Inference at the data's edge: Characterizing and including counterfactual uncertainty with Gaussian Processes

Soonhong Cho, Doeun Kim, Chad Hazlett

Department of Political Science
Department of Statistics and Data Science
University of California, Los Angeles

July 17, 2024

The problem

We rely on models to fit things like $\mathbb{E}[Y(d)|X]$, but use their predictions at values of X that may be at or near the edge of the data.

- Different models may fit the observed data similarly well, but once one is chosen our inferences at new points ignore that.
- Can be exacerbated by more flexible models

The problem

We rely on models to fit things like $\mathbb{E}[Y(d)|X]$, but use their predictions at values of X that may be at or near the edge of the data.

- Different models may fit the observed data similarly well, but once one is chosen our inferences at new points ignore that.
- Can be exacerbated by more flexible models

The "dangers of extreme counterfactuals" are not capture by uncertainty estimates in conventional inference, e.g. for

- adjusted comparisons with poor overlap
- interrupted time-series (ITS)
- regression discontinuity (RD)

The problem

We rely on models to fit things like $\mathbb{E}[Y(d)|X]$, but use their predictions at values of X that may be at or near the edge of the data.

- Different models may fit the observed data similarly well, but once one is chosen our inferences at new points ignore that.
- Can be exacerbated by more flexible models

The "dangers of extreme counterfactuals" are not capture by uncertainty estimates in conventional inference, e.g. for

- adjusted comparisons with poor overlap
- interrupted time-series (ITS)
- regression discontinuity (RD)

We would like an approach that can, under reasonable assumptions, characterize (counterfactual) uncertainty at or beyond edge of data and include it in our inferences.

GP intuition: Ignorance and neighbors

If I tell you Y^* has mean 0 and variance σ^2, all you can believe is.

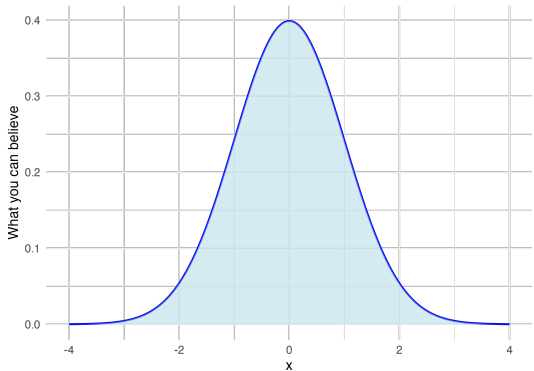


Figure: Total ignorance

GP intuition: Ignorance and neighbors

Now I tell you we see $Y_{obs} = -1$, and $cor(Y^*, Y_{obs}) = .8$. Then you believe:

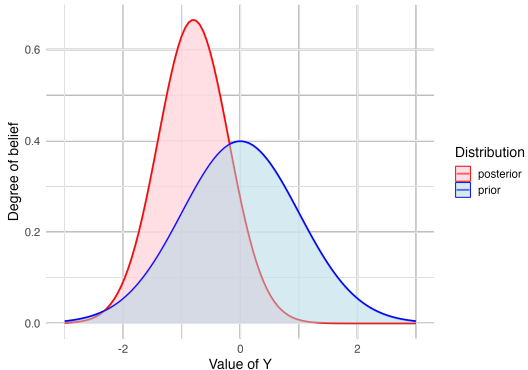


Figure: Informed ignorance

GP intuition: Ignorance and neighbors

Now suppose you have numerous observations of $\{X_i, Y_i\}_i^N$.

GP intuition: Ignorance and neighbors

Now suppose you have numerous observations of $\{X_i, Y_i\}_i^N$.

Big idea: observations nearer each other in X should have similar Y, whatever the values:

- That is, $cov(Y_i, Y_j)$ larger when X_i nearer X_j.
- E.g., $cov(Y_i, Y_j) = k(X_i, X_j) = e^{-\frac{||X_i - X_j||^2}{b}}$

GP intuition: Ignorance and neighbors

Now suppose you have numerous observations of $\{X_i, Y_i\}_i^N$.

Big idea: observations nearer each other in X should have similar Y, whatever the values:

- That is, $cov(Y_i, Y_j)$ larger when X_i nearer X_j.
- E.g., $cov(Y_i, Y_j) = k(X_i, X_j) = e^{-\frac{||X_i - X_j||^2}{b}}$

At values of x^* near to observed points, the high covariance with observed Ys improves our guess about Y^*.

GP intuition: Ignorance and neighbors

Now suppose you have numerous observations of $\{X_i, Y_i\}_i^N$.

