COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

STAT 4224/5224

Bayesian Statistics

Dobrin Marchev

Recall: Gibbs Sampler

- Suppose you have a vector of parameters $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)$.
- Given the current values $\phi_1^{(t)}, \ldots, \phi_p^{(t)}$, obtain new values:

$$\phi_1^{(t+1)} \sim f(\phi_1 \mid \phi_2^{(t)}, \phi_3^{(t)}, \ldots, \phi_p^{(t)})$$
$$\phi_2^{(t+1)} \sim f(\phi_2 \mid \phi_1^{(t+1)}, \phi_3^{(t)}, \ldots, \phi_p^{(t)})$$
$$\phi_3^{(t+1)} \sim f(\phi_3 \mid \phi_1^{(t+1)}, \phi_2^{(t+1)}, \ldots, \phi_p^{(t)})$$
$$\ldots$$
$$\phi_p^{(t+1)} \sim f(\phi_p \mid \phi_1^{(t+1)}, \phi_2^{(t+1)}, \ldots, \phi_{p\text{-}1}^{(t+1)})$$

- The algorithm generates a *dependent* sequence of vectors $\boldsymbol{\phi}^{(t)} = (\phi_1^{(t)}, \phi_2^{(t)}, \phi_3^{(t)}, \ldots, \phi_p^{(t)})$, $t = 1, \ldots, S$.
- Under mild regularity conditions it can be shown that $(\phi_1^{(t)}, \phi_2^{(t)}, \phi_3^{(t)}, \ldots, \phi_p^{(t)})$ *converges* to a draw from the target f.

Exercise 1

Consider the Binomial – Beta – Poisson model:

$$X|n,\theta \sim Binom(n,\theta)$$
$$n|\theta \sim Poisson(\lambda)$$
$$\theta \sim Beta(a,b)$$

Derive the distributions of $\theta|x,n$ and $\theta|x,n$ and simulate a sample from the joint posterior distribution $f(\theta,n|x)$, using a two-stage Gibbs sampler.

Answers:

$\theta|x,n$ is Beta

$\theta|x,n$ is shifted Poisson

Gibbs Sampler Properties

- What does it mean that $(\phi_1^{(t)}, \phi_2^{(t)}, \phi_3^{(t)}, \ldots, \phi_p^{(t)})$ converges to a draw from the target f?
- Formally, the statement is

$$P\left(\boldsymbol{\phi}^{(t)} \in A\right) \rightarrow \int_A f(\boldsymbol{\phi}) d\,\boldsymbol{\phi}\,, as\ t \rightarrow \infty$$

regardless of the starting value $\boldsymbol{\phi}^{(0)}$

- More importantly, for most functions g we have

$$\frac{1}{S}\sum_{i=1}^{S} g\left(\boldsymbol{\phi}^{(i)}\right) \rightarrow E\left(g(\boldsymbol{\phi})\right) = \int g(\boldsymbol{\phi}) f(\boldsymbol{\phi}) d\boldsymbol{\phi}$$

- Notice that $\boldsymbol{\phi}^{(t)}$ depends on the past $\boldsymbol{\phi}^{(0)}, \ldots, \boldsymbol{\phi}^{(t-1)}$ only through $\boldsymbol{\phi}^{(t-1)}$. Such sequences are called *Markov chains*.

What is a Markov chain?

- Most statistical models assume an i.i.d. random sample
- How can we introduce a dependence structure between observations $X_1, \ldots, X_n$
- This can be done with two approaches: *Markov chains* or time series
- Markov models provide a very rich set of tools for handling dependence
- Let's look at an example where the space is discrete.

Prof. Andrei A. Markov (1856-1922) , published his result in 1906.

