Markov chain Monte Carlo

Simon Jackman

Stanford University
http://jackman.stanford.edu/BASS

February 8, 2012

Markov chain Monte Carlo

- Two MCs
- Monte Carlo: we talked about yesterday
- Markov chains: Chapter 4, BASS.
- In general, we simply can't sample directly from the posterior density $p(\boldsymbol{\theta}|\text{data})$: e.g.,
  - $\boldsymbol{\theta}$ is a big object (many parameters).
  - $p(\boldsymbol{\theta}|\text{data})$ is a nasty function, difficult to sample from.
- Sampling from $p(\boldsymbol{\theta}|\text{data})$ in these cases usually require us to ***give up independence*** in the series of sampled values.
- That is, the resulting sequence of sampled values $\{\theta^{(t)}\}$ are "serially dependent".

Results from Markov chain theory

- Simulation consistency results hold even when we don't have independent samples from $p(\boldsymbol{\theta})$.
- Proof relies on results from Markov chain theory
- A Markov chain is a stochastic process: a useful, physical analogy is a particle moving randomly in some space.
- In the context of Bayesian statistics, we have $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$; i.e., the "particle" is $\boldsymbol{\theta}^{(t)}$ and the state space of the Markov chain is $\boldsymbol{\Theta}$.
- Markov chain on $\boldsymbol{\Theta}$: $\{\boldsymbol{\theta}^{(t)}\} = \{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \}$.
- **Ergodic theorem**: *how often* the Markov chain $\{\boldsymbol{\theta}^{(t)}\}$ visits site $\mathcal{A} \in \boldsymbol{\Theta}$ is a simulation-consistent estimate of $\Pr(\boldsymbol{\theta} \in \mathcal{A})$.

Markov chains

- Discrete state space: possible locations/states $\boldsymbol{\Theta}$ is a finite set, say with cardinality D. $\mathbf{p}^{(t)}$ is a D-by-1 vector, with $p_d^{(t)} = \Pr(\boldsymbol{\theta}^{(t)} = d), d \in \boldsymbol{\Theta}$.
- Continuous state space: we will consider the probability of a move from a point $\boldsymbol{\theta}^{(t)}$ to a point $\boldsymbol{\theta}^{(t+1)}$ in a *region* $\mathcal{A} \subseteq \boldsymbol{\Theta}$.
- the move from $\boldsymbol{\theta}^{(t)}$ to $\boldsymbol{\theta}^{(t+1)}$ is governed by the Markov chain's **transition kernel**.
- for a chain on a discrete space: $\mathbf{p}^{(t+1)} = \mathbf{p}^{(t)}\mathbf{K}$, where $\mathbf{K}$ is a *transition matrix*.
- for a chain on a continuous space we have a function, a *transition kernel*, $K(\boldsymbol{\theta}^{(t)}, \cdot)$, and $p(\boldsymbol{\theta})$ a density over $\boldsymbol{\Theta}$.

$$p^{(t+1)}(\boldsymbol{\theta}) = \int_{\boldsymbol{\Theta}} K(\boldsymbol{\theta}^{(t)}, \cdot)p^{(t)}(\boldsymbol{\theta})d\boldsymbol{\theta}^{(t)}$$

- Discrete case:

$$\mathbf{p} = \mathbf{p}\mathbf{K} \Rightarrow \mathbf{p}(\mathbf{I} - \mathbf{K}) = \mathbf{0} \Rightarrow (\mathbf{I} - \mathbf{K})'\mathbf{p}' = \mathbf{0}$$

i.e., an eigenvector of $\mathbf{K}$ gives us the stationary distribution (up to a normalizing factor).

- Continuous case:

$$p(\boldsymbol{\theta}^{(t+1)}) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}^{(t)}) K(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t+1)}) d\boldsymbol{\theta}^{(t)}$$

i.e., we have the same density over p over $\boldsymbol{\Theta}$ irrespective of the value of t.

