1. MCMC: A “New” Approach to Simulation

- Consider the general problem of trying to calculate characteristics of a complicated multivariate probability distribution $f(x)$ on $x = (x_1, \ldots, x_p)$.

- For example, suppose we want to calculate the mean of x_1,

$$\int\int x_1 f(x_1, x_2) dx_1 dx_2$$

where

$$f(x_1, x_2) \propto (1 + x_1^2)^{-1} x_2^{-n} \exp\left\{ -\frac{1}{2x_2^2} \sum_i (y_i - x_1)^2 - x_2 \right\}$$

($y_1, \ldots, y_n$ are fixed constants). Bad news: This calculation is analytically intractable.

- A Monte Carlo approach: Simulate k observations $x^{(1)}, \ldots, x^{(k)}$ from $f(x)$ and use this sample to estimate the characteristics of interest. (Careful: Each $x^{(j)} = (x_1^{(j)}, \ldots, x_p^{(j)})$ is a multivariate observation). For example, we could estimate the mean of x_1 by

$$\bar{x}_1 = \frac{1}{k} \sum_j x_1^{(j)}.$$

- If $x^{(1)}, \ldots, x^{(k)}$ were independent observations (i.e. an iid sample), we could use standard central limit theorem results to draw inference about the quality of our estimate.

- Bad news: In many problems, methods are unavailable for direct simulation of an iid sample from $f(x)$.

- Good news: In many problems, methods such as the Gibbs sampler and the Metropolis-Hastings algorithms can be used to simulate a Markov chain $x^{(1)}, \ldots, x^{(k)}$ which is converging in distribution to $f(x)$, (i.e. as k increases, the distribution of $x^{(k)}$ gets closer and closer to $f(x)$).

- Recall that a Markov chain $x^{(1)}, \ldots, x^{(k)}$ is a sequence such that for each $j \geq 1$, $x^{(j+1)}$ is sampled from a distribution $p(x \mid x^{(j)})$ which depends on $x^{(j)}$ (but not on $x^{(1)}, \ldots, x^{(j-1)}$).

- The function $p(x \mid x^{(j)})$ is called a Markov transition kernel. If $p(x \,|\, x^{(j)})$ is time-homogeneous (i.e. $p(x \,|\, x^{(j)})$ does not depend on j) and the transition kernel satisfies

$$\int p(x \mid x^*) f(x^*) dx^* = f(x),$$

then the chain will converge to $f(x)$ if it converges at all.

- Simulation of a Markov chain requires a starting value $x^{(0)}$. If the chain is converging to $f(x)$, then the dependence between $x^{(j)}$ and $x^{(0)}$ diminishes as j increases. After a suitable "burn in" period of l iterations, $x^{(l)}, \ldots, x^{(k)}$ behaves like a dependent sample from $f(x)$.

- Such behavior is illustrated by Figure 1.1 on page 6 of Gilks, Richardson & Spieglehalter (1995).

- The output from such simulated chains can be used to estimate the characteristics of $f(x)$. For example, one can obtain approximate iid samples of size m by taking the final $x^{(k)}$ values from m separate chains.

- It is probably more efficient, however, to use all the simulated values. For example, $\bar{x}_1 = \frac{1}{k}\sum_j x_1^{(j)}$ will still converge to the mean of x_1.

