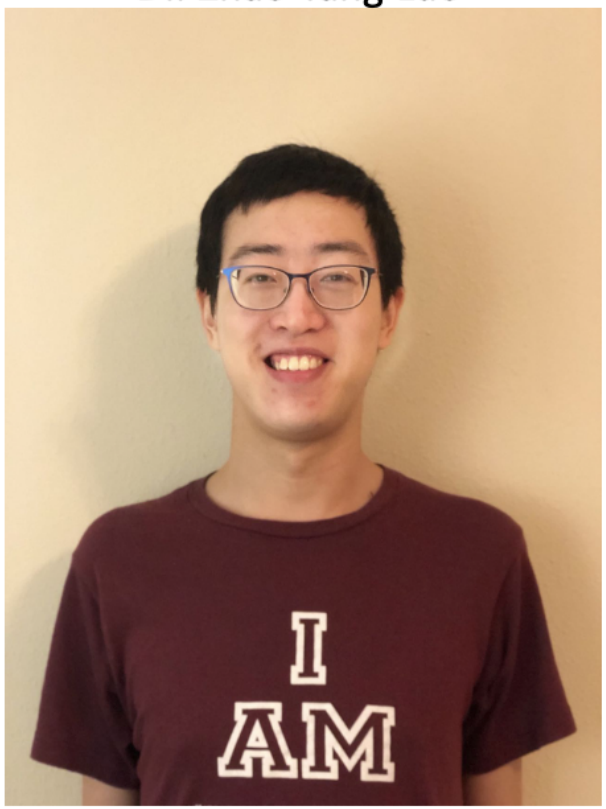
Bayesian Additive Semi-Multivariate Decision Trees

Huiyan Sang

Department of Statistics, Texas AM University

Collaborators

Dr. Zhao Tang Luo



Dr. Bani Mallick

Nonparametric Semi-Structured Regression

- *Semi-Structured Regression*:

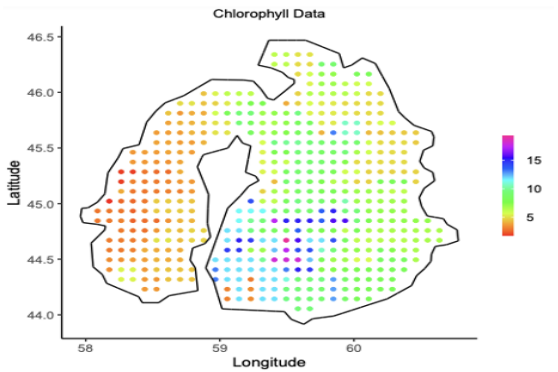
$$Y = f(\mathbf{s}, \mathbf{x}) + \epsilon, \tag{1}$$

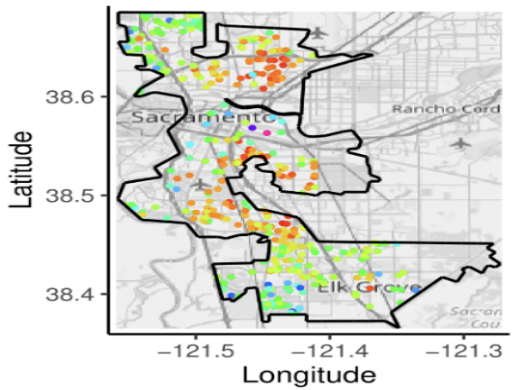
  - $Y \in \mathbb{R}$: response at $\mathbf{s}$ (e.g., housing price)
  - $\mathbf{s} \in \mathcal{M}$: structured multivariate features (e.g., spatial locations)
  - $\mathbf{x} \in \mathcal{X}$: unstructured features (e.g., square footage, housing age)
  - $f : \mathcal{D} \subseteq \mathcal{M} \times \mathcal{X} \to \mathbb{R}$: unknown mean function

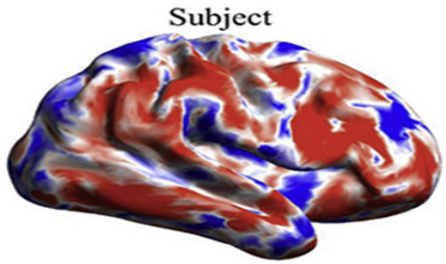
- Goal: non-parametric function estimation of $f(\mathbf{s}, \mathbf{x})$ from noisy observations.
  - $f(\mathbf{s}, \mathbf{x})$: Luo, Sang and Mallick (2022), BAMDT: Bayesian Additive Semi-Multivariate Decision Trees for Nonparametric Regression, ICML, long oral.
  - $f(\mathbf{s})$: Luo, Sang and Mallick (2021), BAST: Bayesian Additive Regression Spanning Trees for Complex Constrained Domain, NeurIPS, 34.

Main Challenges

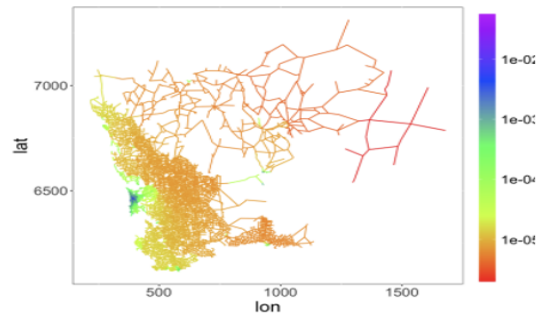
- Structured feature space $\mathcal{M}$ has a complex geometry (e.g., irregular boundary/interior holes, road/river networks, graphs, brain cortical surfaces)
- Irregular function discontinuities/sharp changes (e.g., highways/rivers)
- Potentially high dimensional unstructured features **x**
- Potential interactions between the effects of **s** and **x**







From Joshi et al. NeuroImage, (2018)



Existing Methods

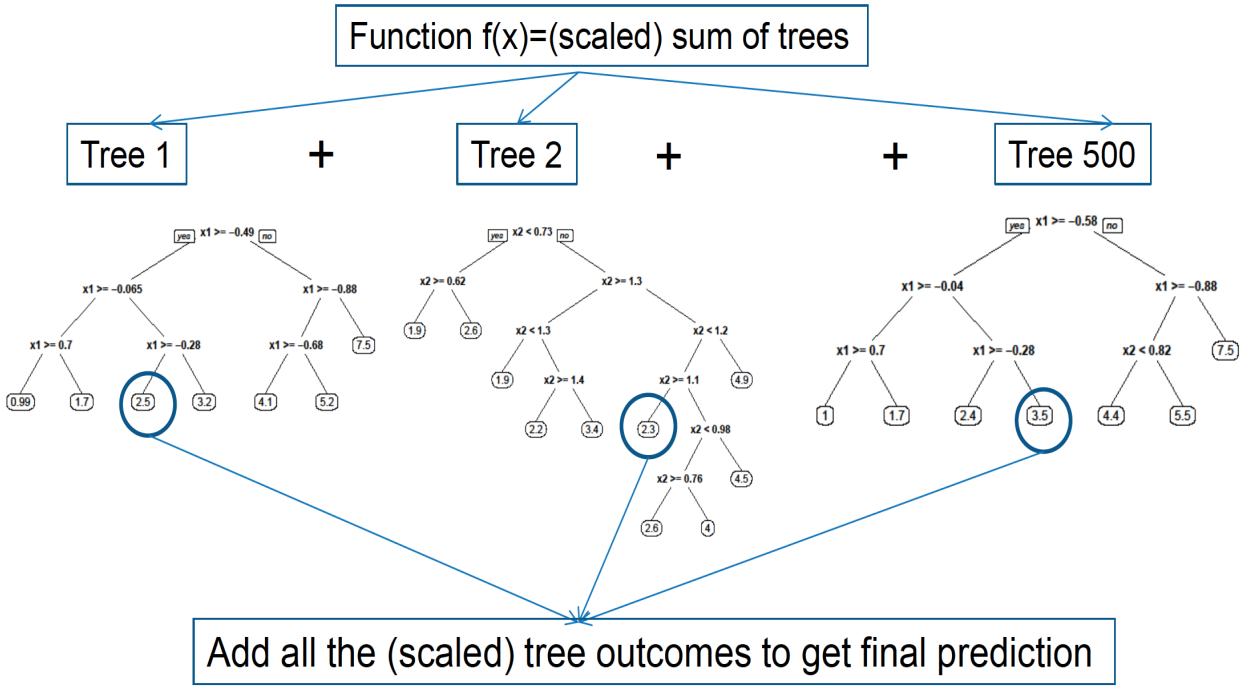
Spline smoothing (Ramsay, 2002; Lai and Schumaker, 2007; Wang and Ranalli, 2007; Wood et al., 2008; Scott-Hayward et al., 2014; Sangalli et al., 2013) and Gaussian process regression (Lin et al., 2019; Niu et al., 2019; Borovitskiy et al., 2020; Dunson et al., 2022):

