

# INTRODUCTION TO BAYESIAN STATISTICS

Sarat C. Dass

Department of Statistics & Probability

Department of Computer Science & Engineering

Michigan State University

## TOPICS

- The Bayesian Framework
- Different Types of Priors
- Bayesian Calculations
- Hypothesis Testing
- Bayesian Robustness
- Hierarchical Analysis
- Bayesian Computations
- Bayesian Diagnostics And Model Selection

## FRAMEWORK FOR BAYESIAN STATISTICAL INFERENCE

- Data:  $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$  (realization:  $\mathbf{y} \in \mathbb{R}^n$ )
- Parameter:  $\Theta = (\theta_1, \theta_2, \dots, \theta_p) \in \mathbb{R}^p$
- Likelihood:  $L(\mathbf{y} | \Theta)$
- Prior:  $\pi_0(\Theta)$

- Thus, the joint distribution of  $\mathbf{y}$  and  $\Theta$  is

$$\pi(\mathbf{y}, \Theta) = L(\mathbf{y} | \Theta) \cdot \pi_0(\Theta)$$

- Bayes formula:  $A$  is a set, and  $B_1, B_2, \dots, B_k$  is a partition of the space of  $(\mathbf{Y}, \Theta)$ . Then,

$$P(B_j | A) = \frac{P(A | B_j) \cdot P(B_j)}{\sum_{j=1}^k P(A | B_j) \cdot P(B_j)}$$

Consider  $A = \{\mathbf{y}\}$  and  $B_j = \{\Theta \in \mathcal{P}_j\}$ , where  $\mathcal{P}_j$  is a partition of  $R^p$ . Taking finer and finer partitions with  $k \rightarrow \infty$ , we get the limiting form of Bayes theorem:

$$\pi(\Theta | \mathbf{y}) \equiv \frac{L(\mathbf{y} | \Theta) \cdot \pi_0(\Theta)}{\int L(\mathbf{y} | \Theta) \cdot \pi_0(\Theta) d\Theta}$$

is called the posterior distribution of  $\Theta$  given  $\mathbf{y}$ .

- We define

$$m(\mathbf{y}) \equiv \int L(\mathbf{y} | \Theta) \cdot \pi_0(\Theta) d\Theta$$

as the marginal of  $\mathbf{y} = P(A)$ , by “summing” over the infinitesimal partitions  $B_j, j = 1, 2, \dots$

- We can also write

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior} \quad (1)$$

$$= L(\mathbf{y} | \Theta) \cdot \pi_0(\Theta), \quad (2)$$

retaining the terms on the RHS that involve  $\Theta$  components. The other terms are constants and cancel out from the numerator and denominator.

## INFERENCE FROM THE POSTERIOR DISTRIBUTION

- The posterior distribution is the MAIN tool of inference for Bayesians.
- Posterior mean:  $E(\Theta | \mathbf{y})$ . This is a point estimate of  $\Theta$ .
- Posterior variance - to judge the uncertainty in  $\Theta$  after observing  $\mathbf{y}$ :  $V(\Theta | \mathbf{y})$
- HPD Credible sets:

Suppose  $\Theta$  is one dimensional. The  $100(1 - \alpha)\%$  credible interval for  $\Theta$  is given by the bounds  $l(\mathbf{y})$  and  $u(\mathbf{y})$  such that

$$P\{l(\mathbf{y}) \leq \theta \leq u(\mathbf{y}) | \mathbf{y}\} = 1 - \alpha$$

Shortest length credible sets can be found using the highest posterior density (HPD) criteria:

Define:  $A_u = \{\theta : \pi(\theta | \mathbf{y}) \geq u\}$  and find  $u_0$  such that

$$P(A_{u_0}) = 1 - \alpha.$$

## SOME EXAMPLES

### EXAMPLE 1: NORMAL LIKELIHOOD WITH NORMAL PRIOR

- $Y_1, Y_2, \dots, Y_n$  are independent and identically distributed  $N(\theta, \sigma^2)$  observations. The mean  $\theta$  is the unknown parameter of interest.

- $\Theta = \{\theta\}$ . Prior on  $\Theta$  is  $N(\theta_0, \tau^2)$ :

$$\pi_0(\theta) = \frac{1}{\tau\sqrt{2\pi}} \exp\left\{-\frac{(\theta - \theta_0)^2}{2\tau^2}\right\}.$$

- $\mathbf{y} = (y_1, y_2, \dots, y_n)$ . Likelihood:

$$L(\mathbf{y} | \theta) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y_i - \theta)^2}{2\sigma^2}\right\}$$

- Posterior:

$$\pi(\theta | \mathbf{y}) \propto L(\mathbf{y} | \theta)\pi_0(\theta)$$

$$\propto \exp\left\{-\sum_{i=1}^n (y_i - \theta)^2 / (2\sigma^2)\right\} \exp\left\{-\frac{\theta^2}{2\tau^2}\right\}.$$

- After some simplifications, we have

$$\pi(\theta | \mathbf{y}) = N(\hat{\theta}, \hat{\sigma}^2)$$

where

$$\hat{\theta} = \left( \frac{n}{\sigma^2} + \frac{1}{\tau^2} \right)^{-1} \left( \frac{n}{\sigma^2} \bar{y} + \frac{1}{\tau^2} \theta_0 \right)$$

and

$$\hat{\sigma}^2 = \left( \frac{n}{\sigma^2} + \frac{1}{\tau^2} \right)^{-1}$$

## POSTERIOR INFERENCE:

- Posterior mean =  $\hat{\theta}$ .
- Posterior variance =  $\hat{\sigma}^2$ .
- 95% Posterior HPD credible set:  $l(\mathbf{y}) = \hat{\theta} - z_{0.975} \hat{\sigma}$  and  $u(\mathbf{y}) = \hat{\theta} + z_{0.975} \hat{\sigma}$ , where  $\Phi(z_{0.975}) = 0.975$ .

## EXAMPLE 2: BINOMIAL LIKELIHOOD WITH BETA PRIOR

- $Y_1, Y_2, \dots, Y_n$  are iid Bernoulli random variables with success probability  $\theta$ . Think of tossing a coin with  $\theta$  as the probability of turning up heads.

- Parameter of interest is  $\theta$ ,  $0 < \theta < 1$ .

- $\Theta = \{\theta\}$ . Prior on  $\Theta$  is  $Beta(\alpha, \beta)$ :

$$\pi_0(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}.$$

- $\mathbf{y} = (y_1, y_2, \dots, y_n)$ . Likelihood:

$$L(\mathbf{y} | \theta) = \prod_{i=1}^n \theta^{I(y_i=1)} (1 - \theta)^{I(y_i=0)}$$

- Posterior:

$$\begin{aligned} \pi(\theta | \mathbf{y}) &\propto L(\mathbf{y} | \theta) \pi_0(\theta) \\ &\propto \theta^{\sum_{i=1}^n y_i + \alpha - 1} (1 - \theta)^{n - \sum_{i=1}^n y_i + \beta - 1}. \end{aligned}$$

Note that this is  $Beta(\hat{\alpha}, \hat{\beta})$  with new parameters  $\hat{\alpha} = \sum_{i=1}^n y_i + \alpha$  and  $\hat{\beta} = n - \sum_{i=1}^n y_i + \beta$ .

## POSTERIOR INFERENCE

$$\text{Mean} = \hat{\theta} = \frac{\hat{\alpha}}{\hat{\alpha} + \hat{\beta}} = \frac{n\bar{y} + \alpha}{n + \alpha + \beta}$$

$$\text{Variance} = \frac{\hat{\alpha}\hat{\beta}}{(\hat{\alpha} + \hat{\beta})^2(\hat{\alpha} + \hat{\beta} + 1)} = \frac{\hat{\theta}(1 - \hat{\theta})}{n + \alpha + \beta}$$

Credible sets: Needs to be obtained numerically. Assume  $n = 20$  and  $\bar{y} = 0.2$ . Set  $\alpha = \beta = 1$ .

$$l(\mathbf{y}) = 0.0692 \quad \text{and} \quad u(\mathbf{y}) = 0.3996$$

## BAYESIAN CONCEPTS

- In Examples 1 and 2, the posterior was obtained in a nice closed form. This was due to *conjugacy*.
- Definition of conjugate priors: Let  $\mathcal{P}$  be a class of densities. The class  $\mathcal{P}$  is said to be conjugate for the likelihood  $L(\mathbf{y} | \Theta)$  if for every  $\pi_0(\Theta) \in \mathcal{P}$ , the posterior  $\pi(\Theta | \mathbf{y}) \in \mathcal{P}$ .
- Other examples of conjugate families include multivariate analogues of Examples 1 and 2:
  1.  $Y_i$ 's are iid  $MVN(\theta, \Sigma)$  and  $\theta$  is  $MVN(\theta_0, \tau^2)$ .
  2.  $Y_i$ 's are iid  $Multi(1, \theta_1, \theta_2, \dots, \theta_k)$  and  $(\theta_1, \theta_2, \dots, \theta_k)$  is  $Dirichlet(\alpha_1, \alpha_2, \dots, \alpha_k)$ .
  3.  $Y_i$ 's are iid Poisson with mean  $\theta$  and  $\theta$  is  $Gamma(\alpha, \beta)$ .
- Improper priors. In order to be completely objective, some Bayesians use improper priors as candidates for  $\pi_0(\Theta)$ .

