Causal Inference through the Method of Direct Estimation*

Marc Ratkovic† Dustin Tingley‡

August 22, 2018

Abstract

Multivariate regression has been the go-to method for data analysis for generations of scholars. While transparent and interpretable, with desirable theoretical properties, the method's simplicity precludes the discovery of complex heterogeneities in the data. We introduce a method that embraces these potential complexities, is interpretable, has desirable theoretical guarantees, and is tailored to causal effect estimation. The proposed method uses a machine learning regression methodology to estimate the observation-level effect of a treatment variable, for either a binary, categorical, or continuous treatment. We illustrate with an instrumental variable analysis, estimating interesting heterogeneities in both the first and second stage. We use our theoretical results to estimate observations for which the treatment is not impacted by the instrument, for which no causal effect is identified. We provide multiple pedagogic introductions to new concepts from the machine learning literature. A technical appendix and extensive simulation evidence establishes the method's utility and use.

Key Words: causal inference, instrumental variables, machine learning, high dimensional Bayesian regression, sure independence screening

---

*We would like to thank Leah Stokes for providing replication data. We would also 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, 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. Not for citation or distribution without permission from the authors.

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

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

1 Introduction

Data analysis in much of political science is often synonymous with multivariate linear regression. In this project, we step back and ask: in what circumstances are political scientists using multivariate linear regression, what makes an estimate *causal*, and can we do better? We assume the researcher confronts an outcome variable, a treatment variable of central interest, and a set of background "control" variables that characterizes each observation's covariate profile. The first question is easy to answer: a slope estimated by a multivariate regression is interpreted as the average marginal effect of a variable on the outcome. Regression is also commonly used to adjust for background covariates when testing a set of pre-specified hypotheses. Both uses are established and long-standing in political science (e.g. Achen, 1986; King, 1998).

Implementing regression for causal inference requires additional assumptions. An estimand is causal if it is defined in terms of differences in observation-level outcomes under different treatment levels (Imbens and Rubin, 2015). In practice, obtaining a causal estimate requires successfully adjusting for confounders, pre-treatment variables which drive both the outcome and the treatment. This adjustment can occur in two ways. The first is "design-based," where a randomized treatment eliminates confounding effects. Examples include lab or survey experiments, where the randomization is controlled by the researcher, or natural experiments such as instrumental variable (IV) analyses where the randomization may simply be observed (e.g. Angrist and Pischke, 2009; Samii, 2016). The second is "model-based," where we assume that we observe all confounders and that we have modeled the treatment and outcome sufficiently well to eliminate bias. Many studies, including our replication of an instrumental variable analysis (Stokes, 2015), encompass elements of both. For example, including additional covariates beyond the instrument and treatment in an IV analysis suggests something beyond the design of the study (e.g., the use of an instrument) is useful in estimating an effect.[1]

But successfully adjusting for confounding requires that the confounding variables are incorporated correctly into the model (Aronow and Samii, 2016; Ho et al., 2007). Typical regression strategies commonly ignore complexity in the data, such as the heterogenous effect of a variable across the sample, or they assume all effects are linear. Departures from typical practice tend to be

[1] Similarly, including control variables in a difference-in-difference study or modeling a propensity score can adjust for confounding.

ad hoc, with maybe one interaction or non-linearity considered. As a result researchers implicitly make many strong assumptions about the data. This is in contrast with advances in the causal inference literature, which works to reduce the impact of modeling choices on causal estimates (Aronow, 2018; Ho et al., 2007). Furthermore, ignoring potentially interesting interactions and nonlinearities can miss adding additional nuance to a researcher's insights (e.g. Hill and Jones., 2014; Beck and Jackman, 1998; Beck, King and Zeng, 2000). Given the limitations of regression models, we think it is worth asking whether an alternative method exists for estimating causal effects.

We propose a framework, the Method of Direct Estimation (MDE), for flexible and reliable causal estimation. The method consists of three parts. First, MDE turns to the theory of causal inference for a causal estimand (Imbens and Rubin, 2015). We focus on the instantaneous causal effect (ICE), which is the observation-level marginal effect of the treatment on the outcome (e.g. Stolzenberg, 1980). Just as a regression coefficient is interpreted as an average marginal effect over the sample, the ICE is interpreted as the "slope" at a given observation, given the value of its pre-treatment variables. We then extend the framework to binary and categorical treatments and then to instrumental variable analysis. Second, MDE advances recent machine learning methods by implementing a flexible, nonparametric regression model to the outcome as a function of the treatment and covariates. The model includes a wide class of nonlinear and complex treatment/covariate interactions, allowing for the recovery of potentially interesting heterogeneities. Third, we use this method to generate predictions of counterfactual outcomes, thereby predicting "direct" estimates of relevant counterfactual quantities. Doing so allows us to recover average effects, by taking the average estimated ICE across a sample, but also reveals underlying heterogeneities lost to standard methods.

MDE combines three different sets of insights. The first, from machine learning, utilizes a high-dimensional, nonparametric regression. MDE generalizes a linear model with fully-saturated interactions between the treatment and covariates. The MDE regression fits a model with a large set of linear and nonlinear interactions between the treatment and covariates, allowing the outcome to vary flexibly and nonlinearly with both the treatment and covariates. The model requires evaluating millions of possible interaction terms, and this procedure, which we describe below, serves as the central estimation contribution.[2]

[2]These bases constitute a larger space than those considered by earlier multivariate nonparametric regression models that inspire our work; see Stone et al. (1997); Friedman (1991) for seminal work and Ravikumar et al. (2009)

Our second insight comes from statistical theory. We review the theory of optimal prediction in a high-dimensional regression model, and show that MDE is also optimal for estimating a treatment effect.[3] We then extend this theory, constructing a bound for estimating which observations that are not impacted by an instrument. This is particularly helpful in the instrumental variables context where we are able to identify observations that are "encouraged" by the treatment.

The third set of insights come from causal inference. Though we cannot observe both an observed and counterfactual outcome (e.g. Holland, 1986), we can use our nonparametric regression model to predict the counterfactual outcome. We then use these observation-level estimates to construct average effect estimates as well as explore the uncovered heterogeneities. This is a form of "predictive causal inference,"[4] and we are the first to use a high-dimensional regression model for causal estimation and prove that the effect estimates achieve desirable theoretical properties.

Just like standard regression or more recent methods like matching we, of course, cannot eliminate the assumption that there are no omitted confounders. Researchers need to address this on a case by case basis. But we can help ensure that the researcher is able to use the observed confounders as effectively and flexibly as possible across a broad range of treatment types and estimands.

As part of this project, we have developed open source software with an efficient and easy-to-use implementation.[5] The analyses below can be implemented using a line or two of code, and computation times are certainly practical. This is important given that we are using a high-dimensional non-parametric regression model that is computationally very intensive.

Throughout the paper we provide intuitions and pedagogic examples in order to introduce readers to foundational tools and ideas that we utilize. This paper is a helpful place to start for political scientists and other social scientists that want to know more about the intersection of causal inference and machine learning. We have a unique and unifying approach but also relate our approach to other methods. The paper proceeds as follows. Section 2 introduces the proposed framework, estimation strategy, a set of illustrative and performance simulations, and theoretical

for recent work.

[3] See, e.g., Buhlmann and van de Geer (2013).

[4] For precursors, see Hill (2011); Hansen (2008); Robins (1989); Rubin (1984); Cochran (1969); Belson (1956); Peters (1941).

[5] Following advances by political scientists to bring feasible computation strategies to researches (e.g., Imai and Olmsted, 2016; Lauderdale and Clark, 2014, esp. Supplemental Information), we use a $C++$ back-end and a computationally efficient estimation strategy.

results. Section 5 discusses the relationship between the proposed method and other work. Section 6 re-analyzes data from a previous paper using instrumental variables. Section 7 concludes and discusses future extensions.

2 The Proposed Method

We estimate the effect of perturbing a treatment variable on an outcome variable allowing for the fact that this effect could be nonlinear, dependent on the values of other individual level variables, or both. We first focus on the case with a continuous treatment, though a binary or categorical treatment poses no additional problems. Later we extend the insights to the instrumental variable setting.

We then discuss our estimation strategy and provide both pedagogic and performance simulation evidence. Statements of several key theoretical properties follow, though we make extensive use of appendices for the technical arguments.

2.1 Continuous Treatment

We assume a continuous treatment variable of central substantive interest. Each observation $i \in \{1, 2, \ldots, n\}$ has a potential outcome function, $Y_i(t)$, which maps some treatment level t to outcome Y_i. The observed treatment level T_i is a realization of random variable $\widetilde{T}_i$, generating observed outcome $Y_i(T_i)$. The treatment variable and outcome might also depend on other variables that precede treatment assignment, X_i. These p variables are observed for every observation i, producing a vector of these variables for each observation. There may be more covariates than observations ($p > n$), and the covariates may include moderators, risk factors, prognostic variables, or simply superfluous variables; the important point is that they are fixed and causally prior to $\widetilde{T}_i$.

The causal effect of moving from one value of the treatment, T_i', to another, T_i'', is

$$\Delta_i(T_i'', T_i') = Y_i(T_i'') - Y_i(T_i'). \tag{1}$$

To identify this causal effect, we make the standard assumptions of no unobserved confounders and non-interference between units (Imbens and Rubin, 2015; Sekhon, 2009).[6]

[6]We first assume the potential outcomes are ignorable given X_i: $Y_i(\widetilde{T}_i) \perp\!\!\!\perp \widetilde{T}_i | X_i, \widetilde{T}_i \in \{T_i'', T_i'\}$. This only need hold near $\widetilde{T}_i \in \{T_i'', T_i'$ rather than over the whole range of $\widetilde{T}_i$. We next assume the counterfactual is realizable with positive probability, $\Pr(\widetilde{T}_i = T_i | X_i, T_i \in \{T_i'', T_i'\}) > 0$. Last, we assume potential outcomes do not depend on

We can observe the realization of at most one potential outcome, so we cannot directly calculate $\Delta_i(T_i'', T_i')$. Rather, we condition on the pretreatment covariates, estimating the conditional effect for observation i as $\mathbb{E}(\Delta_i(T_i'', T_i')|X_i)$, the expected change in outcome for an observation with profile X_i. To reduce concerns over whether T_i' and T_i'' might occur with positive probability, we focus on a counterfactual estimate close to the observed data. We assume $T_i' = T_i$, the observed value, and $T_i'' = T_i + \delta$ for some small perturbation δ.

In the limit, we target a first partial derivative for each observation, $\lim_{\delta \to 0} \nabla_{T_i}(\delta)$, where

$$\nabla_{T_i}(\delta) = \frac{1}{\delta} \mathbb{E}\left\{ Y_i\left(T_i + \delta\right) - Y_i\left(T_i\right) | X_i \right\}. \tag{2}$$

This estimand captures the impact of a ceteris paribus perturbation of the treatment on the outcome.[7] Modeling this local confounding returns an estimate of the "instantaneous causal effect" (ICE, see e.g., Stolzenberg (1980))–a marginal effect for an observation with covariate profile X_i.

The standard means of estimating a marginal effect is a linear model,

$$Y_i = \mu_Y + \beta T_i + X_i^\top \gamma + \epsilon_i;\ \mathbb{E}(\epsilon_i | X_i) = 0 \tag{3}$$

where β, a global slope term, is interpreted as the average marginal effect of T_i on Y_i. The slope term, though, may not be homogenous over the sample, varying with X_i in interesting and informative ways.

We instead model the outcome as

$$Y_i = \mu_Y + f(T_i, X_i) + g(X_i) + \epsilon_i; \tag{4}$$

$$\frac{\partial}{\partial T} f(T, X_i)\big|_{T=T_i} = \lim_{\delta \to 0} \nabla_{T_i}(\delta) \tag{5}$$

where $f(T_i, X_i), g(X_i)$ are nonparametric functions of the covariates and the derivative of $f(T_i, X_i)$ is an ICE for obseration i.[8] The function $f(T_i, X_i)$ allows a non-linear, interactive relationship between the treatment and the covariates. The additional function, $g(X_i)$, is not a function of the treatment. It captures confounding effects, and may also include non-linearities and interactions.

treatment or outcomes of others and there is a single version of each treatment level. See Appendix A for additional details.

[7] We assume in this case that ignorability holds in a continuous, open interval around the observed value T_i, i.e. $Y_i(\widetilde{T}_i) \perp\!\!\!\perp \widetilde{T}_i | X_i, \widetilde{T}_i \in (T_i - \delta^S, T_i + \delta^S)$ for some $\delta^S > 0$ and that $\mathbb{E}(Y_i(T_i)|X_i)$ is differentiable in its manipulable argument at the observed value.

[8] We assume, for identification, that f, g are mean-zero.

We estimate $f(T_i, X_i), g(X_i)$, and hence the ICE, using a high-dimensional, nonparametric regression that we describe in Section 2.3.

2.2 Categorical and Binary Treatments

The preceding discussion extends naturally to binary or categorical treatments. MDE estimates a single outcome model to predict the outcome under different treatment conditions, so it can easily predict the impact on the outcome of moving the treatment from $T_i = 0$ to $T_i = 1$. The estimated average treatment effect is then the sample average of the predicted shift in outcome moving from $T_i = 0$ to $T_i = 1$. The effect on the treated can be calculated by simply taking the average over the treated observations. We provide a more thorough discussion in Appendix B.

2.3 Estimation: High Dimensional Nonparametric Tensor-Spline Regression

We next describe the estimation strategy for MDE. Our central focus is on reducing model dependence, where estimates can be impacted by the inclusion or omission of a nonlinear or interactive term. MDE fits a model for the outcome that allows for flexible, nonparametric interactions between the treatment and covariates. We use "basis functions" to capture nonlinear relationships (a brief introduction to basis functions is given below).

Our estimation approach is novel, in that it fits a flexible model to a large number of nonparametric bases and their interactions. Estimation proceeds in four steps.

1. Generate a large (on the order of millions) set of functions of the treatment and covariates, including linear terms, nonlinear terms, and their interactions. We describe how we generate these non-linear terms and interactions below.

2. Use a screening procedure to select a subset of these millions of nonlinear functions, reducing their number from millions to hundreds or thousands.

3. Use a high-dimensional Bayesian regression to fit the model.

4. Use the estimated model to predict the counterfactual quantities of interest and estimate instantaneous causal effects for each observation.

Before some details, we provide a brief introduction to nonparametric regression and basis functions, and then a simple illustrative simulation (Section 2.4 covers a set of performance simulations compared to other methods).

2.3.1 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*. Basis functions, also known as "blending functions," can be added together to approximate other functions.[9]

In linking machine learning methods and causal estimation, we show that our ignorability assumption implies that the treatment should be modeled as piecewise linear bases; see Appendix A. For intuition, the ICE is a partial derivative of the potential outcome function at the observed value, so using a locally linear model will best capture this derivative. Since we make no ignorability assumption on the background covariates, we model those using a set of both piecewise linear and smooth bases.

The top row of figure 1 provides examples of bases we use to model the treatment. They are flat at zero and elsewhere a triangle or straight line. Where the function departs from 0 is known as a knot; these knots can be widely spaced out (left hand side) or close together (right hand side). These are "degree-2" B-splines, where the degree is one plus the number of non-zero derivatives. Because these functions are locally linear, they only have one non-zero derivative.

Sets of basis functions can be added up in order to approximate a more complicated function. For example, the middle row of figure 1 plots the a simple bivariate relationship between a treatment and an outcome. As can be seen it is highly non-linear, with the drawn arrows indicating the ICE at three selected points. The bottom row plots the result of using an additive set of piecewise linear B-splines to approximate the true non-linear function. For this reason the line is *slightly* jagged as the functions that go into this themselves are all locally linear (straight lines between the knots).

Rather than trying to "guess" the right specific function, *sets* of basis functions can be combined to provide an accurate local prediction. MDE uses a collection of these functions that add up, just like the covariates in a regression model.

[9]For a comparison of our approach and kernel estimation, see Appendix J. See Keele (2008) for a helpful introduction for social scientists.

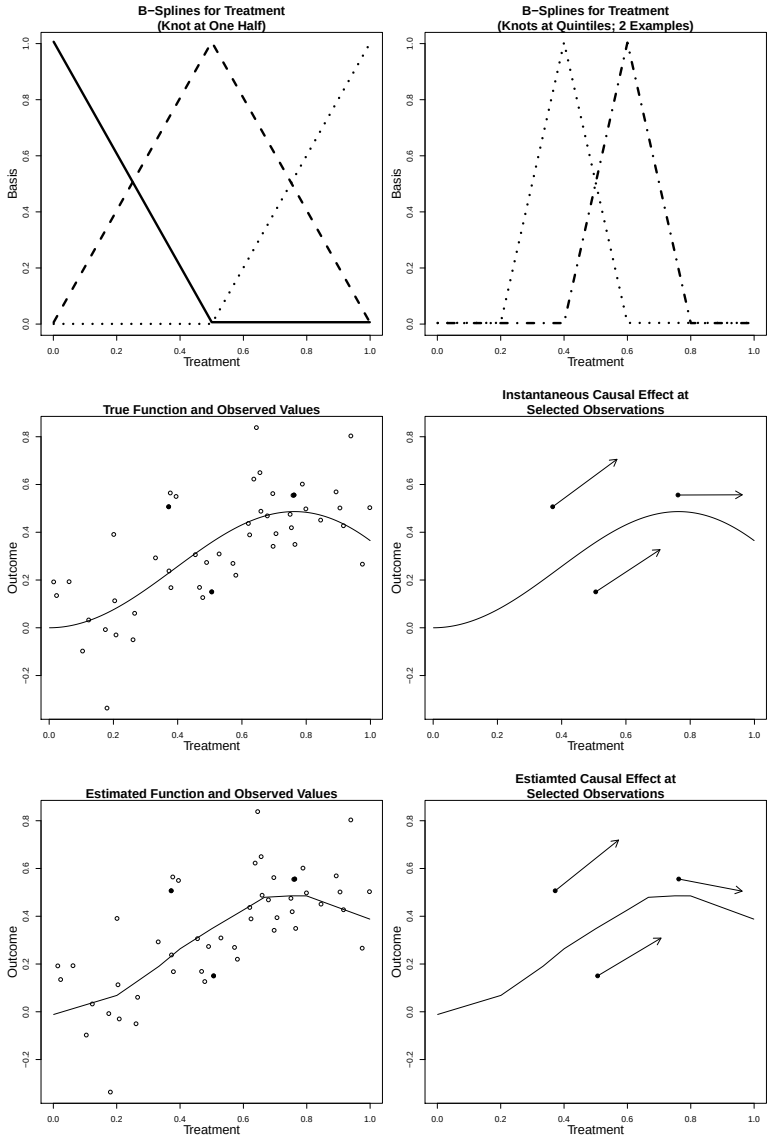

Figure 1: **Combining Basis Functions to Model Nonlinearities.** This figure illustrates how MDE uses basis functions to approximate an outcome. The top row provides examples of basis functions, which are flat at zero and elsewhere a triangle or straight line. Where the function departs from 0 is known as a knot and these knots can be widely spaced out (left hand side) or close together (right hand side). Sets of these basis functions can be added up in order to approximate a more complicated function. The middle row plots an example of a nonlinear bivariate relationship between a treatment and an outcome. Arrows indicate the ICE. The bottom row plots the result of using an additive set of B-splines to approximate the true non-linear function. Rather than assuming a functional form, sets of basis functions can be combined to provide an accurate local prediction.

2.3.2 Illustrative simulation of MDE

Given our intuitions on basis functions from above, we next illustrate the steps in MDE's estimation process. In each of two settings, we have one thousand observations, a continuous treatment

and outcome, and ten covariates, drawn from a standard multivariate normal with correlation 0.2 between each pair.[10]

In the first example, a single linear variable, X_{i1}, confound the relationship between treatment and outcome. The second setting adds a second confounding variable, X_{i2}, but its impact on the treatment is nonlinear and its impact on the outcome is a nonlinear interaction with the treatment ($T_i \times X_{i2}^2$). This simple example shows how MDE handles both nonlinearities and interactions simultaneously. The two models are

$$\text{Setting 1:} \quad Y_i = T_i + \sum_{k=1}^{10} X_{ik}\beta_k + \epsilon_i \tag{6}$$

$$T_i = 1 + X_{i1} + u_i \tag{7}$$

$$\text{Setting 2:} \quad Y_i = T_i + T_i \times X_{i2}^2 + \sum_{k=1}^{10} X_{ik}\beta_k + \epsilon_i \tag{8}$$

$$T_i = 1 + X_{i1} + X_{i1}^2 + u_i \tag{9}$$

The true ICE in the first model is 1 for all observations, and in the second case it is $1 + X_{i2}^2$.

We illustrate our estimation strategy in Figure 2. For each figure, the x-axis contains X_{i2}, the moderating variable in the second setting, and the y-axis contains the ICE. The first setting is in the top row, the second in the bottom. The first column contains the observed data for each setting and below it a subset of the constructed bases. In this specification, we consider over 1.4 million bases, consisting of linear and nonlinear bases. We plot only a subset of these bases at the bottom of the plots in the first column. As we are plotting the moderating variable, X_{i2}, in both cases some bases are smooth (as opposed to locally linear for the treatment variable on its own).

