Estimation and Inference on Nonlinear and Heterogeneous Effects*

Marc Ratkovic† Dustin Tingley‡

August 31, 2022

Abstract

While multiple regression offers transparency, interpretability, and desirable theoretical properties, the method's simplicity precludes the discovery of complex heterogeneities in the data. We introduce the Method of Direct Estimation and Inference (MDEI) that embraces these potential complexities, is interpretable, has desirable theoretical guarantees, and, unlike some existing methods, returns appropriate uncertainty estimates. The proposed method uses a machine learning regression methodology to estimate the observation-level partial effect, or "slope," of a treatment variable on an outcome, and allows this value to vary with background covariates. Importantly, we introduce a robust approach to uncertainty estimates. Specifically, we combine a split-sample and conformal strategy to fit a confidence band around the partial effect curve that will contain the true partial effect curve at some controlled proportion of the data, say 90% or 95%, even in the presence of model misspecification. Simulation evidence and an application illustrate the method's performance.

Key Words: machine learning, statistical inference, conformal inference, cross-fitting, coverage

*We thank Scott de Marchi, Max Gopelrud, Kosuke Imai, Lucas Janson, Shiro Kuriwaki, Lihua Lei, Lisa McKay, Max Farrell, and Brandon Stewart for comments on this paper. A previous unpublished paper, "The Method of Direct Estimation" (2017), worked towards the approach to point estimates used in this paper. On this prior work, we would like to thank Peter Aronow, Scott de Marchi, James Fowler, Andrew Gelman, Max Gopelrud, Kosuke Imai, Gary King, Shiro Kuriwaki, John Londregan, Chris Lucas, Walter Mebane, Rich Nielsen, Molly Roberts, Brandon Stewart, Aaron Strauss, Rocio Titiunik, Tyler VanderWeele, Teppei Yamamoto, Soichiro Yamauchi, and Xiang Zhou, as well as the participants at the Quantitative Social Science Seminar at Princeton, Yale Research Design and Causal Inference seminar, Empirical Implications of Theoretical Models 2018 workshop, and Harvard Applied Statistics workshop.

†Assistant Professor, Department of Politics, Princeton University, Princeton NJ 08544. Phone: 608-658-9665, Email: ratkovic@princeton.edu, URL: http://scholar.princeton.edu/ratkovic

‡Professor of Government, Harvard University, Email: dtingley@gov.harvard.edu, URL: scholar.harvard.edu/dtingley

Table of Contents

Data analysis in much of political science and other social sciences is often synonymous with multiple linear regression. In this project, we assume the researcher confronts an outcome variable, a treatment variable of central interest, and a set of background "control"/"confounding" variables that characterizes each observation's covariate profile. The usefulness of multiple regression in this context depends in part on correctly modeling the influence of the treatment variable while adjusting for the confounding effects of other variables. Typical regression strategies commonly ignore complexity in the data, such as the heterogeneous effect of the treatment across the sample (a treatment by covariate interaction), or they assume all effects are linear (both the treatment and confounders). Departures from typical practice tend to be ad hoc, with maybe one interaction or non-linearity considered. While methods have been introduced for moving beyond multiple regression for finding nonlinearities and interactions, estimating these nonlinearities and interactions is not the same as also returning appropriate uncertainty estimates.

We introduce a novel method for finding nonlinear and heterogeneous effects, and focus on how to appropriately calculate uncertainty in these settings. We propose the Method of Direct Estimation and Inference (MDEI)–that embraces these heterogeneities and nonlinearities while still returning appropriate uncertainty estimates on effects. We focus largely on the case of a continuous treatment variable but also consider the binary case. As with much work in the causal inference literature (Aronow and Miller, 2018; Imbens and Rubin, 2015; Ho et al., 2007; Holland, 1986) we focus on reducing the role of modelling assumptions. However, our approach minimizes the role of assumptions in estimating both point estimates and uncertainty estimates.

We introduce a method that estimates the slope of the treatment variable on the outcome at each datum, the *partial effect* (Wooldridge, 2002, Sec. 2.2.2), allowing this slope to be a function of background covariates. The proposed method flexibly adjusts for background covariates while also allowing for substantial flexibility in the effect of the treatment on the outcome.

The next step–and crucial to this paper–is to generate uncertainty estimates and confidence bands for the results. This is straightforward when there are strong parametric modelling assumptions in place, as with multiple linear regression. The task is much more challenging when we want to allow for more complicated relationships that we do not specify ex-ante. Our goal is to generate a confidence band around any uncovered nonlinearity that will allow us to assess how the estimated curve relates to the true curve. Thinking about uncertainty in this setting requires differentiating between inference at a particular point and inference over a curve. For this, we estimate a confidence band with *average coverage*, meaning we expect the confidence band for our marginal effect to cover the true curve at some proportion, say 90%, of the data. Doing so allows us to use the uncertainty measure around the curve to deduce features of the true underlying curve.

In summary, the MDEI framework provides flexible and reliable estimation *and* inference. The method consists of two parts. First, we estimate the partial effect curve, which is the observation-level effect of the treatment on the outcome. Just as a regression coefficient is interpreted as a marginal effect over the sample, the partial effect is interpreted as the "slope" at a given observation, given the values of observed pre-treatment variables. In generating this estimate, MDEI advances recent machine learning methods by implementing a flexible, nonparametric regression to model the partial effect. The model can detect a wide class of nonlinear and treatment/covariate interactions.

Second, we introduce a confidence band on uncovered nonlinearities and heterogeneities that the researcher can use to assess whether a given effect reflects a systematic pattern in the data. The curve has the *average coverage property* (Nychka, 1988; Wasserman, 2006) that the band will contain the true partial effect at some chosen proportion, say 90% or 95%, of the observed data. In constructing such a curve, we rely on *conformal inference* (Lei and Wasserman, 2014). As discussed below, conformal inference provides a data driven, rather than assumption driven, approach to calculating uncertainty estimates on predicted values. We extend the method from predicted values

to estimating the partial effect of the treatment on the outcome at each point. Bringing all of these things together, researchers can obtain a plot of a partial effect curve that can vary over the covariate space but with a confidence band around it that does not rely on various assumptions.

The MDEI framework draws on tools and ideas that might be new for many readers in political science. Throughout the paper we try to introduce these ideas in an accessible manner and refer readers to a more technical appendix. While the MDEI framework introduced here is new, we relate our approach to existing methods where relevant. Of course, any time parametric and inferential assumptions are relaxed, the importance of having more data increases. Our approach is no different given the data driven, rather than assumption driven, focus of the method.

The paper proceeds as follows. Section 1 lays out the challenge of estimating and conducting inference on partial effects without relying on simplifying assumptions about how the treatment impacts the outcome variable. Section 2 introduces our approach and shows how we estimate both point estimates and uncertainty estimates. Section 3 provides simulations to illustrate our approach while the Appendix compares the performance of MDEI to other cutting edge approaches. Section 4 shows MDEI in action with an applied example. Section 5 concludes. Throughout we discuss related research, but we defer technical details to the online appendix.

1 The Estimation and Uncertainty Challenge

Consider the familiar regression model,

$$y_i = \theta t_i + \mathbf{x}_i^\top \gamma + \epsilon_i; \;\; \mathbb{E}(\epsilon_i | t_i, \mathbf{x}_i) = 0$$

with observations $i \in \{1, 2, \ldots, n\}$, outcome y_i, a variable of theoretical interest t_i, a vector of additional background variables $\mathbf{x}_i$, which includes the intercept, and an error term ϵ_i that is assumed to be mean-independent of the treatment and background variables. This model is adopted by applied researchers for several reasons. First, θ measures the average partial effect, or "slope,"

when characterizing the relationship between the outcome and treatment. Second, $\mathbf{x}_i^\top \gamma$ adjusts for other variables that impact both the treatment and outcome. Third, given the observed data, readily available software can produce an estimate $\widehat{\theta}$ through the method of least squares. Fourth, inference on the average partial effect, θ, uses a confidence interval of the form

$$\widehat{\theta} \pm C_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\widehat{\theta})}$$

where $C_{1-\alpha/2}$ is a critical value that controls the false positive rate (e.g., under mild conditions on the error terms, we can take 1.64 for $\alpha = 0.1$ or 1.96 for $\alpha = 0.05$) given the variance of the estimated slope coefficient on the treatment variable, $\widehat{\theta}$.

While this regression model is useful and versatile, these results rely on assumptions that the model makes about the relationship between the outcome, treatment, and covariates. In this paper we move past this ubiquitous implementation to a more flexible model of the relationship between outcome, treatment, and covariates.[1] For example, we relax the assumption that the covariates in $\mathbf{x}$ enter linearly, and the researcher need not specify how they enter. Rather, we allow this relationship to be learned from the data. We also relax the assumption that the slope θ is homogeneous over the sample. Instead, we allow this value to vary with the value of the treatment variable t_i (e.g., the effect could be a curve rather than a straight line) and pre-treatment covariates $\mathbf{x}_i$. To do this, we will replace the linear component θt_i with a flexible, interactive function which we denote as $\theta(\widetilde{t}_i, \mathbf{x}_i)$, where $\widetilde{t}_i = t_i - \mathbb{E}(t_i|\mathbf{x}_i)$, in order to isolate the nonsystematic fluctuations in the treatment. Then, we can model the effect of t_i on y_i as the partial derivative of $\theta(\widetilde{t}_i, \mathbf{x}_i)$ with respect to $\widetilde{t}_i$, denoted $\tau(\widetilde{t}_i, \mathbf{x}_i)$, which is the "slope" coefficient at a particular value of the treatment and covariates.

[1] See Appendix A for an introductory discussion of work on relaxing these assumptions for the purposes of point estimation.

Doing so will give us a model of the form

$$y_i = \theta(\widetilde{t}_i, \mathbf{x}_i) + f(\mathbf{x}_i) + e_i \tag{1}$$

$$t_i = g(\mathbf{x}_i) + v_i \tag{2}$$

where our aim is estimation and inference on the *partial effect function*, which at a point $(t_i, \mathbf{x}_i)$ is the function

$$\tau(\widetilde{t}_i, \mathbf{x}_i) = \frac{\partial}{\partial t}\theta(t, \mathbf{x}_i)\Big|_{t=\widetilde{t}_i}. \tag{3}$$

The partial effect can be thought of as the "slope" of the treatment at each observed datum, where we allow this slope to vary and be moderated by the covariates in $\mathbf{x}_i$.

One existing body of work focuses on estimating the average partial effect, i.e. $\mathbb{E}(\tau(\widetilde{t}_i, \mathbf{x}_i))$ (Robinson, 1988; Newey and McFadden, 1994).[2] We work instead a literature that has spent a great deal of time developing and testing machine learning for predictive models. Examples include neural networks (Beck, King and Zeng, 2000), averages of trees (Montgomery and Olivella, 2018; Hill, Weiss and Zhai, 2011; Breiman, 2001; Green and Kern, 2012), gradient boosting methods (Kleinberg et al., 2018), or any average of machine learning models (Grimmer, Messing and Westwood, 2017) and, while excellent at prediction, these methods do not return an estimate of the partial effect curve, $\tau(\widetilde{t}_i, \mathbf{x}_i)$.

Two recent methods have focused on modeling the outcome nonparametrically, in a way that allows us to estimate the partial effect curve $\tau(\widetilde{t}_i, \mathbf{x}_i)$: generalized random forests (GRF, Wager and Athey, 2017; Athey, Tibshirani and Wager, 2019) and Kernel Regularized Least Squares (KRLS Hainmueller and Hazlett, 2013).[3]

[2]These works estimate the average partial effect $\mathbb{E}(\tau(\widetilde{t}_i, \mathbf{x}_i)) = \text{Cov}(y_i, t_i|\mathbf{x}_i)/\text{Var}(t_i|\mathbf{x}_i)$. This can be done through weighting, as in Newey and McFadden (1994), or through regressing $y_i - f(\mathbf{x}_i)$ on $t_i - g(\mathbf{x}_i)$, as in Robinson (1988); Chernozhukov et al. (2018). Neither approach, though, estimates heterogeneities in $\tau(t_i, \mathbf{x}_i)$, which is our interest.

[3]For additional work, see (Cattaneo, Farrell and Feng, Forthcoming), though the method does not allow for more

While we achieve competitive performance in terms of point estimation, our real contribution comes from focusing on uncertainty estimation so as to allow for inference on the underlying partial effect curve. Existing methods provide confidence intervals that are overly narrow for at least one of two different reasons. First, they do not account for misspecification, so the intervals will not reflect any systematic error in estimating the underlying partial effect curve. Second, even if there is no misspecification, the curves are constructed to allow for inference at each given point rather than on inference over the entire partial effect curve. We discuss these points in more detail below.

We introduce a confidence interval that can provably and accurately convey information on the true underlying partial effect curve. We illustrate below the shortcomings of existing methods in generating reliable uncertainty estimates, and how our contributions overcome these issues.

We generate a confidence band at each point $(t_i, \mathbf{x}_i)$ of the form

$$\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i) \pm \widehat{C}_{1-\alpha/2} \sqrt{\widehat{\text{Var}}\left\{\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i)\right\}} \tag{4}$$

that can aid the researcher in finding underlying heterogeneities and nonlinearities in the data. The confidence band is constructed to achieve "average coverage" (Nychka (1988) (see also Wasserman (2006) ch. 5.8), meaning that a $100 \times (1-\alpha)\%$ band will cover the true partial effect curve at $100 \times (1-\alpha)\%$ of the observed data. We could use a normal approximation to generate critical value such as $C_{1-0.95/2} = 1.96$. Instead, we show below the value of *estimating* this quantity in a data-driven fashion, so we denote the estimated critical value as $\widehat{C}_{1-\alpha/2}$.

We integrate two recent strategies in order to achieve this band. The first, *repeated cross-fitting* (Chernozhukov et al., 2018), utilizes different subsamples of the data to estimate the effect and conduct inference. The second, *conformal inference* (Lei and Wasserman, 2014; Lei et al., 2018), utilizes a data-driven method to generate the width of the uncertainty interval such that our band

than a few covariates.

will achieve average coverage even if the model is misspecified.[4] We next move onto the proposed method.

2 The Proposed Method

2.1 Overview

We begin with an overview of our approach, with details following below. Estimation of the partial effect curve and its confidence band proceeds in three steps. In the first step, we generate a set of nonlinear and interactive functions of the treatment and covariates that are used to model the partial effect curve, $\tau(\widetilde{t}_i, \mathbf{x}_i)$. These will come from taking the original treatment and covariate vector and constructing a large set of interacted linear and nonlinear functions of these variables. Details are given below, but the goal is to capture any terms that may be driving heterogeneity in the partial effect. In the second, we use the covariates from the earlier set to generate a model for $\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i)$ and estimate its variance, $\widehat{\text{Var}}(\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i))$. In the third, we estimate the width of the confidence band, using conformal inference to estimate a value of $\widehat{C}_{1-\alpha/2}$ that will give us average coverage.

Constructing the partial effect curve and a confidence band that achieves average coverage relies on combining both the split-sample and conformal strategies. The split-sample approach involves taking the observed sample of the data, splitting it into three equally sized subsamples, and conducting each of the three steps above on a separate subsample of the data. The split-sample approach provides a crucial guard against the biases induced by using the same data for each step of the data.[5] We will refer to these subsamples as the *discovery subsample*, the *estimation subsample*,

[4] For a basic conformal inference tutorial for political scientists, see Samii (2019).

[5] Particularly, as described in Athey and Imbens (2016); Wager and Athey (2017); Chernozhukov et al. (2018), a split-sample approach can reduce the biases introduced by using the same data to learn a model and estimate a partial effect. Lei et al. (2018) describe a method for using a split-sample approach to develop a valid conformal

and the *inference subsample*.

We use the discovery subsample to learn a potential set of nonlinearities and heterogeneities, the estimation subsample to estimate the curve and its variance at each point, and then the inference subsample to estimate the width of the confidence band. This three-sample approach combines methods from two existing literatures that have each implemented a split-sample approach, which rely on these methods as a guard against biases that arise when learning and fitting complex models to the same data. The discovery/estimation split allows us to use one subsample of the data to learn the model and another to estimate heterogeneous effects; Wager and Athey (2017); Athey, Tibshirani and Wager (2019) follow a similar strategy, see also Chernozhukov et al. (2018). The estimation/inference split allows us to utilize a split-sample conformal so that we can calibrate the width of our band without making distributional assumptions Lei et al. (2018).

Of course, splitting the data into thirds raises real efficiency concerns, so we implement a *repeated cross-fitting strategy*, where the roles of the subsamples are swapped, such that all the data is used in each step at some point. This process is then repeated, and the final estimate comes from averaging over this process.

In generating the width of the confidence band, we do not rely on a normal approximation, taking critical values of 1.96 or 1.64 for a 95% or 90% interval. Rather, we rely on conformal inference to provide a data-driven means to estimate the width of the confidence interval (Lei and Wasserman, 2014). The basic idea is to expand the interval using the estimates from the second subsample until it contains a set percentage, again say 90% or 95% of the data in the third subsample. We then use this predictive bound to generate a bound on the partial effect curve, $\tau(\widetilde{t}_i, \mathbf{x}_i)$. We show that integrating conformal inference with our split-sample approach for estimating the partial effect and its variance results in asymptotically valid bands; see Section 2.2.2 and Appendix G .

interval.