Example 1: The taxi problem

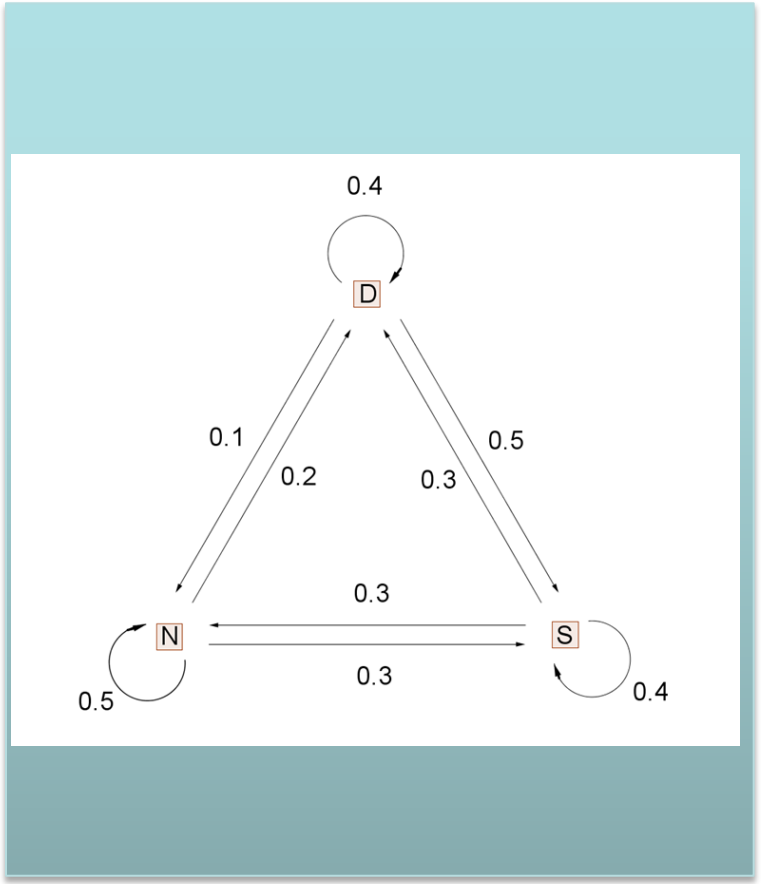
A taxi company has divided the city into three regions – Northside, Downtown, and Southside. By keeping track of pickups and drop-offs, the company has found that of the fares picked up in Northside, 50% stay in that region, 20% are taken to Downtown, and 30% go to Southside. Of the fares picked up Downtown, only 10% go to Northside, 40% stay in Downtown, and 50% go to Southside. Of the fares picked up in Southside, 30% go to each of Northside and Downtown, while 40% stay in Southside.

We would like to know what the distribution of taxis will be over time as they pick up and drop off successive fares. Suppose we want to know the probability that a taxi starting off Downtown, will be Downtown after letting off its seventh fare?

State Diagram

This information can be represented in a state diagram which includes:

- the three states D, N, and S corresponding to the three regions of the city
- the probabilities of a taxi transitioning from one region/state to another



Markov chains

If the location of the taxi at time *n* is denoted X_n, then the sequence X_1, X_2, … consists of *dependent* variables, and can be modeled with a Markov chain. The values X_n can take are known as *states* of the chain.

The probabilities of moving from state to state are constant and independent of the past behavior – this property of the system is called the Markov property. That is

$$P(X_n = s_n | X_{n-1} = s_{n-1}, \dots, X_0 = s_0) = P(X_n = s_n | X_{n-1} = s_{n-1})$$

We assume that a transition – picking up and dropping off a fare – occurs each time the system is observed, and that observations occur at regular intervals. Systems with these characteristics are called Markov chains or Markov processes (when time is continuous).

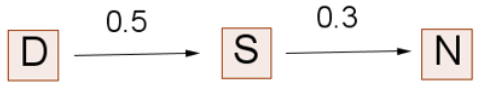
Computing probabilities

- What is the probability that a taxi that starts off Downtown ends up in Northside after two fares?

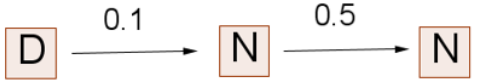
- One possibility is that the taxi stays Downtown for the first fare, and then transitions to Northside for the second. The probability of this occurring is then:

$$= (0.4)(0.1) = 0.04$$

- But we could also have the taxi going to either Northside of Southside first, then transitioning to Northside:



$$= (0.5)(0.3) = 0.15$$



$$= (0.1)(0.5) = 0.05$$

Since the taxi could follow the first, second or third path, the probability of starting and ending Downtown after two fares is

