BML. 3. Intro to MCMC
From Gibbs to Hamilton

DataLab CSIC

Brief description

1. presented key methods in Bayesian inference and basic models and faced 'severe' computational problems quite rapidly (beta-binomial, logistic regression)

2. overviewed Bayesian computational strategies

3. Will focus on MCMC (Markov chain Monte Carlo): Gibbs, Metropolis, Hybrid, Hamiltonian.

Other stuff, including SG-MCMC later on

Labs

3.1 Gibbs
Basic, Re-usable,Paralel chains
Multivariate normal
Normal-gamma inverse
3.2 Metropolis
Basic, Standard, MH
Multivariate normal
Linear regression (ecology)
Poisson regression (ecology)
3.3 Hamiltonian MC
Multivariate normal
Hmclearn: Linear regression, Logistic regression (ecology)

Computational problems in Bayesian analysis

Computing the posterior

Computing the predictive

Finding the optimal alternative

Ignore 2 last ones until Ch 4

$$p(\theta|x) = \frac{p(x|\theta)\, p(\theta)}{p(x)} = \frac{p(x|\theta)\, p(\theta)}{\int p(x|\theta)\, p(\theta)\, d\theta} \propto p(x|\theta)\, p(\theta)$$

$$p(y|x) = \int p(y|\theta)\, p(\theta|x)\, d\theta$$

$$\max_a \int u(a,\theta)\, p(\theta|x)\, d\theta$$

$$\max_a \int u(a,y)\, p(y|x)\, dy$$

a. Gibbs sampling

Sources

French and DRI (2000) Ch 7

DRI et al (2010) Ch4

BDA3 (2015) Ch11, 12

Hoff (2009) Ch 6,7

General idea MCMC

Markov chain X_n with same state space as target and convergent to target distribution g

$$X_n \xrightarrow{d} g$$

Strategy

Initialise x_0, $n=1$

Until convergence, Generate $X_n \mid X_{n-1} = x_{n-1}$, $n=n+1$, n^*

Until n^*+m, Generate and collect $X_n \mid X_{n-1} = x_{n-1}$, $n=n+1$

$$\hat{I}_g \approx \frac{1}{m} \sum_{n^*}^{n^*+m} f(x_i)$$

Generic Gibbs sampler

$$X = (X_1, \dots, X_p) \sim \pi$$

Sample from $X_s \mid X_{-s} = (X_1, \dots, X_{s-1}, X_{s+1}, \dots, X_p)$

Initialize $X_1^0, \dots, X_p^0$, $i = 1$

Iterate

Sample $X_1^i \sim X_1 \mid X_2^{i-1}, \dots, X_p^{i-1}$

Sample $X_2^i \sim X_2 \mid X_1^i, X_3^{i-1}, \dots, X_p^{i-1}$

...

Sample $X_p^i \sim X_p \mid X_1^i, X_2^i, \dots, X_p^{i-1}$

$i = i+1$

Important Gibbs sampling examples

Purpose

Through several important examples reinforce Gibb sampling concept

Motivation

Model and prior

Posterior conditionals

Gibbs sampler

Use

(Implemented in lab or as an exercise)

Sampling the bi-variate normal

Simple example to recall approach

Model. Bivariate normal with unknown means. Variances 1. Known correlation ϱ

1 observation

Prior. Uniform

Use. Expected value and variance of parameters, Expected cross product, Probability that parameter belongs to a set

Auxiliary result

$$\theta \mid \mu, \Sigma \sim N(\mu, \Sigma) \qquad f(\theta) = \frac{1}{(2\pi)^{p/2} \, |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (\theta-\mu)' \Sigma^{-1} (\theta-\mu)\right)$$

$$C = I - \left[\mathrm{diag}\left(\Sigma^{-1}\right)\right]^{-1} \Sigma^{-1}$$

$$\theta_i \mid \theta_j, j \neq i \sim N\left(\mu_i + \sum_{j \neq i} c_{ij} (\theta_j - \mu_j), \left((\Sigma^{-1})_{ii}\right)^{-1}\right)$$

Sampling the bi-variate normal. Model and posterior

$$y = (y_1, y_2) \sim N\left(\begin{pmatrix}\theta_1 \\ \theta_2\end{pmatrix}, \Sigma\begin{pmatrix}1 & \rho \\ \rho & 1\end{pmatrix}\right) \qquad \pi(\theta) \propto K$$

$$\pi(\theta_1, \theta_2 | y) \propto \pi(\theta)\,\pi(y|\theta) \propto \pi(y|\theta)$$

$$\propto \exp\left(-\frac{1}{2}(y-\theta)'\,\Sigma^{-1}(y-\theta)\right) = \exp\left(-\frac{1}{2}(\theta-y)'\,\Sigma^{-1}(\theta-y)\right)$$

$$\theta | y \sim N(y, \Sigma)$$

Sampling the bi-variate normal. Gibbs sampler

$$\theta_1 \mid \theta_2, y \sim N(y_1 + \rho(\theta_2 - y_2), 1 - \rho^2)$$

$$\theta_2 \mid \theta_1, y \sim N(y_2 + \rho(\theta_1 - y_1), 1 - \rho^2)$$

θ_2^0 ARBITRARY, $i = 1$

UNTIL CONVERGENCE

$\theta_1^i \sim N(y_1 + \rho(\theta_2^{i-1} - y_2), 1 - \rho^2)$

$\theta_2^i \sim N(y_2 + \rho(\theta_1^i - y_1), 1 - \rho^2)$

$i = i + 1$

Sampling the bi-variate normal. Answers

$$\hat{E}(\theta_1|y) = \frac{1}{K}\sum_{i=1}^{K} \theta_1^{M+i}$$

$$\hat{E}(\theta_1 \theta_2|y) = \frac{1}{K}\sum_{i=1}^{K} \left(\theta_1^{M+i}\ \theta_2^{M+i}\right)$$

$$Pr(\theta_1 \in A|y) = \frac{\#\{\theta_1^{M+i} \in A\}}{K}$$

Inference for normal-gamma

A very important example. Solution known analytically. Serves as benchmark

Model. Normal with unknown mean and variance. Prior normal-gamma.

Use. Quartiles of mean, Interval of probability 0.7, Test null that mean is bigger than 0

https://en.wikipedia.org/wiki/Inverse-gamma_distribution

https://en.wikipedia.org/wiki/Normal-inverse-gamma_distribution

Inference for normal-gamma. Model

$$x_1, \dots, x_n \mid \theta, \sigma^2 \sim N(\theta, \sigma^2)$$

$$\theta \sim N(\mu_0, \tau_0^2)$$

$$\frac{1}{\sigma^2} = \tilde{\sigma}^2 \sim Ga\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right)$$

$$p(\theta, \sigma^2) = p(\theta)\, p(\sigma^2)$$

$$\mu_n(\tilde{\sigma}^2) = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{x}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$$

$$p(\theta \mid \sigma^2, x_1, \dots, x_n) \sim N(\mu_n, \tau_n^2)$$

$$\tau_n(\tilde{\sigma}^2) = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$$

Inference for normal-gamma. Conditional

$$p(\tilde{\sigma}^2 \mid \theta, x_1, \dots, x_n) \propto p(x_1, \dots, x_n, \theta, \tilde{\sigma}^2) = p(x_1, \dots, x_n \mid \theta, \tilde{\sigma}^2)\, p(\theta \mid \tilde{\sigma}^2)\, p(\tilde{\sigma}^2)$$

$$= p(x_1, \dots, x_n \mid \theta, \tilde{\sigma}^2)\, p(\theta)\, p(\tilde{\sigma}^2) \propto p(x_1, \dots, x_n \mid \theta, \tilde{\sigma}^2)\, p(\tilde{\sigma}^2)$$

$$\propto (\tilde{\sigma}^2)^{n/2} \exp\left(-\tilde{\sigma}^2 \sum_{i=1}^{n} (x_i - \theta)^2/2\right) (\tilde{\sigma}^2)^{\nu_0/2 - 1} \exp\left(-\tilde{\sigma}^2 \nu_0 \sigma_0^2/2\right)$$

$$= (\tilde{\sigma}^2)^{(\nu_0 + n)/2 - 1} \exp\left\{-\tilde{\sigma}^2 \left[\nu_0 \sigma_0^2 + \sum (x_i - \theta)^2\right]/2\right\}$$

$$\nu_n = \nu_0 + n \qquad s_n^2(\theta) = \frac{\sum (x_i - \theta)^2}{2}$$

$$\tilde{\sigma}^2 \mid x_1, \dots, x_n \sim Ga\left(\frac{\nu_n}{2}, \frac{\nu_n \sigma_n^2(\theta)}{2}\right) \qquad \sigma_n^2(\theta) = \frac{1}{\nu_n}\left[\nu_0 \sigma_0^2 + n s_n^2(\theta)\right]$$

Inference for normal-gamma. Sampler

$\tilde{\sigma}^{2\,0}$ ARBITRARY, $i=1$

UNTIL CONVERGENCE

$$\theta_i \sim N\left(\mu_n(\tilde{\sigma}^{2(i-1)}), \tau_n(\tilde{\sigma}^{2(i-1)})\right)$$

$$\tilde{\sigma}^{i} \sim Ga\left(\nu_n/2, \frac{\nu_n \sigma_n^2(\theta_i)}{2}\right)$$

$$i = i+1$$

Inference for normal-gamma. Answers

$$\theta^s_{\cdot} = \mathrm{SORT}(\theta_i)_{M}^{K+M}$$

$$\mathrm{MEDIAN} = \theta^s_{K/2} \quad \ldots$$

$$\mathrm{LB} = \theta^s_{\lfloor 0.15\, K \rfloor}$$

$$\mathrm{UB} = \theta^s_{\lfloor 0.85\, K \rfloor}$$

$$\mathrm{PROB} = \#\{\theta_i \geq 0\} / K$$

IF ($\mathrm{PROB} \geq 0.5$) ACCEPT

ELSE

REJECT

ENDIF

Inference for a mixture of exponentials

An example with mixtures. Very important in modern statistics and machine learning, e.g. Latent Dirichlet allocation. Used in clustering, density estimation,… Also introduces latent variables.

Model. Mixture of k exponentials, gamma priors. Dirichlet-multinomial on labels.

