Causal Inference through the Method of Direct Estimation*

Marc Ratkovic† Dustin Tingley‡

May 23, 2017

Abstract

We propose a method for estimating the causal effect of a treatment, binary or continuous, on an outcome in the presence of a high-dimensional vector of confounders. Rather than estimate an intermediate quantity, like a propensity score or selection model, we fit a tensor-spline regression model for the outcome and then directly estimate counterfactual quantities of interest. Estimation combines a marginal correlation screen and a new sparse Bayesian model that achieves optimal predictive bounds on the outcome and treatment effect. The method extends to instrumental variable, mediation, and sequential-g estimation. A simulation study and two applied examples illustrate the methods.

Key Words: causal inference, controlled direct effects, instrumental variables, mediation, sparse Bayesian modeling, sure independence screening

*We would like to thank Horacio Larreguy for providing replication data. We would also like to thank Peter Aronow, Scott de Marchi, James Fowler, Andrew Gelman, 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, 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 Not for citation or distribution without permission from the authors.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

A central problem for estimating causal effects is adjusting for a potentially high-dimensional vector of confounders. Classic methods deal with the problem in one of two ways. The first reduces the dimensionality of the problem through conditioning on a sufficient statistic of treatment assignment (the "propensity score" (Rosenbaum and Rubin, 1983); for extensions see Imai and Van Dyk (2012); Robins, Hernan and Brumback (2000)). Results can be sensitive, though, to slight or imperceptible misspecification of the treatment model (e.g. Kang and Schafer, 2007). The second approach estimates a low-dimensional structural parameter (e.g. Robins, 1986, 1989). This comes at the cost of averaging over potentially interesting heterogeneities.

We propose a method, the Method of Direct Estimation (MDE), that implements a high-dimensional nonparametric regression model in order to predict observation-level counterfactual quantities. The method sidesteps the above dimension reduction steps and returns both a stable and reliable causal estimate. Through casting causal effect estimation as a prediction problem, the method unifies estimation across a range of treatment regimes and structural models. MDE applies equally to a continuous, binary, or categorical treatment regime, and we extend it to both instrumental variable and mediation models. Heterogeneous effects can be characterized through examining the observation-level estimates.

We are the most recent in a long series of works using cutting edge methods to predict counterfactual outcomes; for precursors, see (see, e.g., Hill, 2011; Hansen, 2008; Robins, 1989; Rubin, 1984; Cochran, 1969; Belson, 1956; Peters, 1941). We extend these works from binary point-treatments to continuous treatments and structural models, working within a nonparametric estimation framework with provably minmax optimal prediction properties. We also contribute to the recent literature on high-dimensional causal inference (see, e.g. Belloni et al., 2017, 2015, 2012; Athey, Tibshirani and Wager, 2016; Hartford et al., 2016), bringing flexible models to bear in causal effect estimation. Our primary contribution is providing a reliable means of estimating fine-grained (observation level) causal effects, as well as extending the method to general treatment regimes and structural models. Rather than using a LASSO screen followed by a regularized least-squares estimate (e.g., Belloni and Chernozhukov, 2013), we implement a one-stage Bayesian estimator that endogenously shrinks coefficients on irrelevant covariates while returning asymptotically unbiased estimates on maintained coefficients. The model is a Bayesian version of the adaptive LASSO of Zou (2006),

but the mean parameter weights are estimated endogenously. Unlike recent sparse Bayesian models with nonseparable, nonconvex priors (Bhattacharya et al., 2015; Rockova and George, Forthcoming; Carvalho, Polson and Scott, 2010; Polson, Scott and Windle, 2014), our key tuning parameter is constructed so that our estimates achieve the minmax optimal prediction rate. In simulations, we show that the improvements in prediction error are sizable.

The proposed estimation procedure consists of four steps. First, we generate a large number of nonparametric tensor-spline bases, on the order of millions, including all *treatment* $\times$ *covariate* $\times$ *covariate* interactions and lower-order terms. These bases constitute a larger space than those considered by earlier multivariate nonparametric regression models that inspire our work (Stone et al., 1997; Friedman, 1991). We include such a dense parameterization so as to directly model the confounding between treatment and covariate. We utilize nonparametric bases as we do not know the functional form this interaction may take, and the nature of the interaction may vary across the covariate space. Second, we use a marginal correlation screen (Fan and Lv, 2008) to reduce the number of bases from millions to hundreds or thousands. Third, we regress the outcome onto the maintained bases using a sparse Bayesian regression. Fourth, we use the estimated model to predict the counterfactual quantities of interest. Pre-processing covariates by generating splines (or other) features is an important step. As show below, variable selection methods alone (such as those discussed above) are not sufficient to reduce specification sensitivity.

We then show how the MDE extends to several structural models and treatment types. In the case of a binary treatment, the proposed method outperforms existing cutting-edge machine learning methods on a benchmark dataset (LaLonde, 1986). In the instrumental variable setting, we show that the method returns an observation level two stage least squares estimate. That is, our approach to the instrumental variables setting estimates two covariances at the observation level: that between the outcome and the instrument and that between the treatment and the instrument, each conditioning on observed covariates. Standard instrumental variables estimates are the average of these covariances, masking heterogeneity in both encouragement and effect. Our approach allows us to relax the standard identification assumptions in instrumental variables from global to local, i.e., monotonicity in the instrument over the entire sample to monotonicity within each strata given by the covariates. Furthermore, we use the oracle bound constructively, rather than simply descriptively, in order to distinguish observations encouraged by the instrument from those not.

This distinction is important as it is only off the encouraged group that we can identify a causal effect. This threshold improves on existing high-dimensional instrumental variables methods which do not estimate the encouraged subsample (e.g. Athey, Tibshirani and Wager, 2016; Hartford et al., 2016; Newey, 2013; Belloni et al., 2012). In a mediation framework, MDE allows for treatment effect heterogeneity in the direct effect, mediated effect, or both. Analytically, MDE maintains a key identity in mediation analysis: that the mediation effect is the product of the treatment effect on mediator and mediator direct effect on outcome, a relationship otherwise lost using existing estimation strategies with nonlinear models. Lastly, we show that MDE extends naturally to the estimation of controlled direct effects via sequential g-estimation (Vansteelandt, 2009; Acharya, Blackwell and Sen, 2016).

Our larger goal is aligning machine learning, particularly high dimensional nonparametric regression, with causal inference, tightly and at a conceptual level. For example, we show that the ignorability assumption common to causal inference actually implies a local linear function for the outcome, guiding our choice of bases. Similarly, we show that our proposed regression method satisfies an oracle inequality not only on the fitted values but also on the instantaneous causal effect. This gives us an optimality condition in predicting the causal effect, which is close but not exactly the same as predicting the outcome well.[1] While the oracle bound has been typically used as a descriptive tool (giving rates of risk convergence) instead we use it constructively by using a feasible version of the bound to estimate which observations are likely not impacted by an instrument in an encouragement design.

The paper proceeds as follows. Section 2 introduces the proposed framework and estimation strategy. Section 3 shows how the method of direct estimation readily extends a binary treatment framework as well as structural models like instrumental variables, mediation, and sequential g-estimation. Section 4 discusses the relationship between the proposed method and earlier work. We include several applications in Section 5. Section 6 concludes by showing how our approach connects disparate literatures and by discussing limitations and future extensions.

[1] As one example, ensemble tree methods will fit the systematic component well (e.g. Hill, Weiss and Zhai, 2011), but since the fits are piecewise constant, they are not well-suited to estimating an instantaneous causal effect.

2 The Proposed Method

We first describe the setup for the proposed method. We next describe each step of the estimation strategy: generating bases, screening, our sparse regression model, and predicting counterfactuals. We then present our oracle bounds on both the fitted values and the instantaneous causal effect.

2.1 Setup

Assume a sample where observation $i \in \{1, 2, \ldots, n\}$ possesses a potential outcome function, $Y_i(T_i)$, which maps treatment level T_i to outcome $Y_i(T_i)$. The treatment T_i is a realization of $\widetilde{T}_i \sim F_{X_i}$, such that $\widetilde{T}_i$ has continuous support and its distribution may vary with X_i, a vector of p observation-level fixed covariates. There may be more covariates than observations, so p may be greater than 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 observed data are $\left\{Y_i, T_i, X_i^\top\right\}^\top$ where $Y_i = Y_i\left(T_i\right)$

We will denote the causal effect of moving from one value of the treatment, T_i', to another, T_i'', as

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

To identify our causal effects, we make three standard assumptions. First, we assume the potential outcomes are ignorable given X_i: $Y_i(\widetilde{T}_i) \perp\!\!\!\perp \widetilde{T}_i | X_i, \widetilde{T}_i \in \{T_i'', T_i'\}$. Second, we assume the counterfactual is realizable with positive probability, $\Pr(\widetilde{T}_i = T_i | X_i, T_i \in \{T_i'', T_i'\}) > 0$. Third, we make stable unit and treatment value assumptions: potential outcomes do not depend on treatment or outcomes of others and there is a single version of each treatment level.

The quantity $\Delta_i(T_i'', T_i')$ poses two challenges. First, as we can observe at most one potential outcome per observation, at least one of the potential outcomes remain hidden. This requires us to instead condition on the 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. A second concern is whether T_i' and T_i'' might occur with positive probability. To reduce concerns over common support, we focus on a counterfactual estimate close to the observed data, $\lim_{\delta \to 0} \nabla_{T_i}(\delta_i)$, where

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

For identification, we assume in this case that ignorability holds in a continuous, open interval around the observed value T_i, i.e. $Y_i(\widetilde{T}_i) \perp\!\!\!\perp \widetilde{T}_i | X_i, \widetilde{T}_i \in (T_i - \delta_i^S, T_i + \delta_i^S)$ for some $\delta_i^S > 0$ and that $\mathbb{E}(Y_i(T_i)|X_i)$ is differentiable in its manipulable argument at the observed value. This estimand captures the impact of a ceteris paribus perturbation of the treatment on the outcome.

For example, researchers are used to fitting a model of the form

$$Y_i = \mu_i + \beta T_i + X_i^\top \gamma + \epsilon_i \tag{3}$$

where β, a global slope term, is interpreted as the marginal effect of T_i on Y_i. We conceive of the outcome in the following form:

$$Y_i = \mu_i + \beta_i T_i + g(X_i) + \epsilon_i; \tag{4}$$

$$\beta_i = f(X_i) = \lim_{\delta \to 0} \nabla_{T_i}(\delta_i) \tag{5}$$

and $f(X_i), g(X_i)$ nonparametric functions of the covariates. That the local effect is only a function of X_i follows directly from our ignorability assumption:

$$Y_i(\widetilde{T}_i) \perp\!\!\!\perp \widetilde{T}_i \big| X_i, \widetilde{T}_i \in (T_i - \delta_i^S, T_i + \delta_i^S) \Rightarrow \tag{6}$$

$$Y_i(T_i + \delta_i) - Y_i(T_i) \perp\!\!\!\perp \delta_i \big| X_i, |\delta_i| < \delta_i^S \tag{7}$$

$$\Rightarrow \lim_{\delta_i \to 0} \mathbb{E}\left\{ \frac{Y_i(T_i + \delta_i) - Y_i(T_i)}{\delta_i} \middle| X_i \right\} = f(X_i) \tag{8}$$

$$= \lim_{\delta_i \to 0} \nabla_{T_i}(\delta_i) \tag{9}$$

where $f(X_i)$ is not a function of T_i. Taking an antiderivative gives the model in Equations 4-5.

We focus on this instantaneous effect under a continuous treatment through the first part of the paper. We will estimate the model Equations 4-5 using a high-dimensional, nonparametric regression. The regression specification allows for the instantaneous effect $\beta_i = \lim_{\delta_i \to 0} \nabla_{T_i}(\delta_i)$ to vary with covariates, as well as including an additional function ($g(X_i)$) that is not a function of the treatment.

Conditioning on the covariates X_i in such a rich, nonparametric regression model achieves two goals: adjusting for confounding bias and characterizing effect heterogeneity. We achieve these by looking at estimated effects within the strata and across the strata defined by the covariates X_i, respectively. In order to reduce confounding bias, we estimate the impact of the treatment within the stratum X_i, i.e., subclassifying on the covariates. Subclassification is a long-established means

of reducing confounding bias (Cochran, 1968; Cochran and Rubin, 1973), though subclassification is normally done on a coarsened sufficient statistic of the treatment (e.g. Rosenbaum and Rubin, 1984; Imai and Van Dyk, 2012). After estimating a treatment effect conditional on X_i, we are able to look across covariate profiles as a means of assessing and characterizing treatment effect heterogeneity. Of course, if an average effect is desired for the sample or some subset, this estimate can just be found by taking the average over the desired subsample.

Modeling this local confounding returns an instantaneous treatment effect estimate–a conditional effect for that observation, or, equivalently, the impact of a tiny "do" operation or the partial derivative of the response function with respect to the treatment.[2] Our emphasis is on implementing a sufficiently rich regression model that uses covariates to consistently and reliably estimate the treatment effect for each observation under our identification assumptions.

Once we have a means of estimating this instantaneous effect, we illustrate several useful ways in which we can extend it to other causal settings. In an instrumental variable setting, we can model first-stage heterogeneity in the encouragement, allowing us to relax the no defiers assumption. In a mediation setting, we show that we can maintain a key identity–that the mediation effect is the product of the treatment effect on mediator and mediator direct effect on outcome–that is otherwise lost in nonlinear models. We also show that the method can estimate controlled direct effects in a single step, but can also extend to a broad class of structural mean models through the use of sequential g-estimation.

We then move past working with partial derivatives to estimating and averaging over the support of $\widetilde{T}_i$. We use a bootstrap-estimated interval to characterize a range over which the treatment effect can be estimated. We also use the bootstrap to move from estimating the effect of the treatment at the observed value, $\lim_{\delta_i \to 0} \nabla_{T_i}(\delta_i)$ to the expected effect of T_i, $\mathbb{E}_{\widetilde{T}_i}\left(\lim_{\delta_i \to 0} \nabla_{T_i}(\delta_i)\right)$. Before exploring these extensions we turn to our estimation procedure.

2.2 Estimation

Our identification assumptions require that we condition on the pre-treatment covariates, but there is no theoretical guidance as to how and in what way to incorporate the covariates. A causal effect estimate is *model dependent* to the extent that choices over how to incorporate covariates into the model, say with interaction terms of higher-order terms, affects the estimated impact of

[2] In Pearl's "do" notation, $\beta_i = \nabla_{T_i}(\delta_i) = \frac{1}{\delta_i}\left\{\mathbb{E}(Y_i|do(T_i + \delta_i), X_i) - \mathbb{E}(Y_i|do(T_i), X_i)\right\}$.

a treatment variable, and nonparametric pre-processing is a now-accepted from of removing this model dependence (Ho et al., 2007). Variable selection methods do not eliminate concerns over what variables to include, as the decision to include only linear terms or include quadratic terms can have a substantive impact on the inferences drawn on the treatment effect–even when using variable selection methods, a point which we illustrate in one of our applied examples below.

In order to reduce concerns over model dependent inference, we fit a model using a high-dimensional, tensor-spline nonparametric regression. Estimation proceeds in four steps. First, we generate a large (on the order of millions) set of nonparametric bases. Second, we use a marginal correlation screen to select a subset of bases, reducing their number from millions to hundreds or thousands. Third, we use a sparse Bayesian regression to fit the model. Fourth, we use the estimated model to predict the counterfactual quantities of interest.

2.2.1 The Assumed Model.

Our ignorability assumption guarantees that conditioning on covariates will allow for estimation of the instantaneous causal effect. The assumption does not guide us in *how* these covariates should enter the outcome model. For generality, we assume the data are generated as

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

with $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, where $R^o(T_i, X_i)$ is a possibly infinite set of basis functions scaled to be mean zero and unit variance. We divide these into three groups: those that survive our covariate screen (described below), $R(T_i, X_i)$; those that we consider in the screen, $R_\infty(T_i, X_i)$; and those that are in the data-generation process but not in our considered model space, $R_\perp(T_i, X_i)$.

We construct our covariate bases so as to satisfy the conditions required for the screening procedure, Conditions C and 1–4 in Fan and Lv (2008). 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 Ratkovic and Tingley, 2017) (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.

