Causal Inference through the Method of Direct Estimation*

Marc Ratkovic† Dustin Tingley‡

February 22, 2017

Abstract

The intersection of causal inference and machine learning is a rapidly advancing field. We propose a new approach, the method of direct estimation, that draws on both traditions in order to obtain nonparametric estimates of treatment effects. The approach focuses on estimating the effect of fluctuations in a treatment variable on an outcome. A tensor-spline implementation enables rich interactions between functional bases allowing for the capture treatment/covariate interactions. We show how new innovations in Bayesian sparse modeling readily handle the proposed framework, and then document its performance in simulation and applied examples. Furthermore we show how the method of direct estimation can easily extend to structural estimators commonly used in a variety of disciplines, like instrumental variables and mediation analysis.

Key Words: causal inference, treatment effects, instrumental variables, mediation, Bayesian LASSO, sure independence screening

*We would like to thank Horacio Larreguy for providing replication data. We would also like to thank Chris Lucas, Peter Aronow, Shiro Kuriwaki, Rich Nielsen, and Aaron Strauss for comments. Not for citation or distribution without permission from the authors.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://www.princeton.edu/∼ratkovic

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

Introduction

Recent work on drawing causal inference from observational and experimental data has emphasized two lessons. First, causal effects are defined at the observation level. Second, causal effects are defined in terms of ceteris paribus manipulations of a treatment. As in most research designs we do not observe the outcome under counterfactual manipulations, a key observation-level quantity remains unobservable (Imbens and Rubin, 2015; Rubin, 2005; Holland, 1986).

Despite these insights, causal estimation has focused on estimating structural parameters that are aggregates of these observation-level effects. For example, inverse probability weights or sub-classification can reduce confounding bias, but only an average effect for the sample is recovered (Rosenbaum and Rubin, 1983; Ho et al., 2007; Iacus, King and Porro, 2011; Diamond and Sekhon, 2012; Imai and Van Dyk, 2012; Hirano and Imbens, 2005). Estimating only an average effect ignores potentially informative underlying heterogeneity. The shortcoming grows more pernicious in more complex designs. In an instrumental variables analysis, for example, a causal effect is estimated but only for the unknown subset actually encouraged by the design (e.g. Angrist, Imbens and Rubin, 1996). Uncertainty over the compliant subset, as well as unverifiable behavioral assumptions that no observation acts contrary to the encouragement, limits the method's internal and external validity (e.g. Deaton, 2010). Additional models for the treatment mechanism, inverse probability weights, and compliance status may help alleviate these problems (e.g Aronow and Carnegie, 2013; Heckman and Vytlacil, 2007*a*,*b*; Abadie, 2003; Hirano et al., 2000) but the models then introduce another layer of modeling assumptions and uncertainty.

Rather than target low-dimensional population-level parameters that correspond with an aggregate of individual effects, we propose a method whereby we directly estimate observation-level quantities of interest. We call this the *method of direct estimation*. The estimates can be analyzed for heterogeneities or aggregated into coarser causal estimands, such as an average effect. In more complex designs, the method offers deeper insight. For example, in an encouragement design, the standard estimate is the Wald estimate, a ratio of two difference-in-means. The sample differences in the numerator and denominator average over both those encouraged and those not encouraged. Instead of a ratio of means, our estimate is a mean of observation-level ratios, allowing the researcher to estimate the subset of the sample actually encouraged.

The method extends directly to general treatment regimes. In the case of a categorical or

binary treatment, we estimate as the difference between the outcome under treatment and under some baseline level. In the case of a continuous treatment, we estimate the partial derivative of the outcome with respect to the treatment, which is simply the analog of the sub-differential used with a the difference in the treatment levels taken to zero. We focus on the impact of a small "perturbation" as this measures the local marginal impact of the treatment on the outcome (e.g., Hardle and Stoker, 1989) and predicting near the observed data allays concerns about lack of common support and the dangers of extreme counterfactuals (King and Zeng, 2006).

A key goal of this project is to connect machine learning and causal inference at a theoretical level. Rather than turn to a generic machine learning method to estimate causal estimands, we use our identification assumptions to drive our modeling choices. For example, we use the ignorability assumption from causal inference to motivate our choice of nonparametric basis. We show that under this assumption, the partial derivative (or subderivative) of the outcome with respect to the treatment is only a function of pre-treatment covariates. This suggests modeling the treatment in terms of a nonparametric basis that is locally linear, i.e. a degree 2 truncated-power basis spline or order 2 B-spline (e.g. Hastie, Tibshirani and Friedman, 2010, ch. 5). We are using the model to adjust for confounding, which is induced by a systematic interaction between the treatment and a prognostic variable. We turn to a tensor-spline basis to account for these interactions. As the covariates may also be interacting with each other, we fit a model with bases that capture *treatment* $\times$ *covariate* $\times$ *covariate* interactions, as well as all lower-order terms.

As a second connection, we use the theory of LASSO estimation to derive a bound on the size of expected prediction error (Buhlmann and van de Geer, 2013). We fit the model using a sparse Bayesian method (Ratkovic and Tingley, 2017), but we use this oracle inequality to drive our prior specification. We show that doing so generates predictions notably better than existing sparse Bayesian methods (Carvalho, Polson and Scott, 2010; Polson, Scott and Windle, 2014). We extend the theory, deriving an oracle bound not only for the prediction error but also for the causal effect. We show that solving a LASSO problem also bounds the error rate on the treatment effect.

Third, we then use this oracle bound constructively rather than simply descriptively. We construct a feasible estimate for this bound, and then we show that the bound is a reliable means of differentiating compliers, non-compliers, and defiers in an instrumental variable analysis. Doing so increases the internal validity of an instrumental variable analysis, as it helps identify off which

observations a causal effect is identified. It also aids external validity, helping to establish what subgroups can be extrapolated to.

Our estimation strategy incorporates two innovations. First is the construction of a nonparametric basis spanning a larger model space than earlier work on multivariate spline models (Stone et al., 1997; Friedman, 1991), as we include bases of multiple degrees and knot locations. As the number of bases places us in an "ultrahigh" dimensional setting, we use a covariate screening method (Fan and Lv, 2008) to return a feasible subset of bases. Second, as described above, we use a sparse Bayesian regression to estimate the model (Ratkovic and Tingley, 2017). The regression model incorporates frequentist Oracle results to motivate hyperprior parameters. The model also includes endogenously estimated adaptive weights (as in Zou, 2006) and a decay parameter that adapts these weights off the estimated global level of sparsity in the model (similar to the nonseparable priors in Bhattacharya et al., 2015; Rockova and George, Forthcoming).

Every method comes with shortcomings. Estimating the outcome model without also modeling the treatment mechanism results in inefficient estimates (Robins and Ritov, 1997). Yet, uncertainties in the treatment model may affect the estimates in nonlinear and unpredictable ways (Kang and Schafer, 2007). We place a greater emphasis on model selection than efficiency. We also note that our model requires assumptions on the differentiability' of the conditional mean function. In our framework, these assumptions are unavoidable, but we work with a dense and varied set of discontinuous bases to accommodate a wide class of functional forms. Our simulations include data generated from models both in and out of our model space, and we find the proposed method works well.

We illustrate the method through simulation studies and two applied examples. We illustrate in a simulation study how our proposed method generally outperforms a variety of cutting-edge machine learning and high-dimensional methods in predicting both fitted values and the marginal responsiveness of the outcome to the treatment. We then illustrate the method using a binary treatment and then continuous treatment. For the binary treatment, we turn to a benchmark dataset in the matching literature but comes from important work in economics (LaLonde, 1986). We show that the method of direct estimation outperforms existing measures in recovering an average treatment effect from an observational dataset. As we focus on individual-level estimates, we also show that our method performs well in recovering individual-level estimates for a held-out

subsample of the data. We next turn to applying the method to an instrumental variables analysis. We show how the method relates to earlier approaches to instrumental variables that focus on structural parameters, evaluate performance of the method against two-stage least squares, and apply the method to recent work on the relationship between education and political participation (Larreguy and Marshall, 2017). We show how the proposed method uncovers individual compliance status as well as an understanding of how causal effects vary with the levels of the instrument. Finally, we analytically show how the method extends to mediation analysis (Baron and Kenny, 1986; VanderWeele, 2015).

The paper proceeds as follows. Section 1 introduces the proposed framework and estimation strategy. Section 2 presents a series of simulations comparing our proposed method of direct estimation to a range of cutting edge machine learning tools. Section 3 shows how the method of direct estimation readily extends to structural models like instrumental variables and mediation. Section 4 presents several illustrative empirical applications. Section 5 concludes by showing how our approach helps to connects disparate literatures and by discussing limitations and future extensions.

1 The Proposed Framework

We next introduce the proposed method, the method of direct estimation (MDE). We first introduce notation for the potential and observed data and then introduce our estimand. Second, we introduce a sparse Bayesian nonparametric regression and discuss estimation and inference. Finally we discuss the relationship of the proposed method to earlier work.

1.1 The Setup

Assume a simple random sample where observation $i \in \{1, 2, \ldots, n\}$ possesses a potential outcome function, $Y_i(\widetilde{T}_i; X_i)$ that maps the treatment level to the outcome. Notationally, we place the random, or manipulable, elements before the semi-colon and the background conditioning effects after. We denote as $\widetilde{T}_i$ a random variable with law F_{X_i}, such that the distribution of $\widetilde{T}_i$ may vary with X_i. We assume X_i is a vector of p individual-level fixed covariates that are causally prior to and not a function of $\widetilde{T}_i$. X_i may contain confounders, moderators, prognostic variables, risk factors, or simply superfluous variables; the important part is they are causally prior to the treatment and outcome. There may be more covariates than observations, i.e., it may be that $p > n$. We consider the case where the distribution F_{X_i} is neither controlled by or known to the researcher. Given T_i a

realization of $\widetilde{T}_i$, we observe $\left\{Y_i, T_i, X_i^\top\right\}^\top$ where $Y_i = Y_i\left(T_i; X_i\right) + \epsilon_i$ with ϵ_i mean-zero, equivariant noise, independent of both $\widetilde{T}_i$ and X_i.

We will denote the individual-level effect as

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

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

in the potential outcomes notation (top, (Rubin, 1974, 2005)) or using Pearl's *do*-notation (below; Pearl, 2000). Both are equivalent formulations of the same idea: the observation-level impact of a fluctuation or manipulation in $\widetilde{T}_i$ of size δ_i on the potential outcomes.

For identification, we make four assumptions. First, we assume X_i is sufficiently rich to render the treatment assignment ignorable and that the counterfactual value is possible. We also assume non-interference among units and there is only one version of each treatment level. We give these assumptions below.

