Metropolis-Hastings Algorithm for Mixture Model and its Weak Convergence

Kengo, KAMATANI

University of Tokyo, Japan

1. Gibbs sampler usually works well.
2. However in certain settings, it works poorly. ex) Mixture model.
3. Fortunately, we found an alternative MCMC method which works better in simulation.

Problem

Both 2 and 3 are uniformly ergodic. Therefore, to compare those methods, we have to calculate the convergence rates. It is very difficult!

Therefore, in Harris recurrence approach, the comparison is difficult.
We take another approach.

Summary of the talk

- Sec. 1 I show a bad behavior of the Gibbs sampler.
- Sec. 2 Define efficiency (consistency) of MCMC. Prove that the Gibbs sampler has a bad convergence property.
- Sec. 3 Propose a new MCMC. Prove that the new MCMC is better than the Gibbs sampler.

Note that...

- Harris recurrence property is also very important for our approach. Without this property, our approach is useless.
- The another motivation of our approach is to divide two different convergence issues 1) convergence to the local area and 2) consistency
- Only the mixture model is considered here. However it may be useful to other models.

Outline

Outline

Bad behavior of the Gibbs sampler

Model description

1. Consider a model

$$p_{X|\Theta}(dx|\theta) = (1-\theta)F_0(dx) + \theta F_1(dx).$$

2. Flip a coin with the proportion of head θ. If the coin is head, generate x from F_1, otherwise, from F_0.
3. We do **not** observe the coin but x.
4. Observation $x_n = (x^1, x^2, \ldots, x^n)$, $x^i \sim p_{X|\Theta}(dx|\theta_0)$.
   Prior distribution $p_\Theta = \mathrm{Beta}(\alpha_1, \alpha_0)$.

We want to calculate the posterior distribution.

Bad behavior of the Gibbs sampler

Gibbs sampler

1. Set $\theta(0) \in \Theta$.
2. $y^i \sim \mathrm{Bi}(1, p_i)$ $(i = 1, 2, \ldots, n)$ where

$$p_i = \frac{\theta(0) f_1(x^i)}{(1 - \theta(0)) f_0(x^i) + \theta(0) f_1(x^i)}.$$

   Count $m = \sum_{i=1}^{n} y^i$. $F_i(dx) = f_i(x)dx$.
3. Generate $\theta(1) \sim \mathrm{Beta}(\alpha_1 + m, \alpha_0 + n - m)$.
4. Empirical measure of $(\theta(0), \theta(1), \ldots, \theta(N-1))$ is an estimator of the posterior distribution.

The next figure is a path of the Gibbs sampler when the true model is F_0, that is, $\theta_0 = 0$.