To summarize, we are going to use each subsample to perform a different element of our estimation and inference. We will use the discovery subsample to learn a set of possible interactions and heterogeneities in the partial effect curve, we will use the estimation sample to estimate the magnitude of these effects, and the inference subsample to construct a confidence interval around the whole curve. Upon conducting each element of our estimation in each subsample, we swap the roles of each subsample so as to generate a fitted value at each datum. This is termed *cross-fitting*. Then, to guard against our results being driven by a particular split of the data into subsamples, we repeat this cross-fitting multiple times, termed *repeated cross-fitting* (Chernozhukov et al., 2018).

2.2 The Method of Direct Estimation and Inference

2.2.1 The Discovery Subsample: Generating Nonlinear and Interactive Covariates

We use the discovery subsample to construct a set of basis functions that can model the outcome and, hence, partial effect curve. The process proceeds in two steps. In the first, using only data in the discovery subsample, we estimate the functions $\widehat{\mathbb{E}}(y_i|\mathbf{x}_i), \widehat{\mathbb{E}}(t_i|\mathbf{x}_i)$.[6] Using these estimated conditional means, we generate the outcome and treatment with the covariates partialed out as

$$\widetilde{y}_i = y_i - \widehat{E}(y_i|\mathbf{x}_i); \;\; \widetilde{t}_i = t_i - \widehat{E}(t_i|\mathbf{x}_i).$$

where the conditional expectations are done using only data in the discovery subsample.

In order to characterize any nonlinearites and interactions in the data, we generate a large set of *basis functions* which we denote $\{\phi_j(\widetilde{t}_i, \mathbf{x}_i)\}_{j=1}^{p}$. A basis function is simply a function, possibly nonlinear and interactive, of the treatment and the covariates. See Appendix B for an introduction to basis functions.

Different choices of basis functions lead to different classes of estimators, including spline models, regularized regression, or neural networks. At this point, less important is the particular choice of

[6]For speed, we use a random forest at this step (Breiman, 2001).

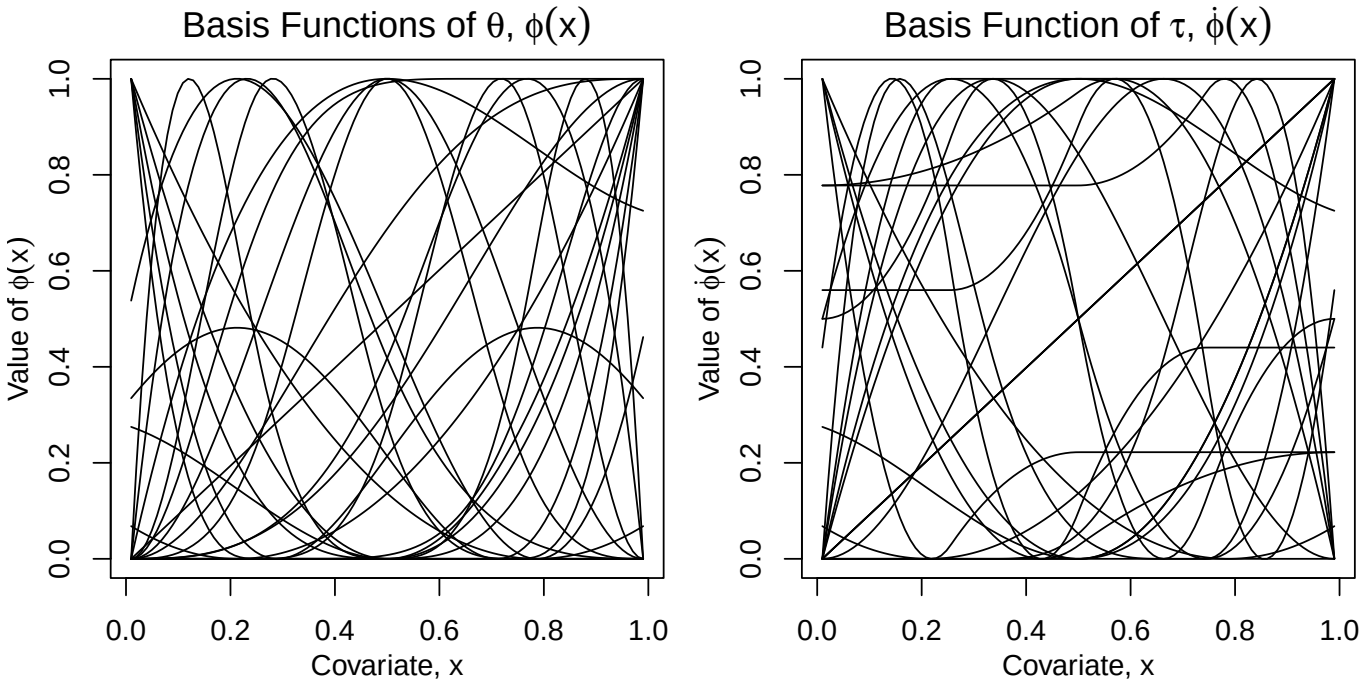

Figure 1: **Basis Functions of Conditional Mean, $\theta(\widetilde{t}, x)$, and Partial Effect $\tau(\widetilde{t}, x) = \partial_t \theta(\widetilde{t}, x)$.**

basis functions but simply that they are sufficiently numerous to approximate a wide array for nonlinearities and interactions between the treatment variable and the covariates.

We show the particular set of basis functions we implement in Figure 1. These bases are a combination of both B-spline bases and orthogonal polynomials in the variable and were selected to account for a wide set of possible nonlinearities in the conditional mean and partial effect, as evident in the density of the bases in the figure. For a precise characterization and discussion of classes of basis functions, see Appendix B.

To generate the set of considered bases, we then interact one of the bases applied to the partialed-out treatment $\widetilde{t}_i$, a potentially different basis of one of the covariates, and a potentially different basis of, potentially, a different covariate. We will use these basis functions to model the function $\theta(t_i, \mathbf{x}_i)$ and then use the partial derivative of these basis functions to construct $\tau(t_i, \mathbf{x}_i)$.

We then implement a *marginal correlation screen* (Fan and Lv, 2008) where, again, restricting ourselves to data in the discovery sample, we calculate the correlation between the partialed-out

outcome, $\widetilde{y}_i$ and each basis. We provide details in Appendix E, but this is the most computationally intensive element of the algorithm; with five covariates, we end up calculating $675,000$ correlations, and with ten covariates, 2.7 million correlations are calculated. We then maintain a set of of these bases with the largest absolute correlation with the partialed-out outcome.[7] We save these selected bases and bring them to the estimation subsample, and will denote the indices of the selected bases as $\mathcal{J}$.[8] We take these maintained bases and bring them to the estimation subsample.

2.2.2 Estimation Subsample: Coefficient and Variance Estimation

We use the estimation subsample to generate coefficients, to estimate the partial effect curve, and variance estimates, to capture our uncertainty in this estimate. We turn to each.

Coefficient Estimation Given the bases from the previous subsample, we assume the model

$$\widetilde{y}_i = \sum_{j \in \mathcal{J}} \phi_j(\widetilde{t}_i, \mathbf{x}_i) c_j + e_i \tag{5}$$

with mean parameters $\{c_j\}_{j \in \mathcal{J}}$. We then use a Bayesian regression model to recover estimates, $\{\widehat{c}_j\}_{j \in \mathcal{J}}$.[9]

We are not interested in modeling $\theta(\widetilde{t}_i, \mathbf{x}_i)$ but $\tau(\widetilde{t}_i, \mathbf{x}_i)$, its partial derivative with respect to the treatment. We have modeled $\theta(\widetilde{t}_i, \mathbf{x}_i)$ in terms of basis functions which are differentiable in the treatment,

$$\dot{\phi}_j(\widetilde{t}_i, \mathbf{x}_i) = \frac{\partial}{\partial t} \phi_j(t, \mathbf{x}_i) \Big|_{t=\widetilde{t}_i} \tag{6}$$

which allows us to generate the partial effect function

$$\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i) = \sum_{j \in \mathcal{J}} \dot{\phi}_j(\widetilde{t}_i, \mathbf{x}_i) \widehat{c}_j. \tag{7}$$

[7] We maintain a proportion of bases growing in sample size, but for sample sizes of $\{100, 250, 500, 1000, 10000\}$ we maintain $25, 63, 125, 250$ and 731 bases. See Appendix E for details.

[8] Importantly, we save these bases at *each* iteration of repeated cross-fitting algorithm, so the maintained bases vary over the course of the entire estimation process.

[9] We use a version of the Bayesian LASSOplus model of Ratkovic and Tingley (2017); see Appendix F

Variance Estimation We turn next to constructing an uncertainty band around our estimated partial effect curve. We formalize below, but the band is constructed around the estimated curve and is designed to inform the researcher on the likely location and characteristics of the true curve. Specifically, we produce a confidence band with the *average coverage* property that the $100\times(1-\alpha)\%$ curve will contain the true curve at $100 \times (1-\alpha)\%$. Doing so allows the researcher to explore the curve and band visually, with confidence that the band will contain the true curve over some proportion of the data.[10]. This band has the nice property that it will contain the true curve at a high percentage of the observed data. It is also narrow enough for applied work, but with provable average coverage properties. Formal derivations of this average coverage can be found in Appendix G.

Constructing the band requires accounting for two separate forms of error: sampling error and misspecification error. The first error captures sample-specific fluctuations of the estimate, and this is the type accounted for in most methods. Importantly, this type of error goes to zero as sample size increases, since more data means our estimate gets more and more precise. The second form of error, *misspecification error*, has been largely ignored. This is the sort of error that does *not* go away in sample size, meaning as we get more and more data, the estimate converges but to the wrong function.

To illustrate this distinction, denote as $\widetilde{\tau}(\widetilde{t}_i, \mathbf{x}_i)$ the limit of our estimator as the sample size grows, i.e.

$$\widetilde{\tau}(\widetilde{t}_i, \mathbf{x}_i) = \lim_{n\to\infty} \widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i). \tag{8}$$

[10]Rather than relying on claims across repeated samples, we follow Nychka (1988) (see also Wasserman (2006) ch. 5.8) and consider *average coverage*, which is the proportion of the sample over which the confidence band contains the true value over the observed sample. A valid band with this property can be written as:
$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left\{\tau(\widetilde{t}_i, \mathbf{x}_i) \in \widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i) \pm \widehat{C}_{1-\alpha/2}\sqrt{\widehat{\mathbb{E}}\left\{\left(\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i) - \tau(\widetilde{t}_i, \mathbf{x}_i)\right)^2\right\}}\right\} \geq 1-\alpha$

In this setting, then, we can decompose the approximation error into sampling error and misspecification error, as

$$\underbrace{\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i) - \tau(\widetilde{t}_i, \mathbf{x}_i)}_{\text{Approximation Error}} = \underbrace{\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i) - \widetilde{\tau}(\widetilde{t}_i, \mathbf{x}_i)}_{\text{Sampling Error}} + \underbrace{\widetilde{\tau}(\widetilde{t}_i, \mathbf{x}_i) - \tau(\widetilde{t}_i, \mathbf{x}_i)}_{\text{Misspecification Error}} \tag{9}$$

Considering the squared error at each point gives us

$$\begin{aligned} &\underbrace{\left(\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i) - \tau(\widetilde{t}_i, \mathbf{x}_i)\right)^2}_{\text{Total Variance}} = \\ &\underbrace{\left(\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i) - \widetilde{\tau}(\widetilde{t}_i, \mathbf{x}_i)\right)^2}_{\text{Sampling Variance}} + \underbrace{2\left(\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i) - \widetilde{\tau}(\widetilde{t}_i, \mathbf{x}_i)\right)\left(\widetilde{\tau}(\widetilde{t}_i, \mathbf{x}_i) - \tau(\widetilde{t}_i, \mathbf{x}_i)\right)}_{\text{Cross-Term}} + \underbrace{\left(\widetilde{\tau}(\widetilde{t}_i, \mathbf{x}_i) - \tau(\widetilde{t}_i, \mathbf{x}_i)\right)^2}_{\text{Misspecification Variance}} \end{aligned} \tag{10}$$

from which we will construct our confidence bands.

Estimating the variance of $\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i)$ involves handling three terms. The first, the sampling variance, is the component used to generate the pointwise standard errors returned by most existing methods. These can be recovered through standard regression calculations. Existing methods generally ignore the latter two terms, and we illustrate the implications of doing so below in Section 3.

In handling the final two terms, we need to address both misspecification error and the cross-term. We address misspecification error through modeling the squared residuals, with details in Appendix G. By capturing systematic patterns in the magnitude of the residuals, we can incorporate model misspecification into our variance estimate.

The cross-product term, though, requires a little more finesse, as it cannot be modeled directly. Instead, we turn to a third subsample, the variance subsample, to evaluate our variance estimates and construct our confidence interval. The cross-product term is a product of two error terms, $\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i) - \widetilde{\tau}(\widetilde{t}_i, \mathbf{x}_i)$ and $\widetilde{\tau}(\widetilde{t}_i, \mathbf{x}_i) - \tau(\widetilde{t}_i, \mathbf{x}_i)$. Any variance in the first term arises from variance in the estimation subsample (this section), so we evaluate them on the next subsample, the inference subsample (Section 2.2.3). Given the bases and partialing out done in the discovery subsample,

these two terms will be uncorrelated as they come from the next two subsamples, driving this term to zero. To complete a single fit, we turn next to the inference subsample.

2.2.3 The Inference Subsample: Conformal Inference

We finally turn to the inference subsample in order to generate our estimated critical value, $\widehat{C}_{1-\alpha/2}$, where we utilize *conformal inference* to generate a curve with average coverage(Lei and Wasserman, 2014; Lei et al., 2018). Conformal inference methods give a means to produce a predictive interval, which will contain a future realization of the outcome some controlled proportion of the time, around a single point. Importantly, it does so through utilizing the estimated residuals in order to construct a band, rather than make distributional assumptions the error terms. The MDEI algorithm innovates here by extending this predictive interval, guaranteed to contain future values of y_i with some a controlled probability (say, 90%), to one containing the true partial effect curve, $\tau(\widetilde{t}_i, \mathbf{x}_i)$, at a controlled proportion of the data (say, 90%).

The insight of the conformal approach comes from using estimated residuals to estimate the critical value on a predictive interval. The method is entirely data-driven, and rather remarkably achieves finite-sample coverage rates on predictive intervals.[11]

We extend this band to cover not just predicted values, but the true conditional mean ($\theta(\widetilde{t}_i, \mathbf{x}_i)$) and partial effect ($\tau(\widetilde{t}_i, \mathbf{x}_i)$). This contribution is original to this work. In estimating the critical value, we are not relying on a normal approximation to achieve valid coverage, allowing our bands to reflect the underlying distribution of the data. We show that these bands, while wide at each data point, can be used to recover valid estimates of the partial effect curve and that, when aggregated over the sample, gives estimates of an average effect competitive with existing methods.

Our use of conformal inference methods proceeds in two steps. In the first, we select a value

[11]Interest in the approach is increasing in other domains of interest to social scientists (e.g., Chernozhukov, Wuthrich and Zhu, working; Lei and Candes, 2020).

around $\widehat{\theta}$ denoted $C^{\widehat{\theta}}_{1-\alpha/2}$ such that it will contain the value $\widetilde{y}_i$ with probability $1-\alpha$,

$$\Pr\left(\widetilde{y}_i \in \widehat{\theta}(\widetilde{t}_i, \mathbf{x}_i) \pm \widehat{C}^{\widehat{\theta}}_{1-\alpha/2}\sqrt{\widehat{\mathbb{E}}\left\{\left(\widehat{\theta}(\widetilde{t}_i, \mathbf{x}_i) - \theta(\widetilde{t}_i, \mathbf{x}_i)\right)^2\right\}}\right) = 1-\alpha$$

Note that this is purely a prediction problem, in that the values of $\widetilde{y}_i$ come from the inference subsample, but the point and variance estimates are constructed from the estimation subsample.

This will allow us to construct a band around $\widehat{\theta}$ such that it will contain values of $\widetilde{y}_i$ with probability $1-\alpha$. Instead, though, we are interested in constructing a bound on $\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i)$ that is likely to contain the true $\tau(\widetilde{t}_i, \mathbf{x}_i)$. We show in Appendix G that if we take as our critical value

$$\widehat{C}_{1-\alpha/2} = 1 + \widehat{C}^{\widehat{\theta}}_{1-\alpha/2} \tag{11}$$

then the interval

$$\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i) \pm \widehat{C}_{1-\alpha/2}\sqrt{\widehat{\mathbb{E}}\left\{\left(\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i) - \tau(\widetilde{t}_i, \mathbf{x}_i)\right)^2\right\}}$$

will contain the true $\tau(\widetilde{t}_i, \mathbf{x}_i)$ with probability at least $1-\alpha$ of the time.

2.3 Repeated Cross-Fitting

While asymptotically valid, splitting the data in thirds raises efficiency concerns, as we only use a third of the data in each step, as well as concerns that our results are driven by a particular split of the data into three subsamples. In order to address these concerns, we follow a *repeated cross-fitting* strategy, recently put forth by Chernozhukov et al. (2018).

