Bayesian Computation

Lin Zhang

Department of Biostatistics
School of Public Health
University of Minnesota

Introduction

- Base of Bayesian inference – posterior distribution

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\mathbf{x}|\boldsymbol{\theta})\pi(\theta)}{m(\mathbf{x})}$$

- However, $p(\boldsymbol{\theta}|\mathbf{x})$ is often NOT analytically tractable.
  - $f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ is not proportional to a "family" density.
  - The normalizing constant

    $$m(\mathbf{x}) = \int_{\boldsymbol{\Theta}} f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$$

    does not have a closed form.

- Solution: approximate the posterior or generate samples from the posterior without knowing $m(x)$.

Bayesian Computational Methods

- Asymptotic approximation methods
  - Normal approximation
  - Laplace approximation
  - – Work for large n, low-dimensional $\boldsymbol{\theta}$
- Non-iterative Monte Carlo methods
  - Direct sampling ← we have seen examples in hierarchical models
  - Indirect sample: rejection sampling, importance sampling
  - – low-dimensional $\boldsymbol{\theta}$, posterior curve vaguely known
- Markov chain Monte Carlo (MCMC) methods
  - Gibbs algorithm
  - Metropolis algorithm
  - Other advance MCMC algorithm
  - – Work for complicated and/or high-dimensional posterior. Most popular!

Asymptotic Normal Approximation

- When n is large, $p(\boldsymbol{\theta}|\mathbf{x})$ will be approximately normal.
- "Bayesian Central Limit Theorem": Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_i(x_i|\boldsymbol{\theta})$, and $\pi(\boldsymbol{\theta})$ is the prior for $\boldsymbol{\theta}$, which may be improper. Further suppose that the posterior distribution is proper and its mode exists. Then as $n \to \infty$,

$$p(\boldsymbol{\theta}|\mathbf{x}) \dot{\sim} N\left(\widehat{\boldsymbol{\theta}}^p, [I^p(\mathbf{x})]^{-1}\right) ,$$

where $\widehat{\boldsymbol{\theta}}^p$ is the posterior mode of $\boldsymbol{\theta}$ obtained by solving

$$\frac{\partial}{\partial \theta_j} \log p^*(\boldsymbol{\theta}|\mathbf{x}) = 0,$$

where $p^*(\boldsymbol{\theta}|\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ is the unnormalized posterior.

$$I^p_{ij}(\mathbf{x}) = -\left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log\left(p^*(\boldsymbol{\theta}|\mathbf{x})\right)\right]_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^p}$$

is minus the inverse Hessian of $\log p^*(\boldsymbol{\theta}|\mathbf{x})$ evaluated at the mode (the "generalized" observed Fisher information matrix).

Example: Beta-Binomial model

Suppose $X|\theta \sim Bin(n, \theta)$ and $\theta \sim Beta(1, 1)$.

- Let $p^*(\theta|x) = f(x|\theta)\pi(\theta)$, we have
$$\ell(\theta) = \log p^*(\theta|x) \propto x \log \theta + (n - x) \log(1 - \theta) \ .$$
Taking the derivative of $\ell(\theta)$ and equating to zero, we obtain $\hat{\theta}^p = \hat{\theta} = x/n$, the familiar binomial proportion.
- The second derivative is
$$\frac{\partial^2 \ell(\theta)}{\partial \theta^2} = \frac{-x}{\theta^2} - \frac{n - x}{(1 - \theta)^2} \ ,$$
such that,
$$\left.\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\right|_{\theta=\hat{\theta}} = -\frac{x}{\hat{\theta}^2} - \frac{n - x}{(1 - \hat{\theta})^2} = -\frac{n}{\hat{\theta}} - \frac{n}{1 - \hat{\theta}} \ .$$

Example: Beta-Binomial model

- Thus

  $$[I^p(x)]^{-1} = \left(\frac{n}{\hat{\theta}} + \frac{n}{1-\hat{\theta}}\right)^{-1} = \left(\frac{n}{\hat{\theta}(1-\hat{\theta})}\right)^{-1} = \frac{\hat{\theta}(1-\hat{\theta})}{n},$$

  which is the usual frequentist expression for $\widehat{Var}(\hat{\theta})$. Thus the Bayesian CLT gives

  $$p(\theta|x) \dot{\sim} N\left(\hat{\theta}\,,\ \frac{\hat{\theta}(1-\hat{\theta})}{n}\right)$$

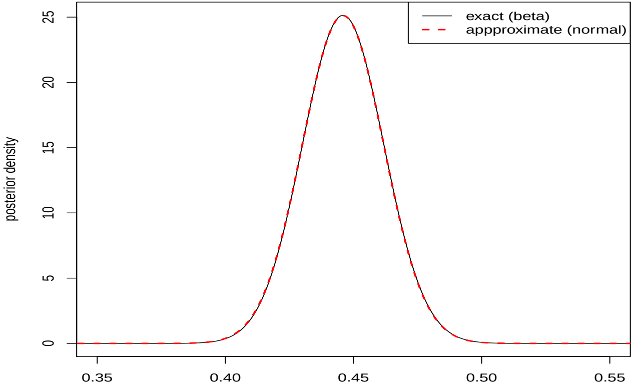
- Notice that a frequentist might instead use MLE asymptotics to write

  $$\hat{\theta}\,|\,\theta \dot{\sim} N\left(\theta\,,\ \frac{\hat{\theta}(1-\hat{\theta})}{n}\right)\ ,$$

  leading to identical inferences for θ, but for different reasons and with different interpretations!

