Automated Local Regression Discontinuity Design Discovery

William Herlands
Carnegie Mellon University
Pittsburgh, PA 15213
herlands@cmu.edu

Edward McFowland III
University of Minnesota
Minneapolis, MN 55455
emcfowla@umn.edu

Andrew Gordon Wilson
Cornell University
Ithaca, NY 14850
andrew@cornell.edu

Daniel B. Neill
New York University
New York, NY 10003
daniel.neill@nyu.edu

ABSTRACT

Inferring causal relationships in observational data is crucial for understanding scientific and social processes. We develop the first statistical machine learning approach for automatically discovering regression discontinuity designs (RDDs), a quasi-experimental setup often used in econometrics. Our method identifies interpretable, localized RDDs in arbitrary dimensional data and can seamlessly compute treatment effects without expert supervision. By applying the technique to a variety of synthetic and real datasets, we demonstrate robust performance under adverse conditions including unobserved variables, substantial noise, and model misspecification.

KEYWORDS

Regression discontinuity; natural experiments; pattern detection

ACM Reference Format:
William Herlands, Edward McFowland III, Andrew Gordon Wilson, and Daniel B. Neill. 2018. Automated Local Regression Discontinuity Design Discovery. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining, August 19–23, 2018, London, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3219819.3219982

1 INTRODUCTION

Understanding causal mechanisms is critical for the social and laboratory sciences. While randomized control trials are the gold standard for identifying causal relationships, such experiments are often time consuming, costly, or ethically inappropriate. In order to exploit the plethora of observational data, econometricians often rely on "natural experiments," fortuitous circumstances of quasi-randomization that can be exploited for causal inference.

Regression discontinuity designs (RDDs) are such a technique. RDDs use sharp changes in treatment assignment for causal inference. For example, it is often difficult to assess the effect of academic interventions since treated students may systematically differ from other students. Yet, if a school intervenes on students who score below some threshold on a test, then students with scores just above or below the threshold are not systematically different and effectively receive random treatment [18]. That threshold induces an RDD that can be used to infer the effect of the intervention.

RDDs require fewer assumptions than most causal inference techniques and are arguably most similar to true randomized experiments [21]. However, identifying RDDs is a painstakingly manual process requiring human intuition and construction, and thus limited by human biases. Indeed, many papers reuse the same or analogous RDDs (e.g., discontinuities at geographic boundaries, or test score cutoffs for school admission) and most of these RDDs are one-dimensional, represented by a threshold value for a single variable. Finally, RDDs often rely on the human "eye" to verify their validity. The "tinkering" that is often done in practice implies that RDDs discovered by humans are subject to multiple testing issues.

To aid in discovering RDDs, we use statistical machine learning techniques to create the first general methodology to discover, quantify, and validate RDDs in data. Our approach can discover new RDDs across arbitrarily high dimensional spaces, enabling us to use RDDs that humans would not be able to identify otherwise. Yet these high dimensional RDDs are still interpretable, and we provide a simple mechanism for ranking how (observed) variables influence the discovered discontinuities. We derive two log likelihood ratio statistics to search for RDDs in potentially heteroskedastic data with either real-valued or binary treatments. Additionally, the technique can seamlessly handle both real-valued and categorical covariates. Finally, we present an integrated validation procedure ensuring rigorous statistical and econometric validity.

We evaluate our approach on synthetic and real data. Using synthetic data we demonstrate robust performance to out of sample discontinuities and model misspecification. For real data we consider three educational and health care settings previously studied in the econometric literature. Our approach can identify the RDDs in these data even with the injection of substantial additional noise.

While this is the first paper we know of that discovers RDDs in general data, Card et al. [7] search for race-based "tipping" points in housing markets using an RDD design. They employ two search methods specific to the problem formulation: one inspired by the shape of curves derived from their data, and one that draws on structural break literature in time series [10]. Beyond RDDs, there is increased interest in integrating econometric and machine learning techniques [4, 24]. For example, deep learning and non-parametric Bayesian methods have been used to predict counterfactuals and compute individualized treatment effects [11, 14, 19]. Additionally,

KDD '18, August 19–23, 2018, London, United Kingdom

ACM ISBN 978-1-4503-5552-0/18/08…$15.00
https://doi.org/10.1145/3219819.3219982

novel approaches have been developed for identifying heterogeneous treatment effects [5, 16]. Within the context of online recommendation systems, Sharma [31] and Sharma et al. [32] develop mechanisms of searching for certain natural experiments.

1.1 Outline

The remainder of the paper proceeds as follows. §2 provides a brief overview of RDDs including their causal assumptions. §3 introduces our local search for RDDs including the search statistics used for the Normal (§3.1) and Bernoulli (§3.2) observation models, neighborhood definitions (§3.3), discontinuity validation (§3.4), and treatment effect estimation (§3.5). §4 discusses the synthetic and real data experiments. §5 ends with concluding remarks.

2 REGRESSION DISCONTINUITY DESIGNS

We provide practical background on RDDs for a computer science audience. There exist excellent papers for details on assumptions, inference, convergence, and model variations [9, 17, 33].

Throughout this paper we consider data, (x, T, y), where $x = \{x_1, ..., x_n\}$, are inputs that can include both categorical and real-valued $x_i \in \mathbb{R}^d$ variables, $T = \{T_1, ..., T_n\}$ is a treatment variable that could either be binary, $T_i \in \{0, 1\}$, or real-valued, $T_i \in \mathbb{R}$, and $y = \{y_1, ..., y_n\}, y_i \in \mathbb{R}$, is an outcome variable. Both x and T are known *a priori* not to be affected by y. Additionally, we consider "forcing variables," z, which are a subset of the real-valued dimensions of x. Typically, the dimensions of $z \in x$ must be specified and validated by the user, but our algorithm does this automatically.

In the most straightforward RDDs, called "sharp RDDs," there is a one-dimensional forcing vector, z, and a cutoff value, c, such that before the cutoff value treatment is never assigned, $E[T|z < c] = 0$, and after the cutoff treatment is always assigned, $E[T|z > c] = 1$. Thus there is a sharp RDD at $z = c$ since at that point T jumps discontinuously from $T = 0$ to $T = 1$. As long as x does not also change discontinuously at $z = c$, there is no reason to believe that the data on either side of the discontinuity are systematically different. Thus, conceptually, at the local area around the discontinuity we can consider T to be randomly assigned. Notice that the RDD is a function of x, z, and T but not y. Indeed, an RDD allows us to investigate the effect of T on multiple different outputs, y.

RDDs appear in real-world settings where thresholds are used to assign treatment. For example, academic punishments given to students whose GPA drops below a specific value [22], or health insurance that covers children until they reach a certain age [2].

