Supplementary Materials of "BAST: Bayesian Additive Regression Spanning Trees for Complex Constrained Domain"

Anonymous Author(s)
Affiliation
Address
email

These appendices provide supplementary details and results of BAST. Appendix A contains additional details on Bayesian estimation and prediction. Supplementary simulation details and results can be found in Appendix B. Hyperparameter selection is also discussed in Appendix B. Finally, Appendix C provides the proof of Proposition 1.

Appendix A Details on Bayesian Inference

Appendix A.1 Estimation

This appendix provides details on the Markov chain Monte Carlo (MCMC) algorithm discussed in Section 3.1. We use $\mathbf{g}_m$ to denote the n-dimensional vector of fitted values at the training locations $\mathcal{S}$ from the mth RST partition, that is, the ith element of $\mathbf{g}_m$ is $g(\mathbf{s}_i|\pi_m, \mathcal{T}_m, k_m, \boldsymbol{\mu}_m)$. Let $\mathbf{X}_{\pi_m}$ be an $n \times k_m$ binary matrix where the (i,j)th element is 1 if and only if $\mathbf{s}_i$ is in the jth cluster under the partition π_m. We write the partial residual term for the mth RST partition as

$$\mathbf{r}_m = \mathbf{Y} - \sum_{\ell \neq m} \mathbf{g}_\ell.$$

Recall that our MCMC algorithm proceeds by successively sampling $(\pi_1, \mathcal{T}_1, k_1, \boldsymbol{\mu}_1)$,..., $(\pi_M, \mathcal{T}_M, k_M, \boldsymbol{\mu}_M)$, and σ^2 from their respective full conditional distributions. To sample from $p(\pi_m, \mathcal{T}_m, k_m, \boldsymbol{\mu}_m|-)$ for each $m = 1, \ldots, M$, we first sample the RST partition with $\boldsymbol{\mu}_m$ analytically integrated out, by performing a birth, a death, a change, or a hyper move with probability $r_b(k_m) = 0.3$, $r_d(k_m) = 0.3$, $r_c(k_m) = 0.3$, and $r_h(k_m) = 0.1$, respectively. Adjustments are made to the probabilities for the boundary cases where $k_m = 1$ and $k_m = \bar{k}$. This probability specification works well in our experiments, but one can modify it if desired. For the first three moves, the Metropolis-Hastings (M-H) acceptance ratio involves the integrated likelihood of $\mathbf{Y}$ given by

$$\mathcal{L}(\mathbf{Y}|\pi_m, \mathcal{T}_m, k_m, -) \propto |\mathbf{P}_{\pi_m}|^{-1/2} \exp\left(-\frac{1}{2}\mathbf{r}_m^\mathsf{T}\mathbf{P}_{\pi_m}^{-1}\mathbf{r}_m\right),$$

where $\mathbf{P}_{\pi_m} = \sigma^2\mathbf{I}_n + \sigma_\mu^2\mathbf{X}_{\pi_m}\mathbf{X}_{\pi_m}^\mathsf{T}$. The Sherman-Woodbury-Morrison formula is applied to simplify the computation of $\mathbf{P}_{\pi_m}^{-1}$ and $|\mathbf{P}_{\pi_m}|^{-1/2}$ as $\mathbf{X}_{\pi_m}\mathbf{X}_{\pi_m}^\mathsf{T}$ has a reduced rank k_m.

Conditional on a sample of $(\pi_m, \mathcal{T}_m, k_m)$, we sample $\boldsymbol{\mu}_m$ from $p(\boldsymbol{\mu}_m|\pi_m, \mathcal{T}_m, k_m, -)$, which is given by

$$[\boldsymbol{\mu}_m|\pi_m, \mathcal{T}_m, k_m, -] \sim \mathrm{N}_{k_m}\left(\mathbf{Q}_m\mathbf{b}_m, \mathbf{Q}_m\right),$$

where $\mathbf{Q}_m = \left(\frac{1}{\sigma^2}\mathbf{X}_{\pi_m}^\mathsf{T}\mathbf{X}_{\pi_m} + \frac{1}{\sigma_\mu^2}\mathbf{I}_{k_m}\right)^{-1}$ and $\mathbf{b}_m = \mathbf{X}_{\pi_m}^\mathsf{T}\mathbf{r}_m/\sigma^2$.

Finally, we sample σ^2 from its inverse-gamma full conditional given by

$$[\sigma^2|-] \sim \mathrm{IG}\left(\frac{n+\nu}{2},\ \frac{1}{2}\left[\nu\lambda_s + \|\mathbf{Y} - \sum_{m=1}^{M}\mathbf{g}_m\|^2\right]\right),$$

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Table S1: Prediction performance of BAST with $M = 20$ weak learners in the U-shape example. Results of BART with various larger numbers of weak learners M are included for comparison. Standard errors are given in parentheses.

