STK4051/9051 Computational statistics

Geir Storvik

Ch 7 - Markov chain Monte Carlo

- Assume now simulating from $f(\mathbf{X})$ is difficult directly
  - $f(\cdot)$ complicated
  - $\mathbf{X}$ high-dimensional
- Markov chain Monte Carlo:
  - Generates $\{\mathbf{X}^{(t)}\}$ sequentially
  - Markov structure: $\mathbf{X}^{(t)} \sim P(\cdot|\mathbf{X}^{(t-1)})$
- Aim now:
  - The distribution of $\mathbf{X}^{(t)}$ converges to $f(\cdot)$ as t increases
  - $\hat{\mu}_{MCMC} = N^{-1}\sum_{t=1}^N h(\mathbf{X}^{(t)})$ converges towards $\mu = E^f[h(\mathbf{X})]$ as t increases

Markov chain theory - discrete case

- Assume $\{X^{(t)}\}$ is a Markov chain where $X^{(t)}$ is a discrete random variable

  $$\Pr(X^{(t)} = y|X^{(t-1)} = x) = P(y|x)$$

  giving the transition probabilities
- Assume the chain is
  - irreducible: It is possible to move from any **x** to any **y** in a finite number of steps
  - reccurent: The chain will visit any state infinitely often.
  - aperiodic: Does not go in cycles
- Then there exists a unique distribution $f(x)$ such that

  $$\lim_{t\to\infty} \Pr(X^{(t)} = y|X^{(0)} = x) = f(y)$$

  $$\hat{\mu}_{MCMC} \to \mu = E^f[X]$$

- How to find $f(\cdot)$ (the stationary distribution): Solve

  $$f(y) = \sum_x f(x)P(y|x)$$

- Our situation: We have $f(y)$, want to find $P(y|x)$
  - Note: Many possible $P(y|x)$!

Markov chain theory - general setting

- Assume $\{\mathbf{X}^{(t)}\}$ is a Markov chain where $\mathbf{X}^{(t)} \in S$

$$\Pr(\mathbf{X}^{(t)} \in A|\mathbf{X}^{(t-1)} = \mathbf{x}) = P(\mathbf{x}, A) = \int_{\mathbf{y}\in A} P(\mathbf{y}|\mathbf{x})d\mathbf{y}$$

  giving the transition densities
- Assume the chain is
  - irreducible: It is possible to move from any $\mathbf{x}$ to any $\mathbf{y}$ in a finite number of steps
  - reccurent: The chain will visit any $A \subset S$ infinitely often.
  - aperiodic: Do not go in cycles
- Then there exists a distribution $f(\mathbf{x})$ such that

$$\lim_{t\to\infty} \Pr(\mathbf{X}^{(t)} \in A|\mathbf{X}^{(0)} = \mathbf{x}) = \int_A f(\mathbf{y})d\mathbf{y}$$

$$\hat{\mu}_{MCMC} \to \mu$$

- How to find $f(\cdot)$ (the stationary distribution): Solve

$$f(\mathbf{y}) = \int_{\mathbf{x}} f(\mathbf{x})P(\mathbf{y}|\mathbf{x})d\mathbf{x}$$

- Our situation: We have $f(\cdot)$, want to find $P(\mathbf{y}|\mathbf{x})$

Detailed balance

- The task: Find a transition probability/density $P(\mathbf{y}|\mathbf{x})$ satisfying

$$f(\mathbf{y}) = \int_{\mathbf{x}} f(\mathbf{x}) P(\mathbf{y}|\mathbf{x}) d\mathbf{x}$$

  Can in general be a difficult criterion to check
- Sufficient criterion:

$$f(\mathbf{x}) P(\mathbf{y}|\mathbf{x}) = f(\mathbf{y}) P(\mathbf{x}|\mathbf{y}) \quad \text{Detailed balance}$$

  We then have

$$\begin{aligned} \int_{\mathbf{x}} f(\mathbf{x}) P(\mathbf{y}|\mathbf{x}) d\mathbf{x} &= \int_{\mathbf{x}} f(\mathbf{y}) P(\mathbf{x}|\mathbf{y}) d\mathbf{x} \\ &= f(\mathbf{y}) \int_{\mathbf{x}} P(\mathbf{x}|\mathbf{y}) d\mathbf{x} = f(\mathbf{y}) \end{aligned}$$

  since $P(\mathbf{x}|\mathbf{y})$ is, for any given $\mathbf{y}$, a density wrt $\mathbf{x}$.
- Note: For $\mathbf{y} = \mathbf{x}$, detailed balance always fulfilled, only necessary to check for $\mathbf{y} \neq \mathbf{x}$.

