# Approximate Bayesian Computation for Model-Free Bayesian Variable Selection

Yi Liu

University of Chicago

*yil@uchicago.edu*

May 28, 2019

Special Thanks: Professor Rockova and Yuexi Wang

# Overview

1. Overview of Problem
   - Definition of Bayesian Variable Selection
   - Theory
   - Implementation

2. ABC Solution
   - Methods
   - Results

# Traditional Bayesian Variable Selection

Assume fixed predictors $\mathbf{x}_i \in \mathbb{R}^p$ and responses

$$Y_i = \mathbf{x}_i'\beta + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0,1), \quad i = 1, \ldots, n. \tag{1}$$

There is an *unknown subset* $\mathcal{S}_0$ of $q_0 < \min\{n, p\}$ active predictors

Bayesian approach to recovering $\mathcal{S}_0$ starts with a **Spike-and-Slab Prior** over all subsets $\mathcal{S}$

$$\gamma_i \equiv I(i \in \mathcal{S}) \sim \mathtt{binomial}(\theta) \quad \text{where } \theta \sim \mathtt{beta}(a, b)$$

$$\Pi(\beta \mid \mathcal{S}) = \prod_{i=1}^{p} [\gamma_i \Pi_1(\beta_i) + (1 - \gamma_i)\Pi_0(\beta_i)]$$

In the linear model (1), $\Pi(\mathcal{S} \mid Y)$ has a closed form so the computation is “easy”.

# Non-parametric Variable Selection with BART

Assume fixed predictors $\mathbf{x}_i \in [0,1]^p$ and responses

$$Y_i = f_0(\mathbf{x}_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0,1), \quad i = 1,\ldots,n. \tag{2}$$

We assume $f_0$ is $\alpha$-Hölder continuous

- $f_0$ depends on an *unknown subset* $\mathcal{S}_0$ of $q_0 < \min\{n,p\}$ predictors
- Given $\mathcal{S}$, we assign a prior on $f_0$ which using forests mappings

$$\textbf{BART}: f_{\mathcal{E},\boldsymbol{B}}(\mathbf{x}_{\mathcal{S}}) = \sum_{t=1}^{T} f_{\mathcal{T}^t,\beta^t}(\mathbf{x}_{\mathcal{S}}) \quad \text{where } \boldsymbol{x}_{\mathcal{S}} = \{x_i : i \in \mathcal{S}\}.$$

Where $\mathcal{E} = \left\{\mathcal{T}^1,\ldots,\mathcal{T}^T\right\}$ are tree partitions and $\boldsymbol{B} = \left[\beta^1,\ldots,\beta^T\right]$ are step coefficients.

Each of $f_{\mathcal{T}^t,\beta^t}(\mathbf{x}_{\mathcal{S}})$ is a tree: we assign Bayesian CART prior.

# We have good theory



## Theorem

*Under the* **spike-and-forest** *prior and some regularity conditions,*

$$\Pi\left[\{\mathcal{S}=\mathcal{S}_0\}\cap\left\{K_{\mathcal{S}_0}\leq\sum_{t=1}^{T}K^t\leq K_n\right\}\mid\mathbf{Y}^{(n)}\right]\to 1$$

*in $P_0^n$ probability as $n\to\infty$ and $p\to\infty$.*
*Where $K=(K^1,\dots,K^T)'\in\mathbb{N}^T$ is the sum of the bottom leaves count of the trees, $K_{\mathcal{S}_0}=\left\lfloor C_K/C_\varepsilon^2 n\varepsilon_{n,\mathcal{S}_0}^2/\log n\right\rfloor$ and $K_n=\left\lceil Cn\varepsilon_{n,s}^2/\log n\right\rceil$ (Liu, Yi and Ročková, Veronika and Wang, Yuexi (2018) )*

Essentially, we get consistency.

# But the marginal likelihood is hard to compute



For the BART case, we notice the following.

1. Marginal Likelihood over all trees $\Pi(Y \mid \mathcal{S})$ is **not available in closed form**.
2. MCMC can be done in principle, but suffers from poor mixing.

# Approximate Bayesian Computation



ABC is method that allows us to sample from the posterior distribution when marginal likelihood is unavailable.

