DISSERTATION DEFENSE

Bayesian approaches for inference after selection and model fitting

Spencer Woody

August 7, 2020

The University of Texas at Austin
Department of Statistics
and Data Sciences
College of Natural Sciences

Acknowledgments

Committee members:

- Prof. James Scott[1 2] (co-advisor)
- Prof. Jared Murray[1 2] (co-advisor)
- Prof. Carlos Carvalho[1 2]
- Prof. Cory Zigler[2 3]
- Prof. Peter Hoff[4]

---

[1]Department of Information, Risk, and Operations Management
[2]Department of Statistics and Data Science
[3]Department of Women's Health, Dell Medical School
[4]Department of Statistical Sciences, Duke University

Acknowledgments

Funding sources:

- Salem Center for Policy, McCombs School of Business
- Eberlin Lab, Department of Chemistry
- Meyers Lab, Department of Integrative Biology

Inference after selection and model fitting

- *Common problem*: using the same data to answer multiple questions can induce bias
- **Part 1:** Use data twice, (i) to select targets of inference, then (ii) form estimates for these targets
- **Part 2:** Fit a model, and then interpret the model through post-hoc exploration [data used once]

Overview

I. Bayes-optimal post-selection inference: selection-adjusted frequentist assisted by Bayes (saFAB)[¶]

II. Model interpretation through posterior summarization[‖]

III. Inference for treatment effects under nested subsets of control variables[‖]

IV. Estimating and interpreting heterogeneous effects of continuous treatments using a new BART-based model[‖]

[¶]Joint work with Prof. James Scott and Prof. Oscar Madrid Padilla (UCLA)
[‖]Joint work with Profs. Carlos Carvalho and Jared Murray

Bayes-optimal post-selection inference

Sparse signal detection:

- $(y_i \mid \theta_i) \sim \mathrm{N}(\theta_i, \sigma^2)$
- Most θ_i are zero or very small

Goal: *Quantify uncertainty for the "interesting" θ_i once we've found them, while adjusting for selection.*

saFAB

We present a method which:

(1) Gives confidence intervals which correctly adjust for selection

(2) Intervals are as short as possible on average while also having uniform coverage (borrowing from Yu and Hoff, 2018)

(3) Estimates the prior when it is unknown and **gives a consistent estimate of the optimal procedure**

- **Update:** new consistency result for nonparametric empirical Bayes procedure
- New version currently under revision for publication in *Biometrika*

Posterior summarization

Introduction

Consider a generic regression model:

$$y_i = f(\mathbf{x}_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}(0, \sigma^2)$$

Suppose we want to accomplish two tasks at once:

1. Estimate f as realistically as possible, and
2. Understand important trends within the data, e.g.
   - Which covariates have strongest effect on prediction?
   - Does covariate importance differ across the covariate space?
   - Are there important interactions?

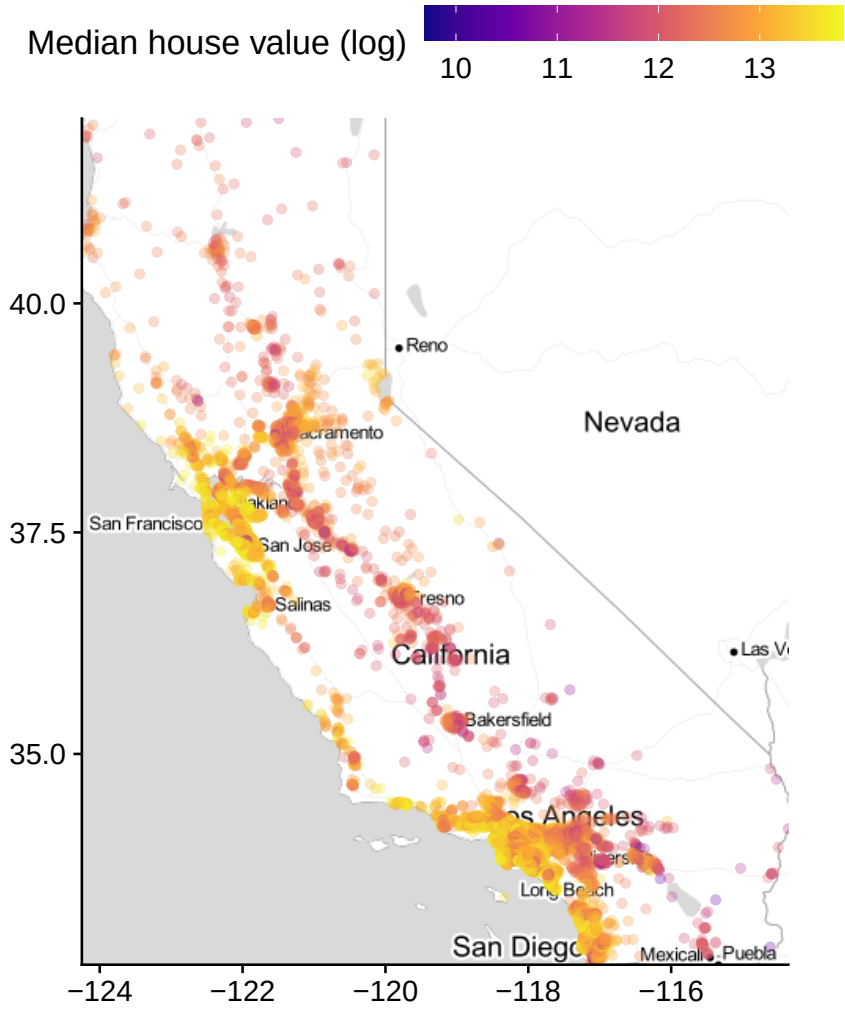
Motivating example: Housing prices in California

Outcome: y_i is census tract-level median house values in California (log)

Predictors:

- log-median household income
- log-population
- median no. of rooms per unit
- longitude
- latitude

$n = 7481,\ p = 5$



Interpretability vs. flexibility

There is a natural tension between fitting...

- Flexible, more realistic, but “black box” models
  - Gaussian process
  - Tree ensembles
- Simple, interpretable, but (presumably) misspecified models
  - $y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i$
  - $y_i = \beta_0 + \sum_{j=1}^{p} g_j(x_{ij}) + \varepsilon_i$

Interpretability vs. flexibility

- Simple models are not believed to be "true" but are rather the *best linear/additive approximation* to the regression surface
- Detecting important trends might involve fitting several models

$$\mathcal{M}_1 : y = \beta_{11}x_1 + \beta_{12}x_2 + \ldots + \varepsilon$$
$$\mathcal{M}_2 : y = \beta_{12}x_1 + \beta_{22}x_2 + \beta_{23}x_1x_2 + \ldots + \varepsilon$$
$$\mathcal{M}_3 : y = \beta_{31}x_1 + \beta_{32}x_2 + \beta_{33}x_1^2 + \ldots + \varepsilon, \quad \text{etc.}$$

  and selecting one for interpretation
- Should worry about model refinement + posterior inference after using the data multiple times ("posterior hacking")

Interpretability vs. flexibility

- Simple models are not believed to be "true" but are rather the *best linear/additive approximation* to the regression surface
- Detecting important trends might involve fitting several models

$$\mathcal{M}_1 : y = \beta_{11} x_1 + \beta_{12} x_2 + \ldots + \varepsilon$$
$$\mathcal{M}_2 : y = \beta_{12} x_1 + \beta_{22} x_2 + \beta_{23} x_1 x_2 + \ldots + \varepsilon$$
$$\mathcal{M}_3 : y = \beta_{31} x_1 + \beta_{32} x_2 + \beta_{33} x_1^2 + \ldots + \varepsilon, \quad \text{etc.}$$

  and selecting one for interpretation

- Should worry about model refinement + posterior inference after using the data multiple times ("posterior hacking")

Separating modeling and interpretation

We propose a two-stage process[vii]:

I. Specify a flexible prior for f and use all available data to best estimate it

II. Perform a *post hoc* investigation of the fitted model using lower-dimensional surrogates as summaries which. . .
   - propagate posterior uncertainty
   - are suited to answer relevant inferential questions, and
   - sufficiently represent the model's predictions

[vii] Woody, Carvalho, and Murray (2020), recently accepted at *JCGS*

Motivating example: GP model for housing prices

Model CA census-tract housing prices with a Gaussian process regression model:

$$(y_i \mid f, \sigma^2) = f(x_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}(0, \sigma^2)$$
$$f \sim \mathrm{GP}(0, k(\cdot, \cdot)), \quad p(\sigma^2) \propto \sigma^{-2}$$

Motivating example: GP model for housing prices

1. **Global summaries**
   Average predictive trends across whole dataset,
   where we project $f(x)$ onto simpler interpretable structures...

   (i) Linear summary

$$\gamma(x) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \varepsilon_i$$

   (ii) Additive summary

$$\gamma(x) = \beta_0 + \sum_{j=1}^{p} h_j(x_j) + \varepsilon_i$$

   (iii) (Mostly) additive summary, allowing for some interactions

$$\gamma(x) = \alpha + h_{kl}(x_k, x_l) + \sum_{j \notin \{k,l\}} h_j(x_j),$$

2. **Local linear summaries** (in paper and supplemental slides)
   Covariate importance within geographic regions

Key point

- The data are *used only once* (in finding posterior for f)
- Therefore, we *retain valid Bayesian inference* even after fitting several summaries

Model interpretation through posterior summarization

Posterior summaries

- Define a class of summaries Γ (e.g. linear, additive)
- User-defined summary loss function

$$\mathcal{L}(f, \gamma, \tilde{X}) = d(f, \gamma, \tilde{X}) + p_\lambda(\gamma)$$

  - $d(\cdot, \cdot, \tilde{X})$ measures predictive difference between f and γ (e.g. squared difference)
  - $\tilde{X}$ are covariate locations of interest
  - $p_\lambda(\cdot)$ penalizes complexity in γ (enforces sparsity/smoothness)

Posterior summaries

- Define a class of summaries Γ (e.g. linear, additive)
- User-defined summary loss function

$$\mathcal{L}(f, \gamma, \tilde{X}) = d(f, \gamma, \tilde{X}) + p_\lambda(\gamma)$$

- The model summary is a functional of f minimizing this loss

$$\gamma(x) = \arg\min_{\gamma' \in \Gamma} \mathcal{L}(f, \gamma', \tilde{X})$$

Projected posterior generated using posterior draws of f:

$$\gamma^{[k]}(x) = \arg\min_{\gamma' \in \Gamma} \mathcal{L}(f^{[k]}, \gamma', \tilde{X}), \quad f^{[k]} \sim p(f \mid y)$$

Point estimate for the summary

- Standard Bayesian decision theory dictates that the optimal *point estimate* is

$$\begin{aligned}\hat{\gamma}(x) &= \arg\min_{\gamma' \in \Gamma} \mathrm{E}[\mathcal{L}(f, \gamma', \tilde{X}) \mid Y, X] \\ &= \arg\min_{\gamma' \in \Gamma} \mathrm{E}[d(f, \gamma', \tilde{X}) \mid Y, X] + p_\lambda(\gamma')\end{aligned}$$

- When $d(\cdot, \cdot, \tilde{X})$ is squared-error loss, this becomes

$$\hat{\gamma}(x) = \arg\min_{\gamma' \in \Gamma} \sum_{i=1}^{\tilde{n}} \left[\hat{f}(\tilde{x}_i) - \gamma'(\tilde{x}_i)\right]^2 + p_\lambda(\gamma)$$

"Fitting the fit" with posterior mean fitted values $\hat{f}(\tilde{x}_i)$.

Application to California housing data

Global additive summary for GP model

- Summary: *Best additive approximation to the model*

$$\Gamma = \left\{ \gamma : \gamma(x) = \alpha + \sum_{j=1}^{p} h_j(x_j) \right\}$$

- The *optimal point estimate* for the summary is:[viii]

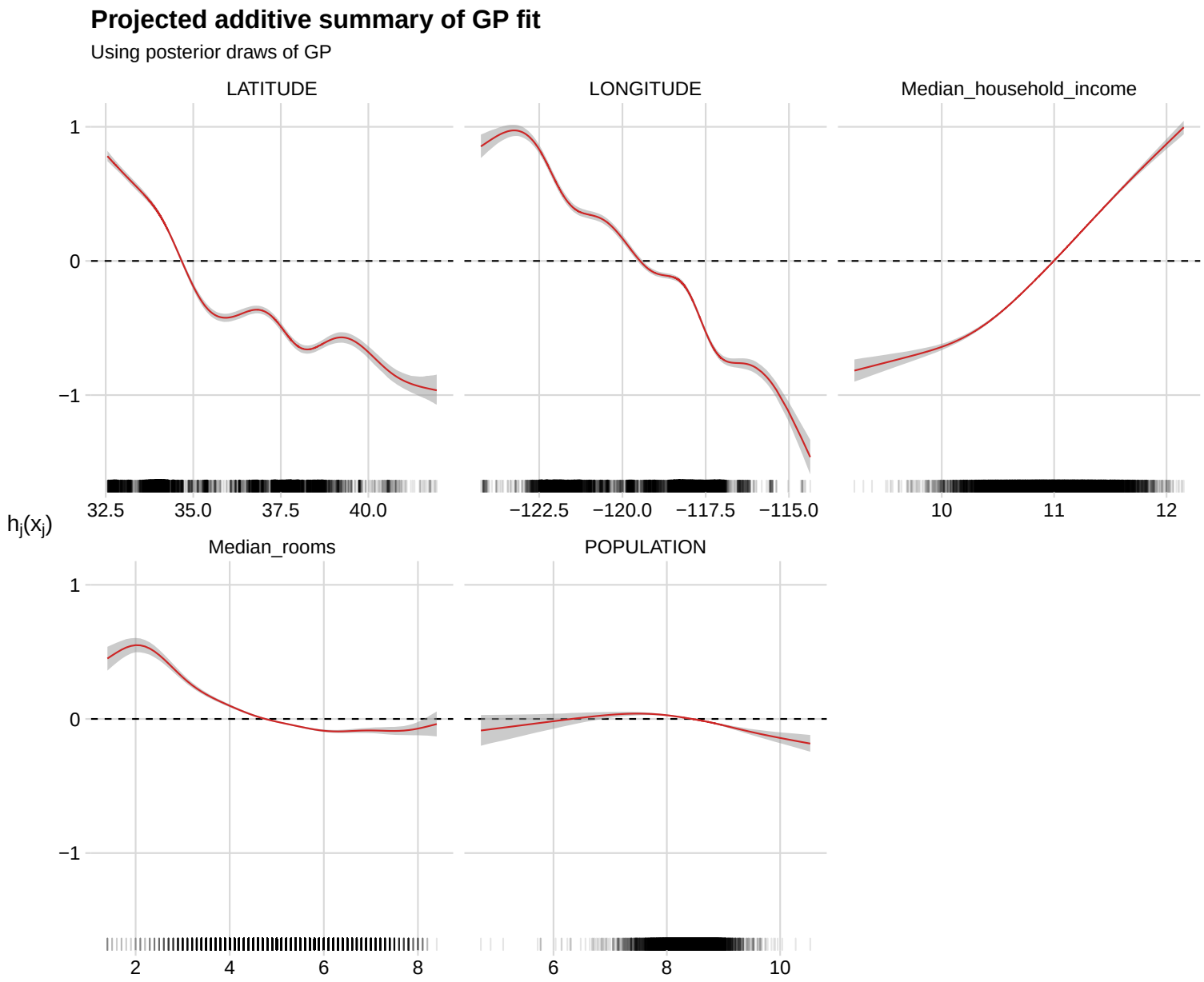
$$\hat{\gamma}(x) = \arg\min_{\gamma' \in \Gamma} \sum_{i=1}^{n} \left[\hat{f}(x_i) - \gamma'(x_i) \right]^2 + \sum_{j=1}^{p} \lambda_j \cdot J(h_j)$$