Metropolis-Hastings algorithms

- $P(\mathbf{y}|\mathbf{x})$ defined through an algorithm:
  1. Sample a candidate value $\mathbf{X}^*$ from a proposal distribution $g(\cdot|\mathbf{x})$.
  2. Compute the Metropolis-Hastings ratio
  $$R(\mathbf{x},\mathbf{X}^*) = \frac{f(\mathbf{X}^*)g(\mathbf{x}|\mathbf{X}^*)}{f(\mathbf{x})g(\mathbf{X}^*|\mathbf{x})}$$
  3. Put
  $$\mathbf{Y} = \begin{cases} \mathbf{X}^* & \text{with probability } \min\{1, R(\mathbf{x},\mathbf{X}^*)\} \\ \mathbf{x} & \text{otherwise} \end{cases}$$
- For $\mathbf{y} \neq \mathbf{x}$:
$$P(\mathbf{y}|\mathbf{x}) = g(\mathbf{y}|\mathbf{x}) \min\left\{1, \frac{f(\mathbf{y})g(\mathbf{x}|\mathbf{y})}{f(\mathbf{x})g(\mathbf{y}|\mathbf{x})}\right\}$$
- Note: $P(\mathbf{x}|\mathbf{x})$ somewhat difficult to evaluate in this case.
- Detailed balance (?)
$$\begin{aligned} f(\mathbf{x})P(\mathbf{y}|\mathbf{x}) =& f(\mathbf{x})g(\mathbf{y}|x) \min\left\{1, \frac{f(\mathbf{y})g(\mathbf{x}|\mathbf{y})}{f(\mathbf{x})g(\mathbf{y}|\mathbf{x})}\right\} \\ =& \min\{f(\mathbf{x})g(\mathbf{y}|\mathbf{x}), f(\mathbf{y})g(\mathbf{x}|\mathbf{y})\} \\ =& f(\mathbf{y})g(\mathbf{x}|\mathbf{y}) \min\left\{\frac{f(\mathbf{x})g(\mathbf{y}|\mathbf{x})}{f(\mathbf{y})g(\mathbf{x}|\mathbf{y})}, 1\right\} = f(\mathbf{y})P(\mathbf{x}|\mathbf{y}) \end{aligned}$$

M-H and unknown constant

- Assume now $f(\mathbf{x}) = c \cdot q(\mathbf{x})$ with c unknown.

$$R(\mathbf{x}, \mathbf{y}) = \frac{f(\mathbf{y})g(\mathbf{x}|\mathbf{y})}{f(\mathbf{x})g(\mathbf{y}|\mathbf{x})} = \frac{c \cdot q(\mathbf{y})g(\mathbf{x}|\mathbf{y})}{c \cdot q(\mathbf{x})g(\mathbf{y}|\mathbf{x})} = \frac{q(\mathbf{y})g(\mathbf{x}|\mathbf{y})}{q(\mathbf{x})g(\mathbf{y}|\mathbf{x})}$$

- Do not depend on c!

Random walk chains

- Popular choice of proposal distribution:

$$\mathbf{X}^* = \mathbf{x} + \boldsymbol{\varepsilon}$$

- $g(\mathbf{x}^*|\mathbf{x}) = h(\mathbf{x}^* - \mathbf{x})$
- Popular choices: Uniform, Gaussian, t-distribution
- Note: If $h(\cdot)$ is symmetric, $g(\mathbf{x}^*|\mathbf{x}) = g(\mathbf{x}|\mathbf{x}^*)$ and

$$R(\mathbf{x}, \mathbf{x}^*) = \frac{f(\mathbf{x}^*)g(\mathbf{x}|\mathbf{x}^*)}{f(\mathbf{x})g(\mathbf{x}^*|\mathbf{x})} = \frac{f(\mathbf{x}^*)}{f(\mathbf{x})}$$

Example

- Assume $f(x) \propto \exp(-|x|^3/3)$
- Proposal distribution $N(x, 1)$
- `Example_MH_cubic.R`

Independent chains

- Assume $g(\mathbf{x}^*|\mathbf{x}) = g(\mathbf{x}^*)$. Then

$$R(\mathbf{x}, \mathbf{x}^*) = \frac{f(\mathbf{x}^*)g(\mathbf{x})}{f(\mathbf{x})g(\mathbf{x}^*)} = \frac{\frac{f(\mathbf{x}^*)}{g(\mathbf{x}^*)}}{\frac{f(\mathbf{x})}{g(\mathbf{x})}},$$

  fraction of importance weights!
- Behave very much like importance sampling and SIR
- Difficult to specify $g(\mathbf{x})$ for high-dimensional problems
- Theoretical properties easier to evaluate than for random walk versions.

M-H and multivariate settings

- $\mathbf{X} = (X_1, ..., X_p)$
- Typical in this case: Only change one or a few components at a time.
  1. Choose index j (randomly)
  2. Sample $X_j^* \sim g_j(\cdot|\mathbf{x})$, put $X_k^* = X_k$ for $k \neq j$
  3. Compute

$$R(\mathbf{x}, \mathbf{X}^*) = \frac{f(\mathbf{X}^*)g(\mathbf{x}|\mathbf{X}^*)}{f(\mathbf{x})g(\mathbf{X}^*|\mathbf{x})}$$

  4. Put

$$\mathbf{Y} = \begin{cases} \mathbf{X}^* & \text{with probability } \min\{1, R(\mathbf{x}, \mathbf{X}^*)\} \\ \mathbf{x} & \text{otherwise} \end{cases}$$

- Can show that this version also satisfies detailed balance
- Can even go through indexes systematic
  - Should then consider the whole loop through all components as one iteration

Example

- Assume $f(\mathbf{x}) \propto \exp(-||\mathbf{x}||^3/3) = \exp(-[||\mathbf{x}||^2]^{3/2}/3) =$
- Proposal distribution
  1. $j \sim \text{Uniform}[1, 2, ..., p]$
  2. $x_j^* \sim N(x_j, 1)$
- `Example_MH_cubic_multivariate.R`

Reparametrization

- Sometimes easier to transform variables to another scale $Y = \bar{h}(X)$
- Two approaches (use $h^{-1}(\cdot) = \bar{h}(\cdot)$ so $X = h(Y)$)
  - Reparametrize $Y = \bar{h}(X)$, simulate from $f_Y(y)$ instead

$$
\begin{aligned}
f_Y(y) =& f_X(h(y))|h'(y)| \\
R(y, y^*) =& \frac{f_X(h(y^*))|h'(y^*)|g_y(y|y^*)}{f_X(h(y))|h'(y)|g_y(y^*|y)} \\
=& \frac{f_X(x^*)|h'(y^*)|g_y(y|y^*)}{f_X(x)|h'(y)|g_y(y^*|y)}
\end{aligned}
$$

