Weak convergence of Markov chain Monte Carlo II

KAMATANI, Kengo

Mar 2011 at Le Mans

Background

- Markov chain Monte Carlo (MCMC) method is widely used in Statistical Science.

- It is easy to use, but difficult to analyze.

- We define efficiency for MCMC and observe its property for some models.

Our main interest is to analyze bad behavior of MCMC. It provides useful insight for constructing alternative methods.

Summary of the talk:

- MCMC and its ergodicity (Harris rec. approach).
- Regular and non-regular convergence (Our approach).
- (Weak) Consistency of MCMC.
- Numerical results.

1. MCMC and its ergodicity (Harris rec. approach)

1. MCMC and its ergodicity (Harris rec. approach)

1.1 Monte carlo method

Consider a Bayesian statistical model:

$$\begin{cases} \theta & \sim & \Pi(d\theta) & \text{prior information (not observed)} \\ x_n & \sim & P_n(dx_n|\theta) & \text{observation} \end{cases}$$

We are interested in evaluation of the integral I_n of $h \in L^1(P_n(\cdot|x_n))$:

$$I_n := I_n(x_n) = \int_\Theta h(\theta) P_n(d\theta|x_n).$$

Unfortunately it does not have a closed form in general. Monte carlo methods generalize above statistical model introducing a new randomness.

Consider the following enlarged model (algorithm):

$$\begin{cases} \theta & \sim & \Pi(d\theta) & \text{prior} \\ x_n & \sim & P_n(dx_n|\theta) & \text{observation} \\ \theta_m = (\theta^0, \ldots, \theta^{m-1}) & \sim & \text{IID } P_n(d\theta|x_n) & \text{new observation} \end{cases}$$

Approximate I_n by

$$I_{n,m} := I_{n,m}(x_n, \theta_m) = m^{-1} \sum_{i=0}^{m-1} h(\theta^i).$$

If $h \in L^1(P_n(\cdot|x_n))$, we have $I_{n,m} \to I_n$ $(m \to \infty)$. There are several methods which extends above algorithm ex. importance sampling, Markov chian Monte carlo methods, sequential Monte carlo methods. Those methods have a lot of descendants such as cross entropy methods, adaptive MCMC and more.

1. MCMC and its ergodicity (Harris rec. approach)

1.2 Markov chain Monte Carlo (MCMC)

We focus on MCMC.

$$\begin{cases} \theta & \sim & \Pi(d\theta) & \text{prior information (not observed)} \\ x_n & \sim & P_n(dx_n|\theta) & \text{observation} \\ \theta_m & \sim & \text{MCMC} & \text{new observation} \end{cases}$$

For MCMC, $\theta_m = (\theta^0, \ldots, \theta^{m-1})$ is a Markov chian with transition kernel $K_n(\theta, d\theta^*|x_n)$. Stability of Markov chain for general state space was well studied in 1970, 80's.

Notation and properties:

1. $K_n(\theta^*, d\theta|x_n)$ is the transition kernel defined by MCMC.

2. The posterior distribution $P_n(d\theta|x_n)$ is the invariant distribution of $K_n(\cdot,\cdot|x_n)$.

3. For $\theta = \theta^0$, $\mathbf{P}_\theta(\theta^1 \in A|x_n) = K_n(x, A|x_n)$. Write $K_n^m(\theta, A|x_n)$ for $\mathbf{P}_\theta(\theta^m \in A|x_n)$.

4. If $K_n(\cdot|x_n)$ is ergodic for fixed n and x_n, $I_{n,m} \to I_n$ $(m \to \infty)$ and $\|K_n^m(\theta, d\theta^*|x_n) - P_n(d\theta^*|x_n)\| \to 0$ as $m \to \infty$.

5 If $K_n(\cdot|x_n)$ is geometrically ergodic for fixed n and x_n,

$$m^{1/2}(I_{n,m} - I_n) \to N(0, \sigma^2_{h,n})$$

for $h \in L^2(P_n(\cdot|x_n))$ by reversibility.

