Recent Advances in Regional Adaptation for MCMC

Radu Craiu

Department of Statistics
University of Toronto

Collaborators:
Yan Bai (Statistics, Toronto)
Antonio Fabio di Narzo (Statistics, Bologna)
Jeffrey Rosenthal (Statistics, Toronto)

AdapSki, January 2011

Outline

Regional AMCMC

- We consider problems in which the random walk Metropolis (RWM) or the independent Metropolis algorithms (IM) are used to sample from the target distribution π with support $\mathcal{S}$.

  - Given the current state of the MC, x, a "proposed sample" y is drawn from a proposal distribution $P(y|x)$ that satisfies symmetry, i.e. $P(y|x) = P(x|y)$.
  - The proposal y is accepted with probability $\min\{1, \pi(y)/\pi(x)\}$.
  - If y is accepted, the next state is y, otherwise it is (still) x.

- The *random walk Metropolis* is obtained when $y = x + \epsilon$ with $\epsilon \sim f$, f symmetric, usually $N(0, V)$.
- If $P(y|x) = P(y)$ then we have the *independent Metropolis* sampler (acceptance ratio is modified).

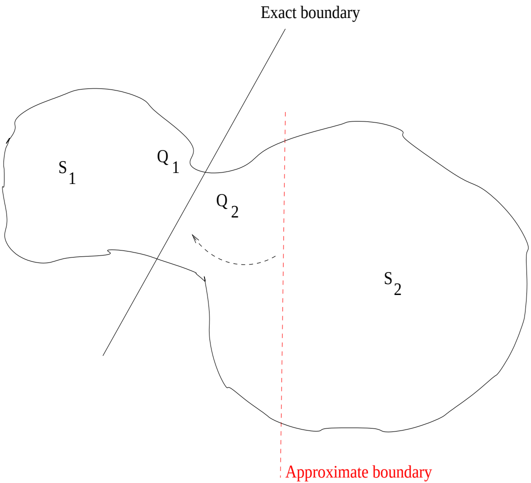
Adaptive MCMC

- Uses an initialization period to gather information about the target π.
- The initial samples are used to produce estimates for the adaption parameters who are subsequently adapted "on the fly" until the simulation is stopped (indefinitely).
- Adaption strategies are adopted based on
  - (i) Theoretical results on the optimality of MCMC, e.g. optimal acceptance rate for MH algorithms.
  - (ii) Other strategies learned by studying the "classical" MCMC algorithms, e.g. annealing, other Metropolis samplers...
  - (iii) Our ability to prove theoretically that the adaptive chain samples correctly from π.
- It is usually easier to do (iii) if we assume $\mathcal{S}$ is compact.

Multimodal targets and regional AMCMC

- Multimodality is a never-ending source of headaches in MCMC.
- "Optimal" proposal may depend on the region of the current state.

Regional Adaptation with Dynamic Boundary



Today: What to do if π is approximated by a mixture of Gaussians.

Mixture representation of the target

- Suppose
$$\tilde{q}_\eta(x) = \sum_{k=1}^{K} \beta^{(k)} N_d(x; \mu^{(k)}, \Sigma^{(k)}),$$
where $\beta^{(k)} > 0$ for all $1 \le k \le K$ and $\sum_{k=1}^{K} \beta^{(k)} = 1$, is a good approximation for the target π.
- At each time n during the simulation process one has available n dependent Monte Carlo samples which are used to fit the mixture $\tilde{q}_\eta$. Can we fit the mixture parameters recursively?
- Given the mixture parameters, define a regional RWM algorithm in which $\mathcal{S}$ is partitioned so that when the chain is in the k-th region we propose from the k-th component of $\tilde{q}_\eta$.

