Estimation and Inference on Nonlinear and Heterogeneous Effects

Marc Ratkovic, Princeton University
Dustin Tingley, Harvard University

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

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

We introduce a novel method for finding nonlinear and heterogeneous effects and focus on how to appropriately calculate uncertainty in these settings. We propose the Method of Direct Estimation and Inference (MDEI), which 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; Ho et al. 2007; Imbens and Rubin 2015), we focus on reducing the role of modeling assumptions. However, our approach minimizes the role of assumptions in estimating both point estimates and uncertainty estimates.

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

The next step, which is crucial to this article, is to generate uncertainty estimates and confidence bands for the results. This is straightforward when there are strong parametric modeling assumptions in place, as with multiple linear regression.

Marc Ratkovic (ratkovic@princeton.edu) is an assistant professor of politics at Princeton University, Princeton, NJ 08540. Dustin Tingley (dtingley@gov.harvard.edu) is a professor of government at Harvard University, Cambridge, MA 02138.

Replication files are available in the *JOP* Dataverse (https://dataverse.harvard.edu/dataverse/jop). The empirical analysis has been successfully replicated by the *JOP* replication analyst. An appendix with supplementary material is available at https://doi.org/10.1086/723811.

Published online March 22, 2023.

The task is much more challenging when we want to allow for more complicated relationships that we do not specify ex ante. Our goal is to generate a confidence band around any uncovered nonlinearity that will allow us to assess how the estimated curve relates to the true curve. Thinking about uncertainty in this setting requires differentiating between inference at a particular point and inference over a curve. For this, we estimate a confidence band with *average coverage*, meaning we expect the confidence band for our marginal effect to cover the true curve at some proportion, say 90%, of the data. Doing so allows us to use the uncertainty measure around the curve to deduce features of the true underlying curve.

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

Second, we introduce a confidence band on uncovered nonlinearities and heterogeneities that the researcher can use to assess whether a given effect reflects a systematic pattern in the data. The curve has the average coverage property (Nychka 1988; Wasserman 2006) that the band will contain the true partial effect at some chosen proportion, say 90% or 95%, of the observed data. In constructing such a curve, we rely on conformal inference (Lei and Wasserman 2014). As discussed below, conformal inference provides a data-driven, rather than assumption-driven, approach to calculating uncertainty estimates on predicted values. We extend the method from predicted values to estimating the partial effect of the treatment on the outcome at each point. Bringing all of these things together, researchers can obtain a plot of a partial effect curve that can vary over the covariate space but with a confidence band around it that does not rely on various assumptions.

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

The article proceeds as follows. First, we lay out the challenge of estimating and conducting inference on partial effects without relying on simplifying assumptions about how the treatment affects the outcome variable. Next, we introduce our approach and show how we estimate both point estimates and uncertainty estimates. The subsequent section provides simulations to illustrate our approach, while the appendix compares the performance of MDEI to other cutting-edge approaches. The next article section shows MDEI in action with an applied example, and then we conclude. Throughout we discuss related research, but we defer technical details to the appendix.

THE ESTIMATION AND UNCERTAINTY CHALLENGE

Consider the familiar regression model,

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

with observations $i \in \{1, 2, \dots, n\}$, outcome y_i, a variable of theoretical interest t_i, a vector of additional background variables $\mathbf{x}_i$ that includes the intercept, and an error term ε_i that is assumed to be mean independent of the treatment and background variables. This model is adopted by applied researchers for several reasons. First, θ measures the average partial effect, or slope, when characterizing the relationship between the outcome and treatment. Second, $\mathbf{x}_i^\top \gamma$ adjusts for other variables that affect both the treatment and outcome. Third, given the observed data, readily available software can produce an estimate $\hat{\theta}$ through the method of least squares. Fourth, inference on the average partial effect, θ, uses a confidence interval of the form

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

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

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 article we move past this ubiquitous implementation to a more flexible model of the relationship between outcome, treatment, and covariates.[1] For example, we relax the assumption that the covariates in $\mathbf{x}$ enter linearly, and the researcher need not specify how they enter. Rather, we allow this

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

relationship to be learned from the data. We also relax the assumption that the slope θ is homogeneous over the sample. Instead, we allow this value to vary with the value of the treatment variable t_i (e.g., the effect could be a curve rather than a straight line) and pretreatment covariates $\mathbf{x}_i$. To do this, we will replace the linear component θt_i with a flexible, interactive function that we denote as $\theta(\tilde{t}_i, \mathbf{x}_i)$, where $\tilde{t}_i = t_i - \mathbb{E}(t_i|\mathbf{x}_i)$, in order to isolate the nonsystematic fluctuations in the treatment. Then, we can model the effect of t_i on y_i as the partial derivative of $\theta(\tilde{t}_i, \mathbf{x}_i)$ with respect to $\tilde{t}_i$, denoted $\tau(\tilde{t}_i, \mathbf{x}_i)$, which is the slope coefficient at a particular value of the treatment and covariates.

Doing so will give us a model of the form

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

$$t_i = g(x_i) + \nu_i, \tag{2}$$

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

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

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

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

Two recent methods have focused on modeling the outcome nonparametrically, in a way that allows us to estimate the partial effect curve $\tau(\tilde{t}_i, \mathbf{x}_i)$: generalized random forests (GRF; Athey, Tibshirani, and Wager 2019; Wager and Athey 2017) and kernel regularized least squares (KRLS; Hainmueller and Hazlett 2013).

