Approximate Bayesian Computation for Model-Free Bayesian Variable Selection

Yi Liu

University of Chicago

yil@uchicago.edu

June 14, 2019

Special Thanks: Professor Rockova and Yuexi Wang

Overview

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 $\mathcal{S}$

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

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

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

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

We assume f_0 is α-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 $\mathcal{E} = \left\{\mathcal{T}^1, \dots, \mathcal{T}^T\right\}$ are tree partitions and $\boldsymbol{B} = \left[\boldsymbol{\beta}^1, \dots, \boldsymbol{\beta}^T\right]$ 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,*

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

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_{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_{Data}
5. If $d(Y_{\text{data}}, Y^*) \leq \epsilon$, accept θ
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; ϵ: 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 $m = m + 1$

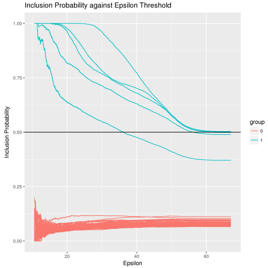
We can obtain *marginal inclusion probabilities* for each of the variable.

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))

References

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

Barbieri M and Berger J O
Optimal predictive model selection.
Annals of Statistics, 2012

The End