## IMPROPER PRIORS

- Improper priors represent lack of knowledge of  $\theta$ . Examples of improper priors include:

1.  $\pi_0(\Theta) = c$  for an arbitrary constant  $c$ . Note that  $\int \pi_0(\Theta) d\Theta = \infty$ . This is not a proper prior. We must make sure that

$$m(\mathbf{y}) = \int L(\mathbf{y} | \Theta) \cdot d\Theta < \infty.$$

For Example 1, we have  $\hat{\theta} = \bar{y}$  and  $\hat{\sigma}^2 = \frac{\sigma^2}{n}$

For Example 2, the prior that represents lack of knowledge is  $\pi_0(\Theta) = \text{Beta}(1, 1)$ .

- Hierarchical priors. When  $\Theta$  is multidimensional, take

$$\begin{aligned} \pi_0(\Theta) = & \pi_0(\theta_1)\pi_0(\theta_2 | \theta_1) \cdot \pi_0(\theta_3 | \theta_1, \theta_2) \\ & \cdots \pi_0(\theta_p | \theta_1, \theta_2, \cdots, \theta_{(p-1)}). \end{aligned}$$

We will see two examples of hierarchical priors later on.

## NON-CONJUGATE PRIORS

- What if we use priors that are non-conjugate?
- In this case the posterior cannot be obtained in a closed form, and so we have to resort to numerical approximations.

### EXAMPLE 3: NORMAL LIKELIHOOD WITH CAUCHY PRIOR

- Let  $Y_1, Y_2, \dots, Y_n \stackrel{i.i.d}{\sim} N(\theta, 1)$  where  $\theta$  is the unknown parameter of interest.

- $\Theta = \{\theta\}$ . Prior on  $\Theta$  is  $C(0, 1)$ :

$$\pi_0(\theta) = \frac{1}{\pi(1 + \theta^2)}.$$

- Likelihood:

$$L(y_1, y_2, \dots, y_n | \theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\{-(y_i - \theta)^2/2\}$$

- The marginal  $m(\mathbf{y})$  is given by

$$m(\mathbf{y}) = \int_{\theta \in \mathcal{R}} \frac{1}{(1 + \theta^2)} \exp\{-n(\bar{y} - \theta)^2/2\} d\theta.$$

- Note that the above integral can not be derived analytically.

- Posterior:

$$\begin{aligned} \pi(\theta | \mathbf{y}) &= \frac{L(\mathbf{y}|\theta)\pi_0(\theta)}{m(\mathbf{y})} \\ &= \frac{1}{m(\mathbf{y})} \exp\{-n(\bar{y} - \theta)^2/2\} \frac{1}{(1 + \theta^2)} \end{aligned}$$

## BAYESIAN CALCULATIONS

- NUMERICAL INTEGRATION

Numerically integrate the quantities

$$\int_{\theta \in R} h(\theta) \pi(\theta | \mathbf{y}) d\theta$$

- ANALYTIC APPROXIMATION

The idea here is to approximate the posterior distribution with an appropriate normal distribution.

$$\begin{aligned} \log(L(\mathbf{y} | \theta)) &\approx \log(L(\mathbf{y} | \theta^*)) \\ &\quad + (\theta - \theta^*) \frac{\partial}{\partial \theta} \log(L(\mathbf{y} | \theta^*)) \\ &\quad + \frac{(\theta - \theta^*)^2}{2} \frac{\partial^2}{\partial^2 \theta} \log(L(\mathbf{y} | \theta^*)) \end{aligned}$$

where  $\theta^*$  is the maximum likelihood estimate (MLE).

Note that  $\frac{\partial}{\partial \theta} \log(L(\mathbf{y} | \theta^*)) = 0$ , and so the posterior is approximately

$$\pi(\theta | \mathbf{y}) \approx \pi(\theta^* | \mathbf{y}) \cdot \exp \left\{ -\frac{(\theta - \theta^*)^2}{2\sigma^2} \right\}$$

where

$$\sigma^2 = - \left( \frac{\partial^2}{\partial^2 \theta} \log(L(\mathbf{y} | \theta^*)) \right)^{-1}$$

Posterior mean =  $\theta^*$  and posterior variance =  $\sigma^2$ .

- Let us look at a numerical example where  $n = 20$  and  $\bar{y} = 0.1$  for the Normal-Cauchy problem. This gives

$$\theta^* = \bar{y} = 0.1 \quad \text{and} \quad \sigma^2 = 1/n = 0.05$$

- MONTE CARLO INTEGRATION (will be discussed later in detail).

## BAYESIAN HYPOTHESIS TESTING

Consider  $Y_1, Y_2, \dots, Y_n$  iid with density  $f(y | \theta)$ , and the following null-alternative hypotheses:

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1$$

- To decide between  $H_0$  and  $H_1$ , calculate the posterior probabilities of  $H_0$  and  $H_1$ , namely,  $\alpha_0 = P(\Theta_0 | \mathbf{y})$  and  $\alpha_1 = P(\Theta_1 | \mathbf{y})$ .
- $\alpha_0$  and  $\alpha_1$  are actual (subjective) probabilities of the hypotheses in the light of the data and prior opinion.

## HYPOTHESIS TESTING (CONT.)

- Working method: Assign prior probabilities to  $H_0$  and  $H_1$ , say,  $\pi_0$  and  $\pi_1$ . Define

$$\begin{aligned} B(\mathbf{y}) &= \frac{\text{Posterior odds ratio}}{\text{Prior odds ratio}} \\ &= \frac{\alpha_0/\alpha_1}{\pi_0/\pi_1} \end{aligned}$$

is called the Bayes factor in favor of  $\Theta_0$ .

- In the case of simple vs. simple hypothesis testing,  $\Theta_0 = \{\theta_0\}$  and  $\Theta_1 = \{\theta_1\}$ , we get

$$\alpha_0 = \frac{\pi_0 f(\mathbf{y} | \theta_0)}{\pi_0 f(\mathbf{y} | \theta_0) + \pi_1 f(\mathbf{y} | \theta_1)},$$

$$\alpha_1 = \frac{\pi_1 f(\mathbf{y} | \theta_1)}{\pi_0 f(\mathbf{y} | \theta_0) + \pi_1 f(\mathbf{y} | \theta_1)}$$

and

$$B = \frac{\alpha_0/\alpha_1}{\pi_0/\pi_1} = \frac{f(\mathbf{y} | \theta_0)}{f(\mathbf{y} | \theta_1)}$$

- Note that  $B$  is the likelihood ratio in the case of simple testing.
- In general,  $B$  depends on prior input. Suppose

$$\pi_0(\Theta) = \begin{cases} \pi_0 \pi_{H_0}(\theta) & \text{if } \theta \in \Theta_0 \\ \pi_1 \pi_{H_1}(\theta) & \text{if } \theta \in \Theta_1 \end{cases}$$

then,

$$B = \frac{\int_{\Theta_0} f(\mathbf{y} | \theta) \pi_{0,H_0}(\theta) d\theta}{\int_{\Theta_1} f(\mathbf{y} | \theta) \pi_{0,H_1}(\theta) d\theta}$$

Also,

$$P(\Theta_0 | \mathbf{y}) = \frac{B}{B + 1}$$

$$P(\Theta_1 | \mathbf{y}) = \frac{1}{B + 1}$$

## BAYESIAN PREDICTIVE INFERENCE

- Let  $Y_1, Y_2, \dots, Y_n$  be independent and identically observations from the density  $f(y | \theta)$ .
- $Z$  is another random variable distributed according to  $g(z | \theta)$ .
- Aim is to predict  $Z$  based on  $\mathbf{Y}$ . MAIN tool of inference is the predictive distribution of  $Z$  given  $\mathbf{Y}$ :

$$\pi(z | \mathbf{y}) = \int_{\theta} g(z | \theta) \pi(\theta | \mathbf{y}) d\theta$$

- Estimate  $Z$  by  $E(Z | \mathbf{y})$  and corresponding variance by  $V(Z | \mathbf{y})$ .

