The Metropolitan-Hastings Algorithm and Extensions

S. Sawyer — Washington University — October 12, 2006

1. Introduction. Suppose that we want to calculate the average value of a function \( F(x) \) with respect to a probability density \( \pi(x) \), specifically

\[
E(F) = \int_X F(x)\pi(x)dx
\]

(1.1)

where \( \pi(x) \) and \( F(x) \) are defined on \( X \). Here \( E(F) \) could also be a sum if \( dx \) is a counting measure.

We assume that we do not know the density \( \pi(x) \) exactly, but that we can calculate \( \pi(x) \) within the normalizing constant. That is,

\[
\pi(x) = Cg(x)
\]

(1.2)

where \( g(x) \) is known or easy to compute, but \( C \) is unknown. An important special case is that of a Bayesian posterior distribution, for which \( g(x) \) is a prior distribution times a likelihood function. In many cases \( g(x) \) is easy to write down but the normalizing constant\n
\[
C = \frac{1}{\int_X g(x) \, dx}
\]

(1.3)

is too complex to compute easily. (See Section 5 below.) The “Metropolis” in the Metropolitan-Hastings (MH) algorithm refers to a paper in the Journal of Chemical Physics (Metropolis et al. 1953). In this paper, \( C \) was related to a partition function in statistical physics.

2. The Metropolis-Hastings Algorithm. Metropolis’ idea is to start with a Markov chain \( X_n \) on the state space \( X \) with a fairly arbitrary Markov transition density \( q(x, y) \) and then modify it to define a Markov chain \( X_n^* \) that has \( \pi(x) \) as a stationary measure. By definition, \( q(x, y) \) is a Markov transition density if \( q(x, y) \geq 0 \) and \( \int_{y \in X} q(x, y) \, dy = 1 \). If the transformed random walk \( X_n^* \) is irreducible and positive recurrent on \( X \), then

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} F(X_k^*) = \int_X F(x)\pi(x)dx \quad \text{a.s.}
\]

(2.1)

by the Birkhoff ergodic theorem for all \( F(x) \) with \( \int_X |F(x)|\pi(x)dx < \infty \). This is the basic idea of what are now known as Markov chain Monte Carlo (MCMC) methods.

The transformed Markov chain \( X_n^* \) is defined as follows. First, given \( X_n = X_n^* = x \), define an “acceptance function” \( a(x, y) \) that satisfies \( 0 \leq a(x, y) \leq 1 \). Given \( x \), the Markov transition density \( q(x, y) \) for \( X_n \) is used to “propose” a new state \( y \). With probability \( a(x, y) \), the proposed state is accepted and \( X_{n+1}^* = y \). Otherwise, the proposal is rejected and \( X_{n+1}^* = X_n^* = x \). (That is, the process stays at the same position for one time step.) Given \( X_{n+1}^* = z \) (where \( z \) is \( x \) or \( y \)), the process is repeated to generate \( X_{n+2}^* \), and so forth. In particular, the transformed Markov chain \( X_n^* \) has the transition density

\[
p(x, y) = a(x, y)q(x, y) + A(x)\delta_x(dy)
\]

where

\[A(x) = 1 - \int_{z \in X} a(x, z)q(x, z) \, dz\]

(2.2)
Here $\delta_x(dy)$ represents a Dirac measure that puts mass one at the point $y = x$ and is otherwise zero and $A(x)$ is the probability that the proposed state is not accepted.

In other words, the Markov chain $X_n^*$ is the same as $X_n$ except for the introduction of “wait states” with $X_{n+1}^* = X_n^* = x$ with probability depending on $a(x, y)$. Hastings (1970) proved that the transition density $p(x, y)$ in (2.2) has the measure $\pi(x)$ in (1.1) as a stationary distribution if

$$a(x, y) = \min \left\{ 1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)} \right\} = \min \left\{ 1, \frac{q(y, x)/\pi(x)}{q(x, y)/\pi(y)} \right\}$$

(2.3)

where $\pi(x) = C g(x)$ as in (1.2). (Of course, $p(x, y)$ is not really a transition density for $x, y \in R^n$ because of the delta function in (2.2). We give a better description of the algorithm below.)

The original Metropolis (1953) algorithm assumed $q(x, y) = q(y, x)$, for which the acceptance function has the simpler form

$$a(x, y) = \min \left\{ 1, \frac{\pi(y)}{\pi(x)} \right\} = \min \left\{ 1, \frac{g(y)}{g(x)} \right\}$$

In this case, the algorithm accepts the proposed new value $y$ with probability one if $g(y) \geq g(x)$ and accepts it with probability $g(y)/g(x)$ if $g(y) < g(x)$.

3. A Question of Notation. In Sections 1 and 2, the implicit measure “$dx$” can be the usual volume measure (Lebesgue measure) in $X$, counting measure that assigns measure one to each point if $X$ is discrete, or surface measure within a subspace of $X$. In many descriptions of the Metropolis-Hastings algorithm, the precise meaning of the implicit measure $dx$ is meant to be understood and can vary from paragraph to paragraph and even from term to term in the same equation. This lack of a good notation can make proofs difficult to follow even when the proofs are basically correct. For this reason, we will adopt a standard notation for measures from probability theory and write expressions like those in (1.1)–(1.3) in the form

$$E(F) = \int_X F(x)\pi(dx)$$

(3.1)

$$\pi(dx) = C g(dx) \quad \text{where} \quad C = 1/\int_X g(dx)$$

(3.2)

We can also write (2.2) in the form

$$p(x, dy) = a(x, y)q(x, dy) + A(x)\delta_x(dy),$$

(3.3)

$$A(x) = 1 - \int_{z \in X} a(x, z)q(x, dz)$$

which now makes better sense since we do not have to pretend that $\delta_x(dy)$ is a function on $X$. (Technically speaking, we assume that all measures on $X$ are $\sigma$-finite and second countable, which is a class that includes all measures that arise in practice.)

In (3.1)–(3.3), $\pi(dx)$ and $g(dx)$ represent arbitrary nonnegative measures and $q(x, dy)$ represents an arbitrary Markov transition function on $X$. By definition, $q(x, dy)$ is a Markov transition function if (i) $q(x, dy)$ is a nonnegative measure for each $x \in X$, (ii) $\int_{y \in X} q(x, dy) = 1$ for each $x \in X$, and (iii) $\int_{y \in X} q(x, dy)$ is a measurable function of $x$ for each measurable set $A \subseteq X$. If the measures $\pi(dx)$, $g(dx)$, and $q(x, y)$ have densities with respect to a measure $dx$ on $X$, then

$$\pi(dx) = \pi_1(x)dx, \quad g(dx) = g_1(x)dx, \quad q(x, dy) = q_1(x, y)dy$$

(3.4)
where \( \pi_1(x), g_1(x), \) and \( q_1(x,y) \) are measurable functions on \( X \) or \( X \times X \). However, the larger family of transition functions \( q(x,dy) \) includes not only examples of ergodic Markov chains on \( X \) that cannot be expressed as \( q(x,dy) \) for Lebesgue measure \( dy \) on \( X \), but also Markov chains on linear or nonlinear subspaces of \( X \) without notational circumlocutions. We will see examples of this below.

### 4. Reversible Transition Functions and the “Detailed Balance Condition”.

The conclusion of the Metropolis-Hastings theorem, which makes equation (2.1) and MCMC algorithms possible, is that the transition function \( p(x,dy) \) in (3.3) has \( \pi(dx) \) as a stationary measure. This condition can be written
\[
\int_{x \in X} \pi(dx)p(x,dy) = \pi(dy)
\]  
(4.1)

An equivalent form of (4.1) is
\[
\int_{x \in X} \int_{y \in Y} \pi(dx)p(x,dy) = \int_{y \in Y} \pi(dy)
\]  
(4.2)

for all measurable functions \( f(y) \geq 0 \). Condition (4.2) is often easier to work with than (4.1). Recall that \( \int f(y)\pi(dy) = \infty \) is allowed in measure theory if \( f(y) \geq 0 \) and \( \pi(dy) \geq 0 \) is a nonnegative measure. That is, most theorems in measure theory are valid if the real numbers are extended in this way. Thus there is no need to require that the integrals in (4.2) are finite.

*Exercise: Prove that (4.2) for all measurable \( f(y) \geq 0 \) is equivalent to (4.1).*

Since
\[
\int\int \pi(dx)p(x,dy) = \int_{x \in X} \int_{y \in Y} \pi(dx)p(x,dy) = 1
\]
whenever \( \pi(dx) \) is a probability measure on \( X \) and \( p(x,dy) \) a Markov transition function on \( X \times X \), the measure \( \pi(dx)p(x,dy) \) is a probability measure on the product space \( X \times X \). Thus one can define random variables \( X \) and \( Y \) with joint distribution
\[
P(X \in dx, Y \in dy) = \pi(dx)p(x,dy)
\]  
(4.3)

The stationarity condition (4.2) is then equivalent to
\[
E(f(X)) = E(f(Y))
\]  
(4.4)
for all measurable functions \( f(y) \geq 0 \).

In most Markov Chain Monte Carlo examples where stationarity (4.1) holds, the relation
\[
\pi(dx)p(x,dy) = \pi(dy)p(y,dx)
\]  
(4.5)
also holds (Tierney 1994). The relation (4.5) is sometimes called the “detailed balance condition” (Chib and Greenberg 1995, Liu 2001). It is equivalent to
\[
E(f(X,Y)) = E(f(Y,X))
\]  
(4.6)
for all measurable functions \( f(x,y) \geq 0 \) on \( X \times X \) for the random variables \( X,Y \) in (4.3).

