Estimation and Inference on Nonlinear and Heterogeneous Effects[*]

Marc Ratkovic[†] Dustin Tingley[‡]

February 1, 2021

Abstract

Multiple regression has been the go-to method for data analysis for generations of scholars due to its transparency, interpretability, and desirable theoretical properties. However, 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 effect of a treatment variable. Importantly, we introduce a robust approach to uncertainty estimates. We provide simulation evidence and an application illustrating the performance of the method.

Key Words: machine learning, causal inference, high dimensional Bayesian regression, sure independence screening, uniform confidence bands, conformal inference

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 modelling the influence of the treatment variable while adjusting for 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 appropriate 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.

Estimating the treatment effect requires developing a method that can return an accurate point estimate of the estimand. This requires flexibly modeling two things: 1) the background covariates/confounders and 2) the treatment effect. Using a novel machine learning method we flexibly adjust 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, like 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, meaning we expect the confidence band for our marginal effect to cover the true curve at some proportion, say 90%, of the data. Doing this 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 instantaneous causal effect, which is the observation-level marginal effect of the treatment on the outcome (e.g. Stolzenberg, 1980). Just as a regression coefficient is interpreted as an average marginal effect over the sample, the instantaneous causal effect is interpreted as the "slope" at a given observation, given the values of observed pre-treatment variables. To perform this estimate, MDEI advances recent machine learning methods by implementing a flexible, nonparametric regression to model the treatment effect. The model can detect a wide class of nonlinear and treatment/covariate interactions. Here, we build on an approach known in the econometrics literature known as seminonparametrics (see e.g., Chen, 2007) where we model the contributions of both the covariates treatment/covariate interactions using flexible machine learning methods.

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 treatment 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 marginal effect of the treatment on the outcome at each point.

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.

The paper proceeds as follows. Section 1 lays out the challenge of estimating and conducting inference on treatment effects without relying on simplifying assumptions about how the treatment impacts the outcome variable. In Section 2 we build up from a model where these assumptions are very strong to the our model where these assumptions are very minimal. Section 3 introduces our approach to inference in this setting where we develop a confidence band utilizing a conformal inference approach. Section 4 then introduces MDEI and shows how we estimate both point esti- mates and uncertainty estimates. Section 5 provides simulations to illustrate our approach while the Appendix compares the performance of MDEI to other cutting edge approaches. Section 6 shows MDEI in action with an applied example. Section 7 concludes. Throughout we discuss related research. We include an extensive online technical appendix.

1   The Estimation and Uncertainty Challenge

Consider the familiar regression model,

$$Y = \theta T + X^\top \gamma + \epsilon; \quad \mathbb{E}(\epsilon | T, X) = 0$$

with outcome Y, a variable of theoretical interest T, a vector of additional background variables X, and an error term ϵ that is assumed to be mean-independent of the treatment and background variables. This model is adopted for several reasons. First, θ measures the average partial effect, or "slope," when characterizing the relationship between the outcome and treatment. Second, $X^\top \gamma$ adjusts for other variables that impact both the treatment and outcome. Third, given the observed data (D_n), 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

$$CI(\widehat{\theta}, \widehat{\mathrm{Var}}(\widehat{\theta}), \mathcal{D}_n, C, \alpha) = (\widehat{\theta} - C_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\widehat{\theta})}, \widehat{\theta} + C_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\widehat{\theta})})$$

where $C_{1-\alpha/2}$ is a critical value that controls the false positive rate (e.g., under a normality assumption 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. For example, we relax the assumption that the covariates in 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 (e.g., the effect could be a curve rather than a straight line) and pre-treatment covariates X. To do this, we will replace the linear component θT with a flexible, interactive function $\theta(T, X)$. Then, we can model the effect of T on Y as the partial derivative of $\theta(T, X)$ with respect to T, denoted $\tau(T, X)$, which is the "slope" coefficient at a particular value of the treatment and covariates.

However, allowing for this flexibility makes characterizing the uncertainty in our estimates ("inference") challenging. Adjusting for covariates nonparametrically and then allowing a flexi-

ble treatment effect falls under a class of models referred to as *seminonparametric* (Chen, 2007). Seminonparametric models raise difficult–but surmountable–issues of how to conduct inference. On the estimation side, a crucial step is a *split-sample* approach, using half the sample to model the covariates and the other half to learn the structure and form of the treatment effect. The roles of the splits are reversed and the process repeated. Doing this serves as a crucial guard against over-fitting, as one subset of the data is used to learn how the covariates enter the model and a separate set to learn the treatment effect.

However, the seminonparametric construction raises even deeper issues for inference. While developing a confidence band for a parametric curve is straightforward, implementing one for an arbitrarily complex, nonparametric curve is rather challenging. In Section 3 we develop a confidence band around $\hat{\tau}$ with an *average coverage* property (Nychka, 1988; Wasserman, 2006). This band has the nice property that, for user-controlled false positive rate $0 < \alpha < 1$, the band will contain the true curve over at least $100 \times (1 - \alpha)\%$ of the observed data, asymptotically. Because we want to consider the treatment effect as a curve, we must conduct inference on the entire curve, not simply at a single point along it. Of course, if the true model is in fact the standard linear model, then our confidence band will contain the same interval listed above (for a given α), but only as the sample size increases.[1] As we discuss below, we turn to the method of *conformal inference* developed by Lei and Wasserman (2014), in order to determine the critical value of the band. We show that integrating conformal inference with a machine learning approach for estimating the treatment effect and its variance results in asymptotically valid bands.

In the next section, we discuss the role of assumptions in regression models and build up to

---

[1]As discussed below, our band will actually be wider than this interval, to allow for uncertainty attributable to model misspecification. Our constructed band will contain the true value more than $100 \times (1 - \alpha)\%$ of the time. Were it arbitrarily complex (which we define specifically below), the band will contain the true value $100 \times (1 - \alpha)\%$ of the time asymptotically.

the seminonparametric approach we use. Section 3 then discusses our approach to uncertainty estimation (inference). These provide the requisite pieces that then guide the MDEI algorithm presented in Section 4.

2   Models and 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. We start with the simplest possible model, the linear regression, and progress to our model.

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 on Y, at some particular point $(T, X) = (T', X')$.

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

$$Y = \theta T + X^\top \gamma + \epsilon, \;\; \mathbb{E}(\epsilon | T, X) = 0$$
$$T = X^\top \beta + u, \;\; \mathbb{E}(u | X) = 0,$$

(1)

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

6

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.

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

$$Y = \theta T + X^\top \gamma + \epsilon, \quad \mathbb{E}(\epsilon | T, X) = 0$$
$$T = g(X^\top \beta) + u, \quad \mathbb{E}(u | X) = 0.$$
(2)

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.[2]

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

$$Y = \theta T + f(X) + \epsilon, \quad \mathbb{E}(\epsilon | T, X) = 0$$
$$T = g(X) + u, \quad \mathbb{E}(u | X) = 0.$$
(3)

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* (Hardle and Stoker, 1989; Härdle et al., 2012), since it is linear in T but nonparametric in the remainder.

Quite a bit of work has been done in this setting, with the aim of estimating an average effect in the presence of unspecified confounding, using machine learning to adjust for the covariates. In the case of a binary treatment, Diamond and Sekhon (2012) introduces such a method, but it has since been extended to continuous or arbitrary treatment variables (Chernozhukov et al., 2018; Kennedy,

---

[2]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 H connects our approach to the binary treatment setting.

Lorch and Small, In Press; Fong, Hazlett and Imai, 2018).[3]

The partially linear model improves on the standard practice captured in Model 2, 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 X 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 with $\theta(T)$, a smooth function of T. Doing so generates the model

$$Y = \theta(T) + f(X) + \epsilon, \mathbb{E}(\epsilon|X,T) = 0;$$

$$T = g(X) + u, \mathbb{E}(u|X) = 0; \ f, g, \theta \text{ unknown}$$

(4)

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).

If we allow θ to be a smooth function of T, estimation can occur following the same cross-fitting described earlier. Rather than simply regressing $\widetilde{Y}$ on $\widetilde{T}$, we could instead model the relationship using a smoothing spline. Under the assumption that $\theta(T)$ 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.

A More Flexible Model We wish to allow for the background covariate specification to be learned from the data *and* the effect of T on Y to be nonlinear and moderated by the covariates.

---

[3]Recovering a valid confidence interval on θ in this setting requires combining machine learning methods for f, g with a regression for θ. A full theoretical discussion of these issues can be found in Chernozhukov et al. (2018); Farrell, Liang and Misra (forthcoming); van der Vaart (1998, ch. 25).

We implement a model of the form

$$Y = \theta(T, X) + f(X) + \epsilon, \quad \mathbb{E}(\epsilon|T, X) = 0$$

$$T = g(X) + u, \quad \mathbb{E}(u|X) = 0. \tag{5}$$

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 marginal effect of T on the outcome Y, 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., $\theta(T, X)$). That is, the impact of the treatment can be both nonlinear and moderated by the covariates.

In the continuous treatment case (see Appendix H 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: $\tau(T, X) = \frac{\partial}{\partial T}\theta(T, X)$. This estimand captures the impact of a ceteris paribus perturbation of the treatment on the outcome.[4] Modeling this local confounding returns an estimate of the "instantaneous causal effect" (ICE), a marginal effect for an observation at treatment level T and covariate profile X (see e.g., Stolzenberg (1980)). This effect can be a curve that varies across values of the treatment and can depend on values of the covariates.[5]

Existing Methods That Fit This Model Before discussing our particular implementation, we first discuss three sets of recent methods that have been used to estimate the same model. The first two, Bayesian Additive Regression Trees (BART) (Montgomery and Olivella, 2018; Hill, Weiss and Zhai, 2011) and Random Forests (Hill, Weiss and Zhai, 2011; Breiman, 2001), use an average of trees to predict an outcome. These methods average a group of tree models to estimate the

---

[4]A causal interpretation requires ignorability holds in a continuous, open ball around T for each X and that θ is differentiable in T. This function τ may be of interest in its own right even in a purely descriptive setting.

[5]See Appendix H for the binary treatment case.

conditional mean, $\mathbb{E}(Y|T, X)$, but these tree models do not allow estimation of the effect curve τ.[6]

Two recent methods have modeled the θ fully nonparametrically, in a way that allows us to estimate τ: generalized random forests (GRF, Wager and Athey, 2017; Athey, Tibshirani and Wager, 2019) and Kernel Regularized Least Squares (KRLS Hainmueller and Hazlett, 2013)[7]. KRLS fits a high-dimensional, nonparametric regression to the model, and since its estimated conditional mean is differentiable in T, point estimates for τ are returned. The method returns excellent point estimates across a range of settings, but as we discuss below in Section 3, its confidence intervals can be misleadingly narrow.

GRF uses a tree method, building a tree from the covariates and then running a bivariate regression of the outcome on the treatment at each leaf. The results are aggregated over multiple trees, bootstrap samples, and randomly selected sets of covariates into a random forest. Since, at each point in a given tree, $\hat{\tau}$ is simply a regression coefficient (conditional on being in that leaf), these can be averaged over the forest. It can provide remarkably accurate estimation and inference on an average effect (as in Model 3).

