

## Markov Chain Monte Carlo

### 22S:138, Bayesian Statistics

Lecture 10

Oct 1, 2007

Kate Cowles

374 SH, 335-0727

### The Poisson distribution (one more one-parameter distribution)

- The Poisson distribution may be appropriate when the data are counts of rare events.
- events occurring at random at a constant rate per unit time, distance, volume, or whatever
- assumption that the number of events that occur in any interval is independent of the number of events occurring in a disjoint interval

- examples:
  - the number of cases of a rare form of cancer occurring in Johnson County in each calendar year
  - the number of flaws occurring in each 100-foot length of yarn produced by a spinning machine
  - the number of particles of pollen per cubic foot of air in this room

- Since the values of a random variable following a Poisson distribution are *counts*, what are the possible values?

- probability mass function for a Poisson random variable

$$p(y|\lambda) = \frac{e^{-\lambda} \lambda^y}{y!}, \quad y = 0, 1, \dots$$

- the count of the number of events occurring in  $m$  time units also follows a Poisson distribution, but with parameter  $m\lambda$

- The conjugate prior distribution for the Poisson rate parameter is the gamma family.

## Markov Chain Monte Carlo Methods

- Goals
  - to make inference about model parameters
  - to make predictions
- Requires
  - integration over possibly high-dimensional integrand
  - and we may not know the integrating constant

### Markov chains

- A Markov chain is a sequence of random variables  $X_0, X_1, X_2, \dots$
- At each time  $t \geq 0$  the next state  $X_{t+1}$  is sampled from a distribution
$$P(X_{t+1}|X_t)$$
that depends only on the state at time  $t$ 
  - called “transition kernel”
- Under certain regularity conditions, the iterates from a Markov chain will gradually converge to draws from a unique *stationary* or *invariant* distribution
  - i.e. chain will “forget” its initial state
  - as  $t$  increases, sampled points  $X_t$  will look increasingly like (correlated) samples from the stationary distribution

## Monte Carlo integration and MCMC

- Monte Carlo integration
  - draw independent samples from required distribution
  - use sample averages to approximate expectations
- Markov chain Monte Carlo (MCMC)
  - draws samples by running a Markov chain that is constructed so that its limiting distribution is the joint distribution of interest

- Suppose:
  - MC is run for  $N$  (large number) iterations
  - we throw away output from first  $m$  iterations
  - regularity conditions are met
- then by *ergodic theorem*
  - we can use averages of remaining samples to estimate means

$$E[f(X)] \simeq \frac{1}{N - m} \sum_{t=m+1}^N f(X_t)$$

## Gibbs sampling: one way to construct the transition kernel

- seminal references
  - Geman and Geman (*IEEE Trans. Pattern. Anal. Mach. Intel.*, 1984)
  - Gelfand and Smith (*JASA*, 1990)
  - Hastings (*Biometrika*, 1970)
  - Metropolis, Rosenbluth, et al. (*J. Chem. Phys.*, 1953)
- subject to regularity conditions, joint distribution is uniquely determined by “full conditional distributions”
  - full conditional distribution for a model quantity is distribution of that quantity conditional on assumed known values of all the other quantities in the model

- break complicated, high-dimensional problem into a large number of simpler, low-dimensional problems

### Example: Inference about normal mean and variance, both unknown

- model

$$y_i | \mu, \sigma^2 \sim N(\mu, \sigma^2) \\ i = 1, \dots, N$$

- priors

$$\mu \sim N(\mu_0, \sigma_0^2) \\ \sigma^2 \sim IG(a_1, b_1)$$

- We want posterior means, posterior medians, posterior credible sets for  $\mu, \sigma^2$

### Full Conditional Distributions for Normal Model

- to extract mathematical form of full conditional for a parameter:
  - write out expression to which joint posterior is proportional
  - pull out all terms containing the parameter of interest

## Gibbs Sampler algorithm for Normal

1. choose initial values  $\mu^{(0)}, \sigma^{2(0)}$
2. at each iteration  $t$ , generate new value for each parameter, conditional on most recent value of all other parameters

## What are BUGS and WinBUGS?

- “Bayesian inference Using Gibbs Sampling”
- general purpose program for fitting Bayesian models
- developed by David Spiegelhalter, Andrew Thomas, Nicky Best, and Wally Gilks at the Medical Research Council Biostatistics Unit, Institute of Public Health, in Cambridge, UK
- BUGS
  - for Unix and DOS platforms
  - written in Modula-2; distributed in compiled form only

- WinBUGS

- for Windows
- written in Component Pascal running in Oberon Microsystems’ Blackbox environment
- able to fit a wider variety of models than BUGS can handle
- undergoing continuing development

- excellent documentation, including two volumes of examples

- Web page:

<http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml>

- OpenBUGS

- open source version of WinBUGS
- interfaces easily with R
- Web page:

<http://mathstat.helsinki.fi/openbugs/>

## What do BUGS and WinBUGS do?

- enable user to specify model in simple Splus-like language
- construct the transition kernel for a Markov chain with the joint posterior as its stationary distribution, and simulate a sample path of the resulting chain
  - determine whether or not the full conditional for each unknown quantity (parameter or missing data) in the model is a standard density.
  - generate random variates from standard densities using standard algorithms.
  - BUGS uses the adaptive rejection algorithm (Gilks and Wild, *Applied Statistics*, 1992) to generate from nonstandard full conditionals; consequently can handle only log-concave or discrete full conditionals

## The Art and Science of MCMC Use

- WinBUGS uses Metropolis algorithm to generate from nonstandard full conditionals and is not subject to this limitation
- Deciding how many chains to run
- Choosing initial values
  - Do not confuse initial values with priors!
  - Priors are part of the *model*. Initial values are part of the computing strategy used to fit the model.
  - Priors must *not* be based on the current data.
  - The best choices of initial values are values that are in a high-posterior-density region of the parameter space. If the prior is not very strong, then maximum likelihood estimates (from the current data) are excellent choices of initial values if they can be calculated.
- In the simple models we have encountered so far, the MCMC sampler will converge quickly even with a poor choice of initial values.
- In more complicated models, choosing initial values in low posterior density regions may make the sampler take a huge number of iterations to finally start drawing from a good approximation to the true posterior.
- Assessing whether sampler has “converged”
  - How many initial iterations need to be discarded in order that remaining samples are drawn from a distribution close enough to the true stationary distribution to be usable for estimation and inference?
  - Once we are drawing from the right distribution, how many samples are needed in order to provide the desired precision in estimation and inference?
- Choosing model parameterizations and MCMC algorithms that will lead to convergence in a reasonable amount of time
- Using correlated samples for estimation and inference
  - adjusting estimates of standard errors
- Our next lab will consider output analysis (items from this list)