BART - Bayesian Additive Regression Trees

Giacomo Petrillo - giacomo.petrillo@unifi.it
Giovanni Poli - giovanni.poli@unifi.it

Università degli Studi di Firenze

March 18, 2022

1. Introduction
2. Model and Prior
3. Prior Interpretation
4. Posterior Inference
5. Examples of posterior inference
6. References

1 Introduction

2 Model and Prior

3 Prior Interpretation

4 Posterior Inference

5 Examples of posterior inference

6 References

BART Definition

BART stands for Bayesian Additive Regression Trees.

We aim to make inference about an unknown function $f(\boldsymbol{x})$ defined on a p dimensional vector $\boldsymbol{x} = (x_1, \ldots, x_p)^\top$.

$$y = f(\boldsymbol{x}) + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

The BART model consists in:

$$f(\boldsymbol{x}) \approx h(\boldsymbol{x}) := \sum_{j=1}^{m} g_j(\boldsymbol{x}) \quad \Longrightarrow \quad y = \sum_{j=1}^{m} g_j(\boldsymbol{x}) + \varepsilon$$

Where each $g_j(\boldsymbol{x})$ denotes a regression tree.

Extension for binary data

Extending BART for binary data is simple by integrating it with a probit Bayesian model.

$$y = \begin{cases} 1 & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases}$$

$$z = h(\boldsymbol{x}) + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0,1) \qquad h(\boldsymbol{x}) = \sum_{j=1}^{m} g_j(\boldsymbol{x}, \mathcal{T}_j, \mathcal{M}_j)$$

from which we derive:

$$z \mid y = 1,\ h(\boldsymbol{x}) \sim \max\left\{\mathcal{N}\big(h(\boldsymbol{x}),1\big), 0\right\} \qquad z \mid y = 0,\ h(\boldsymbol{x}) \sim \min\left\{\mathcal{N}\big(h(\boldsymbol{x}),1\big), 0\right\}$$

$$\mathbb{E}\big[y|h(\boldsymbol{x})\big] = p\big(y = 1\ |h(\boldsymbol{x})\big) = \Phi\big(h(\boldsymbol{x})\big)$$

Bayesian Conceptualization

To use the model in a Bayesian perspective, we must define a prior and a way to sample from the posterior distribution.

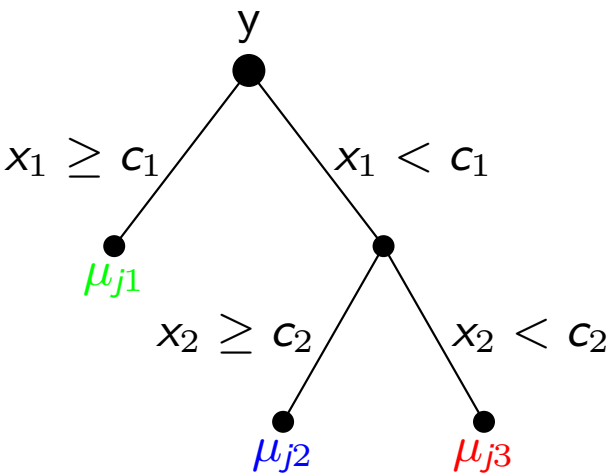
The prior is composed by:

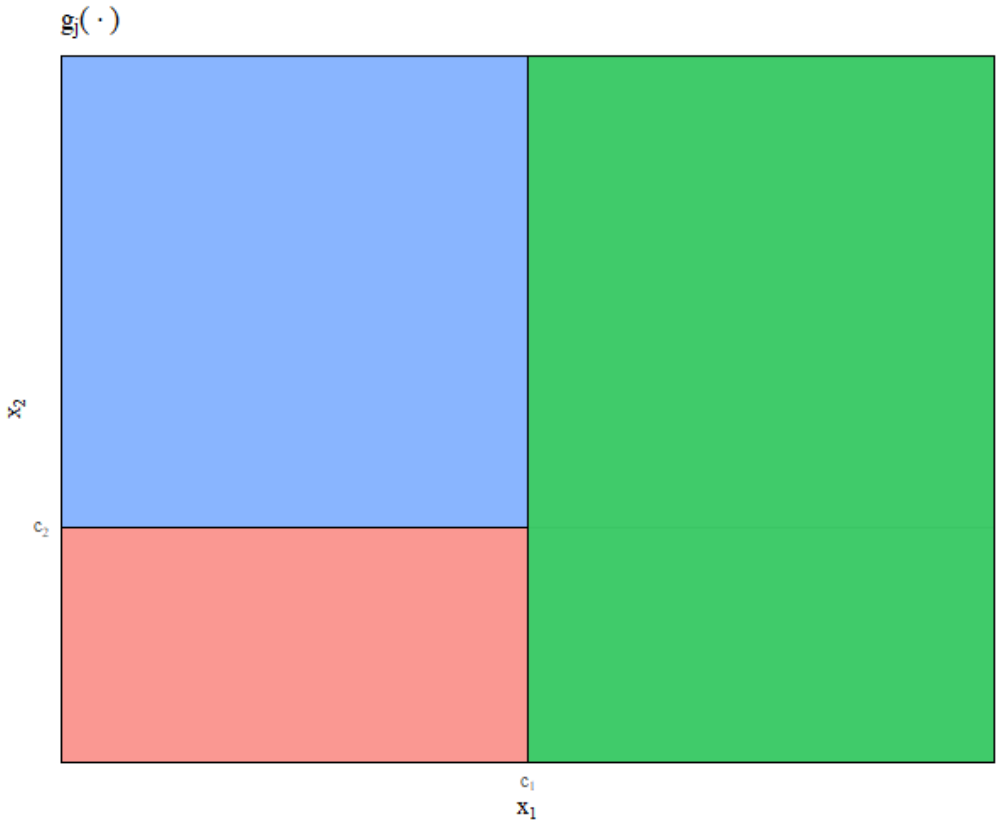
1. Prior distribution for σ^2
2. A prior distribution for each tree $g_j(\boldsymbol{x}, \mathcal{T}_j, \mathcal{M}_j)$, where:
   - $\mathcal{T}_j$ is the structure of the tree (interior node decision rules and terminal nodes).
   - $\mathcal{M}_j = \{\mu_{j1}, \ldots, \mu_{jb_j}\}$ is the set of function values associated to each of the b_j terminal nodes of $\mathcal{T}_j$.

This single-tree prior distribution was originally introduced by Chipman, George ad McCulloch in 1998 [CGM98], who also suggested an algorithm for posterior inference.

Bayesian CART - Chipman, George ad McCulloch (1998)

$$\mathcal{T}_j = \{\{y \in \mathcal{A}_1 \text{ if } x_1 \geq c_1\}, \{y \in \mathcal{A}_2 \text{ if } x_1 < c_1 \cap x_2 < c_2\}, \{y \in \mathcal{A}_3 \text{ otherwise }\}\}$$
$$\mathcal{M}_j = \{\mu_{j1}, \mu_{j2}, \mu_{j3}\}$$





1 Introduction

2 Model and Prior

3 Prior Interpretation

4 Posterior Inference

5 Examples of posterior inference

6 References

The Complete Prior

The prior is defined allowing each tree to have a prior distribution independent from that of the other trees:

$$
\begin{aligned}
p\big((\mathcal{T}_1, \mathcal{M}_1), \ldots, (\mathcal{T}_m, \mathcal{M}_m), \sigma^2\big) &= \left[\prod_{j=1}^{m} p(\mathcal{T}_j, \mathcal{M}_j)\right] p(\sigma^2) \\
&= \left[\prod_{j=1}^{m} p(\mathcal{M}_j|\mathcal{T}_j) p(\mathcal{T}_j)\right] p(\sigma^2) \\
&= \left[\prod_{j=1}^{m} \left[\prod_{i=1}^{b_j} (\mu_{ij}|\mathcal{T}_j)\right] p(\mathcal{T}_j)\right] p(\sigma^2)
\end{aligned}
$$

Where:

- m is the fixed number of trees
- b_j is given by $\mathcal{T}_j$ and is the number of terminal nodes of the j-th tree.

$\mathcal{T}_j$ Distribution

Let's focus on a single tree structure and define a prior on it. The distribution is defined recursively for each node in three steps:

1. The probability for the node of having children is

$$P_{n\mathcal{T}_j} = \alpha(1+d)^{-\beta} \qquad \alpha \in (0,1), \beta \in [0,\infty)$$

   $d =$ node depth (root $= 0, \ldots$).
2. Conditional on having children, which variable to split on (discrete uniform).
3. Conditional on the choice of variable, which splitting point to use (discrete uniform).

$\mu_{ji}|\mathcal{T}_j$ and σ^2 Distribution

We assumed:

$$y_i = \sum_{j=1}^{m} g_j(\boldsymbol{x}_i, \mathcal{T}_j, \mathcal{M}_j) + \varepsilon_i \qquad \text{with } \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

This means that the distribution of y conditional on the values of the parameters and hyperparameters is normal. It is therefore convenient for σ^2 and for each μ_{ji} to have conjugate distributions.

$$\mu_{ji}|\mathcal{T}_j \overset{iid}{\sim} \mathcal{N}(\mu_\mu, \sigma^2_\mu)$$
$$\sigma^2 \sim \lambda \cdot \mathcal{I}\chi^2_\nu$$

Note:

The inverse-chi-squared distribution is very heavy tailed when ν is small (no variance for $\nu \leq 4$).