The primary shortcoming here is that the method will miss any nonlinearities at each leaf. As it fits a simple bivariate regression, it will fail to capture nonlinearities that are missed by the regression at each leaf. For example, if the data is generated as $Y = T^2 + e$ where T, e are each independent and standard normal, the generalized random forest estimate $\hat{\tau}(T, X)$ will completely miss the

---

[6]A tree uses the covariates to generate a set of mutually exclusive and exhaustive indicators over the covariate space. For an observation in each region, the estimated value is just the average outcome at of each group ("leaf"), and forests are averages over leaves; for an excellent overview of tree methods, see Montgomery and Olivella (2018). Since the fitted value is just an average of sample means, though, it does not return a differentiable function, so it cannot estimate τ–though it can be used to estimate τ if the treatment is binary, see Hill, Weiss and Zhai (2011); Green and Kern (2012).

[7]For additional work, see (Cattaneo, Farrell and Feng, Forthcoming), though the method does not allow for more than a few covariates.

curvature. We illustrate this below in Section 5. Before describing the MDEI algorithm, which sidesteps the estimation issues described directly above, we discuss issues pertaining to uncertainty estimation.

3   Uncertainty Estimates and Confidence Bands

As discussed above, there are various options for estimating these heterogeneous effects. Our goal is not simply estimating the treatment effect, τ, but also deriving uncertainty bounds with guaranteed theoretic properties that can guide inference.

When conducting inference on a single coefficient in a regression, we use a confidence interval that contains a segment that is centered on the point estimate whose width is governed by the critical value and standard error estimate. When extending to our seminonparametric setting, we use a *confidence band* instead of a confidence interval. In this case, the point estimate and standard error are each replaced by functions, $\hat{\tau}(T, X)$ and $\sqrt{\widehat{\mathrm{Var}}(\hat{\tau})(T, X)}$, and we wish to connect the confidence band to the true curve, $\tau(T, X)$.

We introduce a band that allows the researcher to assess, both graphically and statistically, whether an estimated nonlinearity or interaction likely reflects a systematic relationship in the data. The curve has the *average coverage property* that, asymptotically, the $100 \times (1 - \alpha)\%$ curve will contain the true treatment effect at $100 \times (1 - \alpha)\%$ of the observed data. By this means, the researcher can gauge the likelihood that an uncovered nonlinearity is differentiable from zero.

We offer three advances in constructing this curve. First, our approach uses a data-driven method to estimate a critical value. We build on a recently proposed method, *conformal inference*, which can be used to generate a curve with an average coverage property on predicted values. The method uses the observed data to generate an band on predicted values with an average coverage property. We extend this band to cover not just predicted values, but the true conditional mean

$(\theta(T, X))$ and treatment effect $(\tau(T, X))$. Second, as we are estimating the critical value, we are not relying on a normal approximation to achieve valid coverage. This allows our critical values to reflect the underlying distribution of the data. Third, we show that these bands, while wide at each data point, can be used to recover valid estimates of average treatment effects. As we discuss in this section and illustrate below in Section 5, uncertainty bands returned by existing software can be misleadingly narrow.

Next, we discuss different types of coverage that can be achieved in nonparametric settings. Then, we introduce our approach, which combines repeated cross-fitting to guard against overfitting, a model of the conditional variance to guard against model misspecification and heteroskedasticity, and a conformal approach that provides a data-driven means to select a critical value.

3.1 Types of Coverage

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.[8] 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', X'), the interval will cover the true value at least $(1 - \alpha) \times 100\%$ of the time.[9]

---

[8]For a single parameter $\tau \in \Re$, then, a valid $100 \times (1 - \alpha)\%$ confidence interval can be characterized as

$\lim_{n \to \infty} \sup_{\tau \in \Re} \Pr(\tau \in CI(\widehat{\tau}, \widehat{\mathrm{Var}}(\widehat{\tau}), \mathcal{D}_n, C, \alpha)) \geq 1 - \alpha$

where the standard critical values of $C = 1.64$ and $C = 1.96$ give the 90% and 95% intervals.

[9]For any given τ that can be well-approximated by the model, this interval can be characterized as:

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 the misleading nature of these curves, 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.[10] Uniform nonparameteric confidence bands have been constructed in several settings (e.g., Genovese and Wasserman, 2005; Robins and van der Vaart, 2006). However, in complex settings, they shrink slowly in sample size and are too wide to be usable (see, e.g., Wahba, 1983).

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.[11] 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.[12]

For any given point (T', X'), $\lim_{n \to \infty} \Pr(\tau(T', X') \in CI(\hat{\tau}(T', X'), \widehat{\mathrm{Var}}(\hat{\tau}(T', X'),), \mathcal{D}_n, C, \alpha)) \geq 1 - \alpha$

[10]For any τ that can be well-approximated by the method, this curve can be characterized as $\lim_{n \to \infty} \Pr(\text{For all points } (T', X'), \ \tau(T', X') \in CI(\hat{\tau}(T', X'), \widehat{\mathrm{Var}}(\hat{\tau}(T', X'),), \mathcal{D}_n, C, \alpha)) \geq 1 - \alpha$

[11]This property can be written as:

$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} ((T'_i, X'_i), \ \tau(T'_i, X'_i) \in CI(\hat{\tau}(T'_i, X'_i), \widehat{\mathrm{Var}}(\hat{\tau}(T'_i, X'_i),), \mathcal{D}_n, C(D_n), \alpha)) \geq 1 - \alpha$

[12]Formal derivations of this average coverage can be found in Appendix F.

13

3.2 A Split-Sample and Repeated Cross-Fitting Approach for Valid Inference

In order to recover valid uncertainty estimates, we implement a *split-sample* strategy. This strategy proves crucial in guarding against over-fitting the data and allowing for valid inference. The basic issue is that using the same data to learn a single model of how the covariates affect the outcome, the treatment, and how they moderate the treatment effect is asking too much of a single data set. Machine learning methods, prone to overfitting, will return estimates with nonnegligible bias and too-narrow standard errors (as we illustrate in Section 5).

The split-sample approach cleaves the problem in two: in half the data, we learn how the covariates drive the treatment and the outcome. With these models in hand, we turn to the other half of the data to learn the treatment effect model τ itself. The split-sample approach, though, comes with clear efficiency concerns, as we are only using half the data to model the treatment effect curve τ.

In order to restore efficiency, we follow a *repeated cross-fitting* strategy, recently put forth by Chernozhukov et al. (2018) in the context of the partially linear model, which is our Equation 3. Moving to *repeated cross-fitting* takes two additional steps. First, cross-fitting involves swapping the roles of the estimation and auxiliary sample, generating an additional estimate. This process is then repeated, averaging over multiple cross-fit estimates. Our MDEI method leverages repeated cross-fitting in its implementation.

3.3 Modeling the Standard Errors as a Guard Against Misspecification

We make two additional moves in generating valid bands. The first is to guard against model misspecification through modeling the conditional variance in the confidence band. The second, which we turn to in the next section, is deriving a critical value from the data rather than through

a normal approximation.

Pointwise intervals are valid only under an assumption that the estimated model is consistent for the true curves θ and τ. This assumption is, in fact, easily violated; see Section 5 for an example. To add another layer of protection against requiring that we have our model correct, we also model the conditional variance $\widehat{\mathrm{Var}}(\hat{\tau}(T, X))$.

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.

One consequence of our approach is that our intervals will, in general, be wider than the pointwise intervals returned by other methods (see Appendix F). 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 misleadingly precise if one of the assumptions does not hold.[14] By modeling the error variance, we are able to recover bands that are robust to model misspecification and heteroskedasticity (see Appendix D.4.1).

3.4 From Predictive to Bands with Average Coverage using Conformal Inference Methods

Our first step away from pointwise intervals was to move to modeling the conditional variance as a means of guarding against model misspecification and heteroskedasticity. Our second step is in deriving a data-driven critical value. Instead of relying on a critical value derived from a normal

---

[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.

[14]As discussed by Aronow (2016) (see also van der Vaart (1998), esp. ch. 5), valid standard errors on regression slope parameters can be recovered with only weak distributional assumptions on the error term and outcome.

approximation (C), we instead allow it to be estimated from the data $(C(D_n))$. We thereby generate a band that contains the true value at some controlled portion of the data that uses the empirical distribution of the data to estimate a critical value. Both steps are needed to achieve a band with average coverage; see Section 5 for an illustration.

The inspiration for our approach comes from recent work on conformal inference (Lei and Wasserman, 2014; Lei et al., 2017). 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 without making distributional assumptions about the error terms. The MDEI algorithm extends this predictive interval to one containing the true treatment effect curve, τ.

Constructing the Interval We start with a predictive interval, which is an interval likely to contain a future realization of the outcome at some point, say (T', X'), with some controlled probability. Constructing these intervals normally relies on distributional assumptions on the error term.[15] The insight of the conformal approach comes from using estimated residuals to derive the width of the confidence interval and critical value on a predictive interval. The method is entirely data-driven, and rather remarkably achieves finite-sample coverage rates on predictive intervals.[16]

Our focus, though, is not on unobserved values of Y, but on a bound that will contain τ evaluated over the sample. We use the same basic insight of the conformal method, a data driven approach for a critical value, but construct an interval around τ. We do so in two steps. First, we use the data to estimate a critical value on the bound around θ then, second, we use the same critical value

---

[15]For example, in the linear model, a predictive interval around the point $\widehat{Y}' = X^\top \widehat{\beta}$ is of the form $\widehat{Y}' \pm \sqrt{(X')^\top \widehat{\text{Var}}(\widehat{\beta})(X') + \widehat{\sigma}^2}$ which is an interval relying on asymptotic normality and recovering an accurate estimate of the error variance $\widehat{\sigma}^2$.

[16]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).

to construct a bound around τ.

Our goal is to derive a bound on the approximation error, which is how far the fitted values are from the true value ($|\widehat{\theta}(T', X') - \theta(T', X')|$), such that over the sample we expect the interval to contain the true value some high proportion of the time. We do so by manipulating two other sources of error in our model. The first is the prediction error, which is how far we expect a future value to be from its prediction ($|Y' - \widehat{\theta}(T', X')|$). Conformal methods give us a confidence interval built from this term. The second is the irreducible error, which is how far a future value will be from the true value ($|Y' - \theta(T', X')|$). Through a chain of inequalities, we show that adding one to the critical value from the predictive conformal interval will generate a valid band on the systematic component, θ (see Appendix F and the discussion in Appendix D).

Finally, we are not interested in bounding $\theta(T, X)$ but instead $\tau(T, X)$, its partial derivative with respect to T. Therefore, we add a further level of approximation, such that our interval around $\widehat{\theta}(T, X)$ will contain $\widehat{\tau}(T, X)$, but, again, we are forced to rely on asymptotic argument; see Appendix F. We fully expect the constructed confidence interval to be conservative, meaning the $100 \times (1 - \alpha)\%$ confidence interval will contain the true systematic component at *more than* $100 \times (1 - \alpha)\%$ of the data, even asymptotically. This is because our method is not exact and is constructed through a set of bounds on an asymptotically exact bound. That is, the predictive conformal band of Lei and Wasserman (2014) will hold exactly, in-sample. Our bound takes two steps past this: first, to the systematic value θ, and then to its partial derivative, τ. In doing so, we move from modeling something that is theoretically observable, a future value of Y, to two functions that are not, θ, τ. Generating estimates of θ, τ, and these intervals pose a real challenge, to which we turn next.

