Gaussian Process Regression Discontinuity

Joseph T. Ornstein*
JBrandon Duck-Mayr†

February 20, 2020

Abstract

In applied settings, regression discontinuity (RD) designs often suffer from noisy data and low power. This tends to produce exaggerated causal effect estimates, typified by implausibly large slope and/or concavity parameters. We propose a new method for estimating causal effects in RD designs called Gaussian Process Regression Discontinuity (GPRD). This approach overcomes the major disadvantages of global polynomial estimators and does so with lower variance than local linear estimators. When applied to several recent empirical examples from the published literature, GPRD yields more modest and plausible treatment effect estimates. We make this new method available through the `R` package `gprd`.

*Postdoctoral Research Associate, Washington University in St. Louis
†PhD Candidate, Washington University in St. Louis

1 Introduction

Regression discontinuity (RD) is an approach to causal inference that leverages a discontinuous change in treatment at a cutoff. The key identification assumption of the RD design is the continuity of other pre-treatment covariates; so long as treatment status is the only variable that changes discontinuously at the cutoff, the causal effect of treatment is identifiable at that point (Hahn, Todd and Van Der Klaauw, 2001). Because such thresholds, cutoffs, and boundaries are a common feature of political institutions, RD has proven a popular research design in political science over the past two decades (de la Cuesta and Imai, 2016).

Estimating the average treatment effect at the cutoff requires finding the limits of the outcome variable as it approaches from the left and right. Traditionally, researchers have estimated these limits in one of two ways. The first approach is to fit a global polynomial regression on either side of the cutoff, then take the difference between the two regressions' predictions at the cutoff. This approach suffers from a disadvantage common to any parametric estimation strategy: if the true data generating process is not captured by the researcher's model specification, then any causal effect estimate is likely to be biased. Gelman and Imbens (2017) catalogue three other disadvantages of this approach: global methods tend to overfit to observations far away from the cutoff, estimates are sensitive to the researcher's choice of polynomial degree, and confidence intervals have poor coverage.

In response to these problems, researchers have developed a more sophisticated nonparametric approach based on local linear regression. An important practical consideration in this literature involves choosing a bandwidth – i.e. how much data to include in the estimation (Imbens and Kalyanaraman, 2012). This choice involves a bias-variance tradeoff: too wide a bandwidth yields biased estimates; too narrow a bandwidth reduces number of observations used to estimate the treatment effect, increasing the variance of the RD estimator.

Calonico, Cattaneo and Titiunik (2014), hereafter CCT, propose a nonparametric approach to RD estimation that selects the bandwidth (h) to minimize the mean squared

error of the RD estimator. Their approach estimates the limits approaching the cutoff using local linear regression weighted by a triangular kernel and adjusting for a bias-correction term, and the authors derive robust standard errors for inference. Confidence intervals from this method produce the best empirical coverage of any method proposed to date, and its estimates perform well in a wide array of simulations.

Because local linear RD estimates focus exclusively on observations near the cutoff, datasets that are sparse or noisy in that neighborhood can yield low-powered hypothesis tests. As the ongoing replication crisis in experimental sciences has demonstrated, such low-powered studies are pernicious when combined with a publication bias towards statistically significant results (Button et al., 2013). The estimated treatment effects from publishable low-powered studies tend to significantly overestimate the true effect, a distortion that Gelman and Carlin (2014) refer to as Type M errors.

To illustrate this problem in the RD case, consider a simulation in which the variables X and Y are generated as follows:

$$X \sim U(-1, 1)$$

$$Y_i = \mathbf{1}(X_i \geq 0)\tau + X_i\beta + \varepsilon_i$$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

where $\mathbf{1}(\cdot)$ is the indicator function, β is a slope parameter, and τ is the true causal effect of treatment. When τ and β equal zero, Y is pure Gaussian noise.

Figure 1 displays two instances of this simulated data when $\tau = 0$, including local linear and local quadratic RD estimates with CCT optimal bandwidths. When the true effect of treatment equals 0, the CCT 95% confidence intervals reject the null hypothesis ($H_0 : \tau = 0$) roughly 7% of the time. Of these rejections, over 90% display a characteristic 'Zig Zag' pattern illustrated in the figure – steep-sloped regression functions on either side of the

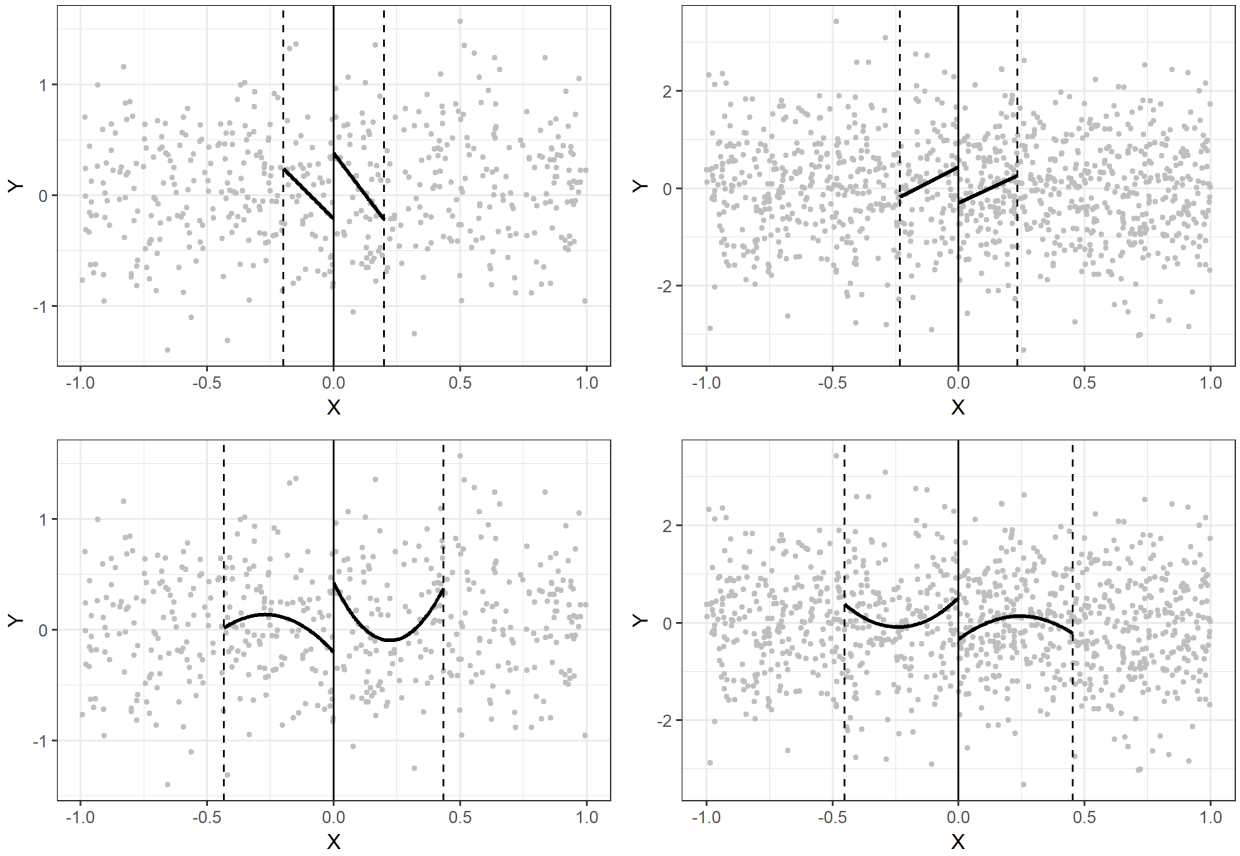


Figure 1: False positive estimates from the Monte Carlo simulation. Dashed vertical lines are the CCT MSE-optimal bandwidth. Solid vertical line is the cutoff. Gray points are the raw data, and black lines are the local low-order polynomial fits. Left two figures generated with parameters $\tau = 0$, $\beta = 0$, $\sigma = \frac{1}{2}$, $n = 500$. Right two figures generated with parameters $\tau = 0$, $\beta = 0$, $\sigma = 1$, $n = 1000$.

cutoff and a treatment effect of the opposite sign.[1] This pattern is emblematic of false positive or exaggerated claims from noisy RD data, and will be present in the empirical applications below.