For this paper we concentrate on "fuzzy RDDs" which generalize the sharp RDD. Fuzzy RDDs exist where T is partially determined by the discontinuity, i.e., where $P(T = 1)$ jumps discontinuously at $z = c$. The special case where that jump is from $P(T = 1) = 0$ to $P(T = 1) = 1$ constitutes a sharp RDD [17]. Given a fuzzy RDD, the treatment effect, τ, with respect to y, is,

$$\tau = \frac{\lim_{\epsilon \to -0} E[y|z = c + \epsilon] - \lim_{\epsilon \to +0} E[y|z = c + \epsilon]}{\lim_{\epsilon \to -0} E[T|z = c + \epsilon] - \lim_{\epsilon \to +0} E[T|z = c + \epsilon]}. \quad (1)$$

The limits in Eq. (1) indicate that although T is effectively random at $z = c$, farther away from the discontinuity T is not expected to be randomly assigned. That said, τ can be considered a weighted average treatment effect across the entire data, where the weights are ex-ante probabilities that a point is in the vicinity of $z = c$ [21].

The fuzzy RDD assumes the following conditions for identification [9] (the first two are also required for the sharp RDD):

- *Imprecise control*: the value of z cannot be precisely controlled to fall at $z = c \pm \epsilon$. If such control did exist, those individuals manipulating z to be just above or just below c are likely to be systematically different than individuals who do not manipulate z, thus invalidating the design.
- *Excludability*: x crossing $z = c$ cannot affect y except through affecting the probability distribution of T.
- *Monotonicity*: x crossing $z = c$ cannot simultaneously cause some data to increase T and other data to decrease T.

These assumptions are relatively light and the first is even testable (see §3.4). Imprecise control replaces the ignorability or unconfoundedness assumptions that are necessary in many causal models. And unlike instrumental variables, RDDs do not assume anything about exogeneity [21]. Thus RDDs are quite suitable for automated discovery since they do not require the onerous, untestable, and often unbelievable assumptions made by other causal inference methods.

3 METHOD

The essential element of an RDD, which our approach aims to discover automatically from data, is the discontinuity, or "unexpected jump," in T. Given a model, $T_i = f(x_i) + \epsilon_i$, this constitutes a special type of local anomaly where $f(x)$ substantially deviates from T both before and after the discontinuity. See Fig. 1 for a 1-D example where $f(x)$ approximates the data well except for the two regions of deviation on either side of the discontinuity. Note that the deviations are of opposite sign and may be of different magnitudes.

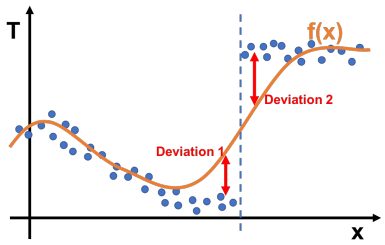

Figure 1: Illustration of a one-dimensional RDD (dashed line). Blue dots are treatment T_i; orange line is $f(x_i)$.

Traditional anomaly detection, such as one-class SVMs [29], focus on identifying individual outliers. Yet an RDD is fundamentally a pattern of multiple data points. Thus we employ anomalous *pattern* detection to search for RDDs. We frame the search as a log likelihood ratio (LLR) comparison between the likelihood of a null model that assumes no RDD exists, and the likelihood of an alternative model that assumes an RDD exists. We locally search for circumscribed neighborhoods that contain a discontinuity. Although any one neighborhood does not necessarily capture the entire discontinuity, it uses local data from around the discontinuity which can provide greater insight. The discovered discontinuities from multiple local neighborhoods can be combined to more precisely measure the treatment effect. Thus our approach is named "**Lo**cal **R**egression **D**iscontinuity **D**esign **D**iscovery" (LoRD3). In a valid RDD, z must be be real-valued since sharp differences are expected to occur between data points with different values of a categorical

variable. Thus, given data (x, T, y), we let all real-valued dimensions of x be forcing variables, z. LoRD3 searches for RDDs as follows[1]:

(1) Model T with smooth model, $f(x)$, such that $T_i = f(x_i) + \epsilon_i$.
(2) Compute the estimated value of $\hat{T}$ using the learned model.
(3) For each neighborhood size, $k = 1, ..., K$:
  (a) For each of the n data points, consider its k-sized neighborhood, $s_{i,k}$, defined by z (see §3.3).
    (i) Compute the likelihood of a null model which assumes that $s_{i,k}$ does not contain an RDD: $L_0(s_{i,k})$.
    (ii) Repeatedly bisect the neighborhood into two mutually exclusive partitions, assigning each point in the neighborhood to one of these two groups (see §3.3). We denote group assignment by $g_{k,j}$. For each grouping, compute the likelihood of an alternative model which assumes that $s_{i,k}$ contains an RDD with each group denoting one side of the discontinuity: $L_1(s_{i,k}, g_{k,j})$.
    (iii) Compute the maximum log likelihood ratio (LLR) over all partitions for that neighborhood,

$$LLR(s_{i,k}) = \max_j LLR(s_{i,k}, g_{k,j}) = \max_j \log \frac{L_1(s_{i,k}, g_{k,j})}{L_0(s_{i,k})}. \quad (2)$$

(4) Test each of the neighborhoods for statistical significance and econometric validity, controlling for multiple hypothesis testing (see §3.4). For each "validated" neighborhood $s_{i,k}$ that passes these tests, record the corresponding $g_{k,j}$.
(5) Estimate the τ using validated neighborhoods (see §3.5).

Notice that in step (a) the local neighborhoods are defined over the potentially multidimensional z. While most research using RDDs considers one-dimensional forcing variables z, even papers that consider multiple dimensional z [26, 34] require human identification and are limited in practice to low dimensions. LoRD3 seamlessly considers z of arbitrary dimension, allowing it to discover more diverse and nuanced RDDs than previously studied.

In §3.1 and §3.2 we detail two observation models and LLR statistics for real-valued treatments and binary treatments respectively.

3.1 Normal residual observation model

Given a model, $T_i = f(x_i) + \epsilon_i$, we would expect $f(x)$ to substantially and systematically deviate from a jump discontinuity in T. Specifically, near the discontinuity $f(x)$ should underestimate the true value of T on one side and overestimate T on the other side. We search for such a pattern using the LLR statistic below.

In principle we can use any regression approach for $f(x)$. Yet the appropriate choice requires $f(x)$ to be expressive enough to faithfully model the data, yet not substantially overfit to a potential discontinuity. For example, deep neural networks are untenable as they can model discrete jumps in data. In order to elucidate LoRD3, we consider polynomial models, $f(x) = \sum_{r=0:R} \gamma_r x^R$, which can be made increasingly expressive by increasing the polynomial order.