*Exercise: Prove that (4.5) for all measurable \( f(x,y) \geq 0 \) is equivalent to (4.4).*

Since (4.3) can be viewed as the probability that \( X \) starts at \( x \) and then goes to \( Y = y \), and the right-hand side of (4.5) can be viewed as the probability that \( X \) starts at \( y \) and then goes to \( X = X \), the relation (4.5) is usually described as saying that the transition function \( p(x,dy) \) is time reversible (or just reversible) with respect to the probability measure \( \pi(dx) \). It follows from (4.5) that
\[
\int_{x \in X} \pi(dx)p(x,dy) = \int_{x \in X} \pi(dy)p(y,dx) = \pi(dy)
\]
This implies that time reversibility (4.5) implies stationarity (4.1), although the reverse may not hold. (Exercise: Construct a proof that reversibility implies stationarity using only the equivalent conditions (4.4) and (4.6).)
5. The Metropolis-Hastings Theorem for Global Densities. As before, let \( \pi(dx) \geq 0 \) be a nonnegative measure and \( q(x, dy) \) a Markov transition function on a space \( X \). Set

\[
p(x, dy) = a(x, y)q(x, dy) + A(x)\delta_x(dy),
\]

\[
A(x) = 1 - \int_{z \in X} a(x, z)q(x, dz)
\]

where \( 0 \leq a(x, y) \leq 1 \). As before, given \( X_n = x \), this corresponds to choosing a proposal value \( y \) according to \( q(x, dy) \) and then moving to \( y \) \((X_{n+1} = y)\) with probability \( a(x, y) \). With probability \( 1 - a(x, y) \), we stay at the same location \((X_{n+1} = X_n = x)\).

**Lemma 5.1.** For \( p(x, dy) \) in (5.1), the reversibility relation

\[
\pi(dx)p(x, dy) = \pi(dy)p(y, dx) \tag{5.2}
\]

holds if and only if the two measures

\[
\pi(dx)a(x, y)q(x, dy) = \pi(dy)a(y, x)q(y, dx) \tag{5.3}
\]

**Proof.** By (5.1),

\[
\begin{align*}
\pi(dx)p(x, dy) &= \pi(dx)a(x, y)q(x, dy) + \pi(dx)A(x)\delta_x(dy) \\
\pi(dy)p(y, dx) &= \pi(dy)a(y, x)q(y, dx) + \pi(dy)A(y)\delta_y(dx)
\end{align*} \tag{5.4}
\]

The two measures on the right-hand side of (5.4) are the same since

\[
\begin{align*}
\int_{x \in X} \int_{y \in X} f(x, y)\pi(dx)A(x)\delta_x(dy) &= \int_{x \in X} f(x, x)A(x)\pi(dx) \\
&= \int_{y \in X} f(y, y)A(y)\pi(dy)
\end{align*}
\]

for all measurable \( f(x, y) \geq 0 \). It follows that (5.3) and (5.2) are equivalent.

**Theorem 5.1.** (Metropolis-Hastings) Suppose that the measure \( \pi(dx) \) and transition function \( q(x, dy) \) satisfy

\[
q(x, dy) = q(x, y)dy \quad \text{and} \quad \pi(dx) = \pi(x)dx
\]

with respect to a nonnegative measure \( dx \) on \( X \). Then

\[
a(x, y) = \min \left\{ 1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)} \right\} \tag{5.5}
\]

satisfies the reversibility condition (5.1), where the right-hand size of (5.5) is replaced by 1 if \( \pi(x)q(x, y) = 0 \). In particular, (5.5) implies that \( \pi(x)dx \) is a stationary measure for \( p(x, y)dy \).

**Corollary 5.1.** The function \( a(x, y) \) in (5.5) is the pointwise maximum value of all functions with \( 0 \leq a(x, y) \leq 1 \) that satisfy the reversibility condition (5.2) for \( p(x, dy) \) in (5.1).

Corollary 5.1 implies that (5.5) has the minimum expected number of wait states among the modified transition functions \( p(x, dy) \) in (5.1) that satisfy the reversibility condition (5.2), and in this sense is the most efficient time-reversible acceptance function for \( p(x, dy) \) and \( \pi(x) \).
Proof of Theorem 5.1. By (5.5),
\[
\pi(x)a(x,y)q(x,y) = \min \{ \pi(x)q(x,y), \pi(y)q(y,x) \}
\]
(5.6)
since the right-hand side of the first equation in (5.6) is a symmetric function of \(x\) and \(y\). This implies (5.3) and hence the reversibility condition (5.2).

Proof of Corollary 5.1. Let \(Q\) be the set of all functions \(b(x,y)\) such that \(0 \leq b(x,y) \leq 1\) and
\[
\pi(x)b(x,y)q(x,dy) = \pi(y)b(y,x)q(y,dx)
\]
(5.7)
for all \(x, y \in X\). Set \(a(x,y) = \max_{b \in Q} b(x,y)\). Then \(a(x,y)\) also satisfies (5.7). It follows from \(0 \leq a(x,y) \leq 1\) that either \(a(x,y) < 1\) or \(a(x,y) = 1\). If \(a(x,y) < 1\), then \(a(y,x) = 1\) by the definition of \(Q\). Similarly, \(a(x,y) < 1\) implies \(\pi(x)q(x,y) > 0\). Thus by (5.7)
\[
a(x,y) = \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}
\]
which implies (5.5).

6. Consequences of the Metropolis-Hastings Theorem. It follows in general that

Theorem 6.1. Let \(X_n\) is an ergodic Markov chain with stationary distribution \(\pi(dx)\) and assume \(F(x) \geq 0\). Then
\[
\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} F(X_k) = \int_X F(x) \pi(x) dx \quad \text{a.s.}
\]
(6.1)
\[
\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} F(X_k)^2 = \int_X F(x)^2 \pi(x) dx \quad \text{a.s.}
\]
and for integers \(m \geq 0\)
\[
\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} F(X_k)F(X_{k+m})
\]
\[
= \int_X F(x) \int_X F(y) \pi(y) p^m(x,dy) \pi(x) dx \quad \text{a.s.}
\]
(6.2)
The function \(p^m(x, dy) = \Pr(X_m \in dy \mid X_0 = x)\) in (6.2) is the \(m\)th power of the transition density \(p(x, dy)\) in (5.1).

The relations (6.1) imply that, in principle, it does not matter what proposal function \(q(x,y)\) we use. The random variables \(F(X_k)\) have the same asymptotic mean and variance in all cases. However, the asymptotic variance of \((1/n) \sum_{k=1}^{n} F(X_k)\) in (6.1) (as opposed to the asymptotic variance of \(F(X_k)\)) depends on the limiting autocovariances (6.2). Proposal functions that minimize these autocovariances will lead to more accurate estimators of the right-hand side of (6.1) for finite \(n\).

Similarly, different ergodic Markov chains \(X_n^*\) take longer to converge or “mix”, so that the minimum \(n\) for which the left-hand side of (6.1) is a reasonable approximation of the integral side can vary even if the asymptotic covariances are the same. Thus, particular choices of the proposal function \(q(x,y)dy\) do matter.

Some particular forms of MCMC algorithms or proposal functions are used often enough to have special names:
7. Random Walk Sampling. If \( q(x, y) = q(y, x) \) is symmetric, then the acceptance function (5.5) simplifies to
\[
a(x, y) = \min \left\{ 1, \frac{\pi(y)}{\pi(x)} \right\}
\]
(7.1)
That is, given a proposed \( y \) for \( \pi(x) \) in (1.2), we accept \( y \) if either \( \pi(y) \geq \pi(x) \) or else accept with probability \( \pi(y)/\pi(x) \) if \( \pi(y) < \pi(x) \). Since \( a(x, y) \) depends only on ratios of \( g(x) \) or \( \pi(x) \) values, the normalization constant \( C \) in (1.2) does not have to be known.

This is Metropolis’ (1953) original form of the algorithm. The usual definition of random walk for \( q(x, y) \) in \( \mathbb{R}^d \) assumes that the proposal random variable \( Y \) satisfies
\[
\{ Y \mid X = x \} \approx x + W
\]
(7.2)
for a random variable \( W \). Note that (7.2) is equivalent to \( q(x, y) = w(x - y) \) where \( w(y) \) is the density for \( W \). Metropolis’ condition \( q(x, y) = q(y, x) \) is equivalent to \( w(x) = w(-x) \), which is equivalent to \( W \approx -W \) or that the distribution of \( W \) is symmetric about 0.