Big idea: observations nearer each other in X should have similar Y, whatever the values:

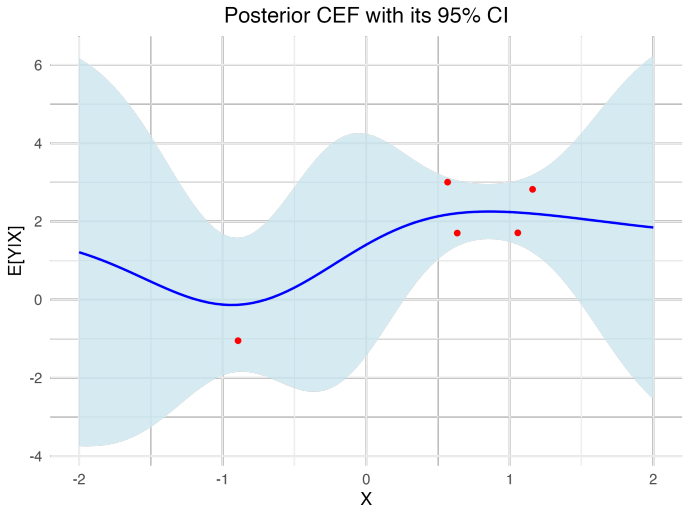
- That is, $cov(Y_i, Y_j)$ larger when X_i nearer X_j.
- E.g., $cov(Y_i, Y_j) = k(X_i, X_j) = e^{-\frac{||X_i - X_j||^2}{b}}$

At values of x^* near to observed points, the high covariance with observed Ys improves our guess about Y^*.

At values of x^* far from observed points, low covariance of observed outcomes with Y^* leaves us more uncertain about $p(Y^*)$.

Inferences given a few observations

Given GP assumptions and 5 observations of $\{X_i, Y_i\}$, ask for predictions at any X:



Written out,

1. $Y \sim \mathcal{N}(\mu, \Sigma)$, deferring how μ and Σ will be determined.

Written out,

1. $Y \sim \mathcal{N}(\mu, \Sigma)$, deferring how μ and Σ will be determined.
2. Choose a kernel function $k(\cdot, \cdot)$, with $cov(Y_i, Y_j) = \sigma_y k(X_i, X_j)$. This can be given a particular form, such as $k(X_i, X_j) = exp(-||X_i - X_j||^2/b)$.

Written out,

1. $Y \sim \mathcal{N}(\mu, \Sigma)$, deferring how μ and Σ will be determined.
2. Choose a kernel function $k(\cdot, \cdot)$, with $cov(Y_i, Y_j) = \sigma_y k(X_i, X_j)$. This can be given a particular form, such as $k(X_i, X_j) = exp(-||X_i - X_j||^2/b)$.
3. Consider the kernel matrix $\mathbf{K}$, containing all the pairwise kernel evaluations, i.e. $\mathbf{K}_{i,j} = k(X_i, X_j)$.

Written out,

1. $Y \sim \mathcal{N}(\mu, \Sigma)$, deferring how μ and Σ will be determined.
2. Choose a kernel function $k(\cdot, \cdot)$, with $cov(Y_i, Y_j) = \sigma_y k(X_i, X_j)$. This can be given a particular form, such as $k(X_i, X_j) = exp(-||X_i - X_j||^2/b)$.
3. Consider the kernel matrix $\mathbf{K}$, containing all the pairwise kernel evaluations, i.e. $\mathbf{K}_{i,j} = k(X_i, X_j)$.
4. Then,

$$Y|X \sim \mathcal{N}(\mu, \sigma_y \mathbf{K})$$

Written out,

1. $Y \sim \mathcal{N}(\mu, \Sigma)$, deferring how μ and Σ will be determined.
2. Choose a kernel function $k(\cdot, \cdot)$, with $cov(Y_i, Y_j) = \sigma_y k(X_i, X_j)$. This can be given a particular form, such as $k(X_i, X_j) = exp(-||X_i - X_j||^2/b)$.
3. Consider the kernel matrix $\mathbf{K}$, containing all the pairwise kernel evaluations, i.e. $\mathbf{K}_{i,j} = k(X_i, X_j)$.
4. Then,
$$Y|X \sim \mathcal{N}(\mu, \sigma_y \mathbf{K})$$
5. That model *goes through every point*, but we expect noise:
$$Y|X \sim \mathcal{N}(\mu, \sigma_y(K + \sigma^2 I))$$

