Markov Chain Monte Carlo for Bayesian Statistical Inference

Sebastien Haneuse

Informal Seminar

14^{th} January, 2002

Overview

- Bayesian Paradigm Overview
- Implementation Methods
- Markov Chain Monte Carlo

Notation

- $\boldsymbol{y} = (y_1, ..., y_n)$ denotes the vector of observed data
- $\boldsymbol{\theta}$ denotes the unobservable vector of population parameters of interest

Bayesian Paradigm

Bayesian Inference

- Treat $\boldsymbol{\theta}$ as a random variable

- Condition on the observed data

- Inference is based upon the conditional distribution of $\boldsymbol{\theta}$ given the data $\boldsymbol{y}$

$$\pi(\boldsymbol{\theta}|\boldsymbol{y})$$

- Statistical conclusions are made in terms of probability statements.

- Joint distribution for $(\boldsymbol{\theta}, \boldsymbol{y})$:

$$\pi(\boldsymbol{\theta}, \boldsymbol{y}) = p(\boldsymbol{y}|\theta)\pi(\boldsymbol{\theta})$$

  - $p(\boldsymbol{y}|\theta)$: *Likelihood* for the data
  - $\pi(\boldsymbol{\theta})$: *Prior distribution*

- The prior distribution is putting relative weights to the values of $\boldsymbol{\theta}$, using previously gathered information.

- Using Bayes' Theorem we can compute the *Posterior distribution*:

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\theta)\pi(\boldsymbol{\theta})}{p(\boldsymbol{y})}$$

- Equivalently we could write down:

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\theta)\pi(\boldsymbol{\theta})$$

  since $p(\boldsymbol{y})$ is just a normalising constant.

Interpretation

- The prior distribution has been revised, via Bayes' Theorem, into the posterior.

Primary Tasks

- Develop models for $p(\boldsymbol{y}|\theta)$ and $\pi(\boldsymbol{\theta})$
- Summarize $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ in appropriate ways:
  - Compute expectations (integrals)

$$\mathrm{E}_{\pi(\boldsymbol{\theta}|\boldsymbol{y})}[g(\boldsymbol{\theta})] \;=\; \int_{\mathcal{S}} g(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y})\; d\boldsymbol{\theta}$$

Special Case:

$$\Pr(\boldsymbol{\theta} \;\in\; \mathcal{B}) \;=\; \int_{\mathcal{S}} I_{[\boldsymbol{\theta}\;\in\;\mathcal{B}]}\pi(\boldsymbol{\theta}|\boldsymbol{y})\; d\boldsymbol{\theta}$$

  - $\mathcal{S}$ parameter space.

Example 1.

- Suppose $Y_1, ..., Y_n$ iid $\mathcal{N}(\theta,\; \sigma^2)$, with σ^2 known

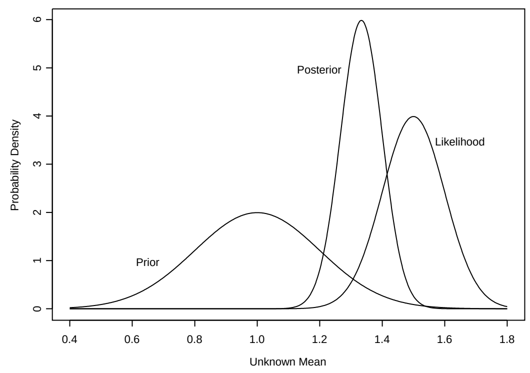
$$\bar{Y}|\theta \;\sim\; \mathcal{N}\left(\theta, \frac{\sigma^2}{n}\right)$$

- Previous data yields the prior:

$$\theta \;\sim\; \mathcal{N}(\mu, \tau^2)$$

- After some algebra we get the following posterior:

$$\theta|\boldsymbol{y} \;\sim\; \mathcal{N}\left(\bar{y}w + \mu(1-w), \frac{\sigma^2}{n}w\right), \quad w = \frac{\tau^2}{\frac{\sigma^2}{n} + \tau^2}$$



- This is an example of *Conjugacy*:

Prior	Likelihood	Posterior
Normal	Normal	Normal
Beta	Binomial	Beta
Gamma	Poisson	Gamma

- For each of these, the posterior is analytically tractible.

