STATISTICS AND MCMC

Jeffrey S. Rosenthal

University of Toronto

jeff@math.toronto.edu

http://probability.ca/jeff/

LMS Durham Symposium

July 27, 2003

Example: Bayesian Statistics

Let $L(\mathbf{y}|\theta)$ be the likelihood function (i.e., density of data y given unknown parameters θ) of a statistical model, where $\theta \in \Theta$. (Usually $\Theta \subseteq \mathbf{R}^d$.) Let the prior density of θ be $p(\theta)$.

Then the posterior distribution of θ given $\mathbf{y}$ is

$$p(\theta|\mathbf{y}) \propto L(\mathbf{y}|\theta)\, p(\theta) \;\; \equiv \pi_u(\theta).$$

[Unnormalised target density; normalisation constant is $p(y)$.]

Thus, the "posterior mean" of any functional f is given by:

$$\pi(f) = \frac{\int_{\mathcal{X}} f(x)\pi_u(x)dx}{\int_{\mathcal{X}} \pi_u(x)dx}\,.$$

So, Bayesians are anxious (desperate?) to compute such integrals.

However, may have Θ high-dimensional, π_u complicated, direct integration difficult.

What to do?

Traditional (Old) Monte Carlo

We're given a (possibly un-normalised) density function π_u. We want to (say) estimate expectations with respect to its normalised version, π:

$$\pi(f) = \mathbf{E}_\pi[f(X)] = \frac{\int_{\mathcal{X}} f(x)\pi_u(x)dx}{\int_{\mathcal{X}} \pi_u(x)dx}\,.$$

Here $\mathcal{X}$ high-dimensional, π_u complicated, direct integration difficult.

The Monte Carlo solution:

1. Simulate i.i.d. random variables $X_1, X_2, \ldots, X_N \sim \pi(x)\,dx$.
2. Estimate $\pi(f)$ by $\hat{\pi}(f) = (1/N)\sum_{i=1}^N f(X_i)$.

Unbiased estimate, variance $= O(1/\sqrt{N})$, errors have limiting normal distribution (CLT). Good!

Problem: Step 1 not always feasible.

Other possible sampling algorithms include:

(1) "Rejection sampling": Suppose g is easy to sample from, and we can find $c < \infty$ such that $\pi_u(x) \le cg(x)$ for all $x \in \mathcal{X}$. Then:

1. Draw $Z \sim g(x)dx$, and $U \sim \text{Uniform}[0,1]$.
2. Output Z if $U \le \pi_u(Z)\,/\,c\,g(Z)$.
3. Otherwise, goto 1.

This algorithm outputs an observation from π. <u>However</u>, we need to find g such that $\sup_x \pi_u(x)/g(x) < \infty$, and such that $\pi_u(Z)\,/\,c\,g(Z)$ is often large. Difficult.

(2) "Importance sampling": Given a sample $X_1, X_2, \ldots X_n \sim g(x)\,dx$, estimate $E_\pi[f(X)]$ by

$$\hat{\pi}(f) = \frac{\sum_{i=1}^n f(X_i)w(X_i)}{\sum_{i=1}^n w(X_i)}$$

where $w(x) = \frac{\pi_u(x)}{g(x)}$. <u>However</u>, $w(x)$ may be small, and estimation errors may be large.

Markov chain Monte Carlo (MCMC)

Markov chain Monte Carlo (MCMC) is a method for using Markov chains to draw samples from π.

Surprisingly, can easily generate Markov chain $X_0, X_1, X_2, \ldots$ with stationary distribution $\pi(\cdot)$. (Reviewed later on.) Then for large B, hopefully(!) we have $\mathcal{L}(X_B) \approx \pi(\cdot)$, i.e. $\mathbf{P}[X_B \in A] \approx \pi(A)$ for all $A \subseteq X$.

Then can estimate $\pi(f)$ by either

$$\hat{\pi}(f) = (1/M) \sum_{i=1}^{M} f(X_B^{[i]})$$

(using M indep. repetitions of Markov chain), or

$$\hat{\pi}(f) = (1/M) \sum_{n=B+1}^{B+M} f(X_n)$$

(one long chain). Simple! Easy! A godsend!

Furthermore, these Markov chains only need π_u for their implementation, i.e. the normalisation constant is not needed. Even better!

