LISA for BART

Radu Craiu

Department of Statistical Sciences
University of Toronto

Joint with
Reihaneh Entezari (Bosch Center for Artificial Intelligence)
and
Jeffrey Rosenthal (Univ. of Toronto)

SSC, Calgary
May, 2019

Outline

Motivation for LISA

- Due to MCMC developments, for 30+ years Bayesian statisticians were *computationally liberated* when thinking about a statistical model.
- Large data and/or intractable likelihoods have brought Bayesian computation at a crossroads.
- The Metropolis-Hastings sampler is one of the most used algorithms in MCMC. It operates as follows:
  - Given the current state of the chain θ, draw $\xi \sim q(\xi|\theta)$.
  - Accept ξ with probability $\min\left\{1, \frac{\pi(\xi|\mathbf{y})q(\theta|\xi)}{\pi(\theta|\mathbf{y})q(\xi|\theta)}\right\}$.
  - If ξ is accepted, the next state is ξ, otherwise it is (still) θ.
- Require calculation of the likelihood at each iteration which is expensive when data is massive.

Motivation for LISA

- Possible remedies: divide and conquer, sequential processing, pseudomarginal, precomputing, etc
- D &C: Divide data into batches, $\mathbf{y}^{(1)} \cup \dots \mathbf{y}^{(K)}$, distribute the sampling from the K sub-posteriors

$$\pi_j(\theta) \propto [L_k(\theta|\mathbf{y}^{(j)})]^a [p_j(\theta)]^b$$

  among K processing units
- Depending on a, b values, design recombination strategies for the π_j-samples to recover the characteristics of the full posterior distribution.
- Aim: minimize the loss of information compared to full posterior analysis.

Example: Consensus Monte Carlo (Scott et al., 2016)

- Consider the full posterior $\pi(\boldsymbol{\theta}|\mathbf{y}) \propto p(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta})$ where $f(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{N} f(y_i|\boldsymbol{\theta})$
- The batch-specific posterior is defined as

$$\pi_{h,CMC} \propto [p(\boldsymbol{\theta})]^{\frac{1}{K}} f(\mathbf{y}^{(h)}|\boldsymbol{\theta})$$

- MCMC samples are obtained independently from each π_h and combined using a weighted average since

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \prod_{h=1}^{K} \pi_{h,CMC}(\boldsymbol{\theta}|\mathbf{y}^{(h)}).$$

- Theory works if the posteriors are Gaussian.
- Motivation: CMC does not perform well for the Bayesian Additive Regression Trees (BART) model.

LISA: Initial targets

- Improve the use of BART for big data.
- Bring the batch-specific likelihood "closer" to the whole-data likelihood.
- Define $\pi_{h,LISA} \propto p(\boldsymbol{\theta})[f(\mathbf{y}^{(h)}|\boldsymbol{\theta})]^K$.
- Intrinsic BF (Berger& Perrichi, JASA '96), Data cloning (Lele & al., JASA '10), Bayesian robustness (Holmes & Walker, Bmka, '17), etc.

LISA: Anchoring Intuitions

- $\hat{\boldsymbol{\theta}}^{(j)}_{n,L}$ and $\hat{\boldsymbol{\theta}}^{(j)}_{n,C}$ denote the *j-th* sub-posterior modes in LISA and CMC
- $\hat{I}^{(j)}_{n,L}$ and $\hat{I}^{(j)}_{n,C}$ denote the negative second derivative for the *j-th* log sub-posterior for LISA and CMC

A1: There exist $\boldsymbol{\theta}_L, \boldsymbol{\theta}_C$ such that if we define $\epsilon^{(j)}_{n,L} = |\hat{\boldsymbol{\theta}}^{(j)}_{n,L} - \boldsymbol{\theta}_L|$ and $\epsilon^{(j)}_{n,C} = |\hat{\boldsymbol{\theta}}^{(j)}_{n,C} - \boldsymbol{\theta}_C|$, then $\max\limits_{1 \leq j \leq K} \epsilon^{(j)}_{n,L} \rightarrow 0$ and $\max\limits_{1 \leq j \leq K} \epsilon^{(j)}_{n,C} \rightarrow 0$ w.p. 1 as $n \rightarrow \infty$.

A2: $|\hat{I}^{(i)}_{n,L} - \hat{I}^{(j)}_{n,L}| \longrightarrow 0$ and $|\hat{I}^{(i)}_{n,C} - \hat{I}^{(j)}_{n,C}| \rightarrow 0$ w.p. 1 $\forall\ i \neq j$ as $n \rightarrow \infty$.