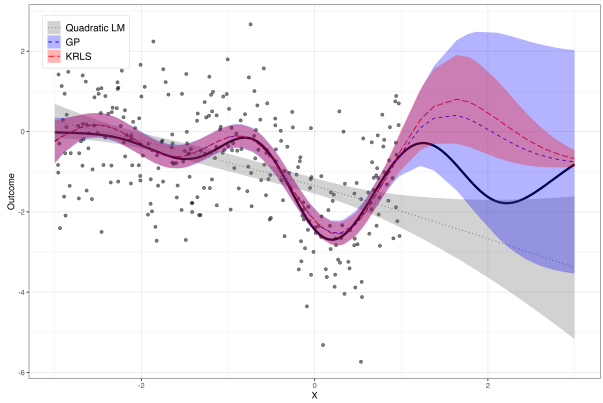
Written out,

1. $Y \sim \mathcal{N}(\mu, \Sigma)$, deferring how μ and Σ will be determined.
2. Choose a kernel function $k(\cdot, \cdot)$, with $cov(Y_i, Y_j) = \sigma_y k(X_i, X_j)$. This can be given a particular form, such as $k(X_i, X_j) = exp(-||X_i - X_j||^2/b)$.
3. Consider the kernel matrix $\mathbf{K}$, containing all the pairwise kernel evaluations, i.e. $\mathbf{K}_{i,j} = k(X_i, X_j)$.
4. Then,
$$Y|X \sim \mathcal{N}(\mu, \sigma_y \mathbf{K})$$
5. That model *goes through every point*, but we expect noise:
$$Y|X \sim \mathcal{N}(\mu, \sigma_y(K + \sigma^2 I))$$
6. After centering and scaling Y we can use $\mu = 0$ and $\sigma_y = 1$,
$$Y|X \sim \mathcal{N}(0, K + \sigma^2 I)$$

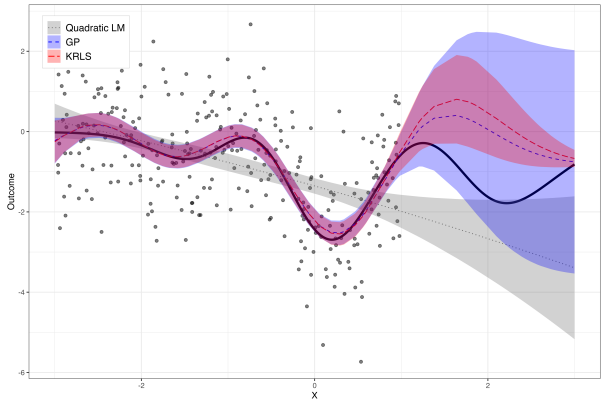
Written out,

1. $Y \sim \mathcal{N}(\mu, \Sigma)$, deferring how μ and Σ will be determined.
2. Choose a kernel function $k(\cdot, \cdot)$, with $cov(Y_i, Y_j) = \sigma_y k(X_i, X_j)$. This can be given a particular form, such as $k(X_i, X_j) = exp(-||X_i - X_j||^2/b)$.
3. Consider the kernel matrix $\mathbf{K}$, containing all the pairwise kernel evaluations, i.e. $\mathbf{K}_{i,j} = k(X_i, X_j)$.
4. Then,
$$Y|X \sim \mathcal{N}(\mu, \sigma_y \mathbf{K})$$
5. That model *goes through every point*, but we expect noise:
$$Y|X \sim \mathcal{N}(\mu, \sigma_y(K + \sigma^2 I))$$
6. After centering and scaling Y we can use $\mu = 0$ and $\sigma_y = 1$,
$$Y|X \sim \mathcal{N}(0, K + \sigma^2 I)$$
7. "Conditioning on the observations", beliefs about the distribution of test points (Y^*) is
$$Y^* \sim \mathcal{N}(K_* \underbrace{(K + \sigma^2 I)^{-1} Y}_{\text{KRLS coefs.}}, \underbrace{K_{*,*} + \sigma^2 I}_{\text{prior var}} - \underbrace{K_*^\top (K + \sigma^2 I)^{-1} K_*}_{\text{reduction due to data}}) \quad (1)$$

Uncertainty and extrapolation in 3 models:

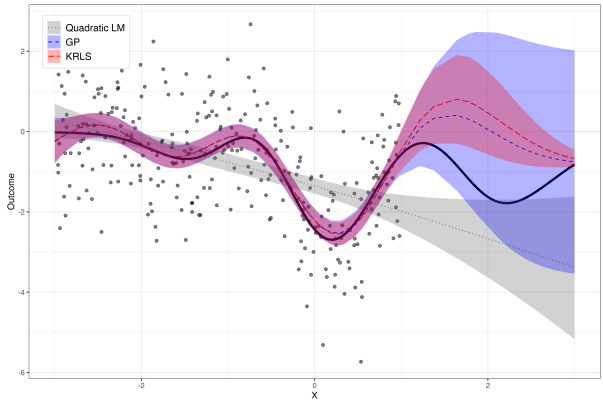


Uncertainty and extrapolation in 3 models:



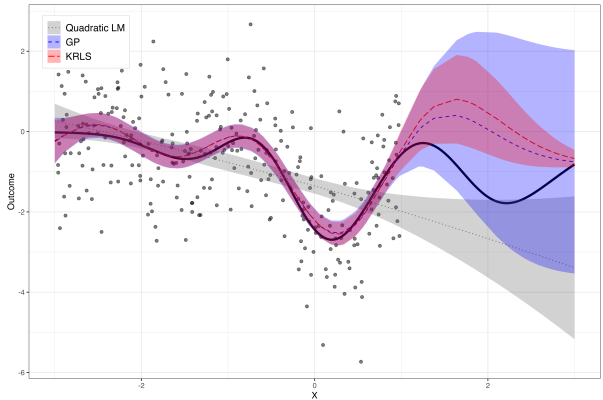
- In linear model, $var(\mathbb{E}[Y^*|X_i]) \propto X_i^2 Var(\beta)$, but is silent on model uncertainty