ASSUMPTION 1 **Identification assumptions for observation-level treatment effect**

Define $\widetilde{T}_i^{\delta_i} = \widetilde{T}_i | \widetilde{T}_i \in \{T_i, T_i + \delta_i\}$. *Our treatment effect is identified by the following assumptions*

1. *Unconfoundedness:* $Y_i(\widetilde{T}_i^{\delta_i}; X_i) \perp\!\!\!\perp \widetilde{T}_i^{\delta_i} | X_i$
2. *Probabilistic treatment:* $1 > \Pr(\widetilde{T}_i^{\delta_i} = T_i | X_i) > 0$
3. *Stable unit:* $Y_i(\widetilde{T}_i^{\delta_i}; X_i) = Y_i(\widetilde{T}_i^{\delta_i}; X_i, \widetilde{T}_{i'}, Y_{i'}(\widetilde{T}_{i'}; X_{i'})) \forall i \neq i'$
4. *Stable treatment value:* $Y_i(\widetilde{T}_i^{\delta_i}; X_i) = Y_i(\widetilde{T}_i^{\delta_i'}; X_i) \forall T_i^{\delta_i'} = T_i^{\delta_i}$

If T_i has continuous support, we can consider taking limits in δ_i. For example, we can take $\lim_{\delta_i \to 0} \nabla_{T_i}(\delta_i)$ for supp $(\widetilde{T}_i) = \Re$. In cases where the treatment variable has a lower bound, such as a monetary expenditure or time spent on some task, we can consider $\lim_{\delta_i \to 0^+} \nabla_{T_i}(\delta_i)$ when T_i is at its lower bound, and for categorical supp $(\widetilde{T}_i) \in \{0, 1, 2, \ldots, K\}$, we can consider the counterfactual manipulations $\nabla_{T_i}(1 \times \mathbf{1}(T_i \neq K))$ or $\nabla_{T_i}(-1 \times \mathbf{1}(T_i \neq 0))$.

1.2 Operationalization

We reformulate the problem of treatment effect estimation as one of estimating a counterfactual functional via modeling the conditional mean. The basic problem is now our ignorance of a func-

tional form connecting the observed outcome to the treatment value and covariates. We address this concern through the use of a high-dimensional, nonparametric regression model.

Our approach proceeds in four steps. First, we construct an extremely rich series expansion around the treatment variable using tensor-product nonparametric bases. This generates hundreds of thousands, if not millions, of possible bases. To reduce this "ultrahigh" number of bases to hundreds or thousands, we implement the Sure Independence Screen (SIS) of Fan and Lv (2008). Third, we use the LASSOPlus, an extension of the Bayesian Lasso, to obtain regularized estimates on the candidates surviving the screen (Ratkovic and Tingley, 2017). Finally, if we see skew in the outcome or residuals, we use bootstrap aggregation (Buhlmann and Yu, 2002).

We assume the data are generated as

$$Y_i = \mu_Y + R^o(T_i, X_i)^\top c^o + \epsilon_i \tag{3}$$

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. 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 our in the data-generation process but not in our considered model space, $R_\perp(T_i, X_i)$. Each will later be crucial components of our bound analysis.

In constructing the set of bases that we will use 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 $|\mathcal{K}_k|$ nonparametric bases for variable X_k evaluated for observation i at the knots in $\mathcal{K}_k$. 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 evaluated at knots in $\mathcal{K}_T$. We construct the mean vector as

$$R_\infty(T_i, X_i) = \{1, b_T(T_i, \kappa_T)\}_{\kappa_T \in \mathcal{K}_T} \otimes \{1, b(X_{ik}, \kappa_k)\}_{\kappa_k \in \mathcal{K}_k} \otimes \{1, b(X_{ik'}, \kappa_{k'})\}_{\kappa_{k'} \in \mathcal{K}_{k'}} \tag{4}$$

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.

Nonparametric basis construction for causal estimation. Our causal identification assumptions on $\nabla_{T_i}(\delta_i)$ can inform our parameterization of the treatment structure in our outcome model, as given in the following proposition.

PROPOSITION 1 *Assume the identification assumptions (1)-(4) hold. For $\widetilde{T}_i$ with support on a compact, open region of $\mathcal{R}$ containing T_i, the partial derivative of the potential outcome function at T_i with respect to $\widetilde{T}_i$ is only a function of X_i. For $\widetilde{T}_i$ binary, the individual causal effect is only a function of X_i.*

Proof. *For δ_i in the compact, open region, assumptions (2)-(4) ensure the counterfactual exists, is of the form $Y_i(T_i; X_i)$, and is well-defined, respectively. Unconfoundedness gives $Y_i(T_i + \delta_i; X_i), Y_i(T_i; X_i) \perp\!\!\!\perp \widetilde{T}_i | X_i \Rightarrow (Y_i(T_i+\delta_i) - Y_i(T_i))/\delta_i \perp\!\!\!\perp \widetilde{T}_i | X_i = 0$. Taking δ_i to zero gives $\lim_{\delta_i \to 0} \nabla_{T_i}(\delta_i)$ is only a function of X_i. For a binary treatment, assumptions (2)-(4) ensure the same properties. Unconfoundedness gives $(Y_i(1; X_i) - Y_i(0; X_i))/(1 - 0)$ is only a function of X_i.*

This proposition implies then that at each observed value the outcome is locally linear in T_i and an arbitrary function of X_i. We therefore model the treatment using a degree 2 thresholded power-basis and 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 the local linearity implied by our proposition, but the proposition will not be satisfied by 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. Specifically, for $k \in \{3, 5, 7, 9\}$, 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, generating $24 = 3 + 5 + 7 + 9$ bases.

We combine the treatment and covariate bases using the tensor product as given above. At our default, we take 28 bases for the treatment and 25 for each covariate. This gives us $(1 + 28) \times (1 + 24 \times p) \times (1 + 24 \times p)$ total basis. Even a modest p we end up with a large number of bases; $p = 10$ gives over $1,680,000$ bases and $p = 20$ over $6,700,000$. This places us in an "ultrahigh" dimensional setting so we implement a Sure Independence Screen (SIS) (Fan and Lv, 2008).[1] The

[1]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.

SIS occurs in two steps. First, we construct all of the tensor product bases in Equation 4. 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 with the outcome.[2]

Our method for constructing and maintaining bases differs from, and improves on, earlier work. Early nonparametric regression models worked through local stepwise selection and deletion of bases (e.g., Friedman, 1991; Stone et al., 1997), and implementation in a causal setting would assume the right bases were known *a priori* (e.g., Newey and Powell, 2003). We improve over these methods by exploring a larger model space, allowing both degree and knot placement to differ. The screen allows us to move from millions of bases to several hundred or thousand, and then we turn to a second stage sparse model to choose amongst these. We also find, in our simulations and applied example below, that our screening+selection approach often outperforms ensemble methods and tree-based methods.

We denote this screened basis vector as $R(T_i, X_i)$, and a counterfactual basis vector as $R(T_i + \delta_i, X_i)$. Given fitted values $\widehat{c}$, we can write

$$\widehat{Y}_i = \widehat{\mu}_Y + R(T_i, X_i)^\top \widehat{c} \tag{5}$$

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

$$\widehat{\nabla}_{T_i}(\delta_i) = \frac{1}{\delta_i} \left\{ \left(R(T_i + \delta_i, X_i) - R(T_i, X_i) \right)^\top \widehat{c} \right\} \tag{6}$$

and we approximate the derivative by choosing δ_i close to zero; we take $\delta_i = 10^{-5}$ in practice. The sample average marginal causal effect can be found through averaging over the sample, $\frac{1}{N} \sum_{i=1}^{N} \widehat{\nabla}_{T_i}(\delta_i)$. We turn next to generating the estimate, $\widehat{c}$.

1.3 A Sparse Bayesian Prior

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

[2]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`.

effects by building on LASSOplus, a sparse Bayesian prior we implemented in earlier work (Ratkovic and Tingley, 2017).

The LASSOplus hierarchy was constructed with several goals in mind. First, we want the estimates to satisfy an oracle bound on the prediction error. Second, we want inverse weights that regularize large effects less aggressively than small effects. Third, we want a global sparsity parameter that helps control the overall level of shrinkage in the model.

The hierarchy is given as

$$Y_i|R(T_i,X_i),c,\sigma^2 \sim \mathcal{N}(R(T_i,X_i)^\top c,\sigma^2) \tag{7}$$

$$c_k|\lambda,w_k,\sigma \sim DE\left(\lambda w_k/\sigma\right) \tag{8}$$

$$\lambda^2|n,p,\rho \sim \Gamma\left(n\times(\{log(n)+2\log(p))-p,\rho\right) \tag{9}$$

$$w_k|\gamma \sim \text{generalizedGamma}(1,1,\gamma) \tag{10}$$

$$\gamma \sim \exp(1) \tag{11}$$

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 B. The prior on λ^2 was constructed to generate to properties for the tuning parameter. First, the prior 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 rate at which the oracle bound holds is controlled by $1-exp\{-\sqrt{\log(n)}\}$, which approaches 1 in the limit. The prior is proper whenever $n\times(\log(n)+2\log(p))-p>0$. For example, with $n\in\{100;1,000;10,000\}$, this requires $p<1,978;27,339;347,529$ respectively, which is sower than the LASSO rate of $n\sim\log(p)$, but clearly allows for $p>n$. This slower growth rate does suggest the utility of the an initial screen for covariates. Last, even with an improper prior, the posterior will be proper, so estimation can still proceed.

The prior weights w_k are constructed so that the MAP estimate is similar to the adaptive LASSO

of Zou (2006). Each mean parameter, c_k, will have its own adaptive penalty $\frac{\widehat{\lambda}\widehat{w}_k}{n\widehat{\sigma}}$.The estimated weights $\widehat{w}_k$ are inversely related to the magnitude of the parameter estimates $\widehat{c}_k$. We prove that the adaptive penalty term is of order $\sqrt{\log(n)/n}$ when $\widehat{c}_k \to 0$, which is not asymptotically negligble. On the other hand, the adaptive penalty is of order $1/n$ when $\widehat{c}_k$ grows in magnitude, and hence is asymptotically negligible.

Third, our global sparsity 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. 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*. Details are provided in Appendix B.

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

$$\begin{aligned} &\frac{1}{n}\left\{||R(T., X.)\widehat{\delta}||_2^2 + \lambda||\sum_{k=1}^{p}\widehat{w}_k\widehat{\delta}_k||_1^1\right\} \leq \\ &\quad C\frac{\widehat{\lambda}^2\widehat{\sigma}^2\overline{\gamma}^2|S|}{n^2\phi_0^2} + C_\infty\frac{||R^o_{\infty/n}(T., X.)c_{\infty/n}||_\infty^2}{n} + C_\perp\frac{||R^o_\perp(T., X.)c_\perp||_\infty^2}{n} \end{aligned} \tag{12}$$

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 C 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 sufficiently dense set of bases.