Hyperparameters

While defining the prior we have introduced a total of 7 hyperparameters:

- α and β, (tree structure $\mathcal{T}_j$)
- μ_μ and σ_μ, (tree leaves $\mathcal{M}_j$)
- m, (number of trees)
- λ and ν. (variance of the error term)

The default parameters proposed by the original paper are:

- $\alpha = 0.95$, $\beta = 3$ to prefer shallow trees.
- μ_μ, σ_μ and λ calibrated on the center and the dispersion of the data.
- $m = 200$ to use many trees.
- $\nu = 3$ to have a very heavy tail that allows large values of σ^2.

1 Introduction

2 Model and Prior

3 Prior Interpretation

4 Posterior Inference

5 Examples of posterior inference

6 References

Prior Generative Model

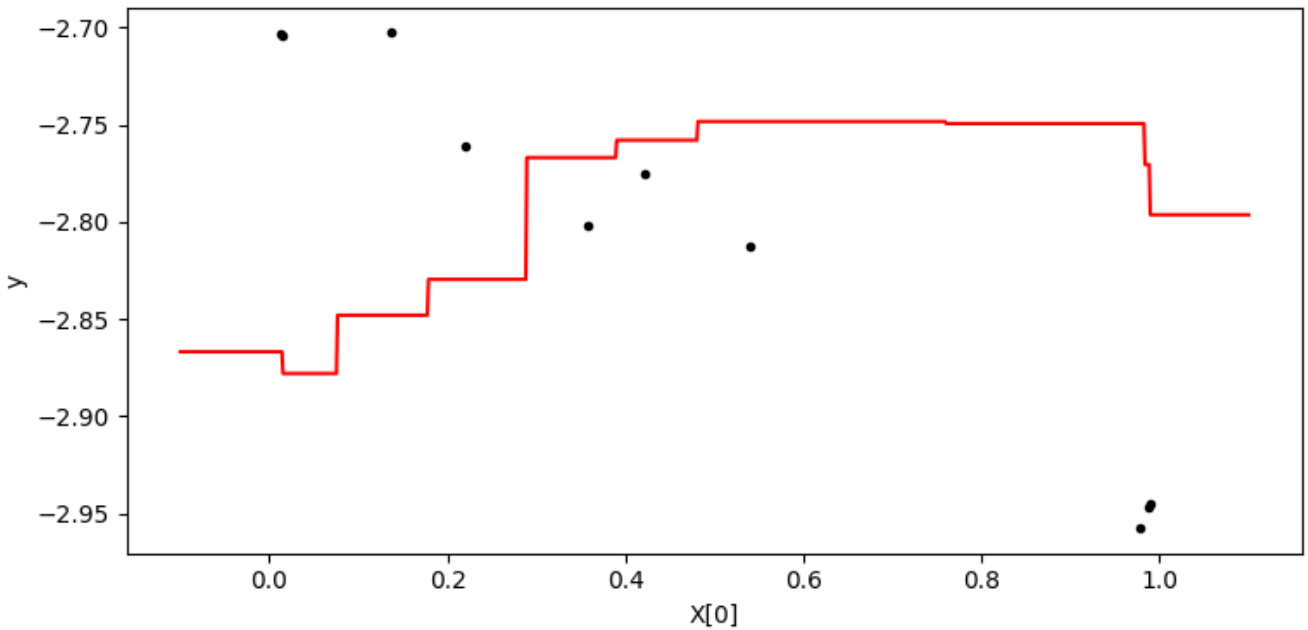
From the math it is not easy to interpret the prior, so in this section we will give a graphical representation.

The BART prior is a distribution on the space of step functions, so we can sample functions from it and plot them.

We will focus exclusively on the sum of trees $\sum_{j=1}^{m} g_j(x, \mathcal{T}_j, \mathcal{M}_j)$, since the error term is easier to interpret.

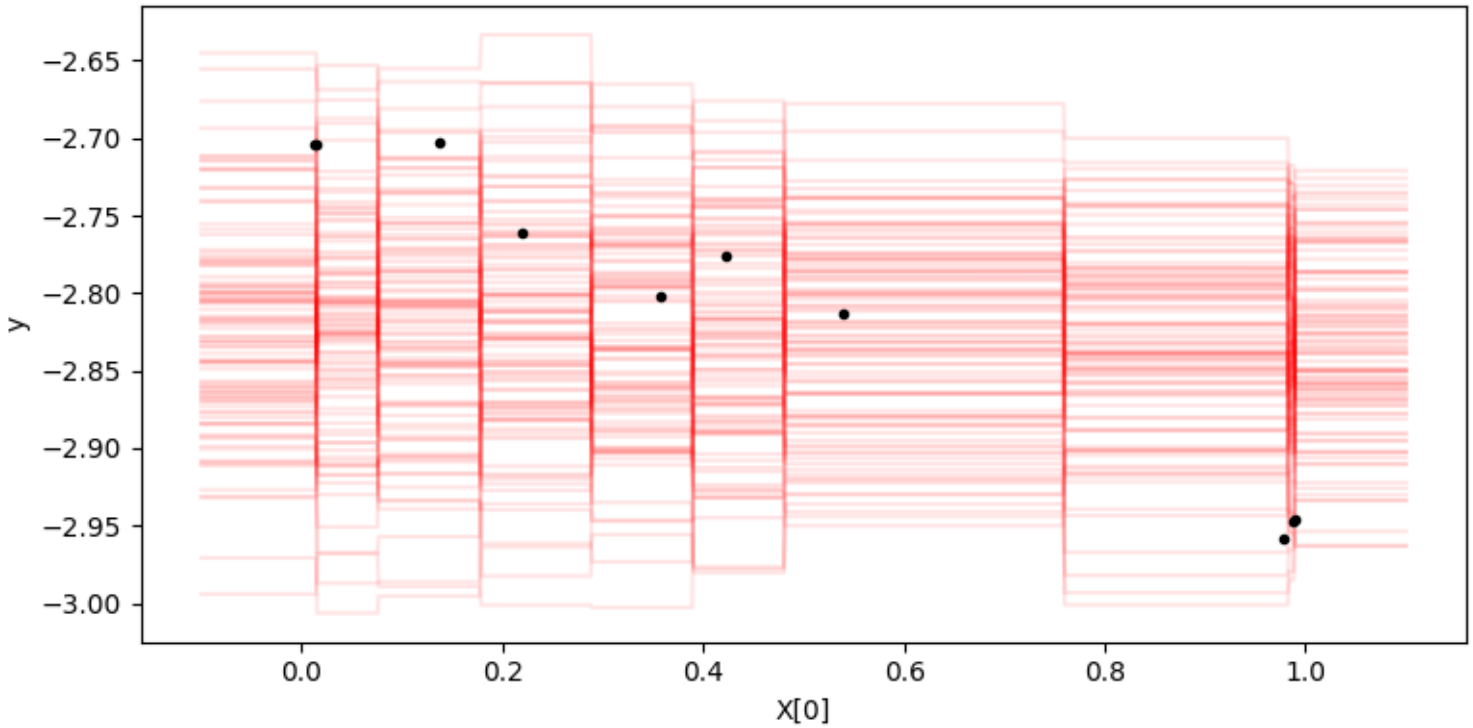
$\sum_j g_j(x, \mathcal{T}_j, \mathcal{M}_j)$ (in 1D)

We generate random data along a line (the data is only used to calibrate the hyperparameters as in the original paper) and draw and draw a function from the prior.



$\sum_j g_j(x, \mathcal{T}_j, \mathcal{M}_j)$ (in 1D) $\times 100$

Now 100 functions.



The Splitting Points

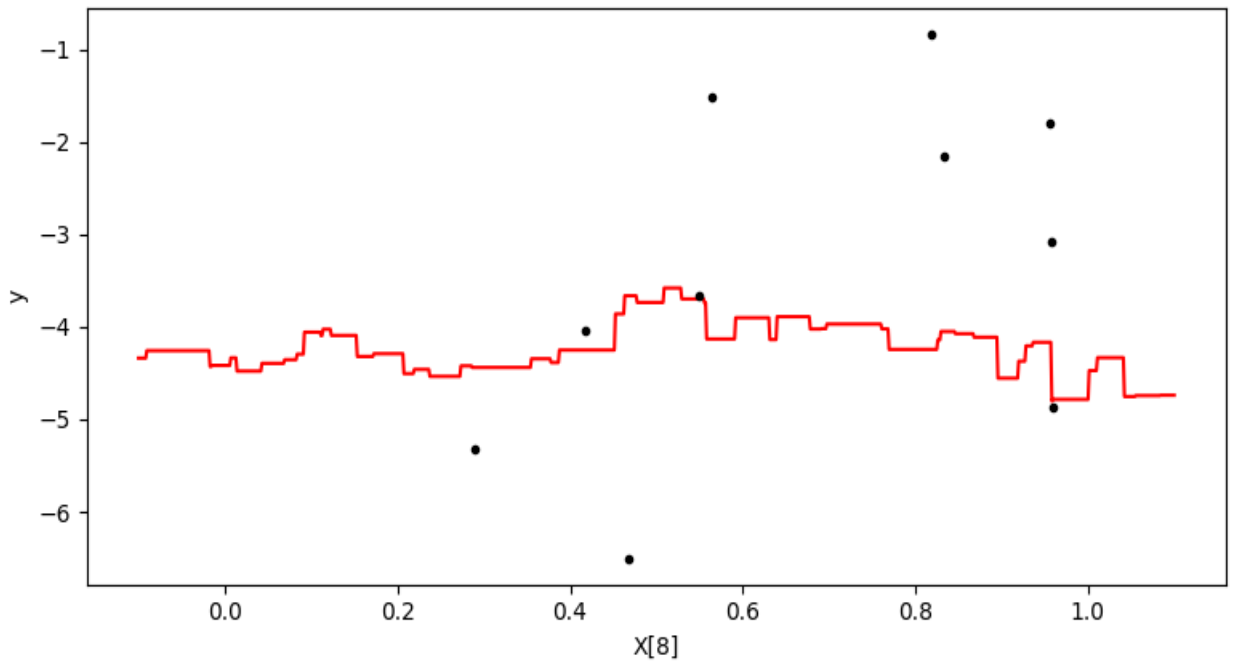
From these graphs it can be seen that the split always takes place halfway between one observation and another. Is this a too narrow restriction on the model?

