Journal of Machine Learning Research 24 (2023) 1-57 Submitted 6/21; Revised 2/23; Published 5/23

Exploiting Discovered Regression Discontinuities to Debias Conditioned-on-observable Estimators

Benjamin Jakubowski BUJ201@NYU.EDU
Machine Learning for Good Laboratory
New York University
Brooklyn, NY 11201, USA

Sriram Somanchi SOMANCHI.1@ND.EDU
IT, Analytics, and Operations
University of Notre Dame
South Bend, IN 46556, USA

Edward McFowland III EMCFOWLAND@HBS.EDU
Technology, Operations, and Management
Harvard University
Boston, MA 02163, USA

Daniel B. Neill DANIEL.NEILL@NYU.EDU
Machine Learning for Good Laboratory
New York University
Brooklyn, NY 11201, USA

Editor: Silvia Chiappa

Abstract

Regression discontinuity (RD) designs are widely used to estimate causal effects in the absence of a randomized experiment. However, standard approaches to RD analysis face two significant limitations. First, they require *a priori* knowledge of discontinuities in treatment. Second, they yield doubly-local treatment effect estimates, and fail to provide more general causal effect estimates away from the discontinuity. To address these limitations, we introduce a novel method for automatically detecting RDs at scale, integrating information from multiple discovered discontinuities with an observational estimator, and extrapolating away from discovered, local RDs. We demonstrate the performance of our method on two synthetic datasets, showing improved performance compared to direct use of an observational estimator, direct extrapolation of RD estimates, and existing methods for combining multiple causal effect estimates. Finally, we apply our novel method to estimate spatially heterogeneous treatment effects in the context of a recent economic development problem.

Keywords: causal inference, natural experiments, heterogeneous treatment effects, regression discontinuity designs, Gaussian processes

1. Introduction

Estimating casual effects from observational data remains a central challenge across empirical sciences. While randomized trials and experimentation allow for causal inference under

minimal assumptions, many important scientific questions are not amenable to study under randomization for practical or ethical reasons.

Where direct randomization is infeasible – but observational data is on hand – empirical scientists are left with two main paths to causal inference. One option is to make the strong assumption that assignment to treatment is conditionally independent of potential outcomes given observables, and to estimate causal effects using a "conditioned-on-observables" estimator (e.g., matching or reweighting estimators, or doubly-robust variants thereof). Unfortunately, if this strong assumption does not hold, unobserved confounders can introduce bias into the estimates of causal effect.

If treatment assignment cannot be assumed to be conditionally independent of potential outcomes, one can alternatively attempt to identify and exploit sources of exogenous, random variation in treatment (natural experiments, or quasi-experimental designs) to obtain causal estimates. While this general approach encompasses a wide range of econometric methods, we focus on one in particular: regression discontinuity (RD) designs. RD designs occur where treatment propensity varies discontinuously across some threshold. Under assumptions given in Dong (2018) and Hahn et al. (2001), such discontinuities can be exploited to obtain unbiased estimates of causal effects at the threshold.

RD designs are well-established, for example, by Hahn et al. (2001), and are widely used in practice. However, as routinely applied, they suffer from two substantial drawbacks. First, and most significantly, most RD analyses exploit discontinuities that are known *a priori*, reducing the applicability of RD designs to emerging domains where discontinuities are as of yet unknown. Second, since RD designs yield causal estimates that are local to the discontinuity, they do not immediately allow for inference across the full population. The combination of the manual, limited, and *ad hoc* nature of current approaches for RD identification, and the inability to extrapolate well from discovered RDs, limits the ability of RD designs to provide unbiased and accurate estimates of heterogeneous treatment effects across a population. In contrast, we wish to discover RDs systematically, automatically, and at scale, and to use these discovered natural experiments to draw broader inferences across the population as to which treatments are effective and for whom.

To address these limitations, we introduce a novel approach, `DEE` (Discover-Estimate-Extrapolate) for the discovery and analysis of RD designs. `DEE` extends LoRD^3, a recently introduced method for discovery of local regression discontinuities from observational data (Herlands et al., 2018), by introducing a complementary approach to extrapolating and integrating information from a set of discovered local discontinuities. After estimating conditional causal effects along the discovered discontinuities, we take a nonparametric approach to extrapolating these estimates throughout the attribute space. We consider two extrapolation strategies: (i) directly extrapolating treatment effect estimates from discovered discontinuities, and (ii) exploiting the discovered discontinuities to debias a conditioned-on-observables estimate. By averaging estimates from these two approaches, we realize gains relative to model selection.

Our approach contributes to three recent areas of research. First, our method extends recent work on discovery of natural experiments (Jensen et al., 2008; Herlands et al., 2018). Jensen et al. (2008) provide one of the earliest methods for automated discovery of natural experiments, describing an automated system for discovering "non-equivalent comparison group designs" under which a pre/post comparison identifies a causal effect. Herlands et al.

(2018) subsequently introduced a local scan method (LoRD^3) for automatic discovery of RD designs. We directly extend Herlands et al. (2018), describing a novel approach for combining multiple discovered, local discontinuities into a global estimator of conditional causal effects.

Second, we contribute to the growing body of literature on generalizing regression discontinuity estimates (Angrist and Rokkanen, 2015; Cattaneo et al., 2020). Angrist and Rokkanen (2015) describe falsification tests for a strong notion of external validity, under which treatment is conditionally independent of potential outcomes. Compared to Angrist and Rokkanen (2015), we adopt and are able to extrapolate regression discontinuity estimates under a weaker restriction, though our method does not offer falsification tests comparable to theirs. Our method is also similar to Cattaneo et al. (2020), in that it exploits multiple discontinuities to extrapolate the treatment effect away from a threshold. However, unlike their method, we do not require (i) that the discontinuities share a common running variable, nor (ii) that each individual unit be assigned to a known discontinuity.

Finally, similar to Kallus et al. (2018), we take the approach of exploiting unbiased treatment effect estimates (in our case, from discovered regression discontinuities) to estimate a bias correction for a conditioned-on-observables estimator. Unlike Kallus et al., we do not assume that this bias follows a parametric form. Moreover, in addition to estimating and extrapolating a bias correction, our method adaptively considers direct extrapolation of RD treatment effect estimates and integrates the two estimates through model averaging.

The remainder of the paper proceeds as follows. Section 2 formally introduces the causal inference problem. Section 3 presents our novel inferential method. Section 4 provides theoretical results in support of model averaging using a novel measure of predictive fit. Section 5 discusses related work. Section 6 presents empirical results on two synthetic problems, and on a recent problem from the economic development literature (Asher and Novosad, 2020). Finally, Section 7 discusses potential extensions and variations of our method, and concludes.

2. Setup

We aim to estimate the heterogeneous effect of a binary treatment $T \in \{0, 1\}$ on a real-valued observed outcome $Y \in \mathbb{R}$, given a set of observed covariates X with support $\mathcal{X} \subset \mathbb{R}^d$. Adopting the Rubin Casual Model and potential outcomes notation (Rubin, 2005), the target estimand is the conditional average treatment effect (CATE) given by

$$\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x],$$

where $Y(1)$ and $Y(0)$ respectively denote the potential outcomes given a unit's assignment to the treatment or control group.[1] The observed outcome (Y) can then be represented in terms of the potential outcomes and treatment as $Y = (1 - T) * Y(0) + T * Y(1)$. We aim to estimate CATE in the absence of experimental data, using just an observational dataset $\mathcal{D} = \{(x_i, t_i, y_i)\}_{i=1}^N$, and as such we face the fundamental problem of causal inference (Holland, 1986), since we only observe one of the two potential outcomes, $y_i = Y_i(t_i)$, for each unit.

1. We adopt the potential outcome notation as it allows us to more easily and clearly express the assumptions and identification in RD designs (Imbens, 2020).

An initial approximation to $\tau(x)$, which we denote as $\tau^{obs}(x)$, can be obtained by conditioning on the observed covariates X:

$$\begin{aligned}\tau^{obs}(x) &= \mathbb{E}[Y \mid X = x, T = 1] - \mathbb{E}[Y \mid X = x, T = 0] \\ &= \mathbb{E}[Y(1) \mid X = x, T = 1] - \mathbb{E}[Y(0) \mid X = x, T = 0].\end{aligned}$$

Assuming probabilistic treatment assignment, such that $0 < P(T = 1 \mid X = x) < 1$ for all x, and strong conditional ignorability[2], such that treatment is conditionally independent of potential outcomes ($\{Y(0), Y(1)\} \perp T \mid X$), the "conditioned-on-observables" estimator $\tau^{obs}(x)$ is in fact equal to the target estimand $\tau(x)$.[3] However, when treatment assignment is not conditionally independent of potential outcomes given observed covariates, the conditioned-on-observables approximation to the CATE can suffer from omitted variables bias.

To address this potential confounding, we follow Kallus et al. (2018), in spirit, and consider a special case where we can generate unconfounded estimates of $\tau(x)$, but only for x within a subset $\mathcal{U}$ (an *unconfounded* subset) of the support $\mathcal{X}$. While Kallus et al. consider the case where a supplemental experimental dataset is available, but only on a restricted region $\mathcal{U} \subset \mathcal{X}$, we instead consider the case where multiple *fuzzy* regression discontinuities[4] – discontinuities in the treatment propensity $P(T = 1 \mid X = x)$ – can be discovered, offering the possibility of directly constructing $\mathcal{U}$ from $\mathcal{D}$.

These regression discontinuities are typically analyzed under a Local Average Treatment Effect (LATE) framework (Imbens and Angrist, 1994). Consider an RD design where treatment propensity is a discontinuous function of Z (the *running variable*), with a discontinuity at $Z = z_0$ (the *threshold*).[5] As a concrete example, Matsudaira (2008) studied the effect of summer school attendance on subsequent achievement scores. In this example, students whose test score (running variable) is less than a predefined cutoff z_0 were encouraged to attend summer school. This is a fuzzy RD design, as there is a probabilistic relationship between test score (running variable) and summer school attendance (treatment), and it is discontinuous at the cutoff z_0.

Under the LATE framework, units are associated with a potential treatment function $T(z)$, which indicates each unit's treatment status based on a given cutoff value z. That is, $T(z) = 1$ if the unit would be treated had the threshold value been set to z, and $T(z) = 0$ otherwise (Imbens and Lemieux, 2008). Considering the earlier example based on Matsudaira (2008), a potential treatment function $T(z)$ maps the cutoff test score z to a student's summer school attendance, i.e., $T(z) = 1$, and increasing the cutoff z means that more students will be encouraged to attend the summer school program. Based on the behavior of $T(z)$, units are assigned a compliance type G. Under Assumptions 1-3 below, *compliers* are those units whose treatment status $T(z)$ is sensitive to the value of the threshold z, while *always-takers* and *never-takers* are characterized by constant potential

2. This assumption has many names in the literature such as no unmeasured confounders, unconfoundedness, selection on observables, exogeneity, or conditional independence.
3. Note that $\tau(x)$ can also be estimated using indirect approaches, including reweighting estimators such as inverse propensity score weighting (Li et al., 2018).
4. We adopt the term "fuzzy regression discontinuity" from Trochim (1984).
5. Note that for simplicity of notation we assume a single running variable. In our approach, we allow for the discovery of complex RDs involving multiple/joint running variables.

treatment functions $T(z) = 1$ and $T(z) = 0$, respectively. For the example based on Matsudaira (2008), compliers are those individuals who would attend the summer school if their test score Z was lower than the cutoff z (i.e., if they were encouraged to attend summer school) and would not attend summer school if their score Z was above the cutoff z. Never-takers and always-takers are respectively those individuals who would never attend and always attend summer school, irrespective of whether their test score Z was above or below the cutoff z.

Given this framework, standard assumptions provide for identification and estimation of the LATE,

$$\tau^{LATE}(Z = z_0) = \mathbb{E}[Y(1) - Y(0) \mid Z = z_0, G = Complier].$$

Several authors (Hahn et al., 2001; Imbens and Lemieux, 2008; Dong, 2018; Angrist and Rokkanen, 2015) give sufficient conditions for identification and estimation of $\tau^{LATE}(Z = z_0)$. These standard assumptions include:

Assumption 1 (Existence of RD) *The treatment propensity $P(T = 1 \mid Z = z)$ is discontinuous at $Z = z_0$. That is, $\lim_{z \downarrow z_0} P(T = 1|Z = z) \neq \lim_{z \uparrow z_0} P(T = 1|Z = z)$.*

Assumption 2 (Monotonicity) *The potential treatment function $T(z)$ is either non-increasing in z for all units, or non-decreasing in z for all units.*

Assumption 3 (Continuity) *The conditional expectations $\mathbb{E}[Y(1) \mid Z = z, G = g]$ and $\mathbb{E}[Y(0) \mid Z = z, G = g]$ and the conditional distributions $P(G = g \mid Z = z)$ are continuous in z at $z = z_0$ for all compliance types $G = g$.*

Assumption 1 is that an RD exists, that is, the probability of treatment is sharply affected at the threshold $Z = z_0$. Assumptions 1 and 2 together imply that there are, in fact, compliers at the threshold $Z = z_0$. Assumption 2 also means that all individuals are affected the same way by the choice of the threshold z, if at all. Finally, Assumption 3 implies the threshold z_0 only impacts the probability of treatment assignment, not the potential outcomes or compliance type.

These standard assumptions allow for identification and estimation of

$$\tau^{LATE}(Z = z_0) = \frac{\lim_{z \downarrow z_0} \mathbb{E}[Y|Z = z] - \lim_{z \uparrow z_0} \mathbb{E}[Y|Z = z]}{\lim_{z \downarrow z_0} \mathbb{E}[T|Z = z] - \lim_{z \uparrow z_0} \mathbb{E}[T|Z = z]}.$$

However, the estimated effect is doubly local: it is both (i) local to the discontinuity $Z = z_0$, and (ii) local to the complier subpopulation. Since we aim to estimate the CATE across all of $\mathcal{X}$, our approach requires two modifications to this standard treatment of RD designs.

First, we aim to estimate the CATE $\tau(x)$ conditioning on all observed covariates, not just the running variable Z. Thus, instead of estimating LATEs of the form $\tau^{LATE}(Z = z_0)$, averaging the treatment effect along the full extent of the discontinuity at $Z = z_0$ (i.e., marginalizing over all $X_{\setminus Z} = X \setminus \{Z\}$), we estimate *conditional* local average treatment effects (CLATEs), $\tau^{CLATE}(Z = z_0, X_{\setminus Z} = x)$, following for example Becker et al. (2013), and defined as

$$\begin{aligned}\tau^{CLATE}(Z = z_0, X_{\setminus Z} = x) &= \mathbb{E}[Y(1) - Y(0) \mid Z = z_0, X_{\setminus Z} = x, G = Complier] \\ &= \frac{\lim_{z \downarrow z_0} \mathbb{E}[Y|Z = z, X_{\setminus Z} = x] - \lim_{z \uparrow z_0} \mathbb{E}[Y|Z = z, X_{\setminus Z} = x]}{\lim_{z \downarrow z_0} \mathbb{E}[T|Z = z, X_{\setminus Z} = x] - \lim_{z \uparrow z_0} \mathbb{E}[T|Z = z, X_{\setminus Z} = x]}.\end{aligned}$$

Second, we require this local treatment effect to generalize to the full population at $(Z = z_0, X_{\setminus Z} = x)$, such that the CLATE, $\tau^{CLATE}(Z = z_0, X_{\setminus Z} = x)$, is equal to the CATE, $\tau(Z = z_0, X_{\setminus Z} = x)$. To this end, in addition to standard fuzzy RD assumptions, we also apply the *conditional constant effects* (CCE) restriction (Angrist, 2004), also called *conditional effect ignorability* (CEI) by Angrist and Fernández-Val (2013).

Assumption 4 (CCE/CEI) *The treatment effect $Y(1) - Y(0)$ is mean independent of compliance type G given observed covariates X:*

$$\mathbb{E}[Y(1) - Y(0) \mid X, G] = \mathbb{E}[Y(1) - Y(0) \mid X] = \tau(X).$$

Note that since the compliance type G (complier, always-taker, or never-taker) dictates the treatment assignment, Assumption 4 leads to $\mathbb{E}[Y(1) - Y(0)] \perp T \mid X$. As noted in Angrist and Fernández-Val (2013), Assumption 4 describes the case when heterogeneity in treatment effects is solely a function of the observed covariates:

$$Y(1) = Y(0) + \tau(x) + \nu,$$

where $\mathbb{E}[\nu \mid G, X] = \mathbb{E}[\nu \mid X] = 0$. That is, the CCE/CEI assumption states that though treatment effects may not be constant, they are the same for two units with the same observed covariates $X = x$, regardless of their compliance type.

Under this assumption, we can extrapolate from the compliers to the full population at the threshold:

$$\tau^{CLATE}(Z = z_0, X_{\setminus Z} = x) = \tau(Z = z_0, X_{\setminus Z} = x).$$

Note that this CCE/CEI assumption is not overly restrictive, as it still allows treatment effects to vary heterogeneously as a function of x. Moreover, the CCE/CEI assumption is similar to assumptions required for other novel machine learning methods for treatment effect estimation, specifically the assumption of additive confounding required for both DeepIV (Hartford et al., 2017), a deep method for instrumental variables regression, and ModeIV (Hartford et al., 2021), a recent method for combining multiple IV estimates.[6]

Under this additional assumption on the underlying regression discontinuities, we can use the observational dataset $\mathcal{D}$ to obtain unconfounded treatment effect estimates at points along discontinuities, and thus construct a set $\mathcal{U}$ directly from $\mathcal{D}$.

Finally, given a set of CATE estimates $\hat{\tau}(x)$ at points falling on regression discontinuities, there are two paths to extrapolating these estimates to the rest of $\mathcal{X}$. First, there is the option of directly extrapolating the CATE estimates. Alternatively, a bias correction $\beta(x)$ can be estimated for a conditioned-on-observables estimand $\tau^{obs}(x)$, that is, $\beta(x) = \tau(x) - \tau^{obs}(x)$. The estimate of the bias can be obtained using the difference

$$\hat{\beta}(x) = \hat{\tau}(x) - \hat{\tau}^{obs}(x).$$

Then, this bias correction can be extrapolated across the full support $\mathcal{X}$. We might expect the former approach to be preferable when the CATE $\tau(x)$ is smoother than the bias

6. Note that CCE/CEI is a reasonable assumption in practice and is the basis of much of empirical work (Angrist, 2004; Hartford et al., 2017). For instance, Hartford et al. (2017) argue that CCE/CEI assumption is a weaker assumption than conditional ignorability, as we could still have $\{Y(1), Y(0)\} \not\perp T \,|\, X$. In this case, techniques based on conditional ignorability, such as matching and propensity score reweighting, will produce biased estimates.

correction $\beta(x)$, and the latter approach to be preferable when the bias correction $\beta(x)$ is smoother than the CATE $\tau(x)$. The fact that there are two candidate approaches to extrapolating from the embedded regression discontinuities to the ambient covariate space $\mathcal{X}$ raises a natural model selection problem, which we address through model averaging.

3. Method

We introduce `DEE` (Discover-Estimate-Extrapolate), a method for estimating the CATE $\tau(x)$ given observational data containing regression discontinuities satisfying the CCE/CEI assumption along with the standard RD assumptions above. At a high level of abstraction, `DEE` has three steps, which correspond to the three methodological questions of *discovering* RDs, *estimating* CATE at the discovered RDs, and *extrapolating* these estimates to the larger population:

1. **Automated RD Discovery**: We first apply LoRD^3 (Herlands et al., 2018) to automatically discover local discontinuities in the treatment propensity $P(T = 1 \mid X = x)$.

2. **CATE Estimation**: Next, we estimate τ^{CLATE} along the discovered regression discontinuities, using the novel repair procedure described below. Under the CCE/CEI assumption, these complier treatment effect estimates generalize to the full population, and are only local in covariate space, and not local to the complier subpopulation.

3. **Extrapolation**: Finally, we apply Gaussian process regression to extrapolate[7] the RD CATE estimates to the full support $\mathcal{X}$, integrating the multiple RD estimates with a conditioned-on-observables estimator through model averaging.

We proceed to describe each of the three steps of `DEE` in greater detail.[8]

3.1 Automated RD Discovery

Given the observational dataset $\mathcal{D}$, `DEE` first applies LoRD^3 (Herlands et al., 2018) to automatically discover local RDs. LoRD^3 characterizes discontinuities as unexpected "jumps" in the probability of treatment, which is assumed to be a smooth function except at the discontinuities. It therefore models $t_i = f(x_i) + \epsilon_i$, where f belongs to a class of functions that are sufficiently expressive to represent the general smooth curvature of the treatment propensity function without fully capturing any potential (non-smooth) discontinuities. Thus, in the absence of a discontinuity, we expect the noise to have constant mean $E[\epsilon_i] = \mu_0$, e.g., if f is unbiased, then $\mu_0 = 0$. However, local to a discontinuity, f will overestimate t_i for the x_i on one side of the discontinuity ($E[\epsilon_i] > \mu_0$), and underestimate those on the other side ($E[\epsilon_i] < \mu_0$). See Figure 1 for graphical intuition.

LoRD^3 exploits the unique residual pattern created by a discontinuity to discover its presence. More precisely, it first computes $\hat{f}$, a global estimate of f such that $t_i = \hat{f}(x_i) + r_i$. Then for each data point $(x_i, t_i, y_i) \in \mathcal{D}$, it builds $s_{i,k}$, the local k-neighborhood around x_i,

7. In our setting, we use Gaussian process regression as it is a standard and natural approach to model correlated data, and we expect the CATE estimates to be spatially correlated. However, our method can allow for any other approach for extrapolation.
8. Our code for `DEE` is available at https://github.com/ssomanch/DEE.

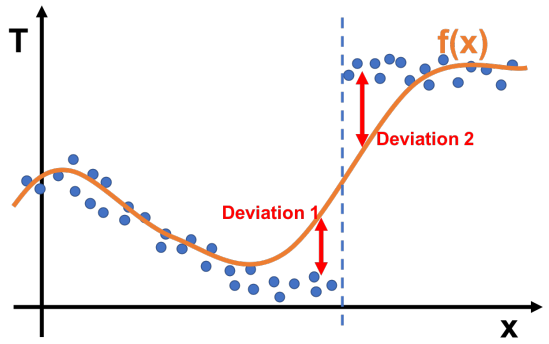

Figure 1: Illustration of a one-dimensional RD design (dashed line) from Herlands et al. (2018). Blue dots: treatment propensity for each x_i. Orange curve: $\hat{f}(x)$.

and selects a bisecting hyperplane $v_{i,k}$ (from a set of $k-1$ options) that it determines is most likely to reflect an underlying regression discontinuity in the neighborhood. Any potential discontinuity will partition the neighborhood into two groups g^1 and g^2. Therefore, LoRD^3 evaluates any candidate hyperplane with a log-likelihood ratio (LLR) test statistic, which computes how much more likely is the evidence for $E[r_j] = \mu_1, \forall j \in g^1$, and $E[r_j] = \mu_2, \forall j \in g^2$ (the alternative hypothesis that the candidate hyperplane creates a pattern of residuals consistent with a discontinuity) than the evidence for $E[r_j] = \mu_0, \forall j \in s_{i,k}$ (the null hypothesis of no discontinuity).

