Model Interpretation through Posterior Summarization

Spencer Woody* Carlos M. Carvalho Jared S. Murray

October 11, 2019

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

TEXAS McCombs
The University of Texas at Austin
McCombs School of Business
Information, Risk, & Operations Management

Introduction

Consider a generic regression model:

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

Suppose we have enough data to estimate f with a nonparametric model.

But we also want to understand how f makes predictions, e.g.

- Which covariates have strongest effect on prediction?
- Does covariate importance differ across the covariate space?
- Are there important interactions?

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
  - $y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i$
  - $y_i = \beta_0 + \sum_{j=1}^{p} g_j(x_{ij}) + \varepsilon_i$

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:

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...

- are suited to answer relevant inferential questions, and
- sufficiently represent the model's predictions

Motivating example: GP model for housing prices

Regress California census tract-level log-median house value on. . .

- log-median household income
- log-population
- median number of rooms per unit
- longitude
- latitude

$n = 7481$, full model:

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

Kernel:

$$k(x_i, x_{i'}) = \tau^2 \cdot \exp\left(-\sum_{j=1}^{p} [x_{ij} - x_{i'j}]^2 / v_j\right) + \sum_{j=1}^{p} a_j x_{ij} x_{i'j}$$

Motivating example: GP model for housing prices

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

This model allows for...

- Nonlinearity
- Nonstationarity
- Interactive effects

Motivating example: GP model for housing prices

1. **Global summaries**
   Average predictive trends across whole dataset,
   where we approximate $f(x)$ with...
   (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**
   Covariate importance within geographic regions

Advantages of our approach

- Can describe both global and local model behavior
- Easier to calculate than existing alternatives, e.g., partial dependence plots
- Summaries come with estimates of posterior uncertainty
- The data are *used only once* (in finding posterior for f) $\rightarrow$ retain valid Bayesian inference even after fitting several summaries
- Rooted in Bayesian decision theory*

*See Hahn and Carvalho (2015); MacEachern (2001)

Model summaries using decision theory

- Assume that we have posterior samples for f
- Action space lower-dimensional class of summary functions Γ
- The optimal summary minimizes the posterior expected loss

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

- 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
  - $\tilde{X}$ are covariate locations of interest
  - $p_{\lambda}(\cdot)$ penalizes complexity in γ

Optimal model summaries

The point estimate for the optimal model summary 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 $\hat{f}(\tilde{x}_i)$.

Global additive summary for GP model

Summary class is set of additive functions, using splines†:

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

The optimal point estimate for the summary is

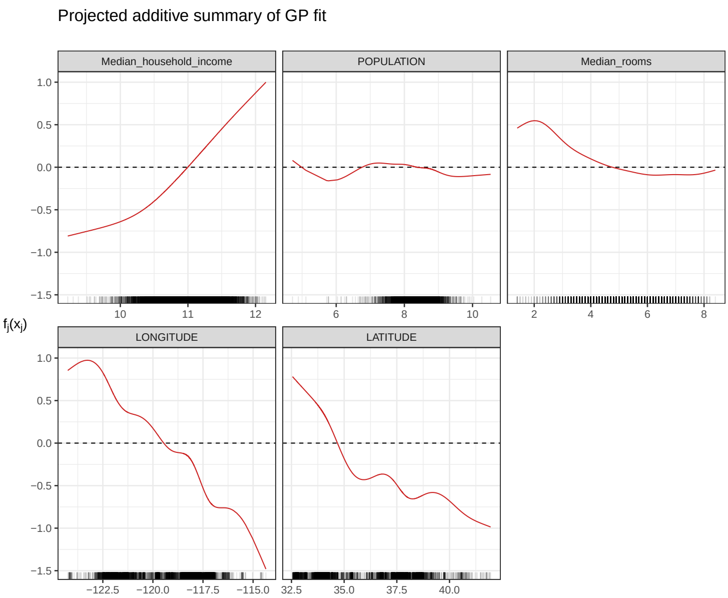
$$\hat{\gamma}(x) = \arg\min_{\gamma \in \Gamma} \sum_{i=1}^{n} \left[\hat{f}(x_i) - \gamma(x_i) \right]^2 + \sum_{j=1}^{p} \lambda_j \cdot J(h_j), \qquad (1)$$

penalty function has terms $J(h_j) = \int h_j''(t)^2 dt$‡.

Best additive approximation to the model

†See Wood (2017)
‡Fit Eq. (1) using GLS, use LOOCV analog for λ_j

Global additive summary for housing price model



Projected posteriors for summary uncertainty

- Optimal point estimate for summary

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

- For posterior uncertainty, we propose using draws of γ using the functional

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

  using posterior draws of f

- Often these are projections (e.g. least squares) of Monte Carlos draws of vector $\{f(\tilde{x}_i)\}_{i=1}^{\tilde{n}}$

Projected posteriors for summary uncertainty

This approach is standard in Bayesian practice, i.e.