- We can write down $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ explicity.
  - Compute probability statements, via integration
  - Mean, median, mode and variance of $\boldsymbol{\theta}$ are known

Questions:

- What if the posterior is not analytically tractible ?
  - We need to compute

$$p(\boldsymbol{y}) = \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta}, \boldsymbol{y}) \, \partial\boldsymbol{\theta}$$

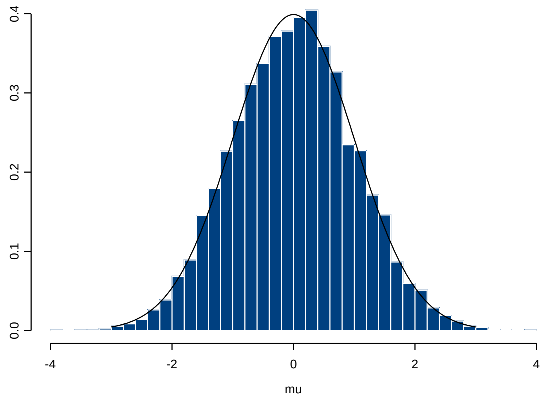
  - This may be very difficult!

- What if we are interested in making inference about $\phi = g(\boldsymbol{\theta})$?

Duality: Density vs Samples

- There is a duality between a density and samples generated by that density
  - Given a density we can generate samples
  - Given samples we can construct a density

- 10,000 samples from $\mathcal{N}(0,1)$:



- Generate samples from $\pi(\boldsymbol{\theta}|\boldsymbol{y})$
  - $\boldsymbol{\theta}^{(1)}$, $\boldsymbol{\theta}^{(2)}$, ..., $\boldsymbol{\theta}^{(n)}$

- Approximate expectations via sample averages

$$\widehat{\mathrm{E}}_{\pi(\boldsymbol{\theta}|\boldsymbol{y})}[g(\boldsymbol{\theta})] \ = \ \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{\theta}^{(i)})$$

$$\widehat{\Pr}(\boldsymbol{\theta} \ \in \ \mathcal{B}) \ = \ \frac{1}{n}\sum_{i=1}^{n} I_{[\boldsymbol{\theta}^{(i)} \ \in \ \mathcal{B}]}$$

- Quantities computed from the samples:
  - Mean : 0.008651355
  - Standard Deviation : 0.9959269

Implementation Methods

- Approximations based on posterior modes:
  - Normal-mixture to approximate the joint posterior.
  - Based on the curvature of the distribution around its modes.
  - Can also use mixtures of multivariate t-distributions.

- Direct simulations:
  - Monte Carlo methods; generate i.i.d samples.
  - Probability integral transforms (inverse cdf).
  - Rejection sampling.
  - Ratio of Uniforms.

- Numerical Integration:
  - Compute the posterior/normalising constant directly.
  - Laplace's Method; based upon multivariate Normal approximations.
  - Importance Sampling; simulating from some convenient density and re-weighting.

- Markov chain simualtion
  - Generate dependent samples.

Example 2.

- Suppose $Y_1, ..., Y_n$ iid $\mathcal{N}(\theta,\ \sigma^2)$, with σ^2 unknown.

- Priors for both unknown parameters:

$$\begin{aligned} \theta &\sim \mathcal{N}(\mu, \tau^2) \\ \sigma^{-2} &\sim \text{Gamma}(a, b) \end{aligned}$$

with μ, τ^2, a, b being specified a priori.

