Output Analysis for MCMC

Galin L. Jones

School of Statistics
University of Minnesota

January 2, 2008

Acknowledgments

- Brian Caffo *Johns Hopkins University*
- Murali Haran *The Pennsylvania State University*
- Ronald Neath *Baruch College*
- James Flegal *University of Minnesota*
- Alicia Johnson *University of Minnesota*

General Setting

Let π be a probability distribution. I want the value of some feature θ. For example, θ might be a quantile, a mode, an interval, or

$$\theta = E_{\pi} g := \int_{\mathsf{X}} g(x)\pi(dx)$$

Assume that θ is analytically intractable.

Treat θ as an unknown parameter and simulate data to estimate it.

- Monte Carlo (MC, GOFMC)
- Markov chain Monte Carlo (MCMC)

Markov Chain Monte Carlo Basics

Simulate a Markov chain $X := \{X_n\}$

$X_i \sim F_i \neq \pi$ and $\text{Cov}(X_i, X_j) > 0$

Use $\hat{\theta}_n = \hat{\theta}(X_0, X_1, \ldots, X_{n-1})$ to estimate θ so that

$$\hat{\theta}_n \to \theta \quad \text{as} \quad n \to \infty$$

Simplest Case

$$\hat{\theta}_n = \bar{g}_n := \frac{1}{n} \sum_{i=0}^{n-1} g(X_i) \stackrel{a.s.}{\to} E_\pi g = \theta \;\; \text{as } n \to \infty$$

Fixed-Width Methodology

When is n large enough?

When is $\hat{\theta}_n$ a good estimate of θ?

<u>Monte Carlo Error</u>: $\hat{\theta}_n - \theta$

<u>Sampling Distribution</u>

$$\tau_n(\hat{\theta}_n - \theta) \xrightarrow{d} J \text{ as } n \to \infty$$

Simulate until

$$[\hat{\theta}_n - c_n, \hat{\theta}_n + c_n]$$

is sufficiently narrow.

Fixed-Width Methodology

Simplest Case

$$\sqrt{n}(\bar{g}_n - E_\pi g) \xrightarrow{d} \text{N}(0, \sigma_g^2) \text{ as } n \to \infty$$

where

$$\sigma_g^2 = \text{Var}_\pi[g(X_0)] + 2\sum_{i=1}^{\infty} \text{Cov}_\pi[g(X_0), g(X_i)]$$

Simulate until

$$t_* \frac{\hat{\sigma}_g}{\sqrt{n}} + a(n) \leq \text{desired half-width}$$

where t_* is an appropriate critical value and $a(n) \downarrow 0$ on $\mathbb{Z}^+$.

Questions

Old Question

- When is $\hat{\theta}_n$ a good estimate of θ?

New Questions

- When does the Monte Carlo error have a limiting distribution?
- How can we construct confidence intervals for θ?
- Will the sequential procedure terminate at a finite time?
- Will the resulting intervals have the desired coverage probability?

Regularity Conditions

$X = \{X_0, X_1, X_2, \ldots\}$ is a Markov chain

- invariant distribution is π
- π-irreducible
- aperiodic
- positive Harris recurrent

$$P^n(x, A) := \Pr(X_{i+n} \in A | X_i = x)$$

As $n \to \infty$

$$\|P^n(x, \cdot) - \pi(\cdot)\| := \sup_A |P^n(x, A) - \pi(A)| \downarrow 0$$

Regularity Conditions

Rate of TV convergence is the key:

$$\|P^n(x,\cdot) - \pi(\cdot)\| \le C(x)\gamma(n)$$

where $C : \mathsf{X} \to [0,\infty]$ and $\gamma(n) \downarrow$ on $\mathbb{Z}^+$.

- uniform/geometric ergodicity means $\gamma(n) = t^n$ for some $0 < t < 1$.
- polynomial ergodicity of order m means $\gamma(n) = 1/n^m$

There exist constructive techniques for establishing the rate of convergence.

Simplest Case

$$\theta = E_\pi g$$

$$\sqrt{n}(\bar{g}_n - E_\pi g) \xrightarrow{d} \mathrm{N}(0, \sigma_g^2) \text{ as } n \to \infty$$

where

$$\sigma_g^2 = \mathrm{Var}_\pi[g(X_0)] + 2\sum_{i=1}^{\infty} \mathrm{Cov}_\pi[g(X_0), g(X_i)]$$

Simulate until

$$t_* \frac{\hat{\sigma}_g}{\sqrt{n}} + a(n) \leq \text{desired half-width}$$

where t_* is an appropriate critical value and $a(n) \downarrow 0$ on $\mathbb{Z}^+$.

Simplest Case: CLT

Suppose at least one of the following conditions hold.

