02/17

- HW #3 : Friday 5pm
- Example codes

- **Example 6.3.2:** (contd) Predictive distribution $f(\underline{y} \mid \boldsymbol{x})$

★★ obtain a sample from the posterior predictive distribution by simulating $y^{(t)} \sim \text{Weibull}(\alpha^{(t)}, \eta^{(t)})$.

$y \perp\!\!\!\perp x$ given θ

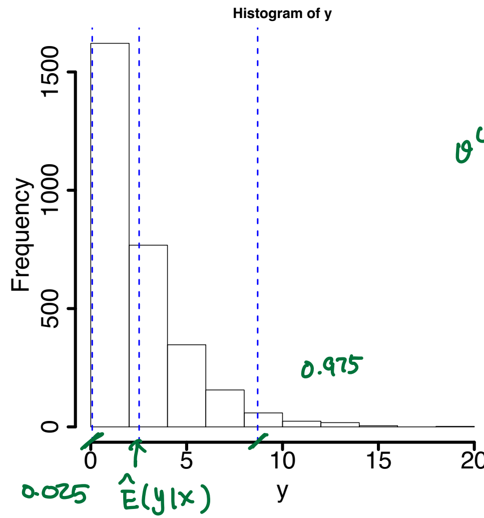
$f(y \mid x) = \int \underbrace{f(y \mid \theta)}_{\text{Weibull}(\alpha, \eta)} \underbrace{\pi(\theta \mid x)} d\theta$

$\theta^{(t)}$

$t = 1, \dots, T$

$\underline{y^{(t)}} \sim f(y \mid \theta^{(t)})$

$f(y \mid x)$




```
> mean(y)
[1] 2.522547
> sd(y)
[1] 2.399689
> quantile(y, prob=c(0.025, 0.975))
      2.5%      97.5%
0.07599534 8.70380222
```

- Let's consider the following, $\int \pi(\theta, \lambda \mid x)\, d\lambda$

$$\underline{\pi(\theta \mid x)} = \int \pi_1(\theta \mid x, \lambda)\underline{\pi_2(\lambda \mid x)}d\lambda.$$

Generating a sample of θ from $\pi(\theta \mid x)$ is equivalent to

⋆⋆ Generating $\underline{\lambda}^{(t)}$ from $\pi_2(\lambda \mid x)$.

⋆⋆ Generating $\theta^{(t)}$ from $\pi_1(\theta \mid x, \underline{\lambda^{(t)}})$

- Example 3: Normal $\times$ IG distribution

  ⋆⋆ Suppose we have

$$p(y) = \int p(x,y)\,dx$$
$$= \int p(x)\,p(y \mid x)\,dx$$

$$\begin{aligned} p(x,y) &= p(x)p(y \mid x) \\ &= \underbrace{\frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left(-\frac{\beta}{x}\right)}_{\mathrm{IG}(x \mid \alpha,\beta)} \underbrace{\frac{1}{\sqrt{2\pi x}} \exp\left(-\frac{(y-m)^2}{2x}\right)}_{\mathrm{N}(y \mid m,x)}. \end{aligned}$$

  ⋆⋆ Obtain a random sample of (x,y) from their joint $p(x,y)$.

  ⋆⋆ (Step 1:) Simulate $x \sim \mathrm{IG}(x \mid \alpha,\beta)$ and $y \mid x \sim \mathrm{N}(y \mid m,x)$.

  ⋆⋆ (Step 2:) Repeat until the target sample size is met.

- Example 3: Normal × IG distribution (contd)

★★ Simulate $(x, y) \sim \underbrace{\mathrm{IG}(x \mid \underbrace{3, 3}_{\nu/2,\ \nu/2})\mathrm{N}(y \mid 0, x)}_{t_\nu}$

B

$x^{(b)} \sim IG(3,3)$

$y^{(b)} \mid x^{(b)} \sim N(0, x^{(b)})$

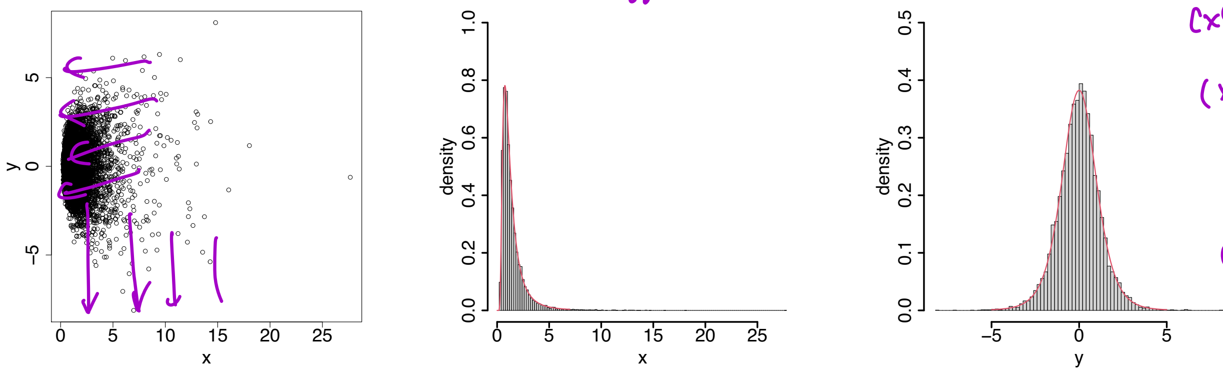
$(x^{(1)}, y^{(1)})$

$(x^{(2)}, y^{(2)})$

$(x^{(3)}, y^{(3)})$

$\vdots$

$(x^{(B)}, y^{(B)})$



(a) (x, y) (b) x (c) y t_6

$IG(3,3)$

- **Example 6.3.4:** Consider $(\theta, \lambda) \in \overset{\{0,\ldots,n\}}{\mathbb{N}} \times [0,1]$ and