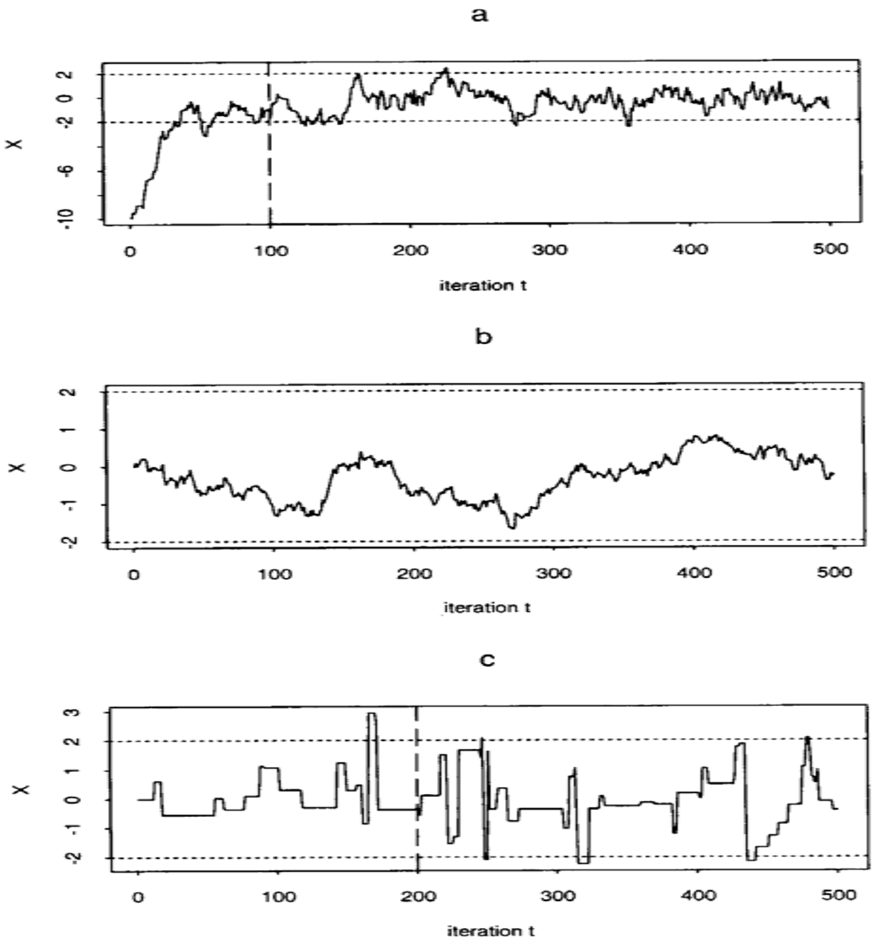

Figure 1.1 *500 iterations from Metropolis algorithms with stationary distribution* N(0, 1) *and proposal distributions (a)* $q(.|X) = \mathrm{N}(X, 0.5)$; *(b)* $q(.|X) = \mathrm{N}(X, 0.1)$; *and (c)* $q(.|X) = \mathrm{N}(X, 10.0)$. *The burn-in is taken to be to the left of the vertical broken line.*

- MCMC is the general procedure of simulating such Markov chains and using them to draw inference about the characteristics of $f(x)$.

- Methods which have ignited MCMC are the Gibbs sampler and the more general Metropolis-Hastings algorithms. As will we now see, these are simply prescriptions for constructing a Markov transition kernel $p(x|x^*)$ which generates a Markov chain $x^{(1)}, \ldots, x^{(k)}$ converging to $f(x)$.

2. The Gibbs Sampler (GS)

- The GS is an algorithm for simulating a Markov chain $x^{(1)}, \ldots, x^{(k)}$ which is converging to $f(x)$, by successively sampling from the full conditional component distributions $f(x_i|x_{-i})$, $i = 1, \ldots, p$, where x_{-i} denotes the components of x other than x_i.

- For simplicity, consider the case where $p = 2$. The GS generates a Markov chain

  $$(x_1^{(1)}, x_2^{(1)}), (x_1^{(2)}, x_2^{(2)}), \ldots, (x_1^{(k)}, x_2^{(k)})$$

  converging to $f(x_1, x_2)$, by successively sampling

  $$\begin{array}{lcl} x_1^{(1)} & \text{from} & f(x_1 \,|\, x_2^{(0)}) \\ x_2^{(1)} & \text{from} & f(x_2 \,|\, x_1^{(1)}) \\ x_1^{(2)} & \text{from} & f(x_1 \,|\, x_2^{(1)}) \\ & \vdots & \\ x_1^{(k)} & \text{from} & f(x_1 \,|\, x_2^{(k-1)}) \\ x_2^{(k)} & \text{from} & f(x_2 \,|\, x_1^{(k)}) \end{array}$$

  (To get started, prespecify an initial value for $x_2^{(0)}$).

- For example, suppose

$$f(x_1, x_2) \propto \left(\begin{array}{c} n \\ x_1 \end{array} \right) x_2^{x_1+\alpha-1}(1-x_2)^{n-x_1+\beta-1}$$

$$x_1 = 0, 1, \ldots, n, \quad 0 \le x_2 \le 1.$$

  The GS proceeds by successively sampling from

$$\begin{aligned} f(x_1 \,|\, x_2) &= \text{Binomial}(n, x_2) \\ f(x_2 \,|\, x_1) &= \text{Beta}(x_1+\alpha, n-x_1+\beta) \end{aligned}$$

- To illustrate the GS for the above, Figure 1 of Casella & George (1992) presents a histogram of a sample of $m = 500$ final values of x_1 from separate GS runs of length $k = 10$ when $n = 16$, $\alpha = 2$ and $\beta = 4$. This is compared with an iid sample from the actual distribution $f(x_1)$, (which here can be shown to be Beta-Binomial).

