MCMC Basics and Gibbs Sampling

Econ 690

Purdue University

February 1, 2010

Outline

A Markov Chain is a sequence of random variables $X_1, X_2, \cdots$ where the probability distribution associated with X_{t+1} depends only on the realization of the last variable in the sequence:

-

If this probability distribution does not depend on t, then the sequence is called a homogenous (Markov) chain.

We limit ourselves to such chains in the context of this discussion.

A description of how the chain moves from state to state (given the previous value of the state of the chain) is summarized through a transition kernel $K(x, y)$ where:

-

In the case of a discrete state space, the transition kernel is simply a probability matrix:

		X_{t+1}			
		1	2	$\cdots$	n
X_t	1	$K(1,1)$	$K(1,2)$	$\cdots$	$K(1,n)$
	2	$K(2,1)$	$K(2,2)$	$\cdots$	$K(2,n)$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	n	$K(n,1)$	$K(n,2)$	$\cdots$	$K(n,n)$

where the rows of the matrix sum to unity.

The probability distribution of X_{t+1} is obtained from the transition kernel and the probability distribution of the current state X_t. Define this (marginal) probability distribution as

-

Then,

-

Let the $1 \times n$ probability distribution vector p_t be defined as

$$p_t = [p_t(1)\ p_t(2)\ \dots\ p_t(n)]$$

and likewise for p_{t+1}. In addition, let K be the $n \times n$ transition matrix:

$$K = \begin{bmatrix} K(1,1) & K(1,2) & \cdots & K(1,n) \\ K(2,1) & K(2,2) & \cdots & K(2,n) \\ \vdots & \vdots & \ddots & \vdots \\ K(n,1) & K(n,2) & \cdots & K(n,n) \end{bmatrix}$$

Then, in matrix form,

-

or, for a continuous state space,

-

- It is typically of interest to determine if there exists a (unique) stationary distribution of the chain.
- By a *stationary distribution*, we mean that if p is the current state distribution, then (under the given transition probability structure), p will also follow as next period's state distribution.
- Formally, for the discrete case, we seek a (unique) solution to the equation

- and similarly, for the continuous case, we seek a (unique) solution to the equation

Consider, for example, solving for q in the following:

-

Multiplying this out implies:

-

or

-

- This last example has a unique stationary distribution. (This is guaranteed when all elements of the transition matrix are positive for a discrete state space). In general, however, there may be more than one stationary distribution.
- If, for example

$$K = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

  than any p will be a stationary distribution of the chain.
- Uniqueness of the stationary distribution depends on the notion of irreducibility of the chain which, in turn, relates to determining if states are "accessible" from other states (i.e., j can be reached from i from some number of steps). Note that K above will not produce an irreducible chain!

- For purposes of posterior simulation, we will want to construct our transition kernel K so that the posterior (or target distribution) is a (unique) stationary distribution of the chain.
- That is, once we arrive at a situation where we are drawing from the posterior, then all subsequent draws produced from the chain will be draws from the posterior as well.
- An as yet unresolved issue is one of convergence - starting from any initial distribution p_0, after successively applying our transition kernel K, we need to eventually converge to the stationary distribution p.
- Convergence is related to the aperiodicity of the chain, which essentially rules out cyclical-type behaviour of the simulations.

Consider, for the sake of illustration, the transition kernel:

$$K = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

It is clear that ...

Also note that

$$K^2 = I_2, \quad K^3 = I_2 K = K, \quad K^4 = K^2 = I_2, \cdots$$

so that all even powers of K are I_2 and all odd powers are K.

Now, suppose you start the chain by obtaining a draw from $p_0 = [a \ b]$.

Then,

-

Thus, the state distribution will alternate between $[a \ b]$ and $[b \ a]$, and will clearly not converge in general. (Of course when we start from the unique stationary distribution of [.5 .5], then we will remain there forever.)

Under the conditions of irreducibility and aperiodicity, one can show (in the discrete state space case) that the chain is ergodic, so that sample averages of functions obtained from the sequence will converge to the expectation of those functions under the stationary distribution p.

In Gibbs sampling, we construct the transition kernel so that the posterior distribution is a stationary distribution of the chain.

In practice, however, it is not guaranteed that such a chain will statisfy conditions like irreducibility and aperoidicity. There have been conditions provided [see Geweke (2005)], but these are seldom checked in empirical practice.

For better or worse, researchers often use a variety of convergence checks and generated data experiments to bolster the case that such an algorithm "works." For the "simple" models discussed in the remainder of this course, these concerns are not substantial.