$$\pi(\theta, \lambda \mid x) \propto \binom{n}{\theta} \lambda^{\theta+\alpha-1}(1-\lambda)^{n-\theta+\beta-1}$$

⋆⋆ Suppose we want to simulate θ from $\pi(\theta \mid x)$.

⋆⋆ We can find that the marginal distribution of θ is a beta-binomial distribution (n, α, β),

$$\pi(\theta \mid x) = \binom{n}{\theta} \frac{B(\alpha+\theta, \beta+n-\theta)}{B(\alpha, \beta)}.$$

It is not easy to simulate θ from $\pi(\theta \mid x)$.

$$\pi(\theta \mid x) = \int \pi(\theta, \lambda \mid x)\, d\lambda$$

$$\propto \int \binom{n}{\theta} \lambda^{\theta+\alpha-1} (1-\lambda)^{n-\theta+\beta-1}\, d\lambda$$

$$\propto \binom{n}{\theta} B(\theta+\alpha, n-\theta+\beta)$$

$\pi(\theta \mid x)$ is the Beta-Binomial distribution w/ (n, α, β)

$$\pi_1(\lambda \mid x) \propto \sum_{\theta=0}^{n} \binom{n}{\theta} \lambda^{\theta+\alpha-1} (1-\lambda)^{n-\theta+\beta-1}$$

$$\propto \lambda^{\alpha-1} \cdot (1-\lambda)^{\beta-1} \underbrace{\sum_{\theta=0}^{n} \binom{n}{\theta} \lambda^{\theta} (1-\lambda)^{n-\theta}}_{=1}$$

$$\propto \lambda^{\alpha-1} (1-\lambda)^{\beta-1}$$

$\pi_1(\lambda \mid x)$ is $Be(\alpha, \beta)$.

$$\pi_2(\theta \mid \lambda, x) \propto \pi(\theta, \lambda \mid x)$$

$$\propto \binom{n}{\theta} \lambda^{\theta+\alpha-1} (1-\lambda)^{n-\theta+\beta-1}$$

$$\propto \binom{n}{\theta} \lambda^{\theta} (1-\lambda)^{n-\theta}$$

$\Rightarrow$ $\pi_2(\theta \mid \lambda, x)$ is $Binom(n, \underline{\lambda})$

- **Example 6.3.4:** (contd) Consider $(\theta, \lambda) \in \mathbb{N} \times [0, 1]$ and

$$\pi(\theta, \lambda \mid x) \propto \binom{n}{\theta} \lambda^{\theta+\alpha-1}(1-\lambda)^{n-\theta+\beta-1}$$

⋆⋆ Alternatively, we utilize the hierarchical structure,

1. Simulate $\lambda^{(t)}$ from $\pi_2(\lambda \mid x) = \text{Be}(\alpha, \beta)$.
2. Simulate $\theta^{(t)}$ from $\pi_1(\theta \mid x, \lambda^{(t)}) = \text{Binom}(n, \lambda^{(t)})$.

⋆⋆ We obtain a sample of (θ, λ) from $\pi(\theta, \lambda \mid x)$.

⋆⋆ A sample of $\{\theta^{(t)}\}$ can be used to infer $\pi(\theta \mid x)$.

† Let's reconsider

$$\pi(\theta \mid x) = \int \underbrace{\pi_1(\theta \mid x, \lambda)}\underbrace{\pi_2(\lambda \mid x)}d\lambda.$$

- How can we simulate θ from $\pi(\theta \mid x)$ if $\underline{\pi_2(\lambda \mid x)}$ is not available?
- Often both $\pi_1(\theta \mid x, \lambda)$ and $\pi_2(\lambda \mid x, \theta)$ can be simulated.
- Possible to simulate θ using the conditionals, $\pi_1(\theta \mid x, \lambda)$ and $\pi_2(\lambda \mid x, \theta)$.

† The Gibbs sampler (CR 6.3.3, BDA Section 11.1 and PH Chapter 6)

1. Start with an arbitrary value $\lambda^{(0)}$.
2. Given $\lambda^{(t-1)}$, $t = 1, \ldots, T$, generate
   2.1 $\theta^{(t)}$ from $\pi_1(\theta \mid x, \lambda^{(t-1)})$.
   2.2 $\lambda^{(t)}$ from $\pi_2(\lambda \mid x, \theta^{(t)})$.