Online EM Updates

At time $n-1$ the current parameter estimates are $\eta_{n-1} = \{\beta_{n-1}^{(k)}, \mu_{n-1}^{(k)}, \Sigma_{n-1}^{(k)}\}_{1\le k\le K}$ and the available samples are $\{x_0, x_1, \ldots, x_{n-1}\}$; when observing x_n we update (see Andrieu and Moulines, Ann. Appl. Probab. 2006)

$$\beta_n^{(k)} = \frac{1}{n+1}\sum_{i=0}^{n} \nu_i^{(k)} = s_{n-1}^{(k)} + \tfrac{1}{n+1}(\nu_n^{(k)} - s_{n-1}^{(k)}),$$

$$\mu_n^{(k)} = \mu_{n-1}^{(k)} + \rho_n\gamma_n^{(k)}\left(x_n - \mu_{n-1}^{(k)}\right),$$

$$\Sigma_n^{(k)} = \Sigma_{n-1}^{(k)} + \rho_n\gamma_n^{(k)}\left((1-\gamma_n^{(k)})(x_n-\mu_{n-1}^{(k)})(x_n-\mu_{n-1}^{(k)})^\top - \Sigma_{n-1}^{(k)}\right),$$

where $\nu_m^{(k)} = \frac{\beta_{m-1}^{(k)} N_d(x_m;\mu_{m-1}^{(k)},\Sigma_{m-1}^{(k)})}{\sum_{k'}\beta_{m-1}^{(k')} N_d(x_m;\mu_{m-1}^{(k')},\Sigma_{m-1}^{(k')})}$, $s_m^{(k)} = \frac{1}{m+1}\sum_{i=0}^{m}\nu_i^{(k)}$,
$\gamma_m^{(k)} = \frac{\nu_m^{(k)}}{(m+1)s_m^{(k)}}$, and $\rho_m = m^{-1.1}$ for all $1\le m\le n$, $1\le k\le K$.

Definitions of Regions

- We would like to define the partition $\mathcal{S} = \cup_{k=1}^{K} \mathcal{S}^{(k)}$ so that, on each set $\mathcal{S}^{(k)}$, π is more similar to $N_d(x; \mu^{(k)}, \Sigma^{(k)})$ than to any other mixture component.
- We maximize the sum of differences between Kullback-Leibler (KL) divergences; when $K = 2$ we want to maximize

$$\mathrm{KL}(\pi, N_d(\cdot; \mu^{(2)}, \Sigma^{(2)}) \mid \mathcal{S}^{(1)}) - \mathrm{KL}(\pi, N_d(\cdot; \mu^{(1)}, \Sigma^{(1)}) \mid \mathcal{S}^{(1)}) + \mathrm{KL}(\pi, N_d(\cdot; \mu^{(1)}, \Sigma^{(1)}) \mid \mathcal{S}^{(2)}) - \mathrm{KL}(\pi, N_d(\cdot; \mu^{(2)}, \Sigma^{(2)}) \mid \mathcal{S}^{(2)}),$$

  where $\mathrm{KL}(f, g|A) = \int_A \log(f(x)/g(x)) f(x) dx$.
- Define

$$\mathcal{S}_n^{(k)} = \{x : \arg\max_{k'} N_d(x; \mu_n^{(k')}, \Sigma_n^{(k')}) = k\}.$$

Proposal Distribution

- The proposal distribution depends on:
  i) The mixture parameters estimated using the online EM,
  ii) The regions defined previously.
- In addition to local optimality (within each modal region) we seek good global traffic (between regions) so we add a global component to the proposal distribution (see also Guan and Krone, Ann. Appl. Probab, 2007).
- Let $\alpha = 0.3$ and $\Sigma_n^{<w>}$ be the sample covariance . Put $\tilde{\Sigma}_n^{<w>} = \delta\Sigma_n^{<w>} + \epsilon \mathbf{I}_d$, $\tilde{\Sigma}_n^{(k)} = \delta\Sigma_n^{(k)} + \epsilon \mathbf{I}_d$, $1 \leq k \leq K$.
- The RAPTOR proposal is then

$$Q_n(x, dy) = (1-\alpha)\sum_{k=1}^{K} 1_{\mathcal{S}_n^{(k)}}(x) N_d(y; x, s_d \tilde{\Sigma}_n^{(k)})dy$$

$$+\alpha N_d(y; x, s_d \tilde{\Sigma}_n^{<w>})dy,$$

Implementation of RAPTOR

- Run in parallel a number (5-10) of RWM algorithms with fixed kernels started in different regions of the sample space (if the local modes are known start there) for an initialization period of M steps.
- At step $M+1$, compute the mixture parameters using the EM algorithm as well as the sample mean and covariance matrix to obtain

$$\Gamma_0 = \{\mu_0^{(1)}, \ldots, \mu_0^{(K)}, \mu_0^{<w>}, \tilde{\Sigma}_0^{(1)}, \ldots, \tilde{\Sigma}_0^{(K)}, \tilde{\Sigma}_0^{<w>}\}$$

.