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

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

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

$$\hat{\tau}(\tilde{t}_i, \mathbf{x}_i) \pm \hat{C}_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}\{\hat{\tau}(\tilde{t}_i, \mathbf{x}_i)\}}, \tag{4}$$

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

We integrate two recent strategies in order to achieve this band. The first, *repeated cross-fitting* (Chernozhukov et al. 2018), uses different subsamples of the data to estimate the effect and conduct inference. The second, *conformal inference* (Lei and Wasserman 2014; Lei et al. 2018), uses a data-driven method to generate the width of the uncertainty interval such that our band will achieve average coverage even if the model is misspecified.[3] We next move on to the proposed method.

THE PROPOSED METHOD

We begin with an overview of our approach, with details following below. Estimation of the partial effect curve and its confidence band proceeds in three steps. In the first step, we generate a set of nonlinear and interactive functions of the treatment and covariates that are used to model the partial effect curve, $\tau(\tilde{t}_i, \mathbf{x}_i)$. These will come from taking the original treatment and covariate vector and constructing a large set of

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

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

interacted linear and nonlinear functions of these variables. Details are given below, but the goal is to capture any terms that may be driving heterogeneity in the partial effect. In the second, we use the covariates from the earlier set to generate a model for $\hat{\tau}(\tilde{t}_i, \mathbf{x}_i)$ and estimate its variance, $\widehat{\text{Var}}(\hat{\tau}(\tilde{t}_i, \mathbf{x}_i))$. In the third, we estimate the width of the confidence band, using conformal inference to estimate a value of $\hat{C}_{1-\alpha/2}$ that will give us average coverage.

Constructing the partial effect curve and a confidence band that achieves average coverage relies on combining both the split-sample and conformal strategies. The split-sample approach involves taking the observed sample of the data, splitting it into three equally sized subsamples, and conducting each of the three steps above on a separate subsample of the data. The split-sample approach provides a crucial guard against the biases induced by using the same data for each step.[4] We will refer to these subsamples as the *discovery subsample*, the *estimation subsample*, and the *inference subsample*.

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

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

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

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

THE METHOD OF DIRECT ESTIMATION AND INFERENCE

The discovery subsample: Generating nonlinear and interactive covariates

We use the discovery subsample to construct a set of basis functions that can model the outcome and, hence, partial effect curve. The process proceeds in two steps. In the first, using only data in the discovery subsample, we estimate the functions $\hat{\mathbb{E}}(y_i|\mathbf{x}_i)$, $\hat{\mathbb{E}}(t_i|\mathbf{x}_i)$.[5] Using these estimated conditional means, we generate the outcome and treatment with the covariates partialed out as

$$\tilde{y}_i = y_i - \hat{E}(y_i|\mathbf{x}_i); \; \tilde{t}_i = t_i - \hat{E}(t_i|\mathbf{x}_i),$$

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

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

Different choices of basis functions lead to different classes of estimators, including spline models, regularized regression,

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

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

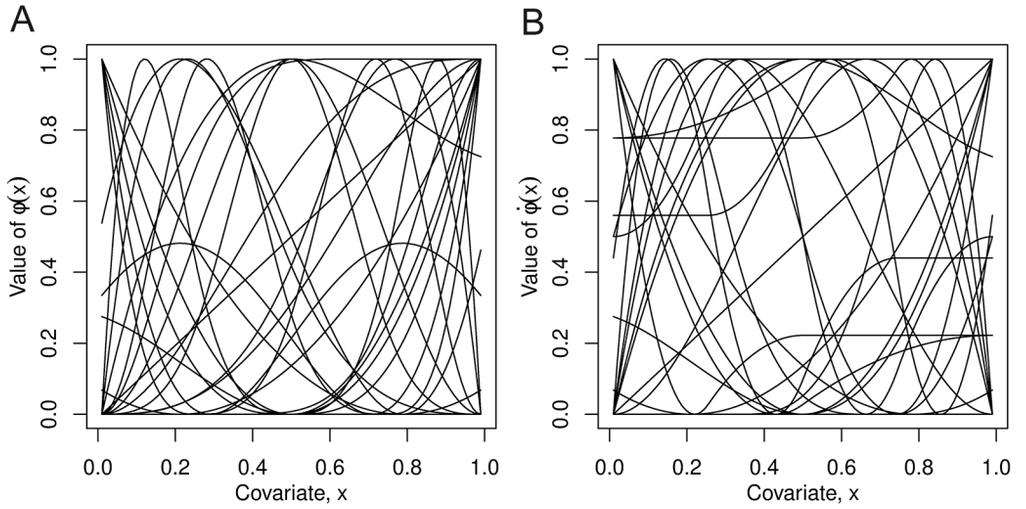

Figure 1. Basis functions of (*A*) conditional mean $\theta(\tilde{t}, x)$ and (*B*) partial effect $\tau(\tilde{t}, x) = \partial_t \theta(\tilde{t}, x)$. See equations (5)–(7) for details.

or neural networks. At this point, less important is the particular choice of basis functions but simply that they are sufficiently numerous to approximate a wide array for nonlinearities and interactions between the treatment variable and the covariates.

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

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

We then implement a marginal correlation screen (Fan and Lv 2008) in which, again, restricting ourselves to data in the discovery sample, we calculate the correlation between the partialed-out outcome, $\tilde{y}_i$ and each basis. We provide details in appendix E, but this is the most computationally intensive element of the algorithm; with five covariates, we end up calculating 675,000 correlations, and with 10 covariates, 2.7 million correlations are calculated. We then maintain a set of these bases with the largest absolute correlation with the partialed-out outcome.[6] We save these selected bases and bring them to the estimation subsample and will denote the indexes of the selected bases as $\mathcal{J}$.[7] We take these maintained bases and bring them to the estimation subsample.