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 basis at least one element covaries with the treatment:

$$R(T_\cdot, X_\cdot) = [R^X(X_\cdot) : R^{TX}(T_\cdot, X_\cdot)]. \tag{13}$$

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 $\Delta R(T_\cdot, X_\cdot) = [\Delta R_{ij}] = [\partial R_{ij}/\partial T_i]$ the elementwise partial derivatve of $R(T_\cdot, X_\cdot)$ with respect to the treatment and note $\widehat{\nabla}_{T_i}(\delta_i) = \Delta R_i(T_i, X_i)^\top \widehat{c}^{TX} = \Delta R_i^{TX}(T_i, X_i)^\top \widehat{c}^{TX}$. Denote as $\widehat{c}^{\widehat{S},TX}$ the subvector of $\widehat{c}^{TX}$ selected in the outcome model.

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

$$\widehat{c}^{S,TX} = \underset{c^{\widehat{S},TX}}{\operatorname{argmin}} \left|\left|\widetilde{\Delta Y} - \Delta R^{\widehat{S},TX} c^{\widehat{S},TX}\right|\right|_2^2 + \left|\left|\widetilde{W}^{\widehat{S},TX} c^{\widehat{S},TX}\right|\right| \tag{14}$$

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

$$\widetilde{\Delta Y_i} = \Delta R_i(T_i, X_i)^\top \widehat{c} + \widehat{\epsilon}_i. \tag{15}$$

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 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

$$\begin{aligned} &\frac{1}{n}\left\{||\Delta R(T_i, X_i)^{\widehat{S},TX}\widehat{\delta}^{\widehat{S},TX}||_2^2 + ||\widetilde{W}^{\widehat{S},TX}\widehat{\delta}||_1^1\right\} \leq \\ &C_\Delta \frac{\widehat{\sigma}^2\overline{\gamma}_\Delta^2|S_\Delta|}{N^2\phi_\Delta^2} + C_\infty \frac{||\Delta R^o_{\infty/\widehat{S}}(T_i, X_i)c_{\infty/n}||_\infty^2}{n} + C_\perp \frac{||\Delta R^o_\perp(T_i, X_i)c_\perp||_\infty^2}{n} \end{aligned} \tag{16}$$

where all the constants in the inequality are analagous to those in Inequality 16 but defined in terms of the design matrix that parameterize the derivative.

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.

Bagging. We have several reasons to be concerned that our estimates may have a large sampling variance. The LASSO problem minimized in Equation 14 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 heteroskedasiticity (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.

1.4 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 the 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*a*), 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 screened 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 also

utilize bagging in order to guard against erratic results attributable to skewed or fat-tailed sampling distributions (Buhlmann and Yu, 2002).

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 combinatorically 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*b*; 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*a*) 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 cubic-spline basis of order one.

Relation to nonparametric and high-dimensional instrumental variable methods. 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. Our work is perhaps closest to Heckman and Vytlacil (2007*a*,*b*, 2005, 1999), Athey, Tibshirani and Wager (2016), and Hartford et al. (2016). Like the Heckman and Vytlacil works, we emphasize observation-level estimates that can reveal heterogeneities and be assembled into substantively meaningful estimates. These works involve a behavioral foundation, and the authors advocate averaging over the entire choice set, whereas we focus on the local responsiveness at the observed treatment level. As well, we focus on estimation and implementation rather than just identification. 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 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*a*). 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 below. 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 implement a bagging procedure to smooth over a possibly erratic sampling distribution.

2 Simulation Evidence

We next present simulation evidence illustrating the proposed method's utility. 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:

$$\text{Linear:} \quad Y_i = T_i + \sum_{k=5}^{8} X_{ik}\beta_k + \epsilon_i \tag{17}$$

$$\text{Interactive:} \quad Y_i = T_i - T_i \times X_{i3} + \sum_{k=5}^{8} X_{ik}\beta_k + X_{i1}X_{i2} + \epsilon_i \tag{18}$$

$$\begin{aligned}\text{Nonlinear:} \quad Y_i = 20 \sin\left(\left(X_{i1} - \frac{1}{2}\right)\frac{T_i}{20}\right) + \frac{1}{2} \times (1 + |X_{i3} \times X_{i4}|_+) \times (1 + T_i) + \\ \sum_{k=5}^{8} X_{ik}\beta_k + \epsilon_i\end{aligned} \tag{19}$$

$$\text{Discontinuous } Y_i = (1 + |T_i|) \times \mathbf{1}(|T_i| > 1/2) \times (X_{i5} + 1) + \sum_{k=5}^{8} X_{ik}\beta_k + X_{i1}X_{i2} + \epsilon_i \tag{20}$$

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

$$T_i = -3 + (X_{i1} + X_{i4})^2 + \epsilon_i^T; \quad \epsilon_i^T \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 4) \tag{21}$$

We consider $n \in \{100, 250, 500, 1000, 2000\}$ and $p \in \{10, 25, 50, 100\}$.

We asses 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. Each of these methods are handed the same nonparametric basis created in our pre-processing step. 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

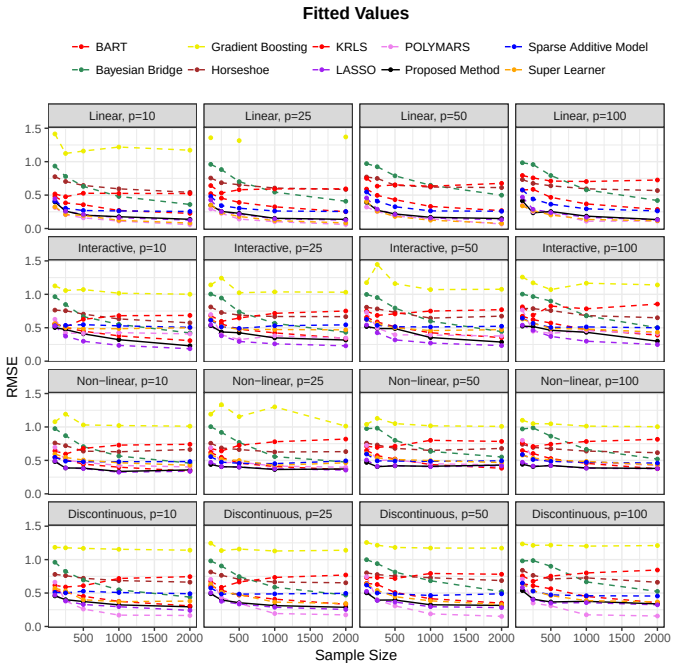

Figure 1: **A Comparison of RMSE for Fitted Values.**

the SuperLearner we include as constituent methods random forests, the LASSO, POLYMARS, and a generalized linear model.

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

$$\text{RMSE (fitted)} : \sqrt{\frac{\sum_{i=1}^{n}(\mu_i - \widehat{Y}_i)^2}{\sum_{i=1}^{n}(\mu_i - \overline{Y}_i)^2}} \tag{22}$$

$$\text{RMSE (derivative)}_{\nabla} : \sqrt{\frac{\sum_{i=1}^{n}\left(\frac{\partial\mu_i}{\partial T_i} - \widehat{\frac{\partial Y_i}{\partial T_i}}\right)^2}{\sum_{i=1}^{n}\left\{\frac{\partial\mu_i}{\partial T_i}\right\}^2}} \tag{23}$$

$$\text{Bias}_{\nabla} : \frac{\left|\sum_{i=1}^{n}\frac{\partial\mu_i}{\partial T_i} - \widehat{\frac{\partial\mu_i}{\partial T_i}}\right|}{\sqrt{\sum_{i=1}^{n}\left\{\frac{\partial\mu_i}{\partial T_i}\right\}^2}} \tag{24}$$

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 1 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. Finally while MDE is in a top performing set for the discontinuous case, POLYMARS had a slight advantage.

Figure 2 presents results on each methods' error in estimating the derivative. In the linear model, we are consistently dominated by POLYMARS and beat SuperLearners as p grows. In the interactive and non-linear settings, MDE is again aligned with the LASSO and POLYMARS, while POLYMARS performs worse than the null model in the non-linear setting, with the horseshoe competitive. POLYMARS performs the best in the discontinuous case but MDE was not far behind. We find that, across settings, LASSOplus is the only method that generally performs first or second best in RMSE, though POLYMARS, particularly in the discontinous case, and cross-validated LASSO are reasonable alternatives.

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

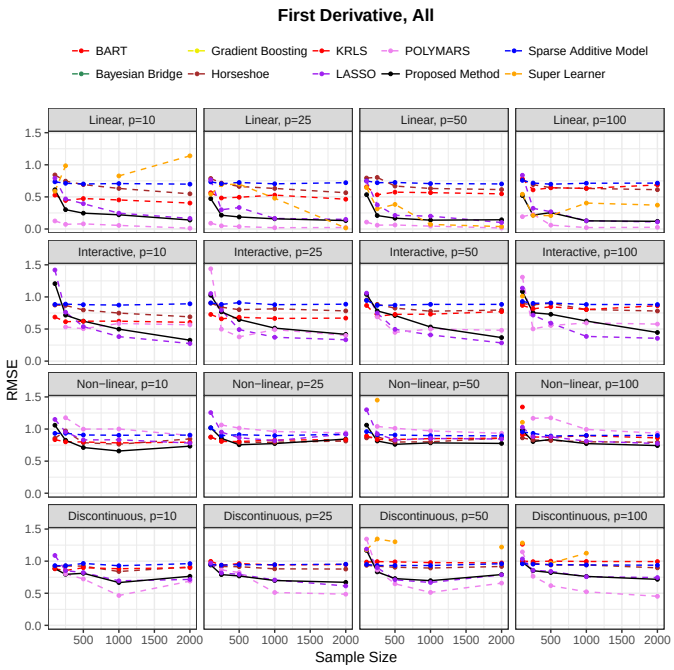

Figure 2: **A Comparison of RMSE for Unit Level Derivatives Across Methods.**

first dimension and the extent to which they capture the residual nonlinear trend. In our first simulation, there is no nonlinear trend in the fitted values, and the first derivative is flat, so any curvature found in the fitted values will show as error in the second component of the RMSE.

Results from this decomposition can be found in Figure 3 and Figure 4 . Figure 3 presents results for the low-dimensional linear trend while Figure 4 presents results for residual non-linear trend terms. We find that the proposed method performs well for both sets of effects. Interestingly in Figure 4 we see that POLYMARS does more poorly in the non-linear setting but better in the discontinuous setting.[3]

3 Extensions

We next extend our framework to two commonly used structural methods in the causal inference literature: instrumental variables and mediation. The literature on both methods are immense (Angrist and Pischke, 2008; VanderWeele, 2015), but there has been less work embedding them in sparse modelling (exceptions include Belloni, Chernozhukov and Hansen, 2011; Chernozhukov,

[3]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.

