Gaussian Process Regression Discontinuity

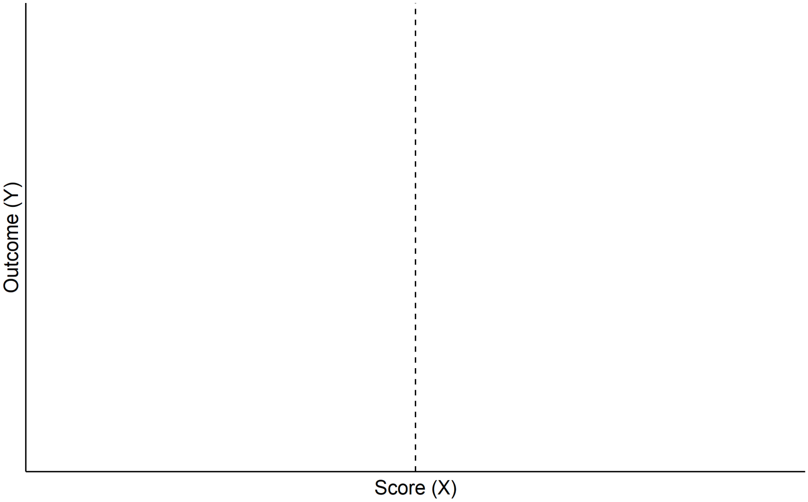
Joseph T. Ornstein   JBrandon Duck-Mayr

Washington University in St. Louis

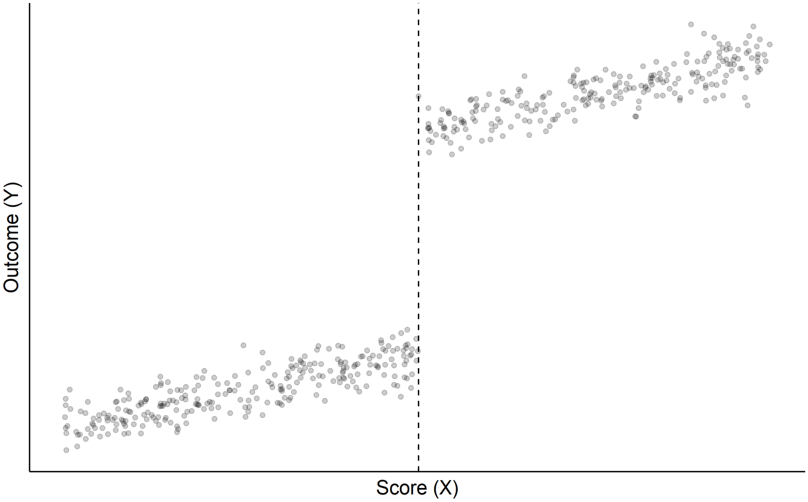
February 21, 2020

The Regression Discontinuity (RD) Design

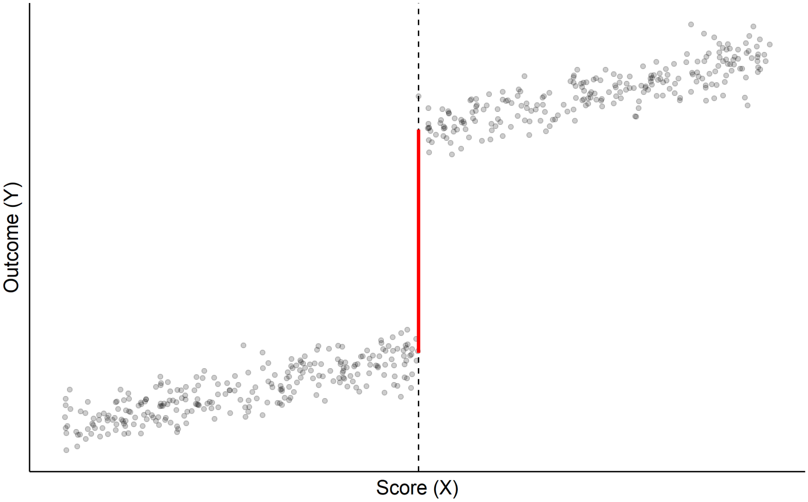
The Regression Discontinuity (RD) Design



