Basics of Markov Chain Monte Carlo

STAT 946: Advanced Bayesian Computing

Motivation

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

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

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

- **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

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

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

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

Markov Chain Monte Carlo (MCMC)

Problem: Let

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

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

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

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

Markov Chain Monte Carlo (MCMC)

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

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

Problem: Drawing $\boldsymbol{x}^{(m)} \stackrel{\text{iid}}{\sim} p(\boldsymbol{x})$ typically very difficult for $d > 2$.

Markov Chain Monte Carlo (MCMC)

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

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

  Problem: Drawing $\boldsymbol{x}^{(m)} \overset{\text{iid}}{\sim} p(\boldsymbol{x})$ typically very difficult for $d > 2$.
- **Solution:** Much easier to design a Markov chain

$$\boldsymbol{x}^{(m)} \sim \mathrm{T}(\boldsymbol{x} \mid \boldsymbol{x}^{(m-1)})$$

  for which the stationary distribution is $p(\boldsymbol{x})$.

Markov Chain Monte Carlo (MCMC)

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

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

  Problem: Drawing $\boldsymbol{x}^{(m)} \overset{\text{iid}}{\sim} p(\boldsymbol{x})$ typically very difficult for $d > 2$.
- **Solution:** Much easier to design a Markov chain

$$\boldsymbol{x}^{(m)} \sim \mathrm{T}(\boldsymbol{x} \mid \boldsymbol{x}^{(m-1)})$$

  for which the stationary distribution is $p(\boldsymbol{x})$.
  Still have $\bar{\tau} \to E[\tau]$, but usually $\mathrm{var}(\bar{\tau}_{\text{iid}}) < \mathrm{var}(\bar{\tau}_{\text{mcmc}})$.

Markov Chain Monte Carlo

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

Markov Chain Monte Carlo

- **Problem:** Let $\tau = g(\boldsymbol{x})$, $\boldsymbol{x} \sim p(\boldsymbol{x})$. Compute $E[\tau]$.
- **MCMC:**
  - Sample from a Markov chain $\boldsymbol{x}^{(m)} \sim \mathrm{T}(\boldsymbol{x} \mid \boldsymbol{x}^{(m-1)})$ for which the stationary distribution is $p(\boldsymbol{x})$.
  - Calculate $\bar{\tau} = \frac{1}{M}\sum_{m=1}^{M} g(\boldsymbol{x}^{(m)}) \to E[\tau]$
- **Transition Density:** How to pick $\mathrm{T}(\boldsymbol{x} \mid \boldsymbol{x}')$?
  Two fundamental concepts:
  1. **REDUCE:** only sample parts of $\boldsymbol{x}$ at a time (Gibbs sampler)

Markov Chain Monte Carlo

- **Problem:** Let $\tau = g(\boldsymbol{x})$, $\boldsymbol{x} \sim p(\boldsymbol{x})$. Compute $E[\tau]$.
- **MCMC:**
  - Sample from a Markov chain $\boldsymbol{x}^{(m)} \sim \mathrm{T}(\boldsymbol{x} \mid \boldsymbol{x}^{(m-1)})$ for which the stationary distribution is $p(\boldsymbol{x})$.
  - Calculate $\bar{\tau} = \frac{1}{M} \sum_{m=1}^{M} g(\boldsymbol{x}^{(m)}) \to E[\tau]$
- **Transition Density:** How to pick $\mathrm{T}(\boldsymbol{x} \mid \boldsymbol{x}')$?
  Two fundamental concepts:
  1. **REDUCE:** only sample parts of $\boldsymbol{x}$ 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 $\boldsymbol{x} \sim p(\boldsymbol{x})$
- Suppose we know how to sample from $p(x_i \mid \boldsymbol{x}_{-i})$ for every $1 \leq i \leq d$.

Input: $\boldsymbol{x}^{(0)}$ ▷ Starting value

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

Output: $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(M)}$

Example: Bivariate Normal

- **Model:**

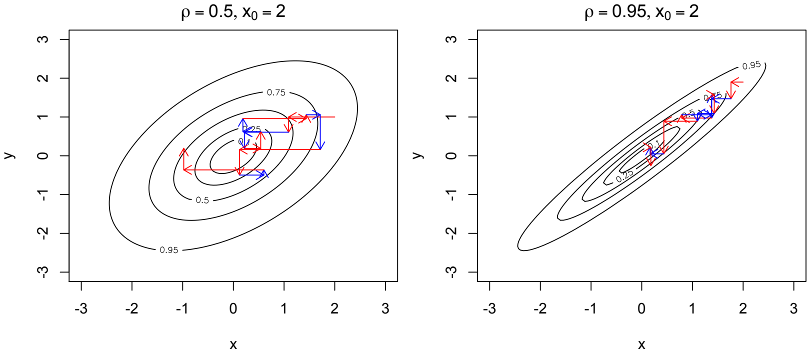
$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}_2 \left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \sigma_x \sigma_y \rho \\ \sigma_x \sigma_y \rho & \sigma_y^2 \end{bmatrix} \right).$$