**Traditional ABC Procedure:**

1. We have some prior $\Pi(\theta)$ and data $Y_{\text{data}}$
2. Sample from the prior $\theta \sim \Pi(\theta)$
3. Generate **pseudo-data** from $Y^* \sim \Pi(Y \mid \theta)$
4. Compare $Y^*$ with the original data $Y_{\text{Data}}$
5. If $d(Y_{\text{data}}, Y^*) \leq \epsilon$, accept $\theta$
6. $\theta_1 \ldots \theta_n$ are approximate samples from the posterior $\Pi(\theta | Y_{data})$.

# ABC in action

We consider a ABC Wrapper around BART. Here we introduce a data splitting process at each iteration.

**Data** $(Y_i^{obs}, \boldsymbol{x}_i)$ for $1 \leq i \leq n$

**Output** $\widehat{\Pi}(j \in \mathcal{S}_0 \mid \boldsymbol{Y}^{(n)})$

**Set** $M$: the number of ABC simulations; $s$: the subsample size; $\epsilon$: the tolerance threshold; $m = 0$ the counter

**While** $m \leq M$

- **(a) Split** data $\boldsymbol{Y}^{obs}$ into $\boldsymbol{Y}^{obs}_{\mathcal{I}_m}$ and $\boldsymbol{Y}^{obs}_{\mathcal{I}^c_m}$
- **(b) Pick** a subset $\mathcal{S}_m$ from $\pi(\mathcal{S})$.
- **(c) Sample** $f^m_{\mathcal{E},\boldsymbol{B}}$ from $\pi(f_{\mathcal{E},\boldsymbol{B}} \mid \boldsymbol{Y}^{obs}_{\mathcal{I}_m}, \mathcal{S}_m)$.
- **(d) Generate** pseudo-data $Y_i^\star = f^m_{\mathcal{E},\boldsymbol{B}}(\boldsymbol{x}_i) + \varepsilon_i$ for each $i \notin I_m$.
- **(e) Compute** discrepancy $\epsilon_m = \|\boldsymbol{Y}^\star_{\mathcal{I}^c_m} - \boldsymbol{Y}^{obs}_{\mathcal{I}^c_m}\|_2$.

**Accept** $\mathcal{S}$ if $\epsilon_m < \epsilon$ and set $m = m + 1$

**Reject** $\mathcal{S}$ if $\epsilon_m \geq \epsilon$ and set $m = m + 1$

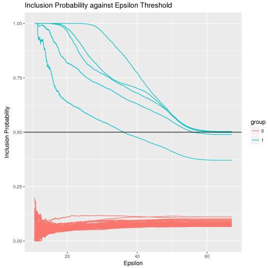
We can obtain *marginal inclusion probabilities* for each of the variable.

# Simulation

## Friedman Data

$$f_0(\boldsymbol{x}_i) = 10\,\sin(\pi\, x_{i1}\, x_{i2}) + 20\,(x_{i3} - 0.5)^2 + 10\, x_{i4} + 5\, x_{i5}, \quad (3)$$

where $x_i \in [0,1]^p$ with $p = 100$ and $n = 500$ are *iid* from a uniform distribution on a unit cube.




We can use the median probability model rule (Barbieri and Berger (2012))

## Current Work



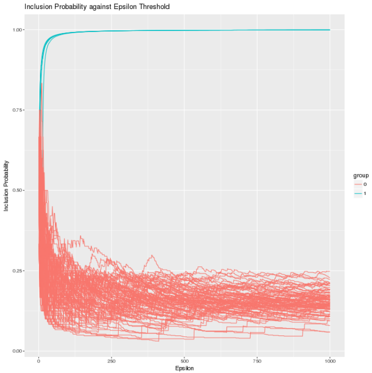
- In ABC, we sample $\mathcal{S} \sim \pi(\mathcal{S})$.
- *Can we sample only the more promising covariates?*
- We regard spike and slab selection as a **bandit problem**.
- Consider a relaxed Spike-and-Slab prior

$$I(i \in \mathcal{S}_0) = \gamma_i \sim \texttt{binomial}(\theta_i) \quad \text{where } \theta_i \sim \texttt{beta}(a_i, b_i)$$

- Each indicator $\gamma_i$ can be regarded as an **arm**.
- Sampling $\mathcal{S}$ is equivalent to playing multiple arms at the same time.
- We can use Thompson Sampling to create a faster algorithm (Liu, Y. and Ročková, V. (coming soon))

# Some Simulation for Thompson Variable Selection

We plot variable importance weights $a_i/(a_i + b_i)$




This is really fast! Separation seen after 200 iterations compared to 10000 iterations needed for ABC.