(i) Specify model

$$\theta \sim \pi(\theta)$$

$$(y \mid \theta) \sim f(y \mid \theta)$$

(ii) Obtain posterior

$$\pi(\theta \mid y)$$

(iii) This implies a posterior distribution for functionals of θ

$$\pi(g(\theta) \mid y)$$

E.g., risk p converted to odds $g(p) = p/(1-p)$

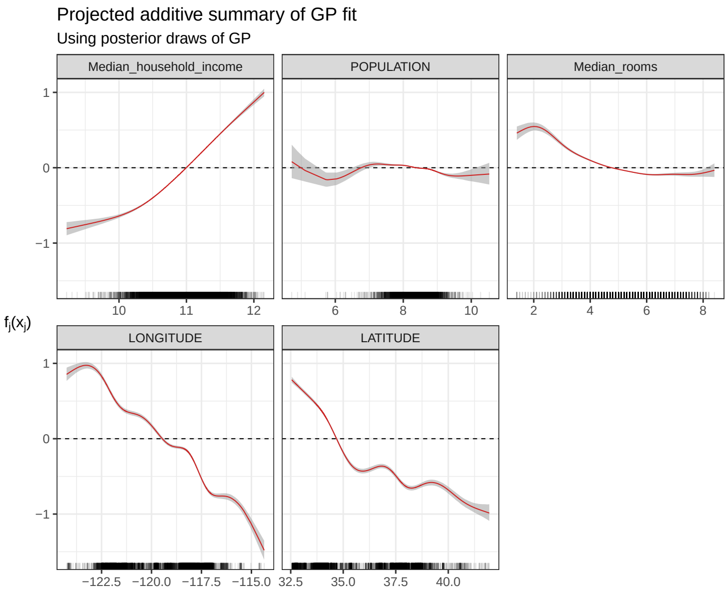
Global additive summary with bands

Solve

$$\begin{aligned}
&\arg\min_{\gamma \in \Gamma} \mathcal{L}(f, \gamma, \tilde{X}) \\
= &\arg\min_{\gamma \in \Gamma} \sum_{i=1}^{n} \left[\ f(x_i) - \gamma(x_i)\ \right]^2 + \sum_{j=1}^{p} \lambda_j \cdot J(h_j)
\end{aligned}$$

using posterior draws of f

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"

- Inflated residual SD:

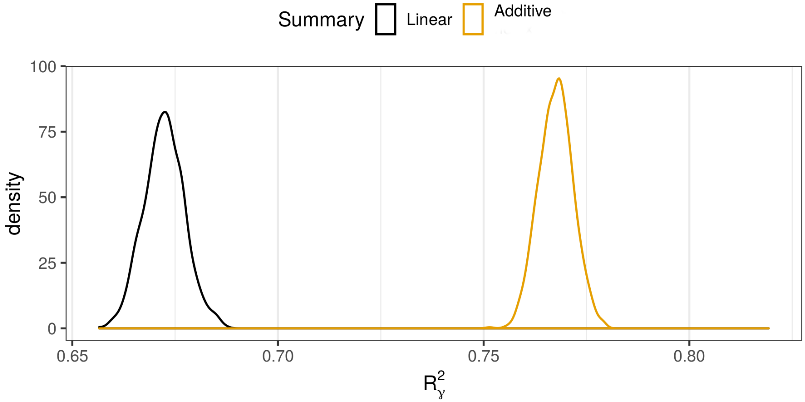
$$\phi_\gamma = \sqrt{\tilde{n}^{-1} \sum_i [\tilde{y}_i - \gamma(\tilde{x}_i)]^2} / \sigma - 1$$

"Inflate predictive intervals by $(\phi_\gamma \times 100)\%$"

- Visually inspect the summary residuals (e.g. with a tree)

$$\hat{f}(\tilde{x}_i) - \hat{\gamma}(\tilde{x}_i)$$

Summary diagnostics for global additive summary



Iterative summary search

Iterative summary search

Sometimes an initial summary isn't sufficient

- We outline an iterative approach; propose, calculate, evaluate, and update the summary as necessary
- Highly flexible. Freedom in the choice of...
  - Regression model for f
  - Error distribution
  - Class of summary
- Global and local summaries available
- *Retain Bayesian interpretation*

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

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

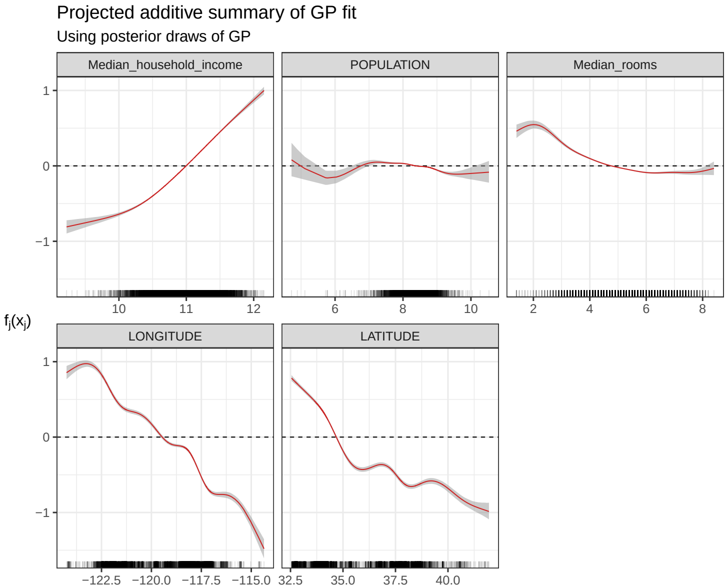
(3) **Evaluate.**
R^2_γ, ϕ_γ, summary residuals $\hat{f}(\tilde{x}_i) - \hat{\gamma}(\tilde{x}_i)$