Probability of female birth given placenta previa

Comparison of this normal approximation to the exact posterior, a $Beta(438, 544)$ distribution (recall $n = 980$):



Overlap with each other!

Higher order approximations

- The Bayesian CLT is a first order approximation, since

$$E(g(\boldsymbol{\theta})) = g(\hat{\boldsymbol{\theta}})\,[1 + O\,(1/n)]\ .$$

- Second order approximations (i.e., to order $O(1/n^2)$) again requiring only mode and Hessian calculations are available via Laplace's Method (BDA3, Chapter 13.3).

- Advantages of Asymptotic Methods:
  - deterministic, noniterative algorithm
  - substitutes differentiation for integration
  - computationally quick

- Disadvantages of Asymptotic Methods:
  - requires well-parametrized, unimodal posterior
  - $\boldsymbol{\theta}$ must be of at most moderate dimension
  - n must be large, *but is beyond our control*

Non-interative Monte Carlo Methods: Direct Sampling

- Suppose $\theta \sim p(\theta|\mathbf{y})$, and we are interested in the posterior mean of $f(\theta)$, which is given by

$$\gamma \equiv E[f(\theta)|\mathbf{y}] = \int f(\theta)p(\theta|\mathbf{y})d\theta.$$

- Approximations to the integral above can be carried out by Monte Carlo integration: Sample $\theta_1, \ldots, \theta_N$ independently from $p(\theta|\mathbf{y})$, and we can estimate γ by

$$\hat{\gamma} = \frac{1}{N}\sum_{j=1}^{N} f(\theta_j)$$

  which converges to $E[f(\theta)|\mathbf{y}]$ with probability 1 as $N \to \infty$ (strong law of large numbers).

- The use of Monte Carlo approximation requires that we are able to directly sample from the posterior distribution $p(\theta|\mathbf{y})$. The quality of the approximation increases as N increases, which we can control!

Example: Normal data with unknown mean and variance

- If $y_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, $i = 1, \ldots, n$, and $\pi(\mu, \sigma^2) = \frac{1}{\sigma^2}$, then the posterior is
$$\begin{aligned} \mu|\sigma^2, \mathbf{y} &\sim N(\bar{y}, \sigma^2/n) \,, \\ \text{and } \sigma^2|\mathbf{y} &\sim \text{inv-Gamma}\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right) \,, \end{aligned}$$
where $s^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)$.

- Draw posterior samples $\{(\mu_j, \sigma_j^2),\ j = 1, \ldots, N\}$ from $p(\mu, \sigma^2|\mathbf{y})$ as:
$$\begin{aligned} \text{sample } \sigma_j^2 &\sim \text{inv-Gamma}\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right) \,; \\ \text{then } \mu_j &\sim N(\bar{y}, \sigma_j^2/n),\ j = 1, \ldots, N \,. \end{aligned}$$

- To estimate the posterior mean: $\hat{E}(\mu|\mathbf{y}) = \frac{1}{N}\sum_{j=1}^{N}\mu_j$.

- Easy to estimate any function of $\boldsymbol{\theta} = (\mu, \sigma^2)$: To estimate the coefficient of variation, $\gamma = \sigma/\mu$, define $\gamma_j = \sigma_j/\mu_j$, $j = 1, \ldots, N$; summarize with moments or histograms!

Direct Sampling

- Monte Carlo integration allows for evaluation of its accuracy for any fixed N: Since $\hat{\gamma}$ is itself a sample mean of independent observations $f(\theta_1), \ldots, f(\theta_N)$, we have

$$Var(\hat{\gamma}) = \frac{1}{N} Var[f(\theta)|\mathbf{y}]$$

  Since $Var[f(\theta)|\mathbf{y}]$ can be estimated by the sample variance of the $f(\theta_j)$ values, a standard error estimate of $\hat{\gamma}$ is given by

$$\hat{se}(\hat{\gamma}) = \sqrt{\frac{1}{N(N-1)} \sum_{j=1}^{N} [f(\theta_j) - \hat{\gamma}]^2}\ .$$

- the CLT implies that $\hat{\gamma} \pm 2\,\hat{se}(\hat{\gamma})$ provides a 95% (frequentist!) CI for γ.

Indirect Methods: Importance Sampling

- Suppose $\theta \sim p(\theta|\mathbf{y})$ which can NOT be directly sampled from, and we wish to approximate
$$E[f(\theta)|\mathbf{y}] = \int f(\theta)p(\theta|\mathbf{y})d\theta = \frac{\int f(\theta)p^*(\theta|\mathbf{y})d\theta}{\int p^*(\theta|\mathbf{y})d\theta} ,$$
where $p^*(\theta|\mathbf{y}) = f(\mathbf{y}|\theta)\pi(\theta)$ is the unnormalized posterior.

- Suppose we can roughly approximate $p(\theta|\mathbf{y})$ by some density $g(\theta)$ from which we can easily sample – say, a multivariate t. Then define the weight function
$$w(\theta) = p^*(\theta|\mathbf{y})/g(\theta)$$

- Draw $\theta_j \overset{\text{iid}}{\sim} g(\theta)$, and we have
$$E[f(\theta)|\mathbf{y}] = \frac{\int f(\theta)w(\theta)g(\theta)d\theta}{\int w(\theta)g(\theta)d\theta} \approx \frac{\frac{1}{N}\sum_{j=1}^{N} f(\theta_j)w(\theta_j)}{\frac{1}{N}\sum_{j=1}^{N} w(\theta_j)} .$$
$g(\theta)$ is called the importance function.