  - Run the MCMC in X-space, but construct proposal through $X = h(Y)$,

$$
\begin{aligned}
g_x(x^*|x) =& g_y(\bar{h}(x^*)|\bar{h}(x))|\bar{h}'(x^*)| \\
R(x, x^*) =& \frac{f_X(x^*)g_y(\bar{h}(x)|\bar{h}(x^*))|\bar{h}'(x)|}{f_X(x)g_y(\bar{h}(x^*)|\bar{h}(x))|\bar{h}'(x^*)|} \\
=& \frac{f_X(x^*)g_y(y|y^*)|h'(y^*)|}{f_X(x)g_y(y^*|y)|h'(y)|}
\end{aligned}
$$

  since $\bar{h}'(x) = 1/h'(y)$

Gibbs sampling

- Assume $\mathbf{X} = (X_1, ..., X_p)$
- Aim: Simulate $\mathbf{X} \sim f(\mathbf{x})$
- Gibbs sampling:
  1. Select starting values $\mathbf{x}^{(0)}$ and set $t = 0$
  2. Generate, in turn

$$
\begin{aligned}
X_1^{(t+1)} &\sim f(x_1|x_2^{(t)}, x_3^{(t)}, ..., x_p^{(t)}) \\
X_2^{(t+1)} &\sim f(x_2|x_1^{(t+1)}, x_3^{(t)}, ..., x_p^{(t)}) \\
&\vdots \\
X_{p-1}^{(t+1)} &\sim f(x_{p-1}|x_1^{(t+1)}, ..., x_{p-2}^{(t+1)}, x_p^{(t)}) \\
X_p^{(t+1)} &\sim f(x_p|x_1^{(t+1)}, ..., x_{p-1}^{(t+1)})
\end{aligned}
$$

  3. Increment t and go to step 2.
- Completion of step 2 is called a cycle

Example - Mixture distribution

- Mixture distribution

$$Y \sim f(y) = \delta\phi(y, \mu_0, 0.5) + (1-\delta)\phi(y, \mu_1, 0.5), \quad \mu_0 = 7, \mu_1 = 10$$

- Prior $\delta \sim \text{Uniform}[0, 1]$
- Aim: Simulate $\delta \sim p(\delta | y_1, ..., y_n)$

$$p(\delta|y_1, ..., y_n) \propto \prod_{i=1}^{n} [\delta\phi(y_i, 7, 0.5) + (1-\delta)\phi(y_i, 10, 0.5)]$$

  Difficult to simulate from directly
- Note, can write model for Y by

$$\begin{aligned} \Pr(Z = z) =& \delta^{1-z}(1-\delta)^z, && z = 0, 1 \\ Y|Z = z \sim& \phi(y, \mu_z, 0.5), && \mu_0 = 7, \mu_1 = 10 \end{aligned}$$

- Note:

$$\begin{aligned} p(\delta|y_1, ..., y_n, z_1, ..., z_n) \propto& \prod_{i=1}^{n} \delta^{1-z_i}(1-\delta)^{z_i}\phi(y_i, \mu_{z_i}, 0.5) \\ \propto& \delta^{n-\sum_{i=1}^n z_i}(1-\delta)^{\sum_{i=1}^n z_i} \\ \propto& \text{Beta}(\delta, n - \sum_{i=1}^{n} z_i + 1, \sum_{i=1}^{n} z_i + 1) \end{aligned}$$

Example - continued

- Aim: Simulate $\delta \sim p(\delta|y_1, ..., y_n)$
- Approach: Simulate from $p(\delta, \mathbf{Z}|y_1, ..., y_n)$
- Gibbs sampling
  1. Initialize $\delta^{(0)}$, set $t = 0$
  2. Simulate $\mathbf{Z}^{(t+1)} \sim p(\mathbf{z}|\delta^{(t)}, \mathbf{y})$
  3. Simulate $\delta^{(t+1)} \sim p(\delta|\mathbf{z}^{(t+1)}, \mathbf{y})$
  4. Increment t and go to step 2.
- Conditional distribution for $\mathbf{z}$:

$$\begin{aligned} p(\mathbf{z}|\delta, \mathbf{y}) \propto & p(\delta)p(\mathbf{z}|\delta)p(\mathbf{y}|\mathbf{z}, \delta) \\ \propto & \prod_{i=1}^{n} \delta^{1-z_i}(1-\delta)^{z_i}\phi(y_i, \mu_{z_i}, 0.5) \end{aligned}$$

- Independence between z_i's:

$$\begin{aligned} \Pr(Z_i = z_i|\delta, y_i) \propto & \delta^{1-z_i}(1-\delta)^{z_i}\phi(y_i, \mu_{z_i}, 0.5) \\ \propto & \begin{cases} \frac{\delta\phi(y_i,\mu_0,0.5)}{\delta\phi(y_i,\mu_0,0.5)+(1-\delta)\phi(y_i,\mu_1,0.5)} & z_i = 0 \\ \frac{(1-\delta)\phi(y_i,\mu_1,0.5)}{\delta\phi(y_i,\mu_0,0.5)+(1-\delta)\phi(y_i,\mu_1,0.5)} & z_i = 1 \end{cases} \end{aligned}$$

- `Mixture_Gibbs_sampler.R`

Example - capture-recapture

- Aim: Estimate population size, N, of a species
- Procedure:
  - At time t_1: Catch $c_1 = m_1$ individuals, each with probability α_1.
    Mark and release
  - At time $t_i, i > 1$: Catch c_i individuals, each with probability α_i.
    Count number of newly caught individuals, m_i, mark the unmarked and release all
- Likelihood:
  - At time t_1:

$$\Pr(C_1 = c_1) = \Pr(C_1 = m_1) = \binom{N}{m_1}\alpha_1^{m_1}(1-\alpha_1)^{N-m_1}$$