Estimation subsample: Coefficient and variance estimation

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

Coefficient estimation. Given the bases from the previous subsample, we assume the model

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

with mean parameters $\{c_j\}_{j \in \mathcal{J}}$. We then use a Bayesian regression model to recover estimates, $\{\hat{c}_j\}_{j \in \mathcal{J}}$.[8]

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

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

which allows us to generate the partial effect function

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

6. We maintain a proportion of bases growing in sample size, but for sample sizes of {100, 250, 500, 1,000, 10,000} we maintain 25, 63, 125, 250, and 731 bases. See app. E for details.

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

8. We use a version of the Bayesian LASSOplus model of Ratkovic and Tingley (2017); see app. F.

Variance estimation. We turn next to constructing an uncertainty band around our estimated partial effect curve. We formalize below, but the band is constructed around the estimated curve and is designed to inform the researcher on the likely location and characteristics of the true curve. Specifically, we produce a confidence band with the average coverage property that the $100 \times (1 - \alpha)\%$ curve will contain the true curve at $100 \times (1 - \alpha)\%$. Doing so allows the researcher to explore the curve and band visually, with confidence that the band will contain the true curve over some proportion of the data.[9] This band has the nice property that it will contain the true curve at a high percentage of the observed data. It is also narrow enough for applied work but with provable average coverage properties. Formal derivations of this average coverage can be found in appendix G.

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

To illustrate this distinction, denote as $\tilde{\tau}(\tilde{t}_i, \mathbf{x}_i)$ the limit of our estimator as the sample size grows; that is,

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

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

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

Considering the squared error at each point gives us

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

from which we will construct our confidence bands.

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

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

The cross-product term, though, requires a little more finesse, as it cannot be modeled directly. Instead, we turn to a third subsample, the variance subsample, to evaluate our variance estimates and construct our confidence interval. The cross-product term is a product of two error terms, $\hat{\tau}(\tilde{t}_i, \mathbf{x}_i) - \tilde{\tau}(\tilde{t}_i, \mathbf{x}_i)$ and $\tilde{\tau}(\tilde{t}_i, \mathbf{x}_i) - \tau(\tilde{t}_i, \mathbf{x}_i)$. Any variance in the first term arises from variance in the estimation subsample (this section), so we evaluate them on the next subsample, the inference subsample. Given the bases and partialing out done in the discovery subsample, these two terms will be uncorrelated as they come from the next two subsamples, driving this term to zero. To complete a single fit, we turn next to the inference subsample.

The inference subsample: Conformal inference

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

The insight of the conformal approach comes from using estimated residuals to estimate the critical value on a predictive interval. The method is entirely data driven, and rather remarkably it achieves finite-sample coverage rates on predictive intervals.[10]

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

10. Interest in the approach is increasing in other domains of interest to social scientists (e.g., Chernozhukov, Wuthrich, and Zhu 2021; Lei and Candes 2021).

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

Our use of conformal inference methods proceeds in two steps. In the first, we select a value around $\hat{\theta}$ denoted $C^{\hat{\theta}}_{1-\alpha/2}$ such that it will contain the value $\tilde{y}_i$ with probability $1 - \alpha$:

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

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

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

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

then the interval

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

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

Repeated cross-fitting

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

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

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

While a single cross-fit estimate has the asymptotic properties we desire, the repeated cross-fitting strategy increases the accuracy of our estimates. It does so by averaging over the choice of which bases to include, so our results are not driven by a particular set of selected bases. Averaging over discrete modeling choices, like inclusion or exclusion of bases, leads to predictive gains (Buhlmann and Yu 2002). Doing so reduces any subsample-particular idiosyncrasies in our estimation, again increasing the predictive accuracy of our estimates.[11]

Modeling the standard errors as a guard against misspecification

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

While our estimation strategy works hard to find the right model—considering nonlinear and interactive effects of a potentially large number of variables—we of course cannot be guaranteed that there will not be some model misspecification. But when we miss, our approach inflates our confidence band so as to maintain average coverage. To see this, recall that we estimate our conditional variance $\widehat{\mathrm{Var}}(\hat{\tau}(\tilde{t}_i, \mathbf{x}_i))$ from the estimation subsample but evaluate it on the inference subsample, using the data-driven conformal approach in the inference subsample to calculate critical values. By modeling the error variance, we are able to recover bands that are robust to model misspecification (see app. G). By construction, this band guarantees us average coverage, as model misspecification will simply generate wider bands around the misspecified component. Existing methods do not do this and instead use auxiliary assumptions, including asymptotic normality and a properly specified model, to reduce the width of their confidence intervals.

One consequence of our approach is that our intervals will, in general, be wider than the intervals returned by other

11. There is no theoretical guidance on how many repeated cross-fits to implement. We recommend 20 for an initial fit, which is the default of our software, but then moving it to at least 100 for publication grade results.

methods (see app. G). We could make these intervals shorter by assuming our model is properly specified or assuming the errors are normal.[12] Making these strong assumptions produces more narrow intervals but comes at the cost of being overly precise if one of the assumptions does not hold.

ILLUSTRATIVE SIMULATIONS

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

Illustration 1: Existing methods in a simple setting

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

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

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

We compare performance of three different methods that return an estimate of the partial effect curve: the proposed method (MDEI), GRF (Athey et al. 2019; Wager and Athey 2017), and KRLS (Hainmueller and Hazlett 2013; Mohanty and Shaffer 2018). GRF and KRLS are prominent and commonly used in the machine learning space.[13] Each method is given the outcome y_i, treatment t_i, and five noise covariates (also independent standard normal) $\mathbf{x}_i$. We report results for $n = 1{,}000$, in order to give a sense of the large-sample behavior.

