignificant estimates close to zero. In contrast, they mainly consist of Republican towns. For all divisions except West North Central, the Republican party had a victory in all of the member states. For the West North Central division, it won in half of the member states. The remaining two states were swing states given the small margin of victory, especially for Iowa. Thus, the Fox News effect appears to be stronger in more Democratic divisions, which is in line with the estimate of DellaVigna and Kaplan (2007) (Table VI column (1)). This makes intuitive sense as well because the share of people that can possibly be convinced to vote Republican is larger in these divisions.

The advantage of more detailed heterogeneity in the Fox News effect can now also be seen from the estimates. The results for level 1 HTEs suggest an insignificant Fox News effect close to zero for the Midwest region, perhaps making us believe that the Fox News effect does not exist in each of the member divisions of the Midwest. The results for level 2 HTEs tell us however that there might be a small but significant positive Fox News effect for the East North Central division. This effect is however masked when including the West North Central division, since the latter shows a negative Fox News effect.

Finally, heterogeneous DML orthogonal debiased LASSO estimates for level 3 HTEs are given in Table 15, but only for states with a significant Fox News effect estimate for at least one ML submethod at the 10% level. Results for the remaining states with insignificant estimates are presented in Table 25 in Section A.1 of the Appendix. We find significant positive estimates for the Democratic states Michigan ($0.7 - 0.8\%$), New York ($0.8 - 1.1\%$) for all ML submethods, and for Hawaii ($0.1 - 0.7\%$), Massachusetts ($0.1 - 0.8\%$), Connecticut ($0.2 - 0.6\%$) for some ML submethods. Hence, it becomes again clear that geographical heterogeneity exists. More importantly, however, is that our results extend the original insight of a larger Fox News effect in Democratic states compared to Republican states, in the sense that we also obtain clear heterogeneity within the group of Democratic states now.

In addition to these Democratic states, we also obtain a significant positive Fox News effect estimate for the Republican states Alabama ($0.5 - 0.9\%$) and Wyoming ($0.5 - 1.5\%$). Furthermore, we find significant negative Fox News effect estimates for Wisconsin and Rhode Island and insignificant slightly negative estimates for Iowa and Minnesota, which are all four Democratic. We do not observe a larger Fox News effect in Democratic states compared to Republican states here. This illustrates that the previously discussed original insight is a bit too simplistic to describe geographical heterogeneity in the Fox News effect. On a more detailed level, geographical differences with respect to the Fox News effect go beyond the distinction between states with another winning party.

Table 13: Orthogonal LS estimates for level 1 Fox News vote share HTEs, $S = 50$

	Random Forest		**Neural Net.**		**SVM**		**Best**	
	Median HTE	se	Median HTE	se	Median HTE	se	Median HTE	se
HTE Model with Modeled Heterogeneity								
Northeast	0.0062^{***}	(0.0017)	0.0053^{***}	(0.0016)	0.0040^{**}	(0.0016)	0.0039^{**}	(0.0017)
South	-0.0069	(0.0064)	-0.0065	(0.0059)	-0.0092^{*}	(0.0055)	-0.0087	(0.0062)
West	0.0115^{*}	(0.0060)	0.0074	(0.0051)	0.0048	(0.0049)	0.0077	(0.0055)
Midwest	0.0000	(0.0022)	-0.0025	(0.0020)	-0.0017	(0.0020)	-0.0022	(0.0021)

Note: Heterogeneity levels are from the hierarchy outlined in Figure 13. Conventional standard errors are given in parentheses. The symbols $*$, $**$ and $***$ denote significance at the 10, 5 and 1% level, respectively. The regions are ordered according to significance strength, i.e. the total amount of asterisks over all ML submethods.

Table 14: Orthogonal debiased LASSO estimates for level 2 Fox News vote share HTEs, $S = 50$

	Random Forest		**Neural Net.**		**SVM**		**Best**	
	Median HTE	se	Median HTE	se	Median HTE	se	Median HTE	se
HTE Model with Modeled Heterogeneity								
Middle Atlantic	0.0048^{***}	(0.0010)	0.0057^{***}	(0.0009)	0.0047^{***}	(0.0009)	0.0044^{***}	(0.0010)
West North Central	-0.0155^{***}	(0.0022)	-0.0106^{***}	(0.0020)	-0.0063^{***}	(0.0020)	-0.0086^{***}	(0.0020)
New England	0.0072^{***}	(0.0012)	0.0013	(0.0013)	0.0015	(0.0011)	0.0034^{***}	(0.0012)
South Atlantic	-0.0077^{***}	(0.0020)	-0.0007	(0.0011)	-0.0036^{**}	(0.0017)	-0.0035^{*}	(0.0019)
Pacific	0.0057^{**}	(0.0024)	0.0017	(0.0017)	0.0027	(0.0019)	0.0053^{**}	(0.0022)
East North Central	0.0038^{***}	(0.0011)	0.0001	(0.0009)	0.0003	(0.0010)	0.0003	(0.0010)
East South Central	-0.0001	(0.0027)	-0.0010	(0.0022)	-0.0011	(0.0022)	-0.0015	(0.0023)
Mountain	0.0002	(0.0026)	-0.0025	(0.0023)	-0.0002	(0.0021)	-0.0013	(0.0022)
West South Central	-0.0002	(0.0038)	0.0009	(0.0027)	-0.0013	(0.0027)	0.0003	(0.0029)

Note: Heterogeneity levels are from the hierarchy outlined in Figure 13. Conventional standard errors are given in parentheses. The symbols $*$, $**$ and $***$ denote significance at the 10, 5 and 1% level, respectively. The regions are ordered according to significance strength, i.e. the total amount of asterisks over all ML submethods.

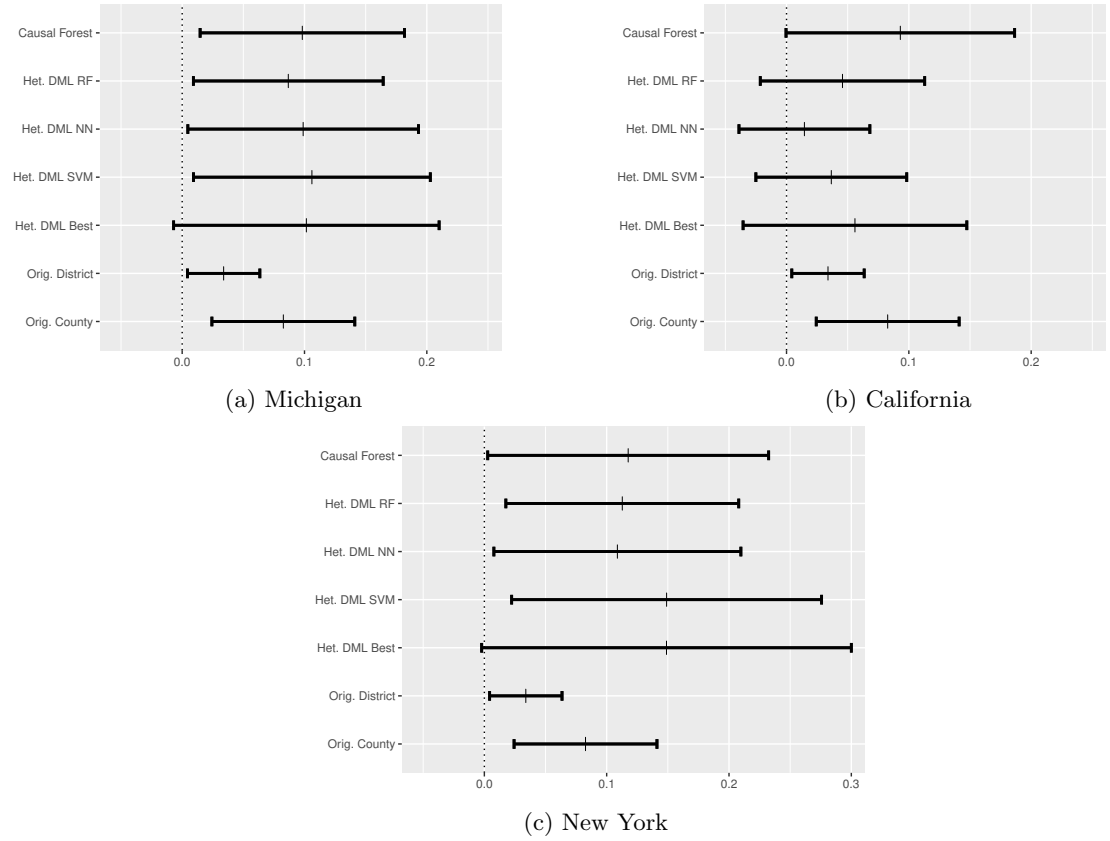
Table 15: Significant orthogonal debiased LASSO estimates for level 3 Fox News vote share HTEs, $S = 50$

	Random Forest		**Neural Net.**		**SVM**		**Best**	
	Median HTE	se	Median HTE	se	Median HTE	se	Median HTE	se
HTE Model with Modeled Heterogeneity								
Michigan	0.0079^{***}	(0.0014)	0.0074^{***}	(0.0012)	0.0078^{***}	(0.0015)	0.0066^{***}	(0.0013)
New York	0.0102^{***}	(0.0010)	0.0082^{***}	(0.0009)	0.0110^{***}	(0.0009)	0.0096^{***}	(0.0010)
Wisconsin	0.0020	(0.0045)	-0.0049^{**}	(0.0021)	-0.0053^{***}	(0.0019)	-0.0047^{**}	(0.0024)
Alabama	0.0092^{**}	(0.0037)	0.0071^{**}	(0.0035)	0.0051^{*}	(0.0028)	0.0055^{*}	(0.0031)
Wyoming	0.0148^{**}	(0.0064)	0.0046	(0.0031)	0.0073^{**}	(0.0035)	0.0061^{*}	(0.0037)
Hawaii	0.0074^{**}	(0.0030)	0.0007^{*}	(0.0004)	0.0045^{*}	(0.0024)	0.0047^{*}	(0.0024)
Massachusetts	0.0080^{***}	(0.0022)	0.0008	(0.0013)	0.0022^{*}	(0.0012)	0.0020	(0.0013)
Rhode Island	0.0021	(0.0019)	-0.0031^{**}	(0.0015)	-0.0027^{*}	(0.0016)	-0.0024	(0.0016)
Connecticut	0.0061^{**}	(0.0024)	0.0015	(0.0019)	0.0015	(0.0018)	0.0016	(0.0016)
South Carolina	-0.0073^{**}	(0.0034)	-0.0003	(0.0020)	-0.0037	(0.0028)	-0.0044	(0.0033)
Iowa	-0.0042^{*}	(0.0023)	-0.0003	(0.0021)	-0.0002	(0.0020)	-0.0012	(0.0020)
Minnesota	-0.0051^{*}	(0.0028)	-0.0022	(0.0027)	-0.0006	(0.0026)	-0.0019	(0.0028)

Note: Heterogeneity levels are from the hierarchy outlined in Figure 13. Conventional standard errors are given in parentheses. The symbols $*$, $**$ and $***$ denote significance at the 10, 5 and 1% level, respectively. We only show regions with a significant estimate, ordered according to significance strength, i.e. the total amount of asterisks over all ML submethods.