The Regression Discontinuity (RD) Design



The Regression Discontinuity (RD) Design

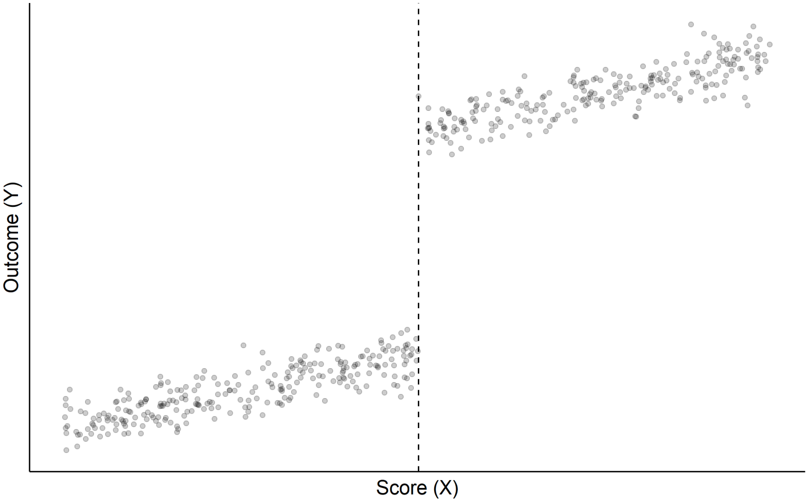


Estimation

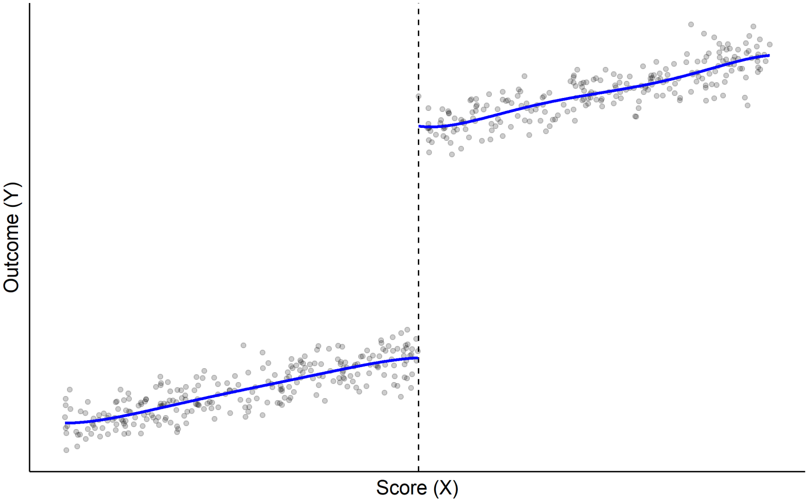
Estimation

1. Global Parametric RD

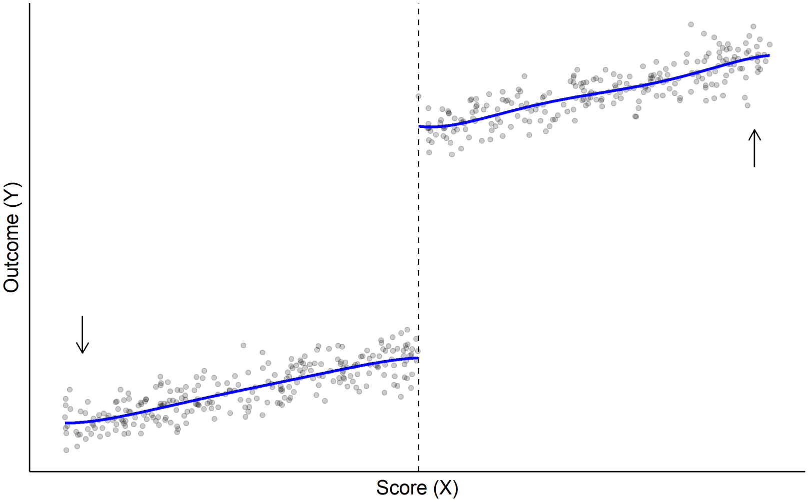
Global Parametric RD



Global Parametric RD



Global Parametric RD



Estimation

1. Global Parametric RD

Estimation

1. Global Parametric RD
   - Specification Bias

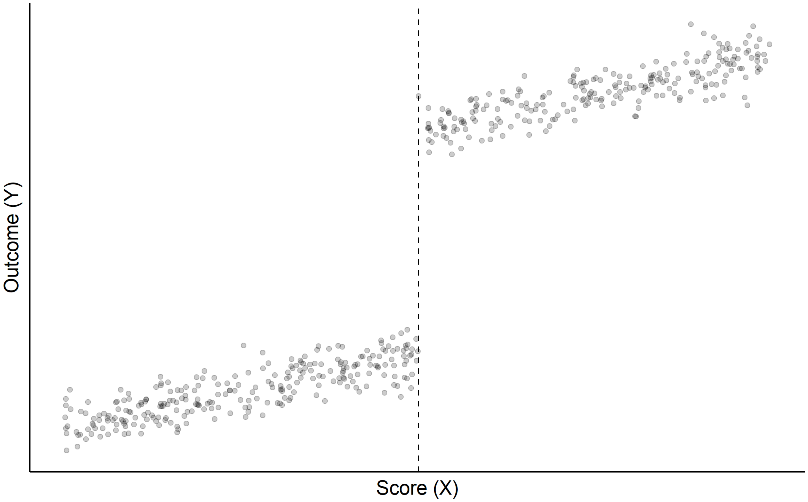
Estimation

1. Global Parametric RD
    - Specification Bias
    - Overfits to observations far from the cutoff

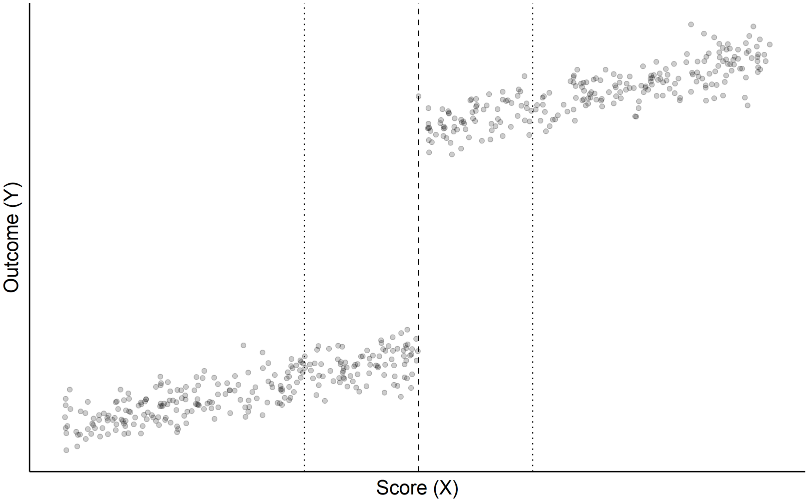
Estimation

1. Global Parametric RD
    - Specification Bias
    - Overfits to observations far from the cutoff
2. Local Nonparametric RD

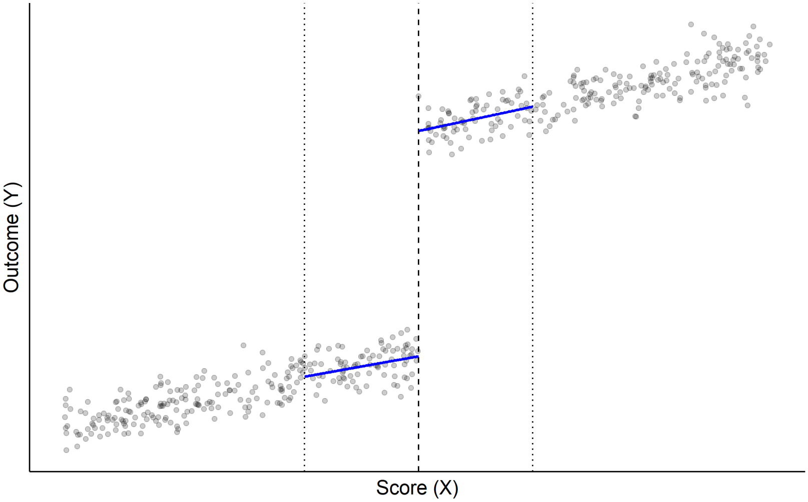
Local Nonparametric RD



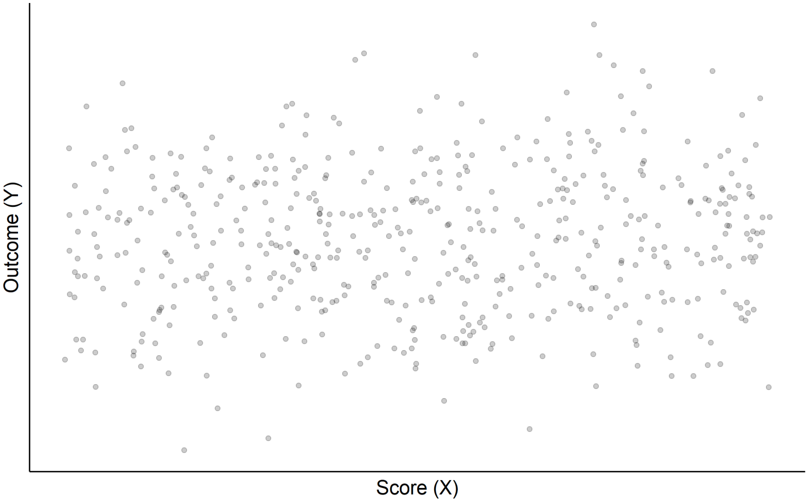
Local Nonparametric RD



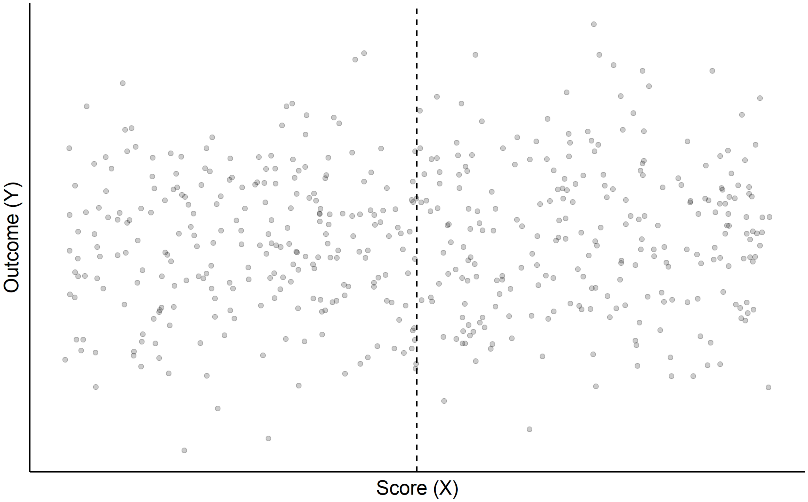
Local Nonparametric RD



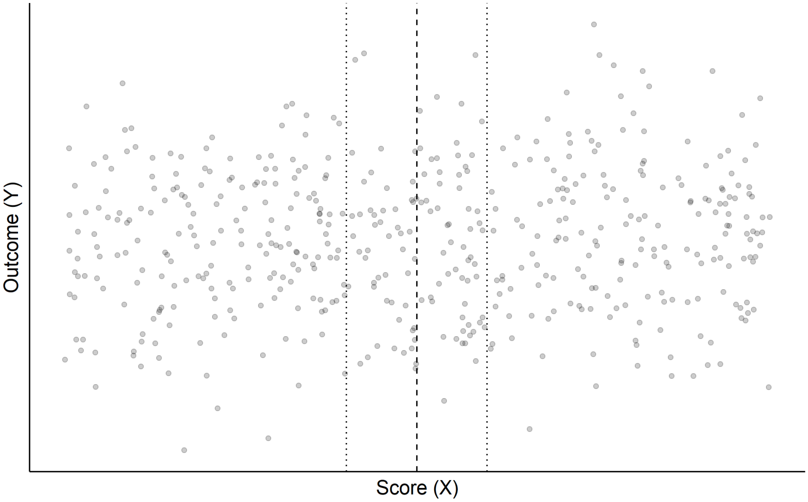
Low power is problematic in practice.



