

22S:166
Computing in Statistics

Review of Bayesian Concepts and Intro to
MCMC

Lecture 14
October 15, 2006

Kate Cowles
374 SH, 335-0727
kcowles@stat.uiowa.edu

Remarks on notation

- I will use $f(\cdot)$, $p(\cdot)$, and $\pi(\cdot)$ to refer to distributions that may be either continuous, discrete, or mixed.
- I will frequently use an integral. When the argument may be a discrete distribution, think summation.
- y will refer to observed, or potentially observable, quantities.
- θ will refer to unobservable quantities.

Bayesian Basics

- Unknown model parameters are random variables.
- Our knowledge / uncertainty about unknown model parameters is appropriately expressed through probability distributions.
- Whenever we observe data, these probability distributions are updated.

Steps in Bayesian data analysis

1. Specify a full probability model
 - joint probability distribution for all observable and unobservable quantities in a problem.
2. Calculate and interpret the *posterior distribution*
 - the conditional probability distribution of the unobserved quantities of interest given the observed data
3. Evaluate model
 - fit to observed data
 - consistency of posterior inference with substantive knowledge
 - sensitivity to model specification

Components of a Bayesian model

- The first stage: the likelihood
 - the probability distribution for the observed data \mathbf{y} conditional on a vector of unknown parameters

$$f(\mathbf{y}|\theta)$$

- The second stage: the priors
 - probability distribution(s) that express our knowledge or uncertainty about model parameters before the data are observed

$$\pi(\theta|\eta)$$

where η is a vector of “hyperparameters”

The marginal likelihood

- the marginal distribution of the data \mathbf{y} given the complete model

$$m(\mathbf{y}) = \int f(\mathbf{y}|\theta^*)\pi(\theta^*|\eta)d\theta^*$$

- also called
 - prior predictive distribution
 - normalizing constant
- Useful in model comparison employing Bayes factors

Bayes theorem and the posterior distribution

- Bayes theorem is the recipe for using the data to update the prior and produce the posterior distribution $p(\theta|\mathbf{y})$:

$$\begin{aligned} p(\theta|\mathbf{y}, \eta) &= \frac{p(\mathbf{y}, \theta|\eta)}{p(\mathbf{y}|\eta)} \\ &= \frac{p(\mathbf{y}, \theta|\eta)}{\int p(\mathbf{y}, \theta^*|\eta)d\theta^*} \\ &= \frac{f(\mathbf{y}|\theta)\pi(\theta|\eta)}{\int f(\mathbf{y}|\theta^*)\pi(\theta^*|\eta)d\theta^*} \end{aligned}$$

- Remarks:
 - Dependence on η in posterior distributions is often suppressed when values of hyperparameters are known constants.
 - In principle, this computation can be done for any valid prior and any well-defined likelihood.
 - The challenge is the integral in the denominator.

Classifications of priors

- informative and noninformative
- proper and improper
 - In general, posterior inference is impossible if the posterior is improper (i.e. if unnormalized posterior does not have a finite integral).
 - * The use of proper priors guarantees a proper posterior.
 - * If improper priors are used, it is important to verify analytically that posterior is proper.
 - * If prior information is minimal and likelihood is complicated, vague but proper priors are advisable.
- conjugate and nonconjugate
 - a class of prior distributions is said to be conjugate for a parameter in a likelihood if the resulting posterior distribution is in the same family as the prior
 - make computation (analytic or by computer) easier, but may not reflect actual prior information
 - for many likelihood functions, there *is* no conjugate prior

- examples: what are conjugate priors for
 - * normal mean (variance assumed known)
 - * normal variance (mean assumed known)

Summarizing Bayesian estimation and inference

- All inference is based on the posterior distribution.
- More integration is required to obtain marginal posterior distributions of parameters of interest (i.e., integrate out nuisance parameters).

$$p(\theta_i|y) = \int p(\theta|y)d\theta_{(-i)}$$

- Point estimates of parameters
 - means and medians of posterior marginals
 - in conjugate one-parameter models
 - * the posterior mean is a weighted average of the prior mean and the mean coming from the likelihood
 - * the posterior variance is smaller than either the prior variance or the variance of the parameter in the likelihood function
 - * example: normal-normal model