Uncertainty and extrapolation in 3 models:



- In linear model, $var(\mathbb{E}[Y^*|X_i]) \propto X_i^2 Var(\beta)$, but is silent on model uncertainty
- GP and KRLS have same CEF/*maximum a posteriori*, including the "return to mean"

Uncertainty and extrapolation in 3 models:



- In linear model, $var(\mathbb{E}[Y^*|X_i]) \propto X_i^2 Var(\beta)$, but is silent on model uncertainty
- GP and KRLS have same CEF/*maximum a posteriori*, including the "return to mean"
- But with GP the uncertainty shows our ignorance farther from the data

Additional details

We quickly found that existing implementations of GP in `R` were ill-suited to social sciences; they give little consideration to σ and are quite hard to tune.

Additional details

We quickly found that existing implementations of GP in `R` were ill-suited to social sciences; they give little consideration to σ and are quite hard to tune.

We can avoid a lot of that with some tricks:

Additional details

We quickly found that existing implementations of GP in `R` were ill-suited to social sciences; they give little consideration to σ and are quite hard to tune.

We can avoid a lot of that with some tricks:

- scale Y so that $\sigma_y = 1$.

Additional details

We quickly found that existing implementations of GP in `R` were ill-suited to social sciences; they give little consideration to σ and are quite hard to tune.

We can avoid a lot of that with some tricks:

- scale Y so that $\sigma_y = 1$.
- then σ_ϵ is interpretable as $1 - R^2$ of the best fitting model in the space, and is (suspiciously) well approximated by a marginal likelihood approach.

Additional details

We quickly found that existing implementations of GP in `R` were ill-suited to social sciences; they give little consideration to σ and are quite hard to tune.

We can avoid a lot of that with some tricks:

- scale Y so that $\sigma_y = 1$.
- then σ_ϵ is interpretable as $1 - R^2$ of the best fitting model in the space, and is (suspiciously) well approximated by a marginal likelihood approach.
- in case of the Gaussian with bandwidth b, choose $\underset{b}{argmax}\ Var(\mathbf{K})$.

Additional details

We quickly found that existing implementations of GP in `R` were ill-suited to social sciences; they give little consideration to σ and are quite hard to tune.

We can avoid a lot of that with some tricks:

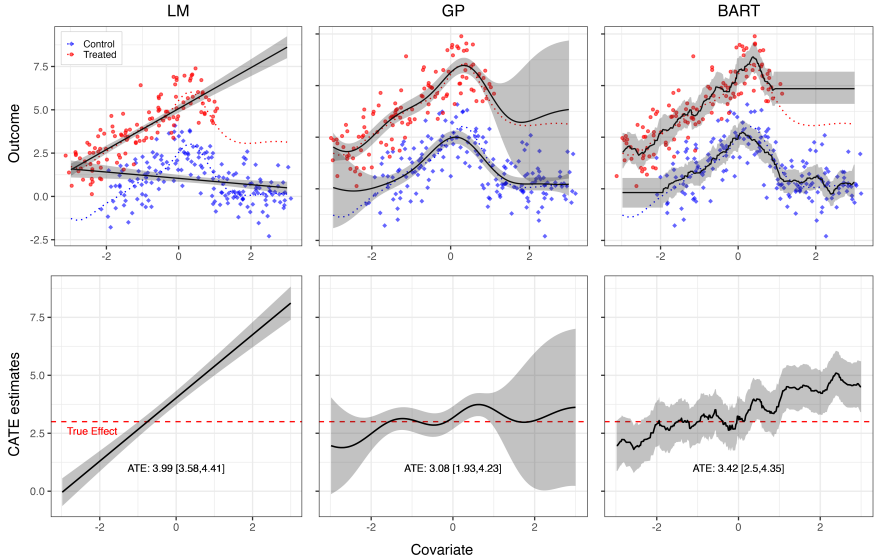
- scale Y so that $\sigma_y = 1$.
- then σ_ϵ is interpretable as $1 - R^2$ of the best fitting model in the space, and is (suspiciously) well approximated by a marginal likelihood approach.
- in case of the Gaussian with bandwidth b, choose $\underset{b}{argmax}\ Var(\mathbf{K})$.
- The resulting procedure avoids all user-chosen hyperparameters and works very well empirically.