- The *projected posterior* is found with

$$\gamma^{[k]}(x) = \arg\min_{\gamma' \in \Gamma} \sum_{i=1}^{n} \left[f^{[k]}(x_i) - \gamma'(x_i) \right]^2 + \sum_{j=1}^{p} \lambda_j \cdot J(h_j)$$

[viii] See paper for details on computation

Global additive summary with bands



Summary diagnostics

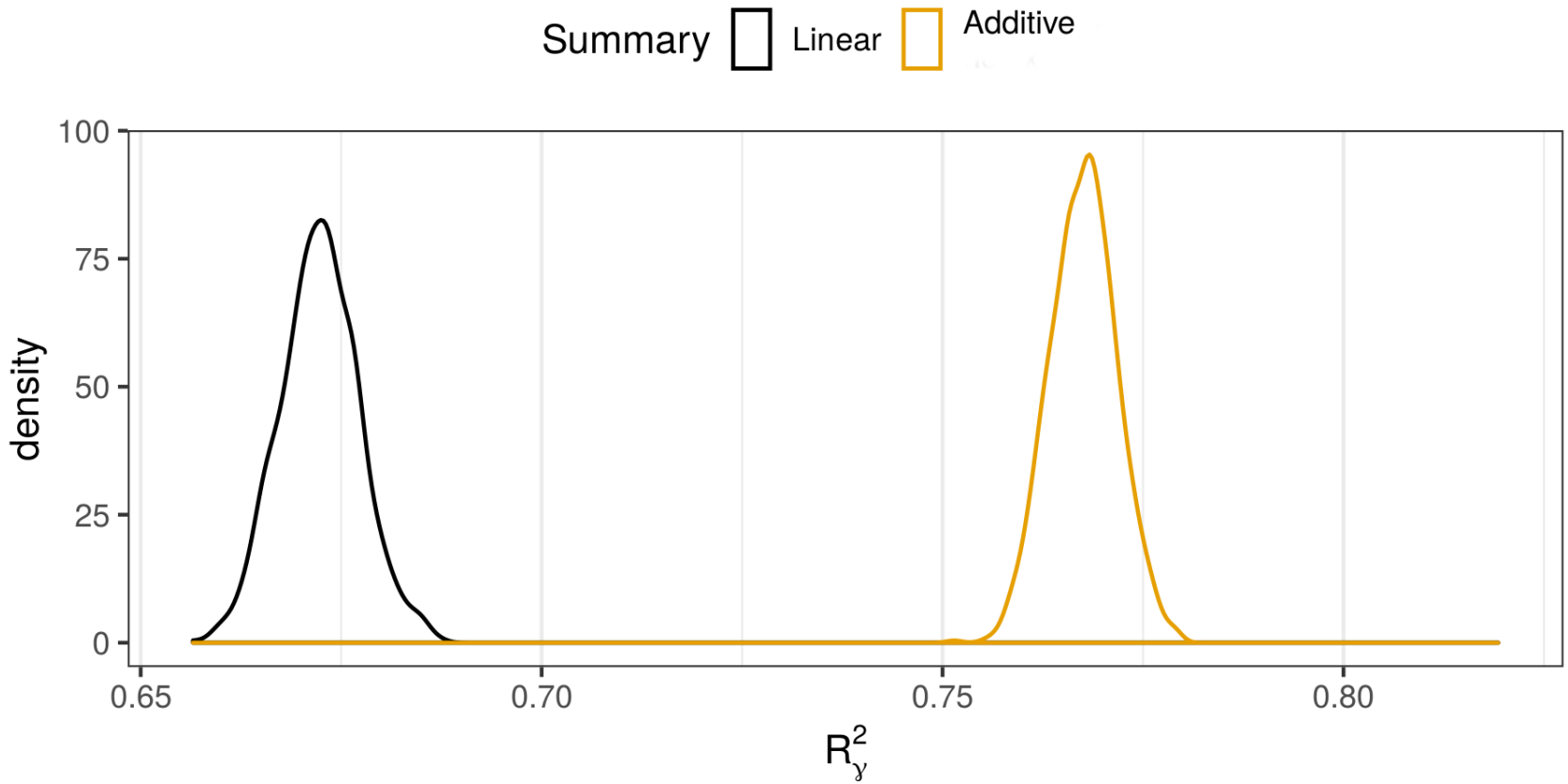
- Summary R^2:

$$R^2_\gamma := 1 - \frac{\sum_i [f(\tilde{x}_i) - \gamma(\tilde{x}_i)]^2}{\sum_i [f(\tilde{x}_i) - \bar{f}]^2},$$

  with $\bar{f} := \tilde{n}^{-1} \sum_i f(\tilde{x}_i)$.
  "Predictive variance explained"

- More in paper...

Summary diagnostics for global additive summary



Iterative summary search

- We can iterate through the summarization process, refining and evaluating the summary each time
- *Retain Bayesian interpretation*

Additive summary with a two-way interaction

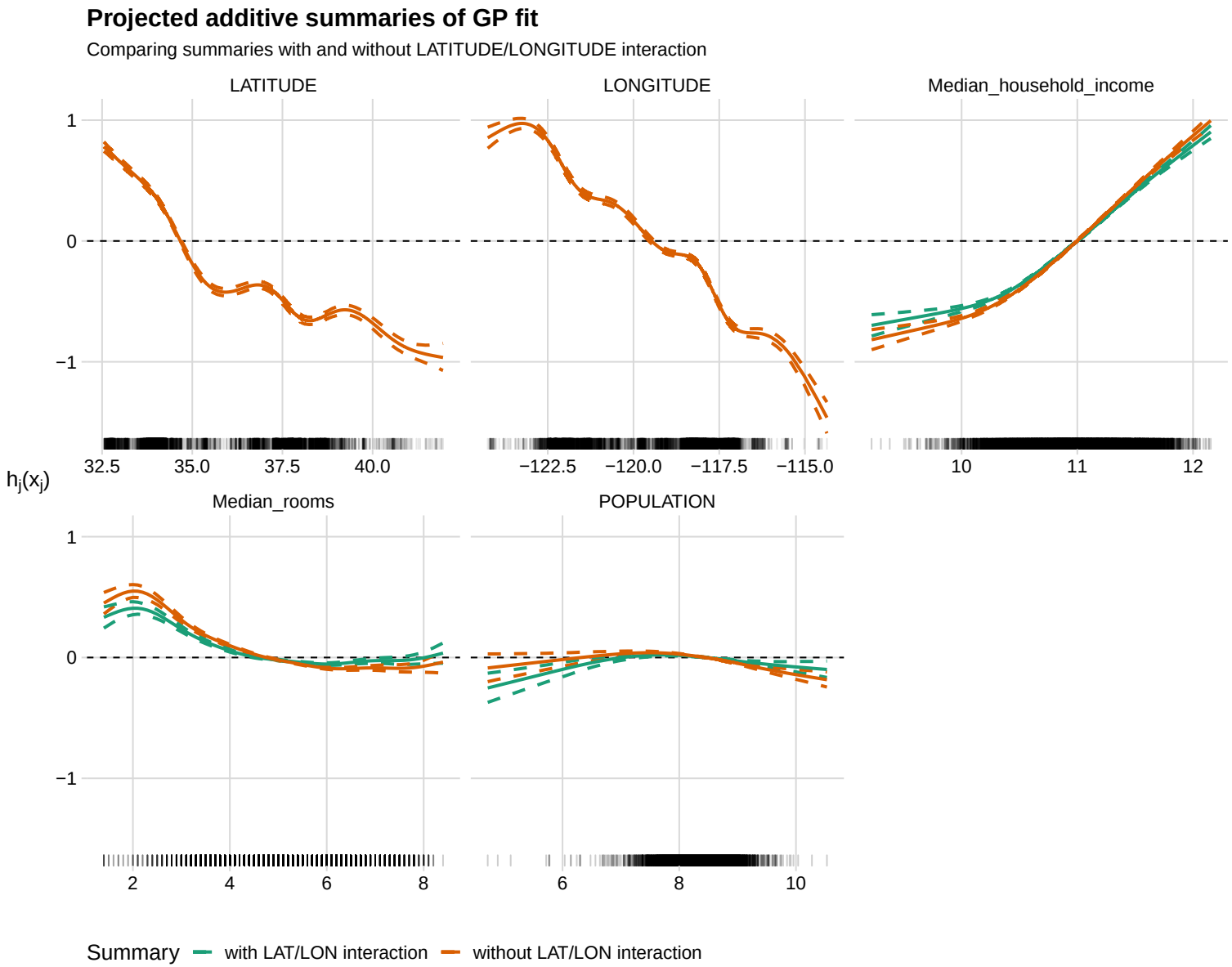
Expand Γ to functions of the form

$$\gamma(x) = \alpha + h_{kl}(x_k, x_l) + \sum_{j \notin \{k,l\}} h_j(x_j),$$

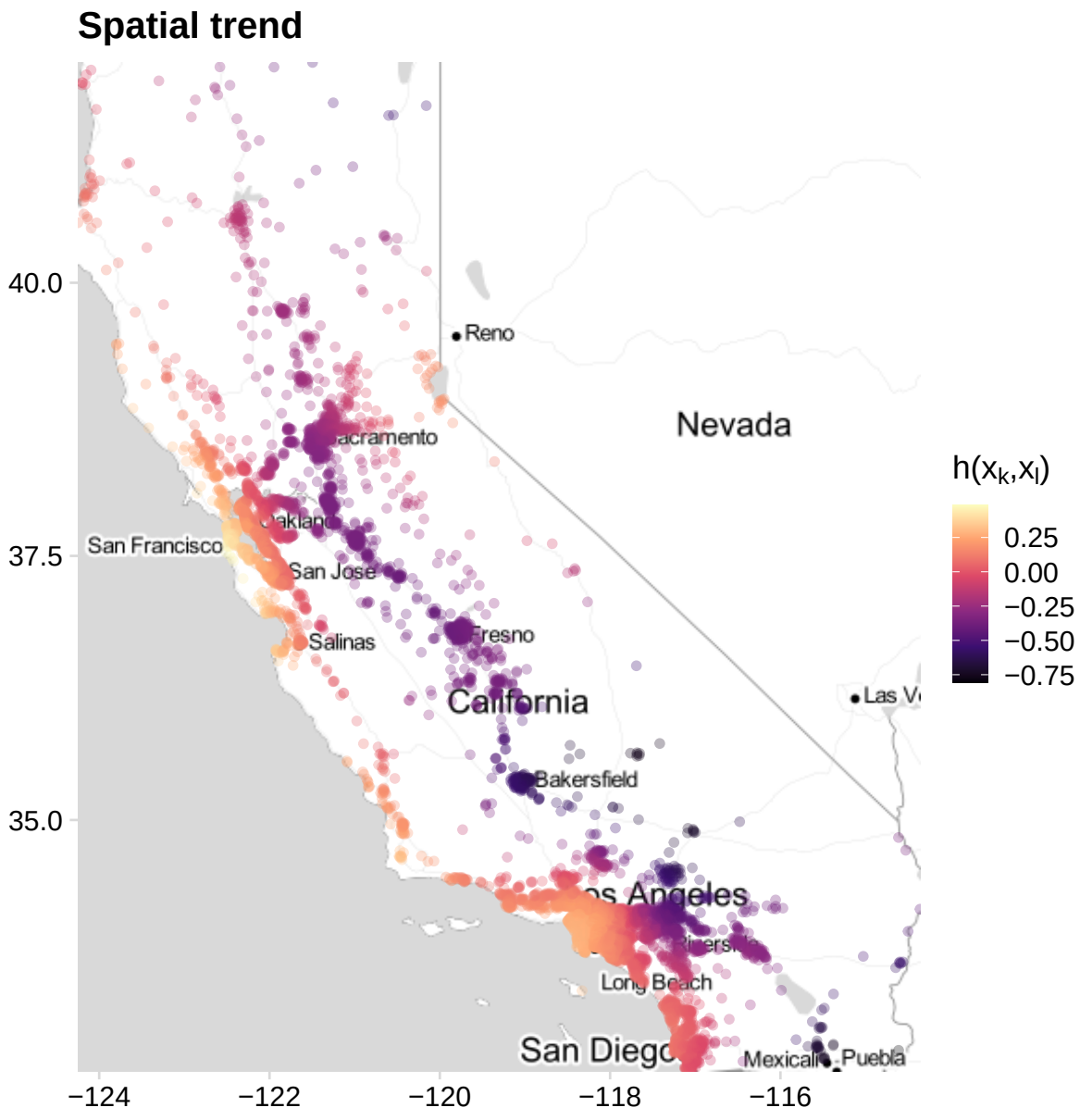
with a 2D smooth function $h_{kl}(x_k, x_l)$ for LON/LAT interaction.

Point estimates and projected posteriors calculated in analogous way to previous summary.