- **Conditional Distributions:**

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

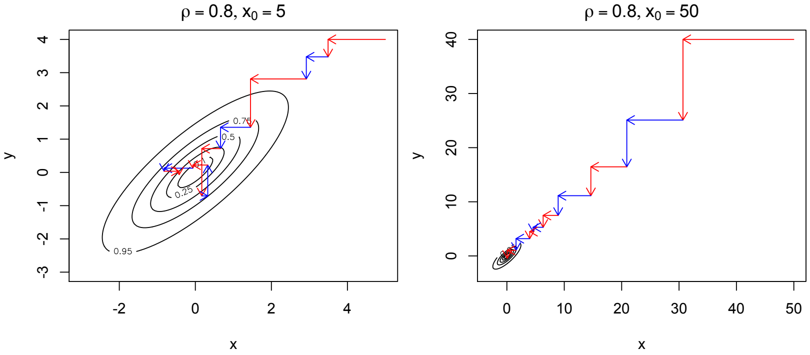
Example: Bivariate Normal

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



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 $x_i \sim p(\boldsymbol{x}_{-i})$. Can do these in any order, even random.
- **Limitations:**
  - Convergence is slow when $\text{cor}(x_i, \boldsymbol{x}_{-i}) \to 1$.

Gibbs Sampler (Continued)

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

Gibbs Sampler (Continued)

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

Metropolis-Hastings Algorithm

- Gibbs sampler requires you to be able to draw from each $p(x_i \mid \boldsymbol{x}_{-i})$.
- What if $p(x_i \mid \boldsymbol{x}_{-i})$ is not easy to draw from?
- MH algorithm requires only a transition density $\mathrm{T}(\boldsymbol{x} \mid \boldsymbol{x}')$ for which:
  1. You can draw $\boldsymbol{x} \sim \mathrm{T}(\boldsymbol{x} \mid \boldsymbol{x}')$

Metropolis-Hastings Algorithm

- Gibbs sampler requires you to be able to draw from each $p(x_i \mid \boldsymbol{x}_{-i})$.
- What if $p(x_i \mid \boldsymbol{x}_{-i})$ is not easy to draw from?
- MH algorithm requires only a transition density $\mathrm{T}(\boldsymbol{x} \mid \boldsymbol{x}')$ for which:
  1. You can draw $\boldsymbol{x} \sim \mathrm{T}(\boldsymbol{x} \mid \boldsymbol{x}')$
  2. You have a closed-form PDF (or PMF) for $\mathrm{T}(\boldsymbol{x} \mid \boldsymbol{x}')$ (i.e., including normalizing constant)

Metropolis-Hastings Algorithm

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

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

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

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

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

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

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

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

  else

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

  end if

end for

Output: $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(M)}$

Metropolis-Hastings Algorithm

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

Metropolis-Hastings Algorithm

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

Metropolis-Hastings Algorithm

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

Common Transition Densities

1. **Random Walk Metropolis:** $\boldsymbol{x}_{\text{prop}} \sim \mathcal{N}\big(\boldsymbol{x}_{\text{curr}}, \text{diag}(\boldsymbol{\sigma}^2_{\text{tune}})\big)$.
   Let $f(\boldsymbol{x})$ denote the PDF of $\mathcal{N}(\mathbf{0}, \text{diag}(\boldsymbol{\sigma}^2_{\text{tune}})$. Then

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

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

Common Transition Densities

1. **Random Walk Metropolis:** $\boldsymbol{x}_{\text{prop}} \sim \mathcal{N}\big(\boldsymbol{x}_{\text{curr}}, \text{diag}(\boldsymbol{\sigma}^2_{\text{tune}})\big)$.
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 $p(x_j \mid \boldsymbol{x}_{-j})$ can't be drawn from directly.

Common Transition Densities

1. **Random Walk Metropolis:** $\boldsymbol{x}_{\text{prop}} \sim \mathcal{N}\big(\boldsymbol{x}_{\text{curr}}, \text{diag}(\boldsymbol{\sigma}^2_{\text{tune}})\big)$.
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:** $\boldsymbol{x}_{\text{prop}} \overset{\text{iid}}{\sim} q(\boldsymbol{x})$.
   Typically this is "mode-quadrature" proposal
   $\mathcal{N}(\hat{\boldsymbol{x}}, -[\frac{\partial^2}{\partial \boldsymbol{x}^2} \log p(\hat{\boldsymbol{x}})]^{-1})$, where $\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} p(\boldsymbol{x})$.

Metropolis-Hastings Algorithm

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

Metropolis-Hastings Algorithm

- **Algorithm Summary:**
  1. Draw $\boldsymbol{x}_{\text{prop}} \sim \mathrm{T}(\boldsymbol{x} \mid \boldsymbol{x}_{\text{curr}} = \boldsymbol{x}^{(m)})$
  2. Let $\alpha = \min\left\{1, \dfrac{p(\boldsymbol{x}_{\text{prop}}) / \mathrm{T}(\boldsymbol{x}_{\text{prop}} \mid \boldsymbol{x}_{\text{curr}})}{p(\boldsymbol{x}_{\text{curr}}) / \mathrm{T}(\boldsymbol{x}_{\text{curr}} \mid \boldsymbol{x}_{\text{prop}})}\right\}$
  3. Set $\boldsymbol{x}^{(m+1)}$ to $\boldsymbol{x}_{\text{prop}}$ with probability α, to $\boldsymbol{x}_{\text{curr}}$ with probability $1 - \alpha$
- **Question:** Why does it work?
- **Theorem:** Suppose that $\boldsymbol{x}^{(m)}$ is drawn from $p(\boldsymbol{x})$, and $\boldsymbol{x}^{(m+1)}$ is an MH update, i.e.,