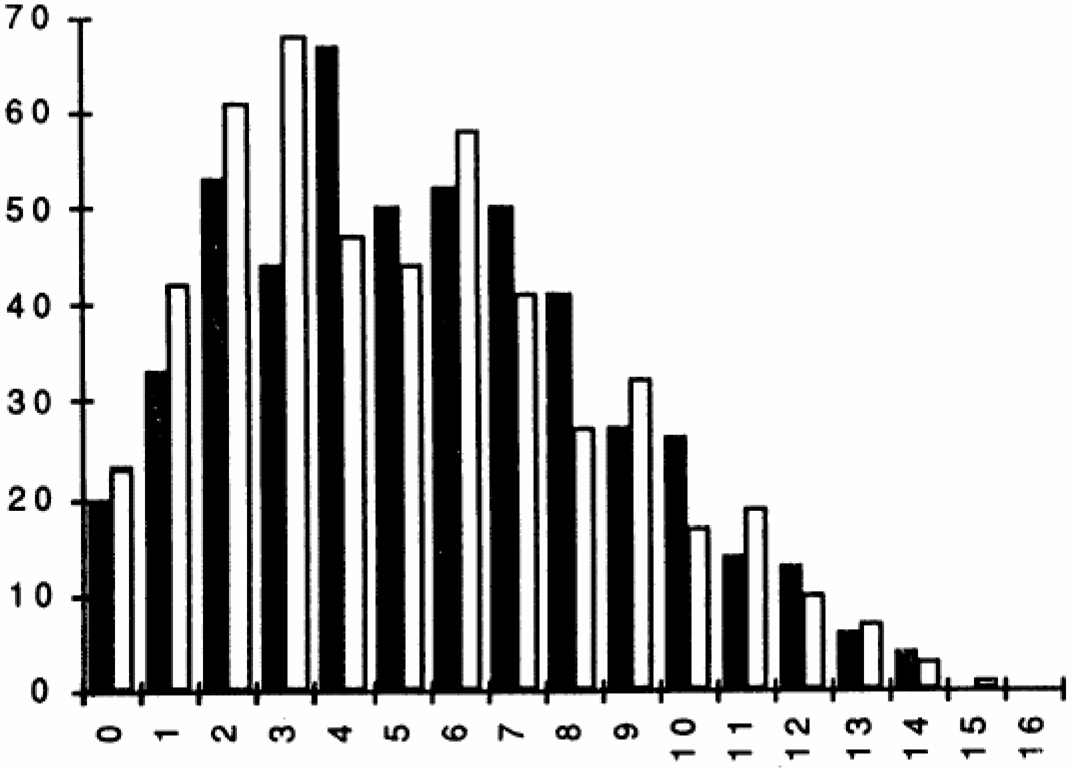

Figure 1. Comparison of Two Histograms of Samples of Size $m = 500$ From the Beta-Binomial Distribution With $n = 16$, $\alpha = 2$, and $\beta = 4$. The black histogram sample was obtained using Gibbs sampling with $k = 10$. The white histogram sample was generated directly from the beta-binomial distribution.

- Note that $f(x_1) = \int f(x_1, x_2)dx_2 = \int f(x_1 \mid x_2)f(x_2)dx_2$. This expression suggests that an improved estimate of $f(x_1)$ in this example can be obtained by inserting the m values of $x_2^{(k)}$ into

  $$\hat{f}(x_1) = \frac{1}{m}\sum_{i=1}^{m} f(x_1 \mid x_2^{(i)}).$$

  Figure 3 of Casella & George (1992) illustrates the improvement obtained by this estimate.

- Note that the conditional distributions for the above setup, the Binomial and the Beta, can be simulated by routine methods. This is not always the case. For example, $f(x_1 \mid x_2)$ from page 2 is not of standard form. Fortunately, such distributions can be simulated using envelope methods such as rejection sampling, the ratio-of-uniforms method or adaptive rejection resampling. As we'll see, Metropolis-Hastings algorithms can also be used for this purpose.