History of MCMC in Statistics

MCMC algorithms first used by physicists Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (J. Chemical Physics 1953). ["Metropolis Algorithm"]

Generalised by Canadian statistician W.K. Hastings (Biometrika 1970). ["Metropolis-Hastings Algorithm"]

Markov chains also used by Geman brothers, to solve Markov random fields for Bayesian image reconstruction (IEEE 1984). ["Simulated Annealing" / "Gibbs sampler"]

Two-variable case "Data Augmentation Algorithm" proposed by Tanner and Wong (JASA 1984).

Gibbs sampler re-discovered in statistics, esp. by Gelfand and Smith (JASA, 1990), applied to various Bayesian inference models. Bayesians were ecstatic.

In the 1990s, MCMC's use in statistics exploded.

Numerous applications, to e.g.:

- medical statistics (e.g. Carlin, Gilks, Richardson, many others)
- pedigree analysis (e.g. Geyer/Thompson)
- spatial statistics (e.g. Møller, Kendall, Thönnes)
- bioinformatics (e.g. Liu)
- image reconstruction (e.g. Besag, Green, ...)
- automated learning [A.I.] (e.g. Hinton, Neal)

etc. Also many new MCMC algorithms, such as:

- Metropolis-coupled (Geyer) [$\approx$ tempering]
- transdimensional MCMC (Green)
- Metropolis-adjusted Langevin Algorithm (Roberts/Tweedie)
- perfect MCMC [using a Markov chain to obtain i.i.d. samples] (Propp/Wilson, Fill, Murdoch/Green, Møller, Kendall, ...)

Also numerous other developments, including:

- MCMC connected to traditional Markov chain theory (Tierney, JASA 1994);
- research-level book (ed. Gilks, Richardson, and Spiegelhalter 1996);
- software "BUGS" for automatic application of Gibbs sampler (Spiegelhalter, Thomas, Best, Gilks 1996–2001);
- "night of the eleven Bayesians" (three read papers and discussion in JRSSB 1993);
- near-majority of talks at ISBA and Valencia;
- mathematical research into convergence properties, optimal design, etc. (e.g. Meyn/Tweedie, Roberts, Geyer, Robert, Moulines, R.)
- nearly 500 research papers currently listed on "MCMC Preprint Server";
- interactions (finally) with physics (Sokal, ...), computer science (Jerrum/Sinclair, Dyer).

But is it popular?

Google hit counts:
"markov chain monte carlo": 33,700
"markov chain monte carlo" + "statistics": 18,400
"markov chain monte carlo" + "bayesian": 18,800
"markov chain monte carlo" + "statistics" + "bayesian": 13,900
"markov chain monte carlo" + "physics": 5,610
"markov chain monte carlo" + "computer science": 4,640
"markov chain monte carlo" + "mathematics": 7,770
"gibbs sampler": 9,070
"gibbs sampler" + "statistics": 6,250
"gibbs sampler" + "bayesian": 6,370
"gibbs sampler" + "statistics" + "bayesian": 5,150
"gibbs sampler" + "physics": 1,350
"metropolis algorithm": 7,440
"metropolis-hastings": 5,440
"hastings-metropolis": 499
"bayesian" + "bugs": 7,980
"cindy crawford": 178,000

How Does MCMC Work?

Let $\pi(\cdot)$ be a target distribution, on some state space $\mathcal{X}$ (e.g. $\mathcal{X} = \mathbf{R}^d$), that we wish to sample from.

We wish to construct a *Markov chain* on $\mathcal{X}$, with transition probabilities $P(x, dy)$, which is easily run on a computer, and which has $\pi(\cdot)$ as its stationary distribution, i.e. $\int_{x \in X} \pi(dx)\, P(x, dy) = \pi(dy)$.

[And hopefully converges rapidly to $\pi(\cdot)$ ...]

But how can we construct $P(x, dy)$?

Note: Many different algorithms are used; different ones work well in different situations. Here "work well" means easily coded and run, and converges in moderate number (not just "order") of iterations.

Reversibility

DEFN: A Markov chain is *reversible* with respect to $\pi(\cdot)$ if

$$\pi(dx)\, P(x, dy) = \pi(dy)\, P(y, dx)\,, \quad x, y \in \mathcal{X}\,.$$

FACT: If Markov chain is reversible with respect to $\pi(\cdot)$, then $\pi(\cdot)$ is stationary.

