BAMDT: Bayesian Additive Semi-Multivariate Decision Trees for Nonparametric Regression

Zhao Tang Luo, Huiyan Sang, and Bani Mallick

Department of Statistics
Texas A&M University

ICML 2022

Nonparametric Semi-Structured Regression

- Notations:
  - $Y \in \mathbb{R}$: response (e.g., housing price)
  - $\mathsf{s} \in \mathcal{M}$: structured features with known multivariate structures (e.g., spatial locations on a constrained domain)
  - $\mathsf{x} \in \mathcal{X}$: unstructured features with unknown or without multivariate structures (e.g., square footage, housing age)
  - $\mathcal{D} \subset \mathcal{M} \times \mathcal{X}$: joint feature space.
- Nonparametric regression models

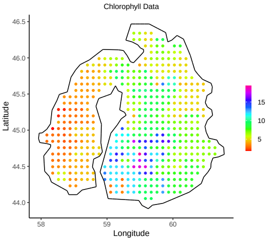
$$Y = f(\mathsf{s}, \mathsf{x}) + \epsilon, \tag{1}$$

where f is an unknown mean function and $\epsilon \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$ with unknown σ^2.

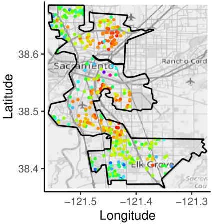
- Goal: estimate unknown f and predict for $(\mathsf{s}_{\text{new}}, \mathsf{x}_{\text{new}})$.

Main Challenges

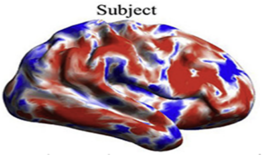
- Structured feature space $\mathcal{M}$ has a (known) non-trivial geometry (e.g., irregular boundary, interior hole, irregular 3-d surfaces)
- Irregular discontinuities in f (e.g., housing price)
- Potentially high-dimensional unstructured features x
- Potential interactions between s and x



(a) Aeal Sea



(b) Cities of Sacramento and Elk Grove, CA



(c) From Joshi et al. (2018)

Figure: Examples of complex constrained domains

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 ✗
  - Usually assume an additive model for x, e.g., $f(\mathrm{s}, \mathrm{x}) = \mathrm{x}^{\mathsf{T}}\boldsymbol{\beta} + f(\mathrm{s})$ ✗
  - Tensor product splines have too many basis functions for high-dimensional x ✗

Existing Methods

- Bayesian additive (univariate decision) regression trees (BART; Chipman et al., 2010):
  - Each tree generates axis-parallel partitions of the feature space
  - Approximate f with summation of simple piecewise constant functions
  - Local adaptivity to discontinuities and different levels of smoothness in f ✓
  - Capture some interaction effects among features ✓
  - Address feature scaling and feature selection issues with high dimensional $\mathbf{x}$ ✓
  - May not fully respect intrinsic geometries in $\mathcal{M}$ or capture irregular discontinuities in f ✗
- Bayesian additive spanning trees (BAST; Luo et al., 2021):
  - Using flexible spanning tree partitions for $\mathcal{M}$, which respects its intrinsic geometry ✓
  - Not straightforward to include unstructured features $\mathbf{x}$ ✗
  - Lack of a coherent model for prediction ✗

Semi-Multivariate Decision Trees (sMDTs)

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 $\{\mathcal{D}_{\eta,1}, \mathcal{D}_{\eta,2}\}$ of $\mathcal{D}_\eta$:
   2.1 With probability p_m, perform a multivariate split using the structured features s.
   2.2 Otherwise, perform a univariate split using one of the unstructured features x.
3. Apply Step 2 to each offspring node of η.

