Basics of Markov Chain Monte Carlo

version: 2018-03-10 · 10:59:22

Motivation

- **Bayesian Inference:**
  - *Posterior Distribution:* $p(\boldsymbol{\theta} \mid \mathbf{y}) \propto \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y}) \times \pi(\boldsymbol{\theta})$, with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$.
  - *Quantity of Interest:* $\tau = g(\boldsymbol{\theta})$.
  - *Point/Interval Estimate:*

$$\hat{\tau} = E[\tau \mid \mathbf{y}] = \int g(\tau) p(\boldsymbol{\theta} \mid \mathbf{y}) \, \mathrm{d}\boldsymbol{\theta}$$

$$\mathrm{CI}_{95}(\tau) = \left(F^{-1}_{\tau \mid \mathbf{y}}(2.5\% \mid \mathbf{y}), F^{-1}_{\tau \mid \mathbf{y}}(97.5\% \mid \mathbf{y}) \right)$$

- **Deterministic Calculation:** Multidimensional integral and Inverse-CDF are typically very difficult for $d > 2$. (any grid method scales terribly with d)

Markov Chain Monte Carlo (MCMC)

Problem: Let

$$\tau = g(\mathbf{x}), \qquad \mathbf{x} = (x_1, \ldots, x_d) \sim p(\mathbf{x}).$$

Compute $E[\tau]$ and $F_\tau^{-1}(\alpha)$.

- **Deterministic calculation:** Typically very difficult for $d > 2$.
- **Monte Carlo:** If we can sample $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)} \overset{\text{iid}}{\sim} p(\mathbf{x})$, then
  - *Point Estimate:* $\bar{\tau} = \dfrac{1}{M} \displaystyle\sum_{m=1}^{M} g(\mathbf{x}^{(m)}) \to \tau.$
  - *Interval Estimate:* Let $\tau^{(m)} = g(\mathbf{x}^{(m)})$ and $\tau^{(1:M)} = (\tau^{(1)}, \ldots, \tau^{(M)})$. Then

$$\hat{q}_\tau(\alpha) = \mathtt{quantile}(\mathbf{x}^{(1:M)}, \mathtt{prob} = \alpha) \to F_\tau^{-1}(\alpha).$$

Markov Chain Monte Carlo (MCMC)

- **Problem:** Let $\tau = g(\mathbf{x})$, $\mathbf{x} \sim p(\mathbf{x})$. Compute $E[\tau]$.
- **Monte Carlo:** If we can sample $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)} \overset{\text{iid}}{\sim} p(\mathbf{x})$, then

$$\bar{\tau} = \frac{1}{M} \sum_{m=1}^{M} \tau^{(m)} \to E[\tau].$$

- **Markov Chain:** Drawing $\mathbf{x}^{(m)} \overset{\text{iid}}{\sim} p(\mathbf{x})$ typically very difficult for $d > 2$.

  Instead, sample from a Markov chain $\mathbf{x}^{(m)} \sim \mathsf{T}(\mathbf{x} \mid \mathbf{x}^{(m-1)})$ for which the stationary distribution is $p(\mathbf{x})$.

  Still have $\bar{\tau} \to E[\tau]$, but usually $\text{var}(\bar{\tau}_{\text{iid}}) < \text{var}(\bar{\tau}_{\text{mcmc}})$.

Markov Chain Monte Carlo

- **Problem:** Let $\tau = g(\mathbf{x})$, $\mathbf{x} \sim p(\mathbf{x})$. Compute $E[\tau]$.
- **MCMC:**
  - Sample from a Markov chain $\mathbf{x}^{(m)} \sim \mathrm{T}(\mathbf{x} \mid \mathbf{x}^{(m-1)})$ for which the stationary distribution is $p(\mathbf{x})$.
  - Calculate $\bar{\tau} = \frac{1}{M} \sum_{m=1}^{M} g(\mathbf{x}^{(m)}) \to E[\tau]$
- **Transition density:** How to pick $\mathrm{T}(\mathbf{x} \mid \mathbf{x}')$?

  Two fundamental concepts:

1. **REDUCE**: only sample parts of $\mathbf{x}$ at a time (Gibbs sampler)
2. **APPROX**: don't try to sample perfectly, as many approximate sampling schemes can be perfectly corrected (Metropolis-Hastings algorithm)

Gibbs Sampler

- **Problem:** sample $\mathbf{x} \sim p(\mathbf{x})$
- **Suppose** we know how to sample from $p(x_i \,|\, \mathbf{x}_{-i})$ for every $1 \leq i \leq d$.

---

input: $\mathbf{x}^{(0)}$ ▷ Starting value

for $m = 1, \ldots, M$ **do**
  $\tilde{\mathbf{x}} \leftarrow \mathbf{x}^{(m)}$
  for $i = 1, \ldots, d$ **do**
    $\tilde{x}_i \sim p(x_i \,|\, \tilde{x}_{-i})$ ▷ Update each rv conditioned on all others
  end for
  $\mathbf{x}^{(m+1)} \leftarrow \tilde{\mathbf{x}}$
end for

output: $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}$

Example: Bivariate Normal

- **Model:**

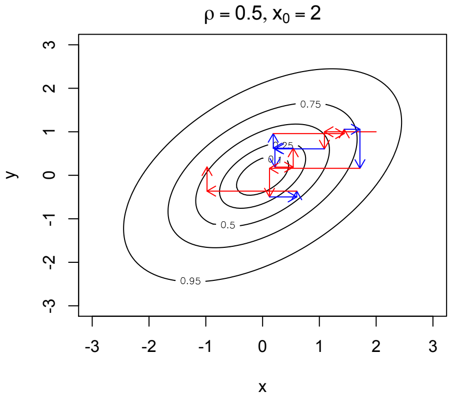
$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}_2 \left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \sigma_x \sigma_y \rho \\ \sigma_x \sigma_y \rho & \sigma_y^2 \end{bmatrix} \right).$$

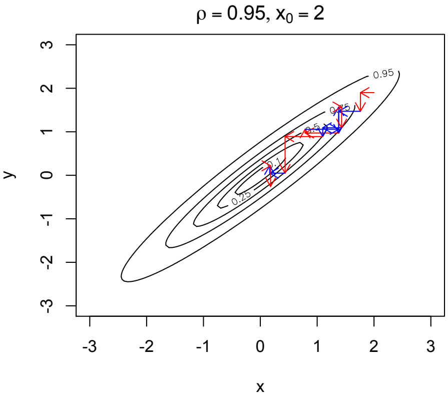
- **Conditional Distributions:**

$$\begin{aligned} x \,|\, y &\sim \mathcal{N}\left(\mu_x + \rho \frac{\sigma_x}{\sigma_y} \times (y - \mu_y), (1 - \rho^2)\sigma_x^2 \right) \\ y \,|\, x &\sim \mathcal{N}\left(\mu_y + \rho \frac{\sigma_y}{\sigma_x} \times (x - \mu_x), (1 - \rho^2)\sigma_y^2 \right). \end{aligned}$$

Example: Bivariate Normal

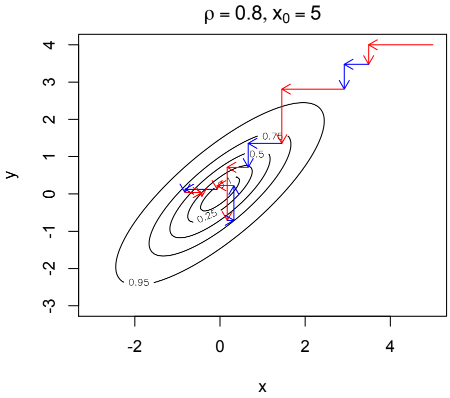
$$\textbf{Model: } \begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right), \qquad \textbf{Starting Point: } x_0.$$