PROOF: If reversible, then

$$\int_{x \in \mathcal{X}} \pi(dx)\, P(x, dy) = \int_{x \in \mathcal{X}} \pi(dy)\, P(y, dx)$$
$$= \pi(dy) \int_{x \in \mathcal{X}} P(y, dx) = \pi(dy)\,.$$

So, suffices to make chain *reversible*.

The Metropolis-Hastings Algorithm

Suppose $\pi(\cdot)$ has a density (w.r.t. something):

$$\pi(dx) = \pi(x)\, dx\,.$$

Suppose $Q(x, \cdot)$ is some other (simple) Markov chain, also having a density:

$$Q(x, dy) = q(x, y)\, dy\,.$$

The Metropolis-Hastings algorithm proceeds as follows.

Given X_n, generate Y_{n+1} from $Q(X_n, \cdot)$.

Then, randomly, either "accept" with probability $\alpha(X_n, Y_{n+1})$ and set $X_{n+1} = Y_{n+1}$, or "reject" with probability $1 - \alpha(X_n, Y_{n+1})$ and set $X_{n+1} = X_n$, where

$$\alpha(x, y) = \min\left[1,\ \frac{\pi(y) q(y, x)}{\pi(x) q(x, y)}\right].$$

FACT: The formula for $\alpha(x, y)$ was chosen "just right", so that the resulting Markov chain $\{X_n\}$ is reversible with respect to $\pi(\cdot)$.

PROOF: Need to show

$$\pi(dx)\, P(x, dy) = \pi(dy)\, P(y, dx)\,.$$

Suffices to assume $x \neq y$ (otherwise trivial).

But for $x \neq y$,

$$\pi(dx)\, P(x, dy) = [\pi(x)\, dx]\, [q(x, y)\, \alpha(x, y)\, dy]$$
$$= \pi(x)\, q(x, y)\ \min\left[1,\ \frac{\pi(y) q(y, x)}{\pi(x) q(x, y)}\right] dx\, dy$$
$$= \min[\pi(x)\, q(x, y),\ \pi(y) q(y, x)]\, dx\, dy$$

and similarly

$$\pi(dy)\, P(y, dx) = \min[\pi(x)\, q(x, y),\ \pi(y) q(y, x)]\, dx\, dy\,.$$

Thus, the Metropolis-Hastings algorithm proposes a new state according to the proposal kernel $Q(x, \cdot)$, and then either accepts or rejects it, with just the right probabilities to make $\pi(\cdot)$ be *reversible* (and hence *stationary*).

To run this algorithm on a computer, we just need to be able to run the proposal chain $Q(x, \cdot)$ [easy, for appropriate choice of Q], and then do the accept/reject step [easy, as long as we can compute the densities at individual points]. Good!

Furthermore we need to compute only *ratios* of densities [e.g. $\pi(y) \,/\, \pi(x)$], so we don't require the *normalising constants*. Good!

But, how to choose the proposal $Q(x, \cdot)$?

Metropolis-Hastings Variations

There are many different ways of choosing the proposal density, such as:

- **Symmetric Metropolis Algorithm.** Here

  $$q(x, y) = q(y, x)$$

  The acceptance probability simplifies to

  $$\alpha(x, y) = \min\left[1,\ \frac{\pi(y)}{\pi(x)}\right]$$

- **Symmetric random walk Metropolis.**

  $$q(x, y) = q(y - x)$$

  [e.g. $Q(x, \cdot) = N(x, \delta^2)$, or $Q(x, \cdot) = \text{Uniform}(x - \delta,\ x + \delta)$, etc.]

- **Independence sampler.** Here

  $$q(x, y) = q(y)\,,$$

  i.e. $Q(x, \cdot)$ does not depend on x.

  [Similar to "rejection sampler" ... but not identical.]

- **Langevin algorithm.**

  Here the proposal is generated by

  $$Y_{n+1} \sim N(X_n + (\delta/2)\, \nabla \log \pi(X_n),\ \delta)\,,$$

  for some $\delta > 0$.

  (Motivated by discrete approximation to a "Langevin diffusion" processes.)

[How to choose scaling δ? Saturday!]

The Gibbs Sampler

[a.k.a. "heat bath" or "Glauber dynamics"]

Suppose that $\pi(\cdot)$ is d-dimensional, i.e. $\mathcal{X} \subseteq \mathbf{R}^d$.

Write $\mathbf{x} = (x_1, \ldots, x_d)$, and
$\mathbf{x}^{(-i)} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$.