In some cases, there may be a good theoretical reason to believe that the treatment effect will move in the opposite direction of the slope of the regression function. For examples, see RD studies on the incumbency disadvantage in Latin America (Klašnja and Titiunik, 2017) or the effect of property tax incentives on homeowners' mobility rates (Ferreira, 2010). But one would rarely expect the slope of the conditional expectation function to diverge sharply

[1] Or, when estimated using a local quadratic regression, a concave regression function on one side of the cutoff, and a convex function on the other side.

only in the neighborhood of the cutoff. Absent such a theoretical motivation, large slope or concavity estimates near the cutoff are indicative of overfitting to noisy data.

Figure 2 plots the estimated RD treatment effect on the y-axis against the sum of the estimated slopes on the x-axis, but only for simulations where the CCT 95% confidence intervals reject the null hypothesis. In the top two panels of Figure 2, the value of τ is small relative to noise (σ). As a result, the test is low-powered, only rejecting the null hypothesis in 9% and 17% of iterations. When an estimate passes the statistical significance threshold, it tends to be many multiples of the true value (Type M error) or have the wrong sign (Type S error). The largest distortions are associated with large estimated slopes on each side of the cutoff. Only when power is high (bottom two panels) are statistically significant estimates roughly centered around the true value of τ.

In the following section, we introduce a Bayesian method for estimating RD treatment effects based on Gaussian Process regression. Because it imposes a prior on the smoothness of the conditional expectation functions, it is a particularly useful method to moderate the exaggerated claims from low-powered RD studies. Gaussian Process Regression Discontinuity (GPRD) performs well in simulations, overcoming the disadvantages of traditional global polynomial approaches while producing lower variance estimates than local linear approaches. In section 3 we apply GPRD to several empirical examples from the published literature where low power has yielded exaggerated treatment effects.

2 Gaussian Process Regression for RD Designs

In regression discontinuity (RD) designs, we attempt to estimate the causal effect of a treatment that is assigned at a specific value of some *forcing variable*, x. We assume the outcomes y are a noisy function of the forcing variable x, and we are interested in the discontinuity in $f(x)$ induced by the treatment. We propose using Gaussian process regression to approximate this conditional expectation function $f(x)$ and estimate causal effects in RD designs.

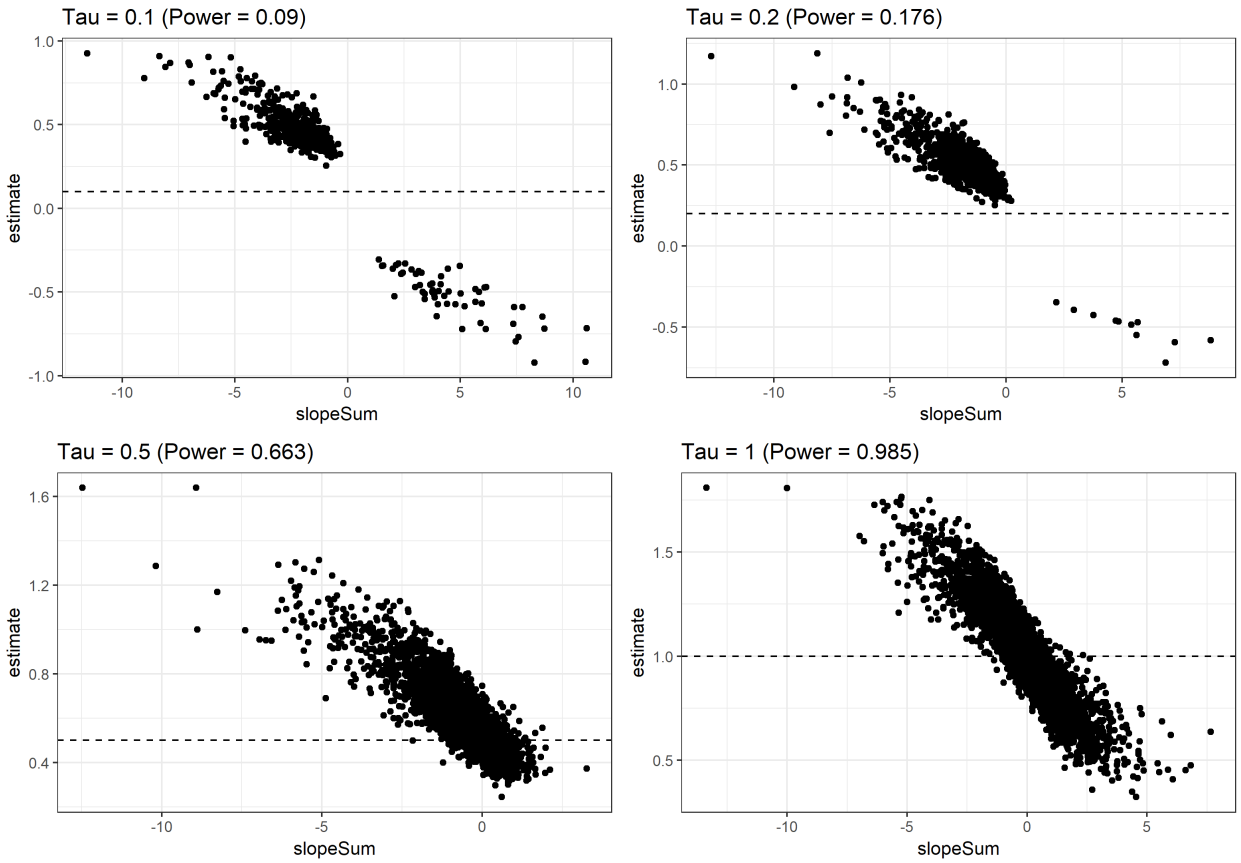


Figure 2: Observed treatment effect and slope estimates from Monte Carlo after applying statistical significance filter $\alpha = 0.05$, for varying values of τ. More extreme slope estimates yield more distorted treatment effect estimates. When the true value of τ is low, the estimated treatment effects from cases that pass statistical significance are greatly exaggerated (Type M error) and many have the wrong sign (Type S error).

This approach was first proposed in Branson et al. (2019); we extend the method to relax some restrictive assumptions (the "same covariance" assumption from that paper) and permit more powerful assumptions (adding a "same mean function" assumption) as befits the researcher, and provide open source software for estimation (an `R` package, `gprd`).

To formalize the RD problem we start by assuming the outcomes are normally distributed,

$$y \sim \mathcal{N}\left(f\left(x\right), \sigma_y^2 I\right), \tag{1}$$

where $f\left(x\right)$ is an unknown function of the forcing variable x. A treatment occurs at a

cutoff value c of x; that is, all outcomes where $x \geq c$ receive the treatment. For simplicity here we will assume $c = 0$.

Alternatively, we might assume that the outcomes are normally distributed but with different mappings between input and output on either side of the cutoff;

$$y_+ \sim \mathcal{N}\left(f_+\left(x_+\right), \sigma_y^2 I\right), \tag{2}$$

$$y_- \sim \mathcal{N}\left(f_-\left(x_-\right), \sigma_y^2 I\right), \tag{3}$$

We will propose two different methods to estimate the treatment effect τ of the intervention using Gaussian process (GP) regression. The first, which we call the *global GPRD estimator*, fits a single GP model for all observations, with a dummy variable to indicate treatment, and estimates the treatment effect using the difference in prediction when $x = c$ and $D = 1$ and $D = 0$. The second, which we call the *piecewise GPRD estimator*, models the data separately on either side of the cutoff and estimates the treatment effect using the difference between predictions at the cutoff from the "right equation" and "left equation."