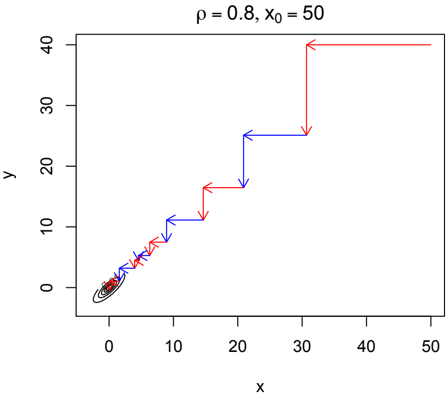




Example: Bivariate Normal

$$\textbf{Model: } \begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right), \qquad \textbf{Starting Point: } x_0.$$





Gibbs Sampler (Continued)

- **Summary:** Cycle through conditional updates $x_i \sim p(\mathbf{x}_{-i})$. Can do these in any order, even random.
- **Limitations:**
  - Convergence is slow when $\mathrm{cor}(x_i, \mathbf{x}_{-i}) \to 1$.
  - Convergence is slow for poorly-chosen initial value $\mathbf{x}^{(0)}$
  - Must be able to sample for each conditional $p(x_i \mid \mathbf{x}_{-i})$.

Metropolis-Hastings Algorithm

- Gibbs sampler requires you to be able to draw from each $p(x_i \mid \mathbf{x}_{-i})$.
- What if $p(x_i \mid \mathbf{x}_{-i})$ is not easy to draw from?
- M-H algorithm requires only a transition density $\mathsf{T}(\mathbf{x} \mid \mathbf{x}')$ for which:

  1. You can draw $\mathbf{x} \sim \mathsf{T}(\mathbf{x} \mid \mathbf{x}')$

  2. You have a closed-form PDF (or PMF) for $\mathsf{T}(\mathbf{x} \mid \mathbf{x}')$

Metropolis-Hastings Algorithm

input: $\mathbf{x}^{(0)}$, $\mathrm{T}(\mathbf{x} \mid \mathbf{x}')$ ▷ Starting value, transition density

for $m = 1, \ldots, M$ **do**

$\mathbf{x}_{\text{curr}} \leftarrow \mathbf{x}^{(m)}$

$\mathbf{x}_{\text{prop}} \sim \mathrm{T}(\mathbf{x} \mid \mathbf{x}_{\text{curr}})$ ▷ Proposal

$\alpha \leftarrow \min\left\{1, \dfrac{p(\mathbf{x}_{\text{prop}}) / \mathrm{T}(\mathbf{x}_{\text{prop}} \mid \mathbf{x}_{\text{curr}})}{p(\mathbf{x}_{\text{curr}}) / \mathrm{T}(\mathbf{x}_{\text{curr}} \mid \mathbf{x}_{\text{prop}})}\right\}$ ▷ Acceptance probability

$U \sim \mathrm{Unif}(0, 1)$

if $U < \alpha$ **then**

$\mathbf{x}^{(m+1)} \leftarrow \mathbf{x}_{\text{prop}}$ ▷ Keep proposal with probability α

else

$\mathbf{x}^{(m+1)} \leftarrow \mathbf{x}_{\text{curr}}$ ▷ Reject proposal with probability $1 - \alpha$

end if

end for

output: $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}$

Metropolis-Hastings Algorithm

- **Algorithm Summary:**
  1. Draw $\mathbf{x}_{\text{prop}} \sim \mathrm{T}(\mathbf{x} \mid \mathbf{x}_{\text{curr}} = \mathbf{x}^{(m)})$
  2. Let $\alpha = \min\left\{1, \dfrac{p(\mathbf{x}_{\text{prop}}) / \mathrm{T}(\mathbf{x}_{\text{prop}} \mid \mathbf{x}_{\text{curr}})}{p(\mathbf{x}_{\text{curr}}) / \mathrm{T}(\mathbf{x}_{\text{curr}} \mid \mathbf{x}_{\text{prop}})}\right\}$
  3. Set $\mathbf{x}^{(m+1)}$ to $\mathbf{x}_{\text{prop}}$ with probability α, to $\mathbf{x}_{\text{curr}}$ with probability $1 - \alpha$
- Requires only a transition density $\mathrm{T}(\mathbf{x} \mid \mathbf{x}')$ for which:
  1. You can draw $\mathbf{x} \sim \mathrm{T}(\mathbf{x} \mid \mathbf{x}')$
  2. You have a closed-form PDF (or PMF) for $\mathrm{T}(\mathbf{x} \mid \mathbf{x}')$
- Only need $r(\mathbf{x}) = p(\mathbf{x})/Z$, where Z is unknown (since $p(\mathbf{x}_{\text{prop}})/p(\mathbf{x}_{\text{curr}}) = r(\mathbf{x}_{\text{prop}})/r(\mathbf{x}_{\text{curr}})$).

  Critical for Bayesian inference, in which case only know $p(\boldsymbol{\theta} \mid \mathbf{y}) \propto \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y})\pi(\boldsymbol{\theta})$

Metropolis-Hastings Algorithm

- **Algorithm Summary:**
  1. Draw $\mathbf{x}_{\text{prop}} \sim \mathrm{T}(\mathbf{x} \mid \mathbf{x}_{\text{curr}} = \mathbf{x}^{(m)})$
  2. Let $\alpha = \min\left\{1, \dfrac{p(\mathbf{x}_{\text{prop}}) / \mathrm{T}(\mathbf{x}_{\text{prop}} \mid \mathbf{x}_{\text{curr}})}{p(\mathbf{x}_{\text{curr}}) / \mathrm{T}(\mathbf{x}_{\text{curr}} \mid \mathbf{x}_{\text{prop}})}\right\}$
  3. Set $\mathbf{x}^{(m+1)}$ to $\mathbf{x}_{\text{prop}}$ with probability α, to $\mathbf{x}_{\text{curr}}$ with probability $1 - \alpha$
- **Transition Density:** Most common choices:
  1. **Random Walk Metropolis:** $\mathbf{x}_{\text{prop}} \sim \mathcal{N}\big(\mathbf{x}_{\text{curr}}, \operatorname{diag}(\boldsymbol{\sigma}^2_{\text{tune}})\big)$. Let $f(\mathbf{x})$ denote the PDF of $\mathcal{N}(\mathbf{0}, \operatorname{diag}(\boldsymbol{\sigma}^2_{\text{tune}}))$. Then

$$\mathrm{T}(\mathbf{x}_{\text{prop}} \mid \mathbf{x}_{\text{curr}}) = f(\mathbf{x}_{\text{prop}} - \mathbf{x}_{\text{curr}}) = f(\mathbf{x}_{\text{curr}} - \mathbf{x}_{\text{prop}}) = \mathrm{T}(\mathbf{x}_{\text{curr}} \mid \mathbf{x}_{\text{prop}}).$$

     Thus, the transition density is symmetric $\implies \alpha = \min\{1, p(\mathbf{x}_{\text{prop}})/p(\mathbf{x}_{\text{curr}})\}$.