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

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

Metropolis-Hastings Algorithm

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

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

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

Metropolis-Hastings Algorithm

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

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

- **Proof:** Consider $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ such that $\alpha = \frac{p(\boldsymbol{x}_a)/\,\mathrm{T}(\boldsymbol{x}_a|\boldsymbol{x}_b)}{p(\boldsymbol{x}_b)/\,\mathrm{T}(\boldsymbol{x}_b|\boldsymbol{x}_a)} < 1$.
  1. Joint distribution of a then b:
     $p(\boldsymbol{x}^{(m)} = \boldsymbol{x}_a, \boldsymbol{x}^{(m+1)} = \boldsymbol{x}_b) = p(\boldsymbol{x}_a) \cdot \mathrm{T}(\boldsymbol{x}_b \mid \boldsymbol{x}_a)$.
  2. Joint distribution of b then a (proposal accepted with probability α)

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

Metropolis-Hastings Algorithm

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

$$\boldsymbol{x}^{(m+1)} \mid \boldsymbol{x}^{(m)} \sim \mathrm{MH}\{\mathrm{T}, \boldsymbol{x}^{(m)}\}$$

- **Proof:** Consider $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ such that $\alpha = \frac{p(\boldsymbol{x}_a) / \mathrm{T}(\boldsymbol{x}_a | \boldsymbol{x}_b)}{p(\boldsymbol{x}_b) / \mathrm{T}(\boldsymbol{x}_b | \boldsymbol{x}_a)} < 1$.
  1. Joint distribution of a then b:
     $p(\boldsymbol{x}^{(m)} = \boldsymbol{x}_a, \boldsymbol{x}^{(m+1)} = \boldsymbol{x}_b) = p(\boldsymbol{x}_a) \cdot \mathrm{T}(\boldsymbol{x}_b \mid \boldsymbol{x}_a)$.
  2. Joint distribution of b then a
     $p(\boldsymbol{x}^{(m)} = \boldsymbol{x}_b, \boldsymbol{x}^{(m+1)} = \boldsymbol{x}_a) = p(\boldsymbol{x}_a) \cdot \mathrm{T}(\boldsymbol{x}_b \mid \boldsymbol{x}_a)$.

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

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

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

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:

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

  - $h(y)$ characterizes distribution, just like $f(y)$ or $F(y)$

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

Weibull Distribution

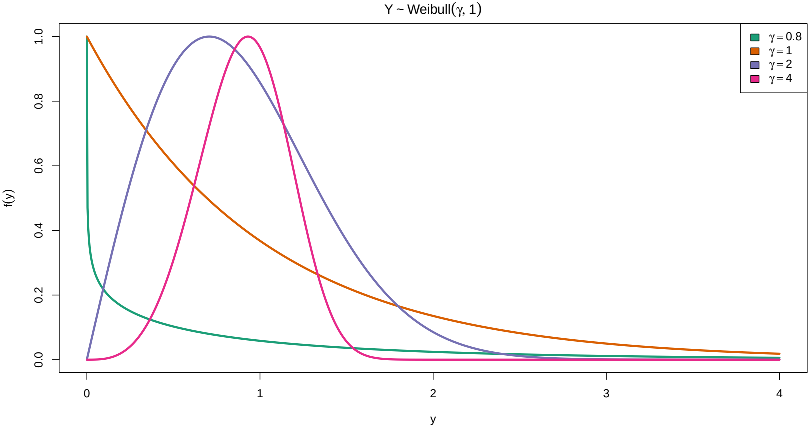
- **Model:** $Y \sim \text{Weibull}(\gamma, \lambda) \iff Y = \lambda X^{\gamma},\ X \sim \text{Expo}(1)$.
- **Utility:** Survival Analysis.
  - *Hazard function:* $\approx$ probability of failing in next instant:
$$h(y) = \lim_{\Delta y \to 0} \frac{\Pr(Y < y + \Delta y \mid Y > y)}{\Delta y} = \frac{f(y)}{1 - F(y)}$$
  - $h(y)$ 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

Weibull Distribution

- **Model:** $Y \sim \text{Weibull}(\gamma, \lambda) \iff Y = \lambda X^{\gamma},\ X \sim \text{Expo}(1)$.
- **Utility:** Survival Analysis.
  - *Hazard function:* $\approx$ probability of failing in next instant:
$$h(y) = \lim_{\Delta y \to 0} \frac{\Pr(Y < y + \Delta y \mid Y > y)}{\Delta y} = \frac{f(y)}{1 - F(y)}$$
  - $h(y)$ 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
  - $\gamma < 1 \implies h(y)$ decreasing
    Ex: infants more and more likely to survive longer as they get older