2.1 Gaussian Process Regression

GP regression is a method used to learn the mapping from x to y when its functional form is not known, accomplished by placing a GP prior over the function space. As this methodology is uncommon in the political science literature, we provide a brief overview here.

A Gaussian process (GP) is an infinite dimensional generalization of the normal distribution. More specifically, it is "collection of random variables, any finite number of which have a joint Gaussian distribution" (Rasmussen and Williams, 2006, 13). The mean and covariance of this normal distribution is given as functions of the inputs, so that we say

$$f(x) \sim \mathcal{GP}(m(x), K(x)). \tag{4}$$

Common examples for the mean function are the "mean zero" function $m(x) = 0$ and the "linear mean" function $m(x) = x\beta$. A common example of a covariance function (also called a kernel) is the isometric squared exponential covariance function, also called simply the squared exponential covariance function or the radial basis function,

$$K(x, x') = \sigma_f^2 \exp\left(-0.5\frac{(x - x')^2}{\ell^2}\right), \tag{5}$$

where σ_f is a hyperparameter called the scale factor, which scales the entire covariance matrix, and ℓ is a hyperparameter called the length scale, which influences how quickly f varies in x. Then the i, j element of the covariance matrix is given by $K(x_i, x_j)$.

This setup allows us to model distributions over *functions* rather than simply distributions over *variables*. Then, rather than assume we know the form of the mapping between the input variables $\mathbf{X}$ and outcomes $\mathbf{y}$—such as a polynomial of a particular order—we can instead use a GP to place a probability distribution over all possible mappings from $\mathbf{X}$ to $\mathbf{y}$. With a Gaussian likelihood for the data given $\mathbf{X}$ and $f(x)$, a posterior distribution over $f(x)$ is then given by application of Bayes' rule utilizing Gaussian identities.[2] Crucially for our purposes, well known results provide the posterior predictive distribution for a test observation x^* as

$$f(x^*) \sim \mathcal{N}(m^*, C^*) \tag{6}$$

$$m^* = m(x^*) + K(x^*, \mathbf{x})[K(\mathbf{x}) + \sigma_y^2 I]^{-1}(\mathbf{y} - m(\mathbf{x})) \tag{7}$$

$$C^* = K(x^*) - K(x^*, \mathbf{x})[K(\mathbf{x}) + \sigma_y^2 I]^{-1}K(\mathbf{x}, x^*). \tag{8}$$

Note that Bayesian linear regression is a special case of Gaussian process regression, using the linear covariance function $K(x, x') = x \cdot x'$.[3] When using other kernels, we simply allow

[2] For a more detailed derivation of this posterior distribution, see Rasmussen and Williams (2006), Chapter 2.

[3] A constant hyperparameter is added to K if an intercept is desired.

non-linearity in the mapping from input to response. In other words, GP regression is a flexible extension to Bayesian linear regression to account for uncertainty in the functional form mapping predictors to response by assuming the covariance between outcomes is a function of the predictor variables. In the case of the common squared exponential covariance function (and its extension discussed in Section 2.2, the automatic relevance determination kernel), we assume that covariance between outcomes is simply a function of distance in the covariate space.

2.2 The Global GPRD Estimator

For the global GPRD estimator, we place a Gaussian process (GP) prior on $f(x)$,

$$p(f) = \mathcal{GP}(X\beta, K(X)), \tag{9}$$

where K is the squared exponential automatic relevance determination covariance function

$$K(X, X') = \sigma_f^2 \exp\left(-0.5 \sum_j \frac{\left(X_{\cdot,j} - X'_{\cdot,j}\right)^2}{\ell_j^2}\right) \tag{10}$$

with hyperparameters σ_f, the scale factor as in the squared exponential covariance function from Equation 5, and ℓ is a length scale *vector* with a separate length scale for each predictor variable, and where $X = \Big[\mathbf{1}|x|D\Big]$, with

$$D = \begin{cases} 1 & \text{if } x \geq c, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

Note that there is never a difference between observations on the value of the intercept column, so the X that goes into the mean function $X\beta$ will include that column, but the X going into the covariance function does not (and accordingly ℓ is of length two, not three);

we suppress this difference in the notation for simplicity.

We use a Gaussian prior for β, with mean $\mathbf{b}$ and covariance B. Then the posterior over β is given by

$$\beta \mid \mathbf{y}, X \sim \mathcal{N}\left(\bar{\beta}, \Sigma_\beta\right), \tag{12}$$

$$\bar{\beta} = \Sigma_\beta \left(X^T K_y^{-1} \mathbf{y} + B^{-1}\mathbf{b}\right), \tag{13}$$

$$\Sigma_\beta = \left(B^{-1} + X^T K_y^{-1} X\right)^{-1}, \tag{14}$$

$$K_y = K(X) + \sigma_y^2 I \tag{15}$$

(see Rasmussen and Williams (2006), section 2.7). Note that for the common case where $\mathbf{b} = \mathbf{0}$, $\bar{\beta}$ simplifies to $\Sigma_\beta(X^T K_y^{-1}\mathbf{y})$.

Then the mean and variance of f at a test point X_* is

$$\bar{f}(X_*) = X_*\bar{\beta} + K_* K_y^{-1}(\mathbf{y} - X\bar{\beta}), \tag{16}$$

$$\mathrm{cov}(f_*) = K_{**} - K_* K_y^{-1} K_*^T + R^T (X^T K_y^{-1} X)^{-1} R, \tag{17}$$

$$K_* = K(X_*, X), \tag{18}$$

$$K_{**} = K(X_*), \tag{19}$$

$$R = X_*^T - X^T K_y^{-1} K_*^T \tag{20}$$

(See Equations 2.24, 2.38, and 2.41 in Rasmussen and Williams 2006). These differ from Equations 7 and 8 because we have incorporated uncertainty in the mean function parameters.[4]

[4]One could instead select mean function coefficients by maximizing the marginal log likelihood as we discuss for the covariance function hyperparameters in Section 2.4, in which case $\bar{f}(X_*)$ and $\mathrm{cov}(f_*)$ would again be given by Equations 7 and 8.

So we are interested in the treatment effect

$$\tau_{GPRD-G} \stackrel{\text{def}}{=} f\left(\begin{bmatrix} 0 & 1 \end{bmatrix}\right) - f\left(\begin{bmatrix} 0 & 0 \end{bmatrix}\right), \tag{21}$$

or the difference between $f(x = 0, D = 1)$ and $f(x = 0, D = 0)$, which is distributed

$$\tau_{GPRD-G} \sim \mathcal{N}(\mu_*, \Sigma_*), \tag{22}$$

$$\mu_* = \bar{f}\left(\begin{bmatrix} 0 & 1 \end{bmatrix}\right) - \bar{f}\left(\begin{bmatrix} 0 & 0 \end{bmatrix}\right), \tag{23}$$

$$\Sigma_* = \operatorname{cov}\left(f\left(\begin{bmatrix} 0 & 1 \end{bmatrix}\right)\right) + \operatorname{cov}\left(f\left(\begin{bmatrix} 0 & 0 \end{bmatrix}\right)\right). \tag{24}$$

Note the key differences between the global GPRD estimator and the global polynomial RD estimator. In the global polynomial model, the treatment effect is taken to be the coefficient on D, which gives the difference in expected outcomes for the treated and untreated observations at any point x. In the GP model, this is not true; the coefficient on D impacts the expected outcome, but the effect of the treatment also runs through the kernel (see Equation 16).