Addressing the first concern, we implement a *cross-fitting strategy* where, given an discovery/estimation/inference split, we swap the roles of the three such that we can recover a point estimate and confidence band at every datum. By rotating each subsample through each role in the estimation process, we can generate a point estimate and band for the estimated partial effect at every datum.

Addressing the second concern, we implement a *repeated cross-fitting* strategy, where we repeatedly implement our cross-fitting strategy over multiple possible discovery/estimation/inference splits. We then report these aggregated results by simply taking the average of the point estimates and band over all repetitions of the cross-fitting.

While a single cross-fit estimate has the asymptotic properties we desire, the repeated cross fitting strategy increases the accuracy of our estimates. It does so by averaging over the choice of which bases to include, so our results are not be driven by a particular set of selected bases. Averaging over discrete modeling choices, like inclusion or exclusion of bases, leads to predictive gains (Buhlmann and Yu, 2002). Second, doing so reduces any subsample-particular idiosyncracies in our estimation, again increasing the predictive accuracy of our estimates.[12]

2.4 Modeling the Standard Errors as a Guard Against Misspecification

Modeling the standard errors as we do serves as a guard against model misspecification. The rationale can be found in the idea that misspecification in the conditional mean may result in systematic patterns in the residuals (see, e.g., King and Roberts (2015); Ratkovic and Eng (2010)). If the model is misspecified in some manner, we have a second chance to get our intervals correct, through using a nonparametric model of the conditional variance. By combining the split-sample approach with the conformal interval, we are able to guarantee that our band will have average coverage.

While our estimation strategy works hard to find the right model–considering non-linear and interactive effects of potentially a large number of variables–we of course can't be guaranteed that there will not be some model misspecification. But when we miss, our approach inflates our confi-

[12]There is no theoretical guidance on how many repeated cross-fits to implement. We recommend twenty for an initial fit, which is the default of our software, but then moving it to at least one hundred for publication grade results.

dence band so as to maintain average coverage. To see this, recall that we estimate our conditional variance $\widehat{\text{Var}}(\hat{\tau}(\widetilde{t}_i, \mathbf{x}_i))$ from the estimation sub-sample (Section 2.2.2) but evaluate it on the inference subsample (Section 2.2.3), using the data driven conformal approach in the inference subsample to calculate critical values. By modeling the error variance, we are able to recover bands that are robust to model misspecification (see Appendix G). By construction this band guarantees us average coverage, as model misspecification will simply generate wider bands around the misspecified component. Existing methods do not do this and instead use auxiliary assumptions, including asymptotic normality and a properly specified model, to reduce the width of their confidence intervals.

One consequence of our approach is that our intervals will, in general, be wider than the intervals returned by other methods (see Appendix G). We could make these intervals shorter by assuming our model is properly specified, or assuming the errors are normal.[13] Making these strong assumptions produces more narrow intervals, but comes at the cost of being overly precise if one of the assumptions does not hold.

3 Illustrative Simulations

We next move to two simulations illustrating the need and decisions underlying the proposed method. In the first simulation, we show that several existing methods produce inaccurate point estimates and overly narrow uncertainty estimates, even in a relatively simple setting. The point estimates and confidence band from the proposed method have the expected properties. In the second simulation, we consider a complex functional form that our model was not designed to estimate, and we show how the constituent pieces of cross-fitting and conformal inference combine to still return a band with average coverage.

[13]This is what other cutting-edge methods like kernel regularized least squares (Hainmueller and Hazlett, 2013) and generalized random forests (Athey et al., 2019) do.

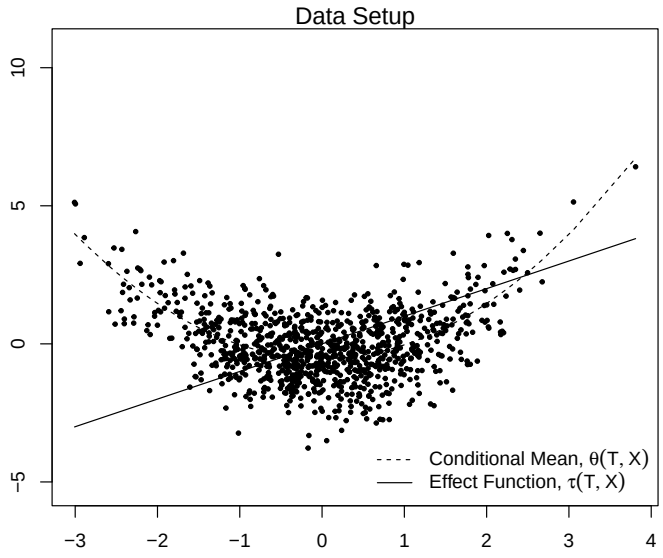

Figure 2: **Data generating process for illustrative simulation.**

3.1 Illustration #1: Existing Methods in a Simple Setting

For this setting, we consider a simple simulation setting in order to evaluate two performance metrics, accuracy and providing an uncertainty interval that captures the distance between the estimated and true curve. Specifically, we generate data as

$$y_i = \frac{1}{2}t_i^2 + e_i; \;\; t_i, e_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \tag{12}$$

where our treatment variable is itself standard normal, but enters the outcome nonlinearly (as a quadratic). In this simple case we have no covariates or interactions. We illustrate the setup in Figure 2, which plots the data around the the true conditional mean ($\theta(\widetilde{t}_i, \mathbf{x}) = t_i^2$, dashed) and the partial effect curve ($\tau(\widetilde{t}_i, \mathbf{x}_i) = 2t_i$, solid).

We compare performance of three different methods that return an estimate of the partial effect curve: the proposed method (MDEI), generalized random forests (GRF, Athey, Tibshirani and Wager, 2019; Wager and Athey, 2017), and Kernel Regularized Least Squares ($KRLS$, Hainmueller and Hazlett, 2013; Mohanty and Shaffer, 2018). GRF and $KRLS$ are prominent and commonly

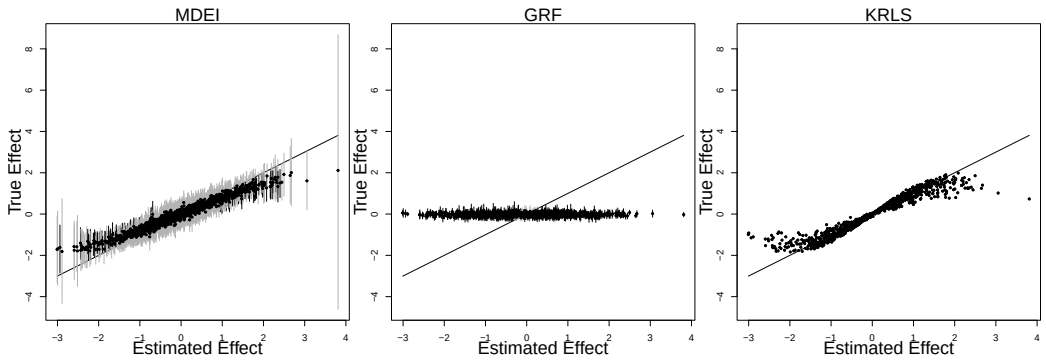

Figure 3: **Effect estimates across MDEI, KRLS and GRF for a quadratic treatment model.** The figure shows the results of each method in estimating the partial effect function given in Figure 2. Black dots are point estimates with gray bars the uncertainty interval at each point. MDEI is the only method of the three to reliably capture both point estimates and uncertainty.

utilized in the machine learning space.[14] Each method is given the outcome, y_i, treatment, t_i, and five noise covariates, also independent standard normal, $\mathbf{x}_i$. We report results for $n = 1000$ in order to give a sense of the large-sample behavior.

We evaluate each method along two dimensions: point estimation and inference. The accuracy of point estimates are simple to assess: in this example, the closer the point estimate (black dots) to the line $2t$, the better the point estimate. The second dimension, inference, asks not how close the point estimates are but whether the uncertainty bands carry some information on the true underlying curve. Is there fidelity between the uncertainty band around our estimated partial effect curve $\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i)$ and the true curve $\tau(\widetilde{t}_i, \mathbf{x}_i)$? We will begin with an intuitive approach, assessing performance graphically.

Figure 2 reports the ability of each method to capture $\tau(\widetilde{t}_i, \mathbf{x}_i)$, reporting results for the proposed method ($MDEI$), generalized random forests (GRF), and kernel regularized least squares ($KRLS$).

[14]For KRLS, we used the software with all tuning parameters set at their defaults. For GRF, we increased the number of trees to $10,000$, as suggested by the documentation, in order to recover accurate estimates of the standard errors over the partial effect curve.

3.1.1 Diagnosing Existing Methods: Point Estimation

Clearly, these existing methods fail in one manner or another. *GRF* returns inaccurate point estimates, wholly missing any curvature in the treatment variable (i.e., any linearity in the partial effect curve). *KRLS* returns accurate point estimates, but its confidence intervals are notably narrow. We turn now to a description of *why* each method performs poorly in this simple simulation, and our proposed fixes for each.

The generalized random forest (*GRF*) provides an estimate of the average partial effect using a forest-based method. The method uses trees constructed from the covariates in order to generate a partialed-out y and t and then the partialed-out outcome is regressed on a partialed-out treatment in the terminal leaf. Results are then aggregated up to a forest.

Mechanically, GRF uses the covariates to fit a tree when all background covariates are noise and then regresses the outcome on the treatment at each leaf. For a simple example, imagine it splits on the first variable at zero, so it regresses the outcome on the treatment for observations with $x_{i1} \geq 0$ and a separate regression for observations with $x_{i1} < 0$. The covariates, though, are pure noise, so it is in effect fitting two lines to a quadratic curve, which the linear terms will miss.[15] GRF estimates the slope at each point, but it can only handle the case where $\tau(\widetilde{t}_i, \mathbf{x}_i)$ is a function of covariates. In the example above, where $\tau(\widetilde{t}_i, \mathbf{x}_i) = t_i$, GRF will miss the partial effect entirely,

[15] Using our notation, generalized random forests are fitting a model of the form

$$y_i = \tau(\mathbf{x}_i)t_i + f(\mathbf{x}_i) + e_i \tag{13}$$

$$t_i = g(\mathbf{x}_i) + v_i \tag{14}$$

where the slope on t_i is allowed to vary in the covariates, parameterized as $\tau(\mathbf{x}_i)$. This model will clearly miss data generated as $y_i = t_i^2 + e_i;\ t_i, e_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ as in our simulation here. The core reason is that this model only captures heterogeneities moderated by a linear treatment variable, rather than a non-linear function of the treatment. Appendix A is provided for readers who wish additional development of these differences

as shown in this figure.

The next method, KRLS, does a better job of recovering an estimated partial effect curve $\tau(\widetilde{t}_i, \mathbf{x}_i)$. KRLS is an example of a nonparametric regression, where it assumes the model

$$y_i = \sum_{p=1}^{P} \phi_p(t_i, \mathbf{x}_i) c_p + e_i$$

where each function ϕ_p is a smooth, nonlinear function of $t_i, \mathbf{x}_i$ and P is some large number, possibly as large or larger than the sample size n. Differences arise in terms of what sorts of basis functions are used and how, precisely, the coefficients are estimated, but the important issue is that these functions are constructed to be differentiable in the treatment.[16]

From a high-level vantage, estimation via KRLS shares some similarity with our estimation strategy. Both are nonparametric regression methods that simply differ on which basis representation is implemented, though the bases are all differentiable in the treatment.[17] In this setting, standard regression calculations can be used to estimate the sampling variance of the coefficients and, hence, of the partial effect curve. The regression approach, given by KRLS and MDEI, captures the curvature in this simple setting accurately.

3.1.2 Diagnosing Existing Methods: Generating Uncertainty Bands

At an intuitive level, we want our estimates of the partial effect curve to be as close as possible to the true curve, and we want our uncertainty estimates to give us a reasonable idea of how far we expect our estimated curve to be from the true curve.

[16] KRLS, in particular, uses *Gaussian radial basis functions*, while we will use interactions among B-splines and orthogonal polynomials, but for the purposes of our method any set of smooth functions that can approximate a wide class of functions will work. See Appendix B for more discussion.

[17] KRLS uses "isotropic" bases in that the bases are constructed from all the covariates $\mathbf{x}_i$, whereas we use "anisotropic" bases, where each basis is a function of only one covariate in $\mathbf{x}_i$, then we interact them. For more on this distinction, see Murphy (2012). We follow our approach for mechanical reasons, in that KRLS requires inverting an $n \times n$ matrix whereas limit ourselves to a few hundred bases.

We construct intervals with the *average coverage* property, such that we can expect that the $(1-\alpha)\times 100\%$ interval will contain the true partial effect curve at $(1-\alpha)\times 100\%$ of the observed data, asymptotically. For assessing an entire curve, we work with this property because it gives an intuitive way of capturing where we suspect the true partial curve may be, given our estimate.

The reasons for the pronounced gap between the confidence intervals and the true partial effect curve returned by KRLS is two-fold. These reasons are not peculiar to the particular method, but instead stems from two problems endemic to many machine learning methods. Importantly, both are addressed through our conformal strategy.

The first reason is the assumption required by existing methods that the model is properly specified. We do not make this assumption, instead using the estimated errors themselves to determine the width of our band. Any misspecification will show up as inflated residuals, relative to the residuals under a properly specified model, and this will just lead us to a wider confidence band.

The second reason is the particular nature and statistical properties of the returned band. For the sake of this point, assume that the model is properly specified. Even if properly specified, existing methods generate what are referred to as *pointwise confidence intervals*. These have a particular property that, given a particular point $(t_i, \mathbf{x}_i)$, then the interval

$$\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i) \pm C_{1-\alpha/2}\sqrt{\widehat{\text{Var}}(\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i))}$$

will contain the true value $\tau(\widetilde{t}_i, \mathbf{x}_i)$ at the point $(t_i, \mathbf{x}_i)$ over $(1-\alpha)\times 100\%$ of repeated samples.[18]

Pointwise confidence interval at every point do not allow for any claims about the whole curve.

[18]Note that we will not achieve average coverage over every single subset of the band. For example, in Figure 3, we achieve coverage near 100% in the middle of the data but lower coverage towards the edge. Different parts of the data and model will likely have different average coverage, but it will achieve the desired proportion over the whole of the data. See Nychka (1988) for more.

To see, imagine the problem from a multiple testing perspective. A 90% confidence interval at every single point is not the same as a 90% band over *all* points. It is likely too narrow; in this simulation, the 90% band for KRLS and GRF contain the true partial effect curve at only 22.3% and 10.4% of the observed data, respectively. In Figure 3, the confidence intervals for GRF and KRLS are clearly concentrating on the wrong partial effect function and, in this example, too small to be visible to the eye.

We correct these issues endemic to pointwise curves and produce informative graphical displays using conformal inference. We turn next to a more complete development of how our strategies combine via a second illustrative simulation.

3.2 Illustrative Simulation #2: Combining Repeated Cross-Fitting and Conformal Methods

We next illustrate the role of repeated cross-fitting and conformal inference in achieving average coverage. Each has a role in achieving coverage: repeated cross-fitting in guarding against over-fitting, and conformal inference in using the observed data to determine the width of our band. As we show next, the two work in conjunction to achieve average coverage.

For this simulation, we again draw five covariates from a standard multivariate normal equicorrelated at 0.5. The first two covariates are used in the model and the last three are noise. From the first covariate, $\mathbf{x}_{i1}$, we generate a new variable $s_i = \text{sgn}(\mathbf{x}_{i1}) \in \pm 1$. This sign function, a discontinuous function of a continuous covariate, will serve to govern the effect heterogeneity: the impact of the treatment on the outcome will vary with whether this first variable is positive or negative. The outcome and the treatment are generated as:

$$t_i = \frac{(\mathbf{x}_{i2} - 1)^2}{4} + u_i;\ u_i \sim \mathcal{N}(0, 1) \tag{15}$$

$$y_i = 2 s_i t_i^2 + \frac{(\mathbf{x}_{i2} - 1)^2}{4} + v\epsilon_i;\ \epsilon_i \sim \mathcal{N}\left(0, \frac{1}{1 + \mathbf{x}_{i2}^2}\right). \tag{16}$$

where v is a scalar selected so that the true $R^2 = 0.5$. This target function is $\tau(\widetilde{t}_i, \mathbf{x}_i) = 4s_i t_i$ for which we want to use our interval to conduct inference. We vary the sample size, $n \in \{250, 500, 1000, 2500, 5000, 10000\}$ and simulations were run 500 times each.

Importantly, our model was not designed to estimate this sort of function: we assume the conditional mean is a smooth function that can be represented by some combination of interacted functions from Figure 1.[19] The function in this simulation is outside of this space due to the discrete break in the first covariate. As a consequence, we developed this to be a challenging case.

Our first step is to vary whether or not we implement our repeated cross-fitting and conformal strategy. When not implementing the repeated cross-fitting strategy, we simply conduct the entire estimation process on the whole of the data. When not implementing the conformal strategy, we simply take our critical value as 1.64, generating a pointwise interval.

Results appear in Figure 4. For each run of the simulation, we calculate a 90% band, and assess at what proportion of the data the true partial effect curve is contained in the constructed band. The y-axis presents average coverage, with sample size on the x-axis. The solid horizontal line at 90% is the expected coverage. Each of the four lines corresponds with the four possible settings for how we construct our confidence bands: with neither a conformal critical value nor repeated cross-fitting; with either repeated cross-fitting or a conformal critical value of 1.64; and both the conformal and repeated cross-fitting strategy.