  - At time $t_i, i > 1$ (number of marked individuals are $\sum_{k=1}^{i-1} m_k$)

$$\begin{aligned}\Pr(C_i =&c_i, M_j = m_i|N, \mathbf{c}_{1:i-1}, \mathbf{m}_{1:i-1})\\ &= \Pr(C_i = c_i|N)\Pr(M_i = m_i|N, C_i = c_i, \mathbf{m}_{1:i-1})\\ &=\binom{N}{c_i}\alpha_i^{c_i}(1-\alpha_i)^{N-c_i}\frac{\binom{N-\sum_{k=1}^{i-1} m_k}{m_i}\binom{\sum_{k=1}^{i-1} m_k}{c_i - m_i}}{\binom{N}{c_i}}\\ &=\alpha_i^{c_i}(1-\alpha_i)^{N-c_i}\binom{N-\sum_{k=1}^{i-1} m_k}{m_i}\binom{\sum_{k=1}^{i-1} m_k}{c_i - m_i}\end{aligned}$$

- Likelihood:

$$
\begin{aligned}
L(N,\boldsymbol{\alpha}|\mathbf{c},\mathbf{m}) \propto& \binom{N}{m_1}\alpha_1^{m_1}(1-\alpha_1)^{N-m_1}\times\\
&\prod_{i=2}^{I}\alpha_j^{c_i}(1-\alpha_i)^{N-c_i}\binom{N-\sum_{k=1}^{i-1}m_k}{m_i}\binom{\sum_{k=1}^{i-1}m_k}{c_i-m_i}\\
\propto&\prod_{i=1}^{I}\alpha_i^{c_i}(1-\alpha_i)^{N-c_i}\binom{N-\sum_{k=1}^{i-1}m_k}{m_i}\\
\propto&\binom{N}{\sum_{k=1}^{I}m_k}\prod_{i=1}^{I}\alpha_i^{c_i}(1-\alpha_j)^{N-c_j}
\end{aligned}
$$

- Prior:

$$
\begin{aligned}
f(N)\propto&1\\
f(\alpha_i|\theta_1,\theta_2)\sim&\text{Beta}(\theta_1,\theta_2)
\end{aligned}
$$

- Can derive ($r=\sum_{k=1}^{I}m_k$):

$$
\begin{aligned}
N|\boldsymbol{\alpha},\mathbf{c},\mathbf{m}\sim& r+\texttt{NegBinom}(r+1,1-\prod_{i=1}^{I}(1-\alpha_i))\\
\alpha_i|N,\boldsymbol{\alpha}_{-i},\mathbf{c},\mathbf{m}\sim&\text{Beta}(c_i+\theta_1,N-c_i+\theta_2)
\end{aligned}
$$

- `Example_7_6.R`

Properties of Gibbs sampler (random scan)

- Gibbs sampling (random scan):
  1. Select starting values $\mathbf{x}^{(0)}$ and set $t = 0$
  2. Sample $j \sim \text{Uniform}\{1, ..., p\}$
  3. Sample $X_j^{(t+1)} \sim f(x_j|\mathbf{x}_{-j}^{(t)})$
  4. Put $X_k^{(t+1)} = X_k^{(t)}$ for $k \neq j$
- The chain $\{\mathbf{X}^{(t)}\}$ is Markov
- Detailed balance:
  - Consider $\mathbf{x}, \mathbf{x}^*$ where $x_j \neq x_j^*$ while $x_k = x_k^*$ for $k \neq j$

$$
\begin{aligned}
f(\mathbf{x})P(\mathbf{x}^*|\mathbf{x}) =& f(\mathbf{x})p^{-1}f(x_j^*|\mathbf{x}_{-j})\\
=& f(\mathbf{x}_{-j})f(x_j|\mathbf{x}_{-j})p^{-1}f(x_j^*|\mathbf{x}_{-j})\\
=& f(\mathbf{x}^*_{-j})f(x_j|\mathbf{x}^*_{-j})p^{-1}f(x_j^*|\mathbf{x}^*_{-j})\\
=& f(\mathbf{x}^*)p^{-1}f(x_j|\mathbf{x}^*_{-j})\\
=& f(\mathbf{x}^*)P(\mathbf{x}|\mathbf{x}^*)
\end{aligned}
$$

Properties of Gibbs sampler (deterministic scan)

- Gibbs sampling (deterministic scan):
  1. Select starting values $\mathbf{x}^{(0)}$ and set $t = 0$
  2. Generate, in turn

$$
\begin{aligned}
X_1^{(t+1)} &\sim f(x_1|x_2^{(t)}, x_3^{(t)}, ..., x_p^{(t)})\\
X_2^{(t+1)} &\sim f(x_2|x_1^{(t+1)}, x_3^{(t)}, ..., x_p^{(t)})\\
&\vdots\\
X_{p-1}^{(t+1)} &\sim f(x_{p-1}|x_1^{(t+1)}, ..., x_{p-2}^{(t+1)}, x_p^{(t)})\\
X_p^{(t+1)} &\sim f(x_p|x_1^{(t+1)}, ..., x_{p-1}^{(t+1)})
\end{aligned}
$$

  3. Increment t and go to step 2.
- The chain $\{\mathbf{X}^{(t)}\}$ is Markov
- Do not fulfill detailed balance (going backwards will revert order of components visited)
- Will still satisfy

$$f(\mathbf{x}^*) = \int_{\mathbf{x}} f(\mathbf{x})P(\mathbf{x}^*|\mathbf{x})d\mathbf{x}$$