$\Rightarrow$ $\{(\theta^{(t)}, \lambda^{(t)}), t = 1, \ldots, T\}$ is a sample of (θ, λ) from their joint distribution.

$\Rightarrow$ $\{\theta^{(t)}, t = 1, \ldots, T\}$ is a sample of θ from its marginal distribution.

$\Rightarrow$ $\{\lambda^{(t)}, t = 1, \ldots, T\}$ is a sample of λ from its marginal distribution.

- **Example 6.3.4:** Consider $(\theta, \lambda) \in \mathbb{N} \times [0,1]$ and

$$\pi(\theta, \lambda \mid x) \propto \binom{n}{\theta} \lambda^{\theta+\alpha-1}(1-\lambda)^{n-\theta+\beta-1}$$

We can simulate θ and λ using Gibbs sampling as follows;

⋆⋆ Recognize

$$\theta \mid x, \lambda \sim \text{Binom}(n, \lambda), \quad \lambda \mid x, \theta \sim \text{Be}(\alpha+\theta, \beta+n-\theta).$$

⋆⋆ Iteratively sample $\theta^{(t)}$ and $\lambda^{(t)}$ from their full conditionals.

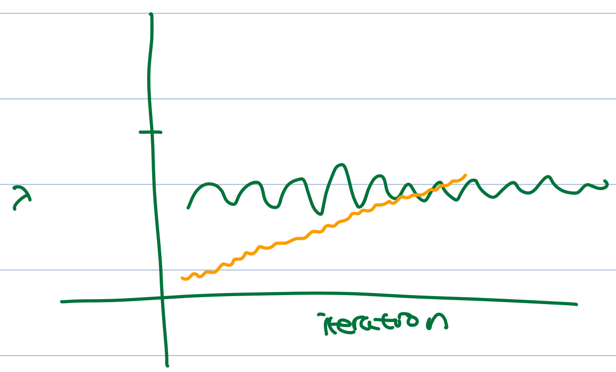
$$\pi(\theta, \lambda \mid x) \propto \binom{n}{\theta} \lambda^{\theta+\alpha-1} (1-\lambda)^{n-\theta+\beta-1}$$

① $\pi_1(\theta \mid \lambda, x) \propto \binom{n}{\theta} \lambda^{\theta} (1-\lambda)^{n-\theta}$

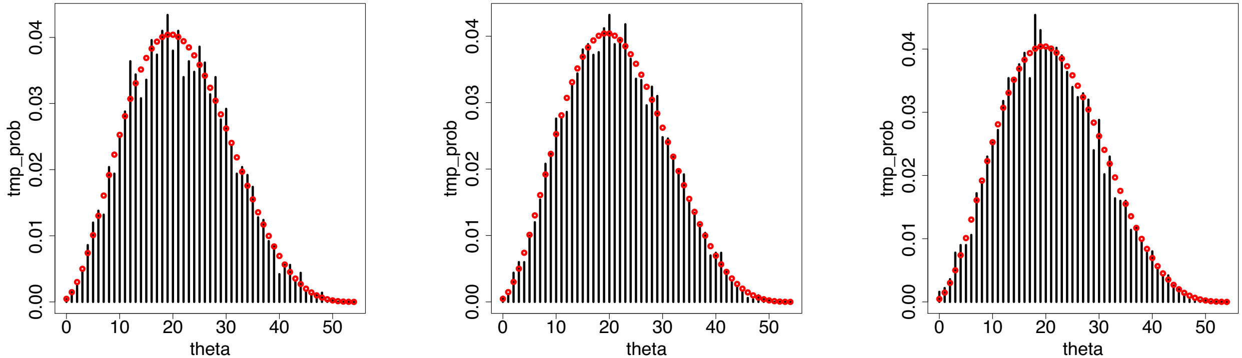
$\Rightarrow \quad \text{Binom}(n, \lambda)$

② $\pi_2(\lambda \mid \theta, x) \propto \lambda^{\theta+\alpha-1} (1-\lambda)^{n-\theta+\beta-1}$

$\Rightarrow \quad Be(\theta+\alpha, \; n-\theta+\beta)$



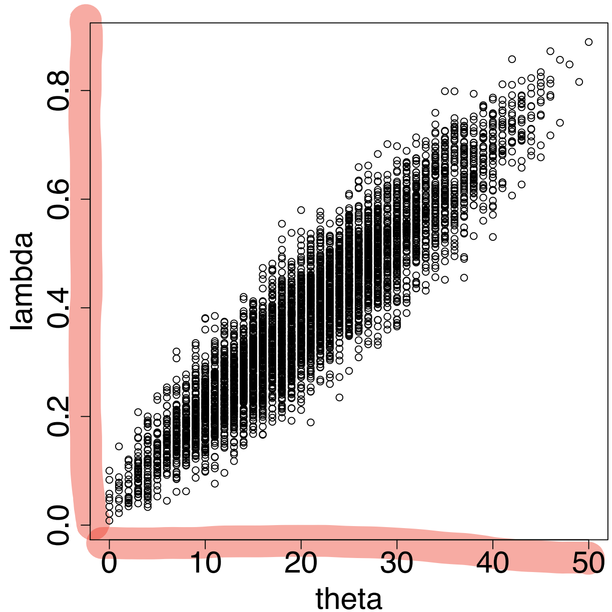
- **Example 6.3.4:** (contd) Marginal distribution $\pi(\theta \mid x)$.



