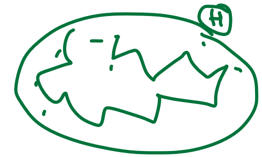
02/15/22

- Midterm 1 : solution ↑
- HW#3 : Due this Friday
- Midterm 2 : 03/09 (Tentative)

$\theta^{(0)}$ $\theta^{(t)} \sim K(\cdot \mid \theta^{(t-1)})$, $t=1,\ldots\ldots$ $\theta^{(t)} \sim \pi(\theta \mid x)$

$\theta^{(1)}, \ldots, \theta^{(T)}$



† Markov chain Monte Carlo (MCMC) methods (CR 6.3)

- *A more general Monte Carlo method* that approximates the generation of random variables from $\pi(\theta \mid x)$.

irreducible
positive recurrent
aperiodic

Ergodic Markov chain

- A Markov chain is a sequence of random variables $\theta^{(1)}, \theta^{(2)}, \ldots,$ where for any t, the distribution of $\theta^{(t)}$ given all previous θ's depends only on the most recent value, $\theta^{(t-1)}$.

  i.e., draw $\theta^{(t)}$ from a transition distribution (the transition kernel of the Markov chain), $K(\theta^{(t)} \mid \theta^{(t-1)})$.

- If $K(\cdot \mid \cdot)$ satisfies certain conditions (*detailed balance condition*), the distribution of $\theta^{(t)}$ converges to a unique stationary distribution that is the posterior distribution as t grows, regardless of where the chain was initiated.

- Markov chain transition kernel K is irreducible & recurrent

  ⇒ the chain visits any state in H w/p 1

- Every irreducible and positive recurrent kernel K has a unique stationary distribution.

- irreducible, positive recurrent & aperiodic

  ⇒ Markov chain is erdogic

- K : irreducible & aperiodic & π : stationary distr.

  ⇒ Regardless of the starting value the Markov chain converges to π

The working principle of MCMC algorithms

- For an arbitrary starting value $\theta^{(0)}$, a chain $(\theta^{(t)})$ is generated using a transition kernel with stationary distribution $\pi(\theta \mid \boldsymbol{x})$.

  Note: we will discuss schemes to produce valid transition kernels associated with arbitrary stationary distributions.

- Markov chain theory asserts that we will eventually sample from the target distribution π.

- Given that the chain is ergodic, the starting value $\theta^{(0)}$ is, in principle, unimportant.

  $\theta^{(0)} \longrightarrow \theta^{(T)} \mid \theta^{(T+1)}$

  $\pi(\theta \mid x)$

  $\pi(\theta \mid x)$

- Draws from the chain are slightly dependent, but independence of $(\theta^{(1)}, \ldots, \theta^{(T)})$ is not critical for an approximation of the form $\mathrm{E}(g(\theta) \mid x) \approx \frac{1}{T}\sum_{t=1}^{T} g(\theta^{(t)})$ (Ergodic Theorem).

† **How to build a transition kernel** such that the Markov chain converges to a unique stationary distribution that is our posterior distribution $\pi(\theta \mid \boldsymbol{x})$.

- Metropolis-Hastings algorithms (CR 6.3.2, PH Chapter 10, BDA Chapter 11.2)
- The Gibbs sampler (CR 6.3.3, PH Chapter 6, BDA Chapter 11.1)
- Building Markov chain algorithms using the Gibbs sampler and Metropolis algorithm

$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{bmatrix}$ $\pi(\theta_j \mid \theta_{-j}, x)$

(MH) $\theta^{(0)} \longrightarrow \theta^{(1)} \longrightarrow \theta^{(2)} \longrightarrow \cdots \longrightarrow \theta^{(T)}$