Use. Predictive distribution, group for first observation

https://en.wikipedia.org/wiki/Multinomial_distribution

https://en.wikipedia.org/wiki/Dirichlet_distribution

https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

Inference for a mixture of exponentials

Model

$$t = (t_1, \dots, t_{n_s})$$

$$f(t|\theta) = q_1 \mu_1 \exp(-\mu_1 t) + \dots + q_k \mu_k \exp(-\mu_k t)$$

$$\theta = (q_1, \mu_1; \dots; q_k, \mu_k) \quad \sum_{i=1}^{k} q_i = 1 \quad q_i > 0$$

$$\mu_i \sim Ga(a_i; p_i) \quad q \sim D(\alpha_1, \dots, \alpha_k) \quad \alpha_i > 0$$

Inference for a mixture of exponentials

Latent variables

$$t_j \longrightarrow z_j = (z_{1j}, \dots, z_{Kj})$$

$$z_j \mid q, \mu \sim \mathcal{M}_K(1; q_1, \dots, q_K)$$

$$\sum z_{ij} = 1 \qquad z_{ij} \in \{0, 1\}$$

$$t_j \mid z_j, q, \mu \sim \mathcal{E}\left(\prod_{i=1}^{K} \mu_i^{z_{ij}}\right)$$

Inference for a mixture of exponentials

Posterior conditionals

$$z_j \mid t_j, q, \mu \sim \mathcal{M}_K\left(1; \frac{q_1 \mu_1 \exp(-\mu_1 t_j)}{\sum q_i \mu_i \exp(-\mu_i t_j)}; \dots; \frac{q_K \mu_K \exp(-\mu_K t_j)}{\sum q_i \mu_i \exp(-\mu_i t_j)}\right) j = \dots n$$

$$\mu_j \mid t, z \sim \mathcal{G}\left(\alpha_j + \sum_{i=1}^{n} z_{ji} t_i, \ \beta_j + \sum_{i=1}^{n} z_{ji}\right), \ j = 1, \dots, K$$

$$q \mid t, z \sim \mathcal{D}\left(\alpha_1 + \sum_{i=1}^{n} z_{1i}, \dots, \alpha_K + \sum_{i=1}^{n} z_{Ki}\right)$$

Inference for a mixture of exponentials

Gibbs sampler

μ^0, z^0 ARBITRARY

UNTIL CONVERGENCE

$$z_j^{i+1} \sim z_j \mid t_j, q^i, \mu^i, \quad j = 1, \dots, n_s$$

$$q^{i+1} \sim q \mid t, z^{i+1}$$

$$\mu_j^{i+1} \sim \mu_j \mid t, z^{i+1}, \quad j = 1, \dots, k$$

$$i = i+1$$

Inference for a mixture of exponentials

Answers

$$P(t_1 \in h) = \frac{\#\{ z_1^i = h \}}{K}$$

FOR $i = 1$ TO K

$$\tau_i \sim \sum_{j=1}^{K} q_j^{M+i} \, \mathcal{E}xp\left(\mu_j^{M+i}\right)$$

Inference for normal multivariate

Another very important model. Basis of much multivariate analysis

Model. Multivariate normal (means and covariance matrix). Normal-Wishart prior

Use. Expected value of maximum, Probability that product of means is bigger than 4

https://en.wikipedia.org/wiki/Wishart_distribution

Inference for normal multivariate

Model

$$Y_1, \dots, Y_n \sim MN(\theta, \Sigma)$$

$$\theta \sim N(\mu_0, \Lambda_0)$$

$$\Sigma^{-1} \sim WISH(\nu_0, S_0^{-1})$$

$$E(\Sigma^{-1}) = \nu_0 S_0^{-1}$$

$$E(\Sigma) = \frac{1}{\nu_0 - p - 1} S_0$$

Inference for normal multivariate

Posterior conditionals

$$\Lambda_n = \left(\Lambda_0^{-1} + n\Sigma^{-1}\right)^{-1}$$

$$\theta \mid y_1, \ldots, y_n, \Sigma \sim N(\mu_n, \Lambda_n)$$

$$\mu_n = \Lambda_n^{-1}\left(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y}\right)$$

$$S_\theta = \sum_{i=1}^{n} (y_i - \theta)(y_i - \theta)^t$$

$$\Sigma \mid y_1, \ldots, y_n, \theta \sim \mathrm{WISH}\left(\nu_0 + n, (S_0 + S_\theta)^{-1}\right)$$

Inference for normal multivariate

Gibbs sampler

Σ^1 "ARBITRARY", $i = 1$

UNTIL CONVERGENCE

COMPUTE Λ_n^i , μ_n^i

$\theta^i \sim N(\mu_n^i, \Lambda_n^i)$

COMPUTE S_θ^i

$\Sigma^i \sim \text{WISH}\left(\nu_0 + n, (S_0 + S_\theta^i)^{-1}\right)$

$i = i+1$

Inference for normal multivariate

Answers

$$\text{MEAN}\left(\text{MAX}\left(\theta^{i}\right)_{i=M}^{M+K}\right)$$

$$\frac{\#\{\theta_1^{i} \cdots \theta_p^{i} \geq 4\}}{M}$$

Gibbs sampling theory

Recall

1. Choose initial values $(\theta_2^0, \ldots, \theta_k^0)$. $i = 1$

2. Until convergence is detected, iterate through

.　　Generate $\theta_1^i \sim \theta_1 | \theta_2^{i-1}, .., \theta_k^{i-1}$

.　　Generate $\theta_2^i \sim \theta_2 | \theta_1^i, \theta_3^{i-1}, ..., \theta_k^{i-1}$

.　　...

.　　Generate $\theta_k^i \sim \theta_k | \theta_1^i, ..., \theta_{k-1}^i$.

.　　$i = i + 1$

Convergence

Kernel

$$p_G(\theta^n, \theta^{n+1}) = \prod_{i=1}^{k} p(\theta_i^{n+1} \mid \theta_j^n, j > i; \theta_j^{n+1}, j < i).$$

Proposition 43 *Suppose that* $D = \{\theta : p(\theta) > 0\}$ *is a product set,* $D = \prod_{i=1}^{k} D_i$. *Then:*

1. $p_{\theta_i}(\theta_i | \theta_j, j \neq i)$ *and* p_G *are well-defined for* $\theta \in D$.

2. p_G *is* p*-irreducible and aperiodic.*

3. p *is invariant with respect to* p_G.

4. $\theta^n \xrightarrow{w} \theta$

b. Metropolis Hastings algos

Brief description

The intro presented key methods in Bayesian inference and basic models and faced ‘severe’ computational problems quite rapidly (beta-binomial, logistic regression)

We have introduced MCMC and Gibbs sampling

We focus here on Metropolis-Hastings and hybrid samplers

More to be followed

Sources

- :

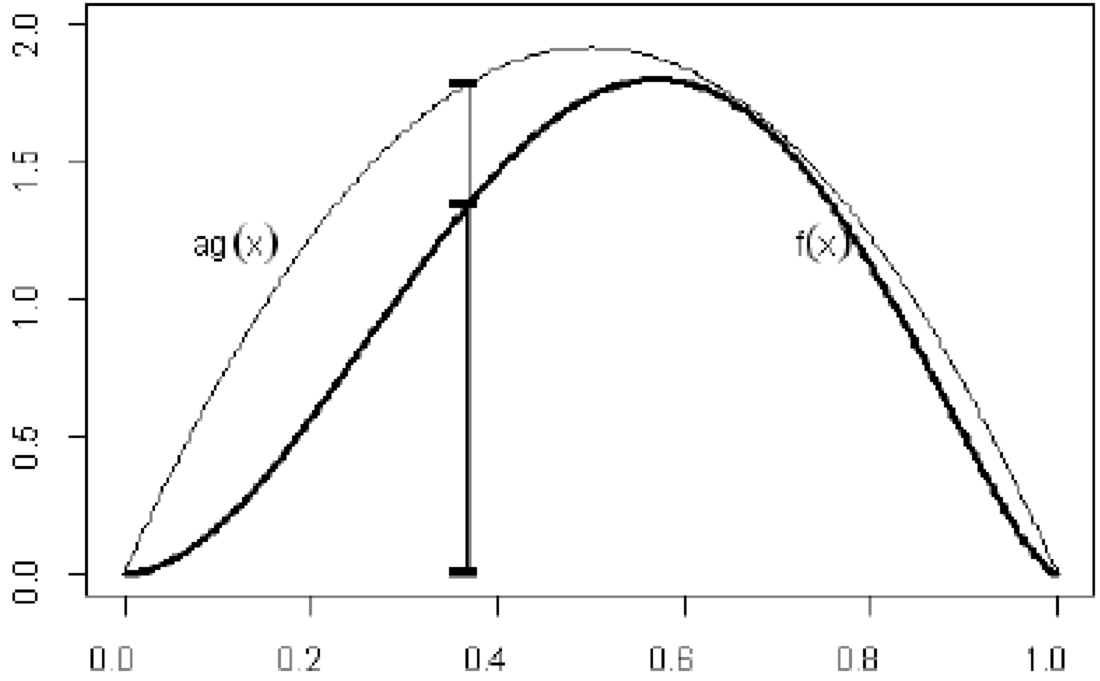
  French and DRI (2000) Ch 7
  DRI et al (2010) Ch4
  BDA3 (2015) Ch11, 12
  Hoff (2009) Ch 6,7

Understanding the Metropolis-Hastings algorithm

Recall: Acceptance-rejection sampling

TARGET DENSITY $\pi(x) = f(x)/K$, K POSSIBLY UNKNOWN

SAMPLING DENSITY $g(x)$: $f(x) \leq a g(x)$, $\forall x$



WHILE $U > f(x) / (a g(x))$

GENERATE $X \sim g$, $U \sim \mathcal{U}[0,1]$

OUTPUT X

$P(X \leq x \mid X \text{ ACCEPTED}) = F(x)$ (F cdf of X)

Metropolis-Hastings rationale I

Transition kernel of Markov chain $P(x,A)$

Invariant distribution $\pi^*(dy) = \int P(x,dy)\,\pi(x)\,dx$

n-th iterate

$$P^{(n)}(x,A) = \int P^{(n-1)}(x,dy)\,P(y,A)$$

Metropolis-Hastings rationale II

Invariant distribution and reversibility. Suppose that for p kernel (x,y)

$$P(x, dy) = p(x,y)\, dy + r(x)\, \delta_x(dy)$$

$$p(x,x) = 0 \qquad \delta_x(dy) = \begin{cases} 1, & \text{if } x \in dy \\ 0, & \text{OTHERWISE} \end{cases}$$

$$r(x) = 1 - \int p(x,y)\, dy$$

$$\pi(x)\, p(x,y) = \pi(y)\, p(y,x) \Rightarrow \pi \text{ INVARIANT FOR } P(y,\cdot)$$

Metropolis-Hastings rationale III

Adjusting a candidate generating distribution

$q(x,y)$ CANDIDATE GENERATING DENSITY

SUPPOSE

$$\pi(x)\, q(x,y) > \pi(y)\, q(y,x)$$

INTRODUCE 'CORRECTION' PROBABILITY OF MOVE

$$\alpha(x,y) < 1$$

$$p_{MH}(x,y) \equiv q(x,y)\, \alpha(x,y) \qquad x \neq y$$

$$\pi(x)\, q(x,y)\, \alpha(x,y) = \pi(y)\, q(y,x)\, \alpha(y,x)$$
$$= \pi(y)\, q(y,x)$$

$$\Longrightarrow \alpha(x,y) = \frac{\pi(y)\, q(y,x)}{\pi(x)\, q(x,y)}$$

Metropolis-Hastings rationale IV

Balance condition

$$\alpha(x,y) = \begin{cases} \min\left(\frac{\pi(y)\,q(y,x)}{\pi(x)\,q(x,y)}, 1\right), & \text{IF } \pi(x)\,q(x,y) > 0 \\ 1 & \text{, OTHERWISE} \end{cases}$$

Observations

- Normalising constant not required
- If q symmetric, Metropolis

$$\alpha(x,y) = \min\left(\frac{\pi(y)}{\pi(x)}, 1\right)$$

Metropolis-Hastings algo

1. Choose initial values θ^0. $i = 0$

2. Until convergence is detected, iterate through

   . Generate a candidate $\theta^* \sim q(\theta|\theta^i)$.