Ergodic Theorem, Proposition 4.7 BASS

Theorem (Pointwise Ergodic Theorem; Law of Large Numbers for Markov chains)

Let $\{\boldsymbol{\theta}^{(t)}\}$ *be a Harris recurrent Markov chain on* $\boldsymbol{\Theta}$ *with a σ-finite invariant measure p. Consider a p-measurable function h s.t.* $\int_{\boldsymbol{\Theta}} |h(\boldsymbol{\theta})| dp(\boldsymbol{\theta}) < \infty$*. Then*

$$\lim_{T \to \infty} T^{-1} \sum_{t=1}^{T} h(\boldsymbol{\theta}^{(t)}) = \int_{\boldsymbol{\Theta}} h(\boldsymbol{\theta}) dp(\boldsymbol{\theta}) \equiv E_p h(\boldsymbol{\theta}).$$

Implications of the Ergodic Theorem for MCMC

- If we can construct a Markov chain the "right way", then:
- the Markov chain will have a unique, limiting distribution, a posterior density that we happen to be interested in, $p \equiv p(\boldsymbol{\theta}|\text{data})$
- no matter where we start the Markov chain, if we let it run long enough, it will eventually wind up generating a random tour of the parameter space, visiting sites in the parameter space $\mathcal{A} \in \boldsymbol{\Theta}$ with relative frequency proportional to $\int_{\mathcal{A}} p(\boldsymbol{\theta}|\text{data}) d\boldsymbol{\theta}$
- the Ergodic Theorem means that averages $\bar{h} = T^{-1} \sum_{t=1}^{T} h(\boldsymbol{\theta}^{(t)})$ taken over the Markov chain output are simulation-consistent estimates of

$$E[h(\boldsymbol{\theta})|\text{data}] = \int_{\boldsymbol{\Theta}} h(\boldsymbol{\theta}) p(\boldsymbol{\theta}|\text{data}) d\boldsymbol{\theta}.$$

- T might have to be big, even "massive"...

Conditions Needed for Ergodicity

- **Harris recurrence** (Definition 4.6, BASS).
  - **irreducibility** (Definition 4.10), BASS: the Markov chain can (eventually) get from regions $\mathcal{A}$ to $\mathcal{B}$, $\forall\, \mathcal{A}, \mathcal{B} \in \boldsymbol{\Theta}$.
  - **uniqueness of invariant distribution**: the Markov chain has a kernel K such that
$$p(\boldsymbol{\theta}^{(t+1)}) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}^{(t)}) K(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t+1)}) d\boldsymbol{\theta}^{(t)}$$
    i.e., iterating the chain doesn't change p.
- almost every Markov chain we encounter in MCMC has these properties

Simulation Inefficiency, §4.4.1

- We pay a price for not having independent draws from the posterior density $p(\boldsymbol{\theta}|\mathbf{y})$.
- estimand $h(\boldsymbol{\theta})$; we estimate $E(h(\boldsymbol{\theta})|\mathbf{y})$ --- the mean of the posterior density of $h(\boldsymbol{\theta})$ --- with the average $\bar{h}_T = T^{-1} \sum_{t=1}^{T} h(\boldsymbol{\theta}^{(t)})$
- Ergodic theorem says we have a simulation consistent estimator
- But the rate at which $\bar{h}_T$ converges on $E(h(\boldsymbol{\theta})|\mathbf{y})$ --- the rate at which the Monte Carlo error of $\bar{h}_T$ approaches zero --- is not as fast as the $\sqrt{T}$ rate we get from an independence sampler.
- Formalizations of this "simulation inefficiency"

Definition (Integrated Correlation Time)

Let $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(T)}$ be realizations from p, the stationary distribution of the Markov chain $\{\boldsymbol{\theta}^{(t)}\}$, and let $h(\boldsymbol{\theta})$ be some (scalar) quantity of interest. If ρ_j is the lag-j autocorrelation of the sequence $\{h(\boldsymbol{\theta}^{(t)})\}$ then

$$\tau_{\text{int}}[h(\boldsymbol{\theta})] = \frac{1}{2} + \sum_{j=1}^{\infty} \rho_j.$$

is the integrated autocorrelation time of the chain.

- n.b., for an independence sampler $\rho_j \approx 0 \, \forall \, j \Rightarrow \tau_{\text{int}}[h(\boldsymbol{\theta})] \approx 1/2$.

"Effective sample size" of an ergodic average

- estimand $h(\boldsymbol{\theta})$; Markov chain $\{h(\boldsymbol{\theta}^{(t)})\}$, stationary distribution p.
- estimated with ergodic average $\bar{h}_T = T^{-1} \sum_{t=1}^{T} h(\boldsymbol{\theta}^{(t)})$.
- $\text{var}_p(h_t) = \sigma^2$.
- But $\text{var}(\bar{h}_T) = \frac{\sigma^2}{T} \times 2 \times \tau_{\text{int}}[h(\boldsymbol{\theta})]$.
- The factor $2 \times \tau_{\text{int}}[h(\boldsymbol{\theta})]$) is a measure of how the dependency inherent in the Markovian exploration of $p(\boldsymbol{\theta}|\mathbf{y})$ is degrading the precision of the summary statistic $\bar{h}$.
- Large and slowly decaying autocorrelations make $\tau_{\text{int}}[h(\boldsymbol{\theta})]$ large.
- for an independence sampler

$$2 \times \tau_{\text{int}}[h(\boldsymbol{\theta})] \approx 2 \times \frac{1}{2} = 1 \Rightarrow \text{var}(\bar{h}_T) \approx \sigma^2 / T$$