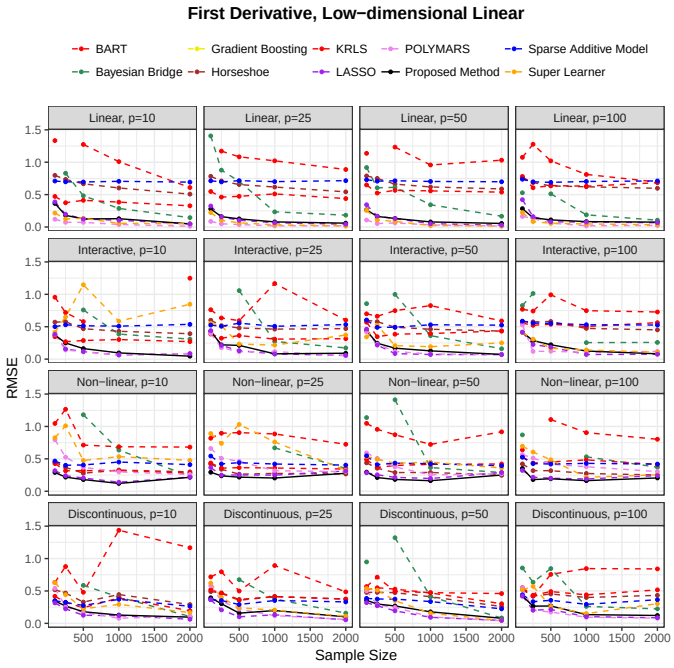

Figure 3: **A Comparison of RMSE Across Methods. Linear terms.**

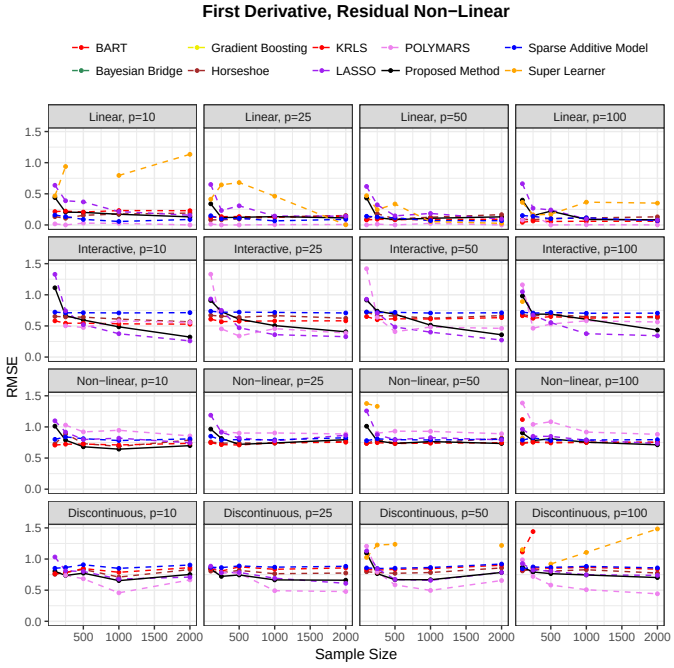

Figure 4: **A Comparison of RMSE Across Methods. Residual non-linear terms**

Hansen and Spindler, 2015; Zhao and Luo, 2016) or nonparameteric frameworks (Newey, 2013; Horowitz and Lee, 2007; Horowitz, 2011; Darolles et al., 2011; Hall, Horowitz et al., 2005; Horowitz, 2014*b*; Imai, Keele and Yamamoto, 2010; Pearl, 2014*a*). And none(?) to our knowledge using non-parametric *and* sparse modelling. Both methods are especially suited in our framework because

they can be approached as a two step estimation strategy where first stage estimates can be plugged into a second stage model.

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, sometimes called an "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.

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.[4] 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 deepen our connection between MDE and the standard two-stage least squares, we note that the Wald estimator or two-stage least squares estimates used in the IV setting is the ratio of two sample (partial) covariances, in effect averaging over compliers and non-compliers before taking a ratio. Our estimator instead is the average of observation-level partial derivatives, so the first stage can be explored directly. The second-stage estimate can be constructed from the estimated compliers, or those observations for which the instrument has a positive effect on the treatment, or some other subset of interest.

To formalize, we assume the treatment is not conditionally independent of the outcome given the covariates, $Y_i(\widetilde{T}_i, X_i) \not\perp\!\!\!\perp \widetilde{T}_i | X_i$, 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 $\widetilde{T}_i = T_i(\widetilde{Z}_i; X_i)$. For identification, we make the exclusion restriction that Z_i has no direct effect on the outcome, so $Y_i(T_i(\widetilde{Z}_i, X_i), \widetilde{Z}_i; X_i) = Y_i(T_i(\widetilde{Z}_i; X_i); X_i)$ or, equivalently, $Y_i(T_i(\widetilde{Z}_i, X_i), \widetilde{Z}_i; X_i) \perp\!\!\!\perp \widetilde{Z}_i | \widetilde{T}_i, X_i$. 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,

[4] Though see Hirano et al. (2000) for work on this limited to a binary treatment/instrument.

we trace the exogenous variation of the instrument through the treatment and onto the outcome. We then denote the observation-level IV causal effect as

$$\nabla_{Z_i}^{IV}(\delta_i^{IV}) = \frac{Y_i(T_i(Z_i + \delta_i^{IV}); X_i) - Y_i(T_i(Z_i); X_i)}{T_i(Z_i + \delta_i^{IV}; X_i) - T_i(Z_i, X_i)}, \tag{25}$$

which we refer to as the local individual causal effect, or LICE. The LICE only exists when the denominator is non-zero. 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). Since our estimand is at the observation level, we do not need to assume homogeneity in the denominator (monotonicity) in order to estimate into which stratum each observation falls (as required by Abadie, 2003; Aronow and Carnegie, 2013).

A nonparametric instrumental variable model. In the IV setting, we utilize nonparametric models for both the outcome and treatment. The outcome model is the same as the treatment model in Eq. 3. We also use a nonparametric model of the treatment in terms of the instrument and covariates as

$$T_i = \mu_T + R_T(Z_i, X_i)^\top c_T + \epsilon_i^T \tag{26}$$

where $R_T(Z_i, X_i)$ is a post-SIS screened set of bases constructed as described above, except taking the treatment as the outcome.

In this setting, the first stage effect of the instrument on the treatment is

$$\nabla_{Z_i}^{IV}(\delta_i^{IV}) = \left(R_T(Z_i + \delta_i^{IV}, X_i) - R_T(Z_i, X_i)\right)^\top c_T \tag{27}$$

and the second stage effect of the instrument-perturbed treatment perturbation on the outcome is

$$\nabla_{T_i}^{IV}(\delta_i^{IV}) = \left\{R\left(R_T(Z_i + \delta_i^{IV}, X_i)^\top \widehat{c}_T, X_i\right) - R\left(R_T(Z_i, X_i)^\top \widehat{c}_T, X_i\right)\right\}^\top c \tag{28}$$

Given estimates for the outcome and treatment models, we can calculate individual-level IV effect estimates using the plug-in estimate of Equation 25,

$$\widehat{\nabla}_{T_i,Z_i}^{IV}(\delta_i^{IV}) = \frac{\nabla_{T_i}^{IV}(\delta_i^{IV})}{\nabla_{Z_i}^{IV}(\delta_i^{IV})} \tag{29}$$

and our estimate of the sample LATE is

$$\widehat{\nabla}_{T_i,Z_i}^{IV} = \sum_{i=1}^{n} \widehat{\nabla}_{T_i,Z_i}^{IV}(\delta_i^{IV}) \times \mathbf{1}\left(\nabla_{Z_i}^{IV}(\delta_i^{IV}) \neq 0\right) \tag{30}$$

The estimate is not feasible without an estimate of each observation's compliance status, $\widehat{\mathbf{1}}\left(\nabla_{Z_i}^{IV}(\delta_i^{IV}) \neq 0\right)$, an issue to which we turn next.

Threshold for Estimating Compliers We next propose a means of estimating observations that are affected by the treatment. We are most interested in an accurate, but conservative, estimate of this subsample because these are the only observations for which a LICE is identified. Our threshold has two components. The first is a plug-in estimate of the Oracle Bound, as given in Inequality 16. This inequality gives a sense of the expected size of an estimate indistinguishable from noise. The Oracle Bound, though, 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

$$\widehat{df}_i = \frac{\partial \widehat{\mu}_i(Y_i)}{\partial Y_i} = \frac{\text{Cov}(Y_i, \widehat{\mu}_i(Y_i))}{\sigma^2} \tag{31}$$

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 and offers several extensions.

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

$$\widehat{df}_i = \frac{\partial [R^X(X_i) : R^{TX}(T_i, X_i)]^\top \widehat{c}}{\partial Y_i} \tag{32}$$

$$= \frac{\partial [R^X(X_i)]^\top \widehat{c}^X}{\partial Y_i} + \frac{\partial [R^{TX}(T_i, X_i)]^\top \widehat{c}^{TX}}{\partial Y_i}. \tag{33}$$

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

$$\widehat{df}_i^{\nabla} = \frac{\partial [R^{TX}(T_i, X_i)]^\top \widehat{c}^{TX}}{\partial Y_i}. \tag{34}$$

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

The adaptive component of our threshold is of the form

$$\widehat{df}_i^{adapt} = \frac{1/n}{1/n + \widehat{df}_i^{\nabla}} \tag{35}$$

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

$$\widehat{\mathbf{1}}\left(\nabla_{Z_i}^{IV}(\delta_i^{IV}) \neq 0\right) = \mathbf{1}\left(\left|\widehat{\nabla}_{Z_i}^{IV}(\delta_i^{IV})\right| > C\frac{\widehat{\lambda}||\widehat{w}_k||_\infty \widehat{\sigma}}{n\widehat{\phi}_0}\widehat{df}_i^{adapt}\right) \tag{36}$$

with $||\widehat{w}_k||_\infty$ the largest weight. C is a user-selected constant, but we show in our simulation that taking $C = 1$ 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 effects.

Comparison with Two-Stage Least Squares Estimate. The proposed method returns observation-level causal effects attributable to perturbations in the instrument. This differs in an important way from the standard Wald or two-stage least squares estimates (TSLS). The TSLS estimate under either a linear structural equation model is calculated from sample covariances as

$$\widehat{\theta}^{TSLS} = \frac{\widehat{\text{Cov}}_{\mathcal{S}}(Y_i, Z_i|X_i)}{\widehat{\text{Cov}}_{\mathcal{S}}(T_i, Z_i|X_i)}, \tag{37}$$

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 25, we see that the TSLS estimate averages the numerator and denominator over compliers, defiers, and observations for which no causal effect is identified.