- X is uniformly ergodic and $E_\pi g^2 < \infty$
- X is geometrically ergodic, reversible and $E_\pi g^2 < \infty$
- X is geometrically ergodic and $E_\pi |g|^{2+\epsilon} < \infty$
- X is polynomially ergodic of order m and $E_\pi |g|^{2+\epsilon} < \infty$ where $m\epsilon > 2 + \epsilon$
- X is polynomially ergodic of order $m > 1$ and there exists a $B < \infty$ such that $|g(x)| \leq B$ π-almost surely.

Then for any initial distribution, as $n \to \infty$

$$\sqrt{n}(\bar{g}_n - E_\pi g) \xrightarrow{d} \mathrm{N}(0, \sigma_g^2)$$

Simplest Case: Estimating σ_g^2

$$\boxed{\sigma_g^2 = \mathrm{Var}_\pi[g(X_0)] + 2\sum_{i=1}^{\infty} \mathrm{Cov}_\pi[g(X_0), g(X_i)]}$$

Batch Means (nonoverlapping, overlapping, spaced)

Regenerative Simulation

Spectral Methods

Subsampling Bootstrap (overlapping batch means)

Time Series Bootstrap

Simplest Case: Batch Means

Split a long run $\{X_0, X_1, \ldots, X_{n-1}\}$ into b_n batches of length a_n:

$$\begin{array}{cc} X_0, \ldots, X_{a_n-1} & \bar{g}_1 = \frac{1}{a_n}\sum_{j=0}^{a_n-1} g(X_j) \\ X_{a_n}, \ldots, X_{a_2-1} & \bar{g}_2 \\ \vdots & \vdots \\ X_{a_n(b_n-1)}, \ldots, X_{n-1} & \bar{g}_{b_n} \end{array}$$

$$\hat{\sigma}^2_{BM} = \frac{a_n}{b_n - 1}\sum_{j=1}^{b_n}(\bar{g}_j - \bar{g}_n)^2$$

Theoretical results

Theorem If X is geometrically ergodic and $E_\pi|g(x)|^{2+\epsilon_1+\epsilon_2} < \infty$ for $\epsilon_1, \epsilon_2 > 0$ then $\hat{\sigma}^2_{BM} \to \sigma^2_g$ w.p. 1 as $n \to \infty$ if

1. $a_n \to \infty$ as $n \to \infty$
2. $b_n \to \infty$ and $b_n/n \to 0$ as $n \to \infty$
3. $b_n^{-1} n^{2\alpha} [\log n]^3 \to 0$ as $n \to \infty$ where $\alpha = 1/(2+\epsilon_1)$
4. there exists $c \geq 1$ such that $\sum_n (b_n/n)^c < \infty$.

Remark $a_n = b_n = \sqrt{n}$ (often) works.

General Case

$\hat{\theta}_n$ approximates θ

Sampling Distribution

$$\tau_n(\hat{\theta}_n - \theta) \xrightarrow{d} J \text{ as } n \to \infty$$

Simulate until

$$[\hat{\theta}_n - c_n, \hat{\theta}_n + c_n]$$

is sufficiently narrow.

General Case: Subsampling Bootstrap

Split a long run $\{X_0, X_1, \ldots, X_{n-1}\}$ into batches of length a_n:

$$\begin{array}{cc} X_0, \ldots, X_{a_n-1} & \hat{\theta}_1 \\ X_1, \ldots, X_{a_n} & \hat{\theta}_2 \\ \vdots & \vdots \end{array}$$

There are $n - a_n + 1$ batches of length a_n. The collection

$$\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_{n-a_n+1}$$

can be used to approximate the sampling distribution of $\hat{\theta}_n$.

Subsampling Bootstrap

<u>Theorem</u> Assume that as $n \to \infty$ $\tau_n \to \infty$ and

$$\tau_n(\hat{\theta}_n - \theta) \xrightarrow{d} J\ .$$

Let J^* be the empirical distribution function of the $\tau_{a_n}(\hat{\theta}_{a_n} - \hat{\theta}_n)$.
If X is geometrically ergodic and as $n \to \infty$

1. $a_n \to \infty$ and $a_n/n \to 0$
2. $\tau_{a_n} \to \infty$ and $\tau_{a_n}/\tau_n \to 0$

then $J^* \to J$ at every continuity point and an "asymptotically valid" $100(1-\alpha)\%$ confidence interval for θ is

$$[\hat{\theta}_n - \tau_n^{-1} {J^*}^{-1}(1-\alpha/2),\ \hat{\theta}_n - \tau_n^{-1} {J^*}^{-1}(\alpha/2)]\ .$$

Toy Example

Goal: Estimate the mean of a Pareto(α, β) ie.

$$\theta = \frac{\alpha\beta}{\beta - 1}$$

We will pretend to require MCMC and use an independence sampler with a Pareto(α, λ) candidate.
$\lambda \leq \beta \Rightarrow$ uniformly ergodic
$\lambda > \beta \Rightarrow$ not even geometrically ergodic
$\lambda > 2\beta \Rightarrow \sigma_g^2 = \infty$.