† Metropolis-Hastings algorithms

1. Start with an arbitrary initial value $\theta^{(0)}$.
2. Update from $\theta^{(t-1)}$ to $\theta^{(t)}$ $(t = 1, 2, \ldots)$ by
   2.1 Generate $\xi \sim q(\xi \mid \theta^{(t-1)})$ $\theta^{(t)}$
   2.2 Define

$$\rho(\theta^{(t-1)}, \xi) = \min\left\{\frac{\pi(\xi)q(\theta^{(t-1)} \mid \xi)}{\pi(\theta^{(t-1)})q(\xi \mid \theta^{(t-1)})}, 1\right\}.$$

   2.3 Take

$$\theta^{(t)} = \begin{cases} \xi & \text{with probability } \rho(\theta^{(t-1)}, \xi), \\ \theta^{(t-1)} & \textit{otherwise}. \end{cases}$$

† Metropolis-Hastings algorithms – contd

$\pi(\theta \mid x)$

- A popular algorithm for drawing from a given distribution $\pi(\theta)$
- The distribution with density $\pi(\theta)$ (can be known upto a normalizing factor) is called the *target* or *objective distribution*.
- The distribution with density $q(\cdot \mid \theta)$ (a conditional density) is the *proposal distribution* (candidate generating, or instrumental distribution).
- The probability $\rho(\theta^{(t-1)}, \xi)$ is called the *Metropolis-Hastings acceptance probability*.

† Metropolis-Hastings algorithms – contd

- An MH algorithm creates a Markov chain with $\pi(\theta)$ as its stationary or limiting distribution.

  ⋆⋆ Generate a state ξ from a candidate transition density $q(\xi \mid \theta^{(t-1)})$

  ⋆⋆ Accept this move with a "corrective" probability $\rho(\theta^{(t-1)}, \xi)$ that

- The algorithm constructs $K(\theta^{(t)} \mid \theta^{(t-1)})$ so that the Markov chain converges to a unique stationary distribution $\pi(\theta)$.

  $\Rightarrow$ if the simulation is run long enough, the distribution of $\theta^{(t)}$ is close enough to $\pi(\theta)$.

† Metropolis-Hastings algorithms – contd

- Conditions for the proposal distribution

  ⋆⋆ The support of $q(\cdot \mid \theta)$ contain the support of π for every θ.

  ⋆⋆ $q(\cdot \mid \theta)$ is positive in a neighborhood of θ of fixed radius.

† Metropolis-Hastings algorithms – contd

- The distribution with density $\pi(\theta)$ (can be known upto a normalizing factor) is called the *target* or *objective distribution*.

- The distribution with density $q(\cdot \mid \theta)$ (a conditional density) is the *proposal distribution* (candidate generating, or instrumental distribution).

- Conditions for the proposal distribution

  ⋆⋆ The support of $q(\cdot \mid \theta)$ contain the support of π for every θ.

  ⋆⋆ $q(\cdot \mid \theta)$ is positive in a neighborhood of θ of fixed radius.

- The probability $\rho(\theta, \xi)$ is called the *Metropolis-Hastings acceptance probability.*

$\theta^{(t)}$, ξ

† Proposal distributions

- A good proposal density q has the following properties:

  ⋆⋆ For any $\xi \in \Theta$, it is easy to sample from $q(\xi \mid \theta^{(t-1)})$.

  ⋆⋆ It is easy to compute ρ

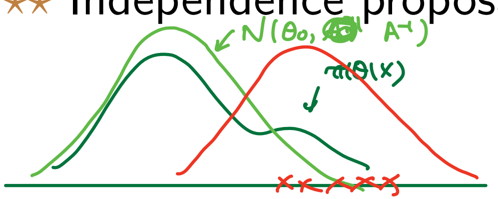
  ⋆⋆ Each move goes a reasonable distance in the parameter space (otherwise the chain moves too slowly)

  ⋆⋆ The jumps are not rejected too frequently (otherwise the chain wastes too much time standing still)

- The infinite number of proposed distributions yield a Markov chain that converges to the distribution of interest.

