Divide and Conquer: A Mixture-Based Approach to Regional Adaptation for MCMC

Yan Bai

Department of Statistics, University of Toronto
Collaborators: V. Radu Craiu (University of Toronto) and Antonio F. Di Narzo (University of Bologna)

Graduate Research Day on April 29, 2010

The Problem

- We assume that the population of interest is heterogenous, or it can be represented as a non-standard density. The posterior distribution can be multimodal.
- MCMC sampling from multimodal distributions can be extremely difficult as the chain can get trapped in one mode due to low probability regions between the modes.

Some approaches:

- Gelman and Rubin (1992); Geyer and Thompson (1995); Neal (1996); Richardson and Green (1997); Kou et al. (2006).
- One possible approach is to approximate the multimodal posterior distribution with a mixture of Gaussians in West (1993) who shows that such an approximation may be useful for computation even if the posterior is skewed and not necessarily multimodal.
- Adaptive MCMC algorithms based on the same natural approach have been developed by Giordani and Kohn (2006), Andrieu and Thoms (2008) and Craiu et al. (2009).

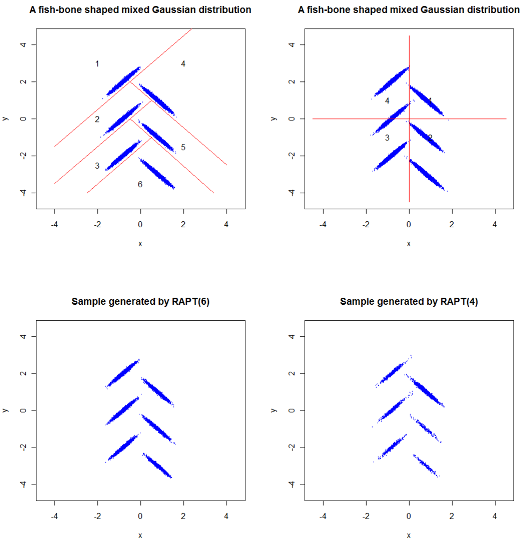
Regional AdaPTive Sampler (RAPT) in Craiu et al. (2009)

- Assume that one has reasonable knowledge about regions where different sampling regions are needed.
- One could use sophisticated methods to detect the modes of a multimodal distribution (see Sminchisescu and Triggs (2001), Neal (2001)), but it is not clear how to use such techniques for defining the desired partition of the sample sample.
- Simply, assume the sample space $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$. RAPT's proposal

$$\tilde{Q}^{(j)}(X_n, \cdot) = \sum_{i=1}^{2} \lambda_i^{(j)} Q_i(X_n, \cdot) \text{ for } j = 1, 2,$$

  where Q_i and the mixture weights $\lambda_i^{(j)}$ are adapted.
- The regions remain unchanged.

RAPT



A fish-bone shaped mixed Gaussian distribution
A fish-bone shaped mixed Gaussian distribution
Sample generated by RAPT(6)
Sample generated by RAPT(4)
x
y

Regional Adaptive with Online Recursion (RAPTOR)

- We consider a different framework allowing the regions to evolve as the simulation proceeds.
- The regional adaptive random walk Metropolis algorithm proposed here relies on the approximation of the target distribution π with a mixture of Gaussians.
- The partition of the sample space used for RAPTOR is defined based on the mixture parameters which are updated using the simulated samples.
- The algorithm 7 in Andrieu and Thoms (2008) differs from RAPTOR in a few important aspects.

RAPTOR - Recursive Adaptation

- Assume that π has K modes in the sample space $\mathcal{S} \subset \mathbb{R}^d$. Consider its approximation by the mixture model

$$\tilde{q}_\eta(x) = \sum_{j=1}^{K} \beta^{(j)} N_d(x, \mu^{(j)}, \Sigma^{(j)}), \tag{1}$$

where $\sum_{j=1}^{K} \beta^{(j)} = 1$ and $N_d(x, \mu, \Sigma)$ is the probability density of a d-variate Gaussian distribution with mean μ and covariance matrix Σ.

- We are facing an online setting in which the parameters need to be updated each time new data are added to the sample.
- Suppose that at time $n-1$ the current parameter estimates are $\eta_{n-1} = \left\{\beta_{n-1}^{(j)}, \mu_{n-1}^{(j)}, \Sigma_{n-1}^{(j)}\right\}_{1 \le j \le K}$ and the available samples are $\{x_0, x_1, \cdots, x_{n-1}\}$.
- We define the mixture indicator Z_n such that given x_n, $\mathbf{P}(Z_n = j \mid x_n, \eta_{n-1}) = \nu_n^{(j)}$

$$\nu_n^{(j)} = \frac{\beta_{n-1}^{(j)} N_d(x_n, \mu_{n-1}^{(j)}, \Sigma_{n-1}^{(j)})}{\sum_{i=1}^{K} \beta_{n-1}^{(i)} N_d(x_n, \mu_{n-1}^{(i)}, \Sigma_{n-1}^{(i)})}, \forall 1 \le i \le n, 1 \le j \le K. \tag{2}$$