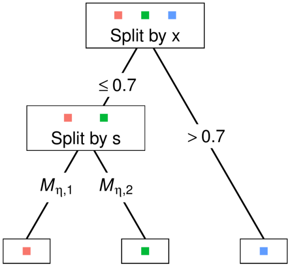
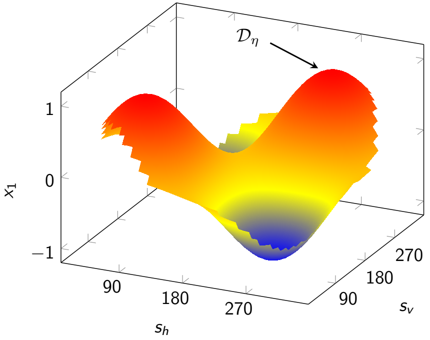


Figure: An example of sMDT

Univariate Split Rules

- A node η in an sMDT represents a subset $\mathcal{D}_\eta \subset \mathcal{D}$.
  - $\mathcal{D}_\eta = \mathcal{D}$ if η is the root node.

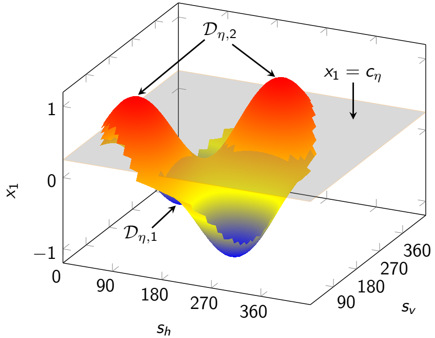


Univariate Split Rules

- A node η in an sMDT represents a subset $\mathcal{D}_\eta \subset \mathcal{D}$.
  - $\mathcal{D}_\eta = \mathcal{D}$ if η is the root node.
- A univariate split rule divides $\mathcal{D}_\eta$ into

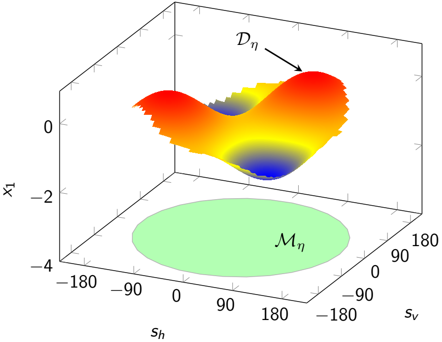
$$\mathcal{D}_{\eta,1} = \{(\mathrm{x}, \mathrm{s}) \in \mathcal{D}_\eta : x_{j(\eta)} \leq c_\eta\}, \quad \mathcal{D}_{\eta,2} = \mathcal{D}_\eta \setminus \mathcal{D}_{\eta,1},$$

for some coordinate $j(\eta) \in \{1, \ldots, p\}$ where $p = \dim(\mathrm{x})$.



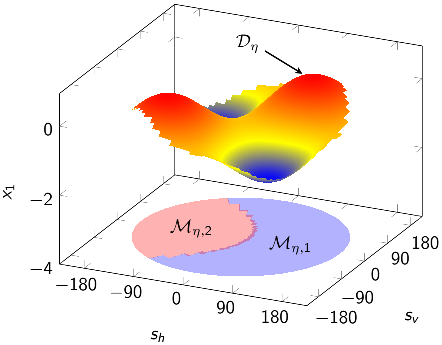
Multivariate Split Rules

- Project $\mathcal{D}_\eta$ to $\mathcal{M}$ to obtain $\mathcal{M}_\eta$. Partition $\mathcal{M}_\eta$ into $\{\mathcal{M}_{\eta,1}, \mathcal{M}_{\eta,2}\}$.



Multivariate Split Rules

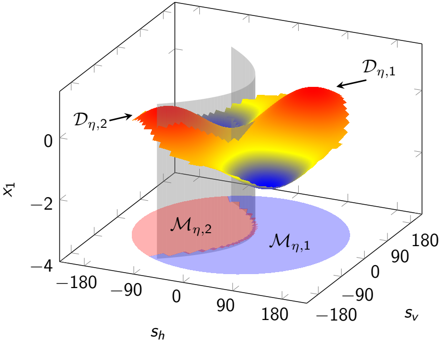
- Project $\mathcal{D}_\eta$ to $\mathcal{M}$ to obtain $\mathcal{M}_\eta$. Partition $\mathcal{M}_\eta$ into $\{\mathcal{M}_{\eta,1}, \mathcal{M}_{\eta,2}\}$.