(4) **If summary is sufficient, perform inference.**

(5) **Otherwise, refine and return to (2).**

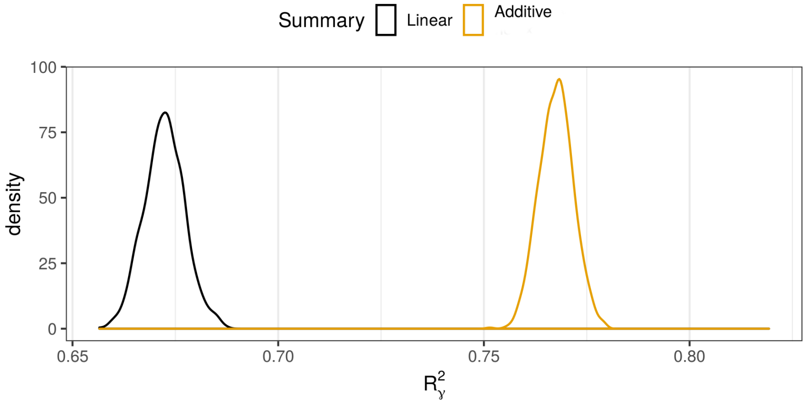
Global summary search

Global additive summary (noninteractive)



Global additive summary (noninteractive)

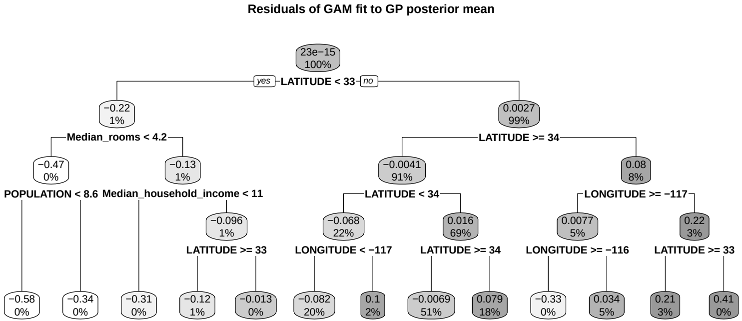
Summary diagnostics



Which interaction to add?

Tree grown on summary residuals

$$\hat{f}(x_i) - \hat{\gamma}(x_i)$$



Additive summary with a two-way interaction

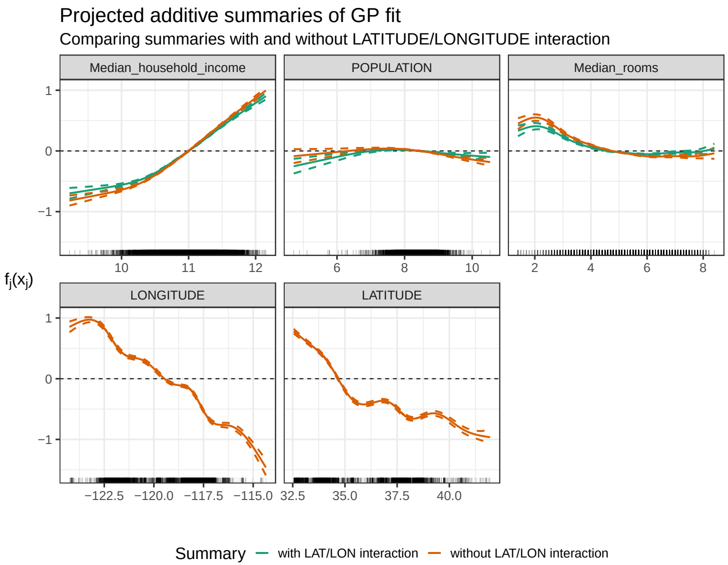
Expand Γ to functions of the form

$$\gamma(x) = \alpha + h_{kl}(x_k, x_l) + \sum_{j \notin \{k,l\}} h_j(x_j), \tag{2}$$

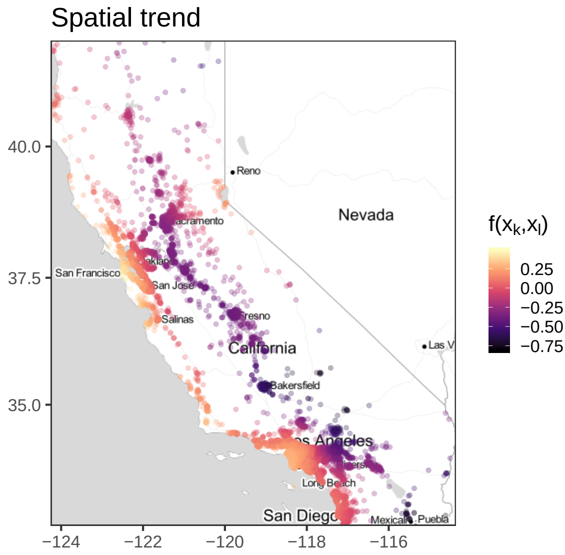
with a 2D smooth function $h_{kl}(x_k, x_l)$ for LON/LAT interaction.