The figure shows that the estimate without cross-fitting and conformal, labeled "neither" on the graph, will be quite narrow and hence not actually containing the true curve at the majority of the data.[20] This band gets worse in sample size, since its overfitting is causing it to converge

[19] See appendix C regarding model spaces.

[20] If the model is correctly specified and hence the third term in the variance decomposition goes to 0, then the MDEI algorithm without split-sample/repeated cross-fitting and without a conformal critical value will produce a pointwise interval.

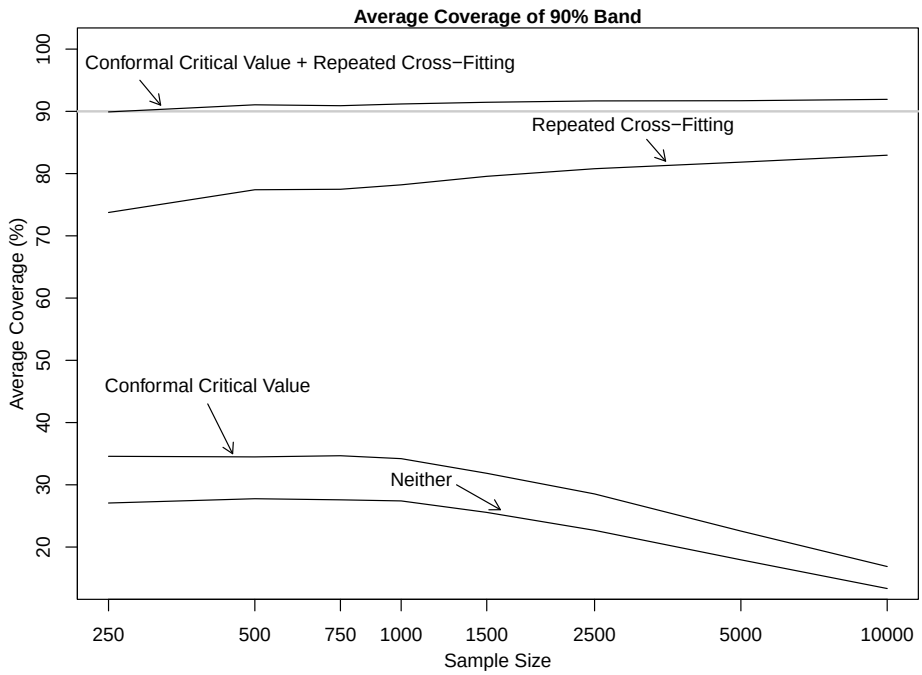

Figure 4: **Average coverage of 90% band.** Each lines corresponds with the four possible settings for our confidence bands. The pointwise interval will be quite narrow, not actually containing the true curve in the majority of the data. Repeated cross-fitting increases coverage, but only when combined with a conformal strategy does average coverage approach nominal.

on the wrong function. The conformal critical value helps somewhat, but it still results in low average coverage. Using repeated cross-fitting with the critical value of 1.64 helps get closer to 90% coverage, but it is only when both are combined that we see average coverage achieved.

Next, we analyze the results using the full MDEI approach for a single draw of the simulation data at two different sample sizes.[21] As a reminder, the function we are trying to fit is outside of the class of models we can handle due to the discontinuity. The result of this discontinuity is that our misspecification error will be higher the further away we get from $t_i = 0$ because this is where the largest gap due to the discontinuity will be. Due to how we construct our confidence bands, this means they should be wider in this region.[22] And furthermore, unlike existing methods, they

[21]Estimation with MDEI is done in **R** through one line of code, s1 ← sparseregTE(Y=y, treat=treat, X=X) where X is simply a matrix of pre-treatment covariates. No additional inputs are required from the user.

[22]Our particular manifestation of heteroskedasticity in this simulation–where variance is smaller at the extremes–will cut against this making simulation results consistent with this observation even more striking.

should not appreciably tighten if we increase the sample size.

Figure 5 displays the results for a case with a sample size of 5000 in the top row and 50000 in the second row. We present results for $s_i = 1$; they are qualitatively similar to $s_i = -1$. The partial effect curve $(\tau(\widetilde{t}_i, \mathbf{x}_i))$ is plotted against its estimate and interval in the left hand pane. Intervals that contain the truth are in grey, and those that are not are in black. The 90% band returned MDEI covers the true value at 92.78% of the data and 89.2%, for the 5000 and 50000 sample sizes respectfully.[23] The middle panes plot the absolute approximation error for each point. As expected the average approximation error increases the further away from $t_i = 0$ we get.

The right hand pane presents the width of the confidence bands. Due to our approach, these bands should be wider in the presence of misspecification error that becomes more extreme the further away from $t = 0$ we get. Indeed, looking at the data and a loess line to illustrate the pattern, we see this pattern. This is despite the fact that the simulations heteroskedastictity shrinks the variance at the extremes. Importantly, these wider bands at the extremes do not radically shrink in the n=50000 case.

4 Applied Example

Bechtel and Hainmueller (2011) explore the impact of an effective policy response to a natural disaster in Germany. They estimate the effect of the government's successful response to the 2002 flooding of the Elbe River on support for the incumbent party, the Social Democrats, in the 2002 federal elections. Using a difference-in-difference design with a regression specification, the authors

[23] In the n=5000 case, for KRLS and GRF, those numbers are 21.7% and 13.6%. The root-mean square error on $\tau(\widetilde{t}_i, \mathbf{x}_i)$ across the methods reveal a similar pattern at n=5000 (MDEI: 3.46, KRLS: 5.81, GRF: 7.97). GRF uses a split sample approach, while KRLS does not, but still achieves a higher error and lower coverage because, as mentioned above, GRF cannot capture curvature in the treatment variable well. KRLS cannot be run at a sample size of 50000, so we omit comparisons at this sample size.

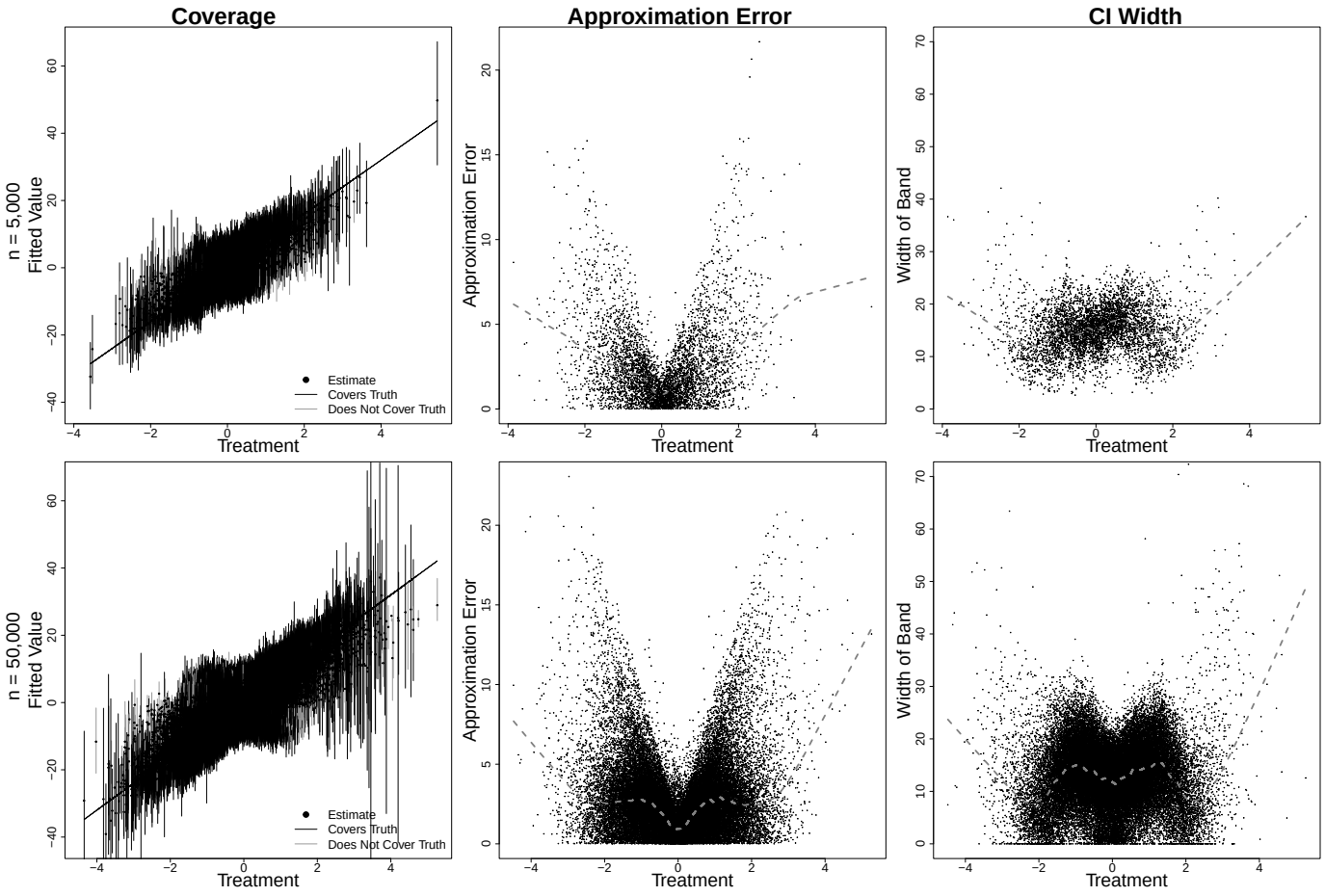

Figure 5: **Partial effect estimates, approximation error, and confidence band width for 5000 (top) and 50000 (bottom) simulations.** Left column presents the partial effect estimates. The dashed line represents the true partial effect, vertical lines represent the 90% uncertainty bands. Middle column represents the approximation error at each point, with a loess line describing the pattern. Right column presents the band width at each point in the curve with a loess line describing the pattern.

estimate an impact of approximately 7 percentage points on the Social Democrats' vote share. Here, the unit of analysis is the district, the outcome is the change in voteshare for the Social Democrats, the treatment variable is whether the region was flooded, and the controls include a battery of covariates that adjust for sociodemographic and economic factors (see Bechtel and Hainmueller, 2011, Table 1, pg. 857).

In a pure difference-in-difference design, the authors could simply estimate the effect as the change in voteshare before and after the flooding of the Elbe, between the flooded and unflooded

districts. Covariates can then adjust for district-level confounders not eliminated through the randomness in the flooding (Sec 5.2.1 Angrist and Pischke, 2009). The authors implement estimate the effect in a regression framework, combining the standard difference-in-difference specification with a smoothing spline in distance from the Elbe with a set of linear, additive controls.

The validity of the results, then, are dependent on a reasonable control specification. To illustrate, we consider the average partial effect on the treated, i.e. the impact of flooding on those districts that were flooded. We start by analyzing this situation, with a binary treatment variable (was a district flooded or not?) in order to build faith in the method (see Appendix I for estimation details in this binary setting).

MDEI returns both a point and uncertainty estimate for each datum, but these can be aggregated up over simply the flooded districts.[24] Estimates of the average effect on change in vote using Bechtel and Hainmueller's data appear in Table 1. The first row contains the results using the control set in the original work. The results from MDEI appear in the second line, using the same control set, outcome, and treatment. To calculate the effect, we took the average effect on all flooded districts, averaged their variances, and used the critical value returned by the method. We find a point estimate lower than the original analysis, though still significant. We find that the discrepancy between the original results and MDEI is likely due to covariate imbalance between treated and untreated regions. If we trim districts further from the Elbe than the treated districts and then run Bechtel and Hainmueller's specification, we recover an estimate that is much closer to that from MDEI. For further verification, we compare the results to generalized random forests (GRF) as well

[24]Formally, let $\mathcal{F}$ denote the flooded districts and $N_{\mathcal{F}}$ the number of flooded districts. The average effect on flooded districts ($\widehat{\tau}_{\mathcal{F}}$) and its standard error ($\widehat{\sigma}_{\mathcal{F}}$) are calculated as

$$\widehat{\tau}_{\mathcal{F}} = \frac{1}{N_{\mathcal{F}}} \sum_{i \in \mathcal{F}} \widehat{\tau}(1, \mathbf{x}_i); \;\; \widehat{\sigma}_{\mathcal{F}} = \frac{1}{N_{\mathcal{F}}} \sqrt{\sum_{i \in \mathcal{F}} \widehat{\sigma}^2(1, \mathbf{x}_i)}$$

	Point Estimate	95% CI
Original Regression	6.91	(5.43, 8.40)
Trimmed Regression	4.87	(2.96, 6.78)
MDEI	4.89	(1.77, 8.01)
GRF	4.56	(3.65, 5.46)
GAM	4.55	(3.27, 5.82)

Table 1: **Estimates Across Model Specifications.** Rows contain the estimated effect on the Social Democrat's voteshare in flooded regions from the original specification, a trimmed regression, *GRF*, and a *GAM* using a trimmed regression but adding a smoothing spline in distance from the river.

as the authors' original specification on the trimmed data, but using a smoothing spline in distance from the Elbe (GAM). We see that all of the methods besides the original regression agree on the magnitude of the effect.

The reason for the improved performance is implicit in the method. The estimand for the difference-in-difference design is the average effect on the treated districts. The difference-in-difference regression coefficient, though, is a weighted average of the difference between treated and untreated units. If the untreated units are not directly comparable with the treated units, the coefficient may be biased. This is what we see here. In contrast, MDEI returns an estimated effect for each point, and then we aggregate only over the treated units in order to estimate the treatment effect on the treated. Doing so reduces concerns over imbalance. Though we use the full data to fit the model and generate confidence bands, we are only evaluating the treatment effect on those observations that were in fact treated.

We next estimate the effect of a continuous treatment on an outcome. Bechtel and Hainmueller argue that the effect of policy response on voteshare decays as the distance from the Elbe increases for regions in which there were flooded districts, which they argue is further evidence that the discovered effect is attributable to disaster response. We reevaluate both claims and present results in Figure 6. We begin with their analysis (see Figure 5 in the original paper),[25] which we present in

[25]We combine both halves of their Figure 5 into one plot for parsimony.

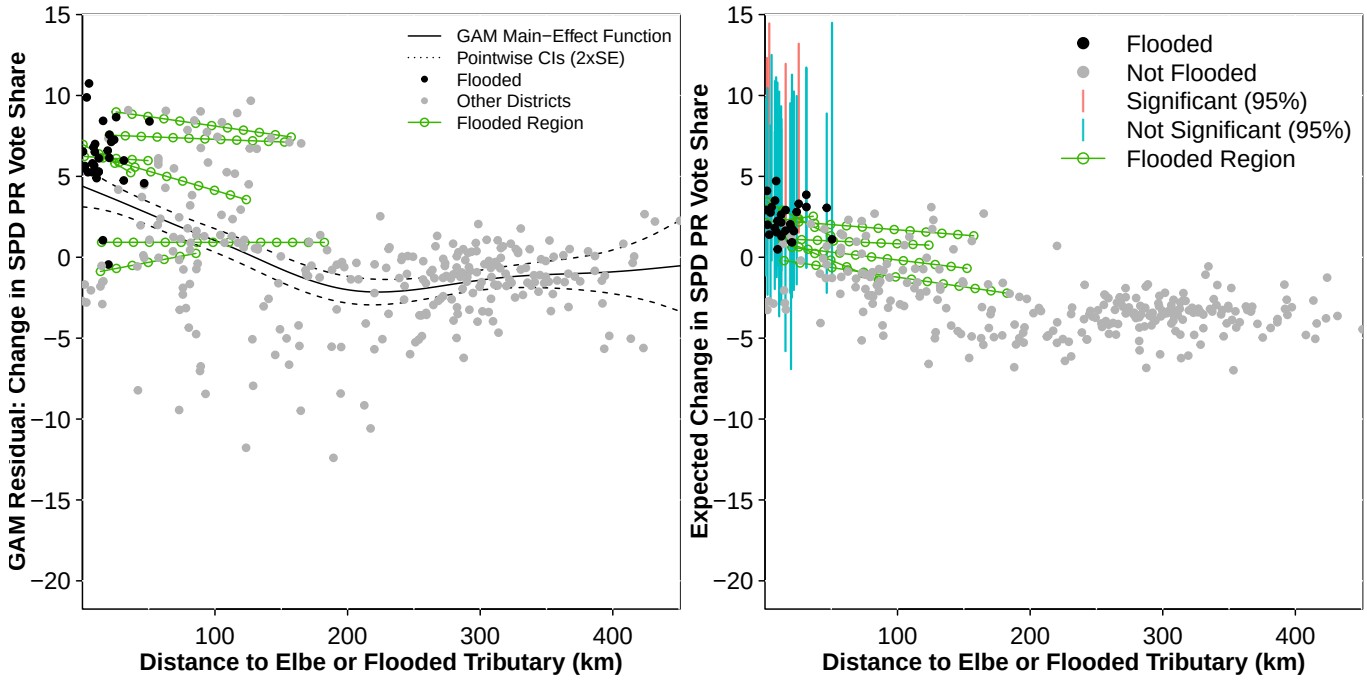

Figure 6: **The Effect of Distance to Elbe on Vote Share.** The left hand figure presents the estimated effect of distance to the Elbe on voteshare using the specification from Bechtel and Hainmueller. Analyzing residuals, flooded districts (solid black dots) are systematically above the trend. Regions with flooding exhibit a negative trend, suggesting that the effect is due to flooding and not some other confounding event. MDEI, on the right, is able to recover similar results as in the original paper, but in one step and with uncertainty bands.

the left-hand panel of Figure 6. The authors fit a smoothing spline (GAM), smooth in distance to the Elbe with the same set of linear controls included as before. Examining the residuals, flooded districts (solid black dots) are systematically above the trend, suggesting that these observations are systematically high. Then, the authors fit lines to the residuals by region containing districts that flooded, which we present as the green lines. The slopes of four of these lines are negative, which they argue suggests the effect is due to flooding and not some other confounding variable or concurrent political event.