- Respect complex domain boundaries and intrinsic geometries in $\mathcal{M}$ ✓
- Assume globally smooth f ✗
- Too many tensor spline basis functions for high dimensional $\mathbf{x}$ ✗
- GP regression usually assumes an additive form: $f(\mathbf{s}) + f(\mathbf{x})$, ✗
  - $f(\mathbf{s}, \mathbf{x}) = \mathbf{x}^{\mathsf{T}}\boldsymbol{\beta} + GP(\mathbf{s})$
  - $f(\mathbf{s}, \mathbf{x}) = GAM(\mathbf{x}) + GP(\mathbf{s})$ (Nandy et al. 2017)
  - $f(\mathbf{s}, \mathbf{x}) = RF(\mathbf{x}) + GP(\mathbf{s})$ (Saha et al. 2021)

Existing Methods: BART

Bayesian additive regression trees (BART; Chipman et al., 2010):

- Each decision tree corresponds to a piecewise (hyper-rectangular) constant function on the feature space.
- Summation of many simple piecewise constant functions can approximate well complex functions with discontinuities (sharp changes) and different levels of smoothness. ✓



From Dr. McCulloch's talk slides on BART

Existing Methods: BART

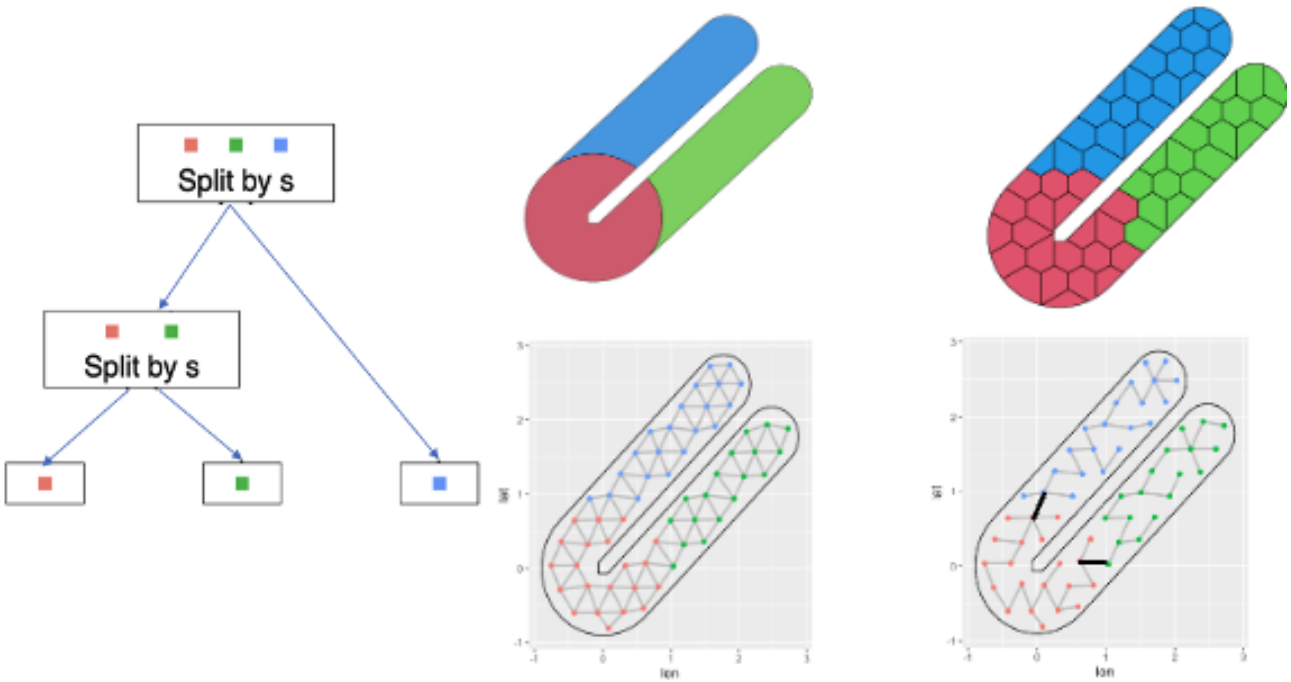
Axis-parallel univariate split rules:

- Address feature scaling/feature selection issues with high dimensional $\mathbf{x}$ ✓
- May not fully respect domain boundaries and capture irregular data discontinuities in $\mathcal{M}$ ✗

From : blog.wilton.com, landolakes.com, cakesrock.blog

Multivariate Decision Trees (MDT) for $f(\mathbf{s})$

To motivate the method for $f(\mathbf{s}, \mathbf{x})$, let's first consider a simpler case $f(\mathbf{s})$.



