Asymptotically Exact, Embarrassingly Parallel MCMC

Willie Neiswanger, Chong Wang, Eric P. Xing

MCMC Journal Club
May 23 2018

Parallel MCMC

- Parallel chains: independent chains on full-data (slow burnin)
- Parallelize single chains: compute on a subset of data and exchange information at each iteration (communication overhead)

Proposed algorithm

- Each machine has access to a portion of data
- Each machine runs independent chains without communication (embarrassingly parallel)
- Each machine can use any type of MCMC to generate samples
- Combine samples to yield asymptotically exact full-data posterior samples

Embarrassingly parallel MCMC

- Partition i.i.d data points x^N into M subsets $\{x^{n_1}, \ldots, x^{n_M}\}$
- For machine $m = 1, \ldots, M$, sample from *subposterior* $p_m(\theta)$

$$p_m(\theta) \propto p(\theta)^{\frac{1}{M}} p(x^{n_m} \mid \theta)$$

- Combine samples to form samples from an estimate of *subposterior density product* $p_1 \ldots p_M$ where

$$p_1 \ldots p_M(\theta) \propto p(\theta \mid x^N)$$

Combine subposterior samples

- Goal: get an estimate of subposterior density product $p_1 \dots p_M(\theta)$, which is proportion to the full-data posterior
- Presented three estimators: parametric, nonparametric, and semiparametric

Parametric estimator

- Bayesian CLT: $p(\theta \mid x^N) \approx \mathcal{N}_d(\theta_0, F_N^{-1})$ as $N \to \infty$
- Estimate each subposterior density with

$$\widehat{p_m} = \mathcal{N}_d(\theta \mid \widehat{\mu}_m, \widehat{\Sigma}_m)$$

- $p_1 \widehat{\ldots p}_M(\theta) = \widehat{p}_1 \ldots \widehat{p}_M(\theta) \propto \mathcal{N}_d(\theta \mid \widehat{\mu}_M, \widehat{\Sigma}_M)$
- fast but asymptotically biased when posterior is non-Gaussian

Nonparametric estimator

- Given T samples $\{\theta_{t_m}^m\}_{t_m=1}^T$, Gaussian KDE of subposterior

$$\widehat{p}_m(\theta) = \frac{1}{T}\sum_{t_m=1}^{T} \mathcal{N}_d(\theta \mid \theta_{t_m}^m, h^2 I_d)$$

- $\widehat{p_1 \dots p_M}(\theta) = \widehat{p}_1 \dots \widehat{p}_M(\theta) \propto \sum_{t_1=1}^{T} \cdots \sum_{t_M=1}^{T} w_{t\cdot}\, \mathcal{N}_d(\theta \mid \bar{\theta}_{t\cdot}, \frac{h^2}{M} I_d)$
- Generate samples using IMG sampler from the mixture
- Asymptotically exact, but slow to converge when d is large

Semiparametric estimator

- product of a parametric estimator $\widehat{f}_m(\theta)$ with a nonparametric estimator $\widehat{r}(\theta)$ of correction function $r(\theta) = \frac{p_m(\theta)}{\widehat{f}_m(\theta)}$

$$\widehat{p}_m(\theta) = \widehat{f}_m(\theta)\widehat{r}(\theta) = \frac{1}{T}\sum_{t_m=1}^{T} \frac{\mathcal{N}_d(\theta \mid \theta^m_{t_m}, h^2 I_d)\mathcal{N}_d(\theta \mid \widehat{\mu}_m, \widehat{\Sigma}_m)}{\mathcal{N}_d(\theta^m_{t_m} \mid \widehat{\mu}_m, \widehat{\Sigma}_m)}$$

- $\widehat{p_1 \dots p_M}(\theta) = \widehat{p}_1 \dots \widehat{p}_M(\theta) \propto \sum_{t_1=1}^{T} \cdots \sum_{t_M=1}^{T} W_{t\cdot}\mathcal{N}_d(\theta \mid \mu_{t\cdot}, \Sigma_{t\cdot})$
- Generate samples using IMG similarly as in nonparametric

Method complexity and density product estimate convergence

- parallel MCMC chains: $O(dTM)$ + combination phase: $O(dTM)$
- MCMC phase and combination phase can also be performed in parallel
- Showed mean square consistency of nonparametric and semiparametric estimator:

$$\sup_{p_1,\ldots,p_M \in \mathcal{P}(\beta,L)} \mathbb{E}\left[\int (\widehat{p_1 \ldots p_M}(\theta) - p_1 \ldots p_M(\theta))^2 d\theta\right] \leq \frac{c}{T^{2\beta/(2\beta+d)}}$$

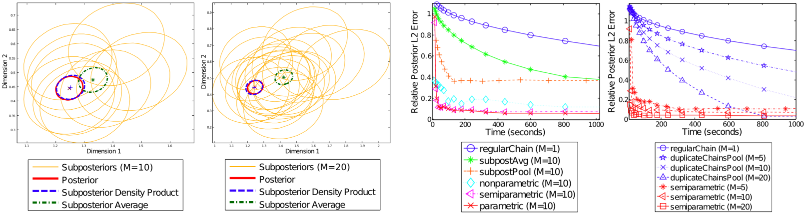
for some $c > 0$ and $0 < h \leq 1$

Method scope

- Posterior distributions over finite-dimensional real spaces
- Not yet extended to infinite dimensional models (nonparametric Bayesian models), distribution over the simplex (LDA)