In the right panel, we present the fitted values from MDEI where we take each district's distance to the Elbe as the treatment. We include the authors' original covariates and add in controls for whether the district flooded and whether the district is in a region that had at least one flooded

district. We find similar trends in regions where districts were flooded, but we find them in the fitted values rather than the residuals. The original work analyzed residuals, after taking out a smooth trend in distance and additive covariates. The righthand panel, using MDEI, uncovers the same effects through the model and the covariates.

Exploring the fitted values is preferable, because we can attribute their values to observed covariates, as compared to estimating with residuals which are, by design, noisy. It also allows us to estimate and analyze effects in one step, looking at fitted values and bands, rather than the two-step process of estimating fitted values and looking at the residual. Using our method, we find a similar pattern: with flooded districts are systematically above zero, meaning the vote share for the Social Democrats went up, and there is less variance in the segments fit to regions where there was flooding than to unflooded regions.

Although we find a similar pattern in the data as Bechtel and Hainmueller, we now want to know whether it is distance from the Elbe, or simply having been flooded, that is driving the estimated effect on voteshare. Figure 7 presents the estimated effect of distance on voteshare at each point, with flooded districts black and non-flooded grey. After adjusting for the other covariates, we find no effect at any observation. Our analysis seems to suggest that the relationship between distance to the Elbe and voteshare is null, after adjusting for flooding and other covariates. The estimated effect seems attributable to whether the district was flooded, and not to its distance from the Elbe.

Our reanalysis has recovered the central finding of Bechtel and Hainmueller, that flooded districts rewarded the Social Democrats. At the same time, we found the effect to be somewhat overstated likely due to the inclusion of non-flooded districts that were not directly comparable to the flooded districts. We then found evidence that the result is being driven by whether a district is flooded, and not its distance from the Elbe. Throughout we are able to entertain non-linear effects as well as recover uncertainty estimates.

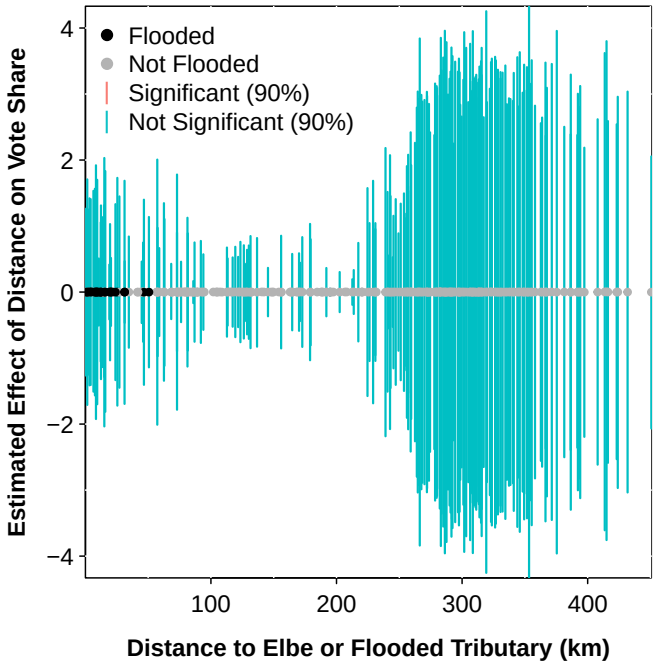

Figure 7: **The Effect of Distance to Elbe on Vote Share.** The point estimate and confidence interval of the marginal effect on distance to the Elbe on support for Social Democrats, by observation. Point estimates for districts in flooded regions are in black, the rest gray. Distance has no discernible effect on support for Social Democrats.

5 Conclusion

A central challenge in regression analysis is correctly modeling how a treatment variable impacts an outcome. Is the effect non-linear? Does it depend on the values of other variables, or a combination of both? Traditional regression models grow increasingly unhelpful given these challenges, especially as the number of variables and potential non-linear relationships increases. We introduce an estimation process that allows for the seminonparametric estimation of a partial effect *and* robust uncertainty estimates.

We hone in on the type of inference that is appropriate when estimating nonlinear relationships when we do not ex ante specify specific nonlinear relationships. The method we propose builds on recent work involving iterated cross-fitting and conformal inference. Simulation evidence shows that the proposed method performs very well.

While we dramatically reduce reliance on ex ante modelling choices, we do of course retain other assumptions required for making causal claims (e.g., no omitted confounders). The approach presented in this paper also does not deal with other challenges to causal inference (e.g., improper confounding strategies such as controlling for post-treatment variables or certain types of pre-treatment variables (Acharya, Blackwell and Sen, 2016; Morgan and Winship, 2014; Glynn and Kashin, 2018), which are research questions that precede the choice of model. In a separate paper we discuss how to extend our framework to the instrumental variables and causal mediation frameworks.

References

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2016. "Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects." *American Political Science Review* 110(3).

Angrist, Joshua D. and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Aronow, Peter and Benjamin Miller. 2018. *Agnostic Statistics*. Cambridge University Press.

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

Athey, Susan, Julie Tibshirani and Stefan Wager. 2019. "Generalized Random Forests." *Annals of Statistics* . Forthcoming.

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

Bechtel, Michael M. and Jens Hainmueller. 2011. "How Lasting Is Voter Gratitude? An Analysis of the Short- and Long-Term Electoral Returns to Beneficial Policy." *American Journal of Political Science* 55(4):851–867.

Beck, Nathaniel, Gary King and Langche Zeng. 2000. "Improving Quantitative Studies of International Conflict: A Conjecture." *American Political Science Review* 94(1):21–35.

Breiman, Leo. 2001. "Random forests." *Machine learning* 45(1):5–32.

Buhlmann, Peter and Bin Yu. 2002. "Analyzing Bagging." *Annals of Statistics* 30(4):926–961.

Cattaneo, Matias D., Max H. Farrell and Yingjie Feng. Forthcoming. "Large Sample Properties of Partitioning-Based Series Estimators." *Annals of Statistics* .

Chernozhukov, Victor, Denis Chetverikov, Esther Demirer, Mertand Duflo, Christian Hansen, Whitney Newey and James Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *The Econometrics Journal* .

Chernozhukov, Victor, Kaspar Wuthrich and Yinchu Zhu. working. "An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls." arXiv:1712.09089.

Fan, Jianqing and Jinchi Lv. 2008. "Sure independence screening for ultrahigh dimensional feature space." *Journal of the Royal Statistical Society: Series B* 70:849–911.

Glynn, Adam N and Konstantin Kashin. 2018. "Front-door versus back-door adjustment with unmeasured confounding: Bias Formulas for front-door and hybrid adjustments with application to a job training program." *Journal of the American Statistical Association* 113(523):1040–1049.

Green, Donald P. and Holger L. Kern. 2012. "Modeling heterogeneous treatment effects in survey experiments with Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76:491–511.

Grimmer, Justin, Solomon Messing and Sean J Westwood. 2017. "Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods." *Political Analysis* 25(4):1–22.

Hainmueller, Jens and Chad Hazlett. 2013. "Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach." *Political Analysis* 22(2):143–168.

Hill, Jennifer, Christopher Weiss and Fuhua Zhai. 2011. "Challenges With Propensity Score Strategies in a High-Dimensional Setting and a Potential Alternative." *Multivariate Behavioral Research* 46(3):477–513.

Ho, Daniel E., Kosuke Imai, Gary King and Elizabeth A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15(3):199–236.

Holland, Paul W. 1986. "Statistics and Causal Inference (with Discussion)." *Journal of the American Statistical Association* 81:945–960.

Imbens, Guido W and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

King, Gary and Margaret E. Roberts. 2015. "How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It." *Political Analysis* 23(2):159–178.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *The Quarterly Journal of Economics* 133(1):237–293.

Lei, Jing and Larry Wasserman. 2014. "Distribution-free prediction bands for nonparametric regression." *Journal of the Royal Statistical Society (Series B)* 76(1):71–96.

Lei, Jing, Max GâSell, Alessandro Rinaldo, Ryan J. Tibshirani and Larry Wasserman. 2018. "Distribution-Free Predictive Inference for Regression." *Journal of the American Statistical Association* 113(523):1094â1111.

Lei, Lihua and Emmanuel J. Candes. 2020. "Conformal Inference of Counterfactuals and Individual Treatment Effects." *working paper* .

Mohanty, Pete and Robert Shaffer. 2018. "bigKRLS: Optimized Kernel Regularized Least Squares." *Political Analysis* 27(2):127–144.

Montgomery, Jacob M and Santiago Olivella. 2018. "Tree-Based Models for Political Science Data." *American Journal of Political Science* 62(3):729–744.

Morgan, Stephen L and Christopher Winship. 2014. *Counterfactuals and causal inference*. Cambridge University Press.

Murphy, Kevin P. 2012. *Machine learning: a probabilistic perspective*. MIT press.

Newey, Whitney K and Daniel McFadden. 1994. "Large sample estimation and hypothesis testing." *Handbook of econometrics* 4:2111–2245.

Nychka, Douglas. 1988. "Bayesian Confidence Intervals for Smoothing Splines." *Journal of the American Statistical Association* 83:1134–1143.

Ratkovic, Marc and Dustin Tingley. 2017. "Sparse Estimation and Uncertainty with Application to Subgroup Analysis." *Political Analysis* 1(25):1–40.

Ratkovic, Marc T and Kevin H Eng. 2010. "Finding jumps in otherwise smooth curves: Identifying critical events in political processes." *Political Analysis* 18(1):57–77.

Robinson, Peter. 1988. "Root-N Consistent Semiparametric Regression." *Econometrica* 56(4):931–954.

Samii, Cyrus. 2019. "Conformal Inference Tutorial." `https://cdsamii.github.io/cds-demos/conformal/conformal-tutorial.html`. Accessed: 2021-12-11.

Wager, Stefan and Susan Athey. 2017. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association* .

Wasserman, Larry. 2006. *All of Nonparametric Statistics*. Springer Texts in Statistics Springer.

Wooldridge, Jeffrey M. 2002. *Economic Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Online Appendix for "Estimation and Inference on Nonlinear and Heterogeneous Effects"

The online appendix for this article contains several sections. Appendix A provides some background information on parametric and seminonparametrics regression models. Appendix B gives an introduction to using basis functions for regression modelling. Appendix C discusses the minimal functional form assumptions made for our model to return consistent estimates. Appendix D discusses pointwise bands and contrasts them to the uniform bands we use in the paper for uncertainty estimation. Appendix E gives an exposition of the MDEI algorithm. Appendix F discusses the specific sparse regression model employed in our estimation steps. Appendix G derives for our variance estimation. Appendix H gives extensive performance simulation evidence leveraging a variety of different data generating processes and comparing MDEI to other relevant methodologies. Finally, Appendix I discusses the case of a binary or categorical treatment variables rather than the continuous treatment variable context.

Appendix

Table of Contents

A Regression Model Background

In this appendix we provide some background on different types of regression models and work up to our proposed approach for calculating point estimates. Starting from the simple linear regression, we progressively relax more and more assumptions until we end up with a specification in which the impact of the treatment, and both the role of the background covariates *and* how they (may) interact with treatment variable, is learned from the data (rather than assumed ex ante). As the models grow more complex, methods for both point estimation and inference require more nuance.

Existing Models Most published work utilizes some version of the simple regression model above: the treatment is entered linearly, is simply included additively along with additional covariates in the outcome and the treatment assignment mechanism is not modeled. This fails to capture what we are interested in modeling, which is the effect of a fluctuation of t_i on y_i, at some particular point $(t_i, \mathbf{x}_i)$.

As a first attempt, we may choose to maintain linearity assumptions, modeling the outcome and treatment with a regression model,

$$\begin{aligned} y_i &= \theta t_i + \mathbf{x}_i^\top \gamma + \epsilon_i, \;\; \mathbb{E}(\epsilon_i | t_i, \mathbf{x}_i) = 0 \\ t_i &= \mathbf{x}_i^\top \beta + u_i, \;\; \mathbb{E}(u_i | \mathbf{x}_i) = 0, \end{aligned} \tag{17}$$

which we will refer to as the outcome model and treatment model, respectively. With standard assumptions such as no omitted confounders and no heterogeneity in the treatment effect, we can interpret θ as a causal effect; absent these assumptions, it is simply an average slope on the treatment (Aronow and Samii, 2016). While the treatment model in this case is not necessary, in more advanced settings, modeling the effect of a one-unit move in the treatment on the outcome will require a flexible model of the treatment variable itself. If we are willing to make narrow assumptions such as linearity and homogeneity in the partial effect, an outcome model alone will do; as we want extend this model into more general settings such as the partially linear model, we will considering the treatment assignment model as well.

A more challenging case emerges when the treatment connects nonlinearly with the covariates through some known function g:

$$\begin{aligned} y_i &= \theta t_i + \mathbf{x}_i^\top \gamma + \epsilon_i, \;\; \mathbb{E}(\epsilon_i | t_i, \mathbf{x}_i) = 0 \\ t_i &= g(\mathbf{x}_i^\top \beta) + u, \;\; \mathbb{E}(u_i | \mathbf{x}_i) = 0. \end{aligned} \tag{18}$$

Then, we can recover a consistent estimate on θ under the assumption that the model is correct: the covariates enter the outcome model linearly and enter the treatment model linearly under link g.[26]

We model the treatment, either parametrically or nonparametrically, for reasons that arise from the definition of the partial effect as the effect of a one unit move in the treatment on the outcome after controlling for covariates. We can work under sets of assumptions, such as linear additivity and homogeneity in the partial effect, that can allow us to estimate the partial effect without modeling the treatment. With a more general setting, partialing the covariates out of the treatment will eliminate the effect of confounders, and also offer efficiency gains, as our estimate at each point does not depend on a function of only the covariates.

[26] In the circumscribed case of a binary treatment $T \in \{0, 1\}$ and $g()$ a logit or probit, a well-developed literature exists using matching and weighting methods. Imbens and Rubin (2015) provide an overview and Sekhon (2009); Ho et al. (2007) provide excellent introductions for political scientists. Appendix I connects our approach to the binary treatment setting.

However, we may not believe that the covariates enter the model in a linear or additive fashion, or that we even know the function g. In this case, we may turn to a model of the form

$$\begin{aligned} y_i &= \theta t_i + f(\mathbf{x}_i) + \epsilon, \;\; \mathbb{E}(\epsilon_i | t_i, \mathbf{x}_i) = 0 \\ t_i &= g(\mathbf{x}_i) + u_i, \;\; \mathbb{E}(u_i | \mathbf{x}_i) = 0. \end{aligned} \tag{19}$$

We now assume that the function f, g are unspecified and must be learned from the data. This is referred to as a *partially linear model* (Chernozhukov et al., 2018; Hardle and Stoker, 1989; Härdle et al., 2012; Robinson, 1988), since it is linear in t_i but nonparametric in the remainder.

The partially linear model improves on the standard practice captured in Model 18, because it allows for the fact that the confounding variables may not be linear. However, the partially linear model still assumes that the treatment enters the outcome model linearly, after adjusting for the covariates nonparametrically. We can relax this assumption, generating a type of *generalized additive model* (Wahba, 1990; Hastie and Tibshirani, 1990; Wood, 2006; Beck and Jackman, 1998), where we replace θt_i with $\theta(\widetilde{t}_i)$, a smooth function of the treatment. Doing so generates the model

$$\begin{aligned} &y_i = \theta(\widetilde{t}_i) + f(\mathbf{x}_i) + \epsilon, \mathbb{E}(\epsilon_i | \mathbf{x}_i, t_i) = 0; \\ &t_i = g(\mathbf{x}_i) + u_i, \mathbb{E}(u_i | \mathbf{x}_i) = 0; \;\; f, g, \theta \text{ unknown} \end{aligned} \tag{20}$$

These GAMs, also known as smoothing spline models, are used in political science and other social sciences (e.g., Beck and Jackman, 1998; Andersen, 2009; Imai, Keele and Tingley, 2010; Carter and Signorino, 2010; Kropko and Harden, 2020).

Worth noting is why we model the treatment in this GAM setting. Of course, the researcher may not, but it will come at some cost. Imagine instead we simply model the outcome, giving us a reduced-form version

$$y_i = \theta(g(\mathbf{x}_i) + u_i) + f(\mathbf{x}_i) + \epsilon, \mathbb{E}(\epsilon_i | \mathbf{x}_i, t_i) = 0; \tag{21}$$

where we have simply substituted our treatment model into the outcome. Our interest is in modeling the effect of a movement in the treatment that cannot be explained by the covariates on the outcome, i.e. of u_i on y_i. Partialing out the covariates from the treatment and the outcome improves the efficiency of our estimate, see Robinson (1988) for foundational work in the area.