4 The MDEI Algorithm

We are now ready to present how we actually estimate and conduct inference on our model, all of which is implemented in open source software. Recall that our model takes the form

$$Y = \theta(T, X) + f(X) + \epsilon, \quad \mathbb{E}(\epsilon | T, X) = 0$$

$$T = g(X) + u, \quad \mathbb{E}(u | X) = 0. \tag{6}$$

In estimating and conducting inference on $\tau = \frac{\partial}{\partial T}\theta(T, X)$, we have three core goals: to allow the instantaneous effect of the treatment to be a nonparametric function of the treatment and interactions with covariates, to adjust for confounders nonparametrically, and to recover uncertainty estimates that allow for inference. We here sketch out the primary pieces. Appendix D provides a detailed step through of the MDEI algorithm. The process involves two distinct stages: a full sample stage and a split-sample stage. The intuition of having two stages is that the full sample stage–where we use all observations–will help to roughly learn the relationships between the outcome and the treatment/covariates. The split-sample stage, where we split the data in half, is used for making specific estimates of parameters and uncertainty around them.

4.1 Full Sample Stage

Using the full set of data, we first want to want to identify a large number of nonlinear transformations of X and T that might help model τ and hence θ. This gives us a model space, rather than committing to a specific model. These transformations are then used in the split-sample stage where we obtain point and uncertainty estimates.

Prior to beginning this step, we partial out individual covariates. We do so by estimating two variables,

$$\tilde{Y} = Y - \mathbb{E}(Y | X); \quad \tilde{T} = T - \mathbb{E}(T | X)$$

18

By this means, we eliminate any systematic relationship $\widetilde{Y}$ and $\widetilde{T}$ that is only a function of the background covariates, i.e. it will involve the treatment variable on its own or moderated by a covariate. Following other work (Chernozhukov et al., 2018), we do this via random forests so as to make extremely minimal assumptions about the relationship between the covariates and the outcome/treatment variables (see Appendix B).

Next we model the treatment effect with nonparametric basis functions, also known as "blending functions," that can be added together to approximate other functions.[17] These capture any complicated interactions between the treatment and covariates, as well as allow the impact of the treatment on its own to be nonlinear. Appendix A provides an introduction to basis functions and our strategy for modeling these.

For an overview, imagine we knew, or suspected, that the outcome had a quadratic nonlinearity in the treatment, so we could simply take the model

$$\theta(T, X) = \widetilde{T} c_1 + \widetilde{T}^2 c_2$$

where c_1, c_2 are slope parameters on the linear and quadratic term of the partialed-out treatment, $\widetilde{T}$. We could then take the partial derivative to recover $\tau(T, X) = c_1 + 2\widetilde{T} c_2$. We of course do not want to restrict ourselves to the consideration of one particular type of nonlinearity (see Appendix B for the minimal functional form assumptions needed for consistency in our estimates) and instead let data rather than model assumptions drive our inferences.

From this additive nonlinear model, we take three additional steps. First, we allow for nonlinear functions of the treatment. We will denote these as $\phi_j(\widetilde{T})$, where each ϕ_j is some transformation, indexed by j, of the covariate. We select a set of basis function transformations which are known to approximate a broad class of possible functions; Appendix A presents details.

If we knew which nonlinear transformation were driving the treatment effect (e.g., quadratic as

---

[17]See Keele (2008) for a helpful introduction for social scientists.

used above), we could model

$$\theta(T, X) = \phi_j(\widetilde{T})c_j$$

with c_j the slope parameter. Taking a partial derivative gives

$$\tau(T, X) = \dot{\phi}_j(\widetilde{T})c_j$$

where

$$\dot{\phi}_j = \partial\phi_j/\partial\widetilde{T}.$$

Second, we may want the covariates to moderate this effect. We could allow for a single moderator, the jth transformation of variable k, denoted $\phi_j(X_k)$, and take the model

$$\theta(T, X) = \phi_j(\widetilde{T})\phi_{j'}(X_{k'})c_{jj'k'}; \quad \tau(T, X) = \dot{\phi}_j(\widetilde{T})\phi_{j'}(X_{k'})c_{jj'k'}.$$

We in fact consider all possible $treatment \times covariate \times covariate$ interactions.[18] These equations capture the idea that we are modeling the treatment effect as a nonlinear function of the treatment interacted with up to two nonlinear interactions of the covariates.

The problem, though, is that we do not know which nonlinear transformations of the treatment and moderators may be driving θ and, hence, τ. Instead, we consider a large set of them. We provide details in Appendix D.2, but for 5 covariates, we consider over $174,000$ possible nonlinear terms and interactions and for 10 covariates, over $615,000$. The full set of these terms allow us to fit a vast number of possible interactions and nonlinearities.

Even with a handful of covariates, the proliferation of candidate basis functions puts us in an extremely high-dimensional space. We reduce this dimensionality in two stages: 1) an iterated Sure

---

[18]In this case of allowing two moderators, we have:

$$\theta(T, X) = \phi_j(\widetilde{T})\phi_{j'}(X_{k'})\phi_{j''}(X_{k''})c_{jj'k'j''k''}; \quad \tau(T, X) = \dot{\phi}_j(\widetilde{T})\phi_{j'}(X_{k'})\phi_{j''}(X_{k''})c_{jj'k'j''k''}.$$

Independence Screen (Fan and Lv, 2008) and 2) a variant of the high-dimensional regression LassoPLUS model (Ratkovic and Tingley, 2017) that we tune for the present context (see Appendix E for details). These steps provide us with a set of candidate basis functions from which a smaller subset will be selected during the split-sample estimation stage.

4.2 Split-sample stage

Armed with a subset of a vast set of candidate basis function transformations learned from the full sample of data, we next split the data into an *auxiliary* and *estimation* sample. The split-sample strategy helps us guard against data dredging by learning a model in the auxiliary sample but then using the estimation sample to conduct inference (van der Vaart, 1998; Athey and Imbens, 2016; Chernozhukov et al., 2018). This split-sample approach is a crucial step in generating proper confidence bands: any overfitting of the machine learning in the auxiliary set will not pollute our inference in the estimation set (see (Chernozhukov et al., 2018) for a recent overview).

Auxiliary sample The auxiliary sample uses the candidate basis functions from the full-sample step in order to learn a plausible specification for θ. As in the full sample step, we partial out the covariates in the auxiliary sample and select bases using the high-dimensional regression model. This provides two things that will be used in the the estimation sample 1) the random forest model learned from the auxiliary sample and 2) the selected bases.

Estimation sample The estimation sample is then used to generate $\widehat{\theta}$, $\widehat{\tau}$, and the confidence band around τ. We first partial out the covariates given the random forest model learned on the auxiliary data. Then we estimate a regression to obtain coefficients on the selected bases that let us obtain point estimates for $\widehat{\theta}$ and $\widehat{\tau}$. We are not just interested in predicting point estimates, but also in getting confidence bands correct. We therefore also use the auxiliary sample to predict the variance at each point, given the value of the treatment and covariates. The square root of this

21

variance is then used to construct our band, at each point. We next use the conformal inference method discussed above to recover a data driven critical value that does not rely on assumptions on our error term (see Appendix D and F for details). In doing so, we extend the work of Lei and Wasserman (2014) from finding a predictive confidence band to constructing a band on our marginal effect curve $\tau(T, X)$.

We then extend our split-sample approach to a *repeated cross-fitting* strategy. This involves two steps: a cross-fitting step that is then repeated. In the cross-fitting step, we alternate the role of the auxiliary and estimation sets, generating two estimates of θ and τ for each split of the data. We then repeat this process of randomly splitting the data and cross-fitting multiple times, averaging over our results; see Appendix D for details.

4.3   Discussion

MDEI is able to recover a nonparametric treatment effect with point estimates competitive with several cutting-edge methods (see below and Appendix G). Furthermore, we provide a principled approach to uncertainty estimation. We are able to recover a confidence band that contains the true value over at least $100 \times (1 - \alpha)\%$ of the data points, asymptotically. In generating a curve that is valid under minimal assumptions on the functional form and error distribution, we expect it to be conservative. Decreasing the width of these bands could be done, but at the cost of additional assumptions on the data. We show in our applied example below that these bands are still narrow enough to be used for inference in applied work. In Appendices D–F we provide extensive technical details for interested readers.

5   Illustrative Simulation

Next we illustrate the role of repeated cross-fitting and conformal inference in achieving average coverage. We also compare MDEI intervals to those returned by to two cutting edge alternative al-

gorithms. We show that MDEI performs well in terms of point estimation and, equally importantly, returns intervals that relay information on where the true systematic curve is likely to lie.[19]

Illustrating the role of split-sample/repeated cross-fitting and conformal approach In the first set of illustrative simulations, we show how repeated cross-fitting and the conformal approach can lead to valid average coverage. Our simulation setting has three important characteristics. First, we include a nonlinear confounder that must be learned from the data. Second, the treatment effect is moderated by one of the covariates in a discontinuous fashion, making it a challenge for our method to uncover. Third, we include heteroskedastic errors in order to illustrate the method's coverage even with systematic components in the error variance.

In each setting, we draw five covariates from a standard multivariate normal equicorrelated at 0.5. From the first, X_{i1}, we generate a new variable $s_i = \text{sgn}(X_{i1}) \in \pm 1$. 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:

$$T_i = \frac{(X_{i2} - 1)^2}{4} + u_i; \; u_i \sim \mathcal{N}(0, 1) \tag{7}$$

$$Y_i = 2s_i T_i^2 + \frac{(X_{i2} - 1)^2}{4} + \epsilon_i; \; \epsilon_i \sim \mathcal{N}\left(0, \frac{1}{1 + X_{i2}^2}\right). \tag{8}$$

This gives us a target of $\tau(T_i, X_i) = 4s_i T_i$ for which we want to use our interval to conduct inference. We vary the sample size, $n \in 100, 250, 500, 1000, 2500, 5000, 10000$ and simulations were run 400 times each.