The bisection that provides the most evidence against H_0 in neighborhood $s_{i,k}$ becomes its most likely discontinuity $v_{i,k}$ with corresponding LLR test statistic $LLR_{i,k}$. This produces $\mathcal{L}^{full} = \{(x_i, v_{i,k}, LLR_{i,k})\}, \forall s_{i,k}$, the full set of discontinuities discovered by LoRD^3. Randomization testing, density ("bunching") testing (McCrary, 2008), and placebo testing are then all carried out to filter out any neighborhood $s_{i,k}$ whose corresponding potential discontinuity $v_{i,k}$ is not statistically significant (adjusting for multiple hypothesis testing) or shows evidence of violating the assumptions of an RD design. This process yields a set of M valid discontinuities discovered by LoRD^3, $\mathcal{L} = \{(x_j, v_j, LLR_j)\}_{j=1}^{M}$, where M can either be prespecified or determined based on the number of discontinuities which are valid and significant at level α. Algorithm 2 in Appendix A provides a summary of the LoRD^3 procedure; we refer readers to Herlands et al. (2018) for a more detailed exposition.

3.2 CATE estimation

Our primary goal is to extrapolate the CATE, $\tau(x)$, over $\mathcal{X}$. We expect that our extrapolation will improve if we can identify all of the discontinuities embedded in $\mathcal{X}$ and estimate $\tau(x)$ over the full extent of each discontinuity. Therefore, given the set $\mathcal{L} = \{(x_j, v_j, LLR_j)\}_{j=1}^{M}$ of validated local discontinuities from LoRD^3, `DEE` proceeds to estimate the CATE along the discovered discontinuities. We treat CATE estimation as a separate step in our inferential procedure, though we could, in theory, combine RD discovery and CATE estimation into a single step. For example, while scanning each data instance $x_i \in \mathcal{D}$ during RD discovery, we could estimate each $\tau(x_i)$ using a local two-stage least squares (2SLS) estimator fit us-

ing the k-nearest neighbors of x_i, and then directly proceed to extrapolate the estimated effects using the discovered local RDs in $\mathcal{L}$. More specifically, a 2SLS estimate of $\tau(x_i)$ would use the data from the neighborhood, s_i, of x_i. Each element $j \in s_i$ is represented by (x_j, t_j, y_j, g_j), where the binary indicator $g_j = \mathbb{1}_{\{j \in g^1\}}$ captures on which side of the discontinuity (v_i) data point x_j lands. The first stage

$$T_j = \nu g_j + f(x_j) + \epsilon_j^{(T)} \tag{1}$$

instruments (the endogenous variable) T_j with (the exogenous variable) g_j, providing

$$\hat{T}_j = \hat{\nu} g_j + \hat{f}(x_j), \tag{2}$$

the predicted value of T_j. Intuitively, $\hat{T}_j$ is not confounded by unobservables because it only uses variation from the exogenous g_j, while controlling for the observed features x_j; all other sources of variation in T_j are contained in the estimate of the residuals, which are excluded from (2). The second stage is then an estimate of

$$y_j = \tau \hat{T}_j + \lambda x_j + \epsilon_j^{(y)}. \tag{3}$$

Intuitively, instead of using all the variation contained in T_j to compute $\hat{\tau}$, the 2SLS procedure first obtains (from (2)) and then uses (in (3)) only the exogenous variation, as estimated by $\hat{T}_j$.

While the neighborhood-based 2SLS procedure allows for the local estimation of $\tau(x_i)$, this approach faces significant drawbacks. That is, $\mathcal{L}$ likely includes a number of very similar local discontinuities[9] (see for example the left panels of Figures 2 and 14). Since the k-nearest neighbors for two similar discontinuities $x_i \approx x_j$ will overlap substantially, the sampling errors in the local 2SLS estimates $\hat{\tau}(x_i), \hat{\tau}(x_j)$ computed using each instance's neighbors will be highly correlated. Using Gaussian process regression to extrapolate directly from such estimates risks learning local correlation in the sampling error (due to overlapping k-NN balls), instead of learning the desired correlation structure in $\tau(x)$. Attempting to address this issue by applying a LLR threshold to select fewer local RDs from $\mathcal{L}$ is insufficient, as this approach still allows selected discontinuities to have substantially overlapping neighborhoods, while also leaving large regions of the true discontinuity uncovered (see for example the center panels of Figures 2 and 14).

As such, instead of using the local RDs in $\mathcal{L}$ directly, `DEE` applies a novel repair procedure that reduces the original set $\mathcal{L}$ of M local discontinuities to a set $\mathcal{U}$ of $M' \leq M$ local discontinuities. While constructing $\mathcal{U}$, the repair procedure assigns each local discontinuity in $\mathcal{U}$ a disjoint *index set* containing those data instances $(x_i, t_i, y_i) \in \mathcal{D}$ that will be used for CATE estimation[10]. Finally, `DEE` applies the local 2SLS estimator to estimate the CATE at each of the local discontinuities in $\mathcal{U}$. We proceed to detail each of these substeps.

9. Two local discontinuities (x_i, v_i, LLR_i) and (x_j, v_j, LLR_j) are similar if $\|x_i - x_j\| \approx 0$ (the discontinuities' centers are near to each other) and $|v_i^T v_j| \approx 1$ (their normal vectors define similar hyperplanes).

10. One can equivalently understand each index set as defining a non-rectangular (but convex) uniform kernel, which we use for a subsequent local 2SLS estimation step.

3.2.1 Repairing $\mathcal{L}$ and Constructing Index Sets

Instead of directly computing the CATE at each local discontinuity included in $\mathcal{L}$, DEE first applies the novel "Voronoi ∩ KNN" repair procedure presented in Algorithm 1. This repair procedure greedily selects a subset of local discontinuities from $\mathcal{L}$, selecting the most likely discontinuities in $\mathcal{L}$ (i.e., those with the highest log-likelihood ratios) while ensuring that the selected discontinuities approximately cover the discovered discontinuities in $\mathcal{L}$. Note that the Voronoi ∩ KNN approach was chosen to satisfy two desired criteria for our repair procedure. First, the KNN procedure ensures we use only the data points near the discontinuity to estimate CATE. Second, the Voronoi procedure minimizes the possibility of choosing very similar discontinuities that could hinder learning the desired correlation structure of $\tau(x)$, as explained above. Therefore, a combination of KNN and Voronoi procedure helps us utilize only the points close to the discontinuity for estimating the treatment effect while simultaneously ensuring that these points are distinct (i.e., forming non-overlapping k-NN balls).

In addition to selecting a subset of the discontinuities in $\mathcal{L}$, this procedure also assigns each selected discontinuity an index set containing those instances from $\mathcal{D}$ that are used for CATE estimation. The index set for discontinuity (x, v, LLR) contains those points in $\mathcal{D}$ which fall in the intersection $VK[x] = V[x] \cap K[x]$ of x's Voronoi cell, $V[x]$ (i.e., those instances in $\mathcal{D}$ that are closer to x than any other discontinuity), and its k-NN ball, $K[x]$ (i.e., the k-nearest neighbors of x in $\mathcal{D}$). Using $VK[x]$ as the index set for discontinuity (x, v, LLR) ensures that the CATE is estimated using points that are local to the discontinuity, and that all of the index sets are disjoint (see, for example, the right panels of Figures 2 and 14). In Appendix D.3, we demonstrate the advantages of using the combined $VK[x]$ in Algorithm 1 as compared to KNN-only and Voronoi-only procedures. Note that while $K[x]$ is static, and can be precomputed for each x, $V[x]$ is dynamically updated during the execution of Algorithm 1 as additional discontinuities are added to $\mathcal{U}$.

3.2.2 Estimating the CATE at Each Discontinuity in $\mathcal{U}$

Algorithm 1 yields a set $\mathcal{U}$ of local discontinuities that approximately cover the original set of discovered discontinuities, and assigns each discontinuity $x \in \mathcal{U}$ a disjoint index set $VK[x]$. To estimate the CATE at $x \in \mathcal{U}$, DEE applies the local 2SLS estimator described in (1)-(3), computing the 2SLS estimate using just those instances in the Voronoi ∩ KNN index set of x. All instances in $VK[x]$ are within x's k-NN ball, so following Becker et al. (2013), eqn. 10, this yields a non-parametric estimate of $\tau^{CLATE}(Z, X_{\setminus Z})$.[11] However, under the CCE/CEI assumption, these CLATE estimates generalize, with $\tau^{CLATE}(Z, X_{\setminus Z}) = \tau^{CATE}(Z, X_{\setminus Z})$. As such, this procedure yields a set of CATE estimates $\mathcal{U} = \{(x_k, \hat{\tau}(x_k), \hat{\sigma}(x_k)\}_{k=1}^{M'}$, where $\hat{\sigma}(x_k)$ is the standard error of the CATE estimate $\hat{\tau}^{CATE}(x_k)$.

3.3 Nonparametric Extrapolation from $\mathcal{U}$ to $\mathcal{X}$

Given the set of discovered, repaired discontinuities and their associated CATE estimates $\mathcal{U} = \{(x_k, \hat{\tau}(x_k), \hat{\sigma}(x_k)\}_{k=1}^{M'}$, the final step of DEE is to extrapolate from $\mathcal{U}$ to the rest of $\mathcal{X}$.

11. Again, this is in contrast to the typical treatment of RD designs, in which an unconditional local average effect is estimated at $Z = z_0$ via local regression with a kernel that is functionally independent of $X_{\setminus Z}$.

Algorithm 1: Voronoi ∩ KNN repair procedure used by `DEE`. Greedily selects discontinuities from the original set $\mathcal{L}$, and assigns each discontinuity an index set.

Input: $\mathcal{L}$
Parameter: k – Number of nearest neighbors for KNN.
Parameter: t – Minimum number of neighbors on each side of the discontinuity.
Output: $\mathcal{U}$

$\mathcal{U} \leftarrow \{\}$
Compute $K[x]$ for each (x, v, LLR) in $\mathcal{L}$.
while $|\mathcal{L}| > 0$ **do**
  Remove the discontinuity (x, v, LLR) with maximum LLR from $\mathcal{L}$.
  $\mathcal{U}' \leftarrow \mathcal{U} \cup \{(x, v, LLR)\}$
  For each $x \in \mathcal{U}'$, update $V[x]$ by computing the Voronoi partition of $\mathcal{D}$ into $\mathcal{U}'$, and update $VK[x] = V[x] \cap K[x]$.
  if *for all* $(x, v, LLR) \in \mathcal{U}'$, $VK[x]$ *includes at least* t *instances on each side of its separating hyperplane* **then**
    $\mathcal{U} \leftarrow \mathcal{U}'$ # Include (x, v, LLR) with maximum LLR in the returned set $\mathcal{U}$
for $(x, v, LLR) \in \mathcal{U}$ **do**
  Center x on the local discontinuity by (i) projecting $VK[x]$ onto the separating hyperplane defined by v, then (ii) updating x to the mean of these projections.

As previously mentioned, there are two natural approaches to extrapolation. We describe each of these approaches in turn, then several candidate strategies for model averaging.

3.3.1 Direct Extrapolation of $\hat{\tau}$

The first approach we consider is direct extrapolation of the treatment effect estimates. For this direct extrapolation, we model the estimated CATE $\hat{\tau}(x_k)$ as a Gaussian process (GP) with noise variance $\hat{\sigma}(x_k)^2$:

$$\begin{aligned} \hat{\tau}(x_k) &= \widetilde{\tau}(x_k) + \epsilon_k, \\ \epsilon_k &\sim \mathcal{N}(0, \hat{\sigma}(x_k)^2), \\ \widetilde{\tau}(x_k) &\sim GP(\mu_\tau, k_{\Theta_\tau}). \end{aligned}$$

We let the GP prior mean μ_τ be a free parameter, along with the kernel parameters Θ_τ, and fit the GP using marginal likelihood maximization on the dataset $\mathcal{U}$, with target $y = \hat{\tau}(x_k)$. To extrapolate the CATE to $x^* \in \mathcal{X}$, we simply compute the posterior distribution of $\widetilde{\tau}(x^*)$, conditioning on the estimates in $\mathcal{U}$.

3.3.2 Extrapolating a Bias Correction

As an alternative to directly extrapolating the treatment effect, we can instead use the CATE estimates in $\mathcal{U}$ to debias a conditioned-on-observables estimate. There are four steps in this approach:

1. Fit a conditioned-on-observables estimator to $\mathcal{D}$ to obtain $\hat{\tau}^{obs}(x)$.

2. Estimate a bias correction for each point in $\mathcal{U}$ as

$$\hat{\beta}(x_k) = \hat{\tau}^{CATE}(x_k) - \hat{\tau}^{obs}(x_k).$$

3. Model this bias correction using a GP:

$$\begin{aligned}\hat{\beta}(x_k) &= \widetilde{\beta}(x_k) + \gamma_k, \\ \gamma_k &\sim \mathcal{N}(0, \hat{\sigma}(x_k)^2), \\ \widetilde{\beta}(x_k) &\sim GP(\mu_\beta, k_{\Theta_\beta}).\end{aligned}$$

   Again, we let the GP prior mean μ_β be a free parameter, along with the kernel parameters Θ_β, and fit the GP using marginal likelihood maximization on the set $\mathcal{U}$, with target $y = \hat{\beta}(x_k)$.

4. To extrapolate the CATE to $x^* \in \mathcal{X}$, (i) compute the posterior distribution of $\widetilde{\beta}(x^*)$, conditioning on $\mathcal{U}$, then (ii) add $\hat{\tau}^{obs}(x^*)$ to the posterior mean to recover the final CATE estimate.

The conditioned-on-observables estimate $\hat{\tau}^{obs}(x)$ can be obtained using a variety of estimators, including direct estimation of the treatment and control regression functions $\mathbb{E}[Y(1) \mid T = 1, X = x]$ and $\mathbb{E}[Y(0) \mid T = 0, X = x]$. Alternatively, tree-based methods can be used, including causal trees (Athey and Imbens, 2016) or causal forests (Wager and Athey, 2018). These methods achieve "honest" estimation through sample splitting, using separate data partitions for tree learning and for effect estimation to ensure that the trees provide unbiased estimates of treatment effects in their leaves. Since causal forests represent the state-of-the-art for CATE estimation when potential outcomes are, in fact, conditionally independent of treatment, we use causal forests to estimate $\tau^{obs}(x)$ in our experiments – though we stress that in our context the causal forest estimates will still be confounded.

As there are two competing approaches to nonparametric extrapolation, we face a model selection and evaluation problem. To address this problem, we consider several potential model averaging strategies, and introduce BLOOCV as a novel measure of predictive fit when nonparametrically extrapolating from points on manifolds into an ambient space.

3.3.3 Model Selection and Averaging

Let $\mathcal{M}_\tau$ denote the direct CATE extrapolation model, and $\mathcal{M}_\beta$ denote the bias correction model. Instead of choosing one of these models *a priori*, we can adaptively blend these models via model averaging. The weight on model $\mathcal{M}_i$ in the average can be expressed as:

$$w(\mathcal{M}_i \mid y) = \frac{m(y \mid \mathcal{M}_i)p(\mathcal{M}_i)}{\sum_{i \in \{\tau, \beta\}} m(y \mid \mathcal{M}_i)p(\mathcal{M}_i)}, \tag{MA}$$

where $m(y \mid \mathcal{M}_i)$ denotes the likelihood and $p(\mathcal{M}_i)$ denotes the model prior. In practice, we use a uniform prior over the two candidate models.

When we let $m(y \mid \mathcal{M}_i, X)$ be the marginal likelihood, Equation MA just describes Bayesian model averaging. However, following Eklund and Karlsson (2007), we can also let $m(y \mid \mathcal{M}_i, X)$ be a predictive likelihood. For example, instead of using the marginal

likelihood, we can use the leave-one-out cross-validation (LOOCV) predictive likelihood. Specifically, let y denote the target of the Gaussian Process (subsuming both $y = \hat{\tau}(x)$ and $y = \hat{\beta}(x)$), and let μ_i and σ_i denote the leave-one-out posterior predictive mean and variance for the i^{th} training instance. Then the LOOCV log predictive likelihood is given by equations 5.10 and 5.11 in Rasmussen and Williams (2005) as

$$ll_{LOOCV}(y) = -\frac{1}{2}\sum_{i=1}^{N}\left[\left(\frac{y_i - \mu_i}{\sigma_i}\right)^2\right] - \underbrace{\sum_{i=1}^{N}\left[\log(\sigma_i\sqrt{2\pi})\right]}_{PPV}, \tag{LOOCV}$$

where PPV is the posterior predictive variance. While the LOOCV likelihood is a commonly used measure of predictive fit, it is not appropriate in our context, since points in $\mathcal{U}$ are not drawn from the underlying marginal distribution $P(X)$ over $\mathcal{X}$. Instead, points in $\mathcal{U}$ fall on discovered RDs embedded in $\mathcal{X}$. As such, the LOOCV likelihood does not meaningfully estimate predictive fit over the underlying marginal distribution $P(X)$.

Furthermore, asymptotically the LOOCV likelihood will fail to concentrate weight on the desired model. Consider the LOOCV log-likelihood ll_{LOOCV} for our two candidate models $\mathcal{M}_\tau$ and $\mathcal{M}_\beta$. We expect $ll_{LOOCV}(y)$ to be larger for the model with smaller posterior predictive variances, due to the contribution of the PPV term. That is, if one of our models yields substantially more precise posterior predictions than the other, we expect that the ll_{LOOCV} will be larger for the more precise model, allowing us to concentrate weight on that model. Unfortunately, in our context, differences in the leave-one-out posterior predictive variances across our two models vanish asymptotically, as shown in Theorem 1 below. Specifically, under the conditions of Lederer et al. (2019), Corollary 3.2, the LOOCV posterior predictive variances at x_k of each of our two models will converge to the shared noise variance $\hat{\sigma}(x_k)^2$. As such, in the limit of $|\mathcal{U}| \to \infty$, LOOCV will fail (in expectation) to differentiate between our two candidate models. To address this issue, we introduce Buffered Leave-One-Out Cross Validation (BLOOCV) as an alternate measure of predictive fit when extrapolating from points on a manifold into an ambient space.

3.3.4 BLOOCV: An Alternate Measure of Predictive Fit

The key difference between LOOCV and BLOOCV is in which data instances are excluded when computing the posterior predictive distributions. In computing the posterior predictive distribution for the k^{th} data instance, LOOCV only excludes that specific data instance, while BLOOCV excludes all instances that fall within a buffer distance b of x_k. To highlight this dependence on b, we denote the BLOOCV log-likelihood for the k^{th} data instance by $ll_{BLOOCV}(y_k; b)$.

Applying BLOOCV requires specifying a method for choosing the buffer radius b. Given that our aim is to approximate generalization performance over $P(X)$, we choose b based on the distance between points sampled from $P(X)$ and points in $\mathcal{U}$. Specifically, fixing $\mathcal{U}$, we let $\hat{D}$ denote the distribution of distances induced by sampling $x \sim P(X)$ and computing $\min_{x_k \in \mathcal{U}} \|x - x_k\|$. Then, for the k^{th} instance in $\mathcal{U}$, we take as our measure of predictive fit the average BLOOCV log-likelihood $\mathbb{E}_{b\sim\hat{D}}[ll_{BLOOCV}(y_k; b)]$, averaging over buffer distances b drawn from this derived distribution $\hat{D}$. Finally, to measure the BLOOCV log-likelihood

over the full set $\mathcal{U}$, we simply take the sum of each of the M' average log-likelihoods

$$\sum_{k=1,\cdots,M'=|\mathcal{U}|} \mathbb{E}_{b\sim\hat{D}}[ll_{BLOOCV}(y_k;b)]. \tag{BLOOCV}$$

We consider two variants of BLOOCV, corresponding to two strategies for estimating $\mathbb{E}_{b\sim\hat{D}}[ll_{BLOOCV}(y_k;b)]$ for the k^{th} instance:

1. **1-MC**: Our first estimation strategy has lower computational expense but higher variance. In this variant, for each instance $k = 1, \cdots, M'$, we approximate $\mathbb{E}_{b\sim\hat{D}}[ll_{BLOOCV}(y_k;b)]$ using a single Monte Carlo sample of b from $\hat{D}$. Specifically, for each $k = 1, \cdots, M'$, we:
   (a) Sample $b' \sim \hat{D}$ by (i) choosing x uniformly at random from $P(X)$, then (ii) letting $b' = \min_{x_j\in\mathcal{U}} \|x - x_j\|$.
   (b) Approximate $\mathbb{E}_{b\sim\hat{D}}[ll_{BLOOCV}(y_k;b)] \approx ll_{BLOOCV}(y_k;b')$.
2. **Weighted variant**: The second approach trades off greater computational expense for lower variance. In this variant, instead of a single Monte Carlo sample from $\hat{D}$, for each k we exactly compute the empirical expectation $E_{b\sim\hat{D}}[ll_{BLOOCV}(y_k;b)]$. This expectation can be computed as a weighted sum of $\mathcal{O}(|\mathcal{U}|)$ buffered log-likelihoods as follows:
   (a) Compute the empirical CDF $F_{\hat{D}}(d) = P(\hat{D} \leq d)$.
   (b) Let $d_0 \leq d_1 \leq \cdots \leq d_{M'-1}$ denote the ordered pairwise distances $\|x_k - x_j\|$ between x_k and $x_j \in \mathcal{U}$. For notational convenience let $d_{M'} = \infty$.
   (c) Then the empirical expectation $E_{b\sim\hat{D}}[ll_{BLOOCV}(y_k;b)]$ can be computed exactly as a weighted sum of $|\mathcal{U}|$ buffered log-likelihoods:

$$E_{b\sim\hat{D}}[ll_{BLOOCV}(y_k;b)] = \sum_{j=0}^{M-1} P(d_j \leq \hat{D} < d_{j+1}) ll_{BLOOCV}(y_k;d_j)$$

   While $E_{b\sim\hat{D}}[ll_{BLOOCV}(y_k;b)]$ can thus be computed as a weighted sum of $\mathcal{O}(|\mathcal{U}|)$ buffered log-likelihoods, computing each log-likelihood requires $\mathcal{O}(|\mathcal{U}|^3)$ time. As such, this approach naively takes $\mathcal{O}(|\mathcal{U}|^5)$ time (ignoring potential optimizations), and is thus potentially impractical for large $\mathcal{U}$.