Given data (x, T), a neighborhood s, and a bisection g of the data points in s into "group 0" ($g_i = 0$) and "group 1" ($g_i = 1$), we consider the residuals, $r_i = T_i - f(x_i)$. The null model H_0 assumes that no discontinuity exists in s. We define the null model as r_i Normally distributed around a single offset parameter, β_0, which accounts for any bias in $f(x)$ over both groups. The alternative model H_1 states that a discontinuity exists between the two groups, and assumes that the r_i are Normally distributed around two distinct mean shifts, one for each group. For maximal applicability, we consider unconstrained heteroskedastic noise, $\epsilon_i \sim N(0, \sigma_i)$:

$$\begin{aligned} H_0 &: r_i \sim N\Big(\beta_0, \sigma_i\Big), \forall i \in s \\ H_1 &: r_i \sim N\Big((1-g_i)\beta_{g_0} + g_i\beta_{g_1}, \sigma_i\Big), \forall i \in s. \end{aligned} \quad (3)$$

Letting the alternative mean, $\mu_i = (1-g_i)\beta_{g_0} + g_i\beta_{g_1}$, for notational simplicity, we can compute the LLR,

$$\begin{aligned} LLR(s, g) &= \log \frac{Lik(H_1(s, g))}{Lik(H_0(s))} \\ &= \log \Big(\prod_{i \in s} P\big(r_i | N(\mu_i, \sigma_i)\big) \Big/ \prod_{i \in s} P\big(r_i | N(\beta_0, \sigma_i)\big) \Big) \\ &= \sum_{i \in s} (2r_i(\mu_i - \beta_0) - \mu_i^2 + \beta_0^2)/(2\sigma_i^2). \end{aligned} \quad (4)$$

For unrestricted heteroskedastic models we cannot directly compute σ_i. Instead, we assume that within a local area around each neighborhood the noise is homoskedastic. Thus we compute σ_i as the empirical variance in the k-neighborhood around each point.

We use the MLE values of β_0, β_{g_0}, and β_{g_1} from their respective heteroskedastic Normal models. Thus for each neighborhood,

$$\begin{aligned} \beta_0^* &= \Big(\sum_{i \in s} \frac{r_i}{\sigma_i^2} \Big) \Big/ \Big(\sum_{i \in s} \frac{1}{\sigma_i^2} \Big) \\ \beta_{g_0}^* &= \Big(\sum_{i \in s \cap (1-g)} \frac{r_i}{\sigma_i^2} \Big) \Big/ \Big(\sum_{i \in s \cap (1-g)} \frac{1}{\sigma_i^2} \Big) \\ \beta_{g_1}^* &= \Big(\sum_{i \in s \cap g} \frac{r_i}{\sigma_i^2} \Big) \Big/ \Big(\sum_{i \in s \cap g} \frac{1}{\sigma_i^2} \Big). \end{aligned} \quad (5)$$

3.2 Bernoulli log-odds observation model

For binary T, the Normal model is inappropriate since the residual between a binary variable and $f(x)$ is rarely Gaussian. Instead, we model T as a Bernoulli distributed random variable and search for discontinuities in the odds ratio [35].

Given a model for probability of treatment, $T_i \sim Bernoulli(p(x_i))$, we would expect $p(x)$ to systematically under- and over-estimate the true data around a jump discontinuity in T. We search for such a pattern using the LLR statistic below. We use a base model of a Logistic regression with polynomial functions, $p(x) = \text{Logit}(\sum_{r=0:R} \gamma_r x^r)$, to model the probability of a data point having $T_i = 1$.

Given data (x, T, s, g) as in §3.1, we consider the odds ratio of $T_i = 1$. The null model assumes that no discontinuity exists in s. We define the null model as a constant multiplicative scaled odds ratio to account for any bias in $p(x)$ over both groups,

$$H_0 : odds(T_i) = \beta_0 \frac{p(x_i)}{1 - p(x_i)}, \forall i \in s. \quad (6)$$

The alternative model assumes that a discontinuity exists between the two groups. Continuing to let $\mu_i = (1 - g_i)\beta_{g_0} + g_i\beta_{g_1}$, we define the alternative model as an odds ratio with two distinct multiplicative scales, one for each region,

$$H_1 : odds(T_i) = \mu_i \frac{p(x_i)}{1 - p(x_i)}, \forall i \in s. \quad (7)$$

[1] Code and data are available at https://gitlab.com/herlands/LORD3

These correspond to the null and alternative models,

$$H_0 : T_i \sim Bernoulli\Big(\frac{\beta_0 p(x_i)}{1-p(x_i)+\beta_0 p(x_i)}\Big), \forall i \in s$$
$$H_1 : T_i \sim Bernoulli\Big(\frac{\mu_i p(x_i)}{1-p(x_i)+\mu_i p(x_i)}\Big), \forall i \in s, \quad (8)$$

with which we can compute the LLR,

$$\begin{aligned} LLR(s,g) &= \log \frac{\prod_{i\in s} P\Big(T_i | Bernoulli\big(\frac{\mu_i p(x_i)}{1-p(x_i)+\mu_i p(x_i)}\big)\Big)}{\prod_{i\in s} P\Big(T_i | Bernoulli\big(\frac{\beta_0 p(x_i)}{1-p(x_i)+\beta_0 p(x_i)}\big)\Big)} \\ &= \sum_{i\in s} T_i \log(\mu_i/\beta_0) + \log(1-p(x_i)+\beta_0 p(x_i)) \\ &\qquad - \log(1-p(x_i)+\mu_i p(x_i)). \end{aligned} \quad (9)$$

Unlike in the Normal case, there is no closed form solution for the MLE of β_0, β_{g_0}, or β_{g_1}. Instead, we solve for their values using a binary search. Eq. (10) provides the derivative of the log likelihood with respect to β_0. Note that the sum is taken over all points in the neighborhood ($i \in s$). Similar results hold for β_{g_0}, summing over group 0 ($i \in s \cap (1-g)$), and β_{g_1}, summing over group 1 ($i \in s \cap g$).

$$\frac{\delta LL(s,g)}{\delta \beta_0} = \sum_{i\in s}\Big(\frac{T_i}{\beta_0} - \frac{p(x_i)}{1-p(x_i)+\beta_0 p(x_i)}\Big) \quad (10)$$

We can then solve $\beta_0 \frac{\delta LL(s,g)}{\delta \beta_0} = 0$ by an efficient binary search, noting that this quantity decreases monotonically with $\beta_0 > 0$.

3.3 Neighborhood definition and bisection

Using only $z \in x$ to measure distance, the local neighborhood of a point includes itself and its $k-1$ nearest neighboring points. Since we are interested in generalizing to arbitrary dimensional RDDs we first compute the vector, $v_{s,i}$, between the center point of neighborhood s and each point $i \in s$. Then we bisect the neighborhood with $k-1$ hyperplanes, each of which passes through the center point, and is orthogonal to a $v_{s,i}$. Within each neighborhood, LoRD3 selects the bisection that maximizes the LLR defined above, testing the alternative hypothesis that there is an RDD for that neighborhood and bisection against the null hypothesis of no RDD.

3.4 Validate RDD neighborhoods

LoRD3 produces $O(n)$ neighborhoods - one centered at each data point - each with a corresponding bisection and $LLR(s)$. We can automatically assess which neighborhoods are statistically and econometrically valid using three techniques:

Randomization testing. As is typical when searching with LLR statistics [20, 25], we use randomization to adjust for multiple testing and determine whether discontinuities are significant at level α. Specifically we use the following procedure:

(1) Draw data $T^{(q)}$ from the null model Q times at the same covariates, x, as the true data.
   - In the Normal observation model, since $T_i = f(x_i) + \epsilon_i$ this corresponds to sampling the noise Q times.
   - In the Bernoulli observation model, each T_i can be drawn directly from the H_0 Bernoulli distribution in Eq. (8).

(2) Run LoRD3 on each $(x, T^{(q)})$. For each run save the value $LLR^{(q)} = \max_s LLR(s)$.

