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

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

A Regression Model Background

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

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

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

$$
\begin{aligned}
y_i &= \theta t_i + \mathbf{x}_i^\top \gamma + \epsilon_i, \quad \mathbb{E}(\epsilon_i | t_i, \mathbf{x}_i) = 0 \\
t_i &= \mathbf{x}_i^\top \beta + u_i, \quad \mathbb{E}(u_i | \mathbf{x}_i) = 0,
\end{aligned}
\tag{1}
$$

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 more advanced settings, modeling the effect of a one-unit move in the treatment on the outcome will require a flexible model of the treatment variable itself. If we are willing to make narrow assumptions such as linearity and homogeneity in the partial effect, an outcome model alone will do; as we want extend this model into more general settings such as the partially linear model, we will considering the treatment assignment model as well.

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

$$
\begin{aligned}
y_i &= \theta t_i + \mathbf{x}_i^\top \gamma + \epsilon_i, \quad \mathbb{E}(\epsilon_i | t_i, \mathbf{x}_i) = 0 \\
t_i &= g(\mathbf{x}_i^\top \beta) + u, \quad \mathbb{E}(u_i | \mathbf{x}_i) = 0.
\end{aligned}
\tag{2}
$$

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.[1]

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

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

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

$$\begin{aligned} y_i = \theta t_i + f(\mathbf{x}_i) + \epsilon, \;\; \mathbb{E}(\epsilon_i|t_i, \mathbf{x}_i) = 0 \\ t_i = g(\mathbf{x}_i) + u_i, \;\; \mathbb{E}(u_i|\mathbf{x}_i) = 0. \end{aligned} \tag{3}$$

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* (Chernozhukov et al., 2018; Hardle and Stoker, 1989; Härdle et al., 2012; Robinson, 1988), since it is linear in t_i but nonparametric in the remainder.

The partially linear model improves on the standard practice captured in Model 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 the covariates nonparametrically. We can relax this assumption, generating a type of *generalized additive model* (Wahba, 1990; Hastie and Tibshirani, 1990; Wood, 2006; Beck and Jackman, 1998), where we replace θt_i with $\theta(\widetilde{t}_i)$, a smooth function of the treatment. Doing so generates the model

$$\begin{aligned} y_i = \theta(\widetilde{t}_i) + f(\mathbf{x}_i) + \epsilon, \mathbb{E}(\epsilon_i|\mathbf{x}_i, t_i) = 0; \\ t_i = g(\mathbf{x}_i) + u_i, \mathbb{E}(u_i|\mathbf{x}_i) = 0; \;\; f, g, \theta \text{ unknown} \end{aligned} \tag{4}$$

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).

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

$$y_i = \theta(g(\mathbf{x}_i) + u_i) + f(\mathbf{x}_i) + \epsilon, \mathbb{E}(\epsilon_i|\mathbf{x}_i, t_i) = 0; \tag{5}$$

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

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

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

$$\begin{aligned} y_i = \theta(\widetilde{t}_i, \mathbf{x}_i) + f(\mathbf{x}_i) + \epsilon_i, \;\; \mathbb{E}(\epsilon_i | t_i, \mathbf{x}_i) = 0 \\ t_i = g(\mathbf{x}_i) + u_i, \;\; \mathbb{E}(u_i | \mathbf{x}_i) = 0. \end{aligned} \tag{6}$$

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

We want to use this model to learn about how the treatment impacts the outcome. In the most simple regression model that we started with, this was just the average partial effect of t_i on the outcome t_i, which was the slope coefficient θ. We want to do something similar, but in our context the slope is not a simple constant. Instead, we want θ to depend on both the value of the treatment and covariates (i.e., $\theta(\widetilde{t}_i, \mathbf{x}_i)$). That is, the impact of the treatment can be both nonlinear and moderated by the covariates.