- intervals: *credible sets*
 - definition: a $100 \times (1 - \alpha)\%$ credible set for a parameter θ is a subset C of the parameter space Θ such that

$$1 - \alpha \leq P(C|\mathbf{y}) = \int_C p(\theta|\mathbf{y})d\theta$$

- interpretation: The probability that θ lies in C given the observed data is (at least) $(1 - \alpha)$.
- equal-tail credible set: the interval between the $\alpha/2$ and $(1 - \alpha/2)$ quantiles of $p(\theta|\mathbf{y})$
- highest posterior density (HPD) credible set: the subset C of Θ such that

$$C = \{\theta \in \Theta : p(\theta|\mathbf{y}) \geq k(\alpha)\}$$

where $k(\alpha)$ is the largest constant satisfying

$$p(C|\mathbf{y}) \geq 1 - \alpha$$

The posterior predictive distribution

- Used to make inferences about a potentially observable but unobserved model quantity \tilde{y} , conditional on data y that have already been observed

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y}, \theta|y)d\theta \\ &= \int p(\tilde{y}|\theta, y)p(\theta|y)d\theta \\ &= \int p(\tilde{y}|\theta)p(\theta|y)d\theta \end{aligned}$$

- last equation follows if \tilde{y} and y are conditionally independent given θ in the model

Example of a simple Bayesian model

- Normal likelihood, mean and variance both unknown
 - precision is inverse of variance

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

- Semi conjugate priors on μ and σ^2

$$\mu \sim N(\mu_0, \sigma_0^2) \quad (1)$$

$$\tau^2 \sim G(a, b) \quad (2)$$

Hierarchical models

- arise if
 - we don't know values of hyperparameters in 2nd stage
 - or we wish to express relationships among parameters
- likelihood is the first “stage” of the model
- 2nd stage is priors on the parameters that appear in the likelihood
- subsequent stages are priors on hyperparameters from the previous stages

Example of hierarchical model

- A hierarchical model is fit to data on failure rates of the pump at each of 10 power plants. The number of failures for the i -th pump is assumed to follow a Poisson distribution:

$$x_i \sim \text{Poisson}(\theta_i t_i), \quad i = 1, \dots, 10$$

where θ_i is the failure rate for pump i and t_i is the length of operation time of the pump (in 1000s of hours).

- A conjugate gamma prior distribution is adopted for the failure rates:

$$\theta_i \sim \text{Gamma}(\alpha, \beta), \quad i = 1, \dots, 10$$

- The following priors are specified for the hyperparameters α and β :

$$\alpha \sim \text{Exponential}(1.0)$$

$$\beta \sim \text{Gamma}(0.10, 1.0)$$

Markov chain Monte Carlo: one method of Bayesian computation

- If we can't analytically do the integration to get the needed joint and marginal posterior distributions, generate random *samples* from the joint posterior
- MCMC: Do this by constructing a Markov chain with the joint posterior distribution as its stationary distribution
 - choose initial values for all parameters
 - run chain until it reaches its stationary distribution
 - outputs from subsequent iterations are (dependent) draws from the joint posterior
 - also, outputs from these iterations for any single parameter are draws from its posterior marginal
 - the Gibbs sampler: one method of constructing the transition kernel for such a Markov chain

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

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

- 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

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

The Gibbs sampler

- breaks up the problem of generating from a complex, high-dimensional posterior into many easier problems of generating from univariate (or low-dimensional distributions)
- algorithm for generating from $p(\theta_1, \theta_2, \dots, \theta_p|y)$
 1. Select initial values $\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_p^{(0)}$
 2. At each iteration k , generate new values for each parameter, one at a time, from their *full conditional distributions*, conditional on most recent values of each of the other parameters

$$\begin{aligned} & p(\theta_1^{(k)} | \theta_2^{(k-1)}, \dots, \theta_p^{(k-1)}, y) \\ & p(\theta_2^{(k)} | \theta_1^{(k)}, \theta_2^{(k-1)}, \dots, \theta_p^{(k-1)}, y) \\ & \vdots \\ & p(\theta_p^{(k)} | \theta_1^{(k)}, \dots, \theta_{p-1}^{(k)}, y) \end{aligned}$$