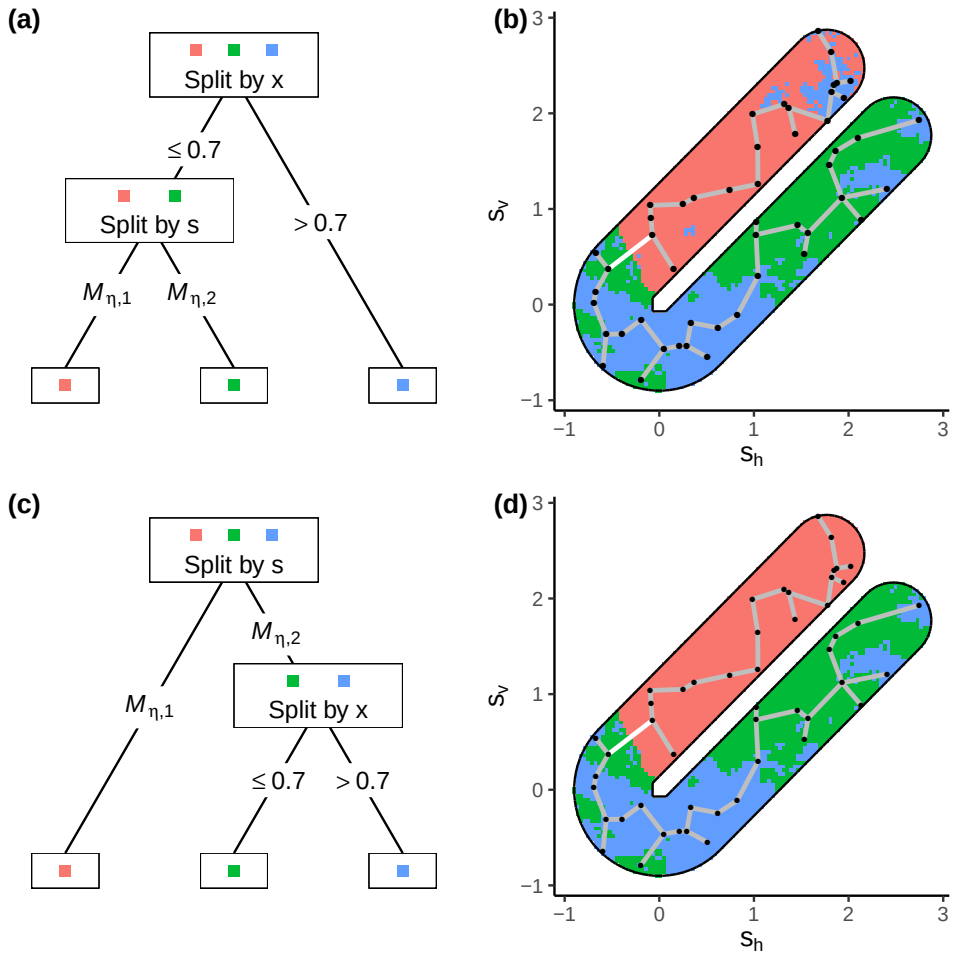
Predictive spanning tree partition prior model (Luo, Sang, Mallick, 2021):

- Prior for the locations of removed edges in spanning tree (where to split)
- Prior for the number of clusters/depth to encourage small MDT
- Prior for the cluster-wise constant to encourage small values at leaf nodes
- Use different reference knots/meshes/spanning trees in each weak MDT learners

Multivariate Decision Trees for $f(\mathbf{s}, \mathbf{x})$

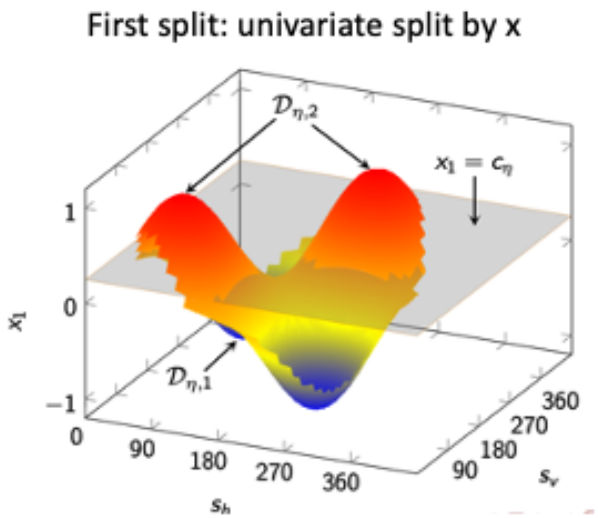
Each node η represents a subset $\mathcal{D}_\eta \subset \mathcal{D}$.

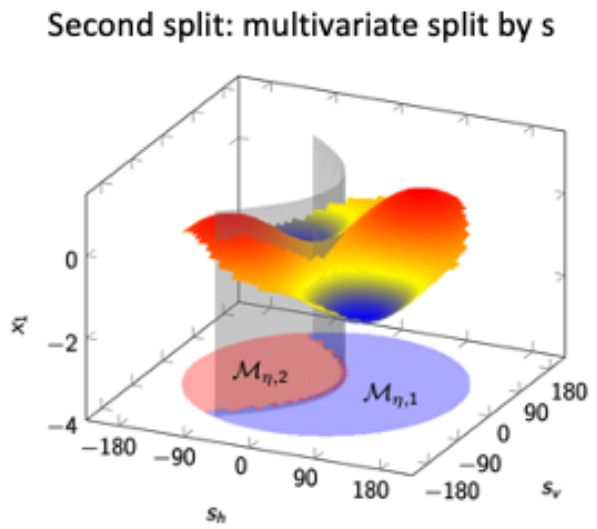
1. Start with a root node representing $\mathcal{D}$.
2. Split a terminal node η with probability $p_{\text{split}}(\eta)$. If η splits, choose one split rule to obtain a bipartition $\{\mathcal{D}_{\eta,1}, \mathcal{D}_{\eta,2}\}$:
   - 2.1 With probability p_m, perform a multivariate split (how?).
   - 2.2 Otherwise, perform a univariate split (same as BART).
3. Apply Step 2 to each offspring node of η.



Multivariate Decision Trees for $f(\mathbf{s}, \mathbf{x})$

- Main idea: at each multivariate node
  1. Project $\mathcal{D}_\eta$ into $\mathcal{M}$, call it $\mathcal{M}_\eta$
  2. Perform a bipartition of $\mathcal{M}_\eta$ following the predictive spanning tree model
  3. Divide $\mathcal{D}_\eta$ into two subsets: $\mathcal{D}_{\eta,k} = \mathcal{D}_\eta \cap (\mathcal{M}_{\eta,k} \times \mathcal{X}), \quad \text{for } k = 1, 2,$
- Main complications: (1) M_η varies with node; (2) how to partition $\mathcal{M}_\eta$ such that both $\mathcal{D}_{\eta,1}$ and $D_{\eta,2}$ contain non-empty sets of observations?





Manifold Bipartitions via Predictive Spanning Trees

- Notations:
  - $\mathcal{S}^*$: Reference knots on $\mathcal{M}$.
  - $\mathcal{G}_T^*$: Fixed undirected spanning tree graph on $\mathcal{S}^*$.

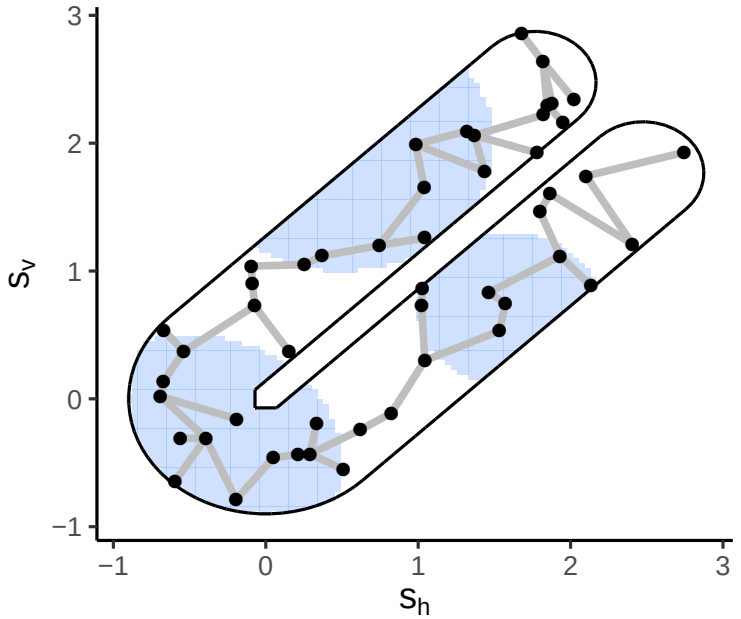


Figure: A predictive spanning tree bipartition

Manifold Bipartitions via Predictive Spanning Trees

- Notations:
  - $\mathcal{S}^*$: Reference knots on $\mathcal{M}$.
  - $\mathcal{G}_T^*$: Fixed undirected spanning tree graph on $\mathcal{S}^*$.
- To obtain a bipartition of $\mathcal{M}_\eta$:
  1. Identify $\mathcal{S}_\eta^*$: Union of the nearest reference knot of each observed point in $\mathcal{M}_\eta$ under geodesic distance d_g.