Bad behavior of the Gibbs sampler

Gibbs sampler

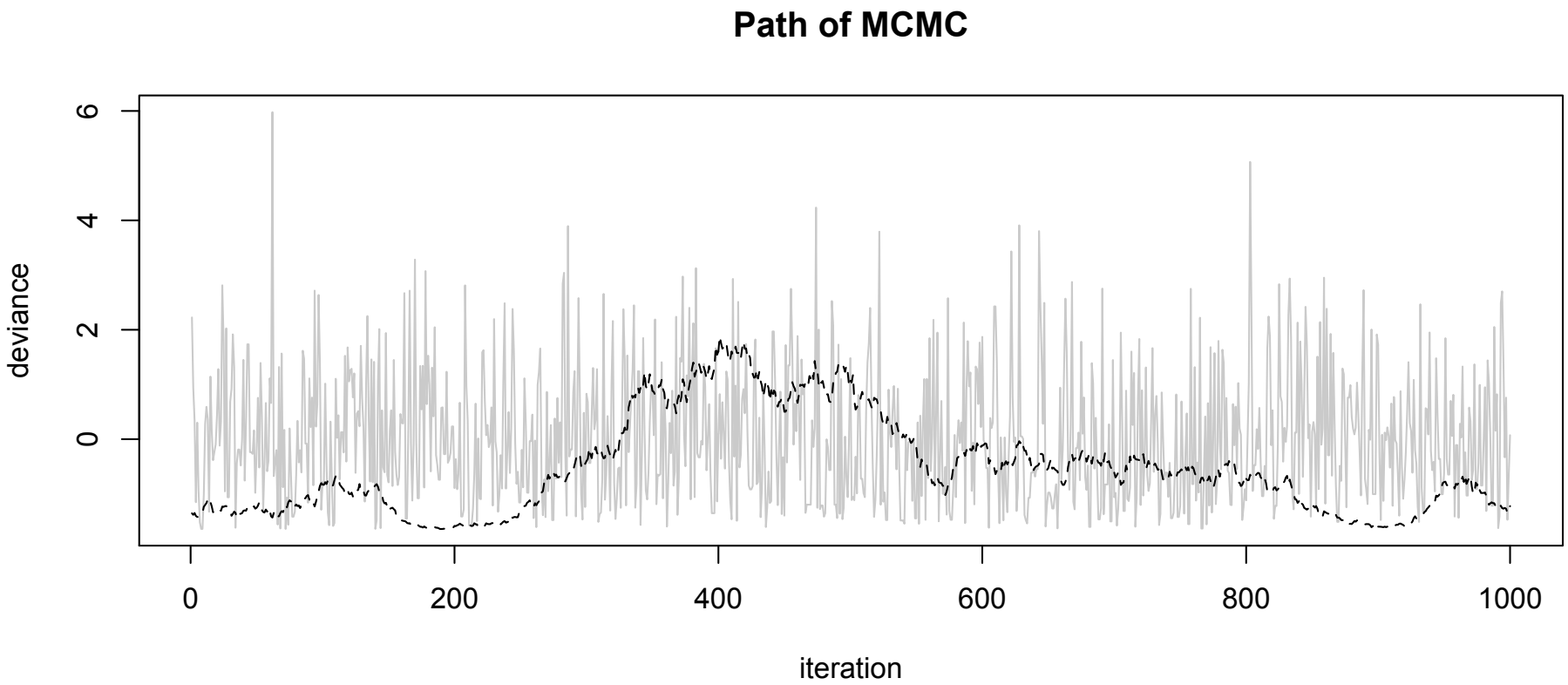

Figure: Plot of paths of MCMC methods for $n = 10^4$. The dashed line is a path from the Gibbs sampler and the solid line is the MH algorithm.

Bad behavior of the Gibbs sampler

How to define efficiency

1. MCMC methods produce complicated Markov chain.
2. We make an approximation of MCMC method.

We observe the behavior of MCMC methods when the sample size $n \to \infty$.

Outline

Weak convergence of MCMC

What is MCMC?

Write s instead of θ.

1. For each observation x, Gibbs sampler produces paths $s = (s(0), s(1), \ldots)$ in S^∞.
2. In other words, for $x \in X$, Gibbs sampler defines a law $G^x \in \mathcal{P}(S^\infty)$.
3. Therefore, a Gibbs sampler is a set of probability measures $G = (G^x; x \in X)$ (Later, we will consider G as a random variable $G(x) = G^x$).

Let $\hat{\nu}_m(s)$ be the empirical measure of $s(0), \ldots, s(m-1)$.
Let ν^x be the target distribution for each x.

We expect that $d(\hat{\nu}_m(s), \nu^x) \to 0$ in a certain sense.

Weak convergence of MCMC

Consistency

1. We expect that as $m \to \infty$,

$$E_G(d(\hat{\nu}_m(s), \nu)) \to 0.$$

But G and ν depend on x!.

2. We expect that as $m \to \infty$,

$$E_{G^x}(d(\hat{\nu}_m(s), \nu^x)) = o_P(1).$$

But G^x and ν^x may depend on n!.

Weak convergence of MCMC

Consistency

Definition

$(M_n = (M_n^x); n \in \mathbf{N})$: sequence of MCMC.
We call $(M_n; n \in \mathbf{N})$ *consistent* for $\nu_n = (\nu_n^x)$ if for **any** $m(n) \to \infty$,

$$E_{M_n^x}(d(\hat{\nu}_{[m(n)]}(s), \nu_n^x)) = o_{P_n}(1).$$

For a regular model, the Gibbs sampler has consistency with scaling $\theta \mapsto n^{1/2}(\theta - \theta_0)$.

Weak convergence of MCMC

Degeneracy

Definition

1. If a measure $\omega \in \mathcal{P}(S^\infty)$ satisfies the following, we call it degenerate:

$$\omega(\{s; s(0) = s(1) = s(2) = \cdots\}) = 1 \tag{1}$$

2. We also call M degenerate (in P) if M^x is **degenerate** a.s. x.
3. If $M_n \Rightarrow M$ and M degenerate, we call M_n degenerate in the limit.

The Gibbs sampler G_n for mixture model is **degenerate** with scaling $\theta \mapsto n^{1/2}\theta$ if $\theta_0 = 0$ as $n \to \infty$.

Weak convergence of MCMC

Degeneracy

In fact, G_n tends to a diffusion process type variable with **time** scaling $0, 1, 2, \ldots \mapsto 0, n^{-1/2}, 2n^{-1/2}, \ldots$!

Under both space and time scaling, G_n^x is similar to the law of

$$dS_t = (\alpha_1 + S_t Z_n - S_t^2 I)dt + S_t dB_t$$

where $Z_n \Rightarrow N(0, I)$ and I is the Fisher information matrix.