"Effective sample size" of an ergodic average

- See function `effectiveSize` in `R` package `coda`
- Suppose we have a 1st order Markov chain:

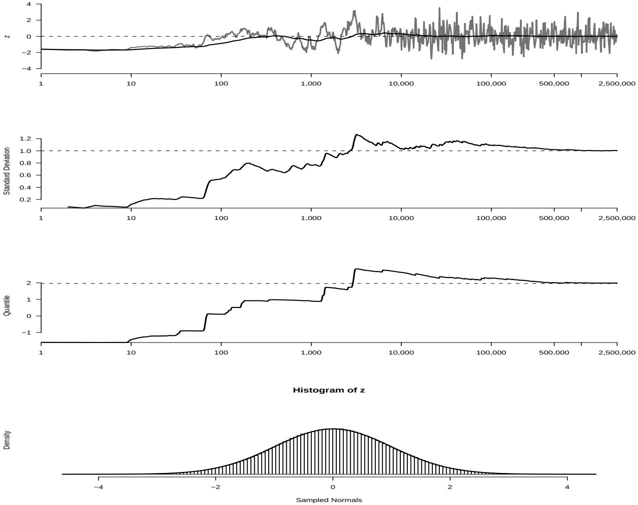
$$E(h_t | h_{t-1}) = \rho h_{t-1}, |\rho| < 1, \text{var}(h_t) = \sigma^2$$

then $\text{var}(\bar{h}_T) = \dfrac{\sigma^2}{T} \dfrac{1+\rho}{1-\rho}$.

- $\rho \to 0$, we tend to the independence sampler
- $\rho \to 1$, the dependency increases, $(1+\rho)/(1-\rho) \to \infty$.
- e.g., $\rho = .9$ and we seek a given level of Monte Carlo error in ergodic average $\bar{h}_T$.
- $(1+.9)/(1-.9) = 1.9/.1 = 19$ or we require $\sqrt{19} \approx 4.36$ as many iterations of the Markov chain to get the same level of Monte Carlo error as we would if we were using an independence sampler.

Example 4.12, highly dependent, stationary series

$z_t = \rho z_{t\text{-}1} + \varepsilon_t, \varepsilon_t \sim N(0, \omega^2), \omega^2 = 1 - \rho^2, \rho = .995.$



Sampling algorithms used in MCMC

- Metropolis-Hastings algorithm; §5.1
- Gibbs sampler; §5.2

Metropolis-Hastings algorithm

sample $\boldsymbol{\theta}^*$ from a "proposal" or "jumping" distribution $J_t(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\theta}^*)$.

$$r \leftarrow \frac{p(\boldsymbol{\theta}^*|\mathbf{y})J_t(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(t-1)})}{p(\boldsymbol{\theta}^{(t-1)}|\mathbf{y})J_t(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\theta}^*)}, \tag{1}$$

$\alpha \leftarrow \min(r, 1)$
sample $U \sim \text{Unif}(0, 1)$
if $U \leq \alpha$ **then**
    $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}^*$
else
    $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}^{(t-1)}$
end if

Theory for the Metropolis sampler §5.1.1

- Transition kernel $K(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t+1)})$ generates a *reversible* Markov chain.
- Reversibility implies $p(\boldsymbol{\theta}|\mathbf{y})$ is the stationary distribution of the Markov chain.
- Ergodicity follows if we can establish irreducibility and aperiodicity. Sufficiently permissive J_t accomplishes this.
- e.g., $J_t(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t+1)}) > 0 \, \forall \, \boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t+1)}$.
- Aperiodicity follows if $\Pr(\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)}) > 0$.