Let $\pi_i(\mathbf{y} \,|\, \mathbf{x}^{(-i)})$ be the conditional density of $\pi(\cdot)$, conditional on knowing that $y_j = x_j$ for $j \neq i$:

$$\pi_i(\mathbf{y} \,|\, \mathbf{x}^{(-i)}) \;=\; \frac{g_{i,x}(\mathbf{y})}{\int g_{i,x}(\mathbf{z})\, dz_i}\,,$$

where $g_{i,x}(\mathbf{y}) = \pi(\mathbf{y}) \mathbf{1}_{\{y_j = x_j \text{ for } j \neq i\}}$.

The i^{th} component Gibbs sampler is defined by

$$P_i(\mathbf{x}, d\mathbf{y}) = \pi_i(\mathbf{y} \,|\, \mathbf{x}^{(-i)})\, dy_i\,.$$

That is, P_i leaves all components besides i unchanged, and replaces the i^{th} component by a draw from the full conditional distribution of $\pi(\cdot)$ conditional on all the other components.

FACT: The i^{th} component Gibbs sampler, P_i, is reversible with respect to $\pi(\cdot)$.

(This follows from the definition of conditional density. In fact, P_i is a special case of a Metropolis-Hastings algorithm, with $\alpha \equiv 1$.)

So, P_i leaves $\pi(\cdot)$ invariant. We then construct the Gibbs sampler out of P_i, as follows:

- The deterministic-scan Gibbs sampler is

$$P \;=\; P_1 P_2 \ldots P_d\,.$$

That is, it does the d different Gibbs sampler components, in order.

- The random-scan Gibbs sampler is

$$P \;=\; \frac{1}{d} \sum_{i=1}^{d} P_i\,.$$

That is, it does one of the d different Gibbs sampler components, chosen uniformly at random.

Either version produces a "zig-zag pattern".

Example: Variance Components Model

MODEL:

$$\begin{array}{ccccccc}
 & & \mu & & & & \\
 & \swarrow & \downarrow & \searrow & & & \\
\theta_1 & \ldots & \ldots & \theta_K & & & \theta_i \sim N(\mu, \sigma_\theta^2) \\
\downarrow & & & \downarrow & & & \\
Y_{11}, \ldots, Y_{1J} & & & Y_{K1}, \ldots, Y_{KJ} & & & Y_{ij} \sim N(\theta_i, \sigma_e^2)
\end{array}$$

PRIORS: $\sigma_\theta^2 \sim IG(a_1, b_1)$; $\quad \sigma_e^2 \sim IG(a_2, b_2)$;
$\mu \sim N(\mu_0, \sigma_0^2)$.

OBSERVED DATA: $Y_{ij} \quad (1 \leq i \leq K,\, 1 \leq j \leq J)$

TARGET DISTRIBUTION:
$\pi(\cdot) = \mathcal{L}(\sigma_\theta^2, \sigma_e^2, \mu, \theta_1, \ldots, \theta_K \,|\, \{Y_{ij}\})$.

How to sample from $\pi(\cdot)$???

Let $\pi_u : \mathbf{R}^{K+3} \to [0, \infty)$ be the (unnormalised) density for $\pi(\cdot)$.

Then taking into account the factors for all the model's various normal (N) and inverse-gamma (IG) dependencies, we obtain

$$\pi_u(\sigma_\theta^2, \sigma_e^2, \mu, \theta_1, \ldots, \theta_K) \propto$$
$$e^{-b_1/\sigma_\theta^2} \sigma_\theta^{2^{-a_1-1}} e^{-b_2/\sigma_e^2} \sigma_e^{2^{-a_2-1}} e^{-(\mu-\mu_0)^2/2\sigma_0^2}$$
$$\times \prod_{i=1}^{K} [e^{-(\theta_i-\mu)^2/2\sigma_\theta^2}/\sigma_\theta] \times \prod_{i=1}^{K} \prod_{j=1}^{J} [e^{-(Y_{ij}-\theta_i)^2/2\sigma_e^2}/\sigma_e]\,.$$

(Here $\{Y_{ij}\}$ are observed data – fixed.)

Messy! [Though simpler than many posterior distributions! Also, π_u is positive throughout $(0, \infty)^2 \times \mathbf{R}^{K+1}$, larger in "center", smaller in "tails"; typical.]

How to sample from π_u??

For Gibbs sampler, need full conditionals ...