- At each step $M+n \geq M+1$ we:
  i) Update the mixture parameters Γ_n
  ii) Construct the partition based on Γ_n and
  iii) Sample the proposal $Y_{M+n} \sim Q_n(X_{M+n}; dy)$.

Theoretical Results

- Assumptions:
  (A1) There is a compact subset $\mathcal{S} \subset \mathbb{R}^d$ 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 $k = 1, \cdots, K$,
  $\Pr(\lim_{i\to\infty} \sup_{l\geq i} \sum_{j=i}^{l} \rho_j \gamma_j^{(k)} = 0) = 1$.
- We work with $\rho_j = j^{-1.1}$ so that **(A2)** and **(A3)** are satisfied.

Convergence

a) Assuming **(A1)** and **(A2)**, RAPTOR is ergodic to π.
b) Assuming **(A2)** and **(A3)**, the adaptive parameter $\{\Gamma_n\}_{n\geq 0}$ converges in probability.

Non-Compactness and Regime-Switching

- RAPTOR assumes $\mathcal{S}$ is compact. Practically, the impact is small but the theoretical gap is vexing.
- What to do if $\mathcal{S}$ is not compact?
- We know that a well-tuned IM has better convergence properties than a RWM. However, it is usually impossible to produce a well-tuned IM using only few samples. The AMCMC literature on adapting IM suggests we first sample using a RWM and then switch to an IM.
- Do things have to happen so suddenly?
- How to decide when it is time to switch?

Non-Compactness

- Strategy: Adapt only within the compact $\mathcal{K} \subset \mathcal{S}$. Use the adapting kernel when the chain is in $\mathcal{K}$ and use a fixed kernel outside $\mathcal{K}$.
- If using the Metropolis algorithm the proposals are assumed to have compact support.
- Proof of ergodicity is direct and requires (pretty much) only diminishing adaptation:

Let $D_n = \sup_{x \in K} ||T_{\gamma_{n+1}}(x, \cdot) - T_{\gamma_n}(x, \cdot)||_{TV}$.
Diminishing Adaptation: $\lim_{n \to \infty} D_n = 0$ in probability.

Regime-Switching

- We propose a more gradual transition between the "accumulation of data" and the "full adaptation" regimes. Usually, the former is done with a RWM and the latter with IM.
- Combine with non-compactness idea and use:
  - A mixture of adapting RWM and IM inside the compact $\mathcal{K}$
  - A fixed RWM outside $\mathcal{K}$

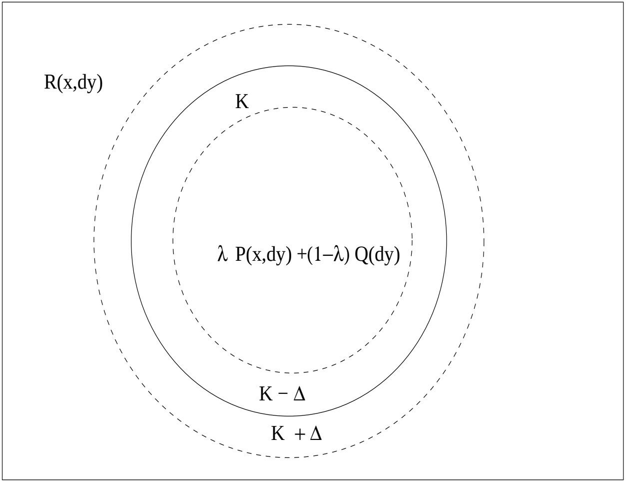
$$\tilde{P}_\Gamma(x, A) = 1_{\mathcal{K}}(x)\left[\lambda_\Gamma P_\Gamma(x, A) + (1 - \lambda_\Gamma) Q_\Gamma(A)\right] + 1_{\mathcal{K}^c}(x) R(x, A),$$

- P_Γ, R are RWM kernels using proposal distributions of compact support (of diameter Δ)
- Q_Γ is a IM kernel using a proposal with support $\mathcal{K}$.