6.1.2 Persuasion rates

We continue with a discussion of heterogeneous persuasion rates, for the states Michigan, New York and California. We choose these states in order to compare our results to other media persuasion rates from the literature. To compute heterogeneous persuasion rates, we use the HTE estimates per state from Section 6.1.1 as $\hat{\beta}_{0,VoteShare}$ in (18). Further, note that we are unable to estimate HTEs in the audience regression, because of the small sample size. We proceed by taking for $\hat{\beta}_{0,Audience}$ the ATE estimated with DML, as given in Table 21 in Section A.1 of the Appendix. The heterogeneous persuasion rates are then presented in Figure 7. For Michigan, we observe persuasion rates of approximately 10% with little difference across ML submethods within heterogeneous DML. For 3 out of 4 ML submethods, the confidence interval does not cover zero, implying a significant persuasion rate. The significant persuasion rates for Michigan exceed the original ATE persuasion rates. It becomes thus clear that in Michigan a higher than average fraction of Fox News watchers is persuaded to vote Republican.



(a) Michigan

(b) California

(c) New York

Figure 7: Estimates and 95% confidence intervals for heterogeneous persuasion rates computed by combining heterogeneous vote share regressions and the audience regression. RF is short for random forest and NN for neural network. The Causal Forest rates are discussed in a later Section.

Continuing with California, we obtain large differences in persuasion rate across ML submethods. The rate is estimated at approximately 5% according to the random forest and Best ML submethod, whereas it is estimated at $1-4\%$ when following the SVM or neural network. We do not find any significant persuasive effects. To highlight the benefit of using HTEs, we compare this effect to the laboratory experiment of Ansolabehere and Iyengar (1995), who assess the persuasive effect of 30 second exposure to a political ad on the vote share for the party sponsoring the ad. In a Southern California testing location, they find a significant persuasion rate of 8.2%. When using the original ATE estimate with county fixed effects, we would conclude that the media persuasive effects of Fox News and political ads do not differ much. Conversely, when adopting the sparsity based heterogeneous DML, we find an insignificant but smaller effect. This implies that the media persuasive effect of political ads might exceed

that of Fox News.

For New York, we find two groups of persuasion rates. The random forest and neural network produce rates of $11 - 12\%$, whereas the SVM or Best ML submethod give rates of approximately 15% with much wider confidence intervals. For 3 out of 4 ML submethods, the persuasion rate is significant. Therefore, we have a wide range of estimates for the persuasion rate in New York of $11 - 15\%$. Hence, the fraction of Fox News watchers that is persuaded to vote Republican is even higher in New York than in Michigan. The relevance of HTEs can now also be seen when we compare the persuasive effect of Fox News to that of other media. Green and Gerber (2001) conducted a randomized field experiment to assess the persuasive effect of phone canvassing, i.e. getting a phone call encouraging you to vote, on voter turnout. In two locations in New York they find persuasion rates of 8.2% point (Albany) and 9.3% point (Stonybrook). Comparing that to the Fox News rate by using the original ATE persuasion rates, we would conclude that the media persuasive effect of Fox News is either lower or around that of phone canvassing. However, when correcting for geographical differences by using the HTE for New York, we arrive at the opposite conclusion: the media persuasive effect of Fox News is larger than that of phone canvassing.

6.2 Causal Forest: Fox News

6.2.1 Vote share regression

Causal Forest estimates for level 1 HTEs of Fox News availability on the Republican vote share in 2000 are presented in Table 16. The Fox News effect estimate is significant for the Northeast, Midwest and West regions. It is greatest for the West region (0.93%), followed by the Northeast region (0.77%) and then the Midwest region (0.67%). The size of the Fox News effect estimate in the Midwest region corresponds to the weighted original ATE estimate with county fixed effects. For the South region, we obtain an insignificant positive estimate. Hence, we find that the significant positive Fox News ATE is mainly driven by the West, Northeast and Midwest region, but perhaps not the South. There is thus evidence in the data for heterogeneity in the Fox News effect across regions.

Both the Causal Forest and heterogeneous DML indicate that the Fox News effect might be lower or even absent in the South region. Furthermore, in the West region the Causal Forest estimate lies within the range of heterogeneous DML estimates across ML submethods ($0.5 - 1.1\%$). There is however no consensus on the Midwest region, since heterogeneous DML estimates for the Midwest are close to zero and insignificant instead of significant and positive, as follows from the Causal Forest. Moreover, the estimates for the Northeast region differ numerically between these methods. Underlying assumptions provide insights into the origin of the discrepancies between the methods. Because it assumes sparsity of the Fox News effect across regions, heterogeneous DML pulls estimates towards zero in some regions (South, Midwest). The differences across regions are then highlighted well, but having null effects in some regions might not be realistic. To the contrary, the Causal Forest gives the same estimate of approximately the ATE for these regions. This leads to less clear differences across regions, but we do not get null effects anymore. Depending on the believed structure of Fox News effect heterogeneity, we can prefer one or another.

Increasing the degree of heterogeneity, we present Causal Forest estimates for level 2 HTEs in Table 17. We find significant estimates for all divisions. The Fox News effect is estimated to be large for the Republican Mountain division (0.82%) and the Democratic Middle Atlantic division (0.78%), but small for the mixed West North Central division (0.60%) and the Republican East South Central division (0.68%). Other divisions get a Fox News effect estimate close to 0.71%. Comparing to heterogeneous DML, it follows that both the Causal Forest and heterogeneous DML with any of the four ML submethods assign the West North Central division the single lowest Fox News effect estimate. Furthermore, the Middle Atlantic division gets the highest estimate with the Causal Forest and heterogeneous DML for the neural network and SVM. Thus, we find some similarity between the two methods. However, we do not observe the sharp distinction between Democratic and Republican divisions anymore with the Causal Forest.

Table 16: Causal Forest estimates for level 1 Fox News vote share HTEs

	Causal Forest	
	Median HTE	se
Conditional ATE Model		
Midwest	0.0067**	(0.0034)
Northeast	0.0077**	(0.0038)
West	0.0093**	(0.0043)
South	0.0069*	(0.0041)

Note: Heterogeneity levels are from the hierarchy outlined in Figure 13. Conventional standard errors are given in parentheses. The symbols $*$, $**$ and $***$ denote significance at the 10, 5 and 1% level, respectively. The regions are ordered according to significance strength, i.e. the amount of asterisks.

Table 17: Causal Forest estimates for level 2 Fox News vote share HTEs

	Causal Forest	
	Median HTE	se
Conditional ATE Model		
East South Central	0.0068***	(0.0025)
Mountain	0.0082***	(0.0031)
Middle Atlantic	0.0078***	(0.0029)
East North Central	0.0070**	(0.0031)
New England	0.0071**	(0.0029)
Pacific	0.0071**	(0.0030)
South Atlantic	0.0071**	(0.0029)
West North Central	0.0060**	(0.0030)
West South Central	0.0072**	(0.0029)

Note: Heterogeneity levels are from the hierarchy outlined in Figure 13. Conventional standard errors are given in parentheses. The symbols $*$, $**$ and $***$ denote significance at the 10, 5 and 1% level, respectively. The regions are ordered according to significance strength, i.e. the amount of asterisks.

Table 18: Causal Forest estimates for level 3 Fox News vote share HTEs

	Causal Forest	
	Median HTE	se
Conditional ATE Model		
Arkansas	0.0073***	(0.0026)
Iowa	0.0070***	(0.0024)
Michigan	0.0076***	(0.0018)
New York	0.0091***	(0.0031)
New Jersey	0.0074***	(0.0025)
Ohio	0.0071***	0.0024
Wisconsin	0.0070***	(0.0019)
Wyoming	0.0073***	(0.0028)
Pennsylvania	0.0071**	(0.0029)
Tennessee	0.0067**	(0.0026)
Remaining states	0.0072***	–

Note: Heterogeneity levels are from the hierarchy outlined in Figure 13. Conventional standard errors are given in parentheses. The symbols $*$, $**$ and $***$ denote significance at the 10, 5 and 1% level, respectively. The regions are ordered according to significance strength, i.e. the amount of asterisks. Remaining states denotes the estimate for other states that are not explicitly indicated. These states and their standard errors are given in Table 26 in Section A.1 of the Appendix.

We move on with the largest degree of heterogeneity. Causal Forest estimates for level 3 HTEs are given in Table 18. A lot of states get an equal Fox News effect estimate of 0.72%. They are listed with the corresponding complete results in Table 26 in Section A.1 of the Appendix. We observe again significant estimates for all states. A distinctly large Fox News effect of 0.91% is obtained for New York. Other states which show a relatively large Fox News effect are Michigan (0.76%), New Jersey (0.74%), Wyoming (0.73%) and Arkansas (0.73%). We find a relatively small Fox News effect for Tennessee (0.67%), Wisconsin (0.70%) and Iowa (0.70%). Except for New York, the Fox News effect does not differ

widely across states. This could be an indication that the data do not fully support heterogeneity to such a high degree. Finally, we compare the results to estimates from heterogeneous DML. Wyoming belongs to the states with a significant positive effect for heterogeneous DML and similarly it gets a relatively large estimate with the Causal Forest. Wisconsin is one of the few states with a significant negative heterogeneous DML estimate and in line with this it gets a relatively small Causal Forest estimate. Further, heterogeneous DML with the neural network, SVM and Best ML submethod obtains the single highest Fox News effect for New York, similar to the Causal Forest. Moreover, for New York the Causal Forest estimate falls within the range of heterogeneous DML estimates across ML submethods (0.8−1.1%). Next, Michigan gets the second highest Fox News effect with both the Causal Forest and heterogeneous DML with the neural network, SVM and Best ML submethod. The Causal Forest estimate additionally resembles the heterogeneous DML estimates for the random forest (0.79%), neural network (0.74%) and SVM (0.78%) closely here. The findings for New York and Michigan seem particularly strong, given that the methods agree on the size of the Fox News effect there, both relative to other states and absolute.

6.2.2 Persuasion rates

Next, we discuss heterogeneous persuasion rates for selected states. Note again that we are unable to estimate HTEs in the audience regression, because of the small sample size. In order to obtain the persuasion rates by using solely the Causal Forest, not DML, we proceed differently than before. We estimate a new Causal Forest for the audience regression after which we average over the HTEs for all observations to obtain an ATE estimate of 0.0231 with a standard error of 0.0084, following the procedure to obtain the ATE via the Causal Forest from Athey et al. (2016b). The recall audience heterogeneous persuasion rate is then presented in Figure 7 in Section 6.1.2. For Michigan, we observe a significant persuasion rate of approximately 10%, conforming to the heterogeneous DML persuasion rates. Hence, the previous conclusion remains valid; evidence for it becomes even stronger.