Results appear in Figure 1. For each run of the simulation, we calculate a 90% interval, and assess what proportion of the data is contained in the constructed interval. 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

---

[19]Appendix G provides a complete set of Monte Carlo performance simulations.
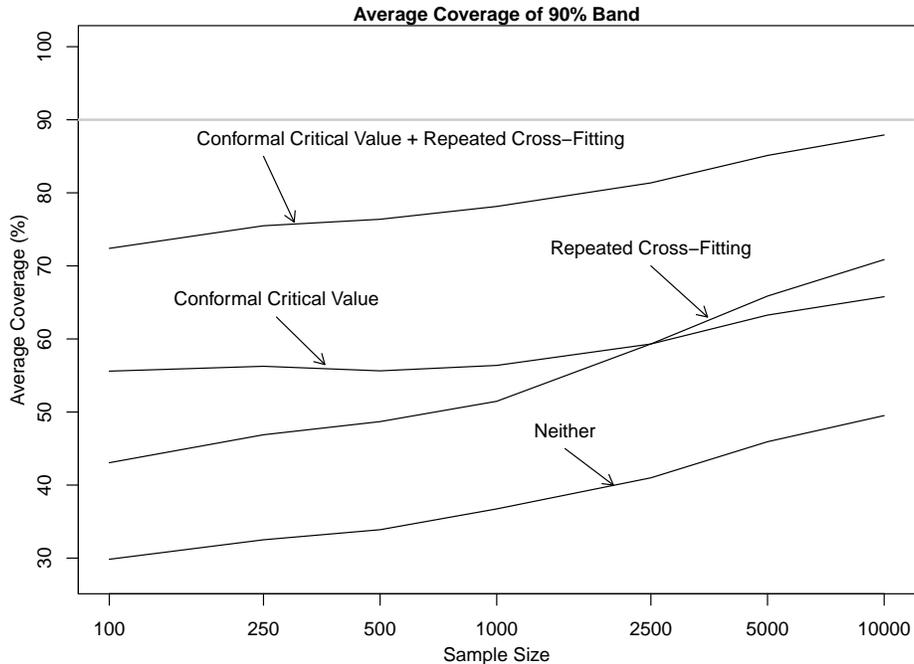
23

Figure 1: **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. The conformal critical value increases coverage, but only when combined with repeated cross-fitting does coverage approach nominal.

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 pointwise interval will be quite narrow, not actually containing the true curve in the majority of the data. The conformal critical value serves to recover bands closer to the expected average coverage, but only when combined with repeated cross-fitting do we find coverage close to what is expected. The result is clearly asymptotic, but even in small sample sizes, our 90% band achieves coverage of $> 80\%$ in this complex setting, suggesting our method can be used in settings with realistic sample sizes.

Illustrating the difference between pointwise and average coverage Next, we compare the pointwise interval returned by existing methods to those with average coverage, returned by
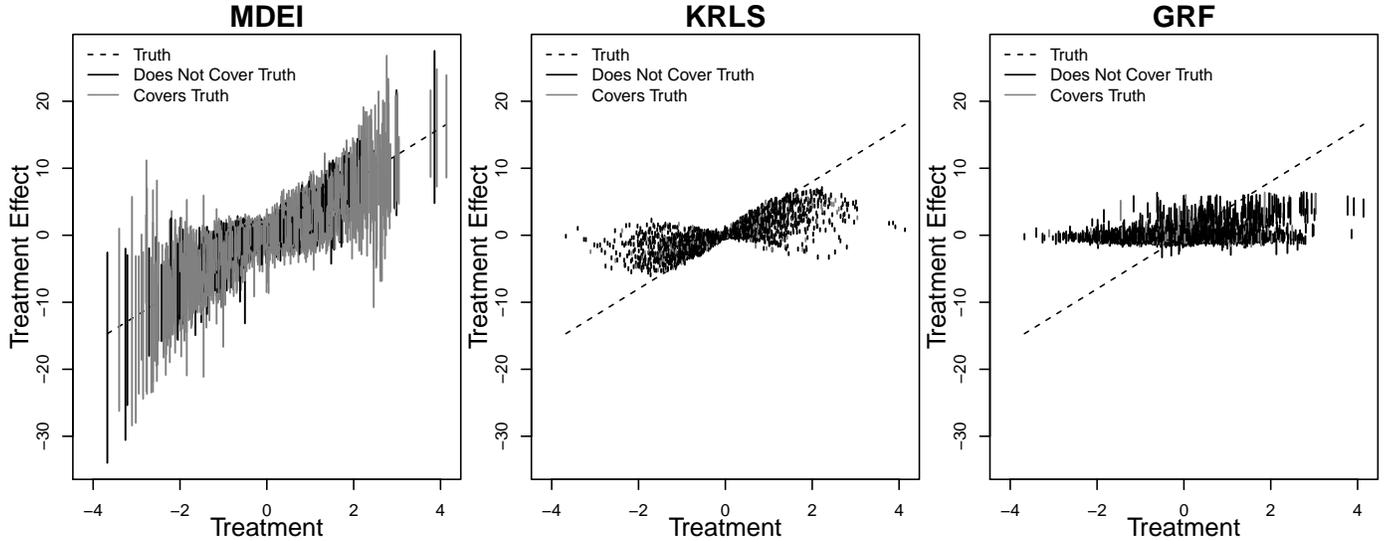
Figure 2: **Treatment effect estimates across MDEI, KRLS and GRF algorithms.** The dashed line represents the true treatment effect, with results by method in each plot. Vertical lines represent the 90% uncertainty intervals from each method. Only MDEI is able to capture the trend in the treatment effect, while the uncertainty estimates from other methods are narrow.

MDEI. We show that the pointwise intervals can be so narrow as to generate misleading inference. Using the same data generation process as before and a sample size of $5,000$, we focus on the contrast between MDEI[20] and both Kernel Regularized Least Squares ($KRLS$, (Hainmueller and Hazlett, 2013)) and Generalized Random Forests (GRF, (Athey et al., 2019)). Both can estimate the marginal effect curve directly.

Figure 2 displays the results, plotting the treatment effect curve (τ) against its estimate and interval, by method. We present results for $s_i = 1$; they are qualitatively similar to $s_i = -1$. 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 84.6% of the data, in this example. For KRLS and GRF, those numbers are 14.7% and 12.2%. The root-mean square error across the methods reveal a similar pattern (MDEI: 1.34, KRLS: 1.98, GRF: 3.31). GRF uses a split sample approach, while

---

[20]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.

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. MDEI, on the other hand, returns point estimates and an uncertainty band that convey the basic nonlinearity in the underlying data as well as containing the true curve about as frequently as would be expected.

6 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 linear regression specification, the authors estimate an impact of approximately 7 percentage points on the Social Democrats' vote share.

Applying MDEI to the same data with the same control set, we find a smaller–while still statistically significant–effect. Here, 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).

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 H for estimation details in this binary setting). Estimates of the 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,

26

	Point Estimate	95% CI
Original Regression	6.91	(5.43, 8.40)
Trimmed Regression	4.87	(2.96, 6.78)
MDEI	4.61	(4.19, 5.03)
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.

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 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, with MDEI providing the narrowest confidence interval.

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, as we are only using observations that were in fact treated when estimating the treatment effect on the 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 3. We begin with their analysis (see Figure 5 in the original paper),[21] which we present in the left-hand panel of Figure 3. 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

---

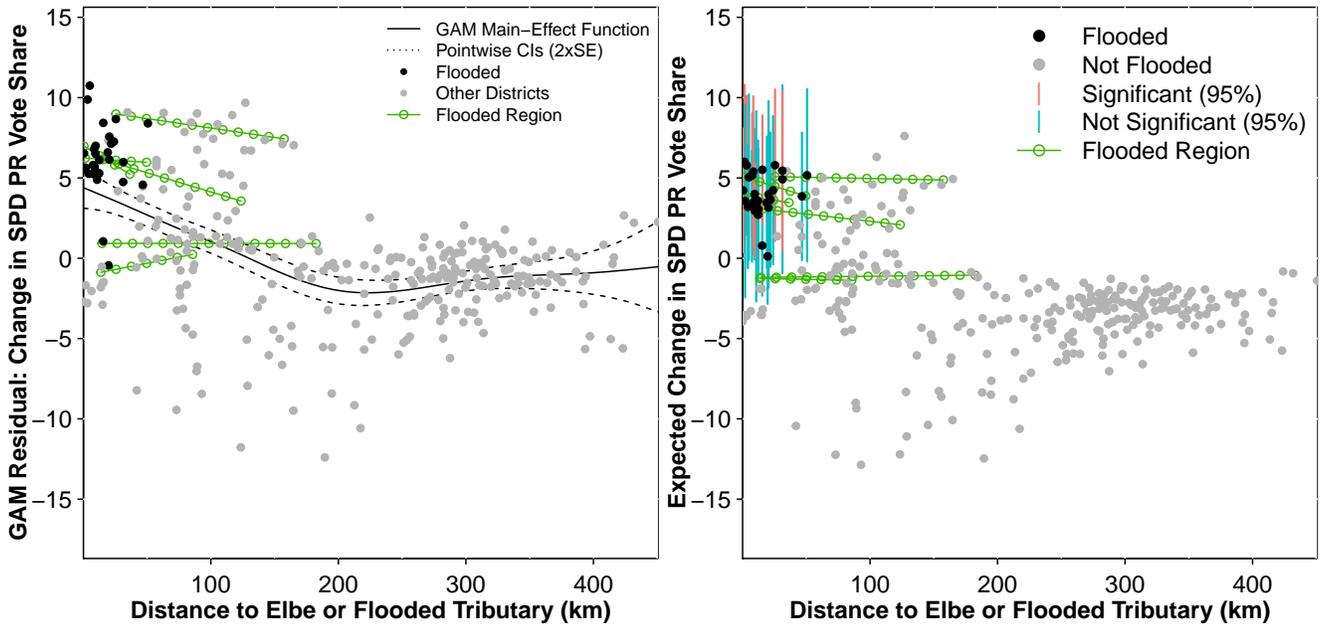[21]We combine both halves of their Figure 5 into one plot for parsimony.

Figure 3: **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.

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 4 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
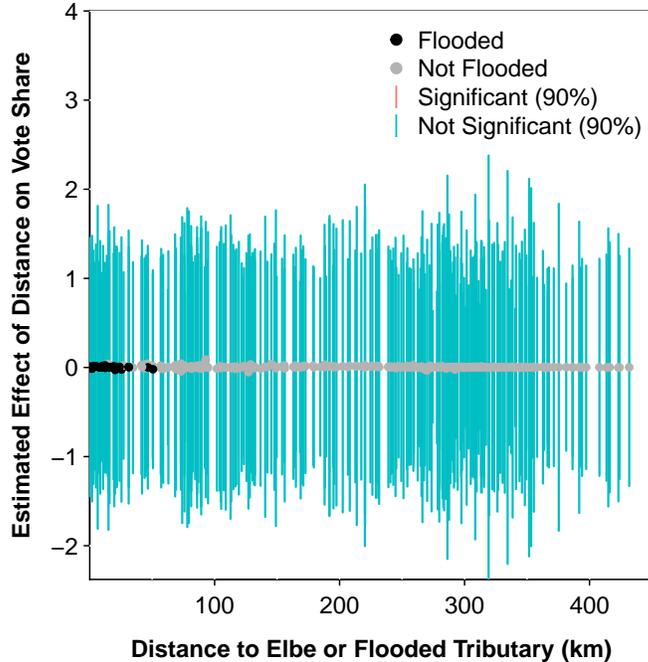
Figure 4: **The Effect of Distance to Elbe on Vote Share.** The point estimate and confidence interval of the marginal effect on distance to the Elbe on support for Social Democrats, by observation. Point estimates for districts in flooded regions are in black, the rest gray. Distance has no discernible effect on support for Social Democrats.

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.

7 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 treatment 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).

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

Aronow, Peter. 2016. "A Note on "How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It."" arXiv Manuscript.