Instead, using the proposed method, this heterogeneity comes to the fore, allowing a more nuanced interpretation and utilization of an instrumental variable analysis. Our estimator recovers

an observation-level Wald estimate. To show this, consider first the continuous case, where we take

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

$$= \frac{\partial Y_i(T_i(Z_i;X_i);X_i)/\partial Z_i}{\partial T_i(Z_i;X_i)/\partial Z_i} \tag{39}$$

$$= \frac{\mathrm{Cov}(Y_i,Z_i|X_i)/\mathrm{Var}(Z_i|X_i)}{\mathrm{Cov}(T_i,Z_i|X_i)/\mathrm{Var}(Z_i|X_i)} \tag{40}$$

$$= \frac{\mathrm{Cov}(Y_i,Z_i|X_i)}{\mathrm{Cov}(T_i,Z_i|X_i)}, \tag{41}$$

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.[5]

We obtain a similar result with a binary instrument and treatment, setting $\delta_i^{IV} = 1 - 2 \times Z_i$:

$$\nabla^{IV}_{T_i,Z_i}(\delta_i^{IV}) = \frac{Y_i(T_i(1;X_i);X_i) - Y_i(T_i(0;X_i);X_i)}{T_i(1;X_i) - T_i(0;X_i)} \tag{42}$$

$$= \frac{\mathbb{E}(Y_i|Z_i=1;X_i) - \mathbb{E}(Y_i|Z_i=0;X_i)}{\mathbb{E}(T_i|Z_i=1;X_i) - \mathbb{E}(T_i|Z_i=0;X_i)} \tag{43}$$

where, again, going from the second to third line comes from the exclusion restriction.

Robust estimation. 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,

$$\widehat{\nabla}^{IV;boot}_{Z_i}(\delta_i^{IV}) = \mathrm{med}_B \left(\frac{1}{\{R_T(Z_i+\delta_i^{IV},X_i) - R_T(Z_i,X_i)\}^\top \widehat{c}_{T,b}} \right) \times$$
$$\mathrm{med}_B \left(\left\{ R\left(R_T(Z_i+\delta_i^{IV},X_i)^\top \widehat{c}_{T,b}, X_i\right) - R\left(R_T(Z_i,X_i)^\top \widehat{c}_{T,b}, X_i\right) \right\}^\top \widehat{c}_b \right) \tag{44}$$

where $\mathrm{med}_B(a)$ 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.

3.1.1 Simulation Evidence

We next present a short set of simulation results to examine the performance of our method applied to the case of an encouragement design (instrumental variable). For simplicity we compare our proposed method to two-stage least squares estimation.[6] We examine first stage and second stage

[5] This is closely related to, yet differs from, Heckman and Vytlacil (2007*a*,*b*, 2005, 1999). We estimate a *pointwise* effect rather than integrating over a first-stage choice set. Also, we are not limited to the binary instrument/treatment

[6] 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

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

$$\text{Null} \;\; T_i = \sum_{k=5}^{8} X_{ik}\beta_k + \epsilon_i^T, \tag{45}$$

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

$$Z_i = -3 + (X_{i1} + X_{i4})^2 + \epsilon_i^T; \;\; \epsilon_i^Z \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 4). \tag{46}$$

Across settings, the outcome is generated as

$$Y_i = |T_i - 2| + \sum_{k \in \{5,6,7,8\}} X_{ik}\beta_k + X_{i1} \times X_{i2} + \epsilon_i^Y. \tag{47}$$

where

$$[\epsilon_i^T, \epsilon_i^Y]^\top \overset{\text{i.i.d.}}{\sim} N\left([0,0]^\top, C \begin{bmatrix} 1, .9 \\ .9, 1 \end{bmatrix}\right) \tag{48}$$

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.

a structural parameter rather than individual effects. Methods used in Section 2 have not been extended to the instrumental variables context except for ensemble tree methods by Athey, Tibshirani and Wager (2016).

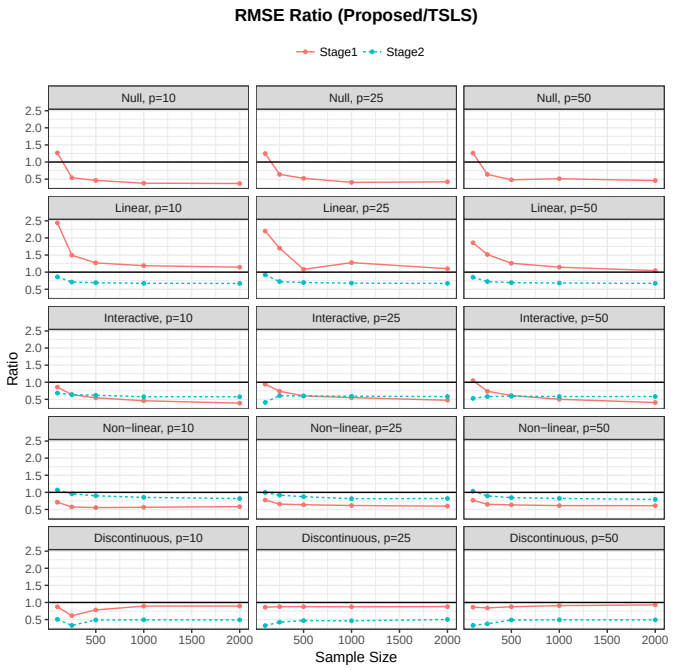

Figure 5: **Ratio of RMSE estimates for proposed method versus 2SLS for first stage and second stage estimation across simulation environments.**

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 2.

Figure 5 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.[7] 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.

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

Compliance Estimates Figure 6 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).[8] 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 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 D. 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.

3.2 Causal Mediation

Mediation analysis examines the pathway through which a treatment variable impacts some outcome through an intermediate variable. Mediation analysis hence examines mechanisms rather than treatment effects. Recently, the causal analysis of mediation has exploded (for a review see VanderWeele, 2015). Here we briefly show how the method of direct estimation can be used for mediation analysis.

Consider first a post-treatment mediator equipped with its own potential outcome function, $\widetilde{M}_i = M_i(\widetilde{T}_i; X_i)$ with law $F^M_{T_i,X_i}$ and observed value $M_i = M_i(T_i; X_i)$. The observed outcome is now $Y_i = Y_i(M_i, T_i; X_i)$ 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

[8] Figure 12 in Appendix D gives the true proportions in the simulated data.

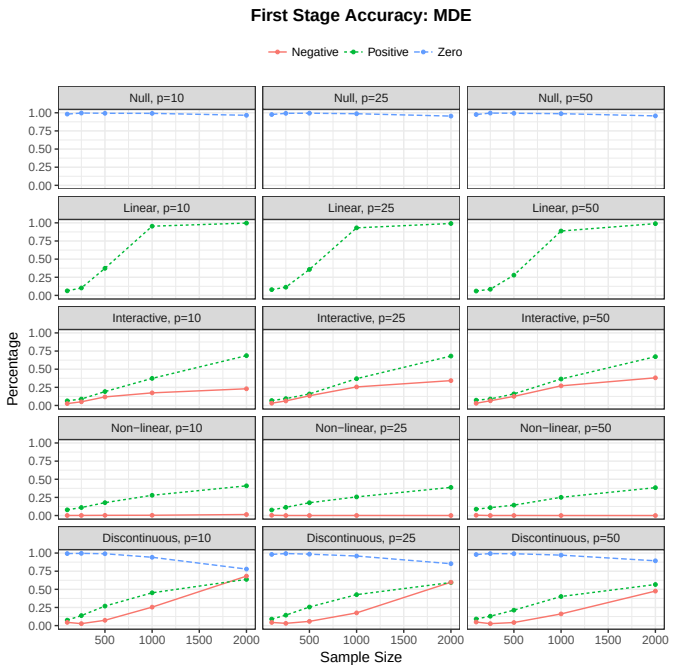

Figure 6: **First stage accuracy for proposed method.** Blue line plots the percentage of in fact positive LICE that we correctly capture. The red line plots the percentage of in fact negative LICE that we correctly capture. The green line plots the percentage of in fact non-identified units that we 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.

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): $Y_i(\widetilde{M}_i, \widetilde{T}_i; X_i) \perp\!\!\!\perp M_i(\widetilde{T}_i; X_i) | \widetilde{T}_i, X_i$ and $\left\{Y_i\left(\widetilde{M}_i, \widetilde{T}_i; X_i\right), M_i(\widetilde{T}_i; X_i)\right\} \perp\!\!\!\perp \widetilde{T}_i | X_i$.

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

$$\text{Total effect: } \nabla_{T_i}^{TE}(\delta_i^{TE}) = \frac{1}{\delta_i^{TE}} \left\{Y_i\left(M_i(T_i + \delta_i^{TE}; X_i), T_i + \delta_i^{TE}\right) - Y_i\left(M_i, T_i; X_i\right)\right\} \tag{49}$$

$$\text{Direct effect: } \nabla_{T_i}^{DE}(\delta_i^{DE}) = \frac{1}{\delta_i^{DE}} \left\{Y_i\left(M_i, T_i + \delta_i^{DE}\right) - Y_i\left(M_i, T_i; X_i\right)\right\} \tag{50}$$

$$\text{Mediated effect: } \nabla_{T_i}^{ME}(\delta_i^{ME}) = \frac{1}{\delta_i^{ME}} \left\{Y_i\left(M_i(T_i + \delta_i^{ME}; X_i), T_i\right) - Y_i\left(M_i, T_i; X_i\right)\right\}. \tag{51}$$

We will also make use of

$$\text{Mediator direct effect: } \nabla_{M_i}(\delta_i^{M}) = \frac{1}{\delta_i^{M}} \left\{Y_i\left(M_i + \delta_i^{M}, T_i; X_i\right) - Y_i\left(M_i, T_i; X_i\right)\right\} \tag{52}$$

$$\text{First stage mediation effect: } \nabla_{T_i}(\delta_i^{T}) = \frac{1}{\delta_i^{T}} \left\{M_i\left(T_i + \delta_i^{T}; X_i\right) - M_i\left(T_i; X_i\right)\right\} \tag{53}$$

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

$$\lim_{\delta_i^{TE}\to 0} \nabla_{T_i}^{TE}(\delta_i^{TE}) = \lim_{\delta_i^{DE}\to 0} \nabla_{T_i}^{DE}(\delta_i^{DE}) + \lim_{\delta_i^{M}\to 0} \nabla_{M_i}(\delta_i^{M}) \times \lim_{\delta_i^{T}\to 0} \nabla_{T_i}(\delta_i^{T}). \tag{54}$$

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

$$\lim_{\delta_i^{TE}\to 0} \nabla_{T_i}^{TE}(\delta_i^{TE}) = \lim_{\delta_i^{DE}\to 0} \nabla_{T_i}^{DE}(\delta_i^{DE}) + \lim_{\delta_i^{ME}\to 0} \nabla_{T_i}^{ME}(\delta_i^{ME}). \tag{55}$$

which gives the well known result that total effect is additive in the direct and mediated effects for linear models. And that the mediated effect is the product of the mediator direct effect and intermediate direct effect (MacKinnon et al., 2002).[9]

A nonparametric mediation model. We estimate two models, one for the outcome and one for the mediator. The models that we fit are

$$Y_i = \mu_Y + R_Y(M_i, T_i, X_i)^\top c_Y + \epsilon_i^Y \tag{56}$$

$$M_i = \mu_M + R_M(T_i, X_i)^\top c_M + \epsilon_i^M \tag{57}$$

where, again, the $R.(\cdot)$ vectors are a post-screened set of bases. We can now recover estimates of the effects in Equations 49–53.