- The Gibbs Sampling algorithm constructs a transition kernel K by *sampling from the conditionals of the target (posterior) distribution.*
- To provide a specific example, consider a bivariate distribution $p(y_1, y_2)$.
- Further, apply the transition kernel

- That is, if you are currently at (x_1, x_2), then the probability that you will be at (y_1, y_2) can be surmised from the conditional distributions $p(y_1|y_2 = x_2)$ and $p(y_2|y_1)$ (where y_1 refers to the value realized from the first step).

It is reasonably straightforward to show that the target distribution $p(y_1, y_2)$ is a stationary distribution under this transition kernel:

To this end, note

- Thus, the target distribution (which in our case is the posterior distribution) is a stationary distribution of the chain.
- What this means is:
  1. If we were fortunate enough to obtain our first draw from $p(\theta|y)$,
  2. and then sampled consecutively from the conditional posterior distributions of the model
  3. then all subsequent draws would be (correlated) draws from $p(\theta|y)$.
- These draws could then be used to calculate posterior means, or other desired features.
- Thus, we can think about this procedure as a "limiting" version of direct sampling, where draws obtained from the Gibbs sampler will (eventually) be draws from the posterior distribution.

The General Gibbs Algorithm

Let θ be a $K \times 1$ parameter vector with associated posterior distribution $f(\theta|y)$ and write

$$\theta = [\theta^1 \ \theta^2 \ \cdots \ \theta^K].$$

(We use superscripts to denote elements of the parameter vector and will employ subscripts to denote iterations in the algorithm.)

The Gibbs sampling algorithm proceeds as follows:

(i) Select an initial parameter vector $\theta_0 = [\theta_0^1 \ \theta_0^2 \ \cdots \ \theta_0^K]$. This initial condition could be arbitrarily chosen, sampled from the prior, or perhaps could be obtained from a crude estimation method such as least-squares.

(1)

(2)

$\vdots$

(K)

(ii) Repeatedly cycle through (1) $\rightarrow$ (K) to obtain $\theta_2 = [\theta_2^1 \ \theta_2^2 \ \cdots \ \theta_2^K]$, θ_3, etc., *always conditioning on the most recent values of the parameters drawn* [e.g., to obtain θ_2^1, draw from $f(\theta^1 | \theta^2 = \theta_1^2, \theta^3 = \theta_1^3, \cdots \theta^K = \theta_1^K, y)$, etc.].

- At convergence, the sequence of draws produced from this algorithm will act as draws obtained from $f(\theta|y)$.
- To implement the Gibbs sampler we require the ability to draw from the posterior conditionals of the model. (Note that some of the previous methods on direct simulation could come in useful in this regard!)
- Although the joint posterior density $f(\theta|y)$ may often be intractable, the complete conditionals $\{f(\theta^j|\theta^{-j}, y)\}_{j=1}^K$, (with θ^{-j} denoting all parameters other than θ^j) prove to be of standard forms in many cases, including:
-
-
-
-
- We will discuss posterior simulation in all of these in the remainder of the course

- We now turn to, perhaps, the simplest example of the Gibbs sampler, and illustrate how the algorithm is implemented within the context of this model.
- We suppose that some problem of interest generates a posterior distribution of the form:

$$p(\theta_1, \theta_2|y) \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right),$$

  where ρ is *known*.
- We will illustrate how the Gibbs sampler can be employed to fit this model, and also discuss how its performance is affected by ρ.

- To begin, we must set a starting value for *either* θ_1 or θ_2.
- It doesn't matter which we choose - the algorithm will work either way. So, let's say that we set $\theta_2 = c$ to start.
- To implement the Gibbs sampler, we must derive the conditional posterior distributions $p(\theta_1|\theta_2, y)$ and $p(\theta_2|\theta_1, y)$. These are readily available using properties of the multivariate normal distribution:

and

So, the first iteration of the Gibbs sampler will proceed as follows:

1.
2.
3.
4.