In the continuous treatment case (see Appendix I for discussion of the binary case) our target of inference is the first partial derivative of the outcome with respect to the treatment, given background covariates: $\tau(\widetilde{t}_i, \mathbf{x}_i) = \frac{\partial}{\partial t}\theta(\widetilde{t}_i, \mathbf{x}_i)$. This estimand captures the impact of a ceteris paribus perturbation of the treatment on the outcome.[2] Modeling this local confounding returns an partial effect, an effect for an observation at treatment level t_i and covariate profile $\mathbf{x}_i$ (see e.g., Stolzenberg (1980) for an explicitly causal interpretation). This effect can be a curve that varies across values of the treatment and can depend on values of the covariates.[3]

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

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

B Introduction to basis functions

We employ a nonparametric regression model that models the outcome as an additive sum of a large number functions of the covariates, called *basis functions*. Each basis is a nonlinear transformation of a covariate, allowing us to model a more flexible set of functions.[4] We illustrate this approach in Figure 1. 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.[5]

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

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

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

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

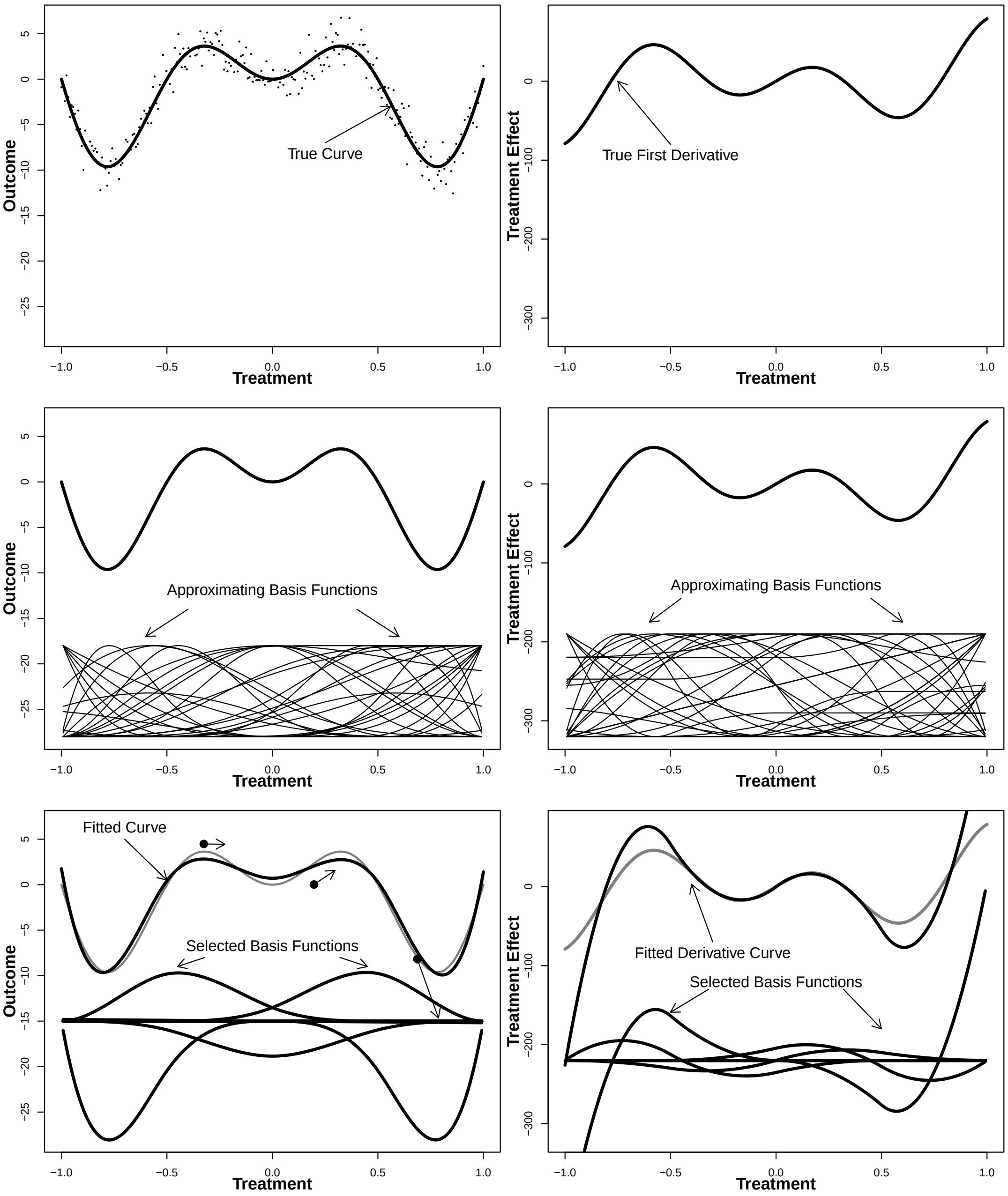

Figure 1: **Combining Basis Functions to Model Nonlinearities.**This figure illustrates how MDEI uses basis functions to approximate an outcome. The left column contains results for the outcome θ and the right for the partial effect $\tau(\widetilde{t}_i, \mathbf{x}_i)$. For simplicity, we assume both are solely a function of the treatment. The top row shows the true curves to be estimated, with the data in the left. The middle middle row illustrates the full set of basis functions that we use to approximate the true curve and its first partial derivative. The bottom row shows the basis functions that were selected, such that when added up they produce the fitted outcome (left) and derivative (right). A set of random points are selected with the estimated derivative at each point in the left figure. Rather than assuming a functional form, MDEI uses sets of basis functions that can be combined and interacted to provide an accurate local prediction.

C Function Classes

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

C.1 Moving beyond parametric functions

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

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

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

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

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

The *basis functions* of a space, which we denote $\{\phi_j\}_{j=1}^{p}$, are a set of functions which can be combined to represent any function in the space. In the parametric setting, the basis functions are simply individual covariates $\{\mathbf{x}_{ij}\}_{j=1}^{p}$. In the nonparametric case discussed below, the basis functions can be more complicated but still combine together to represent some target function like $f(\mathbf{x}_i)$, or, more generically, a conditional mean or even density. As discussed below, different types of approaches to constructing basis functions and estimating their parameters will be used.[6]

[6] For example, cubic smoothing splines are often used to model time (e.g., Ratkovic and Eng, 2010) and other

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

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

C.2 Functions for the treatment versus covariates

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

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

We may want to consider the space

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

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