8. Multiplicative Random Walks. It is sometimes useful to have multiplicative updates
\[
\{ Y \mid X = x \} \approx xW
\]
(8.1)
instead of additive updates (7.2). This allows proposed values \( Y \) to be larger when \( X \) is larger. It follows from (8.1) that
\[
E(f(Y) \mid X = x) = E(f(xW)) = \int f(xy)w(y) \, dy
\]
\[
= \int f(y)q(x, y) \, dy = (1/x) \int f(y)w(y/x) \, dy
\]
for \( f(y) \geq 0 \), where \( w(y) \) is the density for \( W \). This implies that \( q(x, y) = (1/x)w(y/x) \). The proposal function \( q(x, y) \) is said to define a symmetric multiplicative random walk if the random variable \( W \approx 1/W \), which is equivalent to \( w(y) = (1/y)^2 w(1/y) \). (Exercise: Prove this.) Hence
\[
\frac{\pi(y)q(y, x)}{\pi(x)q(x, y)} = \frac{\pi(y)(1/y)w(x/y)}{\pi(x)(1/x)w(y/x)} = \frac{\pi(y)(1/y)(y/x)^2w(y/x)}{\pi(x)(1/x)w(y/x)} = \frac{\pi(y)y}{\pi(x)x}
\]
Thus the acceptance function for an arbitrary multiplicative random walk is
\[
a(x, y) = \min \left\{ 1, \frac{\pi(y)y}{\pi(x)x} \right\}
\]
(8.2)
9. Independence Sampling. If the value proposed for \( X_{n+1} \) is independent of \( X_n \), that is, if \( q(x, y) = q(y) \), then the MH algorithm is called an independence sampler. The acceptance function becomes
\[
a(x, y) = \min \left\{ 1, \frac{\pi(y)q(x)}{\pi(x)q(y)} \right\} = \min \left\{ 1, \frac{\pi(y)/q(x)}{\pi(x)/q(y)} \right\}
\]
(9.1)
That is, given \( X_n = x \) and \( k(y) = \pi(y)/q(y) \), the proposed value \( y \) is accepted if \( k(y) \geq k(x) \) and accepted with probability \( k(y)/k(x) \) if \( k(y) < k(x) \). Note that we do not need to know the normalizing constants of either \( \pi(y) \) or \( q(y) \).

It is very important for independence samplers that \( q(y) \) not be lighter-tailed than \( \pi(y) \) for large values of \( y \) if \( X \) is noncompact, nor than \( q(y) \) be lighter tailed at a singularity of \( \pi(y) \). That is,
we should NOT have \( k(y) = \pi(y)/q(y) \gg 1 \) either for large \( y \) or at a singularity of \( \pi(y) \). In that case, large values of \( Y \) (as measured by \( \pi(y) \)) are rarely proposed and it is easy to find examples in which the independence sampler has truly terrible convergence properties. That is, the left-hand side of (6.1) is a reasonable approximation of \( \int F(x)\pi(x)\,dx \) only for extremely large values of \( n \).

A possible way to make this less likely for an independence sampler for \( \pi(y) \) that might be heavy-tailed for large \( y \) is to use a proposal distribution \( q(y) \) that has a power law \( (q(y) = \max\{1, 1/y^n\}) \) or a Student-\( t \) distribution, but one would have to make sure that \( \pi(y) \) is not more heavy-tailed yet than \( q(y) \) for large \( y \).

If we know how to generate random variables \( X_{n+1} \) with distribution exactly equal to \( \pi(x) = Cg(x) \) and set \( q(x) = \pi(x) \), then \( a(x, y) = 1 \). In this case, there are no wait states and the independence sampler is the same as classical Monte Carlo sampling.

10. Von Neumann’s Rejection Sampling. The independence sampler is similar in spirit to the “rejection method” of von Neumann (1951) for generating random variables with an arbitrary distribution \( \pi(x) \). Von Neumann’s rejection method assumes

\[
\pi(x) = cA(x)q(x)
\]

where \( 0 \leq A(x) \leq 1 \). The algorithm is to sample values \( Y \) from the proposal distribution \( q(y) \) and accept \( Y \) with probability \( A(y) \). If the value is rejected, the algorithm continues to sample from \( q(y) \) until a value is accepted. The final accepted value of \( Y \) has distribution that is exactly \( \pi(x) \). (Exercise: Prove this.)

In contrast with von Neumann’s method, the independence sampler does not have retrials. If a value is rejected, the previous value is used. The Markov chain values \( X_n \) have \( \pi(dx) \) as a stationary distribution, but in general do not have the distribution \( \pi(dx) \) themselves.

If we know how to generate random variables \( X_{n+1} \) with distribution exactly equal to \( \pi(x) = q(x) \), then we can take \( A(x) = 1 \). In this case, there are no rejections and the method is again the same as classical Monte Carlo sampling.

11. Metropolis-Hastings and Importance Sampling. A useful technique to improve the efficiency of any Monte Carlo technique is the following. Suppose in the limiting approximation (5.1)

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} F(X_n) = \int_{X} F(x)\pi(x)\,dx
\]

and that \( F(x) \) is small on a portion of the state space \( X \) that has significant weight under \( \pi(x) \).

Then the Markov chain \( X_n \) (with stationary distribution \( \pi(x)\,dx \)) may spend most of its time in parts of the state space \( X \) that contribute little to the sum and hence little to the estimation of the integral. In general, importance sampling refers to changing any Monte Carlo method so that, first, the sampling is biased towards parts of the state space that are more important for the sum or integral being estimated and, second, corrections are made for the resulting biased sampling.

Let \( H(x) \geq 0 \) be a function on \( X \) that is small where \( F(x) \) is small and relatively large where \( F(x) \) is large but that \( H(x) \) is easier to work with than \( F(x) \). Let \( Y_n \) be the Metropolis-Hastings Markov chain defined in Section 2 with the same proposal function \( q(x, y) \) defined with \( H(x)g(x) \) (or \( H(x)\pi(x) \)) as the stationary distribution instead of \( g(x)\,dx \) or \( \pi(x)\,dx \). (Recall that \( \int \pi(x)\,dx = 1 \) but that \( g(x) \) is unnormalized.) Then the acceptance function for \( Y_n \) is

\[
a(x, y) = \min \left\{ 1, \frac{H(y)g(y)q(y, x)}{H(x)g(x)q(x, y)} \right\}
\]
Metropolis-Hastings substeps. The acceptance function of the $i$th portion of the coordinates $x$ in (12.1), however, it is often possible to update $X_i$ and the Markov chains $Y_n$ use of "full conditional distributions" in (12.2) is just for ease of notation. In fact, the denominators of (11.4) cancel out, so that the terms in the sum in (11.4) should have smaller variance than the corresponding terms in (11.1) without importance sampling.

If the denominator of (11.5) is known or is easy to calculate, this should give a more efficient way to estimate $\int_X F(x)\pi(x)\,dx$.

12. Componentwise Metropolis-Hastings and Gibbs Samplers. In practice, the values $x \in X$ and the Markov chains $X_n$ on $X$ are usually multidimensional and often highly multidimensional. However, it is often possible to update $X_n$ one component at a time. Changing or updating only a portion of the coordinates $x \in X$ is called a block move.

Assume $x, X_n \in R^d$ and that we can find $d$ one-dimensional proposal functions of the form $q_i(x, y) = q_i(x_i, x_{-i}, y), \quad x \in R^d, \ x_i, y \in R, \ x_{-i} \in R^{d-1}$ (12.1)

In (12.1), $x_{-i} \in R^{d-1}$ are the components of $x \in R^d$ other than the $i$th component $x_i$. The vector $x_{-i}$ in (12.1) is viewed as a parameter for the one-dimensional proposal density $q_i(x_i, y)$.

Given (12.1), we update the Markov chain $X_n$ by carrying out $d$ consecutive one-dimensional Metropolis-Hastings substeps. The acceptance function of the $i$th substep is

$$a_i(x_i, x_{-i}, y_i) = \min \left\{ 1, \frac{\pi(y_i \mid x_{-i}) q_i(y_i, x_{-i}, x_i)}{\pi(x_i \mid x_{-i}) q_i(x_i, x_{-i}, y_i)} \right\}$$ (12.2)

Here $\pi(x_i \mid x_{-i}) = \pi(x \mid x_{-i}) = \pi(x) / \int \pi(x)\,dx_{-i}$ for $x = (x_i, x_{-i})$ is the “full conditional distribution” of $x_i$ given $x_{-i}$. That is, $\pi(x_i \mid x_{-i})$ is the distribution of $x_i$ conditional on the other components of $x$. The denominators of $\pi(y_i \mid x_{-i})$ and $\pi(x_i \mid x_{-i})$ in (12.2) cancel out, so that the use of “full conditional distributions” in (12.2) is just for ease of notation. In fact

$$a_i(x_i, x_{-i}, y_i) = \min \left\{ 1, \frac{\pi(y_i, x_{-i}) q(y_i, x_{-i}, x_i)}{\pi(x_i, x_{-i}) q(x_i, x_{-i}, y_i)} \right\}$$ (12.3)

Of course as before, if $\pi(x) = Cg(x)$ where $g(x)$ is simple but the normalizing constant $C$ is not, the constant $C$ cancels out in (12.3).

By convention, $\pi(y \mid x_{-i}) = 0$ unless $y_j = x_j$ for $j \neq i$, so that the densities $\pi(y \mid x_{-i})\,dy$ and $q_i(x, x_{-i}, y)$ are concentrated on lines parallel to the $i$th coordinate axis.
While each of the $d$ component substeps of $X_n$ is a Metropolis-Hastings update in the sense of Sections 2 or 4, the full Markov chain $X_n$ (after $d$ substeps) is generally no longer of this form. Nevertheless, $X_n$ has $\pi(x) \, dx$ as a stationary measure since each substep has $\pi(x) \, dx$ as a stationary measure. The (Birkhoff) ergodic theorem (6.1) still holds for most reasonable choices of componentwise proposal functions (12.1).

In many cases, the individual conditional distributions $\pi(x_i | x_{-i})$ are easy to describe even when the structure of $\pi(x)$ is complex. If the one-dimensional normalizing constant of $\pi(x | x_{-i})$ is known, then it is a probability distribution in $x_i$, unlike $\pi(x)$.

If we know how to generate random variables with the conditional distribution $\pi(y_i | x_{-i})$, then we can use

$$q_i(x_i, y_i) = q_i(x_i, x_{-i}, y_i) = \pi(y_i | x_{-i})$$

(12.4)
as a proposal function in (12.1). Given $x_{-i}$, this is equivalent to independence sampling from the conditional distribution of $x_i$. Since then $a_i(x_{-i}, x_i, y_i) = 1$ in (12.2), there are no wait states or rejection steps. This is called the Gibbs sampler or a Gibbs sampler step for $\pi(x | x_{-i})$. If $d = 1$, there is no dependence on $x_{-i}$ and this is the same as classical Monte Carlo sampling.

