

## Module 3

Bayes computation via Markov Chain Monte Carlo (or "What if I can't do multivariate calculus?" )

Prof. Stephen B. Vardeman  
Statistics and IMSE  
Iowa State University

March 5, 2008

# The Apparent Barrier to Practical Bayes Analyses

Our ability to cook up interesting statistical models and collect data quickly outruns our ability to do the pencil and paper mathematics apparently needed to implement the Bayes paradigm. As a simple example of what would be required, consider a problem where one has  $n$  normal observations and neither the process mean nor the process standard deviation are known. The joint probability density for independent  $N(\mu, \sigma^2)$  random variables  $X_1, X_2, \dots, X_n$  (the likelihood) is

$$f(\mathbf{x}|\mu, \sigma^2) = \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right)$$

and to carry through the Bayes paradigm, one needs to specify a *joint* prior distribution for the pair  $(\mu, \sigma^2)$ , say  $g(\mu, \sigma^2)$ , and to do probability calculations for  $(\mu, \sigma^2)$  based on a joint posterior density

$$g(\mu, \sigma^2|\mathbf{x}) \propto f(\mathbf{x}|\mu, \sigma^2) g(\mu, \sigma^2)$$

Even with fairly "tame"/convenient standard choices for  $g(\mu, \sigma^2)$ , one still can not avoid 2-variable calculus.

# The Apparent Barrier to Practical Bayes Analyses (cont.)

This is already beyond the mathematical background of essentially all users of statistical methods. And the situation clearly gets truly out of hand as the number of parameters in a model grows. (A fairly small multiple regression model with 5 predictors already appears to require 7-dimensional calculus!)

Until the last 15 or 20 years, this technical complexity largely reduced the Bayes paradigm to the status of only a theoretical curiosity, applicable only to "toy" problems. But that has changed with the proliferation of computing power and huge advances in statistical theory that have shown how *simulation methods* can be used to approximate complicated high-dimensional distributions.

# Simulation from Posterior Distributions

What would be the most "obvious" methods of simulation from a posterior distribution of a high-dimensional parameter vector  $\theta = (\theta_1, \theta_2, \dots, \theta_k)$  don't seem to be possible. That is, no one knows general methods of beginning with a function of  $k$  arguments

$$g(\theta_1, \theta_2, \dots, \theta_k | \mathbf{x})$$

and generating *independent* draws from the joint distribution it specifies. But there are now very clever algorithms for generating *dependent* sequences of vectors

$$\theta_1, \theta_2, \theta_3, \dots, \theta_n$$

that (despite their structure, nevertheless) for large  $n$  have relative frequency distributions approximating a target distribution specified by  $g(\theta_1, \theta_2, \dots, \theta_k | \mathbf{x})$ . These methods are called "Markov Chain Monte Carlo" methods, as they employ the theory of "Markov Chains" (a type of probability structure for sequences that involves mathematically tractable dependencies).

# "Gibbs Sampling"

There are by now many variants of MCMC used to do Bayes computations. Here we will talk about the most popular basic method, so called "Gibbs Sampling" or "Successive Substitution Sampling," because 1) it is the easiest method to present and 2) it is the fundamental engine of the WinBUGS software we will use in this workshop.

The fundamental idea of Gibbs sampling is this. For  $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ , from some starting vector  $\theta_0$ , one creates  $\theta_{i+1}$  from  $\theta_i$  by "updating" in succession the  $k$  coordinates of  $\theta_i$  by drawing at random from the conditional distribution of each variable given current values of all the others.

In order to illustrate the basic idea, consider the small artificial example of Gibbs sampling from the jointly discrete distribution of  $\theta = (\theta_1, \theta_2)$  given in Table 1 on the next panel.

# "Gibbs Sampling" (cont.)

**Table 1** A Simple Joint Probability Mass Function  $g$  for  $\theta = (\theta_1, \theta_2)$

	$\theta_2 = 1$	$\theta_2 = 2$
$\theta_1 = 1$	.2	.1
$\theta_1 = 2$	.2	.2
$\theta_1 = 3$	.1	.2

This joint distribution has conditional distributions

$\theta_1$	$g(\theta_1 1)$
1	2/5
2	2/5
3	1/5

$\theta_1$	$g(\theta_1 2)$
1	1/5
2	2/5
3	2/5

and

$\theta_2$	1	2
$g(\theta_2 1)$	2/3	1/3

$\theta_2$	1	2
$g(\theta_2 2)$	2/4	2/4

$\theta_2$	1	2
$g(\theta_2 3)$	1/3	2/3

## "Gibbs Sampling" (cont.)

For sake of example, suppose that one begins Gibbs sampling with

$$\theta_0 = (2, 1)$$

One then generates a replacement for  $\theta_1 = 2$  from  $g(\theta_1|1)$ . For sake of argument, suppose that  $\theta_1 = 3$  is generated. Then one must generate a replacement for  $\theta_2 = 1$  using  $g(\theta_2|3)$ . If, for example,  $\theta_2 = 2$  is generated, after one complete Gibbs cycle

$$\theta_1 = (3, 2)$$

One then generates a replacement for  $\theta_1 = 3$  from  $g(\theta_1|2)$  and so on.