While certainly more flexible than the linear space, we cannot use a finite dataset to estimate an infinite number of parameters (Gyorfi et al., 2002). We can, though, retain a growing or infinite number of basis functions, a nonparametric space, if we constrain the function space in some

continuous covariates (Keele, 2008). Appendix B provides a comprehensive introduction to regression modelling using basis functions in the present context.

manner. Perhaps the simplest example is the "sparsity assumption" that only some finite subset of the parameters $\{c_j\}_{j=1}^{\infty}$ are not zero, formulating the problem as one where the true model is parametric but we just do not know which basis functions constitute the true model (Buhlmann and van de Geer, 2013; Belloni, Chernozhukov and Hansen, 2014).

We implement methods that do not require the sparsity assumption, but constrain the function space so we can still fit models much more complex than a linear model while also recovering a consistent estimate.[7] Below we consider two different classes of functions. We will use the first to model our partial effect, and it will consist of smooth, differentiable functions of the treatment and covariates interacted together. We will fit this component using a high-dimensional regression model. The second class we consider is, roughly, functions that we can approximate well using a random forest. We will use this class to model any confounding or bias introduced from the covariates.[8]

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

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

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

[7] 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.

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

Taking the partial derivative, we can get the space containing $\tau(\widetilde{t}_i, \mathbf{x}_i)$.[9]

Importantly, this subspace subsumes the linear model.[10] For example, if the true model were parametric and linear in basis functions, we would recover the parametric model.

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

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

We turn next to modeling how the covariates affect the outcome and treatment, as represented by the functions f, g respectively. Here, we simply assume that these functions can be well-estimated using a random forest, which places them in a *Lipschitz space*.[11] 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 × covariate interactions,

[9]

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

[10] 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.

[11] 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

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

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

since Lipschitz functions are continuous but need not be differentiable. We add more structure to the space where we look for partial effects, which we operationalize as a derivative in the case of continuous treatment variable.

C.2.3 Consistency

We establishing consistency in our for the curve $\tau(\widetilde{t}_i, \mathbf{x}_i)$ by appealing to a sparsity condition in this class of models. We rely on the condition in Chernozhukov et al. (2018) Remark 4.3, which requires that the number of bases required to approximate the true curve is much less than sample size; formally if $s_{theta,n}$ is the number of bases needed to approximate our θ function uniformly, then we require $s_{\theta,n} << n$.[12] Under these conditions, then $\widehat{\theta}(\widetilde{t}_i, \mathbf{x}_i)$ and hence $\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i)$ is consistent. The B-spline bases will give us consistency in $L_1(P)$ for functions of the form, with p the number of covariates,

$$\begin{aligned}\theta(\widetilde{t}_i, \mathbf{x}_i) =& \widetilde{t}_i\gamma_0 + \\ & \sum_{k=1}^{p}\sum_{k'=k}^{p}\sum_{j=1}^{27}\sum_{j'=1}^{27}\sum_{j''=j'}^{27} \phi_j(\widetilde{t}_i)\phi_{j'}(\mathbf{x}_{ik})\phi_{j''}(\mathbf{x}_{ik'})c_{kk'jj'j''}; \quad \sum \left|c_{kk'jj'j''}\right| < \infty \end{aligned} \tag{7}$$