Example (continued)

$$\mathcal{L}(\sigma_\theta^2 \mid \mu, \sigma_e^2, \theta_1, \ldots, \theta_K, Y_{ij}) \; = \; IG\left(a_1 + \tfrac{1}{2}K,\; b_1 + \tfrac{1}{2}\sum_i (\theta_i - \mu)^2\right) ;$$

$$\mathcal{L}(\sigma_e^2 \mid \mu, \sigma_\theta^2, \theta_1, \ldots, \theta_K, Y_{ij}) \; = \; IG\left(a_2 + \tfrac{1}{2}KJ,\; b_2 + \tfrac{1}{2}\sum_{i,j} (Y_{ij} - \theta_i)^2\right) ;$$

$$\mathcal{L}(\mu \mid \sigma_\theta^2, \sigma_e^2, \theta_1, \ldots, \theta_K, Y_{ij}) \; = \; N\left(\frac{\sigma_\theta^2 \mu_0 + \sigma_0^2 \sum_i \theta_i}{\sigma_\theta^2 + K\sigma_0^2},\; \frac{\sigma_\theta^2 \sigma_0^2}{\sigma_\theta^2 + K\sigma_0^2}\right) ;$$

$$\mathcal{L}(\theta_i \mid \mu, \sigma_\theta^2, \sigma_e^2, \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_K, Y_{ij}) \; = \; N\left(\tfrac{J\sigma_\theta^2 \overline{Y}_i + \sigma_e^2 \mu}{J\sigma_\theta^2 + \sigma_e^2},\; \tfrac{\sigma_\theta^2 \sigma_e^2}{J\sigma_\theta^2 + \sigma_e^2}\right) .$$

The Gibbs sampler proceeds by updating the $K + 3$ variables, in turn (either deterministic or random scan), according to the above conditional distributions.

This is feasible since the conditional distributions are all easily simulated (IG and N).

In fact, it works well! [Gelfand and Smith, 1990]

(Gibbs samplers commonly used for this model and variants. Also studied theoretically by e.g. Mykland, Tierney, Yu 1994; R. 1994.)

Alternatively, can run Metropolis algorithm for this model.

Let $\pi_u : \mathbf{R}^{K+3} \to [0, \infty)$ be the unnormalised density for $\pi(\cdot)$, as above.

Proceed, given X_n, by:

- Choose $Y_{n+1} \sim N(X_n, \sigma^2 I_{K+3})$ (say);
- Choose $U_{n+1} \sim$ Uniform$[0, 1]$.
- If $U_{n+1} < \pi_u(Y_{n+1}) \,/\, \pi_u(X_n)$, then set $X_{n+1} = Y_{n+1}$ (accept). Otherwise set $X_{n+1} = X_n$ (reject).

Also works well, if σ^2 chosen well. [Saturday!]

General Theory

So, now we know how to construct (and run) lots of different MCMC algorithms. Good!

But do they converge to the distribution $\pi(\cdot)$? How quickly?

Write $P^n(x, A)$ for the n-step transition law of the Markov chain:

$$P^n(x, A) = \mathbf{P}(X_n \in A \,|\, X_0 = x) .$$

Big questions are, is $P^n(x, \cdot)$ close to $\pi(\cdot)$ as $n \to \infty$? How large does n need to be?

[Ideally want actual number, e.g. "$n = 140$", not just "polynomially bounded" or even "$O(d^3)$ …]

Recall some standard definitions:

DEFN: A chain is ϕ-irreducible if there exists a non-zero measure ϕ on $\mathcal{X}$ such that for all $A \subseteq \mathcal{X}$ with $\phi(A) > 0$, and for all $x \in \mathcal{X}$, there exists a positive integer $n = n(x)$ such that $P^n(x, A) > 0$.

e.g. if $\phi(A) = \delta_{x_*}(A)$, then this requires that x_* is accessible from any state x; for a continuous Markov chain, $\phi(\cdot)$ might instead be e.g. Lebesgue measure.

DEFN: The chain is aperiodic if there do not exist $d \geq 2$ and disjoint subsets $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_d \subseteq \mathcal{X}$ with $\pi(\mathcal{X}_i) > 0$, such that $P(x, \mathcal{X}_{i+1}) = 1$ for all $x \in \mathcal{X}_i$ $(1 \leq i \leq d-1)$, and $P(x, \mathcal{X}_1) = 1$ for all $x \in \mathcal{X}_d$.