Case 1: Comparing groups with poor overlap

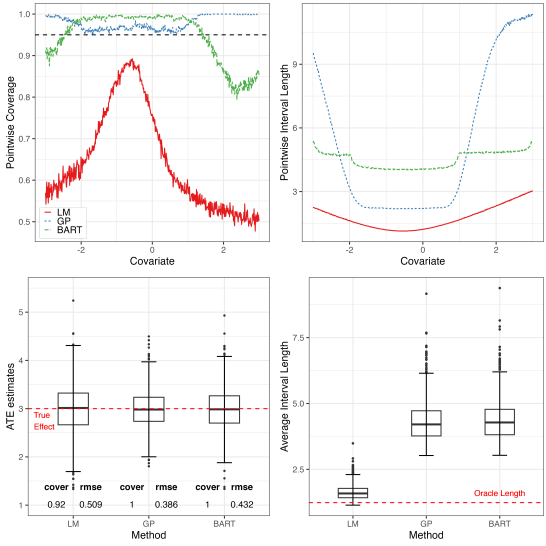
Simulation setting:

- $-3 < X < 3$, but no controls below X=-2, no treated above X=1.
- CEF drawn from a space of random functions on each iteration
- Truth: constant treatment effect of 3
- Compare approaches on ATE, CATE, coverage of both.

Case 1: Checking one draw



Case 1: Comparing groups with poor overlap



Case 2: Interrupted time-series (ITS)

- The ITS design attempts to assess the effect of events/shocks experienced by everyone after a given time

Case 2: Interrupted time-series (ITS)

- The ITS design attempts to assess the effect of events/shocks experienced by everyone after a given time
- Fit model in pre-treatment period, then **extrapolate** it over post-treatment periods

Case 2: Interrupted time-series (ITS)

- The ITS design attempts to assess the effect of events/shocks experienced by everyone after a given time
- Fit model in pre-treatment period, then **extrapolate** it over post-treatment periods
- We can combine kernels to accommodate smooth functions, polynomials growth beyond the data, and periodicity.

Case 2: Interrupted time-series (ITS)

- The ITS design attempts to assess the effect of events/shocks experienced by everyone after a given time
- Fit model in pre-treatment period, then **extrapolate** it over post-treatment periods
- We can combine kernels to accommodate smooth functions, polynomials growth beyond the data, and periodicity.
- For causality need to claim extrapolated pre-treatment trends approximate $\mathbb{E}[Y(0)]$ over time in post-tx era.

Case 2: Interrupted time-series (ITS)

- The ITS design attempts to assess the effect of events/shocks experienced by everyone after a given time
- Fit model in pre-treatment period, then **extrapolate** it over post-treatment periods
- We can combine kernels to accommodate smooth functions, polynomials growth beyond the data, and periodicity.
- For causality need to claim extrapolated pre-treatment trends approximate $\mathbb{E}[Y(0)]$ over time in post-tx era.
- That is, interpret modestly! e.g.,

  "This is our best estimate of how the outcome might look absent the treatment, given the prior trend information, and if nothing else of importance to the outcome were to occur over the post-treatment window"

Case 2: Heller effect (application)

- How did the decision in *District of Columbia v. Heller* (2008) affect gun purchasing behavior? (Thanks to Jack Kappelman!)

Case 2: Heller effect (application)

- How did the decision in *District of Columbia v. Heller* (2008) affect gun purchasing behavior? (Thanks to Jack Kappelman!)
- Treatment event: *Heller* ruling (June 2008)

Case 2: Heller effect (application)

- How did the decision in *District of Columbia v. Heller* (2008) affect gun purchasing behavior? (Thanks to Jack Kappelman!)
- Treatment event: *Heller* ruling (June 2008)
- Period: 3 years before, 1 year after

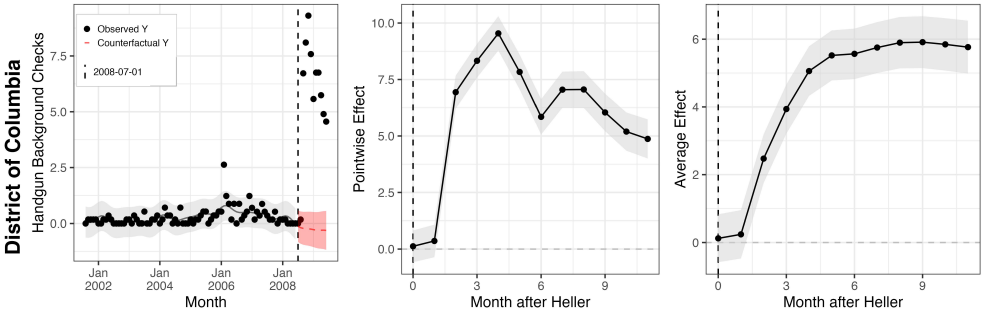
Case 2: Heller effect (application)

- How did the decision in *District of Columbia v. Heller* (2008) affect gun purchasing behavior? (Thanks to Jack Kappelman!)
- Treatment event: *Heller* ruling (June 2008)
- Period: 3 years before, 1 year after
- Outcome: state-month rate of handgun background checks per 100k population (proxy for handgun purchasing)

Case 2: Heller effect (application)