In the second column, we include the true systematic component (grey line) and, at the bottom of each plot, the bases that survive our screening and estimation procedure. These are the bases maintained after the marginal correlation screen and then not zeroed out by our regression model. The sum of the maintained bases gives the estimated ICE, as shown in the third column. We see that in this example, MDE successfully captures both the homogenous effect (top, does not depend on X_{i2}) as well as the heterogenous nonlinear causal effect (bottom, depends on X_{i2}^2). Since these basis functions are constructed from covariates, we can look at the selected basis functions and map them directly back to covariates of interest. Next we turn to discussing these steps in slightly more

[10]In both settings, $\beta_k = 1/k$, $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, and $u_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,9)$.

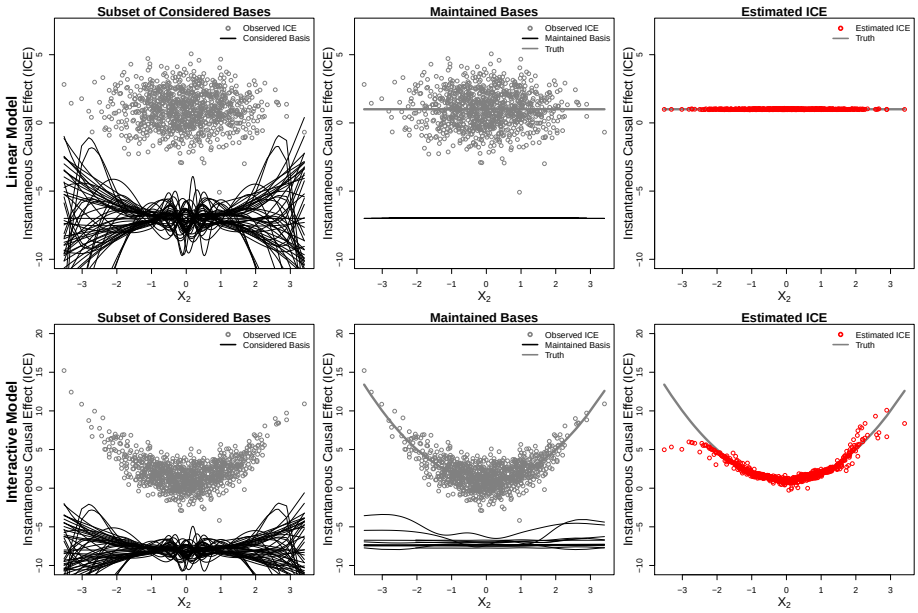

Figure 2: **Pedagogic simulation results for the linear case (top row) and non-linear interactive case (bottom row).** The first column plots the observed outcome against the instantaneous causal effect (ICE), along with a subset of considered bases. The second column plots a subset of bases retained after both screening and estimation procedure. The third column presents the estimated marginal effect for each observation against X_2, which in the bottom row is a moderating variable.

detail.

2.3.3 The Assumed Model

We next describe our estimation procedure. We begin with an assumed model. In a standard regression model, the researcher may assume the data are generated as $Y_i = \mu_Y + T_i\theta + X_i^\top\beta + \epsilon_i$, a standard linear model. The important distinction is that the researcher knows, or at least assumes, the outcome is linear in X_i.

Instead, we assume the data are generated from a (possibly infinite) set of bases ($R^o(T_i, X_i)$) and parameters (c^o), where the o superscript denotes the *true* data generating value. These bases are *not* known a priori, and must be estimated from the data. Given these bases, though, the observed data are generated as

$$Y_i = \mu_Y + R^o(T_i, X_i)^\top c^o + \epsilon_i \tag{10}$$

with $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ and each basis in $R^o(T_i, X_i)$ scaled to be mean zero and unit variance.

It will be convenient at times to split the bases into two components, one that varies with T_i

and one that does not,

$$R^o(T_i, X_i) = [R^o_{TX}(T_i, X_i), R^o_X(X_i)]; \;\; c^o = [c^o_{TX}, c^o_X], \tag{11}$$

as only the first component determines the ICE. In the notation from above, $f(T_i, X_i) = R^o_{TX}(T_i, X_i)^\top c^o_{TX}$ and $g(X_i) = R^o_X(X_i)^\top c^o_X$.

We turn next to our procedure for constructing a plausible estimate of $R^o(T_i, X_i)$ from the data.

2.3.4 Generating Bases

We implement a set of bases that allows us to model the outcome and the ICE as interactions between nonparametric curves. For example, a particular basis function of the treatment can be interacted with some other basis function of a covariate. This lets us explore an extremely broad range of candidates in our quest to recover $(R^o(T_i, X_i))$. In effect, we construct a basis for the outcome that contains nonparametric bases in the treatment and covariates while allowing for *treatment* $\times$ *covariate*, *covariate* $\times$ *covariate*, and *treatment* $\times$ *covariate* $\times$ *covariate* interactions. See Appendix E for a formal characterization.

In terms of Equations 4-5 above, the bases involving the treatment are used to estimate $f(T_i, X_i)$. Bases not involving the treatment are used to estimate the component that is only a function of the pre-treatment covariates, $g(X_i)$.

As described above (see Figure 1 and Appendix A), we construct piecewise linear bases from the treatment. Namely, we implement degree 2 thresholded power-basis and degree 2 B-splines of varying knots, for 24 total bases. We describe precisely how we construct these 24 bases in Appendix E, but our primary goal was in choosing enough bases to flexibly accommodate even complex functional forms, as shown in Figure 1.

For each covariate, we use a B-spline basis of varying knots and degree (see section 2.3.1 for a review and Gyorfi et al. (2002); de Boor (1978) for a more rigorous overview). Unlike the treatment variable, we allow for smooth nonlinearities. Again, we construct 24 bases from each covariate; see Appendix E for details. Note that this is already dramatically more than in current cutting-edge work in machine learning, even before we add interaction terms. For example Ravikumar et al. (2009) implements only three bases per covariate, and does not allow for interactions between the nonparametric curves.[11] Again, we are interested in constructing a sufficiently dense set of bases to accommodate a complex set of functions, as illustrated in Figure 2.

[11] We include this method in our performance simulations.

Given these sets of bases, 24 for the treatment variable and each of the covariates, we construct a massive set of bases, all *treatment* × *covariate* × *covariate* terms and the lower-order terms. Two mechanical aspects of our estimation procedure help MDE's performance at this stage. The first is that all of the bases are "bounded." To see this, consider Figure 2. The bases do not go off to infinity at the boundaries of the data, as a line or quadratic term would. Using bounded bases can dramatically increase the stability of the model. Second, we follow Ratkovic and Tingley (2017) and construct all interactions so that they are uncorrelated with their lower order terms. High dimensional models suffer with highly correlated terms, in much the same way that correlated terms in a regression can inflate the variance of associated estimates, leaving the researcher uncertain as to which variables to include. Interaction terms can be highly correlated with their lower-order components, so constructing interaction terms uncorrelated with their lower order constituent terms reduces the variance in our estimates.

2.3.5 Screening Bases

We allow for nonparametric interactions, a tensor-product regression, by interacting all of the nonparametric bases as described above. Twenty-four bases for the treatment and for each of p covariates gives $(1+24) \times (1+24 \times p) \times (1+24 \times p)$ total bases. Even a modest p generates a large number of bases; $p = 10$ gives over $1,450,000$ bases and $p = 20$ over $5,700,000$. This places us in an "ultrahigh" dimensional setting.

Even modern machine learning methods struggle or fail, for computational or statistical reasons, in this ultrahigh dimensional setting. For that reason, we implement a Sure Independence Screen (SIS) (Fan and Lv, 2008) as a first step. The screen works by taking the massive number of bases and only maintaining those with the largest correlation between the outcome and the basis. We want to select a sufficiently large and rich set of nonparametric bases to approximate a wide set of models; in our software and the analyses below, we maintain the $100 \times (1 + N^{1/5})$ bases with the largest unconditional correlation with the outcome.[12]Appendix D provides additional technical discussion.

[12]Statistical theory suggests a growth rate of $N^{1/5}$ (e.g. Hansen, 2017, 15.10), and we settled on the 100 to make sure a reasonable number of bases were maintained even for a small N. Memory and computational speed are a concern. To preserve memory, we construct millions of bases, but at any given point in their construction we only save the $100 \times (1 + N^{1/5})$ bases with the largest absolute correlation with the outcome.

2.3.6 A High-Dimensional Nonparametric Prediction Model

Even after screening, we still have hundreds or thousands 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.[13] We adapt their prior structure in order to achieve theoretically optimal predictive rates. Details appear in Appendix F. Here we summarize the key takeaways.

We implement a Bayesian regression model that uses a machine learning approach known as the LASSO, and in particular it is a modification of the adaptive LASSO (Zou, 2006). LASSO models are regression models where the model selects many variables to have an effect of zero, known as "regularization." For an introduction to the LASSO, see Hastie, Tibshirani and Friedman. (2013, ch. 3).

Our extension of the LASSOPlus model, which we call Predictive LASSOPlus, ensures that estimates have an important theoretical property discussed below dealing with the prediction error. Essentially, our estimates of observation-level counterfactuals are constructed such that they will l achieve a least upper bound on the possible predictive error (a "minmax" or "oracle" bound; see Section 3 for additional details). The model also has other important theoretical qualities discussed in more detail in Appendix F. For example, the Bayesian prior we use is "proper", meaning we have a well-behaved posterior estimate, even when we have more candidate bases (p) than observations ($p > n$). The prior is also proper when $n > p$. Finally, unlike standard LASSO models we do not have to deal with "tuning" parameters that need to be selected via some external method such as a BIC score or cross-validation (see Appendix F for additional details).[14]

Most importantly, we can use the model returned by MDE's Predictive LASSOPlus model to make counterfactual predictions for each observation. This allows estimation of observation-level causal effects, to which we turn next.

[13] The difference is akin to that of selecting a model on the BIC versus AIC.

[14] We regularize large effects less aggressively than small effects which focuses attention on parameters more likely to have important impacts. This helps to avoid data dredging and forking path problems that can be problematic when estimating heterogeneous effects. See Appendix F for additional details.

2.3.7 Counterfactual Prediction

Central to causal inference is the question, "what if?" What if we changed the value of the treatment for an observation to some other value? This difference between an observed and counterfactual value is the causal effect. We are now at point where we can provide a reliable answer to that question. Specifically, we can use the regresion model from above to estimate the ICE, the impact of a small, counterfactual, perturbation of the treatment on the outcome.

We construct both fitted values and counterfactual estimates from our fitted regression coefficients, $\widehat{c}$. Given these coefficients, we construct

$$\widehat{Y}_i = \widehat{\mu}_Y + R_{TX}(T_i, X_i)^\top \widehat{c}_{TX} + R_X(X_i)^\top \widehat{c}_X, \tag{12}$$

which is simply the fitted value for Y_i.

We can next write use our model to predict counterfactual estimates (superscripted by cf) with the treatment perturbed by δ,

$$\widehat{Y}_i^{cf} = \widehat{\mu}_Y + R_{TX}(T_i + \delta, X_i)^\top \widehat{c}_{TX} + R_X(X_i)^\top \widehat{c}_X. \tag{13}$$

These are our counterfactual predictions.

Our causal effect estimate is defined as the observation-level difference between two potential outcomes. Absent the opportunity to observe these outcomes, we instead use our model to predict the two outcomes. We can then subtract these two outcomes to estimate the instantaneous causal effect,

$$\begin{aligned}
\widehat{\nabla}_{T_i}(\delta) &= \frac{1}{\delta} \widehat{E}\left(Y_i(T_i + \delta) - Y_i(T_i) | X_i\right) && (14)\\
&= \frac{1}{\delta}\left(\widehat{Y}_i^{cf} - \widehat{Y}_i\right) && (15)\\
&= \frac{1}{\delta}\left(R_{TX}(T_i + \delta, X_i) - R_{TX}(T_i, X_i)\right)^\top \widehat{c}_{TX}. && (16)
\end{aligned}$$

We approximate the derivative by choosing δ close to zero; we take $\delta = 10^{-5}$ in practice.[15]

2.4 Performance Simulations

We argue below that MDE works "in theory," but we also want to assess its relative performance "in practice." To that end, we conducted an extensive simulation study, varying the sample size,

[15] See Appendix F for more on estimation.

number of covariates, and the complexity of the underlying true model. We find MDE performs quite well relative to extant machine learning methods in estimating fitted values as well as treatment effects.

The simulations isolate the effect of our two major contributions: the large set of considered bases as well as the Predictive LASSOplus model. To isolate the impact of Predictive LASSOplus, we use the same set of bases but compare MDE to other prior specifications. We look to prominent alternatives, a cross-validated LASSO and two sparse Bayesian priors: the horseshoe and Bayesian Bridge (Carvalho, Polson and Scott, 2010; Polson, Scott and Windle, 2014).

To isolate the impact of our bases, we compare MDE to methods that generate their own basis functions: kernel regularized least squares (Hainmueller and Hazlett, 2013), POLYMARS (Stone et al., 1997), and Sparse Additive Models (Ravikumar et al., 2009). We also use non-regression methods: Bayesian additive regression trees (Chipman, George and McCulloch, 2010), gradient boosted trees (Ridgeway, 1999), and the SuperLearner (Polley and van der Laan, N.d.).

Across settings, MDE is either the best performer or one of the best. Appendix H contains a detailed discussion of the setup and results.

3 Oracle Bounds for Predicted Values and Derivatives

We next present several theoretical results that give us more confidence in MDE. As they are rather technical, we discuss the results and their use cases while deferring a full formal statement to Appendix G. In addition, the following discussion helps to introduce political scientists to some core concepts and recent developments in evaluation metrics used in the field of machine learning.

MDE utilizes a high-dimensional, nonparametric regression for estimating treatment effects. Theoretical guarantees for these models, such as consistency and bounds on predictive accuracy, derive from a formal characterization of the distance from an estimated model to the truth. Heuristically, we characterize this distance by its level of *over-fitting*: the extent to which a model's predicted values reflect peculiarities in the observed sample rather than a real trend in the data. Over-fitting data can lead to a misleading model with poor generalization, since it is explaining only the current data set and not the next one.

The literature in this area has developed a common set of performance expectations (see Buhlmann and van de Geer, 2013, for an overview). Theoretical results involve characterizing the level of over-fitting, developing an upper bound on this level of over-fitting, and developing

estimators that shrink this upper bound as quickly as possible as the sample size grows. These theoretical performance results are important, as it is not immediately obvious which estimators can produce estimates for which functional forms, given that the number of nonparametric bases grows in the sample size. The goal is to generate estimates such that the prediction error diminishes in the sample size at some guaranteed optimal rate.

In slightly more technical terms, over-fitting can be thought of as a difference between how well the observed data is fit by the estimated model versus the true population model. For example, consider the maximum likelihood estimate for a sample mean. Since the maximum likelihood estimate is maximizing the likelihood, as its name implies, the likelihood evaluated at this estimate will be greater than (techincally, no less than) the sample likelihood at any other level–including the true population value. The difference between the log-likelihood maximized on the sample and the log-likelihood evaluated at the true value is an example of "excess risk." This notion of excess risk is a formal representation of over-fitting, capturing the extent to which a sample estimator fits the observed data better than the true parameters.

The goal of the theoretical analysis is to generate a *probabilistic bound* on the estimator's "excess risk". The most ubiquitous example of a probabilistic bound is the confidence interval: we expect a bound constructed from a point estimate plus or minus two standard errors to contain the true value 95% of the time, over repeated samples. The standard 95% width, or 5% error rate, is a user-controlled probability of acceptable error. The confidence interval contains the two elements of a probabilistic bound, the width and the probability of the bound holding. As with confidence intervals, the wider the excess risk bound, the more likely it is to hold, but the less informative it is.

The optimal estimator that achieves the tightest probabilistic bound on the excess risk is called an *oracle estimator*. In other words, an oracle estimator will achieve the minimal possible maximal distance between the true and estimated model, with some fixed probability.[16]

Our theoretical results, presented in Appendix G, provide two probabilistic bounds on the excess risk of the MDE estimator. Our first, and standard, result shows that the MDE regression model is an oracle estimator for the conditional mean of the outcome.[17] This result establishes that as the

[16]For this reason, they are sometimes referred to as "minmax optimal estimates." For a recent overview see Buhlmann and van de Geer (2013).

[17]In fact, we constructed predictive LASSOplus such that the estimated values would be minmax optimal; see

sample size grows, the probabilistic bound on the gap between our model and the truth is shrinking at the fastest rate possible.

The second bound is new and one of the project's innovations. We show that our treatment effect estimate is also an oracle estimator. Treatment effect error captures how close an estimated treatment effect is to the true treatment effect. Error in fitting the data and error in fitting a treatment effect are two distinct concepts. A model optimal for the one need not be optimal for the other (see Athey and Imbens, 2016, for extended discussion). Our theoretical approach provides a sharp distinction from work that focus solely on estimating an outcome. For example, the SuperLearner advocated by (van der Laan and Rose, 2011), explicitly targets predictive accuracy rather than accuracy in causal effect.

This theoretical result is driven by the fact that we interact piecewise linear splines with the covariate bases. When we take the derivative of the fitted values with respect to the treatment, we are left with bases that are (piecewise) in the outcome model. As such, we show that the treatment effect estimate inherits the minmax properties of the outcome model in the submodel, $f(T_i, X_i)$, that predicts the ICE.[18]

4 Extending to Instrumental Variables Estimation

A treatment and outcome may be mutually determined by a confounding variable, biasing linear models. In these cases, an instrument (or "encouragement") can recover a consistent causal estimate under a well understood set of assumptions (e.g., Sovey and Green, 2011): an "exclusion restriction" that the instrument have a direct effect on the treatment but not the outcome, a "monotonicity assumption" that the impact of the instrument on the treatment is not positive for some observations and negative for others, and a "validity assumption" that the instrumen have some impact on the treatment. MDE easily extends to this setting given that IV still uses regression models. MDE allows for a fine-grained analysis of an IV model, while also relaxing the monotonicity assumption and estimating which observations are impacted by the treatment. See Appendix C for details.

Appendix F.

[18] See Appendix G for further intuition and derivations.

4.1 Two-stage least squares

IV analysis is normally motivated by the following model, with Y_i the outcome, T_i the treatment, Z_i the instrument, and covariates X_i.

$$Y_i = \mu_Y + \theta T_i + X_i^\top \beta + \epsilon_i; \tag{17}$$

$$T_i = \mu_T + Z_i \alpha + X_i^\top \gamma + u_i \tag{18}$$

A correlation between the treatment and error ($\text{Cov}(T_i, \epsilon_i) \neq 0$) leads to biased estimation of θ. If the instrument, Z_i, satisfies the exclusion restriction then we can use this estimate to recover a consistent estimate of θ for observations for which $\text{Cov}(T_i, Z_i|X_i) \neq 0$.

Estimation most commonly occurs through two-stage least squares (TSLS).[19] The TSLS estimator can be written as a ratio of the first and second stage covariances,

$$\widehat{\theta}^{TSLS} = \frac{\sum_{i=1}^N \text{Cov}(Y_i, Z_i|X_i)}{\sum_{i=1}^N \text{Cov}(T_i, Z_i|X_i)}. \tag{19}$$

This expression is a ratio of summations,[20] and hence generates a single estimate averaged over the entire sample.

Summing over the sample in the numerator masks heterogeneities in the causal effect, to which we return below. For example, summing over the sample in the denominator masks heterogeneities in compliance. The instrument may affect treatment assignment for some observations ($\text{Cov}(T_i, Z_i|X_i) \neq 0$) and not others ($\text{Cov}(T_i, Z_i|X_i) = 0$). The affected observations are the only observations for which a causal effect is identified. We borrow from the binary treatment/instrument setting and call these observations *compliers* (e.g., Angrist, Imbens and Rubin, 1996). We will refer to observations positively affected by the instrument ($\text{Cov}(T_i, Z_i|X_i) > 0$) as *encouraged* and those negatively affected by the instrument ($\text{Cov}(T_i, Z_i|X_i) < 0$) as *discouraged*. The compliant subset consists of encouraged and discouraged observations.

By allowing the impact of the instrument and the treatment to vary over the sample, MDE provides a more nuanced approach to instrumental variable estimation. MDE accommodates encouraged, discouraged, and non-compliant observations, estimating which observations fall in each

[19] The TSLS estimator equals the OLS estimate of the effect of T_i on Y_i when $Z_i = T_i \; \forall \; i$, i.e., if encouragement is perfect. With a binary treatment and instrument, the TSLS estimator is the Wald estimator.

[20] We denote $\text{Cov}(Y_i, Z_i|X_i) = (Y_i - \widehat{Y}_i)(Z_i - \widehat{Z}_i)$ where $\widehat{Y}_i, \widehat{Z}_i$ are fitted values given X_i.

category. MDE also allows for heterogeneity in the treatment on the outcome, adding empirical nuance to the researcher's analysis.

4.2 Local Instantaneous Causal Effect (LICE)

We now formally introduce our approach for IV analysis.[21] We use MDE to estimate the the Local Instantaneous Causal Effect, or LICE, which is the instrumented causal effect of the treatment on the outcome, for observation with covariate profile X_i. As with TSLS, our model is theorized in two elements. The first is the *instantaneous encouragement* from perturbing the instrument on observation i and is equivalent to ICE of the instrument on the treatment,

$$\nabla_{Z_i}^{IV}(\delta) = \mathbb{E}(T_i(Z_i + \delta) - T_i(Z_i)|X_i), \tag{20}$$

As before, we must condition on X_i in order to to recover an observation-level estimate.