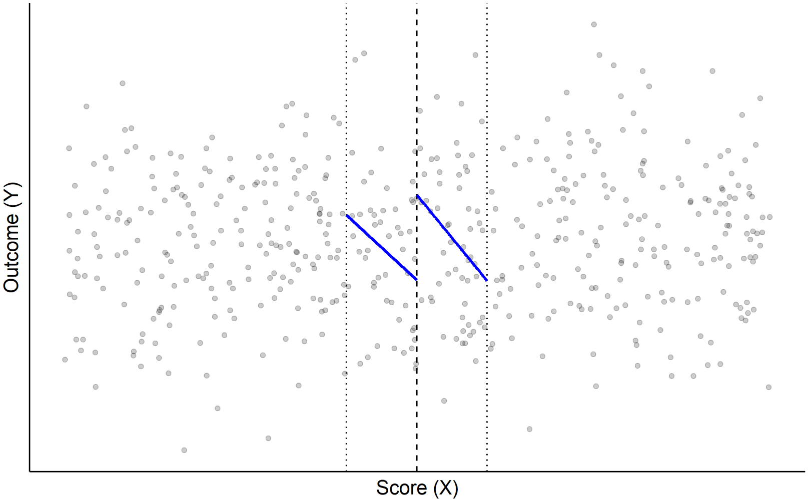
Low power is problematic in practice.



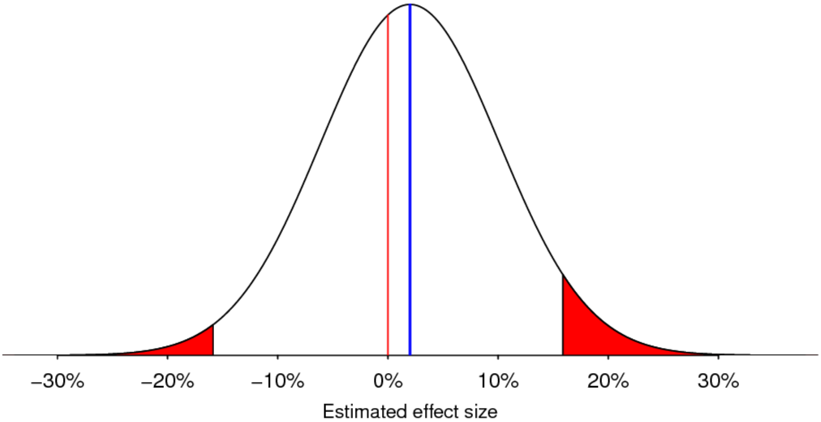
Low power is problematic in practice.



Low power is problematic in practice.



Replication Crisis (Type M Error)



Low power + statistical significance filter → Exaggerated Claims

Estimation

1. Global Parametric RD
   - Specification Bias
   - Overfits to observations far from the cutoff
2. Local Nonparametric RD

Estimation

1. Global Parametric RD
   - Specification Bias
   - Overfits to observations far from the cutoff
2. Local Nonparametric RD
   - Lower bias at the cost of ignoring observations, higher variance

Estimation

1. Global Parametric RD
   - Specification Bias
   - Overfits to observations far from the cutoff
2. Local Nonparametric RD
   - Lower bias at the cost of ignoring observations, higher variance
   - When sample size small, published effect estimates exaggerated

Estimation

1. Global Parametric RD
    - Specification Bias
    - Overfits to observations far from the cutoff
2. Local Nonparametric RD
    - Lower bias at the cost of ignoring observations, higher variance
    - When sample size small, published effect estimates exaggerated
3. **Gaussian Process RD**

Estimation

1. Global Parametric RD
    - Specification Bias
    - Overfits to observations far from the cutoff
2. Local Nonparametric RD
    - Lower bias at the cost of ignoring observations, higher variance
    - When sample size small, published effect estimates exaggerated
3. **Gaussian Process RD**
    - Overcomes these disadvantages in a principled way

Gaussian Process Regression

- We propose and RD estimator based on Gaussian process (GP) regression
- GP regression can be viewed as a simple but flexible extension of Bayesian linear regression
- GP regression helps us avoid strong assumptions about the function mapping the forcing variable x to the outcomes y
- This helps us estimate function values from the left and the right without resorting to local strategies

Gaussian Process Regression

The basic setting is observing inputs x and noisy outputs y that are a function of x.

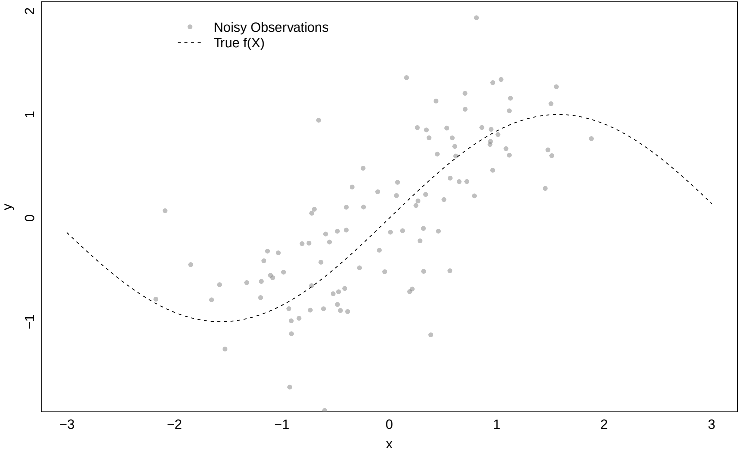


Figure: Learning the Mapping from x to y with a GP Prior

Gaussian Process Regression

But, the problem is we may not know the functional form of $f : x \to y$.

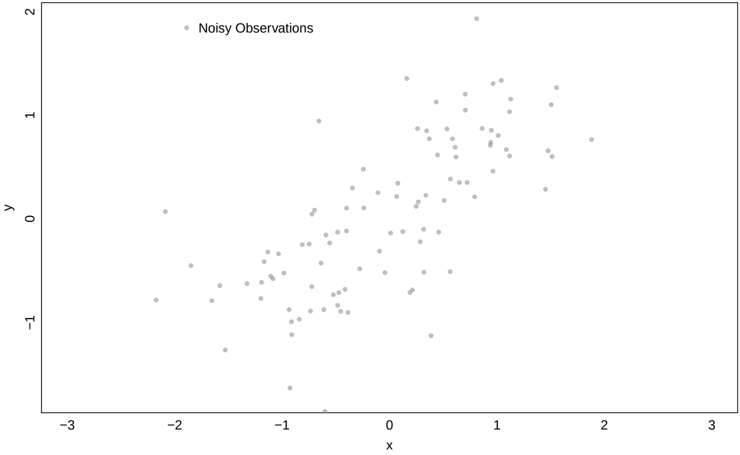


Figure: Learning the Mapping from x to y with a GP Prior

Gaussian Process Regression

- So, we put a GP prior over f.
- Technically, a GP is an infinite-dimensional generalization of the normal distribution.
- Theoretically, in our case, it is a *distribution over functions*.
- Practically, it's fancy way of assuming outputs are distributed normally given the inputs, but crucially the covariance between outputs are a function of the inputs.