A3: π_{Full}, $\pi_{h,LISA}$, and $\pi_{h,CMC}$ are unimodal distributions that have continuous derivatives of order 2.

LISA: Anchoring Intuitions

- Assume:
  - **A1** through **A3** hold
  - $\hat{I}_N^{1/2}(\boldsymbol{\theta}_{Full} - \hat{\boldsymbol{\theta}}_N) \xrightarrow{D} N(0, I)$ as $N \to \infty$ (K is fixed), where $\boldsymbol{\theta}_{Full} \sim \pi_{Full}(\boldsymbol{\theta}|\vec{Y}_N)$ then
- Then if K is fixed and $N \to \infty$

$$\hat{I}_N^{1/2}(\boldsymbol{\theta}_{j,L} - \hat{\boldsymbol{\theta}}_N) \xrightarrow{D} N(0, I)$$
$$\hat{I}_N^{1/2}(\boldsymbol{\theta}_{j,C} - \hat{\boldsymbol{\theta}}_N) \xrightarrow{D} N(0, KI), \;\; \forall \;\; j \in \{1, \ldots, K\}.$$

- Asymptotics suggest that draws from each batch can be used without weighting

LISA: Bernoulli Example

- Consider $\mathbf{y}_N = \{y_1, ..., y_N\}$ to be N i.i.d. Bernoulli (θ)random variables
- Prior $p(\theta) = \text{Beta}(\alpha, \beta)$ for parameter θ
- Set $S = \sum_{i=1}^{N} y_i$ and $S_j = \#$ of 1's in j-th batch. Then:

FULL $\pi_{Full}(\theta|\mathbf{y}_N)$ is $\text{Beta}(S + \alpha, N - S + \beta)$

CMC: $\pi_{j,CMC}(\theta|\mathbf{y}^{(j)})$ is $\text{Beta}\left(S_j + \frac{\alpha-1}{K} + 1, \frac{N}{K} - S_j + \frac{\beta-1}{K} + 1\right)$

LISA: $\pi_{j,LISA}(\theta|\mathbf{y}^{(j)})$ is $\text{Beta}(S_j K + \alpha, (n - S_j)K + \beta)$

- If $S_j = S/K$ and $n = N/K$ then $\pi_{j,LISA}(\theta|\mathbf{y}^{(j)}) = \pi_{Full}(\theta|\mathbf{y}_N)$
- No weighting needed!

BART - Chipman et al. (AOAS, 2010)

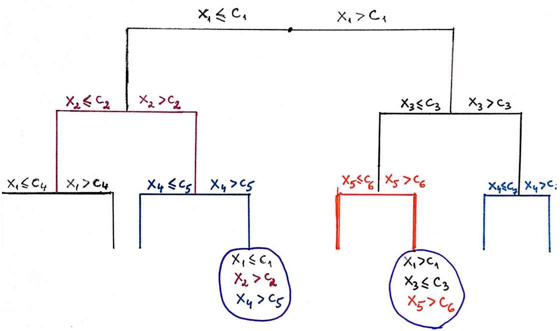
- Flexible Bayesian approach for nonparametric regression
- Regression setting $Y = f(X) + \epsilon$ where predictor $f(X)$ is the sum of (many) regression tree models

$$f(X) = g_1(X, T_1, M_1) + \ldots + g_m(X, T_m, M_m),$$

with $\epsilon \sim N(0, \sigma^2)$.

- Focus is on prediction.

BART - One tree



- A tree T with b terminal nodes has parameters $M = (\mu_1, \ldots, \mu_b)$.
- The splitting rules $\rightarrow$ partition of the covariate space
- BART fits an intercept for data in each marginal node resulting in a piecewise constant approximation of f.

MCMC & BART

The priors are:

- $p(\sigma) = \text{Inv-Gamma}(\frac{\nu}{2}, \frac{\nu\lambda}{2})$,
- $p(\mu_j|\mu_\mu, \sigma_\mu) = N(\mu_\mu, \sigma_\mu)$
- $p(T_j)$, is characterised by three aspects:
  - The probability that a node at depth $d = 0, 1, ...$ is non-terminal.
  - The distribution of the splitting variable at each interior node.
  - The distribution of the splitting rule in each interior node.