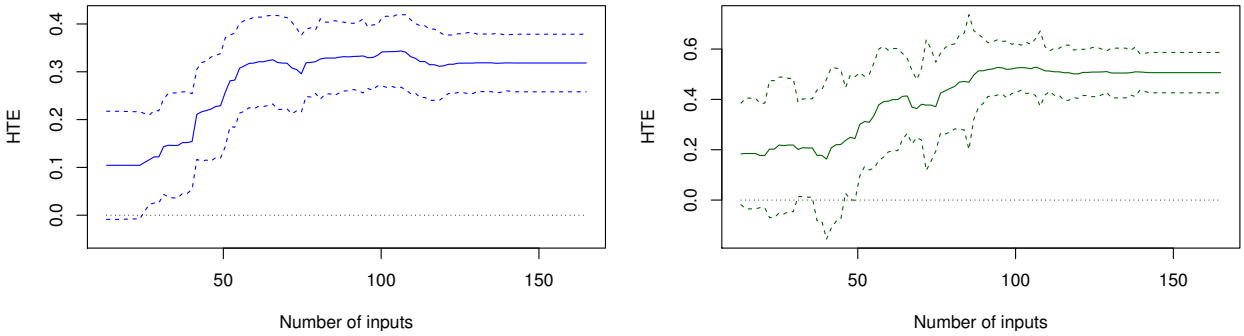
Proceeding to California, we find a significant persuasion rate estimate around 8%. Comparing this to the 8.2% persuasion rate of Ansolabehere and Iyengar (1995), we conclude that the media persuasive effects of Fox News and political ads do not differ much, as opposed to the conclusion from heterogeneous DML. For New York, we find a significant persuasion rate of $11 - 12\%$, similar to heterogeneous DML with the random forest or neural network. Earlier conclusions continue to hold in this case.

6.3 Causal Forest: Contracts and Trade

After having estimated the ATE of contract enforcement quality on trade flows, we extend the original analysis with HTEs. We inspect heterogeneity across the number of inputs for an industry, because the larger the number of inputs, the harder vertical integration becomes. In turn, vertical integration could make good contract enforcement less important for trade flows, since it removes the need to collaborate with suppliers. Thus, HTEs across the number of inputs inform us about the impact of vertical integration on the relationship between contract enforcement and trade flows.

To make a fair comparison with the original result, we firstly compute Causal Forest contracts and trade HTE estimates by using only the country and industry level controls. The results are shown in Figure 8, where HTE estimates are significant if the confidence intervals do not cover zero. The standard Causal Forest of Figure 8a produces an HTE estimate that starts at a low level close to 0.10, for the number of inputs smaller than 15. The HTE estimate increases with the number of inputs next, roughly until the number of inputs reaches 60. Finally, it stabilizes at a level just above 0.30. For any number of inputs larger than approximately 15, the HTE estimate is significant. We interpret this pattern as follows: only for a moderate number of inputs roughly between 15 and 60, we find evidence that easier vertical integration decreases the effect of contract enforcement quality on trade flows.

Originally, when controlling for the levels of trade, Nunn (2007) estimates the contracts and trade effect at 0.186 for industries with the number of inputs below the median across industries and at 0.342 for industries with the number of inputs above it (Table V column (1)). Averaging our Causal Forest results, we get estimates for the below and above median number of input industries of 0.217 and 0.326. Hence, the Causal Forest results are broadly consistent with the original findings, but extend them in the sense that we now obtain a complete functional form for the impact of vertical integration on the contracts and trade effect.



(a) No Instrumental Variables

(b) Instrumental Variables

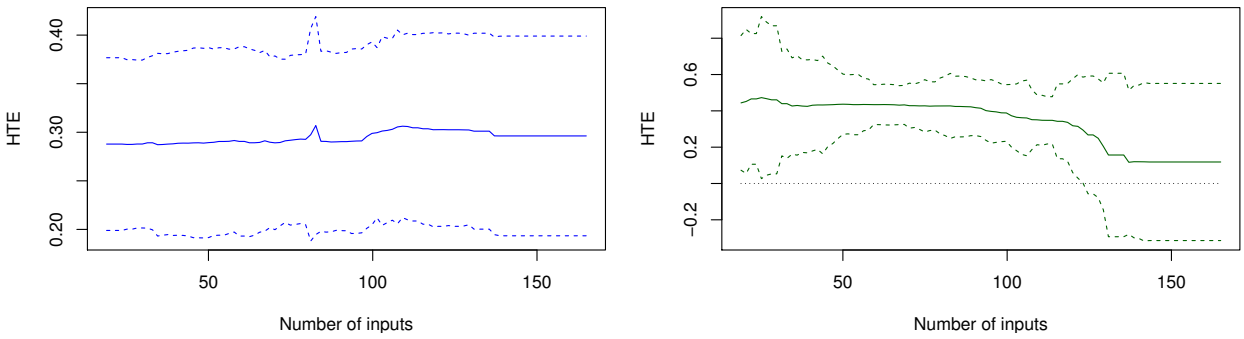
Figure 8: Causal Forest estimates and 95% confidence intervals for HTEs of contract enforcement quality on trade flows across different number of inputs using only the country and industry level control variables.

Another direction in which we extend the original findings is taking into account reverse causality. The previous Causal Forest and original results do not incorporate that causality possibly runs from comparative advantage in relationship important industries to good contract enforcement. Therefore, we apply the instrumental variable version of the Causal Forest, with the results outlined in Figure 8b. The HTE estimate begins with a low level around 0.20, for the number of inputs smaller than 35. Next, we observe that the HTE estimate increases with the number of inputs, roughly until the number of inputs reaches 90. In the end, it stabilizes at a level close to 0.50. For an input amount larger than 50, we find consistent significance of the contracts and trade effect. Our interpretation is that easier vertical integration reduces the effect of contract enforcement quality on trade flows, but only if the number of inputs lies between 50 and 90. The HTE level for the Causal Forest with instrumental variables is generally higher than for the Causal Forest without instrumental variables, for any given number of inputs. Similarly as for the ATE, this suggests that the reverse causal effect found by incorporating IVs into the Causal Forest is negative. It seems that the Causal Forest cannot sufficiently control for alternative channels through which the legal origin instruments affect trade flows. Furthermore, compared to the Causal Forest without instrumental variables, we also find a slightly smaller range for which the number of inputs and accordingly vertical integration affects the contracts and trade effect. Hence, the impact of vertical integration seems to decrease after taking into account reverse causality.

We continue with extending the original findings to the main specification, thereby adding the labor and capital interactions. As such, we control for other determinants of trade flows in order to make our causal effect identification assumptions more plausible. Figure 9 shows Causal Forest contracts and trade HTE estimates for the main specification. The standard Causal Forest of Figure 9a gives an HTE estimate just below 0.30 up to approximately 100 inputs. It then increases to a level slightly above 0.30, until reaching 110 inputs. For more than 110 inputs, the HTE estimate seems to remain around 0.30. It follows that the degree to which easier vertical integration implies a smaller contracts and trade effect drops substantially by including extra controls. What remains is a very narrow region of $100-110$ inputs where vertical integration appears to have a tiny effect. Hence, the original results and the Causal Forest results without the extra controls possibly overstate the impact of vertical integration.

Finally, we turn to instrumental variables to remove potential reverse causal effects. Figure 9b shows the HTE estimates of the instrumental variable version of the Causal Forest. The HTE estimate begins slightly above 0.40 and decreases gradually towards 0.40 for approximately 100 inputs. For more than 100 inputs, we observe a steeper descent until the HTE estimate stabilizes around 140 inputs at a much lower level of roughly 0.15. Additionally, we see that the HTE confidence intervals become very wide for a very small or large number of inputs. The contracts and trade HTE is nonetheless significant approximately until the number of inputs reaches 125. Surprisingly, the results now imply that easier vertical integration leads to a slightly higher effect of contract enforcement quality on trade flows, except for the largest amounts of inputs. Compared to the Causal Forest without IVs, we generally find a higher HTE estimate for any given input amount, indicating again that Causal Forest controlling might not be sufficient in order for the IV assumptions to hold. In addition, any evidence for a lower contracts and trade effect due to vertical integration completely disappears by taking into account reverse causality.

Everything combined, it becomes clear that the negative impact of easier vertical integration on the effect of contract enforcement quality on trade flows might not be as large as found by Nunn (2007) or not even existent. Furthermore, assuming that it still exists, it is only present for industries with certain moderate levels of vertical integration.



(a) No Instrumental Variables

(b) Instrumental Variables

Figure 9: Causal Forest estimates and 95% confidence intervals for HTEs of contract enforcement quality on trade flows across different number of inputs using the main specification.

7 Conclusion

In this paper, we investigate the merits of promising ML-based causal inference methods in empirical applications by revisiting two applied papers. The first estimates the causal effect of the introduction of the Fox News channel on the Republican vote share in the 2000 U.S. presidential elections by assuming unconfoundedness. The second estimates the causal effect of contract enforcement quality on trade flows by using IVs. We target the ATE, like originally, but we also extend the original papers by targeting HTEs. For the ATE, we implement DML and ARB and compare the results both to the original results, that were obtained using established causal inference methods, as well as to each other. For the HTEs, we employ heterogeneous DML and the Causal Forest and examine what additional insights we can get and investigate to what extent the insights from both methods agree.

Starting with ATE estimation under unconfoundedness using DML, we find in the Fox News application for the effect of Fox News availability on the 2000 Republican vote share that DML produces positive estimates, which conforms to our intuition. Conversely, the original difference-in-difference estimator gives negative estimates. This suggests that DML might give a better indication of the ATE than simple causal inference methods. Next, it becomes clear that the DML estimates are consistently larger than the original OLS estimate in both considered specifications. This increase could be the consequence of more adequate correcting for high-dimensional and complex confounding relationships with DML relative to original OLS. This explanation is supported by additional analyses using the LASSO, which suggest that nonlinearities play a large role in the control functions. This result provides evidence that DML might improve on causal inference from standard linear controlling. Finally, in the main specification, the DML estimates show a lot of similarity with more advanced original estimates, i.e. those that use controls and apply weighting and/or fixed effects. DML strengthens the original results here given that DML and the advanced original methods capture confounding very differently. In the dynamic specification, the DML estimates are greater than most of the advanced original estimates, possibly because we are able to correct for nonlinear political trends with DML. DML might improve the original estimates here by adjusting them upwards. Thus, DML can be a very useful alternative to established advanced causal effect estimation methods.

Furthermore, it follows that DML estimates for the fraction of Fox News channel viewers that is persuaded to vote Republican remain close to the main original estimates. Confidence intervals for this persuasion rate are however a lot wider with DML, implying that we cannot exclude that the Fox News channel did in fact not convince any viewers to vote Republican. Economic implications of the changes in estimate due to DML are limited here.

Moving on to ATE estimation under IV-identification using DML, we find in the contracts and trade application with the main specification that DML estimates with IVs are smaller than the original estimate with IVs. Moreover, they are also smaller than the DML estimates without IVs and in majority smaller than the original estimate without IVs. These findings correspond to a positive reverse causal effect, which is in line with intuition and other original estimates from propensity score matching. In contrast, original IV-estimation with a linear control specification leads to a negative reverse causal effect. The difference might indicate that DML is better capable of correcting for alternative channels through which the instruments affect the outcome, making the exclusion assumption more credible. In particular, we might be able to discover complex industry specialization patterns for country variables with DML that could not be found originally. Accordingly, DML might improve on standard causal inference with IVs to a great extent.

The change in ATE estimate due to DML has economic implications. Originally, it was found that the effect of contract enforcement on trade flows exceeds the combined effect of labor and capital variables by large. However, the majority of DML estimates suggest that the contract enforcement effect actually could as well fall in the same range as this combined effect. The size of the causal effect seems to be overstated originally due to improperly taking into account reverse causality.