Gaussian Process Regression

So, we put a GP prior over f (with a Gaussian likelihood)

$$\begin{aligned} \mathbf{y} &= f(\mathbf{x}) + \varepsilon, \\ f(\mathbf{x}) &\sim \mathcal{GP}\left(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x})\right), \\ \varepsilon &\sim \mathcal{N}(0, \sigma_y^2), \end{aligned}$$

where $m(\mathbf{x})$ and K are functions of the inputs.
In this example, we show a common and simple case: $m(\mathbf{x}) = \mathbf{0}$, and the i, j element of the covariance matrix is given by

$$K(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp\left(-0.5 \frac{(\mathbf{x}_i - \mathbf{x}_j)^2}{\ell^2}\right)$$

Gaussian Process Regression

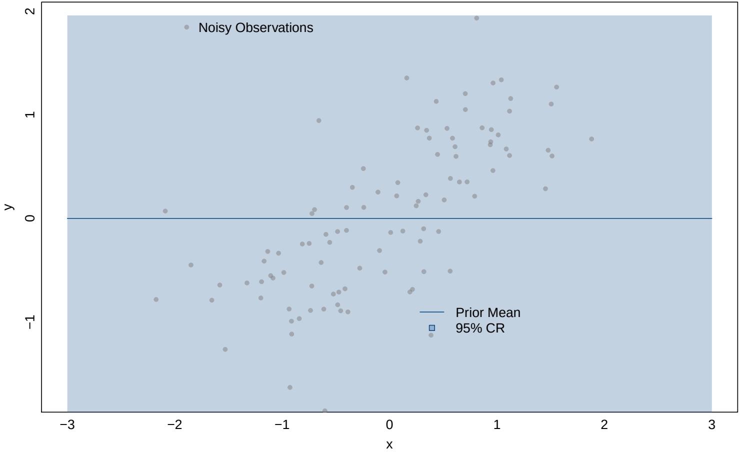


Figure: Learning the Mapping from x to y with a GP Prior

Gaussian Process Regression

The posterior over f is given using well-known Gaussian identities

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K + \sigma_y^2 I & K_*^T \\ K_* & K_{**} \end{bmatrix}\right),$$

$$K = K(\mathbf{x}, \mathbf{x}),$$

$$K_* = K(\mathbf{x}_*, \mathbf{x}),$$

$$K_{**} = K(\mathbf{x}_*, \mathbf{x}_*),$$

$$\mathbf{f}_* \mid \mathbf{x}, \mathbf{y}, \mathbf{x}_* \sim \mathcal{N}\left(K_* \left[K + \sigma_y^2 I\right]^{-1} \mathbf{y}, K_{**} - K_* \left[K + \sigma_y^2 I\right]^{-1} K_*^T\right).$$

Gaussian Process Regression

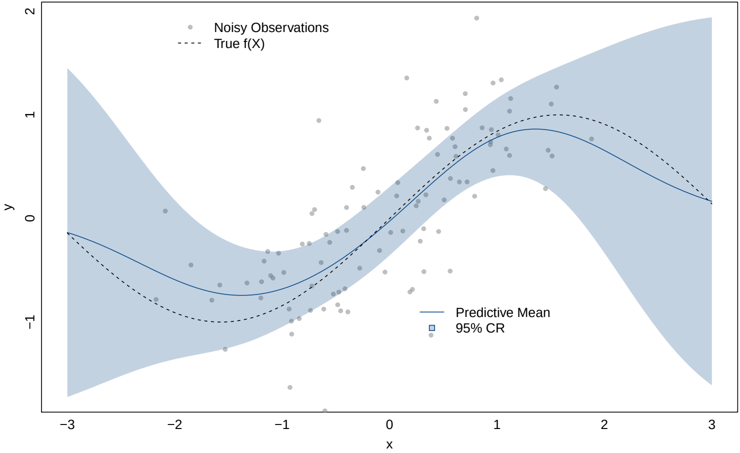


Figure: Learning the Mapping from x to y with a GP Prior

Gaussian Process Regression for Regression Discontinuity Designs

- Two methods to estimate treatment effects in RD designs using GP regression.

Gaussian Process Regression for Regression Discontinuity Designs

- Two methods to estimate treatment effects in RD designs using GP regression.
- First is the *global GPRD estimator*:

Gaussian Process Regression for Regression Discontinuity Designs

- Two methods to estimate treatment effects in RD designs using GP regression.
- First is the *global GPRD estimator*:
  - We fit one GP regression to all the data, with a dummy for treatment

Gaussian Process Regression for Regression Discontinuity Designs

- Two methods to estimate treatment effects in RD designs using GP regression.
- First is the *global GPRD estimator*:
  - We fit one GP regression to all the data, with a dummy for treatment
  - Then the treatment effect is the difference in predictions when x equals the cutoff and the dummy is 1 and 0

Gaussian Process Regression for Regression Discontinuity Designs

- Two methods to estimate treatment effects in RD designs using GP regression.
- First is the *global GPRD estimator*:
  - We fit one GP regression to all the data, with a dummy for treatment
  - Then the treatment effect is the difference in predictions when x equals the cutoff and the dummy is 1 and 0
- Second is the *piecewise GPRD estimator*:

Gaussian Process Regression for Regression Discontinuity Designs

- Two methods to estimate treatment effects in RD designs using GP regression.
- First is the *global GPRD estimator*:
  - We fit one GP regression to all the data, with a dummy for treatment
  - Then the treatment effect is the difference in predictions when x equals the cutoff and the dummy is 1 and 0
- Second is the *piecewise GPRD estimator*:
  - We fit two GP regressions, one to each side of the cutoff

Gaussian Process Regression for Regression Discontinuity Designs

- Two methods to estimate treatment effects in RD designs using GP regression.
- First is the *global GPRD estimator*:
  - We fit one GP regression to all the data, with a dummy for treatment
  - Then the treatment effect is the difference in predictions when x equals the cutoff and the dummy is 1 and 0
- Second is the *piecewise GPRD estimator*:
  - We fit two GP regressions, one to each side of the cutoff
  - Then the treatment effect is the difference in the two GPs' predictions when x equals the cutoff

Global GPRD Estimator

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

$$p(f) = \mathcal{GP}(\mathbf{X}\beta, K(\mathbf{X})),$$
$$\mathbf{X} = \left[\mathbf{1}|\mathbf{x}|D\right],$$

where K is the squared exponential automatic relevance determination covariance function

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

Global GPRD Estimator

So we are interested in the treatment effect

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

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

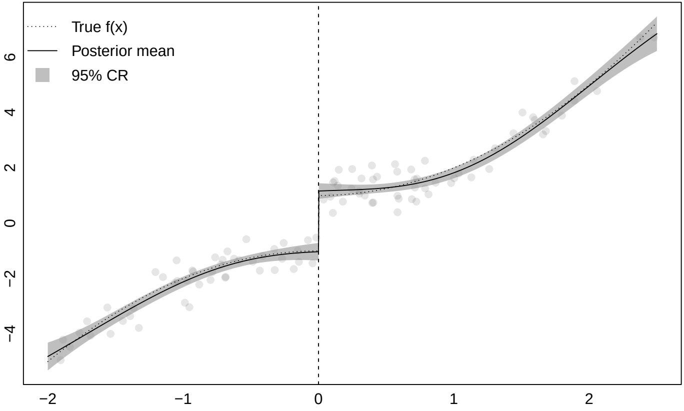
$$\begin{aligned}\tau_{GPRD-G} &\sim \mathcal{N}\left(\mu_*, \Sigma_*\right), \\ \mu_* &= \bar{f}\left(\begin{bmatrix}0 & 1\end{bmatrix}\right) - \bar{f}\left(\begin{bmatrix}0 & 0\end{bmatrix}\right), \\ \Sigma_* &= \operatorname{cov}\left(f\left(\begin{bmatrix}0 & 1\end{bmatrix}\right)\right) + \operatorname{cov}\left(f\left(\begin{bmatrix}0 & 0\end{bmatrix}\right)\right).\end{aligned}$$

Global GPRD Estimator

Here's an example, where

$$x \sim \mathcal{N}(0,1)$$
$$f(x) = \begin{cases} x^2 + 1 & \text{if } x > 0 \\ -x^2 - 1 & \text{otherwise,} \end{cases}$$
$$y = f(x) + \varepsilon,$$
$$\varepsilon \sim \mathcal{N}(0, 0.5^2).$$