You can do a continuous version of the BART without the steps, but it is not necessary because:

- Lots of data $\Longrightarrow$ lots of splitting points
- Many dimensions $\Longrightarrow$ even more splitting points!

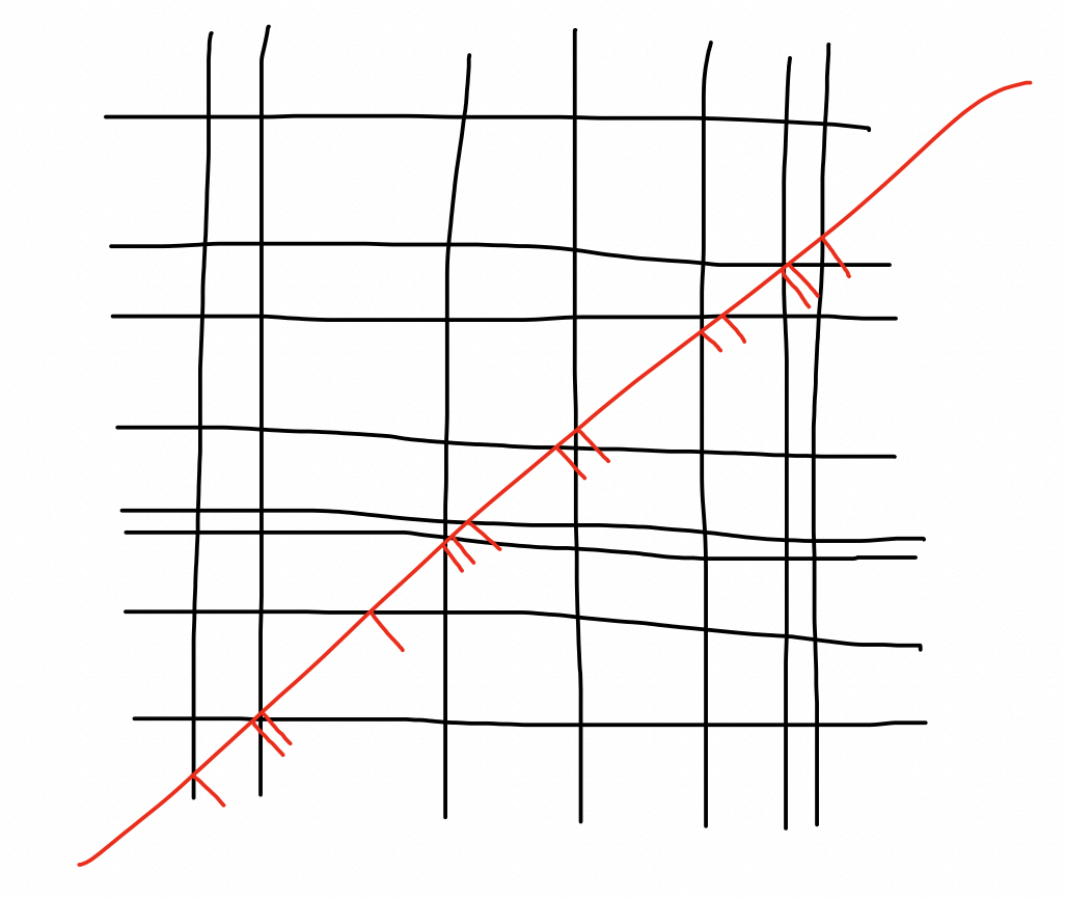
Increasing the Dimensionality

Here the above example has been replicated but with 20 covariates (the section shown is not parallel to any coordinate):



Increasing the Dimensionality

Note that there are many more steps because we are intersecting the function in an oblique direction.



Increasing the Dimensionality

BART is useful when there are many covariates and a lot data (with few covariates you can write an explicit model, with little data you can use a kernel method), so this is the situation we are interested in.

Problems of Dimensionality and Scalability

In general, with p covariates and n observations, the space of the step functions has dimension n^p (the values on each single little ipercube in the p-dimensional grid).

(If there are categorical covariates it is smaller, it is $n_1 n_2 \cdots n_p$.)

Example: with 100 datapoints and 20 covariates, the function space has dimension $100^{20} = 10^{40}$.

It is huge! How can the Markov Chain work efficiently in such an high-dimensional space?

Prior Parametric Space Reduction

Hypothesis:

The prior is reducing the space to explore. Not just the hypervolume, it has to be reducing it to a submainfold with far fewer dimensions.

Prior Parametric Space Reduction

Example of this phenomenon:

$$x_1, \ldots, x_N \overset{iid}{\sim} \mathcal{N}(0, 1)$$

The joint distribution is N-dimensional but is actually concentrated in an $(N-1)$-dimensional space.

$$\chi^2 = x_1^2 + x_2^2 + \cdots + x_N^2$$

And for $N \to \infty$ we know that $\sqrt{\mathbb{V}[\chi^2]}/\mathbb{E}[\chi^2] \to 0$. The χ^2 is the squared distance from the origin, so we're focusing on the $(N-1)$-dimensional sphere.

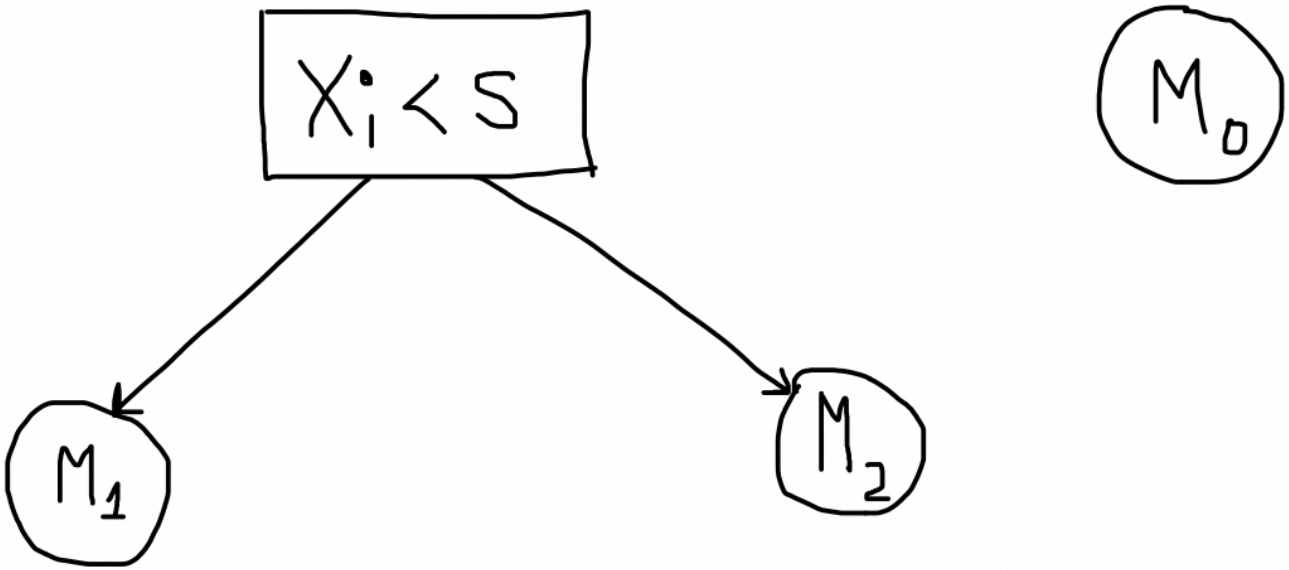
In other words: we have a "soft constraint" on the sum of squares which reduces the effective number of parameters.

$\beta \to \infty$

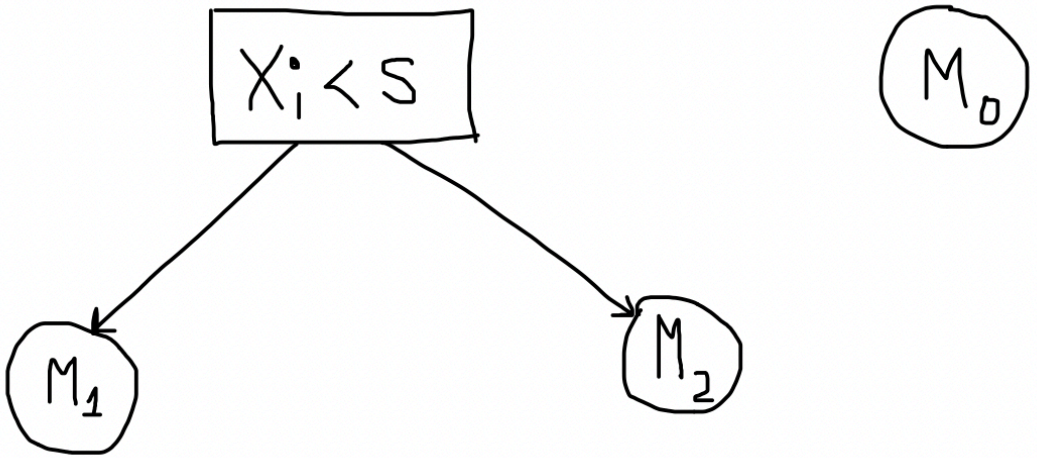
How can we search for something like that in the BART prior distribution which is more complicated?