Metropolis-Hastings algorithm

- proposal density J_t is key to the algorithm
- t subscript for proposal density indicates that the proposal density can evolve, "tuning" the algorithm for an efficient exploration of $p(\boldsymbol{\theta}|\mathbf{y})$
- original paper is Metropolis et al. (1953) with $r_M = p(\boldsymbol{\theta}^*|\mathbf{y})/p(\boldsymbol{\theta}^{(t\text{-}1)}|\mathbf{y})$.
- modification by Hastings (1970) to give the acceptance ratio given on previous slide
- **Random walk M-H:** select a candidate point $\boldsymbol{\theta}^*$ by taking a random perturbation around the current point $\boldsymbol{\theta}^{(t)}$, i.e., $\boldsymbol{\theta}^* = \boldsymbol{\theta}^{(t)} + \boldsymbol{\varepsilon}$. e.g.,
  - $\varepsilon_j \sim \text{Unif}(\text{-}\delta_j, \delta_j), j = 1, \ldots, J$ dimensions of $\boldsymbol{\theta}$.
  - $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Omega})$. Here the key parameter is $\boldsymbol{\Omega}$.
- **Independence M-H:** e.g., $J = N(\hat{\boldsymbol{\theta}}, c \cdot V(\hat{\boldsymbol{\theta}}))$, c a tuning parameter.

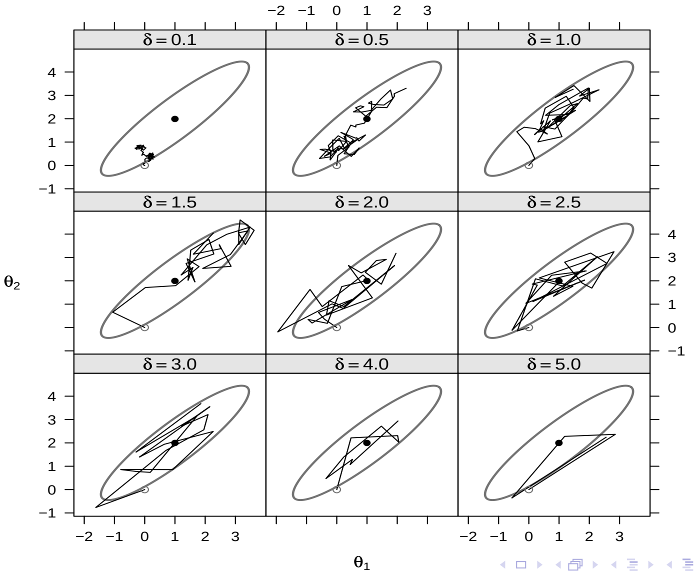
Random Walk Metropolis, Example 5.1

- $\boldsymbol{\theta} \sim N$
- random-walk Metropolis, but what distribution for $\boldsymbol{\varepsilon}$?
- Example 5.1: $\boldsymbol{\theta} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

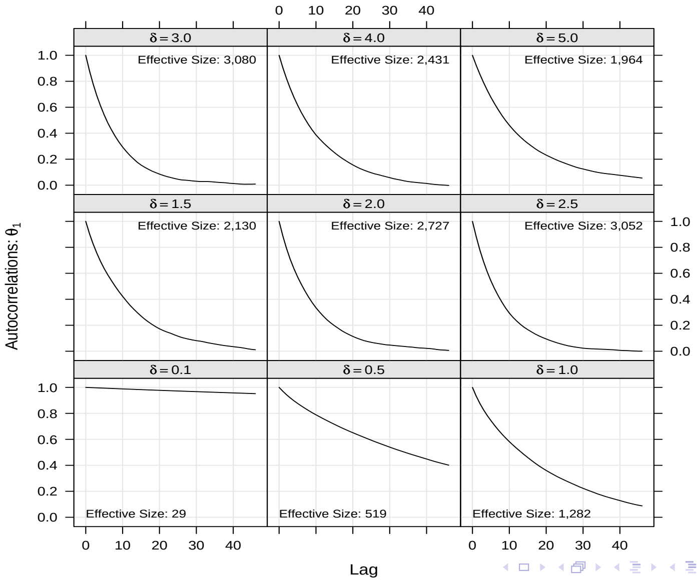
$$\boldsymbol{\mu} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix},$$

  and where $\varepsilon_j \sim \text{Unif}(-\delta, \delta), j = 1, 2.$
- Consider different choices of δ.