There is no requirement that the component steps in the MH algorithm be done in sequence. We could randomly sample from the components of $x$ to form a single-step Markov chain that has $\pi(x) \, dx$ as a stationary measure. The random-component version of $X_n$ is of the MH form of Section 2 with a proposal density $q(x, dy)$ that is concentrated on a set of coordinate axes (that is, a coordinate frame) through the point $x$. (Exercise: Verify this.)

In the more usual case of $d$ substeps in sequence, we can use oversampling for any components of $x$ for which the trajectories of the Markov chain (12.4) are unusually autocorrelated, have an unusually low acceptance rate, or perhaps are just less expensive to compute. By oversampling, we mean that proposals (12.1) with the acceptance functions (12.2) are repeated a fixed number of times in each Markov chain iteration (for example, 5 or 10 or 100 times) and only the last value is used. Several oversampling steps can convert Metropolis updates to an update that is of comparable efficiency to a Gibbs sampler update. Of course, there is no benefit to oversampling a Gibbs sampler substep. (Exercise: Explain why.)

### 13. Improving Convergence by Using Skew Transformations

Suppose $x = (x_1, x_2)$ where $x_1 \in R^d, x_2 \in R^m$, and $x \in R^n$ for $n = d + m$. In some case $x_1$ and $x_2$ (or parts of $x_2$) are naturally highly correlated and that $\pi(dx)$ is stiff (that is, rapidly changing) as a function of $(x_1, x_2)$. It that case, the Metropolis-Hastings algorithm can accept very only small changes in $x_1$ or $x_2$. This can cause the MC Markov chain to take an extremely long time to converge.

One way of improving convergence in this case is to use a proposal function that is, for example, jointly normal in $x_1$ and $x_2$ with a covariance matrix that is estimated from a preliminary run (or from a long initial burnin period of the MC Markov chain whose values are not recorded). Another way (which is the point of this section) is to update $x_1$ and then make an immediate parallel deterministic change in $x_2$.

Specifically, consider a Markov transition proposal function $q(x, dy)$ defined in terms of random variables by

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \rightarrow \begin{pmatrix} Y_1 \\ h(X_1, X_2, Y_1) \end{pmatrix}$$

(13.1)

We assume in (13.1) that $Y_1$ has the conditional density

$$P(Y_1 \in dy_1 \mid X_0 = (x_1, x_2)) = q(x, y_1)dy_1$$

and that $h(x_1, x_2, y_1)$ is a deterministic function of $x_1, x_2$, and $y_1$. We call (13.1) a skew or shear transformation of $X \in R^n$ since a change from $X_1$ to $Y_1$ causes a deterministic change in $X_2$ as if an elastic block is subjected to a shearing stress or transformation.
The Markov chain block updates that we have considered so far have been one-dimensional and also of full rank. That is, the proposal distribution maps a subinterval of a one-dimensional subspace of $\mathbb{R}^n$ into the same subinterval with the remaining coordinates viewed as a parameter. In contrast, (13.1) maps a $d$-dimensional subspace of $\mathbb{R}^n$ into a fairly general $d$-dimensional submanifold of $\mathbb{R}^n$.

The mapping (13.1) is very similar to a transformation that changes only $X_1$ only in a different coordinate system, and then changes coordinates back to the original coordinates. These reparametrization updates are a subset of the skew or shear transformations (13.1) (see below). While it is not known whether all allowable transformations (13.1) are reparametrization updates, reparametrization updates are the easiest to find and work with (see also below).

The cases $d = 0$ and $m = 0$ are not excluded in (13.1): The former is a purely deterministic move in $\mathbb{R}^0$ and the latter a mapping of full rank in $\mathbb{R}^n$. Note that the transformations of $X_1$ and $X_2$ in (13.1) cannot in general be done independently or in sequence: Update proposals must be Markov, and $X_1$ would not be available after updating $X_1$ and $Y_1$ would not be available after it was updated.

We now find general sufficient conditions for the existence of an acceptance function $a(x, y)$ for smooth functions $h(x_1, x_2, y_1)$ such that the update transition function $p(x, dy)$ has a density $\pi(x)dx$ as a stationary measure. (More exactly, such that the update satisfies the detailed balance condition (4.5).) Let $J_2h(x, y_1) = \frac{\partial}{\partial y_1} h(x, x_2, y)$ be the absolute value of the $m \times m$ Jacobian matrix of $h(x_1, x_2, y)$ with respect to $x_2$. Define

$$a(x, y) = \min \left\{ 1, \frac{\pi(y)q(y, x_1)J_2h(x_1, x_2, y_1)}{\pi(x)q(x, y)} \right\}$$

(13.2)

for a density $\pi(x)dx$ on $\mathbb{R}^n$. We then have the following theorem.

**Theorem 13.1.** Let $p(x, dy)$ be the transition function of the Markov process corresponding to the proposal distribution (13.1) and the acceptance function (13.2). Then a sufficient condition for $p(x, dy)$ to satisfy the detailed balance condition

$$\pi(x)dx \ p(x, dy) = \pi(y)dy \ p(y, dx)$$

(13.3)

is that, for $(x_1, x_2, y_1, y_2) \in \mathbb{R}^{2n}$,

$$y_2 = h(x_1, x_2, y_1) \quad \text{if and only if} \quad x_2 = h(y_1, y_2, x_1)$$

(13.4)

**Corollary 13.1.** A sufficient condition for (13.4) is that

$$y_2 = h(x_1, x_2, y_1) \quad \text{if and only if} \quad D(x, y) = 0$$

(13.5)

for some function $D(x, y)$ satisfying $D(x, y) = -D(y, x)$.

**Exercise:** Prove Corollary 13.1. Also prove the converse: That is, for every function $h(x_1, x_2, y_1)$ that satisfies (13.4), there exists a function $D(x, y)$ that satisfies (13.5) and $D(x, y) = -D(y, x)$. Note that in either case the condition $D(x, y) = D(y, x)$ would work just as well as $D(x, y) = -D(y, x)$.

An important special case of Corollary 13.1 is the class of functions $h(x_1, x_2, y_1)$ that satisfy (13.5) with

$$D(x, y) = R(x) - R(y)$$

(13.6)

for some function $R(x)$ on $\mathbb{R}^n$. We will see in the next section that the class (13.5)–(13.6) contains an interesting class of updates defined by temporary reparametrization. Given (13.5) and (13.6),
the Jacobian $J_2 h(x_1, x_2, y_1)$ in the acceptance function (13.2) can be expressed directly in terms of $R(x)$. For fixed $(x_1, x_2, y_1)$ and $y_2 = h(x_1, x_2, y_1)$, (13.6) implies

$$R(x_1, x_2) = R(y_1, h(x_1, x_2, y_1))$$

By the chain rule for Jacobians

$$J_2 R(x_1, x_2) = J_2 R(y_1, y_2) J_2 h(x_1, x_2, y_1)$$

and

$$J_2 h(x_1, x_2, y_1) = J_2 R(x_1, x_2) \frac{J_2 R(y_1, y_2)}{J_2 R(y_1, y_2)} \text{ for } y_2 = h(x_1, x_2, y_1)$$

(13.7)

In this case, the acceptance function (13.2) can be replaced by the more symmetric condition

$$a(x, y) = \min \left\{ 1, \frac{\pi(y) q(y, x_1) J_2 R(x_1, x_2)}{\pi(x) q(x, y_1) J_2 R(y_1, y_2)} \right\}$$

It is conceivable that, for any smooth function $h(x_1, x_2, y_1)$ satisfying (13.4), there exists a function $R(x)$ satisfying (13.5)–(13.6). As far as I know, this is an open question.

We give two examples before proving Theorem 13.1:

**Examples:** (1) Assume $d = 2, n = m + 2$, and write

$$x = (x_1, x_2) = (x_{11}, x_{12}, (x_{2i}))$$

for $1 \leq i \leq m$. Consider the transformation defined by $h(x_1, x_2, y_1)$ in equations (13.5)–(13.6) for

$$R(x) = (x_{2} - x_{11})/x_{12}$$

(13.8)

where $x_2 = \{x_{2i}\}$ and $R(x)$ are $m$-dimensional row vectors. This transformation might be suggested if $x_{11}$ somehow affects the means of $x_{2i}$, for $1 \leq i \leq m$ and $x_{12}$ affects the standard deviations of $x_{2i}$, and that changes in $x_{11}$ or $x_{12}$ by themselves without adjusting $x_{2i}$ cause drastic changes in the density $\pi(dx) = \pi(x)dx$.

Here $h(x_1, x_2, y_1)$ is defined by

$$y_2 = h(x_1, x_2, y_1) \text{ if and only if } R(x_1, x_2) = R(y_1, y_2)$$

where $x_2, h(x, y_1), y_2$, and $R(x)$ are $m$-dimensional. In particular $R(x) = R(y)$ if and only if

$$(y_2 - y_{11})/y_{12} = (x_2 - x_{11})/x_{12}$$

where $x_2, y_2$ are $m$ dimensional. This implies

$$y_2 = h(x_1, x_2, y_1) = y_{11} + (y_{12}/x_{12})(x_2 - x_{11})$$

(13.9)

for $x_2, y_2 \in R^m, x_{11}, x_{12}, y_{11}, y_{12} \in R^1$. Thus the skew transformation (13.1) corresponds to a random update of $(x_{11}, x_{12})$ followed by a “shift and scale” transformation of the component $x_{2i}$ about $x_{11}$.