Let's start from a simple case: the limit for $\beta \to \infty$.

β is the hyperparameter that regulates the probability of childspawn based on the depth of the tree node. $\beta \to \infty$ implies that the tree can have at most depth 1, i.e., either just the root or the root with two children.



$\beta \to \infty$

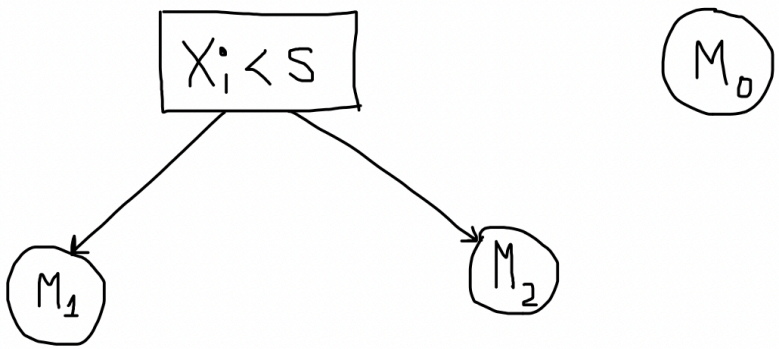


In this case, how many parameters do we need to describe the space of functions implied by the sum of trees?

The possible tree structures are $1 + (n-1)p$, i.e., the childless tree plus all possible two-children trees.

This is also the dimension of the function space, think of the tree fetus as setting an intercept and each bonsai as setting the difference between its two leaves.

$\beta \to \infty$



Another way of looking at this case: each tree can split along only one dimension. So the sum of trees is separable as a sum over dimensions:

$$h(\mathbf{x}) = h(x_1, \ldots, x_p) = h_1(x_1) + \cdots + h_p(x_p).$$

If h was a smooth function, this would be equivalent to the condition that, for $i \neq j$, $\partial^2 h / \partial x_i \partial x_j = 0$, i.e., the hessian is always diagonal. This holds in our case replacing the derivatives with discrete differences.

And When $\beta < \infty$?

So when $\beta \to \infty$ we have super-shallow trees and everything is simple. As β is reduced, we will still have somewhat shallow trees, but not so much.

Hypothesis: when $\beta < \infty$, the constraint becomes soft but remains, i.e., the dimensions that matter are still more or less $1 + (n-1)p$.

How can we verify this?

BART is a Gaussian Process

The BART model is a sum of *many* trees. Default 200, increase if needed.

Summing many things, the distribution tends to a multivariate normal, so the BART is approximately a Gaussian process.

A Gaussian distribution is completely characterized by its covariance matrix, so we can limit ourselves to studying the correlation function.

Correlation Function for $\beta \to \infty$

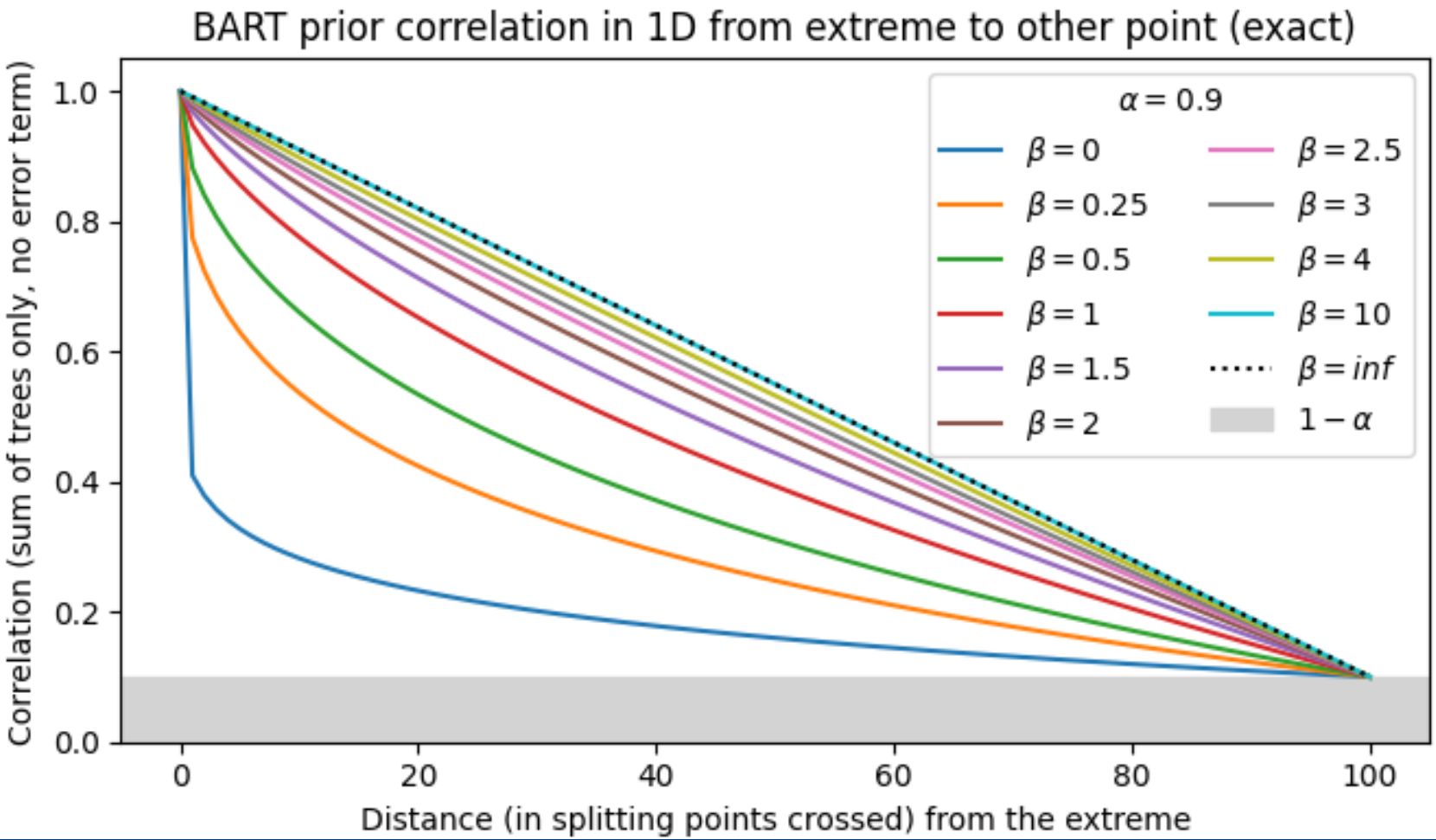
In the $\beta \to \infty$ case the correlation between $h(\mathbf{x})$ and $h(\mathbf{x}')$ can be computed in closed form:

$$\mathrm{Corr}[h(\mathbf{x}), h(\mathbf{x}')] = 1 - \alpha \frac{1}{p} \sum_{i=1}^{p} \frac{n_i^0}{n-1},$$

where n_i^0 is the number of splitting points between $\mathbf{x}$ and $\mathbf{x}'$ along the i-th covariate.

Correlation Function vs. β

Plot of the correlation function varying β. Consider that the default choice is $\beta = 3$.



$\infty = 3$

So in practice the BART model is used with a prior which is close to the case $\beta \to \infty$, implying it keeps the probability mass near a submanifold with $1 + (n-1)p$ dimensions.

(Note: the plot above was made in 1D varying y with x fixed to an extremity, you can check that everything works fine with multiple covariates and with different choices of x and y. In particular the convergence speed w.r.t. β is not spoiled.)

Minimum Correlation

Another interesting thing about the prior:

The minimum correlation is $1 - \alpha$, so 5 % by default.

So with the BART prior your are telling this: if you have even 1000000 datapoints, with the first very far away from the last, you assume that a priori they are all correlated, with the opposite vertices of the hypergrid correlated to 5 %.

(With many covariates the correlation on the farthest pair of datapoints tends to be way larger than that because it is unlikely to find two datapoints so opposite in the grid.)

Linear regression: the correlation of diametrically opposed points tends to -100 %.

Usual kernel methods: the correlation tends to zero as the distance increases.

Warped Distance

Another thing to notice:

The only thing that matters is the number of splitting points between two points.

(Only approximately. There is a slight edge effect. It is exactly true only for $\beta \to \infty$.)

So it is like doing a fit in a deformed space where the distance is given by the density of the observed $\mathbf{x}$s.

In other words, you are using the observed distribution of $\mathbf{x}$ to fix the a priori distribution on y.

1 Introduction

2 Model and Prior

3 Prior Interpretation

4 Posterior Inference

5 Examples of posterior inference

6 References

Backfitting Bayesian Algorithm

A Bayesian context induces the need to sample from a posterior distribution. For BART it means to sample from:

$$p\left((\mathcal{T}_1, \mathcal{M}_1), \ldots, (\mathcal{T}_m, \mathcal{M}_m), \sigma^2 \mid \boldsymbol{y}\right)$$

The original paper refers to the sampler as a backfitting MCMC algorithm. The backfitting procedure is a modular way of fitting an additive mode that cycles through the predictors.

This is equivalent to Gibbs sampling for an appropriately defined Bayesian model (Hastie and Tibshirani (2000)[HT00]).