Continuing with ATE estimation under unconfoundedness using ARB, we find in the Fox News application for the previously considered Fox News effect solely positive estimates with ARB, rather than the negative estimates from difference-in-difference. Hence, ARB might give a better indication of the ATE than simple causal inference methods, similar to DML. Further, it follows that the fully linear ARB estimates become very similar to the original OLS estimate in both specifications. Some of the quadratic ARB estimates are slightly larger. We conclude therefore that adequate correcting for solely high-dimensional confounding does not lead to differences. However, adequately correcting for complex (and high-dimensional) confounding relationships does lead to a slight increase of the ARB estimate compared to the original OLS estimate. There is some evidence that ARB might improve on standard linear controlling, albeit weaker evidence than for DML. Lastly, in both specifications, the ARB estimates are moderately similar to the more advanced original estimates. Both types of methods control for confounding in a very different ways, such that ARB strengthens the original results here. ARB can thus also be a useful alternative to established advanced causal effect estimation methods. Persuasion rate estimates from ARB resemble those of DML closely and lead to similar conclusions.

Comparing DML and ARB more closely by adopting the same model, we find in the main specification that both DML and ARB give insignificant Fox News effect estimates. In the dynamic specification, we obtain significant estimates for both DML and ARB. In that sense, DML and ARB seem to agree here. However, the DML estimates are greater than the ARB estimates in the dynamic specification. This difference possibly reflects the difference in weights that are used to balance covariate distributions across the treatment and control group: ARB weights might be more suitable than DML weights in smaller samples, especially if estimation of the propensity score proves to be difficult.

Next, for HTE estimation under unconfoundedness using heterogeneous DML, we obtain in the Fox News application only a significant positive Fox News effect estimates for the Northeast region, when using the lowest degree of heterogeneity. This gives a first indication that geographical heterogeneity indeed exists. When we increase the degree of heterogeneity, it follows that divisions with the largest estimates all turn out be Democratic, whereas the divisions with the smallest estimate happen to be Republican. Thus, we agree with the original insight that the Fox News effect is broadly driven by towns where the Democratic party won the elections in 2000. For the highest degree of heterogeneity, we find a high Fox News effect in the Democratic states New York ($0.8 - 1.1\%$) and Michigan ($0.7 - 0.8\%$), but also in the Republican states Wyoming ($0.5 - 1.5\%$) and Alabama. The lowest Fox News effect is found for the Democratic state Wisconsin, where it even becomes negative. We firstly conclude that on a more detailed level, geographical heterogeneity of the Fox News effects goes beyond the distinction of states with another winning party in 2000. Secondly, it follows that within the group of Democratic states heterogeneity of the Fox News effect exists as well.

Proceeding to HTE estimation under unconfoundedness using the Causal Forest, we find in the Fox News application for the lowest degree of heterogeneity significant positive Fox News effect estimates for the Northeast, Midwest and West region, but not the South. Hence, we have indication for geographical heterogeneity again. When increasing the degree of heterogeneity, we do not find a sharp distinction in Fox News effect between Democratic and Republican divisions anymore. For the highest degree of

heterogeneity, we find a clear largest Fox News effect in New York (0.91%). Other states with relatively large effects are Michigan (0.76%), New Jersey, Wyoming (0.73%) and Arkansas. For Tennessee, we find the smallest effect, although it remains positive. Other states with a small effect are Wisconsin and Iowa.

Finally, for HTE estimation under IV-identification with the Causal Forest, we find in the contracts and trade application an HTE estimate that increases with the number of inputs of an industry, but only for not too extreme input amounts. This implies that easier vertical integration reduces the causal effect of contract enforcement quality on trade flows, which is in line with the original claim. However, it also follows that this only holds if industries happen to have certain moderate levels of vertical integration. Furthermore, it becomes apparent that taking into account reverse causality by using IVs or controlling for labor and capital variables shrinks the range of levels for which vertical integration does have an influence. When including both, we do not even find an increasing HTE estimate in the number of inputs anymore. Thus, our results refine the original insight regarding vertical integration: it might not have too much of an influence on the relationship between contract enforcement and trade flows.

Comparing heterogeneous DML and Causal Forests, it follows that they actually show some similarity, with respect to the ordering of geographical regions on Fox News HTE as well as the Fox News HTE itself. For the smallest degree of heterogeneity, we find with both methods that the Fox News effect might be lower or absent in the South region. For the largest degree of heterogeneity, we notice for both methods that Wyoming and Wisconsin belong to the states with a relatively high and low Fox News effect, respectively. Furthermore, according to both methods New York has the largest Fox News effect and Michigan the second largest. Moreover, the Causal Forest estimates lie within the range of estimates from heterogeneous DML for these two states. Differences between the heterogeneous DML and Causal Forest estimates seem to arise primarily due to the sparsity assumption on the HTEs for heterogeneous DML. This assumption pulls some heterogeneous DML HTE estimates towards zero, or even makes them negative. Finally, from a literature comparison, it becomes clear that HTE estimates from both methods improve on ATE estimates in the sense that they accommodate valid comparisons of the media persuasion effect of Fox News watching to other media forms.

References

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of econometrics*, 113(2):231–263.

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.

Acemoglu, D., Johnson, S., and Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review*, 91(5):1369–1401.

Angrist, J. D. and Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014.

Ansolabehere, S. and Iyengar, S. (1995). Going negative: How attack ads shrink and polarize the electorate (vol. 8).

Antweiler, W. and Trefler, D. (2002). Increasing returns and all that: a view from trade. *American Economic Review*, 92(1):93–119.

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., Imbens, G., Pham, T., and Wager, S. (2017). Estimating average treatment effects: Supplementary analyses and remaining challenges. *American Economic Review*, 107(5):278–81.

Athey, S., Imbens, G. W., and Wager, S. (2016a). Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. *arXiv preprint arXiv:1604.07125*.

Athey, S., Tibshirani, J., and Wager, S. (2016b). Generalized random forests. *arXiv preprint arXiv:1610.01271*.

Bartlesman, E. and Gray, W. B. (1996). The nber manufacturing productivity database.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

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. K. (2016). Double machine learning for treatment and causal parameters. Technical report, CEMMAP working paper, Centre for Microdata Methods and Practice.

Chernozhukov, V., Goldman, M., Semenova, V., and Taddy, M. (2017). Orthogonal machine learning for demand estimation: High dimensional causal inference in dynamic panels. *arXiv preprint arXiv:1712.09988*.

Chernozhukov, V., Hansen, C., and Spindler, M. (2015). Valid post-selection and post-regularization inference: An elementary, general approach.

Darolles, S., Fan, Y., Florens, J.-P., and Renault, E. (2011). Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565.

DellaVigna, S. and Kaplan, E. (2007). The fox news effect: Media bias and voting. *The Quarterly Journal of Economics*, 122(3):1187–1234.

Dudík, M., Langford, J., and Li, L. (2011). Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*.

Feenstra, R. C. (1996). Us imports, 1972-1994: Data and concordances. Technical report, National Bureau of Economic Research.

Foster, J. C., Taylor, J. M., and Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867–2880.

Gautier, E. and Tsybakov, A. (2011). High-dimensional instrumental variables regression and confidence sets. *arXiv preprint arXiv:1105.2454*.

Green, D. P. and Gerber, A. S. (2001). Getting out the youth vote: Results from randomized field experiments. *Unpublished report to the Pew Charitable Trusts and Yale University's Institute for Social and Policy Studies*.

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

Greene, W. (2002). *Econometric Analysis*. Pearson, 5 edition.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2017). Deep iv: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pages 1414–1423.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Hirshberg, D. A. and Wager, S. (2017). Balancing out regression error: efficient treatment effect estimation without smooth propensities. *arXiv preprint arXiv:1712.00038*.

Hotz, V. J., Imbens, G. W., and Klerman, J. A. (2006). Evaluating the differential effects of alternative welfare-to-work training components: A reanalysis of the california gain program. *Journal of Labor Economics*, 24(3):521–566.

Imai, K., Ratkovic, M., et al. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470.

Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86.

Kaufmann, D., Kraay, A., and Mastruzzi, M. (2004). Governance matters iii: Governance indicators for 1996, 1998, 2000, and 2002. *The World Bank Economic Review*, 18(2):253–287.

La Porta, R., Lopez-de Silanes, F., Shleifer, A., and Vishny, R. (1999). The quality of government. *The Journal of Law, Economics, and Organization*, 15(1):222–279.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620.

Lee, M.-j. (2009). Non-parametric tests for distributional treatment effect for randomly censored responses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):243–264.

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51.

Nunn, N. (2007). Relationship-specificity, incomplete contracts, and the pattern of trade. *The Quarterly Journal of Economics*, 122(2):569–600.

Rauch, J. E. (1999). Networks versus markets in international trade. *Journal of international Economics*, 48(1):7–35.

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

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

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

Rosenblum, M. and van der Laan, M. J. (2011). Optimizing randomized trial designs to distinguish which subpopulations benefit from treatment. *Biometrika*, 98(4):845–860.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222.

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., and Li, B. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10(Feb):141–158.

Tian, L., Alizadeh, A. A., Gentles, A. J., and Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532.

Van Der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1).

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

Warren, A. (2001). Television and cable factbook 2001.

Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17(2):492–514.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922.

A Appendix

A.1 Additional results

Original results

Table 19: Original audience regression ATE estimates with corresponding table numbers and columns from DellaVigna and Kaplan (2007)

	Audience			
	Unweighted		Weighted	
No controls				
Diff-in-diff	0.0219***	Earlier draft	0.0270***	Table VIII column (1)
Main specification				
No f.e. (LS)	0.0228**	Earlier draft	0.0272***	Earlier draft
District f.e.	0.0471***	Earlier draft	0.0371***	Table VIII column (2)
County f.e.	0.0295**	Earlier draft	0.0251***	Table VIII column (3)

Note: Unweighted and weighted estimation corresponds to using the observations directly and after weighting them by the amount of votes cast in 1996, respectively. Diff-in-diff denotes the simple difference-in-difference estimator computed with only the 2000 and 1990 audience shares. No f.e. (LS), district f.e. and county f.e. add controls from the main specification, see Table 1. No f.e. (LS) denotes the least squares estimator without fixed effects, while district f.e. and county f.e. denote different fixed effect estimators. The symbols $*$, $**$ and $***$ denote significance at the 10, 5 and 1% level, respectively. Earlier results indicates an estimate that belongs to earlier original methodology, but where originally a different estimation sample was used. We computed new values by using the final instead of the earlier estimation sample to exclude sample effects from comparisons.

DML

Table 20: 5 Fold DML estimates for the Fox News vote share ATE using the dynamic specification, $S = 50$

	SVM	**LASSO**	**Reg. Tree**	**Boost-ing**	**Random Forest**	**Neural Net.**	**Best**
A. Partially Linear Regression Model							
Fraction best ML method	0.24	0.00	0.00	0.00	0.76	0.00	/
Median ATE (5 fold)	0.0065***	0.0085***	0.0114***	0.0115***	0.0078***	0.0077***	0.0076***
	[0.0025]	[0.0020]	[0.0040]	[0.0042]	[0.0021]	[0.0022]	[0.0021]
	(0.0019)	(0.0018)	(0.0024)	(0.0021)	(0.0019)	(0.0019)	(0.0019)
B. Interactive Regression Model							
Fraction best ML method	0.57	0.00	0.00	0.00	0.43	0.00	/
Median ATE (5 fold)	0.0076**	0.0058**	0.0077**	0.0100***	0.0109***	0.0102*	0.0072*
	[0.0034]	[0.0025]	[0.0034]	[0.0031]	[0.0033]	[0.0062]	[0.0038]
	(0.0028)	(0.0017)	(0.0027)	(0.0019)	(0.0022)	(0.0051)	(0.0032)

Note: Fraction best ML method denotes the fraction of nuisance function estimation problems in which the ML submethod outperforms all the others in terms of out-of-sample MSE, over the folds and S replications. We present conventional standard errors in parentheses and median standard errors that incorporate the variation of sample splitting in square brackets. The symbols $*$, $**$ and $***$ denote significance at the 10, 5 and 1% level, respectively, assessed with median standard errors.