   . If $p_\theta(\theta^i)q(\theta^i \mid \theta^*) > 0$, $\alpha(\theta^i, \theta^*) = \min\left(\frac{p_\theta(\theta^*)q(\theta^*|\theta^i)}{p_\theta(\theta^i)q(\theta^i|\theta^*)}, 1\right)$;

   . else, $\alpha(\theta^i, \theta^*) = 1$.

   . Do

$$\theta^{i+1} = \begin{cases} \theta^* & \text{with prob } \alpha(\theta^i, \theta^*), \\ \theta^i & \text{with prob } 1 - \alpha(\theta^i, \theta^*) \end{cases}$$

   . $i = i + 1$.

Metropolis-Hastings variants I

Random walk chain

Metropolis algorithm

$$q(x,y) = q_1(x-y)$$

$$y = x + z, \quad z \sim q_1 \quad \begin{array}{l} \rightarrow \text{NORMAL} \\ \rightarrow t \end{array}$$

$$\alpha(x,y) = \min\left(\frac{\pi(y)}{\pi(x)}, 1\right)$$

Independence chain

$$q(x,y) = q_2(y) \quad \begin{array}{l} \rightarrow \text{NORMAL} \\ \rightarrow t \end{array}$$

Metropolis-Hastings variants II

Exploiting form of target

$$\pi(t) \propto \psi(t) h(t)$$

h DENSITY

ψ UNIF. BOUND

$$q(x,y) = h(y)$$

$$\alpha(x,y) = \min\left\{\frac{\psi(y)}{\psi(x)}, 1\right\}$$

Acceptance rate not too high, not too low. Tuning parameters

Examples

Sampling the bi-variate normal

Simple example to recall approach

Model. Bivariate normal with unknown means. Variances 1. Known correlation ϱ

1 observation

Prior. Uniform

Use. Expected value and variance of parameters, Expected cross product, Probability that parameter belongs to a set

Sampling the bi-variate normal. Model and posterior

$$y = (y_1, y_2) \sim N\left(\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix}, \Sigma\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right) \qquad \pi(\theta) \propto K$$

$$\pi(\theta_1, \theta_2 | y) \propto \pi(\theta)\, \pi(y|\theta) \propto \pi(y|\theta)$$

$$\propto \exp\left(-\frac{1}{2}(y-\theta)' \Sigma^{-1} (y-\theta)\right) = \exp\left(-\frac{1}{2}(\theta-y)' \Sigma^{-1} (\theta-y)\right)$$

$$\theta | y \sim N(y, \Sigma)$$

Sampling the bi-variate normal. Gibbs sampler

$$\theta_1 \mid \theta_2, y \sim N\left(y_1 + \rho(\theta_2 - y_2),\ 1 - \rho^2\right)$$

$$\theta_2 \mid \theta_1, y \sim N\left(y_2 + \rho(\theta_1 - y_1),\ 1 - \rho^2\right)$$

θ_2^0 ARBITRARY, $i = 1$

UNTIL CONVERGENCE

$$\theta_1^i \sim N\left(y_1 + \rho(\theta_2^{i-1} - y_2),\ 1 - \rho^2\right)$$

$$\theta_2^i \sim N\left(y_2 + \rho(\theta_1^i - y_1),\ 1 - \rho^2\right)$$

$$i = i + 1$$

Sampling the bi-variate normal. Metropolis Hastings

$$\alpha(\theta, z) = \min\left\{ \frac{\exp\left(-\frac{1}{2}(z-y)' \Sigma^{-1}(z-y)\right)}{\exp\left(-\frac{1}{2}(\theta-y)' \Sigma^{-1}(\theta-y)\right)}, 1 \right\}$$

$$= \min\left\{ \frac{\exp\left(-\frac{1}{2}\left((z' \Sigma^{-1} z) - 2 z \Sigma^{-1} y\right)\right)}{\exp\left(-\frac{1}{2}\left((\theta' \Sigma^{-1} \theta) - 2 \theta \Sigma^{-1} y\right)\right)}, 1 \right\}$$

$$z = \theta + u \qquad u \sim N(0, D)$$

Sampling the bi-variate normal. Metropolis Hastings

CHOOSE θ^0, $i=0$

UNTIL CONVERGENCE DETECTED

GENERATE $u \sim N(0, D)$

$z = \theta^i + u$

COMPUTE $\alpha(\theta^i, z)$

DO $\theta^{i+1} = \begin{cases} z, & \text{WITH PROB } \alpha(\theta^i, z) \\ \theta^i, & \text{OTHERWISE} \end{cases}$

$i = i+1$

Sampling the bi-variate normal. Answers….

$$\hat{E}(\theta_1|y) = \frac{1}{K}\sum_{i=1}^{K}\theta_1^{M+i}$$

$$\hat{E}(\theta_1\theta_2|y) = \frac{1}{K}\sum_{i=1}^{K}\left(\theta_1^{M+i}\,\theta_2^{M+i}\right)$$

$$Pr(\theta_1\in A|y) = \frac{\#\{\theta_1^{M+i}\in A\}}{K}$$

Bayesian Probit Regression

$(x_i, y_i) \qquad y_i \in \{0, 1\}$

$$\ell(\beta|y) \propto \prod_{i=1}^{n} \phi(x^{i}\beta)^{y_i} \left(1 - \phi(x^{iT}\beta)\right)^{1-y_i}$$

WITH A FLAT PRIOR $\pi(\beta) \propto 1$

$$\pi(\beta|y) \propto \prod_{i=1}^{n} \phi(x^{i}\beta)^{y_i} \left(1 - \phi(x^{iT}\beta)\right)^{1-y_i}$$

M.H. RANDOM WALK $\tilde{\beta} \sim N(\beta_{t-1}, \tau^2 \hat{\Sigma})$

COMPUTE MLE $\hat{\beta}$, $\hat{\Sigma}$ ASYMPTOTIC COVARIANCE OF $\hat{\beta}$

Bayesian Probit Regression MH

$\beta_0 = \hat{\beta}, \quad i = 1$

UNTIL CONVERGENCE

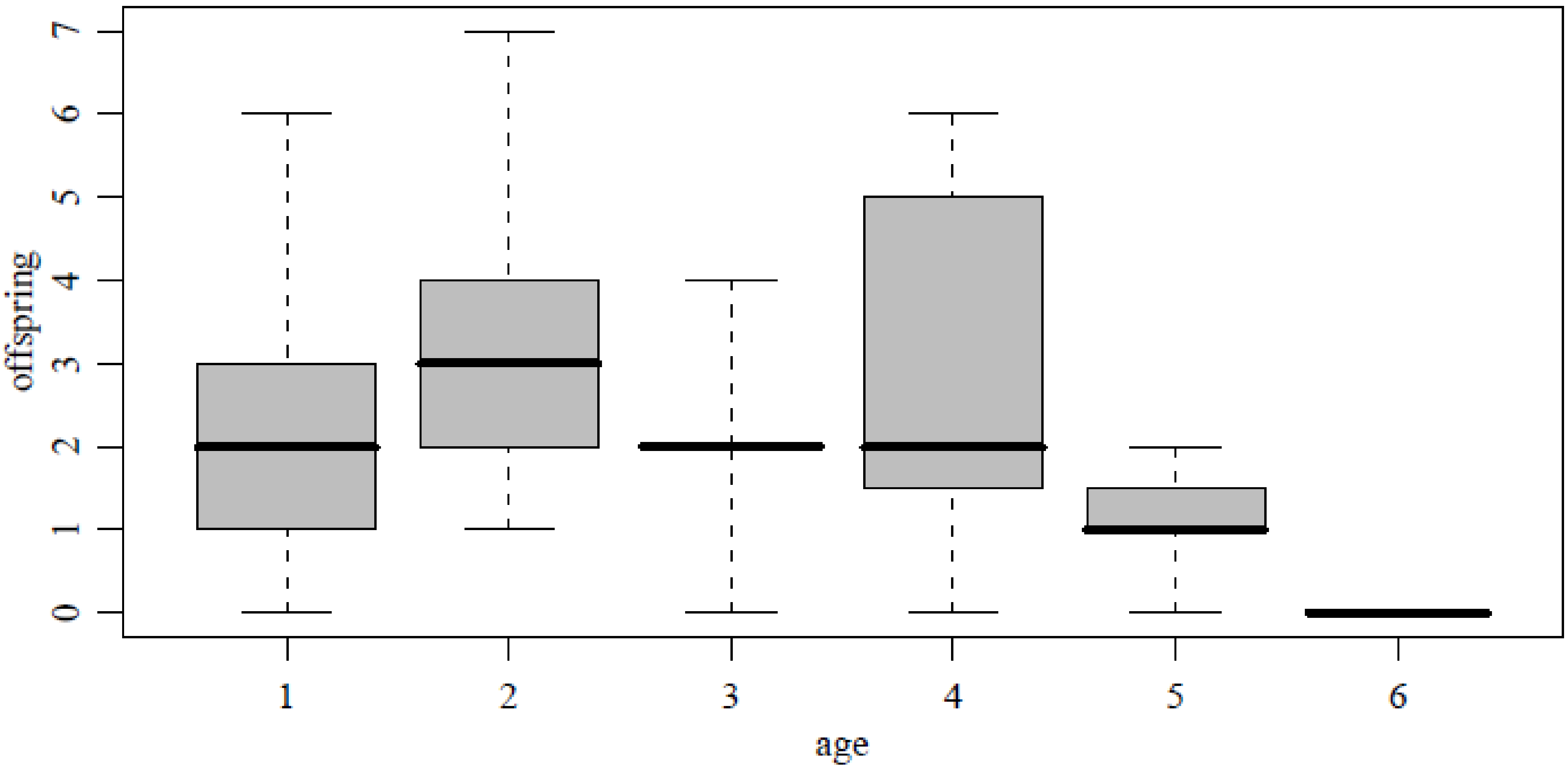
SAMPLE $\tilde{\beta} \sim N(\beta_{i-1}, \tau^2 \hat{\Sigma})$

COMPUTE $\alpha(\beta_{i-1}, \tilde{\beta}) = \min\left(1, \dfrac{\pi(\tilde{\beta} \mid y)}{\pi(\beta_{i-1} \mid y)}\right)$

$$\beta_{i+1} = \begin{cases} \tilde{\beta} & \text{WITH PROB } \alpha(\beta_{i-1}, \tilde{\beta}) \\ \beta_i, & \text{OTHERWISE} \end{cases}$$