RAPTOR - Recursive Adaptation

- The recursive estimator $\eta_n = \left\{\beta_n^{(j)}, \mu_n^{(j)}, \Sigma_n^{(j)}\right\}_{1\le j\le K}$

$$\beta_n^{(j)} = \frac{1}{n+1}\sum_{i=0}^{n} \nu_i^{(j)}$$
$$\mu_n^{(j)} = \mu_{n-1}^{(j)} + \rho_n\gamma_n^{(j)}(x_n - \mu_{n-1}^{(j)}),$$
$$\Sigma_n^{(j)} = \Sigma_{n-1}^{(j)} + \rho_n\gamma_n^{(j)}\left((1-\gamma_n^{(j)})(x_n - \mu_{n-1}^{(j)})(x_n - \mu_{n-1}^{(j)})^\top - \Sigma_{n-1}^{(j)}\right) \tag{3}$$

where $\gamma_n^{(j)} = \frac{\nu_n^{(j)}}{\sum_{i=0}^{n}\nu_n^{(j)}}$ and ρ_n is a non-increasing positive sequence.

- Sample Mean and Sample Covariance: given $\{x_0, x_1, \cdots, x_n\}$

$$\mu_n^{<w>} = \mu_{n-1}^{<w>} + \frac{1}{n+1}(x_n - \mu_{n-1}^{<w>}),$$
$$\Sigma_n^{<w>} = \Sigma_{n-1}^{<w>} + \frac{1}{n+1}\left((1-\frac{1}{n+1})(x_n - \mu_{n-1}^{<w>})(x_n - \mu_{n-1}^{<w>})^\top - \Sigma_{n-1}^{<w>}\right). \tag{4}$$

RAPTOR - Definition of Regions

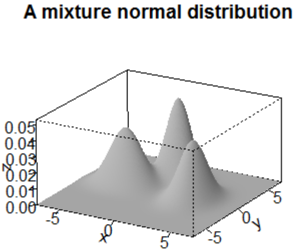
- Suppose that the K-partition of the sample space $\Pi = \{\mathcal{S}^{(1)}, \mathcal{S}^{(2)}, \cdots, \mathcal{S}^{(K)}\}$ satisfying $\mathcal{S} = \mathcal{S}^{(1)} \cup \mathcal{S}^{(2)} \cup \cdots \cup \mathcal{S}^{(K)}$ and $\mathcal{S}^{(i)} \cap \mathcal{S}^{(j)} = \emptyset$ for $i \neq j$.
- Denote the projection of π on the set A by $\pi_A(x) = \pi(x)\mathbb{I}_A(x) / \int_A \pi(y)dy$. We try to find an "optimal" estimator of K-partition minimizing

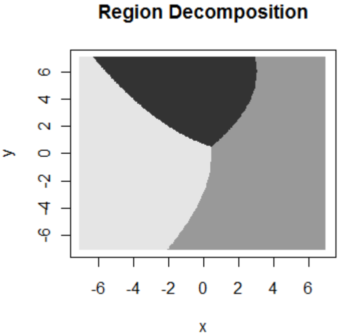
$$\max_{1 \le i \le K} \left| \mathrm{KL}(\pi_{\mathcal{S}^{(i)}}, N_d^{(i)}) \right|$$

where $N_d^{(i)}(x) = N_d(x, \mu^{(i)}, \Sigma^{(i)})$ (defined in Eq. (1)) and the Kullback-Leibler divergence $\mathrm{KL}(f, g) = \int \log(f(x)/g(x)) f(x) dx$.
- With this aim, we define

$$\mathcal{S}_n^{(j)} = \left\{ x \in \mathcal{S} : \arg\max_i N_d(x, \mu_n^{(i)}, \Sigma_n^{(i)}) = j \right\}. \tag{5}$$

RAPTOR - Definition of Regions





RAPTOR - Definition of the Proposal Distribution

- At each time n, the sample $\{x_0, x_1, \cdots, x_n\}$ is obtained, the corresponding parameter estimators $\{\mu_n^{(j)}, \Sigma_n^{(j)} : j = 1, 2, \cdots, K\}$, $\mu_n^{<w>}$, $\Sigma_n^{<w>}$ are computed. The recursive estimates can determine the recursive region partition $\{\mathcal{S}^{(1)}, \mathcal{S}^{(2)}, \cdots, \mathcal{S}^{(K)}\}$.
- Propose the value y_{n+1} from the Proposal distribution

$$Q_n(x_n, dy) = (1-\alpha) \sum_{j=1}^{K} \mathbb{I}_{\mathcal{S}^{(j)}}(x_n) N_d(y; x_n, s_d \tilde{\Sigma}_n^{(j)}) dy + \alpha N_d(y; x_n, s_d \tilde{\Sigma}_n^{<w>}) dy,$$

  where $s_d = 2.38^2/d$, $\tilde{\Sigma}_n^{(j)} = \Sigma_n^{(j)} + \epsilon I_d$, $\tilde{\Sigma}_n^{<w>} = \Sigma_n^{<w>} + \epsilon I_d$, and $\alpha = 1/3$.
- Accept or reject y_{n+1} for x_{n+1} according to Metropolis acceptance rate $\min(1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)})$.
- Compute the recursive parameter estimators indexed by $n+1$.