The second element, the *instantaneous intent to treat on the treated*, measures the impact of the instrument on the outcome,

$$\nabla_{T_i}^{IV}(\delta) = \mathbb{E}\left\{Y_i\left(T_i(Z_i) + \nabla_{Z_i}^{IV}(\delta)\right) - Y_i\left(T_i(Z_i)\right)|X_i\right\}. \tag{21}$$

As in TSLS, we construct a ratio from these two estimates. In particular, we define the Local Instantaneous Causal Effect (LICE) as

$$= \lim_{\delta \to 0} \frac{\nabla_{T_i}^{IV}(\delta)}{\nabla_{Z_i}^{IV}(\delta)}, \tag{22}$$

The estimated LICE is a two-stage least squares estimate that varies with X_i, bringing the underlying heterogeneity to the fore. We show in Appendix C that the LICE can be expressed equivalently as the result from plugging in fitted values of T_i into the second stage model:

$$\lim_{\delta \to 0} \nabla_{T_i, Z_i}^{IV}(\delta) = \lim_{\delta \to 0} \frac{\nabla_{T_i}^{IV}(\delta)}{\nabla_{Z_i}^{IV}(\delta)} \tag{23}$$

$$= \frac{\text{Cov}(Y_i, Z_i|X_i)}{\text{Cov}(T_i, Z_i|X_i)}, \tag{24}$$

We are no longer summing observation level covariances over the sample, as with TSLS (See Equation 19). Instead, MDE will estimate an effect for each observation, given its covariate profile.[22] The

[21] See Appendix C for derivations.

[22] The covariances in the LICE are taken over repeated samples, for an observation with covariate profile X_i. Since we observe only one observation with each profile, estimating this relationship requires pooling or sharing across other covariate profiles; see Robins and Ritov (1997).

LICE is still a ratio, but it varies over the sample. When there is no heterogeneity and everything is entirely linear, the average LICE over the sample equals the parameter estimated by TSLS.

4.3 Using bounds to estimate compliance status

We next use these oracle bounds on the ICE to resolve an important issue in IV analysis: how to discriminate between compliers and noncompliers. A common critique of IV analysis is the gap between the population off which a causal effect is identified (i.e., the compliers) and the overall observed population. Standard tools like TSLS (Equation 19) sum over these distinct subgroups. This generates an external validity problem (Deaton, 2010).

Using our oracle bounds on the ICE, we can construct a bound around a zero first stage effect *at the observation level*, such that estimated effects beyond this range are unlikely due to sampling noise. Returning to the confidence interval analogy, our bound is akin to constructing a predictive confidence interval for the the first derivative of the instrument on the treatment. This enables separating noncompliers from compliers by whether the observations' estimated first-stage effect (ICE) falls inside or outside the interval.[23]

5 Related Work

Our approach to modeling counterfactuals incorporates insights from several different fields, as we discuss next. Appendix J contains a more detailed discussion.

First, we connect with the causal estimation literature. Rather than a two-step procedure as with matching or inverse weighting, we estimate a fully saturated nonparametric model for the outcome. While we still make assumptions (e.g., Robins and Ritov, 1997), we avoid the inevitable and impactful consequences of misspecification of the matching or inverse weights model. We also build out beyond the binary/categorial treatment context that is the focus of most causal inference work. Finally, our focus on targeting a treatment variable is shared with the targeted maximum-likelihood approach in van der Laan and Rose (2011). We depart from this approach in that we do not target an aggregate parameter of the sample and do not use ensemble methods.[24]

Second, we connect with a literature on estimating derivatives developed in engineering as well

[23] Additional technical details are provided in Appendix G.

[24] See Samii, Paler and Daly (2016); Grimmer, Messing and Westwood (2017) for applications of ensemble methods in political science.

as economics; see O'Sullivan (1986); Horowitz (2014) for overviews and Hardle and Stoker (1989); Newey (1994) for seminal work. The economics literature has focused on causal interpretations of a structural model (Haavelmo, 1943; Hansen, 2017; Pearl, 2014; Angrist and Pischke, 2009). We differ from these through the use of non-parametric tensor splines as well as the consideration of a large-p setting.

Third, we connect to the fast growing literature on high-dimensional regression modelling. Inspired by the "POLYMARS" multivariate tensor-spline model of Stone et al. (1997), we differ in considering a much broader model space and consequently a different interaction.[25] Additionally, Hainmueller and Hazlett (2013) introduced kernel-regularized least squares (KRLS, (Rosipal and Trejo, 2001)) to political science. KRLS also uses basis functions to relax parametric assumptions. However, KRLS uses a different family of basis functions constructed from the pairwise distance between observations. Rather than bases that map back to observations, we use bases that map back to covariates. As political science theorizes variable-outcome relationships, we feel this offers a more natural fit for political science. An implication is that KRLS is ill-suited to map estimated effects back to driving variables, without an intermediate estimation and exploration stage, while it is direct and natural for MDE. See Appendix J for more discussion.

Finally, we connect with an emerging literature adapting machine learning methods to structural models, including nonparametric regression models and random forests (Newey, 2013; Belloni et al., 2012; Athey, Tibshirani and Wager, 2016). While these works target an aggregate parameter, we are somewhat closer to Heckman and Vytlacil (2005) who target observation level estimates in their "Local IV" setting. We differ in several respects including using an Oracle bound to estimate non-compliers in the data. In separate work we extend the method to other structural models like mediation or sequential-g estimation.

6 Application

We next illustrate the flexibility of MDE through an IV analysis. The first stage analysis involves estimating the impact of the instrument on the endogenous treatment, which illustrates the use of the ICE introduced in Section 2.1. This stage generalizes the first stage of TSLS, where the endogenous variables is regressed on the instrument, to the MDE model. As such, this example

[25] On the use of interactive nonparametric ("tensor-product") terms see Gu (2002).

illustrates two use cases for MDE, treatment effect and IV estimation. We show that MDE generates more powerful first stage estimates, allows us to estimate compliers, and uncovers substantively interesting heterogeneities in the data.

6.1 Using an Instrument to Estimate Heterogeneity in Political Behavior

In 2009, the Liberal government in Ontario, Canada encouraged the adoption of wind turbines throughout the province. Though broadly favored by the public, the turbines led to local backlash among those closest to the turbines. Stokes (2015) estimates the causal effect of introducing a nearby wind turbine on change in vote share for the Liberal, pro-green party in Ontario between 2009 and 2011, i.e., the treatment effect on the treated. The treatment is binary, for whether a precinct is within 3 km of a turbine, and is instrumented by wind speed. In the original analysis, data are pre-processed via matching, so each treated precinct is paired with a similar untreated precinct. The precincts are matched on demographic variables (average home value, logged; proportion with a university degree; median income, logged; and population density, logged) and the IV model fit using the instrument and geographic variables (minimum distance to a Great Lake, linear and squared; longitude and latitude as linear, square, and interaction terms). Fixed effects per electoral district are included in both stages. We note that the inclusion of the square and interaction terms indicates the author's sensitivity to problems with a simple linear fit, which we accommodate through our nonparametric specification instead of user-specified higher-order terms.

We fit the following model to the observed data, using MDE:

$$Y_i = \mu_Y + R_1(T_i, X_i)^\top c_1 + R_2(X_i)^\top c_2 + \epsilon_i^Y \tag{25}$$

$$T_i = \mu_T + R_3(Z_i, X_i)^\top c_3 + R_4(X_i)^\top c_4 + \epsilon_i^T \tag{26}$$

where Y_i is the change from 2009 to 2011 in Progressive voteshare for the precinct; T_i is binary for whether a turbine is within 3km of the precinct; Z_i is the instrument, wind speed; and X_i are the controls given above.

This model is structurally similar to the standard IV model (as in Equations 17-18), with one model for the outcome and one for the endogenous treatment. We relax the TSLS linearity assumptions rather dramatically, as each of the $R(\cdot)$ elements consist of tensor-product bases that

can accommodate a wide range of possible functional forms. This allows for flexibility and heterogeneity at both the first and second stage.

As with the standard IV model, we assume windspeed satisfies the exclusion restriction. We do not need to assume monotonicity, i.e., that the impact of windspeed on adoption is uniformly positive or negative. We also do not need to assume that the instrument is not weak, since we can estimate the observations for which wind speed appears to have an influence on turbine placement, as described in Section 4.3.

We estimate the model using a two-stage approach. In the first stage, we estimate the impact of the instrument on the treatment ($\widehat{c}_3, \widehat{c}_4$ in Equation 26). We then plug-in $\widehat{T}_i$ into Equation 25 in order to estimate the causal parameters of interest $\widehat{c}_1$ as well as the parameters associated with the excluded exogenous variables ($\widehat{c}_2$). Uncertainty estimates are produced through a nonparametric bootstrap.

Estimating the LICE through MDE offers the same benefits of standard IV analysis, namely estimating an average causal effect for observations impacted by an instrument. MDE offers three additional advantages. First, MDE estimates the compliant subpopulation. This advance improves both the internal validity, as the researcher knows which observations drive the estimate, as well as external validity, as the researcher can characterize to which sorts of observations the estimated effect can be extrapolated. Second, MDE offers a more powerful method for estimating an instrumented effect. A linear model constrains the manner in which an instrument can affect the treatment, and likewise the treatment on the outcome. MDE allows for a richer relationship. We illustrate below through a cross-validation exercise that MDE's interactive, nonlinear model better explains the observed data than a simple linear model. Third, the model allows us to characterize the underlying nonlinearities, revealing interesting substantive insights. We illustrate each of these advantages in turn below.

6.2 Estimating an Average Effect

As described above, MDE can be used to estimate an average effect, similar to two stage least squares. Before moving to the advantages of MDE over TSLS, we compare the two methods' estimates of the average effects.

Table 1 shows that MDE and TSLS return similar average effects. The first column follows the original study (Table 2 in Stokes (2015)), using the matched data. In the second analysis, we

	Model 1	Model 2	Model 3
TSLS	-0.077	-0.103	-0.101
	(-0.129, -0.0247)	(-0.171, -0.0355)	(-0.187, -0.0144)
MDE	-0.033	-0.040	-0.107
	(-0.0799, -0.0145)	(-0.0789, -0.0173)	(-0.131, -0.0659)
MDE-compliers	-0.034	-0.038	-0.107
	(-0.0816, -0.024)	(-0.0658, -0.0114)	(-0.124, -0.0617)
Sample size	718	718	5973

Table 1: Results from TSLS and MDE on Local Average Effect on the Treated. The table reports the estimates and confidence intervals for TSLS, MDE, and MDE on the estimated compliant subset, respectively. Columns correspond with Model 1, on the matched dataset controlling only for geographic variables; Model 2, on the matched dataset including all controls; and Model 3, done on the full dataset with all controls. We see the results from MDE agree with the TSLS estimates in sign and magnitude, though with smaller standard errors.

reintroduce the demographic variables to the matched dataset. In the third analysis, we use all variables but on the full, unmatched dataset. In Models 1 and 2 in Table 1, we have used this matched data in order to directly compare our results with the original study. To calculate an effect on the treated, though, MDE does not require a matching step. MDE can be fit to the whole data and the LICE can be estimated by taking the mean over the observed treated observations, sidestepping the need for matching. For this reason, we analyze the full dataset in subsequent sections.

The first two rows contain the results from two stage least squares and the confidence interval, respectively. The next two rows contain the results from MDE, with the effect calculated using only treated observations. The final rows contain the results calculated from the compliant observations, which are the precincts in which wind speed was estimated to have some impact on the probability of a turbine being placed there. We more fully analyze the compliant subset below, but note that MDE and TSLS largely agree, in terms of sign, magnitude, and significance. This correspondence need not happen in practice, especially if the TSLS model is severely misspecified due to non-linearities and interactions. In what follows we discuss how MDE enables a deeper understanding of the data.

6.3 Estimating the First Stage and Compliant Subset

IV analysis hinges on the reliability of the first stage model. In the first stage, MDE estimates the ICE of the instrument on the endogenous variable, the estimand given in Section 2.1. We first

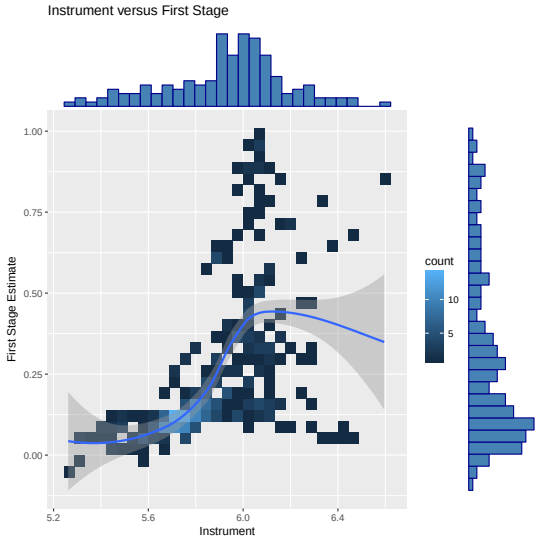

Figure 3: Nonlinear bi-variate relationship between instrument and fitted values of the endogenous variable, for the treated observations. Estimated using MDE. Distribution of instrument and first stage estimate presented via histograms. Loess smoother overlaid to characterize the non-linear pattern.

investigage whether there is a non-linear relationship between the instrument and the treatment. Figure 3 plots the instrument, wind speed, versus the first stage fitted values, the predicted probability of turbine placement using the full data set (similar results hold for the matched data).[26] We find evidence for nonlinearities in the data. The instanteous encouragement is the slope of this curve at each point. The instrument appears to have an impact above about 5.5, being flat below this range.

MDE also allows for estimating compliers utilizing the threshold discussed in Section 4.3. Figure 4 plots the value of the instrument versus the instantaneous encouragement for the 354 treated observations. The solid lines are the estimated threshold, with observations in this threshold band estimated to be unaffected by the instrument. We estimate that 73.2% (259) of observations are encouraged, as expected, 4.2% (15) are discouraged, and the remaining 22.6% (80) are not compliers. This suggests that the study's findings, while valid, are limited to areas with a value on the instrument of above 5.6.

[26] As a validity check, we estimated a Generalized Additive Model (GAM), where the instrument is included as a smooth function and all other variables are as in the original analysis. We find a similar non-linear relationship. A GAM does not model interactions between the instrument and covariates, so it is more general than TSLS but quite a bit less than MDE.

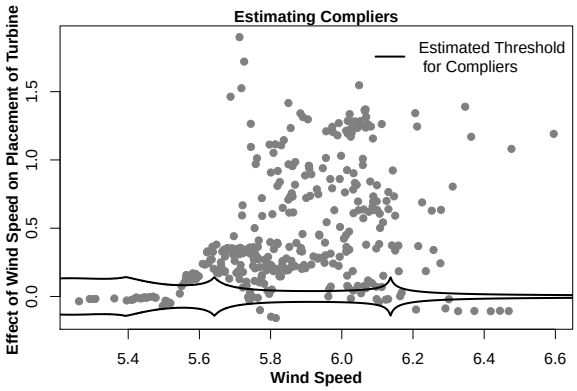

Figure 4: **First Stage Effect versus Instrument with Threshold for Compliers.**. The estimated compliers are the observations that fall outside the threshold line. Eighty-three percent of treated observations were encouraged by wind, such that higher wind speeds would have led to a higher probability of a turbine being placed there. The compliant subset fall roughly in observations with a windspeed between 5.6 and 6. Results come from Model 3; results from Models 1 and 2 are qualitatively similar.

6.4 Effect Heterogeneity

We next establish that a model with nonlinearities and interactions, MDE, better explains the data than a linear model. Having established that there are likely nonlinearities, we analyze them in two stages. First, we compare the first and second stage estimates to explore how the treatment effect and encouragement covary. Second, we characterize some of the interesting heterogeneities uncovered in the data.

6.4.1 Cross-validation to Compare Linear and Nonlinear/Interactive Models

A basic hurdle in high-dimensional modeling is establishing that a complex model fits the data better than a simple linear model (De Marchi, Gelpi and Grynaviski, 2004). Estimating predictive accuracy through cross-validation serves as a standard metric for comparing models. Specifically, the data is split into K subsets of approximately equal size. The model is fit to $K-1$ of these sets and then used to predict the held-out set. This is done in turn for each of the K sets and the cross-validation estimate serves as an unbiased estimate of the predictive accuracy of a model. Since the same data are never used to both fit and evaluate the model, cross-validation serves as a potent check against over-fitting.

Following convention, we take $K = 10$ and compare MDE to least squares. We cross-validate the estimated impact of the instrument on the treatment (i.e., the first stage) in two parts. First, we include only the confounders, in order to assess each method's ability to "control" for background

	Matched Data		Full Data	
	Confounders Only	Instrument + Confounders	Confounders Only	Instrument + Confounders
First Stage OLS	877.17	882.55	858.15	859.71
First Stage MDE	884.91	886.56	859.57	860.85
Reduced Form OLS	321.87	346.83	56.74	68.85
Reduced Form MDE	348.80	367.49	82.22	115.93

Table 2: Results from Cross-Validation Exercise. The cross-validation results for our first stage model using least squares (top) and MDE (bottom). We consider both the matched data (first two columns) and full dataset (right two columns). A larger value indicates the model has a higher predictive likelihood. MDE returns higher values in cross-validation in both the first stage and cross-validation exercise, across settings. Importantly, MDE is dominating OLS in this dataset in terms of the confounders-only model, suggesting that MDE better adjusts for confounders than OLS, producing more reliable estimates.

covariates. We then cross-validate the model using the instrument and the confounders, in order to assess whether each model is improved by adding the instrument.

Results for the cross-validation exercise are presented in Table 2. In order to use a common scale across outcomes, we place all results on a log-likelihood scale, so larger values indicate a better predictive fit. We consider both the matched data (first two columns) and full dataset (right two columns). MDE returns higher values in cross-validation in both the first stage and cross-validation exercise, across settings. Importantly, MDE is dominating OLS in this dataset in terms of the confounders-only model, suggesting that MDE better adjusts for confounders than OLS, producing more reliable estimates. Via a fair predictive criterion, our model allowing for heterogeneities outperforms a linear model in this data.

6.4.2 Local Instantaneous Causal Effect (LICE)

Next, we look at heterogeneities in the Local Instantaneous Causal Effect (LICE). The LICE is an observation-level ratio, so we first plot the first stage (denominator, x-axis) against the second stage (numerator, y-axis). The LICE is the ratio of the y-coordinate to the x-coordinate.

The top row of Figure 5 contains the results; as above, we focus on Model 3 from Table 1, using all the data and all the confounders. The top lefthand side contains the results for only the treated observations that are estimated compliers, while the righthand side contains results for untreated observations that are estimated compliers. Starting with the treated observations that were estimated compliers (274 total), we see a negative effect over the samples, indicated by the

data concentrating in the second and fourth quadrants. The observations in the fourth quadrant (239/274) were those for which an increase in windspeed would have made placement of a turbine more likely, and the turbine hurt the Liberals' voteshare. The observations in the second quadrant (10/274) are the converse–increasing windspeed would have decreased the probability of a turbine, but the turbine helped the Liberals' voteshare. Only five observations are in the third quadrant and twenty in the first.

The top righthand figure can be analyzed similarly, with the second and fourth quadrants observations where a turbine would have hurt the Liberals, and the first and third where a turbine would have helped. The lowess curve presents a more nuanced story for control observations. Near the y-axis, we see a story similar to the treated observations, with the lowess curve running from top left to bottom right. Above a first-stage effect of about 0.7, we see a sharp decrease. Had a turbine been placed with any of the observations with a first-stage effect above 1, the Liberals would have been particularly hurt. We also see more pronounced heterogeneities where the first-stage effect is about 0.5. Here, the effect on Liberals is on-average negative, but there were precincts that did not get a turbine which could have proven much better (top right quadrant) or much worse (lower right quadrant) than where they were actually placed (the lefthand figure). It appears that the turbines were placed in such a way that ultimately hurt the Liberals, but there was opportunity for them to have been placed in a manner that would have either hurt or helped their electoral prospects quite a bit more.

The bottom row of Figure 5 plots the LICE (on the y-axis) against the instrument (x-axis). As expected from examination of the plots in the top row, most observations show a negative LICE. In looking at the treated observations (bottom-left) we see a small non-linearity between the values of 5.75 and 6, where the LICE approaches 0. When examining the control observations there is a slight decreasing trend in the LICE over values of the instrument.

6.4.3 Recovered variables/functional forms

We uncovered heterogeneities in both our first and second stage estimates. Just as in a regression, MDE allows for interpretation in terms of coefficients and nonlinear bases that we are able to map back to variables of interest. Due to space considerations we present the coefficient table for the first and second stage effects in Appendix K, but we present heterogeneities graphically in Figure 6. The lefthand plot contains windspeed on the x-axis against population density on the y-axis.