Controlled Direct Effects 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). In these contexts a two step sequential g-estimation procedure is used. The first step regresses the outcome on the treatment, mediator, and all covariates including those that are pre-treatment and intermediate. The model could be extended to included interactions between the treatment and any pre-treatment covariates. Next, a de-mediated function (or "blip down") function is calculated from this equation. This is just the coefficient on the mediator, multiplied by each observation's value of the mediator, and subtracted from the outcome. The second step then regresses the demediated outcome on the treatment and pre-treatment confounders, and

[9]Imai, Keele and Tingley (2010) show that the product rule above does not apply to estimated coefficients unless the underlying model is linear. Since we are using a local linearization around each point as a function of the treatment and mediator, the product rule can be extended.

the coefficient on the treatment is the estimate of the controlled direct effect. Acharya, Blackwell and Sen (2016) note the strong modelling assumptions required to use this method when dealing with continuous variables (non-parametric identification is shown for binary or limited value treatment and mediators variables in Robins, Hernan and Brumback (2000)). We can relax these modeling assumptions, such as linearity and additivity, by using our nonparametric regression to "de-mediate" the outcome. With the de-mediated outcome, we could then use the method of direct estimation to for estimating the impact of a controlled direct fluctuation of the treatment on the outcome. With this connection to controlled direct effect we open up a broader connection to the literature on sequential g-estimation.

4 Empirical Applications

We now focus on two empirical applications. The first examines the case of a binary treatment in the context of the National Supported Work Study program (LaLonde, 1986). The second moves to a continuous treatment case that uses an instrumental variable structural model (Larreguy and Marshall, 2017).

4.1 Binary Treatment

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.

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

$$Y_i(1; X_i) - Y_i(0; X_i) = \nabla_{T_i}(1 - 2 \times T_i) \tag{58}$$

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

$$ATE = \frac{1}{n} \sum_{i=1}^{n} \nabla_{T_i}(1 - 2 \times T_i) = \tag{59}$$

$$ATT = \frac{1}{n_T} \sum_{i=1}^{n} (\nabla_{T_i}(1 - 2 \times T_i)) \times \mathbf{1}(T_i = 1) \tag{60}$$

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.

Applied Example: The National Supported Work Study. The promise of causal inference lies in the prospect of recovering experimental results from an observational dataset. To that end, we assess our method via a benchmark dataset that allows us to assess the extent to which methods can achieve this goal. We use the design and data first put forward by LaLonde (1986), but has since been utilized in several other works (e.g. 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.[10]

The observational dataset comes from the Panel Study for Income Dynamics, a survey of low-income individuals, with data on the same covariates as the experimental data ($n_P = 2915$). In our first analysis, we assess several different methods' ability to replicate an experimental benchmark. We conduct three separate tests, in each estimating the Average Treatment on the Treated (ATT). In the first, the model is given only the experimental treated and the observational untreated. The goal is to recover the experimental result ($886.30) with access to only the experimental treated and observational untreated group. 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. The "treatment" is now whether the observation was in the experiment. As no one in the treated or untreated group received the treatment, the known true effect is zero. Any deviation from zero 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

[10]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.

Treatment Group Control Grooup	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
MDE, Bagged	960.16	96.66	416.90	639.92
CBPS	453.19	-300.28	872.27	747.42
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 both MDE and the bagged variant perform well across the comparisons.

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. As above, we include the horseshoe estimate fit to the same bases as MDE. We also include LASSO, BART, GBM, and POLYMARS. We include three methods designed to work with a binary treatment: logistic propensity score matching (Propensity, Ho et al., 2007), the covariate balancing propensity score (CBPS, Imai and Ratkovic, 2014) and the targeted maximum likelihood estimate (Gruber and van der Laan, 2011).

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, both the sample version and bagged estimate, 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.

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

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 2: **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.

subsample. Results are presented in Table 2. 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 untreated, then use this model to predict the held-out experimental group. In each case, we include the result from least-squares fit 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.

4.2 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.[11] 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 1.2.[12]. We

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

[12] The original model used a substantial number of fixed effects, in part to establish a difference in difference

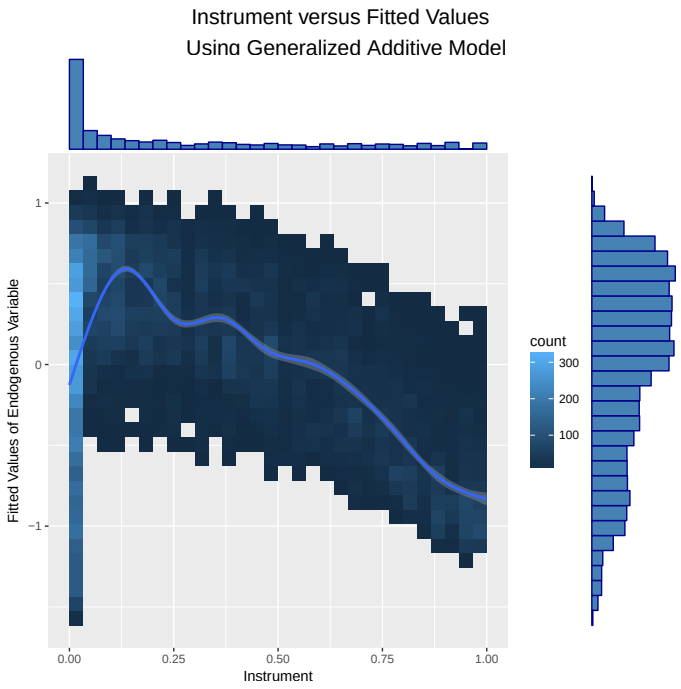

Figure 7: **Nonlinear bi-variate relationship between instrument and fitted values of the endogenous variable. Estimated using generalized additive model.**

present two main sets of results.

First, in Figure 8 we present a plot analogous to Figure 7 which plots for each observation the relationship between the instrument and the covariance between the instrument and and (endogenous) treatment variable. 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 to convey the basic pattern. 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.[13]

Figure 9 plots the "second stage": the treatment variable versus the covariance between the treatment and outcome variable. A piecewise line plots the means across the values of the treatment

identification strategy with an instrumental variable. This poses no problem for our proposed method. However for computational purposes we made some minor modifications that we discuss in Appendix D.

[13]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.

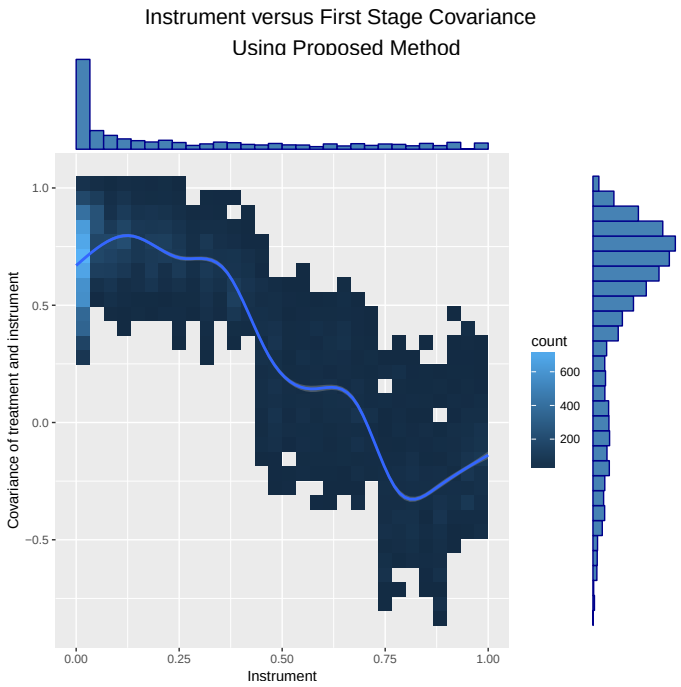

Figure 8: **Relationship between instrument and first stage covariance (covariance between instrument and treatment). Estimated using the proposed method.**

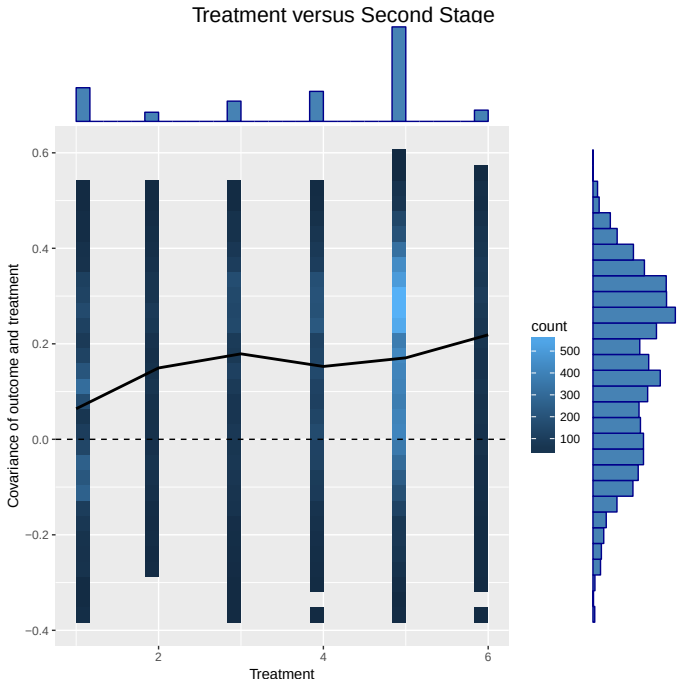

Figure 9: **Treatment variable versus .**

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 8 this helps to give us an expectation of what the LICE will look like over the sample.