The BART Posterior

$$\pi(\theta) = \pi(\theta|Y,X) \propto \underbrace{\left\{(\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{m} g(x_i; M_j, T_j))^2}\right\}}_{\text{Likelihood}} \times$$
$$\underbrace{\left\{\underbrace{(\sigma^2)^{-\frac{\nu}{2}-1} e^{-\frac{\nu\lambda}{2\sigma^2}}}_{\text{Prior of } \sigma^2}\left[\prod_{j=1}^{m} \sigma_\mu^{-b_j}(2\pi)^{-\frac{b_j}{2}} e^{-\frac{1}{2\sigma_\mu^2}\sum_{k=1}^{b_j}(\mu_{kj}-\mu_\mu)^2} p(T_j)\right]\right\}}_{\text{Prior}}. \tag{1}$$

MCMC & BART

- The MCMC sampler has the following steps

Step S Sample σ given $(T_1, M_1), \ldots, (T_m, M_m)$ using
$\sigma^2 \mid (T_1, M_1), ..., (T_m, M_m), Y, X \ \propto$ Inv-Gamma(ρ, γ) where
$\rho = \frac{\nu+n}{2}$ and $\gamma = \frac{1}{2}\ [\ \sum_{i=1}^{n} (y_i - \sum_{j=1}^{m} g(x_i; M_j, T_j))^2 + \lambda\nu\ ]$.

Step R For $1 \leq j \leq m$ sample (T_j, M_j) given $T_{-j}, M_{-j}, X, \mathbf{y}, \sigma$.

Step R

- Sample $T_j|R_j, \sigma$, where $R_j = y - \sum_{k\neq j} g(x; M_k, T_k)$ use Metropolis-Hastings to **GROW**, **PRUNE** and **CHANGE**
- Assume we propose T_*, then the acceptance ratio will be:

$$r = \underbrace{\frac{P(T_* \to T)}{P(T \to T_*)}}_{\text{transition ratio}} \times \underbrace{\frac{P(R \mid T_*, \sigma^2)}{P(R \mid T, \sigma^2)}}_{\text{likelihood ratio}} \times \underbrace{\frac{P(T_*)}{P(T)}}_{\text{tree structure ratio}}$$

- Sample $M_j|T_j, R_j, \sigma$ using
$\mu_{ij} \mid T_j, R_j, \sigma \sim \mathcal{N}\left(\frac{\frac{\sigma^2}{\sigma_\mu^2}\mu_\mu + n_i \bar{R}_{j(i)}}{\frac{\sigma^2}{\sigma_\mu^2} + n_i}, \frac{\sigma^2}{\frac{\sigma^2}{\sigma_\mu^2} + n_i}\right)$ where $\bar{R}_{j(i)}$ denotes the average residual (computed without tree j) at terminal node i with total number of observations n_i.

LISA & BART

- $f(x) = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5$ with $N = 20,000$, $K = 30$, $\sigma = 3$.
- Compare LISA & *Single Machine*:
  - Trees tend to be larger $\rightarrow$ Fewer data in each terminal node
  - σ is severely underestimated
  - Lower acceptance rates for tree moves.

Method	Tree Nodes	Avg $\hat{\sigma}^2$	95% CI for σ^2
LISA (*unif wgh*)	55	0.001	[0.0009 , 0.0011]
SingleMachine	7	9.04	[8.85 , 9.21]

Intermezzo: LISA & Normal Regression Example

- Consider $Y = X\beta + \epsilon$, $\beta \in \mathbf{R}^p$, $X \in \mathbf{R}^{N\times p}$ and $Y, \epsilon \in \mathbf{R}^N$ with $\epsilon \sim N(0, \sigma^2 \mathbf{I}_N)$.
- Consider Jeffrey's prior $p(\beta, \sigma^2) \propto 1/\sigma^2$

Intermezzo: LISA & Normal Regression Example

FULL

$$\sigma^2 \sim \text{Inv-Gamma}\left(\frac{N-p}{2}, \frac{s^2(N-p)}{2}\right)$$
$$\beta|\sigma^2 \sim N(\hat{\beta}, \sigma^2(X^TX)^{-1})$$

with $\hat{\beta} = (X^TX)^{-1}X^TY$ and $s^2 = \frac{(Y-X\hat{\beta})^T(Y-X\hat{\beta})}{N-p}$.