Notation

Defined with:

- $R_j = \boldsymbol{y} - \sum_{k \neq j} g(X, \mathcal{T}_k, \mathcal{M}_k)$ residuals due to $\mathcal{T}_j$
- $\boldsymbol{R_{(j)}} = \{R_k\}_{k \neq j}$ residuals excluding $\boldsymbol{R_j}$
- $\mathcal{T}_{(j)} = \{\mathcal{T}_k\}_{k \neq j}$ trees excluding $\mathcal{T}_j$
- $\mathcal{M}_{(j)} = \{\mathcal{M}_k\}_{k \neq j}$ terminal-node parameters excluding $\mathcal{M}_j$
- $\mathcal{M}, \mathcal{T}, \boldsymbol{R}$ complete sets

So each posterior sample will:

1. $\forall j \in \{1, \ldots, m\}$, sample $(\mathcal{T}_j, \mathcal{M}_j)$ from their full condition.
2. Sample σ^2 from his full condition.

Step I - $\mathcal{T}_j, \mathcal{M}_j$ (1)

We can note that:

$$\mathcal{T}_j, \mathcal{M}_j | \ \mathcal{T}_{(j)}, \mathcal{M}_{(j)}, \sigma^2, \boldsymbol{y} \quad \text{is equal to} \quad \mathcal{T}_j, \mathcal{M}_j | \ \boldsymbol{R}_j, \sigma^2$$

We can sample from the joint distribution

1. sample new value for $\mathcal{T}_j$ from $\mathcal{T}_j | R_j, \sigma^2$
2. sample new value for $\mathcal{M}_j$ from $\mathcal{M}_j | R_j, \sigma^2$

$$\begin{aligned} p(\mathcal{T}_j \, | \boldsymbol{R}_j, \sigma^2) &\propto p(R_j | \mathcal{T}_j, \sigma^2) p(\mathcal{T}_j, \sigma^2) \\ &\propto p(\mathcal{T}_j) \ p(\sigma^2) \int p(R_j, \mathcal{M}_j | \mathcal{T}_j, \sigma^2) d\mathcal{M}_j \\ &\propto p(\mathcal{T}_j) \int p(R_j | \mathcal{M}_j, \mathcal{T}_j, \sigma^2) p(\mathcal{M}_j | \mathcal{T}_j, \sigma^2) d\mathcal{M}_j \end{aligned}$$

Step I - $\mathcal{T}_j, \mathcal{M}_j$ (2)

Since we can analytically calculate the previous expression, we can sample a new value for $\mathcal{T}_j$ through a step of Metropolis–Hastings.

The proposal distribution used is the same of Chipman et Al. (1998) [CGM98]. It propose a new tree choosing amongst:

- Growing a terminal node with $p = 0.25$;
- Pruning a pair of terminal nodes with $p = 0.25$;
- Changing a non-terminal rule with $p = 0.40$;
- Swapping a rule between parent and child with $p = 0.10$.

Step I - $\mathcal{T}_j, \mathcal{M}_j$ (3)

Now that we defined a proposal, given T_j^b, M_j^b and σ^b as the parameters values for the Markov Chain state at b iteration:

1. Sample from $\mathcal{T}_j|\sigma^2$ (for each m)
   - Sample a new value T^* from $\mathcal{T}^*|T_j^b$, where $\mathcal{T}^*$'s mass is induced by the previous probabilities.
   - Calculate acceptance ratio

$$\alpha = \min\left\{1, \frac{p(T^*)\int p(R_j|\mathcal{M}_j, T^*, \sigma^b)p(\mathcal{M}_j|T^*, \sigma^b)d\mathcal{M}_j}{p(T_j^b)\int p(R_j|\mathcal{M}_j, T_j^b, \sigma^b)p(\mathcal{M}_j|T_j^b, \sigma^b)d\mathcal{M}_j}\frac{p(T_j^b|T^*)}{p(T^*|T_j^b)}\right\}$$

   - Accept the transition ($T_j^b = T^*$) with $p = \alpha$ or keep the old value otherwise.
2. Sample from $\mathcal{M}_j|\sigma^b, \boldsymbol{T}^{(b+1)}$ (for each μ_{ji})
   - Given the data partition, set by T_j^b, the above distribution follows a conjugated normal-normal model

Step II - σ^2

As in the previous equations

$$p(\sigma^2|\mathcal{M}, \mathcal{T}, \boldsymbol{y}) \text{ is equal to } p(\sigma^2|\boldsymbol{R})$$

Therefore:

$$p(\sigma^2|\boldsymbol{R}) \propto p(\boldsymbol{R}|\sigma^2)p(\sigma^2)$$

Which follows a conjugated normal-inverse-gamma model.

Therefore step II just needs to sample the new value σ^b from its full conditional $\sigma^2|\boldsymbol{R}$.

Posterior distribution

After an appropriate burn-in period, the distribution sampled through the MCMC algorithm is the target posterior distribution, but the sampled parameters contain little information individually.

All the information of BARTs is contained in posterior distribution $f(\boldsymbol{x})|\boldsymbol{y}$. For example, given B samples from the posterior, a Monte Carlo point-wise estimate for the expected value can be calculated:

$$\mathbb{E}_{MC}\Big[f(\boldsymbol{x})|\boldsymbol{y}\Big] = \frac{1}{B}\sum_{b=1}^{B}\left[\sum_{j=1}^{m} g\left(\boldsymbol{x}, T_j^b, M_j^b\right)\right]$$

Additional tools for inference

A functional of particular interest can be the partial dependence function defined by Friedman (2001) [Fri01], which summarizes a marginal effect of one (or more) predictors on the response. More precisely, defining $f(\boldsymbol{x}) = f(\boldsymbol{x_s}, \boldsymbol{x_c})$, we define our marginal effect as:

$$f_s(\boldsymbol{x_s}) = \frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{x_s}, \boldsymbol{x_{ic}})$$

BART can also be used for variable selection by selecting those variables that appear most often in the fitted sum-of-trees models. Let z_{wb} be the proportion of all splitting rules that use the w-th component of $\boldsymbol{x}$ in the b MCMC sample. Then we can count

$$v_w = \frac{1}{B} \sum_{b=1}^{B} z_{wb}$$

1 Introduction

2 Model and Prior

3 Prior Interpretation

4 Posterior Inference

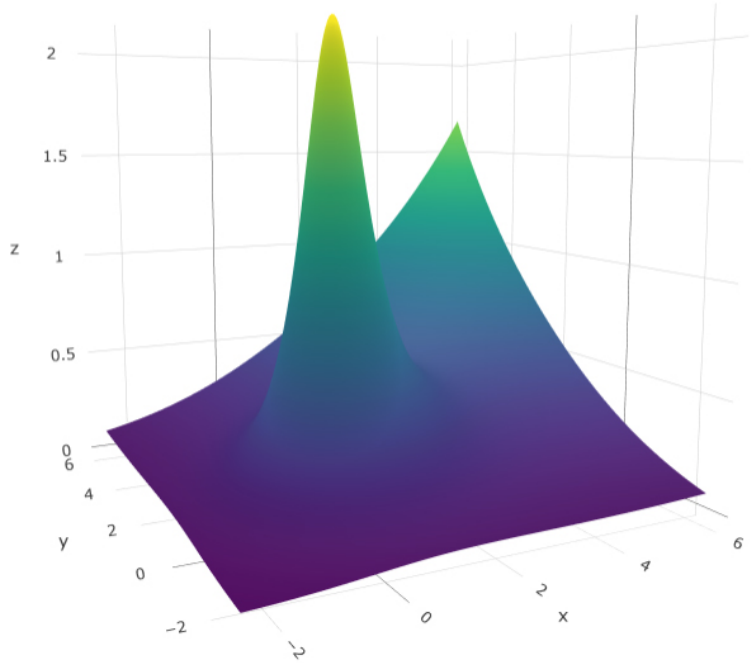
5 Examples of posterior inference

6 References

Test Function

We built a test function to observe the behavior of the BART model in a simple scenario.

$$f(x_1, x_2) = \frac{e^{\frac{4}{10}(x_1+x_2)}}{x_1^2 + x_2^2 + \frac{1}{2}}$$



We used this function to generate data adding a small noise.

Then we tested the model on a real dataset.