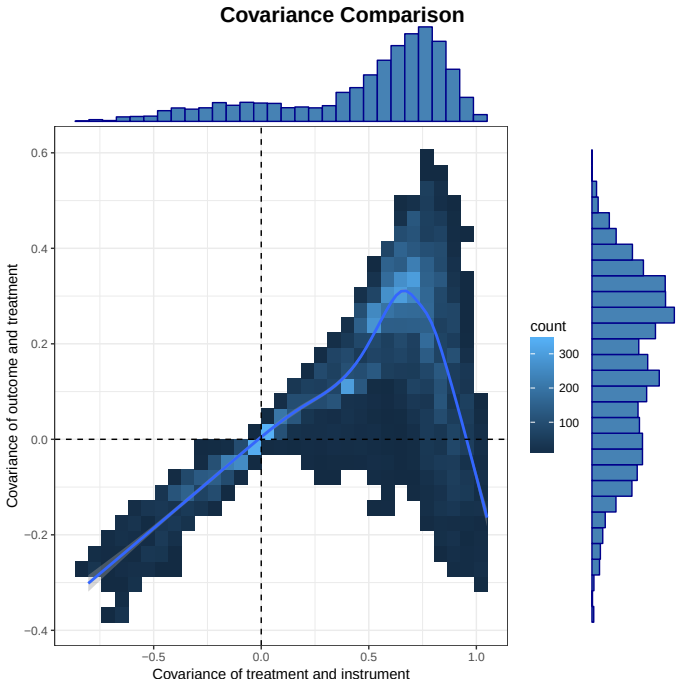

Figure 10: **First stage versus second stage.**

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 10 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. For these observations the LICE will still be positive. On average these observations are likely to be of lower education (see Figure 9). 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 11 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

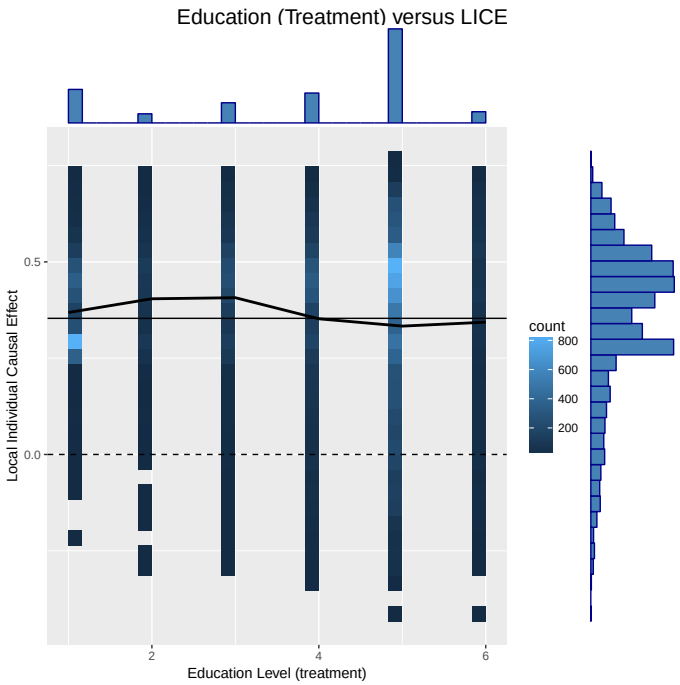

Figure 11: **Local Individual Causal Effect versus treatment variable.**

.35 (95% confidence interval: .20, .49).[14] 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 10.

5 Conclusion

This paper has focused on individual level causal effects. While a core building block of the causal inference literature, highlighted by the use of potential outcomes, most statistical tools used by applied researchers target a structural parameter. We instead directly target individual level causal effect estimates by considering an extremely rich covariate *and* functional space. We utilize recent advances in machine learning to demonstrate that the proposed approach performs extremely well against other cutting edge methods. Finally, we should how our conceptualization easily extends beyond the estimation of simple treatment effects, but can also be used in encouragement designs (instrumental variables) and the analysis of causal mechanisms.

[14]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.

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.[15] 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 that we previously developed (Ratkovic and Tingley, 2017).[16] We are actively extending our approach 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 from partial to Frechet derivatives. As discussed above, out approach appears to naturally fit with what is needed for the calculation of controlled direct effects and sequential-g estimation, but we are applying our approach to real and simulated data in this context. This is current work. Future work could consider how to better incorporate spline bases with discontinuities, so as to capture "jumps" in the data. We

[15]Of note, as shown in Section 2, our expanded basis function approach also enabled the LASSO to beat a number of cutting edge machine learning methods.

[16]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.

also are currently looking to extend the model to longitudinal data.

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. 2008. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

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.

Baron, Reuben M and David A Kenny. 1986. "The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations." *Journal of personality and social psychology* 51(6):1173.

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, Victor Chernozhukov and Christian Hansen. 2011. "LASSO methods for Gaussian instrumental variables models.".

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.".

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

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).

Darolles, Serge, Yanqin Fan, Jean-Pierre Florens and Eric Renault. 2011. "Nonparametric instrumental regression." *Econometrica* 79(5):1541–1565.

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.

Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48(2):424–455.

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*

Hall, Peter, Joel L Horowitz et al. 2005. "Nonparametric methods for inference in the presence of instrumental variables." *The Annals of Statistics* 33(6):2904–2929.

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* .

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2010. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag.

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, 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.

Hirano, Keisuke, Guido W. Imbens, Donald B. Rubin and Xiao-Hua Zhou. 2000. "Assessing the Effect of an Influenza Vaccine in an Encouragement Design." *Biostatistics* 1(1):69–88.

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

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

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

Horowitz, Joel L. 2011. "Applied nonparametric instrumental variables estimation." *Econometrica* 79(2):347–394.

Horowitz, Joel L. 2014*b*. "Adaptive nonparametric instrumental variables estimation: Empirical choice of the regularization parameter." *Journal of Econometrics* 180(2):158–173.

Horowitz, Joel L and Sokbae Lee. 2007. "Nonparametric instrumental variables estimation of a quantile regression model." *Econometrica* 75(4):1191–1208.

Iacus, Stefano M., Gary King and Giuseppe Porro. 2011. "Multivariate Matching Methods That are Monotonic Imbalance Bounding." *Journal of the American Statistical Association* 106:345–361.

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

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.

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.

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*a*. "Interpretation and identification of causal mediation." *Psychological Methods* 19(4):459.

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

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

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

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

Ratkovic, Marc and Dustin Tingley. 2017. "Sparse Estimation and Uncertainty with Application to Subgroup Analysis." *Political Analysis* .

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., 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. 2005. "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions." *Journal of the Americal Statistical Association* 100(469):322–331.

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.

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.

Zhao, Yi and Xi Luo. 2016. "Pathway Lasso: Estimate and Select Sparse Mediation Pathways with High Dimensional Mediators." *arXiv preprint arXiv:1603.07749* .

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

A Instrumental Variables 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.[17] 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:

$$\widehat{F} = \frac{\frac{1}{\widehat{df}^{\nabla}} \sum_{i=1}^{n} \left\{ R_T^{ZX}(Z_i, X_i)^\top \widehat{c}_T^{ZX} - \widehat{\mu}_T \right\}^2}{\frac{1}{n - \widehat{df}^{\nabla}} \sum_{i=1}^{n} \left\{ T_i - R_T(Z_i, X_i)^\top \widehat{c}_T - \widehat{\mu}_T \right\}^2} \tag{61}$$

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.

B LASSOplus Properties

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

$$-\log\left(\Pr\left(c, \{w_k\}_{k=1}^p | \lambda, \sigma^2, \gamma\right)\right) = \frac{1}{\sigma^2}\left\{\frac{1}{2}\sum_{i=1}^{n}(Y_i - R(T_i, X_i)^\top c)^2 + \lambda\sigma\sum_{k=1}^{p} w_k|\beta_k|\right\} + \sum_{k=1}^{p} w_k^\gamma. \tag{62}$$

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 PC, we reintroduce conjugacy in the mean parameters through augmentation:

$$c_k \sim DE(w_k\lambda/\sigma) \Rightarrow c_k|\tau_k^2, \sigma^2 \sim \mathcal{N}(0, \tau_k^2\sigma^2); \;\; \tau_k^2 \sim \exp(\lambda^2 w_k^2/2). \tag{63}$$

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

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

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

$$\widehat{c}\,|\cdot = \underset{c}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i - R(T_i, X_i)^\top c)^2}{2\sigma^2} + \frac{\lambda}{n\sigma}\sum_{k=1}^{p} w_k|c_k|. \tag{64}$$

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

$$\lambda^2|\cdot \sim \Gamma\left(n \times (\log(n) + 2\log(p)), \sum_{k=1}^{p}\tau_k^2/2 + \rho\right) \tag{65}$$

$$\tag{66}$$

which several desirable properties. First, 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.

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

$$\Pr(w_k|\cdot) = \frac{e^{-w_k^\gamma - \frac{\lambda w_k}{\sigma}|c_k|}}{\int e^{-w_k^\gamma - \frac{\lambda w_k}{\sigma}|c_k|}dw_k}; \quad \mathbb{E}(w_k|\cdot) = \int_{w_k=0}^{\infty} w_k \Pr(w_k|\cdot)dw_k \tag{67}$$

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

$$\frac{d\widehat{w}_k}{d|c_k|} = -\widehat{\lambda}\sqrt{\widehat{\frac{1}{\sigma^2}}}\mathrm{Var}(w_k|\cdot) < 0 \tag{68}$$

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).

Next, we 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. The value of this parameter is controlled by $\widehat{\gamma}$, to which we turn next.

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 PC prior with 1/2 the rate parameter.

C 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

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

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

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

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

After some manipulation (e.g., Ratkovic and Tingley, 2017; Buhlmann and van de Geer, 2013), this excess risk generates the inequality

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

which is Equation 12.

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

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

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

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

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

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

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

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

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

This gives us

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

which is a solution to the following LASSO problem

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

where $\Delta R^{\widehat{S},T}$ is the elementwise partial derivative of $R^{\widehat{S},T}$ wrt T_i and $\widetilde{\Delta Y}$ is a pseudo-response constructed from the estimate $\widehat{\Delta Y} = \left[\frac{\widehat{\partial Y_i}}{\partial T_i}\right]$ as

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

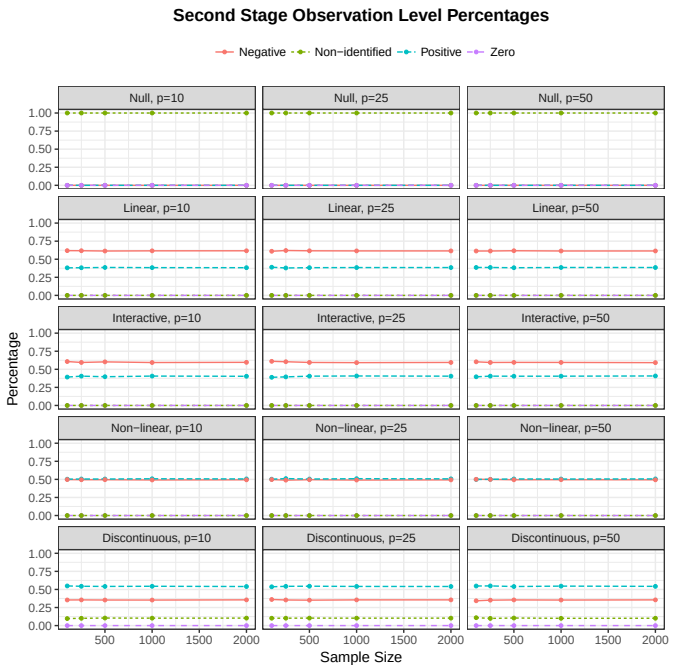

Figure 12: **Percentage of observations in sample by compliance status.**

The model is fit using basis-specific tuning parameter,

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

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

since $\partial \epsilon_i / \partial T_i = 0$. Since each element of ΔR^{TX} is also in R^X, by construction, we get the weights to be of order $\sqrt{n \log(p)}$, from which the oracle result follows directly.