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

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

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

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

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

Chen, Xiaohong. 2007. *Handbook of Econometrics.* Vol. 6 Part B Elsevier chapter Large Sample Sieve Estimation of Semi-Nonparametric Model, pp. 5549–5632.

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.

Diamond, Alexis and Jasjeet Sekhon. 2012. "Genetic Matching for Estimating Causal Effects." *Review of Econonics and Statistics* .

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.

Farrell, Max, Tengyuan Liang and Sanjog Misra. forthcoming. "Deep Neural Networkds for Estimation and Inference." *Econometrica* .

Fong, Christian, Chad Hazlett and Kosuke Imai. 2018. "Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements." *The Annals of Applied Statistics* 12(1):156–177.

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

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.

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.

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.

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.

Kennedy, Edward H., Scott A. Lorch and Dylan S. Small. In Press. "Robust causal inference with continuous instruments using the local instrumental variable curve." *Journal of the Royal Statistical Society: Series B* .

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.

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.

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 G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani and Larry Wasserman. 2017. "Distribution-Free Predictive Inference for Regression." arXiv:1604.04173v2.

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

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.

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.

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

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.

van der Vaart, Aad. 1998. *Asymptotic Statistics.* Vol. 3 of *Cambridge Series in Statistical and Probabilistic Mathematics* Cambridge University Press.

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.

Online Appendix for "Estimation and Inference on Non-linear and Heterogeneous Effects"

The online appendix for this article contains several sections. Appendix A gives an introduction to using basis functions for regression modelling. Appendix B discusses the minimal functional form assumptions made for our model to return consistent estimates. Appendix C discusses pointwise bands and contrasts them to the uniform bands we use in the paper for uncertainty estimation. Appendix D gives a detailed exposition of the MDEI algorithm in order to compliment Section 4. Appendix E discusses the specific sparse regression model employed in our estimation steps. Appendix F derives for our variance estimation. Appendix G gives extensive performance simulation evidence leveraging a variety of different data generating processes and comparing MDEI to other relevant methodologies. Finally, Appendix H discusses the case of a binary or categorical treatment variables rather than the continuous treatment variable context.

Appendix

Table of Contents

A   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 5. 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.[22]

The middle row shows the true systematic component for the outcome and treatment effect and, below it, a set of basis functions that we use to approximate the curve. MDEI uses 23 for each covariate: the linear covariate and then B-splines of degree 3, 4, 5, and 6. 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.

---

[22]For completeness, T is uniform on $[-1, 1]$ with $n = 250$ and the outcome is $\theta(T) = 4\pi T \sin(2\pi T)$ and $\tau(T) = 4\pi \sin(2\pi T) + 8\pi^2 T \cos(2\pi T)$ with normal, mean-zero noise with variance 2.
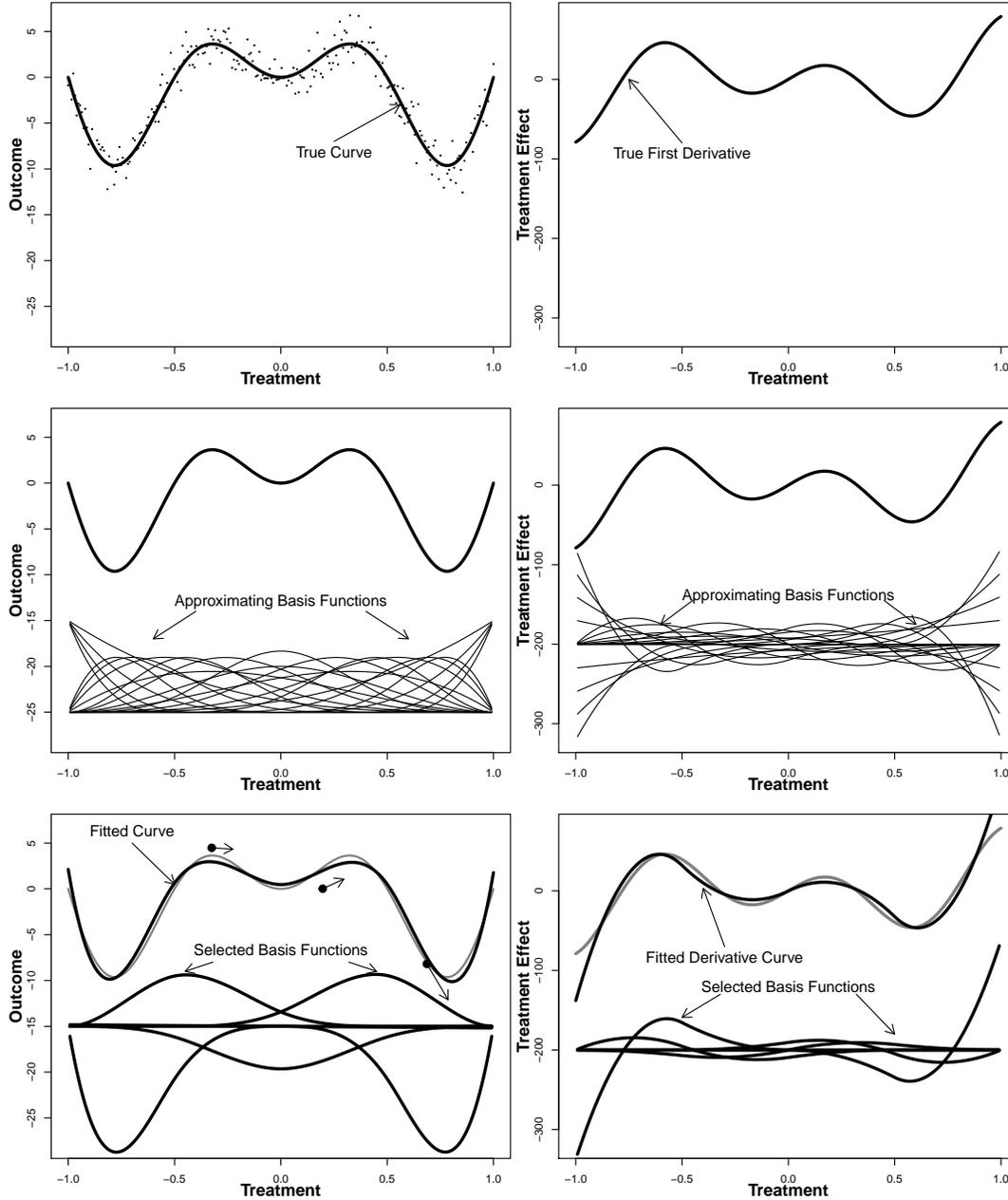
Figure 5: **Combining Basis Functions to Model Nonlinearities.** This figure illustrates how MDEI uses basis functions to approximate an outcome. The left column contains results for the outcome θ and the right for the treatment effect τ. For simplicity, we assume both are solely a function of the treatment. 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.

B   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 (see Section 4).

B.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 K elements of X, 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

$$f(X) = X^\top \gamma$$

with K elements in γ. We are assuming, then, that f lives in the space of functions linear in the elements of X denoted X_k,

$$\{f : f(X) = \sum_{k=1}^{K} \phi_k(X) c_k; \ \phi_k(X) = X_k\}$$

The *basis functions* of a space, which we denote $\{\phi_k\}_{k=1}^{K}$, are a set of functions which can be combined to represent any function in the space. In the parametric setting, the basis functions $\{\phi_k\}_{k=1}^{K}$ are simply individual covariates $\{X_k\}_{k=1}^{K}$. In the nonparametric case discussed below, the basis functions can be more complicated but still combine together to represent some target function like $f(X)$, 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.[23]

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 X.

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, K_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.

---

[23]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 A provides a comprehensive introduction to regression modelling using basis functions in the present context.

B.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

$$\{f : f(X) = X^\top \gamma\}.$$

We may want to consider the space

$$\{f : f(X) = \sum_{k=1}^{K_n} \phi_k(X)c_k\},$$

which is similar to the linear space, except we allow the number of bases K_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 $\{c_k\}_{k=1}^\infty$ 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.[24] Below we consider two different classes of functions. We will use the first to model our treatment 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.[25]

B.2.1 Modeling $\theta(T, X)$ and $\tau(T, X)$

We first consider modeling $\theta(T, X)$, 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 $\phi_j(T, X)$ to be bounded. 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

$$\Theta = \{\theta : \theta(T, X) = \sum_{j=1}^\infty \phi_j(T, X)c_j; \phi_j(T, X) \text{ and } \frac{\partial}{\partial T}\phi_j(T, X) \text{ bounded}; \sum_{j=1}^\infty |c_j| < \infty\}.$$

---

[24]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.

[25]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 τ.[26]

Importantly, this subspace subsumes the linear model.[27] 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 5. 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.

B.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*.[28] 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 treatment effects, which we operationalize as a derivative in the case of continuous treatment variable.

C Pointwise Confidence Bands

A second, and subtly different idea of coverage, is the pointwise band. This band is asymptotically valid at every point,

$$\inf_{\substack{\tau \in \mathcal{T} \\ f,g \in Lipschitz(1) \\ H \in \mathcal{H}}} \lim_{n \to \infty} \Pr_{\mathcal{D}_n}(\tau(T', X') \in CB_{\tau; \mathcal{D}_n}(T', X') \ \forall \ (T', X') \in \mathcal{T} \times \mathcal{X})) \geq 1 - \alpha \qquad (9)$$

which is the uniform band except with the lim and the inf reversed. This is the form of inference most commonly encountered and easiest to establish; see, e.g., Theorem 5 of Athey et al. (2019). This confidence band is constructed to be asymptotically valid at every single point. This means, given a point (T', X'), we would expect the pointwise confidence band to contain the true value $\tau(T', X')$ at this point at least $100 \times (1 - \alpha)\%$ of the time, over repeated samples.

---

[26]
$$\mathcal{T} = \{\tau : \tau(T, X) = \sum_{j=1}^{\infty} \frac{\partial}{\partial T}\phi_j(T, X)c_j; \phi_j(T, X) \text{ and } \frac{\partial}{\partial T}\phi_j(T, X) \text{ bounded}; \sum_{j=1}^{\infty} |c_j| < \infty\}.$$

[27]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.

[28]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

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

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

C.1 Pointwise versus Uniform Bands

The uniform band is to be preferred (e.g., Wasserman, 2006). The argument is subtle but worth explaining. A uniform band ensures some statistical property over the whole of the curve. An asymptotically uniform band, then, guarantees there exists a sample size n_0 such that for all sample sizes $n > n_0$, the band is valid. The pointwise band offers no such guarantee over the whole of the curve, but just that the band is valid at every point.

For example, a $100 \times (1 - \alpha)\%$ uniform band is asymptotically valid if it contains the entire true curve across $100 \times (1 - \alpha)\%$ of repeated samples, in the limit. A $100 \times (1 - \alpha)\%$ pointwise coverage if, at every single point, it contains the true value at this point over $100 \times (1 - \alpha)\%$ of repeated samples, in the limit.

These two claims 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.