Also note the differences between the global GPRD estimator we introduce and the approach taken in Branson et al. (2019). Branson et al. assume shared covariance function parameters while fitting separate GP regressions to observations to the left and right of the treatment cutoff.[5] The global GPRD estimator makes stronger assumptions about the relationship between outputs in the treatment and control in assuming that not only are the covariance function parameters shared, but by fitting only one GP regression to all the data. Then the difference in function values is given *only* by allowing a mean intercept shift for the treated group and added unit covariate space distance in the kernel, rather than fitting entirely separate functions for the treated and control groups. With appropriate selection of

[5]While they state the shared covariance assumption as one that can be used or not, they rely on the assumption in all applications and in most proofs.

the treatment dummy length scale, this modeling choice may be able to improve precision by leveraging more data while allowing enough change in function outputs between treatment and control groups to appropriately model the data.

2.3 The Piecewise GPRD Estimator

Non-parametric regression discontinuity designs fit polynomials on either side of the discontinuity, and estimate the treatment effect as the difference of the limit of the polynomials at the cutoff. The piecewise GP estimation strategy is to simply place Gaussian process (GP) priors over the functions on either side of the cutoff, which will be much more flexible than the polynomial regression approach and less sensitive to overfitting predictions close to the cutoff to observations far away from the cutoff. By design, the fit near the cutoff will rely more on inputs close to the cutoff than those far away.

To formalize, x_+ will be the inputs to the right of the cutoff and y_+ the corresponding outcomes, and analogously for x_- and y_-. Then we will learn two functions, $f_+ : x_+ \to y_+$ and $f_- : x_- \to y_-$. To do so, we place GP priors over the functions,

$$f_+ \sim \mathcal{GP}(X_+\beta_+, K(x_+)), \tag{25}$$

$$f_- \sim \mathcal{GP}(X_-\beta_-, K(x_-)), \tag{26}$$

where $X_\cdot$ prepends the column vector $\mathbf{1}$ to $x_\cdot$, $\beta_\cdot$ is a vector giving the intercept and slope of the linear mean function, and $K(\cdot)$ is the isometric squared exponential covariance function from Equation 5.

We again use a Gaussian prior for β, with mean $\mathbf{b}$ and covariance B, and the mean and variance of $f_\cdot$ at a test point x_* is as given in Equations 16 and 17. Then we are interested in the treatment effect

$$\tau_{GPRD-L} \stackrel{\text{def}}{=} f_+(0) - f_-(0), \tag{27}$$

which is distributed

$$\tau_{GPRD-L} \sim \mathcal{N}(\bar{f}_+(0) - \bar{f}_-(0), \text{cov}(f_+(0)) + \text{cov}(f_-(0))). \tag{28}$$

Note the key differences between the piecewise GPRD estimator and local linear or polynomial RD estimators. The local polynomial estimators use only a portion of the data to either side of the cutoff to avoid undue influence of observations far from the cutoff. In contrast, the GP model is able to use all of the available data because the covariance between outputs decreases with distance in the covariate space, a natural and smooth way to decrease undue influence of observations far from the cutoff while still borrowing *some* information from them. Additionally, the local methods require specification of the degree of the local polynomials, and this researcher-imposed model restriction can significantly impact inference. By contrast, in the GP model, we place a prior over the possible mappings from x to y and learn $f(x)$ from the data. The combination of these differences largely appeases the criticisms of global polynomial RD estimation in (Gelman and Imbens, 2017).

Also note the differences between the piecewise GPRD estimator we introduce and the approach taken in Branson et al. (2019). While Branson et al. largely rely on a shared covariance assumption, the piecewise GPRD estimator is the result of abandoning that assumption and selecting different covariance function hyperparameters for the treatment and control groups. When the covariance function hyperparameters in fact should be shared, parameter selection or sampling routines should be able to recover that. When they are not, and the mapping from forcing variable to outcomes is truly best viewed as potentially wholly different between the treatment and control groups, the piecewise GPRD estimator we introduce may be most appropriate. Taken together, our global and piecewise GPRD estimators can be seen as capturing a fairly wide range of restrictiveness of prior assumptions in GP regression estimation of treatment effects in RD designs. It is also useful to acknowledge one important assumption shared by our approaches, that of *stationarity*, or that covariance

hyperparameters do not themselves also vary as a function of the forcing variable. We leave the relaxation of that assumption to future work.

2.4 Choosing Hyperparameters

Because our inferences are strongly affected by the choice of hyperparameters σ_y, σ_f, and ℓ, we need a theoretically-grounded, automatic procedure for selecting their values. The common practice in GP regression is to use the hyperparameters that maximize the log marginal likelihood. In the case of a linear mean, the log marginal likelihood is

$$\log p(y \mid X, \mathbf{b}, B) = -\frac{1}{2} M^T Q^{-1} M - \frac{1}{2} \log |Q| - \frac{n}{2} \log 2\pi, \tag{29}$$

$$M = X\mathbf{b} - \mathbf{y}, \tag{30}$$

$$Q = K_y + XBX^T, \tag{31}$$

(see Equation 2.43 in Rasmussen and Williams 2006). Again, note that for the common case $\mathbf{b} = \mathbf{0}$, M reduces to $-\mathbf{y}$.

Then the gradient of the log marginal likelihood with respect to the hyperparameters $\theta = (\sigma_y, \sigma_f, \ell)$ in the case of the isometric covariance function is

$$\frac{\partial}{\partial \theta_i} \log p(y \mid X, \mathbf{b}, B) = \frac{1}{2} M Q^{-1} \frac{\partial Q}{\partial \theta_i} Q^{-1} M - \frac{1}{2} \operatorname{Tr}\left(Q^{-1} \frac{\partial Q}{\partial \theta_i} \right), \tag{32}$$

$$\frac{\partial Q}{\partial \theta} = \begin{bmatrix} 2\sigma_y I \\ 2\sigma_f \exp\left(-0.5 K_0\right) \\ \sigma_f^2 \exp\left(-0.5 K_0\right) \circ K_0 \end{bmatrix}, \tag{33}$$

where the i, j element of the matrix K_0 is given by

$$K_{0\,i,j} = \frac{(x_i - x_j)^2}{\ell^2}, \tag{34}$$

and we can use (e.g.) the conjugate gradient method to optimize the hyperparameters. The gradient for the automatic relevance determination case is analogous, but just extended with an element for each element of ℓ accordingly.

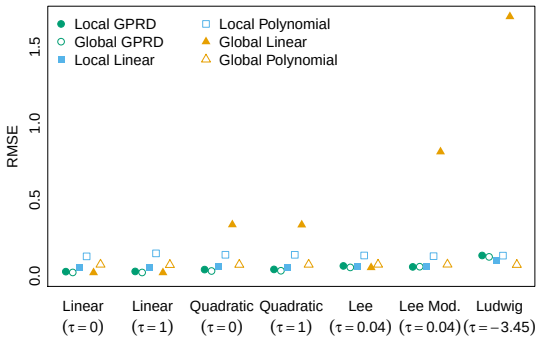
Alternatively, we could simply place a prior over the covariance function hyperparameters to include our uncertainty over them into the posterior distribution of the treatment effect. Note that in this case, however, we can no longer use exact inference but must instead resort to simulating the posterior. Instead, we can use priors with positive support and simulate the posterior with an MCMC sampler. For the remainder of the paper, we use covariance function parameters chosen by maximizing the marginal log likelihood (in particular to save time during the simulations), though an MCMC sampler for a fully Bayesian approach will be available in the `gprd` package.