- Remark: A good match of $g(\theta)$ to $p(\theta|\mathbf{y})$ will produce roughly equal weights, hence a good approximation.

Rejection sampling

- Here, instead of trying to approximate the posterior, we try to "blanket" it: suppose there exists a constant $M > 0$ and a smooth density $g(\theta)$, called the envelope function, such that

$$p^*(\theta|\mathbf{y}) < Mg(\theta)$$

for all θ.

- The algorithm proceeds as follows:

(i) Generate $\theta_j \sim g(\theta)$.

(ii) Generate $U \sim \text{Uniform}(0, 1)$.

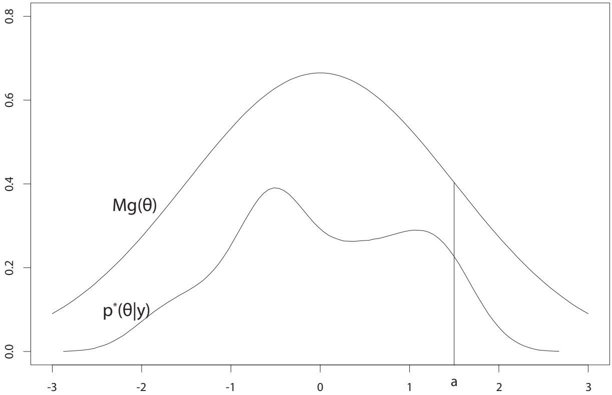
(iii) Accept θ_j if

$$U < \frac{p^*(\theta|\mathbf{y})}{Mg(\theta_j)}.$$

reject θ_j otherwise.

(iv) Repeat (i)-(iii) until the desired sample $\{\theta_j,\ j = 1, \ldots, N\}$ is obtained. The members of this sample will be random variables from the target posterior $p(\theta|\mathbf{y})$.

Rejection Sampling: informal "proof"



- Consider the θ_j samples in the histogram bar centered at a: the rejection step "slices off" the top portion of the bar. Repeat for all a: accepted θ_j mimic the lower curve!
- Remark: Need to choose M as small as possible (so as to maximize acceptance rate), and watch for "envelope violations"!

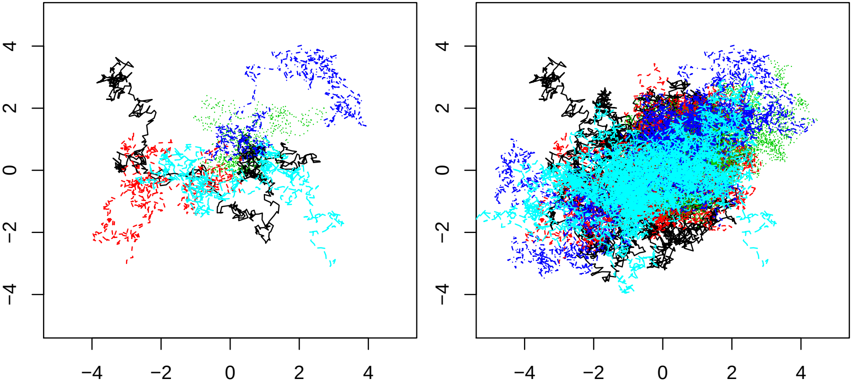
Markov chain Monte Carlo (MCMC) methods

- In many problems, it is difficult or impossible to find a feasible importance or envelope density, especially for high-dimensional $\boldsymbol{\theta}$.
- Luckily, iterative MC methods such as the Metropolis and Gibbs algorithms can be used to draw samples sequentially via Markov chain simulation that converge in distribution to the target posterior $p(\boldsymbol{\theta}|\mathbf{y})$.
- Markov chain is a sequence of random variables $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$, for which, for any $t \geq 1$, $\boldsymbol{\theta}^{(t+1)}$ is sampled from a distribution $T(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ which depends only on $\boldsymbol{\theta}^{(t)}$. $T(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is called the transition kernel distribution.
- In MCMC algorithm, the transition kernel must be constructed so that the Markov chain converges to a unique stationary distribution, which is our target posterior $p(\boldsymbol{\theta}|\mathbf{y})$, i.e.

$$\int T(\boldsymbol{\theta}|\boldsymbol{\theta}^c)p(\boldsymbol{\theta}^c|\mathbf{y})d\boldsymbol{\theta}^c = p(\boldsymbol{\theta}|\mathbf{y}).$$

Example: MCMC chains

Target distribution: $\boldsymbol{\theta} \sim N_2\left(0, \left[\begin{array}{cc} 1 & 0.7 \\ 0.7 & 1 \end{array}\right]\right)$.



Five independent MCMC chains with over-dispersed starting points.
The all converge to the same target bivariate normal distribution!

Metropolis algorithm

– Used when the target posterior $p(\theta|\mathbf{y})$ is not available in closed form, and importance or envelop functions are hard to find.

- Instead, we work with the unnormalized posterior $p^*(\theta|\mathbf{y})$, which is proportional to $p(\boldsymbol{\theta}|\mathbf{y})$ with a (unknown) proportionality constant $m(y)$.