If we take $m(n)n^{-1/2} \to \infty$, the empirical measure converges to the posterior distribution.

We call G_n $n^{1/2}$-weakly consistent.

Outline

MH algorithm converges faster

MH proposal construction

Construct a posterior distribution for another parametric family:

1. Fix $Q \subset \mathcal{P}(X)$.
2. For each θ, set

$$q_{X|\Theta}(dx|\theta) := \mathrm{argmin}_{q \in Q} d(p_{X|\Theta}(dx|\theta), q)$$

   where d is a certain metric. ex) Kullback-Leibler distance.
3. Calculate the posterior $q^n_{\Theta|X_n}(d\theta|x_n)$.

Remark

We assume that we can generate $\theta \sim q^n_{\Theta|X_n}(d\theta|x_n)$ in PC.

This construction is similar to

1. quasi Bayes method (See ex. Smith and Markov 1978)
2. variational Bayes method (See ex. Humphreys and Titterington 2000).

MH algorithm converges faster

MH proposal construction

Construct an independent type Metropolis-Hastings algorithm with target distribution $p^n_{\Theta|X_n}(d\theta|x_n)$.

Step 0 Generate $\theta(0) \sim q^n_{\Theta|X_n}(d\theta|x_n)$. Go to Step 1.

Step i Generate $\theta(i)^* \sim q^n_{\Theta|X_n}(d\theta|x_n)$. Then

$$\theta(i) = \begin{cases} \theta^*(i) & \text{with probability } \alpha(\theta(i), \theta^*(i)) \\ \theta(i-1) & \text{with probability } 1 - \alpha(\theta(i), \theta^*(i)) \end{cases}.$$

Go to Step $i+1$.

MH algorithm converges faster

MH performance (Normal): Mean squared error

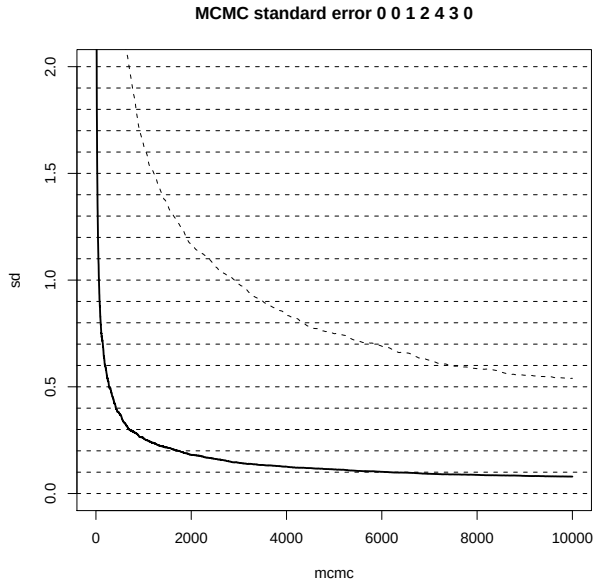

Figure: The dashed line is a path from the Gibbs sampler and the solid line is the MH algorithm for $n = 10$.

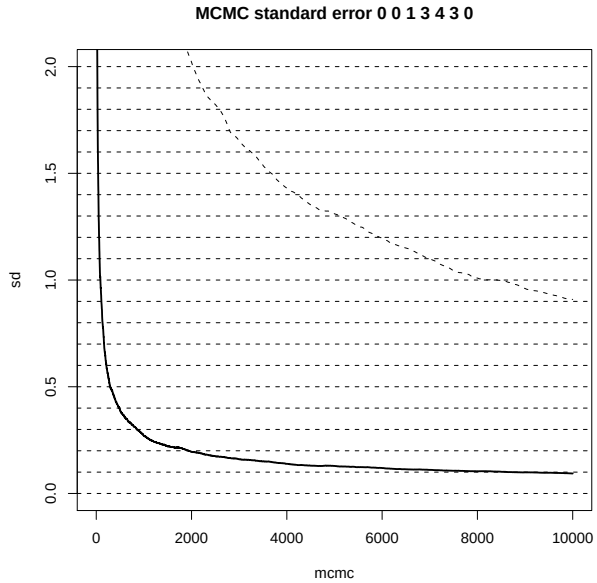

Figure: The same figure as the left. The sample size is 10^2.

MH algorithm converges faster

Remarks

Remarks...

If F_0 and F_1 are similar, the Gibbs sampler becomes even worse. This fact can be verified by setting F_ϵ and $\epsilon \to 0$ instead of $\epsilon \equiv 1$. In this case, we have to take $m(n)\epsilon n^{-1/2} \to \infty$ (G_n is $\epsilon^{-1} n^{1/2}$-weakly consistent).

For other model, there is a case such that $m(n)n^{-1} \to \infty$.

MH algorithm converges faster

Thank you!