Thus by (13.9)

$$J_2 h(x_1, x_2, y_1) = (y_{12}/x_{12})^n$$
The acceptance function (13.2) is then
\[ a(x, y) = \min \left\{ 1, \frac{\pi(y)q(y, x_1)(y_{12}/x_{12})^m}{\pi(x)q(x, y_1)} \right\} \]

Since \( J_2 R(x_1, x_2) = (1/x_{12})^m \), we could obtain the same result from (13.7).

(2) Assume \( d = 1, n = m + 1, x = (x_1, x_2) \) for \( x_2 \in R^n \), and as above
\[ R(x) = x_2 - x_{11}, \]
\[ y_2 = x_2 + (y_{11} - x_{11}) \]

This corresponds to a random update of \( x_{11} \) followed by a shift transformation of the \( m \) components of \( x_2 \). This skew transformation might be suggested if somehow \( x_1 \) affects the mean of \( x_{2i} \), for \( 1 \leq i \leq m \). In this case
\[ J_2 h(x_1, x_2, y_1) = J_2 R(x_1, x_2) = 1 \]
and the Jacobian does not appear in the acceptance function (13.2).

**Proof of Theorem 13.1.** The proposal transition function defined by (13.1) can be written
\[ q(x, dy) = q(x_1, x_2, y_1) dy_1 \delta_{h(x_1, x_2, y)}(dy_2) \]  
(13.11)

where (13.11) means that the measure \( q(x, dy) \) on \( R^n \) satisfies
\[ \int_{y \in X} \phi(y) q(x, dy) = \int \phi(h(x_1, x_2, y_1), y_1) dy_1 \]

for measurable functions \( \phi(y) \geq 0 \). The Markov chain with transition function \( p(x, dy) \) defined in Theorem 13.1 for \( q(x, dy) \) and \( a(x, y) \) is the chain that, for a given value of \( x \), "proposes" a value \( y \) with distribution (13.11) and then either "accepts" the value \( y \) and moves to that point, which takes place with probability \( a(x, y) \), or else "rejects" the value \( y \) and remains at the point \( x \), which takes place with probability \( 1 - a(x, y) \). This means
\[ p(x, dy) = a(x, y)q(x, y_1)dy_1 \delta_{h(x_1, x_2, y)}(dy_2) + A(x)\delta_x(dy) \]

where
\[ A(x) = 1 - \int a(x, y)q(x, dy) = 1 - \int a(x, y_1, h(x_1, x_2, y_1)) q(x, y_1) dy_1 \]

We now find sufficient conditions on \( h(x_1, x_2, y_1) \) in (13.1) and \( a(x, y) \) in (13.2) for the detailed balance condition (13.3).

First, note that the two measures \( \pi(x)dx \cdot p(x, dy) \) and \( \pi(y)dy \cdot p(y, dx) \) in (13.3) are measures in two \((n+d)\)-dimensional submanifolds of \( R^{2n} \). The first condition is that these two manifolds are the same. The first measure is concentrated on the set of points
\[ D_1 = \{ (x_1, x_2, y_1, y_2) : y_2 = h(x_1, x_2, y_1) \} \]  
(13.12)

while the second measure is concentrated on
\[ D_2 = \{ (x_1, x_2, y_1, y_2) : x_2 = h(y_1, y_2, x_1) \} \]  
(13.13)

The condition \( D_1 = D_2 \) is exactly the symmetry condition (13.4).
As in the proof of Lemma 5.1, the detailed balance condition (13.3) follows from the relation
\[ \pi(x) dx a(x, y) q(x, dy) = \pi(y) dy a(y, x) q(y, dx) \] (13.14)
In general, two measures \( \mu_1(dx dy) \) and \( \mu_2(dx dy) \) are the same in \( \mathbb{R}^{2n} \) if and only if
\[ \iint \phi(x, y) \mu_1(dx dy) = \iint \phi(x, y) \mu_2(dx dy) \]
for all measurable functions \( \phi(x, y) \geq 0 \). The integral of \( \phi(x, y) \) with respect to the left-hand measure in (13.14) is
\[ \iint \phi(x, y) \pi(x) a(x, y) q(x, dy) dx \]
\[ = \iint \iint \phi(x_1, x_2, y_1, h(x_1, x_2, y_1)) \pi(x_1, x_2) \times a(x_1, x_2, y_1, h(x_1, x_2, y_1)) q(x_1, x_2, y_1) dx_1 dx_2 dy_1 \] (13.15)
The integral with respect to the second measure in (13.14) is
\[ \iint \phi(x, y) \pi(y) a(y, x) q(y, dx) dy \]
\[ = \iint \iint \phi(x_1, h(y_1, y_2, x_1), y_1, y_2) \pi(y_1, y_2) \times a(y_1, y_2, x_1, h(y_1, y_2, x_1)) q(y_1, y_2, x_1) dy_1 dy_2 dx_1 \] (13.16)
The substitution \( x_2 = h(y_1, y_2, x_1) \) \( (y_2 = h(x_1, x_2, y_1)) \) in (13.16), viewing \( (x_1, y_1) \) as fixed, implies
\[ \iint \phi(x, y) \pi(y) a(y, x) q(y, dx) dy \]
\[ = \iint \iint \phi(x_1, x_2, y_1, h(x_1, x_2, y_1)) \pi(y_1, h(x_1, x_2, y_1)) \times a(y_1, h(x_1, x_2, y_1), x_1, x_2) q(y_1, h(x_1, x_2, y_1), x_1) \times J_2 h(x_1, x_2, y_1) dx_1 dx_2 dy_1 \]
\[ = \int_{\mathcal{D}_1} \phi(x, y) \pi(y) a(y, x) q(y, x) J_2 h(x_1, x_2, y_1) dx_1 dx_2 dy_1 \] (13.17)
where \( J_2 h(x_1, x_2, y_1) \) is the Jacobian function in (13.2). It follows from (13.15) and (13.17) that the two measures in (13.14) are the same if
\[ \pi(x) a(x, y) q(x, y_1) = \pi(y) a(y, x) q(y, x_1) J_2 h(x_1, x_2, y_1) \] (13.18)
for all \( (x, y) \in \mathcal{D} \) for \( \mathcal{D} = \mathcal{D}_1 = \mathcal{D}_2 \) in (13.12)–(13.13). I now claim that
\[ a(x, y) = \min \left\{ 1, \frac{\pi(y) q(y, x_1) J_2 h(x_1, x_2, y_1)}{\pi(x) q(x, y_1)} \right\} \] (13.19)
in (13.2) satisfies (13.18). By (13.4)
\[ h(x_1, h(y_1, y_2, x_1), y_1) = y_2 \]
and the chain rule for Jacobians with $x_2 = h(y_1, y_2, x_1)$ implies
\[ J_2 h(x_1, x_2, y_1) J_2 h(y_1, y_2, x_1) = 1 \] (13.20)

It follows from (13.19) that
\[ \pi(x) a(x, y) q(x, y_1) = \min \{ \pi(x) q(x, y_1), \pi(y) q(y, y_1) J_2 h(x_1, x_2, y_1) \} \]
and hence
\[ \pi(y) a(y, x) q(y, x_1) = \min \{ \pi(y) q(y, x_1), \pi(x) q(x, y_1) J_2 h(y_1, y_2, x_1) \} \]

Hence by (13.20)
\[ \pi(y) a(y, x) q(y, x_1) J_2 h(x_1, x_2, y_1) = \min \{ \pi(y) q(y, x_1) J_2 h(x_1, x_2, y_1), \pi(x) q(x, y_1) \} \]

The relation (13.18) follows from the fact that the right-hand sides of the first and third of the three equations above are the same. In fact, it follows as in the proof of Theorem 5.1 that the maximal pointwise solution of (13.18) for functions $0 \leq a(x, y) \leq 1$ is the acceptance function (13.19).

**Remark.** The notion of skew or shear transformation is very close to the ideas of partial resampling and generalized multigrid methods discussed in Liu and Sabatti (2000) and in Sections 8.1 and 8.3 of Liu (2001). See also the remarks at the end of the next section.

### 14. Temporary Reparametrization as a Skew Transformation.

Let $x = (x_1, x_2)$ for $x_1 \in \mathbb{R}^d$, $x_2 \in \mathbb{R}^m$, and $x \in \mathbb{R}^n$ as in the previous section. In principle, a natural way to carry out a transformation
\[ X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \rightarrow \begin{pmatrix} Y_1 \\ h(X_1, X_2, Y_1) \end{pmatrix} \] (14.1)

where $P(Y_1 \in dy_1 | X_0 = x) = q(x, y_1) dy_1$ is to change coordinates in $\mathbb{R}^m$ in such a way that the update acts only on the first $d$ coordinates and then change coordinates back. Specifically, let
\[ S(x) = S \left[ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \right] = \begin{pmatrix} x_1 \\ S_2(x_1, x_2) \end{pmatrix} \]
be a one-one continuously-differentiable nonsingular mapping of $\mathbb{R}^n$ into itself. Suppose that we make a proposal $X_1 \rightarrow Y_1$ in the new coordinates that changes only the first $d$ coordinates in the new coordinates:
\[ S \left[ \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \right] = \begin{pmatrix} X_1 \\ S_2(X) \end{pmatrix} \rightarrow \begin{pmatrix} Y_1 \\ S_2(X) \end{pmatrix} \] (14.2)