Proof

- Assume $p = 2$: $P(\mathbf{x}^*|\mathbf{x}) = f(x_1^*|x_2)f(x_2^*|x_1^*)$:

$$\begin{aligned}
\int_{\mathbf{x}} f(\mathbf{x})P(\mathbf{x}^*|\mathbf{x})d\mathbf{x} =& \int_{x_2}\int_{x_1} f(\mathbf{x})f(x_1^*|x_2)f(x_2^*|x_1^*)dx_1dx_2\\
=& \int_{x_2}\int_{x_1} f(x_1|x_2)f(x_2)f(x_1^*|x_2)f(x_2^*|x_1^*)dx_1dx_2\\
=& \int_{x_2}\int_{x_1} f(x_1|x_2)f(x_2|x_1^*)f(x_1^*)f(x_2^*|x_1^*)dx_1dx_2\\
=& f(x_1^*,x_2^*)\int_{x_2} f(x_2|x_1^*)\int_{x_1} f(x_1|x_2)dx_1dx_2\\
=& f(x_1^*,x_2^*)\int_{x_2} f(x_2|x_1^*)dx_2\\
=& f(x_1^*,x_2^*) = f(\mathbf{x}^*)
\end{aligned}$$

- Proof similar for general p

Tuning the Gibbs sampler

- Random or deterministic scan?
  - Deterministic scan most common (?)
  - When high correlation, random scan can be more efficient
- Blocking:
  - When dividing $\mathbf{X} = (X_1, ..., X_p)$, each X_j can be vectors
  - Making each X_j as large as possible will typically improve convergence
  - Especially beneficial when high correlation between single components
- Hybrid Gibbs sampling
  - If $f(x_j|\mathbf{x}_{-j})$ is difficult to sample from, use an Metropolis-Hastings step for this component
  - Example ($p = 5$)
    1. Sample $X_1^{(t+1)} \sim f(x_1|\mathbf{x}_{-1}^{(t)})$
    2. Sample $(X_2^{(t+1)}, X_3^{(t+1)})$ through an M-H step
    3. Sample $X_4^{(t+1)}$ through another M.H step
    4. Sample $X_5^{(t+1)} \sim f(x_5|\mathbf{x}_{-5}^{(t+1)})$

Capture-recapture - extended approach

- Assume now a prior $f(\theta_1, \theta_2) \propto \exp\{-(\theta_1 + \theta_2)/1000\}$
- Conditional distributions:

$$N|\cdot \sim r + \texttt{NegBinom}(r+1, 1 - \prod_{i=1}^{I}(1-\alpha_j))$$

$$\alpha_i|\cdot \sim \text{Beta}(c_i + \theta_1, N - c_i + \theta_2)$$

$$(\theta_1, \theta_2)|\cdot \sim k \left[\frac{\Gamma(\theta_1+\theta_2)}{\Gamma(\theta_1)\Gamma(\theta_2)}\right]^I \prod_{i=1}^{I} \alpha_i^{\theta_1}(1-\alpha_i)^{\theta_2} \exp\left\{-\frac{\theta_1+\theta_2}{1000}\right\}$$

- `Example_7_7.R`

Convergence issues of MCMC

- Theoretical properties:

$$\mathbf{X}^{(t)} \xrightarrow{\mathcal{D}} f(\mathbf{x}), \quad \text{as } t \to \infty$$

$$\hat{\theta}_1 = \frac{1}{L}\sum_{t=1}^{L} h(\mathbf{X}^{(t)}) \to E^f[h(\mathbf{X})] \quad \text{as } L \to \infty$$

- Note: We also have

$$\hat{\theta}_2 = \frac{1}{L}\sum_{t=D+1}^{D+L} h(\mathbf{X}^{(t)}) \to E^f[h(\mathbf{X})] \quad \text{as } L \to \infty$$

  - Advantage: Remove those variables with distribution very different from $f(\mathbf{x})$
  - Disadvantage: Need more samples
- Question: How to specify D and L?
  - D: Large enough so that $\mathbf{X}^{(t)} \approx f(\mathbf{x})$ for $t > D$ (bias small)
  - L: Large enough so that $\text{Var}[\hat{\theta}_2]$ is small enough

Mixing

- For $\hat{\theta} = \frac{1}{L}\sum_{t=D+1}^{D+L} h(\mathbf{X}^{(t)})$:

$$\text{Var}[\hat{\theta}] = \frac{1}{L^2}\left[\sum_{t=D+1}^{D+L}\text{Var}[h(\mathbf{X}^{(t)})] + 2\sum_{s=D+1}^{D+L-1}\sum_{t=s+1}^{D+L}\text{Cov}[h(\mathbf{X}^{(s)}), h(\mathbf{X}^{(t)})]\right]$$

Assume D large, so "converged":

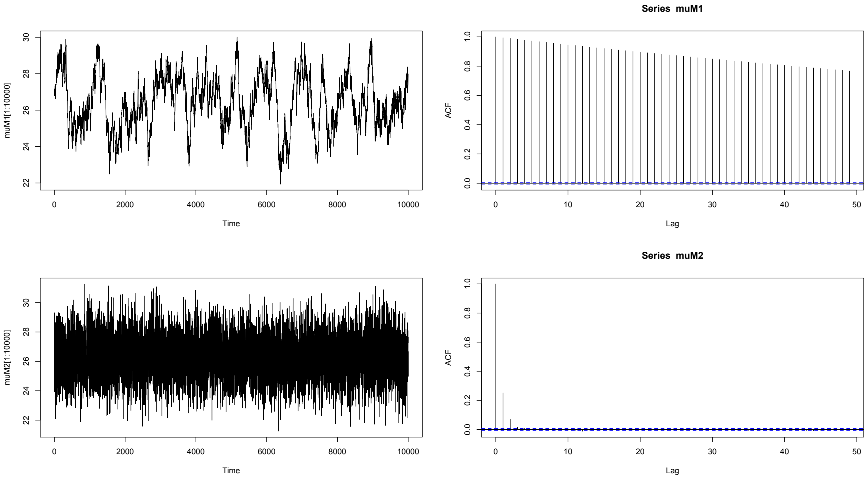
$$\text{Var}[h(\mathbf{X}^{(t)})] \approx \sigma_h^2, \quad \text{Cov}[h(\mathbf{X}^{(s)}), h(\mathbf{X}^{(t)})] \approx \sigma_h^2\rho(t-s)$$