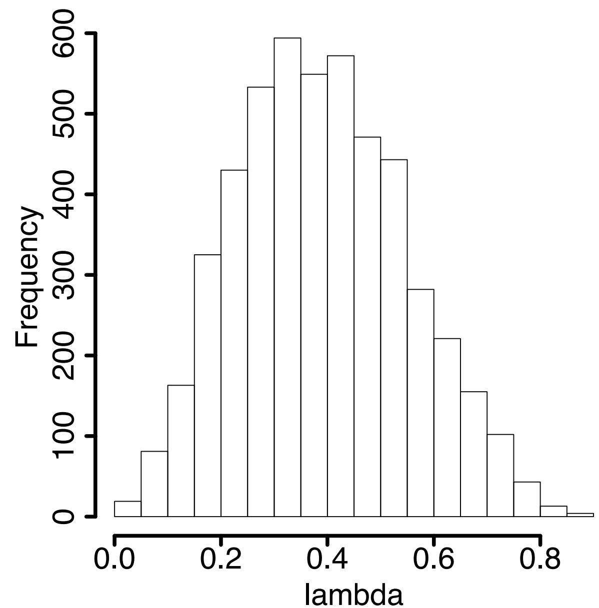
(a) (b) (c)

- (a) Directly from the marginal
- (b) Using the hierarchical structure
- (c) Using Gibbs sampling

- **Example 6.3.4:** (contd) More from Gibbs sampling



(a) Joint of θ and λ



(b) Marginal of λ

† The General Gibbs Sampler (CR 6.3.5)

- *alternating conditional sampling:* Each iterations of the Gibbs sampler cycles through the subvectors of θ, drawing each subset conditional on the value of all the others.
- Suppose the parameter vector θ has been divided into d components, $\theta = (\theta_1, \ldots, \theta_d)$. Let θ_{-j} all the components of θ except θ_j.
- There are d steps in an iteration.
- At each iteration m, an ordering of the d subvectors of θ is chosen and each θ_j is sampled from the conditional distributions given all the other components of θ, $\pi(\theta_j \mid \theta_{-j}, x)$.

For each iteration,

① $\pi(\underline{\theta_1^{(t)}} \mid \theta_2^{(t-1)}, \ldots, \theta_d^{(t-1)}, x)$

② $\pi(\underline{\theta_2^{(t)}} \mid \theta_1^{(t)}, \theta_3^{(t-1)}, \ldots, \theta_d^{(t-1)}, x)$

⋮

③ $\pi(\theta_d^{(t)} \mid \theta_1^{(t)}, \ldots, \theta_{d-1}^{(t-1)}, x)$

- **Example 6.3.9:** (Example 6.3.4 contd) Consider $(\theta, \lambda, n) \in \underset{\{0,\ldots,n\}}{\mathbb{N}} \times [0,1] \times \underset{\{0,\ldots,\infty\}}{\mathbb{N}}$, $n \geq \theta$ and

$$\pi(\theta, \lambda, n \mid x) \propto \binom{n}{\theta} \lambda^{\theta+\alpha-1} (1-\lambda)^{n-\theta+\beta-1} e^{-\xi} \frac{\xi^n}{n!}.$$

★★ The marginal distribution of θ cannot be derived.

★★ To obtain an estimate of $\pi(\theta \mid x)$, we can simulate θ, λ and n using Gibbs sampling. The full conditionals are

$$\begin{aligned}
\theta \mid x, \lambda, \xi &\sim \text{Binom}(n, \lambda), \\
\lambda \mid x, \theta, \xi &\sim \text{Be}(\theta+\alpha, n-\theta+\beta), \\
n - \theta \mid x, \theta, \lambda &\sim \text{Poi}(\xi(1-\lambda)).
\end{aligned}$$

$$\pi(\theta, \lambda, n \mid x) \propto \binom{n}{\theta} \lambda^{\theta+\alpha-1} (1-\lambda)^{n-\theta+\beta-1} e^{-\xi} \frac{\xi^n}{n!}$$

$[\theta | \lambda | n\}$ $d = 3$

①

$$\pi_1(\theta \mid \underline{\lambda}, \underline{n}, x) \propto \pi(\theta, \lambda, n \mid x)$$

$$\propto \binom{n}{\theta} \lambda^{\theta+\alpha-1} (1-\lambda)^{n-\theta+\beta-1} e^{-\xi} \frac{\xi^n}{n!}$$

$$\propto \binom{n}{\theta} \lambda^{\theta} (1-\lambda)^{n-\theta}$$

$Bin(\underline{n}, \underline{\lambda})$

②

$$\pi_2(\underline{\lambda} \mid \underline{\theta}, \underline{n}, x) \propto \pi(\theta, \lambda, n \mid x)$$

$$\propto \underline{\binom{n}{\theta}} \, \underline{\lambda^{\theta+\alpha-1}} \, \underline{(1-\lambda)^{n-\theta+\beta-1}} e^{-\xi} \frac{\xi^n}{n!}$$