The core of a general MATLAB program for fitting this model might look something like this (without worrying about storing any of our simulations yet):

```
theta2draw = c;
rho = .5;
for i=1:100;
theta1draw = rho*theta2draw + sqrt(1-rho^2)*randn(1,1);
theta2draw = rho*theta1draw + sqrt(1-rho^2)*randn(1,1);
end;
```

- It is a good idea to get rid of some of the initial simulations and use the "latter" set of simulations to calculate quantities of interest.
- Remember that this is an iterative algorithm - we must first converge to the target (posterior) distribution, and once we have arrived at this target distrbution, the subsequent draws will be draws from $p(\theta|y)$.
- This "pre-convergence" period is called the burn-in, and the burn-in draws should be discarded.
- A sketch of a MATLAB program that does all of these things is provided on the following page:

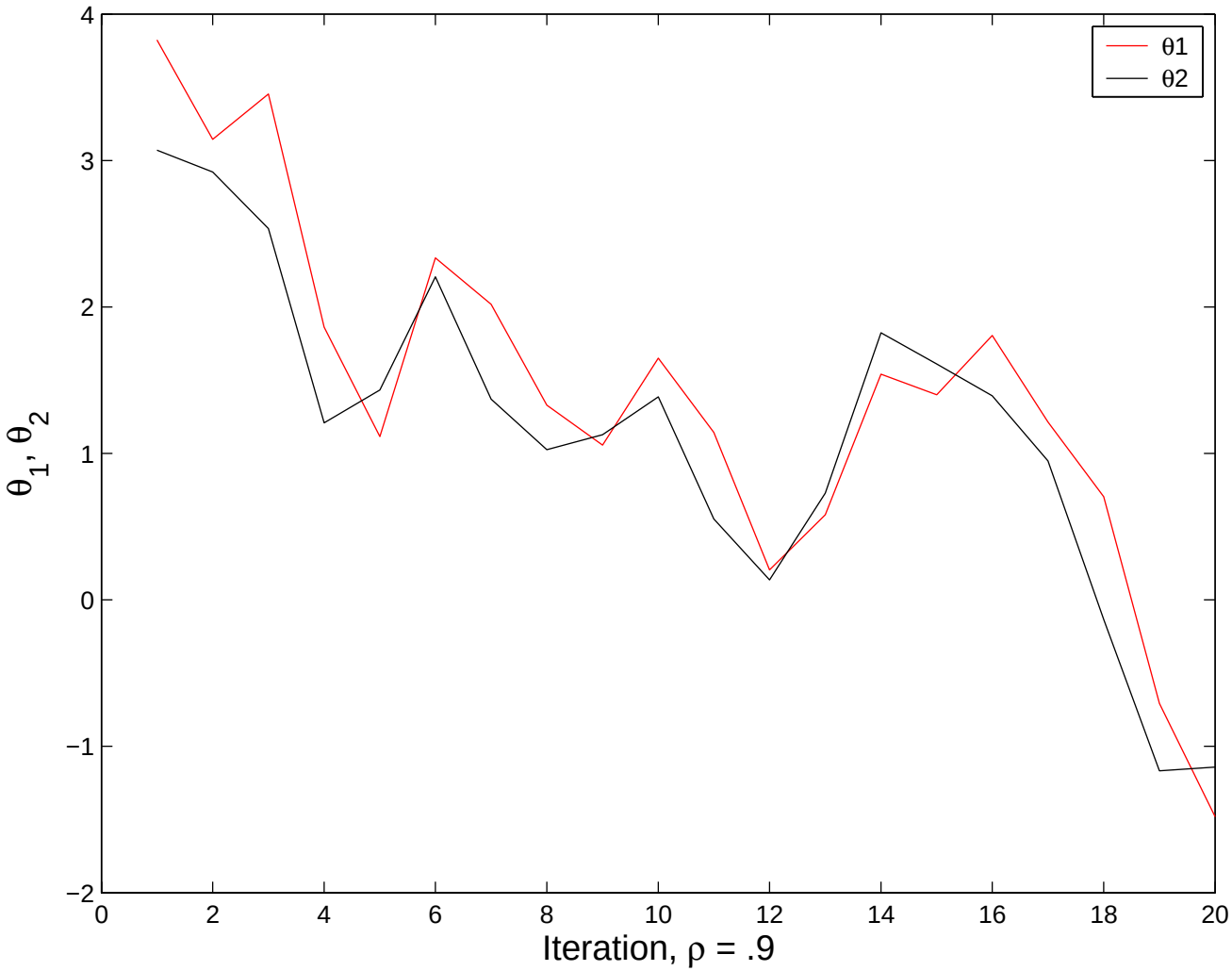
```
rho = c;
iter = 1000;
burn = 100;
theta1keep = zeros(iter-burn,1);
theta2keep = theta1keep;
for i=1:iter;
theta1draw = rho*theta2draw + sqrt(1-rho^2)*randn(1,1);
theta2draw = rho*theta1draw + sqrt(1-rho^2)*randn(1,1);
if i > burn;
theta1keep(i-burn) = theta1draw;
theta2keep(i-burn) = theta2draw;
end;
end;
```