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

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

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

repeated a number of times. The retained bases are then brought to the estimation subsample and used for estimating $\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i)$, as described in Section 2 of the body.

D Coverage and Pointwise Confidence Band Concepts

Uncertainty intervals, like confidence intervals and confidence bands, are designed to have specific coverage properties. For example, in the linear model with a single slope coefficient, then the coverage probability is the proportion of times, over repeated samples that the interval contains the true value.[13] This is the standard confidence interval, as taught in the context of the regression and other parametric models.

There are multiple ways to achieve coverage on a nonparametric curve. Coverage can be achieved pointwise, uniformly, and on average. We turn to each in turn. The first, and most commonly encountered, is the *pointwise confidence interval.* This is the one returned by existing software (e.g., Hainmueller and Hazlett, 2013; Wager and Athey, 2017; Athey et al., 2019). In this case, at *any given* point $(t_i, \mathbf{x}_i)$, the interval will cover the true value at least $(1-\alpha) \times 100\%$ of the time.[14] This pointwise property carry through to averages via a central limit theorem.

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

The second type, the *uniform confidence band*, will contain the *whole* curve $(1-\alpha) \times 100\%$ of the time over repeated sampling. This band *does* contain information on the whole curve, since it will contain the full curve over repeated samples.[15] Uniform nonparameteric confidence bands have been constructed in several specific settings (e.g., Genovese and Wasserman, 2005) but cannot, in

[13] For a single parameter $\tau \in \Re$, then, a valid $100 \times (1-\alpha)\%$ confidence interval given data D_n and critical value C can be characterized as

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

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

[14] For any given $\tau(\widetilde{t}_i, \mathbf{x}_i)$ that can be well-approximated by the model, this interval can be characterized as:

For any given point $(t_i, \mathbf{x}_i)$, $\lim_{n\to\infty} \Pr(\tau(\widetilde{t}_i, \mathbf{x}_i) \in CI(\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i), \widehat{\mathrm{Var}}(\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i)), \mathcal{D}_n, C, \alpha)) \geq 1-\alpha$

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

general, be constructed. Even when feasible they shrink slowly in sample size and are too wide to be usable (see, e.g., Wahba, 1983).

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

We move onto our proposed band which implements the third type of coverage, average coverage. Rather than relying on claims across repeated samples, we instead follow Nychka (1988) (see also Wasserman (2006) ch. 5.8) and consider *average coverage*, which is the probability that a confidence band contains the true value over the observed sample.[16] 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.[17]

E Technical Details for Algorithm

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

E.1 Algorithm Diagram

We outline three algorithms that we use to implement our method. In the first, Algorithm 1, we generate a set of bases that model heterogeneity. The second, Algorithm 2, details how we construct

[16] This property can be written as:

$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(\tau(\widetilde{t}_i, \mathbf{x}_i) \in CI(\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i), \widehat{\mathrm{Var}}(\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i)), \mathcal{D}_n, C(D_n), \alpha)) \geq 1 - \alpha$

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

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 $y_i$, treatment vector $t_i$, length $p$ coveriate vector $\mathbf{x}_i$ with intercept
        in first column, all in the discovery subsample, $n_0$ observations in this subsample
  Functions: 28 basis functions denoted $\phi_j wherebydefault \phi_0(z) = 1, \phi_1(z) = z$
  $\rho(a, b)$ the correlation of $a$ and $b$.
  Result: A set of indices generating nonparametric bases for modeling treatment ×
          covariate interactions
  Using the discovery subsample, generate $\widetilde{y}_i = y_i - \widehat{\mathbb{E}}(y_i|\mathbf{x}_i), \widetilde{t}_i = t_i - \widehat{\mathbb{E}}(t_i|\mathbf{x}_i)$ using random
   forests
  for $j$ in 1 to $p$ do
      for $j'$ in 1 to $p$) do
          for $d$ in 2 to 28 do
              for $d'$ in 1 to 28 do
                  for $d'\prime$ in $d'$ to 28 do
                      Save $\rho(\widetilde{y}_i, \phi_d(\widetilde{t}_i) \times \phi_{d'}(\mathbf{x}_{ij}) \times \phi_{d''}(\mathbf{x}_{ij'}))$
  Return indices corresponding with bases with top $min(100(1 + n_0^2, n_0/4))$ values of $\rho$
```

Algorithm 2: Generating Critical Value