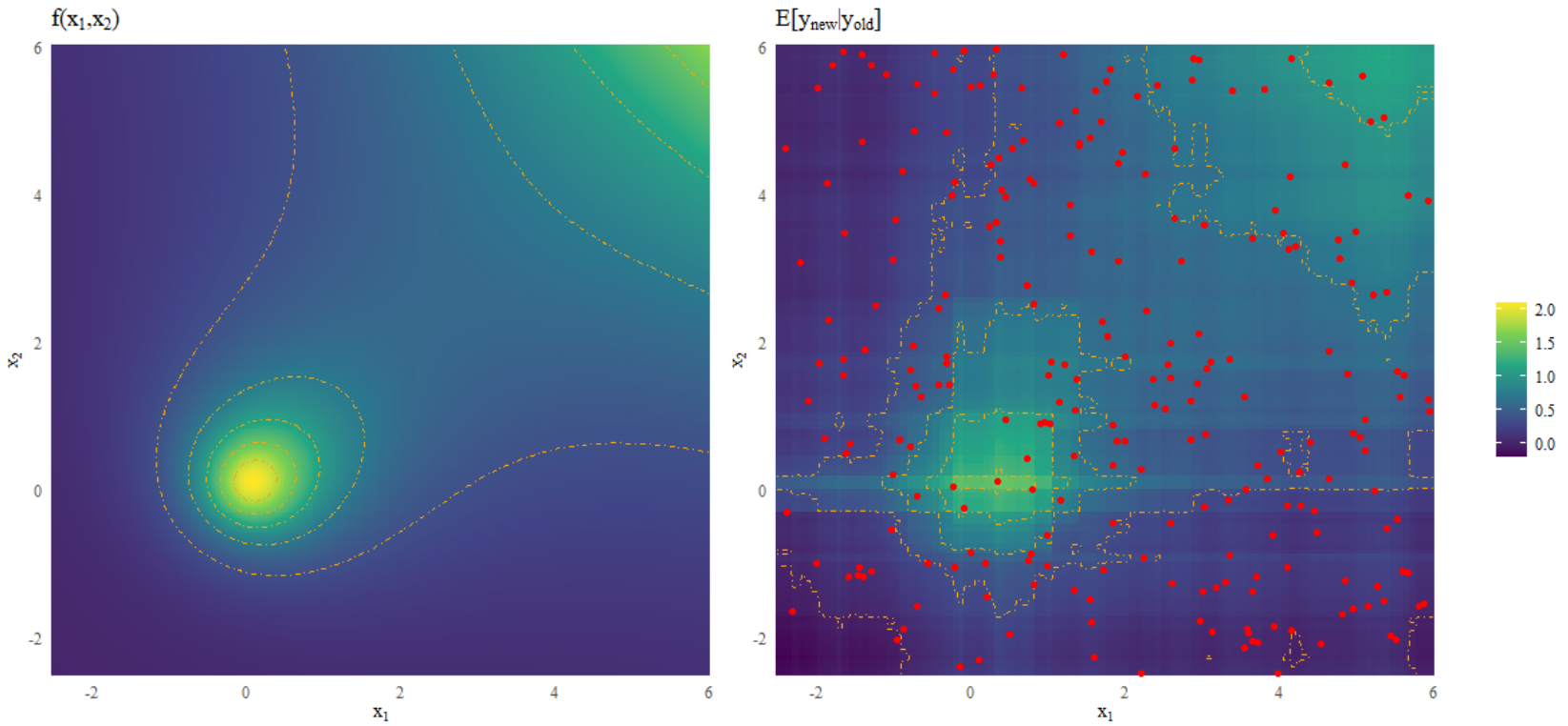
Dimostrative Scenario

We simulated 200 observations following this procedure.

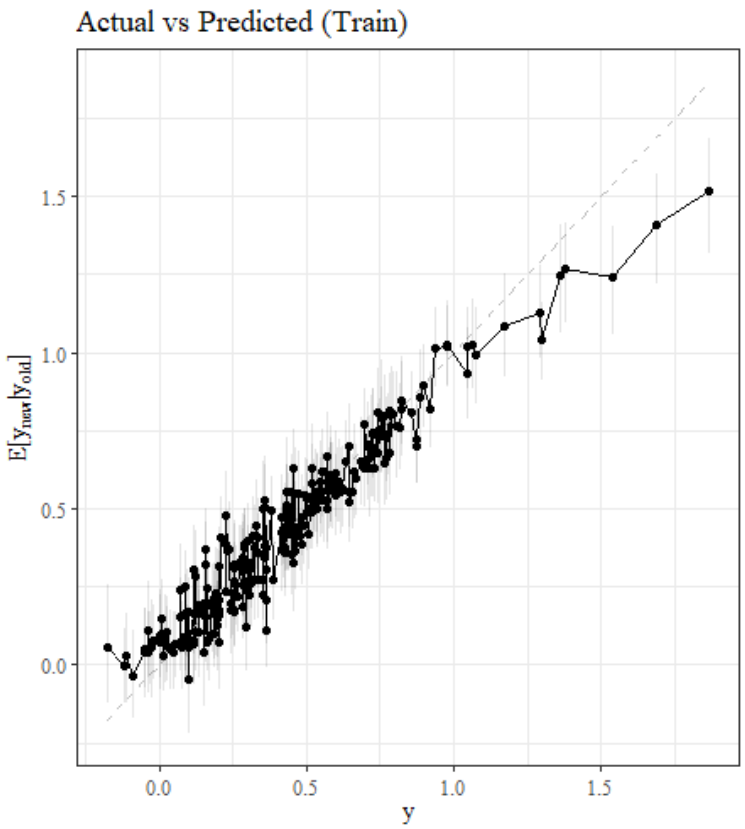
- $x_{1i} \overset{iid}{\sim} \mathcal{U}(-2.5, 6)$
- $x_{2i} \overset{iid}{\sim} \mathcal{U}(-2.5, 6)$
- $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, 0.1^2)$
- $y_i = f(x_{1i}, x_{2i}) + \varepsilon_i$

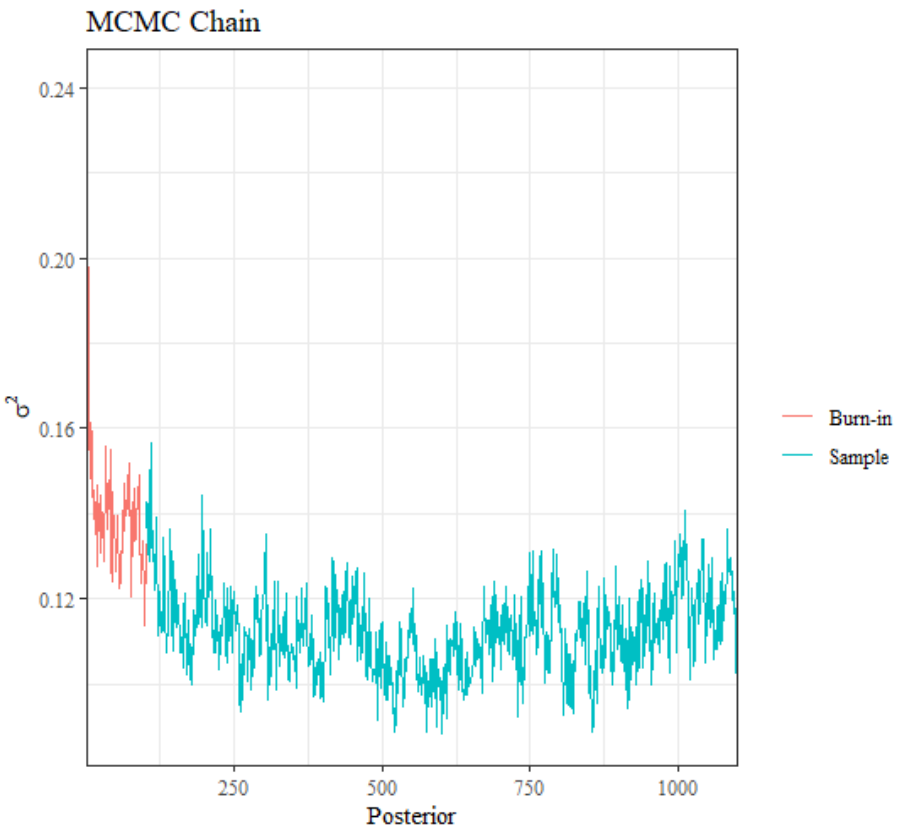
With the data obtained, we estimated the BART with the default algorithm settings and extract the posterior expected value for the function in the whole domain subset.

Posterior Contour Plot



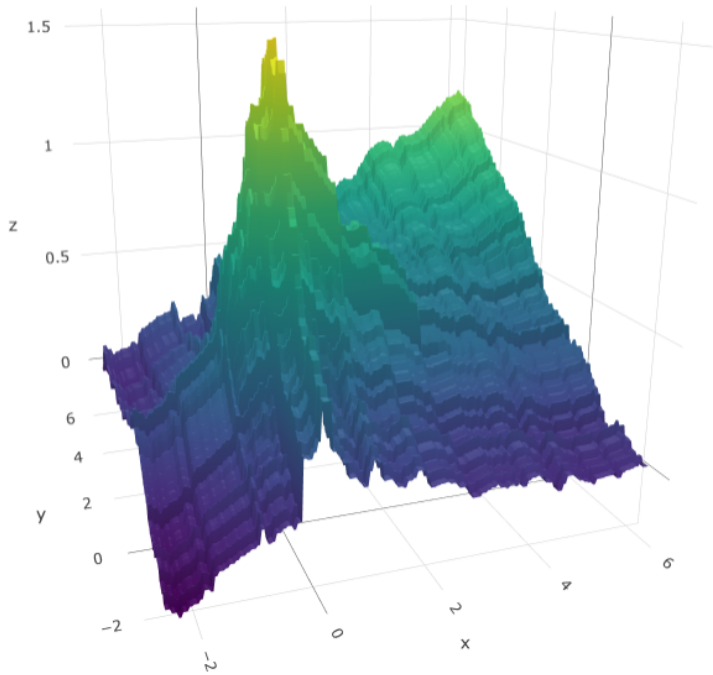
Posterior diagnostics for goodness-of-fit





BART - Posterior of the test function

- BART obtained reliable posterior mean of the true regression function (with default settings).
- The BART predicted function is smoother than a regular decision-tree-based algorithm (averaging over the posterior)
- Predictions are more accurate in areas where the function is flatter.