Joint Posterior:

$$\pi(\theta, \sigma^{-2}|\boldsymbol{y}) \;\propto\; L(\boldsymbol{y}|\theta, \sigma^{-2})\ \pi(\theta)\ \pi(\sigma^{-2})$$

Comments:

(a) Bivariate posterior: $\theta \in (-\infty, \infty)$ and $\sigma^{-2} > 0$.

(b) To compute the density we would need to compute an ulgy bivariate integral!

(c) How can we make inference about the coefficient of variation ?

$$c \;=\; \frac{\theta}{\sigma}$$

Conditional Posteriors

- Consider the conditional posteriors for each parameter:

$$\begin{aligned} \pi(\theta|\sigma^{-2}, \boldsymbol{y}) &\;\propto\; L(\boldsymbol{y}|\theta, \sigma^{-2})\ \pi(\theta) \\ &\;\sim\; \mathcal{N}\left(\bar{y}w + \mu(1-w), \frac{\sigma^2}{n}w\right) \\ & \qquad w = \frac{\tau^2}{\frac{\sigma^2}{n} + \tau^2} \end{aligned}$$

$$\begin{aligned} \pi(\sigma^{-2}|\theta, \boldsymbol{y}) &\;\propto\; L(\boldsymbol{y}|\theta, \sigma^{-2})\ \pi(\sigma^{-2}) \\ &\;\sim\; \text{Gamma}\left(\frac{n}{2} + a, \frac{\sum_{i=1}^n (y_i - \theta)^2}{2} + b\right) \end{aligned}$$

- These are two distributions from which we could draw samples.

Algorithm

(1) Specify some starting value: $\{\theta^{(0)}, \sigma^{-2(0)}\}$

(2) Generate a random draw from the first conditional:

$$\theta^{(1)} \sim \pi(\theta|\sigma^{-2(0)}, \boldsymbol{y})$$

(3) Generate a random draw from the second conditional, using the deviate from part (1):

$$\sigma^{-2(1)} \sim \pi(\sigma^{-2}|\theta^{(1)}, \boldsymbol{y})$$

We have just moved from $\{\theta^{(0)}, \sigma^{-2(0)}\}$ to $\{\theta^{(1)}, \sigma^{-2(1)}\}$

(4) Process continues with

$$\begin{aligned} \theta^{(t)} &\sim \pi(\theta|\sigma^{-2(t-1)}, \boldsymbol{y}) \\ \sigma^{-2(t)} &\sim \pi(\sigma^{-2}|\theta^{(t)}, \boldsymbol{y}) \end{aligned}$$

- This results in a sample of bivariate points:

$$\{\theta^{(0)}, \sigma^{-2(0)}\}, \ldots, \{\theta^{(n)}, \sigma^{-2(n)}\}$$

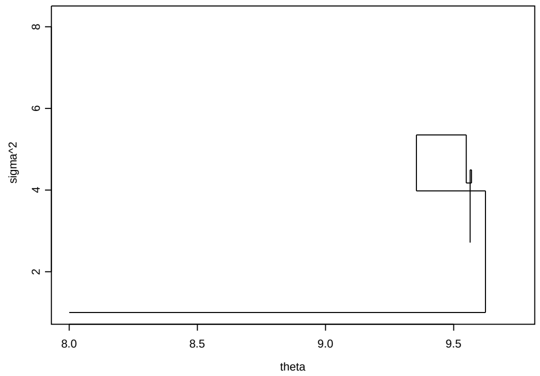


Figure 1: First 5 iterations, component-by-component

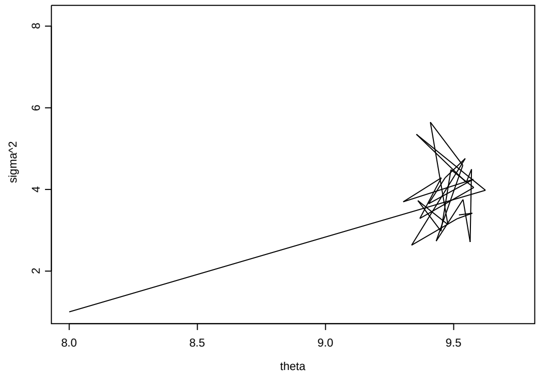


Figure 2: First 25 iterations

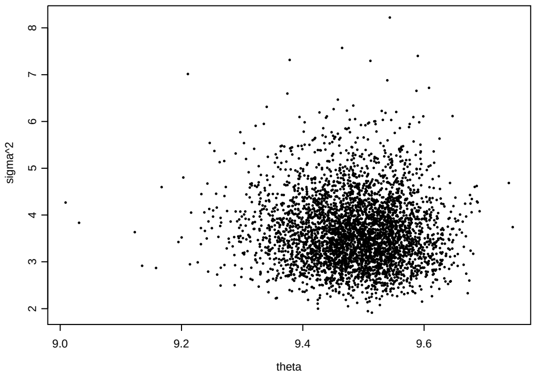


Figure 3: 5000 Bivariate samples; 1000 sample "burn-in"

Quantities of Interest

- We can compute posterior summaries using the samples:

Parameter	Median	95% Credible Interval
θ	9.49	(9.32, 9.62)
σ^2	3.57	(2.49, 5.40)
c	5.02	(4.06, 6.02)

Comments

- This is an example of Gibbs Sampling
- Key is to create a Markov process
  - Limiting stationary distribution is $\pi(\theta, \sigma^{-2}|\boldsymbol{y})$
- Generates dependent samples!
  - Correct procedure would be to sample from:

$$\pi(\theta|\boldsymbol{y}) \;\; \text{and then} \;\; \pi(\sigma^{-2}|\theta, \boldsymbol{y})$$

  - Then $\{\theta, \sigma^{-2}\} \sim \pi(\theta, \sigma^{-2}|\boldsymbol{y})$
  - Generally hard to sample iid deviates from Multivariate distributions.
- Require a "burn-in"
  - Early on, convergence not achieved.
  - Early samples are not distributed according to $\pi(\theta, \sigma^{-2}|\boldsymbol{y})$.
- Run the simulation long enough so that
  - Distribution of current draws $\approx$ stationary distribution
  - Approximation is improved at each iteration

Metroplis Algorithm

- Target distribution, $\pi(\boldsymbol{\theta}|\boldsymbol{y})$, known up to a normalising constant.

- Let $g(\boldsymbol{\theta}_a|\boldsymbol{\theta}_b) = g(\boldsymbol{\theta}_b|\boldsymbol{\theta}_a)$ be a symmetric probability density
  - *Jumping distribution*

- Let $\boldsymbol{\theta}^{(0)}$ denote the initial point.

- At iteration t,
  - Generate a prosposal point $\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)} \sim g(\cdot|\boldsymbol{\theta}^{(t-1)})$
  - Compute the ratio of densities:
$$r = \frac{\pi(\boldsymbol{\theta}^*|\boldsymbol{y})}{\pi(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{y})}$$
  - Accept $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$ with probability $= \min(r, 1)$
  - Otherwise set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$

Comments:

1. r does not depend on the normalising constant

2. Not all points are accepted

Why does this work ?

- Show that the distribution of the deviates at each iteration, $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$ converges to the target distribution:
  - Markov chain with a unique stationary distribution
  - stationary distribution = target distribution

- First step:
  - Markov chain is irreducible, aperiodic and not transient.
  - Consider a discrete system; $X^{(1)}$, $X^{(2)}$, ...
$$X^{(t)} \text{ takes on values in } \mathcal{S} = \{0, 1, \ldots, \text{s}\}$$
  - Transition probability, $i, j \in \mathcal{S}$
$$p_{ij} = P(X^{(t)} = j, X^{(t-1)} = i)$$
  - Assume system has the *Markov Property*