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

Additive summary with a two-way interaction

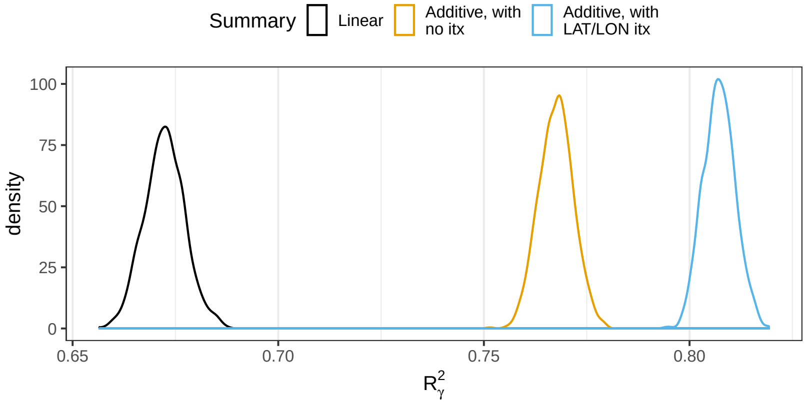


Additive summary with a two-way interaction

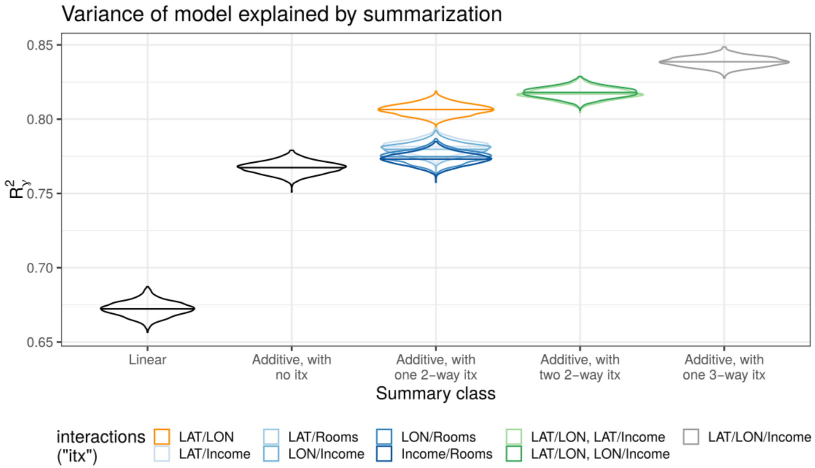


Additive summary with a two-way interaction

Summary diagnostics



Exploring further interactions. . .



*See paper for details

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 is possible. . . check the paper for details

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:

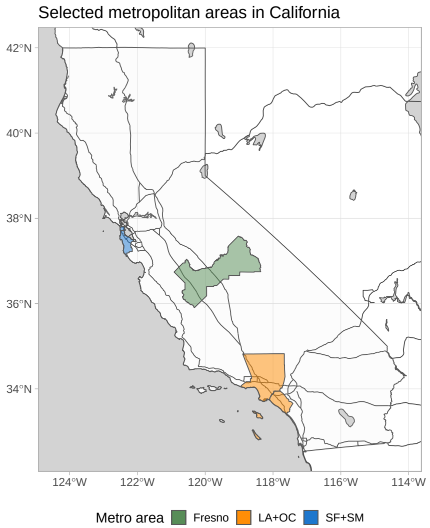
(i) Metropolitan area (compare across MSA's)
(ii) County (compare across and within MSA's)
(iii) Neighborhood (group of tracts; compare within a county)
(iv) Individual tract

Local linear summaries

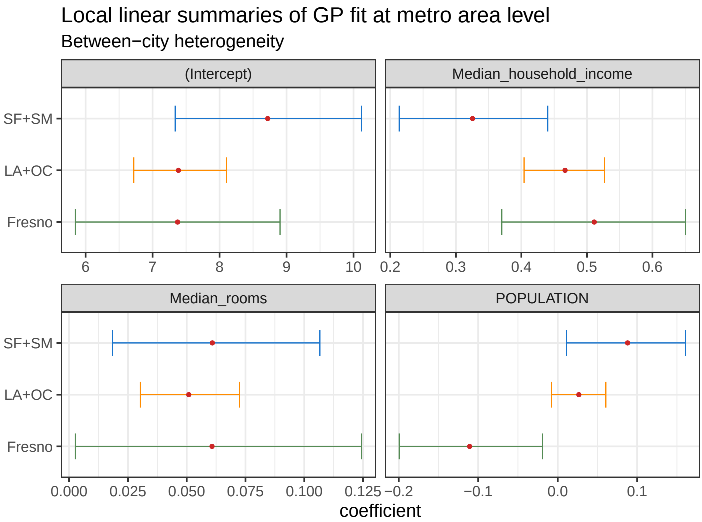
Create synthetic data $\tilde{X}_m$ for each region m by sampling data within neighborhood of x_{m0}