- Metropolis algorithm works by drawing a candidate value, $\boldsymbol{\theta}^*$, from some proposal distribution $q(\theta^*|\theta^{(t-1)})$ that easy to sample, and then using a acceptance/rejection rule to correct the draw so as to better approximate the target distribution.

- Metropolis requires that the proposal density $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)})$ satisfies
$$q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)}) = q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^*) \ ,$$
i.e., q is symmetric in its arguments.

Metropolis algorithm (cont'd)

Given a starting value $\boldsymbol{\theta}^{(0)}$ at iteration $t = 0$, the algorithm proceeds as follows:

- **Metropolis Algorithm:** For $(t = 1, \ldots, T)$, repeat:
  1. Draw $\boldsymbol{\theta}^*$ from $q(\cdot|\boldsymbol{\theta}^{(t-1)})$
  2. Compute the ratio
  $$\alpha = \frac{p(\boldsymbol{\theta}^*|\mathbf{y})}{p(\boldsymbol{\theta}^{(t-1)}|\mathbf{y})} = \frac{p^*(\boldsymbol{\theta}^*|\mathbf{y})}{p^*(\boldsymbol{\theta}^{(t-1)}|\mathbf{y})}.$$
  3. Accept $\boldsymbol{\theta}^*$ and set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$ with probability $\min(\alpha, 1)$;
     Reject $\boldsymbol{\theta}^*$ and set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$ otherwise.
- Then a draw $\boldsymbol{\theta}^{(t)}$ converges in distribution to a draw from the true posterior density $p(\boldsymbol{\theta}|\mathbf{y})$.
- Note: The transition kernel density is
$$T(\boldsymbol{\theta}^*|\boldsymbol{\theta}^c) = q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^c)\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}^c),$$
  which satisfies the stationarity condition. (check!)

Metropolis algorithm (cont'd)

- How to choose the proposal density? The usual approach is to set

$$q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)}) = N(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)}, \widetilde{\Sigma}) \; .$$

- It's crucial to choose an appropriate $\widetilde{\Sigma}$ (moving stepsize):
  - Too large stepsize leads to extremely low acceptance ratio (chain not moving).
  - Too small stepsize results in slow movements (slow convergence).
  - In one dimension, MCMC "folklore" suggests choosing $\widetilde{\Sigma}$ to provide an observed acceptance ratio near 50%.

- Hastings (1970) showed we can drop the requirement that q be symmetric, provided we use

$$\alpha = \frac{p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^{(t-1)} \mid \boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^{(t-1)})q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(t-1)})}$$

  – useful for asymmetric target densities!

  – this form called the Metropolis-Hastings algorithm

Example: beetles under CS_2 exposure

- The data (Bliss, 1935) record the number of adult flour beetles killed after 5 hours of exposure to various levels of CS_2.

Dosage	# killed	# exposed
w_i	y_i	n_i
1.6907	6	59
1.7242	13	60
⋮	⋮	⋮
1.8639	60	60

- Consider the model

$$P(\text{death}|w_i) \equiv g(w_i) = \left[\frac{\exp(x_i)}{1+exp(x_i)}\right]^{m_1}, \qquad x_i = \frac{w_i - \mu}{\sigma}.$$

- Priors:

$$\begin{aligned} m_1 &\sim gamma(a_0, b_0) \\ \mu &\sim N(c_0, d_0) \\ \sigma^2 &\sim IG(e_0, f_0) \end{aligned}$$

Vague priors with $a_0 = .25$, $b_0 = 4$, $c_0 = 2$, $d_0 = 10$, $e_0 = 2$, $f_0 = 1000$.

Example: beetles under CS_2 exposure

- Posterior:

$$\begin{aligned} p(\mu, \sigma^2, m_1 | \mathbf{y}) &\propto f(\mathbf{y}|\mu, \sigma^2, m_1)\pi(\mu, \sigma^2, m_1) \\ &\propto \left\{ \prod_{i=1}^{k} [g(w_i)]^{y_i} [1 - g(w_i)]^{n_i - y_i} \right\} \\ &\quad \times \frac{m_1^{a_0 - 1}}{(\sigma^2)^{e_0+1}} \exp\left\{ -\frac{1}{2}\left(\frac{\mu - c_0}{d_0}\right)^2 - \frac{m_1}{b_0} - \frac{1}{f_0\sigma^2} \right\}. \end{aligned}$$

- Transformation: $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3) = (\mu, \frac{1}{2}\log(\sigma^2), \log(m_1))$. This will be nice for us to work with Gaussian proposal densities.

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{y}) &\propto \left\{ \prod_{i=1}^{k} [g(w_i)]^{y_i} [1 - g(w_i)]^{n_i - y_i} \right\} \times \exp(a_0\theta_3 - 2e_0\theta_2) \\ &\quad \times \exp\left\{ -\frac{1}{2}\left(\frac{\theta_1 - c_0}{d_0}\right)^2 - \frac{\exp(\theta_3)}{b_0} - \frac{\exp(-2\theta_2)}{f_0} \right\}. \end{aligned}$$

- Gaussian Proposal density:

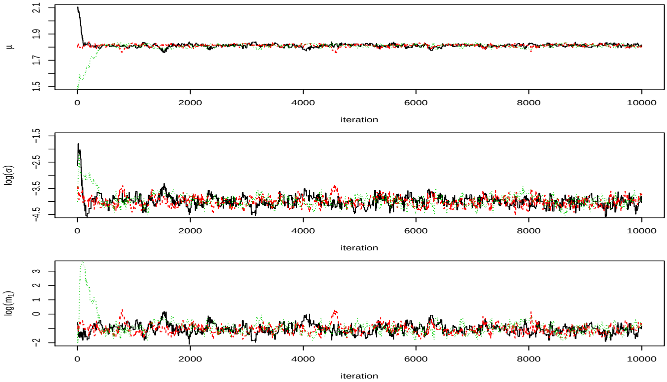
$$q(\boldsymbol{\theta}^* | \boldsymbol{\theta}^{(t-1)}) = N(\boldsymbol{\theta}^{(t-1)}, \tilde{\Sigma}), \quad \tilde{\Sigma} = \mathrm{diag}(.00012, .033, .10).$$

Metropolis algorithm

- Now we have all the components
  1. Proposal density:
  $$q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)}) = N(\boldsymbol{\theta}^{(t-1)}, \tilde{\Sigma}), \quad \tilde{\Sigma} = \text{diag}(.00012, .033, .10).$$
  2. Unnormalized posterior for the transformed paramters $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3) = (\mu, \frac{1}{2}\log(\sigma^2), \log(m_1))$:
  $$\begin{aligned} p^*(\boldsymbol{\theta}|\mathbf{y}) \quad \propto \quad & \left\{ \prod_{i=1}^{k} [g(w_i)]^{y_i} [1 - g(w_i)]^{n_i - y_i} \right\} \times \exp(a_0\theta_3 - 2e_0\theta_2) \\ & \times \exp\left\{ -\frac{1}{2}\left(\frac{\theta_1 - c_0}{d_0}\right)^2 - \frac{\exp(\theta_3)}{b_0} - \frac{\exp(-2\theta_2)}{f_0} \right\}. \end{aligned}$$

- Then run the Metropolis allgorithm: For $(t = 1, \ldots, T)$, repeat:
  (i) Draw $\boldsymbol{\theta}^*$ from $q(\cdot|\boldsymbol{\theta}^{(t-1)})$
  (ii) Compute the ratio
  $$\alpha = \frac{p(\boldsymbol{\theta}^*|\mathbf{y})}{p(\boldsymbol{\theta}^{(t-1)}|\mathbf{y})} = \frac{p^*(\boldsymbol{\theta}^*|\mathbf{y})}{p^*(\boldsymbol{\theta}^{(t-1)}|\mathbf{y})}.$$
  (iii) Accept $\boldsymbol{\theta}^*$ and set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$ with probability $\min(\alpha, 1)$;
  Reject $\boldsymbol{\theta}^*$ and set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$ otherwise.

Example: beetles under CS_2 exposure



Slow convergence due to low acceptance rate (13.5%).

Reason: high correlations $\widehat{corr}(\theta_1, \theta_2) = -.78$, $\widehat{corr}(\theta_1, \theta_3) = -.94$, $\widehat{corr}(\theta_2, \theta_3) = .89$.

Solution: try proposal $N(0, 2\hat{\Sigma})$, where $\hat{\Sigma} = \sum_{t=1}^{T} (\boldsymbol{\theta}^{(t)} - \hat{\boldsymbol{\theta}})(\boldsymbol{\theta}^{(t)} - \hat{\boldsymbol{\theta}})' / T$.

Adaptive MCMC: refine/improve sampling based on early MCMC outputs.

Posterior Inference based on MCMC samples

- For t sufficiently large (say, bigger than t_0), $\{\boldsymbol{\theta}^{(t)}\}_{t=t_0+1}^{T}$ is a (correlated) sample from the true posterior.

- We might therefore use a sample mean to estimate the posterior mean of one parameter θ_i, i.e.,

$$\widehat{E}(\theta_i|\mathbf{y}) = \frac{1}{T - t_0} \sum_{t=t_0+1}^{T} \theta_i^{(t)} \ .$$

- The time from $t = 0$ to $t = t_0$ is commonly known as the *burn-in* period; one can safely adapt (change) an MCMC algorithm during this pre-convergence period, since these samples will be discarded anyway

Posterior Inference based on MCMC samples (cont'd)

- In practice, we may actually run m *parallel* MCMC sampling chains, instead of only 1, for some modest m (say, $m = 5$). Discarding the burn-in period, we obtain

  $$\widehat{E}(\theta_i|\mathbf{y}) = \frac{1}{m(T - t_0)} \sum_{j=1}^{m} \sum_{t=t_0+1}^{T} \theta_{i,j}^{(t)} \ ,$$

  where now the j subscript indicates chain number.

- A posterior density estimate $\hat{p}(\theta_i|\mathbf{y})$ may be obtained by smoothing the histogram of the $\{\theta_{i,j}^{(t)}\}$, or as

  $$\begin{aligned} \hat{p}(\theta_i|\mathbf{y}) &= \frac{1}{m(T - t_0)} \sum_{j=1}^{m} \sum_{t=t_0+1}^{T} p(\theta_i|\theta_{k\neq i,j}^{(t)} \, , \, \mathbf{y}) \\ &\approx \int p(\theta_i|\theta_{k\neq i}, \mathbf{y}) p(\theta_{k\neq i}|\mathbf{y}) d\theta_{k\neq i} \end{aligned}$$

Example: beetles under CS$_2$ exposure

- Posterior mean of μ obtained from 3 parallel chains after discarding the first 1000 as burnin

$$\hat{E}(\mu = \theta_1|\mathbf{y}) = \frac{1}{3 \times 9000} \sum_{j=1}^{3} \sum_{t=1001}^{10000} \theta_{1,3}^{(t)} = 1.81.$$

- Posterior mean of m_1:

$$\hat{E}(m_1 = \exp(\theta_3)|\mathbf{y}) = \frac{1}{3 \times 9000} \sum_{j=1}^{3} \sum_{t=1001}^{10000} \exp(\theta_{3,j}^{(t)}) = 0.37.$$

Gibbs Sampling

– General MCMC procedure for high-dimensional $\boldsymbol{\theta}$.