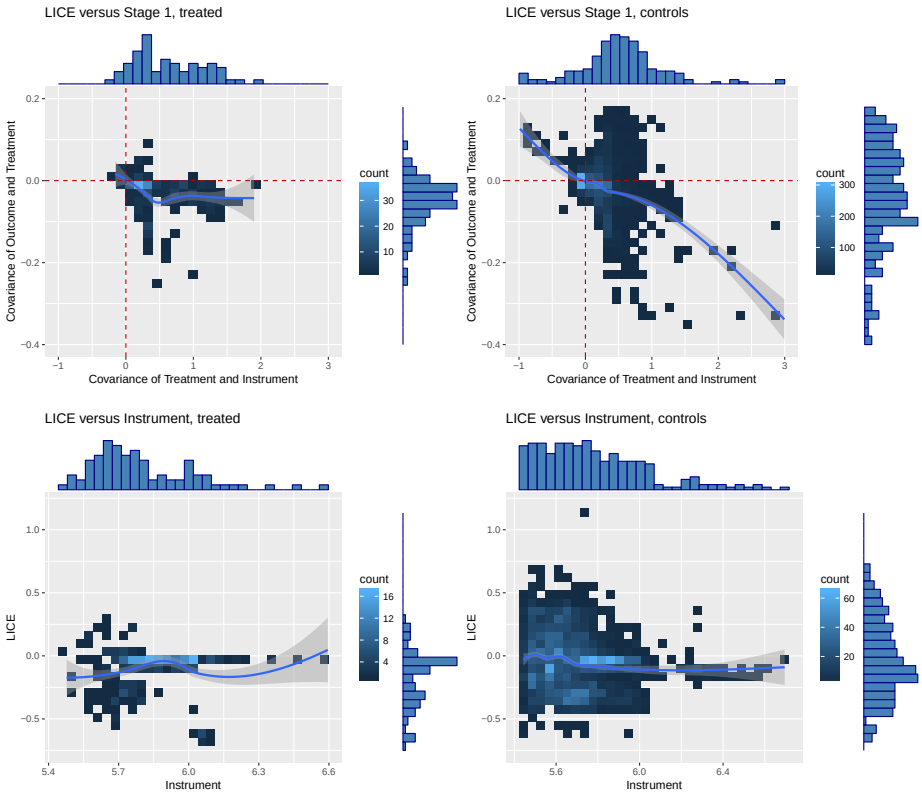

Figure 5: First stage versus second stage ratio and LICE estimates. Top row plots the components of individual level ratio making up the LICE using only the estimated compliers. The treated observations are on the left and control observations are on right. The bottom row plots the LICE against values of the instrument, again using only the compliers and treated observations on the left and control observations on the right.

Darker points have a higher predicted probability of receiving a windmill; lighter points have a lower probability. Comporting with common sense, we find an interaction between windspeed and density: precincts with high windspeeds and low population density are more likely to have received a turbine, while precincts with low windspeeds and high population density were less likely to have received a turbine.

The righthand plot illustrates a second stage interaction, plotting education on the x-axis and the LICE attributable to education on the y-axis. The result shows an interaction between university degree and the treatment, suggesting that Liberals were hurt by the turbine more in higher-educated precincts. Given the necessary planning and involvement in setting up a local opposition group, this heterogeneity seems plausible.[27] Of course, not all recovered coefficients have a straightforward

[27]We ran TSLS on the matched dataset, following the original study, and used the original models specification but included a *turbine* $\times$ *education* interaction. This interaction term is substantively and statistically suggestive

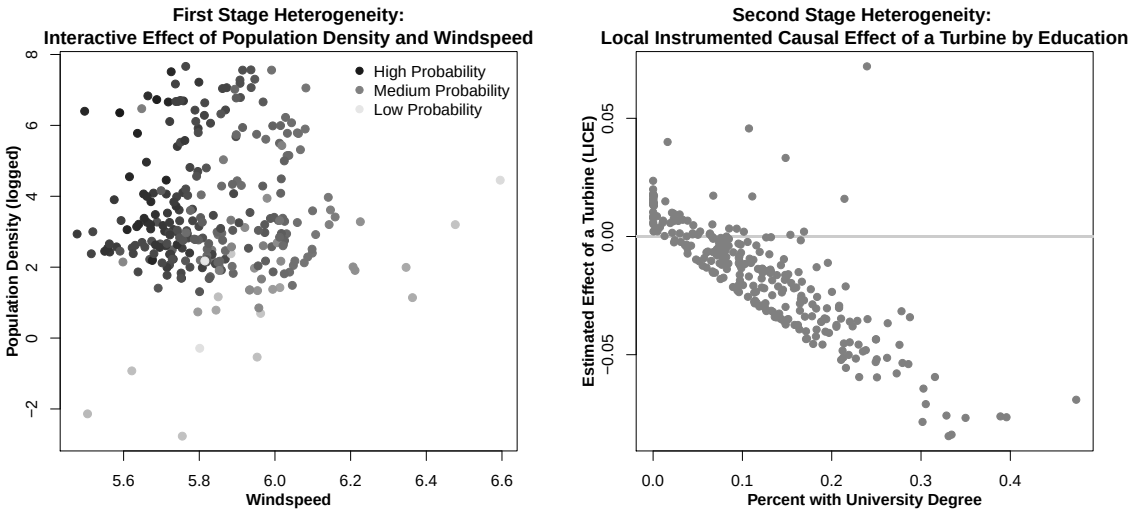

Figure 6: Uncovered Heterogeneities at the First and Second Stage. The lefthand plot contains windspeed on the x-axis against population density on the y-axis. Darker points have a higher predicted probability of receiving a windmill; lighter points have a lower probability. We find that precincts with high windspeeds and low population density are more likely to have received a turbine, while precincts with low windspeeds and high population density were less likely to have received a turbine. The righthand plot illustrates a second stage interaction, plotting education on the x-axis and the LICE attributable to education on the y-axis. The result shows an interaction between university degree and the treatment, suggesting that Liberals were hurt by the turbine more in higher-educated precincts. Given the necessary planning and involvement in setting up a local opposition group, this heterogeneity seems plausible.

interpretation, such as nonlinear interactions between latitude and longitude, but selected bases can be presented and assessed graphically.

7 Conclusion

Scholars face many challenges in producing credible causal estimates. One challenge is correctly modeling how a treatment variable impacts an outcome. Is the effect non-linear, dependent on the values of other variables, or a combination of these two things? Traditional regression models grow increasingly unhelpful given these challenges, especially as the number of variables and potential non-linear relationships increase. We introduce a high-dimensional, flexible regression model that directly targets observation-level causal effect estimates by considering an extremely rich covariate *and* functional form space. This enables us to estimate observation level counterfactuals, which then helps us make observation level causal effect estimates. We further show how our estimation strategy

($est. = -0.259, p = 0.034$). See Appendix K for details.

extends beyond the estimation of simple treatment effects and can also be used in instrumental variable regressions. Simulation evidence, presented in the online appendix, shows that the proposed method performs very well against other cutting-edge methods.

We see this work as providing an integrative thread across a number of literatures. Our approach starts with Rubin's observation that causal inference is a missing data problem: were the counterfactual outcomes known, causal inference would amount to simple descriptive statistics (Holland, 1986). We work towards this goal head-on by directly estimating observation level counterfactuals, avoiding workarounds like inverse weights and propensity scores, which are rarely of direct interest and rely on a reasonable treatment model. Rather than develop methods that are reasonable even in the face of model misspecification (e.g. Robins and Rotnitzky, 2001), we focus our attention on getting the outcome model correct. A high-quality outcome model can generate high-quality counterfactual estimates, allowing for causal inference across a range of scenarios.

Perhaps less familiar to political scientists are some of the machine learning tools we draw on. Given our focus on capturing potential non-linear effects we build off seminal work on multivariate spline models (Stone et al., 1997; Friedman, 1991) but expand the range of basis functions utilized and combine it with recent machine learning tools (Fan and Lv, 2008) in order to focus on the goal of getting the conditional mean correct.[28]

The utility of our approach becomes more clear in the case of the instrumental variable analysis, where we integrate the nonparametric structural equation model (SEM) approach and the potential outcome approach of Angrist, Imbens and Rubin (1996). In other work we show how to extend MDE to other structural models beyond instrumental variables, including estimands from the causal mechanism literature (e.g., mediation, controlled direct effects) which enables the inclusion of variables that are causally post-treatment.[29]

Of course the Method of Direct Estimation does not solve many important issues that confront applied researchers. Reliable causal estimation requires observing all relevant covariates and incorporating them correctly into a model. MDE helps with the second problem, but there is no solution to the first other than actually collecting all theoretically relevant variables. Second, naively entering post-treatment "control" variables or certian types of pre-treatment variables (Acharya,

[28]Of note, as shown in Section H, our expanded basis function and screening approach also enabled the LASSO to beat a number of cutting edge machine learning methods.

[29]For recent work in political science on these topics, see Imai et al. (2011); Acharya, Blackwell and Sen (2016).

Blackwell and Sen, 2016; Morgan and Winship, 2014) will be just as wrong under standard estimation strategies as ours. Finally, while our software includes up to three-way random effects, that is by no means a solution to panel or multilevel modeling. We expect extending to these models will require additional computational and theoretical work.

Our hope is that this paper casts a fresh light on how to think in general about causal inference in political science and other social sciences. We provide pedagogical introductions to advanced strategies for modelling non-linear relationships and treatment effect heterogeneity. We also start to build some intuitions around how scholars should evaluate these new statistical methods from a theoretical perspective. We expect a proliferation of new tools coming from literatures like machine learning, generating a need to develop theoretical and empirical tools that can help choose among them. We feel that MDE should enter the standard toolkit of applied researchers.

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).

Achen, Christopher H. 1986. *The Statistical Analysis of Quasi-experiments*. Berkeley: University of California Press.

Angrist, Joshua D, Guido W Imbens and Donald B Rubin. 1996. "Identification of causal effects using instrumental variables." *Journal of the American statistical Association* 91(434):444–455.

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

Aronow, Peter. 2018. *Agnostic Statistics*. Cambridge University Press.

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

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

Athey, Susan, Julie Tibshirani and Stefan Wager. 2016. "Solving Heterogeneous Estimating Equations with Gradient Forests." *arXiv preprint arXiv:1610.01271* .

Beck, Nathaniel, Gary King and Langche Zeng. 2000. "Improving Quantitative Studies of International Conflict: A Conjecture." *The American Political Science Review* 94(1):21–35.
URL: *http://www.jstor.org/stable/2586378*

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

Belloni, Alexandre, Daniel Chen, Victor Chernozhukov and Christian Hansen. 2012. "Sparse models and methods for optimal instruments with an application to eminent domain." *Econometrica* 80(6):2369–2429.

Belson, William A. 1956. ""A technique for studying the effects of a television broadcast"." *Applied Statistics* 5:195–202.

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

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

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

Cochran, William G. 1969. ""The Use of Covariance in Observational Studies"." *Applied Statistics* 18:270–275.

de Boor, C. 1978. *A Practical Guide to Splines*. New York: Springer.

De Marchi, Scott, Christopher Gelpi and Jeffrey D Grynaviski. 2004. "Untangling neural nets." *American Political Science Review* 98(2):371–378.

Deaton, Angus. 2010. "Instruments, randomization, and learning about development." *Journal of economic literature* 48(2):424–455.

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

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

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

Gu, Chong. 2002. *Smoothing spline ANOVA models*. Springer series in statistics Springer.

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

Haavelmo, Trygve. 1943. "The statistical implications of a system of simultaneous equations." *Econometrica* 11:1–12.

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.

Hansen, Ben B. 2008. "The Prognostic Analogue of the Propensity Score." *Biometrika* 95(2):481–488.

Hansen, Bruce E. 2017. "Econometrics." Unpublished manuscript.

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, Robert Tibshirani and Jerome Friedman. 2013. *The Elements of Statistical Learning.* 10 ed. New York: Springer-Verlag.

Heckman, James and Edward Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica* 73(3):669–738.

Hill, Daniel and Zachary Jones. 2014. "An Empirical Evaluation of Explanations for State Repression." *American Political Science Review* 108(3):661–687.

Hill, Jennifer. 2011. "Bayesian Nonparametric Modeling for Causal Inference." *Journal of Computational and Graphical Statistics* 20(1):217–240.

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

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

Horowitz, Joel. 2014. "Ill-posed inverse problems in economics." *Annual Review of Economics* 6.

Imai, Kosuke, James Lo and Jonathan Olmsted. 2016. "Fast Estimation of Ideal Points with Massive Data." *American Political Science Review,* 110(4):631–656.

Imai, Kosuke, Luke Keele, Dustin Tingley and Teppei Yamamoto. 2011. "Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies." *American Political Science Review* 105(4):765–789.

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

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

King, Gary. 1998. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*. Ann Arbor: University of Michigan Press.

Lauderdale, Benjamin and Tom Clark. 2014. "Scaling Politically Meaningful Dimensions Using Texts and Votes." *American Journal of Political Science* Forthcoming.

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

Newey, Whitney. 1994. "Kernel Estimation of Partial Means and a General Variance Estimator." *Econometric Theory* 10(2):233–253.

Newey, Whitney K. 2013. "Nonparametric instrumental variables estimation." *The American Economic Review* 103(3):550–556.

O'Sullivan, Finbarr. 1986. "A Statistical Perspective on Ill-Posed Inverse Problems." *Statistical Science* 1(4).

Pearl, Judea. 2014. "Trygve Haavelmo and the Emergence of Causal Calculus." *Econometric Theory* .

Peters, Charles. 1941. "A method of matching groups for experiment with no loss of population." *Journal of Educational Research* 34:606–612.

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.

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

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

Robins, James and Andrea Rotnitzky. 2001. "Comment on "Inference for semiparametric models: Some questions and an answer," by P. J. Bickel and J. Kwon." *Statistica Sinica* 11:920–936.

Robins, James M. 1989. *Health Research Methodology: A Focus on AIDS (eds. L. Sechrest, H. Freeman, and A. Mulley).* Washington, D.C.: NCHSR, U.S. Public Health Service chapter The Analysis of Randomized and Non-randomized AIDS Treatment Trials Using a New Approach to Causal Inference in Logitudinal Studies.

Robins, James M and Ya'acov Ritov. 1997. "Toward a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-Parametric Models." *Statistics in Medicine* 16(3):285–319.

Rosipal, Roman and Leonard J Trejo. 2001. "Kernel partial least squares regression in reproducing kernel hilbert space." *Journal of machine learning research* 2(Dec):97–123.

Rubin, Donald B. 1984. "William G. Cochran's Contributions to the Design, Analysis, and Evaluation of Observational Studies". In *W. G. Cochran's Impact on Statistics.* Wiley.

Samii, Cyrus. 2016. "Causal empiricism in quantitative research." *The Journal of Politics* 78(3):941–955.

Samii, Cyrus, Laura Paler and Sarah Zukerman Daly. 2016. "Retrospective Causal Inference with Machine Learning Ensembles: An Application to Anti-recidivism Policies in Colombia." *Political Analysis* 24(4):434–456.

Sekhon, Jasjeet S. 2009. "Opiates for the Matches: Matching Methods for Causal Inference." *Annual Review of Political Science* 12(1):487–508.
URL: *http://dx.doi.org/10.1146/annurev.polisci.11.060606.135444*

Sovey, Allison J and Donald P Green. 2011. "Instrumental variables estimation in political science: A readers' guide." *American Journal of Political Science* 55(1):188–200.

Stokes, Leah C. 2015. "Electoral Backlash against Climate Policy: A Natural Experiment on Retrospective Voting and Local Resistance to Public Policy." *American Journal of Political Science* 60(4):958–974.

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.

van der Laan, Mark J. and Sherri Rose. 2011. *Targeted Learning Causal Inference for Observational and Experimental Data.* Springer.

Zou, Hui. 2006. "The Adaptive Lasso and Its Oracle Properties." *Journal of the American Statistical Association* 101(476):1418–1429.

Online Appendix

We provide an extensive online appendix to provide additional details on our estimands, estimation strategy, theoretical results for the Method of Direct Estimation as well as a set of performance simulations. Some of the topics may be less than familiar to specialists in political scientists. Though we offer a more complete technical literature review in Appendix J, we first note the seminal pieces in applied statistics that guide the analytics underlying this work. For background, especially in regards to technical details, we refer the reader to Park and Casella (2008) for seminal work on the Bayesian LASSO, on which we draw; Buhlmann and van de Geer (2013) for an excellent compilation and overview of proof strategies for high-dimensional modeling; and Stone et al. (1997) for a direct precursor of our method using stepwise variable selection instead of sparse modeling. Our work integrates insights from these works with recent work in political science (Ratkovic and Tingley, 2017), which provides the backbone of the current estimation strategy. The remaining topics, including causal inference from a structural perspective, instrumental variable estimation, and Bayesian modeling are discussed at a level appropriate to specialists in political science.

This appendix proceed as follows. Section A motivates our choice of tensor spline basis through our ignorability assumption. Section B shows how our approach easily accommodates the binary and categorial treatment cases. Section C discusses the instrumental variables case in greater technical detail. Next we move to the estimation strategy. Our estimation strategy involves three steps: the construction of bases and interactions, a Sure Independence Screen, the application of a sparse Bayesian LASSO model. Appendix D discusses the maintained assumptions for using the Sure Independence Screen, Appendix E gives details on the construction of the bases. Appendix F details the properties of the Predictive LASSOPlus estimator. Appendix G presents our set of theoretical results for the Method of Direct Estimation. Following existing practices in evaluating high-dimensional non-parametric models (see Buhlmann and van de Geer (2013) for an introduction), we present a set of Oracle results that give us theoretical guarantees that MDE can recover fitted values and treatment effects. To help readers we present analytical results for a parametric model and then our high-dimensional non-parametric model. We also provide additional details on our ability to recover compliance status in instrumental variables estimation. Next we transition into a set of performance simulations. We consider both the estimation of treatment effects in

Appendix H where we pit our approach against a set of leading alternatives. Appendix J contains a more comprehensive literature review in statistics and machine learning. In Appendix I we evaluate performance in the instrumental variables setting, including our ability to recover correct individual level compliance estimates.

A Local Linearity of Treatment Effect.

We first show that the ignorability assumption in an open region near the observed T_i implies that the outcome model is locally linear in the treatment, with the slope a function of X_i. The result drives our choice to choose locally linear basis functions for the treatment. We assume that ignorability holds at two points, T_i and $T_i + \delta$, and then we will take δ to zero.

We first state our ignorability assumption:

$$Y_i(\widetilde{T}_i) \perp\!\!\!\perp \widetilde{T}_i | X_i, \widetilde{T}_i \in \{T_i + \delta, T_i\} \tag{27}$$

This implies that the ICE $\lim_{\delta \to 0} \frac{Y_i(T_i+\delta) - Y_i(T_i)}{\delta} = \lim_{\delta \to 0} \nabla_{T_i}(\delta_i)$ is not a function of the treatment level, as the two are independent given X_i.

Therefore, we can write the ICE for any given observation as a function of only the covariates, but not the treatment. Denote this function as $q(X_i)$, so

$$\lim_{\delta \to 0} \nabla_{T_i}(\delta_i) = q(X_i) \tag{28}$$

The ICE is just the partial derivative of the potential outcome with respect to the treatment taken at the observed value. Given the ICE, we can then recover the potential outcome function by reversing the derivative operation, i.e. taking an anti-derivative (i.e. an indefinite integral). Taking an anti-derivative allows us to separate the component of the potential outcome that determines the ICE as separable into two functions, a function $r(T_i)$ defined so $\frac{\partial r(T_i)}{T_i}$ is a constant at the observed value T_i and $q(X_i)$ as from above,

$$f(T_i, X_i) = r(T_i) q(X_i). \tag{29}$$

This is exactly the tensor product model we estimate in this project, an interaction between a locally linear function of the treatment and a nonparametric function of X_i.

B Binary and Categorical Treatment Regimes.

Assume a categorical treatment taking values in $T_i \in \{0, 1, 2, \ldots, K_T\}$. The effect of moving from one value, say T', to another, say T'' for observations with profile X_i is

$$\mathbb{E}(Y_i(T_i'') - Y_i(T_i')|X_i) = \mathbb{E}(\Delta_i(T_i'', T_i')|X_i) \tag{30}$$

This notation condenses when we consider a binary treatment, which has been the central focus of most work in the study of causal inference. Under the binary treatment, the causal effect under a binary treatment can be written as

$$\mathbb{E}(Y_i(1) - Y_i(0)|X_i) = \nabla_{T_i}(1 - 2 \times T_i) = \mathbb{E}(\Delta_i(1, 0)|X_i) \tag{31}$$

Estimation is done as in the continuous case except we now consider differences instead of partial derivatives.

The most common estimands with a binary treatment are the sample average treatment effect (ATE) and sample average treatment effect on the treated (ATT),

$$ATE = \frac{1}{n}\sum_{i=1}^{n} \nabla_{T_i}(1 - 2 \times T_i) = \tag{32}$$

$$ATT = \frac{1}{n_T}\sum_{i=1}^{n} (\nabla_{T_i}(1 - 2 \times T_i)) \times \mathbf{1}(T_i = 1) \tag{33}$$

where $n_T = \sum_{i=1}^{n} \mathbf{1}(T_i = 1)$. In our framework we estimate instantaneous causal effects for each observation, which can be aggregated up to these common estimands. The estimated ATE is the sample mean of these differences taken over the entire sample; the estimated ATT is the mean of these differences taken over the observations in the treatment conditions. The ATT is estimated as the mean expected difference in outcome between treatment and control for the observations that were actually treated.

We at a minimum reduce the need for two-stage estimation procedures like matching or weighting (e.g., Ho et al., 2007) and provide several advantages. First, we can estimate treatment effect heterogeneities which are not recovered with matching or weighting. Second, researchers commonly report standard errors on the matched sample, ignoring variance attributable to the matching stage. Since we are fitting a single model, we do not have this additional level of uncertainty in our estimates.

C Instrumental Variables Derivations

To formalize, assume the treatment is not independent of the outcome given the covariates, $Y_i(T_i) \not\perp\!\!\!\perp T_i | X_i$, thereby biasing the estimate of the causal effect. We assume the existence of a pre-treatment instrument Z_i that helps resolve the issue.[30] To generate a consistent estimate, we assume the satisfies an exclusion restriction. [31] Equivalently, $Y_i(T_i(\widetilde{Z}_i), \widetilde{Z}_i) \perp\!\!\!\perp \widetilde{Z}_i | \widetilde{T}_i, X_i$. The observed data is $[Y_i, T_i, Z_i, X_i^\top]^\top$.