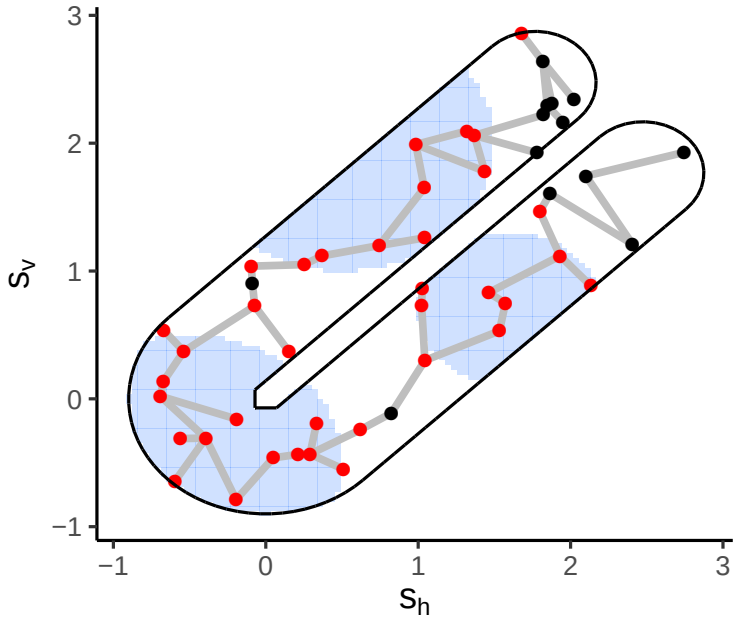


Figure: A predictive spanning tree bipartition

Manifold Bipartitions via Predictive Spanning Trees

- Notations:
  - $\mathcal{S}^*$: Reference knots on $\mathcal{M}$.
  - $\mathcal{G}_T^*$: Fixed undirected spanning tree graph on $\mathcal{S}^*$.
- To obtain a bipartition of $\mathcal{M}_\eta$:
  1. Identify $\mathcal{S}_\eta^*$: Union of the nearest reference knot of each observed point in $\mathcal{M}_\eta$ under geodesic distance d_g.
  2. Randomly sample two knots $\mathbf{s}^*$ and $\mathbf{t}^*$ from $\mathcal{S}_\eta^*$.
  3. Randomly sample an edge e^* from the unique path in $\mathcal{G}_T^*$ connecting $\mathbf{s}^*$ and $\mathbf{t}^*$.

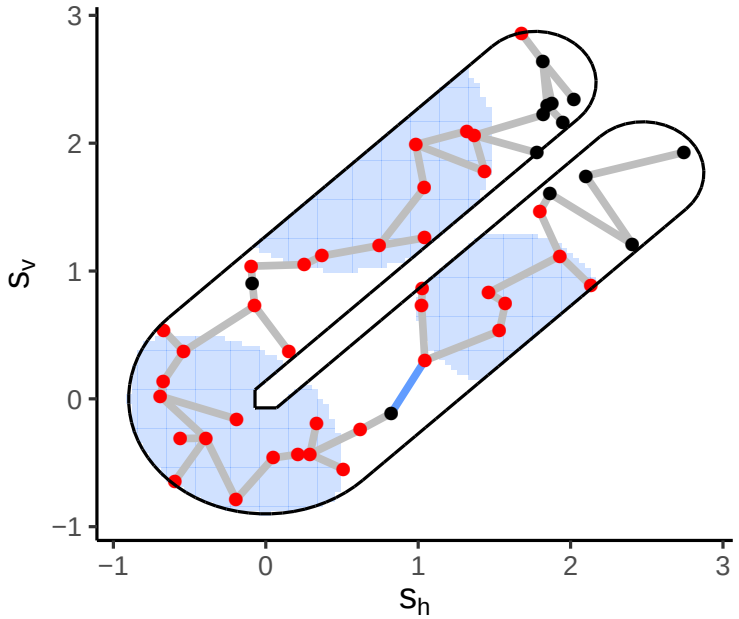


Figure: A predictive spanning tree bipartition

Manifold Bipartitions via Predictive Spanning Trees

- Notations:
  - $\mathcal{S}^*$: Reference knots on $\mathcal{M}$.
  - $\mathcal{G}_T^*$: Fixed undirected spanning tree graph on $\mathcal{S}^*$.
- To obtain a bipartition of $\mathcal{M}_\eta$:
  1. Identify $\mathcal{S}_\eta^*$: Union of the nearest reference knot of each observed point in $\mathcal{M}_\eta$ under geodesic distance d_g.
  2. Randomly sample two knots $\mathbf{s}^*$ and $\mathbf{t}^*$ from $\mathcal{S}_\eta^*$.
  3. Randomly sample an edge e^* from the unique path in $\mathcal{G}_T^*$ connecting $\mathbf{s}^*$ and $\mathbf{t}^*$.
  4. Remove e^* from $\mathcal{G}_T^*$ to obtain bipartitions of $\mathcal{S}_\eta^*$ and $\mathcal{M}_\eta$.

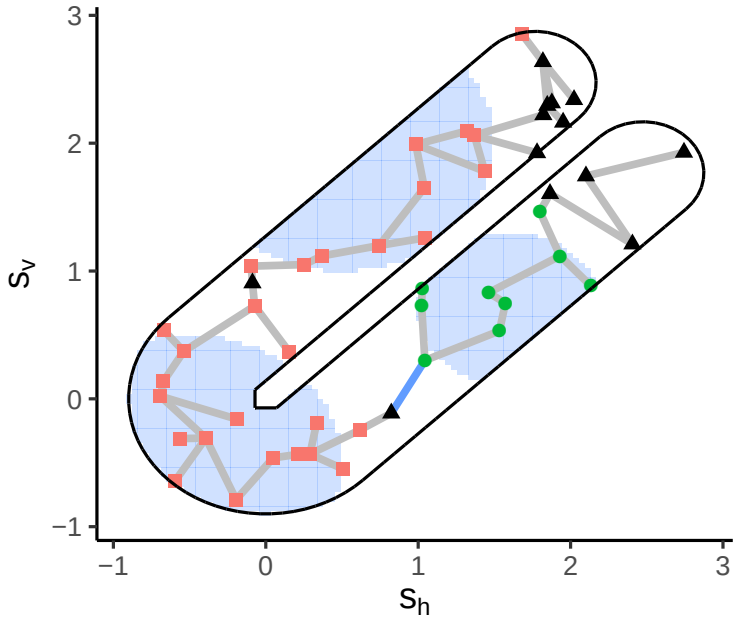


Figure: A predictive spanning tree bipartition

Manifold Bipartitions via Predictive Spanning Trees

- Notations:
  - $\mathcal{S}^*$: Reference knots on $\mathcal{M}$.
  - $\mathcal{G}_T^*$: Fixed undirected spanning tree graph on $\mathcal{S}^*$.
- To obtain a bipartition of $\mathcal{M}_\eta$:
  1. Identify $\mathcal{S}_\eta^*$: Union of the nearest reference knot of each observed point in $\mathcal{M}_\eta$ under geodesic distance d_g.
  2. Randomly sample two knots $\mathbf{s}^*$ and $\mathbf{t}^*$ from $\mathcal{S}_\eta^*$.
  3. Randomly sample an edge e^* from the unique path in $\mathcal{G}_T^*$ connecting $\mathbf{s}^*$ and $\mathbf{t}^*$.
  4. Remove e^* from $\mathcal{G}_T^*$ to obtain bipartitions of $\mathcal{S}_\eta^*$ and $\mathcal{M}_\eta$.