2.2.2 Generating Bases

In constructing the set of bases to model the conditional mean, $R_\infty(T_i, X_i)$, we turn to a tensor-product of spline bases. 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

$$R_\infty(T_i, X_i) = \{1, b_T(T_i, \kappa_T)\}_{\kappa_T \in \mathcal{K}_T} \otimes \{1, b(X_{ik}, \kappa_k)\}_{\kappa_k \in \mathcal{K}_k} \otimes \{1, b(X_{ik'}, \kappa_{k'})\}_{\kappa_{k'} \in \mathcal{K}_{k'}} \tag{11}$$

where $\otimes$ denotes the tensor product. In effect, we construct a basis for the mean 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. 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

In terms of Equations 4-5 above, the bases involving the treatment are used to estimate the instantaneous effect, β_i. The bases that do not involve the treatment are used to estimate the purely prognostic component, $g(X_i)$. Our ignorability assumption carries implications for bases selection for the treatment variable. The ignorability assumption implies that the outcome is locally linear at each T_i with instantaneous effect given by an arbitrary function of X_i. We therefore model the treatment using a degree 2 thresholded power-basis and a degree 2 B-spline. The first is a set of hinge functions radiating off knots, the second is an upside-down V between two knots, and zero elsewhere. Both sets of bases satisfy local linearity, unlike bases with more than one nonzero derivative in T_i (e.g. gaussian radial basis functions, higher degree B-splines).

For each covariate, we use a B-spline basis of varying knots and degree (Eilers and Marx, 1996; de Boor, 1978). The B-spline basis offers a bounded basis for nonparametric modeling, with well-studied optimality properties (Gyorfi et al., 2002). The B-spline basis requires selecting both degree and knot locations, and as we do not know the best choice, we err in favor of selecting too many basis functions. We have two different sets of choices to make: how many bases to generate and how may to maintain after our screen. These choices require balancing theoretical and computational concerns. First, we want the number of bases in $\{1, b(X_{ik}, \kappa_k)\}_{\kappa_k \in \mathcal{K}_k}$ 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.[3] Second, we want the bases to be sufficiently rich to capture a large class of models. Third, we want the number of maintained bases to grow at the nonparametric rate of $N^{1/5}$, but also be sizable with a small sample size. Fourth, we want this to lead to a computationally feasible algorithm.

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, described below. At these default values, computation times are certainly practical; for example, on one of our simulations with a sample size of $n = 1,000$ and $p = 100$ covariates, constructing the bases and recovering point estimates takes between a minute and a minute and a half on a 2.6 GHz Intel Core i7 running **R**.

2.2.3 Screening Bases

We combine the treatment and covariate bases using the tensor product as given above. At our default, we take 24 bases for the treatment and for each covariate. This gives us $(1+24) \times (1+24 \times p) \times (1+24 \times p)$ 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 so we implement a Sure Independence Screen (SIS) (Fan and Lv, 2008).[4] The SIS occurs in two steps. First, we construct all of the tensor product bases in Equation 11. Second, we sort the bases by their absolute correlation with the outcome. We want to select a sufficiently large and rich dictionary of nonparametric bases to approximate a wide set of models; in practice and all of the analyses below, we select $100 \times (1 + n^{1/5})$ bases that have the largest unconditional correlation

[3] 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.

[4] We implement a SIS rather than the grouped SIS strategies of (Fan, Feng and Song, 2012; 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.

with the outcome.[5]

2.2.4 A Sparse Bayesian Model

Even after screening, we are left with hundreds of possible nonparametric bases. In order to enforce some form of regularization, we implement a sparse Bayesian prior. We estimate our treatment effects by extending LASSOplus, a sparse Bayesian prior we implemented in earlier work (Ratkovic and Tingley, 2017), to the nonparametric prediction problem.

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

$$Y_i|R(T_i,X_i),c,\sigma^2 \sim \mathcal{N}(R(T_i,X_i)^\top c,\sigma^2) \tag{12}$$

$$c_k|\lambda,w_k,\sigma \sim DE\left(\lambda w_k/\sigma\right) \tag{13}$$

$$\lambda^2|n,p,\rho \sim \Gamma\left(n\times(\{log(n)+2\log(p))-p,\rho\right) \tag{14}$$

$$w_k|\gamma \sim \text{generalizedGamma}(1,1,\gamma) \tag{15}$$

$$\gamma \sim \exp(1) \tag{16}$$

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.

We summarize several properties of our model and present formal derivations in Appendix A. 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. Second, the

[5]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. Construction of the tensor basis is done in C++ via `Rcpp`.

rate at which the oracle bound holds is controlled by $1 - exp\{-\sqrt{\log(n)}\}$, which approaches 1 in the limit.[6]

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 Appendix A.

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 Appendix A.. 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*.

2.2.5 Counterfactual Prediction

Given fitted values $\widehat{c}$, we can write

$$\widehat{Y}_i = \widehat{\mu}_Y + R(T_i, X_i)^\top \widehat{c} \tag{17}$$

with the intercept chosen as $\widehat{\mu}_Y = \frac{1}{n}\sum_{i=1}^n \left(Y_i - R(T_i, X_i)^\top \widehat{c}\right)$. We then estimate

$$\widehat{\nabla}_{T_i}(\delta_i) = \frac{1}{\delta_i}\left\{\left(R(T_i + \delta_i, X_i) - R(T_i, X_i)\right)^\top \widehat{c}\right\} \tag{18}$$

and we approximate the derivative by choosing δ_i close to zero; we take $\delta_i = 10^{-5}$ in practice.[7] The sample average marginal causal effect can be found through averaging over the sample,

[6]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.

[7]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

$\frac{1}{n}\sum_{i=1}^{n}\widehat{\nabla}_{T_i}(\delta_i)$.

2.3 Theoretical Results

We provide two bounds on the risk of our estimator. First, we provide a bound on the predicted values. Second, we provide a bound on the effect estimate. The first follows almost directly from our prior structure, which guarantees that the key tuning parameter grows at the Oracle rate. The second bound shows that the model will do well in terms of predicting the instantaneous treatment effect as well.

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

$$\begin{aligned}\frac{1}{n}\left\{||R(T.,X.)\widehat{\delta}||_2^2 + \lambda||\sum_{k=1}^{p}\widehat{w}_k\widehat{\delta}_k||_1^1\right\} \leq \\ C\frac{\widehat{\lambda}^2\widehat{\sigma}^2\overline{\gamma}^2|S|}{n^2\phi_0^2} + C_\infty\frac{||R^o_{\infty/n}(T.,X.)c_{\infty/n}||_2^2}{n} + C_\perp\frac{||R^o_\perp(T.,X.)c_\perp||_2^2}{n}\end{aligned} \tag{19}$$

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. For details, see Ratkovic and Tingley (2017). 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 and Fan and Lv (2008). 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 $\nabla_{T_i}(\delta_i)$ We are interested in not only bounding the prediction risk, as above, but also the error on the $\nabla_{T_i}(\delta_i)$. 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

found any value of δ below $\sqrt{\text{Var}(T_i)}/1000$ works well.

basis at least one element covaries with the treatment:

$$R(T.,X.) = [R^X(X.) : R^{TX}(T.,X.)]. \tag{20}$$

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 $\widehat{\nabla}_{T_i}(\delta_i) = \Delta R_i(T_i,X_i)^\top \widehat{c}^{TX} = \Delta R_i^{TX}(T_i,X_i)^\top \widehat{c}^{TX}$, 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.

We show in Appendix G that the mean parameter estimates $\widehat{c}^{\widehat{S},TX}$ are actually a solution to the following LASSO problem:

$$\widehat{c}^{S,TX} = \underset{c^{\widehat{S},TX}}{\operatorname{argmin}} \left|\left|\widetilde{\Delta Y} - \Delta R^{\widehat{S},TX} c^{\widehat{S},TX}\right|\right|_2^2 + \left|\left|\widetilde{W}^{\widehat{S},TX} c^{\widehat{S},TX}\right|\right| \tag{21}$$

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

$$\widetilde{\Delta Y_i} = \Delta R_i(T_i,X_i)^\top \widehat{c} + \widehat{\epsilon}_i. \tag{22}$$

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.

As the coefficients are simultaneously minimizing a LASSO problem on the first derivative, we can establish the oracle bound

$$\begin{aligned}&\frac{1}{n}\left\{||\Delta R^{\widehat{S},TX}(T.,X.)(\widehat{c}^{\widehat{S},TX} - c^{\widehat{S},TX})||_2^2 + ||\widetilde{W}^{\widehat{S},TX}(\widehat{c}_k^{\widehat{S},TX} - c_k^{\widehat{S},TX})||_1^1\right\} \leq \\ &C_\Delta \frac{\widehat{\sigma}^2\overline{\gamma}_\Delta^2|S_\Delta|}{N^2\phi_\Delta^2} + C_\infty \frac{||\Delta R^o_{\infty/\widehat{S}}(T_i,X_i)c_{\infty/n}||_2^2}{n} + C_\perp \frac{||\Delta R^o_\perp(T_i,X_i)c_\perp||_2^2}{n}\end{aligned} \tag{23}$$

where all the constants in the inequality are analogous to those in Inequality 23 but defined in terms of the design matrix that parameterize the derivative.[8]

[8]We make the same assumptions on the model in Equation 22 as we do on the outcome model.

We last note that the model is similar to the relaxed LASSO of Meinshausen (2007), differing in that rather than re-fitting the outcome to first-stage LASSO selected variables, we fit them to the pseudo-observation above. In this formulation, we are conditioning on the fitted residuals, which suggests the utility of bagging to improve estimation (a point we discuss in Appendix C).

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.

2.3.1 Simulation Results

Our estimation contributions are twofold: construction of the tensor-spline nonparametric basis and the sparse Bayesian model for estimation. We conducted a simulation study allowing us to explore both contributions and the performance of MDE, varying sample size and the complexity of the underlying true model. We report these results in Appendix E, and we find MDE performs quite well in terms of RMSE and bias. To isolate the impact of our Bayesian prior, we compare MDE to sparse regression models using our bases, namely the cross-validated LASSO and two sparse Bayesian priors: the horseshoe and Bayesian Bridge (Carvalho, Polson and Scott, 2010; Polson, Scott and Windle, 2014). We also compare MDE to methods that create their own basis functions: kernel regularized least squares (KRLS, Hainmueller and Hazlett, 2013), POLYMARS (Stone et al., 1997), and Sparse Additive Models (SAM, Ravikumar et al., 2009); as well as 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.). Across settings, MDE is either the best performer or in the top set. Due to space considerations we refer the reader to the appendix for a detailed discussion.

3 Extensions

We next consider several extensions of the proposed method. We first extend our local effect estimand to two common structural models: instrumental variable (IV) analyses and mediation analysis. We show that MDE provides new insights into the problems. In the IV setting, MDE lets

us relax the monotonicity assumption, that the instrument uniformly encourage or discourage the treatment across the sample. In the mediation setting, we show that MDE allows us to maintain a useful identity, that the total effect is additive in the direct and mediation effects, even with nonlinear models. We also show how MDE can estimate controlled direct effects via sequential-g estimation. Last, we show that MDE easily accommodates binary treatments, the most studied case in the literature.

3.1 Instrumental Variables

A treatment and outcome may be mutually determined, and thereby a correlative or regression analysis may not recover the causal effect of the treatment. In these cases, an instrument, or "encouragement," can recover a causal effect. The instrument is assumed to have a direct effect on the treatment but no direct effect on the treatment, thereby providing an experimental handle in the study.

Two problems emerge: the instrument may only affect some observations, the compliers, and not others; and the instrument may have a positive impact on some observations and negative on others. These issues pose problems to both internal and external validity. It is unclear which observations are actually impacted by the instrument and hence driving the causal effect estimate. The second concern, that there are "no-defiers," is assumed away in the binary treatment/instrument case, while this assumption is embedded in a linear first-stage specification.

To formalize, we assume the treatment is not conditionally independent of the outcome given the covariates, $Y_i(\widetilde{T}_i) \not\perp\!\!\!\perp \widetilde{T}_i | X_i$, thereby biasing the estimate of the causal effect. We assume the existence of a pre-treatment instrument $\widetilde{Z}_i$ that helps resolve the issue. The instrument, which follows law $F^Z_{X_i}$ and has observed value Z_i, enters the potential outcome function for the treatment as $\widetilde{T}_i = T_i(\widetilde{Z}_i)$. For identification, we make the exclusion restriction that Z_i has no direct effect on the outcome, $Y_i(T_i(\widetilde{Z}_i), \widetilde{Z}_i) = Y_i(T_i(\widetilde{Z}_i), Z_i')$ for all $Z_i' \in \text{supp}\,(\widetilde{Z}_i)$, or, 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 $[Y_i, T_i, Z_i, X_i^\top]^\top$.

In order to recover a consistent estimate of the causal effect of the treatment on the outcome, we trace the exogenous variation of the instrument through the treatment and onto the outcome. The *instantaneous encouragement* from perturbing the instrument on observation i is

$$\nabla^{IV}_{Z_i}(\delta) = \mathbb{E}(T_i(Z_i + \delta) - T_i(Z_i) | X_i), \tag{24}$$

where as above we have to condition on X_i to recover an estimate. The *instantaneous intent to treat on the treated*, or *IITT*, is the impact of this encouragement on the outcome,

$$\nabla_{T_i}^{IV}(\delta) = \mathbb{E}\left\{Y_i\left(T_i(Z_i) + \nabla_{Z_i}^{IV}(\delta)\right) - Y_i\left(T_i(Z_i)\right)|X_i\right\}. \tag{25}$$

The effect $\lim_{\delta\to 0}\nabla_{T_i}(\delta) = \mathbb{E}(dY_i(T_i)/dT_i|X_i)$ is not identified because of the failure of ignorability. Instead we target the causal quantity

$$\mathbb{E}\left(\frac{dY_i(T_i(Z),Z)}{dZ}\middle|X_i, Z = Z_i\right) = \tag{26}$$

$$\mathbb{E}\left(\frac{dY_i(T_i(Z))}{dZ}\middle|X_i, Z = Z_i\right) = \tag{27}$$

$$\mathbb{E}\left(\frac{\partial Y_i(T_i)}{\partial T_i}\frac{\partial T_i(Z)}{\partial Z}\middle|X_i, Z = Z_i\right) \tag{28}$$

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

$$\mathbb{E}\left(\frac{\partial Y_i(T_i)}{\partial T_i}\middle|X_i, Z = Z_i\right) = \text{plim}\,\frac{\mathbb{E}(dY_i(T_i(Z))/dZ|X_i, Z = Z_i)}{\mathbb{E}(\partial T_i(Z)/\partial Z\big|X_i, Z = Z_i)} \tag{29}$$

$$= \lim_{\delta\to 0}\frac{\nabla_{T_i}^{IV}(\delta)}{\nabla_{Z_i}^{IV}(\delta)}, \tag{30}$$

which we refer to as the Local Instantaneous Causal Effect (LICE) and denote $\lim_{\delta\to 0}\nabla_{T_i,Z_i}^{IV}(\delta)$. Like with TSLS, the plug-in estimate is a biased but consistent estimate of the desired causal effect.

The LICE differs in an important way from the standard Wald or two-stage least squares estimates (TSLS). The TSLS estimate under a linear structural equation model, which reduces to the Wald estimate in the binary treatment/encouragement setting, is calculated from sample covariances as

$$\widehat{\theta}^{TSLS} = \frac{\widehat{\text{Cov}}_{\mathcal{S}}(Y_i, Z_i|X_i)}{\widehat{\text{Cov}}_{\mathcal{S}}(T_i, Z_i|X_i)}, \tag{31}$$

a ratio of sample covariances, after X_i has been partialed out. 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. Compared to our estimate in Equation 30, we see that the TSLS estimate averages the numerator and denominator over compliers, defiers, and observations for which no causal effect is identified.