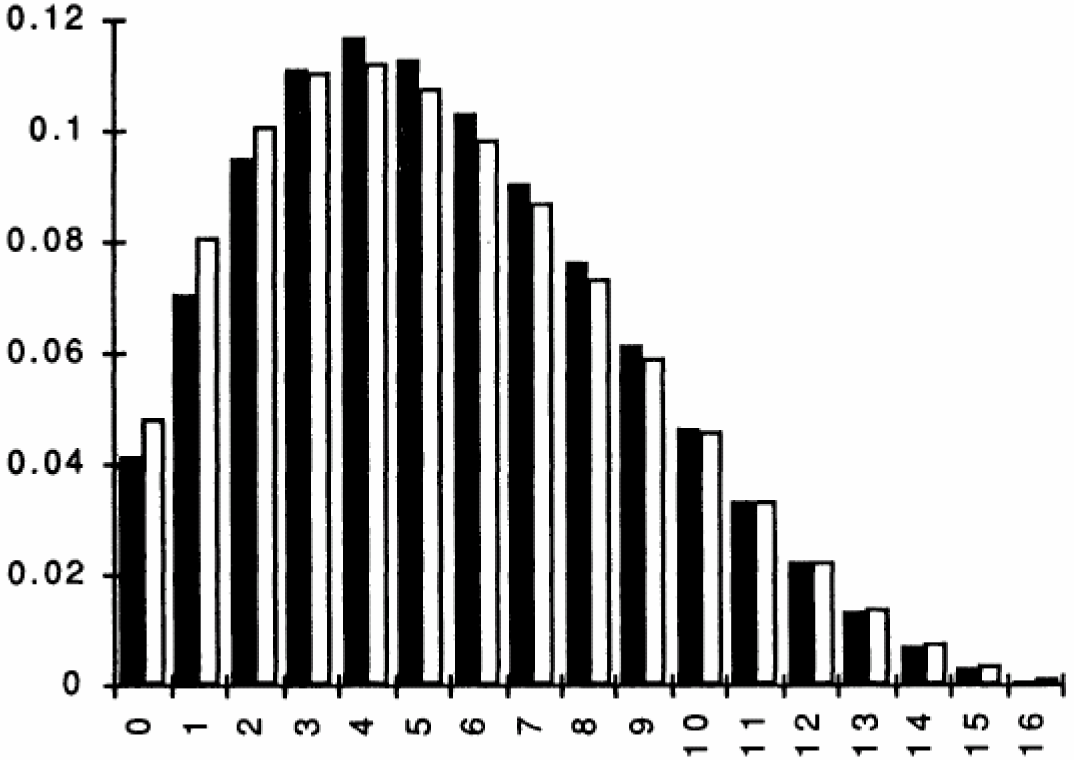

Figure 3. Comparison of Two Probability Histograms of the Beta-Binomial Distribution With n = 16, α = 2, and β = 4. The black histogram represents estimates of the marginal distribution of X using Equation (2.11), based on a sample of Size m = 500 from the pair of conditional distributions in (2.6). The Gibbs sequence had length k = 10. The white histogram represents the exact beta-binomial probabilities.

3. Metropolis-Hastings Algorithms (MH)

- MH algorithms generate Markov chains which converge to $f(x)$, by successively sampling from an (essentially) arbitrary proposal distribution $q(x|x^*)$ (i.e. a Markov transition kernel) and imposing a random rejection step at each transition.

- An MH algorithm for a candidate proposal distribution $q(x \mid x^*)$, entails simulating $x^{(1)}, \ldots, x^{(k)}$ as follows:
  - Simulate a transition candidate x^C from $q(x \mid x^{(j)})$
  - Set $x^{(j+1)} = x^C$ with probability

$$\alpha(x^{(j)}, x^C) = \min \left\{ 1, \frac{q(x^{(j)} \mid x^C)}{q(x^C \mid x^{(j)})} \frac{f(x^C)}{f(x^{(j)})} \right\}$$

Otherwise set $x^{(j+1)} = x^{(j)}$.

- The original Metropolis algorithm was based on symmetric q, (i.e. $q(x \mid x^*) = q(x^* \mid x)$), for which α is of the simple form

$$\alpha(x^{(j)}, x^C) = \min\left\{1, \frac{f(x^C)}{f(x^{(j)})}\right\}.$$

- If $q(x \mid x^*)$ is chosen such that the Markov chain satisfies modest conditions (e.g. irreducibility and aperiodicity), then convergence to $f(x)$ is guaranteed. However, the rate of convergence will depend on the relationship between $q(x \mid x^*)$ and $f(x)$.