Upon generating a long string of  $\theta$ 's in this way, the theory of Markov Chains implies that the relative frequency distribution produced will approximate the distribution in Table 1. Approximately 20% of  $\theta_1, \theta_2, \dots, \theta_n$  will be  $(1, 1)$ , about 10% will be  $(1, 2)$ , etc. So properties of the joint distribution  $g$  can be approximated by corresponding sample properties of  $\theta_1, \theta_2, \dots, \theta_n$ .

# Practical Land Mines and Considerations

The details of exactly how for a more realistic/complicated  $g(\theta_1, \theta_2, \dots, \theta_k | \mathbf{x})$  the sampling from each conditional distribution is done will not be our concern here. What we must discuss are some general cautions about the existence of circumstances under which Gibbs sampling will (completely, or at least for practical purposes) fail to reproduce the target distribution, and practical considerations of using the method.

The next table specifies a hypothetical but instructive discrete joint distribution for  $\theta = (\theta_1, \theta_2)$  that illustrates that Gibbs sampling can fail completely in even a very simple situation.

# Practical Land Mines and Considerations (cont.)

**Table 2** A Simple Joint Probability Mass Function  $g$  for  $\theta = (\theta_1, \theta_2)$  for Which Gibbs Sampling Will Fail

	$\theta_2 = 1$	$\theta_2 = 2$	$\theta_2 = 3$	$\theta_2 = 4$
$\theta_1 = 1$	0	0	.1	.2
$\theta_1 = 2$	0	0	.1	.1
$\theta_1 = 3$	.1	.1	0	0
$\theta_1 = 4$	.2	.1	0	0

The little distribution in Table 2 has two "islands of probability," namely

$$\{(3, 1), (3, 2), (4, 1), (4, 2)\} \text{ and } \{(1, 3), (1, 4), (2, 3), (2, 4)\}$$

If the starting vector  $\theta_0 = (\theta_1, \theta_2)$  is in the first island the chain will never reach the second (since a  $\theta_1$  substitution must leave one in the same column as the starting vector and in a cell with positive probability, and a following  $\theta_2$  substitution moves one in a row between cells with positive probability). And vice versa.

# Practical Land Mines and Considerations (cont.)

A Gibbs sampler will produce one of the distributions

	$\theta_2 = 1$	$\theta_2 = 2$			$\theta_2 = 3$	$\theta_2 = 4$
$\theta_1 = 3$	.2	.2	or	$\theta_1 = 1$	.2	.4
$\theta_1 = 4$	.4	.2		$\theta_1 = 2$	.2	.2

NOT the original joint distribution,  $g$ . What is perhaps worse, is that if one is naive and unaware of the possibility of being fooled by Gibbs sampling, one might produce one of the "island distributions" by Gibbs sampling and think that the whole of  $g$  has been represented. That is, there is nothing to immediately warn the naive user that his or her Gibbs sampler is completely inadequate.

The next table illustrates a case where Gibbs will work in theory but fail in practice.

# Practical Land Mines and Considerations (cont.)

**Table 3** A Simple Joint Probability Mass Function  $g$  for  $\theta = (\theta_1, \theta_2)$  for Which Gibbs Sampling Will Fail in Practice

	$\theta_2 = 1$	$\theta_2 = 2$	$\theta_2 = 3$	$\theta_2 = 4$
$\theta_1 = 1$	0	0	.1	.2
$\theta_1 = 2$	0	0	$.1 - \frac{\epsilon}{2}$	.1
$\theta_1 = 3$	.1	$.1 - \frac{\epsilon}{2}$	$\epsilon$	0
$\theta_1 = 4$	.2	.1	0	0

Suppose that  $\epsilon$  is a very small but positive number. The parameter  $\theta = (3, 2)$  represents a very "narrow" isthmus connecting two islands of probability. As long as  $\epsilon > 0$  in theory a Gibbs sampling algorithm will travel between the islands and eventually represent the distribution in Table 3. But as a practical matter, if  $\epsilon$  is small enough, in a number of iterations that one can afford to complete, one will not get enough transitions across the isthmus to learn the relative (probability) sizes of the continents it represents. Gibbs will fail in practice.