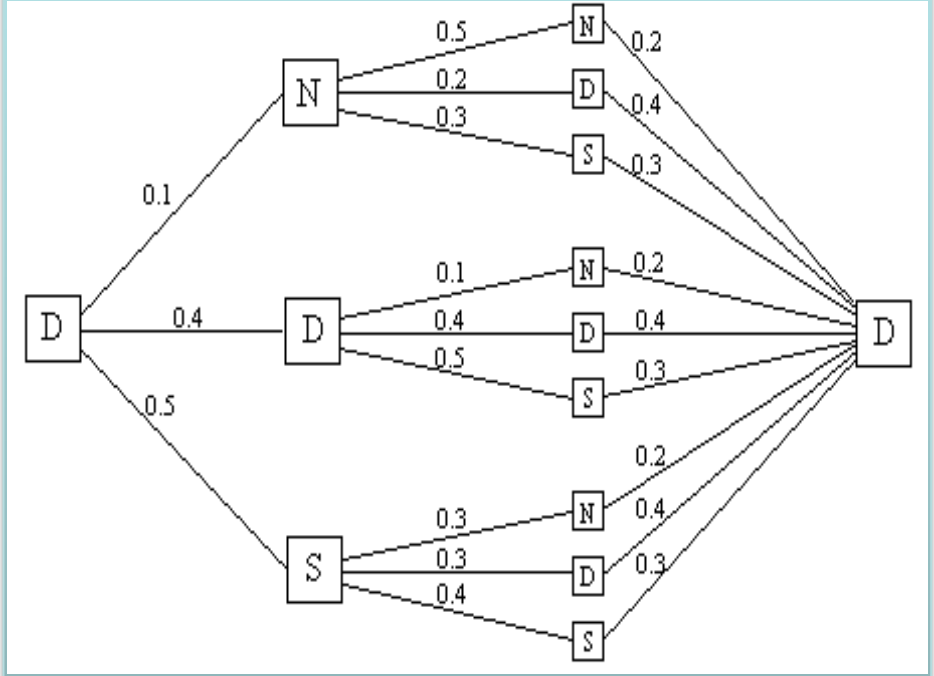
But we could also have the taxi going to either Northside of Southside first, then transitioning to Northside:

$$(0.4)(0.1) + (0.5)(0.3) + (0.1)(0.5) = 0.24$$

$$D \to D \to N \ \underline{\text{or}} \ D \to S \to N \ \underline{\text{or}} \ D \to N \to N$$

Transition in more steps

Tree Diagram



If we multiply along all the paths and sum the results, we find that this probability is 0.309.

- If we want to know the probability of a taxi transitioning from one region to another after just three fares, the computation will have more possible paths.
- Suppose we were interested in the probability of a taxi both starting and ending up Downtown. We can use a tree diagram to represent this calculation.

Transition Matrix

- We can create a square matrix, *P*, called the **transition matrix,** by constructing rows for the probabilities going from Southside and Northside as well.

To

From

$$\begin{array}{c} \\ D \\ S \\ N \end{array} \begin{array}{c} \begin{array}{ccc} D & S & N \end{array} \\ \begin{bmatrix} 0.4 & 0.5 & 0.1 \\ 0.3 & 0.4 & 0.3 \\ 0.2 & 0.3 & 0.5 \end{bmatrix} \end{array}$$

- An entry P_{ij} of this matrix is the probability of a transition from region *i* to region *j*. For example, p_{32}, is the probability of a fare that originates in Northside going to Southside. (Note the sum of entries across rows of *P*.)

Transition Matrix

What results when we multiply the transition matrix by itself?

$$P^2 = \begin{bmatrix} 0.4 & 0.5 & 0.1 \\ 0.3 & 0.4 & 0.3 \\ 0.2 & 0.5 & 0.5 \end{bmatrix}^2$$

$$= \begin{bmatrix} 0.4 & 0.5 & 0.1 \\ 0.3 & 0.4 & 0.3 \\ 0.2 & 0.5 & 0.5 \end{bmatrix} \times \begin{bmatrix} 0.4 & 0.5 & 0.1 \\ 0.3 & 0.4 & 0.3 \\ 0.2 & 0.5 & 0.5 \end{bmatrix} = \begin{bmatrix} 0.33 & 0.43 & 0.24 \\ 0.3 & 0.4 & 0.3 \\ 0.27 & 0.37 & 0.26 \end{bmatrix}$$

The highlighted entry results from the same computation that we already considered for of a taxi going from D to N in two fares.