$E[\beta_{Full}|Y,X] = (X^TX)^{-1}X^TY$ and
$\mathrm{Var}(\beta_{Full}|Y,X) = (X^TX)^{-1}\frac{(N-p)/2}{(N-p)/2-1}s^2 = (X^TX)^{-1}s^2 + O(N^{-1})$.

LISA

$$\sigma^2 \sim \text{Inv-Gamma}\left(\frac{N-p}{2}, \frac{Ks_j^2(n-p)}{2}\right)$$
$$\beta|\sigma^2 \sim N\left(\hat{\beta}_j, \frac{\sigma^2}{K}(X^{(j)\ T}X^{(j)})^{-1}\right),$$

$E[\beta|Y^{(j)},X^{(j)}] = \hat{\beta}_j$ and
$\mathrm{Var}(\beta|Y^{(j)},X^{(j)}) = (X^{(j)\ T}X^{(j)})^{-1}\frac{s_j^2(n-p)/2}{(N-p)/2-1} = (X^{(j)\ T}X^{(j)})^{-1}\frac{s_j^2(n-p)}{(N-p)} + O(N^{-1})$.

LISA: Normal Regression Example

- In order to combine the sub-posterior samples we propose using the weighted average

$$\beta_{LISA} = (\sum_{j=1}^{K} W_j)^{-1} \sum_{j=1}^{K} W_j \beta_j,$$

where $\beta_j \sim \pi_j(\beta | Y^{(j)}, X^{(j)})$ and $W_j = \frac{X^{(j)\ T} X^{(j)}}{\sigma^2}$

Then $E[\beta_{LISA} | Y, X] = \hat{\beta} = (X^T X)^{-1} X^T Y$, and

$$\begin{aligned} \mathrm{Var}(\beta_{LISA} | Y, X) &= (X^T X)^{-1} \frac{n-p}{N-p} \left[\sum_{j=1}^{K} s_j^2 (X^{(j)\ T} X^{(j)}) \right] (X^T X)^{-1} \\ &\approx (X^T X)^{-1} \frac{n-p}{N-p} s^2 \end{aligned}$$

- Modification needed!

LISA: Normal Regression Example

LISA

$$\begin{aligned}
\sigma^2 &\sim \text{Inv-Gamma}\left(\frac{N-p}{2}, \frac{Ks_j^2(n-p)}{2}\right) \\
\beta|\sigma^2 &\sim N\left(\hat{\beta}_j, \frac{\sigma^2}{K}(X^{(j)\ T}X^{(j)})^{-1}\right) \\
w_j &\propto 1
\end{aligned}$$

Mod LISA

$$\begin{aligned}
\sigma^2 &\sim \text{Inv-Gamma}\left(\frac{N-p}{2}, \frac{Ks_j^2(n-p)}{2}\right) \\
\tilde{\sigma} &= \sqrt{K}\sigma \\
\beta|\tilde{\sigma}^2 &\sim N\left(\hat{\beta}_j, \frac{\tilde{\sigma}^2}{K}(X^{(j)\ T}X^{(j)})^{-1}\right) \\
w_j &\propto (X^{(j)\ T}X^{(j)})[s_j^2]^{-1} = \widehat{Var}(\hat{\beta}_j)^{-1}
\end{aligned}$$

Modified LISA & BART

- Introduce an intermediate step between **Step S** and **Step R** in the MCMC algorithm for LISA.
- Adjust the σ draws, i.e. set $\tilde{\sigma}_j = \sqrt{K}\sigma_j$
- Samples from batch j have weights $\propto \hat{\sigma}_j^{-2}$

Modified LISA & BART

Table: Comparing Train & Test RMSE, tree sizes, and average post burn-in $\hat{\sigma}^2$ with 95% CI in each method for $K = 30$ to SingleMachine BART.

Method	TrainRMSE	TestRMSE	Tree Nodes	Avg $\hat{\sigma}^2$	95% CI for σ^2
CMC	2.73	2.94	602	1.91	[1.45 , 2.88]
LISA	1.18	1.19	55	0.001	[0.0009 , 0.0011]
modLISA	0.57	0.59	7	7.97	[7.87 , 8.08]
SingleMachine	0.55	0.56	7	9.04	[8.85 , 9.21]

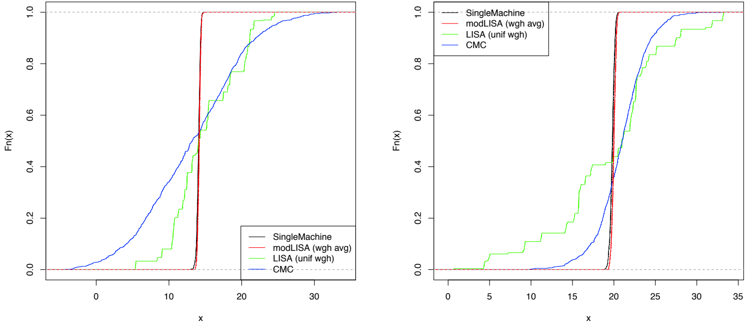
Modified LISA & BART

Average acceptance rates of tree proposal moves.