Set $S_{22}(x_2; x_1) = S_2(x_1, x_2)$ when we view $S_2(x_1, x_2)$ as a transformation of $x \in \mathbb{R}^m$ with $x_1 \in \mathbb{R}^d$ as a parameter. Then the mapping (14.1) in the original coordinates is
\[ X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \rightarrow S^{-1} \left[ \begin{pmatrix} Y_1 \\ S_2(X) \end{pmatrix} \right] = \begin{pmatrix} Y_1 \\ S_{22}^{-1}(S_2(X_1, X_1); Y_1) \end{pmatrix} \] (14.3)

This is a skew transformation of the form (13.1) for
\[ h(X, Y_1) = S_{22}^{-1}(S_2(X_2; X_1); Y_1) \] (14.4)
If $Y_2 = h(X, Y_1)$, then by (14.4) and (14.3)

$$S_2(Y_1, Y_2) = S_2(h(X, Y_1); Y_1) = S_2(X_2, X_1) = S_2(X_1, X_2)$$

Since $S(X)$ is one-one, this implies that if $Y_2 = h(X, Y_1)$ for $Y = (Y_1, Y_2)$ if and only $S_2(Y) = S_2(X)$, or equivalently if and only if

$$D(X, Y) = S_2(Y) - S_2(X) = 0$$

In particular, the function $h(x_1, x_2, y_1)$ in the reparametrization transformation (14.3) satisfies the symmetry conditions (13.4)–(13.5) for $D(x, y) = S_2(x) - S_2(y)$. It then follows from Theorem 13.1 that

**Theorem 14.1.** Let $q(x, dy)$ be the proposal distribution defined by (14.3) for the reparametrization (14.2). Then the associated Markov transition function $p(x, dy)$ for a density $\pi(x)$ and the acceptance function

$$a(x, y) = \min \left\{ 1, \frac{\pi(y)q(y, x_1)J_2S(x_1, x_2)}{\pi(x)q(x, y_1)J_2S(y_1, y_2)} \right\}$$

satisfies the detailed balanced condition (13.3). In particular, the Metropolis-Hastings Markov chain associated with (14.1) has $\pi(x)$ as a stationary measure.

**Remarks.** (1) If $h(X, Y_1)$ in (14.3) is replaced by a Gibbs sampler step from the density $\pi(x)$ restricted to the “fiber”

$$\{ Y : S_2(Y) = S_2(X) \} \quad (14.5)$$

then (14.1)–(14.3) is essentially the same as the covariance-adjusted Markov chain of Liu (1998) (referenced in Chen et al. 2000).

(2) Liu and Sabatti (2000) define a similar procedure called Grouped Move Multi-Grid Monte Carlo by Chen et al. (2000). Here (14.5) are the orbits of a locally compact group $G$ acting on $X$. The transformation $h(X, Y_1)$ in (14.3) is either a Gibbs-sampler step from the measure induced on (14.5) by $\pi(x)dx$ and Haar measure on $G$ or else related Metropolis-Hastings updates (Chen et al. 2000, Liu 2001). See Liu and Sabatti (2000) for examples. Possible choices of the group $G$ for $X = R^n$ would be the Lie group $R^m$ itself or the nilpotent Lie group of shift and scale transformations acting on $R^m$. This in fact is the same as the examples (13.8)–(13.10) except that the transformations in Section 12 are deterministic within fibers rather than random.

**15. Bayesian Models in Statistics.** The Metropolis-Hastings algorithm is often used in Bayesian analyses in statistics. In fact, the MH algorithm may be the main reason for the increased popularity of Bayesian methods in the last 50 years.

For definiteness, assume that one has an observation of a vector-valued random variable $X$ (which might, for example, be repeated observations of vector-valued random variables of lower dimension). Assume that the distribution of $X$ depends on a parameter $\theta$ that we want to estimate. The theoretical distribution of $X$ is given by the likelihood $L(\theta, X)$ where $E(F(X)) = \int L(\theta, x) F(x) dx$ for functions $F(x) \geq 0$. Many statistical methods for estimating of $\theta$ are based on the idea that, for given observed data $X$, those values of $\theta$ for which $L(\theta, X)$ is relatively large are more likely to be close to the “true” value of $\theta$ that generated the random data $X$.

The Bayesian approach is to treat $\theta$ as well as $X$ as a random variable by introducing an arbitrary “prior distribution” $\pi_0(\theta)$ for $\theta$. Then $\pi(\theta, x) = \pi_0(\theta)L(\theta, x)$ is a normalized probability distribution for the joint distribution of $(\theta, X)$ since

$$\int \int \pi(\theta, x) d\theta dx = \int \pi_0(\theta) \int L(\theta, x) dx d\theta = 1$$
In particular, the marginal probability in $\theta$ for $\pi(\theta, x)$ is $\pi_0(\theta)$.

Given $\pi(\theta, x)$, one considers the conditional distribution of the random variable $\theta$ given the observed data $X$, which is

$$
\pi_1(\theta) = \pi(\theta \mid X) = \frac{\pi_0(\theta)L(\theta, X)}{\int \pi_0(s)L(s, X) \, ds} = C_X \pi_0(\theta)L(\theta, X)
$$

(15.1)

where $C_X$ depends only on $X$. Given $X$, the density $\pi_1(\theta)$ should be largest for values of $\theta$ that are closest to being “true”.

For definiteness, the distribution $\pi_1(\theta) = \pi_1(\theta \mid X)$ is called the (Bayesian) “posterior distribution” of $\theta$ given $X$, as opposed to the (Bayesian) “prior distribution” $\pi_0(\theta)$. The density (15.1) is of the form (1.2) with

$$
g(\theta) = \pi_0(\theta)L(\theta, X) \quad \text{and} \quad \pi(\theta) = C_X \pi_0(\theta)L(\theta, X)
$$

(15.2)

The MH algorithm is useful for (15.2) since it does not force us to calculate $C_X$.

Bayesian statistics makes inferences about the unknown value of $\theta$ based on the posterior distribution $\pi_1(\theta) = C_X \pi_0(\theta)L(\theta, X)$. For example, the “Bayes estimator” of $\theta$ is defined as the average of $\theta$ over $\pi_1(\theta)$, which is

$$
\hat{\theta} = \hat{\theta}_B = E(\theta \mid X) = \int \theta \pi_1(\theta) \, d\theta = C_X \int \theta \pi_0(\theta)L(\theta, X) \, d\theta
$$

(15.3)

The main disadvantage of Bayesian methods is that all inferences depend on the prior $\pi_0(\theta)$. Only in rare cases is there a natural candidate for $\pi_0(\theta)$. Thus, in principle, inferences should always be done for more than one choice for $\pi_0(\theta)$. If the resulting inferences are not close, then the problem should be rethought.

The degree of confidence that one might have in a Bayesian estimator such as $\hat{\theta}_B$ can be measured by the distribution of $\pi_1(\theta)$ about $\hat{\theta}_B$. For example, if $Q$ is a set of values of $\theta$ such that

$$
\hat{\theta} \in Q \quad \text{and} \quad \int_Q \pi_1(\theta) \, d\theta \geq 0.95
$$

then we can say that we are “95% posterior sure” that $Q$ contains the true value of $\theta$. A set $Q$ with these properties is called a “95% credible region” for $\theta$. This is the Bayesian analog of the classical 95% confidence interval or confidence region, in which $\theta$ is treated as a parameter and not as an unobserved value of a random variable. However, the terms “credible region” and “confidence region” are sometimes used interchangeably.

16. Arguments For and Against “Improper” Priors. An alternative justification of Bayesian methods can be given as follows. Given a statistical model and data $X$, the likelihood $L(\theta, X)$ should give most of our information about $\theta$. Classical (non-Bayesian) statistical methods are often based on the Maximum Likelihood Estimator (MLE) of $\theta$, which is that value $\hat{\theta} = \hat{\theta}(X)$ at which $L(\theta, X)$ attains its maximum value over $\theta$.

An alternative approach might be to consider the measure $L(\theta, X) \, d\theta$ instead and ask where most of the mass of this measure is concentrated. This leads to the measure

$$
\pi_2(\theta) \, d\theta = \pi_2(\theta \mid X) \, d\theta = L(\theta, X) \, d\theta
$$

(16.1)

instead of $\pi_1(\theta) = \pi_1(\theta \mid X)$. The measure $\pi_2(\theta)$ is called the “Bayesian posterior with uniform improper prior” if $\int X \, d\theta = \infty$ and “with uniform (proper) prior” if $\int X \, d\theta < \infty$ (within the normalization constant $\int X \, d\theta$). Many Bayesian methods fail if

$$
\int \pi_2(\theta) \, d\theta = \int L(\theta, X) \, d\theta = \infty
$$

(16.2)
since then any Markov chain $X_n$ with stationary distribution $C\pi_2(\theta \mid X)$ will be at most null recurrent. If $X_n$ is ergodic, this implies
\[ \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} F(X_k) = 0 \quad \text{a.s.} \quad (16.3) \]
for any function $F(x)$ with $\int \mid F(\theta) \mid \pi_2(\theta) \, d\theta < \infty$. In this case, the Markov chain $X_n$ is said to “converge to infinity”.

If in fact (16.2) is the case, then any Bayesian estimator of a function of $\theta$ can be considered to be an artifact of the prior $\pi_0(\theta)$, since the closer that the prior $\pi_0(\theta)$ is to the “uniform improper prior” $\pi_0(\theta) = 1$, the closer that any estimator based on $\pi_0(\theta)$ is to the situation (16.3). Note that (16.3) implies that median estimators of $\theta$ are infinite as well as mean-based estimators. However, it is still possible for a well-behaved classical MLE $\hat{\theta}(X)$ to exist.