6 For polynomial ergodic case, the same result holds for $h \in L^{2+\alpha}(P_n(\cdot|x_n))$ for some $\alpha > 0$.

There are a lot of literature for sufficient conditions for (geo. poly.) ergodicity for MCMC, ex. Tierney 94 (Independent MH), Mengersen and Tweedie 96, Roberts and Tweedie 96a (Random walk MH) and Roberts and Tweedie 96b (Langevin type MH).

1. MCMC and its ergodicity (Harris rec. approach)
1.3 Analysis of MCMC

MCMC sometimes behaves poorly even if geo ergodicity holds. For fixed n and x_n, the analysis of the complicated transition kernel is quite challenging.

For a concrete model, it may be possible to make an analysis in detail. Sometimes it is possible to obtain almost exact upper bounds of $\|K_n^m(\theta, d\theta^*|x_n) - P_n(d\theta^*|x_n)\|$. See a review [Diaconis 08] in detail. Unfortunately such an analysis is limited to very simple case.

The transition kernel $K_n(\theta, d\theta^*|x_n)$ is similar to the posterior distribution $P_n(d\theta^*|x_n)$. Although $P_n(d\theta^*|x_n)$ is complicated for fixed n and x_n, it is simple in $n \to \infty$. Under mild assumptions, we know

$$\|P_n(\cdot|x_n) - N(\hat{\theta}_n, n^{-1}I(\hat{\theta}_n)^{-1})\| \to 0.$$

See [Le Cam and Yang 00] and [Ibragimov and Hasminskii 81]. Fortunately, $K_n(\theta, d\theta^*|x_n)$ has similar asymptotics.

We only need a simple extension of the Bernstein-von Mises's theorem to study asymptotic property of $K_n(\theta, d\theta^*|x_n)$. We are going to define MCMC and its regular property.

2. Regular and non-regular convergence (Our approach).

2. Regular and non-regular convergence (Our approach).

2.1 Regular behavior

MCMC generate Marov chain realization $\theta_\infty = (\theta^0, \theta^1, \ldots)$.

$$\left\{ \begin{array}{ccc} \theta & \sim & \Pi(d\theta) \quad \text{prior information (not observed)} \\ x_n & \sim & P_n(dx_n|\theta) \quad \text{observation} \\ M_n & : & \text{MCMC} \end{array} \right.$$

MCMC M_n is a random variable $M_n : X_n \to \mathcal{P}(\Theta^\infty)$. Each $M_n(x_n)$ is a Markov measure defined by the transition kernel $K_n(\theta, d\theta^*|x_n)$ and the initial distribution.

To avoid degeneracy, we need a state space scaling

$$\theta \mapsto n^{1/2}(\theta - \widehat{\theta}_n) \in H_n = n^{1/2}(\Theta - \widehat{\theta}_n).$$

The scaled MCMC is $M_n^* : X_n \to \mathcal{P}(H_n^\infty)$.

For IID setting [K08, K10] showed that the Gibbs sampler (which is a kind of MCMC) M_n^* tends to M in distribution for mild assumptions. The random variable $M : (\mathbf{R}^p, \mathcal{B}(\mathbf{R}^p), \Pi) \to \mathcal{P}((\mathbf{R}^p)^\infty)$ is an AR process, more precisely, $M(\theta)$ is the law of

$$h_{i+1} = (I + J(\theta))^{-1} J(\theta) h_i + \epsilon_i.$$

Definition 1 *MCMC is said to be consistent if for any $m(n) \to \infty$, the empirical distribution of $\theta_{m(n)} = (\theta^0, \ldots, \theta^{m(n)-1})$ tends to $P_n(d\theta|x_n)$ in Levy metric, that is,*

$$d(P_{n,m(n)}(d\theta|x_n), P_n(d\theta|x_n)) = o_{P_n}(1)$$

where $P_{n,m}(d\theta|x_n) = m^{-1} \sum_{i=0}^{m-1} \delta_{\theta^i}$.

The scaled Gibbs sampler is consistent under mild assumptions.

2. Regular and non-regular convergence (Our approach).
2.2 Non-Regular behavior

Sometimes above consistency does not hold (degeneracy).