- NORMAL-CAUCHY EXAMPLE: Let  $Z \sim N(\theta, 1)$ .  
Then

$$\begin{aligned} E(Z | \mathbf{y}) &= E(E(Z | \theta) | \mathbf{y}) \\ &= E(\theta | \mathbf{y}) \\ &= \text{Posterior mean of Example 3} \end{aligned}$$

and

$$\begin{aligned} V(Z | \mathbf{y}) &= V(E(Z | \theta) | \mathbf{y}) + E(V(Z | \theta) | \mathbf{y}) \\ &= V(\theta | \mathbf{y}) + 1 \\ &= 1 + \text{Posterior variance of Example 3} \end{aligned}$$

## BAYESIAN ROBUSTNESS

- Prior specification is subjective. How do we assess the influence of the prior on our analysis?
- Consider the Normal-Normal and Normal-Cauchy examples of Examples 1 and 3.
- $Y_1, Y_2, \dots, Y_n$  are iid from  $N(\theta, 1)$ , and we consider two priors:  $\pi_0^N$  and  $\pi_0^C$ , for the Normal and Cauchy priors, respectively.
- Recall the marginal distribution:  $m(\mathbf{y})$  gives higher values to reasonable priors on  $\Theta$ .
- Consider

$$H_0 : \pi_0 = \pi_0^N \quad \text{versus} \quad H_1 : \pi_0 = \pi_0^C$$

Assuming  $\pi_0 = \pi_1 = 0.5$ , we get

$$\begin{aligned} B(\mathbf{y}) &= \frac{\int_{\theta} L(\mathbf{y} | \theta) \pi_0^N(\theta) d\theta}{\int_{\theta} L(\mathbf{y} | \theta) \pi_0^C(\theta) d\theta} \\ &= \frac{m_N(\mathbf{y})}{m_C(\mathbf{y})} \end{aligned}$$

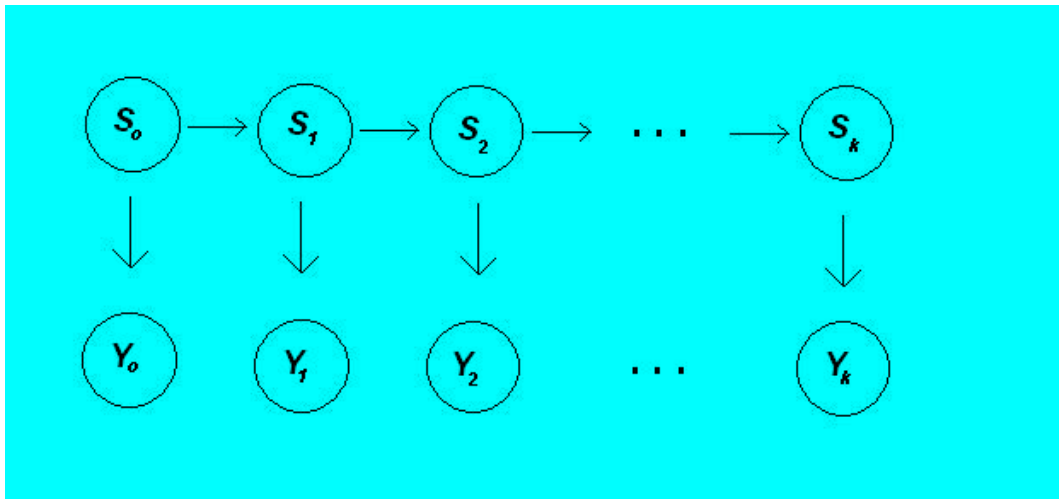
TABLE SHOWING B AS A FUNCTION OF  $\bar{y}$

- Let  $n = 3$ .

$\bar{y}$	0	1	2	3	4
$m_N(\mathbf{y})$	0.1424	0.0954	0.0287	0.0035	0.0001
$m_C(\mathbf{y})$	0.1070	0.0681	0.0282	0.0085	0.0003
$B(\mathbf{y})$	1.3303	1.4005	1.0187	0.4185	0.2358

# HIERARCHICAL BAYESIAN ANALYSIS

## EXAMPLE 4: HIDDEN MARKOV MODELS



The HMM model consists of:

$$\begin{aligned} \pi_0(S_0) &\sim N(0, 1), \\ \pi_0(S_j | s_{j-1}) &\sim N(s_{j-1}, 1), \quad j = 1, 2, \dots, k, \end{aligned}$$

and

$$p(y_j | s_j) \sim N(s_j, 1), \quad j = 0, 1, \dots, k.$$

- $\Theta = (S_0, S_1, \dots, S_k)$
- The likelihood is

$$L(\mathbf{y} | \Theta) = \prod_{j=1}^k p(y_j | s_j).$$

- The prior is

$$\pi_0(S_0, \dots, S_k) = \pi_0(S_0) \prod_{j=1}^k \pi_0(S_j | S_{j-1}).$$

Thus,

$$\begin{aligned} \text{Posterior} &\propto \pi_0(S_0) \prod_{j=1}^k \pi_0(S_j | S_{j-1}) \prod_{j=1}^k p(y_j | S_j) \\ &\propto \exp\left\{-\frac{s_0^2}{2}\right\} \cdot \exp\left\{-\frac{\sum_{j=1}^k (s_j - s_{j-1})^2}{2}\right\} \\ &\quad \cdot \exp\left\{-\frac{\sum_{j=0}^k (y_j - s_j)^2}{2}\right\} \end{aligned}$$

- Again, the posterior is a complicated function of many parameters. Look at the terms in the posterior involving  $s_j$  only.

## A SPECIAL PROPERTY OF NORMAL DENSITIES

- Use the property of the normal distribution that

$$\begin{aligned} & \exp\left\{-\frac{1}{2}(y_j - s_j)^2\right\} \times \exp\left\{-\frac{1}{2}(s_j - s_{j-1})^2\right\} \\ & \times \exp\left\{-\frac{1}{2}(s_{j+1} - s_j)^2\right\} \\ \propto & \exp\left\{-\frac{3}{2}s_j^2 + 2\frac{1}{2}s_j(y_j + s_{j-1} + s_{j+1})\right\} \\ \propto & \exp\left\{-\frac{3}{2}\left(s_j - \frac{(y_j + s_{j-1} + s_{j+1})}{3}\right)^2\right\} \end{aligned}$$

- We get the following conditional densities

$$\begin{aligned} \pi(s_0 \mid s_1, y_0) & \sim N\left(\frac{y_0 + s_1}{2}, \frac{1}{2}\right) \\ \pi(s_j \mid s_{j-1}, s_{j+1}, y_j) & \sim N\left(\frac{s_{j-1} + s_{j+1} + y_j}{3}, \frac{1}{3}\right) \\ \pi(s_k \mid s_{k-1}, y_k) & \sim N\left(\frac{y_k + s_{k-1}}{2}, \frac{1}{2}\right) \end{aligned}$$

## THE INVERSE GAMMA DISTRIBUTION

- Let  $X \sim \text{Gamma}(a, b)$ : the pdf of  $X$  is

$$f(x) = \begin{cases} \frac{1}{b^a \Gamma(a)} x^{a-1} \exp(-x/b) & \text{if } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

- DEFINITION:  $Y = 1/X$  has a  $IG(a, b)$ : the pdf of  $Y$  is

$$f(y) = \begin{cases} \frac{1}{b^a \Gamma(a)} y^{-a-1} \exp(-1/by) & \text{if } y > 0, \\ 0 & \text{otherwise.} \end{cases}$$

- $E(Y) = 1/b(a - 1)$ ,  $V(Y) = 1/(b^2(a - 1)^2(a - 2))$
- The  $IG(a, b)$  prior on  $\sigma^2$  is *conjugate* to the normal likelihood  $N(\mu, \sigma^2)$  where the parameter of interest is  $\sigma^2$ .

## EXAMPLE 5: VARIANCE COMPONENT MODEL

Let

$$y_{ij} = \theta_i + \epsilon_{ij}, \quad i = 1, 2, \dots, K, \quad j = 1, 2, \dots, J$$

where

- $\epsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma_\epsilon^2)$ , and
- $\theta_i \stackrel{iid}{\sim} N(\mu, \sigma_\theta^2)$ .

- $\Theta = (\mu, \sigma_\epsilon^2, \sigma_\theta^2)$ . Prior on  $\Theta$ :

$$\begin{aligned}\pi_0(\sigma_\epsilon^2) &\sim IG(a_1, b_1); \\ \pi_0(\mu | \sigma_\theta^2) &\sim N(\mu_0, \sigma_\theta^2); \\ \pi_0(\sigma_\theta^2) &\sim IG(a_2, b_2).\end{aligned}$$

- Likelihood is

$$\begin{aligned}&L(\mathbf{y} | \Theta) \\ &= \int_{\theta_1, \theta_2, \dots, \theta_K} L(y_{11}, \dots, y_{KJ} | \theta_1, \dots, \theta_K, \sigma_\epsilon) \\ &\times \prod_i \pi_0(\theta_i | \mu, \sigma_\theta^2) d\theta_1 d\theta_2 \dots d\theta_K \\ &= \int_{\theta_1, \theta_2, \dots, \theta_K} \prod_{i,j} \frac{1}{\sqrt{2\pi\sigma_\epsilon}} \exp\{-(y_{ij} - \theta_i)^2 / 2\sigma_\epsilon^2\}\end{aligned}$$

$$\times \prod_i \frac{1}{\sqrt{2\pi}\sigma_\theta} \exp\{-(\theta_i - \mu)^2/2\sigma_\theta^2\} d\theta_1 d\theta_2 \dots d\theta_K$$

- In order to avoid integrating with respect to  $\theta_1, \theta_2, \dots, \theta_K$ , we can extend the parameter space to include them, namely,

$$\Theta = (\theta_1, \theta_2, \dots, \theta_K, \mu, \sigma_\epsilon^2, \sigma_\theta^2)$$