Regime-Switching

- We want λ_Γ to approach zero as Q_Γ gets closer to π on $\mathcal{K}$.
- The samples used to adapt Γ should not be also used for determining the distance between Q_Γ and π.
- Many adaptive strategies that satisfy Diminishing Adaptation can be used for P_Γ and Q_Γ.

Regime-Switching



Regime-Switching

- Given, $y_1, \ldots, y_n$ samples from π and assuming that $\pi(x) = f(x)/M$

$$
\begin{aligned}
KL(\pi, q_\gamma) &= \int \log(\pi(x)/q_\gamma(x))\pi(x)dx = \\
&= A(\pi, q_\gamma) + M \approx \frac{1}{n}\sum_{i=1}^{n} \log(f(x_i)/q_\gamma(x_i)) + M
\end{aligned}
$$

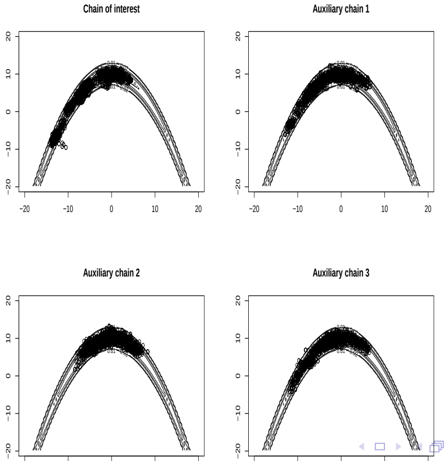
- Assume A is estimated $m = 2h$ times and set

$$
\lambda_m = \min\left\{0.05 + \frac{1}{m^\theta}\frac{\hat{A}_{(\frac{m}{2})} - \hat{A}_{(m)}}{\hat{A}_{(1)} - \hat{A}_{(m)}}, 0.95\right\}
$$

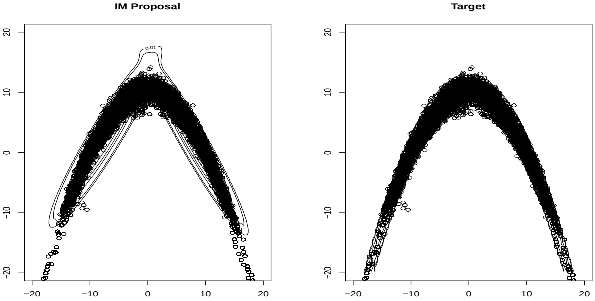
where $\hat{A}_{(1)}, \ldots, \hat{A}_{(m)}$ are the order statistics for the sequence of estimates. We used $\theta \in \{1/10, 1/5\}$ in all examples.

Banana Example

- Let $\pi(x) \propto \exp\left[-x_1^2/200 - \frac{1}{2}(x_2 + Bx_1^2 - 100B)^2 - \frac{x_3^2}{2}\right], \quad B = 0.1.$