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 μ, σ^2

Full conditionals

- Write joint posterior as product of likelihood and all levels of priors
- To obtain full conditional for a particular parameter, examine product of all terms in the joint posterior that contain it

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

Directed graphs

- *directed graphical model* represents all quantities in a statistical model as nodes in a directed graph
- arrows run into nodes from the nodes that directly influence them (i.e. from their *parents*)
- developers of BUGS and WinBUGS recommend drawing directed graphs as part of model development process
- WinBUGS includes DoodleBUGS, which lets you specify models as directed graphs instead of using WinBUGS language
 - limited in what kinds of models can be specified in DoodleBUGS

Types of nodes in directed graphs

- constants
 - fixed by design of study
 - always are *founder notes* (i.e. do not have parents)
 - denoted as single- or sometimes double-edged rectangles
- stochastic nodes
 - variables that are given a distribution
 - may be parents or children or both
 - may be observed data or unobservable parameters
 - generally denoted as circles
 - * although data often denoted as single-edged rectangles
- deterministic nodes
 - logical functions of other nodes

Example: directed graph for Pumps problem

Types of directed links in a directed graph

- stochastic dependence
 - indicated by solid arrow
- logical function
 - indicated by dashed or hollow arrow

Example: An AR(1) time-series with additive normal measurement error

- Assume constant mean μ
- Specify likelihood so as to make observed data points y conditionally independent given parameters

$$f(y_i | \mu, \delta_i, \tau^2) = N\left(\mu + \delta_i, \frac{1}{\tau^2}\right), \quad i = 1, \dots, n$$

- Second stage: Prior on vector of δ s
 - There are several equivalent ways to do this.

- Third stage: Priors on μ , τ^2 , and τ_δ^2

The Metropolis and Metropolis-Hastings algorithms

- can be used to define a transition kernel for a Markov chain
 - may be multivariate and generate a vector of samples from all unknown quantities at once
 - or may be used within Gibbs sampler to generate from full conditionals of single or blocked parameters
- idea:
 - denote the joint unnormalized posterior distribution as $p(\theta_1, \theta_2, \dots, \theta_p | \mathbf{y})$
 - * we are trying to construct a Markov chain with this as its stationary distribution
 - * we don't know normalizing constant
 - denote vector of values from k th iteration of Markov chain as $\theta_1^{(k)}, \theta_2^{(k)}, \dots, \theta_p^{(k)}$
 - at each iteration, “propose” a vector of values $\theta_1^{new}, \theta_2^{new}, \dots$ from a “candidate-generating” density $q(\theta_1^{new}, \theta_2^{new}, \dots, \theta_p^{new} | \theta_1^{(k-1)}, \theta_2^{(k-1)}, \dots, \theta_p^{(k-1)})$
 - compute an “acceptance probability” α , and with probabilities

$$\begin{aligned}
 * \alpha, \text{ set } \theta_1^{(k+1)}, \theta_2^{(k+1)}, \dots, \theta_p^{(k+1)} &= \theta_1^{new}, \theta_2^{new}, \dots, \theta_p^{new} \\
 * 1-\alpha, \text{ set } \theta_1^{(k+1)}, \theta_2^{(k+1)}, \dots, \theta_p^{(k+1)} &= \theta_1^{(k)}, \theta_2^{(k)}, \dots, \theta_p^{(k)}
 \end{aligned}$$

- Metropolis
 - candidate-generating density must be symmetric
 - * e.g. $q(\theta^{(new)} | \theta^{(k-1)})$ is normal with mean $\theta^{(k-1)}$
 - acceptance probability is
- Metropolis-Hastings
 - candidate generating density does not have to be symmetric
 - acceptance probability is

$$\min \left(1, \frac{p(\theta^{(new)} | \mathbf{y})}{p(\theta^{(k-1)} | \mathbf{y})} \right)$$

$$\min \left(1, \frac{p(\theta^{(new)} | \mathbf{y}) \mathbf{q}(\theta^{(k-1)} | \theta^{(new)})}{p(\theta^{(k-1)} | \mathbf{y}) \mathbf{q}(\theta^{(new)} | \theta^{(k-1)})} \right)$$