```
Data: Matrix of fitted values and estimated standard errors at each point;
False positive rate $\alpha$
Result: Fitted values, first derivative, variance estimates, and uniform confidence interval
for i in 1 to numsplits do
    Find smallest critical value such that symmetric confidence interval contains $100 \times (1 - \alpha)\%$ of the data
Return critical value for $\widetilde{y}_i$ and add one to generate critical value for $\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i)$.
```

Algorithm 3: Estimation Algorithm

```
Data: Outcome $y_i$, treatment $t_i$, covariates $\mathbf{x}_i$, indices $\mathcal{I}$ for selected interactive splines, data split into discovery/estimation/inference subsamples.
Result: Fitted values, first derivative, variance estimates, and uniform confidence interval
for 1 in 1 to numsplits do
    Using the discovery subsample, generate partialed-out outcome $\widetilde{y}_i$, partialed-out treatment $\widetilde{t}_i$, and retained bases bases $\mathcal{I}$ from Algorithm 1;
    Using data from the estimation subsample, model $\widehat{\theta}(\widetilde{t}_i, \mathbf{x}_i)$ from regressing $\widetilde{y}_i$ on the retained bases using a sparse regression; Using data from the estimation subsample, model conditional variance by regressing the squared errors on covariates using random forests; Evaluate point estimate, first derivative, and variance of fitted and first derivative at each point in the inference subsample;
    Cross-fit until fitted values, first derivative estimates, variance estimates generated for every datum
Calculate conformal confidence intervals using confidence interval using Algorithm 2.
```

F Sparse Regression Model

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

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

For completeness, we present the full model hierarchy,

$$\begin{aligned} y_i|\mathbf{x}_i, \beta &\sim \mathcal{N}(X_i^\top \beta, \sigma^2) && (8)\\ \beta_k|\lambda, w_k, \sigma &\sim DE\left(\lambda w_k/\sigma\right) && (9)\\ \lambda^2|n, p &\sim \Gamma\left(\alpha, \rho\right) && (10)\\ w_k|\gamma &\sim \text{generalizedGamma}(1, 1, \gamma = 2) && (11)\\ & && (12) \end{aligned}$$

though we return point estimates via an EM algorithm.

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

G Variance Derivation

G.1 Deriving the Conformal Bound

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

$$\Pr(|y_i' - \widehat{\theta}(\widetilde{t}_i, \mathbf{x}_i)| \leq \widehat{C}_{1-\alpha/2}\widehat{\sigma}_{\widehat{\theta}}(\widetilde{t}_i, \mathbf{x}_i)) \geq 1 - \alpha. \tag{13}$$

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

$$\begin{aligned} \widehat{C}_{1-\alpha/2}\widehat{\sigma}_{\widehat{\theta}}(\widetilde{t}_i, \mathbf{x}_i) &\geq |y_i' - \widehat{\theta}(\widetilde{t}_i, \mathbf{x}_i)'| && (14)\\ &\geq |\widehat{\theta}(\widetilde{t}_i, \mathbf{x}_i) - \theta(\widetilde{t}_i, \mathbf{x}_i)| - |\theta(\widetilde{t}_i, \mathbf{x}_i) - y_i'| && (15) \end{aligned}$$