Global GPRD Estimator



True effect: 2; Estimate: 2.19; 95% CI: $[1.77, 2.62]$

Piecewise GPRD Estimator

For the piecewise GPRD estimator, we place GP priors on $f_+(x)$ and $f_-(x)$,

$$p(f_+) = \mathcal{GP}(\mathbf{X}_+\beta_+, K(x_+)),$$
$$p(f_-) = \mathcal{GP}(\mathbf{X}_-\beta_-, K(x_-)),$$
$$\mathbf{X} = \left[\mathbf{1}|\mathbf{x}\right],$$

where K is the isometric squared exponential covariance function

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

Piecewise GPRD Estimator

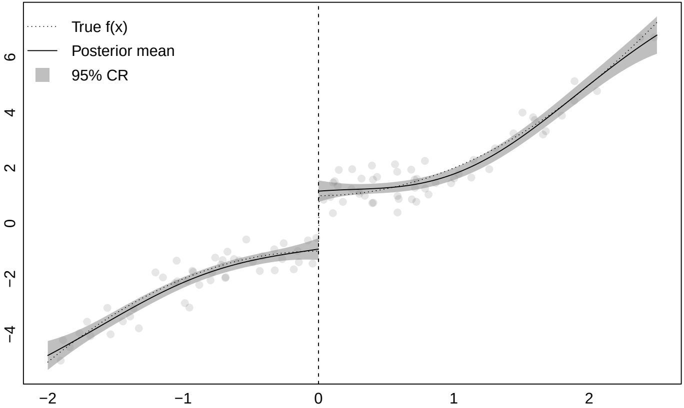
So we are interested in the treatment effect

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

which is distributed

$$\tau_{GPRD-P} \sim \mathcal{N}(\bar{f}_+(0) - \bar{f}_-(0), \mathrm{cov}(f_+(0)) + \mathrm{cov}(f_-(0))).$$

Piecewise GPRD Estimator



True effect: 2; Estimate: 2.09; 95% CI: $[1.54, 2.64]$

Hyperparameter Selection

- Choice of hyperparameters—particularly the length scale (ℓ)—likely to affect our estimates.

Hyperparameter Selection

- Choice of hyperparameters—particularly the length scale (ℓ)—likely to affect our estimates.
  - Length scale in GPRD performs similar role as bandwidth in local linear approach.

Hyperparameter Selection

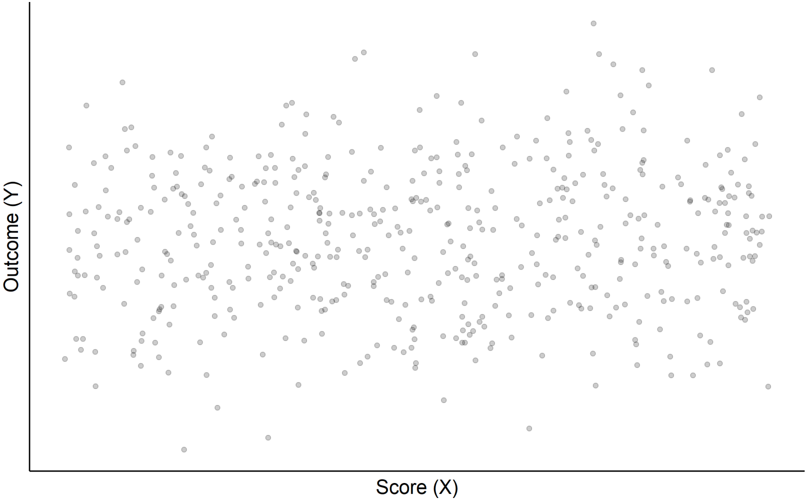
- Choice of hyperparameters—particularly the length scale (ℓ)—likely to affect our estimates.
  - Length scale in GPRD performs similar role as bandwidth in local linear approach.
- Commonly in GP regression, chosen by maximizing marginal log likelihood

Hyperparameter Selection

- Choice of hyperparameters—particularly the length scale (ℓ)—likely to affect our estimates.
  - Length scale in GPRD performs similar role as bandwidth in local linear approach.
- Commonly in GP regression, chosen by maximizing marginal log likelihood
- In simulations and applications shown here, covariance hypers chosen via MLE, prior placed over β then exact inference performed for τ

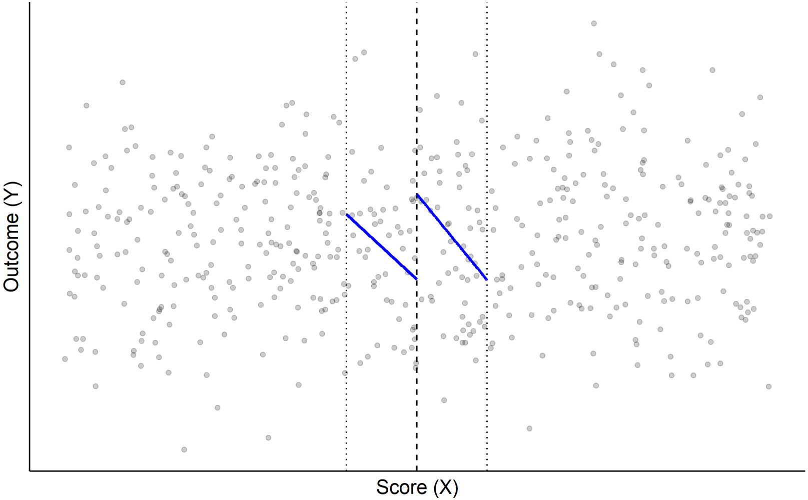
Gaussian Process RD

Returning to our earlier example...



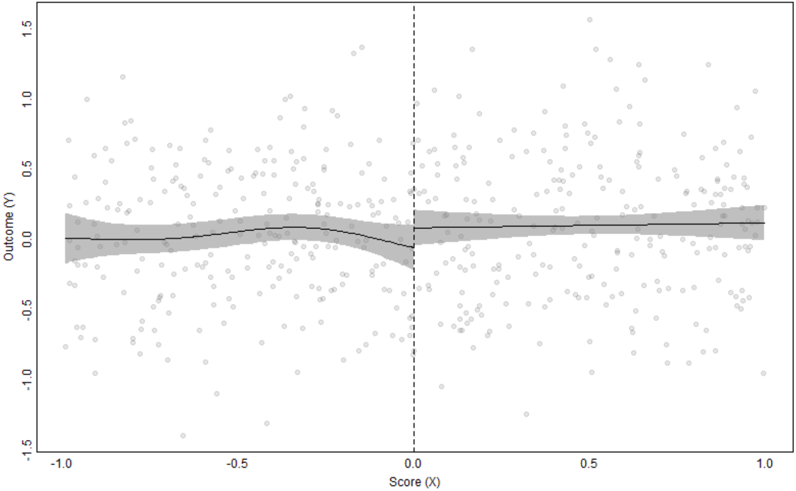
Gaussian Process RD

Returning to our earlier example...



Gaussian Process RD

Now with GPRD...

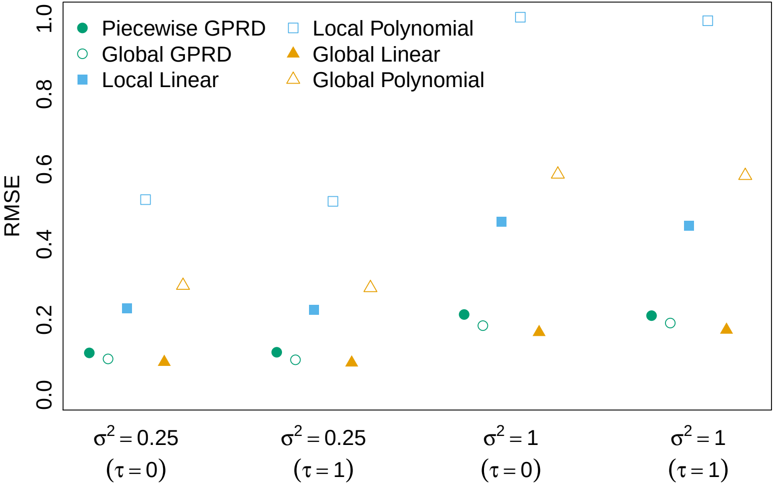


Simulations

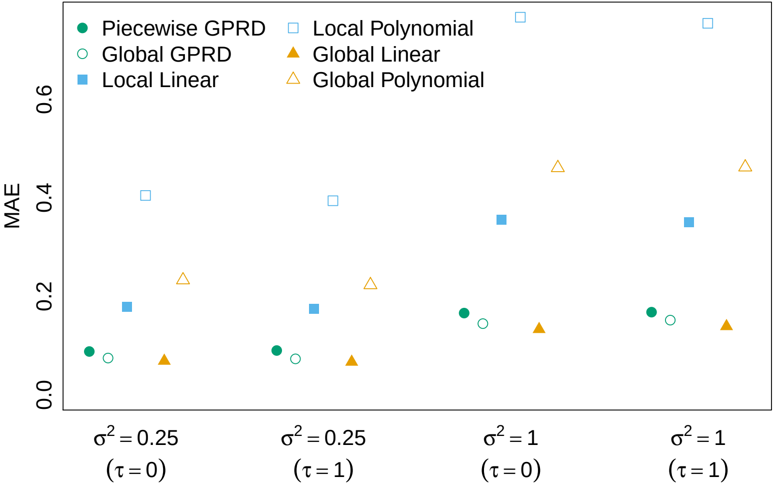
$$\begin{aligned} x &= 2z - 1, \\ z &\sim \mathcal{B}(2,4), \\ f(x) &= x + \tau I(x > 0), \\ y &= f(x) + \varepsilon, \\ \varepsilon &\sim \mathcal{N}(0, \sigma^2), \end{aligned}$$

for $\tau = 0$ and $\tau = 1$, and for $\sigma = 0.5$ and $\sigma = 1$.