D Overview of Algorithm

D.1 Overview

In estimating and conducting inference on τ, we have three core goals: to adjust for confounders nonparametrically, to allow the instantaneous effect of the treatment to be a nonparametric function of the treatment and interactions with covariates, and to recover uncertainty estimates that allow for inference. Given these challenging goals, we must use a number of steps including: a split-sample approach (to avoid over-fitting), partialing out covariates via a random forest (to deal with confounding nonparametrically), a screening step and regression model with nonparametric basis functions (to capture any complicated interactions between the treatment and covariates, as well as to allow the impact of the treatment on its own to be nonlinear), conformal inference methods (to recover a uniform confidence band), and finally repeated cross-fitting (to reduce sensitivity to the particular splits of the data we use at each stage). In what follows we lay out our model of the treatment effect and then work through each of these steps. Appendix D provides a formal algorithm diagram.

D.2 The Implemented Treatment Effect Model

Recall the model that we are estimating is

$$Y = \theta(T, X) + f(X) + \epsilon, \quad \mathbb{E}(\epsilon | T, X) = 0$$
$$T = g(X) + u, \quad \mathbb{E}(u | X) = 0.$$

and our target is the function $\tau(T, X) = \frac{\partial}{\partial T} \theta(T, X)$. As discussed above, doing so requires making modeling choices for each component. We assume that f, g are in a broad class of functions estimated consistently by a random forest, denoted $Lipschitz(1)$. We place θ in a class of bounded,

differentiable functions–so that we can recover τ–such that we can estimate it consistently through a regression model with nonparametric basis functions.

Before proceeding with our exact implementation procedure we first describe the model of τ. We will use a set of nonparametric bases, denoted ϕ_j, where each is a nonlinear function of its input. We use a specific set of bases used in smoothing splines, the B-spline basis, which are a set of bounded, nonlinear transformations commonly used to model complex data.[29] Specifically, we use degree 3, 4, 5, and 6 degree B-spline bases, for a total of 23 bases. We draw on smoothing splines for a very specific reason: we want to allow for the possibility that $\tau(T, X)$ is non-linear in T. For example, the effect of the treatment could be quadratic but given our seminonparametric approach we will not pre-specify that possibility in advance and simply let the data determine the relationship.

We then interact this set of nonparametric bases, denoted $\{\phi_j\}_{j=1}^{23}$, where each basis beyond the first is a nonlinear function of its input. We interact these in order to accommodate functions that are nonlinear, possibly even nonmonotonic, and interactive in the covariates.[30] Because we will be partialing out the X's, we model θ in terms of $\widetilde{T} = T - \widehat{\mathbb{E}}(T|X)$.

Then, for elements of $X = [X_1, X_2, \ldots, X_k, \ldots, X_K]$, we model all *treatment $\times$ covariate $\times$ covariate* interactions, giving us the model for θ:

$$\theta(\widetilde{T}, X) = \widetilde{T}\gamma_0 +$$
$$\sum_{k=1}^{K} \sum_{k'=k}^{K} \sum_{j=1}^{24} \sum_{j'=1}^{24} \sum_{j''=j'}^{24} \phi_j(\widetilde{T})\phi_{j'}(X_k)\phi_{j''}(X_{k'})c_{kk'jj'j''}. \tag{10}$$

Here, the terms $c.$ are simply parameters in a regression model attached to each of the basis functions.

We estimate θ directly, but are able to infer τ by taking the partial derivative with respect to $\widetilde{T}$. This gives us a model for τ as

$$\tau(T, X) = \gamma_0 + \sum_{k=1}^{K} \sum_{k'=k}^{K} \sum_{j=1}^{24} \sum_{j'=1}^{15} \sum_{j''=j'}^{24} \frac{\partial}{\partial \widetilde{T}} \phi_j(\widetilde{T})\phi_{j'}(X_k)\phi_{j''}(X_{k'})c_{kk'jj'j''}. \tag{11}$$

D.3   Full Sample Estimation Stage

Having stated our model of τ we are now ready to work through our estimation procedure. The first stage uses the full sample of data. We want to identify a large number of nonlinear transformations of X and T that we can use to model θ and, consequently, τ. Our goal in this preliminary step is to simply generate a large set of bases that might help model the true θ and, hence, τ. We are not committing to a model here, but a model *space*, i.e., we are simply trying to recover a set of several hundred basis functions that might be informative for the outcome. We then use these basis functions in the subsequent estimation stages. We proceed next with each of the full sample steps.

Full Sample Step 1: Partialling out covariates   Our first step is to partial out the covariates from the treatment and the outcome. To do this, we calculate the expected value of Y and T given the covariates X and then subtract them from the observed values to produce two sets of residuals.

$$\widetilde{Y} = Y - \widehat{\mathbb{E}}(Y|X); \quad \widetilde{T} = T - \widehat{\mathbb{E}}(T|X) \tag{12}$$

---

[29]Appendix A provides a more comprehensive introduction to regression modelling using basis functions.

[30]If $K = 5$ there will be over $175,000$ potential basis functions and $K = 10$ generates over $6,900,00$. This explosion generates one of our greatest estimation challenges. These bases constitute a larger space than those considered by earlier multivariate nonparametric regression models that inspire our work; see Stone et al. (1997); Friedman (1991) for seminal work and Ravikumar et al. (2009) for recent work.

Throughout, when we partial out the covariates, we use a random forest to model the conditional expectation of Y and T (Breiman, 2001; Montgomery and Olivella, 2018).[31] A random forest is particularly helpful in doing this because it is flexible, fast, and does not make strong assumptions about the conditional mean. We will directly model these residuals having partialed out the covariates.

Full Sample Step 2: Iterated SIS Screen We next turn to selecting a set of bases we can use to model θ, τ.

SIS Screen Iteration 1 The first, screening, step is designed to get us from a potentially vast number of bases to a more manageable number. As originally presented in Fan and Lv (2008), the Sure Independence Screen (SIS) selects a subset of bases that have the highest correlation with the outcome. Their method assumed the data are jointly normally distributed, in which case the correlation is an appropriate measure of the relationship between an outcome and basis. We produce a more stable and reliable screening procedure by using a measure of correlation that is robust to outliers and peculiarities of the full sample.[32] In this process, we select $20 \times (1 + n^{1/5})$ bases that have the highest robust correlation with $\widetilde{Y}$.[33]

Intermediate Regression Step A sparse regression (e.g., LASSO) is then fit using the bases returned from our modified SIS screen and the outcome and treatment variables that had been previously residualized by partialling out the covariates (i.e., $\widetilde{Y}, \widetilde{T}$). To do so, we take our outcome $\widetilde{Y}$, and use our Bayesian regression model to regress it on the bases selected in this first stage of the screen (denoted $bases_1$), producing residuals

$$\widetilde{Y}^{SIS1} = \widetilde{Y} - \widehat{\mathbb{E}}^{LASSO}(\widetilde{Y}|bases_1), \tag{13}$$

which we enter into the next SIS iteration.

SIS Screen Iteration 2 As noted in the original work of Fan and Lv (2008), iterating the screening process is effective can improve our estimation accuracy. We use the values from the previous step, $\widetilde{Y}^{SIS1}$, to select a new set of one hundred bases. To do this, we repeat the process as above, except with $\widetilde{Y}$ replaced by $\widetilde{Y}^{SIS1}$. This additional set of $20 \times (1 + n^{1/5})$ gives us one more chance to incorporate relevant bases that may have been missed in the first screen, producing $40 \times (1 + n^{1/5})$ bases.

Summary Our iterative SIS process produces $40 \times (1 + n^{1/5})$ bases for $\theta(T, X)$ that have the property that they have a stable correlation with the outcome. We will then use these bases below

---

[31]We use the implementation from the **R** package `randomForest`. We chose this implementation due to its speed and ubiquity.

[32]To do so, first, we split the data in half, measure the correlation between outcome and basis on each half, and repeat this process five times. We then only consider bases that have the same sign on the correlation with the outcome across all splits. Second, having dropped bases with unstable relationships, we select those where the median value of the correlation across the splits is as large as possible in magnitude. In this way, we are able to guard against correlations that are not robust across splits of the data, and guarantee that the median correlation is still sizeable.

[33]It is important to note that, in our implementation, we are implicitly assuming that the true function θ can be captured by our one hundred recovered bases. This is not a theoretical requirement–we could have the number of bases grow in sample size–but instead an implementation decision that we find works well in practice. If we wish to recover a true sparse model, we need a "beta-min" assumption on the smallest true value. We do not require consistency in variable selection, but consistency in prediction, which requires the much weaker assumption that the sum of the L_1 norm of the mean parameters is finite; see Buhlmann and van de Geer (2013, ch. 6).

for the actual estimation and inference steps.[34][35]

D.3.1 Full Sample Summary

From this full sample portion, we recover a set of candidate basis functions that we will use to model θ, τ below. Our aim was not to select, but to screen–meaning getting the number of bases from hundreds thousands to low hundreds. We then take these low-hundreds set and further select a useful subset (auxiliary sample below) and use these to construct a confidence interval (estimation sample below).

D.4 Split-sample Estimation Stage

In order to avoid any accidental bias due to the over-fitting of our our machine learning method, we implement a *split-sample* strategy. To do so, we first split the sample into two equally sized splits: an *auxiliary sample* and an *estimation sample*. This split-sample approach serves as a crucial guard against data dredging, since we use the auxiliary sample of the data to learn models of the covariates, but the estimation sample to conduct inference (see, e.g., van der Vaart, 1998; Athey and Imbens, 2016; Chernozhukov et al., 2018). This provides a guard such that any incidental over-fitting on the auxiliary split will not carry through to our calculations on the estimation split.

In particular, we split the sample in two, and use the first half to adjust for background covariates and to learn a plausible specification for θ. We then turn to the second half to estimate $\widehat{\theta}$ and, consequently, $\widehat{\tau}$. We then use this split of the data to construct a confidence band around $\widehat{\tau}$. We estimate the standard errors and select a critical value such that the interval contains $100 \times (1-\alpha)\%$ of the observed data.

D.4.1 Auxiliary Sample

Auxiliary Sample: Partialing out covariates using a random forest Using the auxiliary sample, we again partial the covariates out of the treatment and outcome:

$$\widetilde{Y} = Y - \widehat{\mathbb{E}}^{aux}(Y|X); \widetilde{T} = T - \widehat{\mathbb{E}}^{aux}(T|X) \tag{14}$$

where the *aux* subscript denotes that the expectation was estimated on the auxiliary sample. We want to partial out the covariates from the outcome and treatment so we can estimate θ without worrying about the confounding functions f, g. Doing so is also a crucial part of generating narrow confidence bands. Note that since we are using the auxiliary sample to partial out the covariates, the models do not depend on data from the estimation sample, which will allow us to recover the confidence interval.

Auxiliary Sample: Selecting Bases with Sparse Regression Given the nonparametric bases that survived the screen, we next use a sparse regression model to select a subset for modeling θ. A *sparse regression* is a regression that returns point estimates of zero for all but a subset of covariates; we use a slight adaptation the method presented in Ratkovic and Tingley (2017).Appendix E discusses the particulars of the model.[36] Sparse regression offers the ability to select a subset of relevant predictors and estimate their coefficients at the same time. Doing so, though, introduces

---

[34]Since this is our final iteration of the screen, we do not need to regress the residualized outcome $\widetilde{Y}^{SIS1}$ onto the bases selected at this stage.