$$\propto \lambda^{\theta+\alpha-1} (1-\lambda)^{n-\theta+\beta-1}$$

$Be(\underline{\theta}+\alpha, \underline{n}-\underline{\theta}+\beta)$

③

$$\pi_3(n \mid \theta, \lambda, x) \propto \pi(\theta, \lambda, n \mid x)$$

$$\propto \cancel{\binom{n}{\theta}} \lambda^{\theta+\alpha-1} \cancel{(1-\lambda)}^{n-\theta+\beta-1} e^{-\xi} \frac{\xi^n}{\cancel{n!}}$$

$$\propto \frac{\cancel{n!}}{\cancel{\theta!}(n-\theta)!} (1-\lambda)^n \cdot \frac{\xi^n}{\cancel{n!}}$$

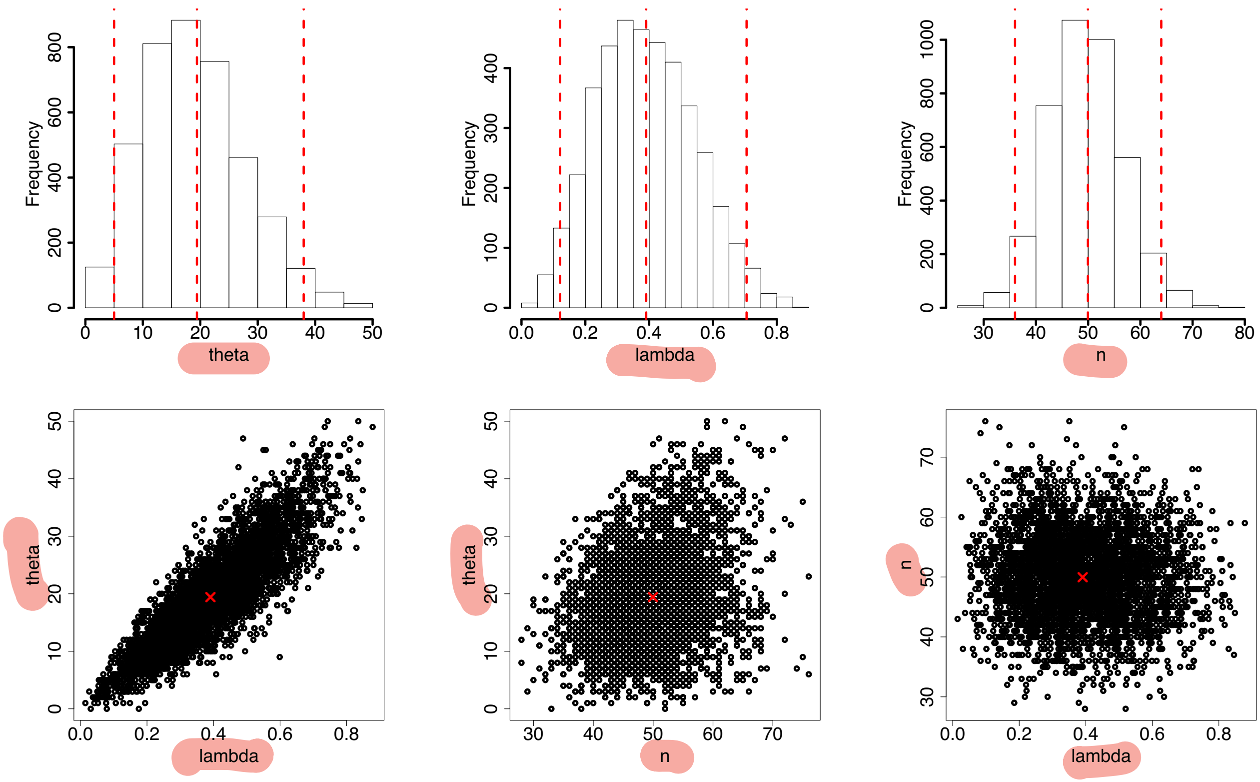
$$\propto \frac{((1-\lambda)\xi)^n}{(n-\theta)!}$$

$$\propto \frac{((1-\lambda)\xi)^{n-\theta}}{(n-\theta)!} \qquad (1-\lambda\xi)^{-\theta}$$

$n-\theta \mid \theta, \lambda, \cancel{X} \sim \mathrm{Poi}\left((1-\lambda)\xi \right)$

$(\theta, \lambda, n)^{(t)}$ from $\pi(\theta, \lambda, n \mid x)$

- **Example 6.3.9:** (contd)



† Building Markov chain algorithms using the Gibbs sampler and Metropolis algorithm

- We use the Gibbs sampler and the Metropolis algorithms as building blocks for simulating from complicated distributions.

  ⋆⋆ Use the Gibbs sampler for conditionally conjugate models.

  ⋆⋆ Use the Metropolis algorithm for models that are not conditionally conjugate.

- The Metropolis algorithm can be

  ⋆⋆ performed in vector form – moving in the multi-dimensional space

  ⋆⋆ embedded within a Gibbs sampler structure, by alternately updating one parameter at a time.