$i = i + 1$

Bayesian Poisson Regression. Offspring of birds given age



Bayesian Poisson Regression

Offspring Y given age?? Per age?? Linear??

$$Y|x \sim \text{POISSON}(\theta_x)$$

$$\theta_x = \beta_1 + \beta_2 x + \beta_3 x^2 ??$$

$$\log \theta_x = \beta_1 + \beta_2 x + \beta_3 x^2$$

$$E(Y|x) = \exp(\beta_1 + \beta_2 x + \beta_3 x^2)$$

$$Y|x \sim \text{POISSON}(\exp(\beta^T x))$$

Bayesian Poisson Regression

Prior and posterior

$$\text{PRIOR} \quad \beta_i \sim N(0, 100)$$

$$\text{POSTERIOR} \propto \left[\prod_{i=1}^{3} N(\beta_i \mid 0, 100)\right] \left[\prod_{i=1}^{n} Po(y_i \mid \beta x)\right]$$

Acceptance probability in MH

$$\alpha(\beta, \hat{\beta}) = \min\left(\frac{\prod_{i=1}^{3} \text{dnorm}(\hat{\beta}_i, 0, 10)}{\prod_{i=1}^{3} \text{dnorm}(\beta_i, 0, 10)} \; \frac{\prod_{i=1}^{n} \text{dpois}(y_i, x_i \hat{\beta})}{\prod_{i=1}^{n} \text{dpois}(y_i, x_i \beta_i)} \; \frac{q(\hat{\beta}, \beta)}{q(\beta, \hat{\beta})}, 1 \right)$$

MH theory

Recall

1. Choose initial values $(\theta_2^0, \ldots, \theta_k^0)$. $i = 1$

2. Until convergence is detected, iterate through

   . Generate $\theta_1^i \sim \theta_1|\theta_2^{i-1}, .., \theta_k^{i-1}$

   . Generate $\theta_2^i \sim \theta_2|\theta_1^i, \theta_3^{i-1}, ..., \theta_k^{i-1}$

   . ...

   . Generate $\theta_k^i \sim \theta_k|\theta_1^i, ..., \theta_{k-1}^i$.

   . $i = i + 1$

Convergence

Kernel $p_{MH}(\theta^n, \theta^{n+1}) = \; q(\theta^n, \theta^{n+1})\alpha(\theta^n, \theta^{n+1})$, if $\theta^n \neq \theta^{n+1}$, and $1 - \int q(\theta^n, z)\alpha(\theta^n, z)dz$, otherwise.

Proposition 44 *The following hold:*

1. *If q is aperiodic, p_{MH} is aperiodic.*
2. *If q is p_{MH}-irreducible and $q(\theta^n, \theta^{n+1}) = 0$ iff $q(\theta^{n+1}, \theta^n) = 0$, p_{MH} is p_θ irreducible.*

Hybrid algos

Typical situation

Frequently

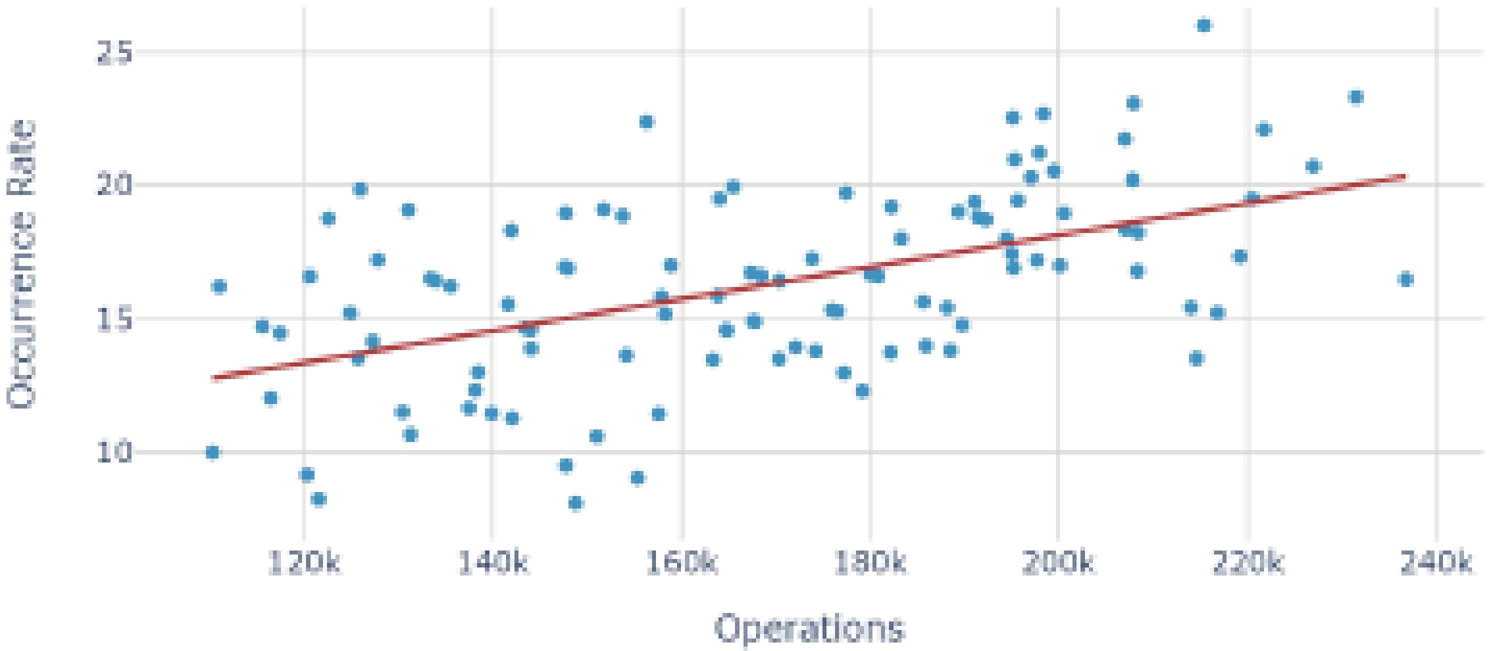
1. Some posterior conditionals available for efficient sampling
2. For others, just known up to a constant

Hybrid samplers

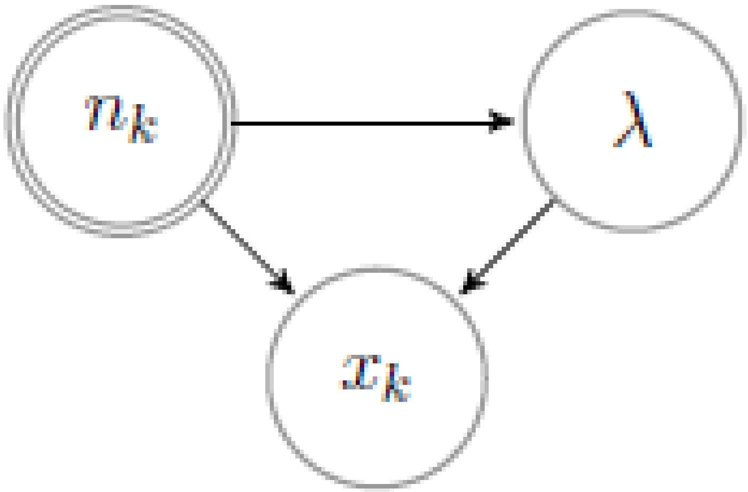
- Gibbs steps for type 1 conditionals
- MH steps for type 2 conditionals

Cse study: Aviation safety

Development of National Aviation Safety Plan. TCAS warnings



Aviation safety. Stress effect



$$x_k|\lambda, n_k \sim Po(\lambda n_k),$$

$$\lambda = a n_k + b + \epsilon_k, \quad \epsilon_k \sim N(0, \sigma^2),$$

$$a \sim N(\mu_a, \sigma_a^2), \quad b \sim N(\mu_b, \sigma_b^2), \quad \sigma^2 \sim Inv\text{-}Gamma(\alpha, \beta)$$

Aviation safety. Stress effect

Posterior proportional to

$$\lambda^{\sum_{i=1}^{k-1} x_i}\sigma^{-3-2\alpha}\exp\left(-\lambda\sum_{i=1}^{k-1} n_i - \frac{(\lambda - a n_k - b)^2}{2\sigma^2} - \frac{(a-\mu_a)^2}{2\sigma_a^2} - \frac{(b-\mu_b)^2}{2\sigma_b^2} - \frac{\beta}{\sigma^2}\right)$$

Posterior conditionals

$$\begin{aligned}
p(a|\lambda, b, \sigma^2, D_k) &\sim N\left(a \,\middle|\, \frac{\sigma^2\mu_a + n_k\sigma_a^2(\lambda - b)}{\sigma^2 + n_k\sigma_a^2}, \frac{\sigma^2}{n_k^2 + \sigma^2/\sigma_a^2}\right),\\
p(b|\lambda, a, \sigma^2, D_k) &\sim N\left(b \,\middle|\, \frac{\sigma^2\mu_b + \sigma_b^2(\lambda - a n_k)}{\sigma^2 + \sigma_b^2}, \frac{1}{1/\sigma_b^2 + 1/\sigma^2}\right),\\
p(\sigma^2|\lambda, a, b, D_k) &\sim \textit{Inv-Gamma}\left(\sigma^2 \,\middle|\, \alpha + \frac{1}{2}, \beta + \frac{1}{2}(b + a n_k - \lambda)^2\right),\\
p(\lambda|a, b, \sigma^2, D_k) &\propto \lambda^{\sum x_i}\exp\left(-\frac{1}{2\sigma^2}\left(\lambda^2 - 2\lambda\left(a n_k + b - \sigma^2\sum n_i\right)\right)\right).
\end{aligned}$$

Aviation safety. Stress effect algo

Algorithm 1: MCMC sampler for *Stress Effect* model

Set $a_0, b_0, \lambda_0, \sigma_0^2, j = 1$;

while *convergence not detected* **do**

Sample $\sigma_j^2 \sim \mathit{Inv\text{-}Gamma}\left(\alpha + \frac{1}{2}, \beta + \frac{1}{2}(b_{j-1} + a_{j-1}n_k - \lambda_{j-1})^2\right)$;

Sample $b_j \sim N\left(\frac{\sigma_j^2 \mu_b + \sigma_b^2(\lambda_{j-1} - a_{j-1}n_k)}{\sigma_j^2 + \sigma_b^2}, \frac{1}{1/\sigma_b^2 + 1/\sigma_j^2}\right)$;

Sample $a_j \sim N\left(\frac{\sigma_j^2 \mu_a + n_k \sigma_a^2(\lambda_{j-1} - b_j)}{\sigma_j^2 + n_k \sigma_a^2}, \frac{\sigma_j^2}{n_k^2 + \sigma_j^2/\sigma_a^2}\right)$;