- The likelihood is  $L(\mathbf{y} \mid \Theta)$  without the integration on  $\theta_1, \theta_2, \dots, \theta_K$ .
- The posterior is

$$\begin{aligned} & \pi(\theta_1, \dots, \theta_K, \sigma_\epsilon^2, \mu, \sigma_\theta^2 \mid \mathbf{y}) \\ \propto & L(y_{11}, \dots, y_{KJ} \mid \theta_1, \dots, \theta_K, \sigma_\epsilon) \times \\ & \left\{ \prod_{i=1}^K \pi_0(\theta_i \mid \mu, \sigma_\theta) \right\} \pi_0(\sigma_\epsilon^2) \pi_0(\sigma_\theta^2) \pi_0(\mu \mid \sigma_\theta^2) \end{aligned}$$

**If we write down every term above, it will be quite long!**

By straight forward calculation, we get the following conditional densities

- $\pi(\theta_i | rest, \mathbf{y}) \sim N\left(\frac{J \bar{y}_{i.} \sigma_\theta^2 + \sigma_\epsilon^2 \mu}{J \sigma_\theta^2 + \sigma_\epsilon^2}, \frac{\sigma_\epsilon^2 \sigma_\theta^2}{J \sigma_\theta^2 + \sigma_\epsilon^2}\right),$
- $\pi(\mu | rest, \mathbf{y}) \sim N\left(\frac{\mu_0 + K \bar{\theta}}{K + 1}, \frac{\sigma_\theta^2}{K + 1}\right),$
- $\pi(\sigma_\theta^2 | rest, \mathbf{y}) \sim IG\left(\frac{2a_2 + K + 1}{2}, \left[\frac{b_2 [\sum_{i=1}^K (\theta_i - \mu)^2 + (\mu - \mu_0)^2] + 2}{2b_2}\right]^{-1}\right)$

and

- $\pi(\sigma_\theta^2 | rest, \mathbf{y}) \sim IG\left(a_1 + \frac{KJ}{2}, \left[\frac{b_1 \sum_{ij} (y_{ij} - \theta_i)^2 + 2}{2b_1}\right]^{-1}\right).$

where  $\bar{y}_{i.} = \frac{1}{J} \sum_{j=1}^J y_{ij}$  and  $\bar{\theta} = \frac{1}{K} \sum_{i=1}^K \theta_i$ .

## BAYESIAN COMPUTATIONS MONTE CARLO INTEGRATION

- Recall that the things of interest from the posterior distribution are  $E(\Theta | \mathbf{y})$ ,  $Var(\Theta | \mathbf{y})$  and HPD sets.
- In general, the problem is to find

$$E_{\pi}(h(X)) = \int_x h(x) \cdot \pi(x) dx$$

where evaluating the integral analytically can be very difficult.

- However, if we are able to draw samples from  $\pi$ , then we can approximate

$$E_{\pi}(h(X)) \approx \bar{h}_N \equiv \frac{1}{N} \sum_{j=1}^N h(X_j)$$

where  $X_1, X_2, \dots, X_N$  are  $N$  samples from  $\pi$ .

- This is called Monte Carlo integration.
- Note that for the examples, we would replace  $\pi(x)$  by  $\pi(\Theta | \mathbf{y})$ .

## JUSTIFICATION OF MONTE CARLO INTEGRATION

- For independent samples, by Law of Large Numbers,

$$\bar{h}_N \rightarrow E_{\pi}(h(X))$$

- Also,

$$\begin{aligned} \text{Var}(\bar{h}_N) &= \frac{\text{Var}_{\pi}(h(X))}{N} \\ &\doteq \frac{\sum_{j=1}^N (h(X_j) - \bar{h}_N)^2}{N^2} \\ &\rightarrow 0, \end{aligned}$$

as  $N$  becomes large.

- But direct independent sampling from  $\pi$  may be difficult.
- Resort to Markov Chain Monte Carlo (MCMC) methods.

## MARKOV CHAINS

A sequence of realizations from a Markov chain is generated by sampling

$$X^{(t)} \sim p(\cdot | x^{(t-1)}), \quad t = 1, 2, \dots$$

•  $p(x_1 | x_0)$  is called the transition kernel of the Markov chain:  $p(x_1 | x_0) = P\{X^{(t)} = x_1 | X^{(t-1)} = x_0\}$ .

•  $X^{(t)}$  depends only on  $X^{(t-1)}$ , and not on  $X_0, X_1, \dots, X^{(t-2)}$ , that is,

$$p(x^{(t)} | x_0, x_1, \dots, x^{(t-1)}) = p(x^{(t)} | x^{(t-1)}).$$

### EXAMPLE OF A SIMPLE MARKOV CHAIN

$$X^{(t)} | x^{(t-1)} = N(0.5x^{(t-1)}, 1)$$

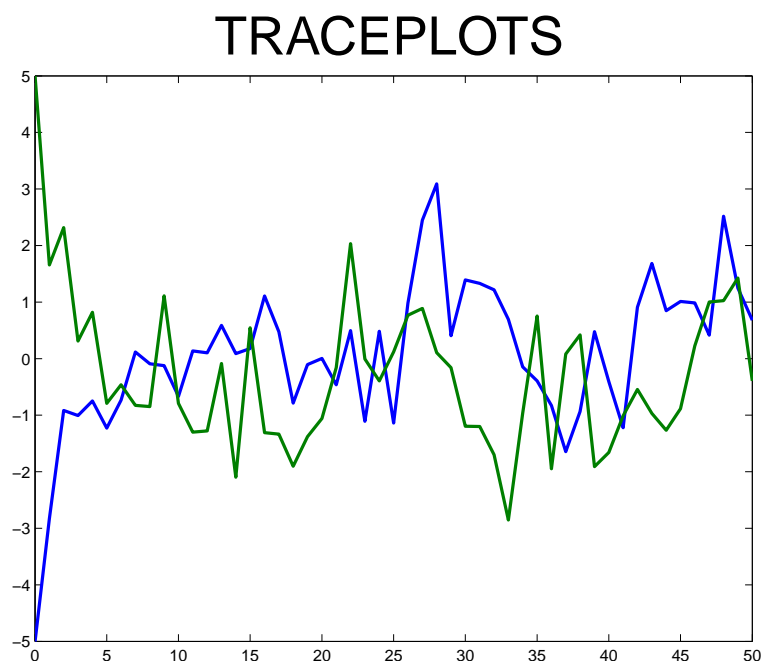
This is called the first order autoregressive process with lag 1 correlation of 0.5.

## MARKOV CHAINS (CONT.)

Simulate from the Markov chain

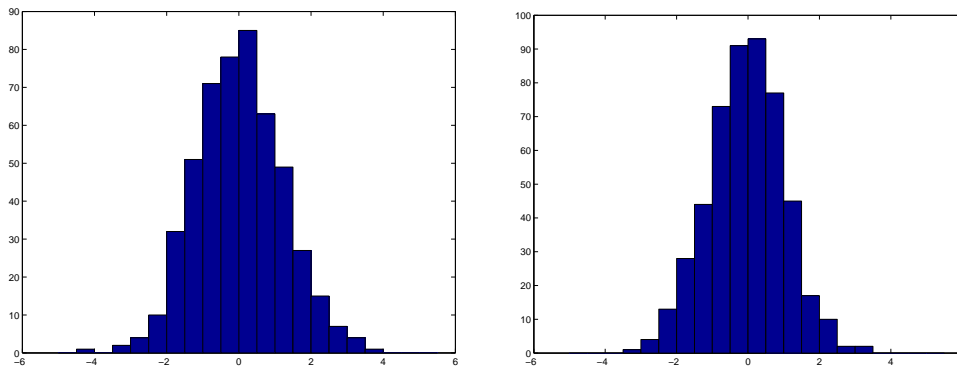
$$X^{(t)} \mid x^{(t-1)} = N(0.5x^{(t-1)}, 1)$$

with two different starting points:  $x_0 = -5$  and  $x_0 = +5$ .



It seems that after 10 iterations, the chains have forgotten their initial starting point  $x_0$ .

## MARGINAL PLOTS



The marginal plots appear normal, centered at 0.

- In fact, the above Markov chain converges to its *stationary distribution* as  $t \rightarrow \infty$ .

- In the above example, the stationary distribution is

$$X^{(\infty)} | x^{(0)} = N(0, 1.333)$$

which does not depend on  $x^{(0)}$ .

- Does this happen for all Markov chains?

# CONDITIONS THAT GUARANTEE CONVERGENCE (1 of 2)

## 1. IRREDUCIBILITY

- The irreducibility condition guarantees that if a stationary distribution exists, it is unique.
- Irreducibility means that each state in a Markov chain can be reached from any other state in a finite number of steps.
- An example of a *reducible* Markov chain: Suppose there are sets  $A$  and  $B$  such that  $p(A | x) = 0$  for every  $x \in B$  and vice versa.

# CONDITIONS THAT GUARANTEE CONVERGENCE (2 of 2)

## 2. APERIODICITY

- A Markov chain with finite number of states is said to be *periodic* with period  $d$  if the return times to a state  $x$  happens in steps of  $kd$ ,  $k = 1, 2, \dots$
- A Markov chain is said to be *aperiodic* if it is not periodic.
- In other words, look at all return times to state  $x$  and consider the greatest common divisor (gcd) of these return times. The gcd of the return times should be 1 for an aperiodic chain (greater than 1 for a periodic chain)
- This can be generalized to general state spaces for Markov chains.