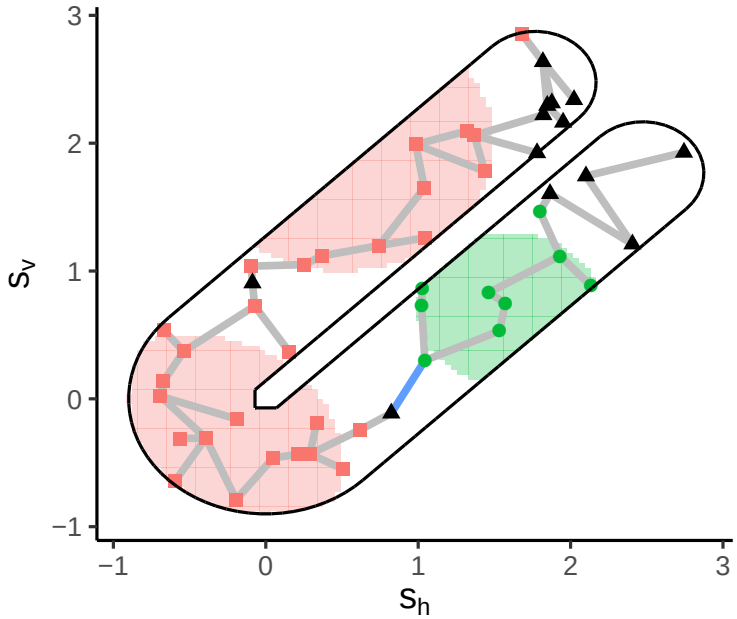


Figure: A predictive spanning tree bipartition

A Bayesian Sum-of-multivariate-decision-trees Model

- Let T denote an sMDT. Define a piecewise constant mapping from $\mathcal{D}$ to $\mathbb{R}$

$$g(\mathbf{s}, \mathbf{x}|\, T, \boldsymbol{\mu}) = \mu_j, \quad \text{if } (\mathbf{s}, \mathbf{x}) \in \mathcal{D}_j.$$

- BAMDT models $Y = f(\mathbf{s}, \mathbf{x}) + \epsilon$ with

$$f(\mathbf{s}, \mathbf{x}) = \sum_{m=1}^{M} g(\mathbf{s}, \mathbf{x}|\, T_m, \boldsymbol{\mu}_m).$$

- Prior specification:

$$p\left(\{T_m, \boldsymbol{\mu}_m\}_{m=1}^{M}, \sigma^2\right) = \left\{\prod_{m=1}^{M} p(\boldsymbol{\mu}_m|\, T_m) p(T_m)\right\} p(\sigma^2),$$

where $\{T_m\}$ is assumed *a priori* to be an iid sample from the sMDT generative process.

Bayesian Inference

- To draw a posterior sample from $[T_m|-]$ with μ_m marginalized out, perform one of the following moves.
  - Grow: Randomly choose a terminal node of T_m and split it following Step 2 of the sMDT generating process.
  - Prune: Randomly choose a node of T_m with two terminal nodes and remove it (and its children) from $\mathcal{T}_m$.

- Importance metric for a feature z:
  - Defined as the proportion of the split rules involving z in the ensemble.
  - z can be $\mathbf{s}$, x_1, ..., or x_p.

Bitten Torus Example: $f(\mathbf{s})$

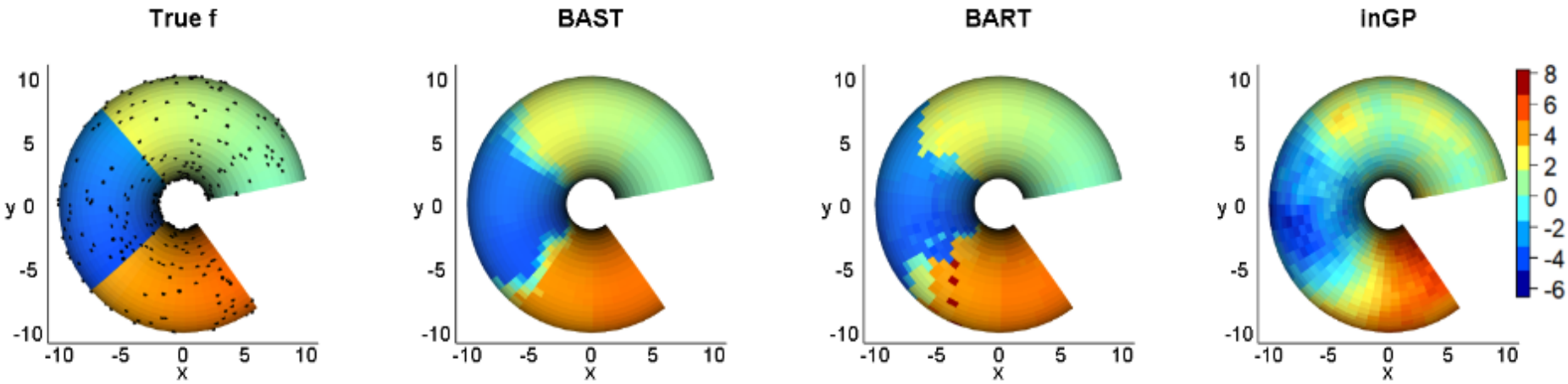


Figure: True function and predictive surfaces

Table: Average prediction performance over 50 replicates

	BAST	BART	inGP (Niu et al., 2019)
MSPE	**0.487**	1.115	2.283
MAPE	**0.307**	0.406	1.159
Mean CRPS	**0.225**	0.355	—