The estimated LICE is a TSLS estimate, but one that varies with X_i, thereby bringing the

underlying heterogeneity to the fore. To see this, consider

$$\lim_{\delta \to 0} \nabla^{IV}_{T_i, Z_i}(\delta) = \lim_{\delta \to 0} \frac{\nabla^{IV}_{T_i}(\delta)}{\nabla^{IV}_{Z_i}(\delta)} \tag{32}$$

$$= \frac{\mathbb{E}(dY_i(T_i(Z))/dZ | X_i, Z = Z_i)}{\mathbb{E}(\partial T_i(Z)/\partial Z \big| X_i, Z = Z_i)} \tag{33}$$

$$= \frac{\text{Cov}(Y_i, Z_i | X_i)/\text{Var}(Z_i | X_i)}{\text{Cov}(T_i, Z_i | X_i)/\text{Var}(Z_i | X_i)} \tag{34}$$

$$= \frac{\text{Cov}(Y_i, Z_i | X_i)}{\text{Cov}(T_i, Z_i | X_i)}, \tag{35}$$

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.

Since our estimand is conditioned by X_i, we do not need to assume homogeneity in the denominator (monotonicity) across values of X_i in order to estimate into which principal stratum each observation falls (as required by Abadie, 2003; Aronow and Carnegie, 2013). We will refer to observations with a positively, negatively, and zero impact of the instrument on the treatment as compliers, defiers, and non-compliers, respectively (Angrist, Imbens and Rubin, 1996). Were we to assume a binary treatment and instrument with underlying latent index models, we would recover the same model used in the Local IV (LIV) estimate in Heckman and Vytlacil (2007*a*,*b*, 2005, 1999). MDE is means of estimation, and as such, either the LIV or LICE can be estimated; we defer a discussion contrasting the two estimands to below.

For estimation, we fit two models: one for the treatment as a function of the instrument and a second for the outcome as a function of the treatment. We do so using the four-step procedure outlined above, and then use the plug-in estimates for equations 24 and 25 to estimate the LICE in equation 30.

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 23, 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

$$\widehat{df}_i = \frac{\partial \widehat{\mu}_i(Y_i)}{\partial Y_i} = \frac{\text{Cov}(Y_i, \widehat{\mu}_i(Y_i))}{\sigma^2} \tag{36}$$

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

$$\widehat{df}_i = \frac{\partial [R^X(X_i) : R^{TX}(T_i, X_i)]^\top \widehat{c}}{\partial Y_i} \tag{37}$$

$$= \frac{\partial [R^X(X_i)]^\top \widehat{c}^X}{\partial Y_i} + \frac{\partial [R^{TX}(T_i, X_i)]^\top \widehat{c}^{TX}}{\partial Y_i}. \tag{38}$$

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

$$\widehat{df}_i^{\nabla} = \frac{\partial [R^{TX}(T_i, X_i)]^\top \widehat{c}^{TX}}{\partial Y_i}. \tag{39}$$

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

The adaptive component of our threshold is of the form

$$\widehat{df}_i^{adapt} = \frac{1/n}{1/n + \widehat{df}_i^{\nabla}} \tag{40}$$

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

$$\widehat{\mathbf{1}}\left(\nabla_{Z_i}^{IV}(\delta_i^{IV}) \neq 0\right) = \mathbf{1}\left(\left|\widehat{\nabla}_{Z_i}^{IV}(\delta_i^{IV})\right| > C \frac{\widehat{\lambda}||\widehat{w}_k||_\infty \widehat{\sigma}}{n\widehat{\phi}_0} \widehat{df}_i^{adapt}\right) \tag{41}$$

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.

We include simulation evidence for the proposed threshold's efficacy and reliability in Appendix F.

3.2 Causal Mediation

Mediation analysis examines the pathway through which a treatment variable impacts some outcome through an intermediate variable, a mediator. Mediation analysis hence examines mechanisms rather than treatment effects; for a review, see VanderWeele (2015).

A key insight of mediation analysis is that the total treatment effect decomposes into the direct effect plus the mediated effect. In a linear SEM framework, the mediated effect is the product of the mediator direct effect and intermediate direct effect (MacKinnon et al., 2002), and it can be estimated as a product of coefficients. Imai, Keele and Tingley (2010) show that the product rule does not apply to estimated coefficients unless the underlying model is linear. We show that MDE allows us to maintain this key identity, as we are fitting a model that is a local linearization around the treatment and mediator at each point.

Consider first a post-treatment mediator equipped with its own potential outcome function, $\widetilde{M}_i = M_i(\widetilde{T}_i)$ with law $F^M_{T_i,X_i}$ and observed value $M_i = M_i(T_i)$. The observed outcome is now $Y_i = Y_i(M_i, T_i)$ and the observed data is $\left[Y_i, M_i, T_i, X_i^\top\right]^\top$. Our goal is to differentiate three effects: the total effect of the treatment on the outcome, the direct effect of the treatment on the outcome, and the mediated effect of the treatment through the mediator on the outcome. For identification, we will assume that the covariates are sufficiently rich to render the outcome, mediator, and treatment sequentially ignorable (Imai, Keele and Yamamoto, 2010): $Y_i(\widetilde{M}_i, \widetilde{T}_i) \perp\!\!\!\perp M_i(\widetilde{T}_i)|\widetilde{T}_i, X_i$ and $\left\{Y_i\left(\widetilde{M}_i, \widetilde{T}_i\right), M_i(\widetilde{T}_i)\right\} \perp\!\!\!\perp \widetilde{T}_i|X_i$. Second, we assume that all considered manipulations of mediator and treatment are on the support of $\widetilde{M}_i|X_i$ and $\widetilde{T}_i|X_i$.

We then estimate the observation-level total effect, direct effect, and mediated effect as

$$\text{Total effect: } \nabla_{T_i}^{TE}(\delta_i^{TE}) = \frac{1}{\delta_i^{TE}}\mathbb{E}\left\{Y_i\left(M_i(T_i+\delta_i^{TE}), T_i+\delta_i^{TE}\right) - Y_i\left(M_i, T_i\right)|X_i\right\} \tag{42}$$

$$\text{Direct effect: } \nabla_{T_i}^{DE}(\delta_i^{DE}) = \frac{1}{\delta_i^{DE}}\mathbb{E}\left\{Y_i\left(M_i, T_i+\delta_i^{DE}\right) - Y_i\left(M_i, T_i\right)|X_i\right\} \tag{43}$$

$$\text{Mediated effect: } \nabla_{T_i}^{ME}(\delta_i^{ME}) = \frac{1}{\delta_i^{ME}}\mathbb{E}\left\{Y_i\left(M_i(T_i+\delta_i^{ME}), T_i\right) - Y_i\left(M_i, T_i\right)|X_i\right\}. \tag{44}$$

We will also make use of

$$\text{Mediator direct effect: } \nabla_{M_i}(\delta_i^{M}) = \frac{1}{\delta_i^{M}}\mathbb{E}\left\{Y_i\left(M_i+\delta_i^{M}, T_i\right) - Y_i\left(M_i, T_i\right)|X_i\right\} \tag{45}$$

$$\text{First stage mediation effect: } \nabla_{T_i}(\delta_i^{T}) = \frac{1}{\delta_i^{T}}\mathbb{E}\left\{M_i\left(T_i+\delta_i^{T}\right) - M_i\left(T_i\right)|X_i\right\} \tag{46}$$

If we assume that the potential outcome functions $Y_i(\cdot)$ and $M_i(\cdot)$ are differentiable in their manipulable arguments, we see that the total effect decomposes into a sum of the direct and mediated effects. The law of the total derivative gives

$$\lim_{\delta_i^{TE}\to 0}\nabla_{T_i}^{TE}(\delta_i^{TE}) = \lim_{\delta_i^{DE}\to 0}\nabla_{T_i}^{DE}(\delta_i^{DE}) + \lim_{\delta_i^{M}\to 0}\nabla_{M_i}(\delta_i^{M}) \times \lim_{\delta_i^{T}\to 0}\nabla_{T_i}(\delta_i^{T}). \tag{47}$$

The second summand on the righthand side can be simplified by the chain rule as

$$\lim_{\delta_i^{TE}\to 0}\nabla_{T_i}^{TE}(\delta_i^{TE}) = \lim_{\delta_i^{DE}\to 0}\nabla_{T_i}^{DE}(\delta_i^{DE}) + \lim_{\delta_i^{ME}\to 0}\nabla_{T_i}^{ME}(\delta_i^{ME}). \tag{48}$$

This gives us two well-known results: the mediated effect is the product of the mediator direct effect and intermediate direct effect (MacKinnon et al., 2002) and that the total effect is additive in the direct and mediated effects.

In our framework we thus estimate two models, one for the outcome and one for the mediator. The models that we fit are

$$Y_i = \mu_Y + R_Y(M_i, T_i, X_i)^\top c_Y + \epsilon_i^Y \tag{49}$$

$$M_i = \mu_M + R_M(T_i, X_i)^\top c_M + \epsilon_i^M \tag{50}$$

where, again, the $R_\cdot(\cdot)$ vectors are a post-screened set of bases. We can now recover estimates of the effects in Equations 42–46.

Interestingly, while Imai, Keele and Tingley (2010) show that the product rule above does not apply to estimated coefficients unless the underlying model is linear, we obtain a different result. Since we are using a local linearization around each point as a function of the treatment and mediator, the product rule can be extended, a point earlier noted by Stolzenberg (1980).

3.3 Controlled Direct Effects and Sequential g-estimation

A natural extension of our method is to the case where controlled direct effects are of interest and there are intermediate confounders (as in Acharya, Blackwell and Sen, 2016; Vansteelandt, 2009). We focus on estimating a controlled direct effect, but show how MDE can then generalize to a far wider class of structural mean models.

The controlled direct effect is the effect of a shift in treatment holding the mediating variable fixed at some level m:

$$\text{Controlled direct effect: } \nabla_{T_i}^{CDE}(\delta_i^{CDE}) = \frac{1}{\delta_i^{CDE}}\mathbb{E}\left\{Y_i\left(M_i = m, T_i + \delta_i^{CDE}\right) - Y_i\left(M_i = m, T_i\right)|X_i\right\} \tag{51}$$

Identification assumptions are nearly identical to the mediation case above, except the CDE allows for post-treatment confounders in the condition set. Decompose $X_i = [X_i^{pre} : X_i^{post}]$ for whether the covariates are observed pre- or post-treatment. The assumptions are, first, $Y_i(\widetilde{M}_i, \widetilde{T}_i) \perp\!\!\!\perp M_i(\widetilde{T}_i)|\widetilde{T}_i, X_i^{pre}, X_i^{post}$ and $\left\{Y_i\left(\widetilde{M}_i, \widetilde{T}_i\right), M_i(\widetilde{T}_i)\right\} \perp\!\!\!\perp \widetilde{T}_i|X_i^{pre}$, and, second, all manipulations are on the support of $\widetilde{M}_i|X_i^{pre}, X_i^{post}$ and $\widetilde{T}_i|X_i^{pre}$.

The estimate of the controlled direct effect is then simply

$$\widehat{\nabla}_{T_i}^{CDE}(\delta_i^{CDE}) = \frac{1}{\delta_i^{CDE}}(R_i(M_i = m, T_i + \delta_i^{CDE}, X_i) - R_i(M_i = m, T_i + \delta_i^{CDE}, X_i))^\top \widehat{c} \tag{52}$$

We note two advances offered by our estimation strategy. First, estimation can be done in one step since we need not model the mediator as a function of treatment, as we did in recovering the mediation effect above. The method of sequential g-estimation, most commonly used in these models, is a multi-stage method, where summaries from earlier models are used as a bias-correction in later models. For MDE, the bias correction is incorporated through the bases: estimation utilizes bases $R_i(M_i, T_i, X_i)$, but prediction at $R_i(m, T_i, X_i)$. Second, we can relax the assumption that post-treatment confounders are independent of the mediator, which is required for the simplest form of g-estimation.

More generally, structural mean models can be estimated using the basic framework of MDE. The basic element of these models, the "blip function," is the conditional counterfactual difference in outcome between two observations. Under an identity link, which we consider in this paper, the

blip function for a point-treatment is

$$\psi^* \left((T_i'', T_i'), X_i; c^o \right) = \mathbb{E}(Y_i(T_i'') - Y_i(T_i')|X_i) \tag{53}$$

where c^o parameterizes the true mean function. This blip function is constructed to satisfy

$$\mathbb{E}\left\{ Y_i(T_i'') - \psi^* \left((T_i'', T_i'), X_i; c^o \right) | X_i \right\} = \mathbb{E}\left\{ Y_i(T_i') | X_i \right\} \tag{54}$$

allowing recovery of an estimate of the outcome at some desired value, T_i'. This technology can be extended to recover consistent treatment effect estimates across a variety of structural settings (see Vansteelandt, 2009, for an excellent review).

In our notation, this blip function $\psi^* \left((T_i'', T_i'), X_i; c^o \right) = \mathbb{E}\left\{ \Delta_i \left(T_i'', T_i' \right) | X_i \right\}$, a generalization of our primary parameter of interest, $\nabla_{T_i}(\delta)$. The methods described above show how MDE can recover a blip function but also assess whether the desired conditional in the blip down has positive support, a particular concern with a continuous treatment where structural mean models and their extensions are most useful (Vansteelandt, 2009, p. 23).

3.4 Binary Treatments

We have so far focused on and emphasized the method's utility with a continuous treatment variable, but there is nothing in our framework that prevents us from predicting counterfactuals with a binary or categorical treatment variable. We next extend MDE to the case of a binary treatment, where most work in the causal literature has taken place, and show that it provides a reasonable means for effect estimation even relative to methods designed explicitly for the binary treatment case.

In our notation, the individual causal effect under a binary treatment can be written as

$$\mathbb{E}(Y_i(1) - Y_i(0)|X_i) = \nabla_{T_i}(1 - 2 \times T_i) = \mathbb{E}(\Delta_i(1, 0)|X_i) \tag{55}$$

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

$$ATE = \frac{1}{n} \sum_{i=1}^{n} \nabla_{T_i}(1 - 2 \times T_i) = \tag{56}$$

$$ATT = \frac{1}{n_T} \sum_{i=1}^{n} \left(\nabla_{T_i}(1 - 2 \times T_i) \right) \times \mathbf{1}(T_i = 1) \tag{57}$$

where $n_T = \sum_{i=1}^{n} \mathbf{1}(T_i = 1)$. Estimation can now proceed as described above, except we now consider differences instead of partial derivatives.

4 Relationship to Earlier Work

Our approach to modeling counterfactuals incorporates insights from several different fields, as we discuss next.

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 Rubin (1974); Rosenbaum and Rubin (1983, 1984); Robins, Rotnitzky and Zhao (1994); Pearl (2000); van der Laan and Rose (2011) for seminal work.

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 face efficiency reductions compared to a model that incorporates a correctly specified propensity score model (e.g., Robins and Ritov, 1997), yet this efficiency loss must be balanced against the modeling uncertainty that comes from not knowing the true model. We show that focusing attention on the outcome model can provide a feasible, robust, and powerful means for causal estimation.

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). We share some of the motivation in Austin (2012); Hill, Weiss and Zhai (2011); Lam (2013), though these works focused on the binary treatment and did not extend to structural models. We find below that the methods advocated by several of these works do not perform as well as our tensor regression model. 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. Unlike this method, we target observation-level estimates rather than an aggregate parameter. We also 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; see Athey and Imbens. (2016) and Horowitz (2014), esp. sec 3.1.

Relationship to High-Dimensional Regression Modeling Our 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 (Ratkovic and Tingley, 2017). Third, unlike existing sparse Bayesian models (e.g. Polson and Scott, 2012; Rockova and George, Forthcoming), we use a frequentist minmax argument to motivate our hyperprior selection as a function of n, p.

We were inspired by work on tensor-product and smoothing spline ANOVA models Gu (2002); Wood (2016, 2006); Currie, Durban and Eilers (2006); Eilers and Marx (1996). 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, we do not specify a penalty function or reproducing kernel, but instead construct tensor-product bases and let the sparse LASSO prior manage the regularization. While this leads to inefficiencies in estimation, current software cannot handle the number of variables and tensor-interactions we are considering here.

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, i.e. the piecewise linear cubic-spline

or B-spline basis.

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; Newey and Powell, 2003). 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 model. We address the proliferation of bases in this model serves as the central issue in estimation (Belloni et al., 2015, Sec. 3.6) 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.