2.5 Comparing Models: Simulation Evidence

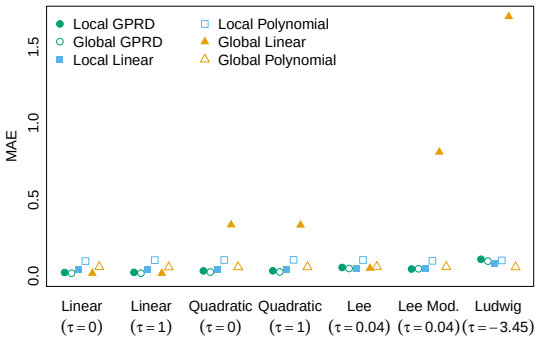
To assess performance of the GPRD estimators compared to existing methods, we engage in two simulation exercises. First, we use a common set of simulations from the RD literature, where the forcing variable x is given by $2z - 1$, with $z \sim \mathcal{B}(2, 4)$, and $y = f_j(x) + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, 0.1295^2)$. To specify the shape function $f_j(x)$, we use the three functions explored in Cattaneo, Frandsen and Titiunik (2014), themselves taken from Lee (2008) and Ludwig and Miller (2007). We additionally use global linear and quadratic functions, $f(x) = x + \tau I(x > 0)$ and $f(x) = x^2 + \tau I(x > 0)$, with $\tau = 0$ and $\tau = 1$, for a total of seven tested data generating processes. For each DGP, we simulate 1,000 datasets of 500 observations.

We estimate treatment effects and confidence intervals for global GPRD, piecewise GPRD, local linear, and fifth-degree global polynomial approaches (the Lee and Ludwig and Miller DGPs both use five-degree polynomials). Figure 3 depicts the root mean squared error, mean absolute error, confidence interval length, and confidence interval coverage averaged

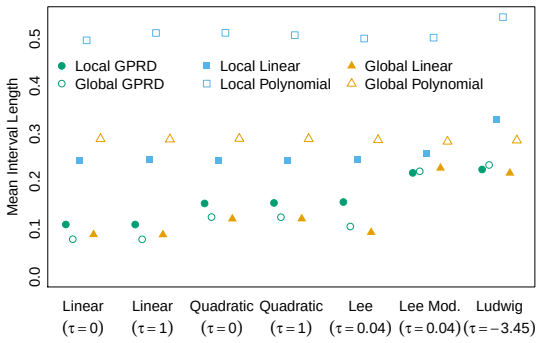
by simulation condition.



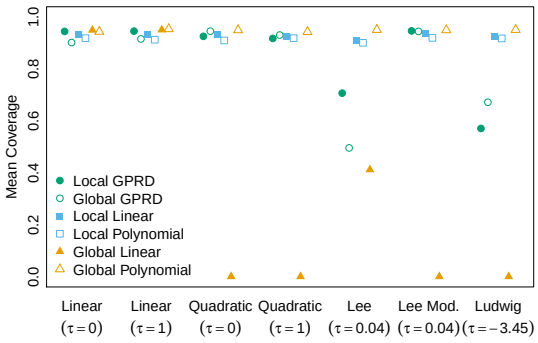
(a) Root mean squared error of treatment effect estimates.



(b) Mean absolute error of treatment effect estimates.



(c) Mean confidence interval length for treatment effect estimates.



(d) Mean 95% confidence interval coverage of true treatment effects.

Figure 3: Fit statistics for treatment effect estimates from the global and piecewise GPRD models (plotted in empty and filled green circles respectively), local linear and polynomial regression (plotted in filled and empty blue squares respectively), and global linear and polynomial regression (plotted in filled and empty yellow triangles).

The GPRD methods generally outperform other methods in terms of error and confidence interval length; when averaged across conditions, the GPRD methods outperform all other methods on these metrics, and for some DGPs, this difference is more pronounced. However, for the Lee and Ludwig and Miller datasets, GPRD method coverage suffered, though error only grew very slightly; this echoes findings in Branson et al. (2019), who note that with these DGPs, the functions' slope increases markedly near the cutoff, violating the stationarity assumption we leverage. This may be cured by using bandwidth cutoffs for the GP regressions' training data, a solution explored in Branson et al. (2019), or by using a non-stationary kernel, an extension we leave for future work. The results averaged across

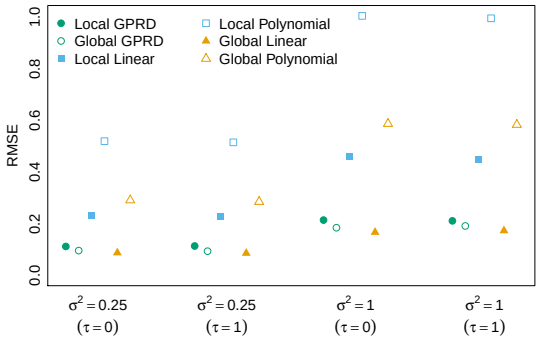
all conditions for all data generating processes, as well as only stationary data generating processes, are given in Table 1.

Table 1: Fit statistics averaged across replicates and data generating processes.

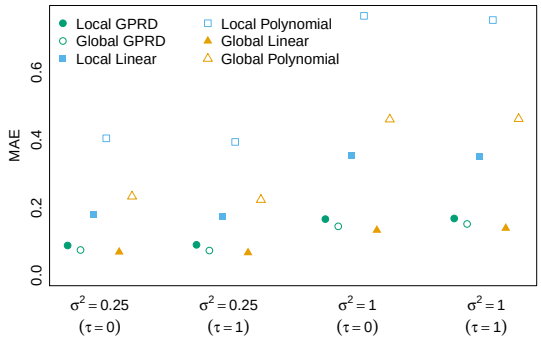
Estimator	MAE	RMSE	Mean CI Length	Mean CI Coverage
All DGPs				
Global GPRD	0.041	0.061	0.138	0.830
Piecewise GPRD	0.047	0.067	0.161	0.851
Local Linear	0.052	0.068	0.259	0.928
Local Polynomial	0.103	0.136	0.511	0.914
Global Linear	0.466	0.735	0.137	0.330
Global Polynomial	0.059	0.075	0.288	0.949
Stationary DGPs				
Global GPRD	0.028	0.038	0.126	0.928
Piecewise GPRD	0.032	0.043	0.149	0.936
Local Linear	0.047	0.060	0.247	0.932
Local Polynomial	0.102	0.137	0.506	0.916
Global Linear	0.303	0.422	0.129	0.380
Global Polynomial	0.059	0.075	0.288	0.948

In the absence of a violation of the stationarity assumption, the GPRD methods show a substantial improvement in terms of reduced bias and increased precision while retaining coverage. Importantly, note that these simulations, standard in the literature, assume very low observation noise. Returning to the setting discussed earlier in the paper where prior methods can produce exaggerated results, we see an even greater difference in performance between the methods. We explore the settings where $\varepsilon \sim \mathcal{N}(0, 0.5^2)$ and $\varepsilon \sim \mathcal{N}(0, 1)$ for $f(x) = x + \tau I(x > 0)$ for $\tau = 0$ and $\tau = 1$ to consider the difference in performance in (perhaps more realistic) noisier environments. We display the RMSE, MAE, interval length, and coverage by condition in Figure 4 and report the pooled results in Table 2.

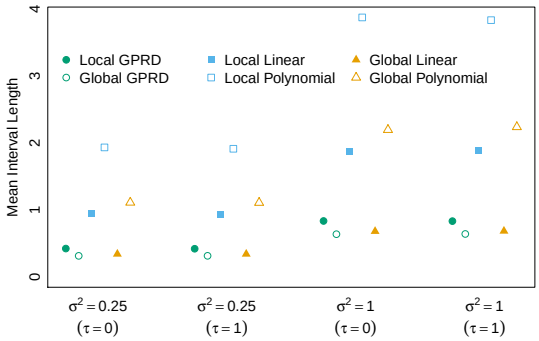
In this setting, the mean absolute error, root mean squared error, and confidence interval length are all more than *twice* as large for the local linear regression method than the piecewise GPRD method, and the global GPRD method edges out the piecewise GPRD



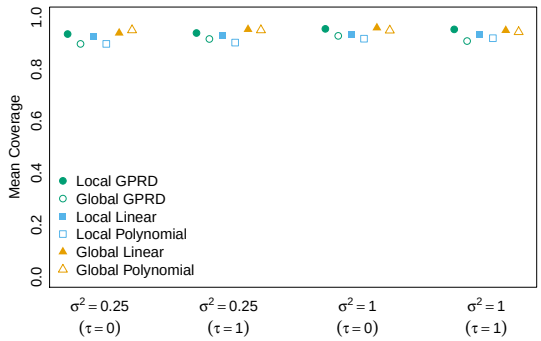
(a) Root mean squared error of treatment effect estimates.