  ⋆⋆ Random-walk proposal: $q(\xi \mid \theta)$ is of the form $f(||\theta - \xi||)$.

  ⋆⋆ Independence proposal: $q(\xi \mid \theta) = h(\xi)$.



$\varepsilon \sim N(0, v^2)$

$\xi = \theta^{(t-1)} + \varepsilon$

$\rho = \min\left\{ 1, \frac{\pi(\xi)}{\pi(\theta^{(t-1)})} \cdot \frac{q(\theta^{(t-1)} \mid \xi)}{q(\xi \mid \theta^{(t-1)})} \right\}$

$= \dfrac{\frac{1}{\sqrt{2\pi v^2}} \cdot \exp\left(-\frac{(\xi - \theta^{(t-1)})^2}{2v^2}\right)}{\frac{1}{\sqrt{2\pi v^2}} \cdot \exp\left(-\frac{(\theta^{(t-1)} - \xi)^2}{2v^2}\right)}$

- M-H with Random-walk Proposal
  - ⋆⋆ Recall $q(\xi \mid \theta)$ is of the form $f(||\theta - \xi||)$.
  - ⋆⋆ $\Rightarrow$ The proposed value ξ is of the form $\xi = \theta^{(t-1)} + \epsilon$, where ϵ is distributed as a symmetric random variable.
  - ⋆⋆ The standard choices for f are uniform, normal or Cauchy.

    $(-1, 1)$ $\quad N(0, v^2)$
  - ⋆⋆ Idea: Perturb the current value of the chain at random, while staying in a neibhborhood of this value and then see if the new value ξ is likely for the distribution of interest.

- M-H with Random-walk Proposal (contd)

  ⋆⋆ Since $q(\theta^{(t-1)} \mid \xi) = q(\xi \mid \theta^{(t-1)})$, the acceptance probability is

$$\rho = \min\left\{ \frac{\pi(\xi)}{\underline{\pi(\theta^{(t)})}}, 1 \right\}.$$

  ⋆⋆ Appears to be the "gold standard" of MCMC techniques.

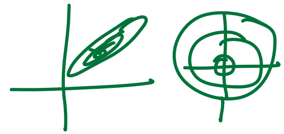
- M-H with Independent Proposal: density $q(\cdot \mid \theta)$ does not depend on θ, $q(\xi \mid \theta) = h(\cancel{\theta}\,\xi)$.

  ⋆⋆ For good performance, $\underline{h}$ should fit the target distribution.

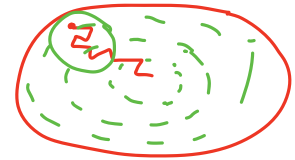
  $\Rightarrow$ limited applicability.

- Read BDA Section 12.2 for Efficient Metropolis jumping rules.

$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \in \mathbb{R}^2$



$\theta^{(t)}$ $\theta^{(t-1)}$



Checking Convergence - BDA Section 11.4

- *Possible problem 1:* If the iterations have not proceeded long enough, the simulations may be grossly unrepresentative of the target distribution.

- *Possible problem 2:* Even when the simulations have reached approximate convergence, the early iterations still are influenced by the starting approximation rather than the target distribution.

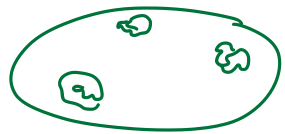
- *Possible problem 3:* Iterative simulation draws have within-sequence correlations which may cause some convergence issues.

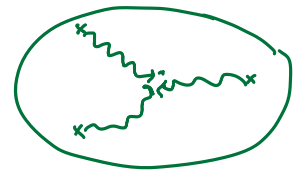
† Checking Convergence - contd.

$\theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \ldots$

- *Burn-in:*

  To diminish the effect of the starting distribution, discard early iterations of the simulation runs.





- *Thin:*

  To diminish the dependence of the iterations in a sequence, thin the sequence by keeping every kth simulation draw and discard the rest.