Our work is perhaps closest to Heckman and Vytlacil (2007*a*,*b*, 2005, 1999)[9], 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.[10] 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 estimated 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

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

[10] See Section D for one way to bridge the two estimands. One could for example calculate the set of observations for which a given level of encouragement could be supported. We also note that, as with our approach the LIV can be aggregated to an average effect. For the Heckman and Vytlacil framework this follows precisely because it is a latent index model with an assumed error (a probit model). In the continuous setting, like we consider here, that expectation requires either assuming the error density or estimating it, as we do.

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. The importance of separate estimates at each stage will allow plug-in estimates for other causal structural models (VanderWeele, 2015), as we illustrate with our discussion of mediation and sequential-g estimation above. Lastly, we are not the first to note that the sampling distribution of instrumental variable estimates can be erratic and non-normal (Bound, Jaeger and Baker, 1995; Imbens and Rosenbaum, 2005). Rather than turning to rank-based estimates, we instead suggest a bagging procedure to smooth over a possibly erratic sampling distribution (see Appendix C).

5 Analyses

5.1 Application: Binary Treatment

As a validation exercise, we test MDE against several existing methods using data first put forward by LaLonde (1986), but see also Smith and Todd (2005); Diamond and Sekhon (2012); Imai and Ratkovic (2014). The data consist of an experimental subset and and observational subset. The experimental subset contains the results from a policy experiment, the National Supported Work Study, with participants randomly assigned to a treated ($n_T = 297$) and untreated ($n_C = 425$) group. The treatment consisted of a job-training program and participants were hard-to-employ individuals across 15 sites in the United States in 1976. The outcome is 1978 earnings and observed covariates include the participants' age, years of schooling, whether the individual received a high school degree, indicators for race (black, hispanic), an indicator for whether the participant is married, previous earnings from 1974 and 1975, and indicators for whether 1974 or 1975 earnings were zero.[11] The observational dataset comes from the Panel Study for Income Dynamics, a survey

[11]The LaLonde data contain several subsets that have been used as a benchmark analysis. One subset, analyzed by Dehejia and Wahba (1999) subsets on the outcome variable, returning a data on which most methods perform quite well. We focus on the full original experimental data, which poses a greater challenge.

of low-income individuals, with data on the same covariates as the experimental data ($n_P = 2915$).

We conduct three separate tests. The first uses the experimental treated and observational untreated, with the goal of recovering the experimental result ($886.30). This is perhaps the most policy-relevant comparison, as it may allow for methods that can estimate the causal impact of a policy intervention in the absence of a randomized control group. The second test is a placebo test, considering the experimental and observational untreated groups. As no one in the treated or untreated group received the treatment, the known true effect is zero, and any non-zero estimate has been termed *evaluation bias* (Smith and Todd, 2005; Imai and Ratkovic, 2014). In the third test, we compare the experimental treated to the experimental untreated, in order to assess the extent to which a causal estimation method can recognize experimental data and recover an effect close to the simple difference-in-means.

We compare the proposed method, in both its point estimate and bagged implementations, to several existing methods. We include the horseshoe model and cross-validated LASSO, both using our nonparametric bases. We also include BART, gradient boosting (GBM), and POLYMARS, an existing nonparametric spline model. We include four methods designed for a binary treatment: logistic propensity score matching (Propensity, Ho et al., 2007), the covariate balancing propensity score (CBPS, Imai and Ratkovic, 2014) the targeted maximum likelihood estimate (Gruber and van der Laan, 2011) and the double-selection OLS post-LASSO estimator of Belloni, Chernozhukov and Hansen (2014); Chernozhukov, Hansen and Spindler (2016).

Results from this analysis can be found in Table 1. The treated and control groups used in the comparison are given in the top two rows. Columns consist of the target, either the experimental benchmark ($886.30) or zero. Results from each method are then listed below and the final column gives the absolute bias across comparisons. Methods are listed in order of performance on this final measure. We see that MDE performs well relative to other methods. CBPS achieves a bias across the three simulations settings comparable to MDE, performing particularly well in the experimental sample. The remaining methods struggle in at least one of the comparisons.

We next illustrate how variable selection alone, without nonparametric preprocessing of covariates, need not reduce model dependent causal effect estimation. The practical point is that a variable selection methodology may suggest estimation is robust to whether higher order terms are included, but we show that this is not necessarily the case.

Treatment Group Control Group	Experimental Observational	Experimental Observational	Experimental Experimental	Total Absloute Bias
Truth	886.30	0.00	886.30	0
MDE	850.82	190.55	572.01	540.31
CBPS	453.19	-300.28	872.27	747.42
OLS Post LASSO	1423.51	803.04	879.01	1347.54
TMLE	-128.11	-520.48	739.26	1681.93
Horseshoe	901.67	1600.79	-93.01	2595.47
BART	-477.87	-1265.07	803.77	2711.77
GBM	-864.43	-1402.73	866.10	3173.66
POLYMARS	-228.30	-1863.41	0.00	3864.32
LASSO	-1204.34	-1726.13	944.94	3875.41
Propensity	-3728.93	-5597.30	1067.96	10394.19

Table 1: **Treatment Effect Estimates, LaLonde Data.** The treated and control groups used in the comparison are given in the top two rows. Columns consist of the target, either the experimental benchmark ($886.30) or zero. Results from each method are then listed below and the final column gives the absolute bias across comparisons. Methods are listed in order of performance on this final measure. We see that MDE performs well across comparisons.

	Treatment Group Control Group	Experimental Observational	Experimental Observational	Experimental Experimental
	Truth	886.30	0.00	886.30
MDE	X	859.10	189.32	832.61
	X reduced	888.41	278.72	720.64
	X, X^2	513.29	614.75	792.76
	X, X^2, interactions	338.08	-105.91	577.83
OLS Post LASSO	X	1423.51	803.04	879.01
	X reduced	46.17	-905.32	893.61
	X, X^2	2306.71	1115.50	756.81
	X, X^2, interactions	8.22×10^{15}	1.33×10^{12}	801.35

Table 2: **Specification Stability, LaLonde Data.** The treated and control groups used in the comparison are given in the top two rows. Columns consist of the target, either the experimental benchmark ($886.30) or zero. Rows show which conditioning set is being used: the complete covariates (X); the covariates dropping covariates constructed from other covariates (i.e. $1974earnings = 0$; X reduced); the complete set of covariates plus square terms (X, X^2); and the covariates, square terms, and two-way interactions (X, X^2, interactions). Setting aside convergence issues in the interactive model, we see that MDE returns values with less variance across different parameterizations of X, each with the same amount of information.

We illustrate by repeating the analysis above, but hold fixed the conditioning set–the amount of information available in the covariates–but vary the form in which the covariates are entered into the model. We first start with the covariate set X, as given above. Next, in row "X, reduced" we drop three covariates that are functions of other covariates, dummy variables for whether a high

MDE	187.25	187.40	188.87	190.54	189.67
Horseshoe	1600.79	1873.87	997.57	1456.29	762.58

Table 3: **Algorithm Stability, MDE versus Horseshoe Prior.** Estimated treatment effects from MDE (top) and a horseshoe regression (bottom) across five random starts, using the same set of nonparametric bases. MDE is much more stable across multiple runs.

school degree is attained ($= \mathbf{1}(schooling_i > 12)$) and indicators for whether earnings in the previous two years were 0. The dropped covariates do not reduce the information in the conditioning set. We then include X and X^2, squared terms for each centered covariate. Again, we are not changing the information in the conditioning set. Lastly, we include X, the square terms, and all interaction terms.

We compard MDE to the post-LASSO OLS treatment effect estimator of Belloni, Chernozhukov and Hansen (2014); results appear in Table 4. We see that MDE is stable across specifications, while the post-LASSO OLS method returns sometimes dramatically different results. For example, when comparing the observational control to the experimental treated set, the effect estimates can vary by over \$2200 dollars, setting aside the failure of convergence on the model with interaction terms. In the experimental data, post-LASSO OLS performs particularly well, as the model is OLS fit to the treatment variable and a set of selected covariates. As the treatment has been randomized, a regression additive in the treatment and any subset of the controls will perform well. In the observational settings, though, model dependence as measured through shifting covariate specifications are dramatically muted by our tensor-spline plus marginal screen method given above.

We next consider algorithm stability, specifically the variance in estimated treatment effects across multiple runs. We compare MDE, with its LASSOplus implementation, to the horseshoe fit via variational inference as implemented in the **R** package `rstanarm`. We focus on the horseshoe estimator due to its popularity in both the applied and theoretical literatures. Both MDE and the horseshoe regression are given the same set of nonparametric bases. Results are in Table 3, where we are only looking at the comparison between untreated experimental units and the untreated observational units (the middle column in Tables 1 and 4), as we know the true effect is \$0. From run to run, MDE stays around $\$188.75 \pm \$2$, whereas the horseshoe estimator has a range of over \$1100.

Our nonparametric basis construction generates a regression model that is robust to how covariates are entered and our Bayesian prior generates estimates robust to different starting values,

while maintaining a high level of accuracy in the LaLonde exercise.

In Appendix H.1 we include an additional LaLonde analysis, looking at how well each method predicts observation-level effects, rather than simply average effects, in a held-out sample. Again, we find that MDE performs competitively.

5.2 Application: Instrumental Variables

Next we examine an empirical application using instrumental variable methods. Here we focus on an application with a continuous instrument in order to highlight how the proposed methodology naturally incorporates continuous as well as non-continuous exogenous variables.

Larreguy and Marshall (2017) study the long term political effects of increased education. To do this they utilize variation in the intensity of a Nigerian government reform, the Universal Primary Education reform of 1976. The authors leverage Afrobarometer survey data to explore a variety of political variables, such as interest in the news and knowledge of politics. Concerned about endogeneity problems present in regressing these political variables onto measures of education, they use an instrumental variables strategy akin to earlier work in development economics on the effects of education (Duflo, 2001). In particular they exploit temporal (the program started in 1976 and so impacted particular cohorts of citizens) and spatial (regional level differences between actual and potential enrollment, see also Bleakley (2010)) variation. They use linear two stage least squares with an interaction between the intensity of the program and whether an individual would have been eligible to benefit from the program as an instrument for education levels. The authors include a range of control variables and fixed effects for cohort, region, and Afrobarometer survey wave. They find strong and robust impacts of education on long term political variables. Using their replication data we implement our proposed method to reevaluate their results.

Part of our interest is in allowing for the instrument to have non-linear effects on the endogenous variable. Hence before displaying results from our analysis, we present evidence of such a non-linear relationship by simply fitting a generalized additive model to the data, allowing for local smoothing on the instrument and including the full set of controls. On the x-axis we present the values of the instrument along with a histogram of its marginal distribution. On the y-axis we present the fitted values of the endogenous variable.[12] Here and below we present this distribution as a 2-dimensional heatmap in order to convey density of observations over the space, and fit a smooth trend line

[12]We re-scaled the endogenous variable, education, which in the original analysis ran from 1 to 5.

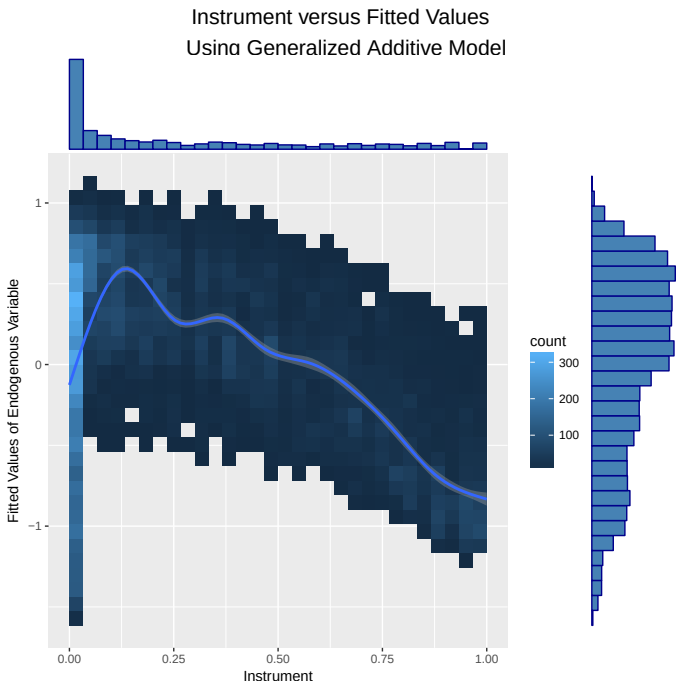

Figure 1: **Nonlinear bi-variate relationship between instrument and fitted values of the endogenous variable. Estimated using generalized additive model and not the proposed method.**

to convey the basic pattern. Generally speaking we see higher effects of the instrument at lower values of the instrument, and lower effects at higher values of the instrument. This gives some evidence that it might be desirable in calculating causal effects to incorporate greater functional form flexibility into the analysis.

We fit our model using a full set of splines and interactions described in Section 2.2.[13] We present two main sets of results.

First, in Figure 2 we present a plot analogous to Figure 1 which plots for each observation the relationship between the instrument and the covariance between the instrument and and (endogenous) treatment variable. This is the "first stage" result. We see a strong positive relationship until the upper end of the instrument distribution, where it begins to get closer to 0. A behavioral interpretation of this pattern is that the returns on education to increasing intensity in the program were lower in areas with greater intensity, perhaps because it would be harder to increase their education levels even higher.[14]

Figure 3 plots the "second stage": the treatment variable versus the covariance between the

[13]The original model used a substantial number of fixed effects, in part to establish a difference in difference identification strategy with an instrumental variable. This poses no problem for our proposed method. However for

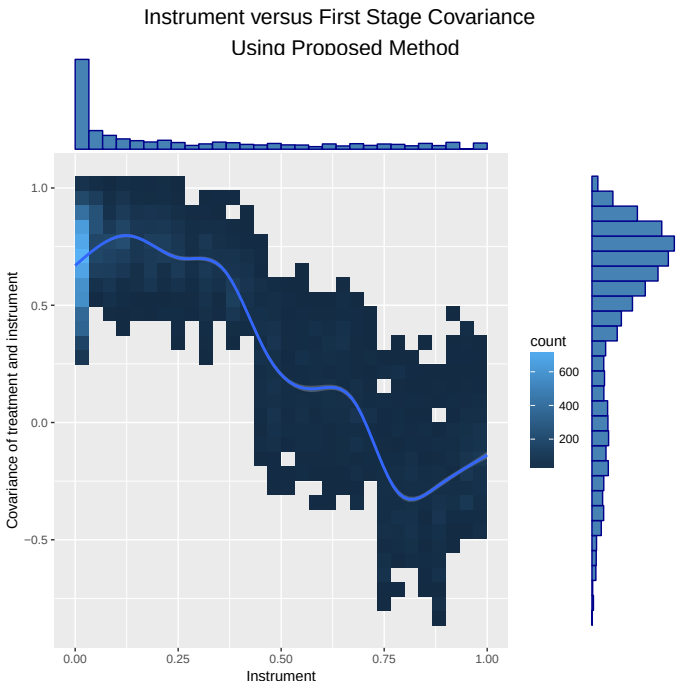

Figure 2: **Relationship between instrument and first stage covariance (covariance between instrument and treatment). Estimated using the proposed method.**

treatment and outcome variable. A piecewise line plots the means across the values of the treatment variable. At lower lower levels of the treatment variable the average covariance was lower than at higher levels of the treatment variable. Combined with the results in Figure 2 this helps to give us an expectation of what the LICE will look like over the sample.

Remembering that the LICE is a ratio estimate, the next step then is to directly plot the numerator of our IV estimate (the covariance between the instrument and treatment) against the denominator (the covariance between the treatment and outcome). Figure 4 plots the results. Several patterns are interesting. First, the majority of observations are in the upper right quadrant, with positive covariances. Second, there is a non-linear relationship. As the first stage covariance increases (that is, the strength of the instrument), there are declining returns to having an effect on the second stage relationship. Substantively this implies a positive relationship overall between education and interest in news, but one that is diminishing at higher levels once endogeneity concerns have been addressed. Third, there are a small number of observations in the bottom left quadrant.

computational purposes we made some minor modifications that we discuss in Appendix H.

[14]We also investigated the first stage fit of our model compared to least squares via cross-validation. We found nearly identical performance, which is impressive for our proposed method given that the sample size is quite large relative to the number of parameters fit by the least squares model.