DEFN: The total variation distance between two probability measures $\nu_1(\cdot)$ and $\nu_2(\cdot)$ is:

$$\|\nu_1(\cdot) - \nu_2(\cdot)\| = \sup_A |\nu_1(A) - \nu_2(A)|\,.$$

Standard M.C. Convergence Theorem

[Doeblin, Orey, Jain/Jameson, Athreya/Ney, Nummelin, . . . ; then Tierney.]

THEOREM: If a Markov chain is ϕ-irreducible and aperiodic, and has a stationary distribution $\pi(\cdot)$, then for π-a.e. $x = X_0 \in \mathcal{X}$,

$$\lim_{n\to\infty} \|P^n(x,\cdot) - \pi(\cdot)\| \;=\; 0\,.$$

In particular,

$$\lim_{n\to\infty} P^n(x, A) \;=\; \pi(A)\,, \quad A \subseteq \mathcal{X}\,.$$

Furthermore, if $h : \mathcal{X} \to \mathbf{R}$ with $\pi(|h|) < \infty$,

$$\lim_{n\to\infty} (1/n) \sum_{i=1}^{n} h(X_i) \;=\; \pi(h) \quad w.p.\ 1$$

Also, "usually" have a central limit theorem:

$$n^{-1/2} \sum_{i=1}^{n} [h(\mathbf{X}_i) - \pi(h)] \;\Rightarrow\; N(0, \sigma^2)$$

for some $\sigma^2 > 0$ [more later].

So, if chain is ϕ-irreducible and aperiodic, and has stationary distribution $\pi(\cdot)$, then it will converge in distribution to $\pi(\cdot)$ from π-a.e. starting value. Good!

Now, in MCMC, always start with $\pi(\cdot)$ stationary. Good.

Furthermore, usually easy to verify that chain is ϕ-irreducible, where e.g. ϕ is Lebesgue measure on appropriate region. Good.

Also, aperiodicity almost always holds, e.g. for virtually any Metropolis algorithm or Gibbs sampler. Good.

So, the theorem guarantees that "most" MCMC algorithms will asymptotically converge to $\pi(\cdot)$. Good.

But questions remain . . .

Time to Stationarity

So now we have a Markov chain, and we know $\mathcal{L}(X_n) \to \pi(\cdot)$. How large should B be, so that $\mathcal{L}(X_B) \approx \pi(\cdot)$? ["Burn-in time."]

Ideally, can prove that $\|\mathcal{L}(X_B) - \pi(\cdot)\| < \epsilon$ for appropriate B. A few of us have ~~wasted~~ spent our lives on this question, with some successes for various non-trivial MCMC algorithms. But for complicated Markov chains, it is very difficult and time-consuming.

Instead, practitioners use "convergence diagnostics", i.e. do statistical analysis of the realised output $X_1, X_2, \ldots, X_B$, to see if the values "seem stable".

Typically, do multiple chain runs from different starting values ["overdispersed starting distribution"], and see if they all converge to approximately the same distribution. (e.g. Gelman/Rubin, 1992)

No guarantees, though!

Example: "Witch's Hat"

$\mathcal{X} = [0,1]^d$ (d large)

$\pi_u(\mathbf{x}) = 1 + \delta^{-d+1} \mathbf{1}_S(\mathbf{x})$, where $\delta > 0$ very small, and

$$S = \{\mathbf{x} \in \mathcal{X} : \; x_i < \delta \; \forall i\} .$$

Then $\pi(S) \approx 1$.

However, unless $X_0 \in S$, or "get lucky" and find $X_n \in S$, then Gibbs Sampler or Metropolis algorithm may well miss S entirely.

Convergence diagnostics would suggest $\pi(\cdot) \approx \text{Uniform}(\mathcal{X})$. Wrong!!

Chain converges extremely slowly, but is still "geometrically ergodic". Misleading!!

Overall, the "convergence time problem" remains largely unresolved ... but usually okay in practice. [This is the motivation for perfect MCMC.]

Qualitative Convergence

DEFN: Say the chain is geometrically ergodic if

$$\|P^n(x,\cdot) - \pi(\cdot)\| \;\leq\; C(x)\,\rho^n\,, \qquad n = 1,2,3,\ldots$$

for some $\rho < 1$, where $C(x) < \infty$ for π-a.e. $x \in \mathcal{X}$.

[Always holds if state space is finite.]

[Other "qualitative rates" also possible.]

If chain is geometrically ergodic, then "probably converges quickly". Good (?).

Also, have CLT whenever $\pi(|f|^{2+\epsilon}) < \infty$: good for estimating distribution of error.