What are the other entries of P^2? What are the entries of P^3? P^n?

Statistical Model

Let X_t denote a process taking values in a state space S = {1, . . . , *s*}. The data is of the form $X_0 = s_0,\ X_{t_1} = s_1, \ldots, X_{t_n} = s_n$, where $0 < t_1 < \ldots < t_n$

In the taxi example, *S* = {N, D, S}

For a sample $X_{t_1}, \ldots, X_{t_n}$ of dependent data we can always write the likelihood (joint pdf) using the multiplication rule for dependent r.v:

$$P\left(X_0 = s_0, \ldots, X_{t_n} = s_n\right)$$
$$= P(X_0 = s_0) \prod_{i=1}^{n} P\left(X_{t_i} = s_{t_i} \middle| X_0 = s_0, X_{t_1} = s_{t_1}, \ldots, X_{t_{i-1}} = s_{i-1}\right)$$

Definition: The (first-order) *Markov property* assumes that "given the present, the future is independent of the past", meaning

$$P\left(X_0 = s_0, \ldots, X_{t_n} = s_n\right) = P(X_0 = s_0) \prod_{i=1}^{n} P\left(X_{t_i} = s_{t_i} \middle| X_{t_{i-1}} = s_{i-1}\right)$$

Stationary Markov chain

Definition: A *Markov chain* is a Markov process in discrete time, so we can simply write $t_i = i$, $i = 1, \ldots, n$.

We will assume the process is *stationary*, i.e.

$$P(X_t = s | X_u = r) = P(X_{t-u} = s | X_0 = r)$$

That is, the conditional probabilities only depend on the time difference.

For a stationary chain X_t observed at a discrete equally spaced times $t = 0, 1, \ldots, n$ we define the *transition probabilities*

$$p_{rs} = P(X_1 = s | X_0 = r), \qquad r, s, \in S$$

Thus, the transition probabilities define the $s \times s$ transition matrix $\boldsymbol{P}$ such that

$$\sum_{j=1}^{s} p_{rj} = \sum_{j=1}^{s} P(X_1 = j | X_0 = r) = 1$$

Transition matrix properties

Let $\boldsymbol{p} = (p_1, \ldots, p_s)$ be the vector of initial probabilities, that is, $p_r = \mathrm{P}(X_0 = r)$

Then the k^{th} element of $\boldsymbol{p}'\boldsymbol{P}$ is $P(X_1 = k) = \sum_{j=1}^{s} p_j p_{jk}$

Similarly, the pmf of X_n is given by $\boldsymbol{p}'\boldsymbol{P}^n$

Definition: A Markov chain with transition matrix P is said to have a *stationary distribution* $\boldsymbol{\pi}$ (aka equilibrium) if $\boldsymbol{\pi}'\boldsymbol{P} = \boldsymbol{\pi}'$

Theorem: Under some conditions (the Markov chain is ergodic, that is, irreducible and aperiodic), the stationary distribution satisfies:

$$\lim_{n\to\infty} \boldsymbol{P}^n = \mathbf{1}'\boldsymbol{\pi}$$

The small print

Definition: A state s_j is said to be *accessible* from state s_i if a chain starting in state s_j has a positive probability to reach state s_j at some future time point *n*. That is, $\exists n > 0 : p_{ij}^{n} > 0$

If s_j is accessible from s_i and s_i is accessible from state s_j then we say that s_i and s_j *communicate*. A *communicating class* is defined to be a set of states that communicate.

Definition: If a discrete time Markov chain is composed by only *one* communicating class (i.e., if all states in the chain communicate), then it is said to be *irreducible*.

Finding the stationary distribution

- Using a computer, you can apply “brute force”, meaning compute some very high power of $\boldsymbol{P}$ and then each row should be approximately equal to $\boldsymbol{\pi}$
- Eigen decomposition: Notice that $\boldsymbol{\pi}'\boldsymbol{P} = \boldsymbol{\pi}'$ implies that $\boldsymbol{\pi}$ is a left eigenvector of $\boldsymbol{P}$, corresponding to eigenvalue 1. Therefore, if you can find the eigenvalues of $\boldsymbol{P}$ it means you can find $\boldsymbol{\pi}$
- You can solve $\boldsymbol{\pi}'\boldsymbol{P} = \boldsymbol{\pi}'$ as a system of linear equations.