		BAST ($M = 20$)	BART ($M = 50$)	BART ($M = 100$)	BART ($M = 200$)
$\sigma = 0.1$	MSPE	**0.189** (0.009)	1.430 (0.343)	1.302 (0.259)	1.219 (0.251)
	MAPE	**0.188** (0.007)	0.408 (0.046)	0.382 (0.033)	0.380 (0.027)
	Mean CRPS	**0.142** (0.008)	0.353 (0.043)	0.324 (0.030)	0.318 (0.024)
$\sigma = 0.5$	MSPE	**0.464** (0.044)	1.694 (0.362)	1.628 (0.277)	1.532 (0.166)
	MAPE	**0.491** (0.025)	0.682 (0.047)	0.695 (0.038)	0.711 (0.035)
	Mean CRPS	**0.371** (0.021)	0.557 (0.043)	0.553 (0.035)	0.554 (0.029)
$\sigma = 1$	MSPE	**1.283** (0.127)	2.546 (0.380)	2.441 (0.246)	2.429 (0.224)
	MAPE	**0.888** (0.049)	1.085 (0.052)	1.099 (0.052)	1.120 (0.050)
	Mean CRPS	**0.693** (0.042)	0.870 (0.047)	0.861 (0.045)	0.870 (0.044)

where $\|\cdot\|$ is the Euclidean norm.

Appendix A.2 Prediction in Two-dimensional Constrained Domains

In this subsection we provide details on specifying the neighbor set $N_{\mathbf{u}}$ for prediction at an unobserved location $\mathbf{u}$ in a constrained domain $\mathcal{D} \subset \mathbb{R}^2$. A constrained Delaunay triangulation (CDT) mesh can be constructed on $\mathcal{D}$ such that every unobserved location of interest is contained in a triangle. In the case where at least one triangle vertex is in $\mathcal{S}$, $N_{\mathbf{u}}$ is specified as those triangle vertices that belong to $\mathcal{S}$. Prediction at $\mathbf{u}$ is then performed as stated in Section 3.2.

In the extreme case where no triangle vertex is in $\mathcal{S}$, we choose $N_{\mathbf{u}}$ to be all the triangle vertices (which lie on the domain boundary). To sample the cluster membership of $\mathbf{u}$, we need to determine the cluster memberships for vertices on the domain boundary, which can be done by, for instance, assigning a boundary vertex to the same cluster as its nearest vertex in $\mathcal{S}$ with respect to the graph distance in the CDT mesh (when the number of vertices in the CDT graph is large, we expect this to well approximate the geodesic distance). Once we obtain the cluster memberships for boundary vertices, we can sample $z_m(\mathbf{u})$ from the cluster memberships of the vertices in $N_{\mathbf{u}}$ as in Section 3.2.

Appendix B Supplementary Simulation Results

We implement BAST in `R` and fit BART and SFS using `R` packages `BART`[1] [2] and `mgcv`[2] [3], respectively. The code for inGP is adopted from `https://github.com/mu2013/Intrinsic-GP-on-complex-constrained-domain`. Experiments are performed on a Linux machine with two 2.4GHz 14-core processors and 64GB memory. Code will be made publicly available upon request of revision or acceptance of the manuscript.

Appendix B.1 U-shape Example

To demonstrate that BAST is more efficient than its binary treed competitors in recovering irregularly shaped regions where discontinuities happen in complex domains, we compare BAST with $M = 20$ to BART with various numbers of weak learners. The experiment setup is the same as in Section 4.1 except for the number of binary decision trees used in BART.

As shown in Table S1, BAST outperforms BART even when BART uses more weak learners, confirming that BART needs much more rectangular partitions to approximate irregularly shaped discontinuity boundaries, while BAST can recover them with only a few RST edge cuts.

Next, we consider selecting hyperparameters of BAST via cross-validation (CV) in the U-shape example with true noise standard deviation $\sigma = 0.1$. More specifically, for each replicate data set, we choose the number of weak learners M, the maximum number of clusters in each RST partition $\bar{k}$, and the shrinkage parameter a that controls prior concentration around zero for $\boldsymbol{\mu}_m$ using 5-fold CV within the training data based on MSPE. The candidate values for each hyperparameter are

[1] License: GPL (>= 2)

[2] License: GPL (>= 2)

Table S2: Candidate values of hyperparameters for CV in the U-shape example.

Method	Hyperparameter	Candidate values
BAST	# of weak learners M	20, 30, 50
	Maximum # of clusters per partition $\bar{k}$	5, 10
	$\boldsymbol{\mu}$-prior shrinkage parameter a	1, 2, 3
BART	# of weak learners M	50, 100, 200
	$\boldsymbol{\mu}$-prior shrinkage parameter a	1, 2, 3

Table S3: Prediction performance of BAST and BART with and without CV in the U-shape example under noise level $\sigma = 0.1$. Standard errors are given in parentheses.