- When parameters are highly correlated in the target distribution, conditional sampling algorithms can be slow.

• **Example 6.3.2:** (cond) Suppose x_i is a random sample of size n from the Weibull distribution

$$f(x \mid \alpha, \eta) \propto \alpha\eta x^{\alpha-1} e^{-x^{\alpha}\eta}.$$

For $\theta = (\alpha, \eta)$, consider the prior

$$\pi(\theta) \propto e^{-\alpha}\eta^{\beta-1}e^{-\xi\eta}.$$

That is, assume a priori independence and place E(1) and Gamma(β, ξ) (with mean β/ξ) for α and η, respectively.

$\pi_1(\alpha \mid \eta, x)$

$\pi_2(\eta \mid \alpha, x)$

- **Example 6.3.2:** (contd)

⋆⋆ Find the posterior distribution of θ.

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

⋆⋆ Simulate θ from $\pi(\theta \mid \boldsymbol{x})$ using the *Gibbs sampler*. First, derive the full conditionals;

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

$z_1 = \log(\alpha)$ $\alpha = e^{z_1}$ $\alpha \in \mathbb{R}^+$

$\tilde{\pi}(z_1 \mid \eta, x) \propto \pi(e^{z_1} \mid \eta, x) \cdot |J|$

$\xi = z_1^{(t)} + \varepsilon$

$\alpha = \min\left\{ \dfrac{\tilde{\pi}(\xi \mid \eta, x)}{\tilde{\pi}(z_1^{(t)} \mid \eta, x)}, 1 \right\}$

$z_1^{(t+1)} = \xi$ w/p α

$= z_1^{(t)}$ w/p $1-\alpha$

$\Rightarrow \alpha^{(t+1)} = e^{z_1^{(t+1)}}$

$\theta \in (0, 1)$

$\log\left(\frac{\theta}{1-\theta}\right) \in \mathbb{R}$

- **Example 6.3.2:** (contd)

1. Start with an arbitrary value $\eta^{(0)}$.

2. Iterate the following steps, $t = 1, \ldots, T$

   2.1 Given $\eta^{(t-1)}$, simulate $\alpha^{(t)}$ from $\pi_1(\alpha \mid x, \eta^{(t-1)})$: use a MH step

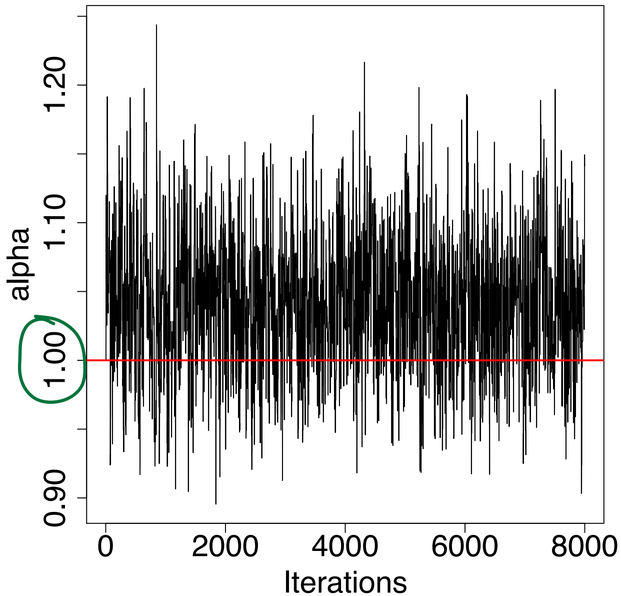
   2.2 Given $\alpha^{(t)}$, simulate $\eta^{(t)}$ from $\pi_2(\eta \mid x, \alpha^{(t)})$: use a MH step

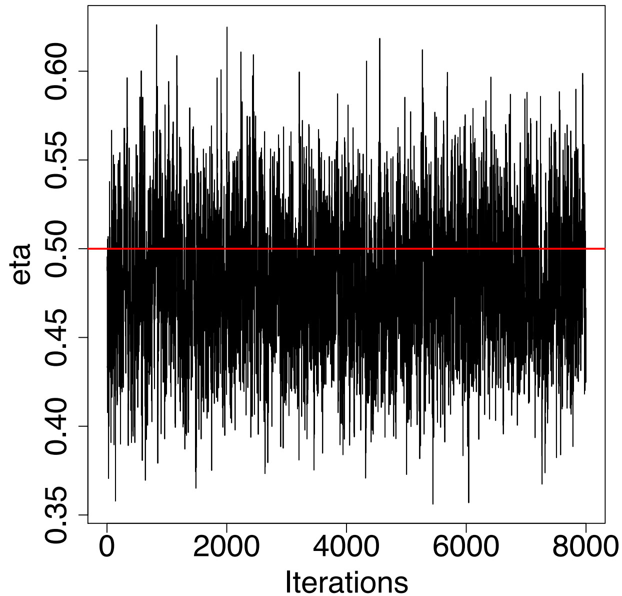
3. Do burn-in and thinning as needed.