If we allow θ to be a smooth function of the treatment, estimation can occur following the same cross-fitting described earlier. Rather than simply regressing $\widetilde{y}_i$ on $\widetilde{t}_i$, we could instead model the relationship using a smoothing spline. Under the assumption that $\theta(\widetilde{t}_i)$ is indeed smooth, the errors are of equal variance, and there are no treatment/covariate interactions, well-established theory and standard software can return a confidence band with average coverage (see Nychka (1988, ch. 4) and the associated **R** package `mgcv` (Wood, 2006)). If the researcher is confident that these assumptions hold, this is an appropriate method.

The MDEI Model We wish to allow for the background covariate specification to be learned from the data *and* the effect of the treatment on the outcome to be nonlinear and moderated by the covariates. We implement a model of the form

$$\begin{aligned} y_i &= \theta(\widetilde{t}_i, \mathbf{x}_i) + f(\mathbf{x}_i) + \epsilon_i, \;\; \mathbb{E}(\epsilon_i | t_i, \mathbf{x}_i) = 0 \\ t_i &= g(\mathbf{x}_i) + u_i, \;\; \mathbb{E}(u_i | \mathbf{x}_i) = 0. \end{aligned} \tag{22}$$

where the functions θ, f, g are all nonparametric.

We want to use this model to learn about how the treatment impacts the outcome. In the most simple regression model that we started with, this was just the average partial effect of t_i on the outcome t_i, which was the slope coefficient θ. We want to do something similar, but in our context

the slope is not a simple constant. Instead, we want θ to depend on both the value of the treatment and covariates (i.e., $\theta(\widetilde{t}_i, \mathbf{x}_i)$). That is, the impact of the treatment can be both nonlinear and moderated by the covariates.

In the continuous treatment case (see Appendix I for discussion of the binary case) our target of inference is the first partial derivative of the outcome with respect to the treatment, given background covariates: $\tau(\widetilde{t}_i, \mathbf{x}_i) = \frac{\partial}{\partial t}\theta(\widetilde{t}_i, \mathbf{x}_i)$. This estimand captures the impact of a ceteris paribus perturbation of the treatment on the outcome.[27] Modeling this local confounding returns an partial effect, an effect for an observation at treatment level t_i and covariate profile $\mathbf{x}_i$ (see e.g., Stolzenberg (1980) for an explicitly causal interpretation). This effect can be a curve that varies across values of the treatment and can depend on values of the covariates.[28]

[27] A causal interpretation requires ignorability holds in a continuous, open ball around t_i for each $\mathbf{x}_i$ and a positivity assumption that $t_i|\mathbf{x}_i$ be nondeterministic. This function $\tau(\widetilde{t}_i, \mathbf{x}_i)$ may be of interest in its own right even in a purely descriptive setting.

[28] See Appendix I for the binary treatment case.

B Introduction to basis functions

We employ a nonparametric regression model that models the outcome as an additive sum of a large number functions of the covariates, called *basis functions*. Each basis is a nonlinear transformation of a covariate, allowing us to model a more flexible set of functions. We illustrate this approach in Figure 8. The top row shows an example of a nonparametric curve with the data (left) and its first derivative (right). For simplicity, we make the curves only a function of the treatment.[29]

The middle row shows the true systematic component for the outcome and partial effect and, below it, a set of basis functions that we use to approximate the curve. MDEI uses 28 for each covariate: the linear covariate and then B-splines of degree 3, integrated B-splines of degree 3 and 5 (which look like sigmoid functions) then a set of degree 2, 3, and 4 Chebyshev polynomials evaluated at x, $x - s.d.(x)$ and $x + sd(x)$. These are illustrated below the true curve on the left. Each is differentiable, so their derivatives are shown on the right. These bases are then interacted with each other to approximate ever more complex functions, while we use a variable selection method to select an approximating subset.

Less important than the particular basis functions is that they are able to approximate a broad set of functions. We illustrate how these basis functions can add up and accurately recover a complex function in the third row. The left shows the estimated conditional mean; the right, the estimated partial derivative. A few points are selected as well, with their estimated derivative. We see that these basis functions can recover the relevant trends in the data generating process.

[29]For completeness, t_i is uniform on $[-1, 1]$ with $n = 250$ and the outcome is $\theta(\widetilde{t}_i) = 4\pi t_i \sin(2\pi t_i)$ and $\tau(\widetilde{t}_i) = 4\pi \sin(2\pi t_i) + 8\pi^2 t_i \cos(2\pi t_i)$ with normal, mean-zero noise with variance 2.

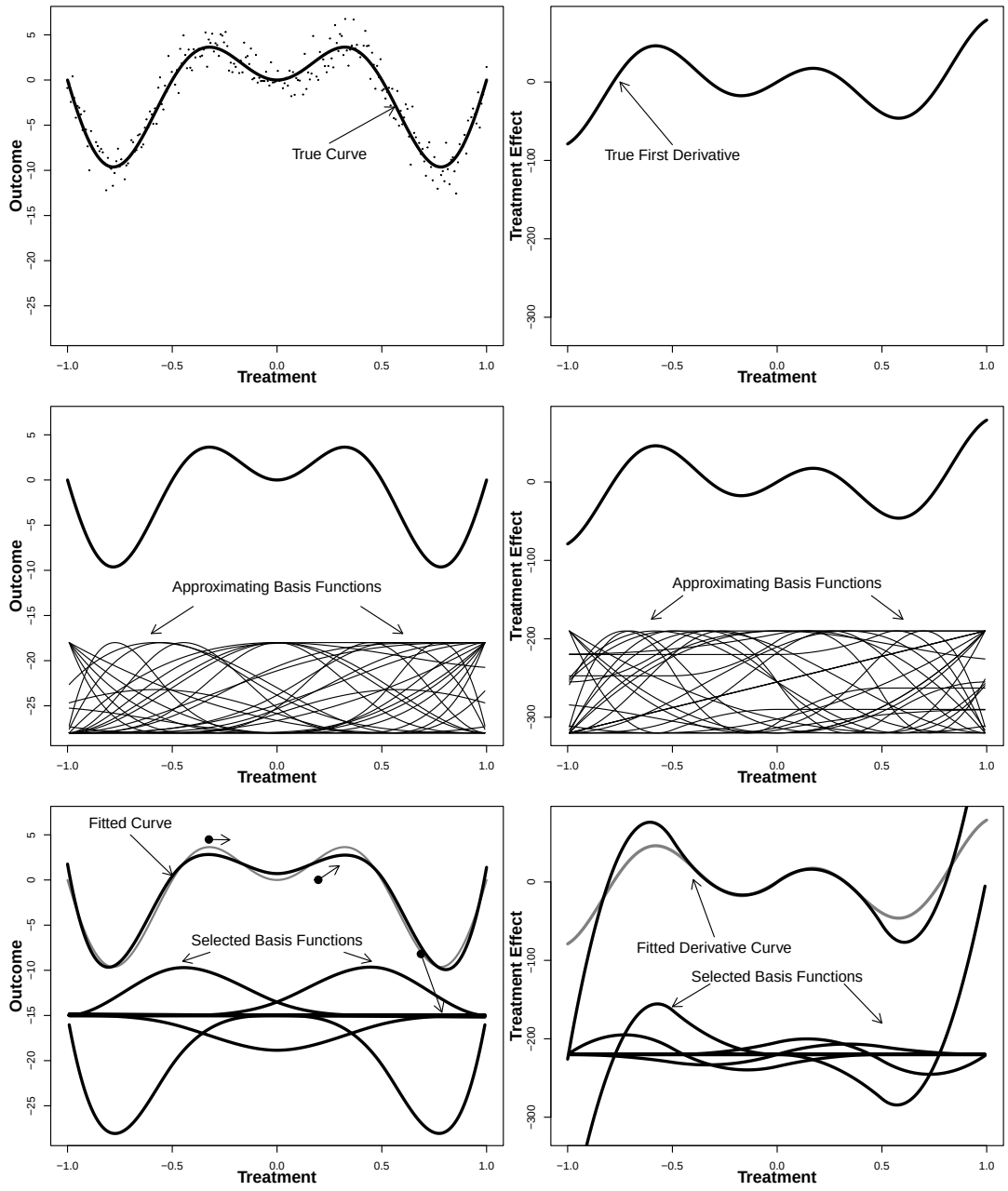

Figure 8: **Combining Basis Functions to Model Nonlinearities.**This figure illustrates how MDEI uses basis functions to approximate an outcome. The left column contains results for the outcome θ and the right for the partial effect $\tau(\widetilde{t}_i, \mathbf{x}_i)$. For simplicity, we assume both are solely a function of the treatment. The top row shows the true curves to be estimated, with the data in the left. The middle middle row illustrates the full set of basis functions that we use to approximate the true curve and its first partial derivative. The bottom row shows the basis functions that were selected, such that when added up they produce the fitted outcome (left) and derivative (right). A set of random points are selected with the estimated derivative at each point in the left figure. Rather than assuming a functional form, MDEI uses sets of basis functions that can be combined and interacted to provide an accurate local prediction.

C Function Classes

In this appendix we lay out additional details about the minimal assumptions we make about the sets of function our approach can handle. Given these assumptions, our estimator is asymptotically consistent as we discuss below Appendix C.2.3.

C.1 Moving beyond parametric functions

Extending past the simple linear regression, and more complicated models like the partially linear model, requires several analytic tools. These tools are necessary to characterize what, exactly, we mean by a complex model; what it means for the estimates of a complex model to be sufficiently "close" to the truth to allow for inference; and what sorts of inferential claims we can make about these curves. Each is crucial in moving beyond the linear model.

For clarity, we illustrate using the function f from our model, which adjusts for covariates in the outcome model. First, we distinguish between a "parametric" and "nonparametric" model. While there is no agreed-upon definition (see, e.g. Wasserman, 2006, p. 1), the distinction relies on the nature of the underlying assumptions. Assuming p elements of $\mathbf{x}_i$, say linear terms and any pre-specified interactions or higher order terms, a *parametric* model is one where the model is specified in advance, such as

$$f(\mathbf{x}_i) = \mathbf{x}_i^\top \gamma$$

with p elements in γ. We are assuming, then, that f lives in the space of functions linear in the elements of $\mathbf{x}_i$,

$$\{f : f(\mathbf{x}_i) = \sum_{j=1}^{p} \phi_j(\mathbf{x}_i) c_j;\ \phi_j(\mathbf{x}_i) = \mathbf{x}_{ij}\}$$

The *basis functions* of a space, which we denote $\{\phi_j\}_{j=1}^p$, are a set of functions which can be combined to represent any function in the space. In the parametric setting, the basis functions are simply individual covariates $\{\mathbf{x}_{ij}\}_{j=1}^p$. In the nonparametric case discussed below, the basis functions can be more complicated but still combine together to represent some target function like $f(\mathbf{x}_i)$, or, more generically, a conditional mean or even density. As discussed below, different types of approaches to constructing basis functions and estimating their parameters will be used.[30]

The crucial characteristic of a *parametric model* is that the number of parameters in the model (the K parameters c_k) are fixed and finite, and we can rely on asymptotics fixed in K with the sample size n growing. We adopt the intuition that a *nonparametric model* is one where we do not assume the model in advance, and instead the model specification is learned from the data. There are several ways to consider this in a regression setting. We may do so by allowing the number of basis functions to grow large, and possibly infinite, in order to accommodate a wide variety of nonlinear, interactive functions in $\mathbf{x}_i$.

The basis function approach itself subsumes the linear model, meaning that if the model is linear, we recover it, but we also allow include a much larger set of covariates in the regression to pick up unanticipated nonlinearities and interactions. The cost of this added flexibility is that our asymptotic analysis grows trickier. Since the number of basis functions, p_n, may be larger than the sample size n, or even infinite, we can no longer rely on parametric asymptotic arguments. In order to conduct inference, we must characterize our model space precisely and guarantee that we can still recover consistent estimates of the functions within it.

[30]For example, cubic smoothing splines are often used to model time (e.g., Ratkovic and Eng, 2010; Carter and Signorino, 2010) and other continuous covariates (Keele, 2008). Appendix B provides a comprehensive introduction to regression modelling using basis functions in the present context.

C.2 Functions for the treatment versus covariates

The *parametric* model will successfully adjust for confounders under the condition that these confounders enter the model linearly and additively. In other words, in the parametric model, our inference requires the conditional mean be in the function space

$$\{f : f(\mathbf{x}_i) = \mathbf{x}_i^\top \gamma\}.$$

We may want to consider the space

$$\{f : f(\mathbf{x}_i) = \sum_{j=1}^{p_n} \phi_j(\mathbf{x}_i) c_j\},$$

which is similar to the linear space, except we allow the number of bases p_n to grow in n and become potentially infinite, in the limit.

While certainly more flexible than the linear space, we cannot use a finite dataset to estimate an infinite number of parameters (Gyorfi et al., 2002). We can, though, retain a growing or infinite number of basis functions, a nonparametric space, if we constrain the function space in some manner. Perhaps the simplest example is the "sparsity assumption" that only some finite subset of the parameters $\{c_j\}_{j=1}^{\infty}$ are not zero, formulating the problem as one where the true model is parametric but we just do not know which basis functions constitute the true model (Buhlmann and van de Geer, 2013; Belloni, Chernozhukov and Hansen, 2014).

We implement methods that do not require the sparsity assumption, but constrain the function space so we can still fit models much more complex than a linear model while also recovering a consistent estimate.[31] Below we consider two different classes of functions. We will use the first to model our partial effect, and it will consist of smooth, differentiable functions of the treatment and covariates interacted together. We will fit this component using a high-dimensional regression model. The second class we consider is, roughly, functions that we can approximate well using a random forest. We will use this class to model any confounding or bias introduced from the covariates.[32]

C.2.1 Modeling $\theta(\widetilde{t}_i, \mathbf{x}_i)$ and $\tau(\widetilde{t}_i, \mathbf{x}_i)$

We first consider modeling $\theta(\widetilde{t}_i, \mathbf{x}_i)$, the part of the outcome explained by the treatment variable. In order to estimate the parameters, we constrain the full function space such that the functions vary, but not too wildly as to render estimation and inference impossible. We do so in three steps. First, we require $\phi_j(y_i, \mathbf{x}_i)$ to be bounded in the data. This allows us to guarantee that no one basis function goes off to infinity, which would leave inference untenable. Second, since we are interested in modeling ceteris paribus shifts in the treatment on the outcome, we require the basis functions to have a bounded partial derivative in the treatment. Third, we require the function to be "simple enough" that we can recover it from the data. We do so by requiring the sum of the absolute values of the parameters $\{c_j\}_{j=1}^{\infty}$ to be finite. This gives our space for θ as

$$\Theta = \{\theta : \theta(\widetilde{t}_i, \mathbf{x}_i) = \sum_{j=1}^{\infty} \phi_j(t_i, \mathbf{x}_i) c_j; \phi_j(t_i, \mathbf{x}_i) \text{ and } \frac{\partial}{\partial t}\phi_j(t, \mathbf{x}_i)\Big|_{t=t_i} \text{ bounded; } \sum_{j=1}^{\infty} |c_j| < \infty\}.$$

[31] Any consistent regression based method will work within our framework. A regression framework is necessary, since taking a derivative is straightforward. We utilize a sparse regression model and give conditions for its consistency below.

[32] We do not use the same function classes for each since recovering the partial effect with respect to the treatment requires restricting attention to differentiable functions. We use the random forest for the covariates due to the method's speed and well-established accuracy.

Taking the partial derivative, we can get the space containing $\tau(\widetilde{t}_i, \mathbf{x}_i)$.[33]

Importantly, this subspace subsumes the linear model.[34] For example, if the true model were parametric and linear in basis functions, we would recover the parametric model.

Before continuing, it is worth a brief mention of what sorts of functions are not in this space. First are those that are discontinuous functions of smooth covariates, since we only consider smooth bases. We present just such an example in Section 3.2. We also do not accommodate complex, erratic functions, meaning those such that the sum of the absolute values of the parameters diverges. For example, if we take $c_j = 1/j$, so the parameters have a long heavy tail, our results would not hold. The closer the model is to sparse, meaning $|c_j|$ decays quickly or even becomes zero, the better we expect our method to perform.

C.2.2 Modeling the functions f, g using a Lipschitz Space

We turn next to modeling how the covariates affect the outcome and treatment, as represented by the functions f, g respectively. Here, we simply assume that these functions can be well-estimated using a random forest, which places them in a *Lipschitz space*.[35] In essence, by using random forests to partial out the covariates–as part of an estimation and inference procedure–we can be operating in the seminonparametric framework.

This Lipschitz assumption is necessary to allow us to use random forests to adjust for the covariates. It is also more general than the space we use for treatment $\times$ covariate interactions, since Lipschitz functions are continuous but need not be differentiable. We add more structure to the space where we look for partial effects, which we operationalize as a derivative in the case of continuous treatment variable.