gives

$$\begin{aligned}\text{Var}[\hat{\theta}] \approx& \frac{1}{L^2}\left[\sum_{t=D+1}^{D+L}\sigma_h^2 + 2\sum_{s=D+1}^{D+L-1}\sum_{t=s+1}^{D+L}\sigma_h^2\rho(t-s)\right]\\ =& \frac{\sigma_h^2}{L}\left[1 + 2\sum_{k=1}^{L-1}\frac{L-k}{L}\rho(k)\right]\end{aligned}$$

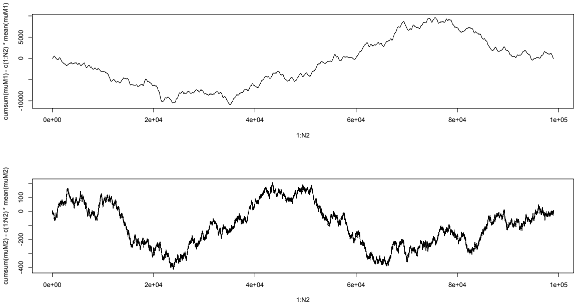
- Good mixing: $\rho(k)$ decreases fast with k!

Example from Exercise 7.8



How to assess convergence?

- Graphical diagnostics:
  - Sample paths:
    - Plot $h(\mathbf{X}^{(t)})$ as function of t
    - Useful with different $h(\cdot)$ functions!
  - Cusum diagnostics
    - Plot $\sum_{i=1}^{t}[h(\mathbf{X}^{(i)}) - \hat{\theta}_n]$ versus t
    - Wiggly and small excursions from 0: Indicate chain is mixing well



The Gelman-Rubin diagnostic

- Motivated from analysis of variance
- Assume J chains run in parallel
- jth chain: $x_j^{(D+1)}, ..., x_j^{(D+L)}$ (first D discarded)
- Define

$$\bar{x}_j = \frac{1}{L}\sum_{t=D+1}^{D+L} x_j^{(t)} \qquad \bar{x}_\cdot = \frac{1}{J}\sum_{j=1}^{J}\bar{x}_j$$

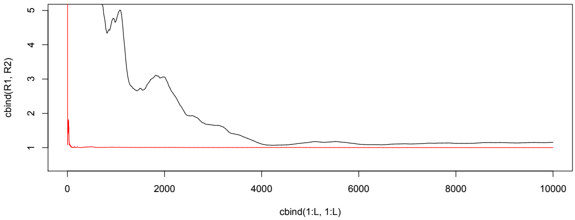
$$B = \frac{L}{J-1}\sum_{j=1}^{J}(\bar{x}_j - \bar{x}_\cdot)^2$$

$$W = \frac{1}{J}\sum_{j=1}^{J} s_j^2 \qquad s_j^2 = \frac{1}{L-1}\sum_{t=D+1}^{D+L}(x_j^{(t)} - \bar{x}_j)^2$$

- If converged, both B and W estimates $\sigma^2 = \text{Var}_f[X]$
- Diagnostic: $R = \frac{\frac{L-1}{L}W + \frac{1}{L}B}{W}$
- "Rule": $\sqrt{R} < 1.1$ indicate D and L are sufficient

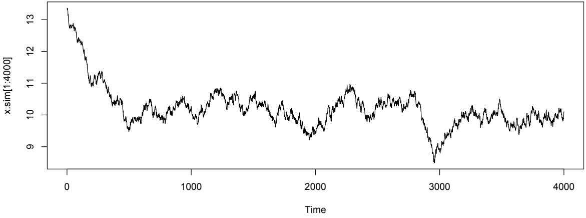
Example: Exercise 7.8

- $D = 100$, $L = 1000$: $\sqrt{R_1} = 1.588$, $\sqrt{R_2} = 1.002$,
- $D = 1000$, $L = 1000$: $\sqrt{R_1} = 1.700$, $\sqrt{R_2} = 1.004$,
- $D = 1000$, $L = 10000$: $\sqrt{R_1} = 1.049$,$\sqrt{R_2} = 1.0008$

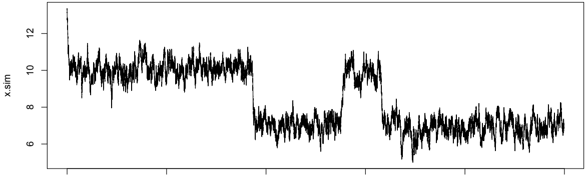


Apparent convergence

- $f(x) = 0.7 \cdot N(7, 0.5^2) + 0.3 \cdot N(10, 0.5^2)$
- Metropolis-Hastings with proposal $N(x^{(t)}, 0.05^2)$
- First 4000 samples (400 discarded)



- Full 10000 samples



M-H: Choice of proposal distribution

- Independence chain:
  - $g(\cdot) \approx f(\cdot)$
  - High acceptance rate
  - Tail properties most important: f/g should be bounded
- Random walk proposal
  - Tune variance so that acceptance rate is between 25 and 50%

Effective sample size for MCMC

- For $\hat{\theta} = \frac{1}{L}\sum_{t=D+1}^{D+L} h(\mathbf{X}^{(t)})$:

$$\text{Var}[\hat{\theta}] = \frac{\sigma_h^2}{L}\left[1 + 2\sum_{k=1}^{L-1}\frac{L-k}{L}\rho(k)\right] \stackrel{L\to\infty}{\to} \frac{\sigma_h^2}{L}[1 + 2\sum_{k=1}^{\infty}\rho(k)]$$

- If independent samples:

$$\text{Var}[\hat{\theta}] = \frac{\sigma_h^2}{L}$$

- Effective sample size: $\frac{L}{1+2\sum_{k=1}^{\infty}\rho(k)}$
- Use empirical estimates $\hat{\rho}(k)$
- Usual to truncate the summation when $\hat{\rho}(k) < 0.1$.