- *Run multiple sequences with overdispersed starting points:*

  Run multiple sequences with different starting points and compare them.

- May check the sample autocorrelation, the effective sample size....

- **Example 4:** Let $\pi(\theta)$ be IG(a, b) with $a = 3$ and $b = 3$ (that is, mean=1.5 and sd=1.5). Simulate θ using a M-H algorithm.

  ⋆⋆ **Strategy 1:** Use with random-walk proposal on $\theta \in \mathbb{R}^+$

  ⋆⋆ **Strategy 2:** Use with random-walk proposal on $\eta = \log(\theta) \in \mathbb{R}$

$$\pi_1(\eta) = \frac{b^a}{\Gamma(a)} (e^\eta)^{-a} \exp\left(-\frac{b}{e^\eta}\right).$$

$\Rightarrow$ draw a sample of η and let $\theta = \log(\eta)$.

$$\pi(\theta) = \frac{b^a}{\Gamma(a)} \theta^{-a-1} e^{-b/\theta}, \quad \theta > 0$$

$$\frac{1}{T}\sum_{t=1}^{T} \theta^{(t)} \approx 1.5$$

$$\pi(\theta) \propto \theta^{-a-1} e^{-b/\theta}$$

$$\rho = \min\left\{1, \frac{\pi(\xi)}{\pi(\theta^{(t-1)})}\right\}$$

$$\frac{\pi(\xi)}{\pi(\theta)} = \frac{\frac{b^a}{\Gamma(a)} \xi^{-a-1} e^{-b/\xi}}{\frac{b^a}{\Gamma(a)} (\theta^{(t-1)})^{-a-1} e^{-b/\theta^{(t-1)}}}$$

- **Strategy 1:** Use with Random-walk Proposal $\theta + \varepsilon$ $\varepsilon \sim N(0, 0.8^2)$

1. Specify a proposal distribution, $q(\xi \mid \theta) = \mathrm{N}(\theta, 0.8^2)$.
2. Let $\theta^{(0)} = 1.0$ for a starting value. $\theta^{(t-1)} = 0.5$
3. Iterate for $t = 1, \ldots, T(= 10{,}000)$ $\xi = -0.3$
   - 3.1 Generate $\xi \sim \mathrm{N}(\theta^{(t-1)}, 0.8^2)$
   - 3.2 Compute the acceptance probability if $\xi > 0$

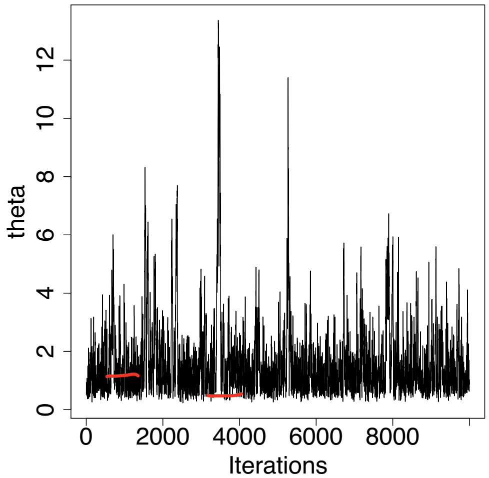
$$\rho = \min\left\{ \frac{\overbrace{\xi^{-a-1}\exp(-b/\xi)}^{=0}}{(\theta^{(t-1)})^{-a-1}\exp(-b/\theta^{(t-1)})}, 1 \right\}$$