## ERGODICITY

- Assume a Markov chain:
  - (a) has a stationary distribution  $\pi(x)$ , and
  - (b) is aperiodic and irreducible.
- Then, we have an *ergodic theorem*:

$$\bar{h}_N \equiv \frac{1}{N} \sum_{t=1}^N h(X^{(t)}) \rightarrow E_{\pi}(h(X))$$

as  $T \rightarrow \infty$ .  $\bar{h}_N$  is called the ergodic average. Also, for such chains with

$$\sigma_h^2 = \text{Var}_{\pi}(h(X)) < \infty$$

- the central limit theorem holds, and
- the convergence has a geometric rate.

## MARKOV CHAIN MONTE CARLO

- Recall that our goal is to sample from the target  $\pi(x)$ .
- Question: How do we construct a Markov chain (aperiodic and irreducible) so that the stationary distribution will be  $\pi(x)$ ?
- Metropolis (1953) showed how. This was generalized by Hastings (1970).
- Henceforth, they are called Markov chain Monte Carlo (MCMC) methods.

## THE METROPOLIS-HASTINGS ALGORITHM

**STEP 1:** For  $t = 1, 2, \dots$ , generate

$$Y | x^{(t-1)} = p(y | x^{(t-1)})$$

(a)  $Y$  is called a candidate point, and

(b)  $p(y | x^{(t-1)})$  is called the proposal distribution.

**STEP 2:** Consider the acceptance probability

$$\alpha(x^{(t)}, y) = \min \left\{ \frac{\pi(y) p(x^{(t)} | y)}{\pi(x^{(t)}) p(y | x^{(t)})}, 1 \right\}$$

**STEP 3:** With probability  $\alpha(x^{(t)}, y)$ ,

set

$$X^{(t+1)} = y \quad (\text{acceptance})$$

else set

$$X^{(t+1)} = x^{(t)} \quad (\text{rejection})$$

## THE METROPOLIS-HASTINGS ALGORITHM (CONT.)

### NOTES:

- The normalization constants in  $\pi(x)$  is not required to run the algorithm. They cancel out from the numerator and denominator.
- The proposal distribution  $p$  is chosen so that it is easy to sample from.
- Theoretically, any  $p$  having the same support as  $\pi$  will suffice but it turns out that *some choices are better than others in practice (implementation issues, see later for more details)*.
- The resulting Markov chain have the desirable properties (irreducibility and aperiodicity) under mild conditions on  $\pi(x)$ .

## BURN-IN PERIOD, $B$

- The early iterations  $x^{(1)}, x^{(2)}, \dots, x^{(B)}$  reflect the initial starting value  $x_0$ .
- These iterations are called burn-in.
- After burn-in, we say that the chain has “converged”.
- Omit the burn-in samples from the ergodic average:

$$\bar{h}_{BN} = \frac{1}{N - B} \sum_{t=B+1}^N h(X^{(t)})$$

and

$$\widehat{Var}(\bar{h}_{BN}) = \frac{1}{(N - B)^2} \sum_{t=B+1}^N (h(X^{(t)}) - \bar{h}_{BN})^2.$$

- Methods for determining  $B$  are called convergence diagnostics, and will be discussed later.

## IMPORTANT SPECIAL CASES: THE INDEPENDENCE SAMPLER

- The independence sampler is based on the choice that

$$p(y | x) = p(y)$$

independent of  $x$ .

- Hence, the acceptance probability has the form

$$\alpha(x, y) = \min \left\{ \frac{\pi(y) p(x)}{\pi(x) p(y)}, 1 \right\}$$

- Choice of  $p$ : For geometric convergence of the algorithm, we must have:

- (a) the support of  $p$  includes the support of  $\pi(x)$ , and
- (b)  $p$  must have heavier tails compared to  $\pi(x)$ .

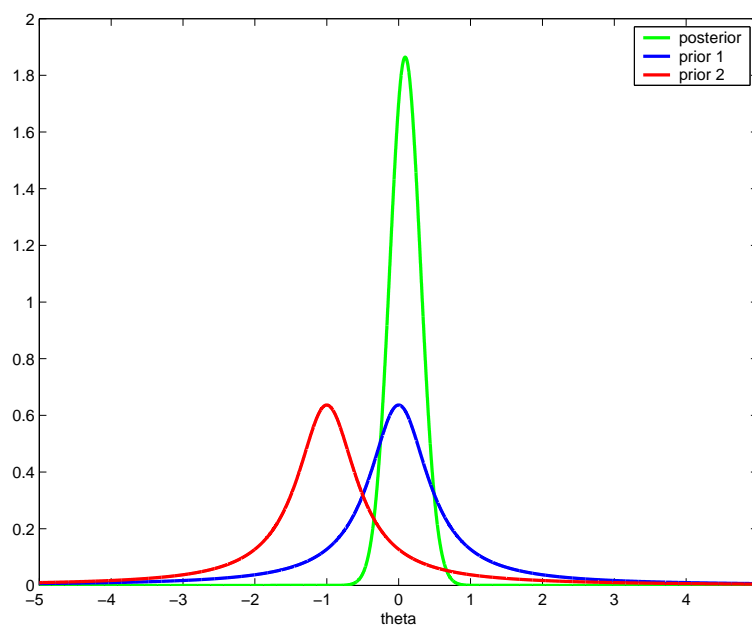
## EXAMPLE 3: NORMAL LIKELIHOOD, CAUCHY PRIOR

- We have  $\bar{y} = 0.1$  and  $n = 20$ . This gives  $C = 1.8813$ .

- Two different candidate densities:

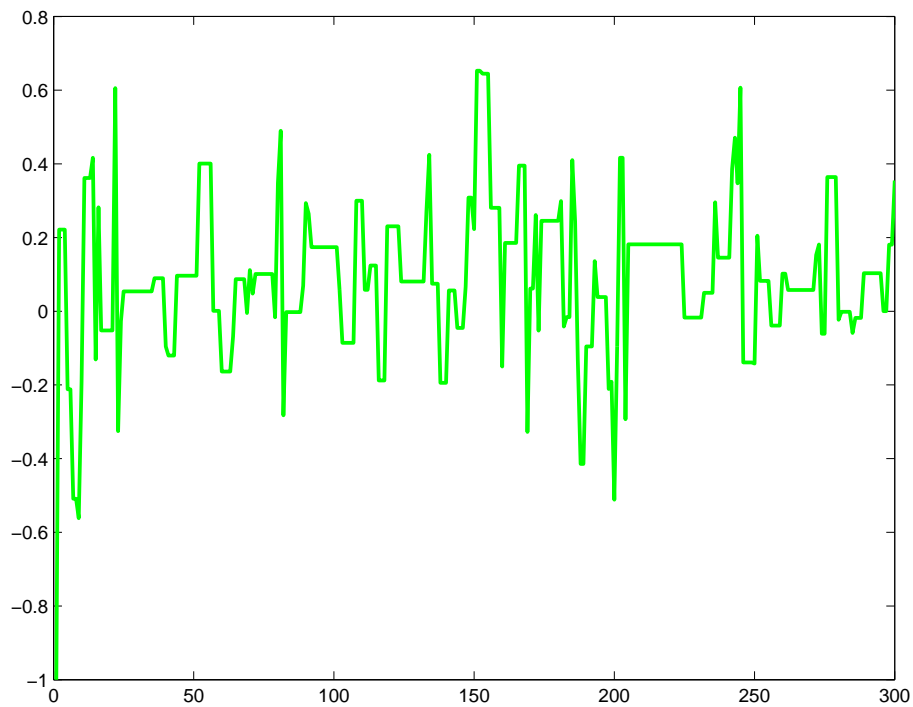
- (a) Cauchy(0,0.5) (blue)

- (b) Cauchy(-1,0.5) (red)



- The posterior mean is 0.0919.
- The posterior variance is 0.0460.

EXAMPLE 3 (CONT.)  
TRACEPLOT USING CAUCHY(0,0.5)

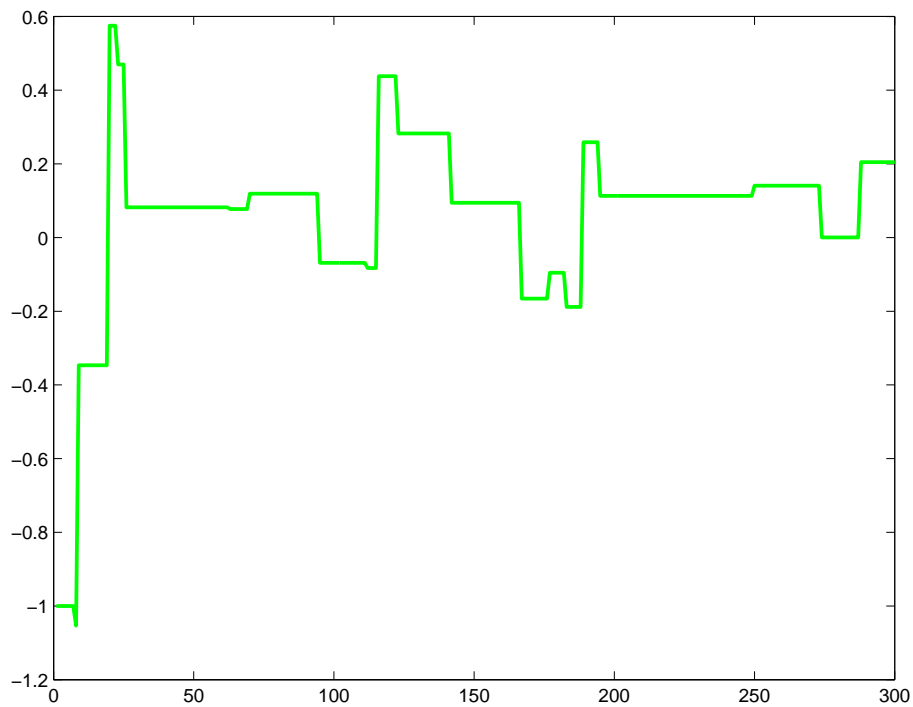


	True	Simulation
Mean	0.0919	0.1006
Variance	0.0460	0.0452

- Starting value is  $-1$ .
- Table entries are based on  $N = 4,000$  iterations of the MC with a burn-in period of  $B = 1,000$ .