We show that the LICE can be written as the ratio of two covariances and, equivalently, that a first-stage estimate of the treatment plugged into the second stage gives a consistent estimate of the causal quantity of interest.

We target the instrumented treatment effect,

$$\mathbb{E}\left(\frac{dY_i(T_i(Z), Z)}{dZ}\middle| X_i, Z = Z_i\right) \tag{34}$$

$$= \mathbb{E}\left(\frac{dY_i(T_i(Z))}{dZ}\middle| X_i, Z = Z_i\right) \tag{35}$$

$$= \mathbb{E}\left(\frac{\partial Y_i(T_i(Z))}{\partial T_i(Z)}\frac{\partial T_i(Z)}{\partial Z}\middle| X_i, Z = Z_i\right) \tag{36}$$

where the second and fourth lines come from the exclusion restriction and the third by the chain rule. We then use the continuous mapping theorem to give

$$\mathbb{E}\left(\frac{\partial Y_i(T_i(Z))}{\partial T_i(Z)}\middle| X_i, Z = Z_i\right) = \text{plim}\ \frac{\mathbb{E}(dY_i(T_i(Z))/dZ | X_i, Z = Z_i)}{\mathbb{E}(\partial T_i(Z)/\partial Z \big| X_i, Z = Z_i)} \tag{37}$$

$$= \lim_{\delta \to 0} \frac{\nabla^{IV}_{T_i}(\delta)}{\nabla^{IV}_{Z_i}(\delta)}, \tag{38}$$

which is the Local Instantaneous Causal Effect (LICE). Like with TSLS, the plug-in estimate is a biased but consistent estimate of the desired causal effect.

We note that our LICE reduces to the standard "two-stage" estimator, where fitted values of the treatment from the first stage model are entered into a second stage model. To see this, consider Equation 41:

$$\frac{\mathbb{E}(dY_i(T_i(Z))/dZ | X_i, Z = Z_i)}{\mathbb{E}(\partial T_i(Z)/\partial Z \big| X_i, Z = Z_i)} = \frac{\partial}{\partial \widehat{T}_i}\mathbb{E}\left(Y_i(\widehat{T}_i)\right)|_{\widehat{T}_i = \mathbb{E}(T_i(Z)|X_i, Z = Z_i)} \tag{39}$$

[30] The instrument is a realization of $\widetilde{Z}_i \sim F^Z_{X_i}$ and has observed value Z_i, enters the potential outcome function for the treatment as $\widetilde{T}_i = T_i(\widetilde{Z}_i)$.

[31] $Y_i(T_i(Z), Z) = Y_i(T_i(Z), Z)$ for all $Z, Z' \in \text{supp}\ (\widetilde{Z}_i)$

where we assume we can freely exchange differentiation and expectation. The second term in this equation is exactly what is calculated in a two-step estimation strategy, the impact of a first-stage estimated value on the outcome.[32] We focus our exposition on the left-hand version of the LICE, since it derives directly from observation level behavioral manipulations. In practice, we implement the right-hand version of the LICE, since the estimated coefficients from the second stage model have a natural causal interpretation.

The estimated LICE is a TSLS estimate, but one that varies with X_i:

$$\lim_{\delta \to 0} \nabla^{IV}_{T_i,Z_i}(\delta) = \lim_{\delta \to 0} \frac{\nabla^{IV}_{T_i}(\delta)}{\nabla^{IV}_{Z_i}(\delta)} \tag{40}$$

$$= \frac{\mathbb{E}(dY_i(T_i(Z))/dZ|X_i, Z = Z_i)}{\mathbb{E}(\partial T_i(Z)/\partial Z \big| X_i, Z = Z_i)} \tag{41}$$

$$= \frac{\text{Cov}(Y_i, Z_i|X_i)/\text{Var}(Z_i|X_i)}{\text{Cov}(T_i, Z_i|X_i)/\text{Var}(Z_i|X_i)} \tag{42}$$

$$= \frac{\text{Cov}(Y_i, Z_i|X_i)}{\text{Cov}(T_i, Z_i|X_i)}, \tag{43}$$

where going from the second to the third line follows from the exclusion restriction and we assume that the outcome and treatment functions are differentiable in the instrument. Hence we have a ratio of individual level covariances rather than an average parameter as in TSLS.

D Sure Independence Screen Assumptions

For the formal conditions for validity of the marginal correlation screen, see see Conditions C and 1–4 in Fan and Lv (2008). To the extent feasible, we have constructed nonparametric bases that satisfy the requirements of the screening procedure. The B-spline basis we implement is bounded, symmetric, and spherical, ensuring that the assumptions on the design are met (Condition C and 2). The number of considered bases is large, but below order $\exp(n)$ (Condition 1), and we reduce multicollinearity in the design through constructing interaction terms that are uncorrelated with their lower order components (see Fan and Lv, 2008) (Condition 4). The remaining assumptions require separable Gaussian noise (Condition 2), which we assume in Equation 10, and a beta-min condition (Condition 3) that is unverifiable absent knowing the true model. We note that selection errors on mean parameters of order $o(n^{\kappa})$ for $\kappa > 0$ will have only a minimal impact on the estimated treatment effect.

[32] We have verified in our estimation that both calculations return the same result. In practice, the two different estimates correlate at above 0.9999 over the sample.

We implement a SIS rather than the grouped SIS strategies of (Fan, Ma and Dai, 2014), as we are not willing to assume the group structure necessary to bring in entire sets of bases, and the grouped SIS has not been extended to tensor product spaces.

E Constructing Bases

.

We next provide a formal characterization of our bases and their construction.

Denote $\{1, b(X_{ik}, \kappa_k)\}_{\kappa_k \in \mathcal{K}_k}$ as an intercept term and a set of $|\mathcal{K}_k|$ mean-zero nonparametric bases for variable X_k evaluated for observation i and each value in $\mathcal{K}_k$ a triplet containing the knot location, degree, and type of spline basis. Similarly, for the treatment, denote $\{1, b_T(T_i, \kappa_T)\}_{\kappa_T \in \mathcal{K}_T}$ as an intercept and $|\mathcal{K}_T|$ bases evaluated for the treatment observed for observation i and $\mathcal{K}_T$ the triplet as given above. We construct the mean vector as

$$R_\infty(T_i, X_i) = \{1, b_T(T_i, \kappa_T)\}_{\kappa_T \in \mathcal{K}_T} \otimes \{1, b(X_{ik}, \kappa_k)\}_{\kappa_k \in \mathcal{K}_k} \otimes \{1, b(X_{ik'}, \kappa_{k'})\}_{\kappa_{k'} \in \mathcal{K}_{k'}} \tag{44}$$

where $\otimes$ denotes the tensor product. Interactions terms are pre-processed through partialling out their lower-order terms, i.e. instead of a two-way interaction, we include the residuals from regressing the two-way interaction on an intercept and its two components. Doing so reduces the multicollinearity of the design matrix.

That is, for $\{1, b(X_{ik}, \kappa_k)\}_{\kappa_k \in \mathcal{K}_k}$ and $\{1, b_T(T_i, \kappa_T)\}_{\kappa_T \in \mathcal{K}_T}$ to be approximately equal, so as not to bias our marginal correlation screen in favor of one variable or another. Were we to include $10,000$ bases for one variable and 3 for the other, bases from the first would be more likely to survive our screen.

In particular, the size of the maintained basis set should be at the nonparametric rate of $N^{1/5}$ (e.g. Hansen, 2017, 15.10) but we also want a sizable set of bases with a small sample size. To this end, at our default and throughout the paper, we model each confounder using k centered B-spline bases of degree k with knots at every $100/(1+k)^{th}$ percentile of the covariate for $k \in \{3, 5, 7, 9\}$, generating $24 = 3 + 5 + 7 + 9$ bases for each covariate. For the treatment, we implement a degree 2 truncated-power basis spline with knots every $100 \times \frac{1}{8}^{th}$ percentile for 6 bases and then degree 2 B-splines with knots at every $100 \times \frac{1}{1+k}^{th}$ percentile $k \in \{3, 4, 5, 6\}$ for $18 = 3 + 4 + 5 + 6$, giving $24 = 18 + 6$ total bases for the treatment variable. We maintain $100 \times (1 + n^{1/5})$ bases after our screening procedure.

F Predictive LASSOplus Properties

Here we describe our extension of the the LASSOPlus model of Ratkovic and Tingley (2017), Predictive LASSOplus. This work also contains additional details on derivations for our bounds and a discussion of particular properties of the LASSOplus estimator, both of which we describe below.

We start with the data generating process. We assume the likelihood component is a normal regression model. The prior structure has three properties. First, it ensures that estimates satisfy an oracle bound on the prediction error. Second, a set of inverse weights on the mean parameters regularize large effects less aggressively than small effects. Third, a global sparsity parameter produces a non-separable prior, allowing for the estimates to adapt to the overall level of shrinkage in the model.

The hierarchy is given as

$$Y_i|R(T_i,X_i),c,\mu_Y,\sigma^2 \sim \mathcal{N}(\mu_Y + R(T_i,X_i)^\top c, \sigma^2) \tag{45}$$

$$c_k|\lambda,w_k,\sigma \sim DE\left(\lambda w_k/\sigma\right) \tag{46}$$

$$\lambda^2|n,p,\rho \sim \Gamma\left(n \times (\log(n) + 2\log(p)) - p, \rho\right) \tag{47}$$

$$w_k|\gamma \sim \text{generalizedGamma}(1,1,\gamma) \tag{48}$$

$$\gamma \sim \exp(1) \tag{49}$$

$$\mu_Y \propto 1 \tag{50}$$

where $DE(a)$ denotes the double exponential density, $\Gamma(a,b)$ denotes the Gamma distribution with shape a and rate b, and generalizedGamma(a,b,c) the generalized Gamma density $f(x;a,d,p) = \frac{p/a^d}{\Gamma(d/p)}x^{d-1}\exp\{-(x/a)^p\}$. We have two remaining prior parameters. We take a Jeffrey's prior $\Pr(\sigma^2) \propto 1/\sigma^2$ on the error variance and we set $\rho = 1$ in the generalized Gamma density. We fit the model using an EM algorithm and use hats to denote the fitted values, i.e. $\widehat{c}$ is the EM estimate for c.

The intercept is estimated such that the mean of the fitted values equals for Y_i is the mean of the observed values, i.e. such that $\overline{Y}_i = \overline{\widehat{Y}_i}$. This is exactly how the intercept is constructed in a standard least squares regression.

We summarize several properties of our sparse Bayesian model. First, the prior on $\widehat{\lambda}^2$ is scaled in n,p such that the estimate $\widehat{\lambda}$ achieves the oracle growth rate of $\sqrt{n\log(p)}$ *ex post* when p is of

order n^{α}, $\alpha > 0$, as in the nonparametric setting here. Please note that p is now the number of bases in our regression, not the number of elements in X_i. Second, the rate at which the oracle bound holds is controlled by $1 - exp\{-\sqrt{\log(n)}\}$, which approaches 1 in the limit. The prior is proper whenever $n \times (\log(n) + 2\log(p)) - p > 0$. For example, with $n \in \{100; 1,000; 10,000\}$, this requires $p < 1,978; 27,339; 347,529$ respectively, which is slower than the LASSO rate of $n \sim \log(p)$, but clearly allows for $p > n$. This slower growth rate does suggest the utility of the an initial screen for covariates. Last, even with an improper prior, the posterior will be proper, so estimation can still proceed.

The prior weights w_k are constructed so that the MAP estimate is similar to the adaptive LASSO of Zou (2006). Each mean parameter, c_k, will have its own adaptive penalty $\frac{\widehat{\lambda}\widehat{w}_k}{n\widehat{\sigma}}$.The estimated weights $\widehat{w}_k$ are inversely related to the magnitude of the parameter estimates $\widehat{c}_k$. The adaptive penalty term is of order $\sqrt{\log(n)/n}$ when $\widehat{c}_k \to 0$, which is not asymptotically negligible. On the other hand, the adaptive penalty is of order $1/n$ when $\widehat{c}_k \not\to 0$, and hence is asymptotically negligible. For proofs, see below.

Third, the parameter $\widehat{\gamma}$ adapts to the global sparsity level of the data. If we take $\gamma \to 0$, then the prior approaches a degenerate 'spike-and-slab' prior uniform over the real number line but an infinite point mass at 0. In this scenario, we are not shrinking any effects except for those exactly zero. At the other extreme, $\gamma \to \infty$, the prior approaches a Bayesian LASSO of Park and Casella (2008), regularizing every term. For proof, see below. The global tuning parameter, $\widehat{\gamma}$, is estimated from the data and adjudicates between these two extremes; the utility of nonseparable priors over the tuning parameters have also been discussed by Bhattacharya et al. (2015); Rockova and George (Forthcoming), though we note that these priors were not tuned to achieve the oracle property (or similar concentration property) *ex post*.

The key difference between the LASSOPlus and Predictive LASSOPlus is the adjustment of the prior on λ^2 from the original work. In this earlier work, the prior was constructed off considerations of correct variable selection in the limit, basically zeroing out irrelevant effects and estimating the non-zero effects. In this work, we are concerned primarily with prediction, which is a conceptually different task than variable selection. To give an intuition of the difference, in prediction, we are willing to tolerate irrelevant variables that correlate with relevant ones, so long as their inclusion increases predictive accuracy, whereas with variable selection this would be considered an error.

Estimation Both the MCMC and EM estimation details are standard and follow that of the original LASSOplus model.

Properties of the tuning parameter. The Oracle rate is achieved when λ/n grows as $\sqrt{\log(p)/n}$, i.e. when λ grows as $\sqrt{n\log(p)}$. Our conditional posterior density of λ^2 is

$$\lambda^2|\cdot \sim \Gamma\left(n \times (\log(n) + 2\log(p)), \sum_{k=1}^{p} \tau_k^2/2 + \rho\right) \tag{51}$$

which possesses several desirable properties. First, each mean parameter c_k is given its own shrinkage parameter, $\frac{\lambda w_k}{n\sigma}$, allowing for differential regularization of large and small effects. Second, for $p \sim n^\alpha, \alpha > 0$, then $\widehat{\lambda} = O\left(\sqrt{n\log(p)}\right)$, as desired. Second, the tuning parameter has the structure given in Buhlmann and van de Geer (2013, Corollary 6.2) so that it provides a consistent estimate of the conditional mean. The term $\log(n)$ controls the probability with which the Oracle bound holds, and it goes to zero in n.

Properties of the adaptive weights. Consider the two limiting cases $|\widehat{c}_k| = 0$ and $|\widehat{c}_k| \to \infty$. Consider first $|\widehat{c}_k| \to \infty$. In this case, the exponent in the conditional kernel of w_k is dominated by the term $-\frac{\lambda w_k}{\sigma}|c_k|$ and approaches an exponential density with mean $\widehat{w}_k \to \widehat{\sigma}/\widehat{\lambda}$. Therefore, as $|c_k|$ grows, the weighted LASSO penalty approaches $1/n$, a negligible term.

When $|\widehat{c}_k| = 0$, the conditional posterior density follows a generalized Gamma density with kernel $\exp\{-w_k^{\widehat{\gamma}}\}$ and has mean $\Gamma(2/\widehat{\gamma})/\Gamma(1/\widehat{\gamma})$ which we denote $\overline{\gamma}$.[33] Therefore, for the zeroed out parameters, the weighted LASSO penalty for $|c_k|$ approaches $\widehat{\lambda}\overline{\gamma}/n$. Since $\widehat{\lambda} = O(\sqrt{n\log(n)})$, the penalty is of order $\log(n)/\sqrt{n}$, which is not negligible.

The global sparsity parameter. The global sparsity parameter γ serves to pool information across the mean parameters (see e.g. Bhattacharya et al., 2015, for a similar insight). To illustrate its workings, we consider the two limiting cases, $\gamma \in \{0, \infty\}$.

First, as γ approaches 0, our prior approaches the spike-and-slab prior for a mean parameter, resulting in no shrinkage but with a point mass at zero. Taking $\gamma \to 0 \Rightarrow \Pr(w_k|\cdot) \to Exp(\lambda|c_k|/\sigma)$. This gives us $\widehat{w}_k = \left(\frac{\widehat{1}}{\sigma^2}\right)^{-1/2}/(\widehat{\lambda}|\widehat{c}_k|)$. Plugging into the prior on c_k gives $\Pr(c_k) \to DE\left(\widehat{\lambda}\widehat{w}_k\left(\frac{\widehat{1}}{\sigma^2}\right)^{1/2}\right) \to DE\left(\widehat{\lambda}\left(\frac{\widehat{1}}{\sigma^2}\right)^{-1/2}/(\widehat{\lambda}|\widehat{c}_k|)\left(\frac{\widehat{1}}{\sigma^2}\right)^{1/2}\right) \propto 1$, the flat Jeffreys prior when $\widehat{c}_k \neq 0$. When $\widehat{c}_k = 0$, the prior has infinite density, to due the normalizing constant of order $1/|\widehat{c}_k|$, giving a spike-and-slab prior with support over the real line.

[33] $\Gamma(\cdot)$ refers to the gamma function (not density)

Taking $\gamma \to \infty \Rightarrow \Pr(w_k|\cdot) \to U(0,1)$, a uniform on $[0,1]$, since any weight greater than 1 has mass proportional to $\exp\{w^{-\gamma}\} = 0$. This gives us $\widehat{w}_k = 1/2$. Plugging into the prior on c_k gives $\Pr(c_k) \to DE\left(\frac{1}{2}\widehat{\lambda}\left(\frac{\widehat{1}}{\sigma^2}\right)^{1/2}\right)$, which is the Park and Casella (2008) prior with 1/2 the rate parameter.

F.1 Estimating the Derivative

While numerical differentiation with precision floating point arithmetic often uses values for δ orders of magnitude lower, we are differentiating an estimated model in a direction that is by construction locally linear in T_i. We have found any value of δ below $\sqrt{\text{Var}(T_i)}/1000$ works well.

G Oracle Results

Our Oracle results are rather technical. To help appendix readers follow their construction we first derive the results with a parametric model and then rederive them for the high-dimensional nonparametric model. Then we connect these results to our strategy for estimating compliers in an instrumental variables setting.

G.1 Results with a Parametric Model

Consider the linear model

$$Y_i = T_i c_T + X_i c_X + \epsilon_i \tag{52}$$

with a treatment of interest T_i and single covariate X_i. We also assume the covariate and treatment are scaled such that $\sum_{i=1}^N Y_i = \sum_{i=1}^N T_i = \sum_{i=1}^N X_i = 0$ and $\frac{1}{N-1}\sum_{i=1}^N T_i^2 = \frac{1}{N-1}\sum_{i=1}^N X_i^2 = 1$. [34].

Our first theoretical result is an "Oracle Inequality," which is a bound on how close the fitted values $T_i\widehat{c}_T + X_i\widehat{c}_X$ are to the true values, $T_i c_T + X_i c_X$. Relying on the central limit theorem for $\widehat{c}_T, \widehat{c}_X$ and some algebra gives the result with the parametric model:

$$\sqrt{\frac{1}{N}\sum_{i=1}^N \left\{T_i(\widehat{c}_T - c) + X_i(\widehat{c}_X - c_X)\right\}^2} \leq C\sqrt{\frac{2\sigma^2}{N}}. \tag{53}$$

The basic structure of this bound extends to our high-dimensional, nonparametric result in four regards. First, the number 2 corresponds both with the number of variables (T_i, X_i) and the number

[34] We assume the errors are mean-zero, equivariant, subgaussian, possess a finite fourth moment, and are uncorrelated.

of parameters in the data-generating process. We will differentiate these two in the nonparametric setting, between the number of variables we are considering (p) and the number of non-zero parameters in the data generating process (S), both of which can grow in the sample size in the nonparametric model. Second, the bound is scaled by the residual error σ. Third, a function of N shows the rate at which this bound goes to zero. This difference goes to zero at the parametric rate $\sqrt{N}$. When we shift to a nonparametric model, we will get more functional form flexibility, but at the cost of moving from a rate of $1/\sqrt{N}$ to $\sqrt{\log(N)/N}$.[35] Lastly, the parameter C is a scalar selected such that the bound holds with some probability.

Our second bound is on the treatment effect. The treatment effect admits a simple form in the simple linear model,

$$\lim_{\delta \to 0} \nabla_{T_i}(\delta) = \lim_{\delta \to 0} \frac{(T_i + \delta)c_T + X_i c_X - (T_i c_T + X_i c_X)}{\delta} = c_T. \tag{54}$$

We can again use standard results to appeal to the central limit theorem and establish

$$\sqrt{(\widehat{c}_T - c_T)^2} \leq C(T, X)\frac{\sigma}{\sqrt{N}}. \tag{55}$$

Here $C(T, X)$ is a constant that is a function of the correlation between the treatment and covariate (i.e. is only a function of the design, but not the outcome). The greater $\text{Corr}(T_i, X_i)$, the greater that collinearity inflates the variance of $\widehat{c}_T$ around c_T, increasing the width of the bound.

Lastly, we generate a threshold only on the treatment effect,

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N} \{T_i\,(\widehat{c}_T)\}^2} \leq C(T, X)\frac{\sigma}{\sqrt{N}} \tag{56}$$

under the assumption $c_T = 0$. In this way, observations that fall outside the threshold under a model with $c_T = 0$ are unlikely to be there due to chance, i.e. there appears to be a treatment effect for these observations. We implement a feasible estimate of $C(T, X)$ to separate observations.[36]

G.2 High-dimensional nonparametric model results

We next present our formal results for the high-dimensional nonparametric case. We assume through this section that the outcome and bases have been scaled as in the discussion of the parametric

[35] For illustration, with $N = 100$, the error is of order $1/(100)^{1/2} = 0.1$. Achieving the same error in a nonparametric model would take $N = 648$, as $\sqrt{\log(648)/648} = 0.1$.