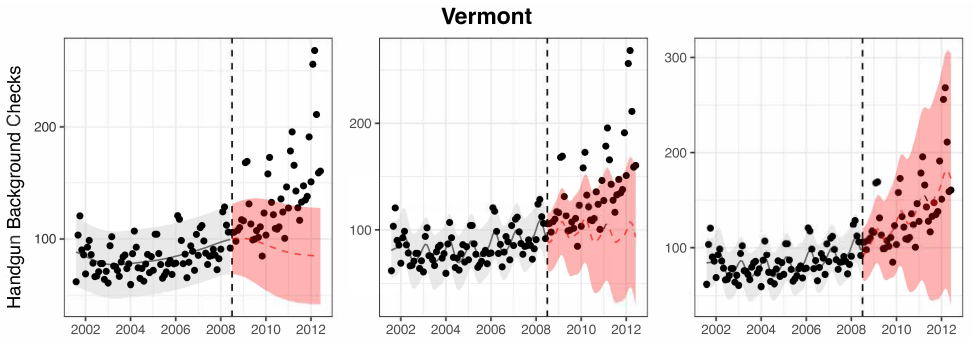
- How did the decision in *District of Columbia v. Heller* (2008) affect gun purchasing behavior? (Thanks to Jack Kappelman!)
- Treatment event: *Heller* ruling (June 2008)
- Period: 3 years before, 1 year after
- Outcome: state-month rate of handgun background checks per 100k population (proxy for handgun purchasing)
- Kernel: Gaussian + periodic + polynomial.

Case 2: Heller and firearms in DC



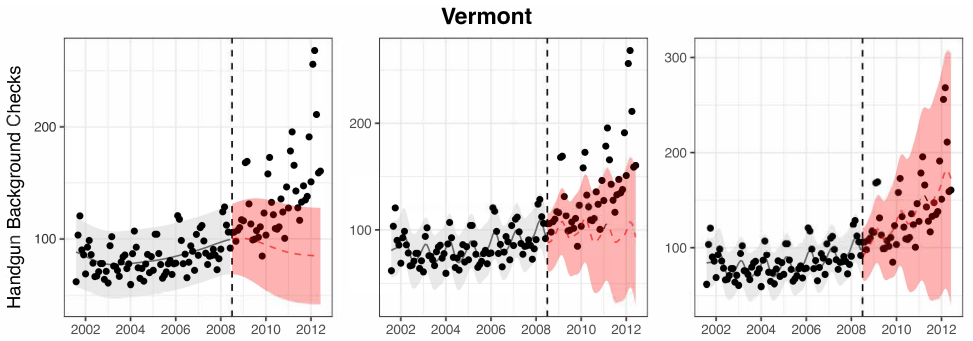
Case 2: Not so clear in Vermont

Gaussian (*left*), Gaussian+periodic+linear (*center*), Gaussian+periodic+quadratic (*right*)



Case 2: Not so clear in Vermont

Gaussian (*left*), Gaussian+periodic+linear (*center*), Gaussian+periodic+quadratic (*right*)



Defensible claims (still) depend on assumptions from outside the data,
here involving the space of functions you cannot rule out for extrapolation.

Case 3: Regression Discontinuity

RD always involves estimating to the edge of the data, or slightly past it:

- $\widehat{Y(0)}_{Z=c}$: Train $\mathbb{E}[Y(0)|Z]$ on $Z < c$ and predict at $Z = c$
- $\widehat{Y(1)}_{Z=c}$: Train $\mathbb{E}[Y(1)|Z]$ on $Z > c$ and predict at $Z = c$

$$\hat{\tau} = \widehat{Y(1)}_{Z=c} - \widehat{Y(0)}_{Z=c}$$

Case 3: Regression Discontinuity

RD always involves estimating to the edge of the data, or slightly past it:

- $\widehat{Y(0)}_{Z=c}$: Train $\mathbb{E}[Y(0)|Z]$ on $Z < c$ and predict at $Z = c$
- $\widehat{Y(1)}_{Z=c}$: Train $\mathbb{E}[Y(1)|Z]$ on $Z > c$ and predict at $Z = c$

$$\hat{\tau} = \widehat{Y(1)}_{Z=c} - \widehat{Y(0)}_{Z=c}$$

Now-standard approaches of optimally-tuned local polynomials are built for relatively large samples (e.g. `rdrobust`)

Case 3: Regression Discontinuity

RD always involves estimating to the edge of the data, or slightly past it:

- $\widehat{Y(0)}_{Z=c}$: Train $\mathbb{E}[Y(0)|Z]$ on $Z < c$ and predict at $Z = c$
- $\widehat{Y(1)}_{Z=c}$: Train $\mathbb{E}[Y(1)|Z]$ on $Z > c$ and predict at $Z = c$

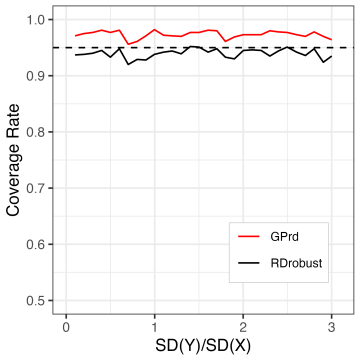
$$\hat{\tau} = \widehat{Y(1)}_{Z=c} - \widehat{Y(0)}_{Z=c}$$