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

MSA-level local linear summaries

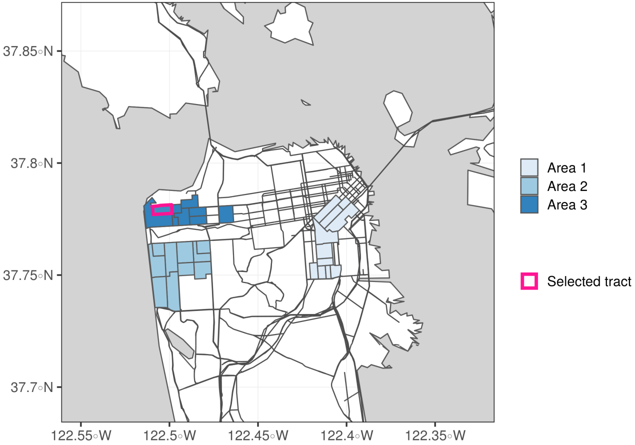


MSA-level local linear summaries

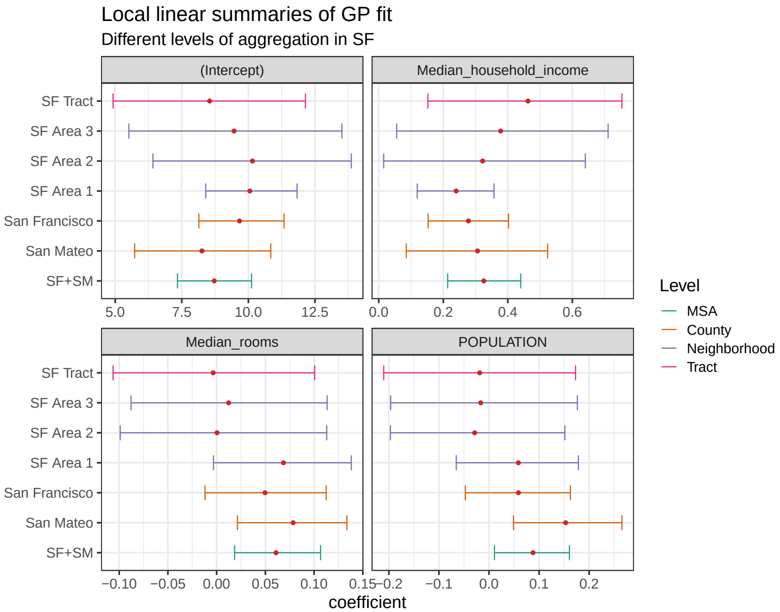


Local linear summaries in San Francisco

Selected areas for summarization

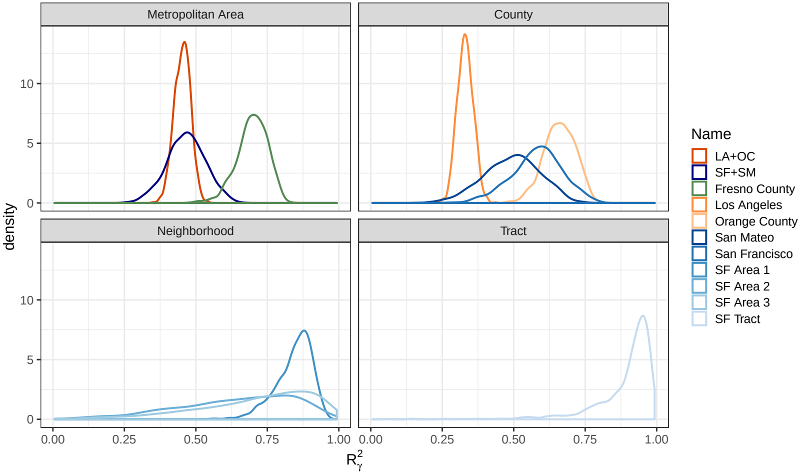


Local linear summaries in San Francisco



Local linear summary diagnostics

R^2_γ



Summarizing effect modification

Summarizing effect modification

Bayesian causal forests model (Hahn, Murray, and Carvalho, 2017) to estimate the causal effect of treatment $Z \in \{0, 1\}$ on a continuous outcome $Y \in \mathbb{R}$.

$$y_i = \mu(x_i) + \tau(x_i) \cdot z_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

with independent priors

$$\mu(x) \sim \text{BART}, \quad \tau(x) \sim \text{BART}$$

Summarizing effect modification

$$y_i = \mu(x_i) + \tau(x_i) \cdot z_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

The average treatment effect (ATE) is

$$\text{ATE} = \mathrm{E}[Y(Z=1) - Y(Z=0)]$$

The conditional average treatment effect (CATE) is

$$\begin{aligned} \text{CATE}(x) := \ & \mathrm{E}[Y(Z=1) - Y(Z=0) \mid x] \\ = \ & \tau(x) \end{aligned}$$

Use same strategies to summarize CATE function

Summarizing effect modification: subgroups

Common goal: find subgroups (non-overlapping partitions of covariate space) with elevated CATE

One approach: summarize $\tau(x)$ with a tree

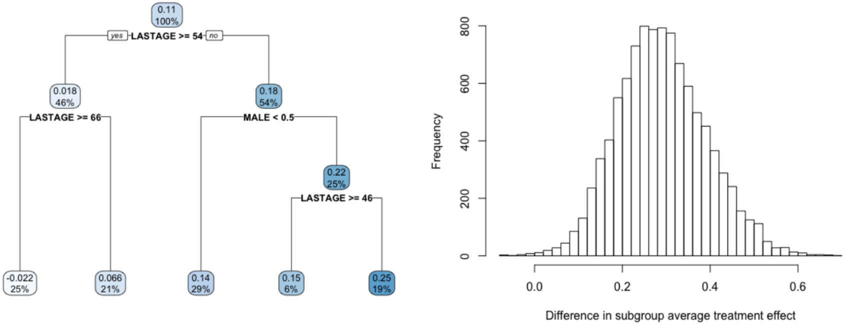
1987 National Medical Expenditure Survey

NMES data

What is the effect of smoking on medical expenditures? Covariates:

- `age`: age in years at the time of the survey
- `smoke-age`: age in years when the individual started smoking
- `gender`: male or female
- `race`: other, black or white
- `marriage-status`: married, widowed, divorced, separated, never married
- `education-level`: college graduate, some college, high school graduate, other
- `census-region`: geographic location, Northeast, Midwest, South, West
- `poverty-status`: poor, near poor, low income, middle income, high income
- `seat-belt`: does patient regularly use a seat belt when in a car

1987 National Medical Expenditure Survey



Right panel: Difference in treatment effect between men ≤ 46 and women ≥ 66

Conclusion

Conclusion

- Possible to interpret nonparametric models via posterior summarization
- Validity of summaries depends on good model fit in first stage
- Confirmatory vs. exploratory analyses; these analyses not confirmatory but still better than fitting multiple models with the same data (“posterior hacking”)
- Closely related to field of interpretable machine learning; see Molnar (2019) for a review

Future work

Prediction case:

- Other classes of summaries?
- Applications to different models
- How can we know which areas to look for heterogenous effects (e.g., in housing data we fixed geographic location)?

Causal inference: selecting confounders, selecting modifiers...

Contact

Website: spencerwoody.github.io

Email: spencer.woody@utexas.edu

Paper by Woody, Carvalho, and Murray (2019):
arxiv.org/abs/1905.07103

References I

Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 04 2010. ISSN 0006-3444. doi: 10.1093/biomet/asq017. URL https://doi.org/10.1093/biomet/asq017.

P. Richard Hahn and Carlos M. Carvalho. Decoupling shrinkage and selection in Bayesian linear models: A posterior summary perspective. *Journal of the American Statistical Association*, 110(509):435–448, 2015. doi: 10.1080/01621459.2014.993077. URL https://doi.org/10.1080/01621459.2014.993077.

P. Richard Hahn, Jared S. Murray, and Carlos M. Carvalho. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects, 2017. URL https://arxiv.org/abs/1706.09523.

Steven N MacEachern. Decision theoretic aspects of dependent nonparametric processes. *Bayesian methods with applications to science, policy and official statistics*, pages 551–560, 2001.

Christoph Molnar. *Interpretable Machine Learning*. Self-published, 2019. https://christophm.github.io/interpretable-ml-book/.

S.N Wood. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2 edition, 2017.

Spencer Woody, Calos M. Carvalho, and Jared S. Murray. Model interpretation through lower-dimensional posterior summarization. *arXiv preprint*, 2019.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006. ISSN 01621459. URL http://www.jstor.org/stable/27639762.

Extra slides...

Projected posterior calculation

Projected posterior calculation

Each function h_j is represented by the linear basis expansion,

$$h_j(x_j) = \sum_{m=1}^{M_j} \delta_{jm}\eta_{jm}(x_j) = \sum_{m=1}^{M_j} \delta_{jm} z_{jm}$$

for some basis functions η_{jm}. Then the vector of output from the additive model is given by $\gamma(X) = \alpha + Z\delta$, where the ith row of the matrix Z represents the linear basis expansion of x_i, and δ is the concatenation of the basis weights δ_{jm}. These weights are estimated using iteratively reweighted least squares, with tuning parameters selected by minimizing the generalized cross validation score. For details on the form of these basis functions and how the model is fit, see Wood (2017).

In the end, the fitted values of the point estimate additive summary can be represented by a linear smoothing of the posterior mean fitted values from f, i.e. $\hat{\gamma}(x) = P\hat{\mathbf{f}}$ where P is an influence matrix. In fact, the fitted values evaluated for each of the additive functions are the result of a linear smoother, i.e. $h(x_j) = P_j\hat{\mathbf{f}}$, where P_j is the subset of rows of the projection matrix P corresponding to the basis expansion for the jth term. This readily provides a way to approximate the projected posterior for the smooth functions using posterior draws of original fitted values $\mathbf{f}^{(k)}$. A single MCMC draw from the projected posterior is calculated simply by $h^{(k)}(x_j) = P_j\mathbf{f}^{(k)}$.

Local linear summary calculation

Local linear summaries: calculation

1. **Create synthetic data specific to locality m**
   (a) Generate $\tilde{n} = 1000$ new geographic locations in the area
   (b) Generate values for other covariates. Draw $\tilde{x}_{i,1:3} \sim \mathcal{N}(\hat{\mu}_m, \hat{\Sigma}_m), i = 1, \ldots, \tilde{n}$ using empirical estimates
   (c) Estimate fitted function at these covariate values $\tilde{X}$. Call this vector $\tilde{\mathbf{f}}$

2. **Calculate projection**
   Point estimate for the linear summary:

$$\hat{\beta} = \arg\min_{\beta} \mathrm{E}[\|\tilde{\mathbf{f}} - \tilde{X}\beta\|_2^2 \mid Y]$$
$$= (\tilde{X}^\intercal \tilde{X})^{-1} \tilde{X}^\intercal \mathrm{E}[\tilde{\mathbf{f}} \mid Y]$$

   Projected posterior draws using

$$\beta^{(k)} = (\tilde{X}^\intercal \tilde{X})^{-1} \tilde{X}^\intercal \tilde{\mathbf{f}}^{(k)}$$