An argument against the improper prior (16.1) or against uniform priors in general is that “$d\theta$” in (16.1) is not invariant under changes of variable unless it is a discrete measure. For example, if $\psi = \theta^2$, then $\theta = \sqrt{\psi}$ and $L(\theta) \, d\theta = L(\psi) \, d\psi / 2\sqrt{\psi}$. Then the “natural” measure $d\theta$ has been changed to $d\psi / 2\sqrt{\psi}$. This problem does not arise with the posterior distribution $\pi_1(\theta) = \pi_0(\theta) L(\theta, X)$ as long as one views the prior and posterior distributions $\pi_0(\theta)\ d\theta$ and $\pi_1(\theta)\ d\theta$ as measures instead of as functions.

Of course, the strongest argument against the improper prior (16.1) given (16.2) is (16.3). In that case, if $n$ is sufficiently large, all parameter values estimated by the sample-path averages (16.3) will be zero.

It often happens in practice that (16.3) appears to converges to reasonable values for most components of a high-dimensional $X_n$ with an improper prior. This is because the components of $X_n$ often converge on different time scales. That is, some component or components of $X_n$ converge to infinity, but other components are nearly independent of the first set of components, and adjust themselves to the first set of components in a fast enough time scale that the averages (16.3) give stable reasonable estimates for those components. This phenomenon is called quasi-stability.

A related problem is that an improper prior for one component, for which the likelihood is obviously integrable without a normalizable prior distribution, can mysteriously lead to unstable behavior in other components even for other components with normalizable priors. This is another aspect of quasi-stability.

17. Sampling from the Prior. Given the form of the distribution $g(\theta) = \pi_0(\theta) L(\theta, X)$ in (15.2), a plausible choice for updating $\theta$ in the MH algorithm might be the proposal function
\[ q(\theta, y) = \pi_0(y) \]
This is independence sampling from the prior distribution for $\theta$ (where we now assume $\int \pi_0(\theta) \, d\theta = 1$). The acceptance function (15.1) becomes
\[ a(\theta, y) = \min \left\{ 1, \frac{g(y)}{g(\theta)} \frac{q(\theta)}{q(y)} \right\} = \min \left\{ 1, \frac{L(y, X)}{L(\theta, X)} \right\} \]
This nicely separates the effect of the prior distribution $\pi_0(\theta)$ and likelihood $L(\theta, X)$ on the Markov chain $X_n$.

18. Conjugate Priors. The joint density $\pi_0(\theta) L(\theta, X)$ takes a simple form in many important cases in statistics. For example, suppose that $X$ has a Poisson distribution with mean $\theta$, so that the likelihood is
\[ L(\theta, X) = e^{-\theta} \frac{\theta^X}{X!}, \quad X = 0, 1, 2, 3, \ldots \quad (18.1) \]
Suppose that we choose a gamma density with parameters \((\alpha, \beta)\) for the prior density \(\pi_0(\theta)\):

\[
\pi_0(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta \theta}, \quad 0 \leq \theta < \infty
\]  

(18.2)

We can write \(\pi_0(\theta) \approx G(\alpha, \beta)\) symbolically. Then the posterior density

\[
\pi_1(\theta | X) = C_X \pi_0(\theta) L(\theta, X) = C_X \theta^{n+\alpha-1} e^{-\beta(\theta + \theta X)} = C_X (\alpha, \beta) \theta^{X+\alpha-1} e^{-(1+\beta)\theta}
\]

is also a gamma distribution. Symbolically

\[
\pi_1(\theta) = \pi(\theta | X) \approx G(\alpha + X, \beta + 1)
\]

In particular, if \(L(\theta, X)\) is the Poisson likelihood (18.1) and \(\pi_0(\theta)\) is the gamma density (18.2), then the posterior density \(\pi_1(\theta)\) is a gamma density with different parameters.

When this happens for an arbitrary family of densities (here, the gamma densities (18.2)), we say that family of densities is a **conjugate prior** for the likelihood, here for the Poisson likelihood (18.1). There are only a few cases where conjugate priors for likelihoods are known, but they cover many of the most important distributions in statistics (see e.g. DeGroot 1989, Chapter 6).

As a second example, suppose that \(\pi_0(\theta)\) is the beta density

\[
\pi_0(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \quad \text{for} \quad 0 \leq \theta \leq 1
\]  

(18.3)

(Symbolically, \(\pi_0(\theta) \approx B(\alpha, \beta)\).) If we toss a biased coin \(n\) times with probability of heads \(\Pr(H) = \theta\) and observe \(X = k\) heads in a particular order, then the joint density of \((\theta, X)\) is

\[
\pi_0(\theta) \binom{n}{X} \theta^X (1-\theta)^{n-X} = C(\alpha, \beta, n, X) \theta^{X+\alpha-1}(1-\theta)^{n-X+\beta-1}
\]

Thus \(\pi_1(\theta | X) \approx B(\alpha + X, \beta + n - X)\). This means that the family of beta densities is a conjugate prior for binomial sampling. It follows from a similar argument that the family of **Dirichlet densities**

\[
\pi_0(\theta) = \frac{\Gamma(\alpha)}{\prod_{i=1}^d \Gamma(\alpha_i)} \prod_{i=1}^d \theta_i^{\alpha_i-1}, \quad \alpha_i > 0, \quad \sum_{i=1}^d \alpha_i = \alpha
\]  

(18.4)

where \(\theta = (\theta_1, \ldots, \theta_d)\) with \(\sum_{i=1}^d \theta_i = 1\) is a conjugate prior for **multinomial sampling**:

\[
\Pr(X = i | \theta) = \theta_i, \quad X \in \{1, 2, \ldots, d\}
\]  

(18.5)

**Exercise:** Prove that (18.4) is a conjugate prior for the likelihood (18.5).

The advantage of conjugate priors in the MH algorithm is that if one knows how to generate independent random variates \(\theta_n\) efficiently from the prior family (for example, the gamma density (18.1) or the beta density (18.3)), then one can do Monte Carlo sampling from the posterior density for any observed \(X\). This means that we always have available at least one numerically efficient candidate for MH sampling. If \(\theta\) is one part of a larger parameter vector and the prior family is the conditional distribution, this would be a Gibbs sampler step. Efficient methods are available for simulating independent random variables with arbitrary gamma, beta, normal, uniform, and exponential distributions. See for example Devroye (1986), Press et al. (1992), and Fishman (1995) in the references.
19. Hidden Variables. In many applications, the likelihood \( L(\theta, X) \) is complex but would be much simpler if an unobserved random variable \( Y \) could be fixed.

For example, suppose that we are given DNA sequence data \( \mathcal{D} \) from \( n \) individuals from the same species. Assume that \( \mathcal{D} \) depends only on the DNA sequence of the most recent common ancestor (MRCA) of the \( n \) individuals and the mutation rate on the pedigree \( Y \) that connects the \( n \) individuals with their MRCA. Assume that there are no repeat mutations at the same site in the pedigree and let \( X \) be the number of DNA sites in the sample that are polymorphic; that is, at which there is more than one base in the \( n \) sequences.

Under reasonable assumptions about DNA mutation, the likelihood \( L(\theta, X) \) for the number of polymorphic sites \( X \) is Poisson with mean \( \theta \text{Len}(Y) \), where \( \text{Len}(Y) \) is the total length of the pedigree \( Y \). Thus

\[
L(\theta, X) = \int_Y L_P(X, \theta \text{Len}(Y)) \pi(\theta, Y) dY \tag{19.1}
\]

where \( L_P(X, A) \) in (19.1) is the Poisson likelihood (18.1) with mean \( A \) (with the variables in the opposite order). The integral \( \pi(\theta, Y) dY \) is the integral over all possible pedigrees of a sample of \( n \) individuals, which is very complex.

If we could observe \( Y \), then we could use the much simpler likelihood \( L_P(X, \theta \text{Len}(Y)) \). However, \( Y \) is not observed. (That is, the pedigree \( Y \) is a hidden variable.)

The MH procedures of the preceding sections can be implemented in this problem in either of two different ways. The direct approach is to apply Theorem 5.1 directly and obtain a Markov chain \( \theta_n \) with stationary distribution

\[
f(\theta) = \pi_0(\theta)L(\theta, X) \tag{19.2}
\]

This requires that we work with the more complex likelihood \( L(\theta, X) \), whose evaluation might require either dozens of integrals or else Monte Carlo simulation in \( Y \) for each proposed value of \( \theta \).

Alternatively, we can consider the pedigree \( Y \) or the total pedigree length \( \text{Len}(Y) \) as a unknown parameter instead of an unobserved random variable. A useful feature of Bayesian methods is that there is no clear distinction between parameters and variables. Parameters are treated essentially the same as if they were unobserved data.

If we want to treat a random variable \( Y \) as a parameter instead of unobserved data, the first step is to choose an appropriate prior for \( Y \). Note that since it is a random variable, it has a density

\[
P(Y \in dy) = f(\theta, y) dy
\]

that in general depends on \( \theta \). We also have to decide how to update both variables \((\theta, Y)\) during each step of an MH Markov chain \( X_n \) for \( Z = (\theta, Y) \). The first problem is answered by

**Theorem 19.1.** Let \( \pi_0(\theta) \) be a probability density in \( \theta \) for a likelihood \( L(\theta \mid X) \) that satisfies

\[
L(\theta \mid X) = \int L(\theta, y, X) f(\theta, y) dy \tag{19.3}
\]

where

\[
\int L(\theta, y, x) dx = 1 \quad \text{for all } (\theta, y) \quad \text{and}
\]

\[
\int f(\theta, y) dy = 1 \quad \text{for all } \theta
\]

Now consider a different probability model with parameters \( z = (\theta, y) \) where the prior distribution and likelihood for \( z \) are

\[
\pi_0(z) = \pi_0(\theta)f(\theta, y), \quad z = (\theta, y)
\]

\[
L(z, X) = L(\theta, y, X) \tag{19.4}
\]
Let \( \pi_1(\theta, y) = \pi_1(\theta, y \mid X) \) be the posterior distribution for (19.4) for an observed value of \( X \). Then the marginal density of the posterior distribution \( \pi_1(\theta, y) \) in \( \theta \) is

\[
\int \pi_1(\theta, y \mid X) \, dy = \pi_1(\theta \mid X) \tag{19.5}
\]

That is, the marginal density in \( \theta \) for \( \pi_1(z \mid X) \) for \( z = (\theta, Y) \) is exactly the same as the original posterior density \( \pi(\theta \mid X) \).