Empirical study

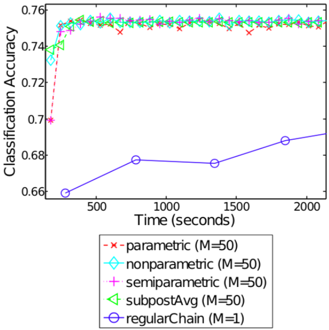
logistic regression: simulated data



$$
\begin{gathered}
N = 5 \times 10^4, \quad d = 50, \quad M = 10 \\
X_{ij} \sim \mathcal{N}(0,1), \quad \beta_j \sim \mathcal{N}(0,1) \\
Y_i \sim \text{Bernoulli}(\text{logit}^{-1}(X_i \boldsymbol{\beta})) \\
d_2(p, \hat{p}) = ||p - \hat{p}||^2 = \left(\int (p(\theta) - \hat{p}(\theta))^2 d\theta\right)^{1/2}
\end{gathered}
$$

Empirical study

logistic regression: real world data *covtype*



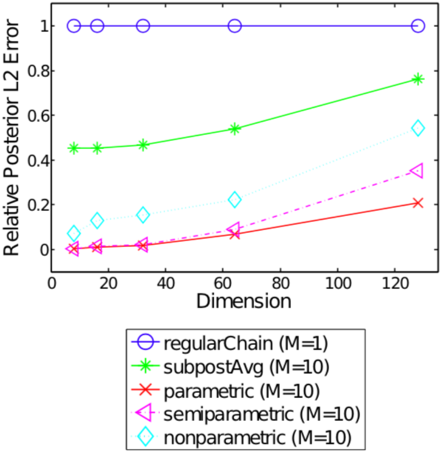
$$N = 581012, \quad d = 54, \quad M = 50$$

$$P(y \mid x, y^N, x^N) \approx \frac{1}{S} \sum_{s=1}^{S} P(y \mid x, \beta_s)$$

$$P(y \mid x, \beta_s) \sim \text{Bernoulli}(\text{logit}^{-1}(x^T \beta_s))$$

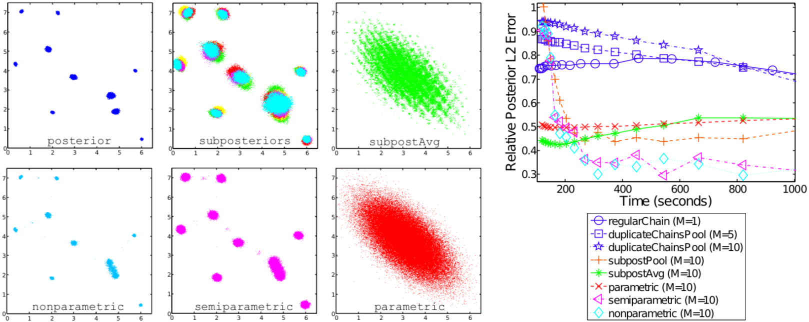
Empirical study

Scalability with dimension



Empirical study

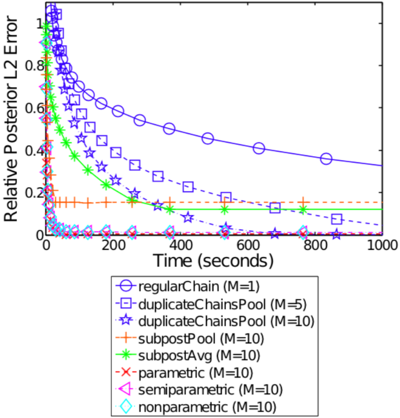
Guassian mixture models



10 2-d Guassians, $\quad N = 5 \times 10^4, \quad M = 10$

Empirical study

HIerarchical Poisson-gamma models



$$N = 5 \times 10^4, \quad M = 10$$

$$a \sim \text{Exponential}(\lambda)$$

$$b \sim \text{Gamma}(\alpha, \beta)$$

$$q_i \sim \text{Gamma}(a, b)$$

$$x_i \sim \text{Poisson}(q_i t_i), \quad i = 1, \ldots, N$$

Summary

- faster burnin by only operating on a subset of data (cf. parallel chains)
- faster sampling since no communication is involved (cf. parallelized single chain)
- ideal for MapReduce settings
- only works when posterior samples are real and unconstrained

IMG procedure

Algorithm 1 Asymptotically Exact Sampling via Non-parametric Density Product Estimation

Input: Subposterior samples: $\{\theta^1_{t_1}\}^T_{t_1=1} \sim p_1(\theta)$, ..., $\{\theta^M_{t_M}\}^T_{t_M=1} \sim p_M(\theta)$

Output: Posterior samples (asymptotically, as $T \to \infty$): $\{\theta_i\}^T_{i=1} \sim p_1 \cdots p_M(\theta) \propto p(\theta | x^N)$

Draw $t\cdot = \{t_1, \ldots, t_M\} \stackrel{\text{iid}}{\sim} \text{Unif}(\{1, \ldots, T\})$
for $i = 1$ **to** T **do**
    Set $h \leftarrow i^{-1/(4+d)}$
    for $m = 1$ **to** M **do**
        Set $c\cdot \leftarrow t\cdot$
        Draw $c_m \sim \text{Unif}(\{1, \ldots, T\})$
        Draw $u \sim \text{Unif}([0, 1])$
        if $u < w_{c\cdot}/w_{t\cdot}$ **then**
            Set $t\cdot \leftarrow c\cdot$
        end if
    end for
    Draw $\theta_i \sim \mathcal{N}_d(\bar{\theta}_{t\cdot}, \frac{h^2}{M} I_d)$
end for