Toy example

Toy example

- Simulate data from

$$
\begin{aligned}
y_i &= f(x_{i1}, x_{i2}) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \\
f(x_1, x_2) &= \frac{1}{1 + \exp(-2x_1 - 2x_2)} + \frac{1}{1 + \exp(-x_1 + 4x_2)},
\end{aligned}
$$

- $\sigma^2 = 0.25$, $n = 2500$ observations on a 2D grid of (x_1, x_2)
- Priors

$$
\begin{aligned}
f &\sim \mathsf{GP}(0, k_{\mathsf{SE}}(\cdot, \cdot)) \\
p(\sigma^2) &\propto \sigma^{-2}
\end{aligned}
$$

Toy example

Two summary classes:

- Linear summary

$$\Gamma_1 = \{\gamma_1 : \gamma_1(x_1, x_2) = \alpha_1 + \beta_1 x_1 + \beta_2 x_2\}$$

- Additive summary (with splines)

$$\Gamma_2 = \{\gamma_2 : \gamma_2(x_1, x_2) = \alpha_2 + h_1(x_1) + h_2(x_2)\}$$

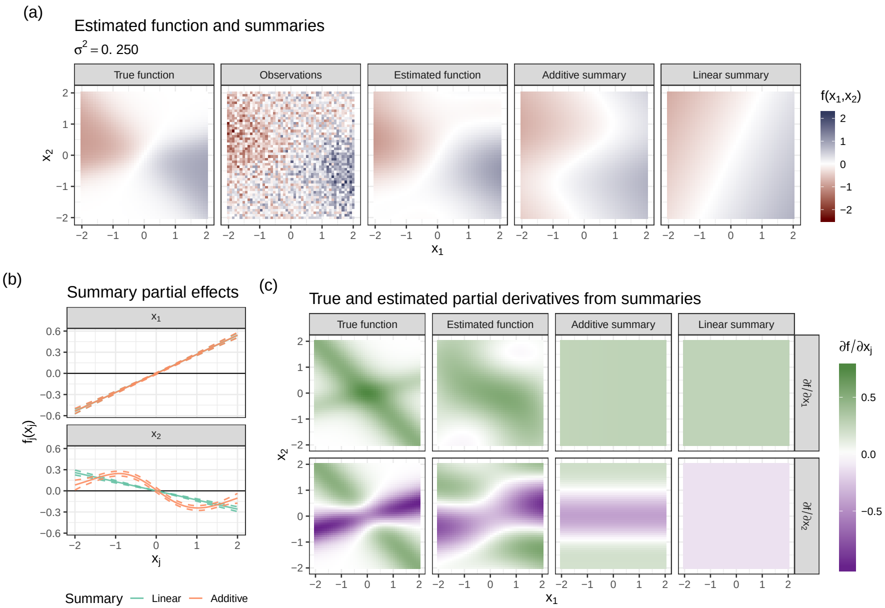
Point estimates:

$$\hat{\gamma}_1(x) = \arg\min_{\gamma_1 \in \Gamma_1} \sum_{i=1}^{n} [\hat{f}(x_i) - \gamma_1(x_i)]^2,$$

$$\hat{\gamma}_2(x) = \arg\min_{\gamma_2 \in \Gamma_2} \sum_{i} [\hat{f}(x_i) - \gamma_2(x_i)]^2 + [\lambda_1 \cdot J(h_1) + \lambda_2 \cdot J(h_2)]$$

$J(h_j) = \int h_j''(t)^2 dt$ enforces smoothness

Toy example



(a)
Estimated function and summaries
σ² = 0. 250
True function
Observations
Estimated function
Additive summary
Linear summary
f(x1,x2)
x2
x1
(b)
Summary partial effects
x1
x2
fj(xj)
xj
Summary
Linear
Additive
(c)
True and estimated partial derivatives from summaries
True function
Estimated function
Additive summary
Linear summary
∂f/∂x1
∂f/∂x2
∂f/∂xj
x2
x1

Lower-dimensional linear model summaries

Summarizing a (high-dimensional) linear model

- Consider the standard multiple linear regression in p variables

$$y \sim \mathcal{N}(X\beta, \sigma^2 \mathcal{I})$$

- Goal: find a sparse set of relevant features $\eta \in \{0, 1\}^p$, i.e.

$$\begin{aligned} f(x) &= x^\intercal \beta \quad \text{is replaced by} \\ \gamma(x) &= x^\intercal \tilde{\beta} \end{aligned}$$

where $\tilde{\beta}_j = 0$ if $\eta_j = 0$

Projected posterior for the sparse summary

- The summary point estimate is

$$\beta_\lambda = \arg\min_{\tilde{\beta}} \|X\bar{\beta} - X\tilde{\beta}\|_2^2 + \lambda \cdot p(\tilde{\beta})$$

  - $\bar{\beta}$ is the posterior mean
  - $p(\cdot)$ enforces sparsity (e.g., ℓ_0, ℓ_1 norm).

- This mirrors Hahn and Carvalho (2015).
- We can also quantify uncertainty around this estimate.

Projected posterior for the sparse summary

- Naive approach: refit with selected covariates ("posterior hacking")
- More appropriate to propagate posterior uncertainty in full model through to the linear summary

Let η_λ be the inclusion vector for β_λ, i.e.

$$(\eta_\lambda)_j = \begin{cases} 0 & \text{if } (\beta_\lambda)_j = 0 \\ 1 & \text{if } (\beta_\lambda)_j \neq 0 \end{cases}$$

X_η is the η-subset of columns of X (dropping λ)

Key: Project the fitted values of from full model $X\beta$ fitted values of summary $X_\eta \beta_\eta$.