Definition 2 *MCMC is said to be degenerate if for some* $m(n) \to \infty$*, the empirical distribution of* $\theta_{m(n)} = (\theta^0, \ldots, \theta^{m(n)-1})$ *tends to* δ_{θ^1} *in Levy metric, that is,*

$$d(P_{n,m(n)}(d\theta|x_n), \delta_{\theta^1}) = o_{P_n}(1).$$

There are many examples which are degenerate. There are also many examples which is uniformly ergodic but degenerate. We can observe these degeneracy by numerical simulation.

Ex. Mixture model, Probit model, Linear model with mixed effect

Even for degenerate MCMC, there are relatively good and bad behaviors. We can measure those by notion of weak consistency.

Definition 3 *MCMC is said to be $\delta(n)$-weakly consistent if for any $m(n) \to \infty$ such that $m(n)/\delta(n) \to \infty$, the empirical distribution of $\theta_{m(n)} = (\theta^1, \ldots, \theta^{(m(n))})$ tends to $P_n(d\theta|x_n)$ in Levy metric, that is,*

$$d(P_{n,m(n)}(d\theta|x_n), P_n(d\theta|x_n)) = o_{P_n}(1).$$

Consistency corresponds to $\delta(n) \equiv 1$. For a mixture model, $\delta(n) = n^{1/2}$ and a simple binomial model, $\delta(n) = n$.

2. Regular and non-regular convergence (Our approach).
2.3 How to show weak consistency?

Weak consistency is proved in the following line. Set a stochastic process $S_n(t)$ $(t \geq 0)$ to be

$$S_n(t) = n^{1/2}(\theta^{[\delta(n)t]} - \widehat{\theta}_n) \ (t \geq 0).$$

Therefore $\{S_n, n = 1, 2, \ldots\}$ is a sequence of Markov processes with transition kernel depending on x_n $(n = 1, 2, \ldots)$. Write the law of S_n given x_n by $M_n^*(x_n)$. Then we can consider a convergence of M_n^*.

3. (Weak) Consistency of MCMC

3. (Weak) Consistency of MCMC
3.1 MCMC (for IID)

Gibbs sampler $G_n : X_n \to \mathcal{P}(\Theta^\infty)$ is a popular subclass of MCMC. Let $\{p(dx|\theta); \theta \in \Theta\}$ and $P(d\theta)$ be the prior. Assume

$$p(dx|\theta) = \int_{y \in Y} p(dx, dy|\theta).$$

Gibbs sampler $M(x_n) \in \mathcal{P}(\Theta^\infty)$ is a Markov measure defined by:

1. $\theta^0 \sim P(d\theta|x_n)$.

2. Generate $y_n \sim P(dy_n|x_n, \theta^0)$ and then $\theta^1 \sim P(d\theta|x_n, y_n)$. Repeat.

Example 4 *Consider*

$$p(dx|\theta) = \theta F_1(dx) + (1-\theta)F_0(dx).$$

In this case,

$$p(dx, dy|\theta) = \theta F_1(dx)\delta_1(dy) + (1-\theta)F_0(dx)\delta_0(dy).$$

$M(x_n)$ has $P(d\theta|x_n)$ as an invariant distribution. If the parametric family has certain regularity, the empirical distribution of $\theta^0, \theta^1, \ldots$ tends to $P(d\theta|x_n)$.

3. (Weak) Consistency of MCMC
3.2 Result

1. $p(dx, dy|\theta)$ is quadratic mean differentiable for any θ and $p(dxdy|\theta)$ and $p(dxdy|\vartheta)$ is equivalent for any $\theta, \vartheta \in \Theta$.

2. Fisher information matrices $I(\theta)$ and $J(\theta) := K(\theta) - I(\theta)$ are strictly positive at any θ.

3. $\theta \mapsto P(dx|\theta)$ is one to one.

4. There exists a uniformly consistent test for $P(dx|\theta)$.

5. Prior distribution is regular.

Theorem 1 *Under the above assumption, the Gibbs sampler with state space scaling is consistent.*

The Gibbs sampler is usually behaved well. However there are known Gibbs samplers which has poor behavior. Some of these can be explained by the approach.