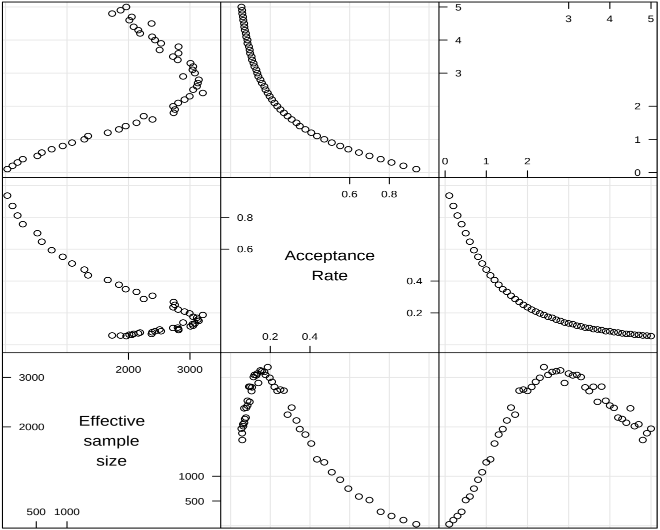
Random Walk Metropolis, Example 5.1



Random Walk Metropolis, Example 5.1



Random Walk Metropolis, Example 5.1



Scatter Plot Matrix

Random Walk Metropolis, Ex 5.2, Poisson regression

- $y_i|\mathbf{x}_i\boldsymbol{\beta} \sim \text{Poisson}(\lambda_i), \lambda_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$
- $y_i \in \{0, 1, 2, \ldots\}$, $\mathbf{x}_i$ is a vector of covariates, $\boldsymbol{\beta}$ is a vector of k unknown coefficients and $i = 1, \ldots, n$ indexes observations.
- likelihood: $f(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\lambda_i^{y_i}}{y_i!} \exp(-\lambda_i)$
- Data: 915 biochemistry graduate students, article counts over last 3 years of PhD studies. Gender differences key.
- Modal number of article counts is zero (30%); 95%-ile is 5, max is 19.
- No conjugate prior for $\boldsymbol{\beta}$; usually just express the posterior in the form it comes from Bayes Rule

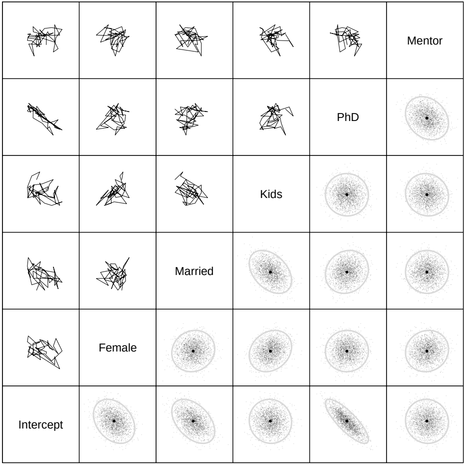
$$p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) \propto p(\boldsymbol{\beta}) f(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta})$$

Random Walk Metropolis, Ex 5.2, Poisson regression

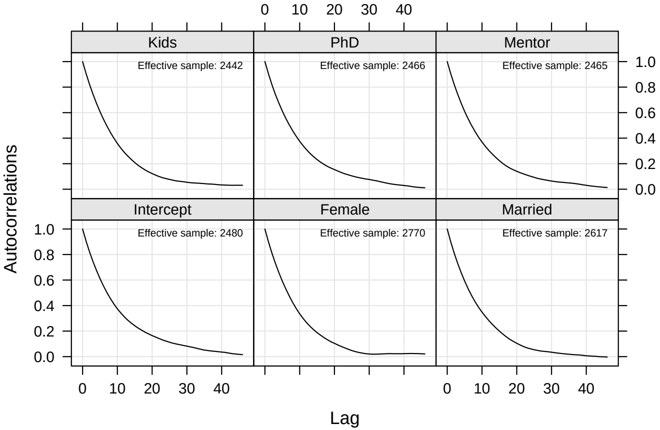
- implementation in `R` package `MCMCpack` with the function `MCMCpoisson`
- multivariate normal prior for $\boldsymbol{\beta}$, $\boldsymbol{\beta} \sim N(\mathbf{b}_0, \mathbf{B}_0^{-1})$.
- Metropolis proposal density is $\boldsymbol{\beta}^* \sim N(\boldsymbol{\beta}^{(t)}, \mathbf{P})$ where $\mathbf{P} = \mathbf{T}(\mathbf{B}_0 + \mathbf{V}^{-1})^{-1}\mathbf{T}$, with $\mathbf{T}$ a k-by-k diagonal, positive definite matrix containing tuning parameters and $\mathbf{V}$ is the large-sample approximation to the frequentist sampling covariance matrix of the maximum likelihood estimates $\hat{\boldsymbol{\beta}}$.
- default is $\mathbf{T} = 1.1 \cdot \mathbf{I}$ and to initialize the random-walk Metropolis algorithm at the MLEs $\hat{\boldsymbol{\beta}}$.
- we use vague priors, with $\mathbf{b}_0 = \mathbf{0}$ and $\mathbf{B}_0 = 10^4 \cdot \mathbf{I}$.