Weibull Distribution

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

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

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

Weibull Distribution

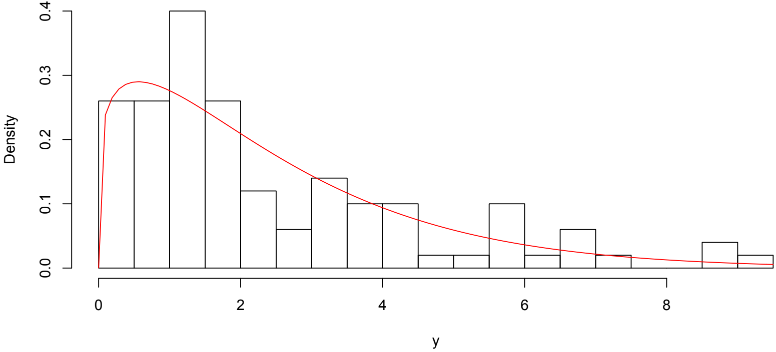
- **Model:** $Y \sim \text{Weibull}(\gamma, \lambda) \quad \iff \quad Y = \lambda X^{\gamma},\ X \sim \text{Expo}(1).$
- **Likelihood:** $\boldsymbol{y} = (y_1, \ldots, y_n) \overset{\text{iid}}{\sim} \text{Weibull}(\gamma, \lambda)$

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

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

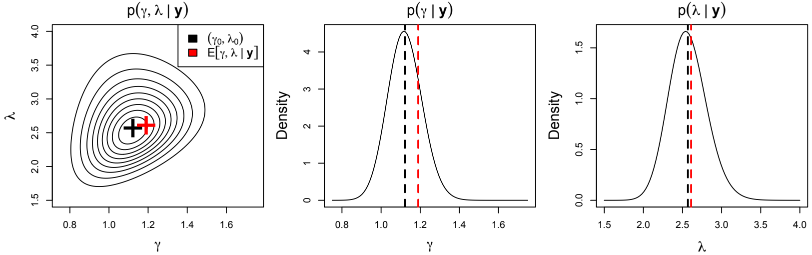
Weibull Distribution

- **Model:** $Y \sim \text{Weibull}(\gamma, \lambda) \iff Y = \lambda X^\gamma,\ X \sim \text{Expo}(1)$.
- **Likelihood:**
  $\ell(\gamma, \lambda \mid \boldsymbol{y}) = n\big[\log(\gamma) - \gamma \log(\lambda)\big] + \sum_{i=1}^n \big[\gamma \log(y_i) - \lambda^{-\gamma} y_i^\gamma\big]$.
- **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:**
  $\ell(\gamma, \lambda \mid \boldsymbol{y}) = n\big[\log(\gamma) - \gamma\log(\lambda)\big] + \sum_{i=1}^n \big[\gamma\log(y_i) - \lambda^{-\gamma} y_i^\gamma\big].$
- **Prior:** $\pi(\gamma, \lambda) \propto 1$ (hopefully won't make much difference)
- **Posterior:** For 2-d problem can compute $p(\gamma, \lambda \mid \boldsymbol{y})$ 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:**
  $\ell(\gamma, \lambda \mid \boldsymbol{y}) = n\big[\log(\gamma) - \gamma\log(\lambda)\big] + \sum_{i=1}^n \big[\gamma\log(y_i) - \lambda^{-\gamma} y_i^\gamma\big].$
- **Prior:** $\pi(\gamma, \lambda) \propto 1$
- **Posterior:** For 2-d problem can compute $p(\gamma, \lambda \mid \boldsymbol{y})$ on a grid, **OR** MCMC on $\boldsymbol{\theta} = (\gamma, \lambda)$:
  1. **Random-Walk Metropolis:** $\boldsymbol{\theta}_{\text{prop}} \sim \mathcal{N}\big(\boldsymbol{\theta}_{\text{curr}}, \text{diag}(\boldsymbol{\sigma}^2_{\text{RW}})\big).$
  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:**
     $\boldsymbol{\theta}_{\text{prop}} \overset{\text{iid}}{\sim} \mathcal{N}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}), \qquad \hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta} \mid \boldsymbol{y})$

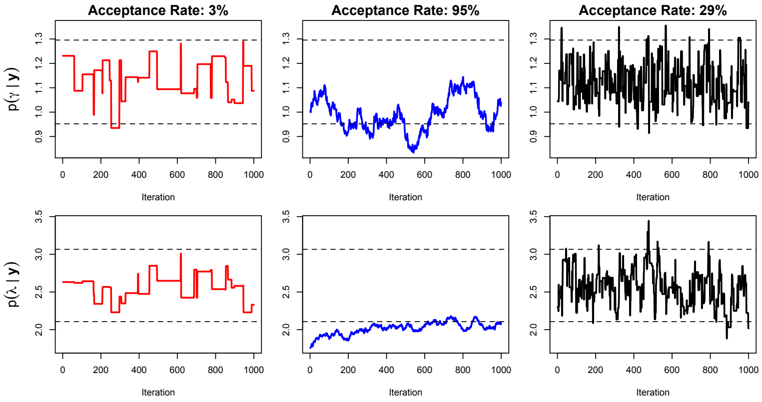
     $$\hat{\boldsymbol{\Sigma}} = -\left[\frac{\partial^2}{\partial\boldsymbol{\theta}^2}\log p(\hat{\boldsymbol{\theta}} \mid \boldsymbol{y})\right]^{-1}$$

Random-Walk Metropolis (RWM)