4. Theoretical Results

Before demonstrating the performance of our `DEE` approach in simulation, we show that BLOOCV addresses the asymptotic degeneracy of LOOCV. First, we describe the asymptotic degeneracy of the LOOCV likelihood. Then, we show in Theorem 1 that BLOOCV does not suffer from the same degeneracy.

To start, consider two zero-mean GPs, GP_1 and GP_2, sharing common homoskedastic noise variance σ_n^2. Moreover, assume these GPs have isotropic Gaussian kernel functions $k_1(q) = \exp\left(-\|q\|^2/(2b_1^2)\right)$ and $k_2(q) = \exp\left(-\|q\|^2/(2b_2^2)\right)$, where the length scale $b_1 > b_2$.

Finally, assume that training instances $\{x_i\}_{i=1}^N$ are sampled from an RD manifold. Then, under the conditions of Lederer et al. (2019), Corollary 3.2, the LOOCV posterior predictive variance at x_* for each of our two models will converge to the shared noise variance σ_n^2. This is problematic, since it implies that in the asymptotic regime, LOOCV likelihood model selection or averaging will fail to differentiate between GP_1 and GP_2. Fortunately, buffering allows us to mitigate this degeneracy and to correctly differentiate between models even in the asymptotic regime.

To show that buffering mitigates this asymptotic degeneracy, we consider a case simplified for analytic tractability. Our simplification is motivated by the following observation: buffering guarantees that there are no training instances x_i within the buffer radius ρ of the held-out instance x_*. We can therefore simplify the problem while retaining this key property by assuming all training instances x_* are exactly at distance ρ from the test instance. By further assuming the training instances are distributed symmetrically on the surface of the radius ρ ball, we can analytically show that, under these simplifying assumptions, increasing the buffer radius ρ addresses the asymptotic degeneracy of LOOCV.

Theorem 1 (Buffering increases the asymptotic PPV difference) *Consider two GPs, GP_1 and GP_2, sharing common homoskedastic noise variance σ_n^2. Moreover, assume that these GPs have isotropic Gaussian kernel functions $k_1(q) = \exp\left(-\|q\|^2/(2b_1^2)\right)$ and $k_2(q) = \exp\left(-\|q\|^2/(2b_2^2)\right)$, where the length scale $b_1 > b_2$. For analytic tractability, assume all training points $\{x_i\}_{i=1}^N$ are at distance ρ from the test point x_*, the distance distribution from each training point to all other training points is the same (assume pairwise distances are drawn $q \sim Q$), and $\mathbb{E}[q] < \rho\sqrt{2}$. Then the difference in posterior predictive variances $\sigma_1^2(x_*) - \sigma_2^2(x_*)$ satisfies the following:*

1. *If $\rho = 0$, then $\lim_{N\to\infty}\left(\sigma_2^2(x_*) - \sigma_1^2(x_*)\right) = 0$.*
2. *If $\rho > 0$, then $\lim_{N\to\infty}\left(\sigma_2^2(x_*) - \sigma_1^2(x_*)\right) > 0$.*
3. *$\lim_{N\to\infty}\left(\sigma_2^2(x_*) - \sigma_1^2(x_*)\right)$ is bounded below by a function $f(\rho)$, which increases from $f(0) = 0$ at $\rho = 0$ through $\rho = b_1 b_2\sqrt{\frac{2\log\left(\frac{b_1}{b_2}\right)}{(2-\theta^2)(b_1^2-b_2^2)}} > 0$ (see proof for definition of θ).*

A proof of Theorem 1 is provided in Appendix B. The key result is implication 3, which states that (under the given conditions) buffering increases the minimum guaranteed difference in the asymptotic posterior predictive variances between GP_1 and GP_2. As such, BLOOCV addresses the degeneracy of LOOCV, and thus provides a measure of predictive fit that is appropriate when extrapolating from points on manifolds into an ambient space.

5. Related Work

There has been a growing literature using statistical learning methods to provide data-driven approaches for estimating heterogeneous treatment effects. Some of this work adapts existing machine learning algorithms, including regularized regression (Imai and Ratkovic, 2013; Tian et al., 2014; Weisberg and Pontes, 2015), regression trees (Su et al., 2009; Athey and Imbens, 2016), random forests (Foster et al., 2011; Wager and Athey, 2018), boosting (Powers et al., 2018), neural networks (Shalit et al., 2017), Bayesian methods (Hill, 2011; Green

and Kern, 2012; Alaa and van der Schaar, 2017), and other ensemble approaches (Grimmer et al., 2017), to causal inference tasks. In addition, there has been a collection of meta-learning methods that allow the use of a broad class of nonparametric estimators; see Curth and van der Schaar (2021) for a unifying framework, taxonomy, and general theory for meta-learning methods. At a high level, these methods detail which relationships to estimate and how to combine them to obtain heterogeneous causal estimates with desirable properties: e.g., double-robustness (Kennedy, 2020), quasi-oracle bounds for binary treatments (Nie and Wager, 2020) or for structured treatments (Kaddour et al., 2021), or adaptation to structural properties such as the sparsity or smoothness of the CATE (Künzel et al., 2019).

While the literature on treatment effect heterogeneity continues to grow, the vast majority of the work (including the papers referenced above) relies on the assumption of strong conditional ignorability (or the even stronger assumption of a randomized, controlled experiment), which assumes that treatment is conditionally independent of potential outcomes, $\{Y(0), Y(1)\} \perp T \mid X$. This assumption is critical for their ability to conduct causal inference, leading to biased treatment effect estimates when the assumption is violated (e.g., when there is non-random selection into treatment). In contrast, `DEE` aims to discover portions of the data for which unbiased estimates can be obtained even when strong conditional ignorability does not hold.

There is a small subset of methods that attempt to justify the conditional independence of the treatment assignment T and the potential outcomes $\{Y(0), Y(1)\}$. As described above, Kallus et al. (2018) assume that supplemental experimental data (restricted to a subset of $\mathcal{X}$) is available, and can be used to unconfound T. Without access to an unconfounded sample, other methods attempt to learn a function to adjust the original covariates, in order to make the treatment conditionally independent of the potential outcomes: Gultchin et al. (2020) carry out an optimization approach to learn a covariate adjustment function over a subset of variables, while Yao et al. (2018) use deep learning to produce a new representation of the original covariate space $\mathcal{X}$. However, Yao et al. (2018) also make the assumption of strong conditional ignorability, while Gultchin et al. (2020) make an alternative strong assumption that there exist an observed auxiliary variable W and a subset of observed covariates $X^* \subseteq X$ for which, conditional on X^*, all paths from W to Y are mediated by the treatment T.

Our `DEE` approach instead discovers RD designs in the data to obtain unbiased local treatment effect estimates, and extrapolates from these to better estimate CATE across the entire covariate space $\mathcal{X}$, as shown by our experimental results below. RDs have been characterized as requiring fewer assumptions than most causal inference techniques and are arguably most similar to true gold-standard randomized experiments (Lee and Lemieux, 2010). Rather than requiring strong conditional ignorability for unbiased estimation, `DEE` relies only on the weaker assumption of CCE/CEI, to generalize from the compliers at each discovered RD to the full population.

6. Experimental Results

To evaluate the performance of our `DEE` method, we present results on two synthetic datasets, then apply `DEE` to a recent problem from the economic development literature (Asher and Novosad, 2020). Within each simulation, we compare our method's performance varying the

prior variances in $\tau(x)$ and in $\beta(x)$. When $\tau(x)$ is less variable than $\beta(x)$, we expect direct extrapolation of $\hat{\tau}$ to outperform bias correction; conversely, when $\beta(x)$ is less variable than $\tau(x)$, we expect bias correction to outperform direct extrapolation. As such, in these cases, we aim to show that model averaging concentrates weight on the model with less variance. Finally, when these functions have similar variances, we anticipate that model averaging will deliver gains over model selection.

In addition to evaluating performance across different $\tau(x)$ and $\beta(x)$ variance regimes, we also benchmark `DEE` against several alternatives. First, we compare our extrapolated estimates to the input causal forest estimates. Second, by treating those instances in the neighborhood of a discovered regression discontinuity as an unconfounded sample, we compare our approach to Kallus et al.'s method of experimental grounding (Kallus et al., 2018).[12] Third, we attempt to apply a method for exploiting multiple discontinuities along a common running variable (Cattaneo et al., 2020). This method requires observation of unit-level threshold assignments – but such threshold assignments are clearly latent when the RDs are themselves unknown *a priori*. We describe how we address this and other implementation challenges in detail in Appendix C. Finally, to highlight the importance of our novel Voronoi $\cap$ KNN repair algorithm, we also compare our procedure to a variant of `DEE` that directly extrapolates from $\mathcal{L}$, instead of applying Algorithm 1 to obtain and extrapolate from $\mathcal{U}$.

6.1 Simulation 1: Smooth τ and β

The data generating process for our first simulation is presented in Appendix D.1. In this simulation, the data generating process includes two binary RD instruments[13], Z_1 and Z_2, each associated with a separate complier subpopulation. Cases in the complier subpopulation C_1 are treated if and only if $Z_1 = 1$, and cases in the complier subpopulation C_2 are treated if and only if $Z_2 = 1$.

As noted in Appendix D.1, this data generating process can be shown to satisfy the CCE/CEI assumption. Beyond simply noting that this assumption is satisfied, this simulation also clarifies its key role in our procedure. In this simulation, our method discovers and estimates conditional local average effects along the discontinuities associated with the RD instruments Z_1 and Z_2. As such, $\mathcal{U}$ contains conditional local effect estimates

$$\hat{\tau}^{CLATE}(Z = z_0, X_{\setminus Z} = x) = \mathbb{E}[Y(1) - Y(0) \mid Z = z_0, X_{\setminus Z} = x,\ \mathit{Complier}]$$

for the two disjoint complier subpopulations $G = C_1$ and $G = C_2$. Since the conditional constant effects assumption holds, it is irrelevant whether a CLATE was associated with $G = C_1$ or $G = C_2$, since both are equivalent to the CATE τ^{CATE}. Thus it is valid to treat these estimates as equivalent in the set $\mathcal{U}$. If conditional constant effects did not hold, then G would no longer be ignorable, and extrapolation from $\mathcal{U}$ would not yield valid estimates for a single, well-defined complier population. Instead, the posterior mean estimates would mix incomparable complier effects from the $G = C_1$ and $G = C_2$ subpopulations.

12. Viewing discovered RDs as local randomized experiments suggests the potential applicability of Kallus et al.'s method of experimental grounding (Kallus et al., 2018). However, while sharp RD designs can be treated as local randomized experiments (Lee and Lemieux, 2010), this approach is expected to fail in fuzzy RD designs where treatment uptake is selective.

13. RD instruments act as traditional instrumental variables at the regression discontinuity. For more information, see Angrist and Keueger (1991) and Imbens and Van Der Klaauw (1995).

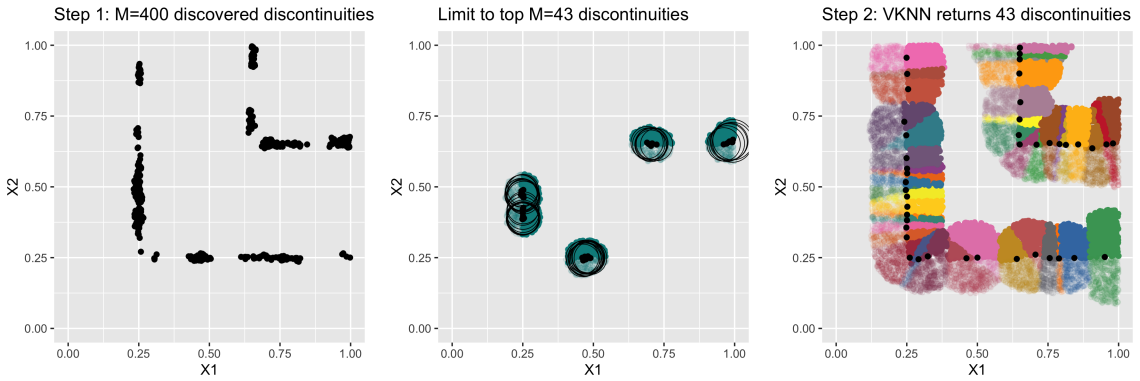

Figure 2: Illustrating the Discover and Estimate steps of DEE. For this example, $X1$ and $X2$ are two covariates in a uniform $[0,1]^2$ grid. The left panel shows the $M = |\mathcal{L}| = 400$ local discontinuities initially selected using LoRD^3, while the right panel shows the subset of discontinuities (and their index sets) output by Algorithm 1. As seen in the middle panel, applying a higher LLR threshold when constructing $\mathcal{L}$ is an insufficient solution to the problems of neighborhood overlap and RD coverage, since (i) the selected local RDs may still overlap, and (ii) the selected local RDs may not cover the full extent of the true discontinuity.

In this simulation the treatment effect $\tau(x)$ and expected bias $\beta(x)$ are both drawn from Gaussian Processes with isotropic Gaussian kernels k_{Θ_τ} and k_{Θ_β}, respectively. In this simulation, we fix the output scale of these kernels to 5, such that $k_{\Theta_\tau}(0) = k_{\Theta_\beta}(0) = 5$, and treat the length scale as the sole kernel parameter. As such, $\Theta_\tau = \{\theta_\tau\}$, and k_{Θ_τ} is

$$k_{\Theta_\tau}(x_i, x_j) = k_{\Theta_\tau}(\|x_i - x_j\|) = 5 \exp\left(-\frac{\|x_i - x_j\|^2}{2\theta_\tau^2}\right),$$

with k_{Θ_β} defined equivalently.

For each parameter configuration in the 2×2 parameter grid ($\theta_\tau \in \{0.2, 0.5\}, \theta_\beta \in \{0.2, 0.5\}$), we draw $N = 20,000$ samples from this DGP and apply our DEE approach [14]. We use the following parameter settings for each step of DEE:

1. **RD Discovery**: LoRD^3 was applied with $k = 200$ nearest neighbors, and a degree-4 polynomial baseline treatment propensity model. We selected the $M = |\mathcal{L}| = 400$ points with the highest LLR. For illustration, we show the 400 x's in $\mathcal{L}$ for one simulated run in the left panel of Figure 2.

2. **CATE Estimation**: The CATE was estimated using

   - Voronoi $\cap$ KNN repair (Algorithm 1) with parameters $k = 1000$ and $t = 30$. For illustration, we show the set $\mathcal{U}$ (and associated index sets) returned by Algorithm 1 in the right panel of Figure 2.

14. We show the results with this specific combination of parameters for illustrative purposes. Our results are robust to various other parameter combinations for θ_τ, θ_β, k, and t.

- Local linear 2SLS, with the following specification: letting Z denote the binary RD instrument, and X_{RD} denote the projection of X onto the RD normal vector, the local linear 2SLS estimator has first stage and reduced form

$$T \sim \alpha_T + \beta_{T,Z} Z + \beta_{T,X_{RD}} X_{RD} + \beta^{(Z)}_{T,X_{RD}} Z X_{RD}, \qquad \text{(First stage)}$$
$$Y \sim \alpha_Y + \beta_{Y,Z} Z + \beta_{Y,X_{RD}} X_{RD} + \beta^{(Z)}_{Y,X_{RD}} Z X_{RD}. \qquad \text{(Reduced form)}$$

3. **Extrapolation**: As an observational estimator, we fit a causal forest (Wager and Athey, 2018). To extrapolate, we use Gaussian Processes fit via marginal likelihood maximization using gpytorch (Gardner et al., 2018), with (i) constant mean parameters, and (ii) isotropic Gaussian (RBF) kernels parameterized by an output scale and length scale.

We then repeat this procedure for $N_{rep} = 50$ independent replications.

Figure 3 presents estimates of the generalization performance of various models over $P(X)$. Specifically, letting $\mu(x)$ denote the posterior mean of the final model, Figure 3 averages $(\tau(x) - \mu(x))^2$ over a 100×100 mesh grid uniformly covering $[0, 1]^2$. The panels each present a different parameter setting $(\theta_\tau, \theta_\beta)$, with $\theta_\tau = \theta_\beta$ on the diagonal, and $\theta_\tau \neq \theta_\beta$ off the diagonal.

As an initial observation, we note that across all parameter settings, all RD methods outperform direct use of the causal forest estimator. Moreover, we observe that our inferential method yields substantially lower mean squared error in the CATE estimates than the methods of Cattaneo et al. (2020) and Kallus et al. (2018). This is unsurprising, as both of these methods require assumptions that do not hold in our context. Specifically, Kallus et al. (2018) requires access to an unconfounded sample, and conditional ignorability $(\{Y(1), Y(0)\} \perp T \,|\, X)$ is a much stronger assumption than the CCE/CEI assumption that $\mathbb{E}[Y(1) - Y(0)] \perp T \,|\, X$. As such, one might expect Kallus's method to do badly in the presence of confounding, as we find in our simulation.[15] Meanwhile, Cattaneo et al.'s treatment effect estimator is biased if there are always-takers, as is the case in our simulation. Moreover, Cattaneo's method requires each data instance to be assigned to one of two thresholds ℓ and h along a running variable, and such threshold assignments are not observed in our observational dataset $\mathcal{D}$. As such, we resort to imputing individual threshold assignments as described in Appendix C, further reducing the accuracy of their method in our context.

Figure 3 also highlights the importance of Voronoi $\cap$ KNN repair (Algorithm 1) in `DEE`. Instead of applying Algorithm 1 to repair the original set of discovered RDs $\mathcal{L}$, we could consider computing treatment effects at each discovered RD in $\mathcal{L}$ using the same $k = 200$ neighborhoods that were used for RD discovery. However, GP extrapolation of these estimates (Figure 3, red trace) achieves significantly worse performance than extrapolation of estimates from the repaired set $\mathcal{U}$ (blue trace).

Next, comparing the MSEs for the bias model $\mathcal{M}_\beta$ and CATE model $\mathcal{M}_\tau$ where $\theta_\tau \neq \theta_\beta$ (off-diagonal), we confirm that the model $\mathcal{M}^*$ with the longer length scale has lower MSE. Moreover, in this $\theta_\tau \neq \theta_\beta$ context, model averaging successfully concentrates weight on $\mathcal{M}^*$. However, all likelihoods (marginal likelihood, LOOCV likelihood, and both BLOOCV

15. Additionally, Kallus's assumption of strong overlap between confounded and unconfounded datasets requires us to split the data and thus lose some accuracy.

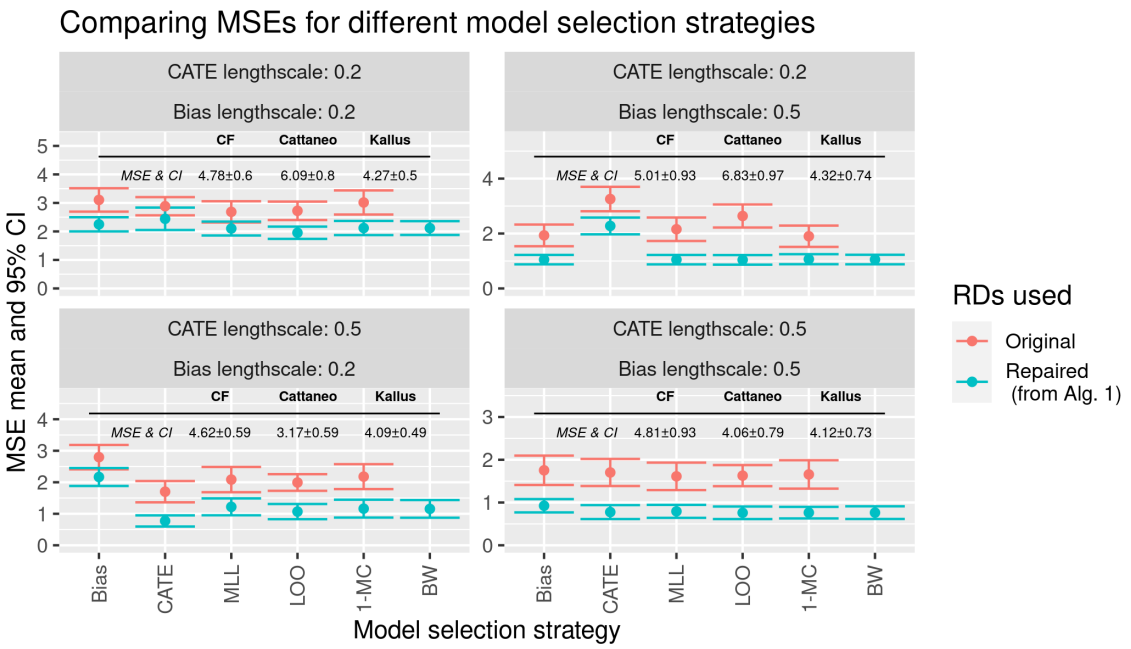

Figure 3: Average MSEs evaluated over a uniform mesh grid, with $N = 100 \times 100$ points, approximating performance over $P(X) = Unif([0,1]^2)$. The $x-$axis gives the model selection strategy. "Bias" and "CATE" refer to the strategies of always selecting the bias correcting model $\mathcal{M}_\beta$, and the direct extrapolation model $\mathcal{M}_\tau$, respectively. "MLL", "LOO", "1-MC", and "BW" refer to model averaging strategies where model weights are computed using the marginal likelihood, the LOOCV likelihood, the 1-MC BLOOCV variant, and the weighted BLOOCV variant, respectively. Benchmark MSEs are given in the inset table. The blue trace shows MSEs for CATE estimates obtained from the full inferential procedure of `DEE`, including the repair procedure in Algorithm 1 (which maps $\mathcal{L} \to \mathcal{U}$). The red trace shows the MSEs obtained when `DEE` does not apply Algorithm 1 and instead directly applies Gaussian process regression to local 2SLS estimates, computed using overlapping k-neighborhoods ($k = 200$), for each discovered local RD in $\mathcal{L}$. Note that we do not use weighted BLOOCV when using direct GP extrapolation from $\mathcal{L}$ due to the $\mathcal{O}(|\mathcal{L}|^5)$ runtime.

likelihood variants) yield model averages with similar MSEs. To explain this result, Figure 4 shows the mean difference in log-likelihoods under the CATE model $\mathcal{M}_\tau$ and bias model $\mathcal{M}_\beta$ in this off-diagonal regime. As expected from Theorem 1, the BLOOCV log-likelihood differences are larger than the LOOCV log-likelihood difference, thus indicating a better ability to distinguish these two models. However, as there are already significant differences between the marginal log-likelihoods under $\mathcal{M}_\tau$ and $\mathcal{M}_\beta$ (and between the LOOCV log-likelihoods as well), the relatively larger log-likelihood difference achieved using BLOOCV does not translate into significant improvements in MSE.