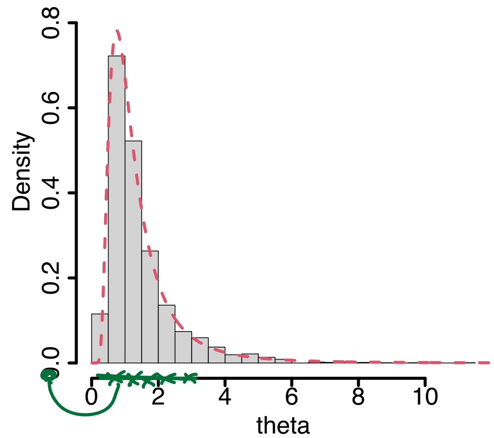
   - 3.3 Generate $r \sim \mathrm{Unif}(0, 1)$ and take

$$\theta^{(t)} = \begin{cases} \xi & \text{if } r < \rho, \\ \theta^{(t-1)} & \textit{otherwise.} \end{cases}$$

4. Discard the first 4000 iterations and keep every other iteration from the remaining. 3,000

- **Example 4:** - Strategy 1 (contd)



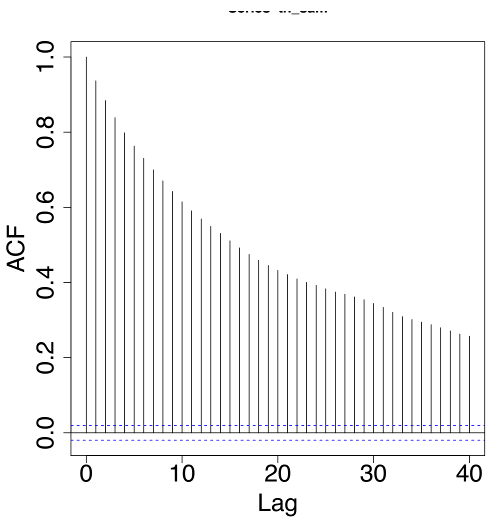



```
> mean(th_sam)
[1] 1.443216
> sd(th_sam)
[1] 1.057938
```

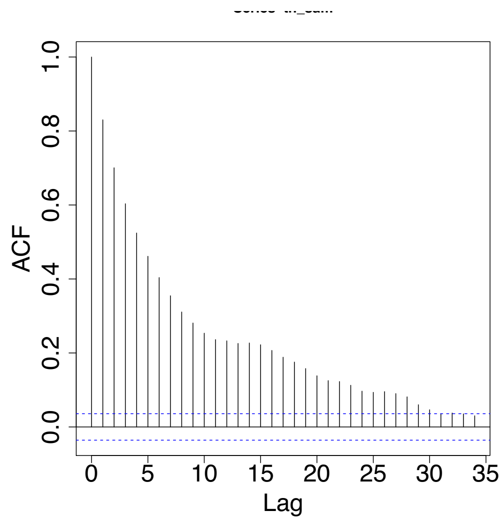
1.5

1.5

- **Example 4:** - Strategy 1 (contd)



(a) Including Burn-in
& before thinning



(b) Discard burn-in
& after thinning

- **Example 4:** - Strategy 1 (contd) Autocorrelation plots

```
> library(coda)
> effectiveSize(th_sam)
    var1
238.1634
```

$*$ The precision of the MCMC approximation to $E(\theta)$ is as good as the precision that would have been obtained by about 238 independent samples of θ.

- **Strategy 2:** Use with Random-walk Proposal for $\eta = \log(\theta)$

1. Specify a proposal distribution, $q(\xi \mid \eta) = \mathrm{N}(\eta, 0.5^2)$.
2. Let $\eta^{(0)} = \log(1.0)$ for a starting value.

   $\xi = \eta^{(t-1)} + \varepsilon$

   $\varepsilon \sim N(0, 0.5^2)$
3. Iterate for $t = 1, \ldots, T(= 10{,}000)$
   3.1 Generate $\xi \sim \mathrm{N}(\eta^{(t-1)}, 0.5^2)$
   3.2 Compute the acceptance probability