(3) Compute an α threshold using the $1-\alpha$ quantile of the $LLR^{(q)}$ values. Any original neighborhoods s with $LLR(s)$ above this threshold are considered statistically significant.

For the unconstrained heteroskedastic model, we estimate each point's σ_i from the variance of data within the k-local neighborhood of that point, as in §3.1 above. Since LoRD3 evaluates $O(kn)$ possible neighborhood bisections, randomization is critical to address multiple testing issues. Since we use the maximum score over s for both original and replica data, this procedure provides an exact test for the highest-scoring neighborhood and a conservative test for secondary neighborhoods.

Density discontinuity. As discussed in §2, RDDs assume that precise manipulation of T is not possible. A violation of this assumption could be reflected in a discontinuous density of z since data might "bunch" in z around the discontinuity to affect treatment status. McCrary [23] provides a commonly used procedure to test for such discontinuities in z. Since this test is limited to one dimension, we map our data to the vector orthogonal to the hyperplane that bisects the two groups in each neighborhood and apply the test on this one-dimensional data [8]. For each s, if the split selected by LoRD3 rejects the null we invalidate this s.

Placebo Testing. In RDDs, placebo testing ensures that the discontinuity in T cannot be explained by a corresponding discontinuity or imbalance in x. While the forcing variables z are continuous within each neighborhood s, any $x \setminus z$, such as categorical variables, may still present issues. In order to be conservative, we run placebo tests on every dimension in x. We iteratively select one observational variable, $x^{(d)}$, and considering data $(T, x \setminus x^{(d)})$, we estimate τ with $x^{(d)}$ as the output (see §3.5 for how to compute $\hat{\tau}$). We ensure that this $\hat{\tau}$ is statistically indistinguishable from zero.

3.5 Estimating the treatment effect τ

Given a validated set of neighborhood discontinuities from LoRD3, practitioners may wish to further investigate the detected regions using domain expertise. Yet, it is also possible to directly use the neighborhood results from LoRD3 to estimate the treatment effect τ of treatment T on some real-valued output y. Below we describe three automated approaches for computing the estimate $\hat{\tau}$ given the results from LoRD3. If LoRD3 detects more than one validated neighborhood s, we compute $\hat{\tau}_s$ for each s and average them for the final estimation. Pooling the regions themselves, such as Bertanha [6] suggests for RDDs with multiple thresholds, is not possible in this case since there is no defined orientation of the two groups.

2SLS estimator. A two-stage least squares estimation of τ first instruments $\hat{T}$ with a validated RDD neighborhood and then regresses $\hat{T}$ on y [3]. Given the data in neighborhood s, and indicators g_i for which group each data point is in, we first estimate,

$$T_i = \nu g_i + f(x_i) + \epsilon_i^{(T)}. \quad (11)$$

Then we use the predicted $\hat{T}$ to regress (where λ is a learned vector),

$$y_i = \hat{\tau}\hat{T}_i + \lambda x_i + \epsilon_i^{(y)}. \quad (12)$$

Non-parametric estimator. Given a neighborhood and bisection, a non-parametric estimator for τ draws on Eq. (1). Assuming that

the neighborhood is sufficiently small to approximate the limit, we use the empirical expectations over y and T to compute,

$$\hat{\tau} = \frac{E[y|g=1] - E[y|g=0]}{E[T|g=1] - E[T|g=0]}. \tag{13}$$

Group instrument. While the 2SLS works generally for RDDs, using LoRD3 we can leverage information about μ to instrument T in each group. For the Normal model we instrument,

$$\hat{T}_i = T_i - \mu_i, \tag{14}$$

while for the Bernoulli model we can instrument $\hat{T}$ as,

$$\hat{T}_i = \frac{\mu_i p(x_i)}{1 - p(x_i) + \mu_i p(x_i)}. \tag{15}$$

Then we can run the second stage regression from Eq. (12).

3.6 Forcing variable influence

When humans identify an RDD it is clear which variables are responsible for the discontinuity. Since we consider potentially high dimensional z it is useful to identify which z variable(s) are most responsible for the RDD. Given a neighborhood, consider v_s, the vector orthogonal to the bisecting hyperplane. After normalizing the individual components of v_s to lie in $[0, 1]$, those components indicate which dimensions of z most influence the discontinuity. For multiple neighborhoods, we average multiple normalized $v_1, ..., v_S$.

3.7 Evaluating discontinuities

Given a known discontinuity in synthetic or real data, we can evaluate how well a neighborhood s and bisection g chosen by LoRD3 correspond to the true discontinuity. Accuracy and precision are not appropriate metrics since there is no defined orientation of the two groups in a neighborhood. Instead, letting $d \in \{0, 1\}$ define the space on either side of the true discontinuity, we compute the information gain (IG) of a k-sized neighborhood,

$$\begin{aligned} IG = k\, H\left(\frac{|s \cap d|}{k}\right) - |s \cap (1-g)|\, H\left(\frac{|s \cap (1-g) \cap d|}{|s \cap (1-g)|}\right) \\ - |s \cap g|\, H\left(\frac{|s \cap g \cap d|}{|s \cap g|}\right), \end{aligned} \tag{16}$$

where $H(p)$ is the entropy, $H(p) = -p\log(p) - (1-p)\log(1-p)$. We then normalize the IG to lie in $[0, 1]$ by dividing by the optimal IG for a neighborhood of size k with a bisection of points into two equally sized groups, $k * H(\frac{1}{2})$. This metric is optimized when the neighborhood bisection overlaps fully with the true discontinuity and when the bisection equally divides the neighborhood points.

We provide measures of the normalized information gain (NIG) for all experiments in §4. Higher NIG is better since it indicates that a neighborhood bisection provides information about the true discontinuity. Lower NIG indicates that either the bisection is misaligned or the neighborhood does not intersect the discontinuity.

3.8 Practical considerations

As a pre-processing step before running LoRD3, we remove any data points with missing values and normalize each real-valued dimension x_j to have zero mean and unit variance. For datasets with categorical variables, we include these in x but not in z. Thus we do not consider heterogeneous treatment effects [15]. By default, all real-valued x are in z, though users may exclude variables based on domain knowledge. Finally, we note that approaches which analyze a known RDD may fit two background functions - one to each side of the discontinuity [21]. As detailed in §3, LoRD3 assumes a single background function in order to enable an efficient search. Additionally, we do not consider nonparametric $f(x)$ models such as local linear regression [17], but these could be easily incorporated.

4 EXPERIMENTS

In order to demonstrate the power and flexibility of LoRD3, we apply the technique to a wide variety of synthetic and real data. RDDs are injected into the synthetic data, while for the real data we consider previously studied settings where known discontinuities exist. Note that the known RDD locations are used for evaluation purposes only, and are not provided to LoRD3. We inject additional noise into the real data to stress-test the search technique and evaluate its performance in the face of increasingly subtle discontinuities. For one-dimensional RDDs, we provide a comparison to existing changepoint detection methods in the literature.