## Practical Land Mines and Considerations (cont.)

Tables 2 and 3 are toy examples. But they illustrate a serious practical problem in using MCMC to do Bayes computations. For some models and data sets, posteriors can be "ill-behaved" and have what are essentially "islands of probability." In such cases, what one gets from a Gibbs sampler will depend upon the starting value  $\theta_0$ , and may fail to really represent the true posterior. It is impossible to simply stare at a given data set and model and tell whether it is one where this might be a problem. What we must then do (in an attempt to protect against being fooled) is

- run samplers/chains from a variety of ("widely dispersed") initial values  $\theta_0$ , and
- do some kind of diagnostic checking to verify that at least after an initial "burn-in" period/number of iterations, all of these chains are producing the same picture of the posterior of interest.

Happily, the WinBUGS software we will use in this workshop has features that facilitate this kind of checking.

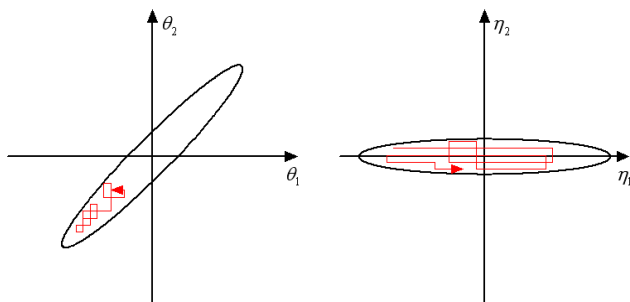
# Considerations Regarding Parameterizations

There is a second possible practical issue with Gibbs sampling related to the fact that every substitution of a single entry of  $\theta = (\theta_1, \theta_2, \dots, \theta_k)$  moves the state of the sampler along a line parallel to a coordinate axis in  $k$ -dimensional space. This means that some choices of parameterizations for a problem will be more efficient (in terms of requiring fewer iterations of a Gibbs sampler to adequately represent a posterior) than others. All other things being equal, one wants parameterizations where the  $k$  individual parameters are "independent" (in either a precise mathematical/probability sense or in a qualitative sense).

# Considerations Regarding Parameterizations (cont.)

Here is an illustration of this point. Imagine that a posterior distribution for  $\theta = (\theta_1, \theta_2)$  is uniform over the ellipse illustrate on the left side of Figure 1.

**Figure 1** Cartoon of the evolution of two Gibbs samplers, one for  $\theta = (\theta_1, \theta_2)$  uniform over the ellipse on the left, and the second for  $\eta = (\eta_1, \eta_2)$  (a transform of  $\theta$ ) uniform over the ellipse on the right



## Considerations Regarding Parameterizations (cont.)

A Gibbs sampler will make substitutions that move the current value of  $\theta$  in steps parallel to the coordinate axes, and by virtue of the orientation of the ellipse ( $\theta_1$  and  $\theta_2$  are far from being "independent") will take many fairly small steps to "cover" the ellipse and thus represent the posterior. On the other hand, if one defines

$$\eta_1 = \sqrt{2}(\theta_1 - \theta_2) \quad \text{and} \quad \eta_2 = \sqrt{2}(\theta_1 + \theta_2)$$

the result is a posterior uniform over the second ellipse (a rigid rotation of the first).  $\eta_1$  and  $\eta_2$  (though not exactly probabilistically independent are uncorrelated and) are approximately independent. A Gibbs sampler *for these transformed parameters*  $\eta_1$  and  $\eta_2$  will be able to take relatively large steps and require relatively fewer steps to cover the ellipse and represent the posterior. By "transforming back" from  $\eta$ 's to  $\theta$ 's one has values that represent the original posterior in far fewer iterations.

(Essentially, one has invented a Gibbs sampler that moves in substitutions parallel to the axes of the ellipse rather than the coordinate axes.)

## Considerations Regarding Parameterizations (cont.)

In extreme cases of this type, there may be a parameterization of a model for which a Gibbs sampler works just fine, while the most obvious parameterization is one in which Gibbs sampling fails in a practical sense.

The take-home message here is then that while MCMC (and in particular) Gibbs sampling is an incredibly helpful and clever idea for Bayesian computation, it is not something to be used mindlessly or without realizing that it has its "land-mines." As the WinBUGS developers say in their user manual:

**Beware: MCMC sampling can be dangerous!**