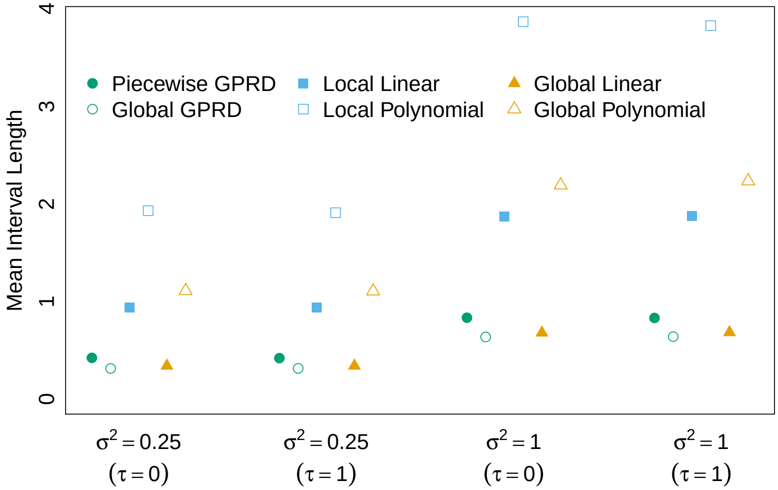
Simulation Results



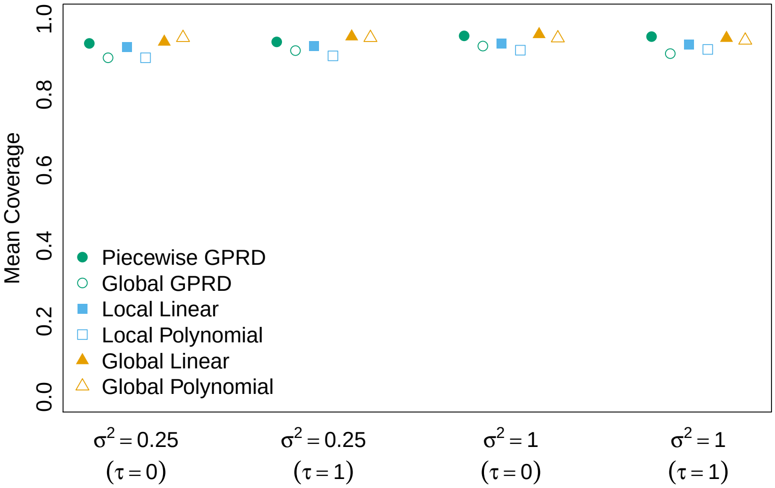
Simulation Results



Simulation Results



Simulation Results

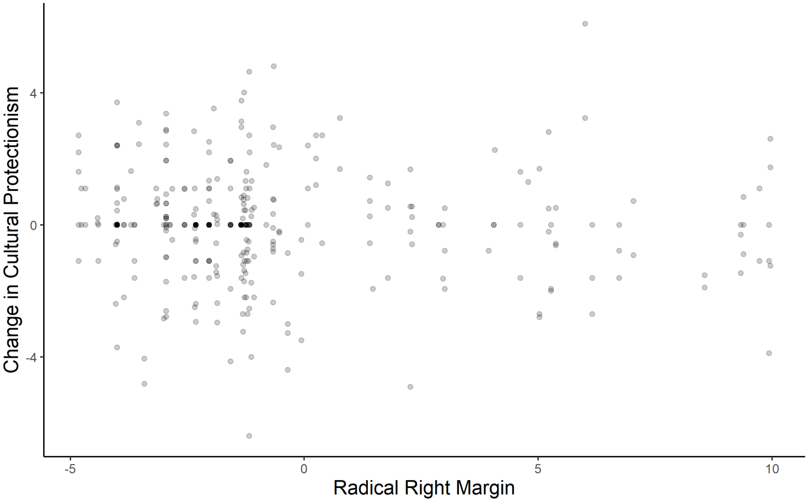


Empirical Applications

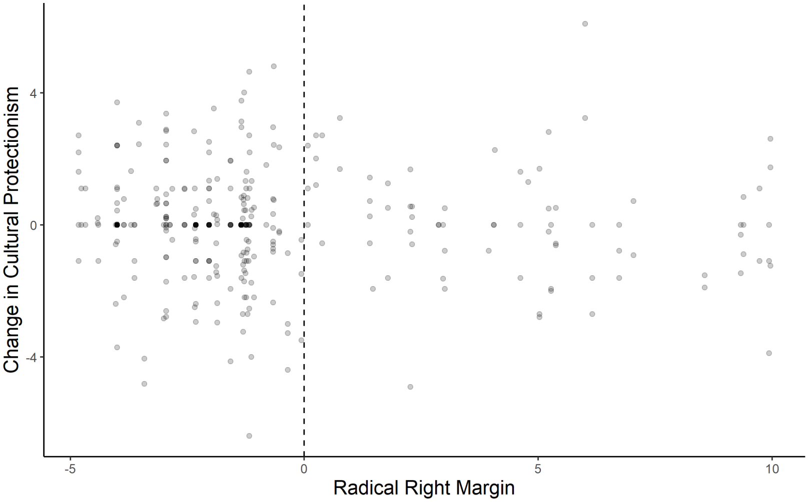
Two Published Examples:

1. The Radical Right and Mainstream Party Platforms (Abou-Chadi & Krause, 2018)
2. Ethnic Diversity on City Councils and Municipal Finance (Beach & Jones, 2017)