Table 21: DML estimates for the Fox News audience ATE using census and cable controls, $S = 50$

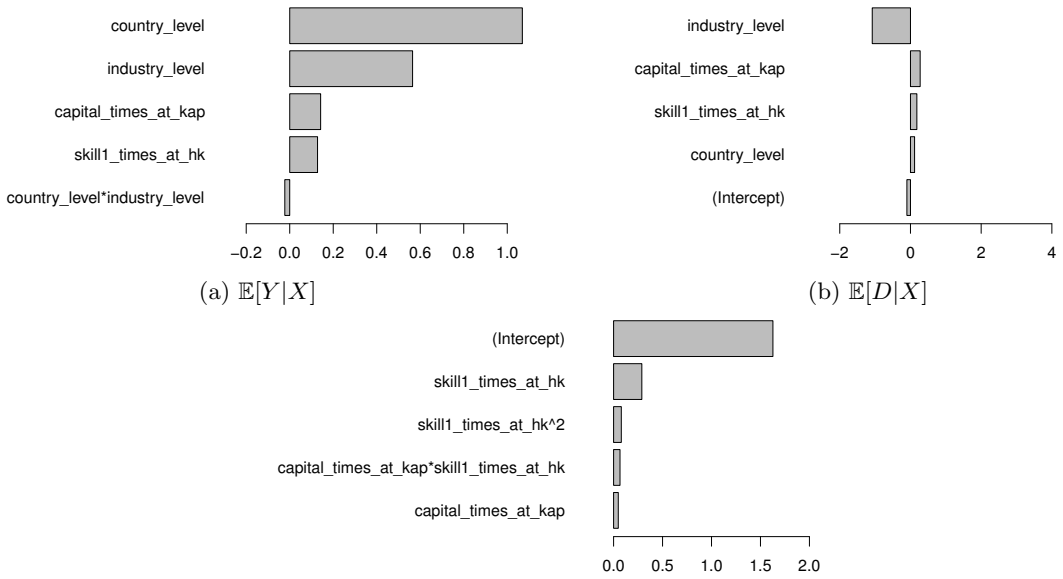
	SVM	**LASSO**	**Reg. Tree**	**Boost-ing**	**Random Forest**	**Neural Net.**	**Best**
A. Partially Linear Regression Model							
Fraction best ML method	0.28	0.13	0.00	0.10	0.00	0.50	–
Median ATE (2 fold)	0.0220**	0.0247**	0.0212**	0.0224**	0.0270**	0.0224*	0.0193*
	[0.0102]	[0.0103]	[0.0099]	[0.0110]	[0.0124]	[0.0119]	[0.0108]
	(0.0094)	(0.0099)	(0.0090)	(0.0106)	(0.0113)	(0.0103)	(0.0098)
B. Interactive Regression Model							
Fraction best ML method	0.40	0.16	0.00	0.11	0.00	0.33	–
Median ATE (2 fold)	0.0291	0.0219***	0.0272	0.0216**	0.0252***	0.0271	0.0265
	[0.0181]	[0.0085]	[0.0227]	[0.0099]	[0.0080]	[0.0447]	[0.0173]
	(0.0152)	(0.0074)	(0.0178)	(0.0084)	(0.0073)	(0.0353)	(0.0149)

Note: Fraction best ML method denotes the fraction of nuisance function estimation problems in which the ML submethod outperforms all the others in terms of out-of-sample MSE, over the folds and S replications. We present conventional standard errors in parentheses and median standard errors that incorporate the variation of sample splitting in square brackets. The symbols $*$, $**$ and $***$ denote significance at the 10, 5 and 1% level, respectively, assessed with median standard errors.

Table 22: 5 Fold DML estimates for the contracts and trade ATE using the main specification, $S = 20$

	SVM	**LASSO**	**Reg. Tree**	**Boost-ing**	**Random Forest**	**Neural Net**	**Best**
A. Partially Linear Regression Model							
Fraction best ML method	0.00	0.00	0.00	0.00	1.00	0.00	–
Median ATE (5 fold)	0.2997***	0.3061**	0.1092	0.1469***	0.1845**	0.2940***	0.1845**
	[0.1147]	[0.1219]	[0.0830]	[0.0393]	[0.0756]	[0.1089]	[0.0781]
	(0.0075)	(0.0076)	(0.0102)	(0.0099)	(0.0151)	(0.0074)	(0.0151)
B. Partially Linear IV Model							
Fraction best ML method	0.00	0.00	0.00	0.00	1.00	0.00	–
Median ATE (5 fold)	0.2203**	0.4701**	0.1047	0.3107***	0.2945**	0.2485***	0.2945**
	[0.0858]	[0.1830]	[0.3581]	[0.0953]	[0.1382]	[0.0760]	[0.1437]
	(0.0422)	(0.0472)	(0.2538)	(0.0718)	(0.1262)	(0.0486)	(0.1262)

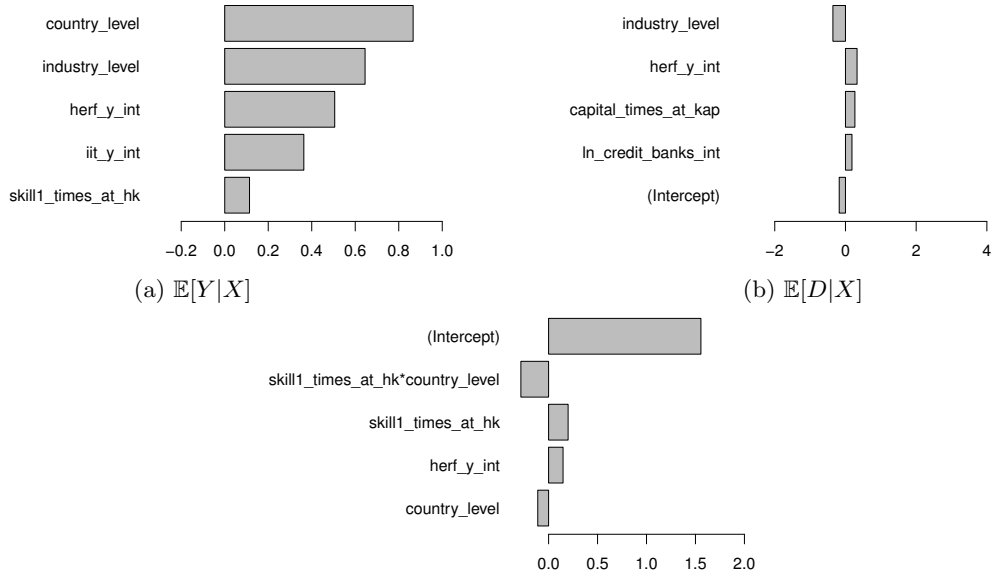
Note: Fraction best ML method denotes the fraction of nuisance function estimation problems in which the ML submethod outperforms all the others in terms of out-of-sample MSE, over the folds and S replications. We present conventional standard errors in parentheses and median standard errors that incorporate the variation of sample splitting in square brackets. The symbols $*$, $**$ and $***$ denote significance at the 10, 5 and 1% level, respectively, assessed with median standard errors. We use fewer replications because computation time is longer here due to the large amount of observations ($n = 10,976$) and folds.



(a) $\mathbb{E}[Y|X]$ (b) $\mathbb{E}[D|X]$

(c) $\mathbb{E}[Z|X]$

Figure 10: Contracts and trade LASSO terms with the largest absolute mean coefficient size over the folds and S replications, in the PLIV using the main specification. * in the name of the term denotes an interaction and $\wedge 2$ denotes a squared term. The abbreviations for the control variables are explained in Table 28 in Section A.2 of the Appendix.



(a) $\mathbb{E}[Y|X]$ (b) $\mathbb{E}[D|X]$

(c) $\mathbb{E}[Z|X]$

Figure 11: Contracts and trade LASSO terms with the largest absolute mean coefficient size over the folds and S replications, in the PLIV using the extended specification. * in the name of the term denotes an interaction and $\wedge 2$ denotes a squared term. The abbreviations for the control variables are explained in Table 28 in Section A.2 of the Appendix.

ARB

Table 23: Fully linear ARB estimates for the Fox News audience ATE

	LASSO			**Elastic Net**		
	$\zeta = 0.3$	$\zeta = 0.5$	$\zeta = 0.7$	$\zeta = 0.3$	$\zeta = 0.5$	$\zeta = 0.7$
Interactive Regression Model						
ATE	0.0219***	0.0218***	0.0210**	0.0219***	0.0218***	0.0210**
	(0.0084)	(0.0085)	(0.0086)	(0.0084)	(0.0085)	(0.0086)

Note: The symbols $*$, $**$ and $***$ denote significance at the 10, 5 and 1% level, respectively. ζ is a tuning parameter that trades off bias against variance, with lower values corresponding to less variance but more bias.

Table 24: Quadratic ARB estimates for the Fox News audience ATE

	LASSO			**Elastic Net**		
	$\zeta = 0.3$	$\zeta = 0.5$	$\zeta = 0.7$	$\zeta = 0.3$	$\zeta = 0.5$	$\zeta = 0.7$
Interactive Regression Model						
ATE	0.0193**	0.0207**	0.0218**	0.0193**	0.0207**	0.0218**
	(0.0087)	(0.0087)	(0.0087)	(0.0087)	(0.0087)	(0.0087)

Note: The symbols $*$, $**$ and $***$ denote significance at the 10, 5 and 1% level, respectively. ζ is a tuning parameter that trades off bias against variance, with lower values corresponding to less variance but more bias.