Additive summary with a two-way interaction

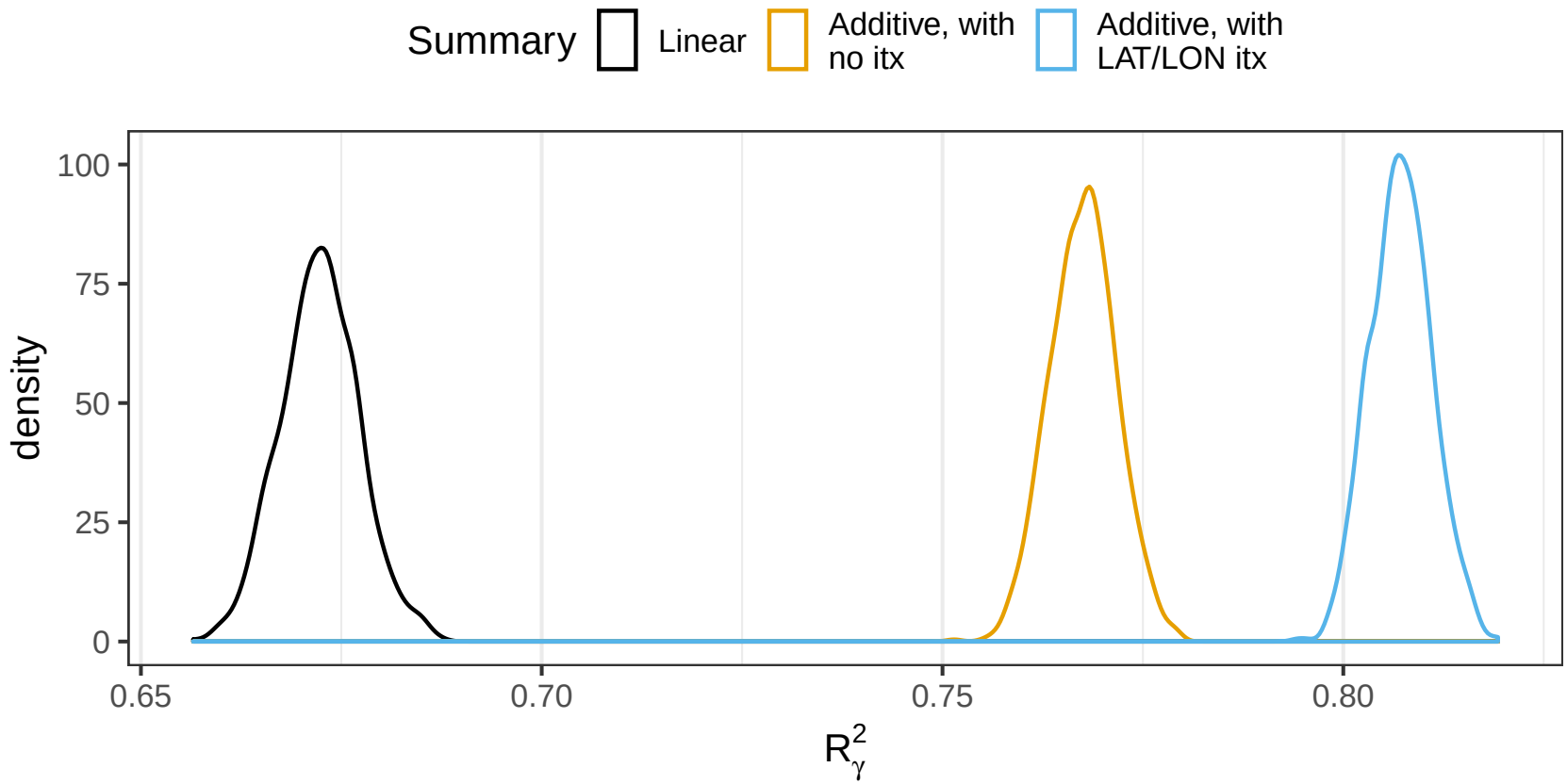


Additive summary with a two-way interaction

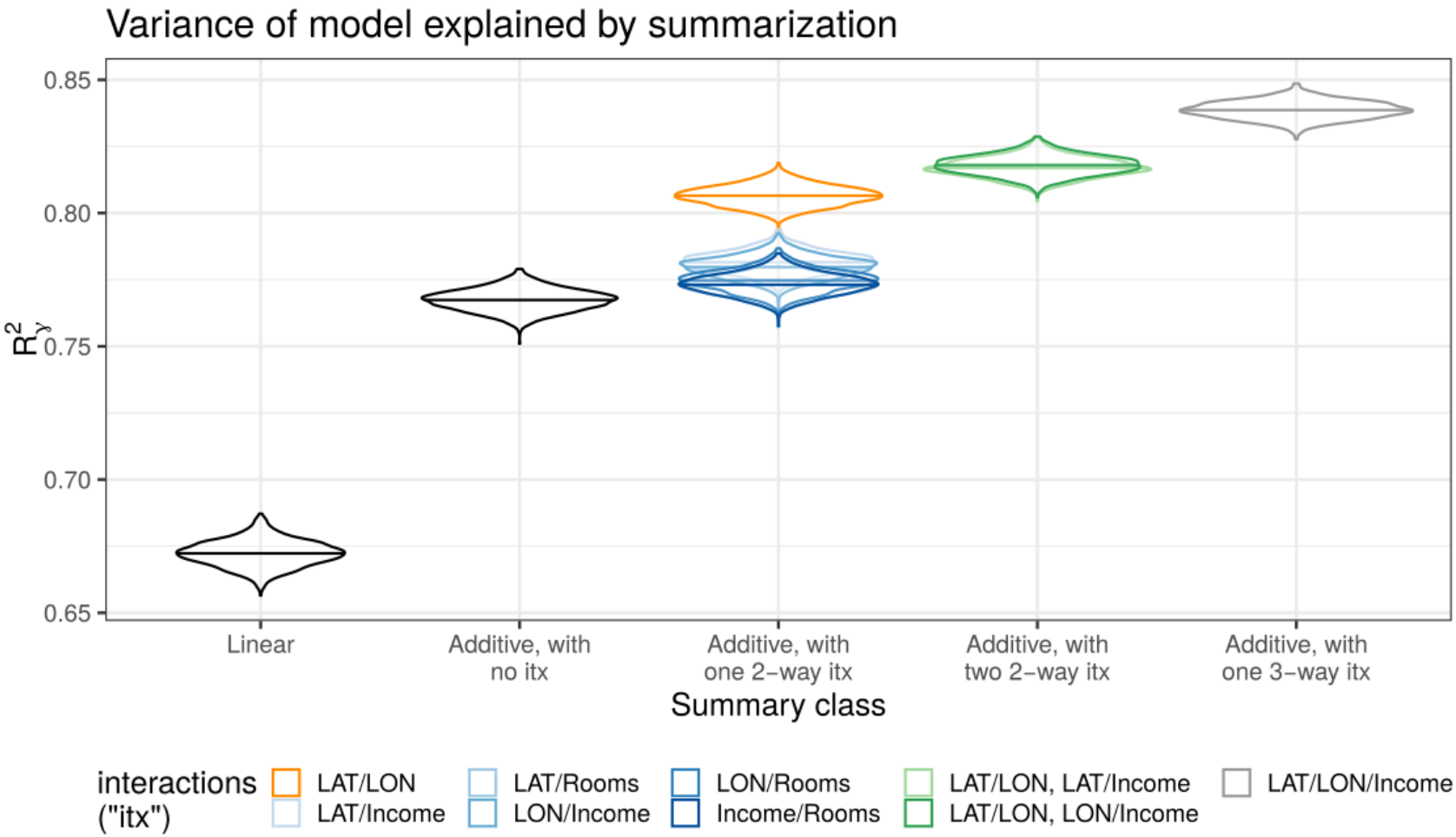


Additive summary with a two-way interaction

Summary diagnostics



Exploring further interactions. . .



Synthesis: global model summary

- The fitted GP regression function is approximately additive, with an important interaction between longitude and latitude
- This summary function explains about 80% of the predictive variance in the fitted model
- More exploration (local summaries) done in paper / extra slides for details

Iterative summary search

(1) **Specify and fit the full model.**
$\mathrm{E}[y_i \mid x_i] = f(x_i)$, assign prior $p(f)$, and compute posterior.

(2) **Summarize.**

- Specify class of summaries Γ and points of interest $\tilde{X}$
- Point estimate

$$\hat{\gamma}(x) = \arg\min_{\gamma \in \Gamma} \mathrm{E}[\mathcal{L}(f, \gamma, \tilde{X}) \mid Y, X]$$

- Posterior around point summary using Monte Carlo draws of f

$$\arg\min_{\gamma \in \Gamma} \mathcal{L}(f, \gamma, \tilde{X})$$

(3) **Evaluate.**
R^2_γ, ϕ_γ, summary residuals $\hat{f}(\tilde{x}_i) - \hat{\gamma}(\tilde{x}_i)$

(4) **Refine and iterate through (2) and (3) as necessary.**

Causal inference under nested models

Introduction

- Apply posterior summarization to causal inference
- **Before:** interpreting model prediction
- **Now:** interpreting a single parameter (treatment effect), and how sensitive it is to model specification

Setup

- **Goal:** estimate causal effect of continuous treatment / exposure $Z \in \mathcal{Z} \subseteq \mathbb{R}$ on some continuous outcome Y
- Use potential outcome framework[ix]:

  Compare $Y(Z = z)$ vs. $Y(Z = z')$ for $z, z' \in \mathcal{Z}$

[ix] see, e.g., Imbens and Rubin (2015)

Identifying assumptions

(i) *Consistency*[x]

$$Z = z \text{ implies } Y = Y(z)$$

(ii) *Weak unconfoundedness*[xi]

$$Y(z) \perp\!\!\!\perp Z \mid X \text{ for all } z \in \mathcal{Z}$$

(iii) *Positivity*[xii]

$$\pi(z \mid x) > 0 \text{ for all } z \in \mathcal{Z}$$

[x] Rubin (1978)
[xi] Imbens (2000)
[xii] Generalized propensity score, Imbens (2000); Hirano and Imbens (2004)

Case study: Abortion-crime hypothesis

- Donohue and Levitt (2001): legalization of abortion in the US in the 1970s helped lead to a dramatic reduction of crime in the 1980s and 1990s.
- Claim a large negative effect after controlling for socioeconomic variables & state- and year-level fixed effects

THE
QUARTERLY JOURNAL
OF ECONOMICS

Vol. CXVI May 2001 Issue 2

THE IMPACT OF LEGALIZED ABORTION ON CRIME*

JOHN J. DONOHUE III AND STEVEN D. LEVITT

D&L control variables

Covariate	Description
police	log-police employment per capita
prison	log-prisoner population per capita
gunlaw	indicator variable for presence of concealed weapons law
unemployment	state unemployment rate
income	state log-income per capita
poverty	state poverty rate
afdc15	generosity to Aid to Families with Dependent Children (AFDC), lagged by 15 years
beer	beer consumption per capita

Common practice: fit multiple models

Variable	ln(Violent crime per capita) (1)	(2)	ln(Property crime per capita) (3)	(4)	ln(Murder per capita) (5)	(6)
"Effective" abortion rate (× 100)	−.137 (.023)	−.129 (.024)	−.095 (.018)	−.091 (.018)	−.108 (.036)	−.121 (.047)
ln(prisoners per capita) ($t-1$)	—	−.027 (.044)	—	−.159 (.036)	—	−.231 (.080)
ln(police per capita) ($t-1$)	—	−.028 (.045)	—	−.049 (.045)	—	−.300 (.109)
State unemployment rate (percent unemployed)	—	.069 (.505)	—	1.310 (.389)	—	.968 (.794)
ln(state income per capita)	—	.049 (.213)	—	.084 (.162)	—	−.098 (.465)
Poverty rate (percent below poverty line)	—	−.000 (.002)	—	−.001 (.001)	—	−.005 (.004)
AFDC generosity ($t-15$) (× 1000)	—	.008 (.005)	—	.002 (.004)	—	−.000 (.000)
Shall-issue concealed weapons law	—	−.004 (.012)	—	.039 (.011)	—	−.015 (.032)
Beer consumption per capita (gallons)	—	.004 (.003)	—	.004 (.003)	—	.006 (.008)
R^2	.938	.942	.990	.992	.914	.918

Figure 1: Table IV from D&L (2001). Treatment effect estimates using state and year dummies, with (*right*) and without (*left*) state controls.

Sensitivity to model specification

- Subsequent studies criticized the functional form of controls
- Belloni et al. (2014) and Hahn et al. (2018) add interactions:
  - state-level controls $\times$ year
  - state-level controls $\times$ year^2
  - state dummies $\times$ year
  - state dummies $\times$ year^2

  After adding these, they claim the causal effect disappears
- Retrospective study by Donohue and Levitt (2019) found that *their predictions from 2001 held up over the next 17 years*

Problem statement

How to reconcile inference under different model specifications without using the outcome data multiple times?

Projected posterior treatment effects

Linear model for estimating the ATE

$$(Y \mid Z, X) = \tau Z + X\beta + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma_\epsilon^2)$$

- **Estimand:** Average treatment effect (ATE):

$$\begin{aligned} \mathrm{ATE}_{z',z} &= \mathrm{E}[Y(Z = z') - Y(Z = z)] \\ &= \tau \cdot (z' - z) \end{aligned}$$

- Agnostic to choice of prior; possible choices: Zigler and Dominici (2014); Wang et al. (2015); Hahn et al. (2018)

Reconciling inference between sets of controls

Use ideas from posterior summarization to consider inference for treatment effects under different specifications

(I) Specify prior and obtain posterior for "full" model

(II) Project this model onto summaries of different specifications (control variables, functional form...)

- Look how posterior for τ changes relative to uncertainty
- Gives valid posterior uncertainty

Application to D&L data

Application to Donahue and Levitt data

- **Outcome:** y_{st} is the log *murder rate* in state s for year t
- **Exposure:** z_{st} is the "effective abortion rate" (D&L, 2001)
  - Lags and weights abortion rates from previous years
- 48 contiguous US states, years 1985–1997 ($N = 624$)
- Denote observations by $i = 1, \ldots, N$

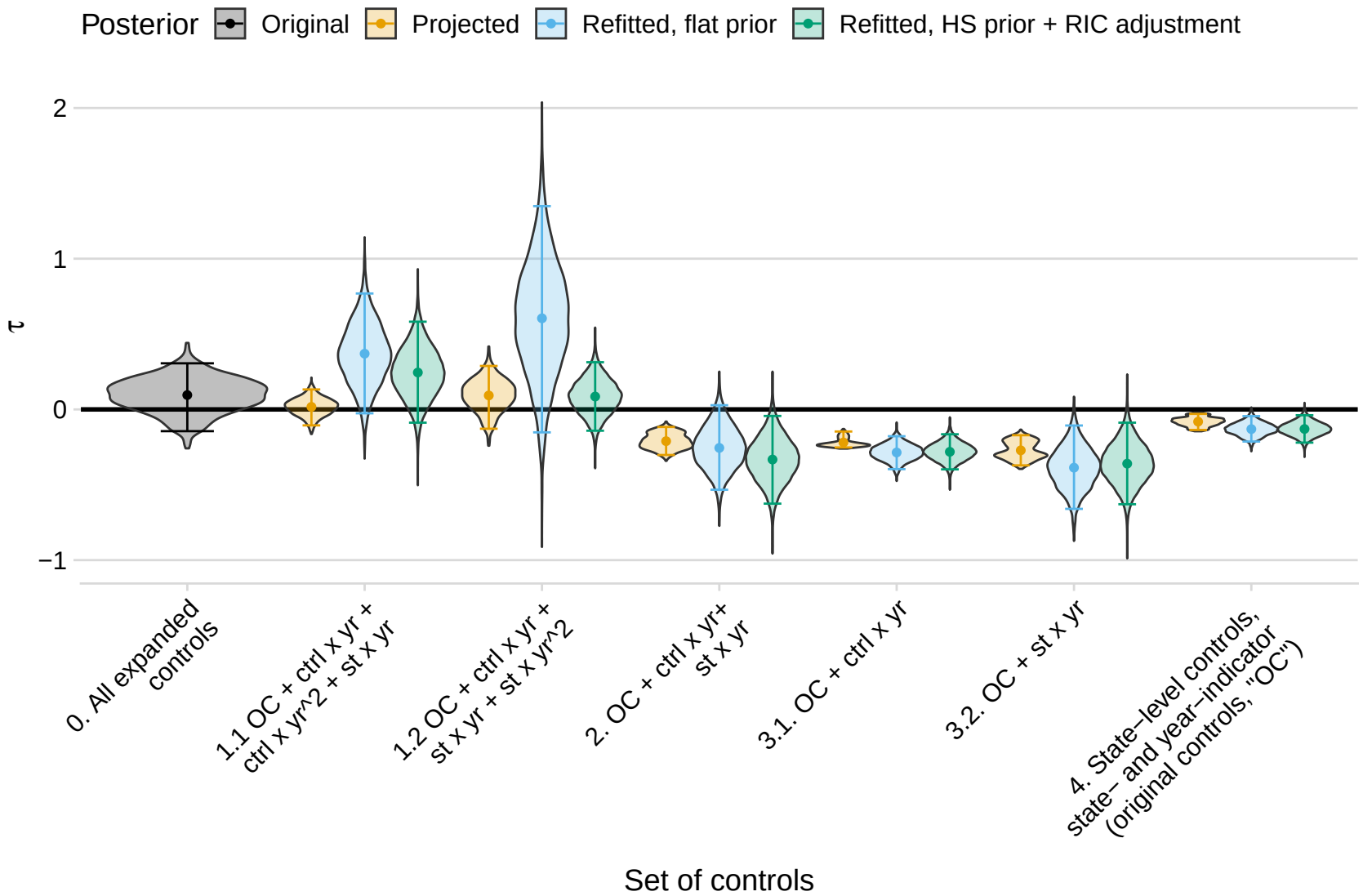
D&L control variables

State-level controls:

Covariate	Description
police	log-police employment per capita
prison	log-prisoner population per capita
gunlaw	indicator variable for presence of concealed weapons law
unemployment	state unemployment rate
income	state log-income per capita
poverty	state poverty rate
afdc15	generosity to Aid to Families with Dependent Children (AFDC), lagged by 15 years
beer	beer consumption per capita

Model specifications

- Donohue and Levitt (2001):
  - state-level controls
  - state dummies
  - year dummies
- Belloni et al. (2014) and Hahn et al. (2018) add interactions:
  - state-level controls $\times$ year
  - state-level controls $\times$ year^2
  - state dummies $\times$ year
  - state dummies $\times$ year^2


Posterior
Original
Projected
Refitted, flat prior
Refitted, HS prior + RIC adjustment
τ
2
1
0
−1
0. All expanded controls
1.1 OC + ctrl x yr + ctrl x yr^2 + st x yr
1.2 OC + ctrl x yr + st x yr + st x yr^2
2. OC + ctrl x yr+ st x yr
3.1. OC + ctrl x yr
3.2. OC + st x yr
4. State−level controls, state− and year−indicator (original controls, "OC")
Set of controls

Moderation of continuous treatments

Methods

Nonparametric control function

- Inference is sensitive to parametric specification (researcher degrees of freedom)
- *Solution*: use a nonparametric function for controls
- Linear model from before. . .

$$(y \mid z, x) = x^{\intercal}\beta + \tau \cdot z + \varepsilon$$

now replaced by. . .

$$(y \mid z, x) = \mu(x) + \tau \cdot z + \varepsilon$$

Addition: heterogeneous treatment effects

- Effect of treatment often depends on context and unit-level qualities
- Closely related to mechanism of the treatment
- *Levitt example:* generous social support systems may reduce impact of abortion on crime by enabling parents to spend more time with their children

Proposed semiparametric model

$$y = \mu(x_C) + \tau(x_M) \cdot z + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2)$$

- $\mu(\cdot)$ is the *control function*
  x_C is vector of the *control variables*
- $\tau(\cdot)$ is the *exposure moderating function*
  x_M is a vector of *moderators*.
- **Main parametric assumption**:
  y is linear in z with slope determined by $\tau(x_M)$

Proposed semiparametric model

$$y = \mu(x_C) + \tau(x_M) \cdot z + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2)$$

Before, we only estimated the average treatment effect (ATE):

$$\mathrm{ATE}_{z',z} = \mathrm{E}[Y(z') - Y(z)]$$

The conditional average treatment effect (CATE) is:

$$\begin{aligned}\mathrm{CATE}_{z',z}(x) &= \mathrm{E}[Y(z') - Y(z) \mid X = x] \\ &= \tau(x_M) \cdot (z' - z)\end{aligned}$$

Proposed semiparametric model

$$y = \mu(x_C) + \tau(x_M) \cdot z + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2)$$

- $\mu(\cdot)$, $\tau(\cdot)$ modeled using Bayesian additive regression trees[xiii]
- Allow for interactions and nonlinearities (no need for *a priori* parametric specification)
- Prior based on Hahn, Murray, and Carvalho (2020), regularize $\tau(\cdot)$ more heavily (shallower trees)
- Interpret effect modification via **posterior summarization**

[xiii] Chipman, George, and McCulloch (2010); review: Hill, Linero, and Murray (2020)

Contribution

Our model. . .

(i) Does not require *a priori* parametric specification for controls

(ii) Identifies effect modification by pre-specified moderators

- Detecting unanticipated effect heterogeneity can *generate novel hypotheses* regarding mechanism, e.g. social support

(iii) Gives interpretable summaries of effect modification using method of *posterior summarization*

Revisiting the D&L data

The data

- Estimate impact of abortion on murder rate (same data...)

The data

Covariate	Description	Used as control?	Used as moderator?
police	log-police employment per capita	Yes	No
prison	log-prisoner population per capita	Yes	No
gunlaw	indicator variable for presence of concealed weapons law	Yes	No
unemployment	state unemployment rate	Yes	Yes
income	state log-income per capita	Yes	Yes
poverty	state poverty rate	Yes	Yes
afdc15	generosity to Aid to Families with Dependent Children (AFDC), lagged by 15 years	Yes	Yes
beer	beer consumption per capita	Yes	Yes
state	categorical variable for state (contiguous US states; 48 levels)	Yes	Yes
year	numeric value for year (1985–1997, inclusive)	Yes	Yes

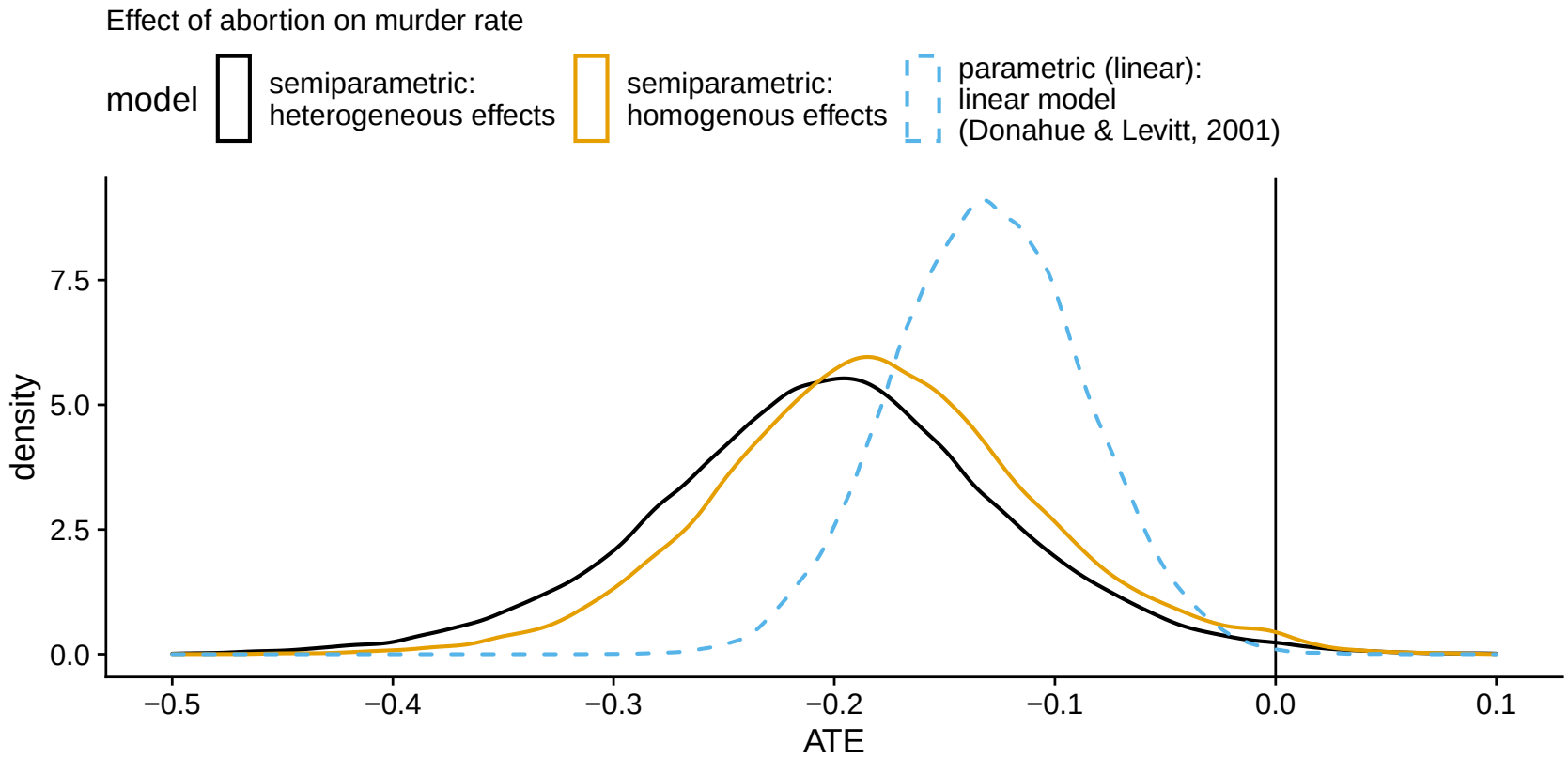
Model definition

$$y = \mu(x_C, s, t) + \tau(x_M, s, t) \cdot z + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2)$$

Comparison with Donohue and Levitt (2001); Belloni et al. (2014); Hahn et al. (2018); and others:

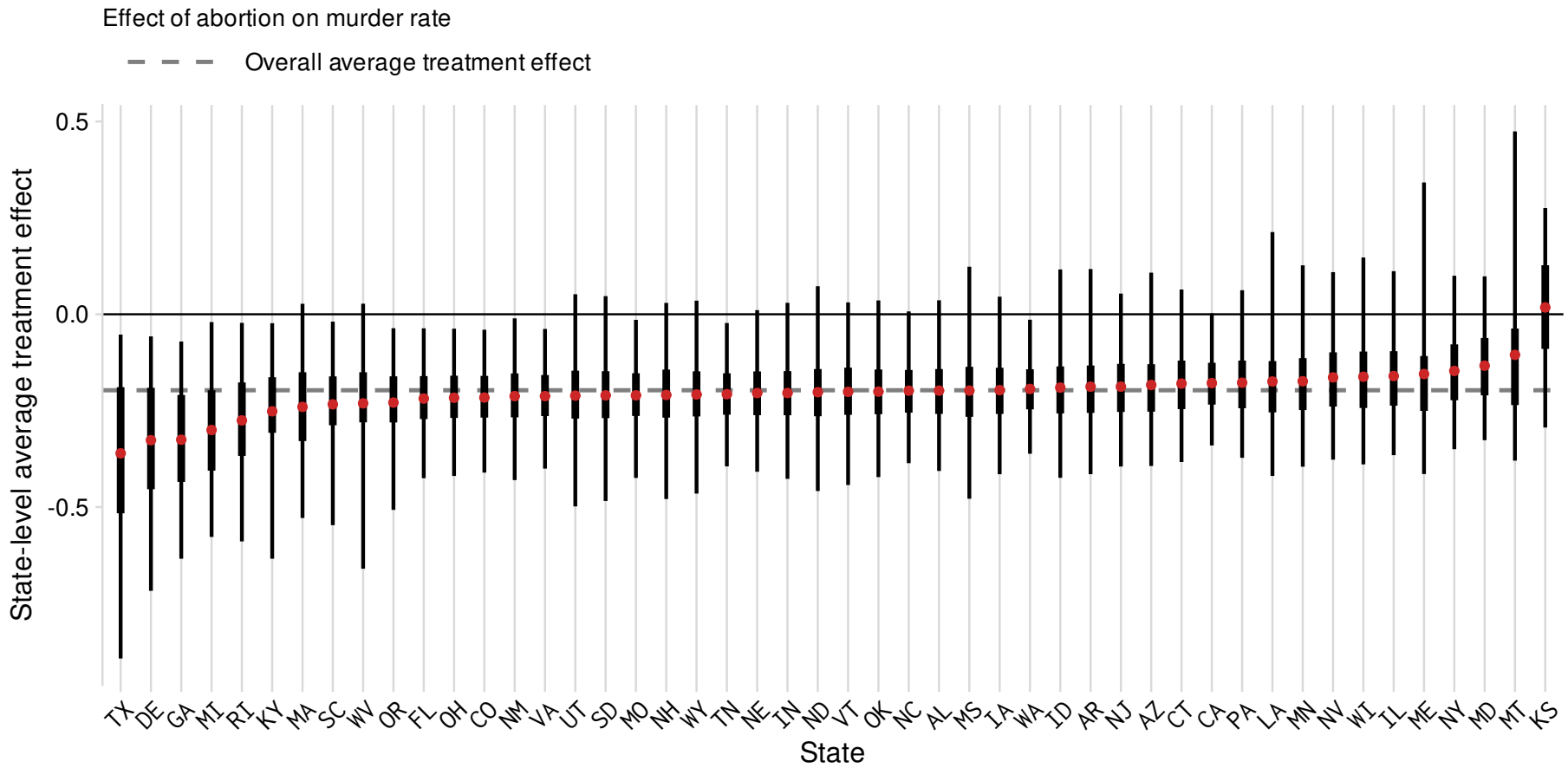
- **Commonality:** Assume linearity of y in z
- **Two departures:**
  - (i) No strict *a priori* parametric specification for controls
  - (ii) Effect heterogeneity through varying slope of treatment effect

ATE estimates



- ATE $= \bar{\tau} = N^{-1} \sum_{i=1}^{N} \tau(x_i)$
- Homogeneous effects model: $\tau(\cdot) \equiv \tau$ fixed

State-level ATEs



Characterizing effect heterogeneity

- High degree of heterogeneity between states
- What about heterogeneity driven by moderators?
- Variation in effect across time?

Posterior summary for effect modification

- Interpret nonparametric function $\tau(\cdot)$ via posterior summarization
- Project $\tau(\cdot)$ down onto a simpler (additive) structure:

$$\tau(\cdot) \approx \gamma(x_i, s_i, t_i) = \bar{\tau} + \sum_{k=1}^{47} b_s \cdot 1(s_i = k) + \sum_{j=1}^{5} h_j(x_{ij}) + h_6(t_i)$$

- Summary communicates treatment effect modification while averaging over possible interactions in $\tau(\cdot)$

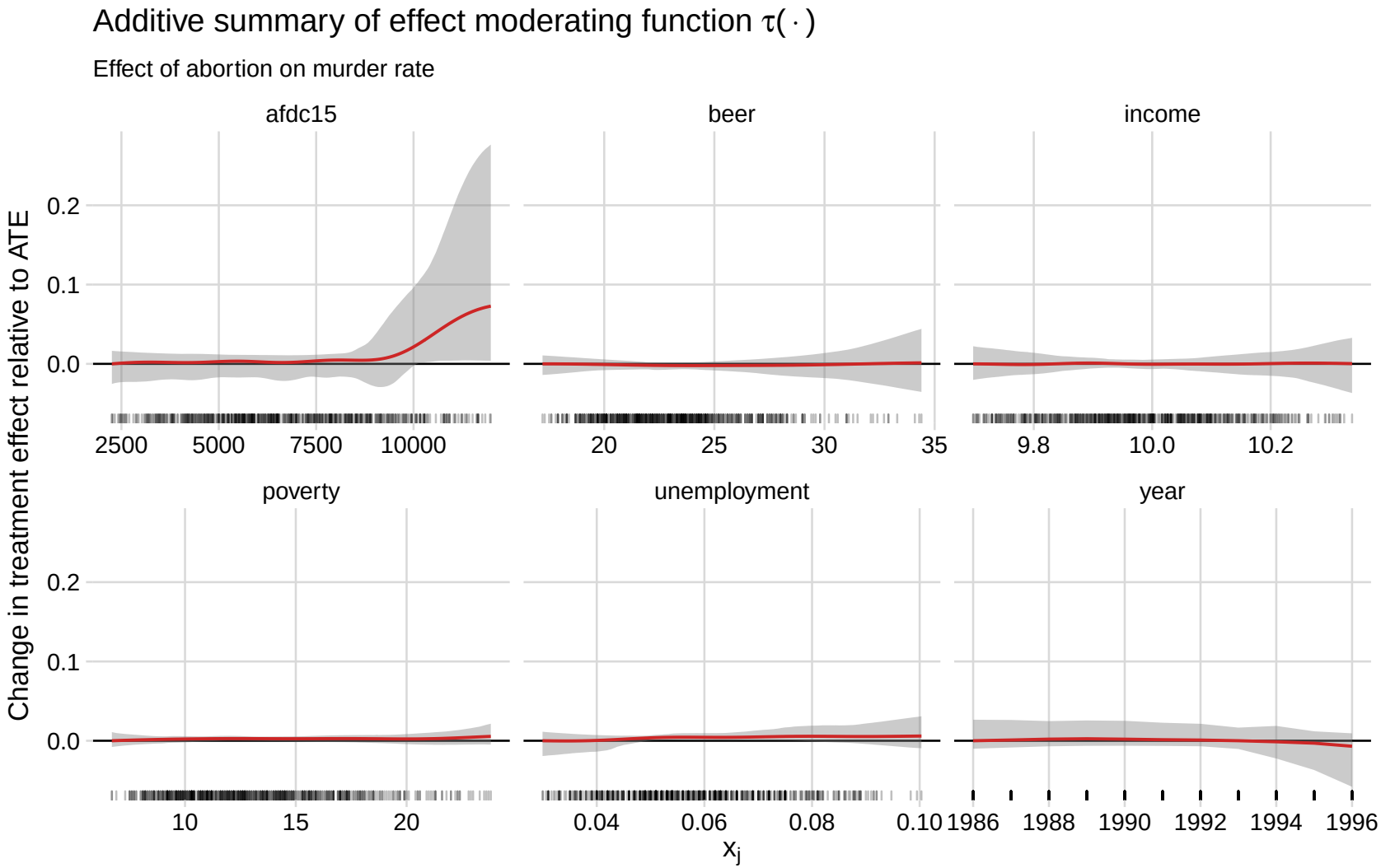
Posterior summary for effect modification