Random Walk Metropolis, Ex 5.2, Poisson regression

50,000 iterations of Metropolis algorithm: upper panels show a trace plot of the algorithm in two dimensions, for the first 250 iterations of the algorithm; lower panels summarize the full 50,000 iterations, plotting the algorithm's history at each of 2,500 evenly-spaced iterations over the full 50,000 iterations.



Random Walk Metropolis, Ex 5.2, Poisson regression



Random Walk Metropolis, Ex 5.2, Poisson regression

	Bayes	MLE
Intercept	0.30	0.30
	[0.088, 0.50]	[0.10, 0.51]
Female	-0.22	-0.22
	[-0.33, -0.12]	[-0.33, -0.12]
Married	0.16	0.16
	[0.044, 0.28]	[0.035, 0.28]
Kids$<$ 5	-0.18	-0.19
	[-0.27, -0.11]	[-0.26, -0.11]
PhD Prestige	0.013	0.014
	[-0.040, 0.065]	[-0.039, 0.065]
Mentor Articles	0.026	0.026
	[0.022, 0.030]	[0.022, 0.029]

Gibbs sampler

Suppose $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)'$.

for $t = 1$ to T **do**
    sample $\boldsymbol{\theta}_1^{(t+1)}$ from $g_1(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2^{(t)}, \boldsymbol{\theta}_3^{(t)}, \ldots, \boldsymbol{\theta}_d^{(t)}, \mathbf{y})$.
    sample $\boldsymbol{\theta}_2^{(t+1)}$ from $g_2(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_3^{(t)}, \ldots, \boldsymbol{\theta}_d^{(t)}, \mathbf{y})$.
    ...
    sample $\boldsymbol{\theta}_d^{(t+1)}$ from $g_d(\boldsymbol{\theta}_d \mid \boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \ldots, \boldsymbol{\theta}_{d\text{-}1}^{(t+1)}, \mathbf{y})$.
    $\boldsymbol{\theta}^{(t+1)} \leftarrow (\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \ldots, \boldsymbol{\theta}_d^{(t+1)})'$.
end for

Gibbs sampler

- "divide and conquer"
- sample from lower dimensional conditional densities, given other elements of $\boldsymbol{\theta}$.
- it works! See theoretical discussion at §5.2.1.
- joint probability densities completely characterized by component conditional densities

Example 5.3, Gibbs sampler for bivariate normal

- $\mathbf{y}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), i = 1, \ldots, n$
- $\boldsymbol{\mu} = (\mu_1, \mu_2)'$ and $\boldsymbol{\Sigma}$ is a known 2-by-2 covariance matrix
- unknown parameters here are $\boldsymbol{\theta} = \boldsymbol{\mu} = (\mu_1, \mu_2)'$.
- Our goal is to compute the posterior density $p(\boldsymbol{\theta}|\mathbf{y})$.
- Independent, conjugate prior densities for each element of $\boldsymbol{\mu}$, say, $\boldsymbol{\mu} = (\mu_1, \mu_2)' \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)'$ with

$$\boldsymbol{\Sigma}_0 = \begin{bmatrix} \sigma^2_{01} & 0 \\ 0 & \sigma^2_{02} \end{bmatrix},$$

- Posterior density for $\boldsymbol{\theta}$ is known in this case; it is bivariate normal.
- But we explore with Gibbs sampler (quite unnecessary for this problem, but helpful for exposition).

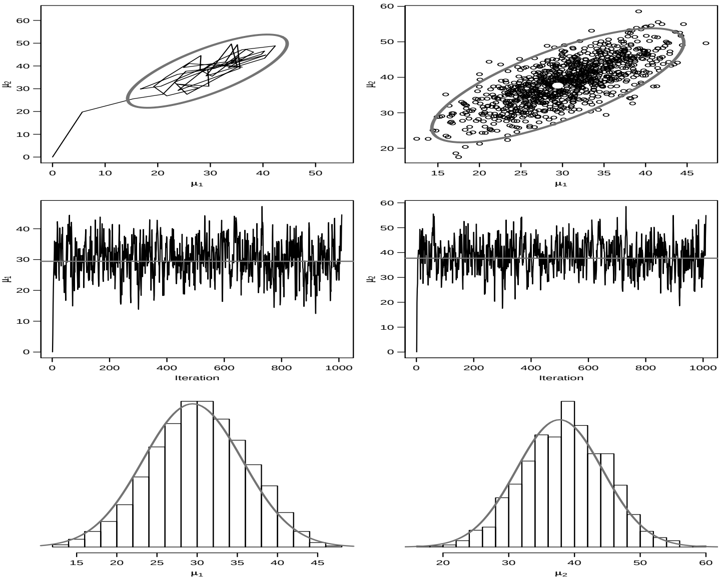
Example 5.3, Gibbs sampler for bivariate normal