### EXAMPLE 3 (CONT.)

#### TRACEPLOT USING CAUCHY(-1,0.5)



	True	Simulation
Mean	0.0919	0.1070
Variance	0.0460	0.0457

- Starting value is  $-1$ .
- Table entries are based on  $N = 4,000$  iterations of the MC with a burn-in period of  $B = 1,000$ .

## THE RANDOM WALK SAMPLER

- THE METROPOLIS ALGORITHM

Proposal is symmetric:  $p(x | y) = p(y | x)$

- RANDOM WALK METROPOLIS

$$p(x | y) = p(|x - y|)$$

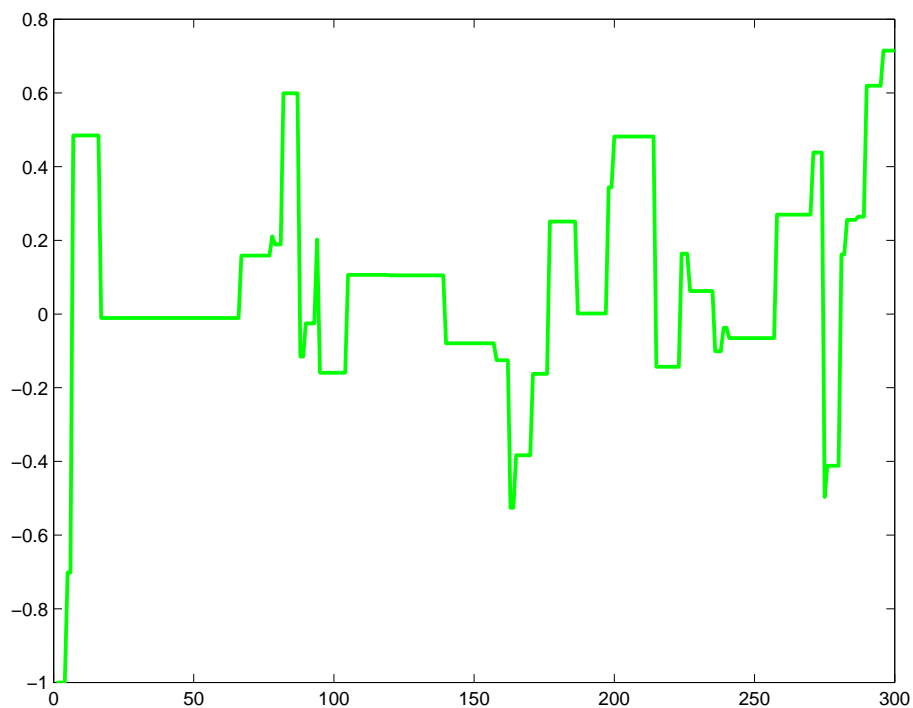
In this case,

$$\alpha(x, y) = \min \left\{ 1, \frac{\pi(y)}{\pi(x)} \right\}$$

BACK TO EXAMPLE 3.

- Three choices of  $p$  were considered:
  - (a) Cauchy(0,2) (large scale)
  - (b) Cauchy(0,0.2) (moderate scale)
  - (c) Cauchy(0,0.02) (small scale)

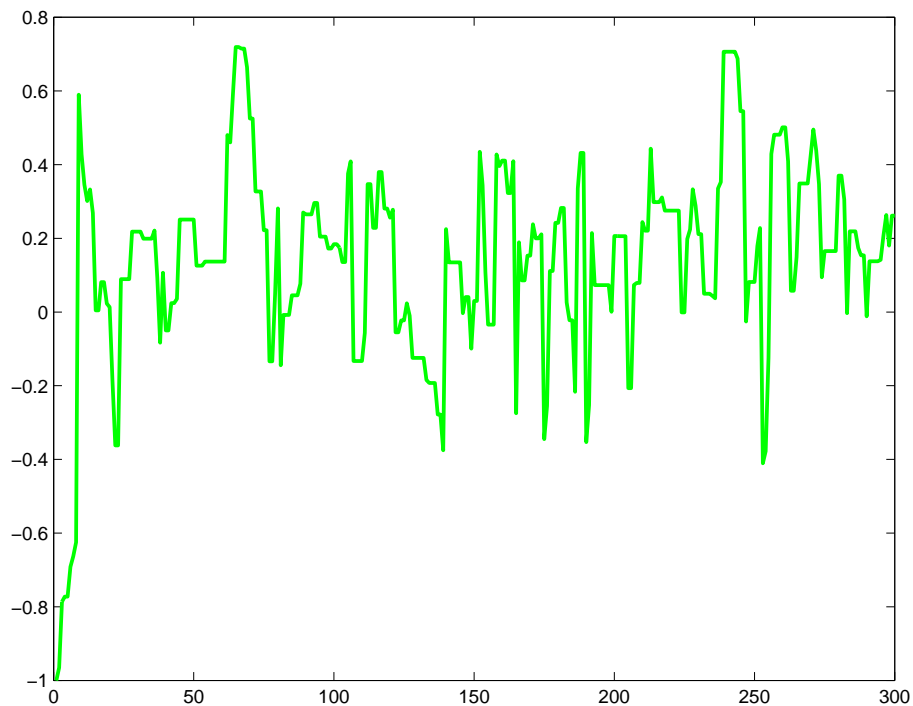
## TRACEPLOT FOR CAUCHY(0,2)



	True	Simulation
Mean	0.0919	0.0773
Variance	0.0460	0.0393

- Starting value is  $-1$ .
- Table entries are based on  $N = 4,000$  iterations of the MC with a burn-in period of  $B = 1,000$ .

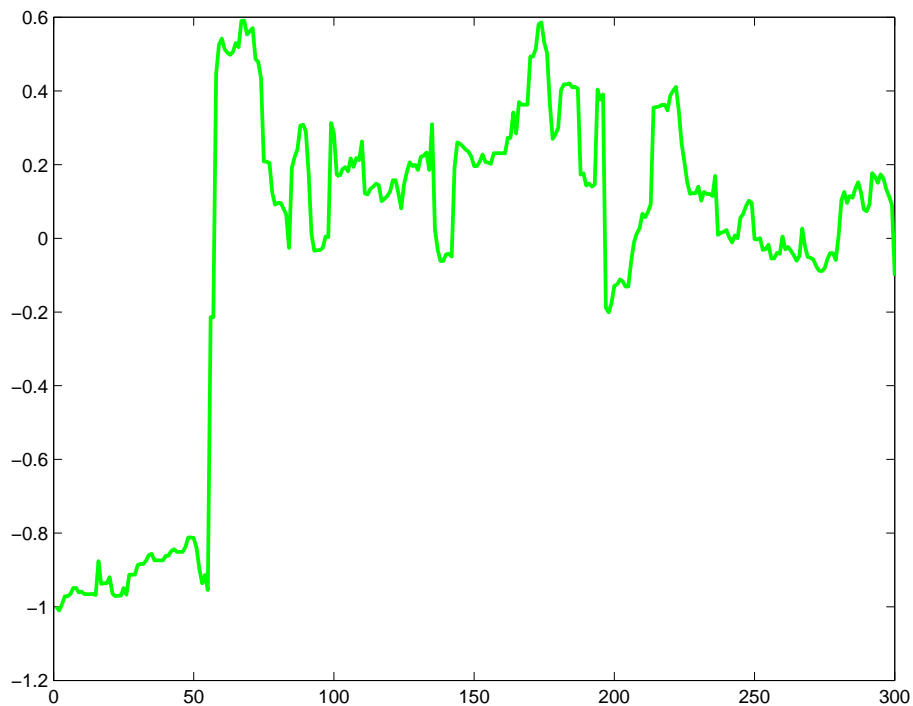
## TRACEPLOT FOR CAUCHY(0,0.2)



	True	Simulation
Mean	0.0919	0.1026
Variance	0.0460	0.0476

- Starting value is  $-1$ .
- Table entries are based on  $N = 4,000$  iterations of the MC with a burn-in period of  $B = 1,000$ .