- Interpret nonparametric function $\tau(\cdot)$ via posterior summarization
- Project $\tau(\cdot)$ down onto a simpler (additive) structure:

$$\tau(\cdot) \approx \gamma(x_i, s_i, t_i) = \bar{\tau} + \sum_{k=1}^{47} b_s \cdot 1(s_i = k) + \sum_{j=1}^{5} h_j(x_{ij}) + h_6(t_i)$$

- Summary communicates treatment effect modification while averaging over possible interactions in $\tau(\cdot)$

Additive summary



Diagnostics of linearity assumption

Diagnostics of linearity assumption

Linear effects model:

$$y = \mu(x) + \tau(x) \cdot z + \varepsilon$$

Subtracting out $\mu(x)$ gives:

$$y - \mu(x) = \tau(x) \cdot z + \varepsilon$$

Diagnostics of linearity assumption

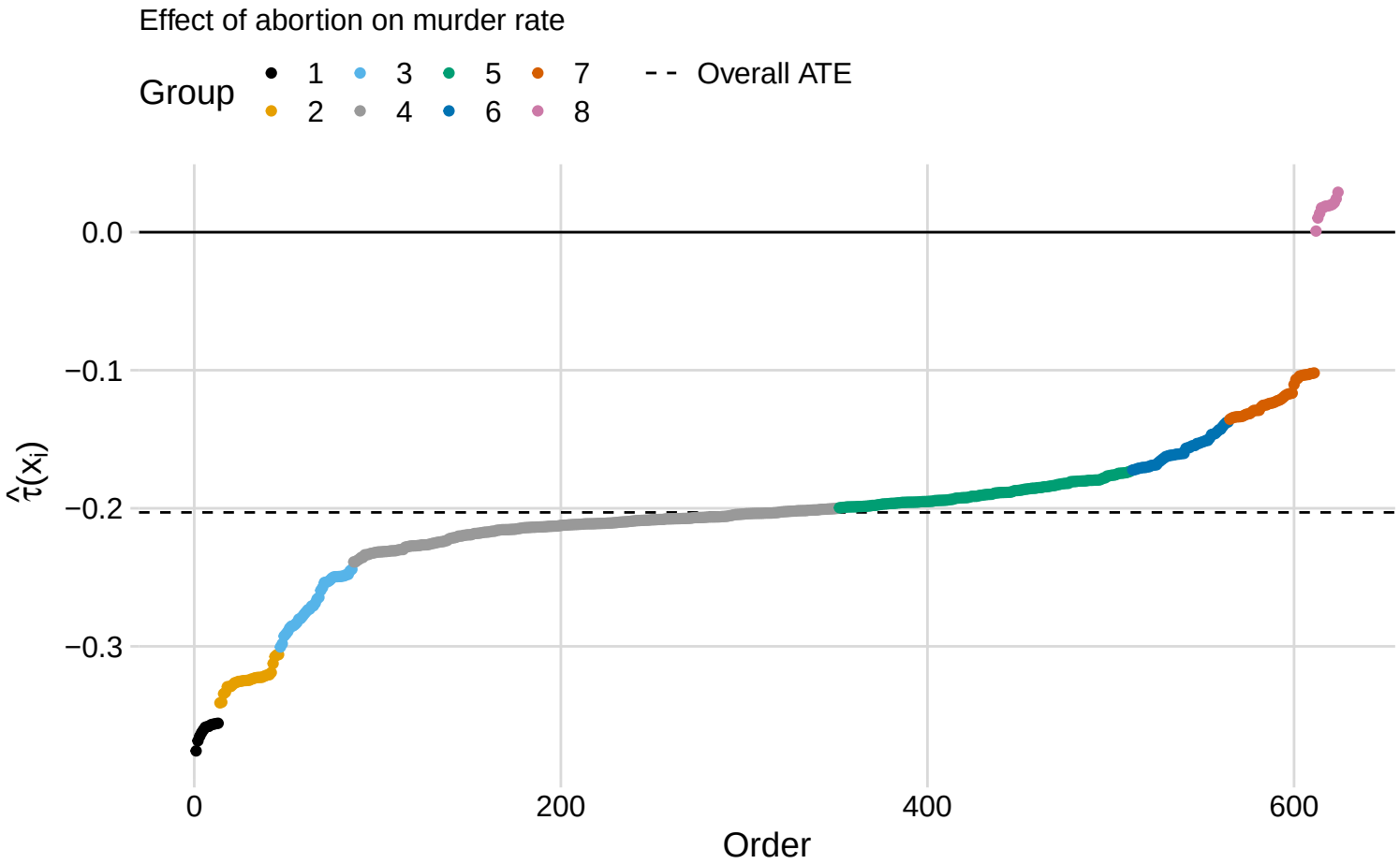
$$y - \mu(x) = \tau(x) \cdot z + \varepsilon$$

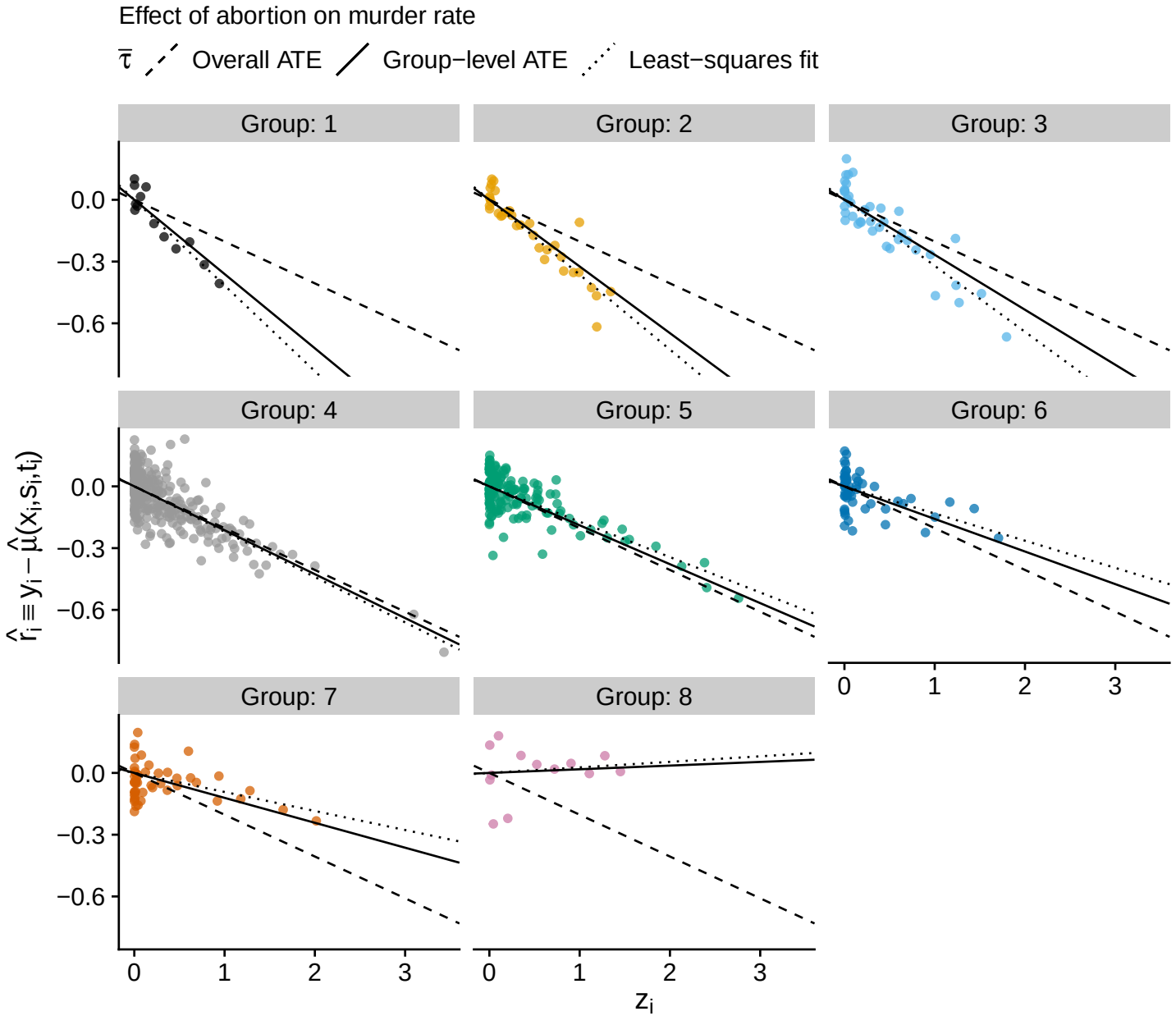
- **Idea:** Combine observations into J disjoint groups g_j such that $\hat{\tau}(x_i) \approx \hat{\tau}(x_{i'})$ for $i, i' \in g_j$, so then

$$\mathrm{E}[y_i - \hat{\mu}(x_i)] \approx \bar{\tau}_{g_j} \cdot z_i \quad \text{for} \quad i \in g_j$$

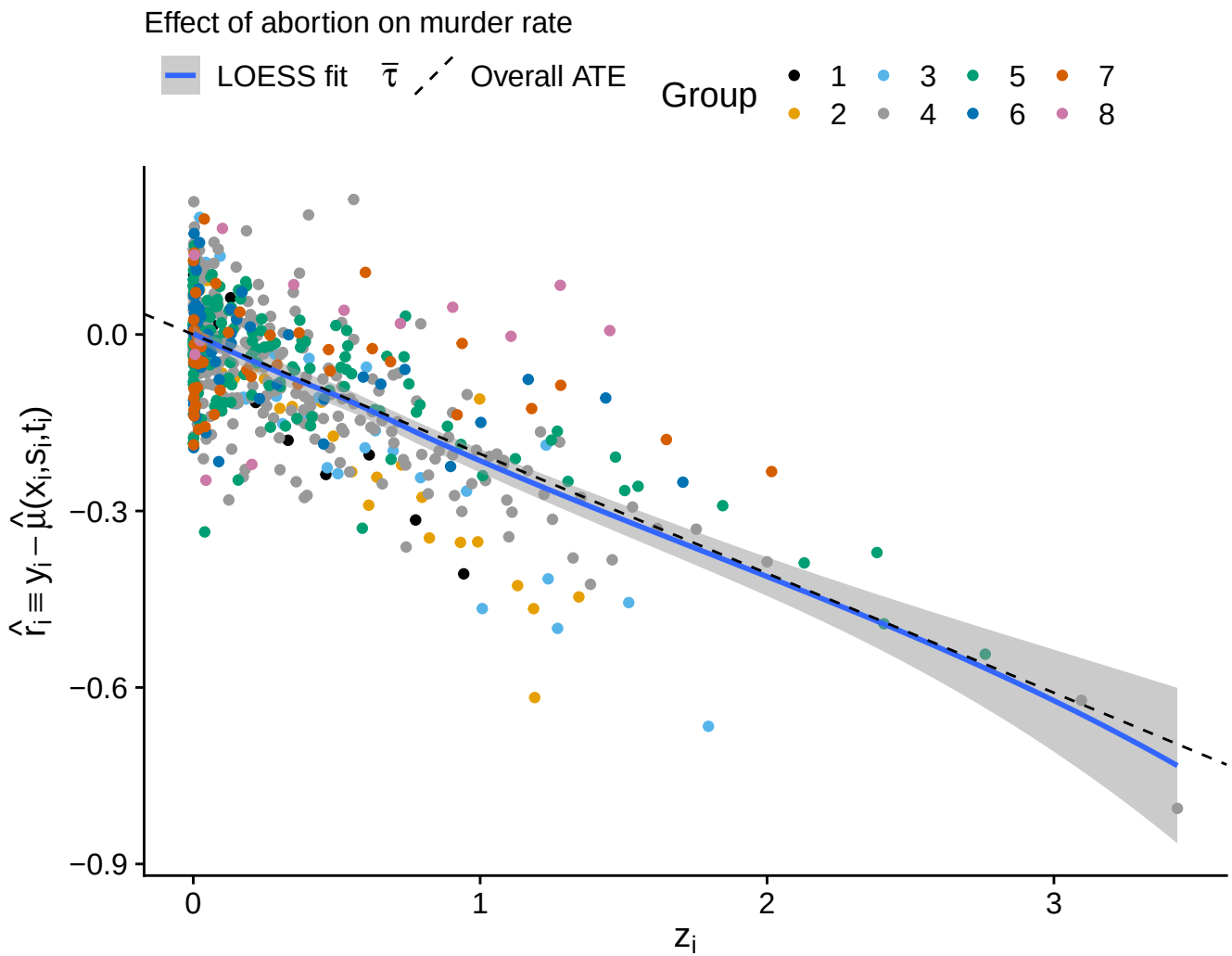
  where $\bar{\tau}_{g_j} = |g_j|^{-1} \sum_{i \in g_j} \hat{\tau}(x_i)$

- Then plot partial residuals $\hat{r}_i \equiv y_i - \hat{\mu}(x_i)$ against z_i to check for linearity within each group


Effect of abortion on murder rate
Group
1
2
3
4
5
6
7
8
Overall ATE
0.0
−0.1
−0.2
−0.3
$\hat{\tau}(x_i)$
0
200
400
600
Order


Effect of abortion on murder rate
$\bar{\tau}$ Overall ATE
Group−level ATE
Least−squares fit
Group: 1
Group: 2
Group: 3
Group: 4
Group: 5
Group: 6
Group: 7
Group: 8
0.0
−0.3
−0.6
0
1
2
3
$\hat{r}_i \equiv y_i - \hat{\mu}(x_i, s_i, t_i)$
z_i

Partial dose response curve



More analyses in the paper. . .