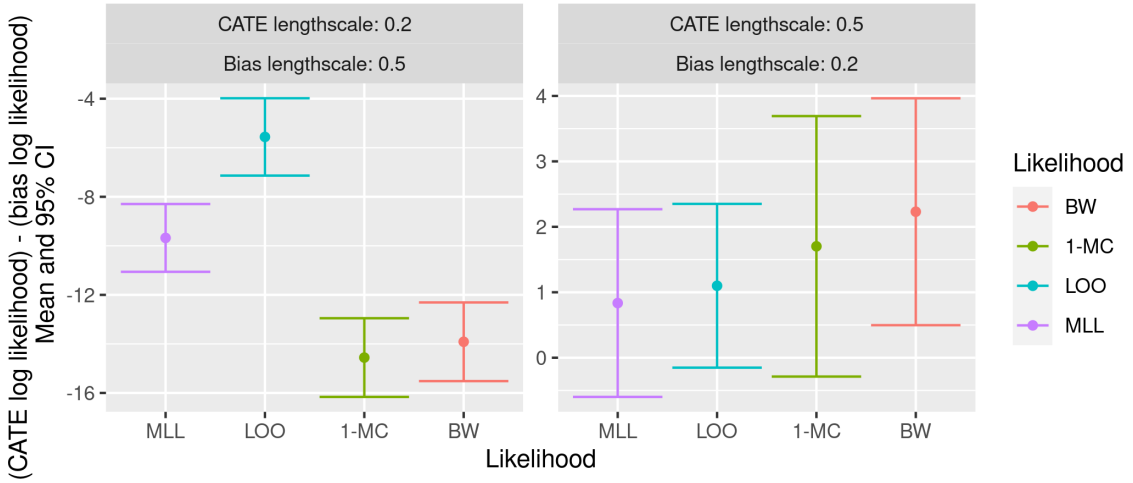

Figure 4: Difference between various log-likelihoods, under the CATE model $\mathcal{M}_\tau$ and the bias model $\mathcal{M}_\beta$, in the off-diagonal regime where $\theta_\tau \neq \theta_\beta$. "MLL", "LOO", "1-MC", and "BW" refer to model averaging strategies where model weights are computed using the marginal likelihood, the LOOCV likelihood, the 1-MC BLOOCV variant, and the weighted BLOOCV variant, respectively. The sign of the difference determines which model is preferred by each approach.

We specifically investigate this further in Appendix D.2, where we increased the number of training samples, thus increasing $|\mathcal{U}|$. We show that both BLOOCV variants are able to identify the significant difference between log-likelihoods under $\mathcal{M}_\tau$ and $\mathcal{M}_\beta$, whereas this difference for both marginal and LOOCV log-likelihoods is close to 0, as predicted by our theoretical results for the large-sample case. This resulted a better MSE for BLOOCV as compared to MLL and LOOCV.

Finally, while Figure 4 considers the off-diagonal regime where $\theta_\tau \neq \theta_\beta$, Figure 5 considers the regime where $\theta_\tau = \theta_\beta$, and shows that model averaging yields gains over model selection.

6.2 Simulation 2: Normal, Latent Index Model

Simulation 1 illustrates the performance of our `DEE` method when $\mathcal{M}_\tau$ and $\mathcal{M}_\beta$ are both correctly specified. In contrast, our second simulation illustrates the performance of `DEE` in the context of a standard econometric model – the normal latent index model (Heckman et al., 2001). Under this model, the bias function $\beta(x)$ is discontinuous at the regression discontinuities. As such, our model is misspecified, with the smooth GP model $\mathcal{M}_\beta$ offering, at best, a biased approximation of $\beta(x)$.

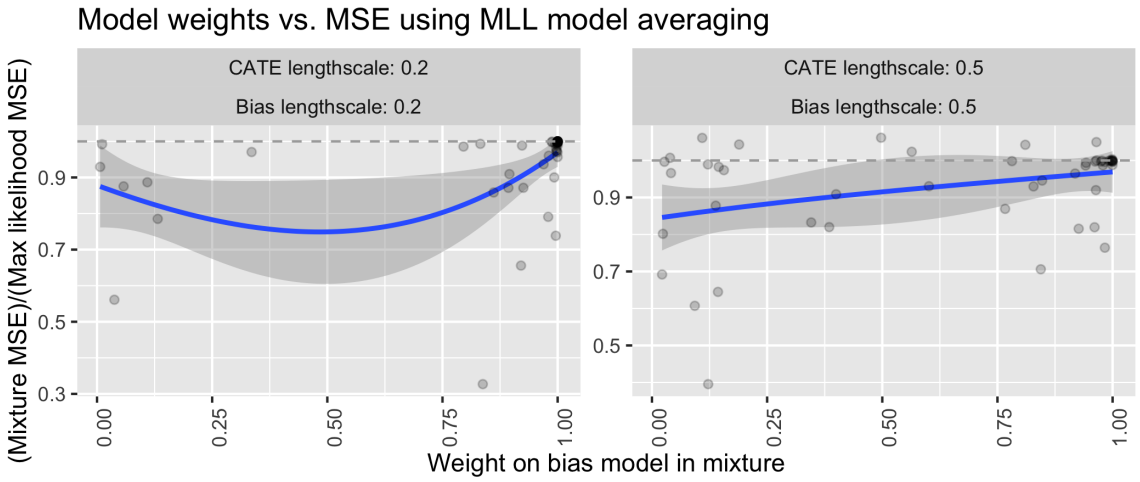

Figure 5: Relative mean squared error versus the $\mathcal{M}_\beta$ weight computed using marginal likelihood weighting (i.e., Bayesian model averaging) in the $\theta_\tau = \theta_\beta$ regime. The y-axis gives the MSE ratio for the marginal likelihood weighted model average versus the individual model ($\mathcal{M}_\tau$ or $\mathcal{M}_\beta$) with the maximum marginal likelihood. In general, model averaging reduces MSE (i.e., the local linear regression estimate, in blue, is smaller than 1, shown as a dashed grey reference line).

The data generating process for this simulation is presented in Appendix E.1. There are two key elements to this normal, latent index model. First, there are two[16] unobserved confounders, U^T and U^Y, drawn from a Gaussian distribution on $\mathbb{R}^2$. U^Y is a confounding term in the structural equations for $Y(1)$ and $Y(0)$, and U^T passes through latent indices to determine potential treatments $T(0)$, $T(1)$, and $T(2)$. If U^T and U^Y are correlated, they correlate the treatment and potential outcomes, and bias the naive approximation $\tau^{obs}(x)$. Moreover, by applying standard results for bivariate normals, this bias can be shown to be a multiple of the covariance between U^Y and U^T (Heckman et al., 2001). Thus, by treating this covariance as a function of x, we can specify the expected bias in τ^{obs} up to a factor that is a piecewise constant function in x.

We apply a similar simulation procedure for this second DGP. For each parameter configuration in the 2×2 parameter grid ($\theta_\tau \in \{0.2, 0.5\}, \theta_{Cov} \in \{0.2, 0.5\}$), we sample a training set $\mathcal{D}$ with $|\mathcal{D}| = N = 20,000$ instances, and then apply `DEE`. In this simulation, we apply Algorithm 1 with parameters $k = 400$ and $t = 30$, and estimate the RD treatment effect using a simple difference-in-means 2SLS estimator, with first stage $T \sim \alpha_T + \beta_T Z$ and reduced form $Y \sim \alpha_Y + \beta_Y Z$. Again, we repeat this procedure for $N_{rep} = 50$ independent replications. Results from these simulations are presented in Appendix E.2, Figures 14-17, and are broadly consistent with our first set of simulations. Again, we observe that exploiting the discovered RD offers improved performance relative to direct use of the causal

16. Readers may be more familiar with the general formulation of a normal latent index model, where there are separate latent variables U^1 and U^0 for $Y(1)$ and $Y(0)$, respectively. To satisfy the conditional constant effects assumption, we restrict the general normal latent index model to $U^1 = U^0 = U^Y$.

forest, that DEE improved on the various benchmarks, that our novel repair procedure improved the performance of DEE, and that model averaging both identified the desired model in the off-diagonal regime and yielded gains in the on-diagonal regime.

6.3 Application: Rural Roads and Economic Development

Having shown that our method outperforms benchmarks on synthetic datasets, we next apply our method to a causal inference problem from the recent economic development literature. In their paper, Asher and Novosad (2020) use a fuzzy RD design to estimate the village-level causal effect of the Pradhan Mantri Gram Sadak Yojana (PMGSY), or the Prime Minister's Village Road Program. This program aimed to connect unconnected rural Indian villages to existing road networks.

Asher and Novosad (2020) use a fuzzy RD design to evaluate this program, exploiting the fact that the program prioritized villages for road building based on their 2001 population. Specifically, it aimed to connect all villages with populations greater than 1,000 by 2003, and all villages with populations greater than 500 by 2007. These population thresholds create discontinuities in the treatment propensity, with villages just above these thresholds being 22 percentage points more likely to be treated than villages just below the threshold.

This problem provides an ideal context to demonstrate our method for two reasons. First, in addition to the two known population discontinuities, it is plausible there are also spatial discontinuities in treatment – whether across geographic boundaries or due to the spatial configuration of the existing road network. Second, inspecting Figure 6, there appears to be overlap in the spatial distribution of treated and untreated villages – allowing for estimation of $\tau^{obs}(x_{geo})$ where $x_{geo} = (Long.,\ Lat.)$. Given these two factors, we proceed by applying our method to estimate the spatial CATE $\tau(x_{geo}) = \mathbb{E}[Y(1) - Y(0) \mid X = x_{geo}]$.

6.3.1 Rural Roads: Dataset

We use a subset of rural villages from the full replication dataset of Asher and Novosad (2020) provided on OpenICPSR.[17] Specifically, we follow Asher and Novosad (2020) in restricting our sample to villages that (i) did not have a paved road in 2001, (ii) were not missing any covariates or outcomes (i.e., were matched across all datasets), and (iii) were in one of six states: Chhattisgarh, Gujarat, Madhya Pradesh, Maharashtra, Orissa, and Rajasthan. While Asher and Novosad restrict their analytic sample to just those villages within the optimal bandwidth of the relevant population threshold, we include all 35,273 villages with populations between 300 and 1,300 that satisfied these initial criteria, to show that DEE can discover and extrapolate from discontinuities without the benefit of this prior knowledge. The left panel of Figure 6 shows the 35,273 villages in our sample, with treated villages in blue and untreated villages in red.

6.3.2 Rural Roads: Method

We applied our method to estimate $\tau(x_{geo}) = \mathbb{E}[Y(1) - Y(0) \mid X = x_{geo}]$, for each of the five main outcomes in Asher and Novosad (2020), as follows:

17. We are grateful to Asher and Novosad for posting full replication data and code on ICPSR.

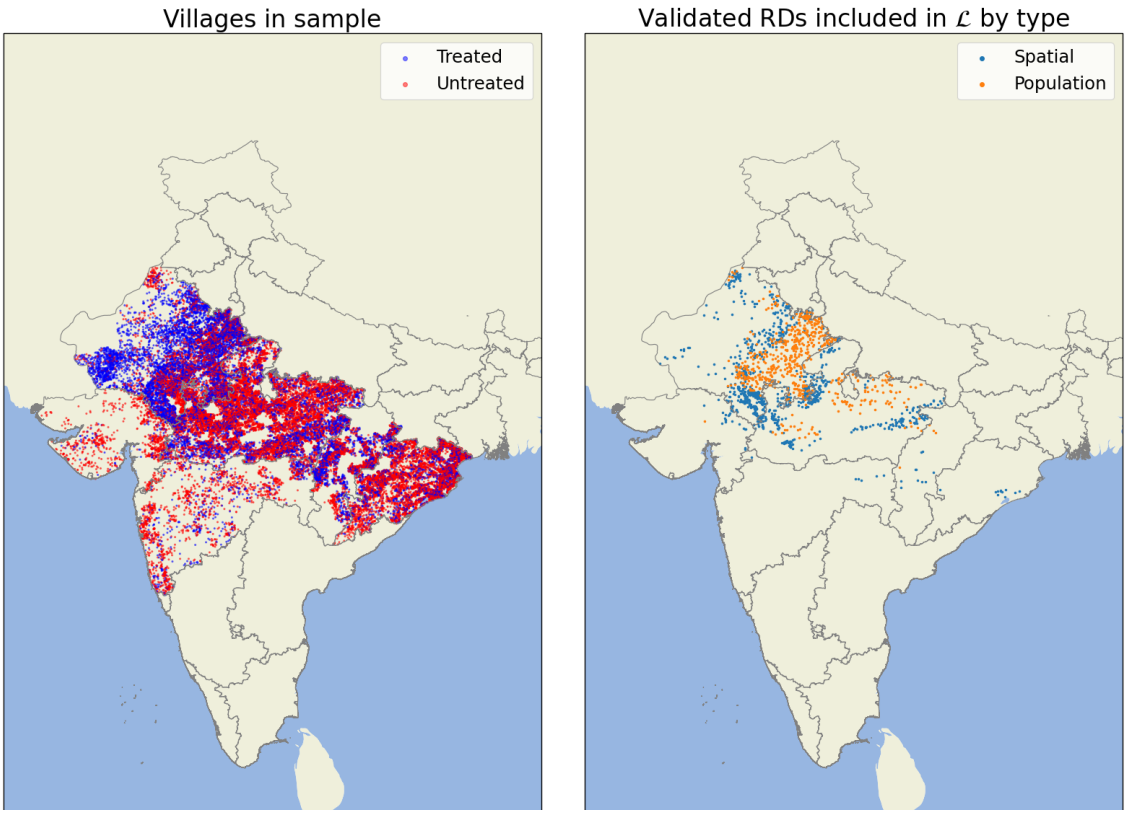

Figure 6: Left panel: 35,273 villages included in our analytic sample, colored by treatment status. Blue villages were treated (i.e., received a new all-weather road), while red villages were untreated. Right panel: Villages selected into $\mathcal{L}$, colored by RD type. Blue points correspond to the center of a detected local spatial RD, while orange points correspond to the center of a detected local population RD. Figure 7 describes the method used to categorize discovered RDs.

1. **RD Discovery**: First, we detect regression discontinuities in

$$(Longitude,\ Latitude,\ Population) \in \mathbb{R}^3$$

by applying LoRD^3 with $k = 200$ nearest neighbors, and a degree-4 polynomial baseline treatment propensity model. We select the $M = 2{,}784$ discovered discontinuities that are significant at $\alpha = 0.05$ using randomization testing (Herlands et al., 2018) into $\mathcal{L}$, then drop discontinuities from $\mathcal{L}$ that fail the covariate balance test described in Appendix F.2.

As shown in Figures 7 and 8, the resulting set $\mathcal{L}$ includes both population and spatial RDs. Since our goal is to estimate the spatial CATE $\tau(x_{geo})$, we proceed under an additional assumption:

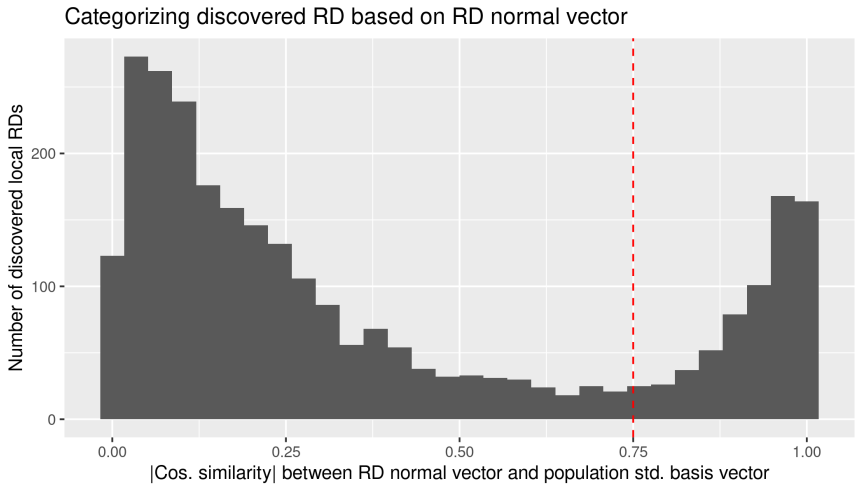

Figure 7: Discovered regression discontinuities are categorized as spatial or population RDs based on the absolute cosine similarity between the RD's normal vector v_j and the population standard basis vector e_{pop}. Discontinuities with $|\cos(v_j, e_{pop})| > 0.75$ are categorized as population discontinuities, and those with $|\cos(v_j, e_{pop})| \leq 0.75$ are categorized as spatial discontinuities.

Assumption 5 *Treatment effects are conditionally mean independent of population size,*

$$\mathbb{E}[Y(1) - Y(0) \mid X_{geo} = x_{geo}, Population] = \mathbb{E}[Y(1) - Y(0) \mid X_{geo} = x_{geo}].$$

Under this assumption, treatment effects estimated at local population discontinuities of 500 or 1,000 generalize to other villages with similar population sizes. (Note that our extrapolations may not be appropriate for megacities with significantly larger population sizes.)

2. **CATE Estimation**: We apply Voronoi ∩ KNN repair (Algorithm 1) with parameters $k = 400$ and $t = 40$ to obtain set $\mathcal{U}$ of disjoint, local discontinuities. Once again we drop discontinuities from $\mathcal{U}$ that fail the covariate balance test described in Appendix F.2. Then, having filtered $\mathcal{U}$ down to only validated RDs, we estimate RD treatment effects using a simple difference-in-means 2SLS estimator, with first stage $T \sim \alpha_T + \beta_T Z$ and reduced form $Y \sim \alpha_Y + \beta_Y Z$.

3. **Extrapolation**: Again we fit a causal forest (Wager and Athey, 2018) as our observational estimator $\hat{\tau}^{obs}(x_{geo})$. To extrapolate, we use Gaussian Processes fit via marginal likelihood maximization using gpytorch (Gardner et al., 2018), with (i) constant mean parameters, and (ii) isotropic Gaussian (RBF) kernels parameterized by an output scale and length scale. We compute our final model average using the weighted BLOOCV approach, described in Section 3.3.4.

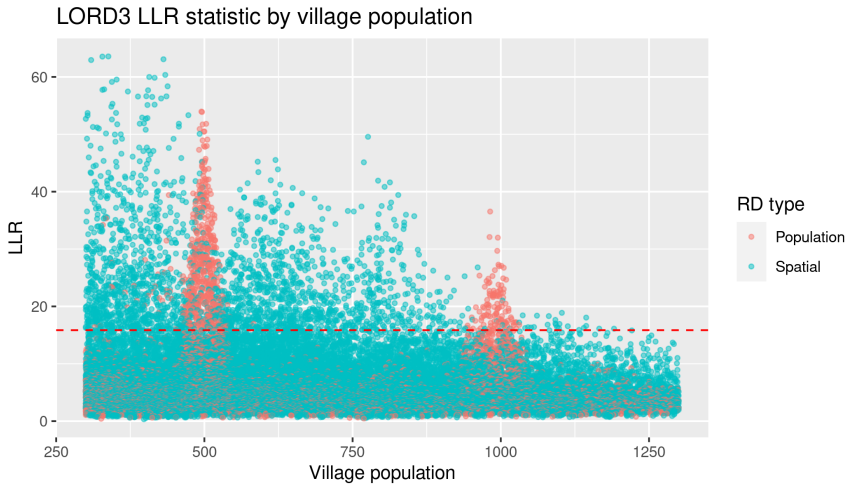

Figure 8: LoRD^3 log-likelihood ratios LLR for each village, by village population. Discontinuities that are categorized as population RDs (as described in Figure 7) are shown in red, while spatial RDs are shown in blue. The $M = 2{,}784$ RDs selected into $\mathcal{L}$ are those RDs that are significant at $\alpha = 0.05$ using randomization testing and fall above the red dashed reference line.

6.3.3 Rural Roads: Results

We evaluate our method's performance on this applied task three ways. First, we evaluate our method's ability to identify known and unknown regression discontinuities. Second, we compare our method's average treatment effect estimates to those obtained using Asher and Novosad's 2SLS specification (Asher and Novosad, 2020). Finally, we comment on the spatial treatment effect heterogeneity detected by our method.

First, we find that `DEE` successfully identifies both known and unknown discontinuities. As shown in Figures 7 and 8, we identify the two known discontinuities in village population at 500 and 1,000. Using Figure 7, we categorize discovered RDs as spatial or population RDs based on the absolute cosine similarity[18] between the RD's normal vector v_j and the population standard basis vector e_{pop}. In addition, as seen in the right panel of Figure 6, we identify several coherent spatial discontinuities. Table 1 provides additional information about discovered RDs, and describes the index sets associated with each validated RD in $\mathcal{U}$. As seen in Table 1, applying Algorithm 1 with $k = 400$ and $t = 40$ yields index sets containing between 81 and 351 villages (mean = 148). Discovered population RDs also have an average effective population bandwidth of 95, similar to the bandwidth of 84 used by Asher and Novosad (2020).

Second, we find that `DEE` yields average treatment effect estimates consistent with those reported in Asher and Novosad (2020). Figure 9 presents 2SLS RD estimates based on Asher and Novosad's original 2SLS specification alongside two effect estimates obtained using `DEE`:

18. The threshold of 0.75 for the cosine similarity was chosen in Figure 7 to clearly distinguish population discontinuities from spatial discontinuities in our visualizations and results. Modifying this parameter between 0.5 and 0.75 has no real impact on our results.