C.2.3 Consistency

We establishing consistency in our for the curve $\tau(\widetilde{t}_i, \mathbf{x}_i)$ by appealing to a sparsity condition in this class of models. We rely on the condition in Chernozhukov et al. (2018) Remark 4.3, which requires that the number of bases required to approximate the true curve is much less than sample size; formally if $s_{theta,n}$ is the number of bases needed to approximate our θ function uniformly, then we require $s_{\theta,n} << n$.[36] Under these conditions, then $\widehat{\theta}(\widetilde{t}_i, \mathbf{x}_i)$ and hence $\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i)$ is consistent. The B-spline bases will give us consistency in $L_1(P)$ for functions of the form, with p the number of

[33]

$$\mathcal{T} = \{\tau : \tau(\widetilde{t}_i, \mathbf{x}_i) = \sum_{j=1}^{\infty} \frac{\partial}{\partial t}\phi_j(t, \mathbf{x}_i)c_j\big|_{t=t_i}; \phi_j(t_i, \mathbf{x}_i) \text{ and } \frac{\partial}{\partial t}\phi_j(t, \mathbf{x}_i)\big|_{t=t_i} \text{ bounded; } \sum_{j=1}^{\infty} |c_j| < \infty\}.$$

[34]This space is a subspace of $L_1(P)$, the space of bounded functions with finite L_1 length of the parameters, consisting of functions partially differentiable in t. We note that standard results normally require working in $L_2(P)$, which contains $L_1(P)$. We use $L_1(P)$ since it leads to "sparser" estimates and handles the setting with a large number of basis functions better than working in $L_2(P)$. In practice, it is the difference between choosing a LASSO prior and a ridge prior on the basis functions.

[35]This is the space of functions where the slope of any secant line between any two points is bounded by some constant, say C. To formalize, this space can be characterized as

$$Lipschitz(\alpha) = \{f : |f(\mathbf{x}_i) - f(\mathbf{x}_i')| \leq C|\mathbf{x}_i - \mathbf{x}_i'|^{\alpha} \text{ for some } C < \infty\},$$

This is the most general space we use, where Linear spaces $\subset \Theta \subset$ Lipschitz functions.

[36]For a deeper conversation explicitly connecting consistency in nonparametric function spaces, particularly $L_1(P)$ considered here, see Buhlmann and van de Geer (2013), esp. Ch 6,8. We note that we have selected differentiable bases, so that the functions we fit are all differentiable, which is a smaller space than Lipschitz spaces (we will do worse on fitting, say, $y_i = |x_i| + e_i$ relative to a random forest), but we do note that we could include bases to accommodate in space.

covariates,

$$\theta(\widetilde{t}_i, \mathbf{x}_i) = \widetilde{t}_i\gamma_0 + \sum_{k=1}^{p}\sum_{k'=k}^{p}\sum_{j=1}^{27}\sum_{j'=1}^{27}\sum_{j''=j'}^{27}\phi_j(\widetilde{t}_i)\phi_{j'}(\mathbf{x}_{ik})\phi_{j''}(\mathbf{x}_{ik'})c_{kk'jj'j''};\ \ \sum\left|c_{kk'jj'j''}\right| < \infty \tag{23}$$

When combined with our algorithm, several issues come into play. First is requiring that the number of retained bases will, in the limit, contain the truth. We have selected this number to grow in sample size, be large, but not so large as to slow down our algorithm (as we will have to invert this matrix). Formal work by Fan and Lv (2008) on the *Sure Independence Screen* sets up conditions where retaining n bases can capture the true model with high probability, in a world where the outcome and bases are all jointly multivariate normal. This creates computational issues for us, as matrix inversion of an $n \times n$ is one of our bottlenecks.

So as to allow the number of bases to grow, but not to choke our algorithm, we retain $min(100 \times (1 + n_0 2), n_0/4)$ bases where n_0 is the number of observations in the discovery subsample $n_0 \approx n/3$. Over the course of all three cross-fits in a single iteration of our algorithm, for the full sample $n \in \{100, 250, 500, 1000, 5000, 100000\}$ we retain $\{9, 21, 42, 84, 417, 607\}$ bases at each cross-fit; i.e. this many bases is retained three times for each split of the data, and then the whole process is repeated a number of times. The retained bases are then brought to the estimation subsample and used for estimating $\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i)$, as described in Section 2 of the body.

D Coverage and Pointwise Confidence Band Concepts

Uncertainty intervals, like confidence intervals and confidence bands, are designed to have specific coverage properties. For example, in the linear model with a single slope coefficient, then the coverage probability is the proportion of times, over repeated samples that the interval contains the true value.[37] This is the standard confidence interval, as taught in the context of the regression and other parametric models.

There are multiple ways to achieve coverage on a nonparametric curve. Coverage can be achieved pointwise, uniformly, and on average. We turn to each in turn. The first, and most commonly encountered, is the *pointwise confidence interval*. This is the one returned by existing software (e.g., Hainmueller and Hazlett, 2013; Wager and Athey, 2017; Athey et al., 2019). In this case, at *any given* point $(t_i, \mathbf{x}_i)$, the interval will cover the true value at least $(1 - \alpha) \times 100\%$ of the time.[38] This pointwise property carry through to averages via a central limit theorem.

The pointwise interval does not contain information on the whole curve. For example, from a multiple testing perspective, a 95% confidence interval at every single point is not the same as a 95% band over *all* points. It is likely too narrow, as we show below in an illustrative simulation. Correcting this, and allowing for more informative and honest graphical displays, is a central goal of the project.

The second type, the *uniform confidence band*, will contain the *whole* curve $(1 - \alpha) \times 100\%$ of the time over repeated sampling. This band *does* contain information on the whole curve, since it will contain the full curve over repeated samples.[39] Uniform nonparameteric confidence bands

[37] For a single parameter $\tau \in \Re$, then, a valid $100 \times (1 - \alpha)\%$ confidence interval given data D_n and critical value C can be characterized as
$\lim_{n\to\infty} \sup_{\tau\in\Re} \Pr(\tau \in CI(\widehat{\tau}, \widehat{\text{Var}}(\widehat{\tau}), \mathcal{D}_n, C, \alpha)) \geq 1 - \alpha$
where the standard critical values of $C = 1.64$ and $C = 1.96$ give the 90% and 95% intervals.

[38] For any given $\tau(\widetilde{t}_i, \mathbf{x}_i)$ that can be well-approximated by the model, this interval can be characterized as:
For any given point $(t_i, \mathbf{x}_i)$, $\lim_{n\to\infty} \Pr(\tau(\widetilde{t}_i, \mathbf{x}_i) \in CI(\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i), \widehat{\text{Var}}(\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i)), \mathcal{D}_n, C, \alpha)) \geq 1 - \alpha$

[39] For any $\tau(\widetilde{t}_i, \mathbf{x}_i)$ that can be well-approximated by the method, this curve can be characterized as $\lim_{n\to\infty} \Pr(\text{For all points } (t_i, \mathbf{x}_i),\ \tau(\widetilde{t}_i, \mathbf{x}_i) \in CI(\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i), \widehat{\text{Var}}(\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i),), \mathcal{D}_n, C, \alpha)) \geq 1 - \alpha$

have been constructed in several specific settings (e.g., Genovese and Wasserman, 2005; Robins and van der Vaart, 2006) but cannot, in general, be constructed. Even when feasible they shrink slowly in sample size and are too wide to be usable (see, e.g., Wahba, 1983).

These two claims, pointwise and uniform, regularly diverge in nonparametric estimation for a subtle reason: not every point along a nonparametric curve will converge at the same rate in sample size. Recall that in order to identify the model, we need to restrict our attention to a particular space. The estimate will converge faster in areas where it is closer to our assumed space, and slower in other spaces. For one example, Leeb and Potscher (2008) show that if you work under a sparsity assumption–that only a finite number of the parameters c_j are non-zero–you can recover standard parameteric pointwise confidence intervals on each coefficient that shrink at the rate $n^{-1/2}$. These intervals are only valid if the model is in-truth sparse, but fall apart otherwise. If there are parameters that converge to zero at a rate of $n^{-1/4}$, the pointwise confidence interval can be arbitrarily misleading, since it may be missing parts of the true curve by a non-negligible amount that will leave our inference asymptotically invalid. Driving the distinction is that, while we may be able to make inferential claims about a given point on a curve, this is not the same as making such a claim along the curve.

We move onto our proposed band which implements the third type of coverage, average coverage. Rather than relying on claims across repeated samples, we instead follow Nychka (1988) (see also Wasserman (2006) ch. 5.8) and consider *average coverage*, which is the probability that a confidence band contains the true value over the observed sample.[40] This band has the nice property that it will contain the true curve at a high percentage of the observed data. It is also narrow enough for applied work, but with provable average coverage properties.[41]

E Technical Details for Algorithm

In this section, we present technical details of our algorithm that we have not placed in the body.

E.1 Algorithm Diagram

We outline three algorithms that we use to implement our method. In the first, Algorithm 1, we generate a set of bases that model heterogeneity. The second, Algorithm 2, details how we construct our intervals, given fitted values and standard errors at each point. The third, Algorithm 3, uses

[40] This property can be written as:
$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(\tau(\widetilde{t}_i, \mathbf{x}_i) \in CI(\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i), \widehat{\mathrm{Var}}(\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i)), \mathcal{D}_n, C(D_n), \alpha)) \geq 1-\alpha$

[41] Formal derivations of this average coverage can be found in Appendix G.

the first two to construct our estimates.

Algorithm 1: Generating Candidate Bases

Data: Outcome vector y_i, treatment vector t_i, length p coveriate vector $\mathbf{x}_i$ with intercept in first column, all in the discovery subsample, n_0 observations in this subsample

Functions: 28 basis functions denoted $\phi_j where by default \phi_0(z) = 1, \phi_1(z) = z$
$\rho(a, b)$ the correlation of a and b.

Result: A set of indices generating nonparametric bases for modeling treatment $\times$ covariate interactions

Using the discovery subsample, generate $\widetilde{y}_i = y_i - \widehat{\mathbb{E}}(y_i|\mathbf{x}_i), \widetilde{t}_i = t_i - \widehat{\mathbb{E}}(t_i|\mathbf{x}_i)$ using random forests

for *j in 1 to p* **do**
  for *j′ in 1 to p)* **do**
    for *d in 2 to 28* **do**
      for *d′ in 1 to 28* **do**
        for *d′′ in d′ to 28* **do**
          Save $\rho(\widetilde{y}_i, \phi_d(\widetilde{t}_i) \times \phi_{d'}(\mathbf{x}_{ij}) \times \phi_{d''}(\mathbf{x}_{ij'}))$

Return indices corresponding with bases with top $min(100(1 + n_0^2, n_0/4))$ values of ρ

Algorithm 2: Generating Critical Value

Data: Matrix of fitted values and estimated standard errors at each point;
False positive rate α
Result: Fitted values, first derivative, variance estimates, and uniform confidence interval
for *i in 1 to numsplits* **do**
  Find smallest critical value such that symmetric confidence interval contains $100 \times (1-\alpha)\%$ of the data

Return critical value for $\widetilde{y}_i$ and add one to generate critical value for $\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i)$.

Algorithm 3: Estimation Algorithm

Data: Outcome y_i, treatment t_i, covariates $\mathbf{x}_i$, indices $\mathcal{I}$ for selected interactive splines, data split into discovery/estimation/inference subsamples.
Result: Fitted values, first derivative, variance estimates, and uniform confidence interval
for *1 in 1 to numsplits* **do**
  Using the discovery subsample, generate partialed-out outcome $\widetilde{y}_i$, partialed-out treatment $\widetilde{t}_i$, and retained bases bases $\mathcal{I}$ from Algorithm 1;
  Using data from the estimation subsample, model $\widehat{\theta}(\widetilde{t}_i, \mathbf{x}_i)$ from regressing $\widetilde{y}_i$ on the retained bases using a sparse regression; Using data from the estimation subsample, model conditional variance by regressing the squared errors on covariates using random forests; Evaluate point estimate, first derivative, and variance of fitted and first derivative at each point in the inference subsample;
  Cross-fit until fitted values, first derivative estimates, variance estimates generated for every datum

Calculate conformal confidence intervals using confidence interval using Algorithm 2.

F Sparse Regression Model

Even after screening, we still have hundreds of nonparametric bases. Regressions of this magnitude, though, can be estimated reliably using existing high-dimensional regression methods. We implement the high-dimensional regression described by Ratkovic and Tingley (2017). This work was focused on variable selection, estimating a subset of bases that are likely non-zero. Our problem is subtly different: we want the best predictive model.

High dimensional regression requires a tuning parameter, λ, that controls the level of shrinkage. We implement an adaptation of the Bayesian sparse regression, from Ratkovic and Tingley (2017). For completeness, we present the full model hierarchy,

$$y_i|\mathbf{x}_i, \beta \sim \mathcal{N}(X_i^\top \beta, \sigma^2) \tag{24}$$
$$\beta_k|\lambda, w_k, \sigma \sim DE\left(\lambda w_k/\sigma\right) \tag{25}$$
$$\lambda^2|n, p \sim \Gamma\left(\alpha, \rho\right) \tag{26}$$
$$w_k|\gamma \sim \text{generalizedGamma}(1, 1, \gamma = 2) \tag{27}$$
$$\tag{28}$$

though we return point estimates via an EM algorithm.

We take $\rho = 1$ but have found our results sensitive to α. Ratkovic and Tingley (2017) took γ as to be estimated, but we instead take $\gamma = 2$ as it leads to tractable updates and then select α via generalized cross-validation (Wahba, 1990).

G Variance Derivation

G.1 Deriving the Conformal Bound

We assume we have a valid conformal bound. Then, for some future value y_i' at $\widetilde{t}_i, \mathbf{x}_i$, variance at this point $\widehat{\sigma}_{\widehat{\theta}}(\widetilde{t}_i, \mathbf{x}_i)$, and critical value $\widehat{C}_{1-\alpha/2}$, we get

$$\Pr(|y_i' - \widehat{\theta}(\widetilde{t}_i, \mathbf{x}_i)| \leq \widehat{C}_{1-\alpha/2}\widehat{\sigma}_{\widehat{\theta}}(\widetilde{t}_i, \mathbf{x}_i)) \geq 1 - \alpha. \tag{29}$$

Now, we bound the inequality inside the probability from the left using $|a+b| - |b| \leq |a|$ where $a + b = \widehat{\theta}(\widetilde{t}_i, \mathbf{x}_i) - \theta(\widetilde{t}_i, \mathbf{x}_i)$, $a = y_i' - \widehat{\theta}(\widetilde{t}_i, \mathbf{x}_i)$ $b = \theta_i - y_i'$. So, with a conformal band, we can ensure the following event occurs with probability at least $1 - \alpha$

$$\widehat{C}_{1-\alpha/2}\widehat{\sigma}_{\widehat{\theta}}(\widetilde{t}_i, \mathbf{x}_i) \geq |y_i' - \widehat{\theta}(\widetilde{t}_i, \mathbf{x}_i)'| \tag{30}$$

$$\geq |\widehat{\theta}(\widetilde{t}_i, \mathbf{x}_i) - \theta(\widetilde{t}_i, \mathbf{x}_i)| - |\theta(\widetilde{t}_i, \mathbf{x}_i) - y_i'| \tag{31}$$

and rearranging, then bounding the term on the right gives

$$|\widehat{\theta}(\widetilde{t}_i, \mathbf{x}_i) - \theta(\widetilde{t}_i, \mathbf{x}_i))| \leq \widehat{C}_{1-\alpha/2}\widehat{\sigma}(\widetilde{t}_i, \mathbf{x}_i) + |\theta(\widetilde{t}_i, \mathbf{x}_i) - y_i'| \tag{32}$$

$$\leq (\widehat{C}_{1-\alpha/2} + 1)\widehat{\sigma}_{\widehat{\theta}}(\widetilde{t}_i, \mathbf{x}_i) \tag{33}$$

since $\widehat{\sigma}(\widetilde{t}_i, \mathbf{x}_i) \leq |\theta(\widetilde{t}_i, \mathbf{x}_i) - y_i'|$, in expectation.

Thus, we should expect the confidence band

$$\widehat{\theta}(\widetilde{t}_i, \mathbf{x}_i) \pm (\widehat{C}_{1-\alpha/2} + 1)\widehat{\sigma}_{\widehat{\theta}}(\widetilde{t}_i, \mathbf{x}_i) \tag{34}$$

to have at least $100 \times (1-\alpha)\%$ average coverage of the systematic component $\theta(\widetilde{t}_i, \mathbf{x}_i)$. We replace the conformal critical value $\widehat{C}_{1-\alpha/2}$ with $\widehat{C}_{1-\alpha/2} + 1$, which has to be widened to better include θ, rather than a future predictive value.

Since our bounds are not exact, we expect it to be conservative for $\theta(\widetilde{t}_i, \mathbf{x}_i)$. The predictive bound is exact, asymptotically, but our bound on the true systematic component comes from bounding this conformal band. Therefore, we expect the coverage of the $100 \times (1-\alpha)\%$ band to be greater than $100 \times (1-\alpha)\%$, but this is the cost we had to incur in moving from bounding the predictive value to the systematic component.