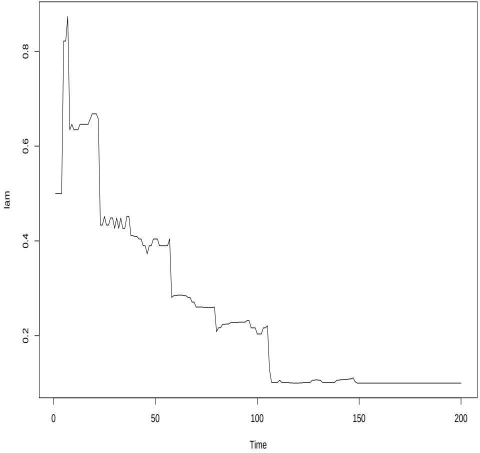


Banana Example



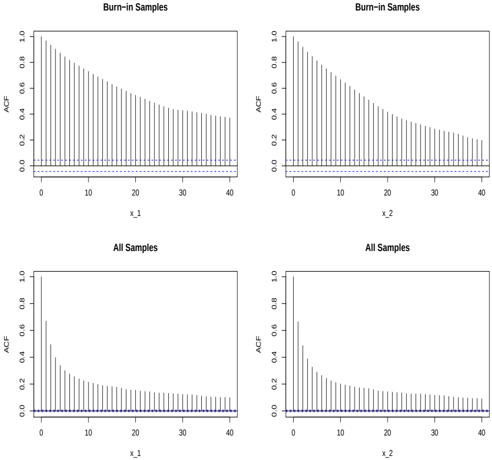
IM proposal (left) and Target (right) 2-dim projections.

Banana Example



Evolution of the lambda coefficient as simulation proceeds.

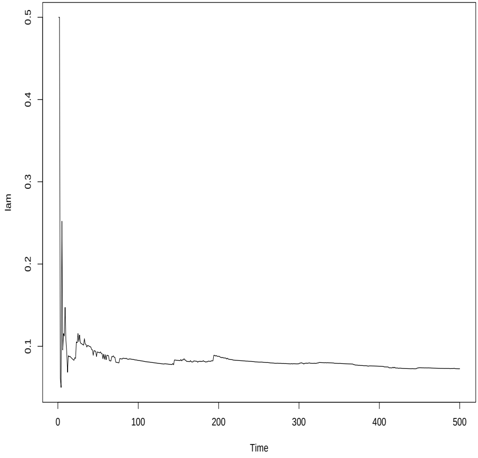
Banana Example



Normal Mixture Example

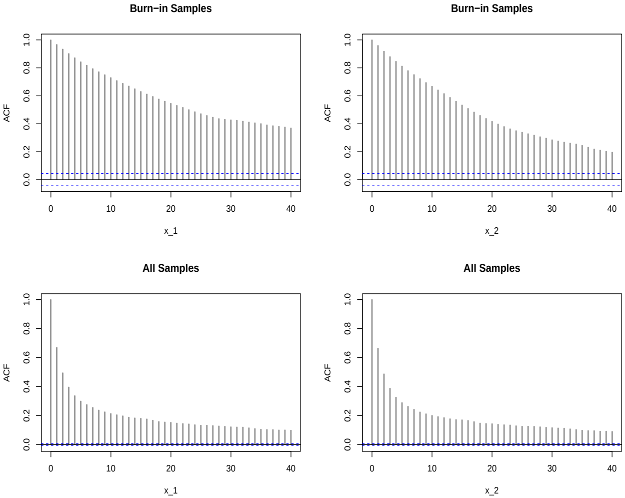
- Let
$$\pi(x) = 0.5N(x|\mu_1, \Sigma_1) + 0.5N(x|\mu_2, \Sigma_2)$$
with $x \in \mathbb{R}^3$, $\mu_1 = (-4, -4, 0)^T$, $\mu_2 = (8, 8, 5)^T$, $\Sigma_1 = 2\mathbb{I}$ and $\Sigma_2 = 4\mathbb{I}$.

Normal Mixture Example



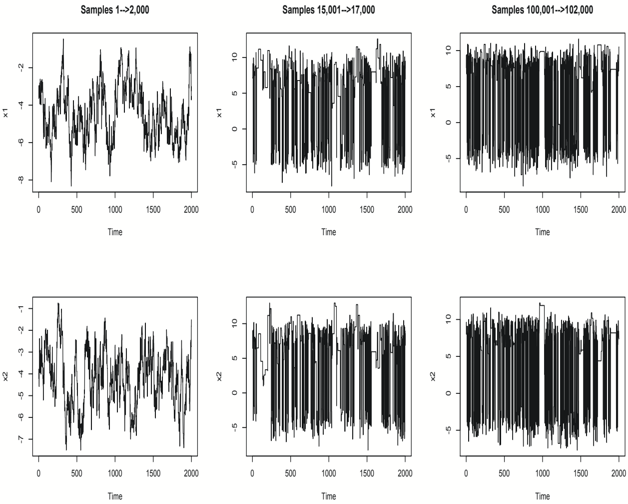
Evolution of the lambda coefficient as simulation proceeds.