# Future Work

We are working on regret analysis for Thompson Sampling for our case.

# References


Liu, Y. and Ročková, V.
Variable Selection via Thompson Sampling
(Manuscript in preparation)

Liu, Y., Ročková, V. and Wang, Y. (2018)
ABC Variable Selection with Bayesian Forests
(Submitted)

Barbieri M and Berger J O
The Median Probability Model.
Annals of Statistics, 2012

# The End

# Theory



Assume $f_0 \in \mathcal{H}_p^\alpha \cap \mathcal{C}(\mathcal{S}_0)$ for some $\alpha \in (0,1]$, $\mathcal{S}_0 \subset \{1,\dots,p\}$ such that $|\mathcal{S}_0| = q_0$ and $\|f_0\|_\infty \lesssim B$. Assume $q_0 \log p \leq n^{q_0/(2\alpha+q_0)}$, where $2 \leq q_0 = \mathcal{O}(1)$ as $n \to \infty$. Furthermore, assume that the design is $\mathcal{S}_0$-regular and that

$$\inf_{\{\mathcal{S} \not\supseteq \mathcal{S}_0\} \cup \{E(Z): Z < K_{\mathcal{S}_0}\}} \inf_{f_{\mathcal{E},\boldsymbol{B}} \in \mathcal{F}_{\mathcal{S}}(\boldsymbol{K})} \|f_{\mathcal{E},\boldsymbol{B}} - f_0\|_n > M\varepsilon_{n,\mathcal{S}_0}$$

Under the spike-and-forest prior,

$$\Pi\left[\{\mathcal{S} = \mathcal{S}_0\} \cap \left\{K_{\mathcal{S}_0} \leq \sum_{t=1}^{T} K^t \leq K_n\right\} \mid \boldsymbol{Y}^{(n)}\right] \to 1$$

in $\mathbb{P}_{f_0}^{(n)}$-probabilities as $n \to \infty$
Where $K = \left(K^1, \dots, K^T\right)' \in \mathbb{N}^T$ is the sum of the bottom leaves count of the trees, $K_{\mathcal{S}_0} = \left\lfloor C_K / C_\varepsilon^2 n\varepsilon_{n,\mathcal{S}_0}^2 / \log n\right\rfloor$ and $K_n = \left\lceil Cn\varepsilon_{n,s}^2 / \log n\right\rceil$

# Thompson Variable Selection Algorithm

| **Algorithm 4:** *Thompson Variable Selection* |
|---|
| Initialize $a_i := 1$, $b_i := 1$ |
| **Choosing Step** |
| C1: **For** $i := 1 \cdots B$ **do:**<br>C1.1: Sample $\theta_i \sim Beta(a_i, b_i)$<br>C1.2: Sample $\gamma_i \sim Binomial(\theta_i)$<br>C1.3: If $\gamma_i = 1$ then $i \in \mathcal{S}$ |
| **Reward Step** |
| R1: Randomly split Data into $(Y^{obs}_{I_m}, X^{obs}_{I_m})$ and $(Y^{obs}_{I^c_m}, X^{obs}_{I^c_m})$<br>R2: Train $\mathcal{T}$ from BART $(Y^{obs}_{I_m}, X^{obs}_{I_m} S)$<br>R3: Predict $\widehat{Y^{obs}_{I^c_m}}$ using $\mathcal{T}$<br>R4: $\epsilon = \frac{1}{I^c_m} \sum_{I^c_m} (\widehat{Y^{obs}_{I^c_m}} - Y^{obs}_{I^c_m})^2$<br>R5: if BART pick variable $k$ and $\epsilon < \frac{1}{I^c_m} \sum_{I^c_m} (\overline{Y^{obs}_{I^c_m}} - Y^{obs}_{I^c_m})^2 - \eta_m$, $R_k = 1$, else $R_k = 0$ |
| **Update Step** |
| U1: If $R_k = 1$, $a_k = a_k + 1$, else $b_k = b_k + 1$ |