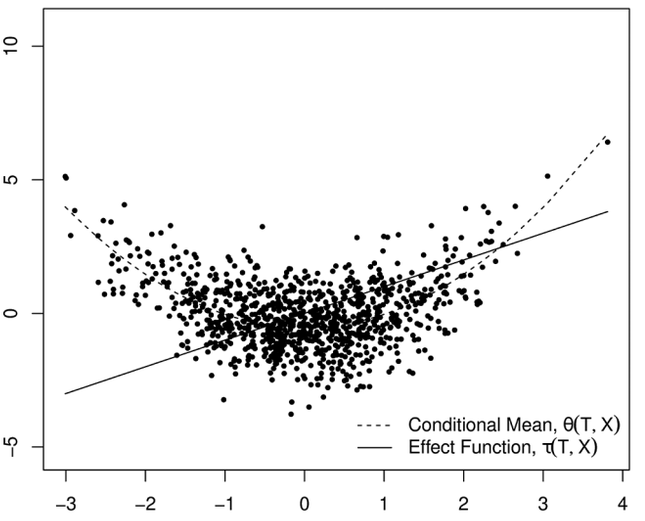

Figure 2. Data-generating process for illustrative simulation

We evaluate each method along two dimensions: point estimation and inference. The accuracy of point estimates is simple to assess: in this example, the closer the point estimate (*black dots*) to the solid line on figure 3, the better the point estimate. The second dimension, inference, asks not how close the point estimates are but whether the uncertainty bands carry some information on the true underlying curve. Is there fidelity between the uncertainty band around our estimated partial effect curve $\hat{\tau}(\tilde{t}_i, \mathbf{x}_i)$ and the true curve $\tau(\tilde{t}_i, \mathbf{x}_i)$? We will begin with an intuitive approach, assessing performance graphically. Figure 3 displays the ability of each method to capture $\tau(\tilde{t}_i, \mathbf{x}_i)$, reporting results for the proposed method (MDEI), GRF, and KRLS.

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

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

Mechanically, GRF uses the covariates to fit a tree when all background covariates are noise and then regresses the outcome on the treatment at each leaf. For a simple example, imagine it splits on the first variable at zero, so it regresses the

12. This is what other cutting-edge methods like KRLS (Hainmueller and Hazlett 2013) and GRF (Athey et al. 2019) do.

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

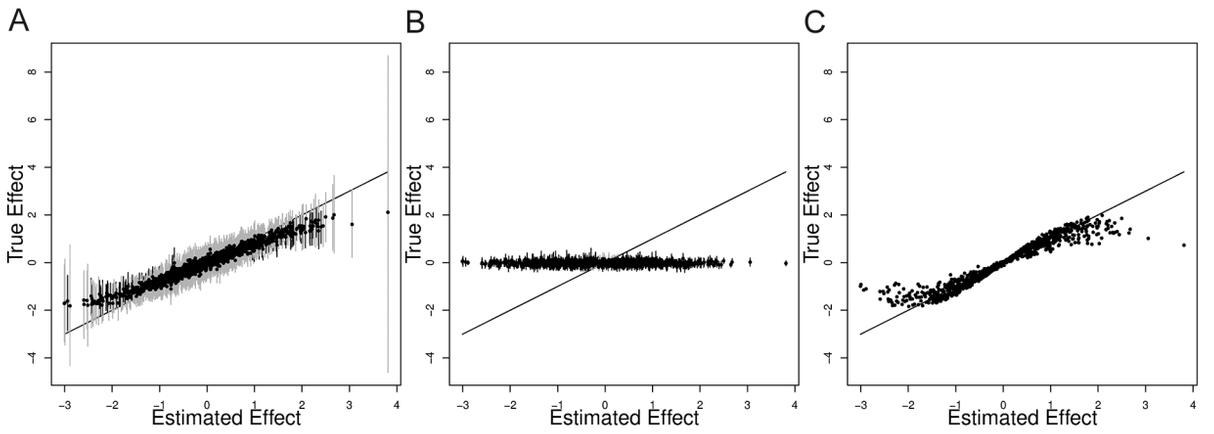

Figure 3. Effect estimates across (*A*) MDEI, (*B*) GRF, and (*C*) KRLS for a quadratic treatment model. Results are of each method in estimating the partial effect function given in figure 2. Black dots are point estimates, with gray bars denoting the uncertainty interval at each point. MDEI is the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

outcome on the treatment for observations with $x_{i1} \geq 0$ and uses a separate regression for observations with $x_{i1} < 0$. The covariates, though, are pure noise, so it is in effect fitting two lines to a quadratic curve, which the linear terms will miss.[14] GRF estimates the slope at each point, but it can only handle the case in which $(\tau(\tilde{t}_i, \mathbf{x}_i))$ is a function of covariates. In the example above, where $\tau(\tilde{t}_i, \mathbf{x}_i) = t_i$, GRF will miss the partial effect entirely, as shown in figure 3*B*.

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

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

where each function ϕ_p is a smooth, nonlinear function of $(t_i, \mathbf{x}_i)$, and P is some large number, possibly as large or larger than the sample size n. Differences arise in terms of what sorts of basis functions are used and how, precisely, the coefficients are estimated, but the important issue is that these functions are constructed to be differentiable in the treatment.[15]

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

Diagnosing existing methods: Generating uncertainty bands. At an intuitive level, we want our estimates of the partial effect curve to be as close as possible to the true curve, and we want our uncertainty estimates to give us a reasonable idea of how far we expect our estimated curve to be from the true curve. We construct intervals with the average coverage property, such that we can expect that the $(1 - \alpha) \times 100\%$ interval will contain the true partial effect curve at $(1 - \alpha) \times 100\%$ of the observed data, asymptotically. For assessing an entire curve, we work with this property because it gives an intuitive way of capturing where we suspect the true partial curve may be, given our estimate.