- Posterior summarization for subgroup identification
- Application to violent crime and property crime
- Simulation results for diagnostics
- ArXiv preprint: arxiv.org/abs/2007.09845

Conclusion

I. saFAB: *frequentist* selection-adjusted confidence sets, leverage Bayesian approach for shorter intervals[xiv]

II. Posterior summarization: *Bayesian* model post-hoc summaries for model interpretation[xv]

III. Valid Bayesian posteriors for treatment effects under nested models[xvi]

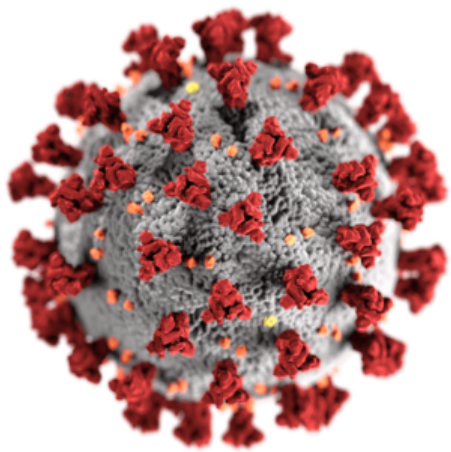
IV. Nonparametric control function, posterior summarization for interpreting heterogeneous treatment effects[xvii]

[xiv] Under revision for *Biometrika*
[xv] To appear in *JCGS*
[xvi] Submitting to *JBES* soon
[xvii] Recently submitted to *JASA ACS*

Future endeavours

- Postdoctoral fellow at lab of Prof. Lauren Ancel Meyers, UT-Austin Department of Integrative Biology
- Modeling spread of COVID-19 and other infectious diseases

meyers lab



Contact

- Email: spencer.woody@utexas.edu
- Website: spencerwoody.github.io

References I

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. Bart: Bayesian additive regression trees. *Ann. Appl. Stat.*, 4(1):266–298, 03 2010. doi: 10.1214/09-AOAS285. URL https://doi.org/10.1214/09-AOAS285.

John J. Donohue and Steven D. Levitt. The Impact of Legalized Abortion on Crime. *The Quarterly Journal of Economics*, 116(2):379–420, 05 2001. ISSN 0033-5533. doi: 10.1162/00335530151144050. URL https://doi.org/10.1162/00335530151144050.

John J Donohue and Steven D Levitt. The impact of legalized abortion on crime over the last two decades. Working Paper 25863, National Bureau of Economic Research, May 2019. URL http://www.nber.org/papers/w25863.

P. Richard Hahn, Carlos M. Carvalho, David Puelz, and Jingyu He. Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Anal.*, 13(1):163–182, 03 2018. doi: 10.1214/16-BA1044. URL https://doi.org/10.1214/16-BA1044.

P. Richard Hahn, Jared S. Murray, and Carlos M. Carvalho. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. *Bayesian Anal.*, 2020. doi: 10.1214/19-BA1195. URL https://doi.org/10.1214/19-BA1195. Advance publication.

Jennifer Hill, Antonio Linero, and Jared Murray. Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and Its Application*, 7:251–278, 2020.

Keisuke Hirano and Guido W. Imbens. *The Propensity Score with Continuous Treatments*, chapter 7, pages 73–84. John Wiley & Sons, Ltd, 2004. ISBN 9780470090459. doi: 10.1002/0470090456.ch7. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/0470090456.ch7.

Guido W. Imbens. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710, 2000. ISSN 00063444. URL http://www.jstor.org/stable/2673642.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.

References II

A. Newton. On a nonparametric recursive estimator of the mixing distribution. *Sankhyā Ser. A*, pages 306–322, 2002.

Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58, 1978.

Surya T. Tokdar, Ryan Martin, and Jayanta K. Ghosh. Consistency of a recursive estimate of mixing distributions. *The Annals of Statistics*, 37(5A):2502–2522, 2009. ISSN 00905364. URL http://www.jstor.org/stable/30243713.

Chi Wang, Francesca Dominici, Giovanni Parmigiani, and Corwin Matthew Zigler. Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models. *Biometrics*, 71(3): 654–665, 2015. doi: 10.1111/biom.12315. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12315.

Spencer Woody, Carlos Carvalho, and Jared Murray. Model interpretation through lower-dimensional posterior summarization. *Journal of Computational and Graphical Statistics*, 2020. doi: 10.1080/10618600.2020.1796684. Preprint at https://arxiv.org/abs/1905.07103.

C Yu and P D Hoff. Adaptive multigroup confidence intervals with constant coverage. *Biometrika*, 105(2):319–335, 2018. doi: 10.1093/biomet/asy009. URL http://dx.doi.org/10.1093/biomet/asy009.

Corwin Matthew Zigler and Francesca Dominici. Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects. *Journal of the American Statistical Association*, 109(505):95–107, 2014.

Extra slides...

saFAB recap

Quick recap: inverting a biased test

Construct confidence sets for θ_i only if $y_i \in S$. Truncated likelihood:

$$f_S(y;\theta) = f(y;\theta) \cdot \mathbf{1}(y \in S) / \int_S f(y;\theta)\mathrm{d}y$$

Inverting unbiased test gives confidence set:

$$A^S(\theta_0) = \{y : F_S^{-1}(\alpha/2;\theta_0) \leq y \leq F_S^{-1}(1-\alpha/2;\theta_0)\}$$
$$C^S(y) = \{\theta : y \in A^S(\theta)\}$$

$$\Pr_\theta(\theta \in C^S(y) \mid y \in S) = \Pr_\theta(y \in A^S(\theta)) = 1-\alpha$$

Quick recap: inverting a biased test

Inverting unbiased test gives confidence set:

$$A^S(\theta_0) = \{y : F_S^{-1}(\alpha/2; \theta_0) \leq y \leq F_S^{-1}(1-\alpha/2; \theta_0)\}$$
$$C^S(y) = \{\theta : y \in A^S(\theta)\}$$

Invert *biased test*, choose quantiles in $F_S(\cdot)$ according to *spending function* $w(\theta) : \mathbb{R} \to [0, 1]$

$$A^S_{w(\theta)}(\theta) = \{y : F_S^{-1}(\alpha w(\theta); \theta) \leq y \leq F_S^{-1}(\alpha w(\theta) + 1 - \alpha; \theta)\}$$
$$C^S_{w(\theta)}(y) = \{\theta : y \in A^S_{w(\theta)}(\theta)\}$$

Quick recap: inverting a biased test

- Choose *optimal* $w(\theta)$ such that confidence sets are as short as possible on average under assumed prior $\pi(\theta)$
- Define frequentist risk

$$R(\theta; w) = \int \int \mathbf{1}(y \in A_w^S(\tilde{\theta})) f_S(y; \theta) \mathrm{d}\tilde{\theta} \mathrm{d}y.$$

- Bayes loss is

$$L(\pi, w) = \int R(\theta; w)\pi(\theta)\mathrm{d}\theta = \ldots = \int \Pr(Y \in A_w^S(\tilde{\theta}))\mathrm{d}\tilde{\theta}$$

- Finding *optimal spending function* is a *minimization problem* with objective function

$$w^\star(\theta) = \arg \min_{w \in [0,1]} H(w; \theta)$$

$$H(w; \theta) \equiv \Pr(Y \in A_w^S(\theta))$$

$$= M_S \left[F_S^{-1}(\alpha w + 1 - \alpha; \theta) \right] - M_S \left[F_S^{-1}(\alpha w; \theta) \right]$$

Nonparametric empirical Bayes saFAB procedure

- Calculating the *optimal spending function* depends on the prior $\pi(\theta)$ through the marginal density M_S

$$w^\star(\theta) = \arg\min_{w \in [0,1]} H(w;\theta)$$

$$H(w;\theta) = M_S\left[F_S^{-1}(\alpha w + 1 - \alpha;\theta)\right] - M_S\left[F_S^{-1}(\alpha w;\theta)\right]$$

- When the prior is unknown, we can use *predictive recursion*[xviii] *to form an estimate* $\hat{\pi}(\theta)$ from the observed data $y_1, \ldots, y_N$
- This gives the *estimated optimal spending function* which minimizes a surrogate objective function

$$\hat{w}^\star(\theta) = \arg\min_{w \in [0,1]} \hat{H}(w;\theta)$$

$$\hat{H}(w;\theta) = \hat{M}_S\left[F_S^{-1}(\alpha w + 1 - \alpha;\theta)\right] - \hat{M}_S\left[F_S^{-1}(\alpha w;\theta)\right]$$

[xviii] Newton (2002); Tokdar et al. (2009)

Convergence of empirical Bayes procedure

$$\underbrace{\Omega_0 = \left\{ w : \min_{\tilde{w} \in [0,1]} H(\tilde{w}) = H(w) \right\}}_{\text{Set of true minimizers}} \qquad \underbrace{\hat{\Omega} = \left\{ w : \min_{\tilde{w} \in [0,1]} \hat{H}(\tilde{w}) = \hat{H}(w) \right\}}_{\text{Set of estimated minimizers}}$$

Theorem: Main consistency result

Under general regularity conditions, the following holds:

(i) Convergence of estimated spending function to true optimal function:

$$\Pr\left(\sup_{\hat{w}^\star \in \hat{\Omega}} \inf_{w^\star \in \Omega_0} \left| \hat{w}^\star - w^\star \right| \geq \epsilon \right) \to 0.$$

(ii) Convergence of objective function:

$$H(\hat{w}^\star) \xrightarrow{\text{a.s.}} H(w^\star) \text{ for some } \hat{w}^\star \in \hat{\Omega}, w^\star \in \Omega_0$$

GP model for CA housing data

Gaussian process model for CA housing data

$$
\begin{aligned}
(y_i \mid f, \sigma^2) &= f(x_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}(0, \sigma^2) \\
f &\sim \mathsf{GP}(0, k(\cdot, \cdot)) \\
p(\sigma^2) &\propto \sigma^{-2} \\
k(x_i, x_{i'}) &= \tau^2 \cdot \exp\left(- \sum_{j=1}^{p} [x_{ij} - x_{i'j}]^2 / v_j \right) + \sum_{j=1}^{p} a_j x_{ij} x_{i'j}
\end{aligned}
$$

Local linear summaries

Local linear summaries

Characterize local behavior of f to answer the question:

How do the determinants of housing prices vary geographically?

Choose **3 metropolitan areas** (MSA's, defined by their counties) in California from **south**, **north**, and **central** regions to compare:

- Greater Los Angeles (LA & Orange Counties)
- Fresno (Fresno County)
- Bay Area (San Francisco and San Mateo Counties)

Local linear summaries

- Greater Los Angeles (LA & Orange Counties)
- Fresno (Fresno County)
- Bay Area (San Francisco and San Mateo Counties)

Summary of f at four levels of resolutions:

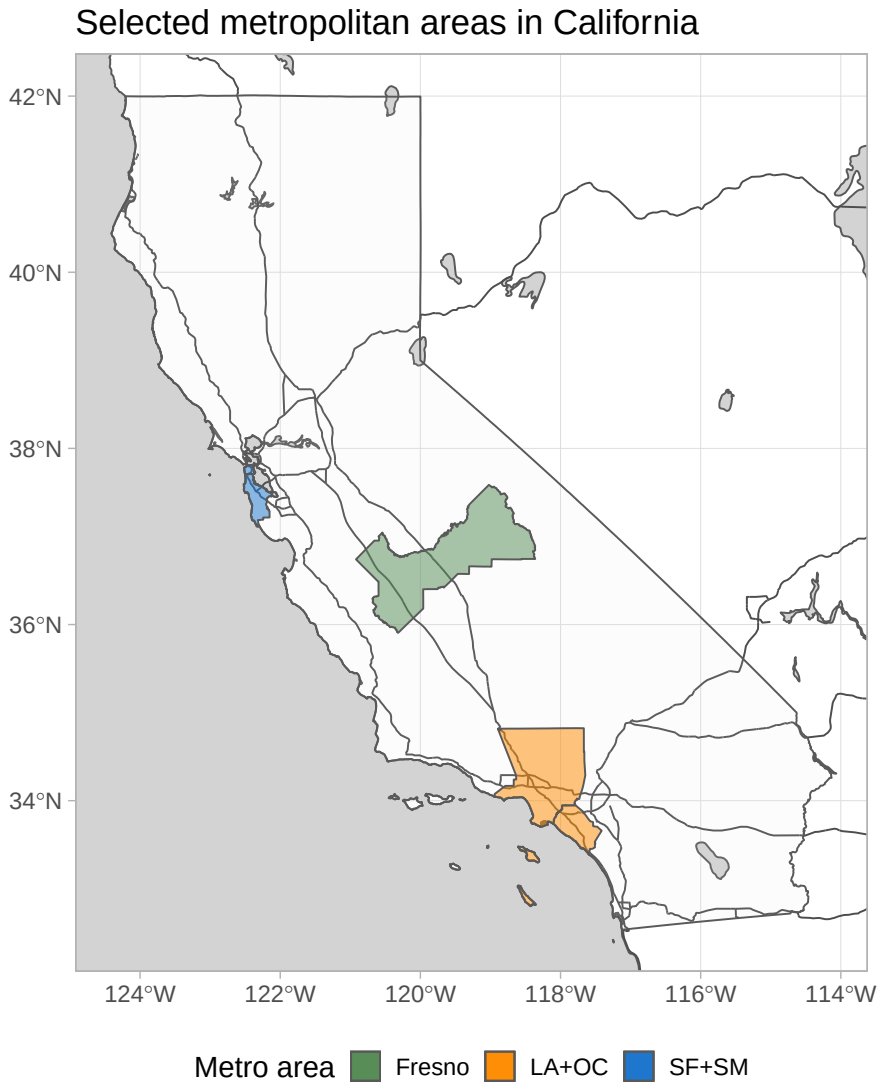
(i) Metropolitan area (compare across MSA's)
(ii) County (compare across and within MSA's)
(iii) Neighborhood (group of tracts; compare within a county)
(iv) Individual tract