- Suppose we have a collection of K random variables (or parameters) $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, and the full conditional distributions
$$\{p_i(\theta_i|\boldsymbol{\theta}_{(-i)}),\ i = 1, \ldots, K\}$$
are available for sampling ("available" means that samples may be directly generated from the distribution). $\boldsymbol{\theta}_{(-i)}$ denotes the components of $\boldsymbol{\theta}$ other than θ_i.

- Under mild conditions, the one-dimensional conditional distributions uniquely determine the full joint distribution of $\boldsymbol{\theta}$.

- Gibbs sampler simulates a Markov chain $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(T)}$ by sampling each element θ_i one at a time from its full conditional distribution $p_i(\theta_i|\theta_{(-i)})$ (while treating other elements as fixed).

Gibbs sampling (cont'd)

Given an arbitrary set of starting values $\{\theta_1^{(0)}, \ldots, \theta_K^{(0)}\}$ at iteration $t = 0$, Gibbs sampling proceeds as follows:

- **Gibbs Sampling:** For $(t = 1, \ldots, T)$, repeat:
  1. Draw $\theta_1^{(t)} \sim p_1(\theta_1 | \theta_2^{(t-1)}, \ldots, \theta_K^{(t-1)})$,
  2. Draw $\theta_2^{(t)} \sim p_2(\theta_2 | \theta_1^{(t)}, \theta_3^{(t-1)}, \ldots, \theta_K^{(t-1)})$,

     $\vdots$

  K. Draw $\theta_K^{(t)} \sim p_K(\theta_K | \theta_1^{(t)}, \ldots, \theta_{K-1}^{(t)})$.

- Under mild conditions,

$$(\theta_1^{(t)}, \ldots, \theta_K^{(t)}) \xrightarrow{d} (\theta_1, \cdots, \theta_K) \sim p \text{ as } t \to \infty \,.$$

- Note: The transition kernel density is

$$\begin{aligned} T(\boldsymbol{\theta}^* | \boldsymbol{\theta}^c) &= p_1(\theta_1^* | \theta_2^c, \ldots, \theta_K^c) \times p_2(\theta_2^* | \theta_1^*, \theta_3^c, \ldots, \theta_K^c) \\ &\quad \times \cdots \times p_K(\theta_K^* | \theta_1^*, \ldots, \theta_{K-1}^*), \end{aligned}$$

  which can be shown to satisfy the stationarity condition.

Pump Example

- Data: Consider a pump dataset about $k = 10$ different systems of a certain nuclear power plant. For each system $i = 1, \ldots, k$, the number of pump failures, Y_i, is observed in s_i thousands of hours.

i	Y_i	s_i	r_i
1	5	94.320	.053
2	1	15.720	.064
$\vdots$	$\vdots$	$\vdots$	$\vdots$
10	22	10.480	2.099

Pump Example: Poisson-gamma model

- Consider the modified Poisson/gamma model

  $$Y_i|\theta_i \overset{ind}{\sim} Poisson(\theta_i s_i),\ \theta_i|\alpha,\beta \overset{ind}{\sim} Gamma(\alpha,\beta).$$

  Add hyperprior

  $$\beta \sim IG(c,d),\ i=1,\ldots,k,$$

  where α, c, d, and the s_i are known.

- Thus we have *the hierarchical model*

  $$\begin{aligned} f(y_i|\theta_i) &= \frac{e^{-(\theta_i s_i)}(\theta_i s_i)^{y_i}}{y_i!},\ y_i \geq 0,\ \theta_i > 0, \\ g(\theta_i|\beta) &= \frac{\theta_i^{\alpha-1}e^{-\theta_i/\beta}}{\Gamma(\alpha)\beta^\alpha},\ \alpha > 0,\ \beta > 0, \\ h(\beta) &= \frac{e^{-1/(\beta d)}}{\Gamma(c)d^c\beta^{c+1}},\ c > 0,\ d > 0. \end{aligned}$$

  Note g is conjugate for f, and h is conjugate for g!

Pump Example: Poisson-gamma model

- The joint posterior distribution

$$p(\boldsymbol{\theta}, \beta|\mathbf{y}) \propto \left[\prod_{i=1}^{k} f(y_i|\theta_i)g(\theta_i|\beta)\right] h(\beta) = p^*(\boldsymbol{\theta}, \beta|\mathbf{y}).$$

- To implement the Gibbs sampler, we require the full conditional distributions of β and each θ_i.

- By Bayes' Rule,

$$\begin{aligned} p(\theta_i|\theta_{j\neq i}, \beta, \mathbf{y}) &= \frac{p(\boldsymbol{\theta}, \beta|\mathbf{y})}{\int p(\boldsymbol{\theta}, \beta|\mathbf{y})d\theta_i} \\ p(\beta|\boldsymbol{\theta}, \mathbf{y}) &= \frac{p(\boldsymbol{\theta}, \beta|\mathbf{y})}{\int p(\boldsymbol{\theta}, \beta|\mathbf{y})d\beta}, \end{aligned}$$

  each is proportional to $p(\boldsymbol{\theta}, \beta|\mathbf{y})$, and thus is also proportional to $p^*(\boldsymbol{\theta}, \beta|\mathbf{y})$.