Multivariate Split Rules

- Project $\mathcal{D}_\eta$ to $\mathcal{M}$ to obtain $\mathcal{M}_\eta$. Partition $\mathcal{M}_\eta$ into $\{\mathcal{M}_{\eta,1}, \mathcal{M}_{\eta,2}\}$.
- A structured multivariate split rule divides $\mathcal{D}_\eta$ into

$$\mathcal{D}_{\eta,k} = \mathcal{D}_\eta \cap (\mathcal{M}_{\eta,k} \times \mathcal{X}), \quad \text{for } k = 1, 2.$$



Multivariate Split Rules

- Project $\mathcal{D}_\eta$ to $\mathcal{M}$ to obtain $\mathcal{M}_\eta$. Partition $\mathcal{M}_\eta$ into $\{\mathcal{M}_{\eta,1}, \mathcal{M}_{\eta,2}\}$.
- A structured multivariate split rule divides $\mathcal{D}_\eta$ into

$$\mathcal{D}_{\eta,k} = \mathcal{D}_\eta \cap (\mathcal{M}_{\eta,k} \times \mathcal{X}), \quad \text{for } k = 1, 2.$$

- Main challenges:
  - $\mathcal{M}_\eta$ varies with nodes η and can be disconnected.
  - How to partition $\mathcal{M}_\eta$ such that both $\mathcal{D}_{\eta,1}$ and $\mathcal{D}_{\eta,2}$ contain non-empty subsets of observations?
  - How to partition $\mathcal{M}_\eta$ into subsets with flexible shapes while respecting its intrinsic geometry?

Manifold Bipartitions via Predictive Spanning Trees

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

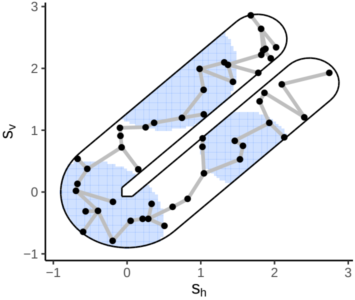


Figure: A predictive spanning tree bipartition

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.

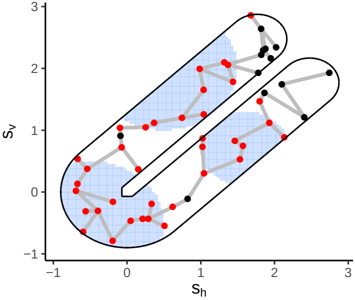


Figure: A predictive spanning tree bipartition

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 s^* and t^* from $\mathcal{S}^*_\eta$.
  3. Randomly sample an edge e^* from the unique path in $\mathcal{G}^*_T$ connecting s^* and t^*.

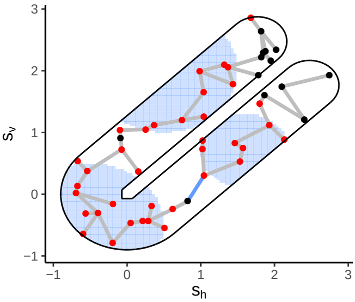


Figure: A predictive spanning tree bipartition

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 s* and t* from $\mathcal{S}_\eta^*$.
  3. Randomly sample an edge e^* from the unique path in $\mathcal{G}_T^*$ connecting s* and t*.
  4. Remove e^* from $\mathcal{G}_T^*$ to obtain bipartitions of $\mathcal{S}_\eta^*$ and $\mathcal{M}_\eta$.

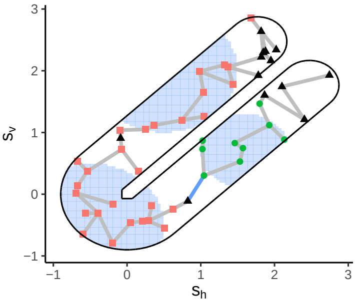


Figure: A predictive spanning tree bipartition

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 s* and t* from $\mathcal{S}_\eta^*$.
  3. Randomly sample an edge e^* from the unique path in $\mathcal{G}_T^*$ connecting s* and t*.
  4. Remove e^* from $\mathcal{G}_T^*$ to obtain bipartitions of $\mathcal{S}_\eta^*$ and $\mathcal{M}_\eta$.