Sample $\lambda_j^* \sim q(\lambda_{j-1})$;

$Ga\left(1 + \sum_{i=1}^{k-1} x_i, \frac{\lambda}{2\sigma^2} + \sum_{i=1}^{k-1} n_i\right)$

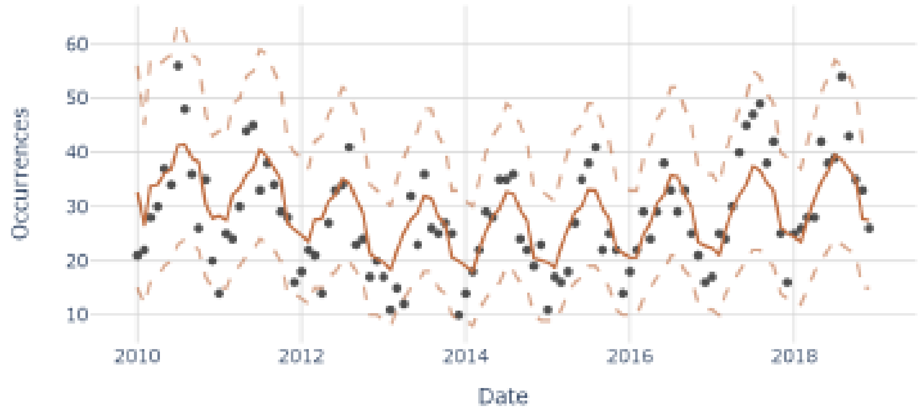
Calculate $\alpha = \min\left(1, \frac{\pi(\lambda_j^*)}{\pi(\lambda_{j-1})} \frac{q(\lambda_{j-1})}{q(\lambda_j^*)}\right)$;

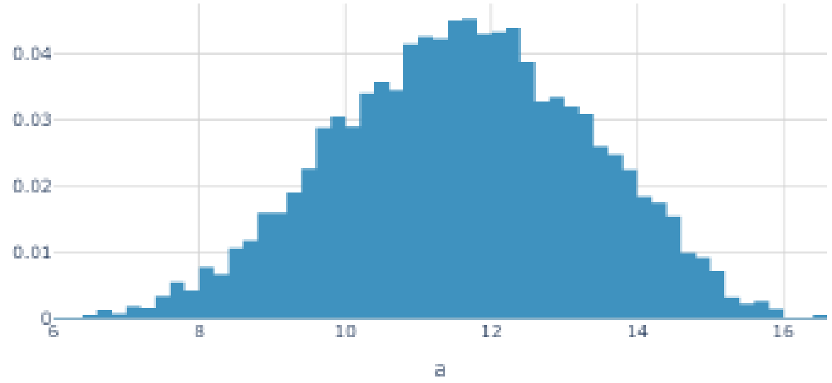
Do $\lambda_j = \begin{cases} \lambda_j^* & \text{with probability } \alpha \\ \lambda_{j-1} & \text{with probability } (1-\alpha) \end{cases}$;

$j \leftarrow j + 1$;

Aviation safety. Uses

$$
\begin{aligned}
Pr(x_k = z|D_k) &= \iiiint Pr(x_k = z|\lambda, n_k)\, p(\lambda|a, b, \sigma^2, D_k)\, p(a, b, \sigma^2|D_k)\, d\lambda\, da\, db\, d\sigma^2 \\
&\approx \frac{1}{N}\sum_{j=1}^{N} Pr(x_k = z|\lambda_j, n_k) = \frac{n_k^z}{N\, z!}\sum_{j=1}^{N} \exp(-\lambda_j n_k)\,(\lambda_j)^z\,,
\end{aligned}
$$





Comments

Gibbs requires more analysis. Not always feasible

Metropolis requires less analysis. Not so fast. More automatic

Recall potentially severe autocorrelation problems, slow mixing, convergence,…

c. Hamiltonian Monte Carlosl

Brief description

We have introduced MCMC, Gibbs sampling and Metropolis-Hastings

We focus here on Hamiltonian Monte Carlo

Sources

Sources:

BDA3 (2015) 12.4 and 12.5

Hamiltonian MC. Basics

HMC. Pros and cons

- Improved computational efficiency over MH et al (specially in high dimensional complex problems)

- Difficulties in implementation…. But Stan is available now: automates tuning of HMC parameters (and can be called from R and Python)
- But Stan a bit of a black box, hmclearn+this sesion before going through it

- Still insufficient for Bayesian analysis of deep learning models… soon

The drawback of MH

Balance condition in MH and M algos. Current x, proposed y

$$\alpha(x,y) = \begin{cases} \min\left(\frac{\pi(y)\, q(y,x)}{\pi(x)\, q(x,y)}, 1\right), & \text{IF } \pi(x)\, q(x,y) > 0 \\ 1, & \text{OTHERWISE} \end{cases}$$

$$\alpha(x,y) = \min\left(\frac{\pi(y)}{\pi(x)}, 1\right)$$

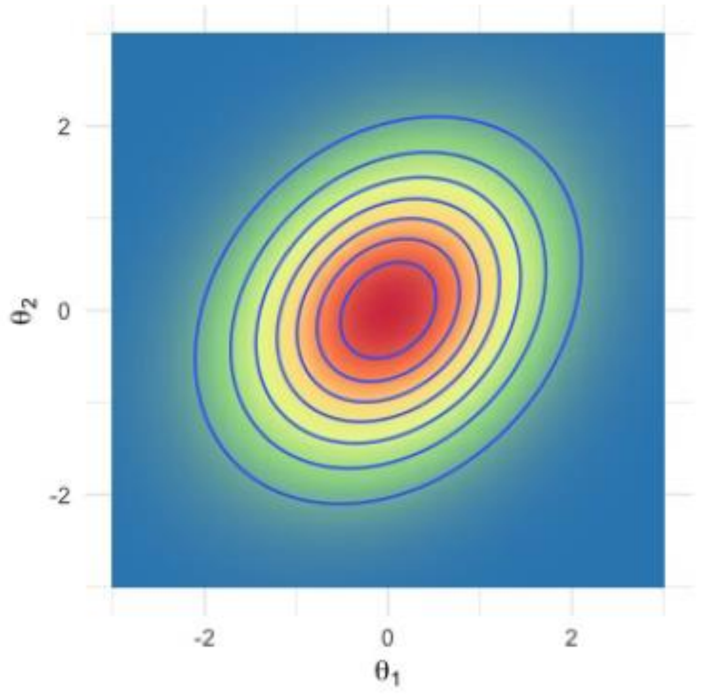
Frequents regions of higher posterior density. Sample from the right region

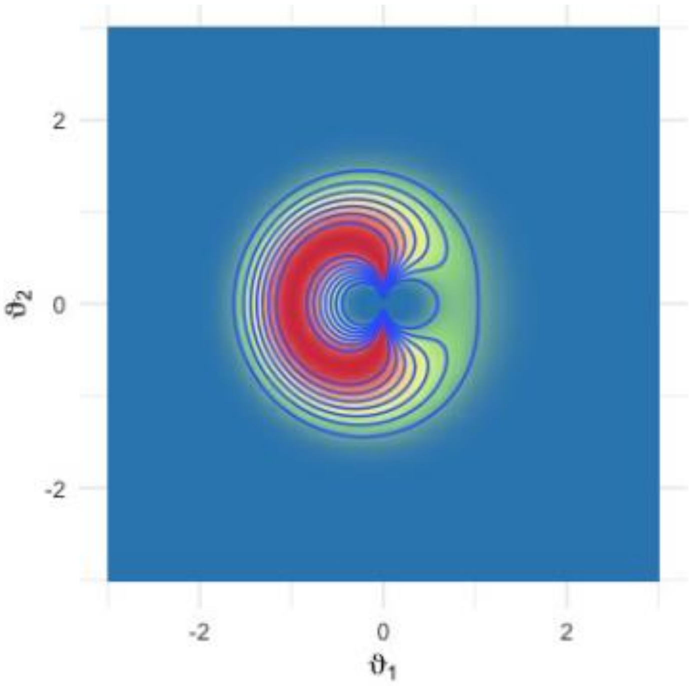
Ocasionally visits low density regions. Fully explore the sample space

As proposals are random, may take quite some time to get in HPD regions

May get stuck

The drawback of MH





HMC. Qualitative description

- A guided proposal generation scheme

- Uses the gradient of log posterior to direct MC towards HPD regions:

A well-tuned HMC accepts proposals at much higher rate than MH

- But still samples the tails properly

HMC. Idea

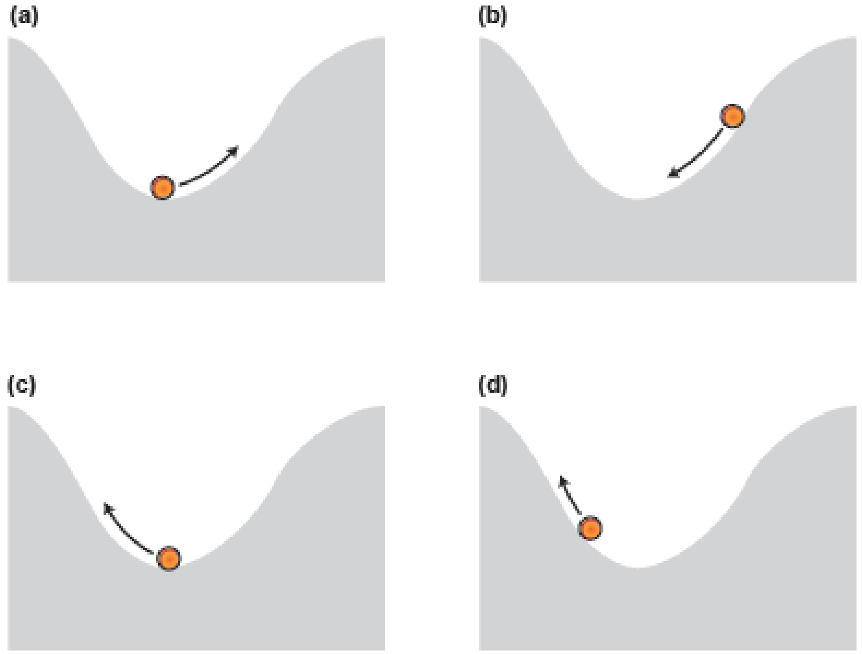
f is the posterior

-log (f) inverse bell-shaped

lower values reached guided by its gradient

In classical mechanics, exchanges between kinetic and potential energy dictate location through hamiltonian equations

(θ, p) horizontal and vertical positions. p is a momentum (auxiliary variable to actually simulate from θ) mass x velocity



Hamiltonian equations and MCMC

Target is posterior. Auxiliary momentum (same dimension)