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

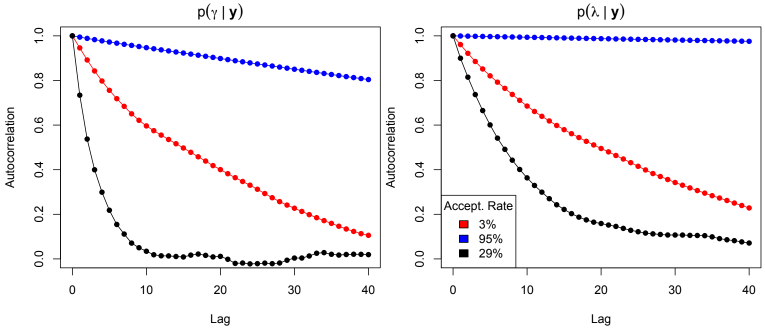
Random-Walk Metropolis (RWM)

- **Transition Density:** $\theta_{\text{prop}} \sim \mathcal{N}\left(\theta_{\text{curr}}, \text{diag}(\boldsymbol{\sigma}^2_{\text{RW}})\right)$
- **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 $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)}$
2. **Autocorrelation Plot:** Ideally would have $\boldsymbol{\theta}^{(m)} \overset{\text{iid}}{\sim} p(\boldsymbol{\theta} \mid \boldsymbol{y})$, but instead draws are correlated.



MCMC Diagnostics

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

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

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

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

MCMC Diagnostics

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

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

Weibull example for $M = 10,000$:

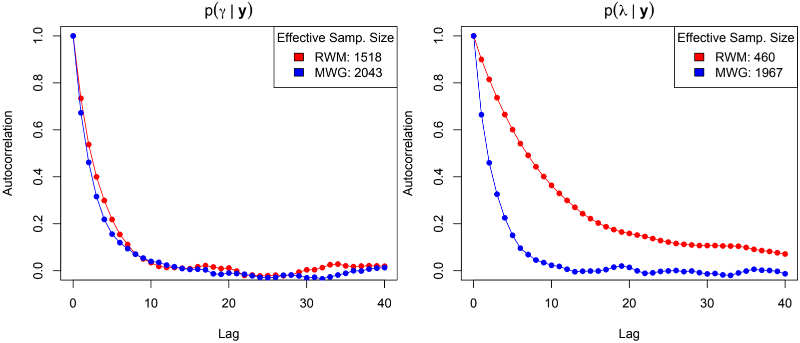
	Accept. Rate		
	3%	95%	29%
γ	286	137	1518
λ	235	125	460

Metropolis-Within-Gibbs (MWG)

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

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

- **Optimal acceptance rate:**

Metropolized IID Sampler (MIID)

- **Transition Density:**

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

- **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).

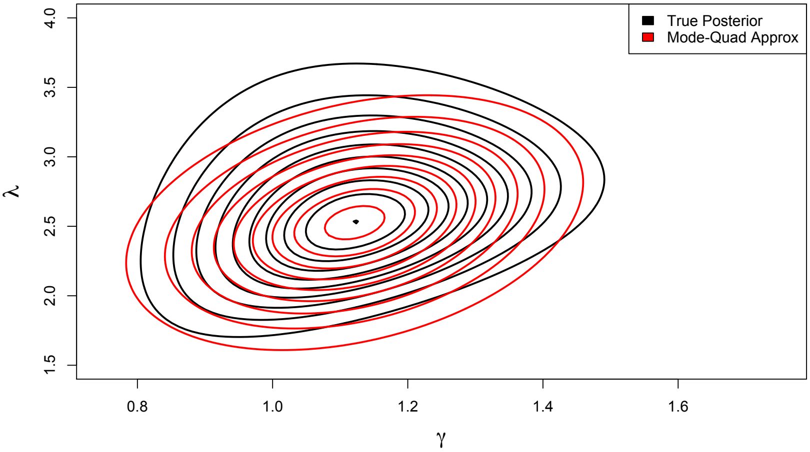
Metropolized IID Sampler (MIID)

- **Transition Density:**

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

- **Optimal acceptance rate:**

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:**
    $\boldsymbol{\theta}_{\text{prop}}^{(RWM)} \sim \mathcal{N}\left(\boldsymbol{\theta}_{\text{curr}}, \text{diag}(\boldsymbol{\sigma}_{\text{RWM}}^2)\right), \quad \theta_{\text{prop},j}^{(MWG)} \sim \mathcal{N}(\theta_{\text{curr},j}, \sigma_{\text{MWG},j}^2)$.
  - **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 $\log p(\boldsymbol{\theta} \mid \boldsymbol{y})$).
  - **Optimal Acceptance Rate:** α_{MIID} as high as possible.
  - **Efficiency:** Calculation of $q(\boldsymbol{\theta}_{\text{prop}})$ and $p(\boldsymbol{\theta}_{\text{prop}} \mid \boldsymbol{y})$ 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:** $Y \sim \text{Weibull}(\gamma, \lambda) \quad \iff \quad Y = \lambda X^{\gamma},\ X \sim \text{Expo}(1).$
- **Loglikelihood:**

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

  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:* $\quad \ell(\eta \mid \gamma, \boldsymbol{y}) = n \log(\eta) - \eta T_{\gamma}.$
  - *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:*
    $$\eta \mid \gamma, \boldsymbol{y} \sim \text{Gamma}(\hat{\alpha}, \hat{\beta}_{\gamma}), \qquad \hat{\alpha} = \alpha + n$$
    $$\hat{\beta}_{\gamma} = \beta + T_{\gamma}.$$