Application to Chlorophyll Data in Aral Sea: $f(\mathbf{s})$

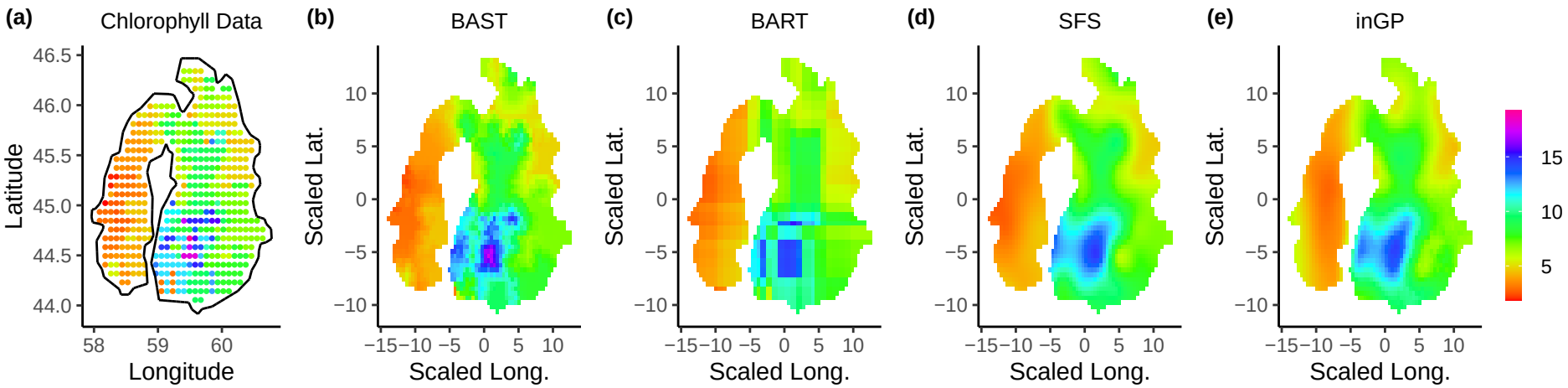


Figure: Observed data and predictive surfaces

Table: Prediction performance from 5-fold cross-validation

	BAST	BART	SFS (Wood et al, 2008)	inGP
MSPE	**2.339**	2.933	2.894	3.191
MAPE	**0.926**	1.172	1.071	1.200
Mean CRPS	**0.641**	0.955	—	—

U-shape and Bitten Torus Examples: $f(\mathbf{s}, \mathbf{x})$

(a) U-shape example with $n = 500$, $p = 2$, and $\sigma = 0.1$

	BAMDT	BART	GP	BAST-s	BAST	GAM-additive	GAM-TP
MSPE	0.374	1.405	0.583	0.912	2.155	0.985	0.418
MAPE	0.281	0.612	0.488	0.543	0.641	0.720	0.345
CRPS	0.219	0.508	0.390	0.398	0.479		

(b) U-shape example with $n = 500$, $p = 10$, and $\sigma = 0.1$

MSPE	0.495	1.219	0.632	0.911	1.214	1.168
MAPE	0.317	0.688	0.537	0.543	0.662	0.751
CRPS	0.252	0.552	0.409	0.398	0.453	

(c) Bitten torus example with $n = 500$, $p = 2$, and $\sigma = 0.1$

MSPE	0.967	1.958	1.621	1.606	1.678
MAPE	0.545	0.693	0.805	0.814	0.666
CRPS	0.431	0.573	0.646	0.591	0.529

In the setting of $p = 10$, the average percentage of splits involving (s, x_1) in BAMDT is 73.98%, while the one in BART is 54.62%

Application to Sacramento Housing Data

- Model housing price using spatial locations, square footage, #bedrooms, and #bathrooms.

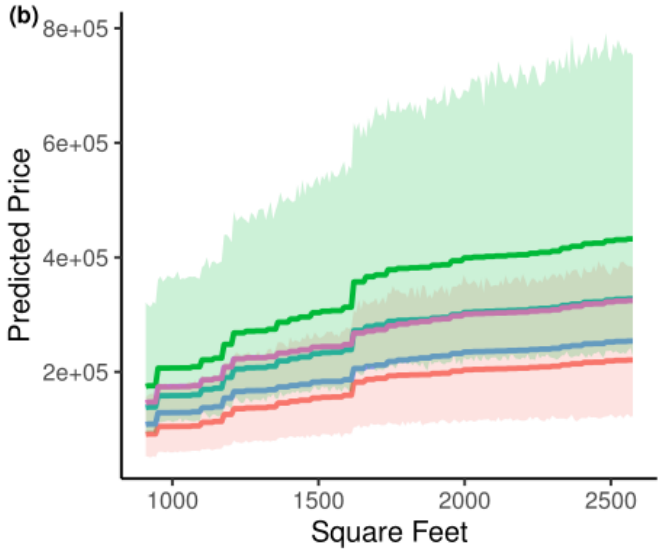


Figure: Predicted price versus square footage.

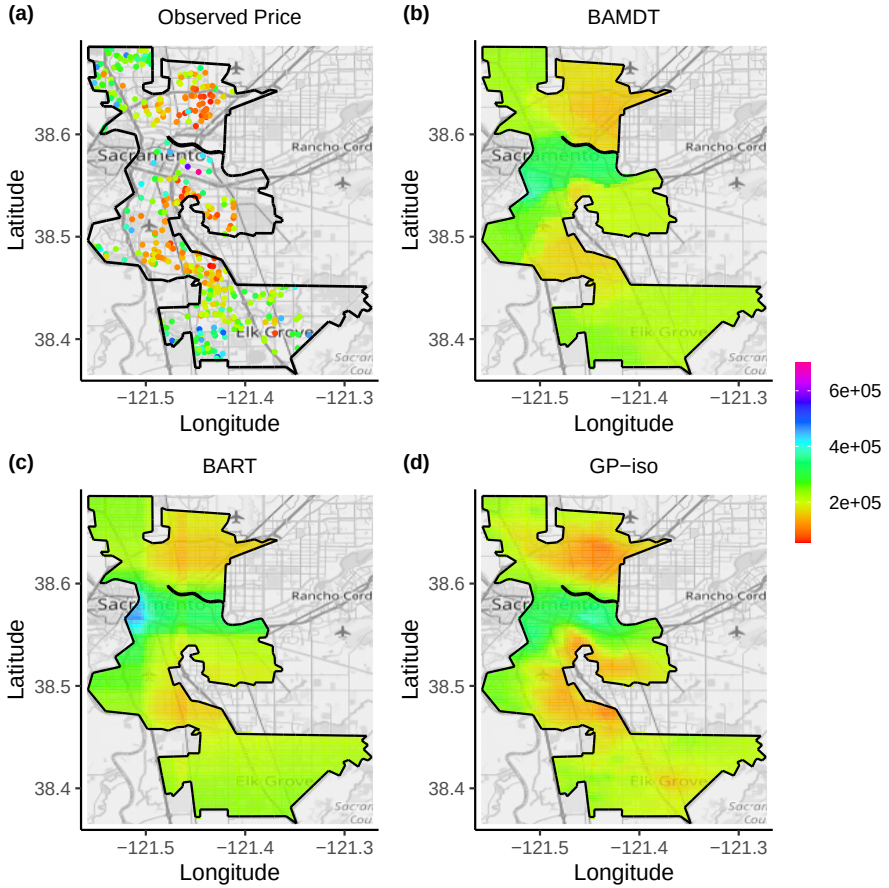


Figure: Observed data and predicted price for a representative house.

Conclusion and Future Work

- A novel Bayesian ensemble model, BAMDT, is developed for nonparametric regression problems on complex constrained domains using the flexible partition models as weak learners.
- Next steps:
  - Extension to unknown manifolds where geodesic distance metrics need to be estimated.
  - Adopting BAMDT as a nonparametric prior model for latent functions in many Bayesian hierarchical modeling settings.
  - Theoretical guarantee such as posterior concentration results.
  - Soft BAST and BAMDT following Linero and Yang, 2018
  - Other partition models? Talk to Changwoo Lee.

Thanks!!