$\theta \sim f(\theta) \qquad \dim(\theta) = \dim(p)$

Hamiltonian as potential + kinetic

$U(\theta) = -\log f(\theta) \qquad p \sim N(0, M)$

$H(\theta, p) = U(\theta) + K(p)$

$H(\theta, p) = -\log f(\theta) + \frac{1}{2} p^{T} M^{-1} p$

Hamiltonian equations

$$\left.\begin{aligned} \frac{dp}{dt} &= -\frac{\partial H(\theta, p)}{\partial \theta} = \nabla_\theta \log(f(\theta)) \\ \frac{d\theta}{dt} &= \frac{\partial H(\theta, p)}{\partial p} = M^{-1} p \end{aligned}\right] (\theta, p)$$

Hamiltonian equations through leapfrog

$$\frac{dp}{dt} = -\frac{\partial H(\theta,p)}{\partial \theta} = \nabla_\theta \log(f(\theta))$$

$$\frac{d\theta}{dt} = \frac{\partial H(\theta,p)}{\partial p} = M^{-1} p$$

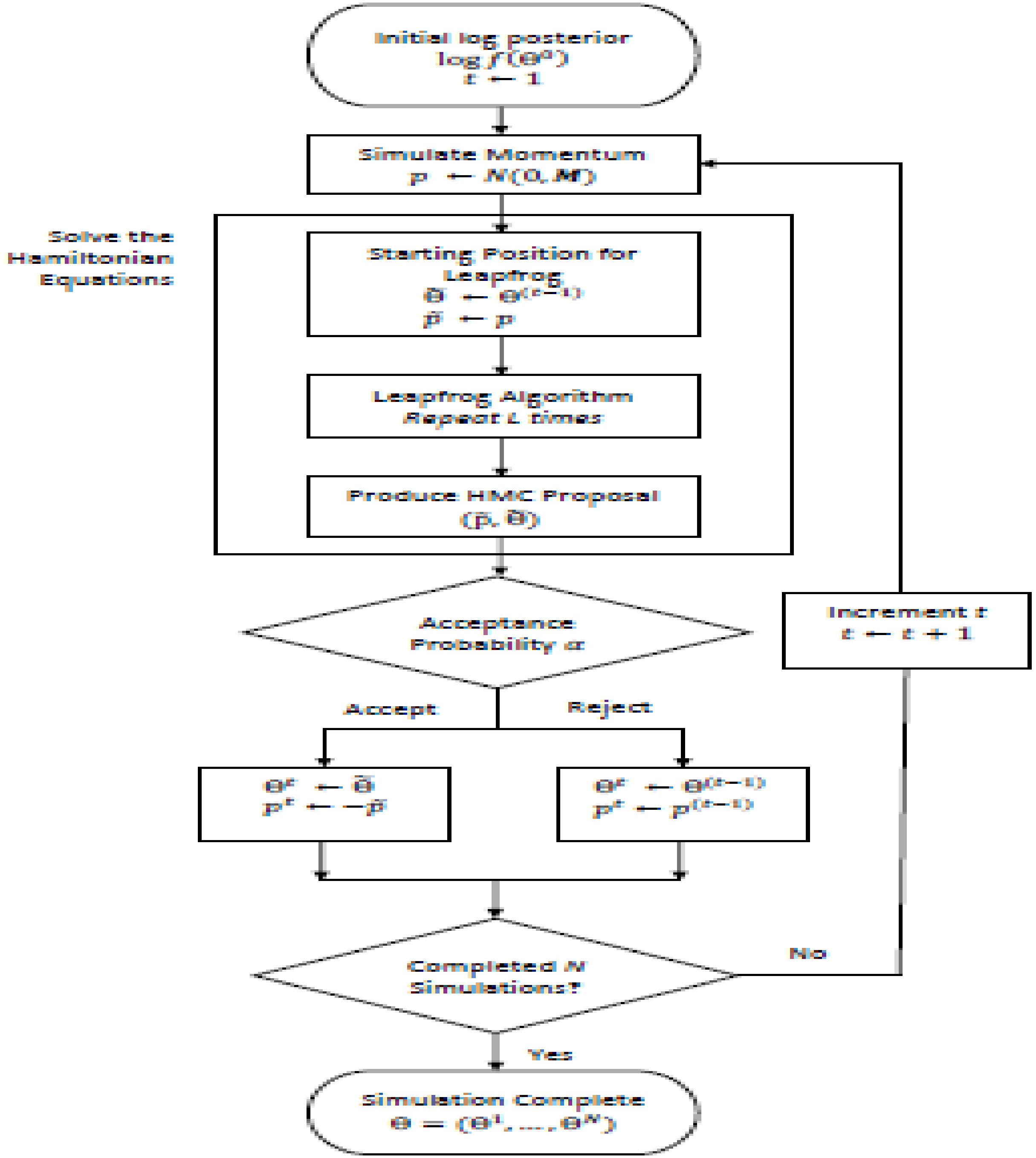
(θ,p)

$$\mathbf{p}(t+\epsilon/2) = \mathbf{p}(t) + (\epsilon/2)\nabla_{\boldsymbol{\theta}} \log f(\boldsymbol{\theta}(t)),$$

$$\boldsymbol{\theta}(t+\epsilon) = \boldsymbol{\theta}(t) + \epsilon \mathbf{M}^{-1}\mathbf{p}(t+\epsilon/2),$$

$$\mathbf{p}(t+\epsilon) = \mathbf{p}(t+\epsilon/2) + (\epsilon/2)\nabla_{\boldsymbol{\theta}} \log f(\boldsymbol{\theta}(t+\epsilon)).$$

HMC. Algo I



Initial log posterior
log f(Θ⁰)
t ← 1
Simulate Momentum
p ← N(0, M)
Solve the Hamiltonian Equations
Starting Position for Leapfrog
θ̃ ← θ^(t−1)
p̃ ← p
Leapfrog Algorithm
Repeat L times
Produce HMC Proposal
(p̃, θ̃)
Acceptance Probability α
Accept
Reject
θᵗ ← θ̃
pᵗ ← −p̃
θᵗ ← θ^(t−1)
pᵗ ← p^(t−1)
Increment t
t ← t + 1
No
Completed N Simulations?
Yes
Simulation Complete
Θ = (Θ¹, …, Θᴺ)

HMC. Algo II

procedure HMC($\boldsymbol{\theta}^{(0)}, \log f(\boldsymbol{\theta}), \mathrm{M}, N, \epsilon, L$)

  Calculate $\log f(\boldsymbol{\theta}^{(0)})$

  for $t = 1, ..., N$ **do**

    $\mathrm{p} \leftarrow N(0, \mathrm{M})$

    $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}^{(t-1)}, \bar{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}^{(t-1)}, \bar{\mathrm{p}} \leftarrow \mathrm{p}$

    for $i = 1, ..., L$ **do**

      $\bar{\boldsymbol{\theta}}, \bar{\mathrm{p}} \leftarrow \text{Leapfrog}(\bar{\boldsymbol{\theta}}, \bar{\mathrm{p}}, \epsilon, \mathrm{M})$

    end for

    $\alpha \leftarrow \min\left(1, \frac{\exp(\log f(\bar{\boldsymbol{\theta}}) - \frac{1}{2}\tilde{\mathrm{p}}^T \mathrm{M}^{-1} \tilde{\mathrm{p}})}{\exp(\log f(\boldsymbol{\theta}^{(t-1)}) - \frac{1}{2}\mathrm{p}^T \mathrm{M}^{-1} \mathrm{p})}\right)$

    With probability α, $\boldsymbol{\theta}^{(t)} \leftarrow \bar{\boldsymbol{\theta}}$ and $\mathrm{p}^{(t)} \leftarrow -\bar{\mathrm{p}}$

  end for

  return $\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(N)}$

  function LEAPFROG($\boldsymbol{\theta}^*, \mathrm{p}^*, \epsilon, \mathrm{M}$)

    $\bar{\mathrm{p}} \leftarrow \mathrm{p}^* + (\epsilon/2)\nabla_{\boldsymbol{\theta}} \log f(\boldsymbol{\theta}^*)$

    $\bar{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}^* + \epsilon \mathrm{M}^{-1} \bar{\mathrm{p}}$

    $\bar{\mathrm{p}} \leftarrow \bar{\mathrm{p}} + (\epsilon/2)\nabla_{\boldsymbol{\theta}} \log f(\bar{\boldsymbol{\theta}})$

    return $\bar{\boldsymbol{\theta}}, \bar{\mathrm{p}}$

  end function

end procedure

HMC Tuning

Step size. Small relative to parameter of interest

x Number of leapfrog steps. Large L.