Cytokines data (1)

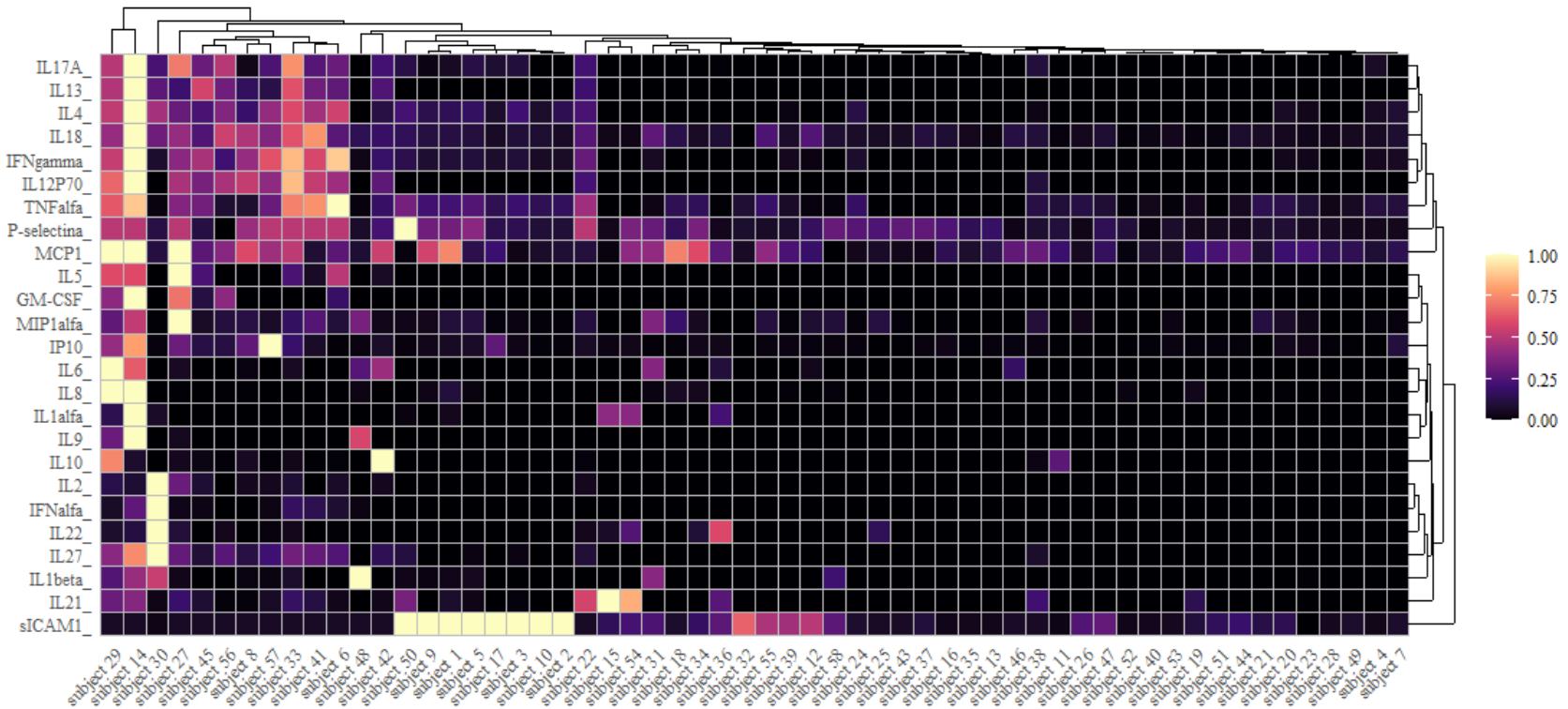
Cytokines are regulatory proteins, produced and secreted by various cells, which are related to immune response, hematopoiesis, inflammation and wound repair and are often studied as biomarkers for certain pathologies.

As data the cytokines framework is often associated with:

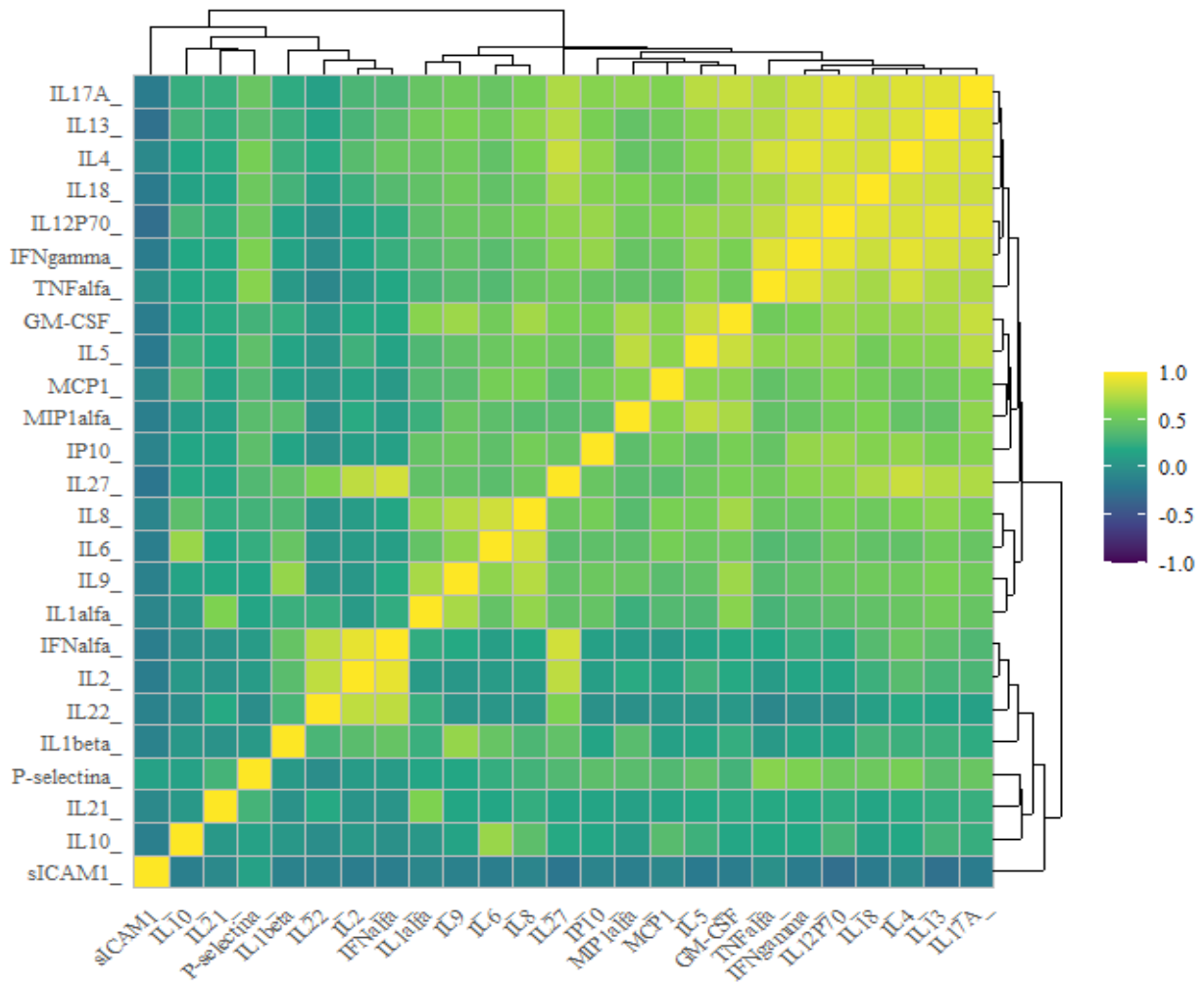
- Small sample size
- High dimensionality
- High correlation
- Strong asymmetry

Our data is composed by 58 subjects, with indicative purposes we have set the goal of predicting $IFN\gamma$ by the rest of the cytokines.

Cytokines data (2)



Cytokines data (3)



Cross Validation for hyper-settings

We performed a k-fold cross validation to choose the values of the hyper-parameters following the original paper, we cycled:

1. 3 pairs for σ^2's prior (ν, q):
   - $(10, 0.75)$ (Conservative)
   - $(3, 0.90)$ (Default)
   - $(3, 0.99)$ (Aggressive)
2. 5 values for inverse scale of μ_{ij}:
   - $k \in \{1, \ldots, 5\}$
3. 4 values or the number of trees:
   - $m \in \{10, 50, 200, 500\}$;

We used 6 folds and we computed MSE and RMSE as accuracy indexes.

Cross Validation Results

Best Model		Train		Test	
		RMSE	MSE	RMSE	MSE
Conservative	$k = 1,\ m = 50$	0.01028	0.00008	0.04256	0.00692