[35]We pass through to the auxiliary and estimation sample stages the linear interactions between the treatment variable and covariates. Below when we estimate our quantities of interest using our high dimensional regression model, we still allow these terms to be zeroed out. That is, we always leave them under consideration rather than screening at this full sample stage.

[36]For a more complete discussion of sparse regression methods, see the discussion in Ratkovic and Tingley (2017) and Buhlmann and van de Geer (2013).

a non-negligible bias–again, learning the model and estimating its coefficients off one data set can introduce a bias that makes inference challenging.

To avoid this problem we use the sparse regression in the auxiliary sample to select a subset of nonparametric bases. We will take these selected bases from the auxiliary sample and apply them to the estimation sample in order to estimate coefficients and confidence bands. The important part here is that we separate learning a subset of bases and estimating their coefficients; doing so serves a crucial role in producing valid confidence bands.

Auxiliary Sample: Modeling the Error Variance   One of our central contributions is introducing not just a flexible model of the treatment effect τ but also confidence bands on its estimate. We proceed in two steps. First, we generate an upper bound on $\mathbb{E}((\hat{\tau} - \tau)^2)$. A central part is considering the case where $\hat{\tau}$ does not converge to τ, due to model misspecification of some type. We denote the function to which $\hat{\tau}$ *does* converge as $\tilde{\tau}$, and incorporate this into our bounds. Second, we describe a means to estimate it. This bound, as we describe below, has an important asymptotic property: if the model for τ is not properly specified, we expect average coverage of our interval to be nominal; if it is properly specified, we expect average coverage to be more than nominal but still tight.

Our strategy relies on generating a bound on our estimation error on θ and then recovering one for τ. Denote as $\tilde{\theta}$ the limiting value of $\hat{\theta}$,

$$\tilde{\theta}(T, X) = \lim_{n \to \infty} \hat{\theta}(T, X), \tag{15}$$

i.e., what we would expect $\hat{\theta}$ to converge to with an infinite amount of data.

An important element of this construction is that we are using the conformal band to recover a conservative band on the true value. This band *could* be shortened by making, say, parametric assumptions on the functional form or distributional assumptions on the error term. Instead, we are striving for the narrowest possible bounds that will work regardless of the true model, so as not to outright rely on these assumptions.

Our construction follows two steps; we place the details in Appendix F. First, we start with a valid conformal band. Specifically, given a future realization Y_i', we can generate a bound on $Y_i' - \hat{\theta}(T_i, X_i)$, such that an interval centered on the fitted value will contain the future realization with some controlled probability. Next, we decompose this future observed value into its systematic and random component, $Y_i' = \theta(T_i, X_i) + \epsilon_i'$. Rearranging terms, we are able to bound the gap between the conditional mean, $\theta(T, X)$, and its estimate $\hat{\theta}(T, X)$, as a function of the error variance. Briefly, we show that if the conformal bound is c at this point, a value of $c+1$ will generate a bound that contains the true value $\theta(T, X)$ with the same probability. Third, we then take derivatives and generate a similar bound on $\tau(T, X)$, adding one to the critical value.

D.4.2   Estimation Sample

From the auxiliary sample, we have gathered: random forest models for how the covariates affect the treatment and outcome; a set of bases selected by the sparse regression for modeling θ; and models for the error variance components. We now take these elements and turn to the estimation sample to construct an estimate.

Estimation Sample: Partialing out Covariates   We take the random forest model from our auxiliary sample and use it to partial out the covariates in the estimation sample. Denote as *est* observations in the estimation sample and *aux* a model fit on the auxiliary sample, this becomes

$$\tilde{Y}^{est} = Y^{est} - \widehat{\mathbb{E}}^{aux}(Y^{est}|X)$$

$$\tilde{T}^{est} = T^{est} - \widehat{\mathbb{E}}^{aux}(T^{est}|X)$$

We then use these values throughout.

Estimation Sample: Estimating Coefficients We are now ready to estimate $\widehat{\theta}$ and, consequently, $\widehat{\tau}$. Turning to the estimation sample, we regress $\widetilde{Y}^{est}$ on the bases evaluated at $(\widetilde{T}^{est}, X)$. Importantly, $\widetilde{Y}^{est}$ and $\widetilde{T}^{est}$ have had X partialed out using a model fit to the auxiliary data but evaluated on the estimation sample. Similarly, we use the estimation sample to estimate coefficients on the selected bases (we used the auxiliary sample to select these bases). Regressing $\widetilde{Y}^{est}$ on the selected bases gives us the coefficients we need to estimate $\widehat{\theta}, \widehat{\tau}$. It also gives us estimates of the error variance and residuals that we use to estimate the sampling variance.

Estimation Sample: Estimating the Error Variance Estimating the error variance for $\widehat{\theta}$ and $\widehat{\tau}$ follows a similar procedure to that for the coefficients. By using the models from the auxiliary sample, we predict the variances on the estimation sample.

Estimation Sample: A Conformal Method We are almost ready to estimate our confidence interval on $\widehat{\tau}$. We have reasonable estimates for every component save the critical value C, which controls how many standard errors above and below the estimate we extend our confidence band. In order to estimate C, we use a simplified version of a data-driven means to select C. We adopt the *conformal inference* approach (see discussion in Section 3 and Lei et al. (2017) for an overview) to generate exact, or nearly exact, confidence intervals that do not rely on normality assumptions. This allows us to recover asymptotically valid confidence bands, yet still have them not rely on distributional assumptions on the error term.

We use a conformal approach to select a critical value such that the symmetric confidence band on $\widehat{\theta}(T, X)$ contains $100 \times (1 - \alpha)\%$ of the data:

$$\widehat{C}_{1-\alpha/2} = \inf\{C : \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left(Y_i \in CI(\widehat{\theta}(T, X), \widehat{E}((\widehat{\theta} - \theta)^2)|T, X), \mathcal{D}_n, \widehat{C}, \alpha)\right) \geq 1 - \alpha\} \quad (16)$$

We then carry this critical value from the outcome to the treatment effect and recover confidence interval

$$CB_{\widehat{\tau}, \mathcal{D}_n} = \widehat{\tau}(T, X) \pm \widehat{C}_{1-\alpha/2}\sqrt{\widehat{s}_\tau^2(T, X) + \widehat{\sigma}_\tau^2(T, X)}. \quad (17)$$

This conformal approach offers us a principled, and data-driven, means to select a critical value on our confidence band. The value is selected to guarantee average coverage, asymptotically, such that we expect the true curve τ to fall in the band at about $100 \times (1-\alpha)\%$ of the observed data. This provides a useful tool, mathematically and visually, for characterizing uncovered heterogeneities in our data.

D.5 Repeated Cross-Fitting

One concern might be that the procedure is inefficient since we are splitting the data. Further, it might be sensitive to which splits of the data we are using at each stage. To ameliorate these concerns, we engage in *repeated cross-fitting*. After using the auxiliary sample to select covariates and the estimation sample to estimate them, we switch their roles, cross-fitting the model to these two splits. We then repeat this process and average over repetitions.

D.6 Summary

In sum, we present a method that offers both advantages and disadvantages. The advantage is that we are able to recover a nonparametric treatment effect with point estimates competitive with several cutting-edge methods. More important, from our point of view, is the uncertainty

estimation. We are able to recover a confidence band that contains the true value over at least $100 \times (1 - \alpha)\%$ of the data points, asymptotically.

We are able to achieve this while making minimal assumptions over the functional form of the treatment effect or the error distribution. This comes at a real cost: our bands are conservative for two separate reasons. First, we make only mild functional form or distributional assumptions. These bands can be narrowed without making functional form or distributional assumptions, but in this manuscript we strive for maximal agnosticism. Second, the bands are constructed from bounds on asymptotically exact (i.e. non-conservative) bands, so we expect slippage through the bounding. At the same time, any shrinkage on these intervals would have to come through assumption.

D.7 Algorithm Diagram

In this section, 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 the first two to construct our estimates.

Algorithm 1: Generating Candidate Bases

Data: Outcome vector Y, treatment vector T, covariate $N \times (K+1)$ covariate matrix X with intercept in first column

Functions: Spline functions ϕ_{jm} the jth B-spline of degree m, standardized.
$\rho(a, b)$ the absolute value of the median of the correlation of a and b taken over 5 splits of half the data.

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

Using the full sample, generate $\widetilde{Y} = Y - \widehat{\mathbb{E}}(Y|X), \widetilde{T} = T - \widehat{\mathbb{E}}(T|X)$ using random forests ;
Standardize $\widetilde{Y}, \widetilde{T}$

for *j in 1 to (K+1)* **do**
    for *j′ in 1 to (K+1)* **do**
        for *d in {3,5,7,9}* **do**
            for *d′ in {3,5,7,9}* **do**
                for *m in 1 to d* **do**
                    for *m′ in 1 to d* **do**
                        Save $\rho(\widetilde{Y}, \widetilde{T} \times \phi_{md}(X_j) \times \phi_{m'd'}(X_{j'})$

Return indices corresponding with bases with top $20(1 + n^{.5})$ 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 mean critical value

Algorithm 3: Estimation Algorithm

Data: Outcome vector Y, treatment vector T, covariate $N \times (K + 1)$ covariate matrix X
> with intercept in first column, indices $\mathcal{I}$ for interactive splines, split sample $\mathcal{S}_0, \mathcal{S}_1$

Result: Fitted values, first derivative, variance estimates, and uniform confidence interval
Generate bases $\mathcal{I}$ from Algorithm 1;
for *1 in 1 to numsplits* **do**
> Generate $\widetilde{Y} = Y - \widehat{\mathbb{E}}^{\mathcal{S}_0}(Y|X), \widetilde{T} = T - \widehat{\mathbb{E}}^{\mathcal{S}_0}(T|X)$ using random forests but a model fit
> using $\mathcal{S}_0$;
> Construct heterogeneity bases from $\mathcal{I}$;
> Using data from $\mathcal{S}_0$ select a subset of the bases using a sparse regression; Using data
> from $\mathcal{S}_0$ model conditional variance by regressing the squared errors on covariates
> using random forests; Regress $\widetilde{Y}_i$ on selected bases on split $\mathcal{S}_1$;
> Evaluate point estimate, first derivative, and variance of fitted and first derivative at
> each point in $\mathcal{S}_1$ using selected bases ;
> Generate predicted values of error variance at each point;
> Switch roles of $\mathcal{S}_0, \mathcal{S}_1$ and cross-fit. Save fitted values, first derivative estimates,
> variance estimates.

Calculate uniform confidence intervals using confidence interval using Algorithm 2.