4.1 Generating synthetic data

For synthetic experiments, we draw observed covariates, $x \in \mathbb{R}^d$, and unobserved covariates, $u \in \mathbb{R}^1$, through independent draws from a Uniform distribution, such that for $i = 1 \ldots n,\ j = 1 \ldots d$,

$$x_{i,j} \sim \text{Uniform}(0, 1), \quad u_i \sim \text{Uniform}(0, 1). \tag{17}$$

We induce a discontinuity by randomly selecting a boundary, $b_j \sim$ Uniform$(0, 1)$ and defining an indicator,

$$D_i = \bigcup_{j=1}^{d} x_{i,j} > b_j. \tag{18}$$

Thus the discontinuous region is a d-dimensional cube and out-of-class for the hyperplanes LoRD3 uses to bisect each neighborhood. Throughout all experiments we consider heteroskedastic noise,

$$\epsilon_i^{(T)}, \epsilon_i^{(p)}, \epsilon_i^{(y)} \sim N\left(0, \frac{1}{d}\sum_j x_{i,j}\right). \tag{19}$$

Real-valued treatment indicators T are generated by selecting the magnitude of the discontinuity, $\zeta \in \mathbb{R}$, and drawing,

$$\begin{aligned} \gamma_T &\sim N(0, I_d) \\ \mu_i &= I(x_i \in D)\frac{\zeta}{2} - I(x_i \notin D)\frac{\zeta}{2} \\ T_i &= x_i\gamma_T + \mu_i + \epsilon_i^{(T)} + u_i. \end{aligned} \tag{20}$$

Binary treatment indicators T are generated by selecting the magnitude of the discontinuity, $\zeta > 0$, and drawing,

$$\begin{aligned} \gamma_p &\sim N(0, I_d) \\ \mu_i &= I(x_i \in D)\exp(\zeta/2) + I(x_i \notin D)\exp(-\zeta/2) \\ p_i &= \text{Logit}(x_i\gamma_p + \mu_i + \epsilon_i^{(p)} + u_i) \\ T_i &\sim \text{Bernoulli}(p_i). \end{aligned} \tag{21}$$

Outputs $y_i \in \mathbb{R}$ are generated by selecting $\tau \in \mathbb{R}$ and drawing,

$$\begin{aligned} \gamma_y &\sim N(0, I_d) \\ y_i &= x_i\gamma_y + T_i\tau + \epsilon_i^{(y)} + u_i. \end{aligned} \tag{22}$$

4.2 Synthetic real-valued treatment results

We generate real-valued T with $x \in \mathbb{R}^2$ and $\tau = 5$. To demonstrate how LoRD3 performs under different signal levels of discontinuity, we vary $\zeta \in [0, 2.5]$. For each ζ value we generate 50 experiments with 1000 data points. For LoRD3 we let $k = 50$, $z = x$, and consider the top scoring neighborhood for evaluation. Base $f(x)$ models are order $r = 1, 2, 4$ polynomials to demonstrate results from both correctly and incorrectly specified models. Randomization testing is performed to determine an $\alpha = .05$ level for significance for each experiment. Finally, throughout the synthetic and real data experiments we have verified the placebo tests, as detailed in §3.4.

Fig. 2 provides an example of an experiment with $\zeta = 3$. The axes are the dimensions of x. The left plot depicts T as colored circles and the square discontinuity is observable in the upper right. The center plot depicts $LLR(s)$ centered on each data point. The outline of the discontinuity has relatively high $LLR(s)$ indicating that LoRD3 has correctly identified neighborhoods along the boundary of the discontinuity. The right plot highlights the two groups from the neighborhood bisection with highest $LLR(s)$.

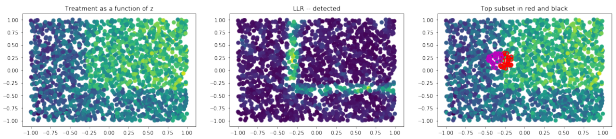

Figure 2: Left plot: synthetic $T \in \mathbb{R}$ as a function of x. Center plot: $LLR(s)$ for each neighborhood using Normal model. Right plot: neighborhood bisection with highest $LLR(s)$.

We present results for NIG and power in Fig. 3. LoRD3 performance improves as ζ increases since higher ζ induce a larger magnitude discontinuity. While the more complex specifications of $f(x)$ have slightly decreased performance due to overfitting, both NIG and power for all models are quite similar, demonstrating that the approach is robust to model misspecification.

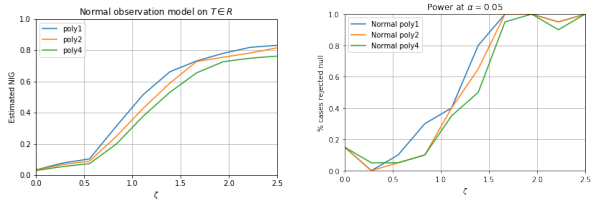

Figure 3: Left: NIG of top neighborhood for $T \in \mathbb{R}$. The x-axis indicates ζ. Right: power to reject the null at $\alpha = 0.05$.

Estimates of the treatment effect $\hat{\tau}$ are plotted in Fig. 4. Due to the data generating process of y in Eq. (22), at low ζ where there is little to no discontinuity, LoRD3 tends to overestimate the true τ. However, all $f(x)$ model specifications (polynomials of degree 1,2,4) yield $\hat{\tau}$ that converge towards the true τ at larger ζ. While the non-parametric and 2SLS approaches converge more slowly, they tend to be more robust to model misspecification.

Varying dimension. Letting $\zeta = 2$ and holding z fixed at two dimensions, we vary the number of covariates from 2 to 20. We apply LoRD3 with the three $f(x)$ models as above and plot the

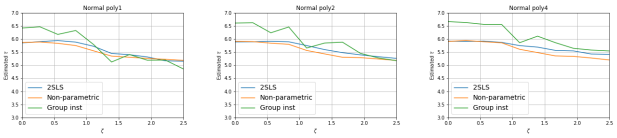

Figure 4: LoRD3 Normal model estimated $\hat{\tau}$ on $T \in \mathbb{R}$. Each plot represents a different $f(x)$ specification. True $\tau = 5$.

resulting NIG in the left panel of Fig. 5. Next we hold x fixed at 10 dimensions and vary the number of dimensions in z from 1 to 10, plotting the results in the right panel of Fig. 5. These results indicate that given the same amount of data LoRD3 performance is robust to large numbers of covariates but reduces in performance over larger spaces of forcing variables.

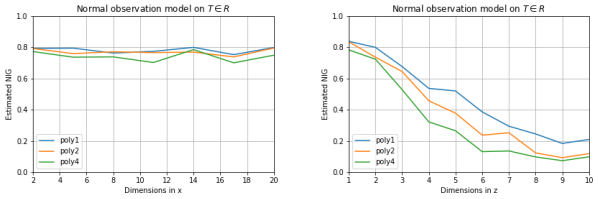

Figure 5: NIG of LoRD3 Normal model for $T \in \mathbb{R}$ with varying the dimensions of x and z in left and right plots, respectively.

4.3 Synthetic binary treatment results

We generate equivalent synthetic tests for $T \in \{0, 1\}$. For each experiment we run LoRD3 with both Normal and Bernoulli observation models. We use $p(x)$ of order $r = 1, 2, 4$ polynomials to demonstrate results from correctly and incorrectly specified models.