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

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

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

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

properly specified model, and this will just lead us to a wider confidence band.

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

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

will contain the true value $\tau(\tilde{t}_i, \mathbf{x}_i)$ at the point $(t_i, \mathbf{x}_i)$ over $(1 - \alpha) \times 100\%$ of repeated samples.[16]

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

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

Illustration 2: Combining repeated cross-fitting and conformal methods

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

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

$$t_i = \frac{(\mathbf{x}_{i2} - 1)^2}{4} + u_i;\ u_i \sim \mathcal{N}(0, 1); \tag{13}$$

$$y_i = 2s_i t_i^2 + \frac{(\mathbf{x}_{i2} - 1)^2}{4} + v\varepsilon_i;\ \varepsilon_i \sim \mathcal{N}\left(0, \frac{1}{1 + \mathbf{x}_{i2}^2}\right), \tag{14}$$

where v is a scalar selected so that the true $R^2 = 0.5$. This target function is $\tau(\tilde{t}_i, \mathbf{x}_i) = 4s_i t_i$ for which we want to use our interval to conduct inference. We vary the sample size, $n \in \{250, 500, 1{,}000, 2{,}500, 5{,}000, 10{,}000\}$, and simulations were run 500 times each.

Importantly, our model was not designed to estimate this sort of function: we assume the conditional mean is a smooth function that can be represented by some combination of interacted functions from figure 1.[17] The function in this simulation is outside of this space due to the discrete break in the first covariate. As a consequence, we developed this to be a challenging case.

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

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

The figure shows that the estimate without cross-fitting or a conformal critical value, labeled "neither" on the graph, will be quite narrow and hence not actually contain the true curve at the majority of the data.[18] This band gets worse in sample size, since its overfitting is causing it to converge on the wrong function. The conformal critical value helps somewhat, but it still results in low average coverage. Using repeated cross-fitting with the critical value of 1.64 helps get closer to 90%

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

17. See app. C regarding model spaces.

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

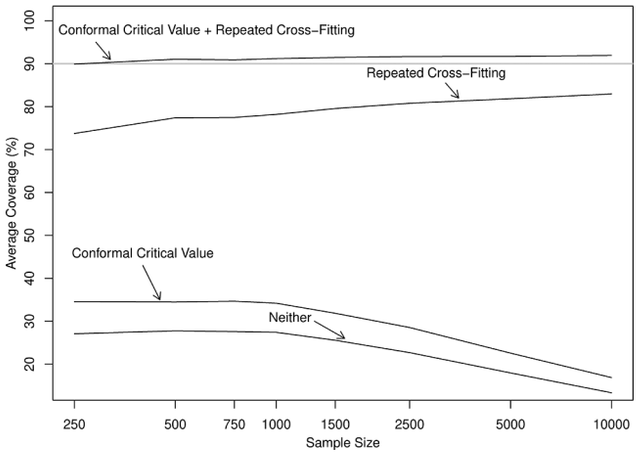

Figure 4. Average coverage of 90% band. Each line corresponds with one of the four possible settings for our confidence bands. The pointwise interval will be quite narrow, not actually containing the true curve in the majority of the data. Repeated cross-fitting increases coverage, but only when combined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

coverage, but it is only when both are combined that we see average coverage achieved.

Next, we analyze the results using the full MDEI approach for a single draw of the simulation data at two different sample sizes.[19] As a reminder, the function we are trying to fit is outside of the class of models we can handle due to the discontinuity. The result of this discontinuity is that our misspecification error will be higher, the farther away we get from $t_i = 0$, because this is where the largest gap due to the discontinuity will be. Because of how we construct our confidence bands, this means they should be wider in this region.[20] And furthermore, unlike existing methods, they should not appreciably tighten if we increase the sample size.

Figure 5 displays the results for a case with a sample size of 5,000 in the top row and 50,000 in the second row. We present results for $s_i = 1$; they are qualitatively similar to $s_i = -1$. The partial effect curve $(\tau(\tilde{t}_i, \mathbf{x}_i))$ is plotted against its estimate and interval in the left-hand panels. Intervals that contain the truth are in gray, and those that do not are in black. The 90%-band-returned MDEI covers the true value at 92.78% and 89.2% of the data, for the 5,000 and 50,000 sample sizes, respectively.[21] The middle panels plot the absolute approximation error for each point. As expected, the average approximation error increases, the farther away from $t_i = 0$ we get.

The right-hand panels of figure 5 present the width of the confidence bands. Because of our approach, these bands should be wider in the presence of misspecification error that becomes more extreme, the farther away from $t = 0$ we get. Indeed, looking at the data and a loess line to illustrate the pattern, we see this result. This is despite the fact that the simulations heteroskedasticity shrinks the variance at the extremes. Importantly, these wider bands at the extremes do not radically shrink in the $n = 50{,}000$ case.

APPLIED EXAMPLE

Bechtel and Hainmueller (2011) explore the impact of an effective policy response to a natural disaster in Germany. They estimate the effect of the government's successful response to the 2002 flooding of the Elbe River on support for the incumbent party, the Social Democrats, in the 2002 federal elections. Using a difference-in-difference design with a regression specification, the authors estimate an impact of approximately 7 percentage points on the Social Democrats' vote share. Here, the unit of analysis is the district, the outcome is the change in vote share 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, 857, table 1).