Rate of convergence

If $\lambda_1 > \cdots > \lambda_s$ are the eigenvalues of $\boldsymbol{P}$, then $\lambda_1 = 1$ and the size of the next eigenvalue λ_2 indicates the rate of the speed of convergence as we approach equilibrium. The reason is because it describes how quickly the largest of the vanishing terms will approach zero in:

$$\boldsymbol{P}^n = \sum_{i=1}^{s} \lambda_i^n \boldsymbol{r}_i \boldsymbol{l}_i'$$

where r_i are the right eigenvectors, and l_i are the left eigenvectors.

Connection with MCMC

- Markov chain theory is most widely used in the Markov chain Monte Carlo (MCMC) methods in statistics.
- However, in MCMC the state space is usually *uncountable*, for example, $S = (-\infty, \infty)$, that is, X_n is a continuous random variable.
- The transition matrix becomes a transition *kernel*

(A matrix can be regarded as an operator from $\mathbb{R}^s$ to $\mathbb{R}^s$ whereas the kernel transforms one density function to another density function)

- The equilibrium distribution is not a vector but a pdf.
- The conditions for ergodicity are more complicated
- The eigenvectors are called eigenfunctions.
- There are infinitely many eigenvalues!

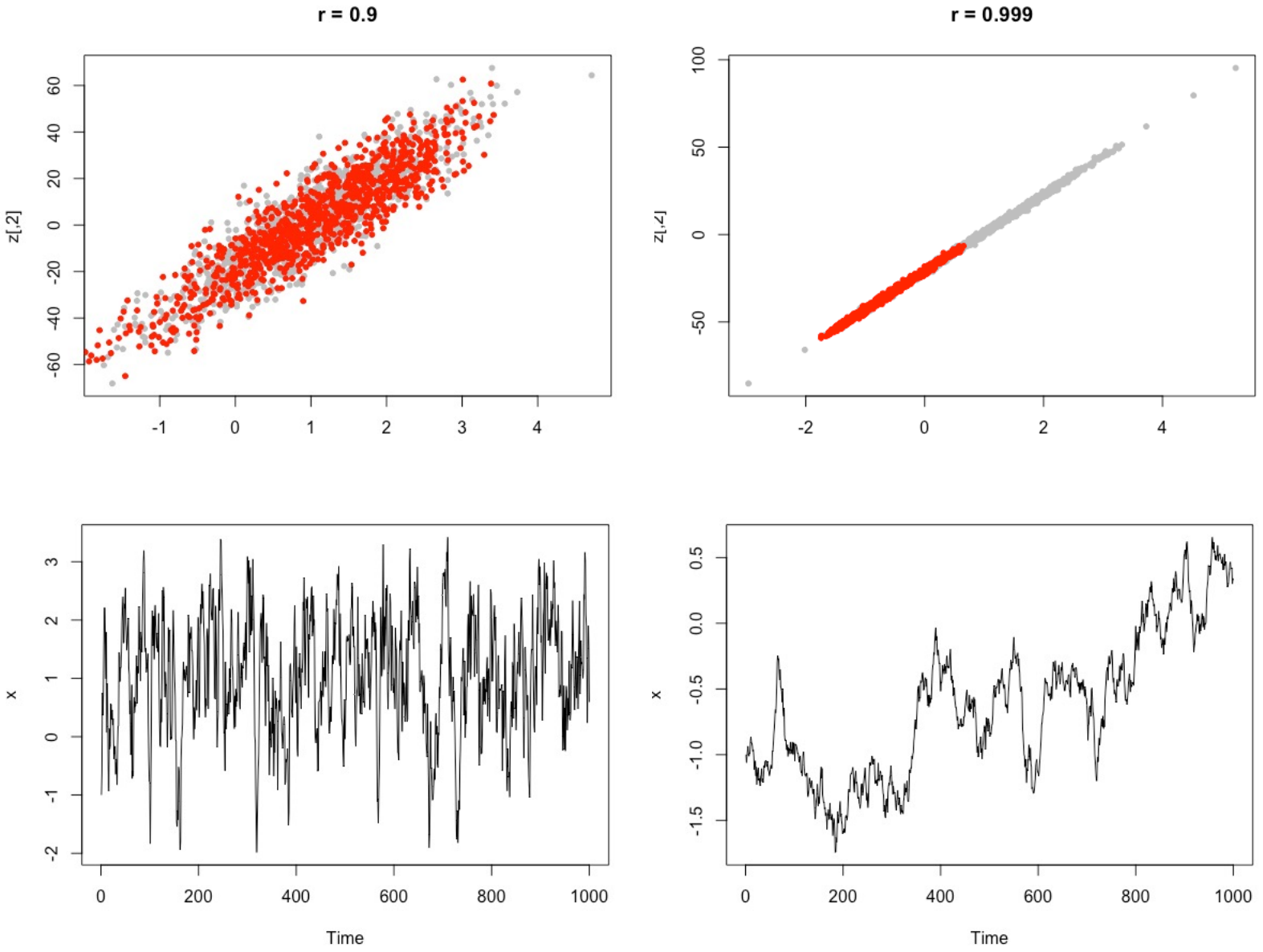
MCM Diagnostics

- We know that

$$\bar{\phi} = \frac{1}{S}\sum_{i=1}^{S} g\left(\boldsymbol{\phi}^{(i)}\right) \approx \int g(\boldsymbol{\phi}) f(\boldsymbol{\phi}) d\boldsymbol{\phi} = \phi_0$$

- But how close is it? In other words, how do we know that the chain has reached *stationarity*?
- We need to check the autocorrelation. It measures how quickly we move around the parameter space, aka the speed of *mixing*.
- An independent MC sampler has perfect mixing: It has zero autocorrelation and can jump between different regions of the parameter space in one step. This is not the case with MCMC.