(b) Mean absolute error of treatment effect estimates.



(c) Mean confidence interval length for treatment effect estimates.



(d) Mean 95% confidence interval coverage of true treatment effects.

Figure 4: Fit statistics for treatment effect estimates from the global and piecewise GPRD models (plotted in empty and filled green circles respectively), local linear and polynomial regression (plotted in filled and empty blue squares respectively), and global linear and polynomial regression (plotted in filled and empty yellow triangles).

Table 2: Fit statistics averaged across replicates and noise and effect sizes.

Estimator	MAE	RMSE	Mean CI Length	Mean CI Coverage
Global GPRD	0.111	0.149	0.476	0.913
piecewise GPRD	0.128	0.170	0.627	0.946
Local Linear	0.265	0.360	1.410	0.931
Local Polynomial	0.580	0.796	2.880	0.910
Global Linear	0.102	0.135	0.511	0.951
Global Polynomial	0.345	0.461	1.660	0.949

method further still. OLS, the global linear model, does very well, as may be expected when the simulation's DGP matches exactly the OLS assumptions. Remarkably, the GPRD methods perform similarly to the model whose assumptions match the DGP. In the common

scenario where noise is appreciable, the GPRD methods outperform existing local methods substantially and do no worse than when the researcher can correctly intuit the precise functional form of the mapping from forcing variable to outcomes.

3 Empirical Applications

We selected the following empirical applications based on three criteria. First, they are all published in top political science and economics journals over the past five years. Second, they all adhere to current best practices for RD studies, employing local polynomial estimators with automated bandwidth selectors and robust standard errors, and conducting extensive falsification and robustness tests. Third, the authors provide sufficient replication materials to reproduce their work. In short, we select these examples not because they are *bad* examples of applied RD, but because they are good examples of careful and rigorous empirical work. Nevertheless, each of these examples exhibits low power stemming from small or noisy datasets, suggesting that the treatment effects estimated using local linear approaches are likely to be exaggerated in some way.

3.1 The Radical Right and Party Manifestos

Abou-Chadi and Krause (2018) study how the presence of radical right parties influence the platforms of mainstream parties. Their causal identification strategy is based on electoral thresholds in parliamentary systems; typically it is required that a party clear some percentage of the total vote before gaining representation in parliament, where the particular threshold varies by country. The authors compare elections where radical right parties barely exceeded the threshold (gaining representation) and barely missed the threshold, observing how mainstream political parties respond.

The dependent variable is change in a measure of Cultural Protection in the party's man-

ifesto during the following election. These data are compiled by the Comparative Manifestos Project (Volkens, Pola Lehmann, Theres Matthieß, Nicolas Merz and Werner, 2015), and the outcome variable is a function of the difference between the number of favorable mentions of cultural diversity and encouragement of integration and cultural homogeneity in the party's platform (Lowe et al., 2011). In their paper, Abou-Chadi and Krause (2018) present estimates from a diverse range of specifications, which range from 3.1 to 4.9. Replicating these results using CCT bias-corrected standard errors and robust confidence intervals yields an estimate of 3.96, with 95% confidence interval $[1.7, 6.2]$.

To get a sense of the relative magnitude of this estimated effect, consider Figure 5. This plots the average value of the Cultural Protection score, by country, for each family of political party since 1980. Although the Cultural Protection score is noisy from election year to election year, averaging across years yields predictable patterns. Right-leaning parties tend to score higher on the measure than left-leaning parties, and Nationalist parties – where they exist – typically score 2 to 3 points higher than the average mainstream party.

In this context, an estimated effect size of 3.9 is enormous. If true, it suggests that not only do mainstream parties respond to Radical Right representation by moving their platforms to the right, but they do so in such a way that their rhetoric completely closes or even overtakes the average gap between mainstream and rightwing nationalist party positions.

We apply our GPRD methodology to this data and find the GPRD methods affirm the finding of a reliable positive effect, but a more *plausible* effect. Table 3 compares the effect sizes estimated by GPRD, local linear regression, and global polynomial regression. Figure 6 visualizes those estimates. We can see the local linear and global polynomial methods drag down the prediction line just to the left of the cutoff substantially, based on a relatively small number of observations in that area. The GPRD methods do not suffer from this weakness and therefore recover a more plausible, but still reliably positive, effect of 1.7 for the piecewise GPRD estimate and 1.9 for the global GPRD estimate.