Metropolis-Hastings Algorithm

- **Algorithm Summary:**

  1. Draw $\mathbf{x}_{\text{prop}} \sim \mathrm{T}(\mathbf{x} \mid \mathbf{x}_{\text{curr}} = \mathbf{x}^{(m)})$
  2. Let $\alpha = \min\left\{1, \dfrac{p(\mathbf{x}_{\text{prop}})/\,\mathrm{T}(\mathbf{x}_{\text{prop}} \mid \mathbf{x}_{\text{curr}})}{p(\mathbf{x}_{\text{curr}})/\,\mathrm{T}(\mathbf{x}_{\text{curr}} \mid \mathbf{x}_{\text{prop}})}\right\}$
  3. Set $\mathbf{x}^{(m+1)}$ to $\mathbf{x}_{\text{prop}}$ with probability α, to $\mathbf{x}_{\text{curr}}$ with probability $1-\alpha$

- **Transition Density:** Most common choices:

  1. **Random Walk Metropolis:** $\mathbf{x}_{\text{prop}} \sim \mathcal{N}\big(\mathbf{x}_{\text{curr}}, \operatorname{diag}(\boldsymbol{\sigma}^2_{\text{tune}})\big).$
  2. **Metropolis-Within-Gibbs:**

---

for $m = 1, \ldots, M$ **do**
  $\mathbf{x}_{\text{curr}}, \mathbf{x}_{\text{prop}} \leftarrow \mathbf{x}^{(m)}$
  for $j = 1, \ldots, d$ **do** ▷ Gibbs loop
    $x_{j,\text{prop}} \sim \mathcal{N}(x_{j,\text{curr}} \mid \sigma^2_{j,\text{tune}})$ ▷ Metropolis step within conditional proposal
    $\alpha \leftarrow \min\{1, p(\mathbf{x}_{\text{prop}})/p(\mathbf{x}_{\text{curr}})\}$ ▷ Symmetric proposal, and note that $\mathbf{x}_{-j,\text{prop}} = \mathbf{x}_{-j,\text{curr}}$
    if `runif(1)` $< \alpha$ **then** $x_{j,\text{curr}} \leftarrow x_{j,\text{prop}}$ ▷ $\implies p(\mathbf{x}_{\text{prop}})/p(\mathbf{x}_{\text{curr}}) = p(x_{j,\text{prop}} \mid \mathbf{x}_{-j,\text{prop}})/p(x_{j,\text{curr}} \mid \mathbf{x}_{-j,\text{curr}})$
    else $x_{j,\text{prop}} \leftarrow x_{j\text{curr}}$ **end if**
  end for
  $\mathbf{x}^{(m+1)} \leftarrow \mathbf{x}_{\text{curr}}$ ▷ Storage at the end of update cycle
end for

Metropolis-Hastings Algorithm

- **Algorithm Summary:**

  1. Draw $\mathbf{x}_{\text{prop}} \sim \mathsf{T}(\mathbf{x} \mid \mathbf{x}_{\text{curr}} = \mathbf{x}^{(m)})$
  2. Let $\alpha = \min\left\{1, \dfrac{p(\mathbf{x}_{\text{prop}}) / \mathsf{T}(\mathbf{x}_{\text{prop}} \mid \mathbf{x}_{\text{curr}})}{p(\mathbf{x}_{\text{curr}}) / \mathsf{T}(\mathbf{x}_{\text{curr}} \mid \mathbf{x}_{\text{prop}})}\right\}$
  3. Set $\mathbf{x}^{(m+1)}$ to $\mathbf{x}_{\text{prop}}$ with probability α, to $\mathbf{x}_{\text{curr}}$ with probability $1 - \alpha$

- **Transition Density:** Most common choices:

  1. **Random Walk Metropolis:** $\mathbf{x}_{\text{prop}} \sim \mathcal{N}\big(\mathbf{x}_{\text{curr}}, \operatorname{diag}(\boldsymbol{\sigma}^2_{\text{tune}})\big)$.
  2. **Metropolis-Within-Gibbs:** $x_{j,\text{prop}} \sim \mathcal{N}(x_{j,\text{curr}}, \sigma^2_{j,\text{tune}}), \quad j = 1, \ldots, d$.

     Like a Gibbs sampler, but each update is RWM if $p(x_j \mid \mathbf{x}_{-j})$ can't be drawn from directly.

Metropolis-Hastings Algorithm

- **Algorithm Summary:**

  1. Draw $\mathbf{x}_{\text{prop}} \sim \mathrm{T}(\mathbf{x} \mid \mathbf{x}_{\text{curr}} = \mathbf{x}^{(m)})$

  2. Let $\alpha = \min\left\{1, \dfrac{p(\mathbf{x}_{\text{prop}})/\,\mathrm{T}(\mathbf{x}_{\text{prop}} \mid \mathbf{x}_{\text{curr}})}{p(\mathbf{x}_{\text{curr}})/\,\mathrm{T}(\mathbf{x}_{\text{curr}} \mid \mathbf{x}_{\text{prop}})}\right\}$

  3. Set $\mathbf{x}^{(m+1)}$ to $\mathbf{x}_{\text{prop}}$ with probability α, to $\mathbf{x}_{\text{curr}}$ with probability $1 - \alpha$

- **Transition Density:** Most common choices:

  1. Random Walk Metropolis: $\mathbf{x}_{\text{prop}} \sim \mathcal{N}\big(\mathbf{x}_{\text{curr}}, \mathrm{diag}(\boldsymbol{\sigma}^2_{\text{tune}})\big).$

  2. Metropolis-Within-Gibbs: $x_{j,\text{prop}} \sim \mathcal{N}(x_{j,\text{curr}}, \sigma^2_{j,\text{tune}}), \quad j = 1, \ldots, d.$

  3. Metropolized IID: $\mathbf{x}_{\text{prop}} \overset{\text{iid}}{\sim} q(\mathbf{x}).$

  Typically this is "mode-quadrature" proposal $\mathcal{N}(\hat{\mathbf{x}}, -[\frac{\partial^2}{\partial \mathbf{x}^2} \log p(\hat{\mathbf{x}})]^{-1})$, where $\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} p(\mathbf{x})$.

Metropolis-Hastings Algorithm

- **Algorithm Summary:**

  1. Draw $\mathbf{x}_{\text{prop}} \sim \mathrm{T}(\mathbf{x} \mid \mathbf{x}_{\text{curr}} = \mathbf{x}^{(m)})$

  2. Let $\alpha = \min\left\{1, \dfrac{p(\mathbf{x}_{\text{prop}})/\,\mathrm{T}(\mathbf{x}_{\text{prop}} \mid \mathbf{x}_{\text{curr}})}{p(\mathbf{x}_{\text{curr}})/\,\mathrm{T}(\mathbf{x}_{\text{curr}} \mid \mathbf{x}_{\text{prop}})}\right\}$

  3. Set $\mathbf{x}^{(m+1)}$ to $\mathbf{x}_{\text{prop}}$ with probability α, to $\mathbf{x}_{\text{curr}}$ with probability $1 - \alpha$

  Why does it work?