In a pure difference-in-difference design, the authors could simply estimate the effect as the change in vote share before and after the flooding of the Elbe, between the flooded and unflooded districts. Covariates can then adjust for district-level confounders not eliminated through the randomness in the flooding (Angrist and Pischke 2009, sec. 5.2.1). Bechtel and Hainmueller implement estimate the effect in a regression framework, combining the standard difference-in-difference specification with a smoothing spline in distance from the Elbe with a set of linear, additive controls.

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

19. 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 pretreatment covariates. No additional inputs are required from the user.

20. Our particular manifestation of heteroskedasticity in this simulation—where variance is smaller at the extremes—will cut against this, making simulation results consistent with this observation even more striking.

21. In the $n = 5{,}000$ case, for KRLS and GRF, those numbers are 21.7% and 13.6%. The root-mean-square error on $\tau(\tilde{t}_i, \mathbf{x}_i)$ across the methods reveals a similar pattern at $n = 5{,}000$ (MDEI: 3.46, KRLS: 5.81, GRF: 7.97). GRF uses a split sample approach, while KRLS does not but still achieves a higher error and lower coverage because, as mentioned above, GRF cannot capture curvature in the treatment variable well. KRLS cannot be run at a sample size of 50,000, so we omit comparisons at this sample size.

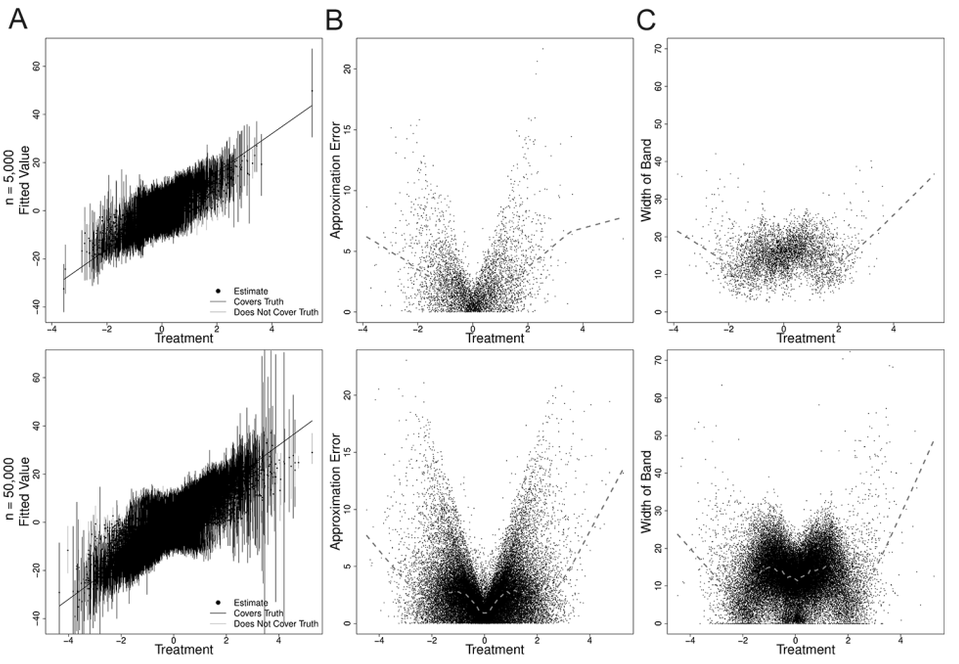

Figure 5. (*A*) Partial effect estimates, (*B*) approximation error, and (*C*) confidence band width for 5,000 (*top*) and 50,000 (*bottom*) simulations. *Left*, partial effect estimates. Solid diagonal lines represent the true partial effect, and vertical lines represent the 90% uncertainty bands. *Middle*, approximation error at each point, with a loess line describing the pattern. *Right*, band width at each point in the curve with a loess line describing the pattern.

MDEI returns both a point and uncertainty estimate for each datum, but these can be aggregated over the flooded districts only.[22] Estimates of the average effect on change in vote using Bechtel and Hainmueller's data appear in table 1. The first row contains the results using the control set in Bechtel and Hainmueller (2011). The results from MDEI appear in the third row, 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, although still significant. We find that the discrepancy between the original results in Bechtel and Hainmueller (2011) and MDEI is likely due to covariate imbalance between treated and untreated regions. If we trim districts farther 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 GRF as well as the authors' original specification on the trimmed data, but using a smoothing spline in distance from the Elbe (generalized additive model; GAM). We see that all of the methods apart from the original regression agree on the magnitude of the effect.

The reason for the improved performance is implicit in the method. The estimand for the difference-in-difference design is the average effect on the treated districts. The difference-in-difference regression coefficient, though, is a weighted average of the difference between treated and untreated units. If the untreated units are not directly comparable with the treated units, the coefficient may be biased. This is what we see here. In contrast, MDEI returns an estimated effect for each point, and then we aggregate only over the treated units in order to estimate the treatment effect on the treated. Doing so reduces

Table 1. Estimates across Model Specifications

	Point Estimate	95% Confidence Interval
Original regression	6.91	5.43, 8.40
Trimmed regression	4.89	1.77, 8.01
MDEI	4.87	2.96, 6.78
GRF	4.57	3.67, 5.47
GAM	4.55	.94, 6.72

Note. Estimated effect on the Social Democrat's vote share in flooded regions from the original specification in Bechtel and Hainmueller (2011), a trimmed regression, MDEI, GRF, and a GAM using a trimmed regression but adding a smoothing spline in distance from the river.