and rearranging, then bounding the term on the right gives

$$\begin{aligned} |\widehat{\theta}(\widetilde{t}_i, \mathbf{x}_i) - \theta(\widetilde{t}_i, \mathbf{x}_i))| &\leq \widehat{C}_{1-\alpha/2}\widehat{\sigma}(\widetilde{t}_i, \mathbf{x}_i) + |\theta(\widetilde{t}_i, \mathbf{x}_i) - y_i'| && (16)\\ &\leq (\widehat{C}_{1-\alpha/2} + 1)\widehat{\sigma}_{\widehat{\theta}}(\widetilde{t}_i, \mathbf{x}_i) && (17) \end{aligned}$$

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

Thus, we should expect the confidence band

$$\widehat{\theta}(\widetilde{t}_i, \mathbf{x}_i) \pm (\widehat{C}_{1-\alpha/2} + 1)\widehat{\sigma}_{\widehat{\theta}}(\widetilde{t}_i, \mathbf{x}_i) \tag{18}$$

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

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

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

$$\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i) \pm (\widehat{C}_{1-\alpha/2} + 1)\widehat{\sigma}_{\widehat{\tau}}(\widetilde{t}_i, \mathbf{x}_i) \tag{19}$$

We estimate the variance using the law of total variance

$$\underbrace{\widehat{\sigma}^2_{\widehat{\theta}}(t_i, \mathbf{x}_i)}_{\text{Total Variance}} = \underbrace{\widehat{s}^2_{\widehat{\theta}}(\widetilde{t}_i, \mathbf{x}_i)}_{\text{Sampling Variance}} + \underbrace{\widehat{\sigma}^2_{\theta}(\widetilde{t}_i, \mathbf{x}_i)}_{\text{Error Variance}} \tag{20}$$

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

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

$$\widehat{\sigma}^2_{\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i)} = \widehat{\mathrm{Var}}(y_i|\mathbf{x}_i) - \widehat{\mathrm{Var}}(y_i|\widetilde{t}_i, \mathbf{x}_i) \tag{21}$$

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

H Performance Simulations

H.1 Data Generating Processes

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

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

In the first four settings, we take

$$t_i = \frac{(\mathbf{x}_{i2} - 1)^2}{4} + \epsilon_i^T; \;\; \epsilon_i^T \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1). \tag{22}$$

In the fifth setting, we introduce a discontinuity by using

$$t_i = \text{sign}\,(\mathbf{x}_{i1}) \times \frac{(\mathbf{x}_{i2} - 1)^2}{4} + \epsilon_i^T; \;\; \epsilon_i^T \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1) \tag{23}$$

where

$$\text{sign}\,(a) = \begin{cases} -1; & a \leq 0 \\ 1; & a > 0 \end{cases} \tag{24}$$

Note that this is outside our model space and a more complex setting than that in our second illustrative simulation in Section 3.2 of the main body.

We then use the following outcome models,

$$\text{1 Linear:} \;\; y_i = y_i + \mathbf{x}_{i1} + \frac{\mathbf{x}_{i2} - 1}{4} + +\epsilon_i \tag{25}$$