- Thus we can find the full conditional distribution for each parameter by dropping irrelevant terms from $p^*(\boldsymbol{\theta}, \beta|\mathbf{y})$, and normalizing!

Pump Example: Poisson-gamma model

$$\begin{aligned}
p(\theta_i|\theta_{j\neq i},\beta,\mathbf{y}) &\propto p^*(\boldsymbol{\theta},\beta|\mathbf{y}) = \left[\prod_{l=1}^{k} f(y_l|\theta_l)g(\theta_l|\beta)\right] h(\beta) \\
&\propto f(y_i|\theta_i)g(\theta_i|\beta) \\
&\propto \theta_i^{y_i+\alpha-1} e^{-\theta_i(s_i+1/\beta)} \\
&\propto G\left(\theta_i \mid y_i+\alpha,(s_i+1/\beta)^{-1}\right)\ , \quad \text{and}
\end{aligned}$$

$$\begin{aligned}
p(\beta|\boldsymbol{\theta},\mathbf{y}) &\propto p^*(\boldsymbol{\theta},\beta|\mathbf{y}) = \left[\prod_{i=1}^{k} f(y_i|\theta_i)g(\theta_i|\beta)\right] h(\beta) \\
&\propto \left[\prod_{i=1}^{k} g(\theta_i|\beta)\right] h(\beta) \propto \left[\prod_{i=1}^{k} \frac{e^{-\theta_i/\beta}}{\beta^{\alpha}}\right] \frac{e^{-1/(\beta d)}}{\beta^{c+1}} \\
&\propto \frac{e^{-\frac{1}{\beta}\left(\sum_{i=1}^{k}\theta_i+\frac{1}{d}\right)}}{\beta^{k\alpha+c+1}} \\
&\propto IG\left(\beta|k\alpha+c,\left(\sum_{i=1}^{k}\theta_i+1/d\right)^{-1}\right)\ .
\end{aligned}$$

Thus the $\{\theta_i^{(t)}\}$ and $\beta^{(t)}$ may be sampled directly!

Pump Example: Poisson-gamma model

- Set $c = 0.1$ and $d = 1.0$ for a vague hyperprior for β.
- We can run the Gibbs Sampling as follows: at each iteration t
  1. Draw $\theta_i^{(t)} \sim Gamma\left(y_i + \alpha, rate = (s_i + 1/\beta^{(t-1)})^{-1}\right)$ for all i's
  2. Draw $\beta^{(t)} \sim IG\left(k\alpha + c, scale = \sum_{i=1}^{k} \theta_i^{(t)} + 1/d\right)$
- If α were also unknown, we use, say, a prior $h(\alpha) = Exp(\mu)$. Then, the full conditional for α

$$\begin{aligned} p(\alpha|\beta, \{\theta_i\}, \mathbf{y}) &\propto \left[\prod_{i=1}^{k} Gamma(\theta_i|\alpha, \beta)\right] \pi_2(\alpha) \\ &\propto \left[\prod_{i=1}^{k} \frac{\theta_i^{\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}}\right] e^{-\alpha/\mu} \end{aligned}$$

  is not proportional to any standard family. We can NOT directly sample α from its full conditional distribution.

Pump Example: Poisson-gamma model

- When the full conditional of certain parameter can NOT be directly sample from, we resort to:
  - adaptive rejection sampling (ARS): provided $p(\alpha|\{\theta_i\}, \beta, \mathbf{y})$ is log-concave, or
  - Metropolis-Hastings sampling

  Note: This is the order the `WinBUGS` software uses when deriving full conditionals!

---

* This is the standard "hybrid approach": Use Gibbs overall, with "substeps" for awkward full conditionals

Pump Example: MH-within-Gibbs Hybrid sampling

- To conduce an MH sampling algorithm, we will use a normal proposal density. However, α is defined on the positive real line, we make a *transformation* on α: $a = \log(\alpha)$. The full conditional for a is then

$$p(a|\beta, \boldsymbol{\theta}, \mathbf{y}) \propto \left[\prod_{i=1}^{k} \frac{\theta_i^{e^a - 1}}{\Gamma(e^a)\beta^{e^a}} \right] e^{-e^a/\mu + a} = p^*(a|\beta, \boldsymbol{\theta}, \mathbf{y})$$

- Now at each iteration t, sample a by:
  (i) Draw a^* from $q(\cdot|a^{(t-1)}) = N(a^{(t-1)}, 0.5^2)$

  (ii) Compute the ratio

$$\alpha_{\text{accept}} = \frac{p^*(a|\beta, \boldsymbol{\theta}, \mathbf{y})}{p^*(a^{(t-1)}|\beta, \boldsymbol{\theta}, \mathbf{y})}.$$

  Accept the a^*, and take $a^{(t)} = a^*$ with probability $\min(\alpha_{\text{accept}}, 1)$; otherwise, take $a^{(t)} = a^{(t-1)}$.

Convergence Monitoring

When it is safe to stop and summarize MCMC output?