D Simulation and Data Appendix

D.1 Additional Simulation Results

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.

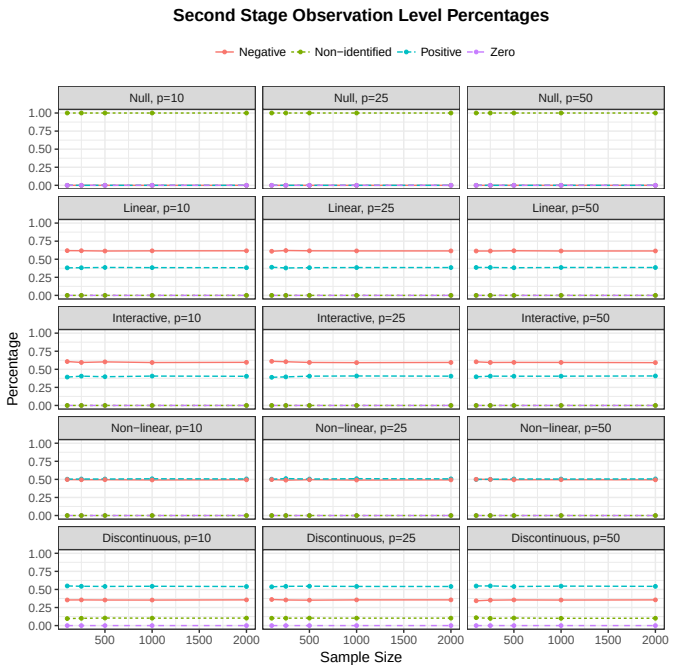

Figure 13: **Percentage of observations in sample by category of sign on instrumented causal effect.**

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 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.

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

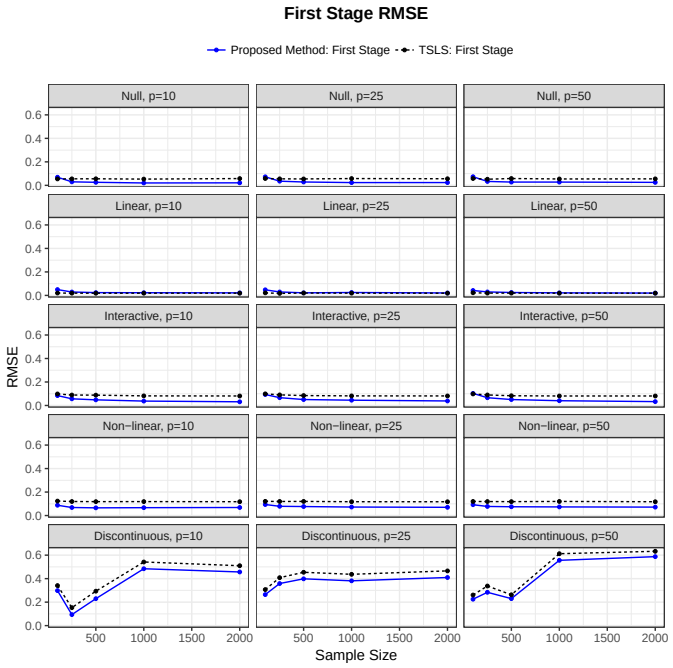

Figure 14: **RMSE estimates for proposed method and 2SLS for first stage estimation across simulation environments.**

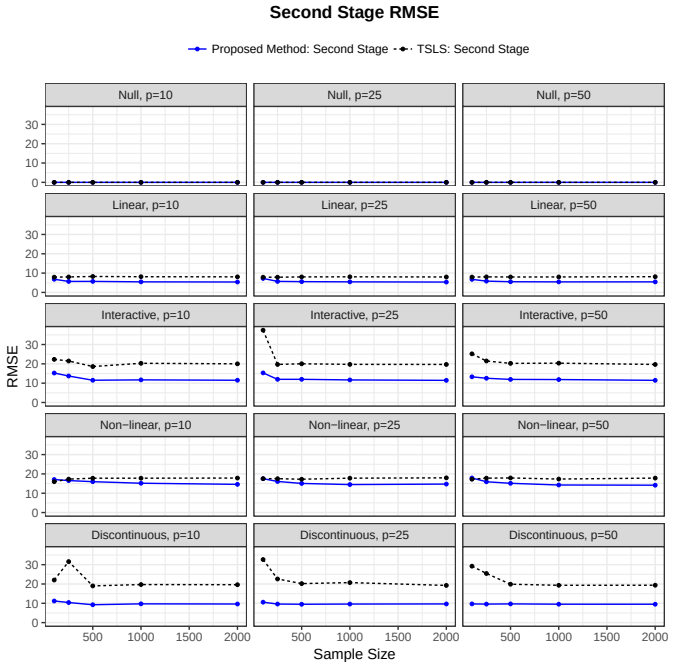

Figure 15: **RMSE estimates for proposed method and 2SLS for second stage estimation across simulation environments.**

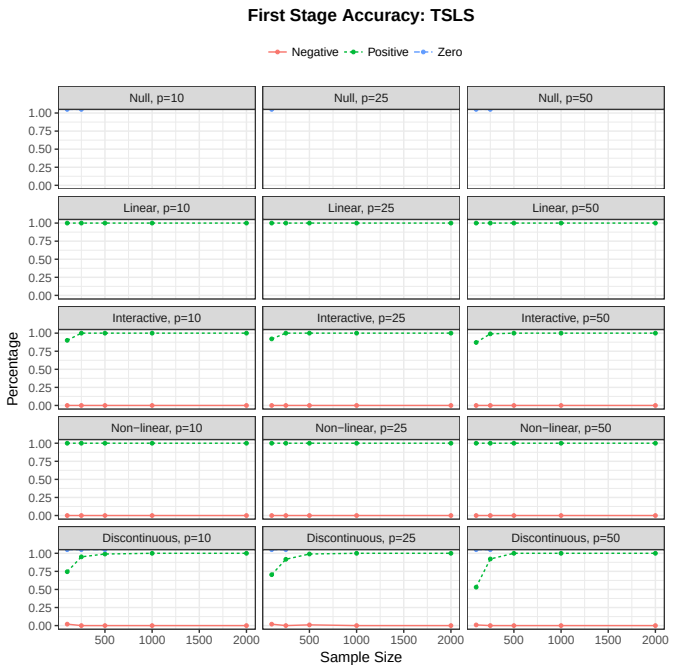

Figure 16: **First stage accuracy for TSLS. Blue line plots the percentage of in fact positive LICE that TSLS correctly captures. The red line plots the percentage of in fact positively encouraged that TSLS correctly captures. The green line plots the percentage of negatively encouraged units that TSLS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

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.

D.2 Larreguy and Marshall Example

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.

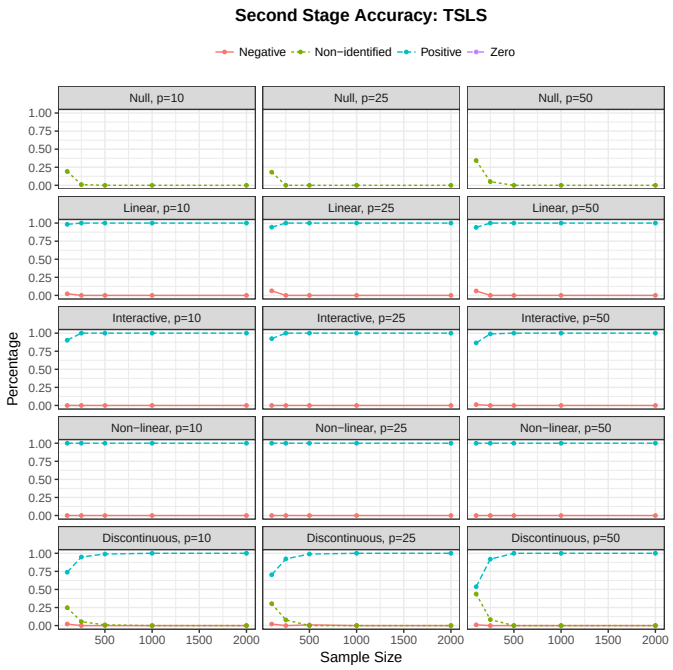

Figure 17: **Second stage accuracy for TSLS. Blue line plots the percentage of in fact positive LICE that TSLS correctly captures. The red line plots the percentage of in fact negative LICE that TSLS correctly captures. The green line plots the percentage of in fact non-identified units that TSLS correctly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

The data we place into our software is then

$$X^{big} = [X_{fe} : X_{var} : H_{fe}X_{var} : (I - H_{fe})X_{var}]. \tag{82}$$

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

D.2.1 Additional Analyses

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.

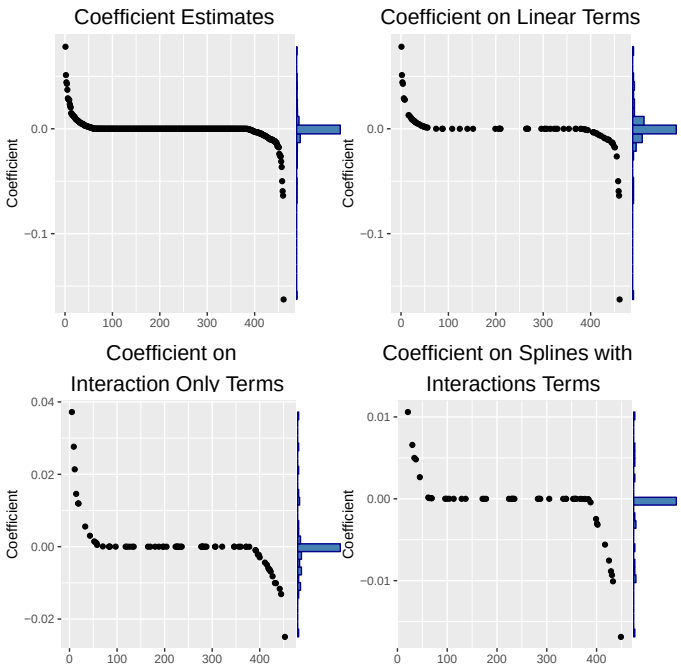

Figure 18: **Coefficient Estimates by Effect Type**