Toy Example

5000 Replications
Target half-width=.005
Nominal 95% confidence interval

Table: Estimated Coverage Probabilities

α	β	λ	BM	SS	BM25
1	4	3	.943	.948	.898
			(.003)	(.002)	(.004)
1	4	9	NA	NA	–

Baseball

Efron and Morris (1975) give a data set consisting of the raw batting averages (based on 45 official at-bats) and a transformation ($\sqrt{45}\arcsin(2x-1)$) for 18 Major League Baseball players during the 1970 season.

Suppose for $i = 1, \dots, K$ that

$$Y_i|\gamma_i \sim \mathrm{N}(\gamma_i, 1) \qquad \gamma_i|\mu,\lambda \sim \mathrm{N}(\mu,\lambda)$$
$$\lambda \sim \mathrm{IG}(2,2) \qquad f(\mu) \propto 1 \, .$$

Block Gibbs Sampler: $(\lambda', \mu', \gamma') \to (\lambda, \mu, \gamma)$

Theorem (Rosenthal,1996) The Markov chain is geometrically ergodic.

Baseball

<u>Goal</u>: Estimate the posterior *median*, θ, of γ_9, the "true" long-run (transformed) batting average of the Chicago Cubs' Ron Santo.

2000 Replications
Target half-width=.005
Nominal 95% confidence interval
Estimated Coverage Probability

SS .951 (.005)

Hierarchical Linear Model

$$
\begin{aligned}
Y|\beta, u, \lambda_R &\sim \mathrm{N}_N(X\beta + Zu, \lambda_R^{-1} I_N) \\
\beta|u &\sim \mathrm{N}_p(\beta_0, B^{-1}) \\
u|\lambda_D &\sim \mathrm{N}_n(0, \lambda_D^{-1} I_n) \\
\lambda_R &\sim \mathrm{Gamma}(r_1, r_2) \\
\lambda_D &\sim \mathrm{Gamma}(d_1, d_2)
\end{aligned}
$$

Block Gibbs Sampler: Let $\xi = (u^T, \beta^T)^T$.

$$(\lambda', \xi') \to (\lambda, \xi') \to (\lambda, \xi)$$

Theorem The Markov chain is geometrically ergodic if $d_1 > 1$.

Hierarchical Linear Model

Measure 2 subjects 5 times each at equal intervals.

$$
\begin{aligned}
Y|\beta, \lambda_R &\sim \mathrm{N}_{10}(x\beta + zu, \lambda_R^{-1} I_{10}) \\
\beta|u &\sim \mathrm{N}(0, 10) \\
u|\lambda_D &\sim \mathrm{N}_2(0, \lambda_D^{-1} I_2) \\
\lambda_R &\sim \mathrm{Gamma}(2, 2) \\
\lambda_D &\sim \mathrm{Gamma}(2, 2)
\end{aligned}
$$

2000 Replications
Target half-width=.02
Nominal 95% confidence interval for $\theta = E[\beta|y]$
Estimated Coverage Probabilities

BM .947 (.005) SS .943 (.005) BM30 .912 (.006)

Summary

- Fixed-width methodology is useful in automating MCMC but requires a strongly consistent estimator of the asymptotic variance / asymptotically valid confidence interval.
- SS is often closer to the nominal level than BM but...
- SS requires more computational effort and can be slower than BM.
- Fixed-width methods require storing the entire simulation.
- There has been no assumption of stationarity.

MLEs for Logistic-Normal

Observable data $Y = \{Y_{ij} : j = 1, \ldots, n_i; i = 1, \ldots, q\}$
Unobservable random effects $U = (U_1, \ldots, U_q)$

$$Y_{ij} | U = u \sim \text{indep Bernoulli}(\pi_{ij})$$

where

$$\log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \beta x_{ij} + u_i$$

where

$$U_1, \ldots, U_q \sim \text{i.i.d. N}(0, \sigma^2)$$

Goal: Likelihood-based inference about (β, σ)

MLEs for Logistic-Normal

Find MLE using MCEM, MCLA, MCNR, etc.

All require simulation from $\pi(u \mid y)$.

Independence sampler: one-at-a-time updates

Update component i for $i = 1, \dots, q$

Proposal: $u_i^* \sim h_i(u_i; \sigma)$

Let $u^* = (u_1, \dots, u_{i-1}, u_i^*, u_{i+1}, \dots, u_q)^T$

Theorem This sampler is uniformly ergodic.

MCEM

Goal: Estimate the Q-function

$$\theta = E[log\pi(y, u \mid \beta, \sigma) \mid y, \beta^{(t)}, \sigma^{(t)}]$$

based on simulated data in Booth and Hobert (1999, *JRSSB*).
1000 Replications
Target half-width=.05
Nominal 95% confidence interval
Estimated Coverage Probabilities

BM .942 (.007)

SS .949 (.007)

BM30 .903 (.009)