## TRACEPLOT FOR CAUCHY(0,0.02)



	True	Simulation
Mean	0.0919	0.0446
Variance	0.0460	0.0389

- Starting value is  $-1$ .
- Table entries are based on  $N = 4,000$  iterations of the MC with a burn-in period of  $B = 1,000$ .

## THE GIBBS SAMPLER

- Suppose that  $x = (x_1, x_2, \dots, x_D)$  is of dimension  $D$ .
- The Gibbs sampler samples from the conditional distributions:

$$\begin{aligned} & \pi(x_u \mid x_1, x_2, \dots, x_{u-1}, x_{u+1}, \dots, x_D) \\ = & \frac{\pi(x_1, x_2, \dots, x_{u-1}, x_u, x_{u+1}, \dots, x_D)}{\int_{x_u} \pi(x_1, x_2, \dots, x_{u-1}, x_u, x_{u+1}, \dots, x_D) dx_u} \end{aligned}$$

- Note that the conditional is proportional to the joint distribution, so collecting  $x_u$  terms in the joint distribution often helps in finding it.

## THE GIBBS SAMPLER

Update componentwise to go from  $t$  to  $t + 1$ :

$$X_1^{(t+1)} \sim \pi(x_1 | x_2^{(t)}, x_3^{(t)} \dots, x_D^{(t)})$$

$$X_2^{(t+1)} \sim \pi(x_2 | x_1^{(t+1)}, x_3^{(t)} \dots, x_D^{(t)})$$

...

$$X_d^{(t+1)} \sim \pi(x_d | x_1^{(t+1)}, x_2^{(t+1)} \dots, \\ x_{d-1}^{(t+1)}, x_{d+1}^{(t)}, \dots, x_D^{(t)})$$

...

$$X_D^{(t+1)} \sim \pi(x_D | x_1^{(t+1)}, x_2^{(t+1)} \dots, x_{D-1}^{(t+1)})$$

- Note how the most recent values are used in the subsequent conditional distributions.

## EXAMPLE 4: HMM

- Recall the conditional densities

$$\begin{aligned}\pi(s_0 \mid s_1, y_0) &\sim N\left(\frac{y_0 + s_1}{2}, \frac{1}{2}\right) \\ \pi(s_j \mid s_{j-1}, s_{j+1}, y_j) &\sim N\left(\frac{s_{j-1} + s_{j+1} + y_j}{3}, \frac{1}{3}\right) \\ \pi(s_K \mid s_{K-1}, y_K) &\sim N\left(\frac{y_K + s_{K-1}}{2}, \frac{1}{2}\right)\end{aligned}$$

- To implement the Gibbs sampler, we took  $K = 4$

## IMPLEMENTATION

- The observations are

$$Y = (0.21, 2.01, -0.36, -2.46, -2.61)$$

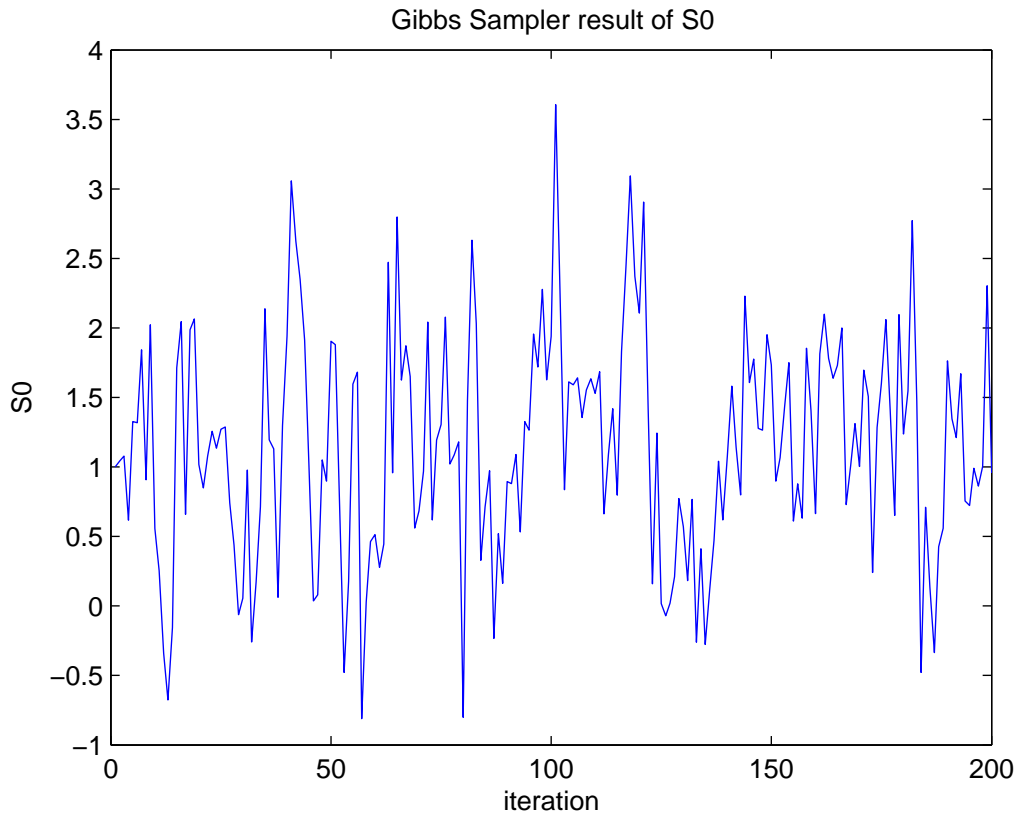
- Next, we chose the starting value

$$S_0 = (0.21, 2.01, -0.36, -2.46, -2.61)$$

and ran the Gibbs sampler.

- We ran  $N = 4,000$  iterations of the MC with a burn-in period of  $B = 1,000$ .

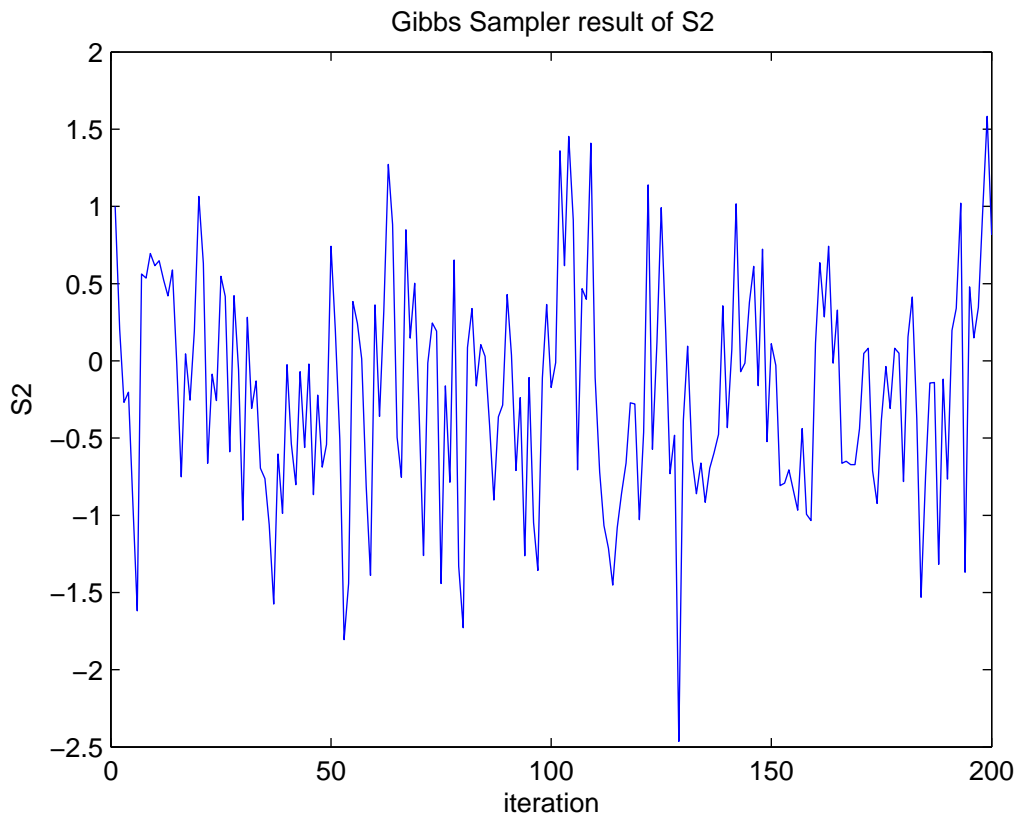
## EXAMPLE 4: TRACEPLOT FOR $S_0$



	Value	Standard error
Posterior Mean	-0.23	0.0475

Table entries are based on  $N = 4,000$  iterations of the MC with a burn-in period of  $B = 1,000$ .