Theoretical Results

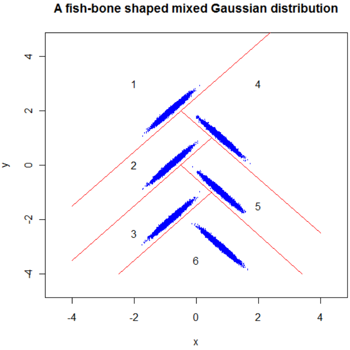
(A1): There is a compact set $\mathcal{S} \subset \mathbb{R}^d$ such that the target density π is continuous on $\mathcal{S}$, positive on the interior of $\mathcal{S}$, and zero outside of $\mathcal{S}$.
(A2): The sequence $\{\rho_j : j \geq 1\}$ is positive and non-increasing.
(A3): For all $k = 1, 2, \cdots, K$,

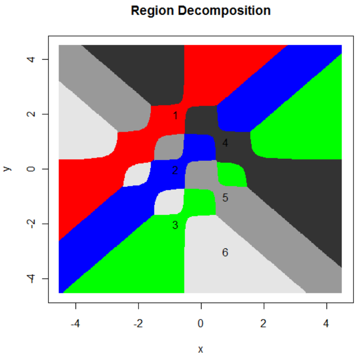
$$\mathbf{P}(\limsup_{i \to \infty} \sup_{l \geq i} \sum_{j=i}^{l} \rho_j \gamma_j^k = 0) = 1.$$

Theorem

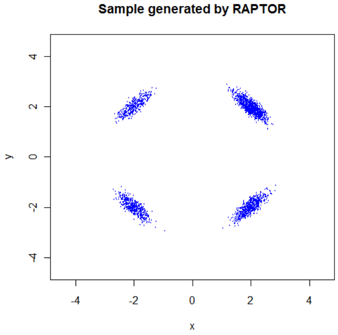
a) Assuming (A1-2), the RAPTOR algorithm is ergodic to π.
b) Assuming (A2-3), the adaptive parameters $\mu_n^{(j)}, \Sigma_n^{(j)}$ converge in probability for any $j \in \{1, 2, \cdots, K\}$.

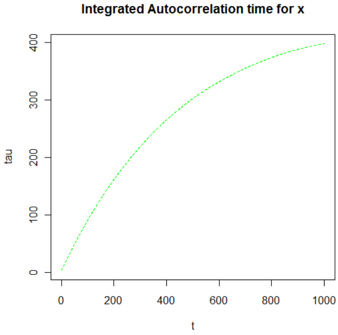
A Fish-Bone-Shaped Distribution





A Square-Shaped Distribution





Summary

- The efficiency of RAPT algorithm is strongly dependent on the decomposition of the state space. If a good decomposition is chosen, the algorithm can perform very well.
- The recursive region study provides a simple way to solve the problem how to decompose the state space for RAPT. But, it takes more time on the computation as the number of modes is large.
- The performance of both algorithms also depends on the pattern of the target distribution.
- Using the mixing parameters $\{\lambda_j^{(i)} : 1 \leq i, j \leq K\}$ for the local adaptive sampler for RAPTOR, it performs better where

$$\lambda_j^{(i)}(n) = \begin{cases} \frac{d_j^{(i)}}{\sum_{l=1}^K d_l^{(i)}} & \text{if } \sum_{l=1}^K d_l^{(i)} > 0; \\ \frac{1}{2} & \text{otherwise} \end{cases} \tag{6}$$

with $d_j^{(i)}(n)$ is the average square jump distance up to iteration n computed every time the accepted proposal was generated from jth regional proposal and the current state of the chain was in $\mathcal{S}_i$.

C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Statist. Comput.*, 18:343–373, 2008.

R.V. Craiu, J. Rosenthal, and C. Yang. Learn from thy neighbor: Parallel-chain adaptive and regional MCMC. *J. Amer. Statist. Assoc.*, 104:1454–1466, 2009.

A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statist. Sci.*, pages 457–511, 1992.

C. J. Geyer and E. A. Thompson. Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Amer. Statist. Assoc.*, 90:909–920, 1995.

P. Giordani and R. Kohn. Adaptive Independent Metropolis-Hastings by fast estimation of mixed normals. *Working paper*, 2006.

S. Kou, Z. Qing, and W. Wong. Equi-energy sampler with applications in statistical inference and statistical mechanics. *Ann. Statist.*, 34:1581–1619, 2006.

R. Neal. Simulating multimodal distributions using tempered transitions. *Statistics and Computing*, 6:353–366, 1996.

R. Neal. Annealed importance sampling. *Statistics and Computing*, 11:125–139, 2001.

S. Richardson and P.J. Green. On Bayesian analysis of mixtures with an unknown number of components. *J. Royal Statist. Society*, 59:731–792, 1997.

C. Sminchisescu and B Triggs. Covariance-scaled sampling for monocular 3D body tracking. *IEEE International Conference on Computer Vision and Pattern Recognition*, 1:Hawaii, 447C454, 2001.