- We illustrate the performance and application of the Gibbs sampler in the following set of experiments.
- We first set $\rho = .9$, and generate Gibbs samples of sizes $M = 1,000$ and $M = 10,000$.
- Posterior means and posterior variances for θ_1 and θ_2 (as well as their correlation) are provided on the following page, where, in each case, the first 100 iterations have been discarded as the burn-in:

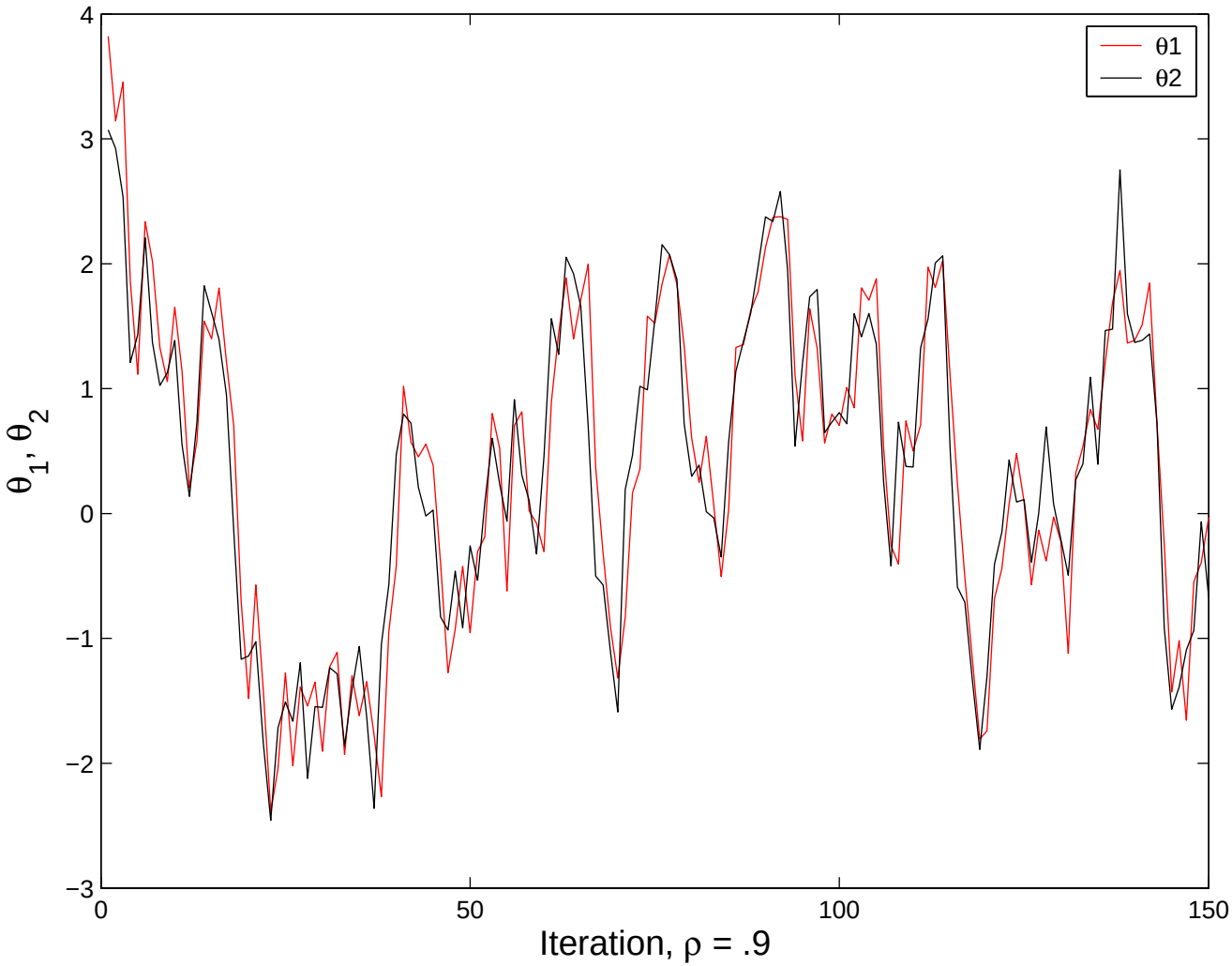
	Estimated Posterior Quantities				
	$E(\theta_1 \mid y)$	$E(\theta_2 \mid y)$	$\text{Var}(\theta_1 \mid y)$	$\text{Var}(\theta_2 \mid y)$	$E(\rho \mid y)$
$M = 1,000$	-.052	-.037	1.18	1.20	.915
$M = 10,000$	-.009	.008	.982	.991	.899

- So, clearly the accuracy of our estimates increases with the Gibbs sample size M.
- The following figures provide plots of the paths of our posterior simulations with $\theta_2 = 4$ initially.