Theorem 19.1 says that if we enlarge the parameter space from \( \theta \) to \( (\theta, y) \) and extend the prior distribution \( \pi_0(\theta) \) and likelihood \( L(\theta \mid X) \) as in (19.4), with in particular the prior of \( Y \) being its probability density in the original model, then we obtain exactly the same posterior distribution \( \pi_1(\theta) \) for \( \theta \). In this sense, treating the “hidden variable” \( Y \) as a parameter instead of a random variable leads to exactly the same model for \( \theta \) as long as we specify the appropriate prior in (19.4) for \( (\theta, Y) \).

**Proof.** By definition

\[
\pi_1(\theta, y \mid X) = \frac{\pi_0(\theta) L(\theta, y, X) f(y)}{\int \pi_0(s) L(s, w, X) f(s, w) \, ds \, dw} \tag{19.6}
\]

This implies

\[
\int \pi_1(\theta, y \mid X) \, dy = \frac{\pi_0(\theta) \int L(\theta, y, X) f(y) \, dy}{\int \pi_0(s) \int L(s, w, X) f(s, w) \, ds \, dw} = \frac{\pi_0(\theta) L(\theta, X)}{\int \pi_0(s) L(s, X) \, ds} = \pi_1(\theta \mid X)
\]

by (19.3). This is all there is to the proof of Theorem 19.1.

**Remarks.** (1) The Metropolis-Hastings Markov chain \( X_n = (\theta_n, Y_n) \) in Theorem 16.1 provides not only estimates of \( \theta \) but also provides estimates of the conditional distribution of \( Y \) given \( X \). That is, any MCMC that treats a “hidden variable” \( Y \) as a parameter also provides estimates of \( Y \). If the hidden variable \( Y \) is of high dimension, as is often the case, this provides a great deal of additional information.

(2) By Theorem 19.1, Metropolis-Hastings Markov chains \( Z_n = (\theta_n^1, Y_n) \) and \( \theta_n^2 \) can both be used to estimate the same posterior density of \( \theta \). The difference is that \( \theta_n^2 \) requires us to carry out the integral in (19.3) for each evaluation of the likelihood \( L(\theta, X) \) which \( \theta_n^1, Y_n \) does not.

In effect, in (19.3), each value \( Y_n \) is a single Monte Carlo simulation of the integral (19.3) for the current value \( \theta_n \) of \( \theta \). When we would not normally evaluate an integral by a single Monte Carlo simulation, the fact that we are averaging over a long trajectory \( \{\theta_n^1, Y_n\} \) corrects for this. Since \( \theta \) and functions of \( \theta \) are being estimated by the average of a large number of simulated values \( \theta_n^1 \) or \( \theta_n^2 \), the fact that each of the integrals (19.3) is evaluated with just one Monte-Carlo simulation appears to be sufficient. In practice, hidden-variable Markov chain \( Z_n = (\theta_n^1, Y_n) \) are nearly always almost as efficient as the “integrated” chain \( \theta_n^2 \), and it is difficult to find examples in which the chain \( \theta_n^1, Y_n \) provides a noticeably less accurate estimate of \( \theta \).

(3) We still have to find an updating procedure for \( Z = (\theta, Y) \) that is, a proposal function \( q(w, z) \) for \( w = (\theta, y) \) and \( z = (\theta, y_1) \). If we already have an updating procedure for \( \theta \), we can use this as a block update for \( \theta \) and we just have to find a block update for \( Y \).

**20. Knowing When to Stop.** Often MH trajectories that start in different parts of the parameter space may stay diverged for a very long period of time. One particular way of measuring the extent to which this may happen is the Gelman scale reduction factor (Gelman et al. 1995, p.331–332, Gilks et al. 1996, specifically Section 8.4 in Chapter 8, p.136–139).
In some cases, computation of the function values \( f(X_n) \) is more expensive than generating the values \( X_n \). We want to assume that the sampled values \( f(X_n) \) are approximately independent, or we want to store fewer values for later estimation of medians and credible intervals. In these situations, we can sample values \( f(X_n) \) only every \( K \)th step. This is equivalent to oversampling the full Markov chain \( X_n \). In this case, we distinguish between samples or sampled values and the steps or iterations of the MC Markov chain.

For definiteness, suppose that we run \( J \) trajectories each of (iteration) length \( nK \), so that each trajectory has \( n \) samples. We assume \( 1 < J \ll n \) and \( K \geq 1 \). These could be either different runs with different starting positions or else consecutive blocks of values in a single run of length \( nKJ \) iterations.

For a particular parameter of interest, let \( \psi_{ij} \) be the \((iK)\)th iteration (or the \(i\)th sampled value) in the \(j\)th trajectory, where \(1 \leq j \leq J\) for separate runs and \(1 \leq i \leq n\) within each trajectory. (That is, \( \psi_{ij} = F(X_{iK,j}) \) for a function \( F(x) \) and \( J \) copies \( X_j(i) = X_{i,j} \) of the MH process in Section 2.) Let
\[
a_j = \frac{1}{n} \sum_{i=1}^{n} \psi_{ij} = \bar{\psi}_{+j}, \quad 1 \leq j \leq J
\]
be the sample mean of the \(j\)th trajectory. Let \( b_j = E(\psi_{ij}) \) be the corresponding theoretical mean. Then
\[
B = \frac{n}{J-1} \sum_{j=1}^{J} (a_j - \bar{a})^2 \quad \text{for} \quad \bar{a} = \frac{1}{J} \sum_{j=1}^{J} a_j
\]
is the numerator of the one-way ANOVA F-test of \( H_0: b_j = b_0 \), which we can call \( B = \text{MSMod} \). In particular
\[
E(B) = \sigma^2 + \frac{n}{J-1} \sum_{j=1}^{J} (b_j - \bar{b})^2 \quad \text{for} \quad \bar{b} = \frac{1}{J} \sum_{j=1}^{J} b_j
\]
if each set of sampled values \( \psi_{ij} \) \((1 \leq i \leq n)\) are approximately uncorrelated with variance \( \sigma^2 \). We call \( B \) the between-sequence variance. Similarly
\[
W = \frac{1}{J} \sum_{j=1}^{J} s_j^2 \quad \text{for} \quad s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\psi_{ij} - a_j)^2
\]
is the denominator MSE of the same one-way ANOVA test. It satisfies \( E(W) = \sigma^2 \) if each trajectory \( \psi_{ij} \) is uncorrelated for fixed \( j \). This is called the within-sequence variance. Then
\[
\frac{E(B) - E(W)}{n} = \frac{1}{J-1} \sum_{j=1}^{J} (b_j - \bar{b})^2 = s_b^2
\]
(20.1)
is the sample variance of the \( J \) theoretical trajectory means \( b_j \). Gelman’s scale-reduction factor is
\[
\hat{GR} = \frac{W + \frac{B-W}{n}}{W} = 1 + \frac{(B-W)}{nW} = 1 + \frac{(F-1)}{n} = \frac{n-1}{n} + \frac{F}{n}
\]
(20.2)
where \( F = \text{MSMod} / \text{MSE} \) is the one-way ANOVA test statistic. This can be viewed an estimator of \((\sigma^2 + s_b^2) / \sigma^2 = 1 + s_b^2 / \sigma^2 \) in (20.1).
Note that one can have \( \hat{GR} < 1 \) or \( B - W < 0 \) due to sampling variation. In fact, if the MH Markov chains converges very strongly, then \( \psi_{i,j} \) for \( i = 1, 2, \ldots \) will be negatively correlated in \( i \) due to their being bound to this stationary distribution. In this case, \( B \) can underestimate the variance \( \sigma^2 \) resulting in \( B < W \).

If \( \hat{GR} \) is close to one, this may be an indication that all \( J \) trajectories are sampling from the stationary distribution in an unbiased manner. Gelman suggests, “In practice, we generally run the simulations until the values of [\( GR \)] are all less than 1.1 or 1.2” (Gelman 1996, p138). Of course, the same calculation must be done for every sampled function \( f(x) \) of interest, since some parameters in an MCMC simulation may converge much faster than others.

A related measure is

\[
R^2 = \frac{n \sum_{j=1}^{J} (a_j - \bar{a})^2}{\sum_{j=1}^{J} \sum_{i=1}^{n} (\psi_{ij} - \bar{a})^2} = \frac{(J - 1)B}{J(n - 1)W + (J - 1)B} \tag{20.3}
\]

For statistical regressions, this is called the “proportion of the total variability of the \( \psi_{ij} \)” that is “explained” by the subchain or trajectory means. Typically \( R^2 \) is small if and only if \( \hat{GR} \) is close to one. Often \( \hat{GR} < 1 \) and \( R^2 \) is tiny when an MCMC Markov chain converges strongly.

References.