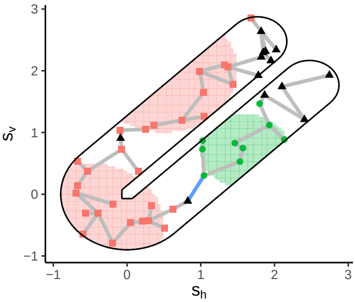


Figure: A predictive spanning tree bipartition

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

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

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

- BAMDT models $f(\mathrm{s}, \mathrm{x})$ with smmation of piecewise constant functions:

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

- Regularization prior:

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

  - Generative prior model for $\{T_m\}$ that encourages shallow sMDTs.
  - Shrinkage Gaussian prior for μ_m.

Bayesian Inference

- To draw a posterior sample from $[T_m|-]$ with $\boldsymbol{\mu}_m$ 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 s, x_1, ..., or x_p.

Bitten Torus Example

- Simulate spatially correlated features x with $p \in \{2, 10\}$.
- The true function only depends on s and x_1.
- When $p = 10$, avg. % of splits involving (s, x_1) in BAMDT is 73% (vs 63% in BART).

Table: Average prediction performance over 50 replicates for $p = 10$.

	MSPE	MAPE	CRPS
BAMDT	1.17	0.62	0.49
BART	2.09	0.79	0.65
GP-iso	1.56	0.80	0.64
GP-aniso	1.60	0.82	0.65
BAST-s	1.61	0.81	0.59
BAST-KNN	2.06	0.85	0.63

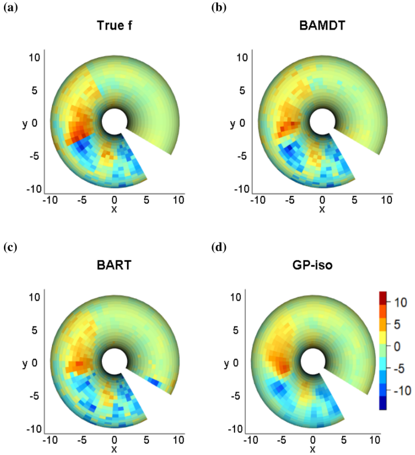


Figure: True function and predictive surfaces on $\mathcal{M}$ in the setting of $p = 2$

Application to Sacramento Housing Data

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

- BAMDT provides more accurate prediction than its competing methods based on 5-fold CV.

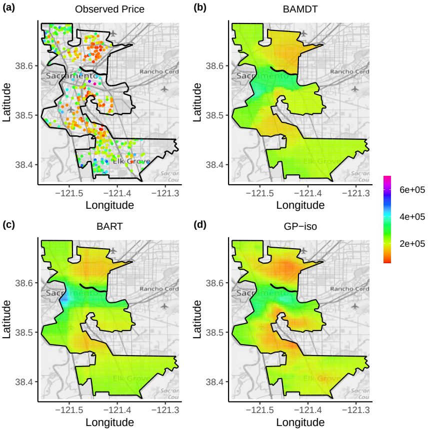


Figure: Observed data and predicted price for a representative house.

Conclusion and Future Work

- A novel Bayesian ensemble model, BAMDT, is developed for nonparametric semi-structured regression problems with complex structured feature spaces using flexible semi-multivariate decision trees 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.

Thanks!!

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.

Joshi, A. A., Chong, M., Li, J., Choi, S., and Leahy, R. M. (2018). Are you thinking what i'm thinking? synchronization of resting fmri time-series across subjects. *NeuroImage*, 172:740–752.

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.

Luo, Z. T., Sang, H., and Mallick, B. (2021). BAST: Bayesian additive regression spanning trees for complex constrained domain. *Advances in Neural Information Processing Systems*, 34.

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.

References II

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.

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.