   $\{(\alpha^{(t)}, \eta^{(t)}), t=1,\ldots,T\}$

- **Example 6.3.2:** (contd)

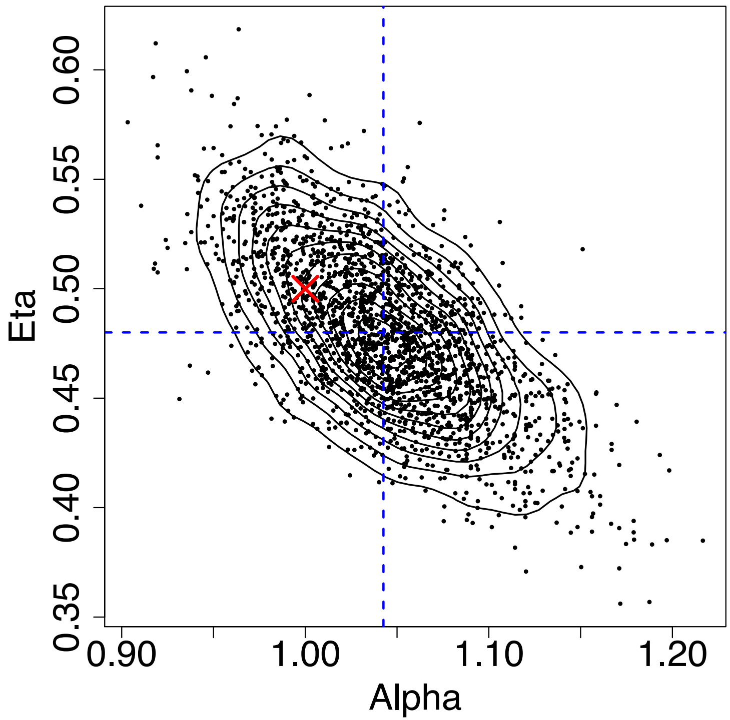
∗ Trace plots to check the MCMC (mixing, convergence...)





- **Example 6.3.2:** (contd)

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



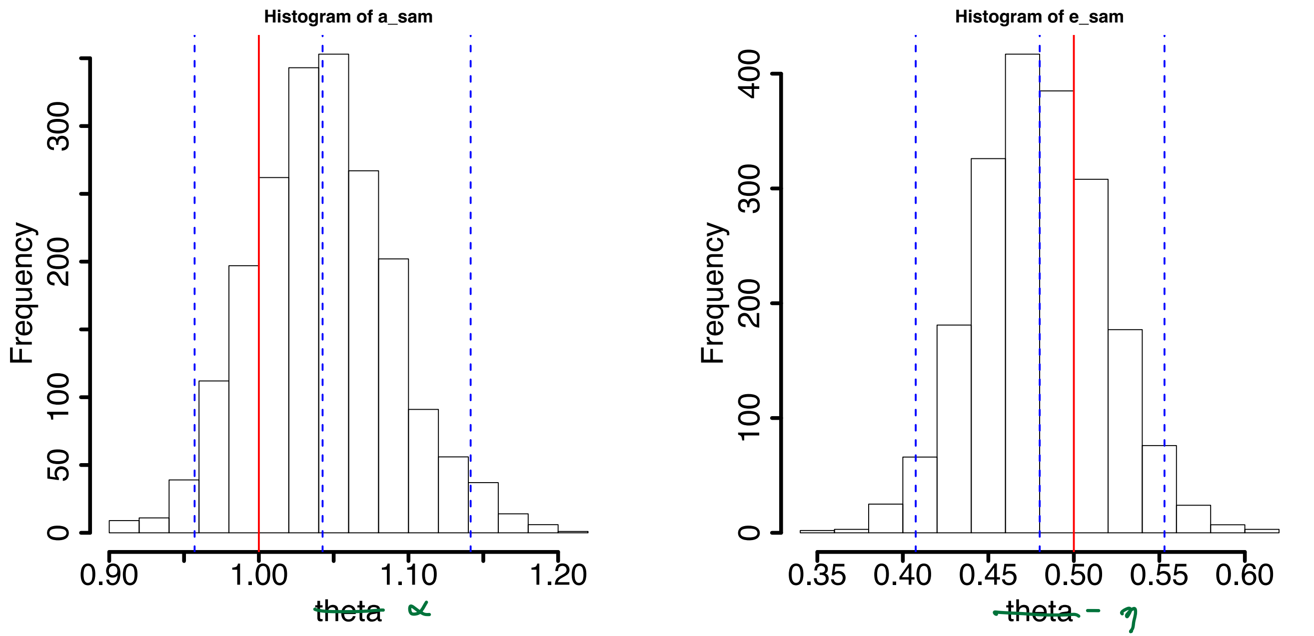
- **Example 6.3.2:** (contd)

⋆⋆ Report a posterior summary.

```
> post_m_a
[1] 1.04259
> post_sd_a
[1] 0.04684268
> ci_a
     2.5%     97.5%
0.9570524 1.1415434
>
> post_m_e
[1] 0.4800338
> post_sd_e
[1] 0.03730137
> ci_e
     2.5%     97.5%
0.4075282 0.5530123
>
```

- **Example 6.3.2:** (contd)

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



STAT 206B
Chapter 4: Bayesian Point Estimation
Chapter 5: Hypothesis Testing & Confidence Regions

Winter 2022

† Bayesian Inference

- The posterior distribution supposedly contains all the available information about θ.