		Population and Spatial RDs			Population RDs
		All	Spatial	Population	All
# RDs	$\mathcal{L}$	2784	2136	648	–
	Validated $\mathcal{L}$	1613	979	634	–
	$\mathcal{U}$	94	75	19	32
	Validated $\mathcal{U}$	62	45	17	30
$VK[x]$	Long BW	1.26	1.31	1.12	1.2
	Lat BW	0.73	0.77	0.61	0.65
	Pop BW	75.69	69.59	91.85	95
	Mean N (min,max)	148 (81-351)	150 (81-351)	141 (96-315)	174 (83-400)

Table 1: Top panel (# RDs): Number of local RDs, by RD type, in each of four sets: all discovered RDs $\mathcal{L}$, covariate balance test validated RDs in $\mathcal{L}$, filtered RDs $\mathcal{U}$, and filtered and covariate balance test validated RDs in $\mathcal{U}$. Bottom panel ($VK[x]$): Descriptive statistics for Voronoi $\cap$ KNN sets associated with discovered, validated RDs $x \in \mathcal{U}$. Rows 1-3 provide the mean effective bandwidth (BW), where a given $VK[x]$'s dimension-specific effective bandwidth is defined as half the max $-$ min difference. Row 4 provides the mean number of villages, and (min, max) range, included in each $VK[x]$.

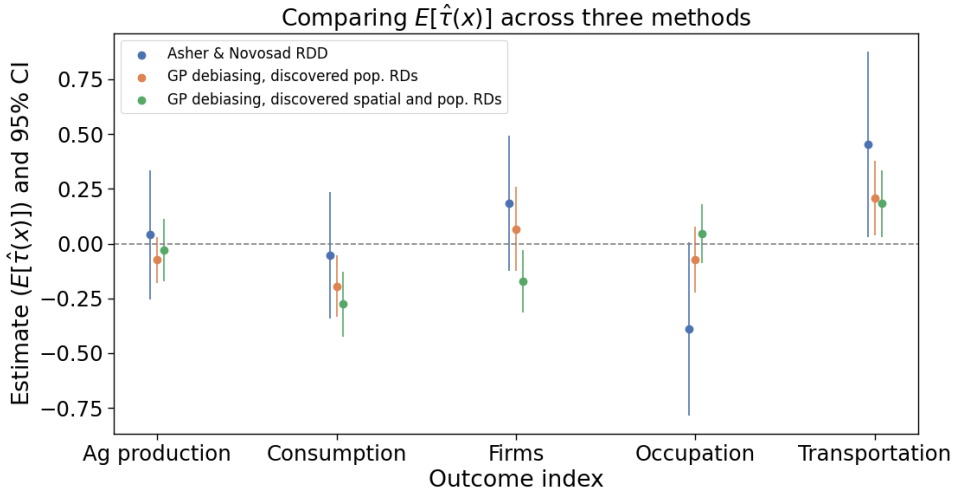

Figure 9: Average treatment effects across the five main outcomes in Asher and Novosad (2020). Asher and Novosad's RD estimates are interpreted as LATEs, while GP debiasing estimates are interpreted as CATEs under assumptions. Note that the reported Asher and Novosad RD estimates are not identical to those in Asher and Novosad (2020), since we refit their 2SLS estimator using a uniform kernel.

(i) using just the discovered population RDs, and (ii) using both discovered spatial RDs and discovered population RDs. Within each of the five indexed outcomes, the confidence intervals for `DEE` overlap with the confidence interval for the 2SLS estimates, indicating that

these estimates are similar – though DEE's confidence intervals are more precise. As in Asher and Novosad (2020), we observe significant (at $\alpha = 0.05$), positive average treatment effects for the impact of rural roads on the availability of transportation services. We also observe significant (at $\alpha = 0.05$), negative average effects on the consumption index.

Finally, Appendix F.1 presents maps displaying spatially heterogeneous treatment effect estimates $\hat{\tau}(x_{geo})$ across all five outcomes. We find relatively little spatial heterogeneity in the effect of road building on transportation availability using just population RDs (see the left panel of Figure 18). However, we find some heterogeneity in treatment effects using spatial and population RDs, especially in the tri-state regions south of Madhya Pradesh (see the right panel of Figure 18). Furthermore, across the remaining outcomes (Figures 19-22), we find potentially greater evidence of spatial heterogeneity – especially when extrapolating from discovered spatial RDs. In particular, it appears that the impacts of rural roads on agricultural yields, consumption, and employment growth in the semi-arid western region may lag behind the possible benefits observed in other regions of India. Further studies are needed to verify these results and to examine potential factors that could explain these regional differences.

7. Discussion

As demonstrated in Section 6, our DEE method improves on a conditioned-on-observables estimate and allows for more accurate extrapolation off of the discovered discontinuities. Moreover, applying DEE to a recent economic development dataset (Asher and Novosad, 2020) demonstrates its ability to detect local regression discontinuities, and to extrapolate heterogeneous effect estimates from these discovered RDs.

Nevertheless, there are several limitations to the method. First, the effective use of DEE requires the presence of a sufficiently large number of RDs in the data to accurately extrapolate CATE estimates, and these estimates will be less precise in regions of the attribute space that are further from the discovered discontinuities. It is not guaranteed that a given dataset will have any RDs to discover. Nevertheless, we believe that large real-world datasets (such as the Rural Roads dataset examined here) are likely to contain numerous RDs because of the many phenomena (such as geographic boundaries, laws, clinical guidelines, etc.) which can result in discontinuous changes in treatment probability. Second, the RDs must be successfully identified before we can use them for estimation and extrapolation, and this may be more difficult for complex, high-dimensional datasets. In their prior work on LoRD^3, Herlands et al. (2018) examine the effects of dimensionality, and conclude that performance of LoRD^3 "is robust to large numbers of covariates but reduced over larger spaces of forcing variables". We note, however, that the estimation and extrapolation steps of DEE could generalize to any set of identified RDs, whether discovered (by LoRD^3 or another procedure) or manually specified, and thus prior domain knowledge could be applied to improve RD discovery in such cases. Third, the ability to obtain unbiased CATE estimates from the discovered RDs depends on the validity of several assumptions noted above. Our Assumptions 1-3 are standard in the RD literature, and necessary for any valid inferences using RDs. While the CCE/CEI assumption is reasonable, common, and weaker than the assumption of strong conditional ignorability that is made in "selection on observables" analyses, it is also untestable, and thus our CATE estimates must still be viewed

with the important caveat, "if CCE/CEI holds". Other potential threats to the validity of the RD estimates, such as discontinuities in covariates or manipulation of treatment at the RD threshold, are testable (and tested) as part of the LoRD^3 discovery procedure.

Given these considerations, there are two potential variations to the method which could be considered. First, there is the potential for data-driven discovery of local RDs to introduce bias into treatment effect estimates $\hat{\tau}(x)$ for $x \in \mathcal{U}$. Using independent samples to discover RDs and to estimate treatment effects (i.e., data splitting) would ameliorate this concern, with the possible tradeoff of increasing variance due to the reduced sample size. Second, our method requires setting a number of hyperparameters, including the parameters t and k in Algorithm 1, and choosing GP kernel functions. While our experimental results suggest that the performance of `DEE` is relatively robust to the choice of hyperparameters, our current approach to choosing hyperparameters is rather arbitrary. Again, BLOOCV or marginal likelihood could be used to guide parameter tuning. Leaving aside these potential variations for future work, we have demonstrated that our method improves on a conditioned-on-observables estimate and allows for effective extrapolation off of the discovered discontinuities.

Acknowledgments

The authors wish to thank William Herlands for discussing the application of the LORD^3 procedure (Herlands et al., 2018) and sharing his code for LORD^3.

References

Ahmed M. Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/6a508a60aa3bf9510ea6acb021c94b48-Paper.pdf`.

Joshua D. Angrist. Treatment effect heterogeneity in theory and practice. *The Economic Journal*, 114(494):C52–C83, 2004. URL `http://www.jstor.org/stable/3590307`.

Joshua D. Angrist and Iván Fernández-Val. *ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework*, volume 3 of *Econometric Society Monographs*, page 401–434. Cambridge University Press, 2013. doi: 10.1017/CBO9781139060035.012.

Joshua D. Angrist and Alan B. Keueger. Does Compulsory School Attendance Affect Schooling and Earnings?*. *The Quarterly Journal of Economics*, 106(4):979–1014, 11 1991. ISSN 0033-5533. doi: 10.2307/2937954. URL `https://doi.org/10.2307/2937954`.

Joshua D. Angrist and Miikka Rokkanen. Wanna get away? regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association*, 110(512):1331–1344, 2015. doi: 10.1080/01621459.2015.1012259. URL `https://doi.org/10.1080/01621459.2015.1012259`.

Sam Asher and Paul Novosad. Rural roads and local economic development. *American Economic Review*, 110(3):797–823, March 2020. doi: 10.1257/aer.20180268. URL https://www.aeaweb.org/articles?id=10.1257/aer.20180268.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016. doi: 10.1073/pnas.1510489113. URL https://www.pnas.org/content/113/27/7353.

Sascha O. Becker, Peter H. Egger, and Maximilian von Ehrlich. Absorptive capacity and the growth and investment effects of regional transfers: A regression discontinuity design with heterogeneous treatment effects. *American Economic Journal: Economic Policy*, 5(4):29–77, 2013. ISSN 19457731, 1945774X. URL http://www.jstor.org/stable/43189353.

Matias D. Cattaneo, Luke Keele, Rocío Titiunik, and Gonzalo Vazquez-Bare. Extrapolating treatment effects in multi-cutoff regression discontinuity designs. *Journal of the American Statistical Association*, 0(0):1–12, 2020. doi: 10.1080/01621459.2020.1751646. URL https://doi.org/10.1080/01621459.2020.1751646.

Alicia Curth and Mihaela van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 1810–1818. PMLR, 2021.

Yingying Dong. Alternative assumptions to identify late in fuzzy regression discontinuity designs. *Oxford Bulletin of Economics and Statistics*, 80(5):1020–1027, 2018. doi: https://doi.org/10.1111/obes.12249. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/obes.12249.

Jana Eklund and Sune Karlsson. Forecast combination and model averaging using predictive measures. *Econometric Reviews*, 26(2-4):329–363, 2007.

Jared C Foster, Jeremy M G Taylor, and Stephen J Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867–2880, Aug 2011.

Jacob Gardner, Geoff Pleiss, Kilian Q Weinberger, David Bindel, and Andrew G Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 7576–7586. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/27e8e17134dd7083b050476733207ea1-Paper.pdf.

Donald P Green and Holger L Kern. Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opinion Quarterly*, 76(3):491–511, Sep 2012.

Justin Grimmer, Solomon Messing, and Sean J. Westwood. Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Analysis*, 25(4):413–434, 2017. doi: 10.1017/pan.2017.15.

Limor Gultchin, Matt Kusner, Varun Kanade, and Ricardo Silva. Differentiable causal backdoor discovery. In *International Conference on Artificial Intelligence and Statistics*, pages 3970–3979. PMLR, 2020.

Jinyong Hahn, Petra Todd, and Wilbert Van der Klaauw. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209, 2001. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2692190.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep IV: A flexible approach for counterfactual prediction. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1414–1423, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/hartford17a.html.

Jason S Hartford, Victor Veitch, Dhanya Sridhar, and Kevin Leyton-Brown. Valid causal inference with (some) invalid instruments. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4096–4106. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/hartford21a.html.

James Heckman, Justin L. Tobias, and Edward Vytlacil. Four parameters of interest in the evaluation of social programs. *Southern Economic Journal*, 68(2):211–223, 2001.

William Herlands, Edward McFowland III, Andrew Gordon Wilson, and Daniel B. Neill. Automated local regression discontinuity design discovery. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 1512–1520, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3219982. URL https://doi.org/10.1145/3219819.3219982.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, Jan 2011.

Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986. ISSN 01621459. URL http://www.jstor.org/stable/2289064.

Kosuke Imai and Marc Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, Mar 2013.

Guido Imbens and Wilbert Van Der Klaauw. Evaluating the cost of conscription in the netherlands. *Journal of Business & Economic Statistics*, 13(2):207–215, 1995. doi: 10.1080/07350015.1995.10524595. URL https://www.tandfonline.com/doi/abs/10.1080/07350015.1995.10524595.

Guido W. Imbens. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, 58(4):

1129–79, December 2020. doi: 10.1257/jel.20191597. URL https://www.aeaweb.org/articles?id=10.1257/jel.20191597.

Guido W. Imbens and Joshua D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2951620.

Guido W. Imbens and Thomas Lemieux. Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635, 2008. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2007.05.001. URL https://www.sciencedirect.com/science/article/pii/S0304407607001091. The regression discontinuity design: Theory and applications.

David D. Jensen, Andrew S. Fast, Brian J. Taylor, and Marc E. Maier. Automatic identification of quasi-experimental designs for discovering causal knowledge. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 372–380, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605581934. doi: 10.1145/1401890.1401938. URL https://doi.org/10.1145/1401890.1401938.

Jean Kaddour, Yuchen Zhu, Qi Liu, Matt J Kusner, and Ricardo Silva. Causal effect inference for structured treatments. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 24841–24854. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/d02e9bdc27a894e882fa0c9055c99722-Paper.pdf.

Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 10888–10897. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/566f0ea4f6c2e947f36795c8f58ba901-Paper.pdf.

Edward H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020. URL https://arxiv.org/abs/2004.14497.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019. doi: 10.1073/pnas.1804597116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1804597116.

Armin Lederer, Jonas Umlauft, and Sandra Hirche. Posterior variance analysis of gaussian processes with application to average learning curves. *arXiv preprint arXiv:1906.01404*, 2019.

David S. Lee and Thomas Lemieux. Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2):281–355, June 2010. doi: 10.1257/jel.48.2.281. URL https://www.aeaweb.org/articles?id=10.1257/jel.48.2.281.

Fan Li, Kari Lock Morgan, and Alan M. Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018. doi: 10.1080/01621459.2016.1260466. URL https://doi.org/10.1080/01621459.2016.1260466.

Jordan D. Matsudaira. Mandatory summer school and student achievement. *Journal of Econometrics*, 142(2):829–850, February 2008. ISSN 0304-4076. doi: 10.1016/j.jeconom.2007.05.015. URL https://www.sciencedirect.com/science/article/pii/S0304407607001194.

Justin McCrary. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of econometrics*, 142(2), 2008.

X Nie and S Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 09 2020.

Scott Powers, Junyang Qian, Kenneth Jung, Alejandro Schuler, Nigam H. Shah, Trevor Hastie, and Robert Tibshirani. Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in Medicine*, 37(11):1767–1787, 2018.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN 026218253X.

Donald B Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331, 2005. doi: 10.1198/016214504000001880. URL https://doi.org/10.1198/016214504000001880.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3076–3085. JMLR.org, 2017.

Xiaogang Su, Chih-Ling Tsai, Hansheng Wang, David M Nickerson, and Bogong Li. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10: 141–158, Dec 2009.

Lu Tian, Ash A Alizadeh, Andrew J Gentles, and Robert Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532, Dec 2014.

William M. K. Trochim. *Research Design for Program Evaluation: the Regression-Discontinuity Approach*. Sage Publications, Beverley Hills, CA, USA, 1984. ISBN 0803920377.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. doi: 10.1080/01621459.2017.1319839. URL `https://doi.org/10.1080/01621459.2017.1319839`.

Herbert I Weisberg and Victor P Pontes. Post hoc subgroups in clinical trials: Anathema or analytics? *Clinical trials*, 12(4):357–364, Aug 2015.

Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. *Advances in neural information processing systems*, 31, 2018.

Appendix A. Pseudocode for LoRD^3 Procedure

Algorithm 2: LoRD^3 procedure, used by `DEE`, to automatically discover local regression discontinuities.

Input: $\mathcal{D} = \{(x_i, t_i, y_i)\}_{i=1}^{N}$
Parameter: K_{min} – The minimum number of neighbors to consider for KNN.
Parameter: K_{max} – The maximum number of neighbors to consider for KNN.
Parameter: Z – The subset of features $Z \subseteq X$ to be considered (jointly) as forcing variables.
Output: $\mathcal{L}$

$\mathcal{L} \leftarrow \{\}$
Compute $\hat{f}(x)$ as an estimate of a properly chosen smooth function $f(x)$, where $t_i = f(x_i) + \epsilon_i$
Compute the estimated residuals $r_i = t_i - \hat{f}(x_i)$ for $i = 1, \ldots, N$.
for $k = K_{min}, ..., K_{max}$ **do**
  for $i = 1, ..., N$ **do**
    Compute $s_{i,k}$, the k-sized neighborhood containing the i^{th} data point and its $k-1$ nearest neighboring points, where only features in Z are used to measure distance.
    Compute $L_0(s_{i,k})$, the likelihood of $s_{i,k}$ under the null hypothesis that it does not contain a regression discontinuity, $H_0 : E[r_j] = \mu_0, \forall j \in s_{i,k}, j \neq i$.
    for $l = 1, ..., k-1$ **do**
      Compute (i) the vector h_l between center point x_i and x_l, where $l \in s_{i,k}$ (again only using Z); and (ii) the hyperplane v_l that passes through x_i and is orthogonal to h_l, bisecting $\{x_j \mid j \in s_{i,k}\}$ into two partitions g^1 and g^2.
      Compute $L_1(s_{i,k}, v_l)$, the likelihood of $s_{i,k}$ under the alternative hypothesis that it contains an regression discontinuity at v_l, $H_1 : E[r_j] = \mu_1 \left(\mathbb{1}_{\{j \in g^1\}}\right) + \mu_2 \left(\mathbb{1}_{\{j \in g^2\}}\right), \forall j \in s_{i,k},\ j \neq i$.
      Compute $LLR(s_{i,k}, v_l) = \log \dfrac{L_1(s_{i,k}, v_l)}{L_0(s_{i,k})}$, the log-likelihood ratio (LLR) testing the alternative hypothesis of a discontinuity at v_l against the null hypothesis of no discontinuity for neighborhood $s_{i,k}$.
    Compute $LLR_{i,k} = \max_l LLR(s_{i,k}, v_l)$ and $v_{i,k} = \arg\max_l LLR(s_{i,k}, v_l)$, by maximizing the log-likelihood ratio (LLR) over all $(k-1)$ partitions of neighborhood $s_{i,k}$.
    if *$LLR_{i,k}$ is statistically significant (i.e., it exceeds the significance threshold for a given level α, obtained by randomization), and the corresponding $s_{i,k}$ and $v_{i,k}$ are econometrically valid (by density and placebo tests)* **then**
      $\mathcal{L} \leftarrow \mathcal{L} \cup \{(s_{i,k}, v_{i,k}, LLR_{i,k})\}$

Note that we typically set $K_{min} = K_{max}$ for computational efficiency. For example, $K_{min} = K_{max} = k = 200$ for our simulation studies.

Appendix B. Proof of Theorem 1

First, we present two lemmas. Then we give the proof of Theorem 1.

Lemma 1 *Under the assumptions of Theorem 1, the posterior predictive variance at a point x_* is given by*

$$\sigma^2(x_*) = \sigma_n^2 + k(0) - \frac{k(\rho)^2}{\frac{\sigma_n^2}{N} + \mathbb{E}[k(q)]}.$$

Proof Following Rasmussen and Williams (2005), the posterior predictive variance at x_*,

$$\sigma^2(x_*) = \sigma_n^2 + k(0) - \mathbf{k}^T K^{-1} \mathbf{k},$$

where $\mathbf{k} = [k(x_*, x_1), \cdots, k(x_*, x_N)]^T$ and K is the $N \times N$ kernel matrix with i, j^{th} element $K_{ij} = k(x_i, x_j)$. Now, since all training instances are distance ρ from the test instance, $k(x_*, x_i) = k(\rho)$ for all i. Thus, we can rewrite

$$\begin{aligned}\sigma^2(x_*) &= \sigma_n^2 + k(0) - \mathbf{k}^T K^{-1} \mathbf{k} \\ &= \sigma_n^2 + k(0) - k(\rho)^2 \mathbf{1}^T K^{-1} \mathbf{1}.\end{aligned}$$

Moreover, recalling that q denotes pairwise distances between training instances, the row sums of K are all equal to

$$r = k(0) + \sigma_n^2 + (N-1)\mathbb{E}_{i,j:i\neq j}[k(x_i, x_j)] = \sigma_n^2 + N\mathbb{E}[k(q)].$$

This means that K has eigenvalue r for eigenvector $\mathbf{1}$, so K^{-1} has eigenvalue $1/r$ for eigenvector $\mathbf{1}$, and thus

$$K^{-1}\mathbf{1} = \frac{1}{\sigma_n^2 + N\mathbb{E}[k(q)]}\mathbf{1}.$$

Thus, we obtain

$$\begin{aligned}\sigma^2(x_*) &= \sigma_n^2 + k(0) - \mathbf{k}^T K^{-1} \mathbf{k} \\ &= \sigma_n^2 + k(0) - k(\rho)^2 \mathbf{1}^T K^{-1} \mathbf{1} \\ &= \sigma_n^2 + k(0) - k(\rho)^2 \mathbf{1}^T \left[\frac{1}{\sigma_n^2 + N\mathbb{E}[k(q)]}\mathbf{1}\right] \\ &= \sigma_n^2 + k(0) - \frac{Nk(\rho)^2}{\sigma_n^2 + N\mathbb{E}[k(q)]} \\ &= \sigma_n^2 + k(0) - \frac{k(\rho)^2}{\frac{\sigma_n^2}{N} + \mathbb{E}[k(q)]}.\end{aligned}$$

■

Lemma 2 *Let Z_1 and Z_2 be defined such that*

$$k_1(Z_1) = \mathbb{E}[k_1(q)]; \qquad k_2(Z_2) = \mathbb{E}[k_2(q)].$$

Then $Z_2 \leq Z_1 \leq \mathbb{E}[q]$, with equality if and only if q is constant.

Proof First, we show $Z_1 \leq \mathbb{E}[q]$, by applying the AM-GM inequality to $\mathbb{E}[k_1(q)]$ and $k_1(\mathbb{E}[q])$.

$$\begin{aligned}\mathbb{E}[k_1(q)] &= \frac{1}{N}\left[\sum_i \exp\left(-\frac{q_i^2}{2b_1^2}\right)\right] \\ k_1(\mathbb{E}[q]) &= \exp\left(\frac{1}{N}\sum_i -\frac{q_i^2}{2b_1^2}\right) \\ &= \left[\prod_i \exp\left(-\frac{q_i^2}{2b_1^2}\right)\right]^{\frac{1}{N}}.\end{aligned}$$

Thus, by the AM-GM inequality, we have

$$\begin{aligned}&\mathbb{E}[k_1(q)] \geq k_1(\mathbb{E}[q]) && \text{(AM-GM)} \\ \iff &k_1(Z_1) \geq k_1(\mathbb{E}[q]) && \text{(definition of } Z_1\text{)} \\ \implies &Z_1 \leq \mathbb{E}[q], && (k_1 \text{ decreasing})\end{aligned}$$

with equality achieved if and only if $\exp(-q_i^2/(2b_1^2))$ is equal for all i, implying q_i is constant.