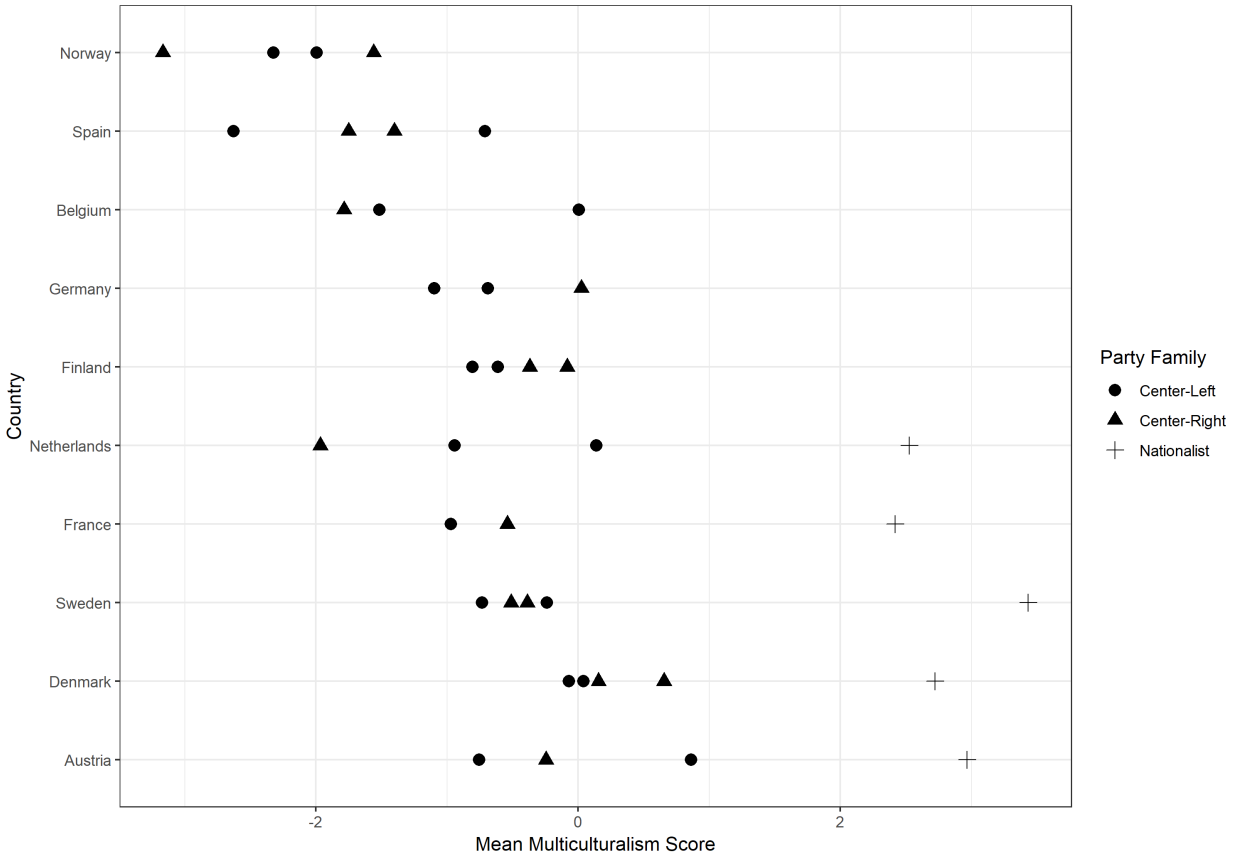


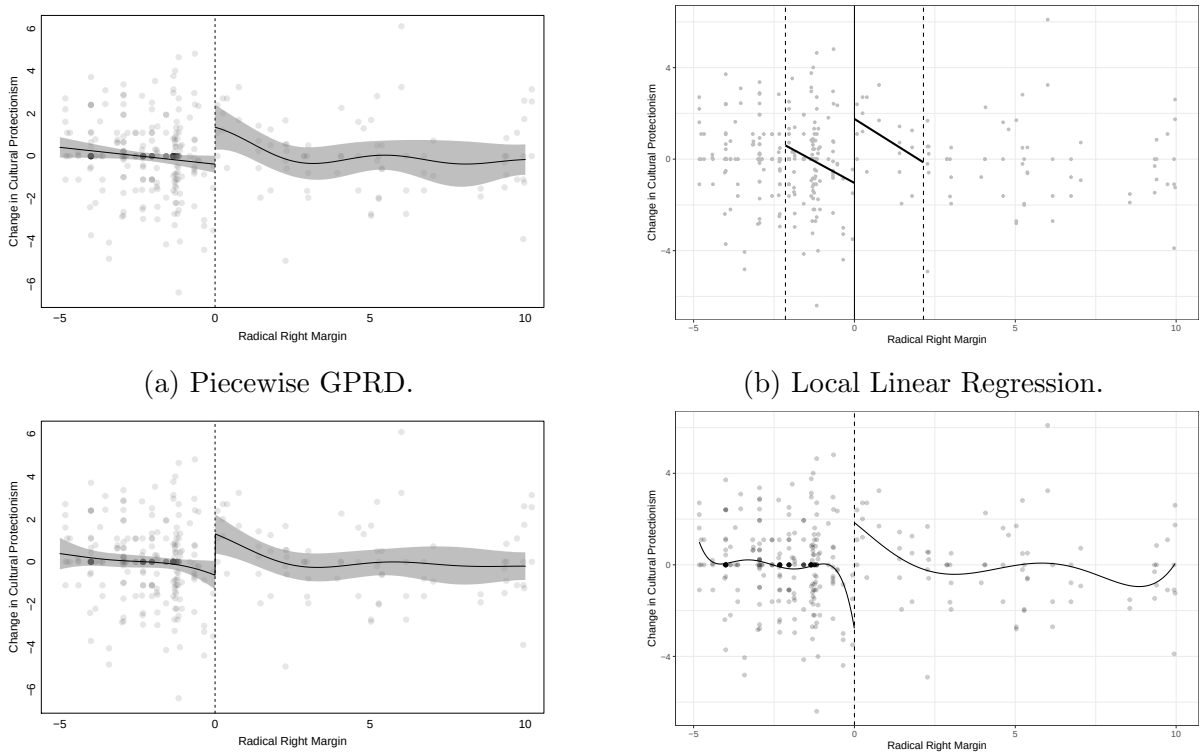
Figure 5: Mean Cultural Protection score by country and party family (all elections post-1980; Center-Left includes Social Democrats and Liberals, Center-Right includes Christian Democrats and Conservatives).

Table 3: Treatment effect comparison between the GPRD methods and local linear and global polynomial regression for the Abou-Chadi and Krause (2018) application.

Estimator	Treatment effect	95% CI
Global GPRD	1.735	[0.804, 3.074]
Piecewise GPRD	1.939	[0.603, 2.867]
Local Linear	3.972	[1.698, 6.238]
Global Polynomial	3.983	[1.707, 6.258]

3.2 Ethnic Diversity and Municipal Public Spending

The second empirical illustration comes from Beach and Jones (2017), who study the effect of diverse city council representation on municipal-level public goods spending. There is a large literature on this topic, dating back to Alesina, Baqir and Easterly (1999) and



(a) Piecewise GPRD.

(b) Local Linear Regression.

(c) Global GPRD.

(d) Global Polynomial Regression.

Figure 6: Fit comparison between the GPRD methods and local linear and global polynomial regression for the Abou-Chadi and Krause (2018) application.

continuing through Hopkins (2011) and Trounstine (2015), finding that in ethnically diverse and segregated cities, municipal governments devote less spending to public goods than in ethnically homogeneous cities. All of these studies rely on cross-sectional or longitudinal regression analysis of public finance data, and so lack a credible causal identification. Beach and Jones (2017) approach the problem through an RD analysis of city council elections in California. Because narrowly elected city councilors from a 'non-modal' ethnicity should increase the diversity of a city council, the discontinuity at the plurality margin allows for the causal identification of an effect of council diversity on public spending.

By regressing per capita public goods spending on the vote margin of 'non-modal' candidates for races where a non-modal candidate faced a modal candidate, Beach and Jones (2017) estimate a treatment effect of -0.31 on a log scale, implying that the election of a councilor that increases ethnic diversity causes a 31% drop in public goods spending.

To consider this estimate in context, a 31% cut in public goods spending amounts to roughly $463 per capita for the average city. This is equivalent to eliminating all spending on police (13% in the average city), fire protection (9%), and roads (9%). By comparison, similar studies suggest much more modest effects. In an RD analyis of Democratic vs. Republican mayoral candidates, de Benedictis-Kessner and Warshaw (2016) find that narrowly-elected Democratic mayors increase public spending by 5% on average. In their original study of US cities, Alesina, Baqir and Easterly (1999) find that ethnically diverse cities spend 6% to 9% lower shares on "productive" public goods than ethnically homogeneous cities.[6] All of these estimates are several multiples smaller than 30%.

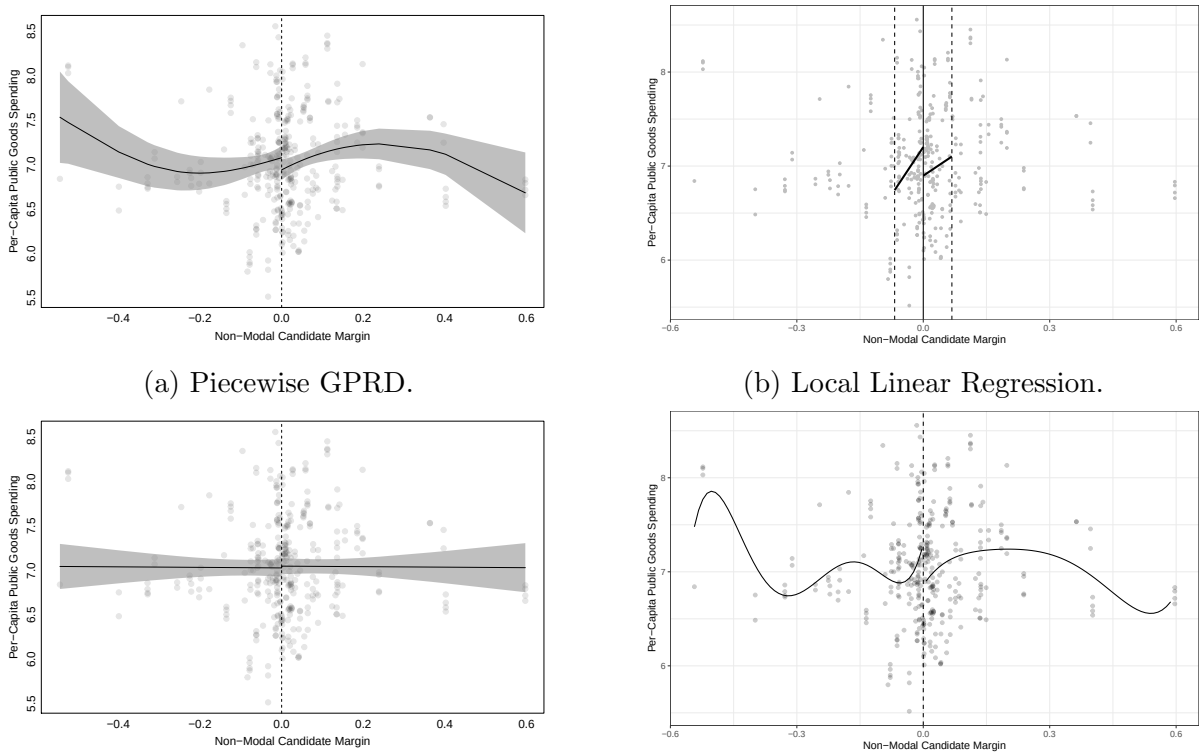
Even assuming that the effect of a non-modal city councilmember is comparable to that of a Democratic mayor,[7] the RD analysis would be too underpowered to reliably detect a more plausible effect. The minimum detectable effect size given the number of observations in this study is 0.32[8], and the RD plots in Figure 7 makes this problem clear: although a substantial fraction of the sample lies within the bandwidth, the variation in public goods spending is very large relative to any plausible effect size. A sufficiently high powered study to detect an effect size of 5% would require many times more observations within the neighborhood of the cutoff.