[36] When we use the threshold in the instance of first stage IV, the "treatment" becomes the instrument (Z_i) and the "outcome" the endogenous variable (T_i). This forms the basis of our approach to estimating compliers.

case.

An Oracle Inequality Denote as $R(T.,X.)$ the $n \times p$ matrix of bases and $\widehat{\delta} = \widehat{c} - c$. Since our tuning parameter grows at the oracle rate $\sqrt{n\log(p)}$, we can bound the excess risk as

$$\begin{aligned} &\frac{1}{n}\left\{ ||R(T.,X.)\widehat{\delta}||_2^2 + \lambda||\sum_{k=1}^{p}\widehat{w}_k\widehat{\delta}_k||_1^1 \right\} \leq \\ &\quad C\frac{\widehat{\lambda}^2\widehat{\sigma}^2\overline{\gamma}^2|S|}{n^2\phi_0^2} + C_\infty\frac{||R^o_{\infty/n}(T.,X.)c_{\infty/n}||_2^2}{n} + C_\perp\frac{||R^o_\perp(T.,X.)c_\perp||_2^2}{n} \end{aligned} \tag{57}$$

with a probability at least $\exp\{-\sqrt{\log(n)}\}$. The bound splits into three components, a bound off the post-screened basis $R(T.,X.)$, a bound attributable to bases that did not survive the screen but would as n grows, $R^o_{\infty/n}(T.,X.)$, and a bound attributable to the portion of the model that falls outside the span of the basis used in the screen, $R^o_\perp(T.,X.)$. In the first component, ϕ_0 is the smallest eigenvalue of the Gram matrix of the submatrix of $R(T_i,X_i)\widehat{W}^{-1/2}$, $|S|$ is the number of non-zero elements of c and we use $C.$ to denote an unknown constant that does not grow in n,p,S. The next two components are attributable to differences between $R^o(T.,X.)$ and $R(T.,X.)$. The second term is the same order of the first–both are of order $O(\log(n)/n)$; see Appendix G.4 and Fan and Lv (2008) for proofs. The third term is irreducible and of order $O(1)$, corresponding with inescapable misspecification error. We minimize our concerns over the third term through selecting a rich set of basis functions.

An Oracle Inequality on $\nabla_{T_i}(\delta)$ We are interested in not only bounding the prediction risk, as above, but also the error on the $\nabla_{T_i}(\delta)$. To do so, we decompose our bases into two components, one submatrix such that no element covaries with the treatment and one submatrix where for each basis at least one element covaries with the treatment:

$$R(T.,X.) = [R^X(X.) : R^{TX}(T.,X.)]. \tag{58}$$

Denote $\widehat{Y}_i = [R^X(X_i) : R^{TX}(T_i,X_i)]^\top\widehat{c}$, and $\widehat{c}^X$ and $\widehat{c}^{TX}$ the subvectors of $\widehat{c}$ associated with $R^X(X_i)$ and $R^{TX}(T_i,X_i)$. We denote as $\Delta R(T.,X.) = [\Delta R_{ij}] = [\partial R_{ij}/\partial T_i]$ the elementwise partial derivative of $R(T.,X.)$ with respect to the treatment and note $\widehat{\nabla}_{T_i}(\delta) = \Delta R_i(T_i,X_i)^\top\widehat{c}^{TX} = \Delta R_i^{TX}(T_i,X_i)^\top\widehat{c}^{TX}$, which is $\widehat{\beta}_i$ in the formulation of equations 4-5. Denote as $\widehat{c}^{\widehat{S},TX}$ the subvector of $\widehat{c}^{TX}$ selected in the outcome model.

Our second bound is on the mean parameter estimates $\widehat{c}^{\widehat{S},TX}$. The result follows from noting that, since the non-zero elements of the first derivative of our tensor spline bases are elements of

$R^X(X_i)$, then first derivative is also satisfying a LASSO problem:

$$\widehat{c}^{S,TX} = \underset{c^{\widehat{S},TX}}{\operatorname{argmin}} \left|\left|\widetilde{\Delta Y} - \Delta R^{\widehat{S},TX} c^{\widehat{S},TX}\right|\right|_2^2 + \left|\left|\widetilde{W}^{\widehat{S},TX} c^{\widehat{S},TX}\right|\right| \tag{59}$$

where $\widetilde{\Delta Y}$ is a pseudo-response generated from the fitted derivative and estimated residuals from the outcome model, $\widehat{\epsilon}_i$, as

$$\widetilde{\Delta Y_i} = \Delta R_i(T_i, X_i)^\top \widehat{c} + \widehat{\epsilon}_i. \tag{60}$$

The weight matrix $\widetilde{W}^{\widehat{S},TX}$ is diagonal with entries $\widetilde{W}_k^{\widehat{S},TX} = \left|\sum_{i=1}^N R_{ik}^{\widehat{S},TX}(T_i, X_i)\frac{\partial \widehat{\epsilon}_i}{\partial T_i}\right|$. For intuition, note that $\partial\widehat{\epsilon}_i/\partial T_i$ is the causal effect of the treatment on the prediction error, $R_i(T_i, X_i)(c - \widehat{c})$. The less correlated the basis $R_k^{\widehat{S},TX}(T_i, X_i)$ with the impact of a treatment perturbation on prediction error, the less that basis's coefficient is penalized in predicting the treatment effect. The more correlated the basis with the sensitivity of prediction error on perturbing the treatment, the more penalized the coefficient on that basis.

Denote $\widetilde{\Delta Y}$ a pseudo-response generated from the fitted derivative and estimated residuals from the outcome model, $\widehat{\epsilon}_i$, as

$$\widetilde{\Delta Y_i} = \Delta R_i(T_i, X_i)^\top \widehat{c} + \widehat{\epsilon}_i. \tag{61}$$

This setup allows us to give our second theoretical result, the bound

$$\begin{aligned} &\frac{1}{n}\left\{||\Delta R^{\widehat{S},TX}(T_\cdot, X_\cdot)(\widehat{c}^{\widehat{S},TX} - c^{\widehat{S},TX})||_2^2 + ||\widetilde{W}^{\widehat{S},TX}(\widehat{c}_k^{\widehat{S},TX} - c_k^{\widehat{S},TX})||_1^1\right\} \leq \\ &C_\Delta \frac{\widehat{\sigma}^2\overline{\gamma}_\Delta^2|S_\Delta|}{N^2\phi_\Delta^2} + C_\infty \frac{||\Delta R^o_{\infty/\widehat{S}}(T_i, X_i)c_{\infty/n}||_2^2}{n} + C_\perp \frac{||\Delta R^o_\perp(T_i, X_i)c_\perp||_2^2}{n} \end{aligned} \tag{62}$$

where all the constants in the inequality are analogous to those in Inequality 62 but defined in terms of the design matrix that parameterize the derivative.[37]

This oracle bound gives us a bound not on the predicted values but on the predicted first derivative. Bounds of this type have been used for primarily descriptive purposes, to establish rates of convergence of different estimates. We use the bound constructively, to help us differentiate values for which the local treatment effect is zero from those where it is not. The bound helps us in the instrumental variable setting to estimate off which observations we are identifying an effect. We introduce a feasible estimate below in the context of an instrumental variables analysis.

[37] We make the same assumptions on the model in Equation 61 as we do on the outcome model.

G.3 Threshold for Estimating Compliers

The LICE only exists for observations for which the encouragement is non-zero. To estimate these observations, we implement a feasible estimate of the Oracle bound, as given in Inequality 62, to allow us to differentiate observations with some systematic perturbation from encouragement estimates that are likely noise. Since the Oracle bound does not vary across the sample, we implement an adaptive bound that narrows where we estimate signal and expands in a noisy region, constructed from a pointwise estimated degree of freedom.

Stein (1981) showed in the normal regression model $Y_i \sim \mathcal{N}(\widehat{\mu}_i(Y_i), \sigma^2)$ for generic conditional mean function $\widehat{\mu}_i(Y) : Y \to \widehat{Y}_i$, a degree of freedom estimate can be recovered as

$$\widehat{df}_i = \frac{\partial \widehat{\mu}_i(Y_i)}{\partial Y_i} = \frac{\text{Cov}(Y_i, \widehat{\mu}_i(Y_i))}{\sigma^2} \tag{63}$$

and the model degree of freedom can be estimated as $\widehat{df} = \sum_{i=1}^{N} \widehat{df}_i$. This definition coincides with the trace of the projection matrix when $\widehat{\mu}_i(Y_i)$ is linear in Y_i.

Decomposing into parts of the design that covary with T_i and those that do not, we can decompose the degrees of freedom into

$$\widehat{df}_i = \frac{\partial [R^X(X_i) : R^{TX}(T_i, X_i)]^\top \widehat{c}}{\partial Y_i} \tag{64}$$

$$= \frac{\partial [R^X(X_i)]^\top \widehat{c}^X}{\partial Y_i} + \frac{\partial [R^{TX}(T_i, X_i)]^\top \widehat{c}^{TX}}{\partial Y_i}. \tag{65}$$

We then take the degree of freedom estimate associated with $\widehat{\nabla}_{T_i}(\delta)$ as

$$\widehat{df}_i^{\nabla} = \frac{\partial [R^{TX}(T_i, X_i)]^\top \widehat{c}^{TX}}{\partial Y_i}. \tag{66}$$

This estimate will be larger the more signal there is at each observation.

The adaptive component of our threshold is of the form

$$\widehat{df}_i^{adapt} = \frac{1/n}{1/n + \widehat{df}_i^{\nabla}} \tag{67}$$

which takes on a value of 1 when there is no signal for observation $i, (\widehat{df}_i^{\nabla} = 0)$. If the true model is additive and linear in the treatment, we would see $\widehat{df}_i^{\nabla} = 1/n$, which gives $\widehat{df}_i^{adapt} = 1/2$. As the model grows more complex at a given point, the threshold will shrink.

We combine the degree of freedom bound and oracle estimate in our threshold as

$$\widehat{\mathbf{1}}\left(\nabla_{Z_i}^{IV}(\delta^{IV}) \neq 0\right) = \mathbf{1}\left(\left|\widehat{\nabla}_{Z_i}^{IV}(\delta^{IV})\right| > C\frac{\widehat{\lambda}||\widehat{w}_k||_\infty \widehat{\sigma}}{n\widehat{\phi}_0}\widehat{df}_i^{adapt}\right) \tag{68}$$

with $||\widehat{w}_k||_\infty$ the largest weight. C is a user-selected constant, but we show in our simulation that taking $C = 2$ helps select observations impacted by the instrument. Lastly, the compatibility constant is the smallest eigenvalue of the covariance of the true model predicting the gradient. We estimate it using columns of the submatrix of the design that contains interactions with the treatment. We find this matrix may be ill-conditioned or even rank-deficient, so we estimate the smallest eigenvalue such that the eigenvalues above it explain 90% of the variance in the selected model.

The geometry of this threshold incorporates a "complexity penalty" in the selection process. If the estimated model for the treatment effect is simple, the design will be low-dimension with a flat spectrum, say a linear model. As the model grows more complex, the design will incorporate more nonparametric bases and its smallest eigenvalues will be closer to zero, inflating the threshold. The adaptive component, $\widehat{df}_i^{adapt}$, serves to adjust for local variation.

G.4 Proofs of Oracle Inequalities

We condition on $\widehat{W} = diag(\widehat{w}_k)$ throughout the proof and note that $\widehat{W} = I_p$ reduces the proof to that of the standard LASSO. Throughout, to ease notation, we write $R = R(T_i, Z_i)$, etc. dropping dependence on T_i, Z_i when it is clear from context. We also assume that Y and R_k have sample mean zero, so we do not have to worry about an intercept.

Start with the LASSO problem

$$\widehat{c} = \underset{\widetilde{c}}{\operatorname{argmin}} \, ||Y - R\widetilde{c}||_2^2 + \lambda||W\widetilde{c}||_1^1. \tag{69}$$

As we are minimizing over the sample, we know the estimator $\widehat{c}$ satisfies

$$||Y - R\widehat{c}||_2^2 + \lambda||W\widehat{c}||_1^1 \leq ||Y - Rc||_2^2 + \lambda||Wc||_1^1. \tag{70}$$

After some manipulation (e.g., Buhlmann and van de Geer, 2013) and evaluating at $\widehat{\lambda}, \widehat{W}$, this excess risk generates the inequality

$$\frac{1}{n}\left\{||R\widehat{\delta}||_2^2 + \widehat{\lambda}||\widehat{W}\widehat{\delta}||_1^1\right\} \leq C\frac{\widehat{\lambda}^2\widehat{\sigma}^2\overline{\gamma}^2|S|}{n^2\phi_0^2} + C_\infty\frac{||R^o_{\infty/n}c_{\infty/n}||_\infty^2}{n} + C_\perp\frac{||R^o_\perp c_\perp||_\infty^2}{n} \tag{71}$$

which is Equation 57.

First order conditions on the LASSO and the derivative. The first-order conditions for the LASSO problem above are

$$2R_k^\top\widehat{\epsilon} = s_k\widehat{\lambda}\widehat{w}_k \tag{72}$$

with $s_k \in [-1, 1]$ and $s_k = 1$ or -1 iff $\widehat{c}_k \neq 0$. We consider the subset of the matrix corresponding with these non-zero estimates

$$R^{\widehat{S}} = \{R_k : \widehat{c}_k \neq 0\} \tag{73}$$

and the submatrices $R^{\widehat{S},X}$ and $R^{\widehat{S},XT}$.

Consider the gradient derivative of equality 72 wrt the treatment at each observation and noting, by the identification assumptions (1)-(4), $\frac{\partial}{\partial T_i}\widehat{s}_k\widehat{w}_k\widehat{\lambda} = 0$, which gives

$$0 = \sum_{i=1}^{N} \frac{\partial}{\partial T_i} \sum_{i=1}^{N} \left(R_{ik}^{\widehat{S}} \widehat{\epsilon}_i \right) \Rightarrow \tag{74}$$

$$0 = \sum_{i=1}^{N} \frac{\partial}{\partial T_i} R_{ik}^{\widehat{S}} \widehat{\epsilon}_i \tag{75}$$

$$= \sum_{i=1}^{N} \frac{\partial R_{ik}^{\widehat{S}}}{\partial T_i} \widehat{\epsilon}_i + R_{ik}^{\widehat{S}} \frac{\partial \widehat{\epsilon}_i}{\partial T_i} \tag{76}$$

This gives us

$$\left| \sum_{i=1}^{N} \frac{\partial R_{ik}^{\widehat{S}}}{\partial T_i} \widehat{\epsilon}_i \right| = \left| \sum_{i=1}^{n} R_{ik}^{\widehat{S}} \frac{\partial \widehat{\epsilon}_i}{\partial T_i} \right| \tag{77}$$

which is a solution to the following LASSO problem

$$\widehat{c}^{S,T} = \underset{c^{S,T}}{\operatorname{argmin}} \, ||\widetilde{\Delta Y} - \Delta R^{\widehat{S},T} c^{\widehat{S},T}||_2^2 + ||\widetilde{W}^{\widehat{S},T} c^{\widehat{S},T}||_1^1 \tag{78}$$

which will satisfy an Oracle Inequality if $\widetilde{W}^{\widehat{S},T}$ is of order $\sqrt{n \log(p)}$. We have denoted as $\Delta R^{\widehat{S},T}$ the elementwise partial derivative of $R^{\widehat{S},T}$ wrt T_i and $\widetilde{\Delta Y}$ is a pseudo-response constructed from the estimate $\widehat{\Delta Y} = \left[\frac{\widehat{\partial Y_i}}{\partial T_i} \right]$ as

$$\widetilde{\Delta Y} = \Delta R^{\widehat{S},T} \widehat{c}^{\widehat{S},T} + \widehat{\epsilon}. \tag{79}$$

The model is fit using basis-specific tuning parameter,

$$\widetilde{w}_k^{\widehat{S},T} = \left| \sum_{i=1}^{n} R_{ik}^{\widehat{S}} \frac{\partial \widehat{\epsilon}_i}{\partial T_i} \right| \tag{80}$$

$$= \left| R_k^{\widehat{S},\top} \Delta R^{TX} (c^T - \widehat{c}^T) \right| \tag{81}$$

since $\partial \epsilon_i / \partial T_i = 0$.

Then,

$$\left| R_k^{\widehat{S},\top} \Delta R^{TX} (c^T - \widehat{c}^T) \right| \leq ||R_k^{\widehat{S},\top}||_2 ||\Delta R^{TX} (c^T - \widehat{c}^T)||_2 \tag{82}$$

by Cauchy-Schwarz. The first term on the righthand side is order $\sqrt{n}$ and the second is of order of the square root of the oracle bound on the prediction since each element of ΔR^{TX} is also in R^X, which gives a bound on the weights of order $\sqrt{n \log(p)}$, from which the oracle result follows directly.

H Treatment Effect Simulations

We next present simulation evidence illustrating the MDE's utility in estimating treatment effect. We include four sets of simulations presented in increasing complexity, a linear model, a low-dimensional interactive model, a nonlinear model, and a model with interactions and discontinuous breaks, respectively:

$$\text{Linear:} \quad Y_i = T_i + \sum_{k=5}^{8} X_{ik}\theta_k + \epsilon_i \tag{83}$$

$$\text{Interactive:} \quad Y_i = T_i - T_i \times X_{i3} + \sum_{k=5}^{8} X_{ik}\theta_k + X_{i1}X_{i2} + \epsilon_i \tag{84}$$

$$\text{Nonlinear:} \quad Y_i = 20 \sin\left(\left(X_{i1} - \frac{1}{2}\right)\frac{T_i}{20}\right) + \frac{1}{2} \times (1 + |X_{i3} \times X_{i4}|_+) \times (1 + T_i) + \sum_{k=5}^{8} X_{ik}\theta_k + \epsilon_i \tag{85}$$

$$\text{Discontinuous } Y_i = (1 + |T_i|) \times \mathbf{1}(|T_i| > 1/2) \times (X_{i5} + 1) + \sum_{k=5}^{8} X_{ik}\theta_k + X_{i1}X_{i2} + \epsilon_i \tag{86}$$

where the the error is independent, identical Gaussian such that the true R^2 in the outcome model is 0.5. All covariates are from a multivariate normal with variance one and covariance of 0.5 between all pairs and the elements of β are drawn as independent standard normal. The treatment is generated as

$$T_i = -3 + (X_{i1} + X_{i4})^2 + \epsilon_i^T; \quad \epsilon_i^T \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 4) \tag{87}$$

We consider $n \in \{100, 250, 500, 1000, 2000\}$ and $p \in \{10, 25, 50, 100\}$. In each case, the true data is generated from the treatment and the first 8 covariates, but the dimensionality of the underlying model may be much higher. For example, the smooth curve in simulation 4 is infinite dimensional (a sine curve).

We assess methods across two dimensions, each commensurate with our two estimation contributions: the nonparametric tensor basis for causal estimation and the sparse Bayesian method. We want to assess, first, how well sparse regression methods perform given the same covariate basis. We include in this assessment the cross-validated LASSO, as a baseline, as well as two sparse Bayesian priors: the horseshoe and Bayesian Bridge (Carvalho, Polson and Scott, 2010; Polson, Scott and Windle, 2014), as well as LASSOplus, described above.[38] Each of these methods are handed the same nonparametric basis created in our pre-processing step. To date, the use of these methods have not taken advantage of our nonparametrics basis construction and screening steps. We compare these methods to regression-methods that generate their own bases internally, kernel regularized least squares (KRLS, Hainmueller and Hazlett, 2013), POLYMARS (Stone et al., 1997), and Sparse Additive Models (SAM, Ravikumar et al., 2009). These methods are simply given the treatment and covariates, not our nonparametric basis. The last comparison set are non-regression methods, bayesian additive regression trees (BART, (Chipman, George and McCulloch, 2010)), gradient boosted trees (GBM, Ridgeway, 1999), and the SuperLearner (Polley and van der Laan, N.d.). For the SuperLearner we include as constituent methods random forests, the LASSO, POLYMARS, and a generalized linear model. All simulations were run 500 times.

H.1 Results

We report results on each methods' ability to recover the conditional mean $\mu_i = \mathbb{E}(Y_i|X_i, T_i)$ and partial derivative $\partial\mu_i/\partial T_i$. To summarize the performance, we calculate three statistics measuring error and bias. The statistics are constructed such that they are 0 when $\mu_i = \widehat{\mu}_i$ and $\partial\mu_i/\partial T_i = \widehat{\partial}\mu_i/\partial T_i$ and take a value of 1 when $\widehat{\mu}_i = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}_i$.

[38] We included the OLS post-LASSO estimator of Belloni and Chernozhukov (2013) implemented in **R** package `hdm`. We found the method performed practically identically to cross-validated LASSO, when applied to our bases.