Method	GROW	PRUNE	CHANGE
CMC	21%	0.03%	34%
LISA	1.8%	0.5%	1.6%
modLISA	20%	26%	19%
SingleMachine	9%	10%	6%

Modified LISA & BART

Empirical CDF for $\hat{f}(x)$



Left: Test $f(x^*) = 14.4$; Right: Training $f(x) = 19.8$

Modified LISA & BART

Method	Avg Time per iteration (Secs)	Speed-up
CMC	11.99	31%
LISA	5.04	71%
modLISA	1.81	90%
SingleMachine	17.28	—

Running times for CMC, LISA, modLISA and SingleMachine when $K = 30$.

Interval Coverage

- Consider two types of intervals:
  - Let $\hat{J}_{y_i}$ is the $1-\alpha$ Prediction Interval (PI) for y_i
  - Coverage for $\hat{J}_{y_i}$ is given by the average over train/test data

$$\frac{\#\{\tilde{y}_j \in \hat{J}_{y_i} :\ \tilde{y}_j \overset{iid}{\sim} N(f(x_i), \sigma^2),\ 1 \leq j \leq 1000\}}{1000}.$$

  - Credible Interval (CI) Coverage

$$\frac{\#\{f(x_i) \in \hat{I}_{f(x_i)} :\ 1 \leq i \leq N\}}{N}$$

  where $\hat{I}_f(x_i)$ is the CI for $f(x_i)$.
- Both are considered for Test and Train Data.

Interval Coverage

Method	Predictive Int Train	Predictive Int Test	Credible Int Train	Credible Int Test
CMC	45.71 %	47.83 %	81.95 %	99.99 %
LISA	1.54 %	1.54 %	100 %	100 %
modLISA	92.93 %	92.91 %	60.88 %	58.45 %
SingleMachine	94.67 %	94.65 %	71.58 %	71.54 %

- PI's are influenced by $\hat{\sigma}^2$ and $\widehat{\mathrm{Var}}(\hat{f}(x))$.
- CI's are influenced by $\widehat{\mathrm{Var}}(\hat{f}(x))$.

Alternative Model

$$f(x) = \mathbf{1}_{[0,0.2)}(x_1)+2\cdot\mathbf{1}_{[0.2,0.4)}(x_1)+3\cdot\mathbf{1}_{[0.4,0.6)}(x_1)+4\cdot\mathbf{1}_{[0.6,0.8)}(x_1)+5\cdot\mathbf{1}_{[0.8,1)}(x_1)$$

Method	Test RMSE	Test Credible
CMC	1.35	100 %
LISA (unif wgh)	0.94	100 %
modLISA (wgh avg)	0.24	90.16 %
SingleMachine	0.15	98.76 %

Housing Data

- Data consist of variables related to people and housing units.
- Predict a person's total income based on variables such as sex, age, education (at least a BA degree), class of worker, living state, and citizenship status
- $N = 437,297$, $K = 100$, Monte Carlo sample size is $M = 1500$, time > 1 day for *Single Machine*.

Method	TestRMSE	Avg $\hat{\sigma}^2$	Tree Nodes	Speed-up
modLISA (*wgh avg*)	0.71	0.488	7	90%
SingleMachine	0.70	0.485	23	–

Conclusions

- ModLISA combined with better mixing chains for BART (Pratola, BA 2016) exhibits similar gains.
- Despite attractive asymptotic properties, fine-tuning of LISA-like samplers is still needed.
- Theoretical validation may rely on approximate & noisy MCMC and perturbation errors (e.g., Mithrophanov 2005, Pillai and Smith 2015, Johndrow et al. 2017, Negrea and Rosenthal 2017).
- Important questions about batch-sample design $\rightarrow$ Extension to non-iid case is an important future direction.
- Promising alternatives include the use of core-sets or non-reversible Markov chains.