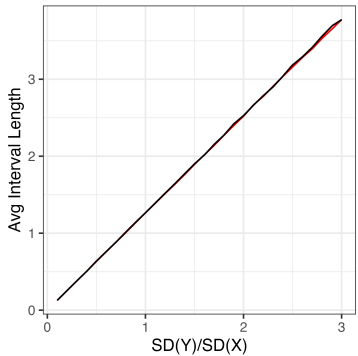
Now-standard approaches of optimally-tuned local polynomials are built for relatively large samples (e.g. `rdrobust`)

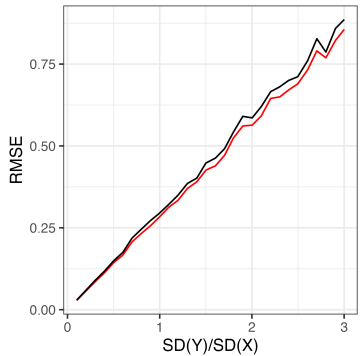
For investigators with small sample RDs, can GP help?

RD: random Z and Y, large sample

$\tau = 3$ (or $\tau = 0$), $N = 500$, `rdrobust` at defaults (robust estimate), `GPrd` at defaults

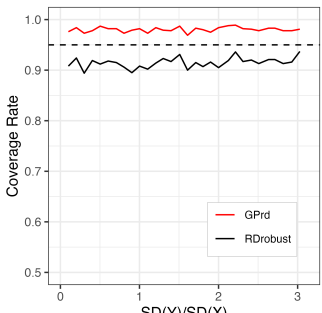


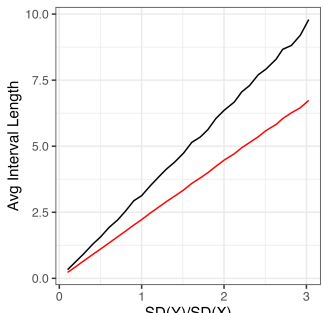


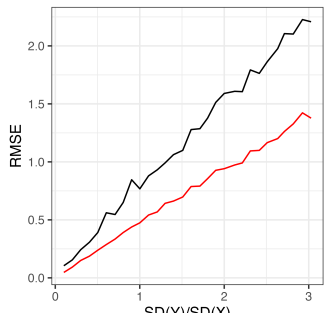


RD: random Z and Y, small sample

$\tau = 3$ (or $\tau = 0$), $N = 100$, `rdrobust` at defaults (robust estimate), `GPrd` at defaults

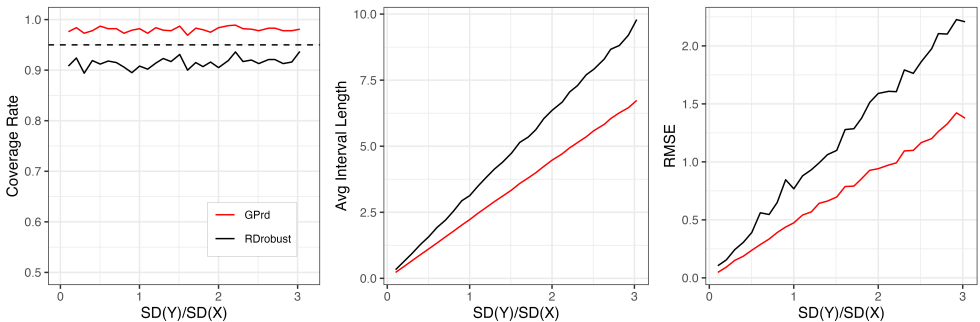






RD: random Z and Y, small sample

$\tau = 3$ (or $\tau = 0$), $N = 100$, `rdrobust` at defaults (robust estimate), `GPrd` at defaults



(Similar story if you posit a latent variable confounding running variable and outcome)

Placebo cutoffs in close elections (application)

Lee (2004) close election setting:

- Forcing variable: Democrats' vote share in US House election (1948 - 1990)
- Outcome variable: ADA score (liberal vote score of each representative)

Placebo cutoffs in close elections (application)

Lee (2004) close election setting:

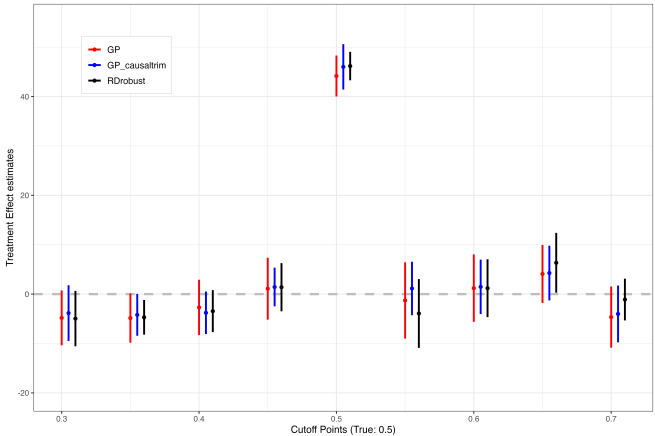
- Forcing variable: Democrats' vote share in US House election (1948 - 1990)
- Outcome variable: ADA score (liberal vote score of each representative)

Another twist: Should our RD estimates be informed by some *causal assumptions* regarding comparability?