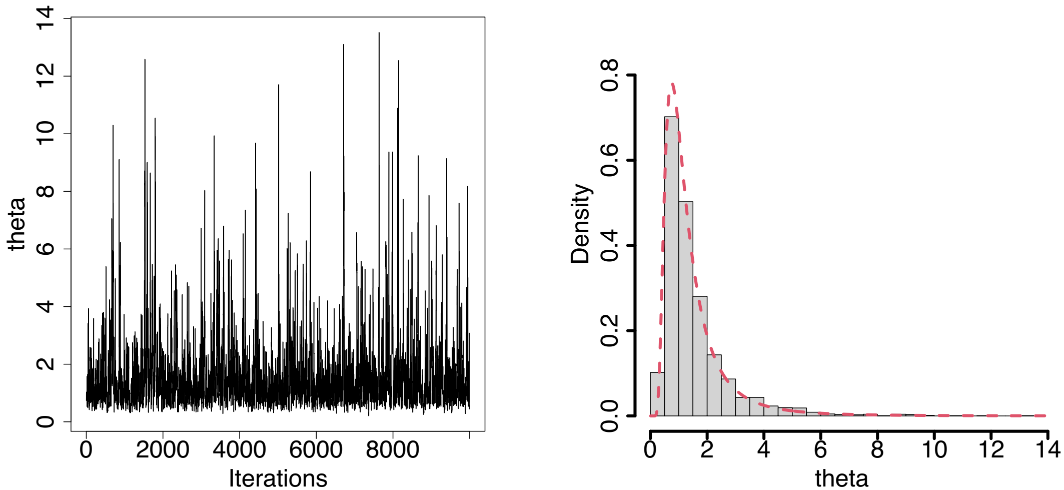
$$\rho = \min\left\{ \frac{(e^{\xi})^{-a} \exp(-b/e^{\xi})}{(e^{\eta^{(t-1)}})^{-a} \exp(-b/e^{\eta^{(t-1)}})}, 1 \right\}$$

   3.3 Generate $r \sim \mathrm{Unif}(0, 1)$ and take

$$\eta^{(t)} = \begin{cases} \xi & \text{if } r < \rho, \\ \theta^{(t-1)} & \textit{otherwise}. \end{cases}$$

4. Let $\underline{\theta^{(t)}} = e^{\eta^{(t)}}$ $\theta^{(1)}, \ldots, \theta^{(T)}$
5. Discard the first 4000 iterations and keep every other iteration from the remaining.

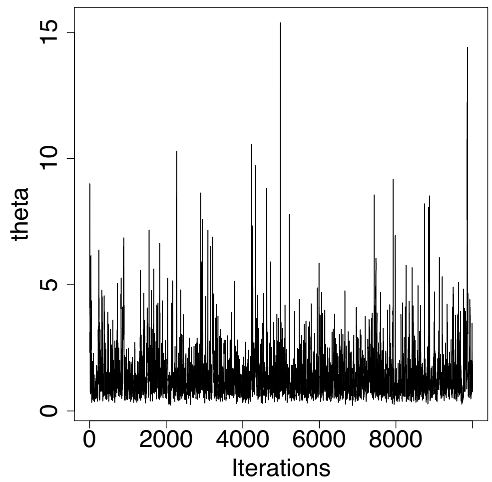
- **Example 4:** - Strategy 2 (contd)

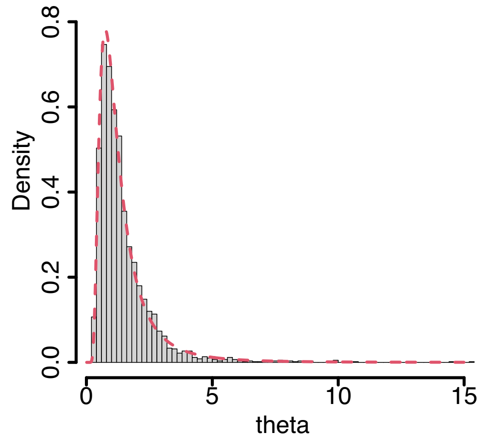



```
> mean(exp(eta_sam))
[1] 1.537653
> sd(exp(eta_sam))
[1] 1.250824
> effectiveSize(exp(eta_sam))
    var1
474.8557
```