Heterogeneous DML

Table 25: Insignificant orthogonal debiased LASSO estimates for level 3 Fox News vote share HTEs, $S = 50$

	Random Forest		**Neural Net.**		**SVM**		**Best**	
	Median TE	se	Median TE	se	Median TE	se	Median TE	se
HTE Model with Modeled Heterogeneity								
Alaska	−0.0016	(0.0056)	−0.0002	(0.0020)	−0.0014	(0.0032)	−0.0021	(0.0030)
Arkansas	0.0004	(0.0050)	0.0000	(0.0034)	−0.0026	(0.0037)	−0.0002	(0.0038)
California	0.0041	(0.0026)	0.0011	(0.0020)	0.0027	(0.0020)	0.0036	(0.0024)
Idaho	−0.0081	(0.0058)	−0.0030	(0.0047)	−0.0031	(0.0034)	−0.0062	(0.0040)
Maine	−0.0004	(0.0020)	−0.0018	(0.0016)	−0.0014	(0.0014)	−0.0007	(0.0017)
Missouri	−0.0041	(0.0044)	−0.0007	(0.0033)	−0.0025	(0.0036)	−0.0005	(0.0043)
Montana	0.0048	(0.0042)	−0.0034	(0.0031)	−0.0002	(0.0032)	−0.0008	(0.0033)
New Hampshire	0.0004	(0.0025)	0.0006	(0.0007)	0.0004	(0.0016)	0.0024	(0.0024)
New Jersey	0.0028	(0.0021)	0.0012	(0.0019)	−0.0005	(0.0018)	0.0011	(0.0019)
North Dakota	−0.0019	(0.0069)	0.0025	(0.0024)	0.0005	(0.0037)	0.0011	(0.0037)
Ohio	0.0002	(0.0012)	−0.0011	(0.0010)	−0.0014	(0.001)	−0.0013	(0.001)
Pennsylvania	−0.0004	(0.0012)	0.0015	(0.0012)	−0.0005	(0.0012)	−0.0004	(0.0012)
Tennessee	−0.0049	(0.0039)	−0.0022	(0.0028)	−0.0026	(0.0027)	−0.0019	(0.0029)
Utah	−0.0053	(0.0040)	−0.0023	(0.0036)	−0.0050	(0.0035)	−0.0053	(0.0036)
Vermont	0.0015	(0.0028)	−0.0011	(0.0018)	−0.0008	(0.0020)	−0.0028	(0.0025)
Virginia	0.0012	(0.0021)	−0.0004	(0.0021)	−0.0008	(0.0024)	−0.0016	(0.0028)

Note: Heterogeneity levels are from the hierarchy outlined in Figure 13. Conventional standard errors are given in parentheses. The symbols $*$, $**$ and $***$ denote significance at the 10, 5 and 1% level, respectively. The states are ordered alphabetically.

Causal Forest

Table 26: Remaining states' Causal Forest estimates for level 3 Fox News vote share HTEs

	Causal Forest	
	Median TE	se
Conditional ATE Model		
Utah	0.0072***	(0.0025)
Alaska	0.0072***	(0.0026)
Alabama	0.0072***	(0.0026)
Hawaii	0.0072***	(0.0026)
Idaho	0.0072***	(0.0026)
Massachusetts	0.0072***	(0.0026)
Missouri	0.0072***	(0.0026)
Montana	0.0072***	(0.0026)
North Dakota	0.0072***	(0.0026)
New Hampshire	0.0072***	(0.0026)
Rhode Island	0.0072***	(0.0026)
Virginia	0.0072***	(0.0026)
Vermont	0.0072***	(0.0026)
South Carolina	0.0072***	(0.0026)
Connecticut	0.0072***	(0.0025)
California	0.0072***	(0.0026)
Minnesota	0.0072***	(0.0026)
Maine	0.0072***	(0.0027)

Note: Heterogeneity levels are from the hierarchy outlined in Figure 13. Conventional standard errors are given in parentheses. The symbols $*$, $**$ and $***$ denote significance at the 10, 5 and 1% level, respectively. The states are ordered according to significance strength, i.e. the amount of asterisks.

A.2 Supporting material

Table 27: Description of the individual Fox News control variables for the LASSO term importance figures.

Variable	Control set	Description
reppresfv2p1996	$v_{k,1996}^{Rep}$	Republican vote share in 1996
reppresfv2p1992	$v_{k,1992}^{Rep}$	Republican vote share in 1992
reppresfv2p1988	$v_{k,1988}^{Rep}$	Republican vote share in 1988
black2000 and black00m90	$X_{k,2000}$ and $X_{k,00-90}$	Fraction of the town African-American
hisp2000 and hisp00m90	$X_{k,2000}$ and $X_{k,00-90}$	Fraction of the town Hispanic
empl2000 and empl00m90	$X_{k,2000}$ and $X_{k,00-90}$	Fraction of the town employed
unempl2000 and unempl00m90	$X_{k,2000}$ and $X_{k,00-90}$	Fraction of the town unemployed
male2000 and male00m90	$X_{k,2000}$ and $X_{k,00-90}$	Fraction of the town male
married2000 and married00m90	$X_{k,2000}$ and $X_{k,00-90}$	Fraction of the town married
urban2000 and urban00m90	$X_{k,2000}$ and $X_{k,00-90}$	Fraction of the town urban
hs2000 and hs00m90	$X_{k,2000}$ and $X_{k,00-90}$	Fraction of the town with high school
hsp2000 and hsp00m90	$X_{k,2000}$ and $X_{k,00-90}$	Fraction of the town with some college
college2000 and college00m90	$X_{k,2000}$ and $X_{k,00-90}$	Fraction of the town with a degree
noch2000d2 until noch2000d9	$C_{k,2000}$	Decile 2 until 9 in the number of channels across towns
poptot2000d2 until poptot2000d10	$C_{k,2000}$	Decile 2 until 9 in voting age population reached across towns

Note: The variables with *and* are available for 2000 and 1990. The ending 2000 indicates the variable for 2000, whereas 00m90 indicates the difference between the variables for 2000 and 1990. Control set denotes the set that is described in Section 4.1 to which the variable belongs.

Table 28: Description of the individual contracts and trade control variables for the LASSO term importance figures.

Variable	Control set	Description
country_level	Country trade level	Trade level variable within a country
industry_level	Industry trade level	Trade level variable within an industry
capital_times_at_kap	$k_i K_c$	Capital
skill1_times_at_hk	$h_i H_c$	Human capital
herf_y_int	$(1 - hf_i) ln(y_c)$	Input variety
iit_y_int	$iit_i ln(y_c)$	Intra-industry trade
ln_credit_banks_int	$k_i CR_c$	Financial development

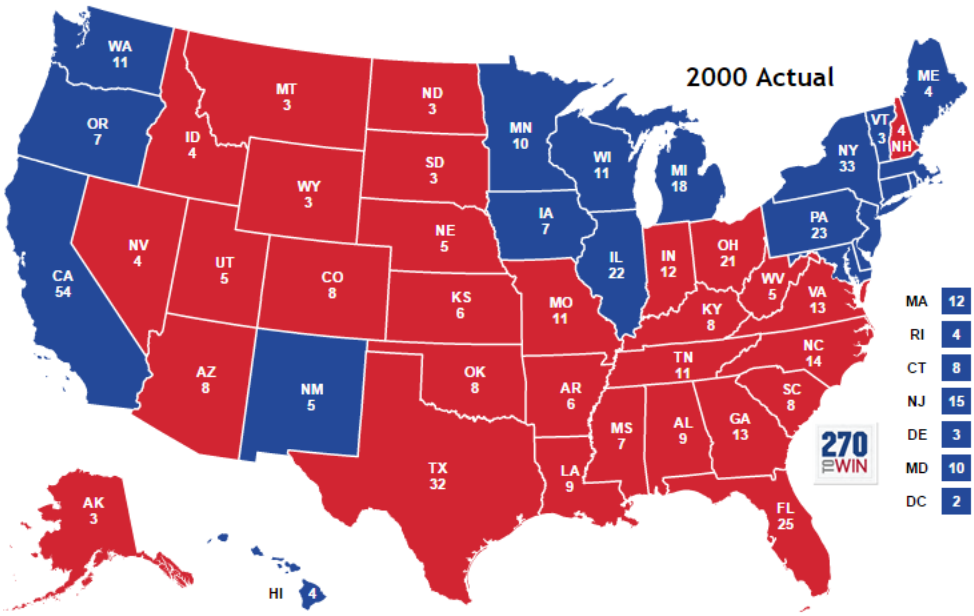


Figure 12: Result of the 2000 U.S. presidential elections; red corresponds to a Republican win and blue to a Democratic win.

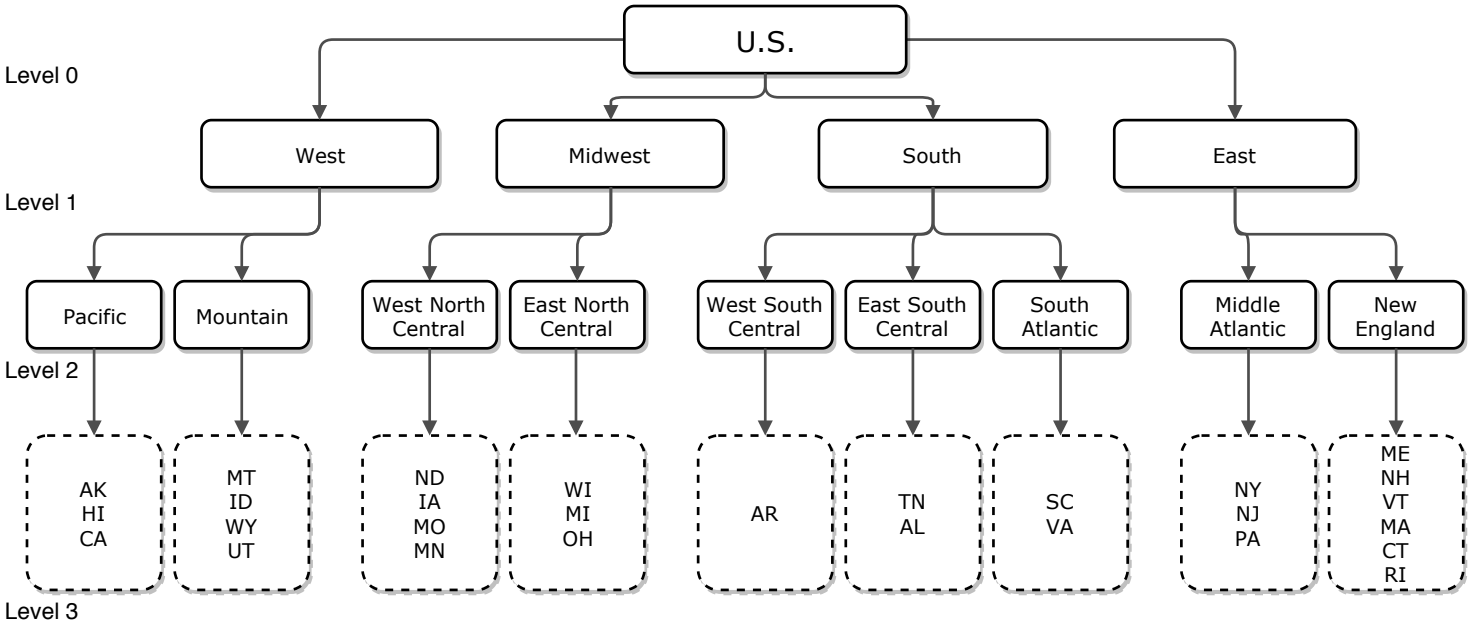


Figure 13: Geographical hierarchy for Fox News effect heterogeneity

A.3 ML submethods

Regression Tree

Regression trees recursively split the covariate space into different regions. A tree can in this way be constructed according to the covariate splits, with subgroups of observations as final nodes or leafs. For each leaf, we estimate a different effect on the response variable. The splitting variable and the splitting point are commonly chosen to minimize impurity. Impurity from a split is the weighted average of the node impurities of the daughter nodes, with weights proportional to the number of daughter node observations. We calculate the node impurity measure by running the regression with only a constant by using the observations in the node that we are considering. Node impurity is then the sum of squared errors of this regression. To avoid overfitting, we eventually remove part of the tree by using cost-complexity pruning. We grow a tree without stopping criterion and compute for each subtree up to the split the cost-complexity measure, i.e. out-of-sample predictive accuracy penalized by complexity. In the end, starting from the leafs, we remove all branches with cost-complexity above the threshold of the minimum value plus one standard deviation.