At iteration t,

1. sample $\mu_1^{*(t)}$ from its conditional distribution $g_1(\mu_1^*|\mu_2^{*(t-1)}, \mathbf{\Sigma}^*, \mathbf{y})$, a normal density with mean $\mu_1^* + \frac{\sigma_{12}^*}{\sigma_{22}^*}(\mu_2^{*(t-1)} - \mu_2^*)$ and variance $\sigma_{11}^* - \sigma_{12}^{*2}/\sigma_{22}^*$
2. sample $\mu_2^{*(t)}$ from its conditional distribution $g_2(\mu_2^*|\mu_1^{*(t)}, \mathbf{\Sigma}^*, \mathbf{y})$, a normal density with mean $\mu_2^* + \frac{\sigma_{12}^*}{\sigma_{11}^*}(\mu_1^{*(t)} - \mu_1^*)$ and variance $\sigma_{22}^* - \sigma_{12}^{*2}/\sigma_{11}^*$.

Note that we condition on $\mu_2^{*(t-1)}$ when sampling $\mu_1^{*(t)}$; then, given the sampled value $\mu_1^{*(t)}$, we condition on it when sampling $\mu_2^{*(t)}$.

Example 5.3, Gibbs sampler for bivariate normal



Example 5.3, Gibbs sampler for bivariate normal

	Analytic	1 000 iterations	50 000 iterations
$E(\mu_1 \mid \mathbf{y})$	29.44	30.34	29.38
$E(\mu_2 \mid \mathbf{y})$	37.72	38.63	37.65
$V(\mu_1 \mid \mathbf{y})$	38.41	35.54	38.91
$V(\mu_2 \mid \mathbf{y})$	43.17	40.12	43.84
$C(\mu_1, \mu_2 \mid \mathbf{y})$	30.59	28.07	31.12

Conditional distributions for the Gibbs sampler

Theorem

If a statistical model can be expressed as a directed acyclic graph (a DAG) $\mathcal{G}$, then the conditional density of node θ_j in the graph is

$$\begin{aligned} f(\theta_j|\mathcal{G} \setminus \theta_j) &\propto f(\theta_j|parents[\theta_j]) \times \\ &\prod_{w \in children[\theta_j]} f(w|parents[w]), \end{aligned} \tag{2}$$

where $\mathcal{G} \setminus \theta_j$ stands for all nodes in $\mathcal{G}$ other than θ_j.

Proof.

See Spiegelhalter and Lauritzen (1990). □

Example 5.6, 2-level hierarchical model

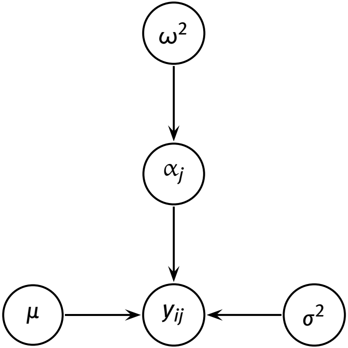
- We have multiple observations $i = 1, \ldots, n$ for each of $j = 1, \ldots, J$ units (e.g., students indexed by i in schools indexed by j) on a real-valued variable y_{ij}.
- Model:

$$\begin{aligned} y_{ij}|\mu, \alpha_j, \sigma^2 &\sim N(\mu + \alpha_j, \sigma^2) \\ \alpha_j &\sim N(0, \omega^2) \end{aligned}$$

- Likelihood: $f(\mathbf{Y}|\mu, \alpha, \sigma^2) = \prod_{i=1}^{n} \prod_{j=1}^{J} \phi\left(\frac{y_{ij}\text{-}\mu\text{-}\alpha_j}{\sigma}\right)$
- The hyper-parameter ω^2 is referred to as the *between* unit variance, while σ^2 is the *within* unit variance.
- Priors on μ, ω^2 and σ^2: *a priori* independence for these parameters, $p(\mu, \omega^2, \sigma^2) = p(\mu)p(\omega^2)p(\sigma^2)$.
- $\boldsymbol{\theta} = (\mu, \alpha, \sigma^2, \omega^2)'$