We then use this critical value to construct a band around $\widehat{\tau}$ as

$$\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i) \pm (\widehat{C}_{1-\alpha/2} + 1)\widehat{\sigma}_{\widehat{\tau}}(\widetilde{t}_i, \mathbf{x}_i) \tag{35}$$

We estimate the variance using the law of total variance

$$\underbrace{\widehat{\sigma}^2_{\widehat{\theta}}(t_i, \mathbf{x}_i)}_{\text{Total Variance}} = \underbrace{\widehat{s}^2_{\widehat{\theta}}(\widetilde{t}_i, \mathbf{x}_i)}_{\text{Sampling Variance}} + \underbrace{\widehat{\sigma}^2_{\theta}(\widetilde{t}_i, \mathbf{x}_i)}_{\text{Error Variance}} \tag{36}$$

We calculate the sampling variance as the variance in the fitted values over repeated cross-fits. We then estimate the error variance using a random forest on the squared residuals, and these estimates are also averaged over split-samples.

We then construct our error on $\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i)$ using the same formula. The sampling variance can be calculated from the cross-fit sample variance over the estimates. For the second variance term, we estimate the variance of y_i attributable to $\widetilde{t}_i$, but not $\mathbf{x}_i$, which we estimate as

$$\widehat{\sigma}^2_{\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i)} = \widehat{\text{Var}}(y_i|\mathbf{x}_i) - \widehat{\text{Var}}(y_i|\widetilde{t}_i, \mathbf{x}_i) \tag{37}$$

where the estimates are constructed using random forests on the estimation subsample. In the limit, this estimate should be nonnegative; in practice, we instead take its absolute value.

H Performance Simulations

H.1 Data Generating Processes

We next present simulation evidence illustrating MDEI's utility in estimating a partial effect. We include four sets of simulations presented in increasing complexity, a linear model, a low-dimensional interactive model, a high-dimensional interactive model, and a model with a nonlinearity, respectively:

In each setting, we generate five covariates $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{i5}$ from a standard multivariate normal with correlation 0.5.

In the first four settings, we take

$$t_i = \frac{(\mathbf{x}_{i2} - 1)^2}{4} + \epsilon_i^T; \;\; \epsilon_i^T \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1). \tag{38}$$

In the fifth setting, we introduce a discontinuity by using

$$t_i = \text{sign}\,(\mathbf{x}_{i1}) \times \frac{(\mathbf{x}_{i2} - 1)^2}{4} + \epsilon_i^T; \;\; \epsilon_i^T \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1) \tag{39}$$

where

$$\text{sign}\,(a) = \begin{cases} -1; & a \leq 0 \\ 1; & a > 0 \end{cases} \tag{40}$$

Note that this is outside our model space and a more complex setting than that in our second illustrative simulation in Section 3 of the main body.

We then use the following outcome models,

$$\text{1 Linear:} \;\; y_i = y_i + \mathbf{x}_{i1} + \frac{\mathbf{x}_{i2} - 1}{4} + + \epsilon_i \tag{41}$$

$$\text{2 Partially Linear:} \;\; y_i = t_i + \mathbf{x}_{i1} + \frac{(\mathbf{x}_{i2} - 1)^2}{4} + \epsilon_i \tag{42}$$

$$\text{3 Additive Linear:} \;\; y_i = 4\sin(t_i) + \mathbf{x}_{i1} + \frac{(\mathbf{x}_{i2} - 1)^2}{4} + \epsilon_i \tag{43}$$

$$\text{4 Interactive:} \; y_i = 4\sin(t_i) \times \mathbf{x}_{i1} + \frac{(\mathbf{x}_{i2} - 1)^2}{4} + \epsilon_i \tag{44}$$

$$\text{5 Discontinuity} \; y_i = 4\sin(y_i) \times \text{sign}\,(\mathbf{x}_{i1}) + \frac{(\mathbf{x}_{i2} - 1)^2}{4} + \epsilon_i \tag{45}$$

where the the error is independent, identical Gaussian such that the true R^2 in the outcome model is 0.5. We consider $n \in \{250, 500, 1000, 2000\}$.

We continue to contrast with the kernel regularized least squares model (Hainmueller and Hazlett, 2013) and Generalized Random Forests (Athey et al., 2019). We select these two models for comparison because they offer both point estimates and uncertainty estimates for the partial effect curve, $\tau(\widetilde{t}_i, \mathbf{x}_i)$.[42]

[42] We also do not include other candidate approaches. Many existing models focus on uncertainty and estimates for the fitted values, not the partial effect curve; for example, POLYMARS (Stone et al., 1997), Sparse Additive Models (Ravikumar et al., 2009), Bayesian additive regression trees (Chipman, George and McCulloch, 2010) and boosting (Ridgeway, 1999), and the SuperLearner (Polley and van der Laan, N.d.). Any of these could have been used for partialing out the covariates; we implemented random forests for simplicity. Other possible sparse estimators could have included the horseshoe, and Bayesian Bridge (Carvalho, Polson and Scott, 2010; Polson, Scott and Windle, 2014); we found our EM implementation to offer a more stable estimate than the variational implementation of these. Cattaneo, Farrell and Feng (Forthcoming) offer an alternative estimtation strategy, though it does not accommodate more than a handful of covariates.

H.1.1 Evaluation Metrics

We assess methods across two dimensions, each commensurate with our two estimation contributions: point estimation and coverage rates on $\tau(\widetilde{t}_i, \mathbf{x}_i)$. For the former, we use the mean absolute error,

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i) - \tau(\widetilde{t}_i, \mathbf{x}_i)| \tag{46}$$

and the sample average coverage probability,

$$SACP = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left\{\tau(\widetilde{t}_i, \mathbf{x}_i) \in CB_{\tau,\mathcal{D}_n}(t_i, \mathbf{x}_i)\right\}.$$

All simulations were run 500 times.

H.1.2 Results

Results from the simulations can be found in Figure 9. Each row corresponds with our simulation setting, from the additive linear model in row one to the complex, discontinuous model in row 5. In each figure, the x-axis shows outcomes by sample size. The first column presents the bias on the average partial effect, by method. Across settings, all methods do well, with GRF doing the best overall. KRLS misses the average effect in the simplest model, due to its shrinkage, but with any complexity, all models do well. The second column presents mean absolute error, a measure of accuracy of our estimates over the whole of the curve. All methods perform well in the simplest settings, but KRLS and MDEI perform the best in the most complex settings. We suspect that there are conditional mean specifications where any of the methods presented will outperform others; our take away here is that all three methods perform passably well.

The third column, presenting the sample coverage, is the most important. The horizontal line at 0.9 is the nominal rate, so values above this line are denote conservative bands and values below it denote an invalid band. In the simplest settings, all methods are valid, and MDEI is quite wide. This is to be expected, of course, since we construct our bands to be valid even if the model is wrong. As the models get more complex, in rows 3-5, we see that coverage plummets for KRLS and GRF. Basically, in settings 3 and 4, and especially 5, the confidence bands returned by these method provide little information on the location of the true curve. The cost of this coverage shows in the last columns, which contains the average width of the interval, by method. We see that the our confidence intervals are notably wider, as expected. Narrower bands can be achieved, but at the cost of only covering simple models.

I Binary and Categorical Treatment Regimes.

In the paper we focus on the continuous treatment case. A mature literature examines the case with binary and categorical treatments. This manuscript does not treat the binary or categorical treatment regime as a separate setting. Instead, we note that our approach carries through to the binary treatment setting. Rather than modeling the propensity score, or conditional probability of treatment, we instead model the conditional mean of the treatment. The key distinction is that the former is constrained to fall in $[0, 1]$, and the probabilites are used to match or generate inverse probability weights.

Rather than adjust through matching or weighting, we are instead adjusting the conditional mean. So, instead of fitting a logistic or probit regression, we are instead fitting using nonlinear least squares. The benefit is that we are not working with inverse probability weights, which can be unstable, nor relying on distributional assumptions of the treatment variable or outcome. We lose, though, efficiency gains that come from making distributional assumptions.

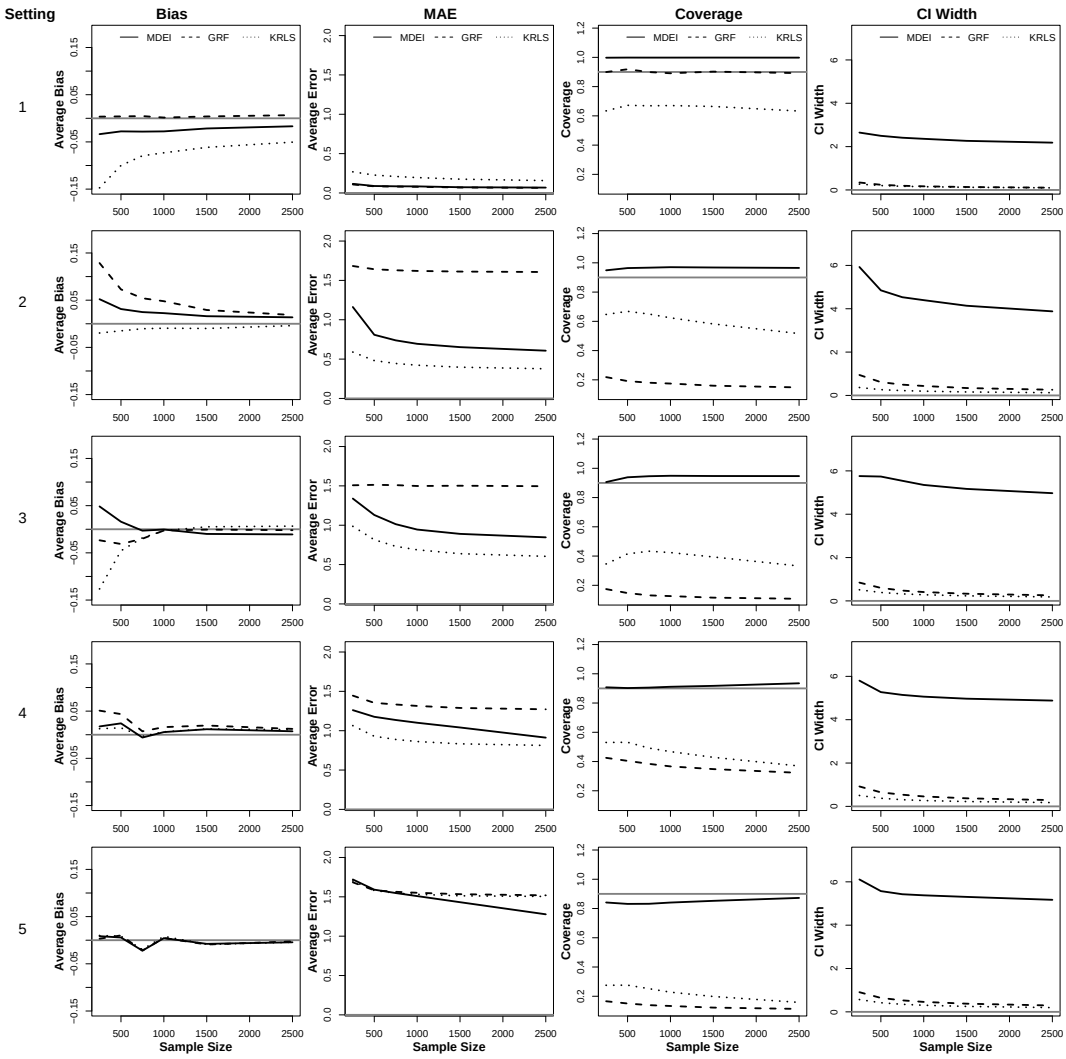

Figure 9: Performance Simulation Results

References

Andersen, Robert. 2009. "Nonparametric methods for modeling nonlinearity in regression analysis." *Annual Review of Sociology* 35:67–85.

Aronow, Peter and Cyrus Samii. 2016. "Does Regression Produce Representative Estimates of Causal Effects?" *American Journal of Political Science* 60(1):250–267.

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

Beck, Nathaniel and Simon Jackman. 1998. "Beyond linearity by default: Generalized additive models." *American Journal of Political Science* pp. 596–627.

Belloni, Alexandre, Victor Chernozhukov and Christian Hansen. 2014. "Inference on Treatment Effects after Selection among High-Dimensional Controls." *Review of Economic Studies* 81(2):608–650.

Buhlmann, Peter and Sara van de Geer. 2013. *Statistics for High-Dimensional Data*. Berlin: Springer.

Carter, David B and Curtis S Signorino. 2010. "Back to the future: Modeling time dependence in binary data." *Political Analysis* 18(3):271–292.

Carvalho, C, N Polson and J Scott. 2010. "The Horseshoe Estimator for Sparse Signals." *Biometrika* 97:465–480.

Cattaneo, Matias D., Max H. Farrell and Yingjie Feng. Forthcoming. "Large Sample Properties of Partitioning-Based Series Estimators." *Annals of Statistics* .

Chernozhukov, Victor, Denis Chetverikov, Esther Demirer, Mertand Duflo, Christian Hansen, Whitney Newey and James Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *The Econometrics Journal* .

Chipman, Hugh A, Edward I George and Robert E McCulloch. 2010. "BART: Bayesian additive regression trees." *The Annals of Applied Statistics* pp. 266–298.

Fan, Jianqing and Jinchi Lv. 2008. "Sure independence screening for ultrahigh dimensional feature space." *Journal of the Royal Statistical Society: Series B* 70:849–911.

Genovese, Christopher R. and Larry Wasserman. 2005. "Confidence sets for nonparametric wavelet regression." *Annals of Statistics* 33(2):698–729.

Gyorfi, Laszlo, Michael Koholor, Adam Krzyzak and Harro Walk. 2002. *A Distribution-Free Theory of Nonparametric Regression*. New York: Springer.

Hainmueller, Jens and Chad Hazlett. 2013. "Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach." *Political Analysis* 22(2):143–168.

Härdle, Wolfgang Karl, Marlene Müller, Stefan Sperlich and Axel Werwatz. 2012. *Nonparametric and semiparametric models*. Springer Science & Business Media.

Hardle, Wolfgang and Thomas M. Stoker. 1989. "Investigating Smooth Multiple Regression by the Method of Average Derivatives." *Journal of American Statistical Association* 84:986–95.

Hastie, Trevor and Robert Tibshirani. 1990. *Generalized additive models*. Wiley Online Library.

Ho, Daniel E., Kosuke Imai, Gary King and Elizabeth A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15(3):199–236.

Imai, Kosuke, Luke Keele and Dustin Tingley. 2010. "A general approach to causal mediation analysis." *Psychological methods* 15(4):309.

Imbens, Guido W and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Keele, Luke John. 2008. *Semiparametric regression for the social sciences*. John Wiley & Sons.

Kropko, Jonathan and Jeffrey J. Harden. 2020. "Beyond the Hazard Ratio: Generating Expected Durations from the Cox Proportional Hazards Model." *British Journal of Political Science* 50(1):303–320.

Leeb, Hannes and Benedikt Potscher. 2008. "Sparse Estimators and the Oracle Property, or the Return of Hodges Estimator." *Journal of Econometrics* 142:201–211.

Nychka, Douglas. 1988. "Bayesian Confidence Intervals for Smoothing Splines." *Journal of the American Statistical Association* 83:1134–1143.

Polley, Eric and Mark van der Laan. N.d. "SuperLearner: super learner prediction, 2012." *URL http://CRAN. R-project. org/package= SuperLearner. R package version*. Forthcoming.

Polson, Nicholas G, James G Scott and Jesse Windle. 2014. "The bayesian bridge." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76(4):713–733.

Ratkovic, Marc and Dustin Tingley. 2017. "Sparse Estimation and Uncertainty with Application to Subgroup Analysis." *Political Analysis* 1(25):1–40.

Ratkovic, Marc T and Kevin H Eng. 2010. "Finding jumps in otherwise smooth curves: Identifying critical events in political processes." *Political Analysis* 18(1):57–77.

Ravikumar, Pradeep, John Lafferty, Han Liu and Larry Wasserman. 2009. "Sparse Additive Models." *Journal of the Royal Statistcal Society, Series B* 71(5):1009–1030.

Ridgeway, Greg. 1999. "The state of boosting." *Computing Science and Statistics* 31:172–181.

Robins, James and Aad van der Vaart. 2006. "Adaptive Nonparametric Confidence Sets." *Annals of Statistics* 34(1):229–253.

Robinson, Peter. 1988. "Root-N Consistent Semiparametric Regression." *Econometrica* 56(4):931–954.

Sekhon, Jasjeet S. 2009. "Opiates for the Matches: Matching Methods for Causal Inference." *Annual Review of Political Science* 12(1):487–508.

Stolzenberg, Ross. 1980. "The Measurement and Decomposition of Causal Effects in Nonlinear and Nonadditive Models." *Sociological Methodology* 11:459–488.

Stone, Charles J., Mark H. Hansen, Charles Kooperberg and Young K. Truong. 1997. "Polynomial Splines and Their Tensor Products in Extended Linear Modeling." *The Annals of Statistics* 25(4):1371–1470.

Wager, Stefan and Susan Athey. 2017. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association* .

Wahba, Grace. 1983. "Bayesian "Confidence Intervals" for the Cross-Validated Smoothing Spline." *Journal of the Royal Statirtical Society, Ser. B* 45:133–150.

Wahba, Grace. 1990. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics.

Wasserman, Larry. 2006. *All of Nonparametric Statistics*. Springer Texts in Statistics Springer.

Wood, Simon N. 2006. *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC Texts in Statistical Science.