3. (Weak) Consistency of MCMC

3.3 Mixture model

Consider

$$p(dx|\theta) = \theta F_1(dx) + (1-\theta)F_0(dx).$$

with $p(dxdy|\theta)$ defined before. When true value is $\theta_0 = 0$, G_n has a poor behavior. If F_0 and F_1 satisfies some condition, the model $p(dx|\theta)$ is quadratic mean differentiable. On the other hand, $p(dx, dy|\theta)$ is not Q.M.D.

Proposition 5 *The Gibbs sampler for the model (with scaling) is degenerate. Moreover, it is* $n^{1/2}$*-weakly consistent (*$n^{1/2}$ *times worse than regular MCMC).*

3. (Weak) Consistency of MCMC

3.4 Binomial model

Consider

$$P(y = 1|\theta, x) = G(\alpha + \beta^T x)$$

and consider extended model

$$z \sim G(\cdot + \beta^T x),\ y = 1(z \leq \alpha).$$

Then $p(dxdy|\theta)$ is almost regular but $p(dxdydz|\theta)$ is not.

Proposition 6 *The Gibbs sampler for the model (with scaling) is degenerate. Moreover, if $\beta = 0$ and known, it is n-weakly consistent (n times worse than regular MCMC).*

Weak consistency is proved in the following line. Under suitable scaling, the Gibbs sampler defines a Markov process S_n with observation dependent transition kernel $K_n(h, dh^*|x_n)$. We show the (locally uniformly) convergence of

$$\int (h^*)^k K_n(h, dh^*|x_n)$$

for $k = 1, 2, 4$. Since the model is simple, we can calculate the above integrals.

For the mixture case, G_n tends to $G(z)$, which is the law of

$$dS_t = (\alpha_1 + S_t z - S_t^2 I) + \sqrt{2S_t} dW_t. \tag{1}$$

The law of z corresponds to the (scaled) law of m.l.e.

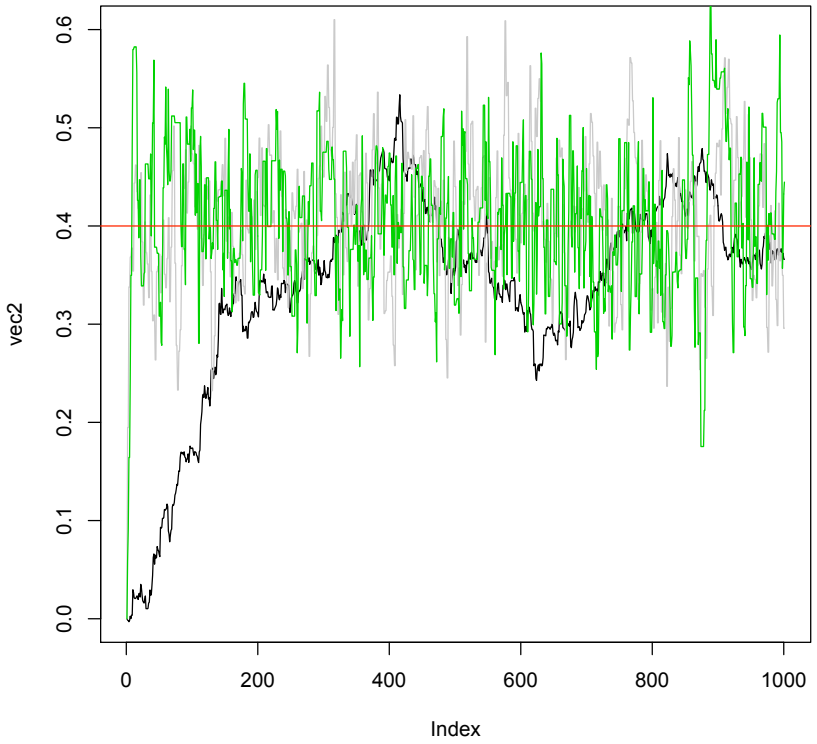
For the probit case, G_n tends to $G(\theta)$, which is the law of