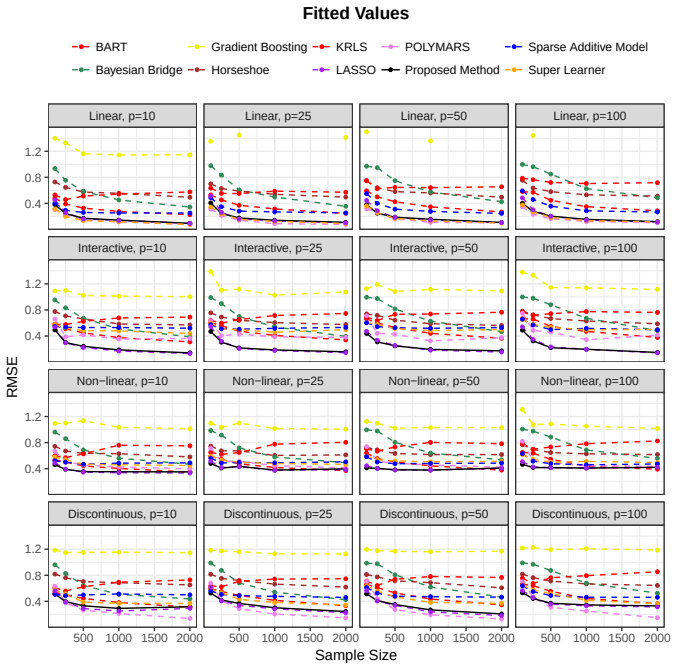

Figure 7: **A Comparison of RMSE for Fitted Values.**

$$\text{RMSE}: \sqrt{\frac{\sum_{i=1}^{n}(\mu_i - \widehat{Y}_i)^2}{\sum_{i=1}^{n}(\mu_i - \overline{Y}_i)^2}} \tag{88}$$

$$\text{RMSE}_{\nabla}: \sqrt{\frac{\sum_{i=1}^{n}\left(\frac{\partial \mu_i}{\partial T_i} - \frac{\widehat{\partial Y_i}}{\partial T_i}\right)^2}{\sum_{i=1}^{n}\left\{\frac{\partial \mu_i}{\partial T_i}\right\}^2}} \tag{89}$$

$$\text{Bias}_{\nabla}: \frac{\left|\sum_{i=1}^{n}\frac{\partial \mu_i}{\partial T_i} - \widehat{\frac{\partial \mu_i}{\partial T_i}}\right|}{\sqrt{\sum_{i=1}^{n}\left\{\frac{\partial \mu_i}{\partial T_i}\right\}^2}} \tag{90}$$

In the figures, a value of 1 can be interpreted as performing worse than the sample mean, in either a mean-square or bias sense, and values closer to 0 as being closer to the truth. A value above 1 when estimating the derivative means the method did worse than simply estimating 0 for each value, the first derivative of the null model $\widehat{\mu}_i = \overline{Y}_i$. For presentational clarity we remove any results greater than 1.5 for the RMSE results and .2 for the bias results.

Figure 7 reports the RMSE for the fitted values, by method. We can see that MDE, the proposed method, performs well in each situation. The first column contains results from the simple additive linear model, where MDE, POLYMARS, and the SuperLearner are all competitive. We suspect

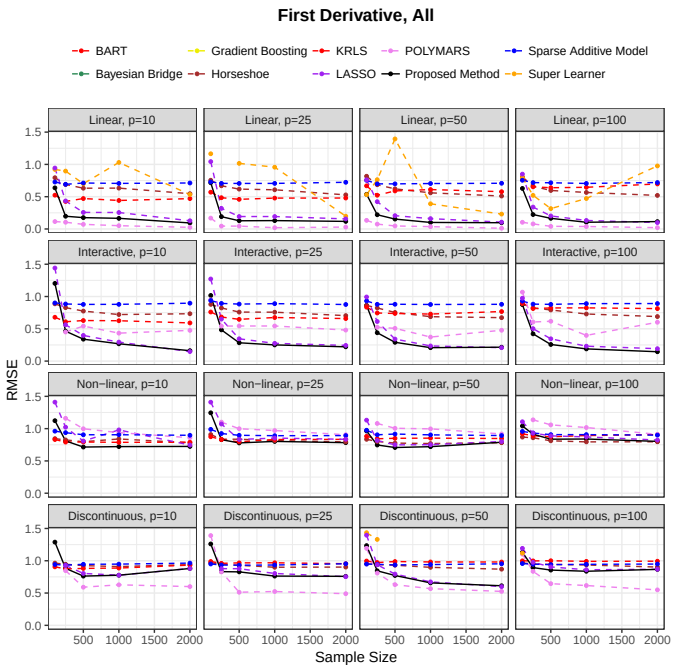

Figure 8: **A Comparison of RMSE for Unit Level Derivatives Across Methods.**

SuperLearner is competitive here because the true model is in the model space for OLS, one of the components included in the SuperLearner. MDE, POLYMARS and LASSO are nearly tied for the best performance in the interactive and non-linear settings.[39] Finally while MDE is in a top performing set for the discontinuous case, POLYMARS had a slight advantage.

Figure 8 presents results on each methods' error in estimating the derivative. In the linear model, we are consistently dominated by POLYMARS and beat SuperLearners as p grows. In the interactive and non-linear settings, MDE is again aligned with the LASSO and POLYMARS, while POLYMARS performs worse than the null model in the non-linear setting, with the horseshoe competitive. POLYMARS performs the best in the discontinuous case but MDE was not far behind. We find that, across settings, LASSOplus is the only method that generally performs first or second best in RMSE, though POLYMARS, particularly in the discontinous case, and cross-validated LASSO are reasonable alternatives.

This highlights how estimating fitted values well need not result in recovering the treatment effect well. Most of the systematic variance in our simulations is attributable to the background covariates, and a method could perform well simply by getting these effects correct but missing

[39]LASSO methods with a linear specification in the covariates, i.e. not using our nonparametric bases, performed poorly so were not included.

the impact of the treatment. For example, BART and Sparse Additive Models perform relatively well in estimating fitted values in the interactive, and nonlinear but not much better than the null model in estimating the partial derivative.

We perform well in our simulations, and we share the concern that this performance is the result of favorable decisions we made in the simulation design. To help alleviate these concerns, we use ordinary least squares to decompose the first derivative into two different components, one that correlates with the treatment and one that does not. Examining the two separate components helps identify the extent to which each methods captures the low-dimensional linear trend in the first dimension and the extent to which they capture the residual nonlinear trend. In our first simulation, there is no nonlinear trend in the fitted values, and the first derivative is flat, so any curvature found in the fitted values will show as error in the second component of the RMSE.

Results from this decomposition can be found in Figure 9 and Figure 10 . Figure 9 presents results for the low-dimensional linear trend while Figure 10 presents results for residual non-linear trend terms. We find that the proposed method performs well for both sets of effects. Interestingly in Figure 10 we see that POLYMARS does more poorly in the non-linear setting but better in the discontinuous setting.[40]

I Instrumental Variables Simulations

I.1 Simulations

We next present a short set of simulation results to examine the performance of our method applied to the case of an encouragement design (instrumental variable). For simplicity we compare our proposed method to two-stage least squares estimation.[41] We examine first stage and second stage performance, as well as how well the proposed method helps to unpack individual level compliance estimates.

[40]Beyond individual effects, sample average effects though, may be of interest to the researcher or policy maker. For that reason, we also considered the level of the bias and generally find similar results though the Bayesian Bridge performed better than in the previous results.

[41]The conclusion discusses connections with Chernozhukov, Hansen and Spindler (2015); Belloni et al. (2012) that investigate the case of variable selection, rather than variable and functional selection, in the estimation of a structural parameter rather than individual effects. Methods used in Section H have not been extended to the instrumental variables context except for ensemble tree methods by Athey, Tibshirani and Wager (2016).

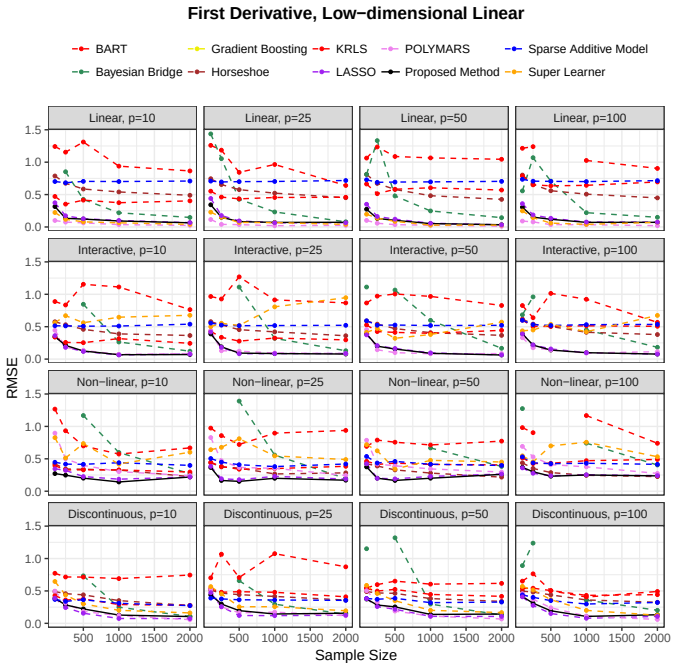

Figure 9: **A Comparison of RMSE Across Methods. Linear terms.**

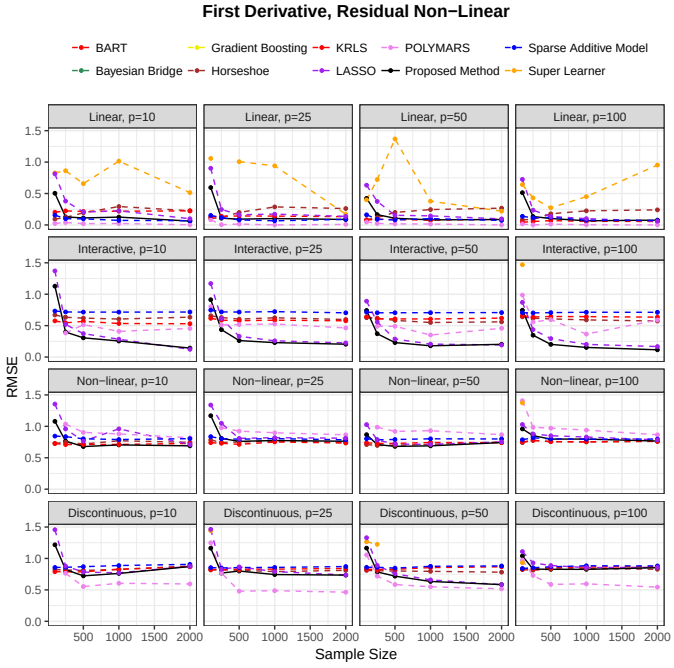

Figure 10: **A Comparison of RMSE Across Methods. Residual non-linear terms**

Simulation Environments We consider four different simulation settings. We fix the second stage (outcome) model but vary the first stage (treatment) model across five settings. The first is a null model, in which the instrument does not impact the treatment and therefore no observation complies with the instrument. In the second, the instrument has an additive linear effect on the

treatment. The third scenarios includes an interaction term between the instrument and one of the covariates. The fourth scenario adds an additional nonlinearity in the instrument.

In each case, the instrument Z_i is generated as

$$Z_i = -1 + X_{i1} + X_{i1}^2 + \epsilon_i^Z; \;\; \epsilon_i^Z \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 9). \tag{91}$$

and, also in each case, the outcome is generated as

$$Y_i = T_i \times (1 + X_{i2})^2 + X_i^\top \beta + \epsilon_i^Y. \tag{92}$$

where $\beta = [1, 1/2, \ldots, 1/K]$ in all cases. We generate the treatment using four different models of increasing complexity:

$$\text{Null: } T_i = X_i^\top \beta + \epsilon_i^T \tag{93}$$

$$\text{Linear: } T_i = Z_i + X_i^\top \beta + \epsilon_i^T \tag{94}$$

$$\text{Interactive: } T_i = Z_i \times (1 + X_{i2})^2 + X_i^\top \beta + \epsilon_i^T \tag{95}$$

$$\text{Nonlinear: } T_i = max(Z_i + 2, 0) \times (1 + X_{i2})^2 + X_i^\top \beta + \epsilon_i^T \tag{96}$$

$$\tag{97}$$

In each, the errors are simulated such that

$$[\epsilon_i^T, \epsilon_i^Y]^\top \overset{\text{i.i.d.}}{\sim} N\left([0,0]^\top, C \begin{bmatrix} 1, .9 \\ .9, 1 \end{bmatrix}\right) \tag{98}$$

where C is selected such that the second stage has a signal to noise ratio of 1.

Results In each setting, we compare the performance of MDE to TSLS. We consider three sets of performance measures. The first two are RMSE performance for the first and second stages. The second stage RMSE is only measured off the observations with an in-truth identified causal effect. Third, we consider the ability of MDE to differentiate compliers from defiers and for our threshold to identify non-compliers.

Figure 15 presents RMSE estimates for the first stage model. In the null simulation environment, the first stage in TSLS (OLS) slightly out performs the proposed method at very low sample sizes and for cases with very few in truth covariates. This advantage quickly goes away. MDE becomes competitive with the linear simulation environment in a similar way. MDE dominates once there are interactions and non-linearities in the data generating process.

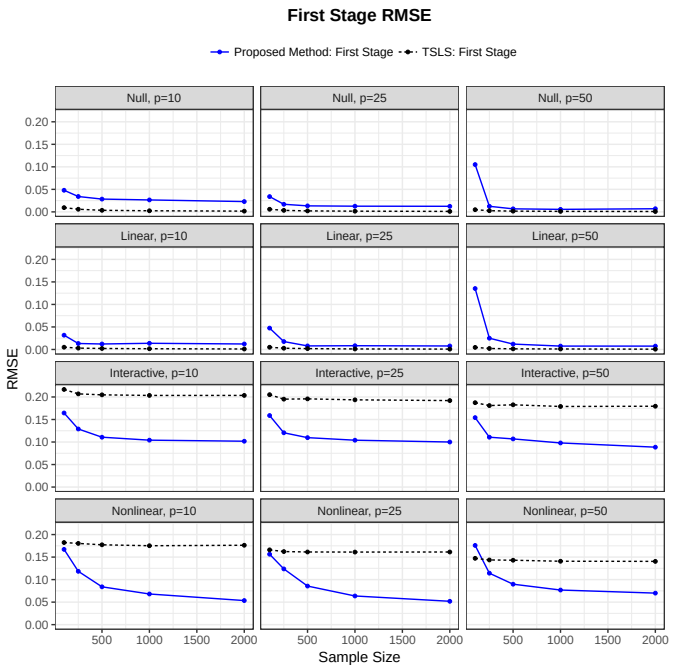

Figure 11: **RMSE estimates for proposed method and 2SLS for first stage estimation across simulation environments.**

Figure 12 presents the second stage results. Here we remove the null model results as the causal effect was not identified for any observations. In the subsequent settings the RMSE is lower in almost all cases for the proposed method. For example, even when the first stage is linear, the fact that the second stage has a non-linear and interactive relationship means MDE will beat TSLS. This highlights the importance of having to ex ante get the model "right" in both stages when using conventional methods.

Compliance Estimates Figure 13 presents the *true* percentage of cases by compliance status in the first stage. This will let us evaluate how well compliance status is recovered. As designed the null model had only types that were unaffected by the instrument (zero types), the linear and interactive setting had only positive types, and the threshold model had mostly positive effects but some zero effects.

Figure 14 examines how well we estimate the correct sign for first stage using our threshold (as used in Figure 4 of the main paper). In our simulations we know if someone was positively encouraged by the instrument, negatively encouraged, and not encouraged (zero). We then recorded the model's individual level estimates and calculate the percentage of observations correctly classified

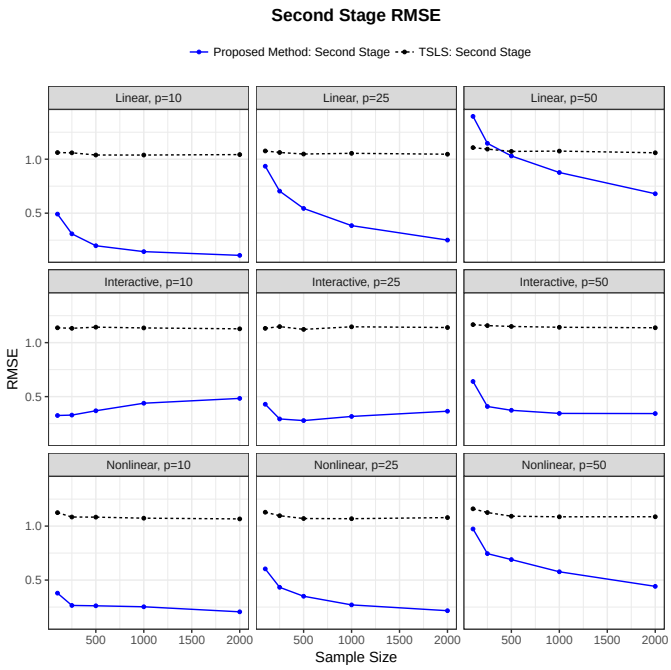

Figure 12: **RMSE estimates for proposed method and TSLS for second stage estimation across simulation environments.**

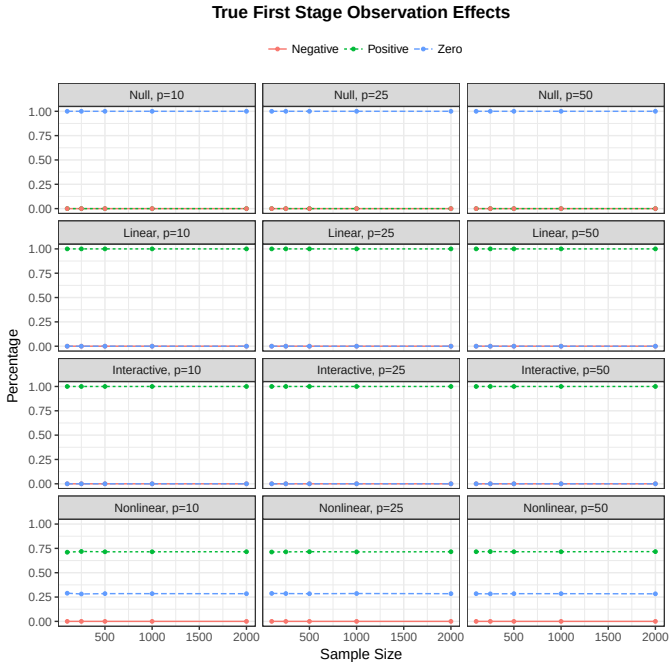

Figure 13: **Percentage of observations in sample by effect type.**

into each bin. This is defined only in cases where there were some observations in the category. In the first row there are in-truth no observations encouraged by design. For this reason, only the

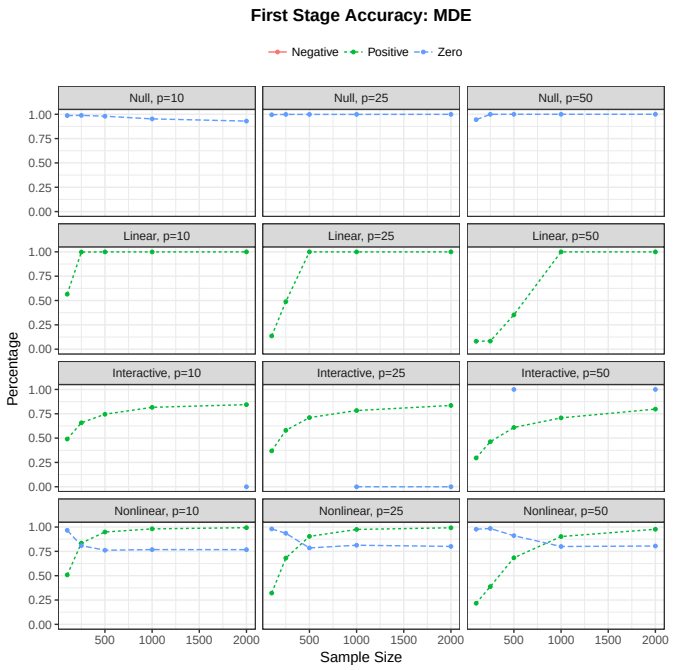

Figure 14: **First stage accuracy for proposed method.** Green line plots the percentage of in fact positive LICE that we correctly capture. The red line plots the percentage of in fact negative LICE that we correctly capture. The blue line plots the percentage of in fact non-encouraged units that we correctly capture.

plot for the "zero effect" types are plotted. The model recovers these effect types nearly perfectly. The second linear model had only positive effect types and the model quickly pegs these effects in sample size. In the third, interactive simulation design, the model recovers the positive effect types with increasing sample.[42] Finally the last simulation had both positive and zero effect types. Here these types are recovered thought the performance is slightly worse for the rarer (see Figure 13) zero group.

TSLS of course is unable to differentiate compliance status at the observation level. TSLS can only return one compliance estimate for the sample. We recorded all observations as not complying if the first-stage F statistic on the instrument was less than 10. Results are in Figure 15. When TSLS did recover a compliance type correctly, it was only for those with a positive instrumented causal effect. In these simulations it records everything as falling in this category and so it picks

[42]While in this case the simulation was tuned to produce no "zero" effects, an extremely small number of observations (less than 5) occasionally were produced due to extreme draws. In these very rare cases the model did not correctly identify their type.