Number of chains

- Assume possible to perform N iterations
  - One long chain of length N, or
  - J parallel chains, each of length N/J?
- Burnin:
  - One long chain: Only need to discard D samples
  - Parallel chains: Need to discard $J \cdot D$ samples
- Check of convergence
  - Easier with many parallel chains
- Efficiency
  - Parallel chains give more independent samples
- Computational issues
  - Possible to utilize multiple cores with parallel chains

Data uncertainty and Monte Carlo uncertainty

- Parameter: $\theta = E^f[h(\mathbf{X})]$
- Estimator: $\hat{\theta} = \frac{1}{L}\sum_{t=D+1}^{D+L} h(\mathbf{X}^{(t)})$:
- Two types of uncertainty
  - Variability in $h(\mathbf{X})$: $\sigma_h^2 = \text{Var}^f[h(\mathbf{X})]$
    - Estimator: $\hat{\sigma}_h^2 = \frac{1}{L}\sum_{t=D+1}^{D+L}[h(\mathbf{X}^{(t)}) - \hat{\theta}]^2$
  - MC variability in $\hat{\theta}$:
    - Estimator: Divide data into batches of size $b = \lfloor L^{1/a} \rfloor$, make estimates $\hat{\theta}$ within each batch and variance from these
- Recommendation: Specify L so that MC variability is less than 5% of variability in $h(\mathbf{X})$.

- Adaptive MCMC: Automatic tuning of proposal distributions
  - Main challenge: Specifying proposal based on history of chain breaks down the Markov property
  - Solution: Reduce the amount of tuning as the number of iterations increases
- Reversible Jump MCMC
  - Assume several models $\mathcal{M}_1, ..., \mathcal{M}_K$
  - Corresponding parameters $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K$ of different dimensions!
  - Aim: Simulate $\mathbf{X} = (\mathcal{M}, \boldsymbol{\theta}_{\mathcal{M}})$
  - RJMCMC: M-H method for moving between spaces of different dimensions
  - Main challenge: When changing $\mathcal{M} \to \mathcal{M}^*$, how to propose $\boldsymbol{\theta}_{\mathcal{M}^*}$?
- Simulated tempering
  - Define $f^i(\mathbf{x}) \propto f(\mathbf{x})^{1/\tau_i}$, $1 = \tau_1 < \tau_2 < \cdots < \tau_m$
  - Simulate $(\mathbf{X}, I)$, where I changes distribution
  - Easier to move around when $\tau_i > 1$
  - Keep samples for which $I = 1$
- Multiple-Try M-H
  - Generate k proposals $\mathbf{X}_1^*, ..., \mathbf{X}_k^*$ from $g(\cdot|\mathbf{x}^{(t)})$
  - Select $\mathbf{X}_j^*$ with probability $w(\mathbf{x}^{(t)}, \mathbf{X}_j^*) = f(\mathbf{x}^{(t)})g(\mathbf{X}_j^*|\mathbf{x}^{(t)})\lambda(\mathbf{x}^{(t)}, \mathbf{X}_j^*)$, λ symmetric
  - Sample $\mathbf{X}_1^{**}, ..., \mathbf{X}_{k-1}^{**}$ from $g(\cdot|\mathbf{X}_j^*)$, put $\mathbf{X}_k^{**} = \mathbf{x}^{(t)}$
  - Use Generalized M-H ratio

$$R_g = \frac{\sum_{i=1}^k w(\mathbf{x}^{(t)}, \mathbf{X}_i^*)}{\sum_{i=1}^k w(\mathbf{X}_j^*, \mathbf{X}_i^{**})}$$

Hamiltonian MC

- Common trick in Monte Carlo: Introduce auxiliary variables
- Hamiltonian MC (Neal et al., 2011):

$$\begin{aligned}
\pi(\boldsymbol{q}) &\propto \exp(-U(\boldsymbol{q})) && \text{Distribution of interest}\\
\pi(\boldsymbol{q},\boldsymbol{p}) &\propto \exp(-U(\boldsymbol{q}) - 0.5\boldsymbol{p}^T\boldsymbol{p}) && \text{Extended distribution}\\
&= \exp(-H(\boldsymbol{q},\boldsymbol{p})) && H(\boldsymbol{q},\boldsymbol{p}) = U(\boldsymbol{q}) + 0.5\boldsymbol{p}^T\boldsymbol{p}
\end{aligned}$$

- Note
  - $\boldsymbol{q}$ and $\boldsymbol{p}$ are independent
  - $\boldsymbol{p} \sim N(\mathbf{0}, \boldsymbol{I})$.
  - Usually dim($\boldsymbol{p}$)= dim($\boldsymbol{q}$)
- Algorithm ($\boldsymbol{q}$) current value
  1. Simulate $\boldsymbol{p} \sim N(\mathbf{0}, \boldsymbol{I})$
  2. Generate $(\boldsymbol{q}^*, \boldsymbol{p}^*)$ such that $H(\boldsymbol{q}^*, \boldsymbol{p}^*) \approx H(\boldsymbol{q}, \boldsymbol{p})$
  3. Accept $(\boldsymbol{q}^*, \boldsymbol{p}^*)$ by a Metropolis-Hastings step
- Main challenge: Generate $(\boldsymbol{q}^*, \boldsymbol{p}^*)$