- Transition matrix

$$\mathbb{P} = \begin{bmatrix} p_{00} & p_{01} & \cdots & p_{0s} \\ p_{10} & p_{11} & \cdots & p_{1s} \\ \vdots & \vdots & \ddots & \vdots \\ p_{s0} & p_{s1} & \cdots & p_{ss} \end{bmatrix}$$

- Stationarity; $\mathbb{P}$ does not depend on t

- n-step transition probability

$$p_{ij}^{(n)} = \sum_{k=0}^{s} p_{ik}^{(n-1)} p_{kj}$$

$\Rightarrow$ state i to state j in n steps

- n-step transition matrix

$$\mathbb{P}^{(n)} = \left[p_{ij}^{(n)}\right]_{i,j=0}^{s} = \mathbb{P}^n = \underbrace{\mathbb{P}\cdots\mathbb{P}}_{\text{n times}}$$

- Steady state matrix: as $n \to \infty$

$$\mathbb{P}^{(n)} = \mathbb{P}^n \to \mathbb{P}^\infty$$

- If $\vec{\pi}$ is a probability vector satisfying *general balance*; $\vec{\pi}\mathbb{P} = \vec{\pi}$

$\Rightarrow$ $\vec{\pi}$ is the stationary distribution for $\mathbb{P}$.

- $\vec{\pi}$ vector of long-run probabilities of being in any particular state

- If the transition matrix, $\mathbb{P}$, is

(1) Irreducible: $p_{ij} > 0 \; \forall \; i, j = 1, \cdots, s$

(2) Aperiodic: no cycles

then $\vec{\pi}$ is unique and

$$\mathbb{P}^\infty = \begin{bmatrix} \vec{\pi} \\ \vdots \\ \vec{\pi} \end{bmatrix}$$

- Regardless of the starting distribution

$\Rightarrow$ Regardless of how you pick the initial points in the chain.

- Second step:

- Consider starting the algorithm at iteration $(t-1)$, with a draw $\boldsymbol{\theta}^{(t-1)}$ from $\pi(\boldsymbol{\theta}|\boldsymbol{y})$.

- Consider any two points, $\boldsymbol{\theta}_a$ and $\boldsymbol{\theta}_b$ labelled so that

$$\pi(\boldsymbol{\theta}_b|\boldsymbol{y}) \geq \pi(\boldsymbol{\theta}_a|\boldsymbol{y}).$$

- Unconditional probability of transition from θ_a to θ_b is
$$\pi(\boldsymbol{\theta}^{(t-1)} = \boldsymbol{\theta}_a,\ \boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}_b) \;=\; \pi(\boldsymbol{\theta}_a|\boldsymbol{y})g(\boldsymbol{\theta}_b|\boldsymbol{\theta}_a)$$
where acceptance probability $= 1$.

- Unconditional probability of transition from θ_b to θ_a is
$$\begin{aligned}\pi(\boldsymbol{\theta}^{(t-1)} = \boldsymbol{\theta}_b,\ \boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}_b) \;&=\; \pi(\boldsymbol{\theta}_b|\boldsymbol{y})g(\boldsymbol{\theta}_a|\boldsymbol{\theta}_b)\left(\frac{\pi(\boldsymbol{\theta}_a|\boldsymbol{y})}{\pi(\boldsymbol{\theta}_b|\boldsymbol{y})}\right) \\ &=\; \pi(\boldsymbol{\theta}_a|\boldsymbol{y})g(\boldsymbol{\theta}_b|\boldsymbol{\theta}_a).\end{aligned}$$
where acceptance probability $= r$.

- Same transition probabilities, since $g(\cdot|\cdot)$ is symmetric.

- Joint is symmetric
$\Rightarrow \boldsymbol{\theta}^{(t-1)}$ and $\boldsymbol{\theta}^{(t)}$ same marginal distribution; $\pi(\boldsymbol{\theta}|\boldsymbol{y})$

Note:

- More generally
  - Construct a Markov transition probability matrix, whose stationary distribution is the desired posterior.