- **Theorem:** Suppose that $\mathbf{x}^{(m)}$ is drawn from $p(\mathbf{x})$, and $\mathbf{x}^{(m+1)}$ is an MH update, i.e.,

$$
\begin{aligned}
\mathbf{x}^{(m)} &\sim p(\mathbf{x}) \\
\mathbf{x}^{(m+1)} \mid \mathbf{x}^{(m)} &\sim \mathrm{MH}\{\mathrm{T}, \mathbf{x}^{(m)}\} \\
&= \alpha \cdot \mathrm{T}(\mathbf{x} \mid \mathbf{x}^{(m)}) + (1 - \alpha) \cdot \mathbb{1}\{\mathbf{x} = \mathbf{x}^{(m)}\}.
\end{aligned}
$$

  Then the marginal distribution of $\mathbf{x}^{(m+1)} \sim p(\mathbf{x})$. In other words, the MH algorithm generates a Markov chain with stationary distribution $p(\mathbf{x})$.

Metropolis-Hastings Algorithm

- **Algorithm Summary:**
  1. Draw $\mathbf{x}_{\text{prop}} \sim \mathrm{T}(\mathbf{x} \mid \mathbf{x}_{\text{curr}} = \mathbf{x}^{(m)})$
  2. Let $\alpha = \min \left\{ 1, \dfrac{p(\mathbf{x}_{\text{prop}}) / \mathrm{T}(\mathbf{x}_{\text{prop}} \mid \mathbf{x}_{\text{curr}})}{p(\mathbf{x}_{\text{curr}}) / \mathrm{T}(\mathbf{x}_{\text{curr}} \mid \mathbf{x}_{\text{prop}})} \right\}$
  3. Set $\mathbf{x}^{(m+1)}$ to $\mathbf{x}_{\text{prop}}$ with probability α, to $\mathbf{x}_{\text{curr}}$ with probability $1 - \alpha$
- **Theorem:**

$$\begin{aligned} \mathbf{x}^{(m)} &\sim p(\mathbf{x}) \\ \mathbf{x}^{(m+1)} \mid \mathbf{x}^{(m)} &\sim \mathrm{MH}\{\mathrm{T}, \mathbf{x}^{(m)}\} \end{aligned} \qquad \Longrightarrow \qquad \mathbf{x}^{(m+1)} \sim p(\mathbf{x}).$$

- **Proof:** Consider $\mathbf{x}_a$ and $\mathbf{x}_b$ such that $\alpha = \frac{p(\mathbf{x}_a) / \mathrm{T}(\mathbf{x}_a \mid \mathbf{x}_b)}{p(\mathbf{x}_b) / \mathrm{T}(\mathbf{x}_b \mid \mathbf{x}_a)} < 1$.
  1. Joint distribution of a then b: (proposal automatically accepted)

$$p(\mathbf{x}^{(m)} = \mathbf{x}_a, \mathbf{x}^{(m+1)} = \mathbf{x}_b) = p(\mathbf{x}_a) \cdot \mathrm{T}(\mathbf{x}_b \mid \mathbf{x}_a).$$

Metropolis-Hastings Algorithm

- **Algorithm Summary:**

  1. Draw $\mathbf{x}_{\text{prop}} \sim \mathrm{T}(\mathbf{x} \mid \mathbf{x}_{\text{curr}} = \mathbf{x}^{(m)})$
  2. Let $\alpha = \min\left\{1, \dfrac{p(\mathbf{x}_{\text{prop}}) / \mathrm{T}(\mathbf{x}_{\text{prop}} \mid \mathbf{x}_{\text{curr}})}{p(\mathbf{x}_{\text{curr}}) / \mathrm{T}(\mathbf{x}_{\text{curr}} \mid \mathbf{x}_{\text{prop}})}\right\}$
  3. Set $\mathbf{x}^{(m+1)}$ to $\mathbf{x}_{\text{prop}}$ with probability α, to $\mathbf{x}_{\text{curr}}$ with probability $1 - \alpha$

- **Theorem:**

$$\left.\begin{aligned} \mathbf{x}^{(m)} &\sim p(\mathbf{x}) \\ \mathbf{x}^{(m+1)} \mid \mathbf{x}^{(m)} &\sim \mathrm{MH}\{\mathrm{T}, \mathbf{x}^{(m)}\} \end{aligned}\right. \quad \Longrightarrow \quad \mathbf{x}^{(m+1)} \sim p(\mathbf{x}).$$

- **Proof:** Consider $\mathbf{x}_a$ and $\mathbf{x}_b$ such that $\alpha = \frac{p(\mathbf{x}_a) / \mathrm{T}(\mathbf{x}_a \mid \mathbf{x}_b)}{p(\mathbf{x}_b) / \mathrm{T}(\mathbf{x}_b \mid \mathbf{x}_a)} < 1$.

  1. Joint distribution of a then b: $p(\mathbf{x}^{(m)} = \mathbf{x}_a, \mathbf{x}^{(m+1)} = \mathbf{x}_b) = p(\mathbf{x}_a) \cdot \mathrm{T}(\mathbf{x}_b \mid \mathbf{x}_a)$.
  2. Joint distribution of b then a: (proposal accepted with probability α)

$$\begin{aligned} p(\mathbf{x}^{(m)} = \mathbf{x}_b, \mathbf{x}^{(m+1)} = \mathbf{x}_a) &= p(\mathbf{x}_b) \cdot \mathrm{T}(\mathbf{x}_a \mid \mathbf{x}_b) \cdot \frac{p(\mathbf{x}_a) / \mathrm{T}(\mathbf{x}_a \mid \mathbf{x}_b)}{p(\mathbf{x}_b) / \mathrm{T}(\mathbf{x}_b \mid \mathbf{x}_a)} \\ &= p(\mathbf{x}_a) \cdot \mathrm{T}(\mathbf{x}_b \mid \mathbf{x}_a). \end{aligned}$$

Metropolis-Hastings Algorithm

- **Algorithm Summary:**

  1. Draw $\mathbf{x}_{\text{prop}} \sim \mathrm{T}(\mathbf{x} \mid \mathbf{x}_{\text{curr}} = \mathbf{x}^{(m)})$
  2. Let $\alpha = \min\left\{1, \dfrac{p(\mathbf{x}_{\text{prop}})/\ \mathrm{T}(\mathbf{x}_{\text{prop}} \mid \mathbf{x}_{\text{curr}})}{p(\mathbf{x}_{\text{curr}})/\ \mathrm{T}(\mathbf{x}_{\text{curr}} \mid \mathbf{x}_{\text{prop}})}\right\}$
  3. Set $\mathbf{x}^{(m+1)}$ to $\mathbf{x}_{\text{prop}}$ with probability α, to $\mathbf{x}_{\text{curr}}$ with probability $1-\alpha$

- **Theorem:**

$$\begin{rcases} \mathbf{x}^{(m)} \sim p(\mathbf{x}) \\ \mathbf{x}^{(m+1)} \mid \mathbf{x}^{(m)} \sim \mathrm{MH}\{\mathrm{T}, \mathbf{x}^{(m)}\} \end{rcases} \implies \mathbf{x}^{(m+1)} \sim p(\mathbf{x}).$$

- **Proof:** Consider $\mathbf{x}_a$ and $\mathbf{x}_b$ such that $\alpha = \frac{p(\mathbf{x}_a)/\ \mathrm{T}(\mathbf{x}_a \mid \mathbf{x}_b)}{p(\mathbf{x}_b)/\ \mathrm{T}(\mathbf{x}_b \mid \mathbf{x}_a)} < 1$.