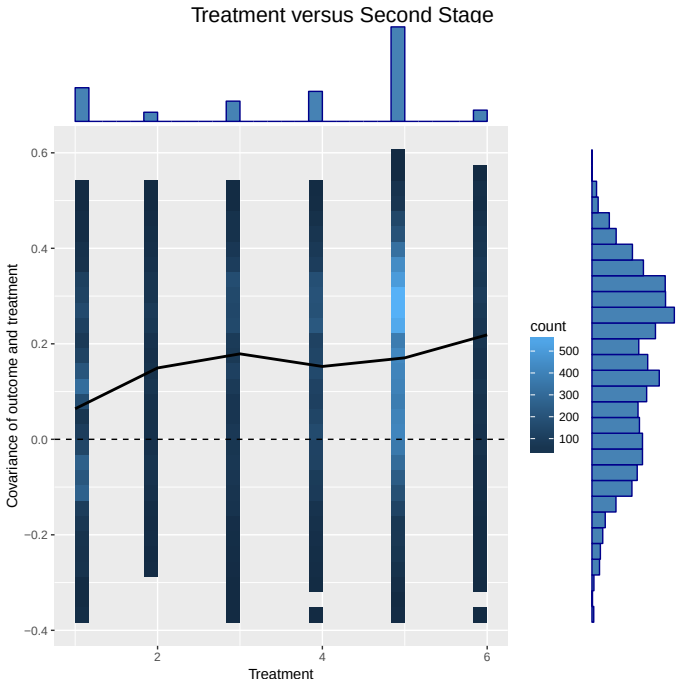

Figure 3: **Treatment variable versus covariance between treatment and outcome.**

For these observations the LICE will still be positive. On average these observations are likely to be of lower education (see Figure 3). Behaviorally these individuals were negatively encouraged but also saw a decrease in interest in news. Finally, there are no observations in the top left quadrant, and few observations in the bottom right. Individuals in the bottom right were positively encouraged by the reforms but saw a decrease in news interest; for these individuals we would expect a negative LICE

Finally in Figure 5 we plot the distribution of local individual causal effects (LICE) against the treatment. The dashed horizontal lines represents the average LICE, which we estimate to be .35 (95% confidence interval: .20, .49).[15] While there was non-linearity in the first stage estimate, we observe more stability over the sample in the LICE. This directly follows from the results in Figure 4.

[15]We bootstrapped the confidence intervals using 100 bootstrap runs. Using TSLS the original analysis returned a point estimate of .62. Larreguy and Marshall (2017) cluster standard errors at the state level. However, re-analysis of their specification shows this made no difference. In fact, the confidence intervals were slightly wider without clustering. We did not cluster standard errors, though a block bootstrap approach would accomplish this goal.

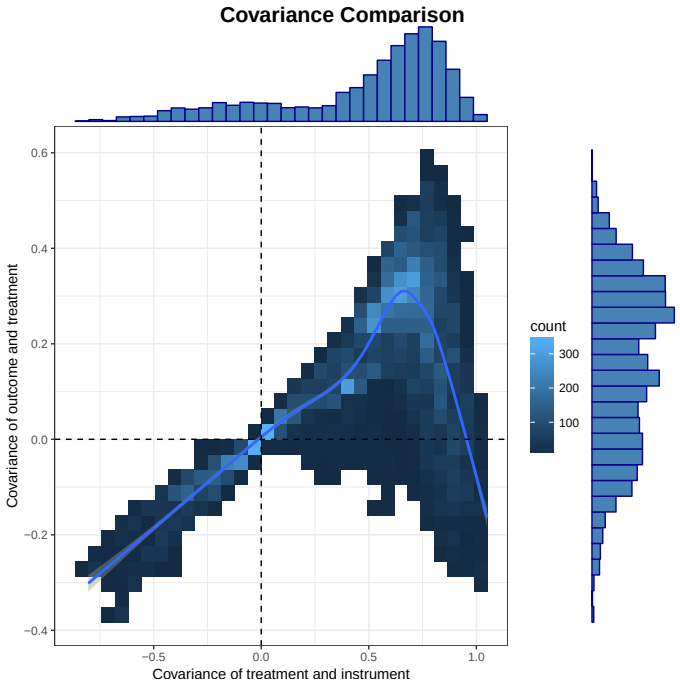

Figure 4: **First stage versus second stage.**

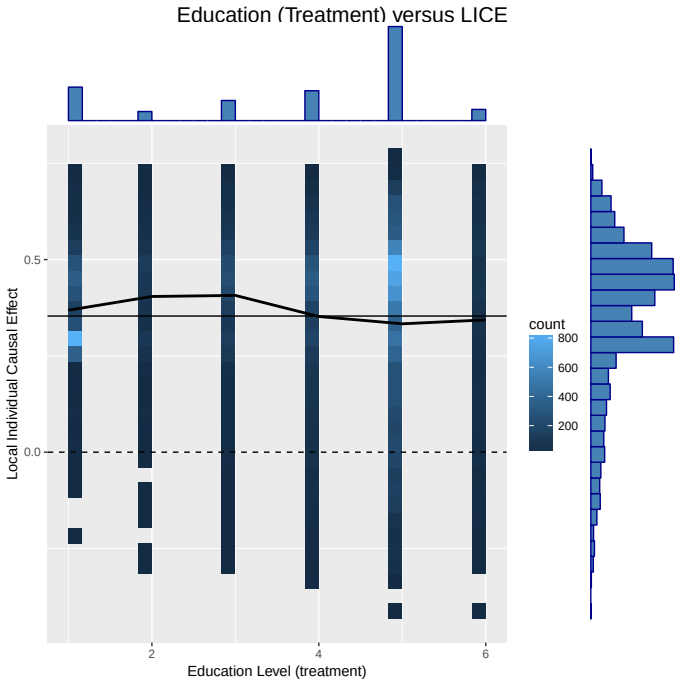

Figure 5: **Local Individual Causal Effect versus treatment variable.**

6 Conclusion

We focus on using a high-dimensional, flexible regression model directly target observation-level causal effect estimates by considering an extremely rich covariate *and* functional space. We utilize

recent advances in machine learning to estimate our model and show that the proposed approach performs extremely well against other cutting edge methods. Finally, we show how our estimation strategy extends beyond the estimation of simple treatment effects and can also be used in instrumental variable or other structural models.

We see this work as providing an integrative thread across a number of seminal contributions. 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. We work towards this goal head-on, avoiding workarounds like inverse weights and propensity scores, which add another layer of modeling uncertainty and are rarely subjects of direct interest. An early stumbling block in causal inference was acknowledged by Robins, in that causal inference is only as persuasive as the underlying model. Rather than develop methods that are reasonable even in the face of model misspecification, we focus our attention on getting the conditional mean correct. 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; Ratkovic and Tingley, 2017) in order to focus on this goal.[16] Though we started the project in terms of the potential outcomes approach, we found that our estimation strategy combined thinking in terms of observation-level counterfactuals as well as structural "do" manipulations, as developed by Pearl. 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).

The are a number of areas for future work. One thing we have left out of the current paper is the set of advantages of using the Bayesian LASSOPlus model that we previously developed (Ratkovic and Tingley, 2017).[17] We are actively extending the method of direct estimation to cases with multiple separate treatment variables, instruments (e.g., Chernozhukov, Hansen and Spindler, 2015; Belloni et al., 2012), or mediators. Our current approach considers the impact of perturbing a single instrument, but extending it to multiple instruments and treatments will involve moving

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

[17]These advantages include straightforward ways to incorporate binary or truncated outcome variables, random effects, uncertainty estimates with desirable coverage properties, and not relying on arbitrary tuning parameter selection.

from partial to Frechet derivatives. We have also explored a means for estimating the range of values for the treatment over which the model could predict. We can then use this approach to change our estimand to the expected value of the local effect over the range of treatment values, $\mathbb{E}_{\widetilde{T}_i}(\lim_{\delta \to 0} \nabla_{T_i}(\delta))$ (see Appendix D). Appendix C explores the use of bagging in contexts where there will be high sampling variability. Finally, future work could also consider how to extend the model to the longitudinal setting or how to incorporate spline bases with discontinuities so as to capture "jumps" in the data, and how to estimate the effect in a regression discontinuity design.

References

Abadie, Alberto. 2003. "Semiparametric instrumental variable estimation of treatment response models." *Journal of Econometrics* 113(231–263).

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

Alhamzawi, Rahim, Keming Yu and Dries F Benoit. 2012. "Bayesian adaptive Lasso quantile regression." *Statistical Modelling* 12(3):279–297.

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 M and Allison Carnegie. 2013. "Beyond LATE: Estimation of the average treatment effect with an instrumental variable." *Political Analysis* 21(4):492–506.

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* .

Austin, Peter C. 2012. "Using Ensemble-Based Methods for Directly Estimating Causal Effects: An Investigation of Tree-Based G-Computation." *Multivariate Behavioral Research* 47:115–135.

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

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.

Belloni, Alexandre, Victor Chernozhukov, Ivan Fernandez-Val and Christian Hansen. 2017. "Program Evaluation and Causal Inference With High-Dimensional Data." *Econometrica* 85(1):233–298.

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

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.

Bleakley, Hoyt. 2010. "Malaria eradication in the Americas: A retrospective analysis of childhood exposure." *American Economic Journal: Applied Economics* 2(2):1–45.

Bound, John, David A Jaeger and Regina M Baker. 1995. "Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak." *Journal of the American statistical association* 90(430):443–450.

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

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.".

Chernozhukov, Victor, Christian Hansen and Martin Spindler. 2016. "hdm: High-Dimensional Metrics." *The R Journal* 8(2):185–199.

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. 1968. "The effectiveness of adjustment by subclassification in removing bias in observational studies." *Biometrics* 24:295–313.

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

Cochran, William G. and Donald B. Rubin. 1973. "Controlling bias in observational studies: A review." *Sankhya: The Indian Journal of Statistics, Series A* 35:417–446.

Currie, I. D., M. Durban and P. H. C. Eilers. 2006. "Generalized linear array models with applications to multidimensional smoothing." *Journal of the Royal Statistcal Society, Series B* 68(2).

Davidson, Russel and Emmanuel Flachaire. 2008. "The wild bootstrap, tamed at last." *Journal of Econometrics* 146(1):162–169.

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

Dehejia, Rajeev H and Sadek Wahba. 1999. "Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs." *Journal of the American statistical Association* 94(448):1053–1062.

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

Duflo, Esther. 2001. "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment." *American Economic Review* 91(4):795–813. **URL:** *http://www.aeaweb.org/articles?id=10.1257/aer.91.4.795*

Eilers, Paul H. C. and Brian D. Marx. 1996. "Flexible smoothing with B-splines and penalties." *Statistical Science* 11(2):89–121.

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, Yang Feng and Rui Song. 2012. "Nonparametric independence screening in sparse ultra-high-dimensional additive models." *Journal of the American Statistical Association* .

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.

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

Griffin, J. E. and P. J. Brown. 2010. "Inference with normal-gamma prior distributions in regression problems." *Bayesian Analysis* 5(1):171–188.

Griffin, J. E. and P. J. Brown. 2012. "Structuring shrinkage: some correlated priors for regression." *Biometrika* 99(2):481–487.

Gruber, Susan and Mark J van der Laan. 2011. "tmle: An R package for targeted maximum likelihood estimation.".

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.
URL: *+ http://dx.doi.org/10.1093/pan/mpt019*

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.

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. 1999. "Local instrumental variables and latent variable models for identifying and bounding treatment effects." *Proceedings of the National Academy of Sciences* 96:4730–4734.

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

Heckman, James and Edward Vytlacil. 2007*a*. "Econometric Evaluation of Social Programs, Part 1: Causal Models,Structural Models, and Econometric Policy Evaluation." *Handbook of Econometrics* 6b:4779–4874.

Heckman, James and Edward Vytlacil. 2007*b*. "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments." *Handbook of Econometrics* 6b:4875–5143.

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

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

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* .

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

Imai, Kosuke, Luke Keele and Teppei Yamamoto. 2010. "Identification, inference and sensitivity analysis for causal mediation effects." *Statistical Science* pp. 51–71.

Imai, Kosuke and Marc Ratkovic. 2014. "Covariate Balancing Propensity Score." *Journal of the Royal Statistical Society Series B* 76(1):243–263.

Imbens, G.W. and P.R. Rosenbaum. 2005. "Robust, accurate confidence intervals with a weak instrument: quarter of birth and education." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 168(1):109–126.

Kang, Jian and Jian Guo. 2009. "Self-adaptive Lasso and its Bayesian Estimation." Working Paper.

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* .

King, Gary and Langche Zeng. 2006. "The dangers of extreme counterfactuals." *Political Analysis* 14(2):131–159.

Kleibergen, Frank. 2002. "Pivotal statistics for testing structural parameters in instrumental variables regression." *Econometrica* 70(5):1781–1803.

LaLonde, Robert J. 1986. "Evaluating the econometric evaluations of training programs with experimental data." *The American economic review* pp. 604–620.

Lam, Patrick. 2013. Estimating Individual Causal Effects PhD thesis Harvard University.

Larreguy, Horacio and John Marshall. 2017. "The Effect of Education on Civic and Political Engagement in Non-Consolidated Democracies: Evidence from Nigeria." *Review of Economics and Statistics* .

Leng, Chenlei, Minh-Ngoc Tran and David Nott. 2014. "Bayesian Adaptive LASSO." *Annals of the Institute of Statistical Mathematics* 66(2):221–244.

MacKinnon, David P, Chondra M Lockwood, Jeanne M Hoffman, Stephen G West and Virgil Sheets. 2002. "A comparison of methods to test mediation and other intervening variable effects." *Psychological methods* 7(1):83.

Meinshausen, Nicolai. 2007. "Relaxed LASSO." *Computational Statistics and Data Analysis* 52(1):374–393.

Morgan, Kari Lock and Donald B. Rubin. 2012. "Rerandomization to improve covariate balance in experiments." *Annals of Statistics* 40(2):1263–1282.

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

Newey, Whitney and James Powell. 2003. "Instrumental Variable Estimation of Nonparametric Models." *Econometrica* 71(5):1565–78.

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. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambrdige University Press.

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.

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. 1986. "A new approach to causal inference in mortality studies with sustained exposure periods: Application to control of the healthy worker survivor effect." *Mathematical Modeling* 7:1393–1512.

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., Andrea Rotnitzky and Lue Ping Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors are Not Always Observed." *Journal of the American Statistical Association* 89(427):846–866.

Robins, James M, Miguel Angel Hernan and Babette Brumback. 2000. "Marginal structural models and causal inference in epidemiology." *Epidemiology* pp. 550–560.

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* .

Rosenbaum, Paul R. and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational studies for Causal Effects." *Biometrika* 70(1):41–55.

Rosenbaum, Paul R. and Donald B. Rubin. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79(387):516–524.

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology* 66(5):688.

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.

Smith, Jeffrey A and Petra E Todd. 2005. "Does matching overcome LaLonde's critique of nonexperimental estimators?" *Journal of Econometrics* 125(1):305–353.

Staiger, Douglas O and James H Stock. 1994. "Instrumental variables regression with weak instruments.".

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

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.

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

Vansteelandt, Stijn. 2009. "Estimating direct effects in cohort and case–control studies." *Epidemiology* 20(6):851–860.

Wood, Simon. 2006. "Low-Rank Scale-Invariant Tensor Product Smooths for Generalized Additive Mixed Models." *Biometrics* 62(4):1025–1036.

Wood, Simon. 2016. "P-splines with derivative based penalties and tensor product smoothing of unevenly distributed data." https://arxiv.org/pdf/1605.02446.pdf.

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

Appendix

The outline of the appendix is as follows.

Section A details properties of the LASSOPlus estimator. Interested readers can consult Ratkovic and Tingley (2017) for additional details. Section B proves validity of our bootstrap estimate. Section C covers how bagging can be used and Section D proposes a method for calculating the extrapolation range. Section F conducts a simulation study on the IV compliance threshold and proposes a diagnostic for instrumental variables akin to the first-stage F-statistic. Section G gives the proofs of our Oracle results. Section H discusses our processing of the Larreguy and Marshall (2017) data as well as plots of variable effect estimates. Section H.1 includes additional exercises using the LaLonde data.