22. Formally, let $\mathcal{F}$ denote the flooded districts and $N_{\mathcal{F}}$ the number of flooded districts. The average effect on flooded districts ($\hat{\tau}_{\mathcal{F}}$) and its standard error ($\hat{\sigma}_{\mathcal{F}}$) are calculated as $\hat{\tau}_{\mathcal{F}} = 1/N_{\mathcal{F}} \sum_{i \in \mathcal{F}} \hat{\tau}(1, \mathbf{x}_i)$; $\hat{\sigma}_{\mathcal{F}} = 1/N_{\mathcal{F}} \sqrt{\sum_{i \in \mathcal{F}} \hat{\sigma}^2(1, \mathbf{x}_i)}$.

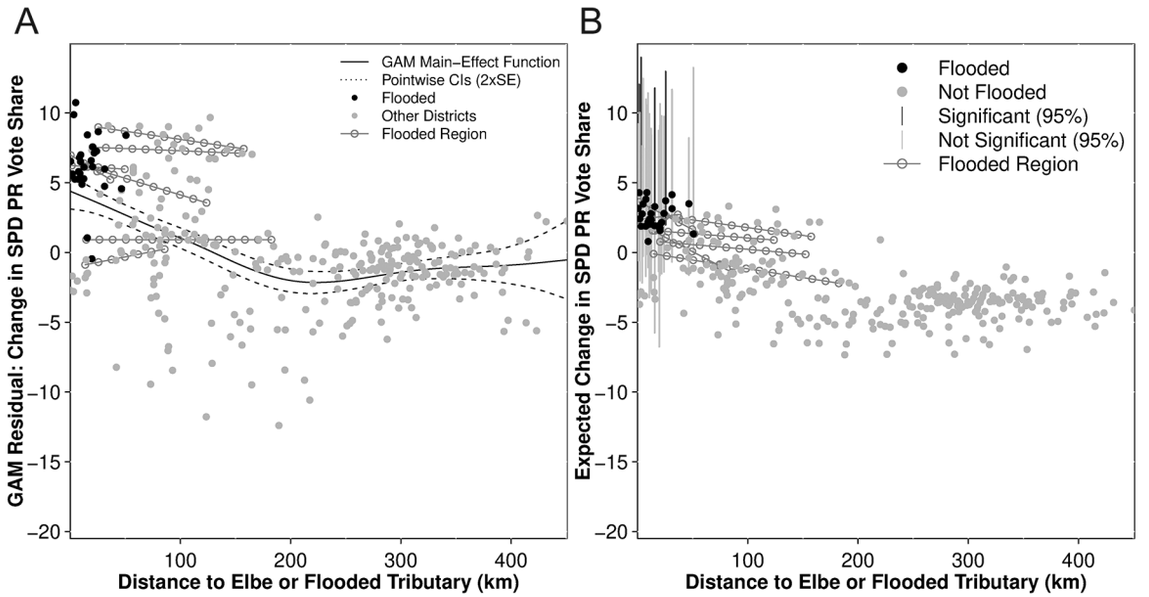

Figure 6. *A*, Estimated effect of distance to the Elbe on vote share using the specification from Bechtel and Hainmueller (2011). Analyzing residuals, flooded districts (*solid black dots*) are systematically above the trend. Regions with flooding exhibit a negative trend, suggesting that the effect is due to flooding and not some other confounding event. *B*, MDEI is able to recover similar results as in Bechtel and Hainmueller (2011) but in one step and with uncertainty bands.

concerns over imbalance. Although we use the full data to fit the model and generate confidence bands, we are only evaluating the treatment effect on those observations that were in fact treated.

We next estimate the effect of a continuous treatment on an outcome. Bechtel and Hainmueller argue that the effect of policy response on vote share decays as the distance from the Elbe increases for regions in which there were flooded districts, which they argue is further evidence that the discovered effect is attributable to disaster response. We reevaluate both claims and present results in figure 6. We begin with their analysis (see Bechtel and Hainmueller 2011, fig. 5), which we present in figure 6*A*.[23] 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 regions containing districts that flooded, which we present as the open circles. 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 figure 6*B*, we present the fitted values from MDEI in which we take each district's distance to the Elbe as the treatment. We include Bechtel and Hainmueller's (2011) 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. Bechtel and Hainmueller analyzed residuals, after taking out a smooth trend in distance and additive covariates. Figure 6*B*, 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: flooded districts are systematically above zero, meaning the vote share for the Social Democrats went up, and there is less variance in the segments fit to regions where there was flooding than to unflooded regions.

Although we find a similar pattern in the data as Bechtel and Hainmueller, we now want to know whether it is distance from the Elbe, or simply having been flooded, that is driving the estimated effect on vote share. Figure 7 presents the estimated effect of distance on vote share at each point, with flooded districts black and nonflooded gray. 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 vote share is null, after adjusting for flooding and other covariates. The estimated effect seems