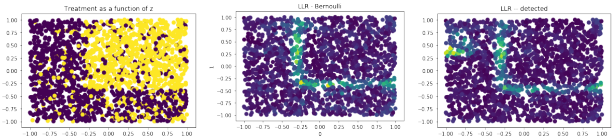

Figure 6: Left shows $T \in \{0, 1\}$ as a function of x, center shows $LLR(s)$ of Bernoulli model, $LLR(s)$ of Normal model.

Fig. 6 provides an example of an experiment with $\zeta = 4$ and $LLR(s)$ using both the Bernoulli and Normal models. While both models discover neighborhoods with high LLR around the discontinuity boundary, the Normal model detects spuriously high LLR elsewhere in the space. The advantage of the Bernoulli model for binary T is also seen through the NIG results in Fig. 7 where all $p(x)$ specifications using the Bernoulli model outperform the Normal model. We plot $\hat{\tau}$ from the Bernoulli model in Fig. 8 where all $p(x)$ specifications have $\hat{\tau}$ that converge to the true $\tau = 5$ at larger ζ.

4.4 Comparison to changepoint detection

In one dimension, RDD discovery is similar to changepoint detection, where the objective is to identify points between regimes with persistent changes in mean or covariance structure. We consider competitive changepoint methods that utilize Binary Segmentation

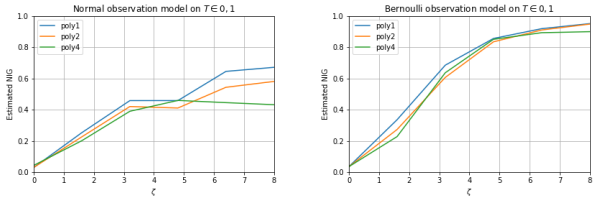

Figure 7: NIG of top neighborhood for $T \in \{0, 1\}$. Left plot: Normal model. Right plot: Bernoulli model.

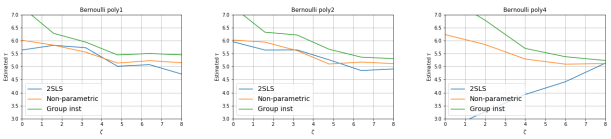

Figure 8: LoRD3 Bernoulli model $\hat{\tau}$ on $T \in \{0, 1\}$. Each plot represents a different $p(x)$ specification. True $\tau = 5$.

cluster analysis [30], parametric methods using Bartlett [13] and Student-t [12] test statistics, and non-parametric methods using Mann-Whitney [28] and Kolmogorov-Smirnov [27] test statistics.

We generate one-dimensional $T \in \mathbb{R}$ using $\zeta \in [0, 2.5]$ (see §4.1). For each ζ value we generate 50 experiments with 1000 data points. We apply all changepoint methods and LoRD3 with the Normal model and three $f(x)$ specifications. Mean squared error from the true discontinuity is used to evaluate the results in Fig. 9.

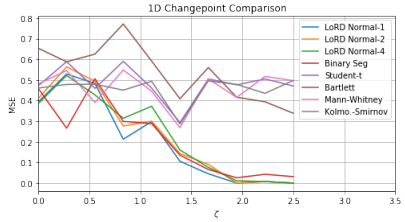

Figure 9: Comparison of LoRD3 with changepoint methods.

We observe that all LoRD3 configurations are superior to changepoint methods for high ζ. Binary Segmentation equals the performance of LoRD3 MSE at low ζ, but has worse MSE than LoRD3 as ζ increases. Moreover, we note that these changepoint methods are limited to one dimension. LoRD3 advances into new territory by discovering RDDs in arbitrary dimensions and thus may be considered a generalization of changepoints to multiple dimensions.

4.5 Student test score data

Jacob et al. [18] consider the effect of an educational intervention on math test scores. We use their student test score dataset which is based on seventh-grade math assessments. It contains two sets of scores: "pre-test" scores that reflect student achievement before a potential intervention, and "post-test" scores after the intervention. Only students who received below 215 on the pre-test were intervened upon. Thus there is a sharp RDD at pre-test score 215.

The data has 2,606 observations and eight covariates, x, for each student. Six covariates are binary indicators for gender, special education status, eligibility for reduced-price lunch, English as second language status, and ethnicity (Black, White, Hispanic or Asian). Two of the covariates are real-valued: age of student and pre-test score. We use both real-valued variables as z even though only pre-test score is the true relevant variable. The intervention status is T and the post-test score is y. The true value of τ is 10.

We apply LoRD3 with the Normal and Bernoulli models, $k = 100$, and a 1-degree polynomial for $f(x)$ and $p(x)$. $LLR(s)$ is depicted in Fig. 10. The strip of high LLR around pre-test score 215 indicates that LoRD3 was able to locate the discontinuity with both observation models.

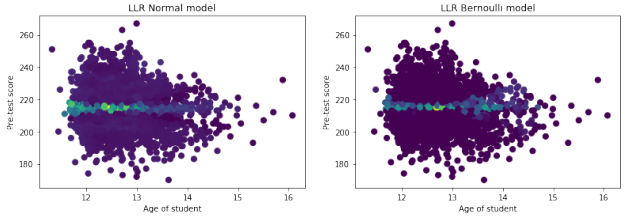

Figure 10: $LLR(s)$ with student age as x-axis and pre-test score as y-axis. Normal model on left, Bernoulli model on right.

Table 1 lists the NIG, influence of the two z dimensions, and $\hat{\tau}$ over the top ten scoring neighborhoods. Both observation models yield high NIG and correctly identify pre-test score as the primary discontinuity variable. While the 2SLS and group instrument methods generally correctly yield $\hat{\tau}$ within the standard error of 10, the non-parametric method underestimates τ in both models.

Table 1: NIG, influence of z, and $\hat{\tau}$ for the student test data.

	Normal model	Bernoulli model
NIG	0.92 ± 0.02	0.93 ± 0.04
Influence: pre-test score	1.0 ± 0.0	1.0 ± 0.0
Influence: age of student	0.0 ± 0.0	0.0 ± 0.0
$\hat{\tau}$ 2SLS	8.89 ± 1.11	9.88 ± 0.98
$\hat{\tau}$ non-parametric	6.66	5.91
$\hat{\tau}$ Group inst	9.57 ± 1.18	6.30 ± 1.14

While the true data contains a sharp RDD, we inject synthetic noise to increase the difficulty of the search problem. We generate noisy treatment, T_ρ, such that $P(T_{\rho,i} = T_i) = \rho$, where $\rho \in [0.5, 1]$. Thus when $\rho = 1$, $T_\rho = T$, and the data contains a sharp RDD. When $\rho = 0.5$, $T_{\rho,i}$ is 0 or 1 with equal probability, resulting in no signal. Between those two extremes, the data exhibits a fuzzy RDD at pre-test score 215. Fig. 11 depicts T_ρ at $\rho \in \{0.5, 0.75, 1\}$ to provide intuition for the magnitude of ρ noise.