[FACT: If chain reversible and geometrically ergodic, then have CLT whenever $\pi(|f|^2) < \infty$. What if not reversible?? Open question!]

But does it matter??

Example #1: RWM for Cauchy

Let $\mathcal{X} = \mathbf{R}$, and let $\pi_1(x) = 1/(1+x^2)$ be (unnormalised) Cauchy distribution.

Then RWM for π_1 (with, say, $X_0 = 0$ and $Q(x,\cdot) = \text{Uniform}[x-1, x+1]$) is ergodic but not geometrically ergodic.

Now let $\pi_2(x) = \pi_1(x)\,\mathbf{1}(|x| < 10^{100})$.

Then RWM for π_2 (again with, say, $X_0 = 0$ and $Q(x,\cdot) = \text{Uniform}[x-1, x+1]$) is geometrically ergodic.

And yet, RWM on π_1 or on π_2 is indistinguishable when run on any physical computer for any remotely feasible amount of time.

[Similarly, "witch's hat" example is geometrically ergodic but still converges very poorly.]

So, geometric ergodicity doesn't matter??

Example #2: Independence sampler

Let $\pi(x) = e^{-x}$, with proposal $q(x,y) = ke^{-ky}$ for some $k > 0$.

We consider two possible choices:

1. $k = 0.01$ (geometrically ergodic, CLT)
2. $k = 5$ (not geometrically ergodic, no CLT)

Both algorithms were run for one million iterations started at the mean value of π (i.e. $X_0 = 1$). The experiment was repeated 55 times for each case producing the following results.

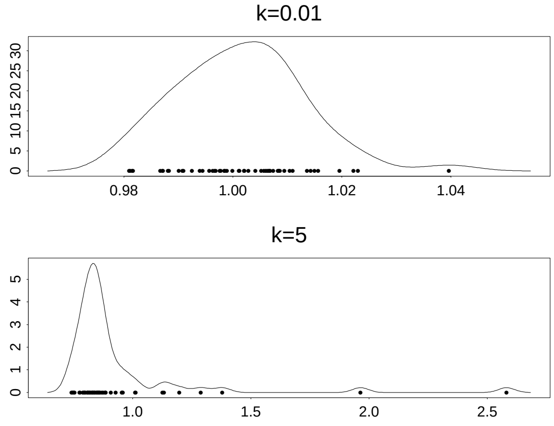

Figure 1: Sample means and kernel density estimators of true mean (1.0). Case $k = 0.01$ has small, symmetric error. Case $k = 5$ has large, skewed error.

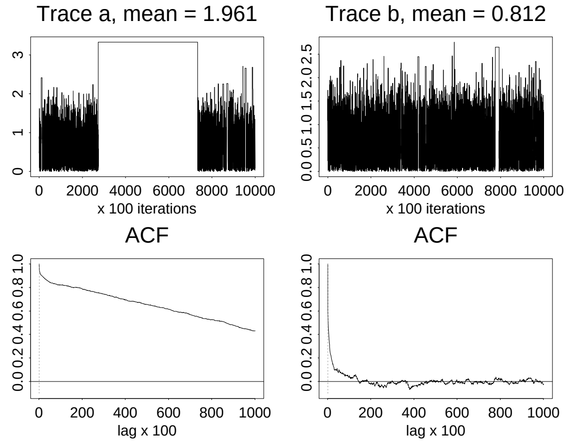

Figure 2: Two different sample paths from the $k = 5$ simulation study.

How to establish geometric ergodicity?

DEFN: A subset $C \subseteq \mathcal{X}$ is **small** if there exists a positive integer n_0, $\epsilon > 0$, and a probability ν such that the following **minorisation condition** holds ["overlap"]:

$$P^{n_0}(x, A) \geq \epsilon \nu(A) \,, \qquad x \in C, \quad A \subseteq \mathcal{X} \,.$$

THEOREM: If chain ϕ-irreducible and aperiodic with invariant measure π, and there exists a small set C, and constants $0 < \lambda < 1$, $b < \infty$ and a π-a.e. finite function $V : \mathcal{X} \to [1, \infty]$, with **drift condition**

$$\int_{\mathcal{X}} P(x, dy) V(y) \leq \lambda V(x) + b \mathbf{1}_C(x) \,,$$

then chain is geometrically ergodic.

Idea: Drift condition forces many returns to C. Then minorisation condition gives probability ϵ of "forgetting past" each time chain is in C.