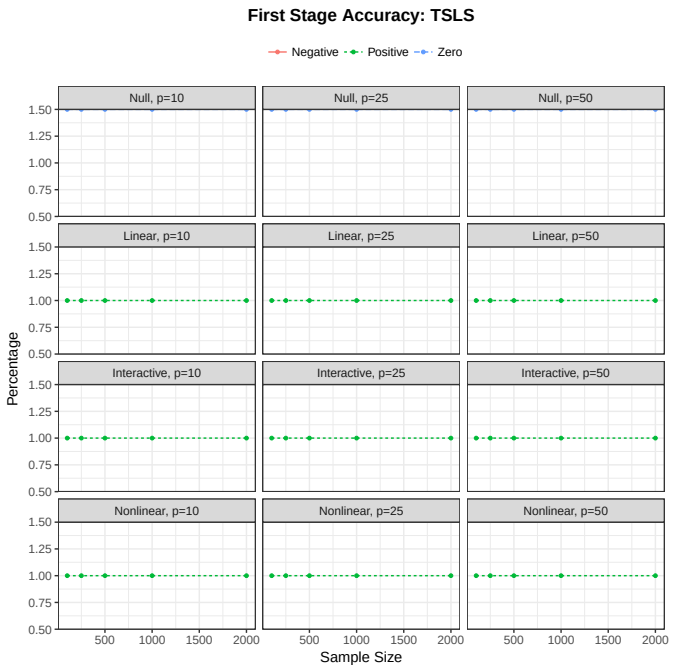

Figure 15: **First stage accuracy for TSLS. TSLS pegs all observations to be positively encouraged and so correctly captures these observations. No other compliance types are identified.**

up correctly those positively encouraged units. This is obviously misleading.

J Related Work

We provide here a more comprehensive literature review of our work, particularly as it relates to the broader statistical and machine learning literatures.

Relationship to causal inference Causal estimation generally involves a two-step procedure. In the first, a model of the treatment is used to characterize any confounding. In the second, some summary statistic from the first stage, such as a the density estimate or conditional mean of the treatment, is incorporated into the outcome model through matching, subclassification, or inverse weighting. See Imbens and Rubin (2015) for an overview.

Misspecification of this treatment model is both inevitable and impactful (Kang and Schafer, 2007). We sidestep the entire endeavor and directly target a fully saturated, nonparametric conditional mean function. In doing so we must make assumptions on the nonparametric model (e.g., Robins and Ritov, 1997), yet we have worked to make the weakest possible assumptions: that the outcome model is additive in some set of a massive set of flexible, interactive nonparametric bases.

The bulk of the literature on causal inference has focused on the case of the case of binary or categorical treatment regimes (For exceptions see Imai and Van Dyk, 2012; Hirano and Imbens, 2005, though these works focus on average, not conditional, effects). The idea of using machine learning to directly estimates counterfactuals with a binary treatment has been explored (e.g. Hill, 2011), though these works did not explore structural models or continuous treatments.

We also share a motivation with the targeted maximum-likelihood approach of van der Laan and Rose (2011). The authors use a cross-validation tuned ensemble of methods, or "SuperLearner," to predict the treatment density, then incorporate these estimates in the second stage model so as to achieve semiparametrically efficient estimates of a treatment parameter.[43] Unlike this method, we target observation-level estimates rather than an aggregate parameter. In simulations we find the SuperLearner performs poorly in estimating the derivative of the loss function, as minimizing the predictive risk may result in poor estimates of a derivative or treatment effect.[44]

Estimation of first derivatives. In the case of a continuous or binary treatment, our method reduces to estimating the partial derivative or subderivative, respectively, of the response function with respect to the treatment variable. The econometric literature has long recognized this manipulation-based, and hence causal, interpretation of a structural equation model (Haavelmo, 1943; Hansen, 2017; Pearl, 2014; Angrist and Pischke, 2009). We differ by targeting observation-level counterfactuals rather than a parameter in a structural model. Even absent a causal interpretation, estimating derivatives has been motivated by problems in engineering as well as economics; see O'Sullivan (1986); Horowitz (2014) for overviews and Hardle and Stoker (1989); Newey (1994) for seminal work. We differ primarily in estimation, through considering a large-p setting as well as models that are nonparametric in all covariates. These methods also stay silent on how to choose a basis, whether B-splines, Hermite polynomials, radial basis functions, and so on. We leverage our ignorability assumption, which is central to causal inference, to show that the appropriate basis function for a causal effect is locally linear almost everywhere.[45]

[43]For application of ensemble methods in political science, see Samii, Paler and Daly (2016); Grimmer, Messing and Westwood (2017).

[44]See Athey and Imbens. (2016) and Horowitz (2014), especially Section 3.1 for more on the distinction between minimizing predictive error and minimizing error in the causal estimate.

[45]I.e., the piecewise linear cubic-spline or B-spline basis.

Relationship to High-Dimensional Regression Modeling The introduction of kernel-regularized least squares (KRLS, (Rosipal and Trejo, 2001)) to political science by Hainmueller and Hazlett (2013) is a close and welcome cousin. We differ in several ways. First, while both of our approaches make use of nonparametric bases (to avoid relying particular functional form assumptions), we differ in the types of bases considered. KRLS constructs "isotropic" nonparametric bases, which means the bases are a function of the distance between a given observation's covariates and the covariates of all other observations. This means the number of basis functions is equal to the sample size and, importantly, because a distance metric is used all variables are weighted equally. Put differently, with isotropic bases each parameter is associated with a single observation. MDE uses "anisotropic" bases which constructed from interaction between the treatment and other control variables, as well as interactions between other control variables. This means that the number of basis function used is dramatically higher in MDE and that a particular variable is targeted (the treatment variable). Each parameter is associated with a function of the covariates and unlike KRLS the Method of Direct Estimation targets a particular parameter. An implication is that KRLS is not tailored for the estimation of individual causal effects in the case of a continuous treatment variable.[46] Following from this, the theoretical guarantees for KRLS's predictions do not extend to the estimation of an ICE. A final difference of course is that, in order to deal with the very large number of basis functions and interactions considered, we use a two stage screening/sparse estimation strategy yielding a small subset of nonzero estimates associated with interpretable bases. The bases may be local nonparametric functions, but we can map these bases directly back to substantive variables of interest. As a result, with MDE, researchers can directly examine whether the model contains interactive effects. With KRLS only average partial derivatives for individual variables are returned. Researchers need to explore predicted values from the model under different covariate configurations to identify heterogeneities, with no guidance from the model.

While political scientists might be more familiar with valuable methods like Hainmueller and

[46]In the case of a binary treatment, the methods become slightly more similar as the KRLS implementation in (Mohanty and Shaffer, 2017) allows for treatment by covariate interactions as a special option. Here too there exist differences in terms of bandwidth selection. KRLS assumes that the distance between two observations is calculated the same everywhere, as controlled by a *single* bandwidth parameter. MDE includes a large number of bases all with differing bandwidths. Selecting a bandwidth well is more important in practical estimation than the particular shape of the bases (Hansen, 2017, ch. 14.6).

Hazlett (2013), our anisotropic spline basis is closest to the nonparametric specification in the "POLYMARS" multivariate tensor-spline model of Stone et al. (1997). We differ from POLYMARS in two regards. First, our screening step allows us to include an "ultra-high" number of candidate bases (Fan and Lv, 2008). We include spline bases of multiple degrees and interactions, then reduce the millions of possible bases to hundreds or thousands. Second, rather than conduct Rao and Wald tests for basis inclusion/exclusion, we use a sparse model to fit all of the maintained bases at once. Third, unlike existing sparse Bayesian models (e.g. Polson and Scott, 2012; Rockova and George, Forthcoming), we use a frequentist minmax ("oracle") argument to motivate our hyperprior selection as a function of n, p.

We were inspired by work on regression models that allow for interactive nonparametric ("tensor-product") terms (see Gu, 2002, for an overview). We differ from these methods primarily in scale. We are considering tens or hundreds of covariates, a combinatorially growing number of tensor products, while focusing on the impact of only a single treatment. Due to the complexity of the problem, the proper specification of these models is computationally infeasible.[47]

Another popular form of nonparametric modeling is kernel estimation. Kernel estimators fit a local, and normally linear model, but incorporate nonlinear weights so observations near a point of interest are upweighted relative to distant observations (Härdle et al., 2012). While useful and powerful, any nonlinearities are incorporated into observation-level weights. If the distance is constructed from several covariates, it requires some post-processing, such as post-hoc analysis of fitted values, to map uncovered nonlinearities to covariates of interest.

Relation to nonparametric and high-dimensional structural models Recent work has extended nonparametric and high-dimensional estimation to the instrumental variables setting, through either series estimation or LASSO selection (Newey, 2013; Belloni et al., 2012). Like these methods, we use a regularized series estimator to recover a conditional mean. We differ from these methods in that we construct our estimator from observation-level predictions under counterfactual manipulations, rather than targeting structural parameters. Recent work by Belloni et al. (2015) develop theory for both estimation and inference on a broad class of models under least squares estimation; included in their analysis is inference on the instantaneous causal effect in a tensor-spline

[47] These models would require inverting a matrix of order millions by millions, while also doing a grid search for dozens of tuning parameterd, which is infeasible given current and even foreseeable computational power.

model. The proliferation of bases in this model serves as the central issue in estimation (Belloni et al., 2015, Sec. 3.6), which we address through our screen and regularization procedure. We also extend the approach to additional quantities of interest beyond those considered in Belloni et al. (2015), such as those studied in instrumental variable and mediation analyses. We last note that these models focus on root-N unbiased and efficient estimation of a finite-dimensional parameter in the presence of high-dimensional confounding. Instead, we are interested in consistent estimates of heterogeneous, and possibly, high dimensional effects. While we gain in recovering a more complex model, our average estimates are plagued by persistent root-N bias (Robins and Ritov, 1997), though this is not a concern for us.

Our approach to instrumental variables is perhaps closest to Heckman and Vytlacil (2005, e.g.)[48], Athey, Tibshirani and Wager (2016), and Hartford et al. (2016). Heckman and Vytlacil develop the Local IV (LIV) in the binary treatment/encouragement setting as the impact of the treatment on those indifferent between control and treatment at a given level of encouragement. The LICE considers the impact of a perturbation to the encouragement at the observed data. In a policy setting, where the encouragement can be controlled, the LIV may be preferred, while in an observational setting, where interest is on the data as observed, the LICE is likely to be preferred. We emphasize, though, that MDE can be used to estimate either the LICE or LIV, using a rich mean model unavailable to these previous papers, and handles settings beyond a binary treatment/encouragement.

In additional related work, Athey, Tibshirani and Wager (2016) use a random forest to estimate a kernel density for each observation, then use the density weights in a two-stage least-squares calculation. As with us, Athey, Tibshirani and Wager (2016) generate observation-level effect estimates and use a high-dimensional model to avoid the curse of dimensionality. Like Hartford et al. (2016), we use a flexible regression model to estimate conditional mean functions. While we generate pointwise counterfactuals at the two stages, Hartford et al. (2016) uses the first stage for a second-stage residual correction, which is the "control function" approach described in Horowitz (2014). Our primary additions over Athey, Tibshirani and Wager (2016) and Hartford et al. (2016) come from both returning a first-stage estimate, which offers important insight into internal validity, and using the oracle bound to identify non-compliers in the data.[49]

[48] See Kennedy, Lorch and Small (2017) for concurrent, and fascinating, work.

[49] The importance of separate estimates at each stage will allow plug-in estimates for other causal structural models (VanderWeele, 2015). In separate work, we illustrate how the MDE extends to mediation and sequential-g estimation.

K Recovered variables/functional forms

One of the advantages of using a sparse linear model over other methods, such as a regression tree or neural network, is the ability to match these effects back to covariates of interest. We present the largest estimated effects at the first (Table 3) and second stage (Table 4) for coefficients that interact with the instrument in the first stage or treatment in the second stage. Our software automatically standardizes all terms, parametric and nonparametric, so the coefficients discussed below are all on a comparable scale within either the first stage model or second stage model. Under the assumptions that the instrument is valid and monotonic within each stratum given by the covariate profile, the second stage coefficients that drive the effect of the turbine on voteshare can be interpreted causally.

The main body discusses several interactions of interest. We note here that some interactions will be less interpretable such as interactions between spline components for latitude, longitude, and distance to the nearest Great Lake. At a minimum these serve as a nonparametric means of "controlling" for other effects that could confound variables of interest. Interactions with district fixed effects (districts 36 and 41) suggest some unmodeled heterogeneity across districts.

Tables 3 and 4 report bootstrapped confidence intervals. To calculate these on coefficients in a structural model like instrumental variables using the MDE is more challenging than in a single stage context.

One result of this is that sometimes estimated coefficients will lay outside of the bootstrap interval. This is the case, especially, for some of the second stage coefficients. We note, though, that the confidence interval and point estimates are never on opposite sides of zero. In standard parametric models, the point estimates, which maximize the likelihood, also center the confidence intervals, which have asymptotic coverage. The two are connected through the central limit theorem and especially the finite-dimensional nature of the problem. Our model has neither property. Our model grows in sample size, and our prior does not disappear quickly enough, to be negligible in the limit. Unfortunately, many of our intuitions fail in this setting. There is little reason to suspect that the point estimates that maximize the posterior will, even in the limit, fall within an estimated bootstrap confidence or credible interval, which are constructed to achieve either coverage or concentration on the true value. Such results can be guaranteed through complexity assumptions on the true underlying function class and prior density (i.e. concentration rates of the ϵ-net, through

either an entropy bound or P-Donsker assumption; the model space being considered is close to the true model space; sufficient density in the prior near the true model, as measured by KL divergence; see Ghosal, Ghosh and van der Vaart (2000, esp. 2.2–2.4) or Ghosal and van der Vaart (2017, ch. 9–12)). We have found that when we return back to a low-dimensional ("parametric") mode of inference by looking at effects averaged over the sample, this problem disappears; see Table 1 for an illustration.

Bases	Coefficient	Bootstrap CI
Population Density, log × Average Wind Power, log	-0.027	(-0.0186, -0.00683)
University Degree (%) × Average Wind Power, log	-0.022	(-0.0168, -0.00817)
b(Average Wind Power, log) × b(Distance to Great Lake) × longitude	-0.0135	(-0.0347, -0.0022)
b(Average Wind Power, log) × b(Distance to Great Lake) × longitude	-0.0134	(-0.0273, -0.00156)
b(Average Wind Power, log) × b(Distance to Great Lake) × Precinct 14	0.0181	(0.00643, 0.0313)
b(Average Wind Power, log) × b(Longitude)	0.0227	(0.0146, 0.0303)

Table 3: **Selected stage 1 coefficients.**

Bases	Coefficient	Bootstrap CI
b(University Degree (%)) × Proposed Turbine (Treatment)	-0.00499	(-0.00231, -0.000048)
b(Proposed Turbine (Treatment))	0.00605	(0.00084, 0.00442)
b(Proposed Turbine (Treatment))	0.0145	(0.0101, 0.0152)
b(University Degree (%)) × Proposed Turbine (Treatment)	-0.00499	(-0.00231, -0.000048)
Proposed Turbine (Treatment) × Precinct 41	-0.00374	(-0.00529, -0.00169)
Proposed Turbine (Treatment) × Precinct 29	-0.0023	(-0.0000299, 0.0000394)
Proposed Turbine (Treatment)	-0.00208	(-0.00573, -0.00101)
Proposed Turbine (Treatment) × Precinct 85	0.00252	(0.000324, 0.003)
b(Proposed Turbine (Treatment)) × Precinct 29	0.00272	(0.00000144, 0.00133)
b(Proposed Turbine (Treatment)) × Precinct 36	0.00605	(0.00084, 0.00442)

Table 4: **Selected stage 2 coefficients.**

Finally, Table 5 presents the Two-Stage Least Squares model discussed in the text that uses the treatment/education interaction.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.567	16.528	0.216	0.829
Proposed Turbine (Treatment)	-0.069	0.026	-2.695	0.007
Proposed Turbine (Treatment) × University Degree (%)	-0.259	0.122	-2.127	0.034

Table 5: **Two-Stage Least Squares Results With Uncovered Education Interaction.** The result of fitting a TSLS model using the specification from Stokes (2015) but including the treatment/education interaction uncovered by MDE. The estimated treatment effect on the turbine is comparable to that from the original specification (-0.077, see 1). The interaction term uncovered by MDE (Proposed Turbine (Treatment) × University Degree (%)) is statistically significant in this sample.

References

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

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

Athey, Susan, Julie Tibshirani and Stefan Wager. 2016. "Solving Heterogeneous Estimating Equations with Gradient Forests." *arXiv preprint arXiv:1610.01271* .

Belloni, Alexandre, Daniel Chen, Victor Chernozhukov and Christian Hansen. 2012. "Sparse models and methods for optimal instruments with an application to eminent domain." *Econometrica* 80(6):2369–2429.

Belloni, Alexandre and Victor Chernozhukov. 2013. "Least squares after model selection in high-dimensional sparse models." *Bernoulli* 19(2):521–547.

Belloni, Alexandre, Victor Chernozhukov, Denis Chetverikov and Kengo Kato. 2015. "Some new asymptotic theory for least squares series: Pointwise and uniform results." *Journal of Econometrics* 186(2):345–366.

Bhattacharya, Anirban, Debdeep Pati, Natesh S Pillai and David B Dunson. 2015. "Dirichlet–Laplace priors for optimal shrinkage." *Journal of the American Statistical Association* 110(512):1479–1490.

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

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

Chernozhukov, Victor, Christian Hansen and Martin Spindler. 2015. "Instrumental Variables Estimation with Very Many Instruments and Controls.".

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.

Fan, Jianqing, Yunbei Ma and Wei Dai. 2014. "Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models." *Journal of the American Statistical Association* 109(507):1270–1284.

Ghosal, Subhashish and Aad van der Vaart. 2017. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press.

Ghosal, Subhashish, Jayanta Ghosh and Aad W. van der Vaart. 2000. "Convergence Rates of Posterior Distributions." *Annals of Statistics* 28(2):500–531.

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

Gu, Chong. 2002. *Smoothing spline ANOVA models*. Springer series in statistics Springer.

Haavelmo, Trygve. 1943. "The statistical implications of a system of simultaneous equations." *Econometrica* 11:1–12.

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.

Hansen, Bruce E. 2017. "Econometrics." Unpublished manuscript.

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.

Hartford, Jason, Greg Lewis, Kevin Leyton-Brown and Matt Taddy. 2016. "Counterfactual Prediction with Deep Instrumental Variables Networks." *https://arxiv.org/abs/1612.09596* .

Heckman, James and Edward Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica* 73(3):669–738.

Hill, Jennifer. 2011. "Bayesian Nonparametric Modeling for Causal Inference." *Journal of Computational and Graphical Statistics* 20(1):217–240.

Hirano, Keisuke and Guido Imbens. 2005. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. John Wiley and Sons, Ltd, Chichester, UK chapter The Propensity Score with Continuous Treatments.

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.

Horowitz, Joel. 2014. "Ill-posed inverse problems in economics." *Annual Review of Economics* 6.

Imai, Kosuke and David A Van Dyk. 2012. "Causal inference with general treatment regimes." *Journal of the American Statistical Association* .

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

Kang, Joseph DY and Joseph L Schafer. 2007. "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data." *Statistical science* pp. 523–539.

Kennedy, Edward H., Scott A. Lorch and Dylan S. Small. 2017. "Robust causal inference with continuous instruments using the local instrumental variable curve." *https://arxiv.org/abs/1607.02566* .

Mohanty, Pete and Robert Shaffer. 2017. *bigKRLS: Optimized Kernel Regularized Least Squares*. R package version 2.0.4.
URL: *https://CRAN.R-project.org/package=bigKRLS*

Newey, Whitney. 1994. "Kernel Estimation of Partial Means and a General Variance Estimator." *Econometric Theory* 10(2):233–253.

Newey, Whitney K. 2013. "Nonparametric instrumental variables estimation." *The American Economic Review* 103(3):550–556.

O'Sullivan, Finbarr. 1986. "A Statistical Perspective on Ill-Posed Inverse Problems." *Statistical Science* 1(4).

Park, Trevor and George Casella. 2008. "The bayesian lasso." *Journal of the American Statistical Association* 103(482):681–686.

Pearl, Judea. 2014. "Trygve Haavelmo and the Emergence of Causal Calculus." *Econometric Theory* .

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.

Polson, Nicholas and James Scott. 2012. "Local shrinkage rules, Levy processes and regularized regression." *Journal of the Royal Statistical Society, Series B* 74(2):287–311.

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

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

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

Robins, James M and Ya'acov Ritov. 1997. "Toward a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-Parametric Models." *Statistics in Medicine* 16(3):285–319.

Rockova, Veronica and Edward George. Forthcoming. "The Spike-and-Slab LASSO." *Journal of American Statistical Association* .

Rosipal, Roman and Leonard J Trejo. 2001. "Kernel partial least squares regression in reproducing kernel hilbert space." *Journal of machine learning research* 2(Dec):97–123.

Samii, Cyrus, Laura Paler and Sarah Zukerman Daly. 2016. "Retrospective Causal Inference with Machine Learning Ensembles: An Application to Anti-recidivism Policies in Colombia." *Political Analysis* 24(4):434–456.

Stein, Charles. 1981. "Estimation of the mean of a multivariate normal distribution." *Annals of Statistics* 9:1135–51.

Stokes, Leah C. 2015. "Electoral Backlash against Climate Policy: A Natural Experiment on Retrospective Voting and Local Resistance to Public Policy." *American Journal of Political Science* 60(4):958–974.

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

van der Laan, Mark J. and Sherri Rose. 2011. *Targeted Learning Causal Inference for Observational and Experimental Data.* Springer.

VanderWeele, Tyler. 2015. *Explanation in causal inference: methods for mediation and interaction.* Oxford University Press.

Zou, Hui. 2006. "The Adaptive Lasso and Its Oracle Properties." *Journal of the American Statistical Association* 101(476):1418–1429.