Jointly acceptance rate of 65%

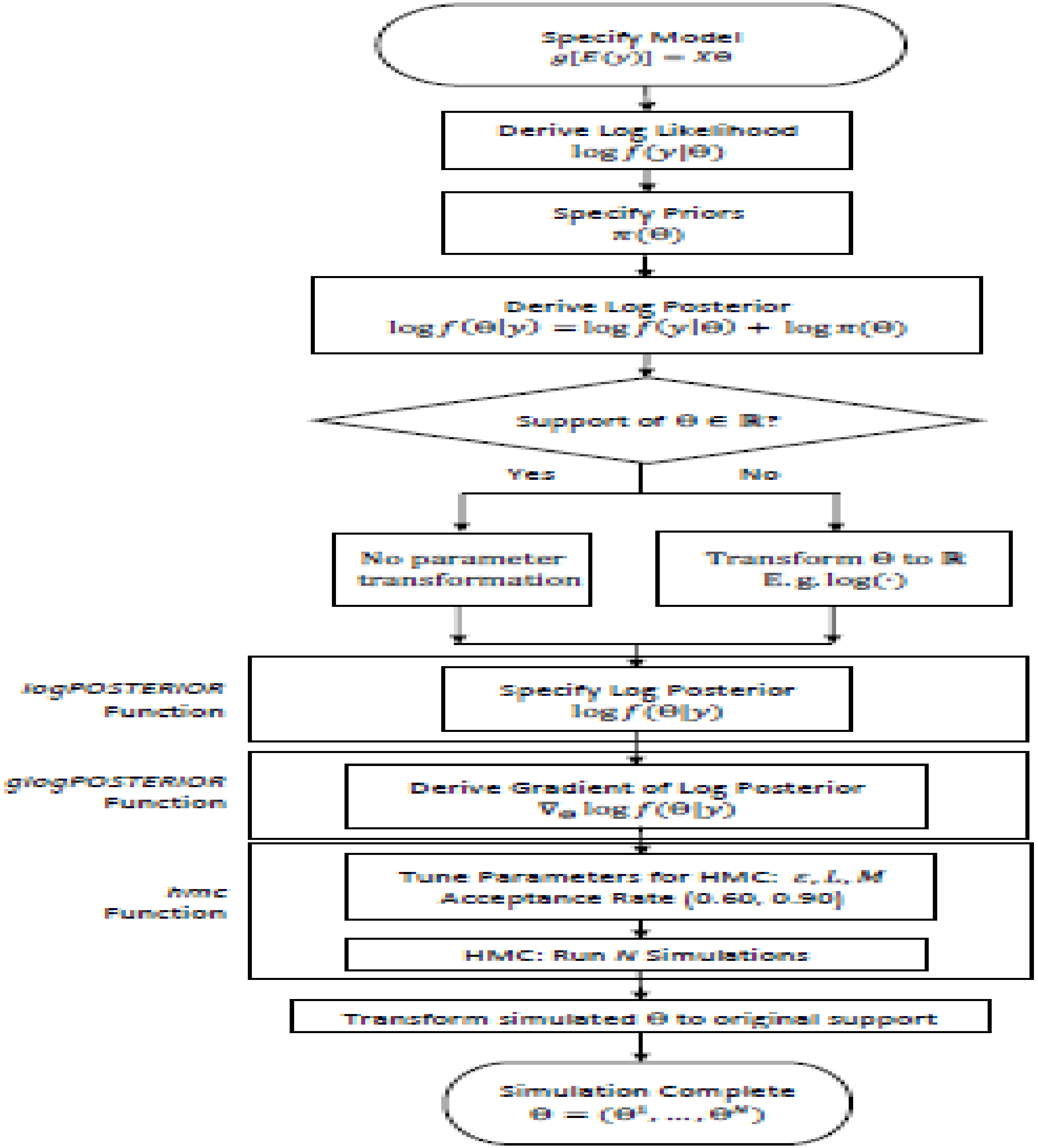
Examine for correlations

Adaptively select L as in No U-turn Sampler (NUTS) . To be seen with Stan

Covariance matrix M

Hamiltonian MC. Examples

Typical setup



Sampling the bi-variate normal

Simple example to recall approach

Model. Bivariate normal with unknown means. Variances 1. Known correlation ϱ

1 observation

Prior. Uniform

Use. Expected value and variance of parameters, Expected cross product, Probability that parameter belongs to a set

Sampling the bi-variate normal. Model and posterior

$$y = (y_1, y_2) \sim N\left(\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix}, \Sigma\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right) \qquad \pi(\theta) \propto K$$

$$\pi(\theta_1, \theta_2 \mid y) \propto \pi(\theta)\, \pi(y \mid \theta) \propto \pi(y \mid \theta)$$

$$\propto \exp\left(-\frac{1}{2}(y-\theta)' \Sigma^{-1} (y-\theta)\right) = \exp\left(-\frac{1}{2}(\theta-y)' \Sigma^{-1} (\theta-y)\right)$$

$$\theta \mid y \sim N(y, \Sigma)$$

Sampling the bi-variate normal. Gibbs sampler

$$\theta_1 | \theta_2, y \sim N(y_1 + \rho(\theta_2 - y_2), 1 - \rho^2)$$

$$\theta_2 | \theta_1, y \sim N(y_2 + \rho(\theta_1 - y_1), 1 - \rho^2)$$

θ_2^0 ARBITRARY, $i = 1$

UNTIL CONVERGENCE

$$\theta_1^i \sim N(y_1 + \rho(\theta_2^{i-1} - y_2), 1 - \rho^2)$$

$$\theta_2^i \sim N(y_2 + \rho(\theta_1^i - y_1), 1 - \rho^2)$$

$$i = i + 1$$

Sampling the bi-variate normal. Metropolis Hastings

$$\alpha(\theta, z) = \min\left\{ \frac{\exp\left(-\frac{1}{2}(z-y)'\,\Sigma^{-1}(z-y)\right)}{\exp\left(-\frac{1}{2}(\theta-y)'\,\Sigma^{-1}(\theta-y)\right)}, 1 \right\}$$

$$= \min\left\{ \frac{\exp\left(-\frac{1}{2}\left((z'\Sigma^{-1}z) - 2z\,\Sigma^{-1}y\right)\right)}{\exp\left(-\frac{1}{2}\left((\theta'\Sigma^{-1}\theta) - 2\theta\,\Sigma^{-1}y\right)\right)}, 1 \right\}$$

$$z = \theta + u \qquad u \sim N(0, D)$$

Sampling the bi-variate normal. Metropolis Hastings

CHOOSE θ^0, $i=0$

UNTIL CONVERGENCE DETECTED

GENERATE $u \sim N(0, D)$

$z = \theta^i + u$

COMPUTE $\alpha(\theta^i, z)$

DO $\theta^{i+1} = \begin{cases} z, & \text{WITH PROB } \alpha(\theta^i, z) \\ \theta^i, & \text{OTHERWISE} \end{cases}$

$i = i+1$

Sampling the bi-variate normal. Hamiltonian MC

procedure HMC($\boldsymbol{\theta}^{(0)}, \log f(\boldsymbol{\theta}), \mathbf{M}, N, \epsilon, L$)

  Calculate $\log f(\boldsymbol{\theta}^{(0)})$

  for $t = 1, ..., N$ **do**

    $\mathbf{p} \leftarrow N(0, \mathbf{M})$

    $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}^{(t-1)}, \bar{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}^{(t-1)}, \bar{\mathbf{p}} \leftarrow \mathbf{p}$

    for $i = 1, ..., L$ **do**

      $\bar{\boldsymbol{\theta}}, \bar{\mathbf{p}} \leftarrow \text{Leapfrog}(\bar{\boldsymbol{\theta}}, \bar{\mathbf{p}}, \epsilon, \mathbf{M})$

    end for

    $\alpha \leftarrow \min\left(1, \frac{\exp(\log f(\bar{\boldsymbol{\theta}}) - \frac{1}{2}\bar{\mathbf{p}}^T \mathbf{M}^{-1} \bar{\mathbf{p}})}{\exp(\log f(\boldsymbol{\theta}^{(t-1)}) - \frac{1}{2}\mathbf{p}^T \mathbf{M}^{-1} \mathbf{p})}\right)$

    With probability α, $\boldsymbol{\theta}^{(t)} \leftarrow \bar{\boldsymbol{\theta}}$ and $\mathbf{p}^{(t)} \leftarrow -\bar{\mathbf{p}}$

  end for

  return $\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(N)}$

  function LEAPFROG($\boldsymbol{\theta}^*, \mathbf{p}^*, \epsilon, \mathbf{M}$)

    $\bar{\mathbf{p}} \leftarrow \mathbf{p}^* + (\epsilon/2) \nabla_{\boldsymbol{\theta}} \log f(\boldsymbol{\theta}^*)$

    $\bar{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}^* + \epsilon \mathbf{M}^{-1} \bar{\mathbf{p}}$

    $\bar{\mathbf{p}} \leftarrow \bar{\mathbf{p}} + (\epsilon/2) \nabla_{\boldsymbol{\theta}} \log f(\bar{\boldsymbol{\theta}})$

    return $\bar{\boldsymbol{\theta}}, \bar{\mathbf{p}}$

  end function

end procedure

$\theta | y \sim N(y, \Sigma)$

$\log(f(\theta)) \propto -\frac{1}{2} (\theta - y)' \Sigma^{-1} (\theta - y)$

$-\log(f(\theta)) \propto \frac{1}{2} (\theta - y)' \Sigma^{-1} (\theta - y)$

$-\nabla_\theta (\log(f(\theta))) \propto (\theta - y)' \Sigma^{-1}$

Sampling the bi-variate normal. Hamiltonian MC. Bonus

$$H(\theta,p) = \frac{1}{2}(\theta-y)^t \Sigma^{-1}(\theta-y) + \frac{1}{2} p^t M^{-1} p$$

$$\frac{dp}{dt} = -\frac{\partial H(\theta,p)}{\partial\theta} = (\theta-y)'\Sigma^{-1}$$

$$\frac{d\theta}{dt} = \frac{\partial H(\theta,p)}{\partial p} = M^{-1}p$$

$$p(t+\tfrac{\varepsilon}{2}) = p(t) + \frac{\varepsilon}{2}(y-\theta(t))'\Sigma^{-1}$$

$$\theta(t+\varepsilon) = \theta(t) + \varepsilon M^{-1} p(t+\varepsilon/2)$$

$$p(t+\varepsilon) = p(t+\tfrac{\varepsilon}{2}) + \frac{\varepsilon}{2}(y-\theta(t+\tfrac{\varepsilon}{2}))'\Sigma^{-1}$$

Sampling the bi-variate normal. Answers (Whatever the method used)

$$\hat{E}(\theta_1|y) = \frac{1}{K}\sum_{i=1}^{K} \theta_1^{M+i}$$

$$\hat{E}(\theta_1\theta_2|y) = \frac{1}{K}\sum_{i=1}^{K} \left(\theta_1^{M+i}\,\theta_2^{M+i}\right)$$

$$Pr(\theta_1 \in A|y) = \frac{\#\{\theta_1^{M+i} \in A\}}{K}$$

Linear regression model

$$y_i = x_i^t \beta + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \quad i = 1, \dots, n$$

$$y = (y_1, \dots, y_n)$$

$$X = (x_1^t, \dots, x_n^t)$$

$$\log f(y \mid \beta, \sigma^2) \propto -n \log \sigma - \frac{1}{2\sigma^2} (y - X^t \beta)(y - X\beta)$$

Normal Prior for weights

$$f(\beta \mid \sigma_\beta^2) \propto \exp\left(-\frac{\beta^t \beta}{2\sigma_\beta^2}\right)$$

$$f(\sigma_\varepsilon^2 \mid a, b) = \frac{b^a}{\Gamma(a)} (\sigma_\varepsilon^2)^{-a-1} \exp\left(-\frac{b}{\sigma_\varepsilon^2}\right)$$

Inv-gamma prior for variance

$$\gamma = \log \sigma_\varepsilon^2 \qquad \sigma_\varepsilon^2 = e^\gamma$$

$$f(\gamma \mid a, b) = \frac{b^a}{\Gamma(a)} \exp\left(-a\gamma - \frac{b}{e^\gamma}\right)$$

$$\log\left(f(\gamma \mid a, b)\right) \propto -a\gamma - b e^{-\gamma}$$

Linear regression model

$$\log f(\beta, \gamma \mid y, X, \sigma_\beta^2, a, b) \propto -\left(\frac{n}{2} + a\right)\gamma - \frac{e^{-\gamma}}{2}(y - X\beta)^T(y - X\beta) - \frac{\beta^T\beta}{2\sigma_\beta^2} - be^{-\gamma}$$

$$\nabla_\beta \log f(\beta, \gamma \mid y, X, \sigma_\beta^2, a, b) \propto e^{-\gamma} X^T (y - X\beta) - \beta / \sigma_\beta^2$$

$$\nabla_\gamma \log f(\beta, \gamma \mid y, X, \sigma_\beta^2, a, b) \propto -\left(\frac{n}{2} + a\right) + \frac{e^{-\gamma}}{2}(y - X\beta)^T(y - X\beta) + be^{-\gamma}$$

Logistic regression model

$$P(y_i = 1 \mid x_i, \beta) = \frac{1}{1 + \exp(-x_i^t \beta)}$$

$$X = (x_1^t, \dots, x_n^t)^t \qquad \beta = (\beta_0, \dots, \beta_p)$$

$$\log f(y \mid X, \beta) = \beta^T X^T (y - 1_n) - 1_n^t \left(\log \left(1 + e^{-x_i \beta} \right) \right)_{\dots}$$

$$\beta \sim N(0, \sigma_\beta^2 I)$$

Logistic regression model

$$\log f(\beta \mid y, X, \sigma_\beta^2) \propto \beta^T X^T (y - 1_n) - 1_n^t \left(\log\left(1 + e^{-x_i^t \beta}\right) \right)_{n \times 1} - \frac{\beta^T \beta}{2\sigma_\beta^2}$$

$$\nabla_\beta \log f(\beta \mid y, X, \sigma_\beta^2) \propto X^T \left(y - 1_n + \left[\frac{e^{-x_i^t \beta}}{1 + e^{-x_i^t \beta}} \right]_{n \times 1} \right) - \beta / \sigma_\beta^2$$

Final comments

Final comments

Gibbs. Lots of hard prior work. Not always implementable. If so may work well.