Local linear summaries

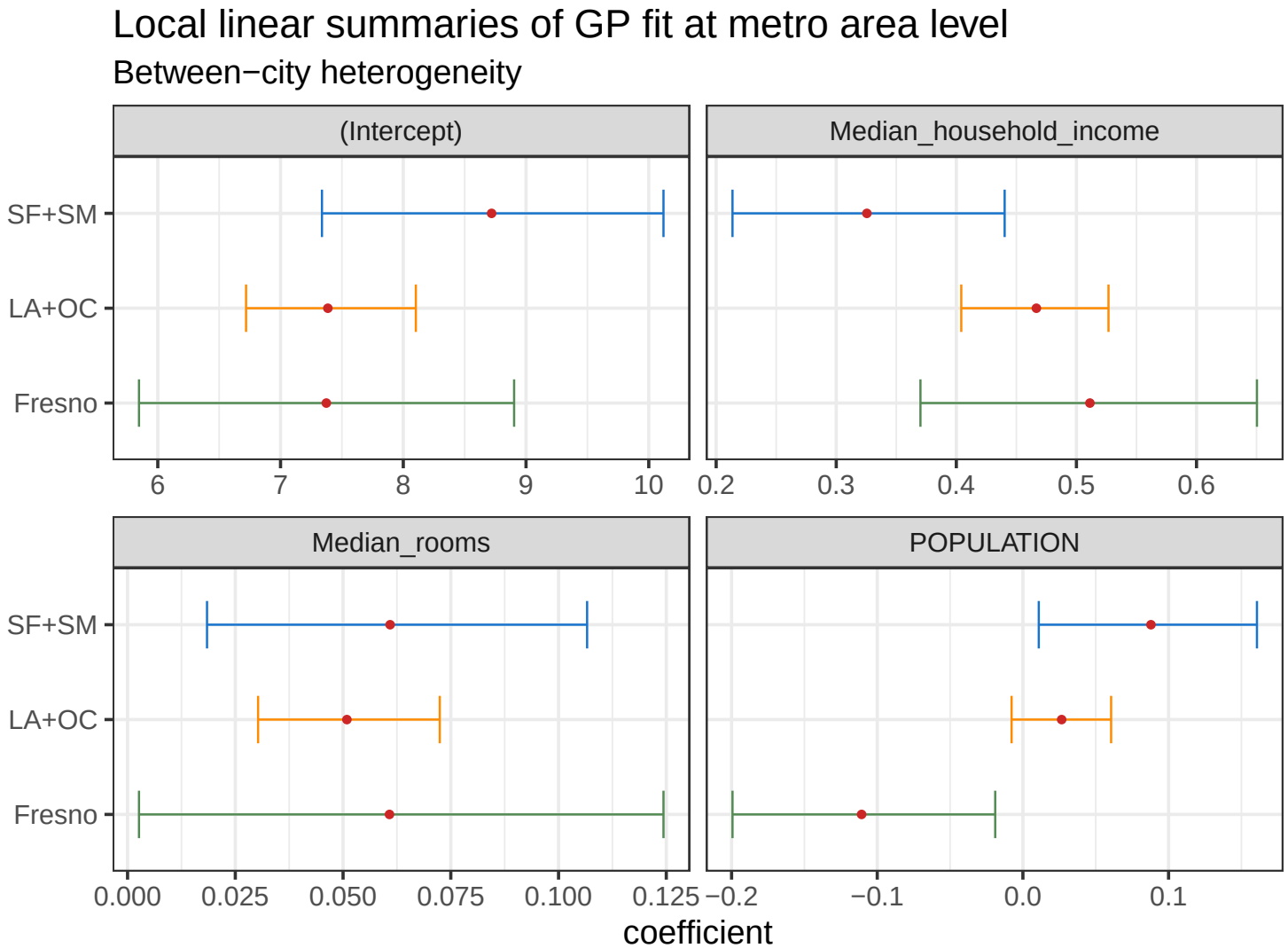
Create synthetic data $\tilde{X}_m$ for each region m by sampling data within neighborhood of x_{m0}

Linear summary for predictions made by model at points $f(\tilde{x}_{mi})$

MSA-level local linear summaries

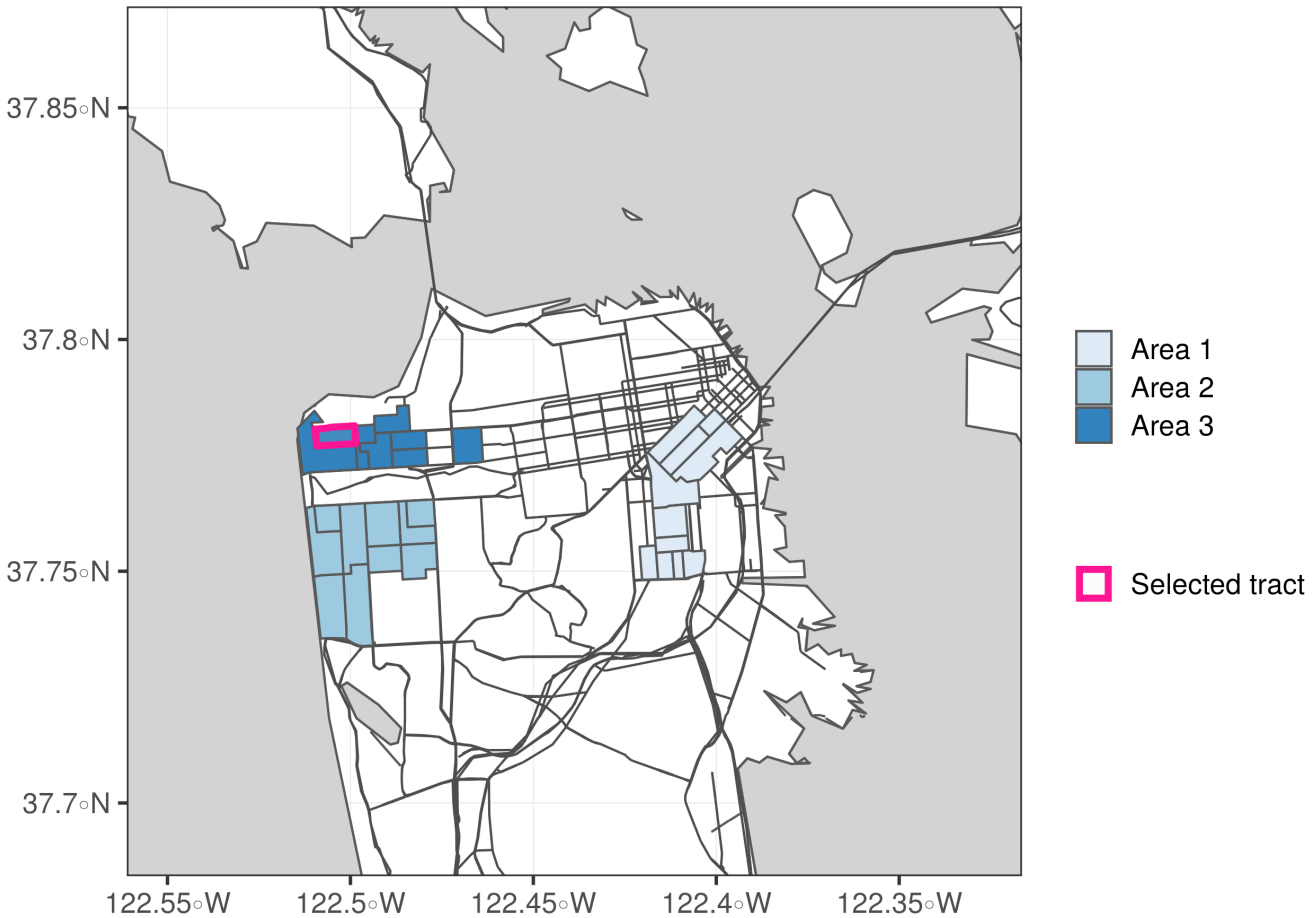


MSA-level local linear summaries

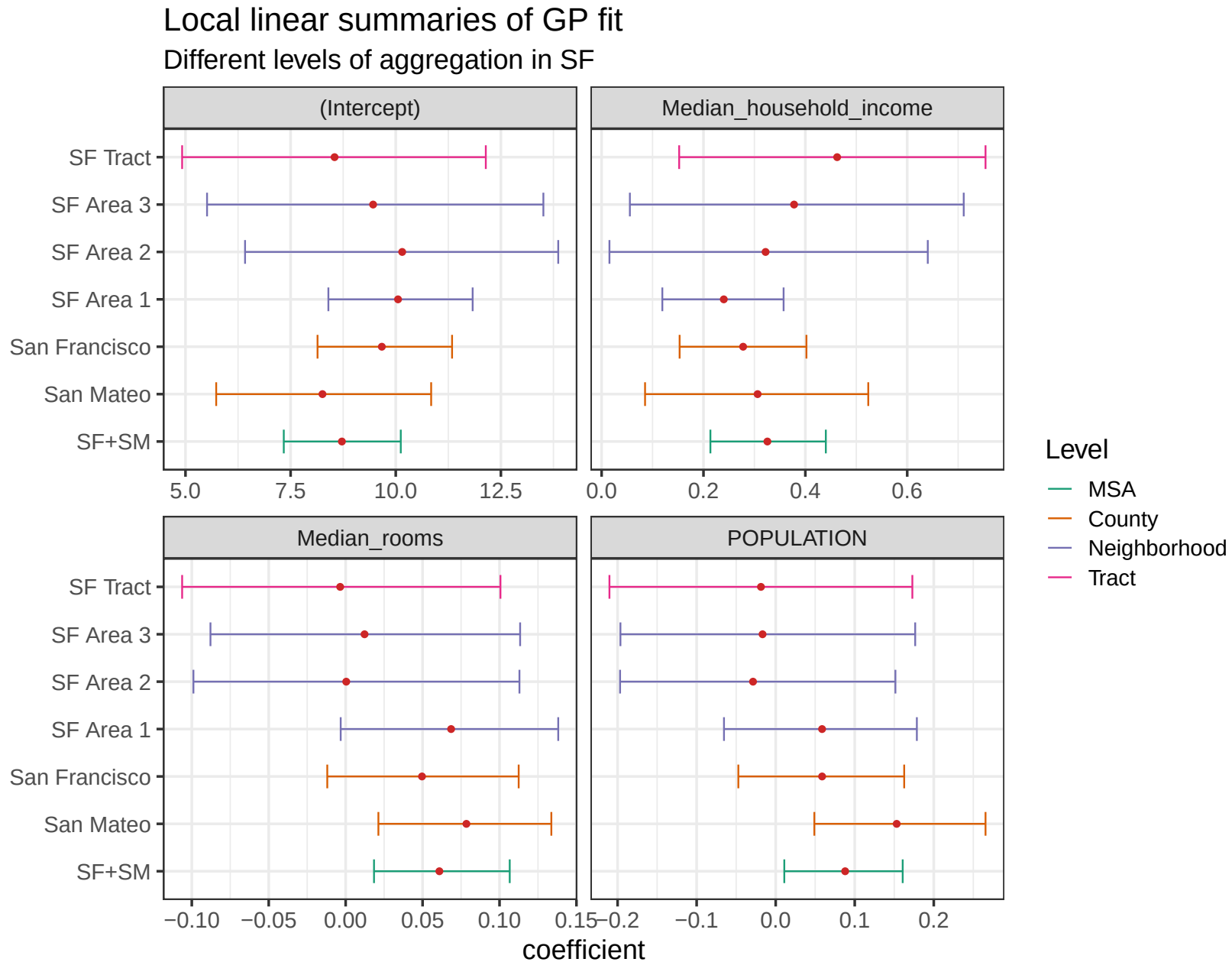


Local linear summaries in San Francisco

Selected areas for summarization

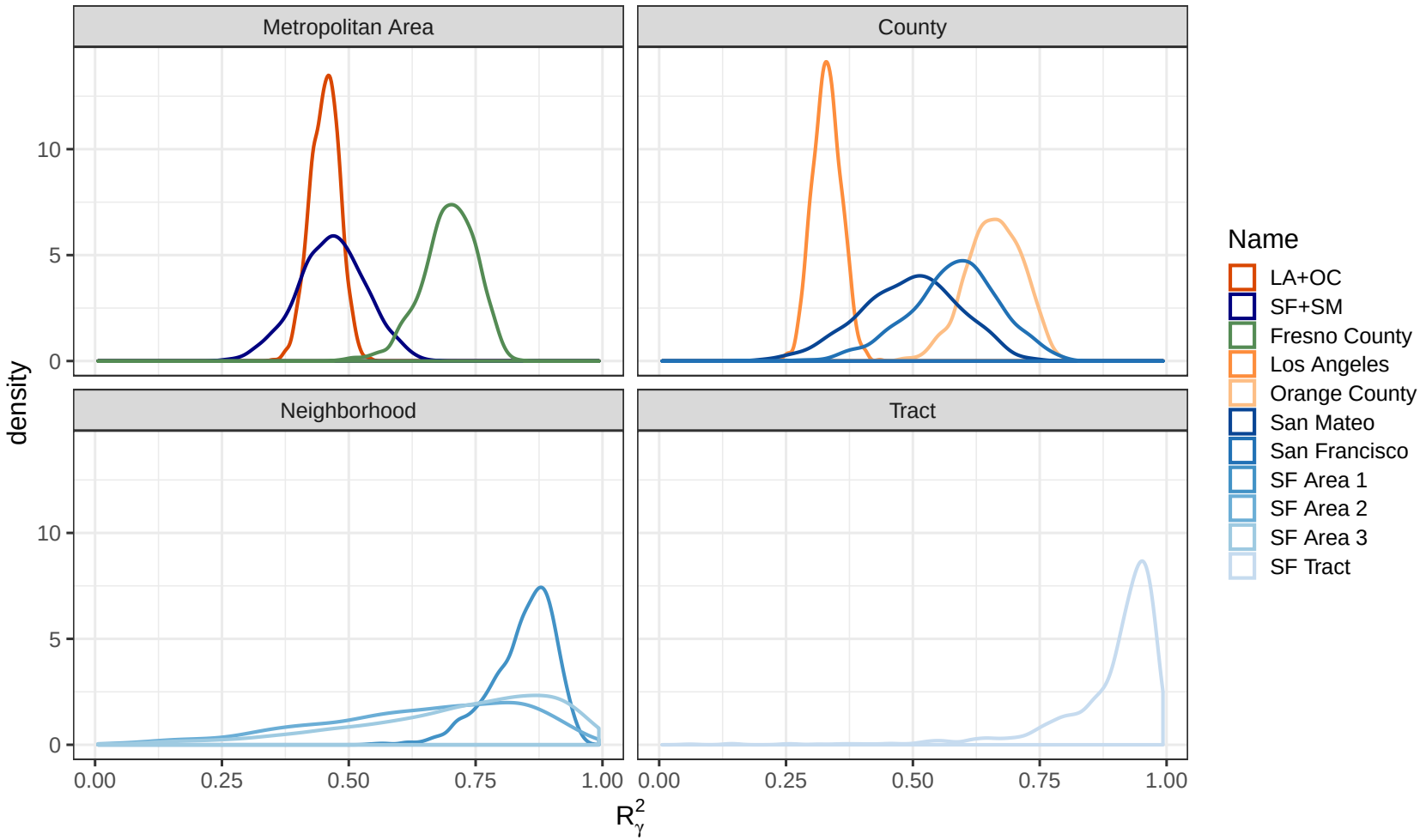


Local linear summaries in San Francisco



Local linear summary diagnostics

R^2_γ



Calculating projected posterior treatment effects

Projected posterior treatment effects

(I) *Posterior for "full" model*

$$(Y \mid Z, X) = \tau Z + X\beta + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma_\epsilon^2)$$

- Define. . .
  - concatenated data matrix $W = [Z \quad X]$
  - condensed coefficient vector $\psi = [\tau \quad \beta^\intercal]^\intercal$

  so outcome model becomes

$$Y = W\psi + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma_\epsilon^2).$$

- Obtain posterior for ψ

Projected posterior treatment effects

(II) *Inference under submodels*

Consider nested subset of controls described by $\phi \in \{0,1\}^p$, $|\phi| < p$

- Restricted covariate matrix $X_\phi = \{X_j\}_{j \in \phi}$, coefficient vector β_ϕ
- $W_\phi = [Z \quad X_\phi]$ and $\psi_\phi = [\tau \quad \beta_\phi]$
- **Goal**: Compare model specifications

$$\mathcal{M}_0 : Y = W\psi + \epsilon$$
$$\mathcal{M}_\phi : Y = W_\phi\psi_\phi + \epsilon$$

Projected posterior treatment effects

- **Goal**: Compare model specifications

$$\mathcal{M}_0 : Y = W\psi + \epsilon$$
$$\mathcal{M}_\phi : Y = W_\phi\psi_\phi + \epsilon$$

- *Projected posterior* for τ under the ϕ subset of controls:

$$\psi_\phi = (W_\phi^\intercal W_\phi)^{-1} W_\phi^\intercal W\psi$$

Projected posterior treatment effects

- **Goal**: Compare model specifications

$$\mathcal{M}_0 : Y = W\psi + \epsilon$$
$$\mathcal{M}_\phi : Y = W_\phi\psi_\phi + \epsilon$$