Now, we show $Z_2 \leq Z_1$ using the generalized power mean inequality. We can write

$$\exp(-Z_1^2) = \mathbb{E}[(x_i)^{1/(2b_1^2)}]^{2b_1^2}$$

and

$$\exp(-Z_2^2) = \mathbb{E}[(x_i)^{1/(2b_2^2)}]^{2b_2^2},$$

where $x_i = \exp(-q_i^2)$. Since $\frac{1}{2b_1^2} < \frac{1}{2b_2^2}$, we know $\exp(-Z_1^2) \leq \exp(-Z_2^2)$ by the power mean inequality, and thus $Z_1 \geq Z_2$, again with equality if and only if q_i is constant. ■

Now we proceed to restate and prove Theorem 1.

Theorem 1 (Buffering increases the asymptotic PPV difference) *Consider two GPs, GP_1 and GP_2, sharing common homoskedastic noise variance σ_n^2. Moreover, assume that these GPs have isotropic Gaussian kernel functions $k_1(q) = \exp\left(-\|q\|^2/(2b_1^2)\right)$ and $k_2(q) = \exp\left(-\|q\|^2/(2b_2^2)\right)$, where the length scale $b_1 > b_2$. For analytic tractability, assume all training points $\{x_i\}_{i=1}^N$ are at distance ρ from the test point x_*, the distance distribution from each training point to all other training points is the same (assume pairwise distances are drawn $q \sim Q$), and $\mathbb{E}[q] < \rho\sqrt{2}$. Then the difference in posterior predictive variances $\sigma_1^2(x_*) - \sigma_2^2(x_*)$ satisfies the following:*

1. *If $\rho = 0$, then $\lim_{N\to\infty}\left(\sigma_2^2(x_*) - \sigma_1^2(x_*)\right) = 0$.*
2. *If $\rho > 0$, then $\lim_{N\to\infty}\left(\sigma_2^2(x_*) - \sigma_1^2(x_*)\right) > 0$.*
3. *$\lim_{N\to\infty}\left(\sigma_2^2(x_*) - \sigma_1^2(x_*)\right)$ is bounded below by a function $f(\rho)$, which increases from $f(0) = 0$ at $\rho = 0$ through $\rho = b_1 b_2\sqrt{\frac{2\log\left(\frac{b_1}{b_2}\right)}{(2-\theta^2)(b_1^2-b_2^2)}} > 0$ (see proof for definition of θ).*

Proof
Implication 1:
Applying Lemma 1, and noting $k_1(0) = k_2(0)$, we obtain

$$\begin{aligned}\sigma_2^2(x_*) - \sigma_1^2(x_*) &= \left[\sigma_n^2 + k_2(0) - \frac{k_2(\rho)^2}{\frac{\sigma_n^2}{N} + \mathbb{E}[k_2(q)]}\right] - \left[\sigma_n^2 + k_1(0) - \frac{k_1(\rho)^2}{\frac{\sigma_n^2}{N} + \mathbb{E}[k_1(q)]}\right] \\ &= \frac{k_1(\rho)^2}{\frac{\sigma_n^2}{N} + \mathbb{E}[k_1(q)]} - \frac{k_2(\rho)^2}{\frac{\sigma_n^2}{N} + \mathbb{E}[k_2(q)]}.\end{aligned}$$

Then taking the limit as $N \to \infty$, the noise variance drops out:

$$\lim_{N\to\infty}\left(\sigma_2^2(x_*) - \sigma_1^2(x_*)\right) = \lim_{N\to\infty}\left[\frac{k_1(\rho)^2}{\frac{\sigma_n^2}{N} + \mathbb{E}[k_1(q)]} - \frac{k_2(\rho)^2}{\frac{\sigma_n^2}{N} + \mathbb{E}[k_2(q)]}\right] = \frac{k_1(\rho)^2}{\mathbb{E}[k_1(q)]} - \frac{k_2(\rho)^2}{\mathbb{E}[k_2(q)]}.$$

If $\rho = 0$, then the pairwise distances q are uniformly 0, so this reduces to

$$\frac{k_1(0)^2}{k_1(0)} - \frac{k_2(0)^2}{k_2(0)} = k_1(0) - k_2(0) = 0.$$

Implication 2:
Now let $z_i = k_i^{-1}(\mathbb{E}[k_i(q)])$. Then this expression simplifies to

$$\begin{aligned}\frac{k_1(\rho)^2}{\mathbb{E}[k_1(q)]} - \frac{k_2(\rho)^2}{\mathbb{E}[k_2(q)]} &= \frac{k_1(\rho)^2}{k_1(z_1)} - \frac{k_2(\rho)^2}{k_2(z_2)} \\ &\geq \frac{k_1(\rho)^2}{k_1(z_2)} - \frac{k_2(\rho)^2}{k_2(z_2)},\end{aligned}$$

since $z_2 \leq z_1$ by Lemma 2. Next, let $\theta = \frac{z_2}{\rho}$, and note that $z_2 \leq \mathbb{E}[q] \leq \rho\sqrt{2}$ implies $\theta = z_2/\rho \leq \sqrt{2}$. Then, we can write $k_i(z_2) = k_i(\rho)^{\theta^2}$, simplifying the lower bound to

$$\begin{aligned}\frac{k_1(\rho)^2}{k_1(z_2)} - \frac{k_2(\rho)^2}{k_2(z_2)} &= \frac{k_1(\rho)^2}{k_1(\rho)^{\theta^2}} - \frac{k_2(\rho)^2}{k_2(\rho)^{\theta^2}} \\ &= \exp\left(\frac{-\rho^2(2-\theta^2)}{2b_1^2}\right) - \exp\left(\frac{-\rho^2(2-\theta^2)}{2b_2^2}\right).\end{aligned}$$

But then noting $-\rho^2(2-\theta^2) \leq 0$, we have

$$\begin{aligned}b_1 > b_2 \implies & \quad 1/b_2 > 1/b_1 \\ \implies & \quad \frac{-\rho^2(2-\theta^2)}{2b_2} \leq \frac{-\rho^2(2-\theta^2)}{2b_1} \\ \implies & \quad \exp\left(\frac{-\rho^2(2-\theta^2)}{2b_2^2}\right) \leq \exp\left(\frac{-\rho^2(2-\theta^2)}{2b_1^2}\right) \\ \implies & \exp\left(\frac{-\rho^2(2-\theta^2)}{2b_1^2}\right) - \exp\left(\frac{-\rho^2(2-\theta^2)}{2b_2^2}\right) \geq 0,\end{aligned}$$

with equality only if $\theta = \sqrt{2}$, or equivalently if $z_2^2 = 2\rho^2$. But, by Lemma 2, this cannot be the case if $0 < \mathbb{E}[q] < \rho\sqrt{2}$, completing the proof of the second implication.

Implication 3:

We now show that the asymptotic difference in posterior variances is bounded below by a function $f(\rho)$, which increases from $f(0) = 0$ at $\rho = 0$, through $\rho = b_1 b_2 \sqrt{\frac{2\log\left(\frac{b_1}{b_2}\right)}{(2-\theta^2)(b_1^2 - b_2^2)}}$. Specifically, continuing from above, let

$$f(\rho) = \exp\left(\frac{-\rho^2(2-\theta^2)}{b_1^2}\right) - \exp\left(\frac{-\rho^2(2-\theta^2)}{b_2^2}\right),$$

and consider its first and second derivatives with respect to ρ.

$$\begin{aligned}
f'(\rho) &= -\frac{2\rho(2-\theta^2)}{b_1^2}\exp\left(\frac{-\rho^2(2-\theta^2)}{b_1^2}\right) + \frac{2\rho(2-\theta^2)}{b_2^2}\exp\left(\frac{-\rho^2(2-\theta^2)}{b_2^2}\right) \\
&= 2\rho(2-\theta^2)\left(\frac{1}{b_2^2}\exp\left(\frac{-\rho^2(2-\theta^2)}{b_2^2}\right) - \frac{1}{b_1^2}\exp\left(\frac{-\rho^2(2-\theta^2)}{b_1^2}\right)\right); \\
f''(\rho) &= 2(2-\theta^2)\left(\frac{1}{b_2^2}\exp\left(\frac{-\rho^2(2-\theta^2)}{b_2^2}\right) - \frac{1}{b_1^2}\exp\left(\frac{-\rho^2(2-\theta^2)}{b_1^2}\right)\right) + \\
&\quad 4\rho^2(2-\theta^2)^2\left(\frac{1}{b_1^4}\exp\left(\frac{-\rho^2(2-\theta^2)}{b_1^2}\right) - \frac{1}{b_2^4}\exp\left(\frac{-\rho^2(2-\theta^2)}{b_2^2}\right)\right).
\end{aligned}$$

So $f'(\rho)$ at $\rho = 0$ is 0, and $f''(\rho)$ at $\rho = 0$ is $2(2-\theta^2)(1/b_2^2 - 1/b_1^2) > 0$. Hence, as we increase ρ from 0, the lower bound on the difference in posterior predictive variances increases. Moreover, we find a $\rho > 0$ for which $f'(\rho) = 0$:

$$\begin{aligned}
& 0 = \frac{1}{b_2^2}\exp\left(\frac{-\rho^2(2-\theta^2)}{b_2^2}\right) - \frac{1}{b_1^2}\exp\left(\frac{-\rho^2(2-\theta^2)}{b_1^2}\right) \\
\iff \qquad & \rho = b_1 b_2 \sqrt{\frac{2\log\left(\frac{b_1}{b_2}\right)}{(2-\theta^2)(b_1^2 - b_2^2)}}.
\end{aligned}$$

■

Appendix C. Applying Kallus et al. (2018) and Cattaneo et al. (2020)

C.1 Kallus et al.

Kallus et al.'s method of experimental grounding (Kallus et al., 2018) requires two data samples: an observational sample $\mathcal{D}^{Conf}$, and an unconfounded sample $\mathcal{D}^{Unc}$. However, in our context, we only have access to a single observational dataset $\mathcal{D} = \{(x_i, t_i, y_i)\}_{i=1}^N$. As such, in order to apply Kallus's method, we treat the discovered RDs as local randomized experiments. Specifically, we:

1. Apply LoRD^3 to the full dataset $\mathcal{D}$ to discover a set $\mathcal{L}$ of local regression discontinuities, each associated with a k-NN ball.

2. Randomly partition the full dataset $\mathcal{D}$ into two equally-sized subsamples $\mathcal{D}^1$ and $\mathcal{D}^2$.
3. Treat $\mathcal{D}^1$ as the confounded sample $\mathcal{D}^{Conf}$.
4. Treat the instances in $\mathcal{D}^2$ that fall within one of the discovered RD's k-NN balls (i.e. instances in $\mathcal{D}^2$ that are local to a discovered RD) as the "unconfounded" set $\mathcal{D}^{Unc}$.

Again, Kallus's method could prove effective if the discovered RDs in fact served as local randomized experiments, such that $\{Y(1), Y(0)\} \perp T \mid X$ for instances in the neighborhood of the discovered RDs. However, while sharp RD designs can be viewed as local randomized experiments (Lee and Lemieux, 2010), viewing fuzzy RD designs as randomized experiments requires considering the potential for selective compliance. As such, in our context, we do not expect Kallus's method to offer gains over the confounded observational estimator.

C.2 Cattaneo et al.

Cattaneo et al. (2020) provide a method for extrapolating a LATE-type parameter in the fuzzy RD setting. In order to apply this method, we must satisfy two requirements. First, Cattaneo's method is applicable when there are at least two discontinuities, ℓ and h, in the treatment propensity along a common running variable. Second, their method requires a dataset where each unit (z_i, t_i, y_i) is associated with a known discontinuity $c_i \in \{\ell, h\}$. This second requirement poses a substantial challenge in our context, since individual threshold assignments c_i are clearly latent where RDs are themselves unknown.

To meet the first requirement, we restrict our application of Cattaneo's method to Simulation 1 (see Appendix D.1), where there are multiple discontinuities along a common running variable (see Figure 10). In particular, in Simulation 1 we can apply Cattaneo's method twice, first to discontinuities at $x_1 \in \{\ell, h\}$ in $P(T = 1 \mid X_1 = x_1)$, and second to discontinuities at $x_2 \in \{\ell, h\}$ in $P(T = 1 \mid X_2 = x_2)$, where $\ell = 0.25$ and $h = 0.65$. We also simplify our problem: instead of requiring RD discovery, we provide Cattaneo's method with oracle access to the true underlying RDs.

To meet the second requirement, we must assign each instance an imputed threshold $\hat{c}_i \in \{\ell, h\}$. We desire an imputation strategy that assigns compliers falling between the two thresholds ℓ and h to the correct threshold. We consider the behavior of compliers for each of the two discontinuities $c = \{\ell, h\}$ across different values of x:

1. **Case 1**: If $x \leq \ell$, then compliers for both discontinuities are treated, and treatment uptake is uninformative.
2. **Case 2**: If $\ell < x \leq h$, then compliers for the $c = h$ discontinuity are treated, and compliers for the $c = \ell$ discontinuity are untreated.
3. **Case 3**: If $x > h$, then compliers for both discontinuities are untreated, and treatment uptake is uninformative.

As such, the following randomized imputation $\hat{c}_i(x_i, t_i)$ is guaranteed to assign compliers between the two thresholds ℓ and h to the correct threshold, such that $\hat{c}_i(x_i, t_i) = c_i$:

$$\hat{c}_i(x_i, t_i) = \begin{cases} P(\hat{c}_i = \ell) = P(\hat{c}_i = h) = 0.5 & x_i \leq \ell & \text{(Case 1)} \\ \ell & \ell < x_i \leq h, t_i = 0 & \text{(Case 2: Untreated compliers)} \\ h & \ell < x_i \leq h, t_i = 1 & \text{(Case 2: Treated compliers)} \\ P(\hat{c}_i = \ell) = P(\hat{c}_i = h) = 0.5 & x_i > h & \text{(Case 3)} \end{cases}$$

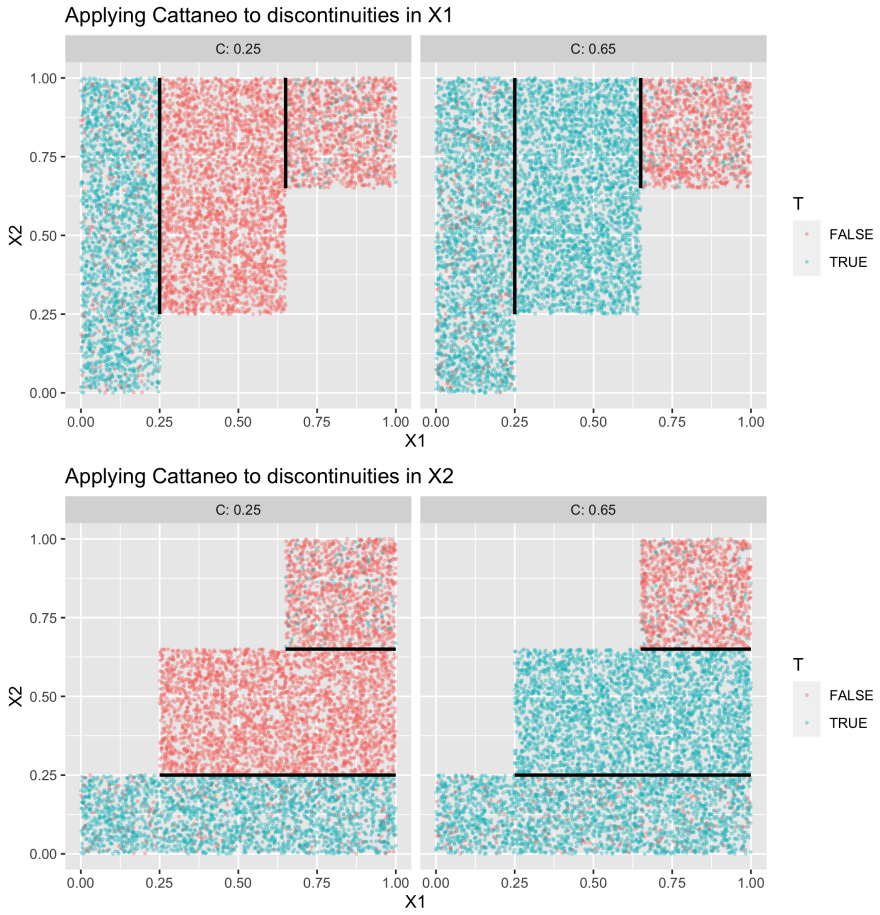

Figure 10: Imputed threshold assignments c_i and data subsets used to extrapolate from multiple discontinuities in X_1 (top row) and X_2 (bottom row). Color shows the unit's treatment status.

Figure 10 shows the imputed threshold assignments $\hat{c}_i$ (columns) for discontinuities along each of the two dimensions (rows), along with each unit's treatment status (color).

Given this setup, we apply Cattaneo's method twice. First, we apply Cattaneo's method along X_1 to estimate an extrapolated LATE-type parameter

$$\tau^{Catt,X_1}(x_1) = \mathbb{E}[Y(1) - Y(0) \mid X_1 = x_1, \hat{c} = h,\ Complier] \text{ for } x_1 \in (\ell, h).$$

Applying Cattaneo's method along X_2 yields a similar estimate for

$$\tau^{Catt,X_2}(x_2) = \mathbb{E}[Y(1) - Y(0) \mid X_2 = x_2, \hat{c} = h,\ Complier] \text{ for } x_2 \in (\ell, h).$$

While Cattaneo's method yields estimates for these LATE-type parameters, our goal is to estimate the CATE $\tau(x)$. To that end, we approximate the CATE using a simple average of $\hat{\tau}^{Catt,X_1}(x_1)$ and $\hat{\tau}^{Catt,X_2}(x_2)$:

$$\frac{1}{2}\left(\hat{\tau}^{Catt,X_1}(x_1) + \hat{\tau}^{Catt,X_2}(x_2)\right) \approx \tau(x_1, x_2) \text{ for } (x_1, x_2) \in (\ell, h)^2.$$

For instances $(x_1, x_2) \notin (\ell, h)^2$, we project onto the set $(\ell, h)^2$ before computing this average. In practice, we do not necessarily expect this method to yield a competitive CATE estimate, since Simulation 1 does not satisfy the assumptions required for Cattaneo's method to yield unbiased estimates for the LATE-type parameters $\tau^{Catt,X_1}(x_1)$ and $\tau^{Catt,X_2}(x_2)$, let alone prove a context where this simple average will yield an unbiased estimate of the CATE.

Appendix D. Simulation 1

D.1 Simulated DGP

The simulated DGP for our first simulation is as follows:

$$
\begin{aligned}
X &\sim Unif([0,1]^2) && \text{(Covariates)}\\
\tau(X) &\sim GP(0, k_{\Theta_\tau}) && \text{(CATE GP prior)}\\
\beta(X) &\sim GP(0, k_{\Theta_\beta}) && \text{(Bias GP prior)}\\
P(G=g) &= \begin{cases} 0.1 & A \text{ (Always-taker)}\\ 0.4 & C_1 \text{ (Complier type 1)}\\ 0.4 & C_2 \text{ (Complier type 2)}\\ 0.1 & N \text{ (Never-taker)} \end{cases} && \text{(Treatment uptake group)}\\
Z_1 &= (X_1 \le 0.65) \vee (X_2 \le 0.65) && \text{(First RD instrument)}\\
Z_2 &= (X_1 \le 0.25) \vee (X_2 \le 0.25) && \text{(Second RD instrument)}\\
T &= \begin{cases} 1 & G = A\\ 1 & (G = C_1) \wedge (Z_1 = 1)\\ 1 & (G = C_2) \wedge (Z_2 = 1)\\ 0 & \text{Otherwise} \end{cases} && \text{(Observed treatment)}\\
g(X,G) &= \begin{cases} \frac{1}{2}\beta(X) & G = A\\ 0 & G = C_1\\ -\frac{4}{5}\beta(X) & G = C_2\\ -\frac{13}{10}\beta(X) & G = N \end{cases} && \text{(Group-specific shift in potential outcomes)}\\
\epsilon &\sim \mathcal{N}(0,1) && \text{(Outcome noise)}\\
Y(1) &= \frac{1}{2}\tau(X) + g(X,G) + \epsilon && \text{(Potential outcome)}\\
Y(0) &= -\frac{1}{2}\tau(X) + g(X,G) + \epsilon && \text{(Potential outcome)}\\
Y &= Y(T) && \text{(Observed outcome)}
\end{aligned}
$$

Note that this DGP satisfies the CCE/CEI assumption, since

$$\mathbb{E}[Y(1) - Y(0) \mid X = x, G = g] = \tau(x) \ \forall g.$$

Moreover, we note that $\beta(x)$ is in fact the bias in a conditioned-on-observables approximation to $\tau^{obs}(x)$. This can be confirmed using iterated expectations, iterating over G.