- The *entire* posterior distribution $\pi(\theta \mid x)$ is the extensive summary of the information available on the parameter θ.

- A visual inspection of the graph of the posterior will often provide the best insight concerning θ (at least in low dimensions)

- More standard uses of the posterior are still helpful e.g. point estimation, interval estimation, testing, prediction...

- CR Chapter 4 and JB Chapter 4.3

† **Bayesian Point Estimation:** the simplest inferential use of the posterior distribution

- Report a point estimate for $h(\theta)$, with an associated measure of accuracy

  $\Rightarrow$ Find $\pi(h(\theta) \mid x)$ and then the *Bayes rule* d, i.e., a solution of

  $$\min \mathsf{E}^{\pi}\left\{L(\theta, d) \mid x\right\} \quad \text{for } d \in \mathcal{D} \text{ and } \theta \in \Theta.$$

  ⋆⋆ Recall we found the Bayes actions under standard loss functions such as the quadratic loss, the absolute error loss and the 0-1 loss.

  ⋆⋆ The mean and median of the posterior are frequently better estimates of θ than the mode (i.e., MAP).

† Estimation Error

- We evaluate the precision of $\delta^\pi(x)$
- For example, we may use the posterior squared error:

$$E^\pi[(\underbrace{\delta^\pi(x)} - h(\theta))^2 \mid x].$$

⋆⋆ If we use $E^\pi[h(\theta) \mid x]$ as the estimate of $h(\theta)$, report $\sqrt{\mathrm{Var}^\pi(h(\theta) \mid x)}$ as the standard error (posterior standard deviation).

- **JB Example 1** (p136) Consider the situation wherein a child is given an intelligence test. Assume that the tet result X is $\mathrm{N}(\theta, 100)$, where θ is the true IQ (intelligence) level of the child, as measured by the test. Assume also that, in the population as a whole, θ is distributed according to a $\mathrm{N}(100, 225)$ distribution. Suppose that we observe a student who scores 115 on the test.

  $x=115$

⋆⋆ We can find

$$\theta \mid x \sim \mathrm{N}((1/100+1/225)^{-1}(\overset{115}{x}/100+100/225), (1/100+1/225)^{-1}).$$

$\Rightarrow$ $\mu^{\pi}(115) = 110.39$ and $\sqrt{V^{\pi}(115)} = \sqrt{69.23} = 8.32$.

- **JB Example 8**(p137) Assume $X \sim \mathrm{N}(\theta, \sigma^2)$ (σ^2 known) and the noninformative prior $\pi(\theta) = 1$ is used, then the posterior distribution of θ given x is $\mathrm{N}(x, \sigma^2)$. Hence the posterior mean is $\mu^\pi(x) = x$ and the posterior variance and standard deviation are σ^2 and σ, respectively.

  ⋆⋆ The same as the usual classical estimate with standard error.

  ⋆⋆ Their interpretations are different!

- Sampling Properties
  - ⋆⋆ Sampling properties: behavior of an estimator under hypothetically repeatable surveys or experiments.
  - ⋆⋆ Suppose θ_0 = the true value of the population mean.
  - ⋆⋆ To evaluate how close an estimator $\delta(x)$ is likely to be to θ_0, we use the mean square error(MSE)

$$
\begin{aligned}
\mathrm{MSE}(\delta \mid \theta_0) &= \mathrm{E}\{(\delta - \theta_0)^2 \mid \theta_0\} \\
&= \mathrm{E}\{(\delta - m)^2 \mid \theta_0\} + \mathrm{E}\{(m - \theta_0)^2 \mid \theta_0\} \\
&= \mathrm{Var}(\delta \mid \theta_0) + \mathrm{Bias}^2(\delta \mid \theta_0),
\end{aligned}
$$

where $m = \mathrm{E}(\delta \mid \theta_0)$

- **PH p82** Recall the IQ example (similar but different!).
  - ⋆⋆ $X \sim \mathrm{N}(100, 225)$ for the general population.
  - ⋆⋆ Suppose that we sample n individuals from a particular town and estimate θ, the town-specific mean IQ score based on the sample of size n.
  - ⋆⋆ In fact, people in the town are extremely exceptional so $\theta_0 = 112$ and $\sigma^2 = 169$.
  - ⋆⋆ Consider $x_i \mid \theta \overset{iid}{\sim} \mathrm{N}(\theta, \sigma^2)$, where $\sigma^2 = 169$ but θ is unknown
  - ⋆⋆ Assume $\theta \sim \mathrm{N}(\mu_0, \tau_0^2)$, where $\tau_0 = \sigma / \sqrt{\kappa_0}$
  - ⋆⋆ For Bayesian inference, we lack the information about the town a natural choice of $\mu_0 = 100$.

- $$\delta_n^{\pi}(x) = \frac{\tau_0^2}{\tau_0^2 + \sigma^2/n}\,\bar{x} + \frac{\sigma^2/n}{\tau_0^2 + \sigma^2/n}\,\mu_0$$

$$\delta_n(x) = \bar{x}$$