E 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,

$$Y_i|X_i, \beta, Z_i, b\sigma^2 \sim \mathcal{N}(X_i^\top \beta, \sigma^2) \tag{18}$$
$$\beta_k|\lambda, w_k, \sigma \sim DE\left(\lambda w_k/\sigma\right) \tag{19}$$
$$\lambda^2|N, K \sim \Gamma\left(\alpha, \rho\right) \tag{20}$$
$$w_k|\gamma \sim \text{generalizedGamma}(1, 1, \gamma) \tag{21}$$
$$\gamma \sim \exp(1) \tag{22}$$
$$\tag{23}$$

Taking $\alpha = \sqrt{N}K$ gives the model of Ratkovic and Tingley (2017). In that model, the goal was consistent model recovery, meaning zeroing out irrelevant covaraites and maintaining relevant

ones. This required selecting α so that the model would successfully distinguish between the two types. We instead adjust the prior on λ to $\alpha = n \log p$, which will optimize prediction, rather than model recovery. They differ in that, for prediction, we are more willing to tolerate irrelevant variables so long as they help improve prediction. Setting λ to the "oracle rate" $\lambda = O\left(\sqrt{n \log p}\right)$ will minimize the model's predictive error (see, e.g. Buhlmann and van de Geer, 2013, ch. 6). In this manuscript, we are analyzing predicted values, not selected variables. To this end, we select our tuning parameter for optimal prediction.

F   Variance Derivation

F.1   Deriving the Conformal Bound

We assume we have a valid conformal bound. Then, for some future value Y_i' with covariates X_i, variance at this point $\widehat{\sigma}_{\widehat{\theta}}(X_i)$, and critical value $\widehat{C}_{1-\alpha/2}$, we get

$$\Pr_X(|Y_i' - \widehat{\theta}(X_i)| \le \widehat{C}_{1-\alpha/2}\widehat{\sigma}_{\widehat{\theta}}(X_i)) \ge 1 - \alpha. \tag{24}$$

Now, we bound the inequality inside the probability from the left using $|a + b| - |b| \le |a|$ where $a + b = \widehat{\theta}(X_i) - \theta_i$, $a = Y_i' - \widehat{\theta}(X_i)$ $b = \theta_i - Y_i'$. So, with a conformal band, we can ensure the following event occurs with probability at least $1 - \alpha$

$$\widehat{C}_{1-\alpha/2}\widehat{\sigma}_{\widehat{\theta}}(T_i, X_i) \ge |Y_i' - \widehat{\theta}(T_i, X_i)'| \tag{25}$$
$$\ge |\widehat{\theta}(T_i, X_i) - \theta_i| - |\theta_i - Y_i'| \tag{26}$$

and rearranging, then bounding the term on the right gives

$$|\widehat{\theta}(T_i, X_i) - \theta_i)| \le \widehat{C}_{1-\alpha/2}\widehat{\sigma}(T_i, X_i) + |\theta_i - Y_i'| \tag{27}$$
$$\le (\widehat{C}_{1-\alpha/2} + 1)\widehat{\sigma}_{\widehat{\theta}}(T_i, X_i) \tag{28}$$

since $\widehat{\sigma}(X_i) \ge |\theta_i - Y_i'|$, in expectation.

Thus, we should expect the confidence band

$$CB_{\widehat{\theta}, \mathcal{D}_n} = \widehat{\theta}(T_i, X_i) \pm (\widehat{C}_{1-\alpha/2} + 1)\widehat{\sigma}_{\widehat{\theta}}(T_i, X_i) \tag{29}$$

to have at least $100 \times (1 - \alpha)\%$ average coverage of the systematic component $\theta(T, X)$. 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 $\theta(T_i, X_i)$. 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

$$CB_{\widehat{\tau}, \mathcal{D}_n} = \widehat{\tau}(T_i, X_i) \pm (\widehat{C}_{1-\alpha/2} + 1)\widehat{\sigma}_{\widehat{\tau}}(T_i, X_i) \tag{30}$$

We estimate the variance using the law of total variance

$$\underbrace{\widehat{\sigma}_{\widehat{\theta}}^2(T_i, X_i)}_{\text{Total Variance}} = \underbrace{\widehat{s}_{\widehat{\theta}}^2(T_i, X_i)}_{\text{Sampling Variance}} + \underbrace{\widehat{\sigma}_{\theta}^2(T_i, X_i)}_{\text{Error Variance}} \tag{31}$$

14

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 $\hat{\tau}$ using the same formula. The sampling variance can, again, be calculated from the cross-fit sample variance over the estimates. For the second variance term, we estimate the variance of Y attributable to T, but not X, which we estimate as

$$\hat{\sigma}^2_{\hat{\tau}(T,X)} = \widehat{\text{Var}}^{aux}(Y|X) - \widehat{\text{Var}}^{aux}(Y|T,X). \tag{32}$$

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

G   Performance Simulations

G.1   Data Generating Processes

We next present simulation evidence illustrating MDEI's utility in estimating treatment 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 $X_{i1}, \ldots, X_{i5}$ from a standard multivariate normal with correlation 0.5.

In the first four settings, we take

$$T_i = \frac{(X_{i2} - 1)^2}{4} + \epsilon_i^T; \quad \epsilon_i^T \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1). \tag{33}$$

In the fifth setting, we introduce a discontinuity by using

$$T_i = \text{sign}(X_{i1}) \times \frac{(X_{i2} - 1)^2}{4} + \epsilon_i^T; \quad \epsilon_i^T \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \tag{34}$$

where

$$\text{sign}(a) = \begin{cases} -1; & a \leq 0 \\ 1; & a > 0 \end{cases} \tag{35}$$

We then use the following outcome models,

$$\text{1 Linear: } Y_i = T_i + X_{i1} + \frac{X_{i2} - 1}{4} + + \epsilon_i \tag{36}$$

$$\text{2 Partially Linear: } Y_i = T_i + X_{i1} + \frac{(X_{i2} - 1)^2}{4} + \epsilon_i \tag{37}$$

$$\text{3 Additive Linear: } Y_i = 4\sin(T_i) + X_{i1} + \frac{(X_{i2} - 1)^2}{4} + \epsilon_i \tag{38}$$

$$\text{4 Interactive: } Y_i = 4\sin(T_i) \times X_{i1} + \frac{(X_{i2} - 1)^2}{4} + \epsilon_i \tag{39}$$

$$\text{5 Discontinuity } Y_i = 4\sin(T_i) \times \text{sign}(X_{i1}) + \frac{(X_{i2} - 1)^2}{4} + \epsilon_i \tag{40}$$

where the the error is independent, identical Gaussian such that the true R^2 in the outcome model is 0.5. We consider $n \in \{250, 500, 1000, 2000\}$. The discontinuous model is the one reported in the body of the paper and falls outside the space covered by our model.

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 marginal effect curve, τ.[37]

G.1.1 Evaluation Metrics

We assess methods across two dimensions, each commensurate with our two estimation contributions: point estimation and coverage rates on τ. For the former, we use the mean absolute error,

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{\tau}(T_i, X_i) - \tau(T_i, X_i)| \tag{41}$$

and the sample average coverage probability,

$$SACP = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1} \left\{ \tau(T_i, X_i) \in CB_{\tau, \mathcal{D}_n}(T_i, X_i) \right\}.$$

We note that we expect to get good coverage on the marginal effect, but this will come at the cost of conservative (i.e. wide) confidence intervals.

All simulations were run 200 times.

G.1.2 Results

Results from the simulations can be found in Figure 6. 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 treatment 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 way 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.

---

[37]We also do not include other candidate approaches. Many existing models focus on uncertainty and estimates for the fitted values, not the marginal 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.
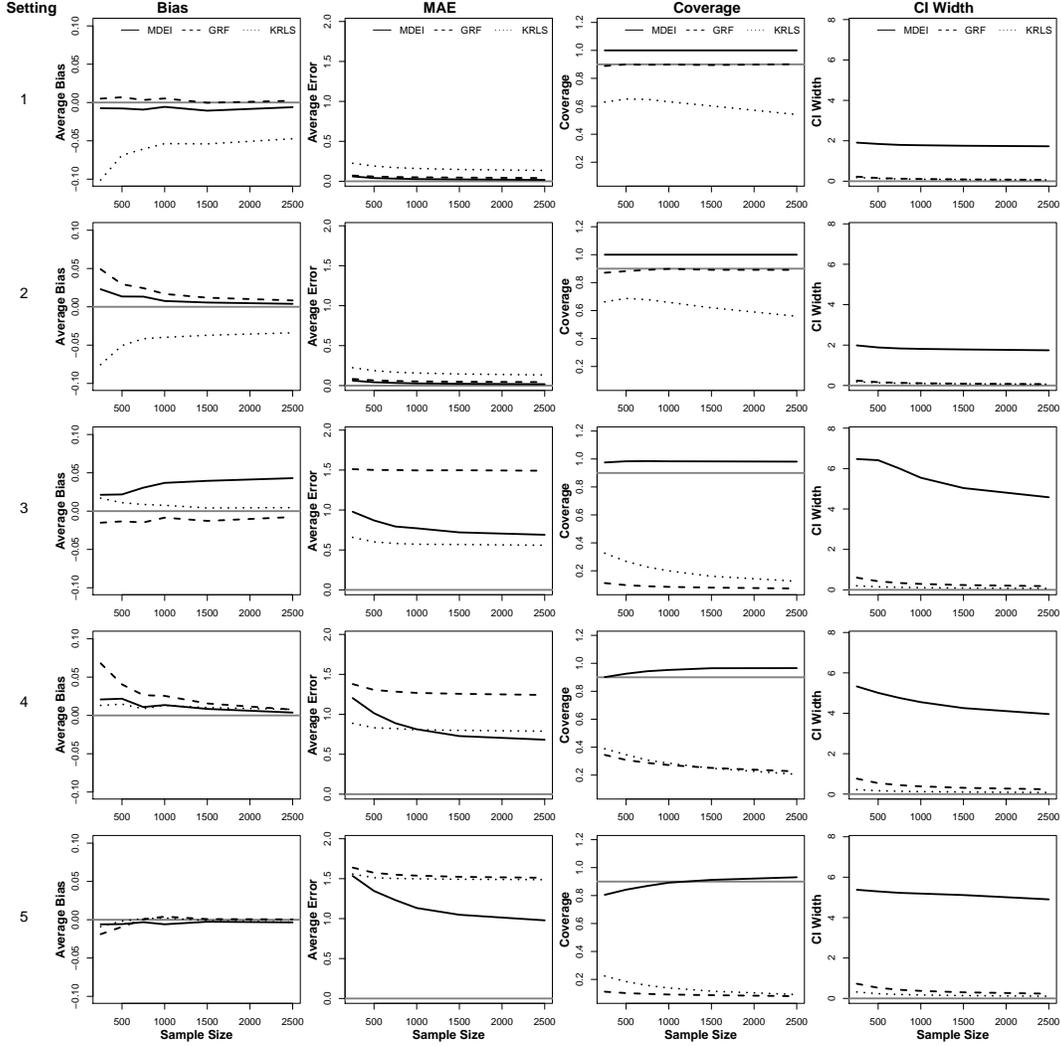
Figure 6: Performance Simulation Results

H   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.

References

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

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.

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

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.

Friedman, Jerome H. 1991. "Multivariate adaptive regression splines." *The Annals of Statistics* pp. 1–67.

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.

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

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

Lei, Jing, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani and Larry Wasserman. 2017. "Distribution-Free Predictive Inference for Regression." arXiv:1604.04173v2.

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

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.

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.

van der Vaart, Aad. 1998. *Asymptotic Statistics*. Vol. 3 of *Cambridge Series in Statistical and Probabilistic Mathematics* Cambridge University Press.

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