Metropolis Hastings. Less priori work. Quite general. May work slowly.

Hamiltonian. Even less work. Quite general. Works more efficiently…. Yet suffers in large scale problems

More when introducing Stan

More when introducing SG-MCMC and Variational inference

d. Complements

Schedule

Sources:

BDA3 (2015) 11 and 12

More when we check Stan

MCMC recall

Core computational problem in Bayesian analysis

Computing the posterior

$$f(\theta|x) = \frac{f(x|\theta)\, f(\theta)}{f(x)} = \frac{f(x|\theta)\, f(\theta)}{\int f(x|\theta)\, f(\theta)\, d\theta} \propto f(x|\theta)\, f(\theta)$$

General idea MCMC

Markov chain X_n with same state space as target and convergent to target distribution g

$$X_n \xrightarrow{d} g$$

Strategy

Initialise x_0, $n=1$

Until convergence, Generate $X_n \mid X_{n-1}=x_{n-1}$, $n=n+1$, n^*

Until n^*+m, Generate and collect $X_n \mid X_{n-1}=x_{n-1}$, $n=n+1$

$$\hat{I}_g \approx \frac{1}{m} \sum_{n^*}^{n^*+m} f(x_i)$$

Gibbs sampling

1. Choose initial values $(\theta_2^0, \ldots, \theta_k^0)$. $i = 1$

2. Until convergence is detected, iterate through

. Generate $\theta_1^i \sim \theta_1 | \theta_2^{i-1}, .., \theta_k^{i-1}$

. Generate $\theta_2^i \sim \theta_2 | \theta_1^i, \theta_3^{i-1}, ..., \theta_k^{i-1}$

. ...

. Generate $\theta_k^i \sim \theta_k | \theta_1^i, ..., \theta_{k-1}^i$.

. $i = i + 1$

Metropolis-Hastings algo

1. Choose initial values θ^0. $i = 0$

2. Until convergence is detected, iterate through

. Generate a candidate $\theta^* \sim q(\theta|\theta^i)$.

. If $p_\theta(\theta^i)q(\theta^i \mid \theta^*) > 0$, $\alpha(\theta^i, \theta^*) = \min\left(\frac{p_\theta(\theta^*)q(\theta^*|\theta^i)}{p_\theta(\theta^i)q(\theta^i|\theta^*)}, 1\right)$;

. else, $\alpha(\theta^i, \theta^*) = 1$.

. Do

$$\theta^{i+1} = \begin{cases} \theta^* & \text{with prob } \alpha(\theta^i, \theta^*), \\ \theta^i & \text{with prob } 1 - \alpha(\theta^i, \theta^*) \end{cases}$$

. $i = i + 1$.

HMC

procedure HMC($\boldsymbol{\theta}^{(0)}, \log f(\boldsymbol{\theta}), \mathrm{M}, N, \epsilon, L$)

  Calculate $\log f(\boldsymbol{\theta}^{(0)})$

  for $t = 1, ..., N$ **do**

    $\mathrm{p} \leftarrow N(0, \mathrm{M})$

    $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}^{(t-1)}, \bar{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}^{(t-1)}, \bar{\mathrm{p}} \leftarrow \mathrm{p}$

    for $i = 1, ..., L$ **do**

      $\bar{\boldsymbol{\theta}}, \bar{\mathrm{p}} \leftarrow \mathrm{Leapfrog}(\bar{\boldsymbol{\theta}}, \bar{\mathrm{p}}, \epsilon, \mathrm{M})$

    end for

    $\alpha \leftarrow \min\left(1, \frac{\exp(\log f(\bar{\boldsymbol{\theta}}) - \frac{1}{2}\bar{\mathrm{p}}^T \mathrm{M}^{-1}\bar{\mathrm{p}})}{\exp(\log f(\bar{\boldsymbol{\theta}}^{(t-1)}) - \frac{1}{2}\mathrm{p}^T \mathrm{M}^{-1}\mathrm{p})}\right)$

    With probability α, $\boldsymbol{\theta}^{(t)} \leftarrow \bar{\boldsymbol{\theta}}$ and $\mathrm{p}^{(t)} \leftarrow -\bar{\mathrm{p}}$

  end for

  return $\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(N)}$

  function LEAPFROG($\boldsymbol{\theta}^*, \mathrm{p}^*, \epsilon, \mathrm{M}$)

    $\bar{\mathrm{p}} \leftarrow \mathrm{p}^* + (\epsilon/2)\nabla_{\boldsymbol{\theta}} \log f(\boldsymbol{\theta}^*)$

    $\bar{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}^* + \epsilon \mathrm{M}^{-1}\bar{\mathrm{p}}$

    $\bar{\mathrm{p}} \leftarrow \bar{\mathrm{p}} + (\epsilon/2)\nabla_{\boldsymbol{\theta}} \log f(\bar{\boldsymbol{\theta}})$

    return $\bar{\boldsymbol{\theta}}, \bar{\mathrm{p}}$

  end function

end procedure

Further variants

Further variants

There's many other variants

Slice sampling, simulated tempering, parallel tempering,…

We sketch here two: reversible jump sampling and particle filtering

Later we mention SG-MCMC variants

Reversible jump

In many complex problems we need to do trans-dimensional Markov chain simulation

- Mixtures with unknown number of components
- Shallow neural nets with unknown number of hidden nodes
- Model averaging
- Bartmachine

Parameters=(indicator of model, parameters of such model)

Within the model, a 'standard' Markov chain (Gibbs, MH, HMC, etc…)

Between models, Metropolis Hastings with reversible moves (jumps, collapsing and splitting models…)

See classic paper by Peter Green in VC

Particle filtering

For nonlinear sequential problems, MCMC gets complex

Generate initial sample at time t=0

Let them evolve (and learn) according to nonlinear sequential model

Introduce rules to avoid collapse of particles

Inference and assessing convergence

Inference

Once convergence detected, collect samples from posterior and perform inference (point estimates, intervals, hypothesis tests, predictions and expected utility computations) via Monte Carlo (recall uncertainty associated, they are stochastic algos!!!)

Difficulties

If iterations have not proceeded long enough, target is not approximated well, samples are unrepresentative of target!!!

Early iterations may bias results

Autocorrelation impacts precision of estimates and the effective number of samples may be smaller than the one actually drawn (as if we'd be using a smaller number of samples)

Solutions

Runs to allow for effective monitoring of convergence (based on multiple chains, recall labs)

Monitor convergence by comparing variation within and between simulated sequences (until within and between variation are similar)

Modifying the algorithm by reparameterising or learning good parameterisations, if efficiency is very low (algo too slow)

Discarding initial values

Thinning

Take into account AC when estimating precisions

Discarding early iterations

Rule of thumb: Discard first half of iterations (warm-up, burn-in) and use the second half of iterations for inference and prediction

Warm-up fraction may vary

Warm-up fraction may adapt

- Example. Run 200. Discard first 100. If second 200, do not suggest convergence, run 200 more and discard first 200.

Thinning iterations

Mitigate autocorrelation: keep every k-th simulation draw and discard the rest (to avoid storage problems also!!!)

Rule of thumb: Choose k to save up to 1000 draws

Assessing convergence

Comparing different simulated sequences from multiple chains with overdispersed starting points

- Compare variance within each sequence with variance between sequences
- Monitoring various scalar estimands (and other quantities of interest like the logposterior) separately
- Checking mixing and stationarity
  - Split each series in two parts
    - after discarding burn-in

    - Check R

- coda
- monitor

$\psi \longrightarrow \psi_{ij} \quad \begin{pmatrix} i = 1, \dots, n & \text{\# length of each chain} \\ j = 1, \dots, m & \text{\# of chains} \end{pmatrix}$

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\bar{\psi}_{.j} - \bar{\psi}_{..})^2 \qquad \bar{\psi}_{.j} = \frac{1}{n} \sum_{i=1}^{n} \psi_{ij} \qquad \bar{\psi}_{..} = \frac{1}{m} \sum_{j=1}^{m} \bar{\psi}_{.j}$$

$$W = \frac{1}{m} \sum_{j=1}^{m} s_j^2 \qquad s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\psi_{ij} - \bar{\psi}_{.j})^2$$

$$\widehat{var}(\psi|y) = \frac{n-1}{n} W + \frac{1}{n} B \qquad \hat{R} = \sqrt{\frac{\widehat{var}(\psi|y)}{W}} \downarrow 1$$

Effective number of simulation draws

Once simulated sequences judged to have mixed, estimate effective number of simulation draws for any estimand of interest

If draws within each sequence truly independent, B would be an unbiased estimator of the posterior variance and we'd have mn independent observations

But typically will be autocorrelated and E(B) will be larger than the posterior variance

Effective number of simulation draws

We have

$$\bar{\psi}_{..} = \frac{1}{m}\sum_{j=1}^{m}\bar{\psi}_{.j} \qquad \lim_{n\to\infty} mn\,\mathrm{var}(\bar{\psi}_{..}) = \left(1+2\sum_{t=1}^{\infty}\rho_t\right)\mathrm{var}(\psi|y)$$

And define the effective simple size as

$$n_{eff} = \frac{mn}{1+2\sum_{t=1}^{\infty}\rho_t}$$

We estimate it through

$$V_t = \frac{1}{m(n-t)}\sum_{j=1}^{m}\sum_{i=t+1}^{n}\left(\psi_{i,j}-\psi_{i-t,j}\right)^2$$

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

$$E(\psi_i-\psi_{i-t})^2 = 2(1-\rho_t)\,\mathrm{var}(\psi) \longrightarrow \hat{\rho}_t = 1-\frac{V_t}{2\,\widehat{\mathrm{var}}^{+}}$$

$$\hat{n}_{eff} = \frac{mn}{1+2\sum_{t=1}^{T}\hat{\rho}_t} \qquad T \overset{\text{first}}{:=} \hat{\rho}_{T+1}+\hat{\rho}_{T+2}<0$$

Final comments

Check lab on Stan (monitor+effective sample size)

SG-MCMC to be seen

Variational Bayes to be seen