U-shape Example: $f(\mathbf{x}, \mathbf{x})$

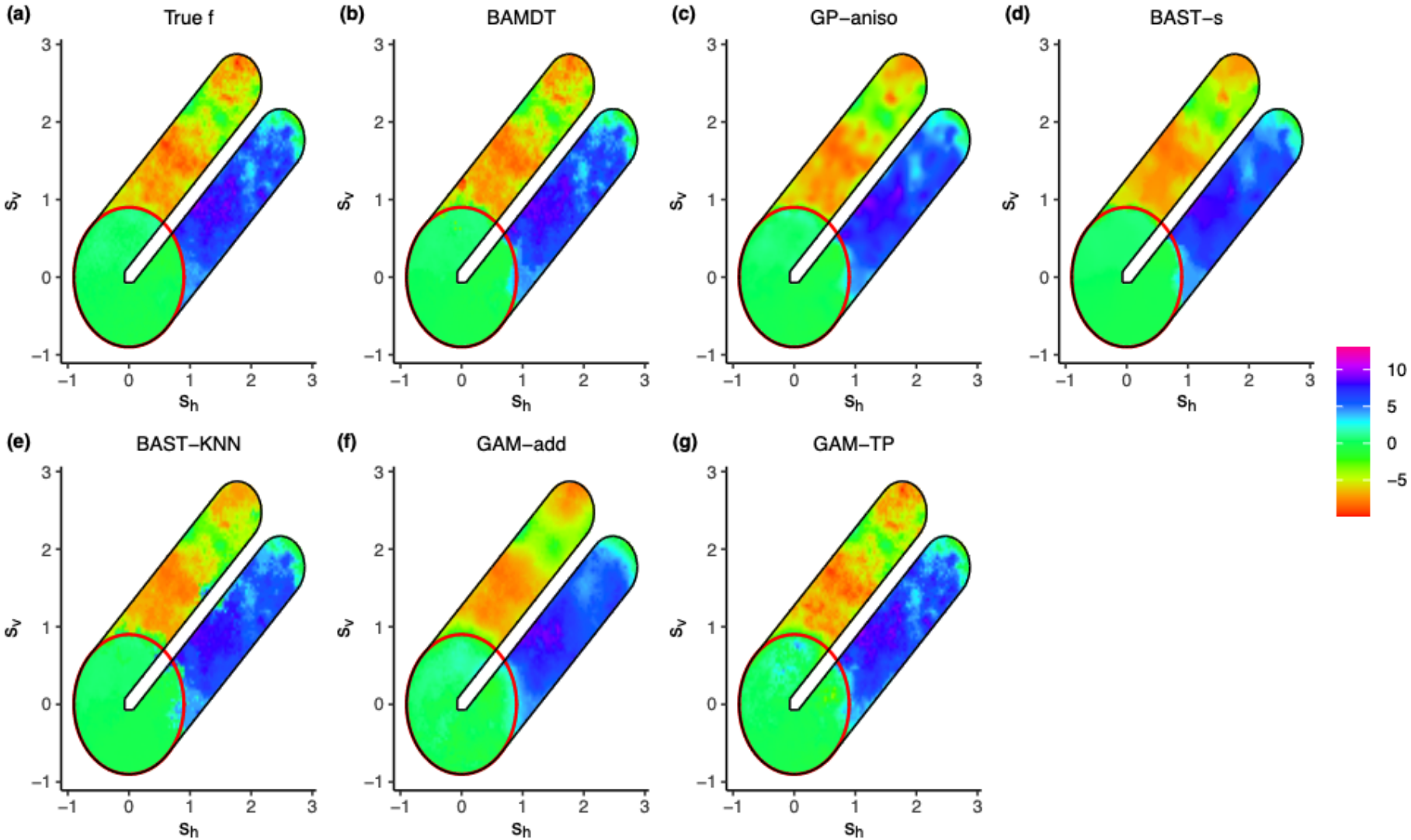


Figure: Posterior predictive surfaces

BAMDT vs BART

Table: Average prediction performance metrics over 50 replicate data sets of BAMDT and BART with various numbers of weak learners

	BAMDT	BART			
	$M = 50$	$M = 50$	$M = 100$	$M = 200$	$M = 400$
MSPE	**0.374**	1.405	1.234	1.162	1.126
	(0.106)	(0.249)	(0.217)	(0.161)	(0.069)
MAPE	**0.281**	0.612	0.590	0.586	0.627
	(0.025)	(0.056)	(0.051)	(0.029)	(0.024)
CRPS	**0.219**	0.508	0.475	0.458	0.481
	(0.023)	(0.052)	(0.046)	(0.026)	(0.018)