We confirm by considering three cases:

1. **Case 1:** $Z_1 = 0$ and $Z_2 = 0$. In this region, the treated subpopulation is composed entirely of always-takers $G = A$, and the untreated subpopulation mixes never-takers and both complier types $G \in \{C_1, C_2, N\}$.

$$\begin{aligned}
\tau^{obs}(x) &= \mathbb{E}[Y(1) \mid T = 1, X = x] - \mathbb{E}[Y(0) \mid T = 0, X = x] \\
&= \underbrace{\mathbb{E}_G[\mathbb{E}[Y(1) \mid T = 1, X = x, G = g]]}_{G=A} - \underbrace{\mathbb{E}_G[\mathbb{E}[Y(0) \mid T = 0, X = x, G = g]]}_{G\in\{C_1,C_2,N\}} \\
&= \mathbb{E}\left[\frac{1}{2}\tau(X) + g(X,G) + \epsilon \mid X = x, G = a\right] \\
&\quad - \frac{4}{9}\mathbb{E}\left[-\frac{1}{2}\tau(X) + g(X,G) + \epsilon \mid X = x, G = C_1\right] \\
&\quad - \frac{4}{9}\mathbb{E}\left[-\frac{1}{2}\tau(X) + g(X,G) + \epsilon \mid X = x, G = C_2\right] \\
&\quad - \frac{1}{9}\mathbb{E}\left[-\frac{1}{2}\tau(X) + g(X,G) + \epsilon \mid X = x, G = N\right] \\
&= \underbrace{\frac{1}{2}\tau(x) + \frac{1}{2}\beta(x)}_{G=A} - \underbrace{\frac{4}{9}\left(-\frac{1}{2}\tau(x)\right)}_{G=C_1} - \underbrace{\frac{4}{9}\left(-\frac{1}{2}\tau(x) - \frac{4}{5}\beta(x)\right)}_{G=C_2} - \underbrace{\frac{1}{9}\left(-\frac{1}{2}\tau(x) - \frac{13}{10}\beta(x)\right)}_{G=C_2} \\
&= \tau(x) + \beta(x)
\end{aligned}$$

2. **Case 2:** $Z_1 = 1$ and $Z_2 = 0$. In this region, the treated subpopulation mixes always-takers and complier type C_1, while the untreated subpopulation mixes never-takers and complier type C_2.

$$\begin{aligned}
\tau^{obs}(x) &= \mathbb{E}[Y(1) \mid T = 1, X = x] - \mathbb{E}[Y(0) \mid T = 0, X = x] \\
&= \underbrace{\mathbb{E}_G[\mathbb{E}[Y(1) \mid T = 1, X = x, G = g]]}_{G\in\{A,C_1\}} - \underbrace{\mathbb{E}_G[\mathbb{E}[Y(0) \mid T = 0, X = x, G = g]]}_{G\in\{C_2,N\}} \\
&= \frac{1}{5}\mathbb{E}\left[\frac{1}{2}\tau(X) + g(X,G) + \epsilon \mid X = x, G = a\right] \\
&\quad + \frac{4}{5}\mathbb{E}\left[\frac{1}{2}\tau(X) + g(X,G) + \epsilon \mid X = x, G = C_1\right] \\
&\quad - \frac{4}{5}\mathbb{E}\left[-\frac{1}{2}\tau(X) + g(X,G) + \epsilon \mid X = x, G = C_2\right] \\
&\quad - \frac{1}{5}\mathbb{E}\left[-\frac{1}{2}\tau(X) + g(X,G) + \epsilon \mid X = x, G = N\right] \\
&= \underbrace{\frac{1}{5}\left(\frac{1}{2}\tau(x) + \frac{1}{2}\beta(x)\right)}_{G=A} + \underbrace{\frac{4}{5}\left(\frac{1}{2}\tau(x)\right)}_{G=C_1} - \underbrace{\frac{4}{5}\left(-\frac{1}{2}\tau(x) - \frac{4}{5}\beta(x)\right)}_{G=C_2} - \underbrace{\frac{1}{5}\left(-\frac{1}{2}\tau(x) - \frac{13}{10}\beta(x)\right)}_{G=C_2} \\
&= \tau(x) + \beta(x)
\end{aligned}$$

3. **Case 3:** $Z_1 = 1$ and $Z_2 = 1$. In this region, the treated subpopulation mixes always-takers and both complier types C_1 and C_2, while the untreated subpopulation is composed entirely of never-takers.

$$
\begin{aligned}
\tau^{obs}(x) &= \mathbb{E}[Y(1) \mid T=1, X=x] - \mathbb{E}[Y(0) \mid T=0, X=x] \\
&= \underbrace{\mathbb{E}_G[\mathbb{E}[Y(1) \mid T=1, X=x, G=g]]}_{G \in \{A, C_1, C_2\}} - \underbrace{\mathbb{E}_G[\mathbb{E}[Y(0) \mid T=0, X=x, G=g]]}_{G=N} \\
&= \frac{1}{9}\mathbb{E}\left[\frac{1}{2}\tau(X) + g(X,G) + \epsilon \mid X=x, G=a\right] \\
&\quad + \frac{4}{9}\mathbb{E}\left[\frac{1}{2}\tau(X) + g(X,G) + \epsilon \mid X=x, G=C_1\right] \\
&\quad + \frac{4}{9}\mathbb{E}\left[\frac{1}{2}\tau(X) + g(X,G) + \epsilon \mid X=x, G=C_2\right] \\
&\quad - \mathbb{E}\left[-\frac{1}{2}\tau(X) + g(X,G) + \epsilon \mid X=x, G=N\right] \\
&= \underbrace{\frac{1}{9}\left(\frac{1}{2}\tau(x) + \frac{1}{2}\beta(x)\right)}_{G=A} + \underbrace{\frac{4}{9}\left(\frac{1}{2}\tau(x)\right)}_{G=C_1} + \underbrace{\frac{4}{9}\left(\frac{1}{2}\tau(x) - \frac{4}{5}\beta(x)\right)}_{G=C_2} - \underbrace{\left(-\frac{1}{2}\tau(x) - \frac{13}{10}\beta(x)\right)}_{G=C_2} \\
&= \tau(x) + \beta(x)
\end{aligned}
$$

In all three cases, $\tau^{obs}(x) = \tau(x) + \beta(x)$, implying that $\beta(x)$ is the bias in the conditioned-on-observables approximation of $\tau(x)$.

Finally, we note that LoRD^3 does not require access to Y. As such, in simulation, we apply the following procedure to sample $N = 20,000$ training instances, and evaluate $\tau(x)$ and $\beta(x)$ at each (i) $x \in \mathcal{D}$, (ii) $x \in \mathcal{U}$, and (iii) x on our test grid:

1. Sample X and G, and compute T, for all $20,000$ training instances in $\mathcal{D}$.
2. Run LoRD^3 to compute $\mathcal{L}$, then apply Algorithm 1 to map $\mathcal{L} \to \mathcal{U}$.
3. Sample $\tau(x)$ and $\beta(x)$ from the GP priors and evaluate at all x in $\mathcal{D}, \mathcal{U}$, and the test grid.
4. Compute $Y(0)$ and $Y(1)$ for each instance in the training set $\mathcal{D}$.
5. Proceed to estimate $\hat{\tau}^{obs}(x)$ and to estimate $\hat{\tau}(x)$ for each $x \in \mathcal{U}$.

Empirical results for Simulation 1 are provided in the main text.

D.2 Improvements from BLOOCV in Large Sample

In Section 4, we showed theoretically that BLOOCV addresses the asymptotic degeneracy of LOOCV. Specifically, we showed that as the number of repaired discontinuities ($\mathcal{U}$), from Algorithm 1 increases, LOOCV will fail to differentiate between our two candidate models $\mathcal{M}_\tau$ and $\mathcal{M}_\beta$. We further showed that our novel BLOOCV-based predictive log-likelihood would not suffer from the same degeneracy as $|\mathcal{U}| \to \infty$. In this section, we empirically

validate these results and further demonstrate the effect on MSE of the extrapolated CATE estimates using the BLOOCV method.

We use the same setup as Simulation 1 for the off-diagonal regime with $\theta_\tau = 0.2$ and $\theta_\beta = 0.5$, where we saw a significant difference in MSE between the direct CATE extrapolation model $\mathcal{M}_\tau$ and the bias-correcting model $\mathcal{M}_\beta$ (see Figure 3). In order to increase $|\mathcal{U}|$, we increase the number of training samples $N = 200,000$.[19] From Figure 11, we observe that the difference between LOOCV log-likelihoods under $\mathcal{M}_\tau$ and $\mathcal{M}_\beta$ is close to 0. Similarly, the difference between marginal log-likelihoods between $\mathcal{M}_\tau$ and $\mathcal{M}_\beta$ is close to 0. More importantly, we further observe that both BLOOCV variants are able to identify the significant difference between log-likelihoods under $\mathcal{M}_\tau$ and $\mathcal{M}_\beta$. This corresponded to a significant improvement in MSE for the BLOOCV variants as compared to MLL and LOOCV, as shown in Figure 12.

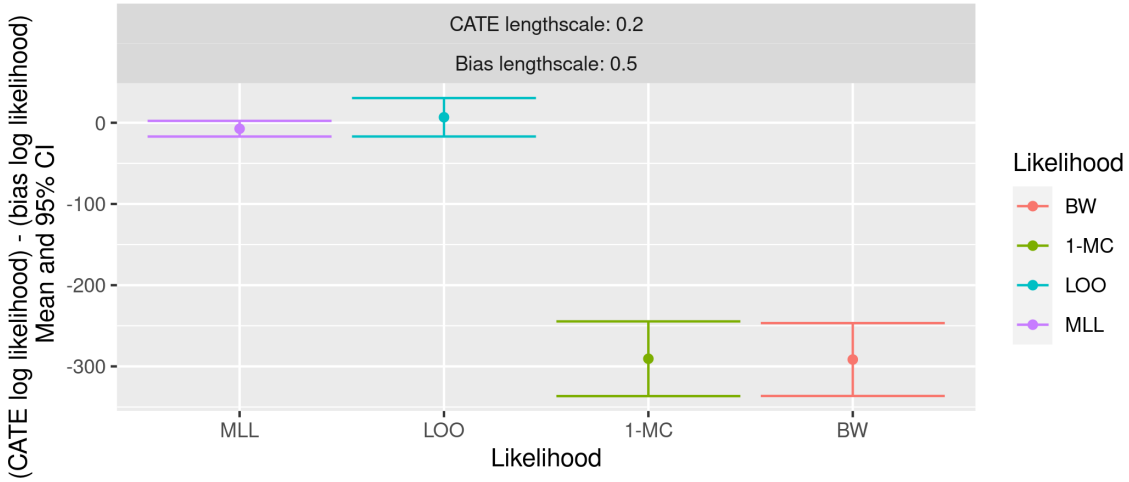

Figure 11: Difference between various log-likelihoods, under the CATE model $\mathcal{M}_\tau$ and the bias model $\mathcal{M}_\beta$, in a large sample setting, $N = 200,000$, under the off-diagonal regime with $\theta_\tau = 0.2$ and $\theta_\beta = 0.5$.

19. We generated more training samples at the discontinuities to increase $|\mathcal{U}|$ without significantly increasing N, which increases the memory requirements during the sampling from the GPs in step 3 of the simulation. As a result, $|\mathcal{U}|$ changed from around 50 to 60 in our main paper (across 50 independent replications) to around 500 to 600.

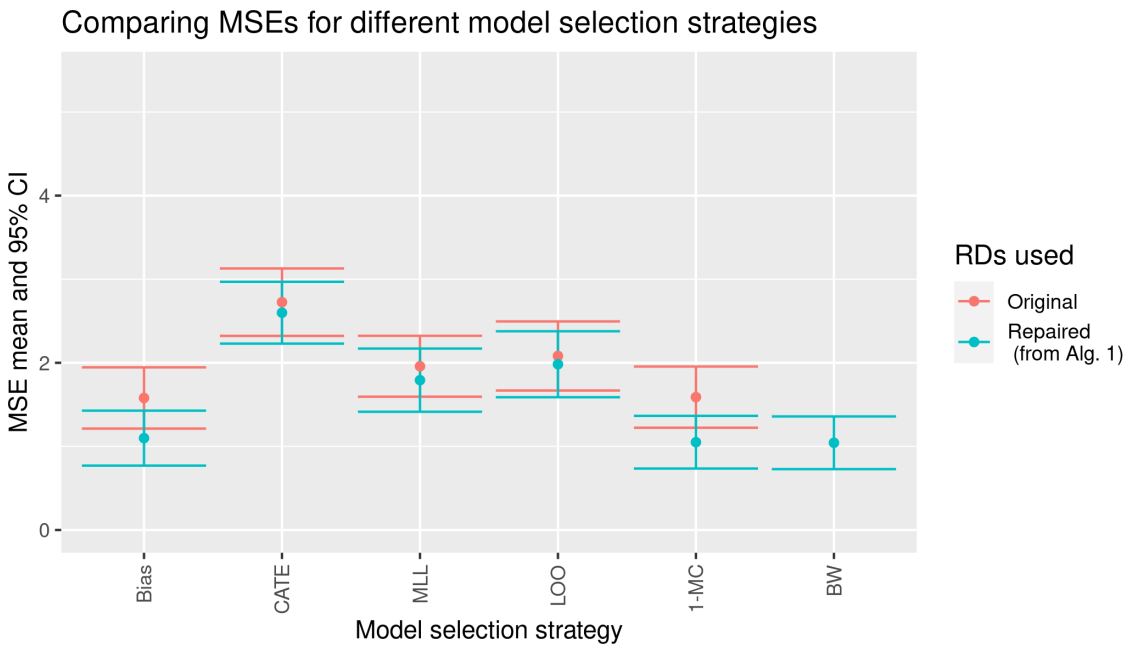

Figure 12: Average MSEs evaluated over a uniform mesh grid, with $N = 100 \times 100$ points, approximating performance over $P(X) = Unif([0,1]^2)$ in a large sample setting, $N = 200,000$, under the off-diagonal regime with $\theta_\tau = 0.2$ and $\theta_\beta = 0.5$. The x-axis gives the model selection strategy. "Bias" and "CATE" refer to the strategies of always selecting the bias correcting model $\mathcal{M}_\beta$, and the direct extrapolation model $\mathcal{M}_\tau$, respectively. "MLL", "LOO", "1-MC", and "BW" refer to model averaging strategies where model weights are computed using the marginal likelihood, the LOOCV likelihood, the 1-MC BLOOCV variant, and the weighted BLOOCV variant, respectively. The blue trace shows MSEs for CATE estimates obtained from the full inferential procedure of `DEE`, including the repair procedure in Algorithm 1 (which maps $\mathcal{L} \to \mathcal{U}$). The red trace shows the MSEs obtained when `DEE` does not apply Algorithm 1 and instead directly applies Gaussian process regression to local 2SLS estimates, computed using overlapping k-neighborhoods ($k = 200$), for each discovered local RD in $\mathcal{L}$. Note that we do not use weighted BLOOCV when using direct GP extrapolation from $\mathcal{L}$ due to the $\mathcal{O}(|\mathcal{L}|^5)$ runtime.

D.3 Using KNN-only and Voronoi-only procedures in Algorithm 1

Empirically, performing the KNN-only procedure (i.e., including just the index sets $K[x]$) in Algorithm 1 will result in the same set of discontinuities as not running Algorithm 1 (the original RDs), as shown in the right panel of Figure 13. This is because the index sets with KNN-only remain static and do not change over the iterations of Algorithm 1. As we can see from the red-colored error bars in Figure 3, empirically, this yields worse MSEs as compared to a combination of Voronoi and KNN procedures. Finally, performing the Voronoi-only

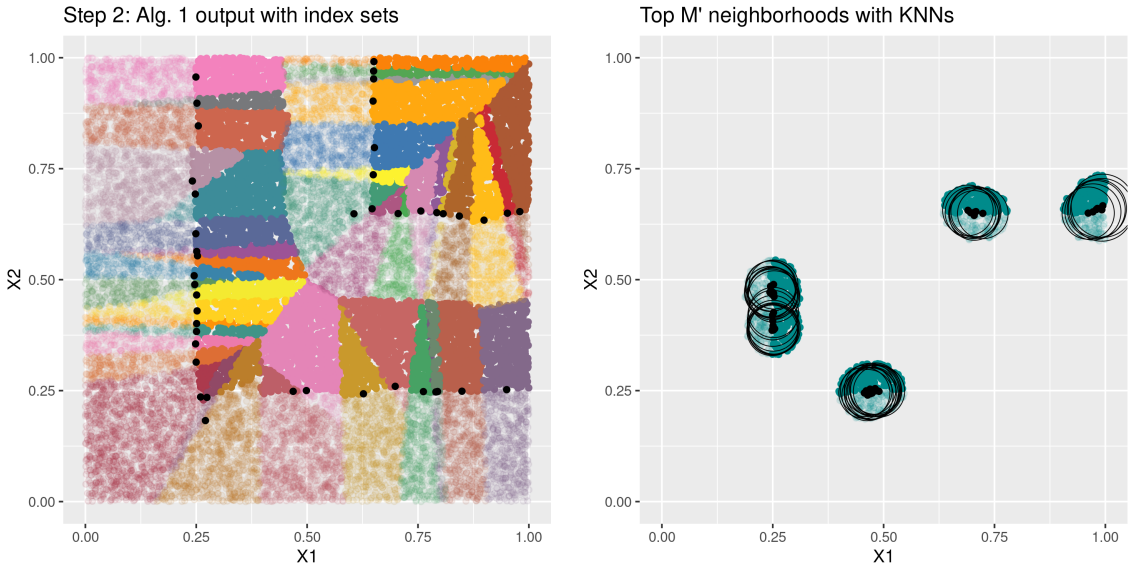

Figure 13: Index sets for Algorithm 1 with Voronoi-only procedure (left panel) and KNN-only procedure (right panel).

procedure (i.e., including just the index sets $V[x]$) in Algorithm 1 is shown in the left panel of Figure 13. We observe that though the discovered RDs are distinct, their index sets include points far away from the discontinuity, which can bias our estimation of treatment effects at the RDs. Therefore, the Voronoi ∩ KNN procedure we used in Algorithm 1 in §3.2.1 ensures that we include only the points local to the discontinuity (based on KNN) and that the index sets are disjoint (based on Voronoi), and the resulting treatment effect estimates are valid generally at RDs.

Appendix E. Simulation 2

E.1 Simulated DGP

Before giving the DGP for our second simulation, we introduce a number of auxiliary functions. Let Φ denote the Gaussian cumulative density function (cdf), and ϕ denote the Gaussian probability density function (pdf). We construct the CATE $\tau(x)$ and the (scaled) covariance function $Cov(x)$ by first sampling Gaussian Process priors $GP(0, k_{\Theta_\tau})$ and $GP(0, k_{\Theta_{Cov}})$ at a set of inducing points (a 50×50 uniform mesh grid on $[0,1]^2$). We then define $\tau(x)$ and $Cov(x)$ as the posterior means of the same Gaussian process, conditioned on these samples. Again, we let the length scales θ_τ and θ_{Cov} be free parameters in our simulation.

While in Simulation 1 we set the prior output scale $k_{\Theta_\tau} = k_{\Theta_\beta}$, in Simulation 2 we do not directly specify a GP prior over the bias function $\beta(x)$; instead, we specify the GP prior over the covariance function $Cov(x)$. Thus, we cannot simply use the same output scale for the prior kernels k_{Θ_τ} and k_{Θ_β}. However, we note that $Cov(x) = \rho(x)\sigma$, so the bias

$\beta(x) = Cov(x)\frac{\phi(\gamma)}{\Phi(\gamma)(1-\Phi(\gamma))}$ for some x-dependent γ. Given this observation, we define the output scales $k_\tau(0)$ and $k_{Cov}(0)$ as follows. First, we let $k_{Cov}(0) = \frac{1}{4}$. Then, we let c denote the weighted average of the multiplier $\frac{\phi(\gamma)}{\Phi(\gamma)(1-\Phi(\gamma))}$, where we average over the distribution on γ induced by $X \sim Unif([0,1]^2)$. Finally, we let $k_\tau(0) = \frac{c^2}{4}$.

Given these functions, the DGP is:

$$
\begin{aligned}
X &\sim Unif([0,1]^2) && \text{(Observed covariates)}\\
\gamma_0 &= \Phi^{-1}(0.1), \gamma_1 = 0, \gamma_2 = \Phi^{-1}(0.9) && \text{(Latent treatment index thresholds)}\\
Z_1 &= (\max\{X_1, X_2\} > 0.5) \wedge (\min\{X_1, X_2\} < 0.5) && \text{(First RD instrument)}\\
Z_2 &= (\min\{X_1, X_2\} > 0.5) && \text{(Second RD instrument)}\\
\begin{bmatrix} U^T(x) \\ U^Y(x) \end{bmatrix} &\sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho(x)\sigma \\ \rho(x)\sigma & \sigma^2 \end{bmatrix}\right) && \text{(Unobserved confounders)}\\
T(0) &= \mathbb{1}[\gamma_0 > U^T] && \text{(Potential treatment)}\\
T(1) &= \mathbb{1}[\gamma_1 > U^T] && \text{(Potential treatment)}\\
T(2) &= \mathbb{1}[\gamma_2 > U^T] && \text{(Potential treatment)}\\
T &= T(1)Z_1 + T(2)Z_2 + T(0)(1-Z_1)(1-Z_2) && \text{(Observed treatment)}\\
Y(0) &= -\frac{1}{2}\tau(X) + U^Y && \text{(Potential outcome)}\\
Y(1) &= \frac{1}{2}\tau(X) + U^Y && \text{(Potential outcome)}\\
Y &= Y(T) && \text{(Observed outcome)}
\end{aligned}
$$

Note that, in this DGP, the variance σ^2 of U^Y is an unspecified constant, and the correlation function $\rho(x)$ is also (up to this point) undefined. In simulation, we compute σ^2 and define $\rho(x)$ as follows:

1. First, we draw the $N = 20,000$ training set x_i's.
2. We then compute $Cov(x_i)$ for each training x_i.
3. We then set σ to
$$\sigma = \frac{\max_{i=1\ldots20,000} |Cov(x_i)|}{0.9}$$
and define the correlation function $\rho(x) = Cov(x)/\sigma$. This ensures that the correlation at each training x_i is between -0.9 and 0.9.

E.2 Empirical Results for Simulation 2

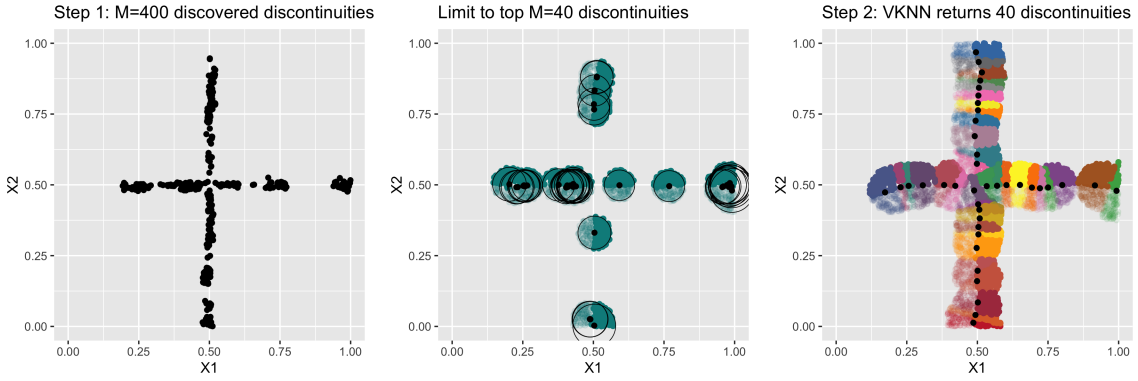

Figure 14: Illustrating the Discover and Estimate steps of `DEE` for Simulation 2. The left panel shows the $M = |\mathcal{L}| = 400$ local discontinuities initially selected using LoRD^3, while the right panel shows the subset of discontinuities (and their index sets) output by Algorithm 1. As seen in the middle panel, applying a higher LLR threshold when constructing $\mathcal{L}$ fails to solve the problems of neighborhood overlap and RD coverage.