- An MCMC algorithm is said to have converged at time T if its output can be "safely" thought of as coming from the true stationary distribution $p(\boldsymbol{\theta}|\mathbf{y})$ for all $t > T$.
- However, we do not know $p(\boldsymbol{\theta}|\mathbf{y})$; all we can hope to see is $\int |\hat{p}_t(\boldsymbol{\theta}) - \hat{p}_{t+k}(\boldsymbol{\theta})| d\boldsymbol{\theta} < \epsilon$!
- Common cause of convergence failure: Nonidentifiability (due to over overparameterization!) Example:

$$y_i|\theta_1, \theta_2 \overset{iid}{\sim} N(\theta_1 + \theta_2, 1).$$

- Overparameterization also typically lead to high posterior correlations amongst parameters, resulting in slow convergence.
- One remedy is to reparameterize, but hard to do in general!

Convergence Diagnostics Statistics

Gelman and Rubin (1992, *Statistical Science*)

1. Run a small number (m) of parallel chains with overdispersed starting points
2. Run the m chains for $2N$ iterations each, and we then compare the **variation within chains** to the **total variation across chains** during the latter N iterations.
3. Specifically, we monitor convergence by the estimated scale reduction factor

$$\sqrt{\hat{R}} = \sqrt{\left(\frac{N-1}{N} + \frac{m+1}{mN}\frac{B}{W}\right)\frac{df}{df-2}},$$

where B/N is the variance between the means from the m parallel chains, W is the average of the m within-chain variances, and df is the degrees of freedom of an approximating t density to the posterior.

$\sqrt{\hat{R}} \to 1$ as $N \to \infty$. Thus $\sqrt{\hat{R}}$ close to 1 suggests good convergence.

Convergence diagnosis strategy

- Run a few (3 to 5) parallel chains, with starting points believed to be overdispersed
  - say, covering ±3 prior standard deviations from the prior mean
- Overlay the resulting sample traces for a representative subset of the parameters
  - say, most of the fixed effects, some of the variance components, and a few well-chosen random effects)
- Annotate each plot with Gelman and Rubin diagnostics and lag 1 sample autocorrelations
  - autocorrelation close to 0 → near-independence → fast convergence
  - autocorrelation close to 1 → "stuck" chain
- Investigate bivariate plots and cross-correlations among parameters suspected of being nonidentifiable.

Other sampling algorithm

- Blocked Gibbs sampler: sample a set of parameters from their joint conditional posterior
- Slice Sampler: alternative to Metropolis steps and have excellent convergence properties
- Hamilton Monte Carlo algorithm: used when there are a large number of parameters that do not have closed-form full conditional posteriors

Variance estimation

How good is our MCMC estimate once we get it?

- Suppose a single long chain of (post-convergence) MCMC samples $\{\lambda^{(t)}\}_{t=1}^{N}$. A simple estimator of $E(\lambda|\mathbf{y})$ is

$$\hat{E}(\lambda|\mathbf{y}) = \hat{\lambda}_N = \frac{1}{N}\sum_{t=1}^{N}\lambda^{(t)} \; .$$

- Analogously, we could attempt to estimate $Var(\hat{\lambda_N})$ as

$$\widehat{Var}_{iid}(\hat{\lambda}_N) = s_\lambda^2/N = \frac{1}{N(N-1)}\sum_{t=1}^{N}(\lambda^{(t)} - \hat{\lambda}_N)^2 \; .$$

  But this is likely an underestimate due to positive autocorrelation in the MCMC samples.

Variance estimation (cont'd)

- To avoid wasteful parallel sampling or "thinning," compute the *effective sample size*,

  $$ESS = N/\kappa(\lambda) \ ,$$

  where $\kappa(\lambda)$ is the *autocorrelation time*,

  $$\kappa(\lambda) = 1 + 2\sum_{k=1}^{\infty} \rho_k(\lambda),$$

  where $\rho_k(\lambda)$ is lag k autocorrelation for λ. We may estimate $\kappa(\lambda)$ using MCMC samples, and cut off the sum when $\rho_k(\lambda) < \epsilon$.

- Then

  $$\widehat{Var}_{ESS}(\hat{\lambda}_N) = s_\lambda^2/ESS(\lambda) = \frac{\kappa(\lambda)}{N(N-1)}\sum_{t=1}^{N}(\lambda^{(t)} - \hat{\lambda}_N)^2 \ .$$

  Note: $\kappa(\lambda) \geq 1$, so $ESS(\lambda) \leq N$, and so we have that $\widehat{Var}_{ESS}(\hat{\lambda}_N) \geq \widehat{Var}_{iid}(\hat{\lambda}_N)$, in concert with intuition.

Variance estimation (cont'd)

- Another alternative is Batching: Divide the run into m successive batches of length k with batch means $b_1, \ldots, b_m$. Obviously $\hat{\lambda}_N = \bar{b} = \frac{1}{m}\sum_{i=1}^{m} b_i$, and

$$\widehat{Var}_{batch}(\hat{\lambda}_N) = \frac{1}{m(m-1)} \sum_{i=1}^{m} (b_i - \hat{\lambda}_N)^2 \ ,$$

  provided that k is large enough so that the batch means are nearly independent and m is large enough to reliably estimate $Var(b_i)$.

- Check lag 1 autocorrelation of b_i to verify independence of batch means.

- For any $\widehat{V}$ used to approximate $Var(\hat{\lambda}_N)$, a 95% CI for $E(\lambda|\mathbf{y})$ is then given by

$$\hat{\lambda}_N \pm z_{.025}\sqrt{\widehat{V}} \ .$$