Minorisation and drift conditions also used to bound time to stationarity [Meyn/Tweedie, Rosenthal, Roberts/Tweedie, ...]. (Take $n_0 = 1$.)

Imagine running two copies $\{X_n\}$ and $\{X'_n\}$.

Start with $X_0 = x$ and $X'_0 \sim \pi(\cdot)$.

Given X_n and X'_n, choose X_{n+1} and X'_{n+1} by:

1. If $X_n = X'_n$, choose $X_{n+1} = X'_{n+1} \sim P(X_n, \cdot)$.
2. Else, if $(X_n, X'_n) \in C \times C$, then:
   (a) w.p. ϵ, choose $X_{n+1} = X'_{n+1} \sim \nu(\cdot)$;
   (b) else, w.p. $1 - \epsilon$, independently choose
$$X_{n+1} \sim \frac{1}{1-\epsilon}[P(X_n, \cdot) - \epsilon\, \nu(\cdot)] \,,$$
$$X'_{n+1} \sim \frac{1}{1-\epsilon}[P(X'_n, \cdot) - \epsilon\, \nu(\cdot)] \,.$$
3. Else, just independently choose $X_{n+1} \sim P(X_n, \cdot)$ and $X'_{n+1} \sim P(X'_n, \cdot)$.

Coupling inequality says
$\|P^n(x,\cdot) - \pi(\cdot)\| \le \mathbf{P}[X_n \ne X'_n]$.

Can use this to explicitly bound $\|P^n(x,\cdot) - \pi(\cdot)\|$ in terms of ϵ, λ, b, and $\sup_{x \in C} PV(x)$. [R., JASA 1995]

Idea: Drift condition forces many returns to $C \times C$. Then minorisation condition gives probability ϵ of coupling, each time $(X_n, X'_n) \in C \times C$.

For Gibbs sampler on version of variance components model, with published data ($K = 15$), for appropriate x, get $\|P^{140}(x,\cdot) - \pi(\cdot)\| < 0.01$. [R., Stat. and Comput. 1996]

Good! Explicit!

But in general, difficult to obtain explicit bounds in complicated models.

Harris Recurrence

Why just "from π-a.e. starting value"?

EXAMPLE:

Let P be any ϕ-irreducible, aperiodic Markov chain on $\mathcal{X} = \mathbf{R}$, with densities $P(x, dy) = p(x,y)\, dy$, and with continuous stationary distribution $\pi(\cdot)$.

Let P' be defined as follows. Let $P'(x,\cdot) = P(x,\cdot)$ whenever x is not a positive integer. For x a positive integer, let

$$P'(x,\cdot) = (1/x^2)\pi(\cdot) + (1 - 1/x^2)\, \delta_{x+1}(\cdot)\,.$$

Then $\pi(\cdot)$ is stationary for P', and P' is still ϕ-irreducible and aperiodic.

But if $X_0 = 3$ (say), then could have $X_n = n + 3$ for all n, so that $\|\mathcal{L}(X_n) - \pi(\cdot)\| \not\to 0$.

[Not "Harris recurrent".]

DEFN: Say a chain is Harris recurrent if for all $B \subseteq \mathcal{X}$ with $\pi(B) > 0$, and all $x \in \mathcal{X}$,

$$\mathbf{P}[\exists n;\ X_n \in B \,|\, X_0 = x] = 1\,.$$

(Stronger than π-irreducibility.)

THEOREM: If chain Harris recurrent, then convergence theorem holds from every starting point (not just π-a.e. starting point).

For example, this always holds if $P(x, dy) = p(x,y)\, \pi(dy)$, or for any Metropolis algorithm with a π-irreducible proposal.

Virtually all MCMC algorithms used in practice are Harris recurrent; the issue has received a lot of [too much?] attention in statistics.

Summary

- Markov chains used very often in statistics (esp. Bayesian) to approximately sample from complicated distributions $\pi(\cdot)$.
- Applications to many applied areas.
- MCMC algorithms very easily constructed (Metropolis-Hastings, Gibbs sampler).
- Asymptotic convergence usually guaranteed (by ϕ-irreducibility and aperiodicity).
- "Time to stationarity" a big issue; can sometime prove bounds, otherwise use diagnostics, or just hope (usually okay).
- CLT's used to estimate error distributions.
- Qualitative convergence rates (esp. geometric ergodicity) often important; established by minorisation and drift conditions.