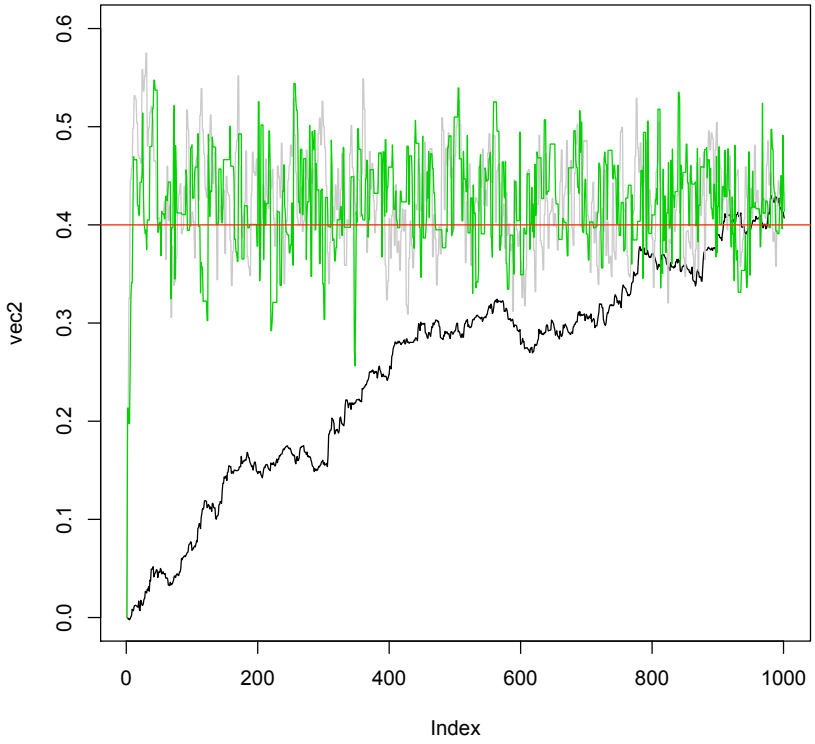
$$dS_t = -\frac{S_t}{2\theta(1-\theta)} + W_t. \tag{2}$$

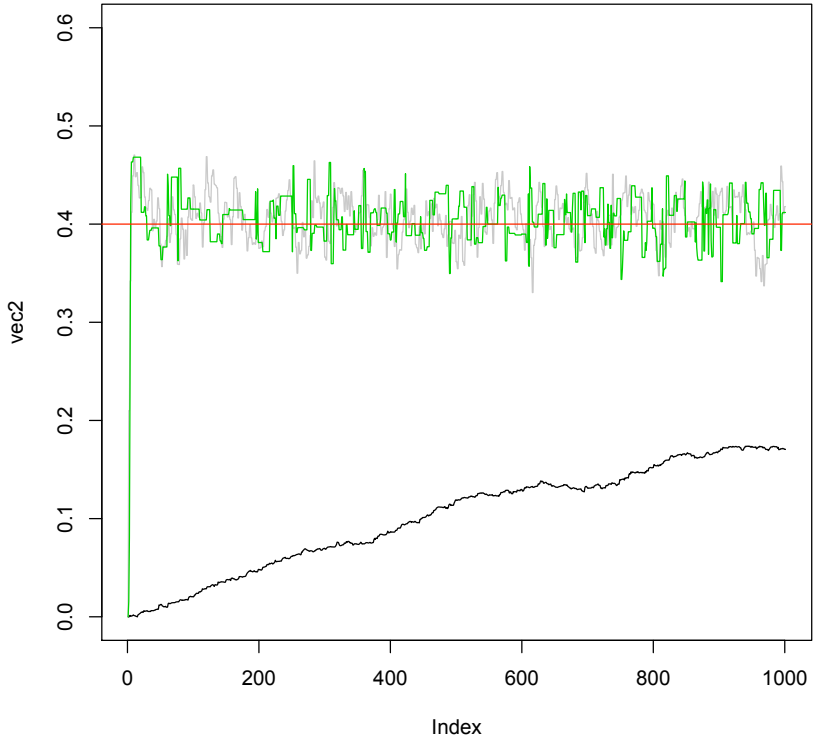
with $\theta \sim P(d\theta)$.

4. Simulation result

4.1 Trajectory of the MCMC methods for fixed n (Plot)




vec2
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
200
400
600
800
1000
Index


vec2
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
200
400
600
800
1000
Index