- When x is continuous, a popular choice for $q(x \mid x^*)$ is $x = x^* + z$ where $z \sim N_p(0, \Sigma)$. The resulting chain is called a random walk chain. Note that the choice of scale Σ can critically affect the mixing (i.e. movement) of the chain. Figure 1.1 on page 6 of Gilks, Richardson & Spieglehalter (1995) illustrates this when $p = 1$. Other distributions for z can also be used.

- Another useful choice, called an independence sampler, is obtained when the proposal $q(x \mid x^*) = q(x)$ does not depend on x^*. The resulting α is of the form

  $$\alpha(x^{(j)}, x^C) = \min\left\{1, \frac{q(x^{(j)})}{q(x^C)} \frac{f(x^C)}{f(x^{(j)})}\right\}.$$

  Such samplers work well when $q(x)$ is a good heavy-tailed approximation to $f(x)$.

- It may be preferable to use an MH algorithm which updates the components $x_i^{(j)}$ of x one at a time. It can shown that the Gibbs sampler is just a special case of such a single-component MH algorithm where q is chosen so that $\alpha \equiv 1$.

- Finally, to see why MH algorithms work, it is not too hard to show that the implied transition kernel $p(x \mid x^*)$ of any MH algorithm satisfies
$$p(x \mid x^*)f(x^*) = p(x^* \mid x)f(x),$$
a condition called detailed balance or reversibility. Integrating both sides of this identity with respect to x^* yields
$$\int p(x \mid x^*)f(x^*)dx^* = f(x),$$
showing that $f(x)$ is the limiting distribution when the chain converges.

4. The Model Liberation Movement

- Advances in computing technology have unleashed the power of Monte Carlo methods, which in turn, are now unleashing the potential of statistical modeling.

- Our new ability to simulate from complicated multivariate probability distributions via MCMC is having impact in many areas of Statistics, but most profoundly for Bayesian approaches to statistical modeling.

- The Bayesian paradigm uses probability to characterize **ALL** uncertainty as follows:

  - $Data$ is a realization from a model $p(Data \mid \Theta)$, where Θ is an unknown (possibly multivariate) parameter.
  - Θ is treated as a realization from a prior distribution $p(\Theta)$.
  - Post-data inference about Θ is based on the posterior distribution

$$p(\Theta \mid Data) = \frac{p(Data \mid \Theta)p(\Theta)}{\int p(Data \mid \Theta)p(\Theta)d\Theta}$$

- In the past, analytical intractability of the expression for $p(\Theta|Data)$ severely stymied realistic practical Bayesian methods. Unrealistic, oversimplified models were too often used to facilitate calculations. MCMC has changed this, and opened up vast new realms of modeling possibilities.

- My initial example

$$f(x_1, x_2) \propto (1 + x_1^2)^{-1} x_2^{-n} \exp\left\{ -\frac{2}{x_2^2} \sum_i (y_i - x_1)^2 - x_2 \right\}$$

was a just a disguised posterior distribution for the Bayesian setup

$$y_1, \ldots, y_n \text{ iid} \sim N(\mu, \sigma^2)$$

$$\mu \sim \text{Cauchy}(0, 1) \qquad \sigma \sim \text{Exponential}(1).$$

The posterior of the parameters μ and σ is

$$p(\mu, \sigma \mid Data) \propto (1 + \mu^2)^{-1} \sigma^{-n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_i (y_i - \mu)^2 - \sigma \right\}.$$

- In the above example, $f(x)$ can only be specified up to a norming constant. This is typical of Bayesian formulations. A huge attraction of GS and MH algorithms is that these norming constants are not needed.

- The previous example is just a toy problem. MCMC is in fact enabling posterior calculation for extremely complicated models with hundreds and even thousands of parameters.

- Going even further, the Bayesian approach can be used to obtain posterior distributions over model spaces. Under such formulations, MCMC algorithms are leading to new search engines which automatically identify promising models.

References For Getting Started

Casella, G. & George, E.I. (1992) Explaining the Gibbs Sampler, *The American Statistician*, 46, 167-174.

Chib, S. & Greenberg, E. (1995) Understanding the Metropolis-Hastings Algorithm, *The American Statistician*, 49, 327-335.

Gilks, W. R., Richardson, S. & D.J. Spieglehalter (1995) *Markov Chain Monte Carlo in Practice*, Chapman & Hall, London.

Robert, C.P. & Casella, G. (2004) *Monte Carlo Statistical Methods, 2nd Edition*, Springer, New York.