$$\text{2 Partially Linear:} \;\; y_i = t_i + \mathbf{x}_{i1} + \frac{(\mathbf{x}_{i2} - 1)^2}{4} + \epsilon_i \tag{26}$$

$$\text{3 Additive Linear:} \;\; y_i = 4\sin(t_i) + \mathbf{x}_{i1} + \frac{(\mathbf{x}_{i2} - 1)^2}{4} + \epsilon_i \tag{27}$$

$$\text{4 Interactive:} \; y_i = 4\sin(t_i) \times \mathbf{x}_{i1} + \frac{(\mathbf{x}_{i2} - 1)^2}{4} + \epsilon_i \tag{28}$$

$$\text{5 Discontinuity} \; y_i = 4\sin(y_i) \times \text{sign}\,(\mathbf{x}_{i1}) + \frac{(\mathbf{x}_{i2} - 1)^2}{4} + \epsilon_i \tag{29}$$

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

We continue to contrast with the kernel regularized least squares model (Hainmueller and Hazlett, 2013) and Generalized Random Forests (Athey et al., 2019). We select these two models for comparison because they offer both point estimates and uncertainty estimates for the partial effect curve, $\tau(\widetilde{t}_i, \mathbf{x}_i)$.[18]

H.1.1 Evaluation Metrics

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

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\widehat{\tau}(\widetilde{t}_i, \mathbf{x}_i) - \tau(\widetilde{t}_i, \mathbf{x}_i)| \tag{30}$$

and the sample average coverage probability,

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

All simulations were run 500 times.

H.1.2 Results

Results from the simulations can be found in Figure 2. Each row corresponds with our simulation setting, from the additive linear model in row one to the complex, discontinuous model in row 5. In each figure, the x-axis shows outcomes by sample size. The first column presents the bias on the average partial effect, by method. Across settings, all methods do well, with GRF doing the best overall. KRLS misses the average effect in the simplest model, due to its shrinkage, but with any complexity, all models do well. The second column presents mean absolute error, a measure of accuracy of our estimates over the whole of the curve. All methods perform well in the simplest

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

settings, but KRLS and MDEI perform the best in the most complex settings. We suspect that there are conditional mean specifications where any of the methods presented will outperform others; our take away here is that all three methods perform passably well.

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

I Binary and Categorical Treatment Regimes.

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

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

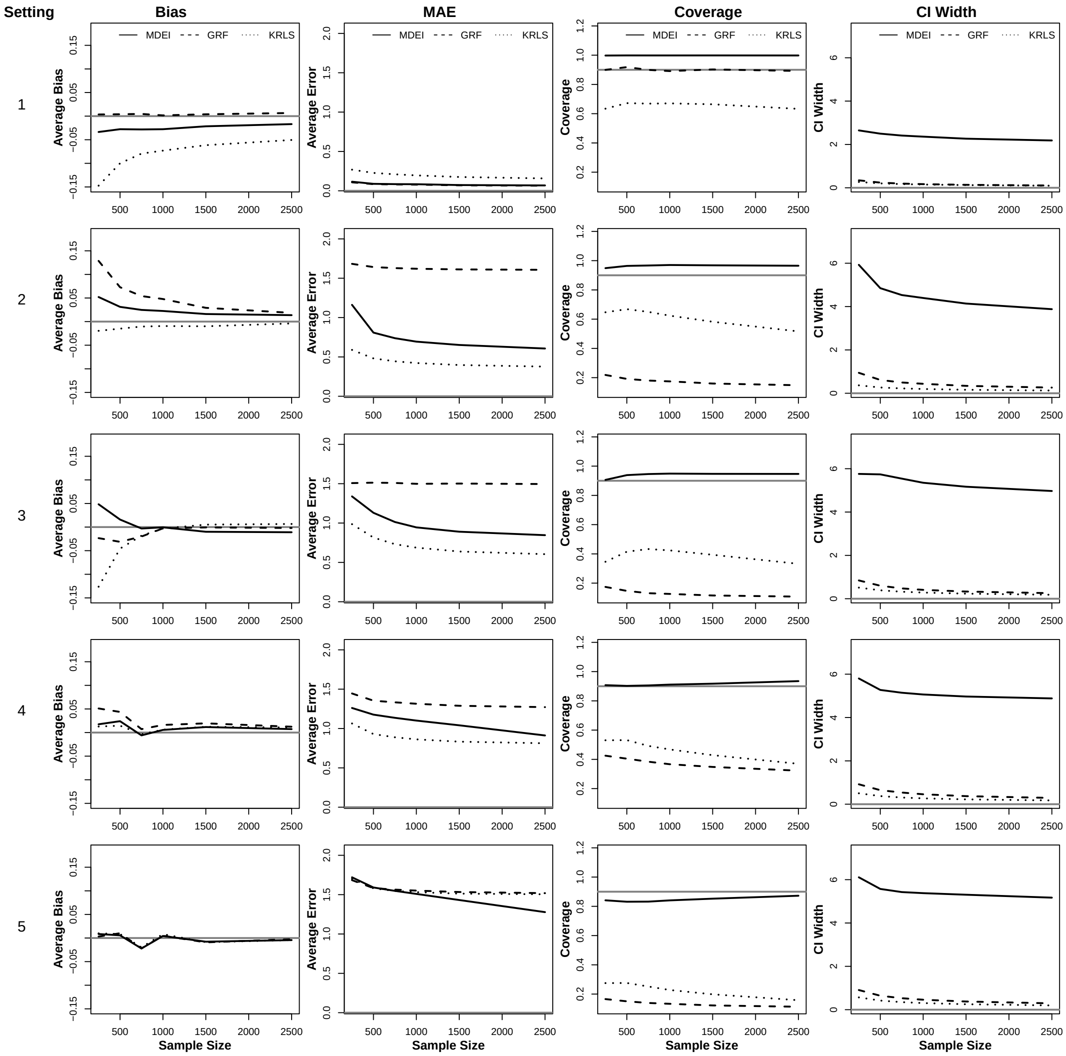

Figure 2: Performance Simulation Results

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Wager, Stefan and Susan Athey. 2017. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association* 113(523):1228–1242.

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.