- *Projected posterior* for τ under the ϕ subset of controls:

$$\psi_\phi^{[k]} = (W_\phi^\intercal W_\phi)^{-1} W_\phi^\intercal W\psi^{[k]}, \quad \psi^{[k]} \sim p(\psi \mid y)$$

(using Monte Carlo draws from posterior,

take first element of ψ_ϕ)

Toy example

Generate data from

$$(Z \mid X) = X\gamma + \nu, \quad \nu \sim \mathrm{N}(0, 1)$$
$$(Y \mid Z, X) = \tau Z + X\beta + \epsilon, \quad \epsilon \sim \mathrm{N}(0, 1)$$

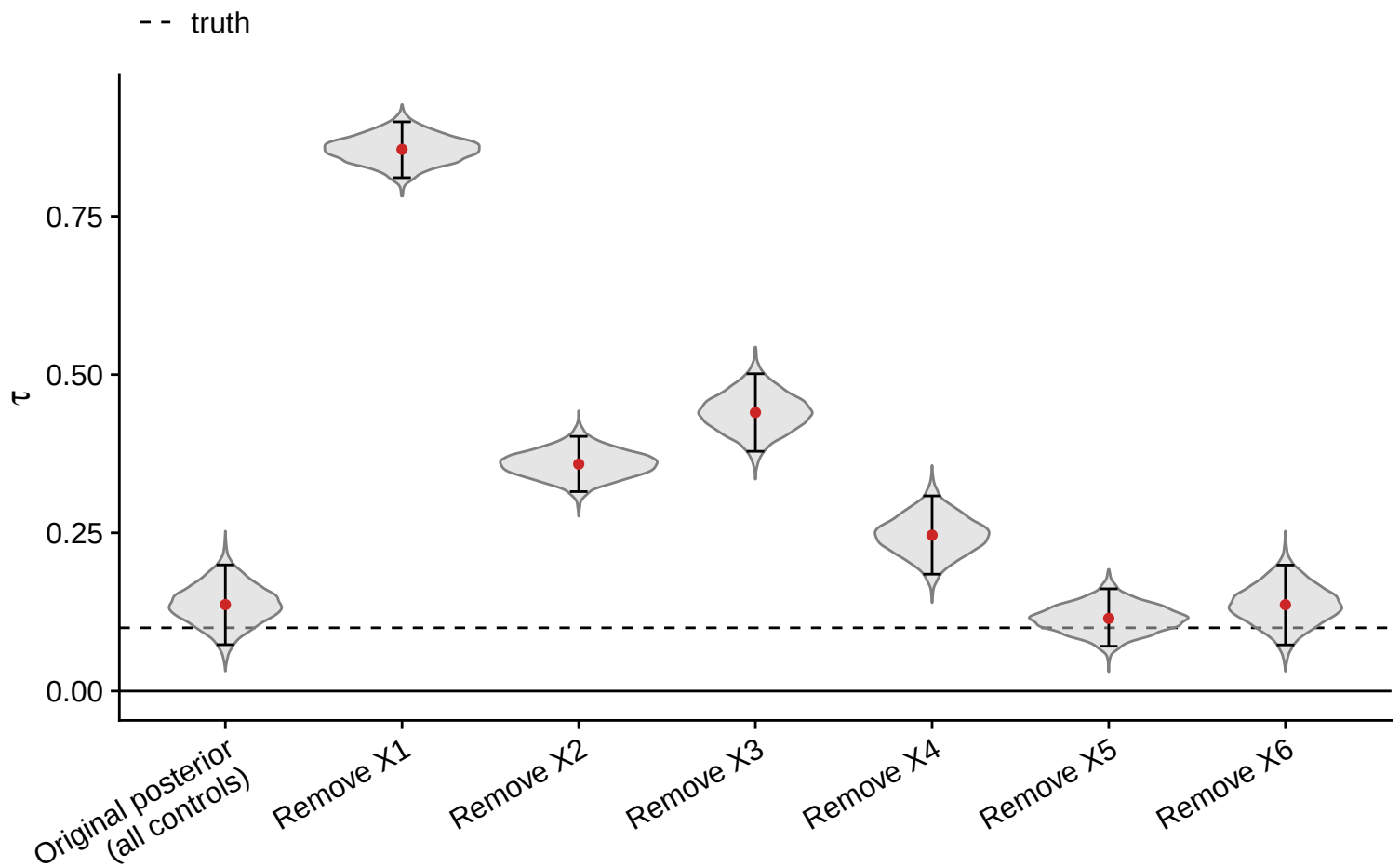
with $X \sim \mathrm{N}(0, I)$

$$\gamma = \begin{bmatrix} 1.0 & 1.0 & 0.2 & 0.2 & 1.0 & 0.0 \end{bmatrix}^\intercal$$
$$\beta = \begin{bmatrix} 1.5 & 0.5 & 1.5 & 0.5 & 0.0 & 0.0 \end{bmatrix}^\intercal$$
$$\tau = 0.1.$$

Determinant of...	X_1	X_2	X_3	X_4	X_5	X_6
Exposure Outcome	Strong Strong	Strong Weak	Weak Strong	Weak Weak	Strong None	None None
Variable type	Strong conf.	Weak conf.	Weak conf.	Weaker conf.	Instrument	Noise

Toy example

Determinant of...	X_1	X_2	X_3	X_4	X_5	X_6
Exposure	Strong	Strong	Weak	Weak	Strong	None
Outcome	Strong	Weak	Strong	Weak	None	None
Variable type	Strong conf.	Weak conf.	Weak conf.	Weaker conf.	Instrument	Noise



Stepwise procedure

Stepwise procedure

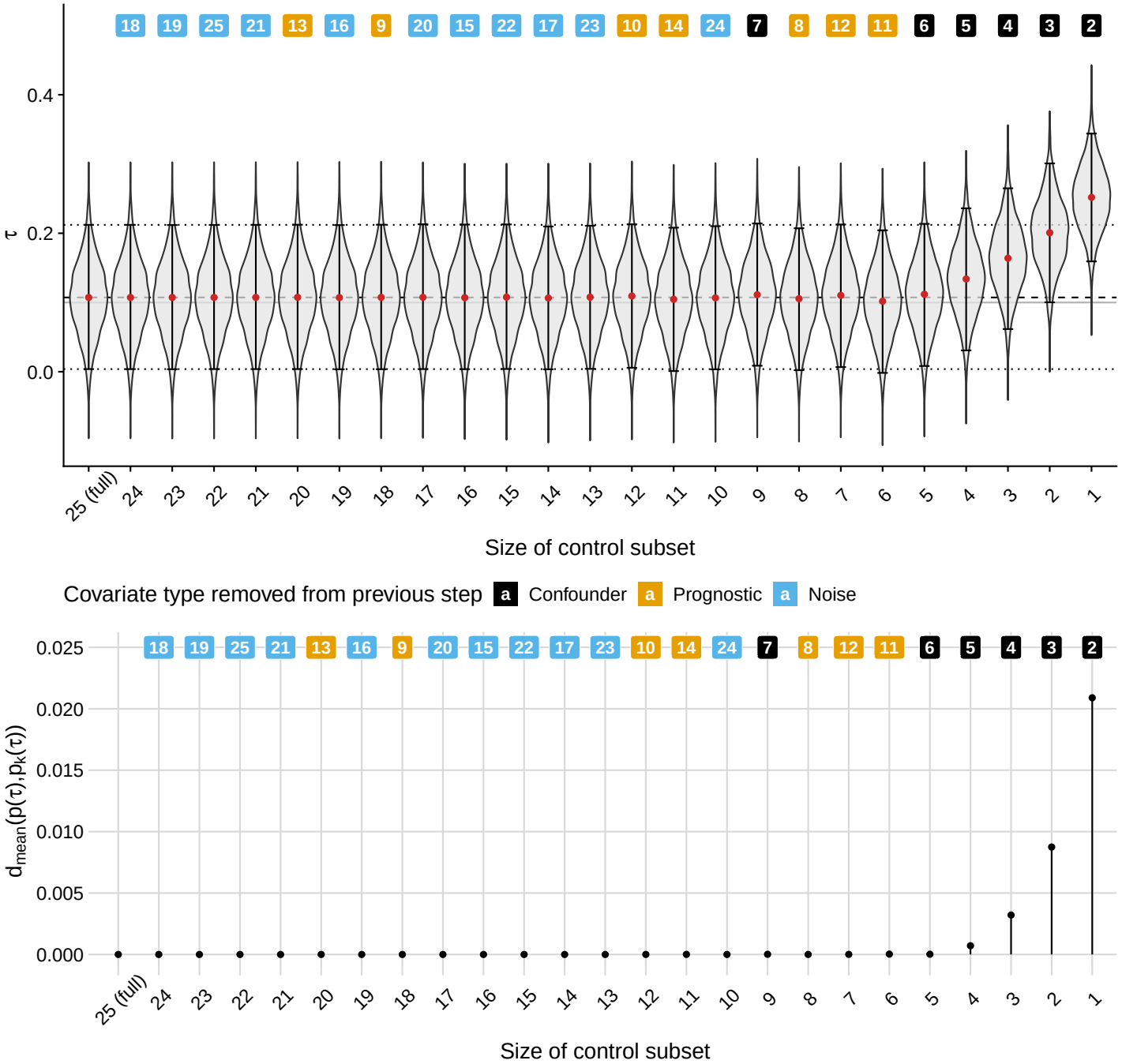
- Can remove control terms in a backward stepwise manner
- Remove one control at a time in a way that minimizes change in posterior mean for τ
- This ranks the control terms according to apparent confoundingness

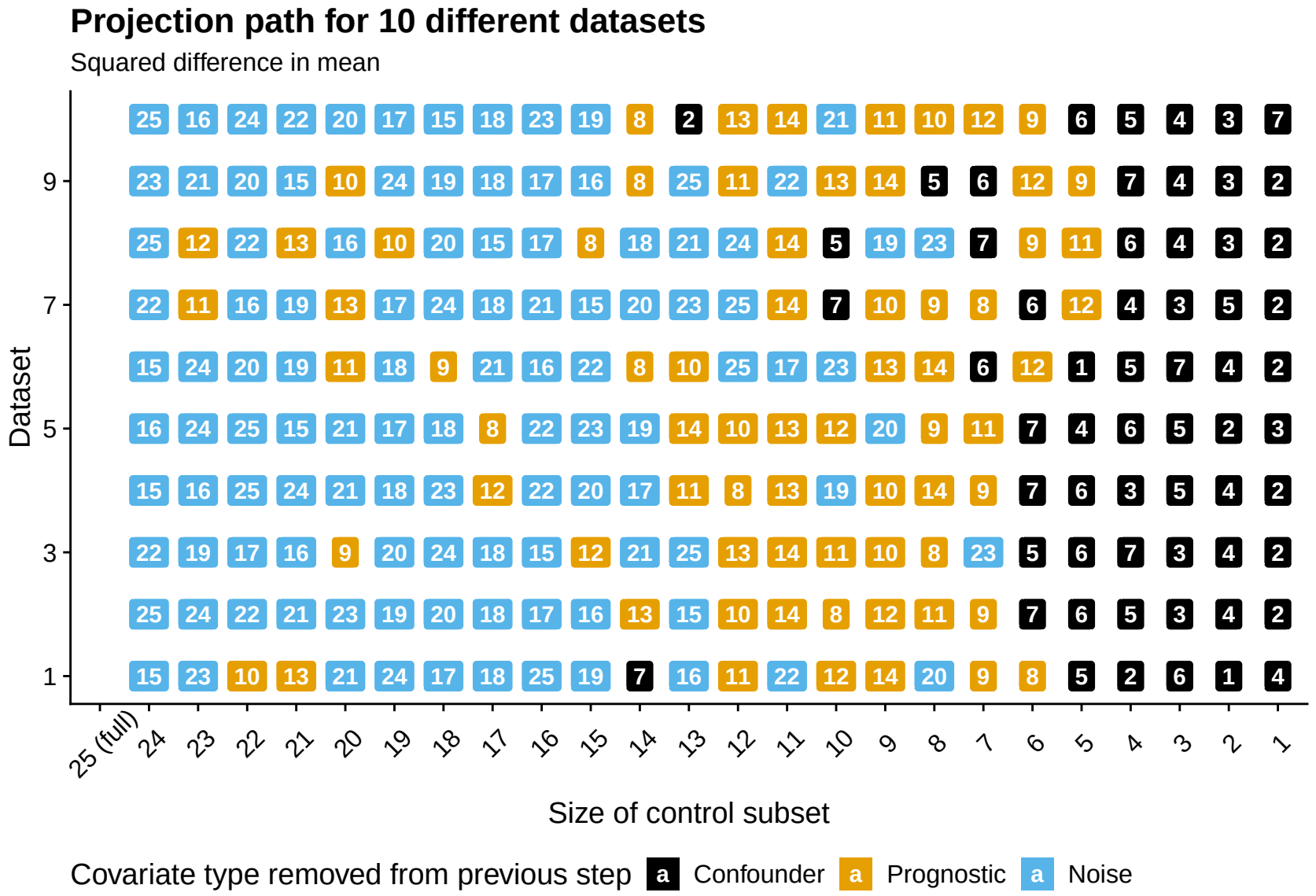
Stepwise procedure: simulation example

We generate $n = 1000$ observations from the model

$$Y = \tau Z + \beta_1 X_1 + \ldots + \beta_{14} X_{14} + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0, 1)$$
$$\tau = \beta_1 = \ldots = \beta_{14} = 0.1$$

- $X_1, \ldots, X_7$: confounders
- $X_8, \ldots, X_{14}$: prognostic variables
- $X_{15}, \ldots, X_{25}$: noise variables


18 19 25 21 13 16 9 20 15 22 17 23 10 14 24 7 8 12 11 6 5 4 3 2
τ
0.4
0.2
0.0
25 (full) 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
Size of control subset
Covariate type removed from previous step
Confounder
Prognostic
Noise
0.025
0.020
0.015
0.010
0.005
0.000
$d_{mean}(p(\tau), p_k(\tau))$
Size of control subset


Projection path for 10 different datasets
Squared difference in mean
Dataset
9
7
5
3
1
25 16 24 22 20 17 15 18 23 19 8 2 13 14 21 11 10 12 9 6 5 4 3 7
23 21 20 15 10 24 19 18 17 16 8 25 11 22 13 14 5 6 12 9 7 4 3 2
25 12 22 13 16 10 20 15 17 8 18 21 24 14 5 19 23 7 9 11 6 4 3 2
22 11 16 19 13 17 24 18 21 15 20 23 25 14 7 10 9 8 6 12 4 3 5 2
15 24 20 19 11 18 9 21 16 22 8 10 25 17 23 13 14 6 12 1 5 7 4 2
16 24 25 15 21 17 18 8 22 23 19 14 10 13 12 20 9 11 7 4 6 5 2 3
15 16 25 24 21 18 23 12 22 20 17 11 8 13 19 10 14 9 7 6 3 5 4 2
22 19 17 16 9 20 24 18 15 12 21 25 13 14 11 10 8 23 5 6 7 3 4 2
25 24 22 21 23 19 20 18 17 16 13 15 10 14 8 12 11 9 7 6 5 3 4 2
15 23 10 13 21 24 17 18 25 19 7 16 11 22 12 14 20 9 8 5 2 6 1 4
25 (full) 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
Size of control subset
Covariate type removed from previous step
a Confounder
a Prognostic
a Noise