Marginal MCMC

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

$$\gamma \sim \pi(\gamma)$$
$$\eta \mid \gamma \sim \text{Gamma}(\alpha, \beta).$$

- **Conditional Posterior:**

$$\eta \mid \gamma, \boldsymbol{y} \sim \text{Gamma}(\hat{\alpha}, \hat{\beta}_\gamma), \qquad \hat{\alpha} = \alpha + n$$
$$\hat{\beta}_\gamma = \beta + T_\gamma.$$

- **Marginal Posterior:**

$$p(\gamma \mid \boldsymbol{y}) = \frac{p(\gamma, \eta \mid \boldsymbol{y})}{p(\eta \mid \gamma, \boldsymbol{y})} \propto \frac{\mathcal{L}(\gamma, \eta \mid \boldsymbol{y}) \pi(\gamma, \eta)}{\texttt{dgamma}(\eta \mid \hat{\alpha}, \hat{\beta}_\gamma)}$$
$$= \exp\left\{\log \Gamma(\hat{\alpha}) - \hat{\alpha} \log(\hat{\beta}_\gamma) + n \log(\gamma) + \gamma S\right\} \times \pi(\gamma$$

$\implies$ can do 1-d MCMC to get $\gamma^{(m)} \sim p(\gamma \mid \boldsymbol{y})$, 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:
  ESS(Scheme 1) $\leq$ ESS(Scheme 2) $\leq$ ESS(Scheme 3).

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:
  $\text{ESS(Scheme 1)} \leq \text{ESS(Scheme 2)} \leq \text{ESS(Scheme 3)}$.
- **Practical Considerations:**
  - Result only holds for exact Gibbs sampler, i.e., if all schemes above use Metropolis-within-Gibbs, then usually $\text{ESS(Scheme 1)} \geq \text{ESS(Scheme 2)}$, as the effectiveness of RW multivariate proposals decreases exponentially with number of dimensions.

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:
  $\text{ESS(Scheme 1)} \leq \text{ESS(Scheme 2)} \leq \text{ESS(Scheme 3)}$.
- **Practical Considerations:**
  - Result only holds for exact Gibbs sampler, i.e., if all schemes above use Metropolis-within-Gibbs, then usually $\text{ESS(Scheme 1)} \geq \text{ESS(Scheme 2)}$, 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 $\text{ESS(Scheme 1)} \geq \text{ESS(Scheme 3)}$ if number of parameters is large and few are being collapsed.

A Receipe for Basic MCMC

- **Goal:** Sample from
$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \rho(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{y}) \times \pi(\boldsymbol{\theta}), \qquad \boldsymbol{\theta} = (\theta_1, \ldots, \theta_d).$$
- **Receipe:**
  1. *Gibbs Moves:* Carefully inspect $\log \rho(\boldsymbol{\theta})$ to see if any of the variables have conjugate distributions, e.g.,
$$\rho(\boldsymbol{\theta}) = -\frac{[\theta_i - \mu(\boldsymbol{\theta}_{-i})]^2}{2\sigma^2(\boldsymbol{\theta}_{-i})} + g(\boldsymbol{\theta}_{-i}) \quad \Longrightarrow \quad p(\theta_i \mid \boldsymbol{\theta}_{-i}, \boldsymbol{y}) \sim \mathcal{N}\big(\mu(\boldsymbol{\theta}_{-i}), \sigma^2(\boldsymbol{\theta}$$
(Sometimes advantageous to change $\pi(\boldsymbol{\theta})$ in order for this work)

A Receipe for Basic MCMC

- **Goal:** Sample from $p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \rho(\boldsymbol{\theta})$
- **Receipe:**
  1. *Gibbs Moves:* Carefully inspect $\log \rho(\boldsymbol{\theta})$ to see if any of the variables have conjugate distributions, e.g.,

     $$\rho(\boldsymbol{\theta}) = -\frac{[\theta_i - \mu(\boldsymbol{\theta}_{-i})]^2}{2\sigma^2(\boldsymbol{\theta}_{-i})} + g(\boldsymbol{\theta}_{-i}) \quad \Longrightarrow \quad p(\theta_i \mid \boldsymbol{\theta}_{-i}, \boldsymbol{y}) \sim \mathcal{N}\big(\mu(\boldsymbol{\theta}_{-i}), \sigma^2(\boldsymbol{\theta}$$

  2. *Metropolis-within-Gibbs:* For the remaining variables,

     2.1 Suppose $\rho(\boldsymbol{\theta}_{\boldsymbol{J}} \mid \boldsymbol{\theta}_{-\boldsymbol{J}})$ can be *easily* maximized. Then use Metropolized-IID proposal

     $$\boldsymbol{\theta}_{\boldsymbol{J},\text{prop}} \sim \mathcal{N}(\hat{\boldsymbol{\theta}}_{\boldsymbol{J}}, -\boldsymbol{Q}_{\boldsymbol{J}}^{-1}), \qquad \boldsymbol{Q}_{\boldsymbol{J}} = \frac{\partial^2}{\partial \boldsymbol{\theta}_{\boldsymbol{J}}{}^2} \log \rho(\hat{\boldsymbol{\theta}}_{\boldsymbol{J}}, \boldsymbol{\theta}_{-\boldsymbol{J}})$$