- **Example 4:** - Strategy 2 (contd)

$*$ different initial value, $\underline{\eta^{(0)} = 10}$.






```
> mean(exp(eta_sam))
[1] 1.450118
> sd(exp(eta_sam))
[1] 1.141774
> effectiveSize(exp(eta_sam))
    var1
505.196
```

- **Example 6.3.2:** Weibull distributions are used extensively in reliability and other engineering applications, partly for their ability to describe different hazard rate behaviors, and partly for historic reasons. Suppose x_i is a random sample of size n from the Weibull distribution

$$f(x \mid \alpha, \eta) \propto \alpha\eta x^{\alpha-1} e^{-x^{\alpha}\eta}. \qquad x \in \mathbb{R}^+$$

For $\theta = (\alpha, \eta) \in (\mathbb{R}^+, \mathbb{R}^+)$, consider the prior

$$\pi(\theta) \propto \underbrace{e^{-\alpha}}_{= \pi_1(\alpha)} \underbrace{\eta^{\beta-1} e^{-\xi\eta}}_{= \pi_2(\eta)}.$$

That is, assume a priori independence and place E(1) and Gamma(β, ξ) (with mean β/ξ) for α and η, respectively. Let $\beta = 1$ and $\xi = 0.01$.
Simulate θ from $\pi(\theta \mid \boldsymbol{x})$ using a Metropolis-Hastings algorithm.

$x_1, \dots, x_n, \qquad x_i \in \mathbb{R}^+ \qquad \begin{cases} \alpha^* = 1 \\ \eta^* = 0.5 \end{cases}$

$$f(x_i \mid \alpha, \eta) = \alpha \eta \, x_i^{\alpha-1} e^{-x_i^\alpha \eta},$$

$\alpha \in \mathbb{R}^+ \quad \& \quad \eta \in \mathbb{R}^+$

$$\pi(\alpha, \eta) = \underbrace{\pi_1(\alpha)}_{Exp(1)} \underbrace{\pi_2(\eta)}_{Ga(1,\, 0.01)}$$

$$\pi(\alpha, \eta \mid x) \propto \underbrace{\prod_{i=1}^{n} f(x_i \mid \alpha, \eta) \cdot \pi(\alpha, \eta)}$$

- **Example 6.3.2:** (contd)

★★ Find the posterior distribution of θ.

$$\begin{aligned}\pi(\alpha,\eta \mid \boldsymbol{x}) &\propto f(\boldsymbol{x} \mid \alpha,\eta)\pi(\alpha,\eta)\\ &\propto \prod_{i=1}^{n}\left\{\alpha\eta x_i^{\alpha-1}e^{-x_i^{\alpha}\eta}\right\}e^{-\alpha}\eta^{\beta-1}e^{-\xi\eta}\\ &\propto \alpha^n\eta^{n+\beta-1}\prod_{i=1}^{n}x_i^{\alpha-1}\exp\left\{-\eta\sum_{i=1}^{n}x_i^{\alpha}-\alpha-\xi\eta\right\}.\end{aligned}$$

★★ Let $\underline{z_1} = \log(\alpha) \in \mathbb{R}$ and $\underline{z_2} = \log(\eta) \in \mathbb{R}$ and find

$$\begin{aligned}\underline{\pi_1(\boldsymbol{z} \mid \boldsymbol{x})} &\propto (e^{z_1})^{(n+1)}(e^{z_2})^{(n+\beta)}\\ &\quad \prod_{i=1}^{n}x_i^{e^{z_1}-1}\exp\left\{-e^{z_1}\sum_{i=1}^{n}x_i^{e^{z_1}}-e^{z_1}-\xi e^{z_1}\right\},\end{aligned}$$