Random Forest

Another way to avoid overfitting is to combine multiple trees into a random forest. We construct a random forest prediction by averaging the predictions of multiple unpruned trees into a new prediction. For classification, we take a minimum of 3 observations in a leaf and for regression 5. We obtain for each tree a different data set by bootstrapping from the original sample. Furthermore, since it creates variation in the trees, we randomly draw at each splitting point r of the covariates. Next, we only consider the drawn variables for splitting instead of the complete set. We follow common practice and apply $r = \sqrt{p}$. Additionally, we grow 300 trees for each forest, which is a reasonable amount that does not result in too much computation time.

Boosting

Regression trees often have difficulty distinguishing multiple overlapping regions of the covariate space. Forests solve this by combining the predictions of many simultaneously grown trees. Boosting employs a different strategy by sequentially growing trees, where the residuals of a tree serve as input for the next tree. Hence, to improve predictions, it focuses on the observations that are difficult to predict. Boosting methods minimize $\sum_{i=1}^{n} L(Y_i; \gamma)$, where $L(\cdot)$ is the squared loss function for regression and the AdaBoost exponential loss function for classification. First, we calculate pseudo residuals $r_{im} = -[\partial L(Y_i; F(x_i))/\partial F(x_i)]$ evaluated at the tree $F(x) = F_{m-1}(x)$ instead of normal residuals, to be able to handle complex loss functions. These are based on the direction of steepest descent. Next, we grow a regression tree $h_m(x)$ on the pseudo residuals, using exactly $s = 5$ leaf nodes. Then, we compute the update of our regression tree: $F_m(x) = F_{m-1}(x) + \nu\gamma_m h_m(x)$, where γ_m are computed as $\arg\min_\gamma L(Y_i, F_{m-1}(x) + \nu\gamma h_m(x))$ and $\nu = 0.01$ is a regularization factor. Finally, we note that we randomly select a fraction $f = 0.5$ of the training sample observations to fit the $h_m(x)$ in each iteration, while we use the test sample to compute the loss function. We use 2-fold cross-validation to get training and test samples. We repeat the whole procedure $M = 300$ times and use standard initialization routines.

Neural Network

Another class of ML methods uses weighted combinations of the covariates to predict the response. Neural networks belong to this class. neural networks make weighted linear combinations of input covariates or nodes and transform them with a logistic transformation to an end node with probabilities. The logistic transformation permits us to capture nonlinear relationships. The weights are optimized using a gradient descent algorithm on a least squares loss function, using as training data class observations for classification and probabilities for regression. These probabilities are calculated by standardization with the minimum and maximum value of a variable. We use a decay regularization parameter $\lambda = 0.02$ in the gradient descent algorithm. In practice, adding an additional layer between the input and end node often improves the predictions. Moreover, it enables us to model covariate interactions. Therefore, we

choose to add a hidden layer with $d = 2$ additional nodes. The input covariate nodes are transformed to probabilities by the hidden layer nodes and these serve as input data for the end node.

LASSO & Elastic Net

The last class of ML methods modifies standard regression methods by applying regularization. Popular implementations are the LASSO and the elastic net. Both shrink regression coefficients towards zero by adding a penalty term. The LASSO uses an ℓ_1-penalty term, which leads to sparse estimates. It is computed by minimizing $\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'\beta)^2 + \frac{\lambda}{n}||\hat{\Gamma}\beta||_1$ over β, where $\Gamma = \text{diag}(\hat{\gamma}_1, \ldots, \hat{\gamma}_p)$ is a penalty loading matrix and λ a penalty parameter. The elastic net brings additional stability by using a combination of the ℓ_2- and ℓ_1-penalty term. That is, it adds a penalty term $\frac{\lambda}{n}(\alpha||\beta||_2^2 + (1-\alpha)||\beta||_1)$ instead of the previously added $\frac{\lambda}{n}||\hat{\Gamma}\beta||_1$. Therefore, the elastic net can be seen as generalization of the LASSO. For DML, we use the LASSO with data-driven choice of λ, as proposed by Belloni et al. (2012), because it is robust to nonnormality and heteroskedasticity. The algorithm initializes the penalty components as $\hat{\gamma}_l = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2(y_i - \bar{y})^2}, l = 1, \ldots, p$ and specifies $\lambda = 2c\sqrt{n}\Phi^{-1}(1 - \gamma/2p)$, with $c = 1.1$ and $\gamma = 0.1/\log(\min(n, p))$. Next, it computes the LASSO estimate and its residuals $\hat{v}_i$. Finally, we update the penalty loadings setting $\hat{\gamma}_l = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2\hat{v}_i^2}, l = 1, \ldots, p$ and compute a new LASSO estimate. We iterate the last steps $K = 15$ times. For ARB, we use the LASSO but we simply pick the identity matrix for $\hat{\Gamma}$ and estimate λ with cross-validation. Additionally, we apply the elastic net with cross-validated λ and $\alpha = 0.5$. In both cases, we pick λ that produces the smallest cross-validated error plus 1 standard deviation.

Best

The best method from Chernozhukov et al. (2016) picks for each prediction problem the estimates from the ML method that gives the smallest mean squared prediction error or misclassification error, for regression and classification respectively. Hence, if there is no ML method that outperforms the others in all prediction problems, different ML methods are used for each prediction problem. This hybrid method combines ML methods optimally after estimation.

Support Vector Machine

Support vector machines also use weighted combinations of covariates, but in a slightly different fashion than neural networks. Firstly, for regression, the idea is to find the flattest function $w'x + b$ with at most ϵ deviation from Y_i, which reflects the tradeoff between regularization and overfitting. However, to overcome infeasibility, we add slack variables ξ_i to ϵ to relax deviation restrictions when needed. Hence, formally we optimize the leftmost primal problem over w and b:

$$\min \frac{1}{2}||w||_2^2 + C\sum_{i=1}^{n}\xi_i \qquad \max \frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i^* - \alpha_i)(\alpha_j - \alpha_j^*)x_i'x_j - \epsilon\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) + \sum_{i=1}^{n} y_i(\alpha_i - \alpha_i^*)$$

$$\text{s.t.} \quad |Y_i - w'x_i - b| \leq \epsilon + \xi_i \qquad \text{s.t.} \quad \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0$$

where $C > 0$ is a cost parameter that controls the size of deviations. In practice, due to ease of computation, we turn to the dual problem that is given on the right, with $\alpha_i, \alpha_i^* \in [0, C]$ the dual variables. Afterwards, we use $w = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)x_i$ to construct predictions according to $f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)x_i'x + b$, where b is calculated from KKT conditions belonging to the dual problem.

SVMs capture nonlinear relationships by preprocessing, that is we obtain a new set of covariates $\Phi(x)$ with added nonlinearities like quadratic and cross terms. However, computations become quickly infeasible for either high order nonlinearities or many covariates in this case. The solution is to apply a kernel trick: since we only compute $x_i'x_j$ in the dual problem, it is sufficient to have a function that gives $k(x_i, x_j) = \Phi_i(x)'\Phi_j(x)$ rather than $\Phi(x_i)$ itself. Then, we substitute $x_i'x_j$ in the dual problem and prediction function $f(x)$ for $k(x_i, x_j)$. We apply the widespread radial basis function: $k(u, v) = \exp(-\gamma||u - v||_2^2)$, because it only has a single parameter that needs tuning. Here, we differ from neural networks, since we do not use the logistic kernel. In initial runs, we tuned the parameters C and γ by

cross validation with ranges 2^c with $c = \{-10, \ldots, 0\}$ for γ and $c = \{0, \ldots, 6\}$ for C. This results in the choice of $\gamma = 2^{-9}$ and $C = 2^4$.

For classification, the idea is to find a hyperplane that separates the binary class observations as well as possible. For that, we recode the classes as 1 and -1. Then, the projections of observations on a hyperplane are given by $q_i = w'x_i + b$. Ideally, we would get a hyperplane for which all class 1 observations have $q_i \geq 1$ and all class -1 observations $q_i \leq -1$. However, this is not feasible in most cases, so again we need to use slack variable $\xi_i \geq 0$ to permit deviations. The primal optimization problem is given on the left:

$$\begin{aligned} &\min \frac{1}{2}||w||_2^2 + C\sum_{i=1}^{n}\xi_i && \max \frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i\alpha_j y_i y_j x_i' x_j) - \sum_{i=1}^{n}\alpha_i \\ &\text{s.t.} \quad Y_i(w'x_i + b) \geq 1 - \xi_i && \text{s.t.} \quad \sum\nolimits_{i=1}^{n} Y_i\alpha_i = 0 \end{aligned}$$

For class 1 observations with $q_i \geq 1$ or class -1 observations with $q_i \leq -1$, we set $\xi_i = 0$. Otherwise, we have to correct using the slack variable ξ_i. The restriction on w ensures normalization, it validates the choice of 1 and -1 as boundary points. Again, for computation of the solution, we employ the dual formulation on the right, with variable $\alpha_i \in [0, C]$. Furthermore, by using $w = \sum_{i=1}^{n} y_i\alpha_i x_i$ we calculate decision values $f(x) = \sum_{i=1}^{n} y_i\alpha_i x_i' x + b$, where b is obtained from the KKT conditions. Probability predictions follow from fitting a logistic distribution to the decision values using maximum likelihood. The kernel trick with radial basis function can be applied similarly as for regression. With the same initial ranges, we get $\gamma = 2^{-6}$ and $C = 2^3$. For regression and classification, we follow standard practice by using $\epsilon = 0.1$. We also standardize all variables before training SVMs.

A.4 Proofs and derivations

DML

Consider the PLR model and Robinson (1988) style score function. We apply GMM and solve the corresponding sample moment conditions as follows:

$$
\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\phi(W_i;\theta_0,\eta_0)=0\\
\implies &\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\{Y_i-l_0(X_i)-\beta_0(D_i-m_0(X_i))\}\{D_i-m_0(X_i)\}=0\\
\implies &\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\{Y_i-l_0(X_i)\}\{D_i-m_0(X_i)\}-\beta_0\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\{D_i-m_0(X_i)\}^2=0\\
\implies &\beta_0\sum_{k=1}^{K}\sum_{i\in I_k}\{D_i-m_0(X_i)\}^2=\sum_{k=1}^{K}\sum_{i\in I_k}\{Y_i-l_0(X_i)\}\{D_i-m_0(X_i)\}\\
\implies &\hat{\beta}_0^{DML}=\Big(\sum_{k=1}^{K}\sum_{i\in I_k}\{D_i-m_0(X_i)\}^2\Big)^{-1}\Big(\sum_{k=1}^{K}\sum_{i\in I_k}\{D_i-m_0(X_i)\}\{Y_i-l_0(X_i)\}\Big),
\end{aligned}
$$

which indeed equals the expression $(X'X)^{-1}(X'y)$ for OLS estimation of residualized Y_i on residualized D_i.