     2.2 Otherwise, do *single*-component Random-Walk proposal

     $$\theta_{j,\text{prop}} \sim \mathcal{N}(\theta_{j,\text{curr}}, \sigma^2_{j,\text{RW}})$$

Random Walk with Constraints

- **Posterior Distribution:** $p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \rho(\boldsymbol{\theta})$
- **Proposal:** $\theta_{j,\text{prop}} \sim \mathcal{N}(\theta_{j,\text{curr}}, \sigma^2_{j,\text{RW}})$
- **Question:** What to do if $\theta_j > 0$?
  1. Immediately reject proposal if $\theta_{j,\text{prop}}$. Easiest solution.
  2. Reparametrize $\eta_j = \log(\theta_j)$. Most effective solution, but don't forget to apply change-of-variables to prior:
  $$\pi(\eta_j, \boldsymbol{\theta}_{-j}) = \pi(\boldsymbol{\theta}) \times \frac{\mathrm{d}\theta_j}{\mathrm{d}\eta_j} = \pi(\boldsymbol{\theta}) \times \exp(\eta_j).$$
  3. Propose from truncated normal:
  $\theta_{j,\text{prop}} \sim \mathcal{N}(\theta_{j,\text{curr}}, \sigma^2_{j,\text{RW}}) \times \mathbb{1}\{\theta_{j,\text{prop}} > 0\}$.
  Careful: Acceptance rate is
  $$\alpha = \frac{\rho(\boldsymbol{\theta}_\text{prop})/T(\boldsymbol{\theta}_\text{prop} \mid \boldsymbol{\theta}_\text{curr})}{\rho(\boldsymbol{\theta}_\text{curr})/T(\boldsymbol{\theta}_\text{curr} \mid \boldsymbol{\theta}_\text{prop})} = \underbrace{\frac{\rho(\boldsymbol{\theta}_\text{prop})}{\rho(\boldsymbol{\theta}_\text{curr})}}_{\text{no truncation}} \times \underbrace{\frac{\texttt{pnorm}\left(\frac{\theta_{j,\text{prop}} - \theta_{j,\text{curr}}}{\sigma_{j,\text{RW}}}\right)}{\texttt{pnorm}\left(\frac{\theta_{j,\text{curr}} - \theta_{j,\text{prop}}}{\sigma_{j,\text{RW}}}\right)}}_{\text{trunc. prop. not reversible}},$$

A Receipe for Basic MCMC

- **Goal:** Sample from $p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \rho(\boldsymbol{\theta})$
- **Receipe:**
  1. *Gibbs Moves:* Carefully inspect $\log \rho(\boldsymbol{\theta})$ to see if any of the variables have conjugate distributions, as these can be drawn analytically.
  2. *Metropolis-within-Gibbs:* For the remaining variables,
     2.1 Suppose $\rho(\boldsymbol{\theta}_{\boldsymbol{J}} \mid \boldsymbol{\theta}_{-\boldsymbol{J}})$ can be *easily* maximized. Then use Metropolized-IID proposal

$$\boldsymbol{\theta}_{\boldsymbol{J},\text{prop}} \sim \mathcal{N}(\hat{\boldsymbol{\theta}}_{\boldsymbol{J}}, -\boldsymbol{Q}_{\boldsymbol{J}}^{-1}), \qquad \boldsymbol{Q}_{\boldsymbol{J}} = \frac{\partial^2}{\partial \boldsymbol{\theta}_{\boldsymbol{J}}{}^2} \log \rho(\hat{\boldsymbol{\theta}}_{\boldsymbol{J}}, \boldsymbol{\theta}_{-\boldsymbol{J}})$$

     2.2 Otherwise, do *single*-component Random-Walk proposal

$$\theta_{j,\text{prop}} \sim \mathcal{N}(\theta_{j,\text{curr}}, \sigma^2_{j,\text{RW}})$$

Question: How to set the tuning parameters $\boldsymbol{\sigma}_{\text{RW}}$?

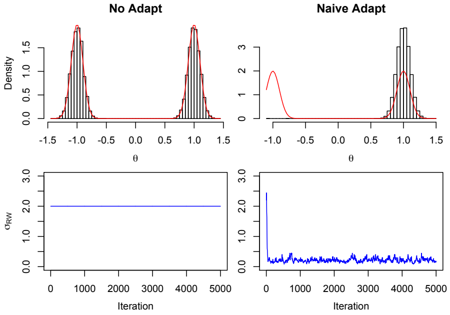
Adaptive MCMC

- **Random-Walk Proposal:** $\theta_{j,\text{prop}} \sim \mathcal{N}(\theta_{j,\text{curr}}, \sigma^2_{j,\text{RW}})$
- **Question:** How to set the tuning parameters $\boldsymbol{\sigma}_{\text{RW}}$?
  (Ideally want acceptance rate $\approx 45\%$)
- **Method 1 – Trial-and-Error.:** Can make this part of burn-in.
- **Method 2 – Adaptive MCMC:** Increase/Decrease $\sigma_{j,\text{RW}}$ at each step, depending on whether previous draw was accepted/rejected.
  Example:
  - *Target Distribution:* Mixture-Normal $\theta \sim \frac{1}{2}\mathcal{N}(-1, .1^2) + \frac{1}{2}\mathcal{N}(1, .1^2)$
  - *Update Rule:* At step $m+1$, $\sigma^{(m+1)}_{\text{RW}} = \exp(\log(\sigma^{(m)}_{\text{RW}}) \pm \delta)$, depending on whether draw at step m was accepted/rejected.