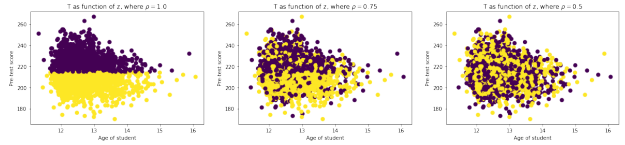

Figure 11: Student data with pre-test score on y-axis, age of student on x-axis, and T indicated by circle color. Left plot is $\rho = 1$ (true T), center plot is $\rho = 0.75$, and right plot is $\rho = 0.5$.

For each value of $\rho \in [0.5, 1]$, we generate 25 experiments with 2000 randomly sampled data points. We apply LoRD3 as above and

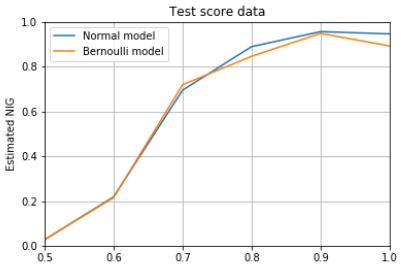

Figure 12: NIG of top LoRD3 neighborhood on student test score data using Normal and Bernoulli observation models.

Table 2: NIG and influence of z for full university GPA data.

	Normal model	Bernoulli model
NIG	0.59 ± 0.05	0.71 ± 0.06
Influence: GPA cutoff	0.79 ± 0.37	1.0 ± 0.0
Influence: HS grade pct	0.59 ± 0.36	0.11 ± 0.13
Influence: credits yr 1	0.20 ± 0.40	0.0 ± 0.0
Influence: age of student	0.0 ± 0.0	0.0 ± 0.0

show results from the top scoring neighborhood in Fig. 12. Both observation models converge to nearly $NIG = 1$ well before $\rho = 1$, demonstrating that they can identify RDDs in noisy data.

4.6 College GPA data

Lindo et al. [22] analyze the effect of academic probation on students at a Canadian university with three campuses. Students are placed on probation if their first year GPA is below a cutoff value. This cutoff induces an RDD that Lindo et al. [22] use to determine the causal effect of academic probation on educational outcomes.

The data has 44,362 observations and nine covariates, x, for each student. Five of the covariates are binary indicators for gender, English as a first language, being born in North America, and two variables to indicate which campus the student attended. Four covariates are real-valued: matriculation age, credits attempted in first year, high school grade percentile, and distance of GPA from the GPA cutoff. We use all four real-valued variables in z even though only distance from the GPA cutoff is the relevant factor. The intervention status is T. There are five outcomes of interest: decision to leave after the first academic term, GPA in the next academic term, and whether the student graduated within 4, 5 or 6 years.

We apply LoRD3 with Normal and Bernoulli models, $k = 100$, and $f(x)$ as a 1-degree polynomial. Table 2 lists the NIG and influence of the z dimensions over the top ten scoring neighborhoods. The Bernoulli model yields substantially higher NIG than the Normal model, as expected in data with binary T. Both methods correctly rank GPA cutoff as the most influential dimension of z, but the importance of this variable is less pronounced in the Normal results. Although [22] estimates $\hat{\tau}$ for each outcome using all students within 0.6 grade points of the cutoff GPA, these are not ground truth. Table 3 provides these values as well as LoRD3 $\hat{\tau}$ values using the methods in §3.5. Though we expect deviations between these estimates, most values, and nearly all signs, are quite similar.

Table 3: Estimated $\hat{\tau}$ on university GPA data.

	Leave	GPA y2	Grad y4	Grad y5	Grad y6
Lindo et al. [22]	0.018	0.233	-0.020	-0.044	-0.024
Normal LoRD3 model					
2SLS	0.066	0.255	-0.211	-0.207	-0.116
Non-para	0.030	-0.025	-0.056	-0.189	-0.172
Group inst	0.058	0.188	-0.198	-0.187	-0.094
Bernoulli LoRD3 model					
2SLS	0.021	0.219	-0.273	-0.245	-0.050
Non-para	0.017	0.125	-0.076	-0.036	0.105
Group inst	0.020	0.055	-0.293	-0.098	0.144

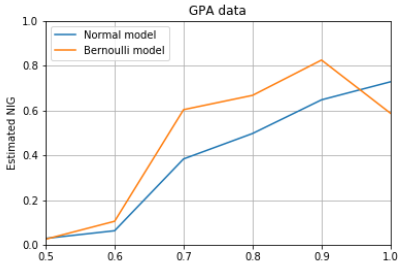

Figure 13: NIG of top LoRD3 neighborhood on university GPA data using Normal and Bernoulli observation models.

In order to increase the difficulty of detection, we inject increasingly high ρ noise, as described in §4.5. For each ρ value we generate 25 experiments with 2000 randomly sampled data points. We apply LoRD3 with the same parameters as above and show NIG results for the top scoring neighborhood in Fig. 13. In this case, there is substantial improvement using the Bernoulli model. While both models improve at higher values of ρ the Bernoulli increases to NIG= 0.8 while the Normal model only reaches NIG= 0.7.

4.7 Emergency department usage

We consider aggregate emergency department (ED) patient data used to study the impact of health insurance on ED usage [1, 2]. Data come from 2.2 million ED visits between 2002-2009 in Arizona, California, Iowa, New Jersey, and Wisconsin. The only covariate is patient age and previous studies identified RDDs at ages 19 and 23. The existence of multiple discontinuities in this data is particularly interesting and we apply LoRD3 to see which RDDs it can detect. Note that due to endogeneity issues Anderson et al. [1] develop a specialized τ estimation approach that is not replicated here.

Letting $x = z$ be ED patient age, we separately consider T as percentage of ED patients with private insurance and T as percentage of ED patients without insurance. In both cases we use $f(x)$ as a 3-degree polynomial and run 1000 randomization tests. We depict data, $LLR(s)$, and the $\alpha = 0.05$ significance threshold in Fig. 14.

The most prominent RDD peaks at 23 years 3 months for both T. This corresponds to the RDD used in Anderson et al. [2] and reflects that health insurance plans at the time allowed full-time students to remain on their parents' plans until age 23. Both setups also identify an RDD at age 19 corresponding to the RDD used in Anderson et al. [1]. This reflects that non-students were allowed to

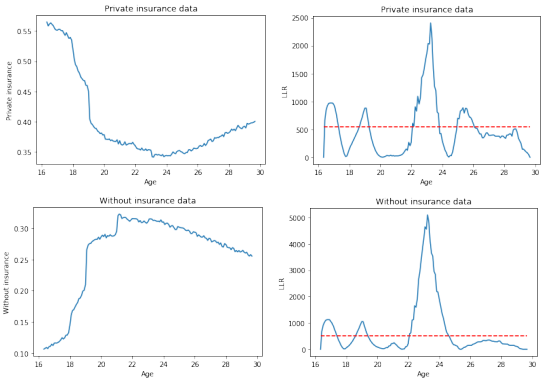

Figure 14: ED patients with private insurance on top, without insurance on bottom. Left: % of patients vs. age. Right: $LLR(s)$ centered at each age. Red line indicates $\alpha = 0.05$ level.