- Should we worry if an automated RD procedure chooses a very wide bandwidth?
- In the `gp_causaltrim` approach we:
  1. use small bandwidth to limit comparison range: e.g. within 2% vote share, cov=cor=0.9 $\rightarrow$ b=0.005
  2. trim: remove data outside useful range ($\pm$10%, so scaling and σ^2 choice are focused)

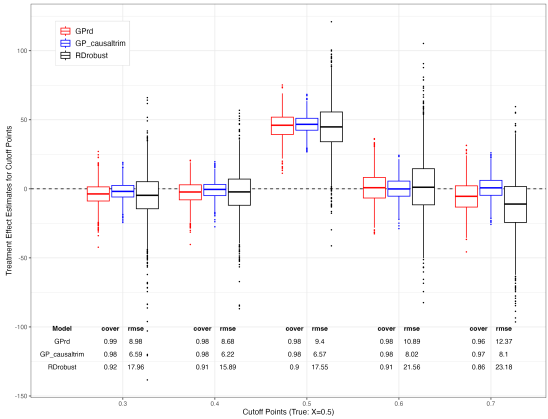
Placebo cutoffs in close elections: Large sample

Full sample: $N = 13,577$ (though each cutoff analysis sees only part of this)



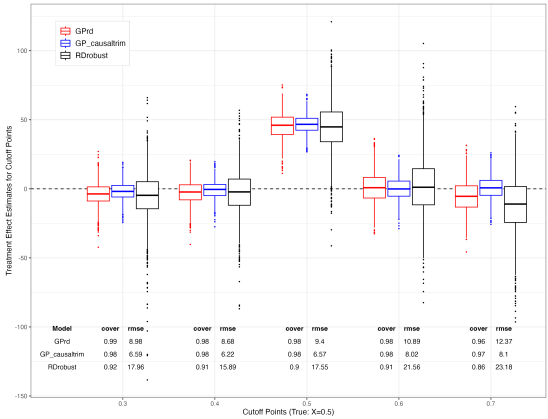
Placebo cutoffs in close elections: Small samples

Repeated sampling of sub-samples: $N = 200$ in each segment



Placebo cutoffs in close elections: Small samples

Repeated sampling of sub-samples: $N = 200$ in each segment



`GP_causaltrim`: conveniently does best while being a bit more transparent, restrictive.

Wrap-up: Entailing counterfactual uncertainty

In conventional approaches,

Wrap-up: Entailing counterfactual uncertainty

In conventional approaches,

- once you pick and fit a model, your SEs are conditional on that specification and fit

Wrap-up: Entailing counterfactual uncertainty

In conventional approaches,

- once you pick and fit a model, your SEs are conditional on that specification and fit
- when extrapolated, uncertainty estimates do not acknowledge our model uncertainty

GPs offer one reasoned solution:

- for the price of the GP model assumptions, gives a principled, "ignorance inclusive" accounting of uncertainty at new points

Wrap-up: Entailing counterfactual uncertainty

In conventional approaches,

- once you pick and fit a model, your SEs are conditional on that specification and fit
- when extrapolated, uncertainty estimates do not acknowledge our model uncertainty

GPs offer one reasoned solution:

- for the price of the GP model assumptions, gives a principled, "ignorance inclusive" accounting of uncertainty at new points

Wrap-up: Entailing counterfactual uncertainty

In conventional approaches,

- once you pick and fit a model, your SEs are conditional on that specification and fit
- when extrapolated, uncertainty estimates do not acknowledge our model uncertainty

GPs offer one reasoned solution:

- for the price of the GP model assumptions, gives a principled, "ignorance inclusive" accounting of uncertainty at new points
- manages the "danger of extreme counterfactuals" into "reasoned counterfactual uncertainty" baked into our uncertainty estimates for QOIs.

Wrap-up: Entailing counterfactual uncertainty

In conventional approaches,

- once you pick and fit a model, your SEs are conditional on that specification and fit
- when extrapolated, uncertainty estimates do not acknowledge our model uncertainty

GPs offer one reasoned solution:

- for the price of the GP model assumptions, gives a principled, "ignorance inclusive" accounting of uncertainty at new points
- manages the "danger of extreme counterfactuals" into "reasoned counterfactual uncertainty" baked into our uncertainty estimates for QOIs.
- limitations: violation of $Y \sim \mathcal{N}(0, K)$; uncertainty ultimately limited to σ^2

We also hope to improve accessibility through our explanations and simplified implementation...

Give it a go!

https://doeunkim.org/gpss/