	BAST-cv	BAST-default	BART-cv	BART-default
MSPE	**0.186** (0.011)	0.189 (0.009)	1.277 (0.306)	1.541 (0.530)
MAPE	**0.182** (0.008)	0.188 (0.007)	0.390 (0.034)	0.436 (0.068)
Mean CRPS	**0.135** (0.014)	0.142 (0.008)	0.331 (0.032)	0.380 (0.066)

summarized in Table S2, and a total of 18 hyperparameter combinations are considered for BAST. For comparision, we also choose the number of weak learners and the prior shrinkage parameter of $\boldsymbol{\mu}_m$ for BART using 5-fold CV, and their candidate values can be also found in Table S2.

Table S3 shows the performance of BAST and BART using the hyperparameters chosen by CV (referred to as BAST-cv and BART-cv, respectively). As a benchmark, the performance metrics for BAST and BART using the hyperparameters in Section 4.1 are also included (referred to as BAST-default and BART-default, respectively). The fine-tuned BAST-cv achieves better performance than BAST-default as expected, but the performance of them is close to each other, suggesting that BAST is robust to the choices of hyperparameters in this example. Both versions of BAST outperform BART with and without hyperparameter selection.

Appendix B.2 Bitten Torus Example

We consider the bitten torus example in Section 4.2 but with noise levels $\sigma = 0.5$ and $\sigma = 1$. The results are summarized in Table S4. Consistent to the finding under the noise level $\sigma = 0.1$, BAST performs the best among all three methods.

As in Appendix B.1, we also experiment with choosing hyperparameters via 5-fold CV for the data sets with true noise level $\sigma = 0.1$. In addition to the BAST hyperparameters in Table S2, we also select K, the size of the predictive neighbor set $N_{\mathbf{u}}$ discussed in Section 3.2, from its candidate values $\{3, 4, 5, 6\}$. As shown in Table S5, BAST outperforms BART in both CV and default settings. Our results again confirm that BAST performs reasonably well even without hyperparameter tunning.

Appendix C Proof of Proposition 1

Proof 1 *For any spatially continuous partition $\pi(\mathcal{S})$ with k clusters, it follows from Propositions 2 of Luo et al. [1] that there exists a spanning tree $\mathcal{T}$ of $\mathcal{G}$ and a set of $k-1$ edges in $\mathcal{T}$ that induce $\pi(\mathcal{S})$. Hence, conditional on $\mathcal{T}$, the conditional probability for $\pi(\mathcal{S})$ is strictly positive due to* (2) *and* (4)*. To show $\mathcal{T}$ is within the support of* (3)*, note that $\mathcal{T}$ is the MST of $\mathcal{G}$ given the edge weights satisfying $\omega_e \in (0, 1/2)$ if $e \in \mathcal{E}_{\mathcal{T}}$ and $\omega_e \in (1/2, 1)$ if $e \notin \mathcal{E}_{\mathcal{T}}$. This completes the proof.*

References

[1] Luo, Z. T., Sang, H., and Mallick, B. (2021). A Bayesian contiguous partitioning method for learning clustered latent variables. *Journal of Machine Learning Research*, 22(37):1–52.

[2] McCulloch, R., Sparapani, R., Gramacy, R., Spanbauer, C., and Pratola, M. (2019). *BART: Bayesian Additive Regression Trees*. R package version 2.7.

Table S4: Prediction performance of BAST and its competing methods in the bitten torus example under different noise levels. Standard errors are given in parentheses.

		BAST	BART	inGP
$\sigma = 0.5$	MSPE	**0.754** (0.053)	1.358 (0.270)	2.601 (0.234)
	MAPE	**0.584** (0.024)	0.682 (0.045)	1.240 (0.069)
	Mean CRPS	**0.405** (0.022)	0.567 (0.043)	—
$\sigma = 1$	MSPE	**1.568** (0.140)	2.378 (0.354)	4.628 (3.117)*
	MAPE	**0.960** (0.048)	1.092 (0.062)	1.648 (0.469)*
	Mean CRPS	**0.706** (0.041)	0.904 (0.060)	—

* The results for inGP under $\sigma = 1$ are based on 49 replicates due to numerical errors in one replicate data set.

Table S5: Prediction performance of BAST and BART with and without CV in the bitten torus example under noise level $\sigma = 0.1$. Standard errors are given in parentheses.

	BAST-cv	BAST-default	BART-cv	BART-default
MSPE	**0.463** (0.055)	0.487 (0.012)	0.850 (0.141)	1.115 (0.287)
MAPE	**0.287** (0.029)	0.307 (0.006)	0.370 (0.031)	0.406 (0.062)
Mean CRPS	**0.216** (0.022)	0.225 (0.017)	0.310 (0.027)	0.355 (0.059)

[3] Wood, S. (2017). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2nd edition.