Hamiltonian dynamics

- Consider $(\boldsymbol{q}, \boldsymbol{p})$ as a time-process $(\boldsymbol{q}(t), \boldsymbol{p}(t))$
- Hamiltonian dynamics: Change through

$$
\begin{aligned}
\frac{dq_i}{dt} &= \frac{\partial H}{\partial p_i} \\
\frac{dp_i}{dt} &= -\frac{\partial H}{\partial q_i}
\end{aligned}
$$

  This gives

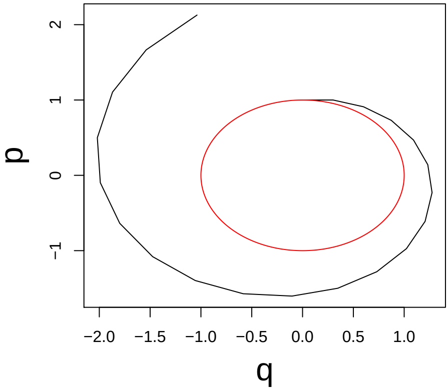
$$
\begin{aligned}
\frac{dH}{dt} &= \sum_{i=1}^{d} \left[\frac{\partial H}{\partial q_i} \frac{dq_i}{dt} + \frac{\partial H}{\partial p_i} \frac{dp_i}{dt} \right] \\
&= \sum_{i=1}^{d} \left[\frac{\partial H}{\partial q_i} \frac{\partial H}{\partial p_i} - \frac{\partial H}{\partial p_i} \frac{\partial H}{\partial q_i} \right] = 0
\end{aligned}
$$

- If we can change $(\boldsymbol{q}, \boldsymbol{p})$ exactly by the Hamiltonian dynamics, H will not change!
- In practice, only possible to make numerical approximations

- Assume

$$\begin{aligned}
p_i(t+\varepsilon) =& p_i(t) + \varepsilon \frac{dp_i}{dt}(t) \\
=& p_i(t) - \varepsilon \frac{\partial U}{\partial q_i}(q_i(t)) \\
q_i(t+\varepsilon) =& q_i(t) + \varepsilon \frac{dq_i}{dt}(t) \\
=& q_i(t) + \varepsilon p_i(t)
\end{aligned}$$

- Note: Derivatives of $U(\boldsymbol{q})$ are used.
- However, not very exact.

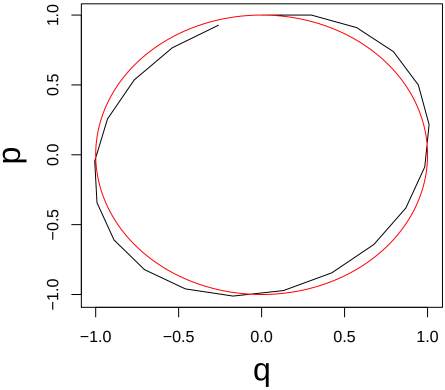


Hamiltonian dynamics - the modified Eulers method

- Assume

$$p_i(t+\varepsilon) = p_i(t) - \varepsilon \frac{\partial U}{\partial q_i}(q(t))$$
$$q_i(t+\varepsilon) = q_i(t) + \varepsilon p_i(t+\varepsilon)$$

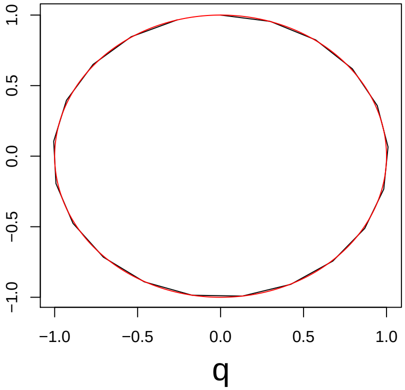
- Better than Eulers method.



Hamiltonian dynamics - the Leapfrog method

- Assume

$$
\begin{aligned}
p_i(t+\tfrac{\varepsilon}{2}) =& p_i(t) - \frac{\varepsilon}{2}\frac{\partial U}{\partial q_i}(q(t))\\
q_i(t+\varepsilon) =& q_i(t) + \varepsilon p_i(t+\tfrac{\varepsilon}{2})\\
p_i(t+\varepsilon) =& p_i(t+\tfrac{\varepsilon}{2}) - \frac{\varepsilon}{2}\frac{\partial U}{\partial q_i}(q_(t+\varepsilon))
\end{aligned}
$$



- Quite exact!
- Idea: Use this L steps

Example - 2-dimensional Gaussian

- Assume $\boldsymbol{x} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.95 \\ 0.95 & 1 \end{pmatrix}$
- $H(\boldsymbol{x}, \boldsymbol{p}) = 0.5\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + 0.5\boldsymbol{p}^T\boldsymbol{p}$
- Use $L = 5$ leapfrog steps, with stepsize $\varepsilon = 0.1$
- `leapfrog_Gauss2.R`

Example - mixture Gaussians

- Assume

$$\pi(x) = pN(x; \mu_1, \sigma_1^2) + (1 - p)N(x; \mu_2, \sigma_2^2)$$

- $H(\boldsymbol{x}, \boldsymbol{p}) = -log(\pi(x) + 0.5\boldsymbol{p}^T\boldsymbol{p}$
- Use $L = 5$ leapfrog steps, with stepsize $\varepsilon = 0.1$
- `leapfrog_mixture.R`

R. M. Neal et al. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2, 2011.