  1. Joint distribution of a then b: $p(\mathbf{x}^{(m)} = \mathbf{x}_a, \mathbf{x}^{(m+1)} = \mathbf{x}_b) = p(\mathbf{x}_a) \cdot \mathrm{T}(\mathbf{x}_b \mid \mathbf{x}_a)$.
  2. Joint distribution of b then a: $p(\mathbf{x}^{(m)} = \mathbf{x}_b, \mathbf{x}^{(m+1)} = \mathbf{x}_a) = p(\mathbf{x}_a) \cdot \mathrm{T}(\mathbf{x}_b \mid \mathbf{x}_a)$.

$$\implies p(\mathbf{x}^{(m)} = \mathbf{x}_a, \mathbf{x}^{(m+1)} = \mathbf{x}_b) = p(\mathbf{x}^{(m)} = \mathbf{x}_b, \mathbf{x}^{(m+1)} = \mathbf{x}_a).$$

  Since joint distribution is *symmetric*, each marginal must be *identical*

$$\implies p(\mathbf{x}^{(m+1)}) = p(\mathbf{x}^{(m)}) = p(\mathbf{x}).$$

Example: Weibull Distribution

Definition: If $X \sim \text{Expo}(1)$, then

$$Y = \lambda X^{\gamma} \sim \text{Weibull}(\gamma, \lambda).$$

The PDF of Y is

$$f(y) = \frac{\gamma}{\lambda}\left(\frac{y}{\lambda}\right)^{\gamma-1} e^{-(y/\lambda)^{\gamma}}, \qquad y > 0.$$

Weibull Distribution

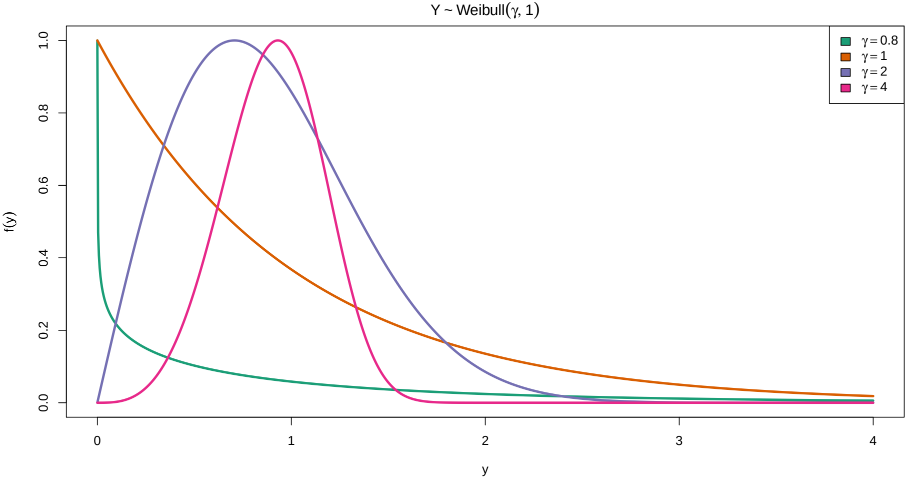
- **Model:** $Y \sim \text{Weibull}(\gamma, \lambda) \quad \iff \quad Y = \lambda X^\gamma,\ X \sim \text{Expo}(1).$
- **Utility:** Survival analysis
  - *Hazard function:* $\approx$ probability of failing in next instant:

$$h(y) = \lim_{\Delta y \to 0} \frac{\Pr(Y < y + \Delta y \,|\, Y > y)}{\Delta y} = \frac{f(y)}{1 - F(y)}$$

  - $h(y)$ characterizes distribution, just like $f(y)$ or $F(y)$
- **Weibull Hazard:** $h(y) = (\frac{\gamma}{\lambda^\gamma}) \cdot y^{\gamma - 1}$
  - $\gamma = 1 \implies h(y) = \text{const} \implies Y \sim \lambda \cdot \text{Expo}(1)$
    memoriless property (chance of failing constant through time)
  - $\gamma > 1 \implies h(y)$ increasing
    Ex: elderly patients more and more likely to die soon as they get older
  - $g < 1 \implies h(y)$ decreasing
    Ex: infants more and more likely to survive longer as they get older

Weibull Distribution

- **Model:** $Y \sim \text{Weibull}(\gamma, \lambda) \quad \iff \quad Y = \lambda X^{\gamma},\ X \sim \text{Expo}(1).$
- **Hazard Function:** $h(y) \propto y^{\gamma - 1}$



Weibull Distribution

- **Model:** $Y \sim \text{Weibull}(\gamma, \lambda) \quad \iff \quad Y = \lambda X^\gamma,\ X \sim \text{Expo}(1).$
- **Likelihood:** $\mathbf{y} = (y_1, \ldots, y_n) \overset{\text{iid}}{\sim} \text{Weibull}(\gamma, \lambda)$

$$\ell(\gamma, \lambda \mid \mathbf{y}) = n\big[\log(\gamma) - \gamma \log(\lambda)\big] + \sum_{i=1}^n \gamma \log(y_i) - \lambda^{-\gamma} \sum_{i=1}^n y_i^\gamma.$$

Not an Exponential Family (because of y_i^γ).

Weibull Distribution

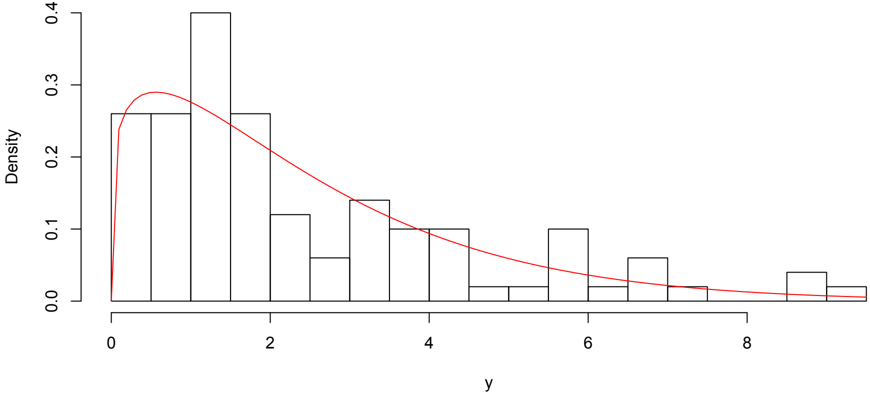
- **Model:** $Y \sim \text{Weibull}(\gamma, \lambda) \quad \iff \quad Y = \lambda X^{\gamma},\ X \sim \text{Expo}(1).$
- **Likelihood:** $\mathbf{y} = (y_1, \ldots, y_n) \overset{\text{iid}}{\sim} \text{Weibull}(\gamma, \lambda)$

$$\ell(\gamma, \lambda \mid \mathbf{y}) = n\big[\log(\gamma) - \gamma \log(\lambda)\big] + \gamma \sum_{i=1}^{n} \log(y_i) - \lambda^{-\gamma} \sum_{i=1}^{n} y_i^{\gamma}.$$

Not an Exponential Family (because of y_i^{γ}).