A LASSOplus Properties

Our joint conditional density on c, w produces a problem similar to the adaptive LASSO of Zou (2006):

$$-\log\left(\Pr\left(c, \{w_k\}_{k=1}^p|\lambda, \sigma^2, \gamma\right)\right) = \frac{1}{\sigma^2}\left\{\frac{1}{2}\sum_{i=1}^n (Y_i - R(T_i, X_i)^\top c)^2 + \lambda\sigma\sum_{k=1}^p w_k|c_k|\right\} + \sum_{k=1}^p w_k^\gamma. \quad (58)$$

where the weights and global sparsity parameter enter the log-posterior as w_k^γ rather than as separate terms, as in Leng, Tran and Nott (2014); Alhamzawi, Yu and Benoit (2012); Griffin and Brown (2012, 2010); Kang and Guo (2009).

Estimation Following Park and Casella (2008), we reintroduce conjugacy in the mean parameters through augmentation:

$$c_k \sim DE(w_k\lambda/\sigma) \Rightarrow c_k|\tau_k^2, \sigma^2 \sim \mathcal{N}(0, \tau_k^2\sigma^2); \;\; \tau_k^2 \sim \exp(\lambda^2 w_k^2/2). \quad (59)$$

Both the MCMC and EM estimation details are standard, see Ratkovic and Tingley (2017).

The tuning parameter. We first focus on the conditional posterior density of the tuning parameter, λ. Rescaling by n reveals that we are recovering the MAP estimate

$$\widehat{c}\,|\cdot = \underset{c}{\text{argmin}}\, \frac{1}{n}\sum_{i=1}^n \frac{(Y_i - R(T_i, X_i)^\top c)^2}{2\sigma^2} + \frac{\lambda}{n\sigma}\sum_{k=1}^p w_k|c_k|. \quad (60)$$

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

$$\lambda^2|\cdot \sim \Gamma\left(n \times (\log(n) + 2\log(p)), \sum_{k=1}^{p} \tau_k^2/2 + \rho\right) \tag{61}$$

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. For a full discussion of bounds, see Ratkovic and Tingley (2017).

The weights. To see the impact of the weights, consider the posterior conditional density and mean of w_k,

$$\Pr(w_k|\cdot) = \frac{e^{-w_k^\gamma - \frac{\lambda w_k}{\sigma}|c_k|}}{\int e^{-w_k^\gamma - \frac{\lambda w_k}{\sigma}|c_k|} dw_k}; \quad \mathbb{E}(w_k|\cdot) = \int_{w_k=0}^{\infty} w_k \Pr(w_k|\cdot) dw_k \tag{62}$$

with $\widehat{w}_k$ simply the conditional mean of w_k under this density.

The derivative of $\widehat{w}_k$ with respect to $|c_k|$ evaluated at the EM estimate gives

$$\frac{\partial \widehat{w}_k}{\partial |c_k|} = -\widehat{\lambda}\sqrt{\widehat{\frac{1}{\sigma^2}}}\mathrm{Var}(w_k|\cdot) < 0 \tag{63}$$

showing that the weights are inversely related to the magnitude of the estimates, $|c_k|$. This inverse relationship between the weights and the estimate gives us the same properties as the adaptive LASSO of Zou (2006).

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 $\overline{\gamma}$.[18] 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.

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

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.

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.

B Proof of Validity of Bootstrap Estimate

Denote as T_{B_N} the complete set of all possible bootstrap estimates. We need to show that the bootstrap estimate T_{B_N} converges in distribution to $\widetilde{T}_i$. We exploit the independence between the bootstrapped fitted values, $\widehat{T}_{B_N}$ and permuted bootstrapped errors $\widehat{\epsilon}_{B'_N}$, as well as Slutsky's Theorem, to show that a bootstrapped approximation to the convolution formula converges to its population analog. We assume that the distribution of $\widetilde{T}_i|X_i$ is Glivenko-Cantelli with bounded density a.e. and that $\epsilon_i \perp\!\!\!\perp X_i$. Then, with F_{B_N} the empirical bootstrap distribution over replicates

$$F_{B_N}(z) = \Pr(T_{ib} \leq z) = \Pr(\widehat{T}_{ib} + \widehat{\epsilon}_{ib'} \leq z) \tag{64}$$

$$= \sum_{i=1}^{B_N} \sum_{x \in \widehat{\epsilon}_{B'_N}} \frac{1}{B_N} \mathbf{1}(x = \widehat{\epsilon}_{ib}) \mathbf{1}(\widehat{T}_{ib} + x \leq z) \xrightarrow{\text{p}} \int f_{\epsilon_i}(x) F_{\widetilde{T}_i}(z - x) dx \tag{65}$$

$$= F(\widetilde{T} \leq z) \tag{66}$$

by Glivenko-Cantelli and Slutsky. In practice T_{B_N} is too large to compute, so we use T_B with large B.

C Bagging

We have several reasons to be concerned that our estimates may have a large sampling variance. The LASSO problem minimized in Equation 21 conditions on the estimated residuals. This suggests that results may be sensitive to sample-specific outlying residuals. Second, we are fitting a high-dimensional nonparametric regression, which is known to generate high-variance estimates near the boundaries of the covariate space. Beyond the generic issues of nonparametric regression, several of our examples and simulations involve cases where we confront highly-skewed or fat-tailed distributions. For example, in an instrumental variable setting, the effect estimate is a ratio estimator that may have such fat tails that the sampling distribution may have no finite moments. In one of our empirical examples, the outcome is earnings, which is highly right-skewed.

To reduce the variance of our estimates, we turn to bootstrap aggregation (bagged, Buhlmann and Yu, 2002). We implement bagging use a Rademacher-wild bootstrap, to account for possible heteroskedasticity (e.g., Davidson and Flachaire, 2008). Bagging normally involves taking the mean across bootstrapped samples, but as we are worried about skew or whether the mean is even finite, we take as our bagged estimate the median across bootstrap samples for each observation. We show below that the method can lead to a decrease root-mean-squared error with only little increase in bias.

We confront a similar problem with our IV estimator. The LICE is a ratio estimator, and we worry that the estimator's sampling distribution may be Cauchy or approximately so. In order to stabilize the estimation, we utilize median bagged estimates,

$$\begin{aligned}\widehat{\nabla}^{IV;boot}_{Z_i}(\delta^{IV}_i) = \text{med}_B \left(\frac{1}{\left\{R_T(Z_i+\delta^{IV}_i, X_i) - R_T(Z_i, X_i)\right\}^\top \widehat{c}_{T,b}} \right) \times \\ \text{med}_B \left(\left\{ R\left(R_T(Z_i+\delta^{IV}_i, X_i)^\top \widehat{c}_{T,b}, X_i\right) - R\left(R_T(Z_i, X_i)^\top \widehat{c}_{T,b}, X_i\right)\right\}^\top \widehat{c}_b \right)\end{aligned} \tag{67}$$

where $\text{med}_B(a)$ refers to the median over bootstrap samples $b \in \{1, 2, \ldots, B\}$. This median-bagging is effective at reducing the impact of wild or erratic estimates.

D Extrapolation Ranges and Expected Effects

We have so far focused on the partial derivative of an outcome with respect to a treatment as our primary estimand. The researcher may want to know the causal effect of a larger shift in the treatment, or perhaps expected outcomes over a range of treatment values. As a model of the conditional

mean, MDE certainly *can* estimate the outcome under any treatment level. Estimates are only reliable, though, over the range of treatment values that have non-zero probability, supp $(\widetilde{T}_i|X_i)$. Doing so requires an estimate of the density of $\widetilde{T}_i|X_i$

We suggest using a bootstrap estimate. Denote as $\widehat{T}_i$ the estimate of the treatment from our regression model; $\widehat{\epsilon}_i$ the estimated residual; $\mathbb{E}(T_i|X_i)$ the population value; $\widehat{T}_{ib}, \widehat{\epsilon}_{ib}$ estimated fitted values and residuals from a bootstrap replicate; and $\widehat{T}_B = \{\widehat{T}_{ib}\}_{b=1}^B, \widehat{\epsilon}_B = \{\widehat{\epsilon}_{ib}\}_{b=1}^B$, and $q(A, \alpha)$ the α^{th} quantile of set A.

We cannot construct a bootstrapped estimate from $\widehat{T}_{ib}+\widehat{\epsilon}_{ib}$ since this value is T_i for each replicate. Instead, we construct our estimates from $T_{ib} = \widehat{T}_{ib} + \widehat{\epsilon}_{ib'}$, for some $b \neq b'$, breaking the correlation between fitted value and residuals for a given observation over bootstrap replicates. Under the assumption that the error is independent of the systematic component, the set $T_B = \{T_{ib}\}_{b=1}^B$ can be used to construct a confidence interval. For example, we can take use the percentile bootstrap, $[q(T_b, \alpha/2), q(T_B, 1-\alpha/2)]$ for false positive rate α. For proof, see Appendix B.

This bootstrap estimate of the sampling distribution offers three separate advances. First, we can use it to estimate a reasonable range of extrapolation for each covariate profile. Doing so helps give a sense of what range of counterfactual predictions are supported by the data.

Second, the bootstrap estimated support can be used *ex ante*, say in an experiment, or *ex post*, in an observational study, to find observations with common support. In an experimental study, a randomization with poor in-sample balance on prognostic variable, can be rerun (e.g. Morgan and Rubin, 2012). In a field study, where experimental units may be quite heterogeneous, the researcher may have concerns over the ability to compare the treated and control group in a given randomization. The support for the treated and control can be estimated, and consequently compared, as a means of assessing whether the treatment and control groups are indeed comparable. In an observational study, the estimated support can be used for trimming. Rather than trim off the marginal covariates (e.g. King and Zeng, 2006), the bootstrap estimates can be used for trimming off the actual estimated support of the treatment variable.

Lastly, the bootstrap estimate can be used to extend from a local effect at the observed value, $\lim_{\delta \to 0} \nabla_{T_i}(\delta)$, to an expected effect. The local effect conditions on a particular value of the treatment, $\widetilde{T}_i = T_i$. It could very well be, though, that the effect at the observed T_i may be different than an expected value over all possible T_i. The researcher interested in the expected effect over

the treatment, $\mathbb{E}_{\widetilde{T}_i}(\lim_{\delta\to 0}\nabla_{T_i}(\delta))$, can use T_B to approximate this integral,

$$\widehat{\mathbb{E}}_{\widetilde{T}_i}\left\{\lim_{\delta\to 0}\nabla_{T_i}(\delta)\right\} = \frac{1}{B}\sum_{b=1}^{B}\lim_{\delta\to 0}\nabla_{T_{ib}}(\delta), \tag{68}$$

the expected effect of the treatment.

E 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:

$$\text{Linear:}\quad Y_i = T_i + \sum_{k=5}^{8} X_{ik}\theta_k + \epsilon_i \tag{69}$$

$$\text{Interactive:}\quad Y_i = T_i - T_i \times X_{i3} + \sum_{k=5}^{8} X_{ik}\theta_k + X_{i1}X_{i2} + \epsilon_i \tag{70}$$

$$\begin{aligned}\text{Nonlinear:}\quad Y_i = 20\sin\left(\left(X_{i1}-\frac{1}{2}\right)\frac{T_i}{20}\right) + \frac{1}{2}\times(1+|X_{i3}\times X_{i4}|_+)\times(1+T_i) + \\ \sum_{k=5}^{8} X_{ik}\theta_k + \epsilon_i\end{aligned} \tag{71}$$

$$\text{Discontinuous } Y_i = (1+|T_i|)\times\mathbf{1}(|T_i|>1/2)\times(X_{i5}+1) + \sum_{k=5}^{8} X_{ik}\theta_k + X_{i1}X_{i2} + \epsilon_i \tag{72}$$

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

$$T_i = -3 + (X_{i1}+X_{i4})^2 + \epsilon_i^T;\quad \epsilon_i^T \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,4) \tag{73}$$

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.[19] 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.

E.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 $\widehat{\mu}_i = \frac{1}{n}\sum_{i=1}^n Y_i = \overline{Y}_i$.

$$\text{RMSE}: \sqrt{\frac{\sum_{i=1}^n(\mu_i - \widehat{Y}_i)^2}{\sum_{i=1}^n(\mu_i - \overline{Y}_i)^2}} \tag{74}$$

$$\text{RMSE}_\nabla: \sqrt{\frac{\sum_{i=1}^n\left(\frac{\partial\mu_i}{\partial T_i} - \widehat{\frac{\partial Y_i}{\partial T_i}}\right)^2}{\sum_{i=1}^n\left\{\frac{\partial\mu_i}{\partial T_i}\right\}^2}} \tag{75}$$

$$\text{Bias}_\nabla: \frac{\left|\sum_{i=1}^n \frac{\partial\mu_i}{\partial T_i} - \widehat{\frac{\partial\mu_i}{\partial T_i}}\right|}{\sqrt{\sum_{i=1}^n\left\{\frac{\partial\mu_i}{\partial T_i}\right\}^2}} \tag{76}$$

[19] 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.

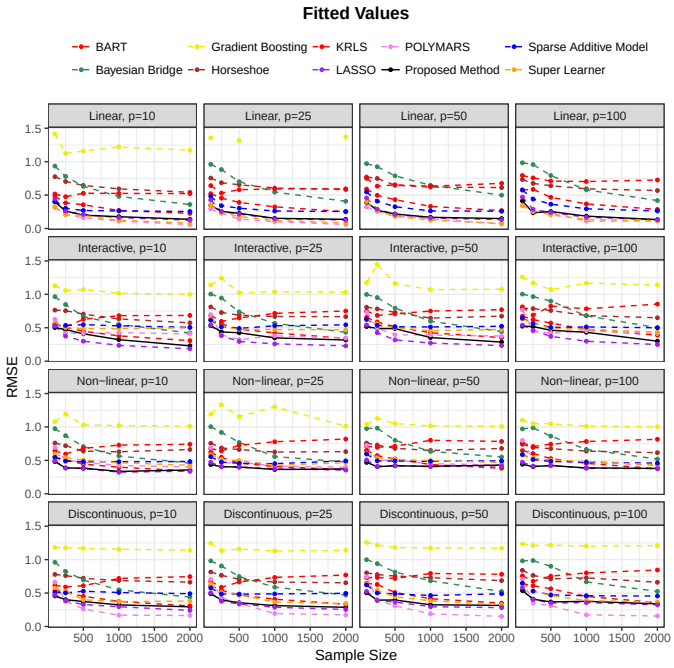

Figure 6: **A Comparison of RMSE for Fitted Values.**

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 6 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 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.[20] Finally while MDE is in a top performing set for the discontinuous case, POLYMARS had a slight advantage.