where $\boldsymbol{z} = (z_1, z_2)$

- **Example 6.3.2:** (contd) Use MH with Random-walk Proposal

1. Specify a proposal distribution, $q(\xi \mid \boldsymbol{z}) = \mathrm{N}(z_1, 0.05)\mathrm{N}(z_2, 0.1)$.
2. Let $\boldsymbol{z}^{(0)} = (1.0, 1.0)$ for a starting value.
3. Iterate for $t = 1, \ldots, T$ (= 10,000)
   - 3.1 Generate $\xi_1 \sim \mathrm{N}(z_1^{(t-1)}, 0.05)$ and $\xi_2 \sim \mathrm{N}(z_2^{(t-1)}, 0.1)$ and let $\boldsymbol{\xi} = (\xi_1, \xi_2)$.
   - 3.2 Compute the acceptance probability

$$\rho = \min\left\{ \frac{\pi(\boldsymbol{\xi} \mid \boldsymbol{x})}{\pi(\boldsymbol{z}^{(t-1)} \mid \boldsymbol{x}))}, 1 \right\}$$

$\{ (\alpha^{(t)}, \eta^{(t)}) \}$

3,000

   - 3.3 Generate $r \sim \mathrm{Unif}(0, 1)$ and take

$$\boldsymbol{z}^{(t)} = \begin{cases} \boldsymbol{\xi} & \text{if } r < \rho, \\ \boldsymbol{z}^{(t-1)} & \textit{otherwise}. \end{cases}$$

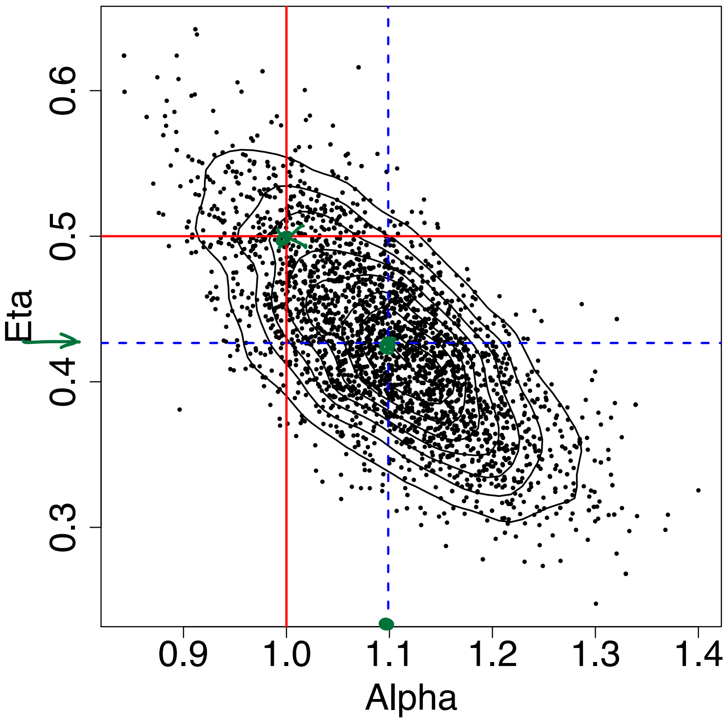
4. Let $\alpha^{(t)} = \exp(z_1^{(t)})$and $\eta^{(t)} = \exp(z_2^{(t)})$
5. Discard the first 4000 iterations and keep every other iteration from the remaining.

- **Example 6.3.2:** (contd)

$*$ Joint posterior distribution $\pi(\alpha, \eta \mid \boldsymbol{x})$

$\frac{1}{T}\sum_{t=1}^{T} \eta^{(t)} = 0.42$

$\approx E(\eta \mid x)$



$\frac{1}{T}\sum_{t=1}^{T} \alpha^{(t)} = 1.09 \approx E(\alpha \mid x)$

- **Example 6.3.2:** (contd)

$*$ Marginal posterior distributions $\pi(\alpha \mid \boldsymbol{x})$ & $\pi(\eta \mid \boldsymbol{x})$