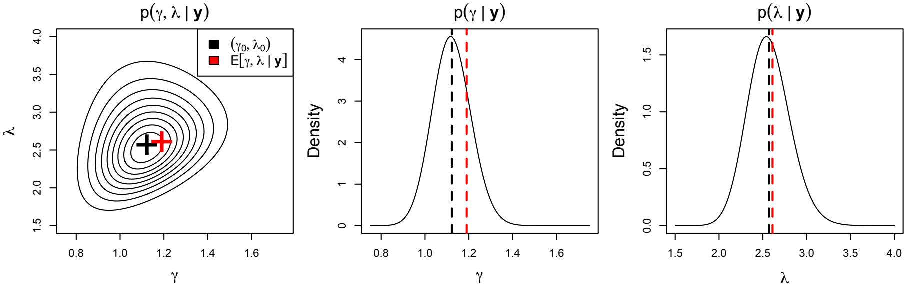
Weibull Distribution

- **Model:** $Y \sim \text{Weibull}(\gamma, \lambda) \quad \iff \quad Y = \lambda X^{\gamma},\ X \sim \text{Expo}(1)$.
- **Likelihood:** $\ell(\gamma, \lambda \mid \mathbf{y}) = n\big[\log(\gamma) - \gamma\log(\lambda)\big] + \sum_{i=1}^n \big[\gamma\log(y_i) - \lambda^{-\gamma} y_i^{\gamma}\big]$.
- **Simulated Data:** $\gamma = 1.19, \lambda = 2.61, n = 100$



Weibull Distribution

- **Model:** $Y \sim \text{Weibull}(\gamma, \lambda) \quad \iff \quad Y = \lambda X^{\gamma}, \ X \sim \text{Expo}(1).$
- **Likelihood:** $\ell(\gamma, \lambda \mid \mathbf{y}) = n\big[\log(\gamma) - \gamma \log(\lambda)\big] + \sum_{i=1}^{n} \big[\gamma \log(y_i) - \lambda^{-\gamma} y_i^{\gamma}\big].$
- **Prior:** $\pi(\gamma, \lambda) \propto 1$ (hopefully won't make much difference)
- **Posterior:** For 2-d problem can compute $p(\gamma, \lambda \mid \mathbf{y})$ on a grid



Weibull Distribution

- **Model:** $Y \sim \text{Weibull}(\gamma, \lambda) \quad \iff \quad Y = \lambda X^{\gamma},\ X \sim \text{Expo}(1).$
- **Likelihood:** $\ell(\gamma, \lambda \mid \mathbf{y}) = n\big[\log(\gamma) - \gamma\log(\lambda)\big] + \sum_{i=1}^{n}\big[\gamma\log(y_i) - \lambda^{-\gamma} y_i^{\gamma}\big].$
- **Prior:** $\pi(\gamma, \lambda) \propto 1$
- **Posterior:** For 2-d problem can compute $p(\gamma, \lambda \mid \mathbf{y})$ on a grid,

  OR MCMC on $\boldsymbol{\theta} = (\gamma, \lambda)$:

  1. **Random-Walk Metropolis:** $\boldsymbol{\theta}_{\text{prop}} \sim \mathcal{N}\big(\boldsymbol{\theta}_{\text{curr}}, \operatorname{diag}(\boldsymbol{\sigma}^2_{\text{RW}})\big).$
  2. **Metropolis-Within-Gibbs:** $\theta_{j,\text{prop}} \sim \mathcal{N}(\theta_{j,\text{curr}}, \sigma^2_{j,\text{RW}}), \quad j = 1, 2.$
  3. **Metropolized IID:** $\boldsymbol{\theta}_{\text{prop}} \overset{\text{iid}}{\sim} \mathcal{N}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}), \qquad \hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta} \mid \mathbf{y})$

     $$\hat{\boldsymbol{\Sigma}} = -\left[\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \log p(\hat{\boldsymbol{\theta}} \mid \mathbf{y})\right]^{-1}$$

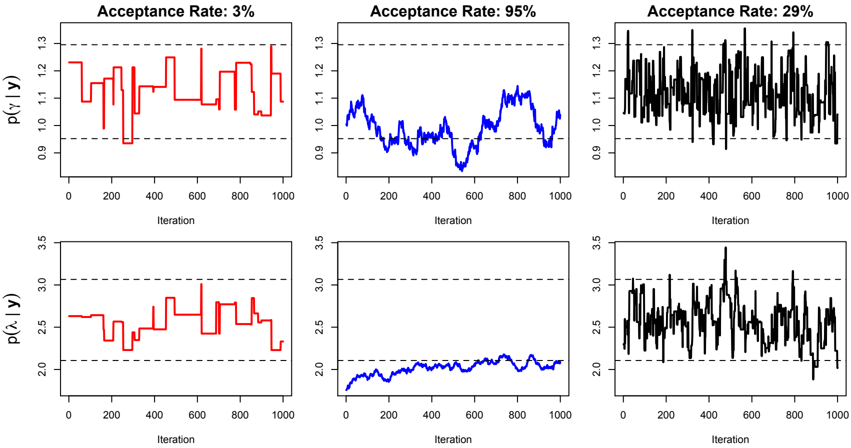
Random-Walk Metropolis (RWM)

- **Transition Density:** $\boldsymbol{\theta}_{\text{prop}} \sim \mathcal{N}\big(\boldsymbol{\theta}_{\text{curr}}, \text{diag}(\boldsymbol{\sigma}^2_{\text{RW}})\big)$
- **Tuning Parameters:** coordinate-wise "jump size" $\boldsymbol{\sigma}_{\text{RW}}$.
- **Question:** How to pick $\boldsymbol{\sigma}_{\text{RW}}$?

Random-Walk Metropolis (RWM)

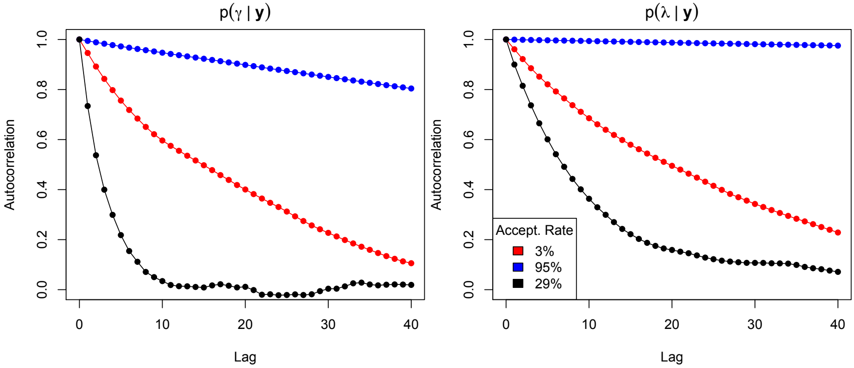
- **Transition Density:** $\theta_{\text{prop}} \sim \mathcal{N}\left(\boldsymbol{\theta}_{\text{curr}}, \text{diag}(\boldsymbol{\sigma}^2_{\text{RW}})\right)$
- **Tuning Parameters:** coordinate-wise "jump size" $\boldsymbol{\sigma}_{\text{RW}}$.

  "Optimal" acceptance rate: $\approx 25\%$.



MCMC Diagnostics

1. **Trace Plot:** Time series of MCMC output $\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(M)}$
2. **Autocorrelation Plot:** Ideally would have $\boldsymbol{\theta}^{(m)} \overset{\text{iid}}{\sim} p(\boldsymbol{\theta} \mid \mathbf{y})$, but instead draws are correlated.