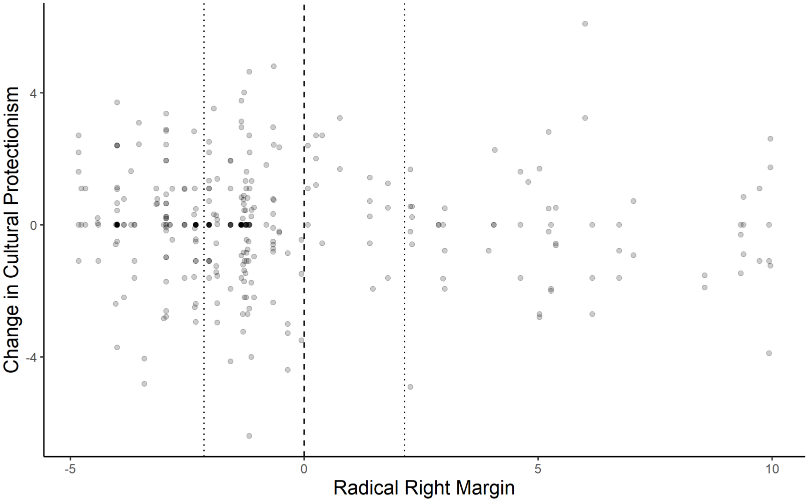
The Radical Right in Parliament



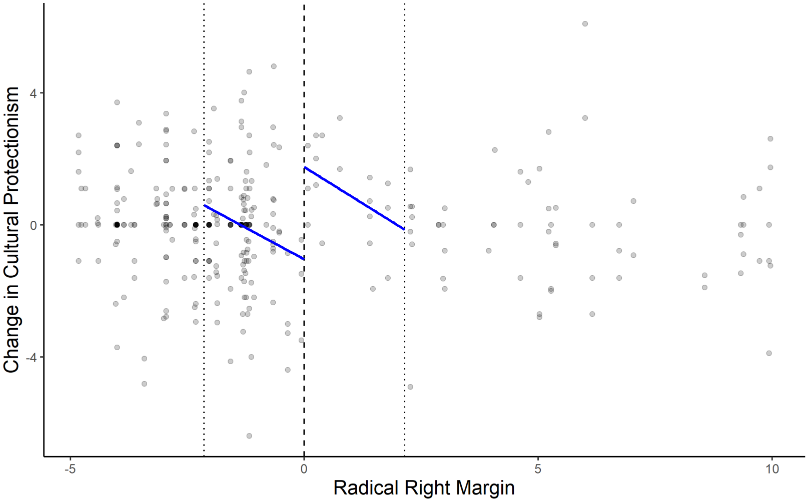
The Radical Right in Parliament



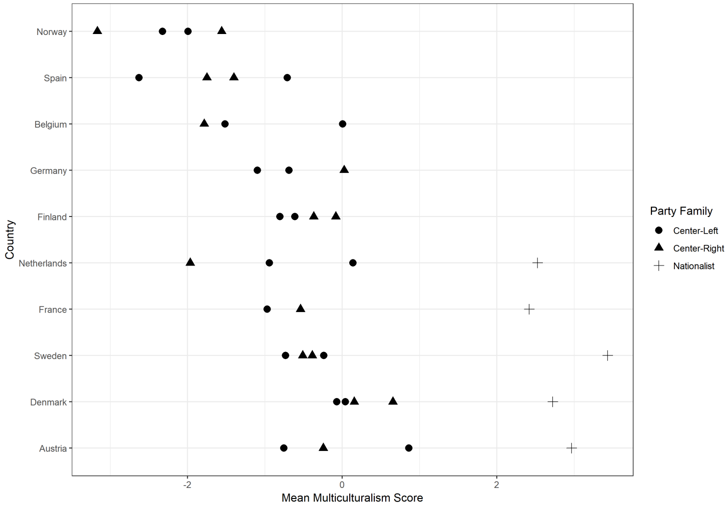
The Radical Right in Parliament



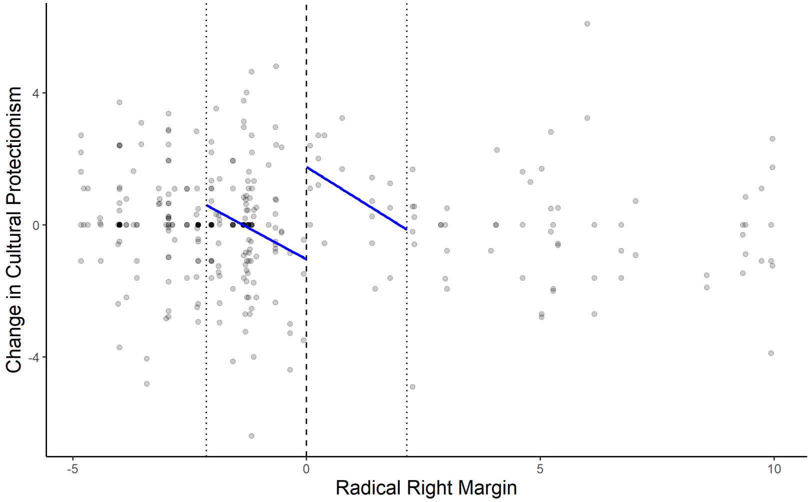
The Radical Right in Parliament



The Radical Right in Parliament

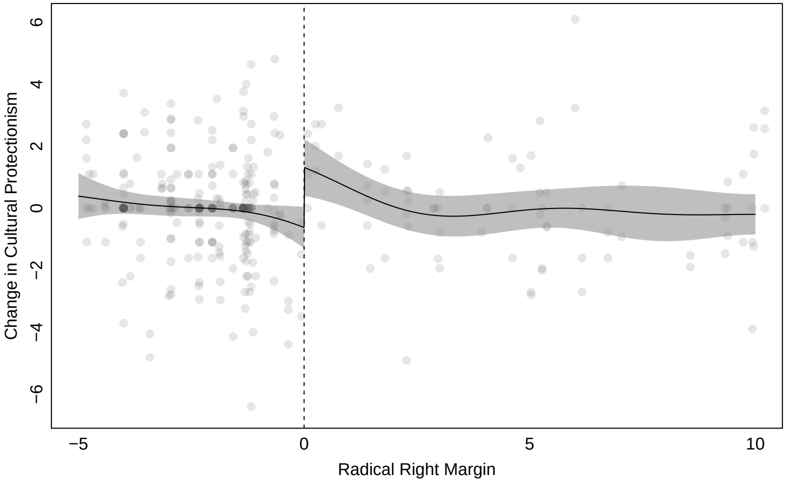


The Radical Right in Parliament



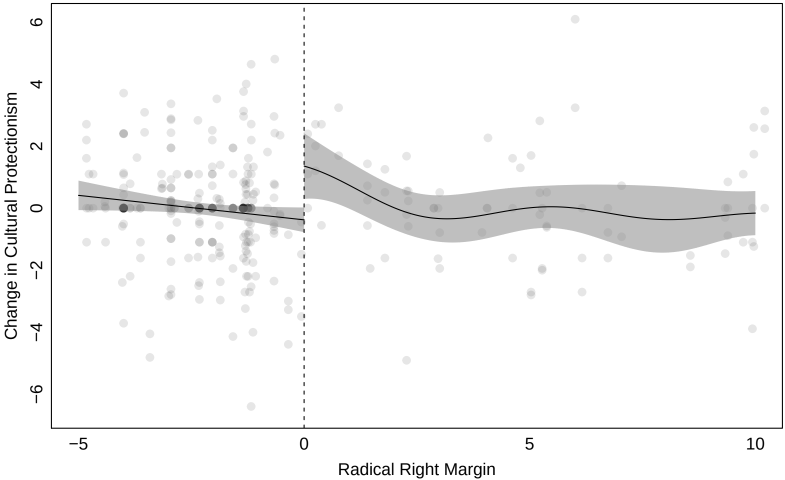
The Radical Right in Parliament

Global GPRD ($\hat{\tau} = 1.7$, 95% CI: $[0.8, 3.1]$)

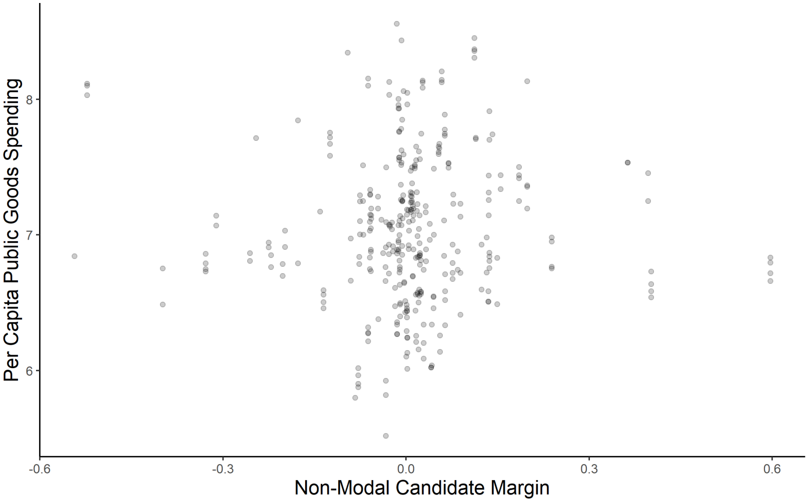


The Radical Right in Parliament

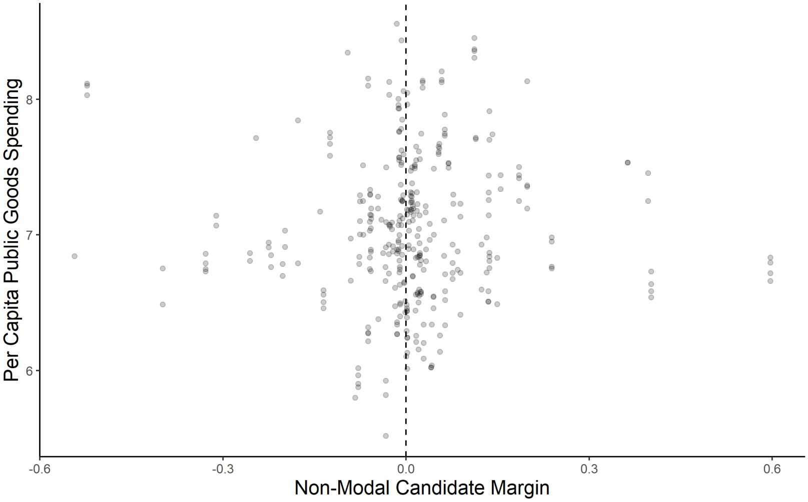
Piecewise GPRD ($\hat{\tau} = 1.9$, 95% CI: $[0.6, 2.8]$)