Recall: Gibbs sampler example



MCMC Variance

Recall that for independent Monte Carlo samples

$$Var_{MC}(\bar{\phi}) = E[(\bar{\phi} - \phi_0)^2] = \frac{Var(\phi)}{S}$$

We can use this result and CLT to conclude that roughly 95% of the approximations are within $\pm 2\sqrt{Var_{MC}(\bar{\phi})}$ of the target value.

However, in the MCMC the samples are positively correlated and

$$Var_{MCMC}(\bar{\phi}) = E\left\{\left[\frac{1}{S}\sum_{i=1}^{S}(\phi^s - \bar{\phi})^2\right]\right\}$$

$$= \frac{1}{S^2}\sum_{i=1}^{S} E[(\phi^s - \bar{\phi})^2] + \frac{1}{S^2}\sum_{i\neq j} E\left[\left(\phi^{(i)} - \phi_0\right)\left(\phi^{(j)} - \phi_0\right)\right]$$

$$= Var_{MC}(\bar{\phi}) + \frac{1}{S^2}\sum_{i\neq j} E\left[\left(\phi^{(i)} - \phi_0\right)\left(\phi^{(j)} - \phi_0\right)\right]$$

So the MCMC variance is equal to the MC variance plus a term that depends on the correlation of samples within the Markov chain.

Effective Sample Size (ESS)

The higher the autocorrelation, the more MCMC samples we need to attain a given level of precision for our approximation.

$$Var_{MCMC}(\bar{\phi}) = \frac{Var_{MC}(\bar{\phi})}{ESS}$$

where

$$ESS = \frac{S}{1 + 2\sum_{k=1}^{\infty} acf(k)}$$

It can be interpreted as the number of independent Monte Carlo samples necessary to give the same precision as the MCMC samples.

Rule of thumb: if $\frac{ESS}{S} < 0.1$ be alarmed!

Ref: "Practical Markov chain Monte Carlo" (Geyer, 1992)

Other Diagnostics

- Look at the traceplot aka the hairy caterpillar ocular inspection test.
- If needed allow a burn in period or warmup of the chain.
- Check the R-hat values (requires special packages).
- Run multiple chains!
- Crudely, ratio of variance between chains to variance within chains
- Should be around 1.
- Rule of thumb be suspicious even if it is more than 1.1