MCMC Diagnostics

1. **Trace Plot:** Time series of MCMC output $\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(M)}$
2. **Autocorrelation Plot:** Ideally would have $\boldsymbol{\theta}^{(m)} \overset{\text{iid}}{\sim} p(\boldsymbol{\theta} \mid \mathbf{y})$, but instead draws are correlated
3. **Effective Sample Size:** For given $\tau = g(\boldsymbol{\theta})$, M draws from MCMC are roughly equivalent to $\mathsf{ESS}(\tau)$ iid draws, where

$$\mathsf{ESS}(\tau) = \frac{M}{1 + 2 \times \sum_{t=1}^{\infty} \gamma_t}, \qquad \gamma_t = \mathrm{cor}(\tau^{(m)}, \tau^{(m+t)}).$$

   That is, if $\hat{\tau}_{\mathsf{MCMC}}$ and $\hat{\tau}_{\mathsf{IID}}$ are sample means of M draws from MCMC and IID sampler, then

$$\frac{\mathrm{var}(\hat{\tau}_{\mathsf{IID}})}{\mathrm{var}(\hat{\tau}_{\mathsf{MCMC}})} \approx \frac{1}{1 + 2 \times \sum_{t=1}^{\infty} \gamma_t}.$$

MCMC Diagnostics

1. **Trace Plot:** Time series of MCMC output $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)}$
2. **Autocorrelation Plot:** Ideally would have $\boldsymbol{\theta}^{(m)} \overset{\text{iid}}{\sim} p(\boldsymbol{\theta} \mid \mathbf{y})$, but instead draws are correlated
3. **Effective Sample Size:** For given $\tau = g(\boldsymbol{\theta})$, M draws from MCMC are roughly equivalent to $\mathsf{ESS}(\tau)$ iid draws, where

$$\mathsf{ESS}(\tau) = \frac{M}{1 + 2 \times \sum_{t=1}^{\infty} \gamma_t}, \qquad \gamma_t = \mathrm{cor}(\tau^{(m)}, \tau^{(m+t)}).$$

Weibull example for $M = 10,000$:

	Accept. Rate 3%	95%	29%
γ	286	137	1518
λ	235	125	460

Metropolis-Within-Gibbs (MWG)

- **Transition Density:** $\theta_{\mathsf{prop},j} \sim \mathcal{N}(\theta_{\mathsf{curr},j}, \sigma^2_{\mathsf{RW},j})$

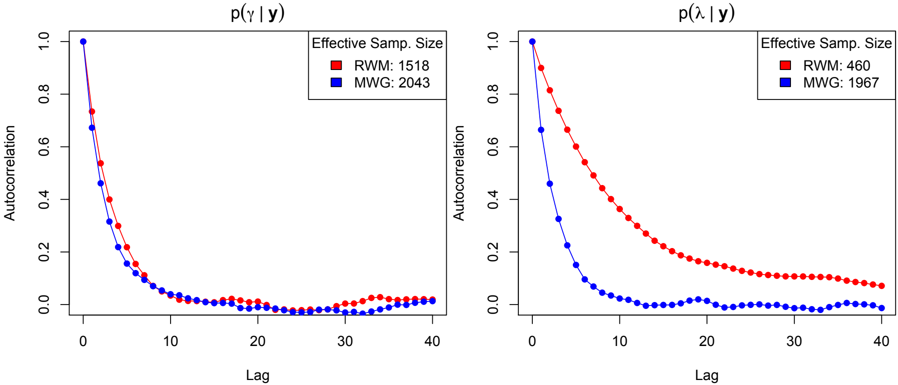
  Contrast with RWM, which proposes all of $\boldsymbol{\theta}$ at once.

- **Tuning Parameters:** "Optimal" coordinate-wise acceptance rate $\approx 45\%$.

  Contrast with RMW, for which optimal acceptance rate $\approx 25\%$.

Metropolis-Within-Gibbs (MWG)

- **Transition Density:** $\theta_{\text{prop},j} \sim \mathcal{N}(\theta_{\text{curr},j}, \sigma^2_{\text{RW},j})$
- **Tuning Parameters:** "Optimal" coordinate-wise acceptance rate $\approx 45\%$.



Metropolized IID Sampler (MIID)

- **Transition Density:**

$$\boldsymbol{\theta}_{\text{prop}} \overset{\text{iid}}{\sim} \mathcal{N}\left(\hat{\boldsymbol{\theta}}, -\left[\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \log p(\hat{\boldsymbol{\theta}} \mid \mathbf{y})\right]^{-1}\right), \qquad \hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta} \mid \mathbf{y}).$$

- **Optimal acceptance rate:**

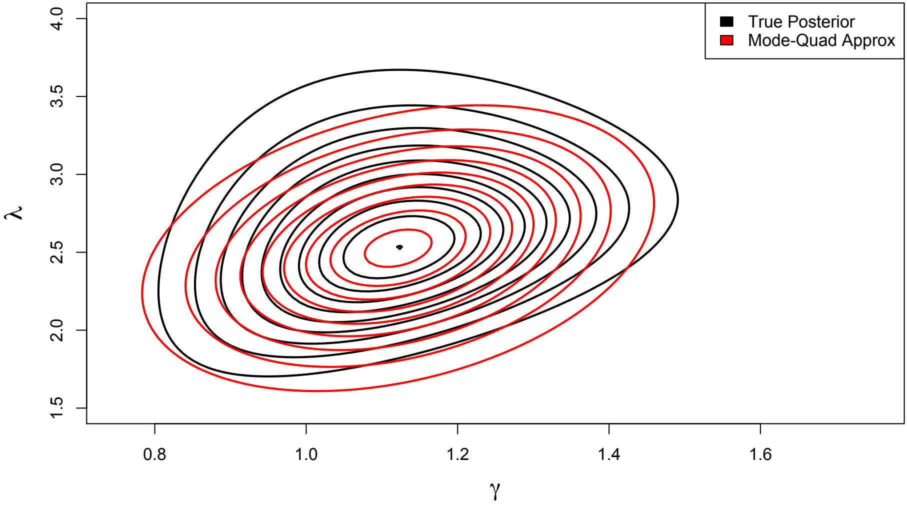
Metropolized IID Sampler (MIID)

- **Transition Density:**

$$\boldsymbol{\theta}_{\text{prop}} \overset{\text{iid}}{\sim} \mathcal{N}\left(\hat{\boldsymbol{\theta}}, -\left[\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \log p(\hat{\boldsymbol{\theta}} \mid \mathbf{y})\right]^{-1}\right), \qquad \hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta} \mid \mathbf{y}).$$

- **Optimal acceptance rate:** 100%!
  - MIID has no tuning parameters: no need to tune (good), but also stuck with whatever acceptance rate the proposals have (bad).
  - Since proposals are IID, all of $p(\boldsymbol{\theta}_{\text{prop}} \mid \mathbf{y})$ and $q(\boldsymbol{\theta}_{\text{prop}})$ can be precomputed before entering the MCMC $\implies$ can parallelize these calculations, and write a generic and lightweight MIID sampler directly in **R** (see `miid.sampler` in `mcmc-functions.R` on LEARN).
  - Works extremely well when number of parameters is $d \sim 10 - 20$. But for large d acceptance rate typically goes to 0.
  - I usually resort to MWG if MIID acceptance rate is $< 10 - 25\%$, or if mode-finding algorithm is unreliable, etc.