23. We combine both halves of their fig. 5 into one plot for parsimony.

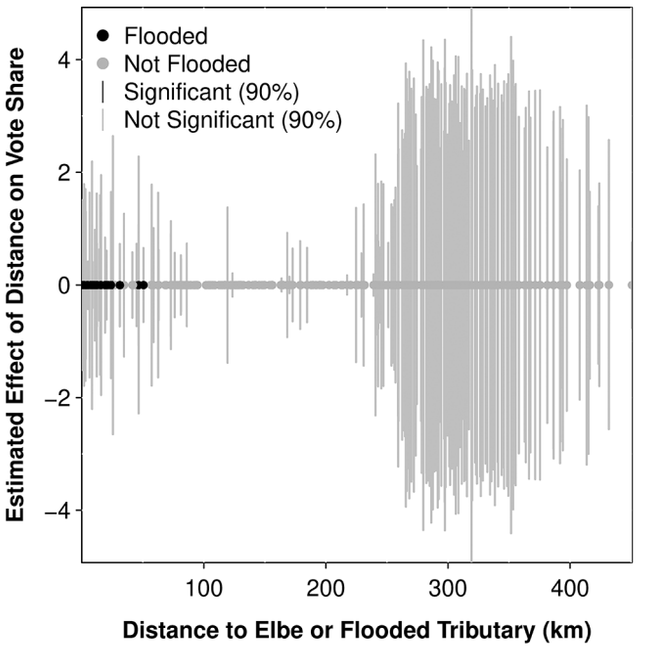

Figure 7. Point estimates and confidence intervals of the marginal effect on distance to Elbe on support for Social Democrats, by observation. Point estimates for districts in flooded regions are in black; the rest, gray. Distance has no discernible effect on support for Social Democrats.

attributable to whether the district was flooded, not to its distance from the Elbe.

Our reanalysis has recovered the central finding of Bechtel and Hainmueller, that flooded districts rewarded the Social Democrats. At the same time, we found the effect to be somewhat overstated, likely due to the inclusion of nonflooded 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, not by its distance from the Elbe. Throughout we are able to entertain nonlinear effects as well as recover uncertainty estimates.

CONCLUSION

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

We hone in on the type of inference that is appropriate when estimating nonlinear relationships when we do not ex ante specify 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 modeling choices, we do of course retain other assumptions required for making causal claims (e.g., no omitted confounders). The approach presented in this article also does not deal with other challenges to causal inference (e.g., improper confounding strategies such as controlling for posttreatment variables or certain types of pretreatment variables (Acharya, Blackwell, and Sen 2016; 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.

ACKNOWLEDGMENTS

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

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): 512–29.

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

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

Athey, Susan, and Guido Imbens. 2016. "Recursive Partitioning for Heterogeneous Causal Effects." *Proceedings of the National Academy of Sciences of the USA* 113 (27): 7353–60.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized Random Forests." *Annals of Statistics* 47 (2): 1148–78.

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): 851–67.

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

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

Buhlmann, Peter, and Bin Yu. 2002. "Analyzing Bagging." *Annals of Statistics* 30 (4): 926–61.

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." *Econometrics* 21 (1): C1–C68.

Chernozhukov, Victor, Kaspar Wuthrich, and Yinchu Zhu. 2021. "An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls." *Journal of the American Statistical Association* 116 (536): 1849–64.

Fan, Jianqing, and Jinchi Lv. 2008. "Sure Independence Screening for Ultrahigh Dimensional Feature Space." *Journal of the Royal Statistical Society* B 70:849–911.

Glynn, Adam N., and Konstantin Kashin. 2018. "Front-Door versus Back-Door Adjustment with Unmeasured Confounding: Bias Formulas for Front-Door and Hybrid Adjustments with Application to a Job Training Program." *Journal of the American Statistical Association* 113 (523): 1040–49.

Grimmer, Justin, Solomon Messing, and Sean J. Westwood. 2017. "Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods." *Political Analysis* 25 (4): 1–22.

Hainmueller, Jens, and Chad Hazlett. 2013. "Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach." *Political Analysis* 22 (2): 143–68.

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.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press.

King, Gary, and Margaret E. Roberts. 2015. "How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do about It." *Political Analysis* 23 (2): 159–78.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *Quarterly Journal of Economics* 133 (1): 237–93.

Lei, Jing, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. 2018. "Distribution-Free Predictive Inference for Regression." *Journal of the American Statistical Association* 113 (523): 1094–111.

Lei, Jing, and Larry Wasserman. 2014. "Distribution-Free Prediction Bands for Nonparametric Regression." *Journal of the Royal Statistical Society* B 76 (1): 71–96.

Lei, Lihua, and Emmanuel J. Candes. 2021. "Conformal Inference of Counterfactuals and Individual Treatment Effects." *Journal of the Royal Statistical Society* B 83 (5): 911–38.

Mohanty, Pete, and Robert Shaffer. 2018. "bigKRLS: Optimized Kernel Regularized Least Squares." *Political Analysis* 27 (2): 127–44.

Montgomery, Jacob M., and Santiago Olivella. 2018. "Tree-Based Models for Political Science Data." *American Journal of Political Science* 62 (3): 729–44.

Newey, Whitney K., and Daniel McFadden. 1994. "Large Sample Estimation and Hypothesis Testing." *Handbook of Econometrics* 4:2111–245.

Nychka, Douglas. 1988. "Bayesian Confidence Intervals for Smoothing Splines." *Journal of the American Statistical Association* 83:1134–43.

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

Ratkovic, Marc T., and Kevin H. Eng. 2010. "Finding Jumps in Otherwise Smooth Curves: Identifying Critical Events in Political Processes." *Political Analysis* 18 (1): 57–77.

Robinson, Peter. 1988. "Root-N Consistent Semiparametric Regression." *Econometrica* 56 (4): 931–54.

Samii, Cyrus. 2019. "Conformal Inference Tutorial." https://cdsamii.github.io/cds-demos/conformal/conformal-tutorial.html (accessed December 11, 2021).

Wager, Stefan, and Susan Athey. 2017. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." *Journal of the American Statistical Association* 113 (523): 1228–42.

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

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