- Ideas extend to continuous & multivariate systems.

Metroplis-Hastings Algorithm

- Generalises the basic Metropolis algorithm
  - Jumping distribution need not be symmetric.
  - Can increase the speed on convergence.

- To account for asymmetry, replace r with
$$r \;=\; \frac{\pi(\boldsymbol{\theta}^*|\boldsymbol{y})/g(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)})}{\pi(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{y})/g(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^*)}$$

Efficiency

- Ideally, we choose $g(\boldsymbol{\theta}^*|\boldsymbol{\theta}) \;=\; \pi(\boldsymbol{\theta}|\boldsymbol{y})$

$\Rightarrow$ ratio $r = 1$

- Generally,
  - For any $\boldsymbol{\theta}$ it is easy to sample from $g(\boldsymbol{\theta}^*|\boldsymbol{\theta})$
  - Easy to compute r
  - Jumps go a reasonable distance in the parameter space
  - Jumps not rejected too frequently

Gibbs Sampler

- Gibbs is a special case of the Metropolis-Hastings algorithm
- $\boldsymbol{\theta}$ divided into d subvector, $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_d)$
- Each iteration cycles through the subvectors
  - d Metropolis steps in each iteration
- For $j = 1, ..., d$, at iteration t, $\theta_j^{(t)}$ is sampled from

$$g_j(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)}) \;=\; \pi(\theta_j^*|\boldsymbol{\theta}_{-j}^{(t-1)}, \boldsymbol{y})$$

  - $\boldsymbol{\theta}^*$ differs from $\boldsymbol{\theta}^{(t-1)}$ only in the j^{th} component
  - $\theta_{-j}^{(t-1)}$ represents all other components in their current state:

$$\boldsymbol{\theta}_{-j}^{(t-1)} \;=\; (\theta_1^{(t)}, \cdots, \theta_{j-1}^{(t)}, \theta_{j+1}^{(t-1)}, \cdots, \theta_d^{(t-1)})$$

  - All jumps are accepted since $r = 1$.
- If sampling is impossible from some, or all, of the conditionals
  - Metropolis-Hastings algorithm

Assessing Convergence

Diffculties from iterative simulation

- If iterations have not proceeded long enough:
  - Simulations unrepresentative of target distribution
  - Especially early iterations
- Within-sequence correlation
  - Inference is less precise, than from the same number of independent draws.
  - Number of "effective" draws will be fewer.
  - Variances will be under-estimated.

Multiple Sequences

- Simulate $m \geq 2$ independent sequences of length n
  - Starting points from an overdispersed distribution
- Monitor each scalar parameter (or any other quantity)
  - Compare the between and within-sequence variances ($\mathbb{B}$ and $\mathbb{W}$) for each scalar, ψ.
  - Estimate the marginal posterior variance, $\mathrm{var}(\psi|\boldsymbol{y})$ via

$$\widehat{\mathrm{var}}^+(\psi|\boldsymbol{y}) \;=\; \frac{n-1}{n}\mathbb{W} \;+\; \frac{1}{n}\mathbb{B}$$

  - Estimate factor by which scale of current distribution we reduce if $n \;\to\; \infty$
  - As $n \;\to\; \infty$, the *potential scale reduction*

$$\sqrt{\widehat{R}} \;=\; \sqrt{\frac{\widehat{\mathrm{var}}^+(\psi|\boldsymbol{y})}{\mathbb{W}}} \;\to\; 1.$$

  - GCSR (1997) suggest that values below 1.2 are acceptable.

Other considerations

- Constructing efficient algorithms
  - Parameterisation
  - Efficient jumping rules
- Frequentist methods.

References

- Bayesian Data Analysis,
  - *Gelman, Carlin, Stern & Rubin*, 1997.
- Markov Chain Monte Carlo for Statistical Inference,
  - *Julian Besag*, CSSS Working Paper # 9.
  - Available on the web.