Adaptive MCMC

- **Random-Walk Proposal:** $\theta_{j,\text{prop}} \sim \mathcal{N}(\theta_{j,\text{curr}}, \sigma^2_{j,\text{RW}})$
- **Adaptive MCMC:** Increase/Decrease $\sigma_{j,\text{RW}}$ at each step, depending on whether previous draw was accepted/rejected.
  Example:

- *Target Distribution:* Mixture-Normal $\theta \sim \frac{1}{2}\mathcal{N}(-1, .1^2) + \frac{1}{2}\mathcal{N}(1, .1^2)$
- *Update Rule:* At step $m+1$, $\sigma^{(m+1)}_{\text{RW}} = \exp(\log(\sigma^{(m)}_{\text{RW}}) \pm \delta)$, depending on whether draw at step m was accepted/rejected.
- *Initialization:* $\theta_0 = 1$, $\sigma_{\text{RW}} = 2$, $\delta = .1$

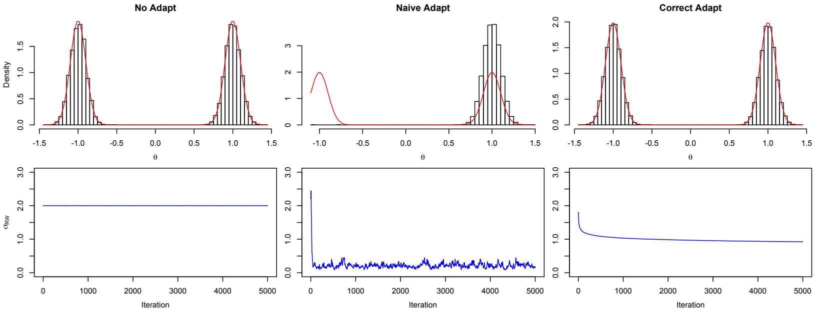


Adaptive MCMC

- **Random-Walk Proposal:** $\theta_{j,\text{prop}} \sim \mathcal{N}(\theta_{j,\text{curr}}, \sigma^2_{j,\text{RW}})$
- **Adaptive MCMC:** Increase/Decrease $\sigma_{j,\text{RW}}$ at each step, depending on whether previous draw was accepted/rejected. Careful: "Naive" adaptation rules typically don't preserve the MCMC stationary distribution.
  - *Naive update rule:* $\sigma_{\text{RW}}^{(m+1)} = \exp(\log(\sigma_{\text{RW}}^{(m)}) \pm \delta)$
  - *Correct update rule:* $\sigma_{\text{RW}}^{(m+1)} = \exp(\log(\sigma_{\text{RW}}^{(m)}) \pm \delta/m)$
    $\implies$ Amount of adaptation $\to 0$.

Adaptive MCMC

- **Random-Walk Proposal:** $\theta_{j,\text{prop}} \sim \mathcal{N}(\theta_{j,\text{curr}}, \sigma^2_{j,\text{RW}})$
- **Adaptive MCMC:** Increase/Decrease $\sigma_{j,\text{RW}}$ at each step, depending on whether previous draw was accepted/rejected.
  - *Naive update rule:* $\sigma_{\text{RW}}^{(m+1)} = \exp(\log(\sigma_{\text{RW}}^{(m)}) \pm \delta)$
  - *Correct update rule:* $\sigma_{\text{RW}}^{(m+1)} = \exp(\log(\sigma_{\text{RW}}^{(m)}) \pm \delta/m)$



Adaptive Metropolis-within-Gibbs

- **Goal:** Sample from $p(\theta_1, \ldots, \theta_d \mid \boldsymbol{y}) \propto \rho(\boldsymbol{\theta})$.
- **Random-Walk-within-Gibbs Proposal:** At step m,

$$\theta_{j,\text{prop}} \sim \mathcal{N}\left(\theta_{j,\text{curr}}, (\sigma_{j,\text{RW}}^{(m)})^2\right)$$

- **Adaptive jump size:**

$$\sigma_{j,\text{RW}}^{(m+1)} = \exp(\log(\sigma_{j,\text{RW}}^{(m)}) \pm \delta^{(m)}), \qquad \delta^{(m)} = \min\{\delta_0, 1/m^{1/2}\}.$$

  Increase/decrease depending on whether cumulative fraction of accepted draws is greater/smaller than 45%.
- **Caution:** This won't fix everything, i.e., won't work well when either $\boldsymbol{\sigma}_{\text{RW}}^{(0)}$ or $\boldsymbol{\theta}^{(0)}$ is way off. Still, it's a great receipe for MCMC which I use all the time.

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.
- **Cython:** A language very similar to Python which gets translated into C/C++ that interfaces directly with the Python environment. In other words, Cython lets you write MCMC algorithms in something very close to Python but which is orders of magnitude faster, and which you can use directly from within Python.
- **Numba:** A just-in-time (JIT) compiler for Python. Unlike Cython it has zero learning curve, but it's not quite as flexible.
- **reticulate:** An R package for calling Python code or libraries from within R.