Given the Monte Carlo simulation evidence, it may be that the effect is entirely spurious and due to the noisy nature of the data; our simulation results in the context of noisy data give us hope that the GPRD estimators can uncover a more plausible effect if present, or signal the lack of evidence of a reliable effect. Table 4 and Figure 7 compares the GPRD methods with local linear and global polynomial regression. Both GPRD estimators fail to find a reliable effect while both traditional methods find reliable and implausibly large effects.

[6]This is the estimated effect of changing the Herfindahl ethnic fractionalization index from 0 to 1 (i.e. from perfect homogeneity to perfect heterogeneity).

[7]Not an unreasonable assumption, since in many municipalities the mayor does not wield strong executive power, and is essentially the chair of the city council.

[8]See Calonico et al. (2018) for derivation of power calculations in the RD context



(a) Piecewise GPRD.

(b) Local Linear Regression.

(c) Global GPRD.

(d) Global Polynomial Regression.

Figure 7: Fit comparison between the GPRD methods and local linear and global polynomial regression for the Beach and Jones (2017) application.

Table 4: Treatment effect comparison between the GPRD methods and local linear and global polynomial regression for the Beach and Jones (2017) application.

Estimator	Treatment effect	95% CI
Global GPRD	0.019	$[-0.115, 0.154]$
Piecewise GPRD	-0.137	$[-0.311, 0.038]$
Local Linear	-0.344	$[-0.595, -0.093]$
Global Polynomial	-0.438	$[-0.734, -0.142]$

4 Conclusion

In this paper, we have demonstrated how low-powered regression discontinuity analyses can yield misleading and exaggerated causal effect estimates, characterized by an implausibly large divergence in the slope of the conditional expectation function in the neighborhood of the cutoff. We introduce a novel RD estimator, Gaussian Process Regression Discontinuity (GPRD), to address these problems. GPRD performs well in simulations and provides more

plausible treatment effect estimates in empirical applications.

In future work we hope to extend GPRD in several directions. First, we plan to develop novel methods for optimizing the covariance kernel hyperparameters in the RD context, placing greater weight on observations closer to the cutoff when estimating the length scale and introducing a fully Bayesian MCMC approach. Second, we plan to expand GPRD to handle fuzzy RD designs and the inclusion of pre-treatment covariates, both natural extensions to the current framework. All of these will be made available for researchers in the forthcoming `R` package `gprd`.

References

Abou-Chadi, Tarik and Werner Krause. 2018. "The Causal Effect of Radical Right Success on Mainstream Parties' Policy Positions: A Regression Discontinuity Approach." *British Journal of Political Science* pp. 1–19.

Alesina, Alberto, Reza Baqir and William Easterly. 1999. "Public Goods and Ethnic Divisions." *The Quarterly Journal of Economics* 114(4):1243–1284.

Beach, Brian and Daniel B. Jones. 2017. "Gridlock: Ethnic Diversity in Government and the Provision of Public Goods." *American Economic Journal: Economic Policy* 9(1):112–136.

Branson, Zach, Maxime Rischard, Luke Bornn and Luke W. Miratrix. 2019. "A nonparametric Bayesian methodology for regression discontinuity designs." *Journal of Statistical Planning and Inferenc* 202:14–30.

Button, Katherine S., John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson and Marcus R. Munafò. 2013. "Power failure: why small sample size undermines the reliability of neuroscience." *Nature Reviews Neuroscience* 14(5):365–376.

Calonico, Sebastian, Matias D. Cattaneo, Max H. Farrell and Rocio Titiunik. 2018. "Re-

gression Discontinuity Designs Using Covariates." *Review of Economics and Statistics* pp. 1–30.

Calonico, Sebastian, Matias D. Cattaneo and Rocio Titiunik. 2014. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica* 82(6):2295–2326.

Cattaneo, Matias D., Brigham Frandsen and Rocío Titiunik. 2014. "Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate.".

de Benedictis-Kessner, Justin and Christopher Warshaw. 2016. "Mayoral Partisanship and Municipal Fiscal Policy." *The Journal of Politics* 78(4):1124–1138.

de la Cuesta, Brandon and Kosuke Imai. 2016. "Misunderstandings About the Regression Discontinuity Design in the Study of Close Elections." *Annual Review of Political Science* 19:375–396.

Ferreira, Fernando. 2010. "You can take it with you: Proposition 13 tax benefits, residential mobility, and willingness to pay for housing amenities." *Journal of Public Economics* 94(9-10):661–673.

Gelman, Andrew and Guido Imbens. 2017. "Why high-order polynomials should not be used in regression discontinuity designs." *Journal of Business & Economic Statistics* .

Gelman, Andrew and John Carlin. 2014. "Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors." *Perspectives on Psychological Science* pp. 1–11.

Hahn, Jinyong, Petra Todd and Wilbert Van Der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica* 69(1):201–209.

Hopkins, Daniel J. 2011. "The limited local impacts of ethnic and racial diversity." *American Politics Research* 39(2):344–379.

Imbens, Guido and Karthik Kalyanaraman. 2012. "Optimal bandwidth choice for the regression discontinuity estimator." *The Review of Economic Studies* 79(3):933–959.

Klašnja, Marko and Rocio Titiunik. 2017. "The Incumbency Curse: Weak Parties, Term Limits, and Unfulfilled Accountability." *The American Political Science Review* 111(1):129–148.

Lee, David S. 2008. "Randomized experiments from non-random selection in US House elections." *Journal of Econometrics* 142(2):675–697.

Lowe, Will, Kenneth Benoit, Slava Mikhaylov and Michael Laver. 2011. "Scaling policy preferences from coded political texts." *Legislative Studies Quarterly* 36(1):123–155.

Ludwig, J and D L Miller. 2007. "Does Head Start improve children's life chances?" *The Quarterly Journal of Economics* 122:159–208.

Rasmussen, Carl Edward and Christopher K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. MIT Press.

Trounstine, Jessica. 2015. "Segregation and Inequality in Public Goods." *American Journal of Political Science* 60(3):709–725.

Volkens, Andrea, Sven Regel Pola Lehmann, Theres Matthieß, Nicolas Merz and Annika Werner. 2015. *The Manifesto Data Collection: Manifesto Project (MRG/CMP/MARPOR). Version 2015a.* Berlin: .