Normal Mixture Example



ACF plots for burn-in samples (top) and all samples (bottom).

Normal Mixture Example



Trace plots obtained at different stages in the simulation.

RAPTOR: Simulation Setup

- Target distribution is

$$\pi(x; m, s) \propto 1_{\mathbf{C}_d}(x) \left[0.5 N_d(x; -m \times \mathbf{1}, \mathbf{I_d}) + 0.5 N_d(x; m \times \mathbf{1}, s \times \mathbf{I_d})\right]$$

where $\mathbf{C}_d = [-10^{10}, 10^{10}]^d$

- We consider the scenarios given by the following ten combinations of parameter values

$$(d, m, s) \in \{(2,1,1), (5,0.5,1), (2,1,4), (5,0.5,4), (2,0,1), (5,0,1), (2,0,4), (5,0,4), (2,2,1), (5,1,1)\}.$$

RAPTOR: Simulation Results

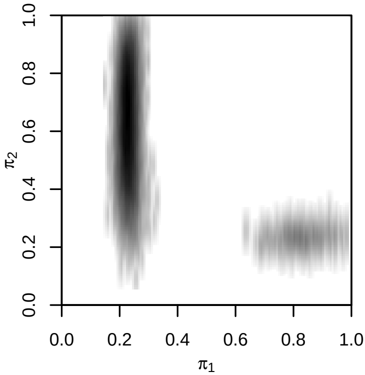
(m, s)	RAPTOR	RRWM	RAPT
d=2			
$(1, 1)$	21	21	22
$(1, 4)$	43	39	46
$(0, 1)$	10	8	11
$(0, 4)$	25	20	28
d=5			
$(0.5, 1)$	30	22	41
$(0.5, 4)$	72	62	108
$(0, 1)$	23	18	29
$(0, 4)$	51	48	62

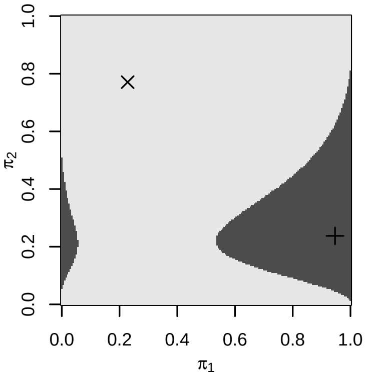
RAPTOR: Genetic Instability of Esophageal Cancers

- Cancer cells suffer a number of genetic changes during disease progression, one of which is *loss of heterozygosity (LOH)*.
- Chromosome regions with high rates of LOH are hypothesized to contain genes which regulate cell behavior and may be of interest in cancer studies.
- We consider 40 measures of frequencies of the event of interest (LOH) with their associated sample sizes. The model adopted for those frequencies is a mixture model

$$X_i \sim \eta \,\text{Binomial}(N_i, \pi_1) + (1-\eta)\,\text{Beta-Binomial}(N_i, \pi_2, \gamma).$$

RAPTOR: Graphical Results





References

- Regional Adaptation
  - Andrieu, C. and Thoms, J. (2008). *Statistics and Computing*, **18**, 343-373.
  - Bai, Y., C.R.V. and Di Narzo, A. F. (201?) *Journal of Computational and Graphical Statistics* (to appear).
  - C.R.V., Rosenthal, J. and Yang, C. (2009). *JASA*, **104**, 1454-1466.
  - C.R.V. and Rosenthal, J. (201?). Velvet Revolution in Adaptive MCMC: The Benefits of Smooth Regime Change.
- Online EM
  - Andrieu, C. and Moulines, E. (2006). *Annals of Applied Probability*, **16**, 1462-1505.
  - Cappé, O. and Moulines, E (2009). *JRSS-B*, **71**, 593 -613.