Example 5.6; Figure 5.11

- y_{ij} is a child of μ, α_j, σ^2; μ etc are the parents of y_{ij} etc.
- α_j is a child of ω^2.
- y_{ij} is conditionally independent of ω^2 given α_j; i.e., $\{y_{ij} \perp\!\!\!\perp \omega^2\}|\alpha_j, \forall\, i, j$.
- $\{\alpha_j \perp\!\!\!\perp \alpha_k\}|\omega^2, \forall\, j \neq k$.



1. sample $\sigma^{2(t)}$ from $g(\sigma^2|\mathcal{G}_{-\sigma^2}) = g(\sigma^2|\mathbf{Y}, \mu, \alpha) \propto f(\mathbf{Y}|\mu, \alpha, \sigma^2)p(\sigma^2)$
2. sample $\omega^{2(t)}$ from $g(\omega^2|\mathcal{G}_{-\omega^2}) = g(\omega^2|\alpha)$, noting that $\{\omega^2 \perp\!\!\!\perp (\mathbf{Y}, \mu, \sigma^2)\}|\alpha$. Thus, $g(\omega^2|\alpha) \propto f(\alpha|\omega^2)p(\omega^2)$.
3. for $j = 1, \ldots, J$, sample $\alpha_j^{(t)}$ from $g(\alpha_j|\mathcal{G}_{-\alpha_j}) = g(\alpha_j|\mathbf{y}_j, \sigma^2, \omega^2, \mu)$, where $\mathbf{y}_j$ is a vector of the observations from unit j, and noting that $\{\alpha_j \perp\!\!\!\perp \alpha_k\}|\omega^2 \; \forall \, j \neq k$; i.e.,

$$\begin{aligned} g(\alpha_j|\mathbf{y}_j, \sigma^2, \omega^2, \mu) &\propto f(\mathbf{y}_j|\mu, \alpha_j, \sigma^2)p(\alpha_j|\omega^2) \\ &= \prod_{i=1}^{n} \phi\left(\frac{y_{ij} - \mu - \alpha_j}{\sigma}\right) \cdot \phi\left(\frac{\alpha_j}{\omega}\right) \end{aligned}$$

4. sample $\mu^{(t)}$ from $g(\mu|\mathcal{G}_{-\mu}) = g(\mu|\mathbf{Y}, \alpha, \sigma^2)$, since $\{\mu \perp\!\!\!\perp \omega^2\}|\alpha$; i.e.,

$$g(\mu|\mathbf{Y}, \alpha, \sigma^2) \propto f(\mathbf{Y}|\mu, \alpha, \sigma^2)p(\mu) = \prod_{i=1}^{n} \prod_{j=1}^{J} \phi\left(\frac{y_{ij} - \mu - \alpha_j}{\sigma}\right) \cdot p(\mu)$$

Implementation in JAGS

JAGS code

```
model{
        ## loop over data frame
        for(i in 1:N){
              ## expression for E(y[i])
              ## note double-subscript on alpha
              ymu[i] <- mu + alpha[j[i]]

              ## sampling model for y[i]
              y[i] ~ dnorm(ymu[i],tau.sigma)
        }

        ## hierarchical model for alphas
        for(i in 1:J){
              alpha[i] ~ dnorm(0,tau.omega)
        }

        ## predictions for a future election?
        for(i in 1:J){
              muFuture[i] <- mu + alpha[i]
              yFuture[i] ~ dnorm(muFuture[i],tau.sigma)
        }

        ## prior for mu
        mu ~ dnorm(0,.01)

        ## prior for standard deviations (not variances!)
        sigma ~ dunif(0,10)
        omega ~ dunif(0,10)
        tau.sigma <- pow(sigma,-2) ## precision!
        tau.omega <- pow(omega,-2) ## precision!
}
```

References

Hastings, W. K. 1970. "Monte Carlo sampling methods using Markov chains, and their applications." *Biometrika* 57:97--109.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller. 1953. "Equations of state calculations by fast computing machines." *Journal of Chemical Physics* 21:1087--91.

Spiegelhalter, David J. and S. L. Lauritzen. 1990. "Sequential updating of conditional probabilities on directed graphical structures." *Networks* 20:579--605.