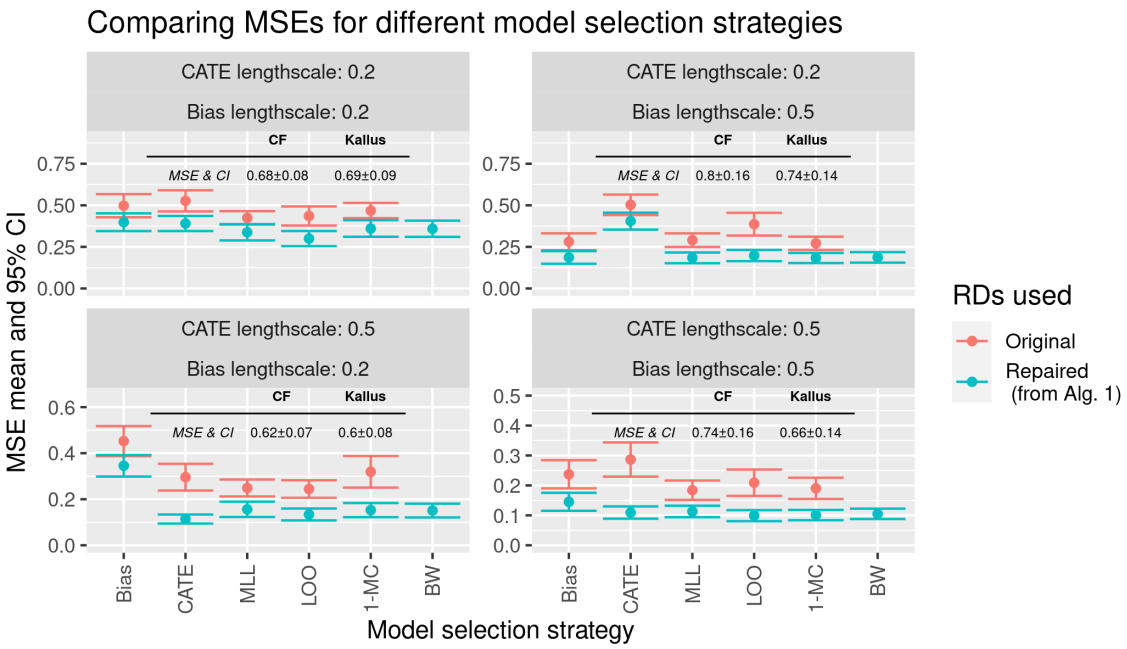

Figure 15: Average MSEs evaluated over a uniform mesh grid, with $N = 100 \times 100$ points, approximating performance over $P(X) = Unif([0,1]^2)$ for Simulation 2. The $x-$axis gives the model selection strategy. "Bias" and "CATE" refer to the strategies of always selecting the bias correcting model $\mathcal{M}_\beta$, and the direct extrapolation model $\mathcal{M}_\tau$, respectively. "MLL", "LOO", "1-MC", and "BW" refer to model averaging strategies where model weights are computed using the marginal likelihood, the LOOCV likelihood, the 1-MC BLOOCV variant, and the weighted BLOOCV variant, respectively. Benchmark MSEs are given in the inset table. The blue trace shows MSEs for CATE estimates obtained from the full inferential procedure of `DEE`, including the repair procedure in Algorithm 1 (which maps $\mathcal{L} \to \mathcal{U}$). The red trace shows the MSEs obtained when `DEE` does not apply Algorithm 1, and instead directly applies Gaussian process regression to local 2SLS estimates, computed using overlapping k-neighborhoods ($k = 200$), for each discovered local RD in $\mathcal{L}$. Note that we do not use weighted BLOOCV when using direct GP extrapolation from $\mathcal{L}$ due to the $\mathcal{O}(|\mathcal{L}|^5)$ runtime.

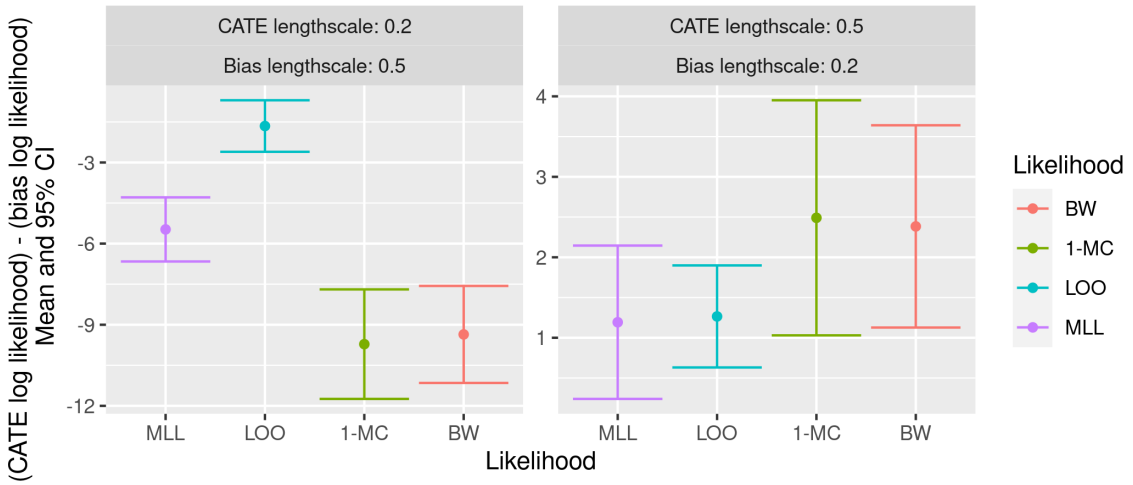

Figure 16: Difference between various log-likelihoods for Simulation 2, under the CATE model $\mathcal{M}_\tau$ and the bias model $\mathcal{M}_\beta$, in the off-diagonal regime where $\theta_\tau \neq \theta_\beta$. "MLL", "LOO", "1-MC", and "BW" refer to model averaging strategies where model weights are computed using the marginal likelihood, the LOOCV likelihood, the 1-MC BLOOCV variant, and the weighted BLOOCV variant, respectively. The sign of the difference determines which model is preferred by each approach.

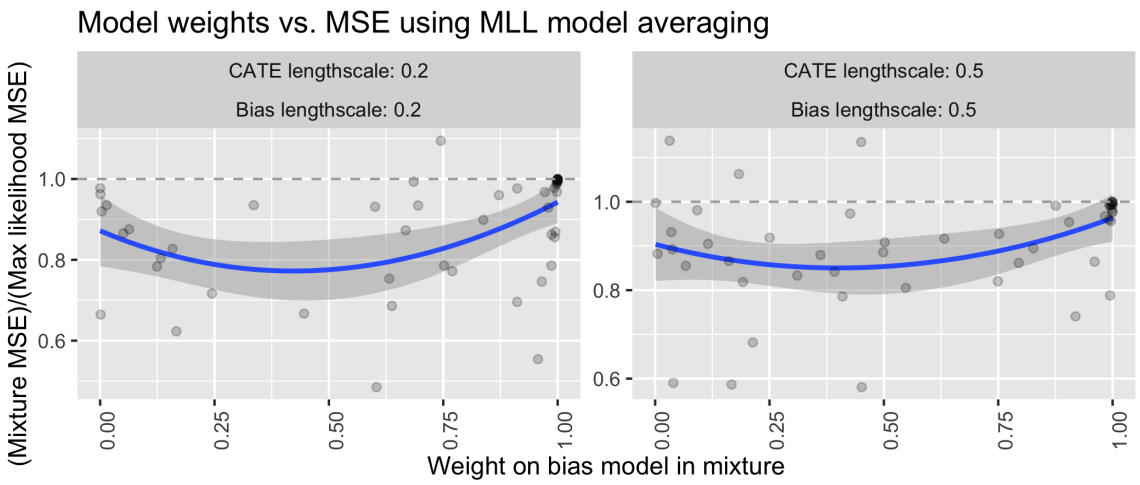

Figure 17: Relative mean squared error versus the $\mathcal{M}_\beta$ weight computed using marginal likelihood weighting (i.e., Bayesian model averaging) in the $\theta_\tau = \theta_\beta$ regime for Simulation 2. The y-axis gives the MSE ratio for the marginal likelihood weighted model average, versus the individual model ($\mathcal{M}_\tau$ or $\mathcal{M}_\beta$) with the maximum marginal likelihood. In general, model averaging reduces MSE (i.e., the local linear regression estimate, in blue, is smaller than 1, shown as a dashed grey reference line).

Appendix F. Application: Rural Roads and Economic Development (Asher and Novosad, 2020)

F.1 Rural Roads and Economic Development: Additional Results

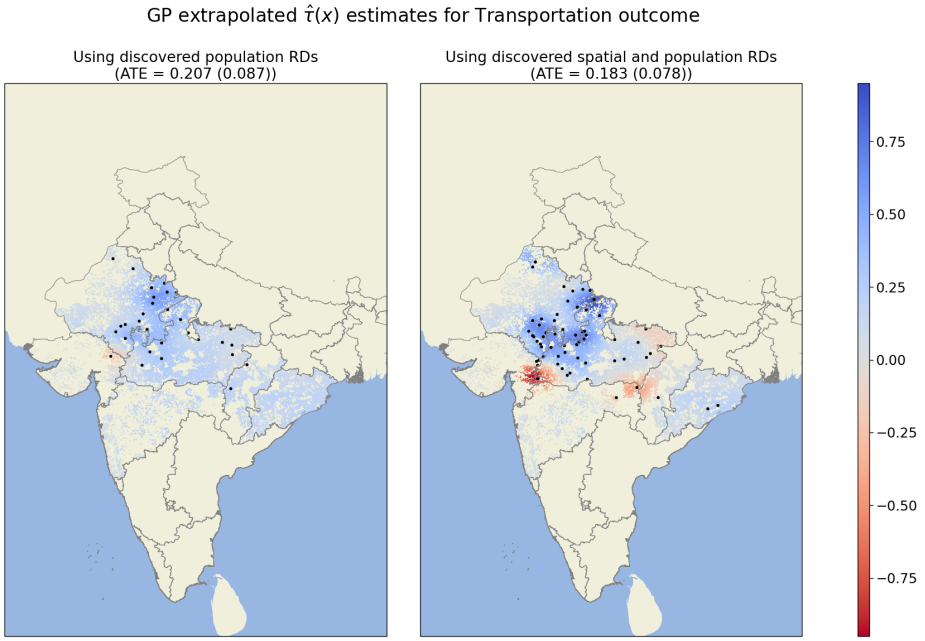

Figure 18: Applying GP extrapolation to estimate the spatially heterogeneous effect of road building on the availability of transportation services.

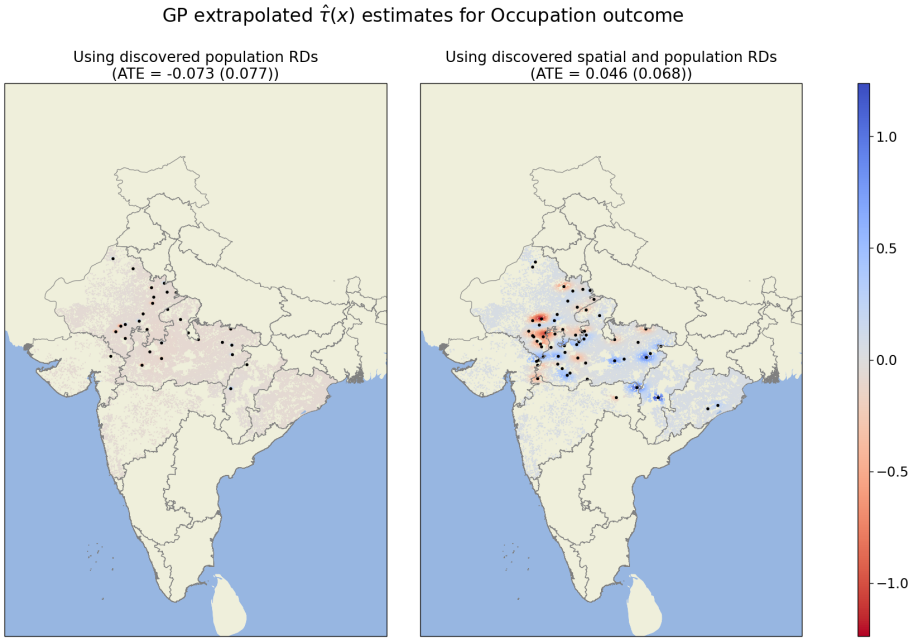

Figure 19: Applying GP extrapolation to estimate the spatially heterogeneous effect of road building on employment in agriculture.

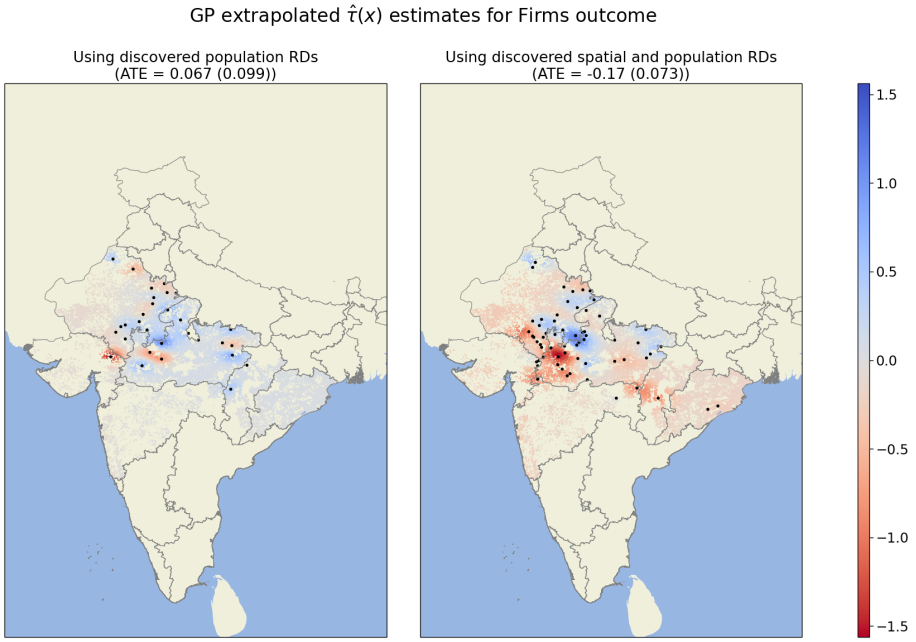

Figure 20: Applying GP extrapolation to estimate the spatially heterogeneous effect of road building on employment growth in village firms.

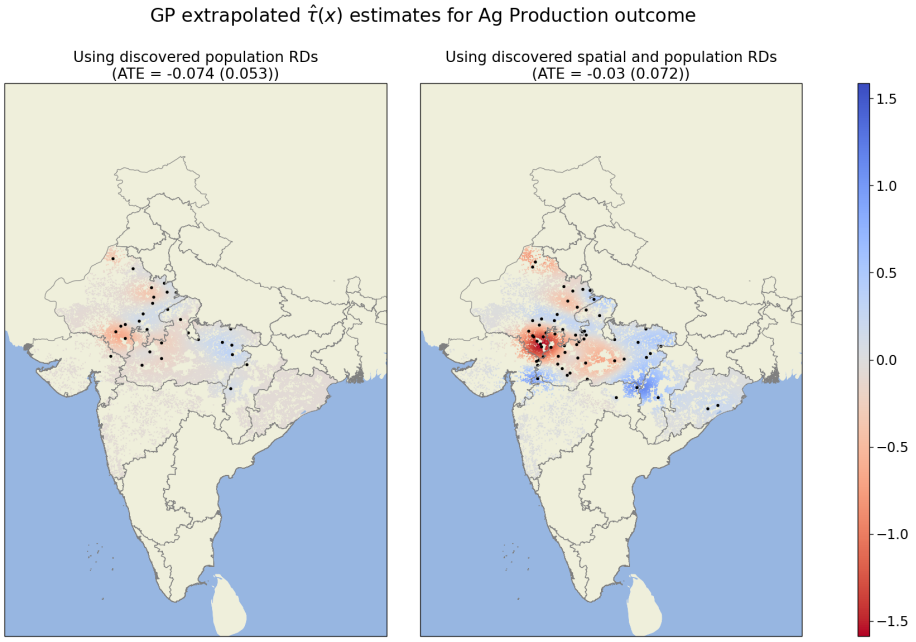

Figure 21: Applying GP extrapolation to estimate the spatially heterogeneous effect of road building on agriculture yields and investments.

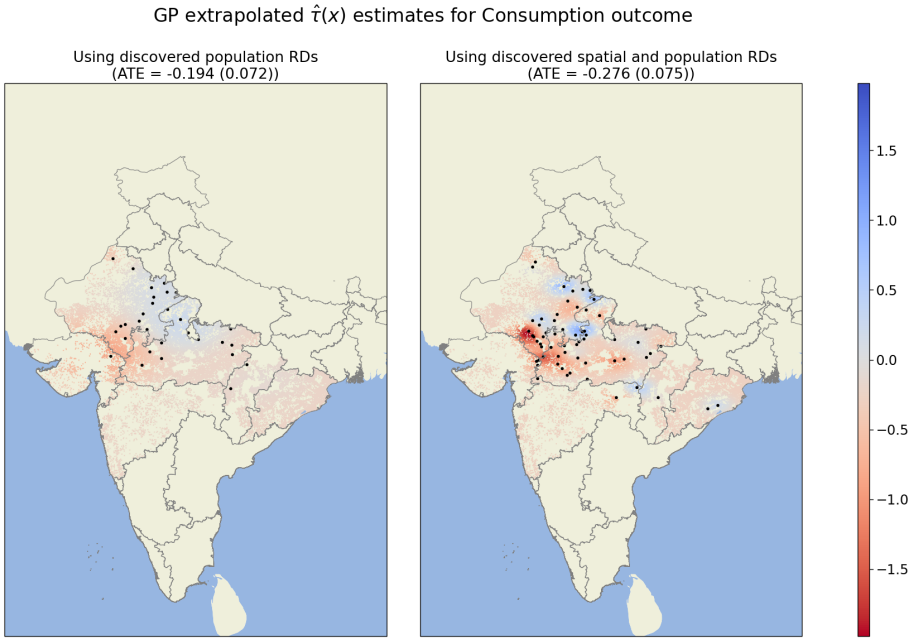

Figure 22: Applying GP extrapolation to estimate the spatially heterogeneous effect of road building on an asset/consumption index.

F.2 Rural Roads and Economic Development: Balance Test

Using LoRD^3, we identify a set of $M = 2{,}784$ candidate local regression discontinuities that are significant at $\alpha = 0.05$ using randomization testing (Herlands et al., 2018). However, instead of simply selecting all of these local RDs into $\mathcal{L}$ and proceeding with our inferential procedure, we first apply a covariate balance test to validate the discovered RDs. Specifically, for each baseline covariate B in the set of baseline covariates included in Table 1 of Asher and Novosad (2020), we test for continuity at the threshold using the regression specification

$$B \sim \alpha_Y + \beta_{B,Z} Z + \beta_{B,X_{RD}} X_{RD} + \beta^{(Z)}_{B,X_{RD}} Z X_{RD},$$

where, as previously, Z denotes the binary RD instrument and X_{RD} denote the projection of X onto the RD normal vector. This set includes the following 11 baseline covariates: having a primary school, having a medical center, electrification, distance from the nearest town, land irrigated, in land area, literate population, scheduled caste, land ownership, subsistence agriculture, and household income above INR 250. We test for continuity in B across the discontinuity using the p-value associated with the t-test of $\beta_{B,Z}$. Since we are testing 11 covariates for continuity, we adjust for multiple testing by rejecting discontinuities where $\min_B p < \alpha/11 = 0.05/11$. Note that after applying Algorithm 1 we once again apply this balance test to validate the RDs included in the final set $\mathcal{U}$.

Figure 23 shows the distribution of $\min_B p$ across the two types of discovered RDs, population and spatial. As expected, we reject very few of the (known, and previously validated) population RDs, but reject a large number of the spatial RDs based on baseline covariate imbalance.

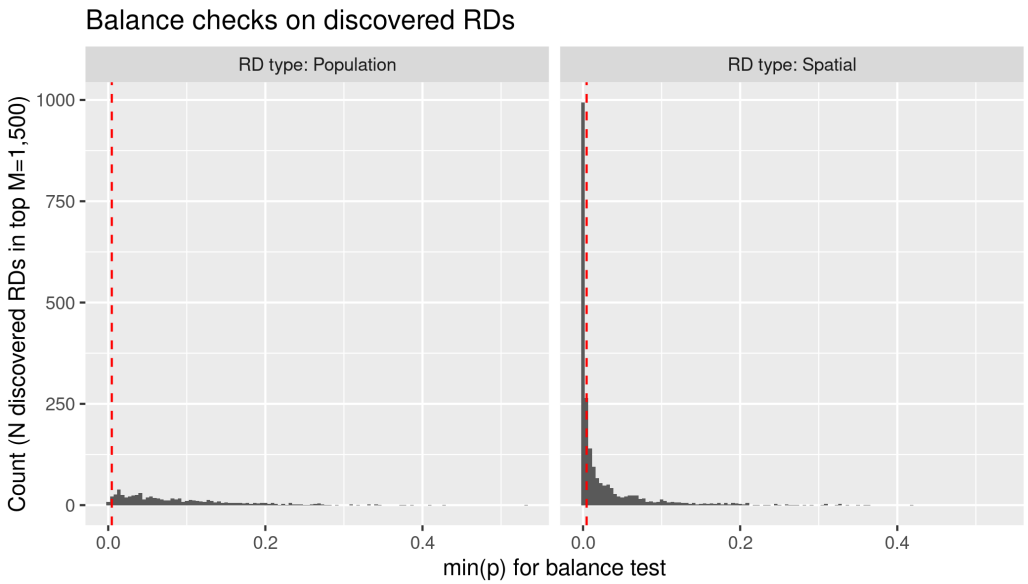

Figure 23: Testing balance in baseline covariates across the $M = 2{,}784$ discontinuities with significantly high LLR statistics using randomization testing. $\alpha/11$ is shown by the vertical red line.