First 20 Iterations of Gibbs sampler

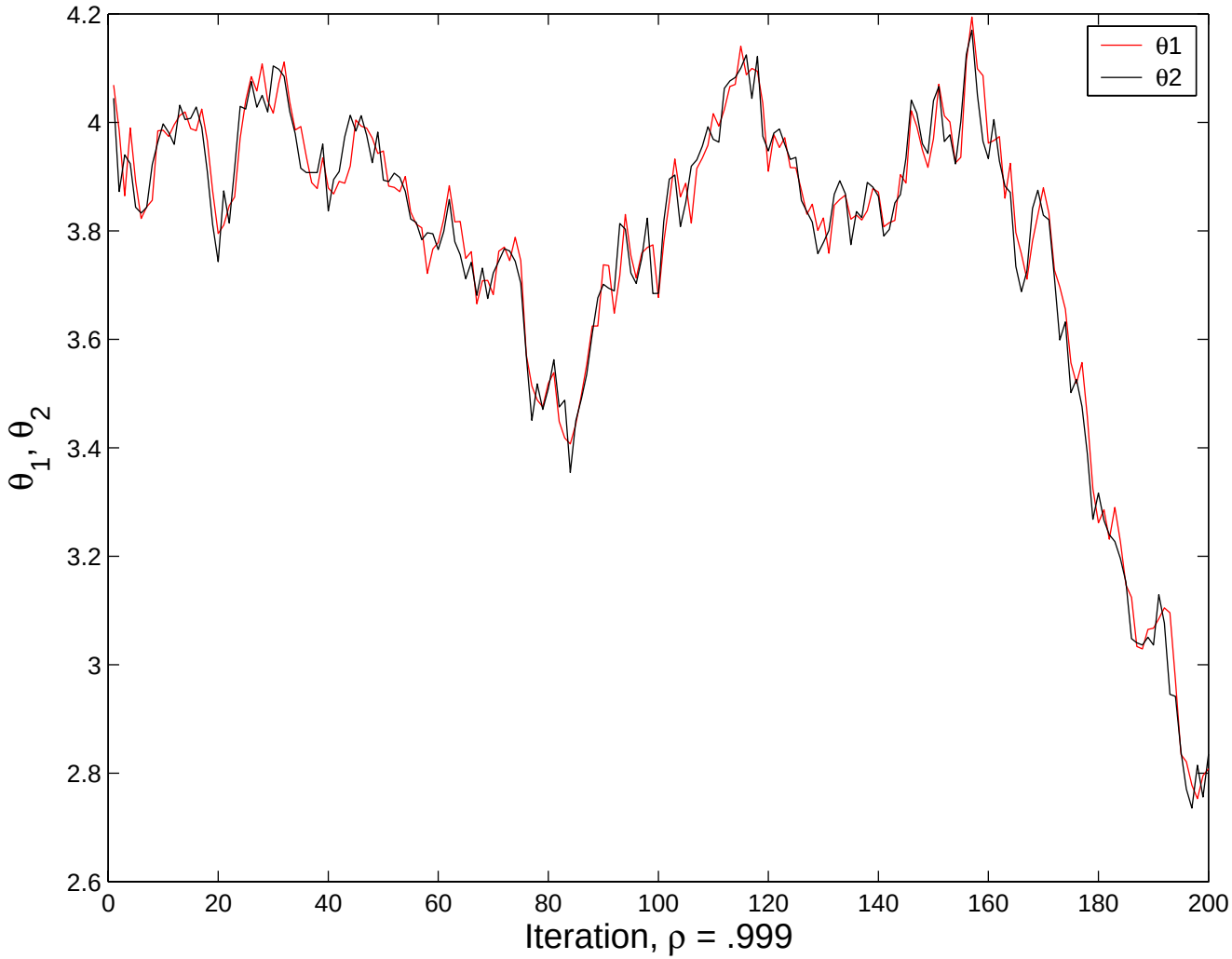


First 150 Iterations of Gibbs Sampler

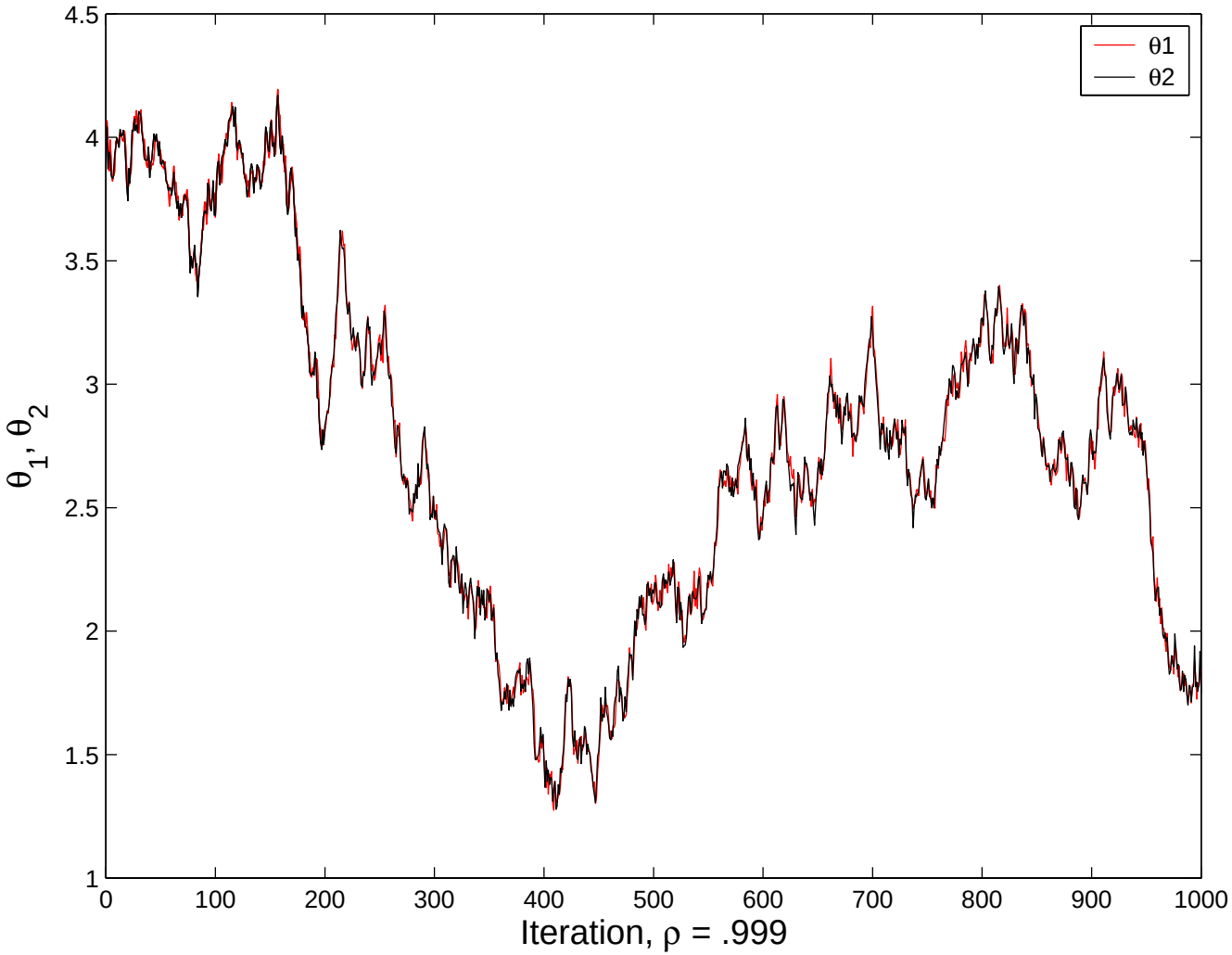


- This example is also suggestive of the fact that the performance of our method will, in general, depend on the degree of correlation between the elements of the posterior.
- Intuitively, if θ_1 and θ_2 are highly correlated, then consecutive iterations will tend to produce little movement.
- This is termed slow mixing of the parameter chain, and suggests that a large amount of draws may be necessary to produce a reasonable level of numerical accuracy.
- Note that, if $\rho = 0$, then sampling from the conditionals is like sampling from the marginals, whence Gibbs reduces to direct Monte Carlo integration.
- The following examples are constructed to illustrate the potential of a slow mixing problem when $\rho = .999$, and the chain is again initialized at $\theta_2 = 4$.