Figure 7 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,

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

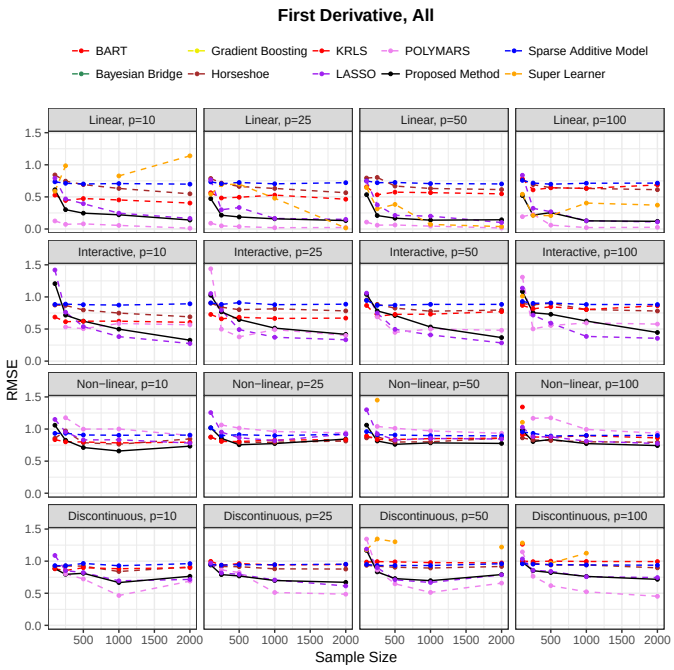

Figure 7: **A Comparison of RMSE for Unit Level Derivatives Across Methods.**

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 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

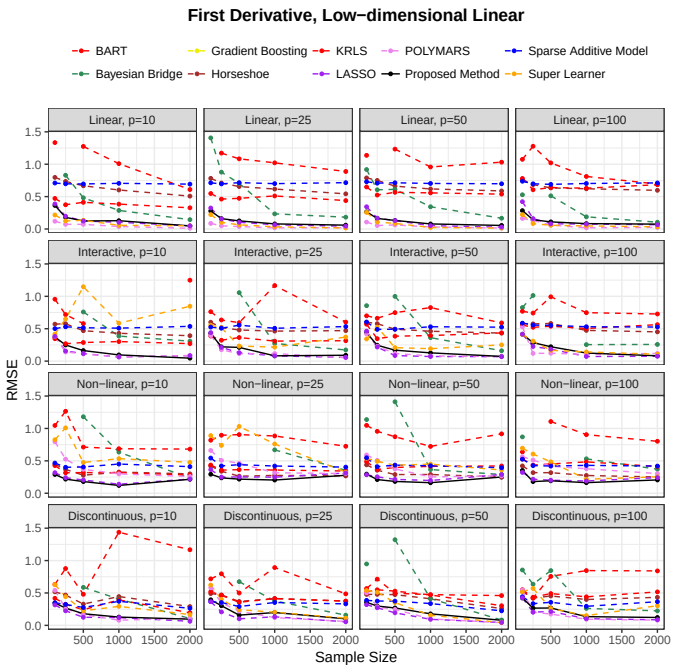

Figure 8: **A Comparison of RMSE Across Methods. Linear terms.**

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 8 and Figure 9 . Figure 8 presents results for the low-dimensional linear trend while Figure 9 presents results for residual non-linear trend terms. We find that the proposed method performs well for both sets of effects. Interestingly in Figure 9 we see that POLYMARS does more poorly in the non-linear setting but better in the discontinuous setting.[21]

F Instrumental Variables Diagnostics and Simulations

F.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

[21]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.

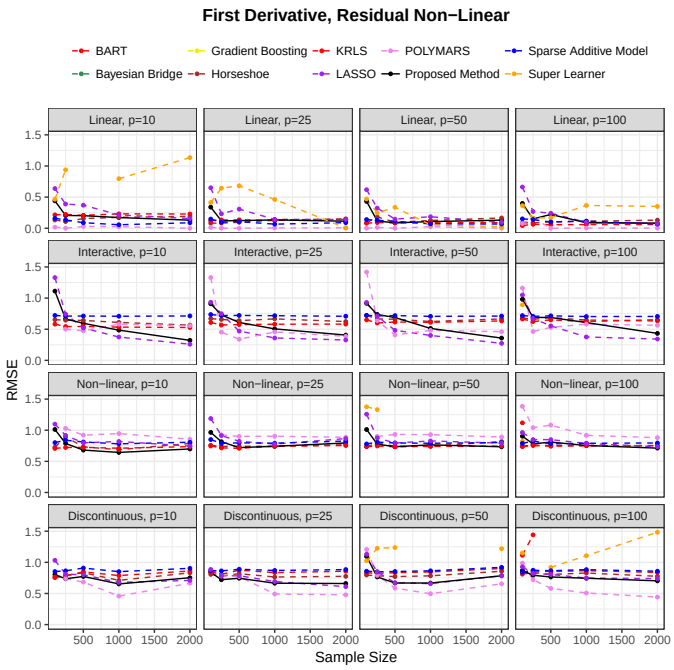

Figure 9: **A Comparison of RMSE Across Methods. Residual non-linear terms**

proposed method to two-stage least squares estimation.[22] We examine first stage and second stage performance, as well as how well the proposed method helps to unpack individual level compliance estimates

Simulation Environments We consider five different simulation settings. We fix the second stage (outcome) model but vary the first stage (treatment) model across five settings. The fist 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 contains a mixture of compliers, non-compliers, and defiers. The fourth scenario is the non-linear model from the earlier simulation setting, and the fifth contains a discontinuity in the instrument. We note that, in this last setting, the true function is not in the model space of our spline bases.

Specifically, we use the exact same settings as in the previous simulations to generate the treat-

[22]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 E have not been extended to the instrumental variables context except for ensemble tree methods by Athey, Tibshirani and Wager (2016).

ment from the instrument. We add an additional null model,

$$\text{Null} \;\; T_i = \sum_{k=5}^{8} X_{ik}\theta_k + \epsilon_i^T, \tag{77}$$

resulting in five first stage models. The instrument is also generated some confounding from covariates

$$Z_i = -3 + (X_{i1} + X_{i4})^2 + \epsilon_i^T; \;\; \epsilon_i^Z \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 4). \tag{78}$$

Across settings, the outcome is generated as

$$Y_i = |T_i - 2| + \sum_{k \in \{5,6,7,8\}} X_{ik}\theta_k + X_{i1} \times X_{i2} + \epsilon_i^Y. \tag{79}$$

where

$$[\epsilon_i^T, \epsilon_i^Y]^\top \overset{\text{i.i.d.}}{\sim} N\left([0,0]^\top, C \begin{bmatrix} 1, .9 \\ .9, 1 \end{bmatrix}\right) \tag{80}$$

where C is selected such that the second stage has a signal to noise ratio of 1. We generated the instrument similar to above, with nonlinear confounding with two pre-treatment covariats. The error structure was induced so endogeneity bias would be severe.

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.

We focus on two ways to evaluate our results. First, we examine the RMSE on the first and second stage. The RMSE in the second stage is particularly important because this evaluates the extent to which we are capturing the LICE. Results for the RMSE in the first stage are still helpful to show, though they are analogous to the insights provided in Section E.

Figure 10 plots the ratio of the proposed method's RMSE to the RMSE of 2SLS. Values less than one indicate better performance for the proposed method.[23] In all settings except for the linear model, and for both the first and the second stage, the method of direct estimation returns a superior RMSE. As the sample size increases in the linear model, MDE quickly catches up.

[23]Figures 16 and 17 in Appendix H plot the actual RMSE's for both methods.

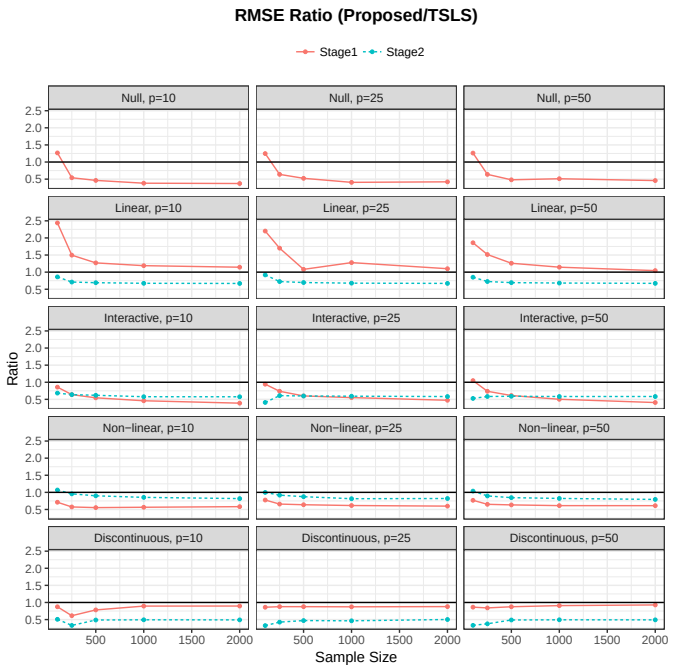

Figure 10: **Ratio of RMSE estimates for proposed method versus 2SLS for first stage and second stage estimation across simulation environments.**

Compliance Estimates Figure 11 examines how well we estimate the correct sign for first stage. In our simulations we know if someone was positively encouraged by the instrument, negatively encouraged, and not encouraged (zero).[24] We then recorded the model's individual level estimates and calculate the percentage of observations correctly classified into each bin. We make several observations on the figure. First, where there are in-truth no observations encouraged, as in the first row, the proposed threshold does indeed return this value. In the second row, the first stage estimate is an additive linear model in which all observations comply positively, and again, the threshold separates compliers from others even at a modest sample size. In the interactive setting the performance is increasing in sample size for identifying both positive and negative compliers. In the non-linear setting our performance for positive observations increases in sample but the proposed method performs poorly for the other types. Finally in the discontinuous setting positive and negative observations are increasingly better identified. Zero types are estimated with high accuracy, though there is a slight degradation as the sample increases.

We also recorded compliance estimates for TSLS. TSLS can only return one compliance estimate

[24]Figure 12 in Appendix H gives the true proportions in the simulated data.

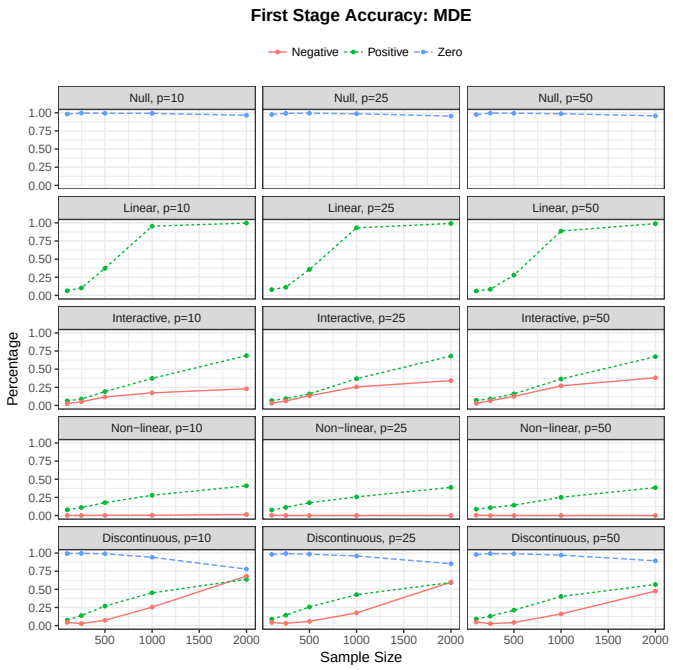

Figure 11: **First stage accuracy for proposed method.** Blue line plots the percentage of in fact positive LICE that we correctly capture. The red line plots the percentage of in fact negative LICE that we correctly capture. The green line plots the percentage of in fact non-identified units that we correctly capture. In setting 1 no one is identified. In our simulations there are no observations with in truth 0 LICE and so the purple line for this case is not present.

for each observation, and we recorded all observations as not complying if the first-stage F statistic on the instrument was less than 10. Results are in Figure 16 of Appendix H. When TSLS did recover compliance percentages correctly, it was only for positively encouraged units. Given the large number of other types in the simulations this performance is not desirable.

F.1.1 True Compliance Rates by Simulation Setting

Figure 12 presents the true percentage for the first stage compliance status. This shows how in setting 1 no observations were encouraged. Settings 2-5 mix the ratio in different ways across positive, negative, and in truth 0.

Figure 12 presents the true percentage each sign for the instrumented causal effect of observations. This shows how in setting 1 all observations do not have an identified sign in the second stage. Settings 2-5 mix the ratio in different ways across positive, negative, in truth 0, and unidentified.

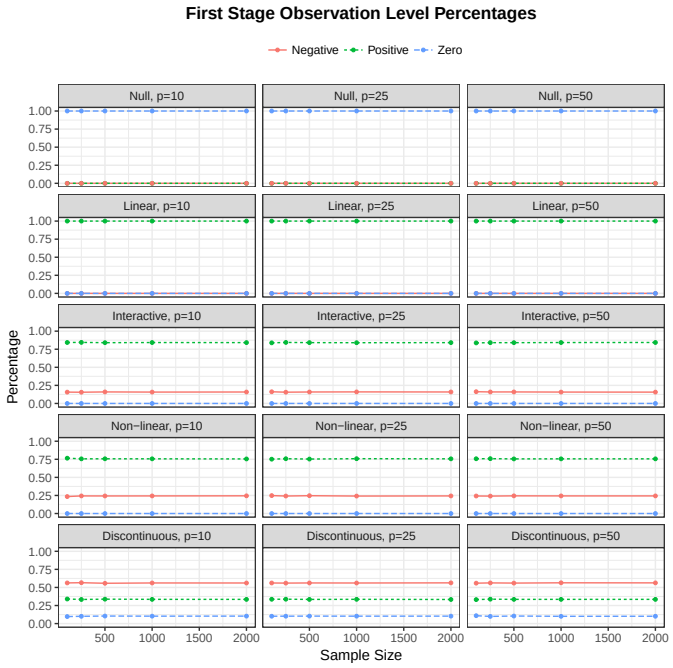

Figure 12: **Percentage of observations in sample by compliance status.**

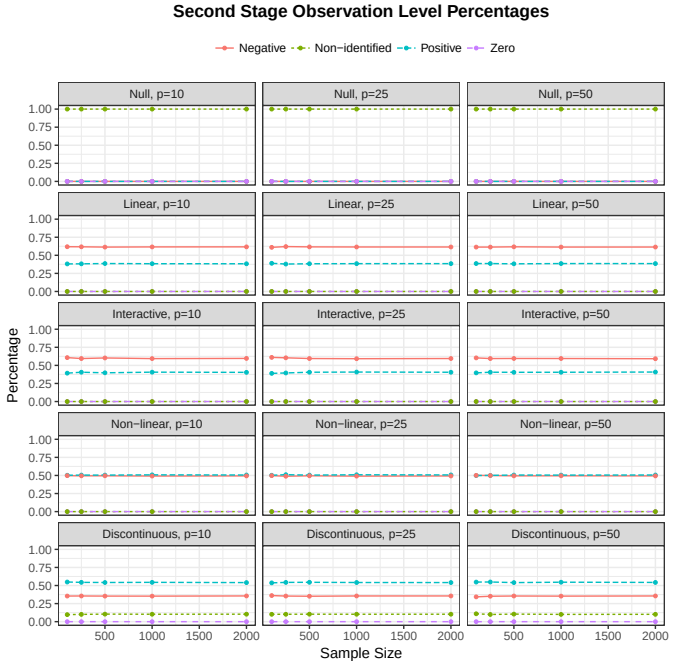

Figure 13: **Percentage of observations in sample by category of sign on instrumented causal effect.**

F.1.2 RMSE

Figure 16 presents RMSE estimates for the first stage model. Consistent with the previous simulation results, the proposed method performs well across a range of settings. The standard least

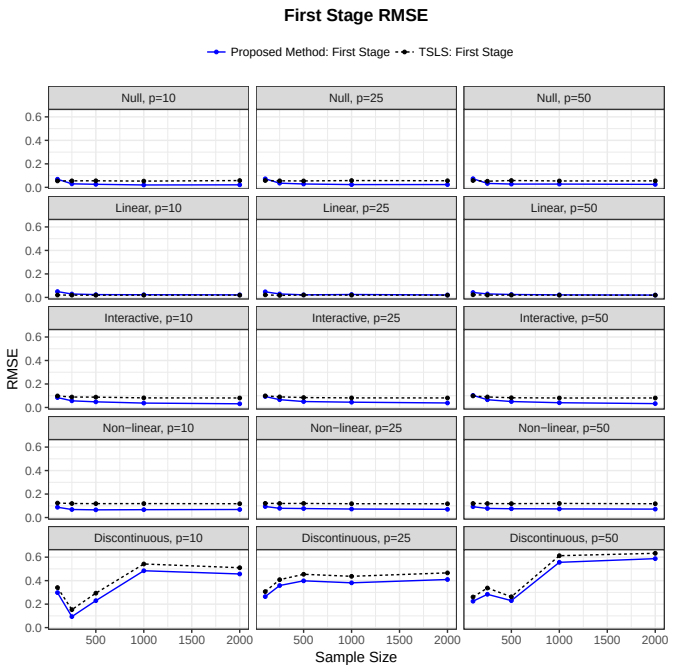

Figure 14: **RMSE estimates for proposed method and 2SLS for first stage estimation across simulation environments.**

squares model, which is misspecified, generally does not perform as well, especially in the third simulation environment.

Figure 15 presents the second stage results. We again see improved performance by the proposed method. RMSE's are lower than 2SLS estimates. In the simulation context we consider there are mixtures of individuals compliance types (discussed further below). Not capturing this heterogeneity means our estimate of the LICE is more accurate than the 2SLS.