U-shape Example: $f(\mathbf{x}, \mathbf{x})$

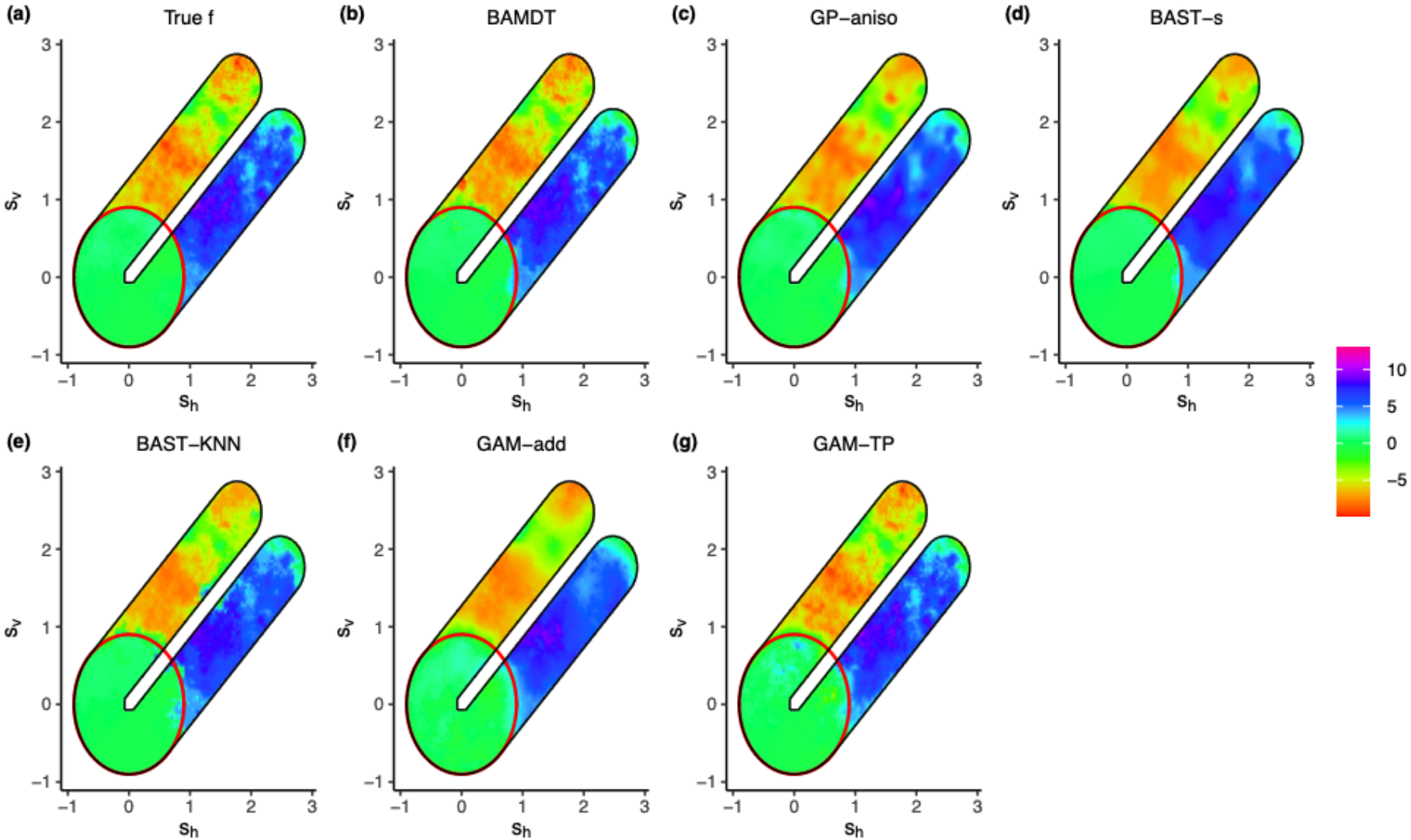


Figure: Posterior predictive surfaces

BAMDT - Sacramento Housing Data

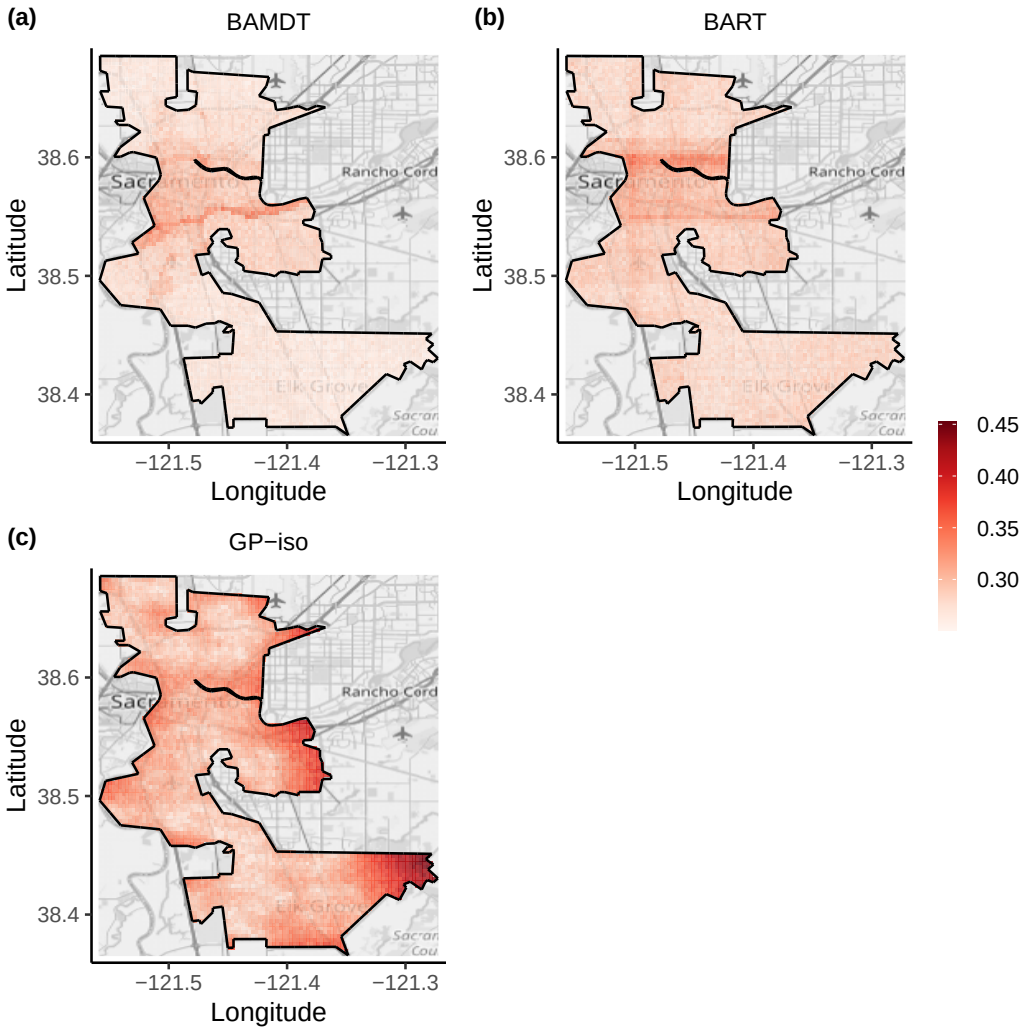


Figure: Posterior predictive standard deviation.

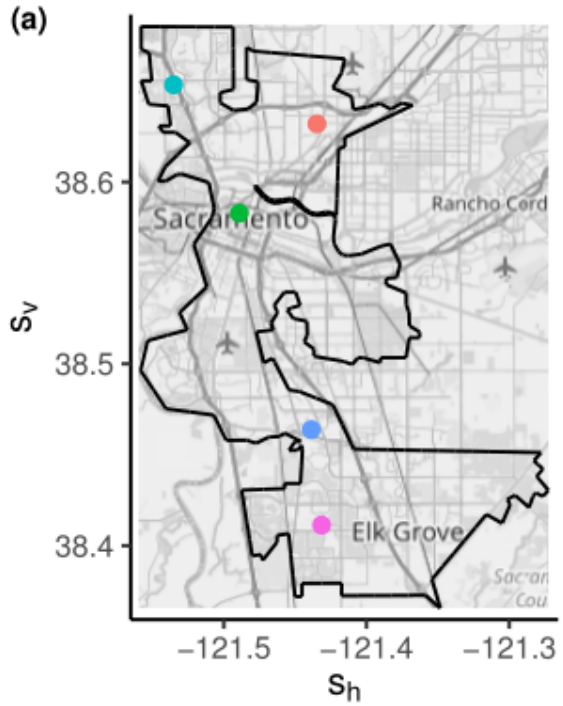


Figure: Map of five representative locations.

References I

Borovitskiy, V., Terenin, A., Mostowsky, P., and Deisenroth, M. P. (2020). Matérn Gaussian processes on Riemannian manifolds. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Dunson, D. B., Wu, H.-T., Wu, N., et al. (2022). Graph based Gaussian processes on restricted domains. *Journal of the Royal Statistical Society Series B*, 84(2):414–439.

Lai, M.-J. and Schumaker, L. L. (2007). *Spline functions on triangulations*, volume 110. Cambridge University Press.

Lin, L., Mu, N., Cheung, P., and Dunson, D. (2019). Extrinsic Gaussian processes for regression and classification on manifolds. *Bayesian Analysis*, 14(3):887–906.

Niu, M., Cheung, P., Lin, L., Dai, Z., Lawrence, N., and Dunson, D. (2019). Intrinsic Gaussian processes on complex constrained domains. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(3):603–627.

Ramsay, T. (2002). Spline smoothing over difficult regions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2):307–319.

Sangalli, L. M., Ramsay, J. O., and Ramsay, T. O. (2013). Spatial spline regression models. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 681–703.

Scott-Hayward, L. A. S., MacKenzie, M. L., Donovan, C. R., Walker, C., and Ashe, E. (2014). Complex region spatial smoother (CReSS). *Journal of Computational and Graphical Statistics*, 23(2):340–360.

References II

Wang, H. and Ranalli, M. G. (2007). Low-rank smoothing splines on complicated domains. *Biometrics*, 63(1):209–217.

Wood, S. N., Bravington, M. V., and Hedley, S. L. (2008). Soap film smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):931–955.