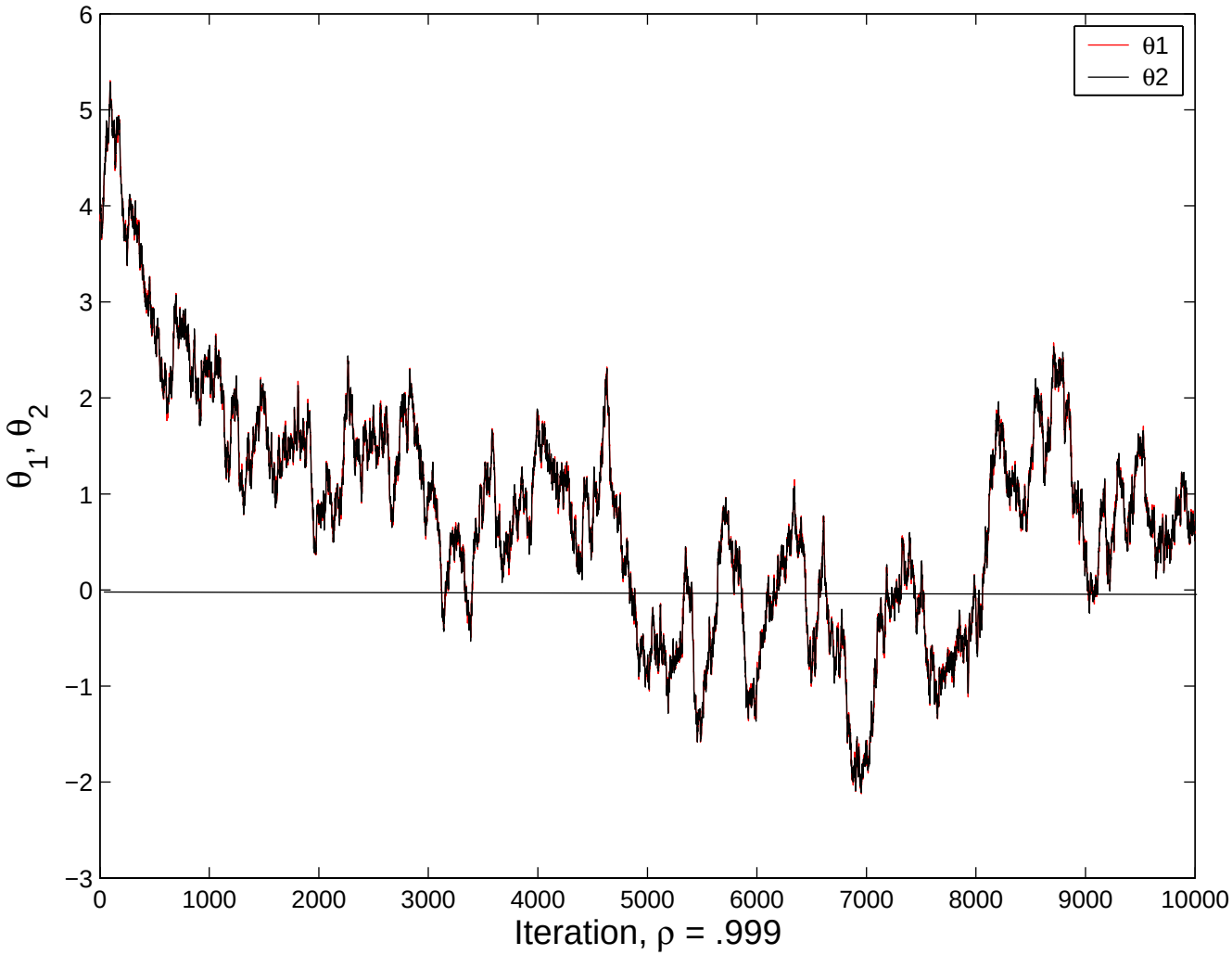
First 200 Iterations of Gibbs Sampler



First 1,000 Iterations of Gibbs Sampler

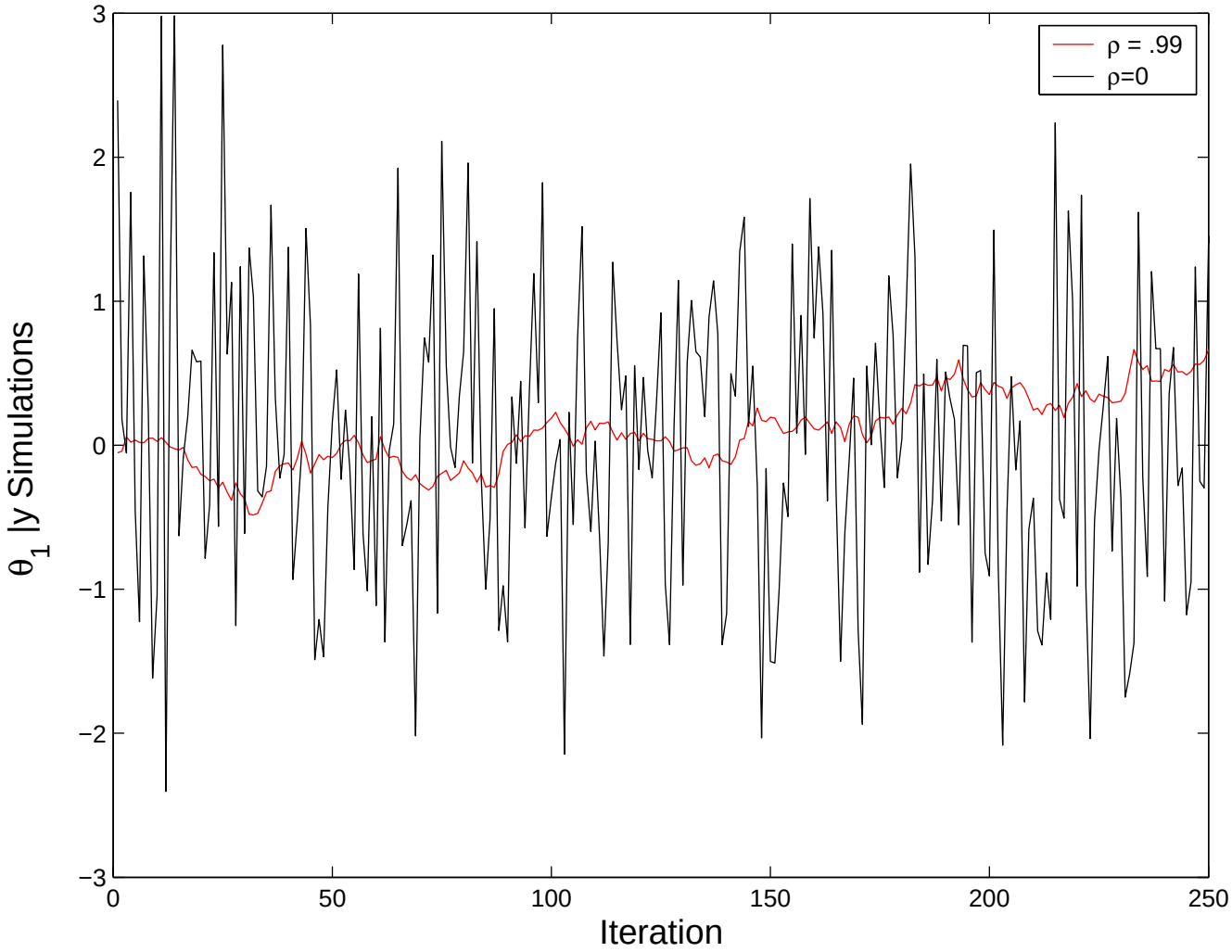


First 10,000 Iterations of Gibbs Sampler



- Thus, under very high correlation between θ_1 and θ_2, we see that it may take many iterations (approximately 5,000) for the chain to "converge" toward the mean value of 0, given a starting value equal to 4.
- Thus suggests that, for a given problem, one must be careful to ensure that an adequate choice for the burn-in is made, and that the post-convergence sample is adequately large to ensure reasonably accurate estimates of posterior moments of interest.
- We will return to this issue when we discuss the calculation of numerical standard errors and other diagnostics.
- The final graph on the next page also illustrates the slow mixing problem, as we plot the path of two θ_1 chains under two Gibbs algorithms with $\rho = 0$ and $\rho = .999$.

250 Post-Convergence Simulations



- When $\rho = .999$ the last graph shows the slow movements and long cyclical patterns in our simulated variates.
- Using the last 5,000 of 10,000 simulated draws with $\rho = .999$, we obtain

Estimated Posterior Quantities

$E(\theta_1 \mid y)$	$E(\theta_2 \mid y)$	$\text{Var}(\theta_1 \mid y)$	$\text{Var}(\theta_2 \mid y)$	$E(\rho \mid y)$
-.637	-.638	.714	.714	.999

which shows that when parameters of the posterior distribution are highly correlated, estimated quantities of interest can be inaccurate, and a large number of posterior simulations is called for.

Further Reading


Casella, E. and E.I. George
Explaining the Gibbs Sampler.
The American Statistician, 1992.


Geweke, J.
Using Simulation Methods for Bayesian Econometric Models: Inference, Development and Communication
Econometric Reviews 1999.


Geweke, J.
Comtemporary Bayesian Econometrics and Statistics,
Wiley, 2005, sections 4.3, 4.5-4.7


Tierney, L.
Markov Chains for Exploring Posterior Distributions
Annals of Statistics, 1994.