Metropolized IID Sampler (MIID)



Metropolized IID Sampler (MIID)

Effective sample size for $M = 10,000$:

	Algorithm (acc. rate)		
	RMW (25%)	MWG (45%)	MIID (90%)
γ	1518	2043	8892
λ	460	1967	4195

Summary

- **RWM vs MWG:**
  - **Transition Density:** $\boldsymbol{\theta}_{\text{prop}}^{(RWM)} \sim \mathcal{N}(\boldsymbol{\theta}_{\text{curr}}, \text{diag}(\boldsymbol{\sigma}^2_{\text{RWM}}))$, $\quad \theta_{\text{prop},j}^{(MWG)} \sim \mathcal{N}(\theta_{\text{curr},j}, \sigma^2_{\text{MWG},j})$.
  - **Almost always** use MWG instead of RWM.
    - MWG almost always converges faster.
    - Price to pay is more log-posterior evaluations.
  - **Optimal Acceptance Rates:** $\alpha_{\text{RWM}} \approx 25\%$ and $\alpha_{\text{MWG}} \approx 45\%$.
- **MIID:**
  - **Transition Density:** $\boldsymbol{\theta}_{\text{prop}} \overset{\text{iid}}{\sim} q(\boldsymbol{\theta})$ (typically a normal with mode-quadrature matching $\log p(\boldsymbol{\theta} \mid \mathbf{y})$).
  - **Optimal Acceptance Rate:** α_{MIID} as high as possible.
  - **Efficiency:** Calculation of $q(\boldsymbol{\theta}_{\text{prop}})$ and $p(\boldsymbol{\theta}_{\text{prop}} \mid \mathbf{y})$ can be easily vectorized (unlike RWM and MWG).
  - Can be combined with MWG, but recalculating mode-quadrature within each Gibbs step can be very expensive.

Marginal MCMC

- **Model:** $Y \sim \text{Weibull}(\gamma, \lambda) \iff Y = \lambda X^\gamma,\ X \sim \text{Expo}(1)$.
- **Loglikelihood:**

$$\ell(\gamma, \lambda \mid \mathbf{y}) = n\big[\log(\gamma) - \gamma\log(\lambda)\big] + \sum_{i=1}^n \big[\gamma \log(y_i) - \lambda^{-\gamma} y_i^\gamma\big]$$
$$= n\big[\log(\gamma) + \log(\eta)\big] + \gamma S - \eta T_\gamma,$$

where $\eta = \lambda^{-\gamma}$, $S = \sum_{i=1}^n \log(y_i)$, and $T_\gamma = \sum_{i=1}^n y_i^\gamma$.

- **Conditionally Conjugate Prior:** For fixed γ:
  - *Conditional Likelihood*: $\ell(\eta \mid \gamma, \mathbf{y}) = n\log(\eta) - \eta T_\gamma$.
  - *Conjugate Prior:* $\pi(\eta \mid \gamma) \sim \text{Gamma}(\alpha, \beta)$

    $$\iff \log \pi(\eta \mid \gamma) = (\alpha - 1)\log(\eta) - \eta\beta.$$
  - *Conditional Posterior:* $\eta \mid \gamma, \mathbf{y} \sim \text{Gamma}(\hat{\alpha}, \hat{\beta}_\gamma)$, $\quad \hat{\alpha} = \alpha + n$

    $$\hat{\beta}_\gamma = \beta + T_\gamma.$$

Marginal MCMC

- **Model:** $Y \sim \text{Weibull}(\gamma, \lambda) \iff Y = \lambda X^{\gamma},\ X \sim \text{Expo}(1)$.
- **Loglikelihood:** $\ell(\gamma, \lambda \mid \mathbf{y}) = n\big[\log(\gamma) + \log(\eta)\big] + \gamma S - \eta T_{\gamma}$,

  where $\eta = \lambda^{-\gamma}$, $S = \sum_{i=1}^{n} \log(y_i)$, and $T_{\gamma} = \sum_{i=1}^{n} y_i^{\gamma}$.
- **Conditionally Conjugate Prior:** $\pi(\gamma, \eta)$ such that
$$\begin{aligned} \gamma &\sim \pi(\gamma) \\ \eta \mid \gamma &\sim \text{Gamma}(\alpha, \beta). \end{aligned}$$
- **Conditional Posterior:** $\eta \mid \gamma, \mathbf{y} \sim \text{Gamma}(\hat{\alpha}, \hat{\beta}_{\gamma})$,
$$\begin{aligned} \hat{\alpha} &= \alpha + n \\ \hat{\beta}_{\gamma} &= \beta + T_{\gamma}. \end{aligned}$$
- **Marginal Posterior:**
$$\begin{aligned} p(\gamma \mid \mathbf{y}) = \frac{p(\gamma, \eta \mid \mathbf{y})}{p(\eta \mid \gamma, \mathbf{y})} &\propto \frac{\mathcal{L}(\gamma, \eta \mid \mathbf{y}) \pi(\gamma, \eta)}{\texttt{dgamma}(\eta \mid \hat{\alpha}, \hat{\beta}_{\gamma})} \\ &= \exp\Big\{ \log \Gamma(\hat{\alpha}) - \hat{\alpha} \log(\hat{\beta}_{\gamma}) + n \log(\gamma) + \gamma S \Big\} \times \pi(\gamma). \end{aligned}$$

  $\implies$ can do 1-d MCMC to get $\gamma^{(m)} \sim p(\gamma \mid \mathbf{y})$, followed by $\eta^{(m)} \overset{\text{ind}}{\sim} \text{Gamma}(\hat{\alpha}, \hat{\beta}_{\gamma^{(m)}})$.

Efficiency of Gibbs Sampling Schemes

- **Theorem:** Consider three Gibbs sampling schemes on $p(x, y, z)$:

  1. **Single-Component Gibbs:** $x \leftrightharpoons y \leftrightharpoons z$
  2. **Block Gibbs:** $x \leftrightharpoons (y, z)$
  3. **Collapsed Gibbs:** first $x \leftrightharpoons y$, then $z \sim p(z \mid x, y)$.

  Then we have: $\quad \text{ESS(Scheme 1)} \le \text{ESS(Scheme 2)} \le \text{ESS(Scheme 3)}.$

- **Practical Considerations:**
  - Result only holds for exact Gibbs sampler, i.e., if all schemes above use Metropolis-within-Gibbs, then usually ESS(Scheme 1) $\ge$ ESS(Scheme 2), as the effectiveness of RW multivariate proposals decreases exponentially with number of dimensions.
  - If all schemes are MWG, then Scheme 3 (if available) is always better than the other two. However, if Scheme 1 is exact Gibbs and Scheme 3 is MWG, then often ESS(Scheme 1) $\ge$ ESS(Scheme 3) if number of parameters is large and few are being collapsed.

Resources

- **Julia Programming Language:** MCMC is for-loop intensive, and these are very slow in **R**. **Julia** is very similar to **R** and **Matlab**, but it can execute for-loops extremely fast (see here for technical details). Moreover, the **R** package `JuliaCall` allows you to interface **Julia** code directly from **R**.
- **Adaptive MCMC:** RWM and MWG algorithms work best when the acceptance rate is 25% or 45%. Instead of finding jump sizes $\boldsymbol{\sigma}_{\mathsf{RW}}$ to achieve this by trial-and-error, it is possible to do so automatically (e.g., decrease jump size if fraction of accepted proposals so far is $< 45\%$, increase otherwise).

  However, doing this naively typically produces draws $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$ which do not come from $p(\boldsymbol{\theta} \mid \mathbf{y})$. For several examples of valid adaptive MCMC methods, see here.