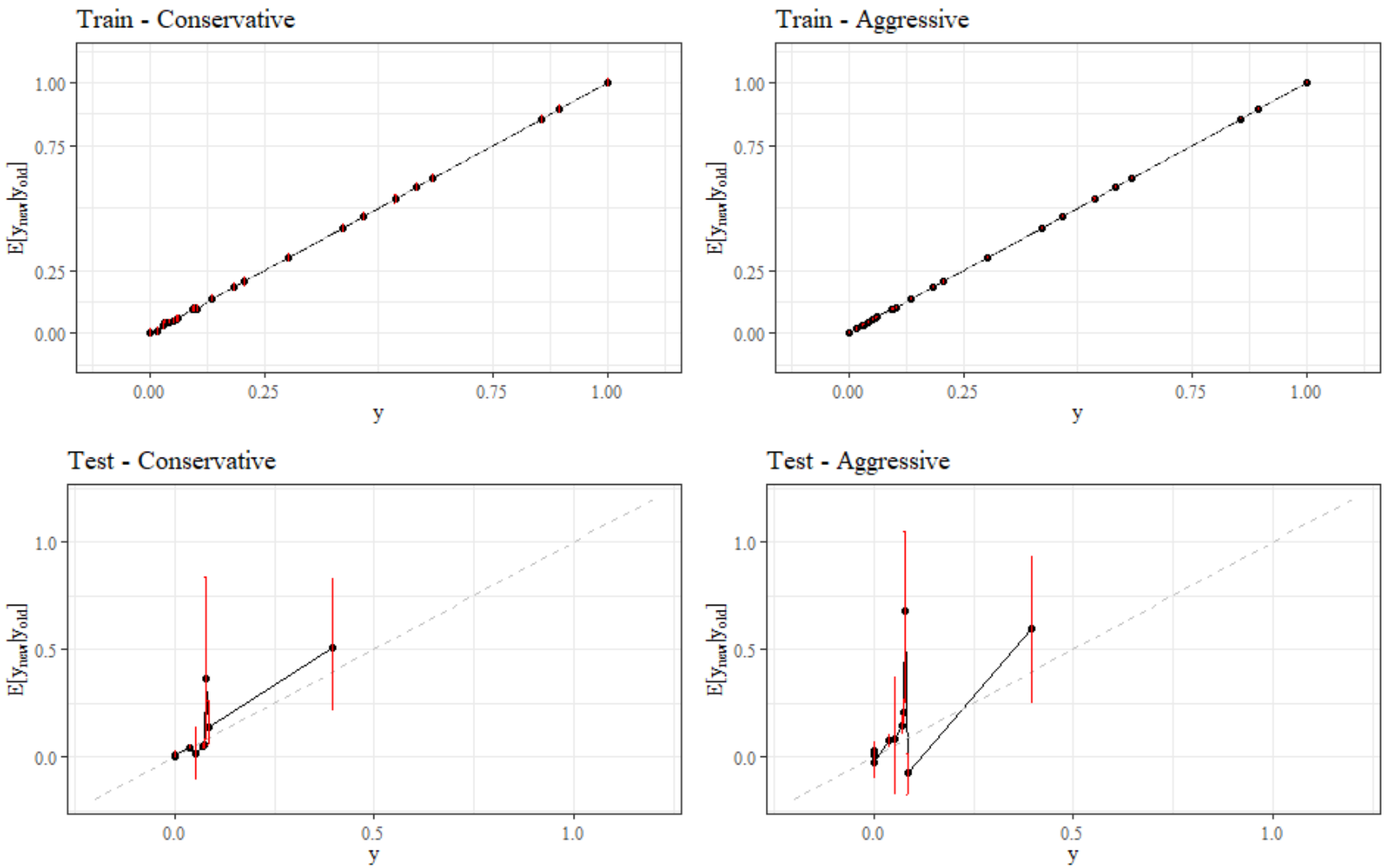
Others Model		Train		Test	
		RMSE	MSE	RMSE	MSE
conservative	$k = 1,\ m = 10$	0.02177	0.00026	0.06985	0.01682
default	$k = 1,\ m = 50$	0.00616	0.00003	0.05248	0.01170
default	$k = 3,\ m = 200$	0.00255	0.00001	0.06113	0.01341
aggressive	$k = 1,\ m = 50$	0.00486	0.00002	0.08119	0.02335
aggressive	$k = 5,\ m = 500$	0.00079	0.00000	0.07480	0.01818

Variable Importance - (Conservative vs Aggressive)

	Stato	*MIP*1α	*IL*27	*IL*1β	*IL*2	*IL*4	*IL*5	*IP*10	*IL*6	*IL*8	*IL*10	*IL*12*P*70	*IL*13
	Best Model - ($m = 50,\ k = 1,\ \nu = 10,\ q = 0.75$)												
Fold.1	6.45	3.24	4.49	3.34	3.98	6.33	6.44	6.10	4.28	4.02	0.55	8.02	3.69
Fold.2	3.85	2.66	3.79	3.63	4.90	6.48	7.51	6.18	4.55	3.54	2.66	5.01	4.89
Fold.3	2.82	3.00	3.18	1.85	3.99	5.76	4.30	5.08	2.48	7.37	2.43	6.47	5.91
Fold.4	3.12	2.22	2.63	8.98	6.89	4.21	4.06	5.49	4.60	5.75	1.53	3.64	3.34
Fold.5	2.91	2.96	5.08	3.49	6.72	3.43	4.51	2.56	3.99	4.54	1.96	8.50	3.66
Fold.6	5.31	3.79	6.51	4.71	7.18	6.20	4.71	3.46	3.45	4.48	3.58	5.65	4.54
	*IL*17*A*	*GMCSF*	*TNF*α	*IFN*α	*MCP*1	*IL*9	*Pselectina*	*IL*1α	*IL*18	*IL*21	*sICAM*1	*IL*22	
Fold.1	4.94	4.21	3.29	5.41	1.96	4.95	4.67	2.56	3.66	0.67	0.51	2.24	
Fold.2	3.97	5.24	4.81	3.36	1.25	3.84	3.13	3.71	3.72	2.78	2.73	1.83	
Fold.3	4.34	3.57	4.57	4.41	1.62	5.19	4.00	5.73	4.00	3.99	2.20	1.72	
Fold.4	3.17	4.73	7.68	8.34	1.26	4.64	0.85	3.27	4.74	0.97	1.72	2.15	
Fold.5	5.09	6.28	3.43	6.29	3.20	4.96	3.43	2.30	4.76	0.91	2.96	2.07	
Fold.6	5.80	2.90	6.75	3.55	1.72	3.44	3.19	2.63	1.96	1.10	1.37	2.02	

	Aggressive Model - ($m = 500,\ k = 6,\ \nu = 3,\ q = 0.99$)												
	Stato	*MIP*1α	*IL*27	*IL*1β	*IL*2	*IL*4	*IL*5	*IP*10	*IL*6	*IL*8	*IL*10	*IL*12*P*70	*IL*13
Fold.1	3.64	4.01	4.04	3.99	3.99	4.04	4.15	4.16	3.93	3.72	4.14	4.17	3.82
Fold.2	3.49	3.90	3.88	3.90	4.01	4.20	4.15	4.19	3.91	3.73	4.07	4.36	4.18
Fold.3	3.46	3.93	3.91	3.94	3.84	4.18	4.04	4.44	3.88	3.95	4.08	4.75	4.07
Fold.4	3.50	3.55	3.76	3.87	5.27	4.74	3.89	4.60	3.86	3.48	4.47	4.47	3.76
Fold.5	3.72	3.86	4.16	4.04	3.94	4.20	4.18	3.86	3.88	3.80	4.14	4.41	4.30
Fold.6	3.50	3.99	4.15	3.92	3.99	4.22	4.20	4.22	3.88	3.85	4.08	4.41	4.02
	*IL*17*A*	*GMCSF*	*TNF*α	*IFN*α	*MCP*1	*IL*9	*Pselectina*	*IL*1α	*IL*18	*IL*21	*sICAM*1	*IL*22	
Fold.1	4.07	3.85	4.57	4.03	3.90	3.98	4.31	3.83	4.07	3.99	3.71	3.90	
Fold.2	4.14	3.96	4.43	3.88	3.94	3.82	4.17	4.05	4.12	4.06	3.55	3.95	
Fold.3	3.98	3.81	4.41	4.00	3.80	3.86	4.26	3.82	4.20	4.12	3.40	3.87	
Fold.4	3.82	3.75	5.27	4.46	3.33	4.00	3.68	3.70	3.32	4.66	2.99	3.78	
Fold.5	3.96	3.85	4.60	3.97	3.89	3.94	4.24	3.80	3.85	4.09	3.41	3.91	
Fold.6	4.20	3.81	4.43	3.84	3.91	3.88	4.19	3.78	4.11	3.94	3.51	3.97	

Predictive differences



BART advantage

BART main characteristics are:

- Soft assumptions and great flexibility.
- Not excessive dependence on the choice of hyper-parameters.
- Excellent predictive ability.

Chipman et al. (2010) [CGM10] compared the performance of BART relative to several machine learning competitors.

- Similar performance between BART and the other popular machine learning algorithms with default prior specification.
- BART performance was noticeably better than the rest when cross-validation was used to set BART hyper-parameters.

BART drawbacks and extension

BART shares typical defects of tree-based models:

- Interpretability;
- High dependence on the sample covariates (local model);

In addition to its own critical aspects:

- Irregularity of the estimated function;
  - Soft-BART (Linero & Yang (2018)[LY18])
- High dimensionality;
  - Step-wise variable selection;
  - Model DART for in-model variable selection (Linero (2018)[Lin18])

See Hill, Linero and Murray (2020) [HLM].

1 Introduction

2 Model and Prior

3 Prior Interpretation

4 Posterior Inference

5 Examples of posterior inference

6 References

[CGM98] Hugh A. Chipman, Edward I. George, and Robert E. McCulloch.
Bayesian cart model search.
Journal of the American Statistical Association, 93(443):935–948, 1998.

[CGM10] Hugh A. Chipman, Edward I. George, and Robert E. McCulloch.
BART: Bayesian additive regression trees.
The Annals of Applied Statistics, 4(1):266 – 298, 2010.

[Fri01] Jerome H. Friedman.
Greedy function approximation: A gradient boosting machine.
The Annals of Statistics, 29(5):1189 – 1232, 2001.

[HLM] Jennifer Hill, Antonio Linero, and Jared Murray.
Bayesian additive regression trees: A review and look forward.
Annual Review of Statistics and Its Application, 7(1).

[HT00] Trevor Hastie and Robert Tibshirani.
Bayesian backfitting.
Statistical Science, 15(3):196–213, 2000.

[Lin18] Antonio R. Linero.
Bayesian regression trees for high-dimensional prediction and variable selection.
Journal of the American Statistical Association, 113(522):626–636, 2018.

[LY18] Antonio R Linero and Yun Yang.
Bayesian regression tree ensembles that adapt to smoothness and sparsity series b statistical methodology.
2018.