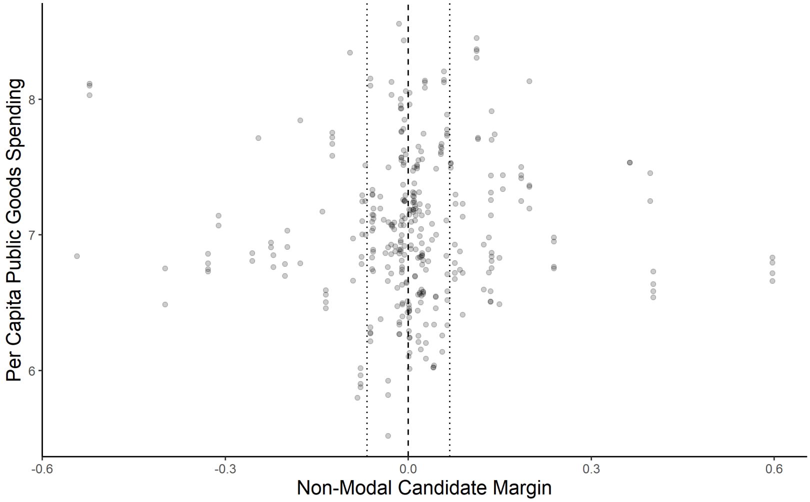
Ethnic Diversity on City Councils



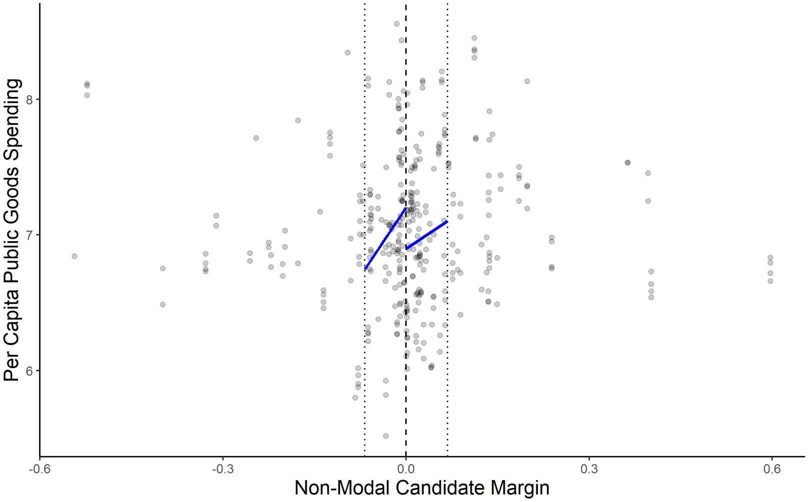
Ethnic Diversity on City Councils



Ethnic Diversity on City Councils

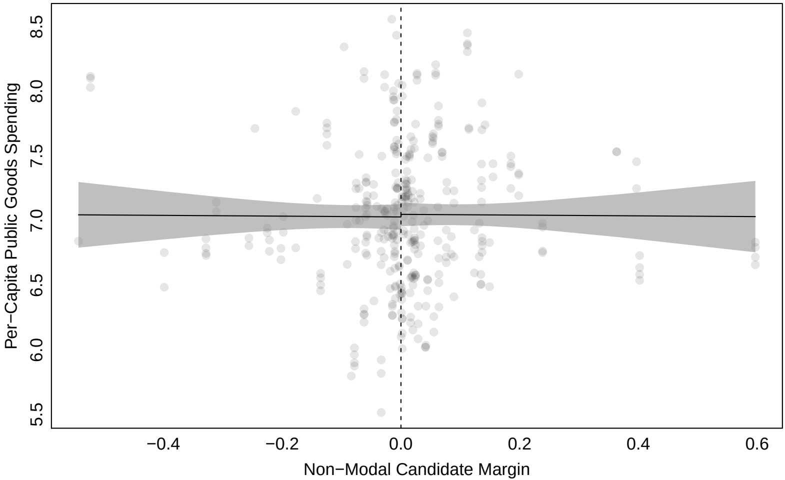


Ethnic Diversity on City Councils



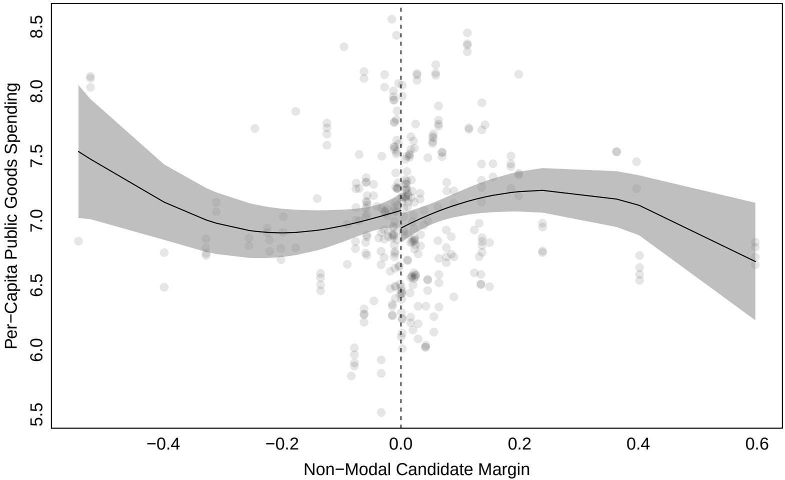
Ethnic Diversity on City Councils

Global GPRD ($\hat{\tau} = 0.02$, 95% CI: $[-0.115, 0.154]$)



Ethnic Diversity on City Councils

Piecewise GPRD ($\hat{\tau} = -0.13$, 95% CI: $[-0.311, 0.038]$)



Conclusion

Conclusion

- GPRD overcomes several disadvantages of current approaches

Conclusion

- GPRD overcomes several disadvantages of current approaches
- Performs well in simulation, particularly for noisy or low powered datasets

Conclusion

- GPRD overcomes several disadvantages of current approaches
- Performs well in simulation, particularly for noisy or low powered datasets
- Provides more plausible estimates in empirical applications

Conclusion

- GPRD overcomes several disadvantages of current approaches
- Performs well in simulation, particularly for noisy or low powered datasets
- Provides more plausible estimates in empirical applications
- Research in progress:

Conclusion

- GPRD overcomes several disadvantages of current approaches
- Performs well in simulation, particularly for noisy or low powered datasets
- Provides more plausible estimates in empirical applications
- Research in progress:
  - Alternative hyperparameter optimization approaches

Conclusion

- GPRD overcomes several disadvantages of current approaches
- Performs well in simulation, particularly for noisy or low powered datasets
- Provides more plausible estimates in empirical applications
- Research in progress:
  - Alternative hyperparameter optimization approaches
  - Extend to include pre-treatment covariates, fuzzy designs, multiple cutoffs

Conclusion

- GPRD overcomes several disadvantages of current approaches
- Performs well in simulation, particularly for noisy or low powered datasets
- Provides more plausible estimates in empirical applications
- Research in progress:
  - Alternative hyperparameter optimization approaches
  - Extend to include pre-treatment covariates, fuzzy designs, multiple cutoffs
- `R` package in development (`gprd`)