F.1.3 TSLS Peformance

Next we give compliance estimates for TSLS in the first stage. TSLS can only return one compliance estimate for each observation, and we recorded all observations as not complying if the first-stage F statistic on the instrument was less than 10. Results are in Figure 16. When TSLS did recover compliance percentages correctly, it was only for those with a positive instrumented causal effect. Essentially it records everything as falling in this category. Given the large number of other types in the simulations this performance is not desirable.

We also recorded compliance estimates for TSLS in the second stage. TSLS can only return one compliance estimate for each observation, and we recorded all observations as not complying if the

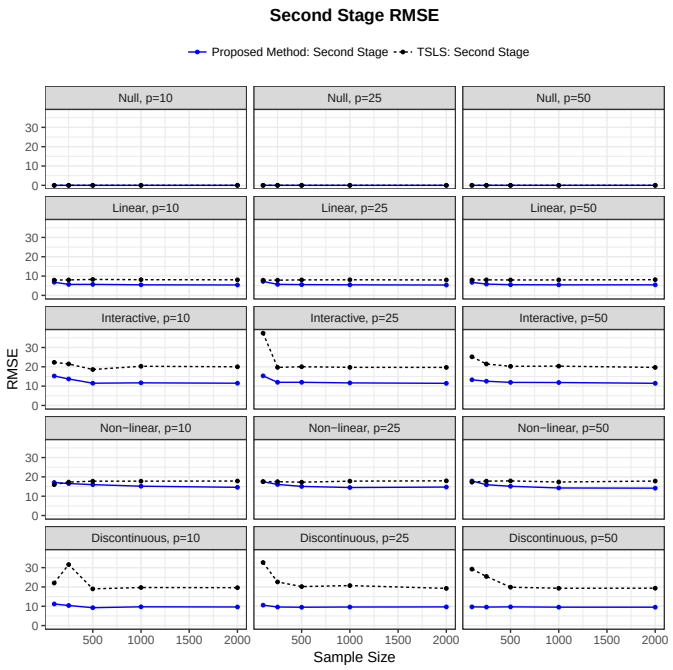

Figure 15: **RMSE estimates for proposed method and 2SLS for second stage estimation across simulation environments.**

first-stage F statistic on the instrument was less than 10. Results are in Figure 15. When TSLS did recover compliance percentages correctly, it was only for those with a positive instrumented causal effect. Given the large number of other types in the simulations this performance is not desirable.

F.2 Diagnostics

Instrumental variables estimation is reliable to the extent that the encouragement of the instrument has a strong effect on the treatment (Bound, Jaeger and Baker, 1995; Staiger and Stock, 1994). The standard measure of this effect in the TSLS case is the F-statistic.[25] Many researcher use a threshold, like 10, to decide whether their instrument is strong enough.

Using the notation in the paper, the first-stage F-statistic for our proposed method is:

$$\widehat{F} = \frac{\frac{1}{\widehat{df}^{\nabla}} \sum_{i=1}^{n} \left\{ R_T^{ZX}(Z_i, X_i)^\top \widehat{c}_T^{ZX} - \widehat{\mu}_T \right\}^2}{\frac{1}{n-\widehat{df}^{\nabla}} \sum_{i=1}^{n} \left\{ T_i - R_T(Z_i, X_i)^\top \widehat{c}_T - \widehat{\mu}_T \right\}^2} \tag{81}$$

where $R_T^{ZX}(Z_i, X_i)$ is the the subvector of $R_T(Z_i, X_i)$ that fluctuates with Z_i and c_T^{ZX} the corresponding parameters.

[25] Though see Kleibergen (2002) for one alternative proposal.

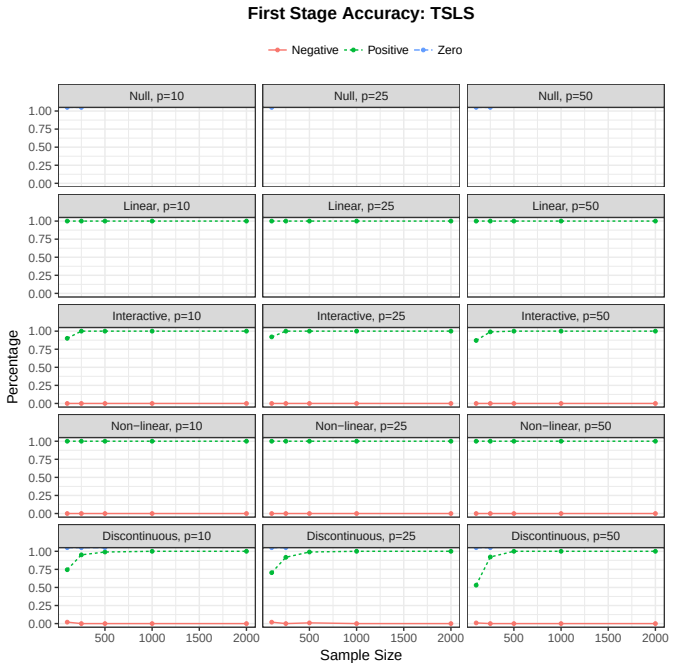

Figure 16: **First stage accuracy for TSLS. Blue line plots the percentage of in fact positive LICE that TSLS correctly captures. The red line plots the percentage of in fact positively encouraged that TSLS correctly captures. The green line plots the percentage of negatively encouraged units that TSLS correctly captures and the blue line captures the percentage of non-encouraged observations correctly identified.**

G Oracle Proofs

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

$$\widehat{c} = \underset{\widetilde{c}}{\operatorname{argmin}} \, ||Y - R\widetilde{c}||_2^2 + \lambda ||W\widetilde{c}||_1^1. \tag{82}$$

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

$$||Y - R\widehat{c}||_2^2 + \lambda ||W\widehat{c}||_1^1 \leq ||Y - Rc||_2^2 + \lambda ||Wc||_1^1. \tag{83}$$

After some manipulation (e.g., Buhlmann and van de Geer, 2013; Ratkovic and Tingley, 2017) and

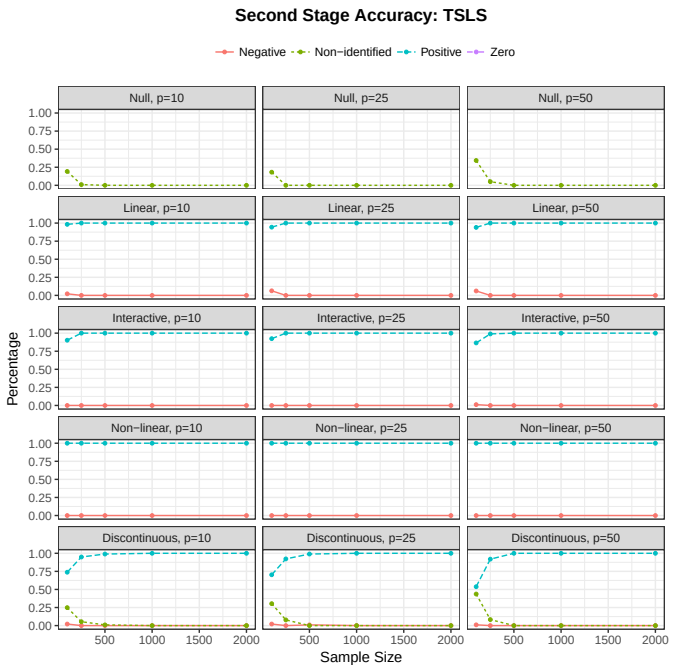

Figure 17: **Second stage accuracy for TSLS. Blue line plots the percentage of in fact positive LICE that TSLS correctly captures. The red line plots the percentage of in fact negative LICE that TSLS correctly captures. The green line plots the percentage of in fact non-identified units that TSLS correctly captures.**

evaluating at $\widehat{\lambda}, \widehat{W}$, this excess risk generates the inequality

$$\frac{1}{n}\left\{||R\widehat{\delta}||_2^2 + \widehat{\lambda}||\widehat{W}\widehat{\delta}||_1^1\right\} \leq C\frac{\widehat{\lambda}^2\widehat{\sigma}^2\overline{\gamma}^2|S|}{n^2\phi_0^2} + C_\infty\frac{||R^o_{\infty/n}c_{\infty/n}||_\infty^2}{n} + C_\perp\frac{||R^o_\perp c_\perp||_\infty^2}{n} \tag{84}$$

which is Equation 19.

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

$$2R_k^\top\widehat{\epsilon} = s_k\widehat{\lambda}\widehat{w}_k \tag{85}$$

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

$$R^{\widehat{S}} = \{R_k : \widehat{c}_k \neq 0\} \tag{86}$$

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

Consider the gradient derivative of equality 85 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

$$0 = \sum_{i=1}^{N} \frac{\partial}{\partial T_i} \sum_{i=1}^{N} \left(R_{ik}^{\widehat{S}} \widehat{\epsilon}_i \right) \Rightarrow \tag{87}$$

$$0 = \sum_{i=1}^{N} \frac{\partial}{\partial T_i} R_{ik}^{\widehat{S}} \widehat{\epsilon}_i \tag{88}$$

$$= \sum_{i=1}^{N} \frac{\partial R_{ik}^{\widehat{S}}}{\partial T_i} \widehat{\epsilon}_i + R_{ik}^{\widehat{S}} \frac{\partial \widehat{\epsilon}_i}{\partial T_i} \tag{89}$$

This gives us

$$\left| \sum_{i=1}^{N} \frac{\partial R_{ik}^{\widehat{S}}}{\partial T_i} \widehat{\epsilon}_i \right| = \left| \sum_{i=1}^{n} R_{ik}^{\widehat{S}} \frac{\partial \widehat{\epsilon}_i}{\partial T_i} \right| \tag{90}$$

which is a solution to the following LASSO problem

$$\widehat{c}^{S,T} = \underset{c^{S,T}}{\operatorname{argmin}} \, ||\widetilde{\Delta Y} - \Delta R^{\widehat{S},T} c^{\widehat{S},T}||_2^2 + ||\widetilde{W}^{\widehat{S},T} c^{\widehat{S},T}||_1^1 \tag{91}$$

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

$$\widetilde{\Delta Y} = \Delta R^{\widehat{S},T} \widehat{c}^{\widehat{S},T} + \widehat{\epsilon}. \tag{92}$$

The model is fit using basis-specific tuning parameter,

$$\widetilde{w}_k^{\widehat{S},T} = \left| \sum_{i=1}^{n} R_{ik}^{\widehat{S}} \frac{\partial \widehat{\epsilon}_i}{\partial T_i} \right| \tag{93}$$

$$= \left| R_k^{\widehat{S},\top} \Delta R^{TX} (c^T - \widehat{c}^T) \right| \tag{94}$$

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

Then,

$$\left| R_k^{\widehat{S},\top} \Delta R^{TX} (c^T - \widehat{c}^T) \right| \leq ||R_k^{\widehat{S},\top}||_2 ||\Delta R^{TX} (c^T - \widehat{c}^T)||_2 \tag{95}$$

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 Data Appendix

H.1 Larreguy and Marshall Data

Data Preparation and Transformation The original analysis selected a set of state $\times$ covariate interactions to place into a TSLS model. Rather than select the interactions a priori, we instead create a model of fully saturated interactions and then use the SIS screen to winnow them down. As the data have a hiearchical structure, we include each variable three times: first, the original variable; second, the mean of the variable by state; and third, the state-centered version of this variable. Specifically, assume X_{var} are the matrix of variables and X_{fe} the matrix of state fixed effects. Denote as H_{fe} the hat matrix from the fixed effects and I the commensurate $n \times n$ identity matrix.

The data we place into our software is then

$$X^{big} = [X_{fe} : X_{var} : H_{fe}X_{var} : (I - H_{fe})X_{var}]. \tag{96}$$

All two way interactions and spline transformations of this data are then calculated as normal.

Types of coefficients Figure 18 plots estimates for several different types of variables that get included in our model using data from Larreguy and Marshall (2017). The top left plots all coefficients that came out of the Sure Independence Screen. The top right plots coefficients on the linear terms, the bottom left on terms that had an interaction, but no non-linear basis function. Finally, the bottom right plots coefficients on variables constructed via an interaction term with a non-linear basis function. We provide Figure 18 for descriptive purposes. Analysts could use this information, for example, to explore sources of heterogeneity in the LICE.

LaLonde Predictive Performance

Throughout this study, we have focused not just on sample-average estimates but individual-level effect estimation. We next compare several methods on their ability to predict a held-out subsample. Results are presented in Table 4. In the first comparison (columns 2-3), we fit a model to the experimental data and then use this model to predict outcomes in the observational data. As an experimental sample may not resemble a target population of interest, this prediction exercise indicates the extent to which methods can generalize from an experiment to observational data. In the second comparison (columns 4-5), we fit a model to the experimental treated and observational

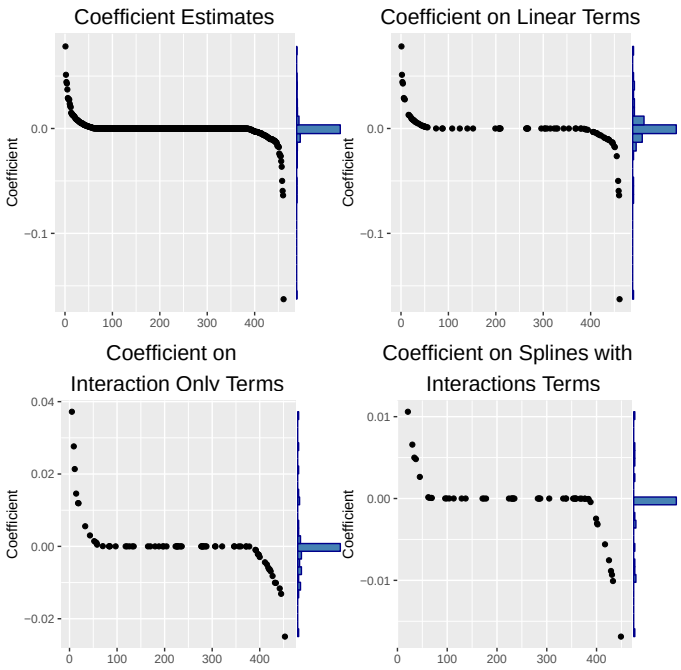

Figure 18: **Coefficient Estimates by Effect Type**

Outcome	Observational Untreated		Experimental Untreated	
	Bias	RMSE	Bias	RMSE
OLS, in-sample	0.00	10173.27	0.00	5433.39
MDE	-15054.61	21267.99	428.78	7040.68
MDE, bagged	-255.94	10804.69	660.84	6865.75
Horseshoe	-16104.93	22371.13	14733.41	15798.17
BART	-11049.12	17046.63	1125.95	6896.35
POLYMARS	-13035.59	18957.33	996.27	6646.84
SuperLearner	-13407.01	19474.82	2057.12	6684.54
LASSO	-12942.43	19186.10	1689.17	6751.50

Table 4: **Prediction Exercise, LaLonde Data.** Columns contain the results from predicting the outcome on held-out subsets of the LaLonde data (top row). We fit a model to the experimental data and then use this model to predict outcomes in the observational data (columns 2-3). Next, we fit a model to the experimental treated and observational untreated, then use this model to predict the held-out experimental group (columns 4-5). In each case, we include the result from least-squares fit to the held-out sample as a baseline. We see that the bagged MDE estimate performs well in both settings, in terms of both bias and RMSE. In the second setting, POLYMARS achieves the lowest RMSE, but at the cost of a large bias.

untreated, then use this model to predict the held-out experimental group. In each case, we include the result from least-squares fit directly to the held-out sample as a baseline.

We see again that bagged MDE performs well in both exercises, particularly the first, where it achieves a predictive RMSE nearly that of OLS fit to the held-out data. In the first simulation, MDE

performs less well, achieving a higher RMSE than all but the horseshoe. In the second analysis, predicting the experimental untreated, bagged MDE achieves a RMSE and bias lower than BART. MDE achieves the lowest bias, but with a somewhat higher RMSE than the remaining methods except for the horseshoe. POLYMARS and the SuperLearner both achieve a lower RMSE than MDE and bagged MDE in the second analysis, but at the cost of a higher variance.