Table 4: LoRD3 and changepoint comparisons for ED data.

	Private insurance	Without Insurance
LoRD3	16.83, 19, 23.25, 25.33	16.83, 19, 23.25
Binary Seg	18, 19.08, 22, 26.67	19.08, 21.08, 24.42, 27.08
Student-t	17.42	17.42
Bartlett	17.42	17.33
Mann-Whitney	16.92	17.25
Kolmo.-Smirnov	16.92	17.25

remain on their parents' insurance plans until age 19. Interestingly, both setups also identify an additional RDD centered at 16 years 10 months which may provide useful information for research. Finally, the setup with private insurance as T identifies a weaker RDD that peaks at 25 years 4 months. The identification of both known and unexplored discontinuities confirms the ability of LoRD3 to identify RDDs and to provide potentially policy-relevant insights.

We compare these results to the changepoint methods from §4.4. Binary Segmentation, which can find multiple changepoints, correctly identified the discontinuity at age 19, but was not able to discern the discontinuity at age 23. The remainder of the methods seem to corroborate that there is a discontinuity around age 17, though the precise values they detect differ slightly from LoRD3.

5 CONCLUSION

In this paper we described an automated statistical machine learning method for discovering RDDs in observation settings that is domain-agnostic and applicable to multi-dimensional data. We derive observation models for both real-valued and binary treatments as well as an automated validation and treatment estimation framework. After demonstrating robust performance in a variety of synthetic settings, we apply the approach to three real datasets illustrating the method's ability to discover both previously known, and potentially unexplored, RDDs.

ACKNOWLEDGMENTS

The authors thank Akshaya Jha. This work is supported by NSF GRFP DGE-1252522, NSF IIS-0953330, and NSF IIS-1563887.

REFERENCES

[1] M Anderson, C Dobkin, and T Gross. 2012. The effect of health insurance coverage on the use of medical services. *AEJ: Economic Policy* 4, 1 (2012).
[2] Michael L Anderson, Carlos Dobkin, and Tal Gross. 2014. The effect of health insurance on emergency department visits: Evidence from an age-based eligibility threshold. *Review of Economics and Statistics* 96, 1 (2014), 189–195.
[3] Joshua D Angrist and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion.* Princeton University Press.
[4] Susan Athey. 2015. Machine learning and causal inference for policy evaluation. In *Proceedings of the 21th ACM SIGKDD Conference.* ACM, 5–6.
[5] Susan Athey and Guido Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *Proc. of the National Academy of Sciences* 113, 27 (2016), 7353–7360.
[6] M Bertanha. 2016. Regression discontinuity design with many thresholds. (2016).
[7] David Card, Alexandre Mas, and Jesse Rothstein. 2008. Tipping and the Dynamics of Segregation. *The Quarterly Journal of Economics* 123, 1 (2008), 177–218.
[8] Drew Dimmery. 2016. *rdd: Regression Discontinuity Estimation.* R package v0.57.
[9] J Hahn, P Todd, and W Van der Klaauw. 2001. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* 69, 1 (2001).
[10] Bruce E Hansen. 2000. Sample splitting and threshold estimation. *Econometrica* 68, 3 (2000), 575–603.
[11] Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. 2016. Counterfactual Prediction with Deep Instrumental Variables Networks. *arXiv preprint arXiv:1612.09596* (2016).
[12] Douglas M Hawkins, Peihua Qiu, and Chang Wook Kang. 2003. The changepoint model for statistical process control. *Journal of quality technology* 35, 4 (2003).
[13] Douglas M Hawkins and KD Zamba. 2005. A change-point model for a shift in variance. *Journal of Quality Technology* 37, 1 (2005), 21.
[14] Jennifer L Hill. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20, 1 (2011), 217–240.
[15] Yu-Chin Hsu, Shu Shen, et al. 2016. *Testing for treatment effect heterogeneity in regression discontinuity design.* Technical Report. Academia Sinica, Taiwan.
[16] Edward McFowland III, Sriram Somanchi, and Daniel B. Neill. 2018. Efficient Discovery of Heterogeneous Treatment Effects in Randomized Experiments via Anomalous Pattern Detection. *Working paper* (2018).
[17] Guido W Imbens and Thomas Lemieux. 2008. Regression discontinuity designs: A guide to practice. *Journal of econometrics* 142, 2 (2008), 615–635.
[18] Robin Jacob, Pei Zhu, Marie-Andrée Somers, and Howard Bloom. 2012. A Practical Guide to Regression Discontinuity. *MDRC* (2012).
[19] Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *ICML*. 3020–3029.
[20] Martin Kulldorff. 1997. A spatial scan statistic. *Communications in Statistics-Theory and methods* 26, 6 (1997), 1481–1496.
[21] David S Lee and Thomas Lemieuxa. 2010. Regression discontinuity designs in economics. *Journal of economic literature* 48, 2 (2010), 281–355.
[22] J M Lindo, N J Sanders, and P Oreopoulos. 2010. Ability, gender, and performance standards: Evidence from academic probation. *AEJ: Applied Economics* 2, 2 (2010).
[23] Justin McCrary. 2008. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of econometrics* 142, 2 (2008).
[24] Sendhil Mullainathan and Jann Spiess. 2017. Machine learning: an applied econometric approach. *Journal of Economic Perspectives* 31, 2 (2017), 87–106.
[25] Daniel B Neill. 2012. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74, 2 (2012), 337–360.
[26] Sean F Reardon and Joseph P Robinson. 2012. Regression discontinuity designs with multiple rating-score variables. *JREE* 5, 1 (2012), 83–104.
[27] Gordon J Ross and Niall M Adams. 2012. Two nonparametric control charts for detecting arbitrary distribution changes. *Journal of Quality Technology* (2012).
[28] Gordon J Ross, Dimitris K Tasoulis, and Niall M Adams. 2011. Nonparametric monitoring of data streams for changes in location and scale. *Techno.* 53, 4 (2011).
[29] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. 2001. Estimating the support of a high-dimensional distribution. *Neural computation* 13, 7 (2001), 1443–1471.
[30] Andrew Jhon Scott and M Knott. 1974. A cluster analysis method for grouping means in the analysis of variance. *Biometrics* (1974), 507–512.
[31] Amit Sharma. 2016. *Necessary and probably sufficient test for finding valid instrumental variables.* Technical Report. working paper, Microsoft Research, NY.
[32] Amit Sharma, J M Hofman, and D J Watts. 2016. Split-door criterion for causal identification: Automatic search for natural experiments. *arXiv:1611.09414* (2016).
[33] Wilbert Van der Klaauw. 2008. Regression–discontinuity analysis: a survey of recent developments in economics. *Labour* 22, 2 (2008), 219–245.
[34] Vivian C Wong, Peter M Steiner, and Thomas D Cook. 2013. Analyzing regression-discontinuity designs with multiple assignment variables: A comparative study of four estimation methods. *Journal of Educational and Behavioral Statistics* (2013).
[35] Zhe Zhang and Daniel B Neill. 2016. Identifying significant predictive bias in classifiers. *arXiv preprint arXiv:1611.08292* (2016).