Projected posterior for the sparse summary

With Monte Carlo draws of $\beta^{(k)} \sim p(\beta \mid y)$,

$$\tilde{\beta}_{\eta}^{(k)} = (X_{\eta}^{\intercal} X_{\eta})^{-1} X_{\eta}^{\intercal} X \beta^{(k)}$$

$p(\tilde{\beta} \mid y)$ is called the projected posterior.

Example: `MASS::UScrime` dataset

- $n = 47$, $p = 15$, use horseshoe prior (Carvalho et al., 2010)
- Use adaptive lasso (Zou, 2006) penalty

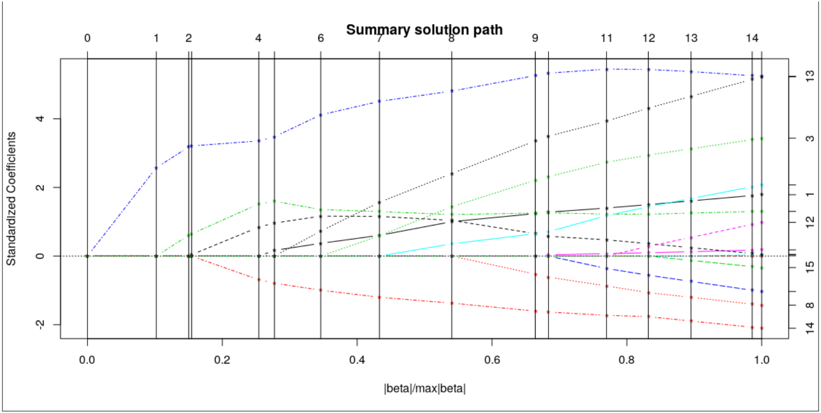
$$p(\tilde{\beta}) = \sum_j w_j^{-1} |\tilde{\beta}_j|$$

  with weights $w_j = |\bar{\beta}_j|$ from posterior mean.

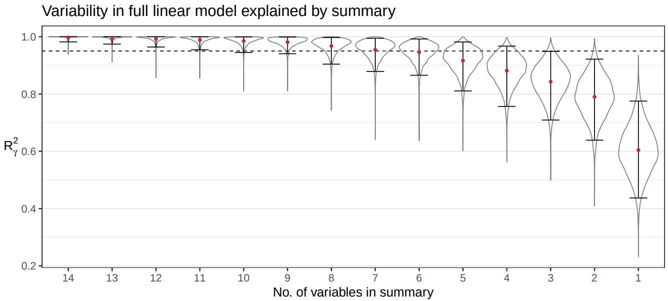
- Solution path of summary parsimony determined by λ
- Compare posteriors:
  - Projected posterior
  - Refitting with selected variables using flat prior
  - Marginal posteriors from original (horseshoe) posterior

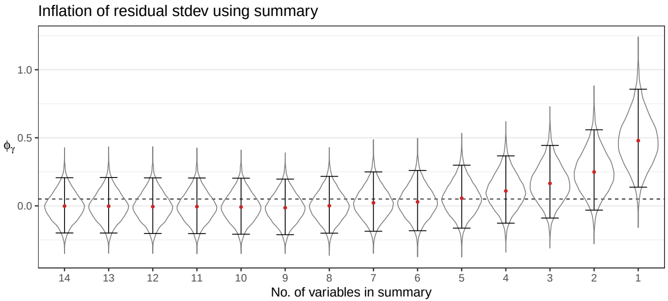
Point estimates for summaries

$$\beta_\lambda = \arg\min_{\tilde{\beta}} \|X\bar{\beta} - X\tilde{\beta}\|_2^2 + \lambda \sum_j w_j^{-1} |\tilde{\beta}_j|$$



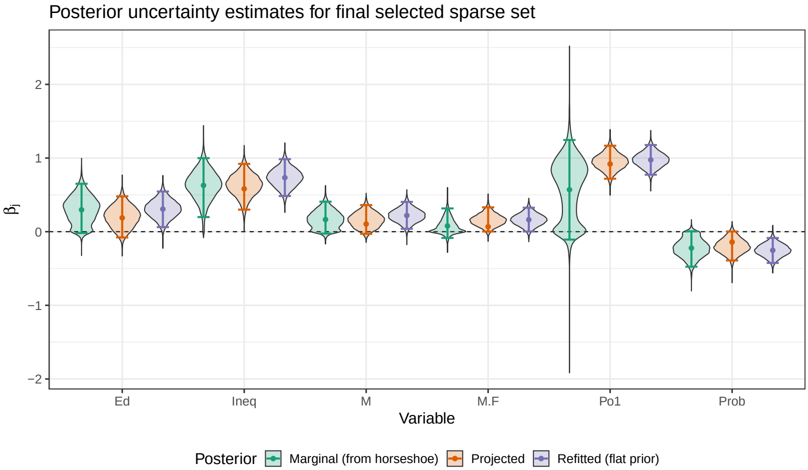
Summary diagnostics: choose summary size 6





Comparing posteriors from different methods

$$\text{Projection: } \tilde{\beta}_\eta = (X_\eta^\intercal X_\eta)^{-1} X_\eta^\intercal X\beta$$



Projected posteriors for two covariates