Consider the PLR model and Robinson (1988) style score function. We fill in the expressions in order to compute the DML variance estimates:

$$
\begin{aligned}
&\widehat{J}_0=\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\{D_i-m_0(X_i)\}^2\\
\implies &\widehat{\Sigma}=(\widehat{J}_0^{-1})\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\phi(W_i;\hat{\theta}_0,\hat{\eta}_{0,k})\phi(W_i;\hat{\theta}_0,\hat{\eta}_{0,k})'(\widehat{J}_0^{-1})'\\
&\quad=\Big(\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\{D_i-m_0(X_i)\}^2\Big)^{-1}\Big(\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\{D_i-m_0(X_i)\}^2\{Y_i-l_0(X_i)-\hat{\beta}_0^{DML}(D_i-m_0(X_i))\}^2\Big)\\
&\quad\cdot\Big(\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\{D_i-m_0(X_i)\}^2\Big)^{-1}.\\
\implies &\mathrm{Var}[\hat{\beta}_0^{DML}]=\mathrm{Var}[(1/\sqrt{n})\widehat{\Sigma}]=\frac{1}{n}\Big(\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\{D_i-m_0(X_i)\}^2\Big)^{-1}\\
&\quad\cdot\Big(\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\{D_i-m_0(X_i)\}^2\{Y_i-l_0(X_i)-\hat{\beta}_0^{DML}(D_i-m_0(X_i))\}^2\Big)\Big(\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\{D_i-m_0(X_i)\}^2\Big)^{-1},
\end{aligned}
$$

which indeed equals the expression $(X'X)^{-1}(X'\mathrm{diag}\{\hat{u}_1^2,\ldots,\hat{u}_1^2\}X)(X'X)^{-1}$ for White's heteroscedasticity consistent standard errors using residualized Y_i and residualized D_i.

Consider the interactive model and Robins and Rotnitzky (1995) style score function.

$$
\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\phi(W_i;\theta_0,\eta_0)=0\\
\Longrightarrow\ &\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\Big(g_0(1,X_i)-g_0(0,X_i)+\frac{D_i\{Y_i-g_0(1,X_i)\}}{m_0(X_i)}-\frac{(1-D_i)\{Y_i-g_0(0,X_i)\}}{1-m_0(X_i)}-\beta_0\Big)=0\\
\Longrightarrow\ &\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\beta_0=\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\Big(g_0(1,X_i)-g_0(0,X_i)+\frac{D_i\{Y_i-g_0(1,X_i)\}}{m_0(X_i)}-\frac{(1-D_i)\{Y_i-g_0(0,X_i)\}}{1-m_0(X_i)}\Big)\\
\Longrightarrow\ &\hat{\beta}_0^{DML}=\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\Big(g_0(1,X_i)-g_0(0,X_i)+\frac{D_i\{Y_i-g_0(1,X_i)\}}{m_0(X_i)}-\frac{(1-D_i)\{Y_i-g_0(0,X_i)\}}{1-m_0(X_i)}\Big),
\end{aligned}
$$

which indeed equals the mean over $\psi(W_i;\eta_0)$ from (3.1.1), denoted $\overline{\psi}(W_i;\eta_0)$ in the following.

Consider the interactive model and Robins and Rotnitzky (1995) style score function. We fill in the expressions in order to compute the DML variance estimates:

$$
\begin{aligned}
&\widehat{J}_0=\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}1=1\\
\Longrightarrow\ &\widehat{\Sigma}=(\widehat{J}_0^{-1})\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\phi(W_i;\hat{\theta}_0,\hat{\eta}_{0,k})\phi(W_i;\hat{\theta}_0,\hat{\eta}_{0,k})'(\widehat{J}_0^{-1})'\\
&\quad=\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\Big(g_0(1,X_i)-g_0(0,X_i)+\frac{D_i\{Y_i-g_0(1,X_i)\}}{m_0(X_i)}-\frac{(1-D_i)\{Y_i-g_0(0,X_i)\}}{1-m_0(X_i)}-\hat{\beta}_0^{DML}\Big)^2\\
&\quad=\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\Big(\psi(W_i;\eta_0)-\overline{\psi}(W_i;\eta_0)\Big)^2\\
\Longrightarrow\ &\mathrm{Var}[\hat{\beta}_0^{DML}]=\mathrm{Var}[(1/\sqrt{n})\widehat{\Sigma}]=\frac{1}{n}\Big(\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\Big(\psi(W_i;\eta_0)-\overline{\psi}(W_i;\eta_0)\Big)^2\Big)
\end{aligned}
$$

which indeed suggests that DML standard errors are computed as the standard deviation of $\psi(W_i;\eta_0)$ divided by $\sqrt{n}$.

Consider the PLIV model and Robinson (1988) style score function. We apply GMM and solve the corresponding sample moment conditions as follows:

$$
\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\phi(W_i;\theta_0,\eta_0)=0\\
\implies &\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\{Y_i-l_0(X_i)-\beta_0(D_i-r_0(X_i))\}\{Z_i-m_0(X_i)\}=0\\
\implies &\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\{Y_i-l_0(X_i)\}\{D_i-r_0(X_i)\}-\beta_0\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\{Z_i-m_0(X_i)\}\{D_i-r_0(X_i)\}=0\\
\implies &\beta_0\sum_{k=1}^{K}\sum_{i\in I_k}\{Z_i-m_0(X_i)\}\{D_i-r_0(X_i)\}=\sum_{k=1}^{K}\sum_{i\in I_k}\{Y_i-l_0(X_i)\}\{D_i-r_0(X_i)\}\\
\implies &\hat{\beta}_0^{DML}=\Big(\sum_{k=1}^{K}\sum_{i\in I_k}\{Z_i-m_0(X_i)\}\{D_i-m_0(X_i)\}\Big)^{-1}\Big(\sum_{k=1}^{K}\sum_{i\in I_k}\{D_i-m_0(X_i)\}\{Y_i-l_0(X_i)\}\Big),
\end{aligned}
$$

which indeed equals the expression $(Z'X)^{-1}(Z'y)$ for IV estimation of residualized Y_i on residualized D_i using as instrument residualized Z_i.

Consider the PLIV model and Robinson (1988) style score function. We fill in the expressions in order to compute the DML variance estimates:

$$
\begin{aligned}
&\widehat{J}_0=\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\{D_i-r_0(X_i)\}\{Z_i-m_0(X_i)\}\\
\implies &\widehat{\Sigma}=(\widehat{J}_0^{-1})\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\phi(W_i;\hat{\theta}_0,\hat{\eta}_{0,k})\phi(W_i;\hat{\theta}_0,\hat{\eta}_{0,k})'(\widehat{J}_0^{-1})'\\
&=\Big(\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\{Z_i-m_0(X_i)\}\{D_i-m_0(X_i)\}\Big)^{-1}\Big(\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\{Z_i-m_0(X_i)\}^2\\
&\cdot\{Y_i-l_0(X_i)-\hat{\beta}_0^{DML}(D_i-m_0(X_i))\}^2\Big)\Big(\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\{Z_i-m_0(X_i)\}\{D_i-m_0(X_i)\}\Big)^{-1}.\\
\implies &\mathrm{Var}[\hat{\beta}_0^{DML}]=\mathrm{Var}[(1/\sqrt{n})\widehat{\Sigma}]=\frac{1}{n}\Big(\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\{Z_i-m_0(X_i)\}\{D_i-m_0(X_i)\}\Big)^{-1}\\
&\cdot\Big(\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\{Z_i-m_0(X_i)\}^2\{Y_i-l_0(X_i)-\hat{\beta}_0^{DML}(D_i-m_0(X_i))\}^2\Big)\\
&\cdot\Big(\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\{Z_i-m_0(X_i)\}\{D_i-m_0(X_i)\}\Big)^{-1},
\end{aligned}
$$

which indeed equals the expression $(Z'X)^{-1}(Z'\mathrm{diag}\{\hat{u}_1^2,\ldots,\hat{u}_1^2\}Z)(Z'X)^{-1}$ for White's heteroscedasticity consistent standard errors IV estimation using residualized Y_i, residualized D_i and residualized Z_i.

Causal Forest

Consider the CATE model with the associated score function. We apply weighted GMM with the corresponding system of sample moment conditions:

$$\sum_{i=1}^{n} \psi(O_i; \theta_0(x^*), \eta_0(x^*)) = 0$$
$$\implies \begin{cases} \sum_{i=1}^{n} & \alpha_i(x^*)(Y_i - \theta_0(x^*)D_i - \eta_0(x^*) = 0 \\ \sum_{i=1}^{n} & \alpha_i(x^*)(Y_i - \theta_0(x^*)D_i - \eta_0(x^*)D_i = 0 \end{cases}$$

We start with solving the top equation:

$$\sum_{i=1}^{n} \alpha_i(Y_i - \theta_0 D_i - \eta_0) = 0$$
$$\implies \eta_0 \sum_{i=1}^{n} \alpha_i = \sum_{i=1}^{n} \alpha_i Y_i - \theta_0 \sum_{i=1}^{n} \alpha D_i$$
$$\implies \hat{\eta}_0^{CF} = \overline{Y}_\alpha - \theta_0 \overline{D}_\alpha,$$

because $\sum_{i=1}^{n} \alpha_i = 1$. Inserting this into the second equation yields:

$$\sum_{i=1}^{n} \alpha_i(Y_i - \theta_0 D_i - \eta_0)D_i = 0$$
$$\implies \sum_{i=1}^{n} \alpha_i(Y_i - \theta_0 D_i - (\overline{Y}_\alpha - \theta_0 \overline{D}_\alpha))D_i = 0$$
$$\implies \sum_{i=1}^{n} \alpha_i((Y_i - \overline{Y}_\alpha) - \theta_0(D_i - \overline{D}_\alpha))D_i = 0.$$

Next, we add the following term:

$$\sum_{i=1}^{n} \alpha_i((Y_i - \overline{Y}_\alpha) - \theta_0(D_i - \overline{D}_\alpha))\overline{D}_\alpha$$
$$= \overline{D}_\alpha \sum_{i=1}^{n} \alpha_i(Y_i - \overline{Y}_\alpha) - \overline{D}_\alpha \theta_0 \sum_{i=1}^{n} \alpha_i(D_i - \overline{D}_\alpha)$$
$$= \overline{D}_\alpha(\overline{Y}_\alpha - \overline{Y}_\alpha) - \overline{D}_\alpha \theta_0(\overline{D}_\alpha - \overline{D}_\alpha) = 0,$$

in order to obtain:

$$\sum_{i=1}^{n} \alpha_i((Y_i - \overline{Y}_\alpha) - \theta_0(D_i - \overline{D}_\alpha))(D_i - \overline{D}_\alpha) = 0$$
$$\implies \theta_0 \sum_{i=1}^{n} \alpha_i(Y_i - \overline{Y}_\alpha)(D_i - \overline{D}_\alpha) = \sum_{i=1}^{n} \alpha_i(D_i - \overline{D}_\alpha)^2$$
$$\implies \hat{\theta}_0^{CF} = \Big(\sum_{i=1}^{n} \alpha_i(D_i - \overline{D}_\alpha)^2 \Big)^{-1} \Big(\sum_{i=1}^{n} \alpha_i(D_i - \overline{D}_\alpha)(Y_i - \overline{Y}_\alpha) \Big)$$

which indeed equals the expression that is given in the main text.