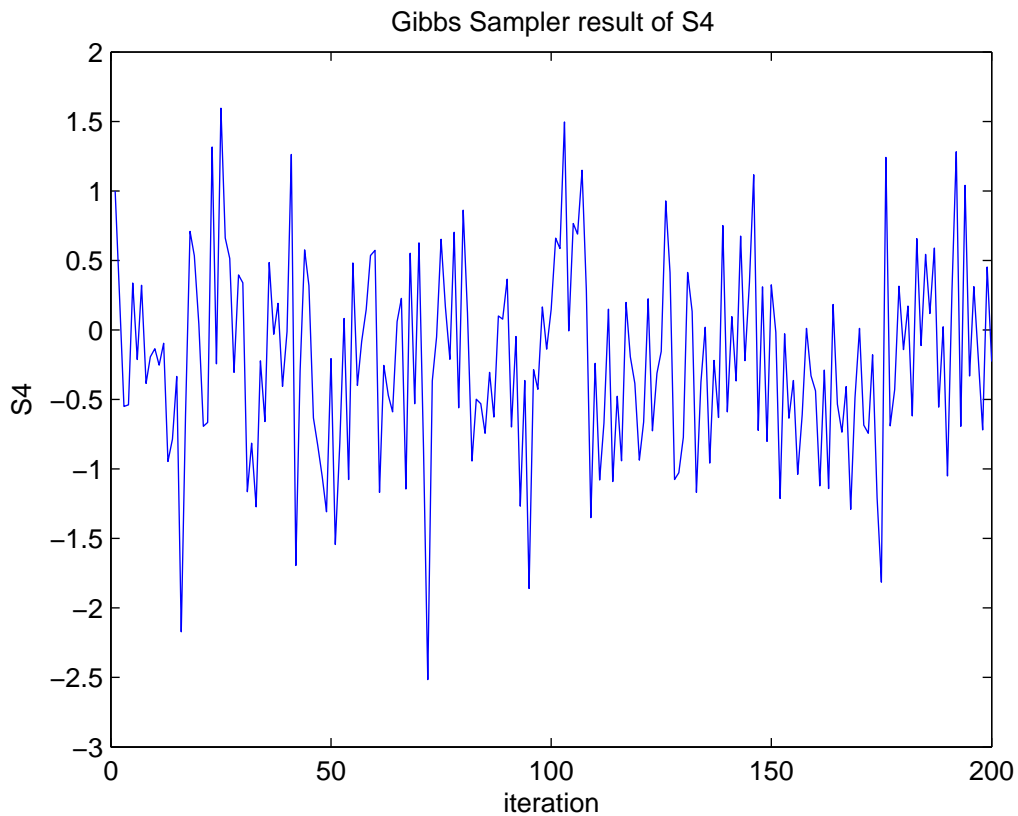
## EXAMPLE 4 TRACEPLOT FOR $S_2$



	Value	Standard error
Posterior Mean	-0.1300	0.0388

Table entries are based on  $N = 4,000$  iterations of the MC with a burn-in period of  $B = 1,000$ .

## EXAMPLE 4 TRACEPLOT FOR $S_4$



	Value	Standard error
Posterior Mean	0.0339	0.0413

Table entries are based on  $N = 4,000$  iterations of the MC with a burn-in period of  $B = 1,000$ .

## EXAMPLE 5: VARIANCE COMPONENT MODEL

RECALL:

$$y_{ij} = \theta_i + \epsilon_{ij}, \quad i = 1, 2, \dots, K, j = 1, 2, \dots, J$$

RECALL THE CONDITIONAL DISTRIBUTIONS:

- $\pi(\theta_i | rest, \mathbf{y}) \sim N\left(\frac{J\bar{y}_{i\cdot}\sigma_\theta^2 + \sigma_\epsilon^2\mu}{J\sigma_\theta^2 + \sigma_\epsilon^2}, \frac{\sigma_\epsilon^2\sigma_\theta^2}{J\sigma_\theta^2 + \sigma_\epsilon^2}\right),$
  - $\pi(\mu | rest, \mathbf{y}) \sim N\left(\frac{\mu_0 + K\bar{\theta}}{K + 1}, \frac{\sigma_\theta^2}{K + 1}\right),$
  - $\pi(\sigma_\theta^2 | rest, \mathbf{y}) \sim IG\left(\frac{2a_2 + K + 1}{2}, \left[\frac{b_2[\sum_{i=1}^K (\theta_i - \mu)^2 + (\mu - \mu_0)^2] + 2}{2b_2}\right]^{-1}\right)$
- and
- $\pi(\sigma_\epsilon^2 | rest, \mathbf{y}) \sim IG\left(a_1 + \frac{KJ}{2}, \left[\frac{b_1 \sum_{ij} (y_{ij} - \theta_i)^2 + 2}{2b_1}\right]^{-1}\right).$

## HOW DO YOU CHOOSE STARTING VALUES IN GENERAL

- In practice, the values of the true parameters will be unknown.
- How do you select good starting values in such cases?
- I usually use an ad-hoc estimate of the parameters.
- Illustrate!

## IMPLEMENTATION

- We took  $K = 2, J = 5$

- The data is  $Y =$

$$(1.70, 1.30, 3.53, 1.14, 3.15, \\ -2.25, 2.29, 1.77, -3.80, 3.36)$$

- To implement the Gibbs sampler, we took:

(a)  $a_1 = 2; b_1 = .02;$

(b)  $a_2 = 3; b_2 = 0.03;$

- STARTING VALUES:

- (a) Note that

$$E\left(\sum_{i=1}^K \sum_{j=1}^J (y_{ij} - \bar{y}_{i\cdot})^2\right) = K(J - 1)\sigma_\epsilon^2$$

- So, we set

$$\sigma_{\epsilon,0}^2 = \frac{1}{K(J - 1)} \sum_{i=1}^K \sum_{j=1}^J (y_{ij} - \bar{y}_{i\cdot})^2.$$

(b) Note that

$$E(\bar{y}_{..}) = \mu,$$

• We set

$$\mu_0 = \bar{y}_{..}$$

(c) Note that

$$E\left(J \sum_{i=1}^K (y_{i.} - \bar{y}_{..})^2\right) = J(K-1)\sigma_{\theta}^2 + (K-1)\sigma_{\epsilon}^2,$$

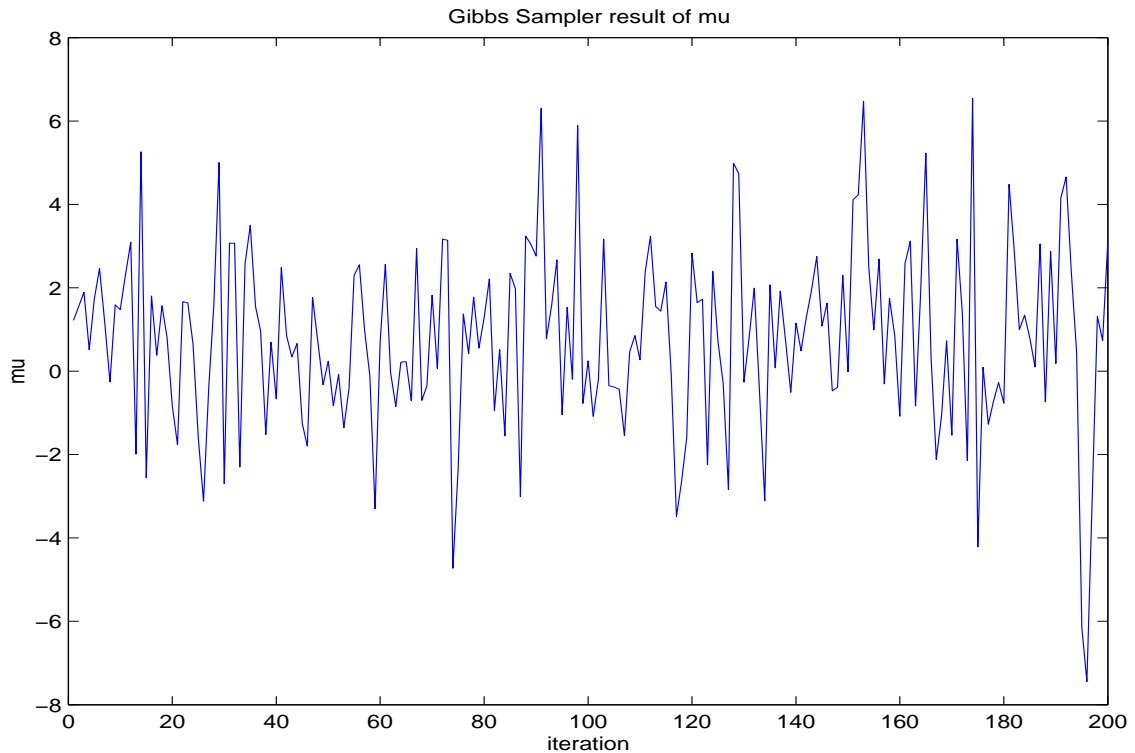
• We set

$$\sigma_{\theta,0}^2 = \frac{J \sum_{i=1}^K (y_{i.} - \bar{y}_{..})^2 - (K-1)\sigma_{\epsilon,0}^2}{J(K-1)}$$

(d)  $\theta_1, \theta_2$  are iid  $N(\mu_0, \sigma_{\theta,0}^2)$ :

• So, we generate  $\theta_{1,0}$  and  $\theta_{2,0}$ ; or, you can set  $\theta_{1,0} = \bar{y}_1$ . and  $\theta_{2,0} = \bar{y}_2$ .

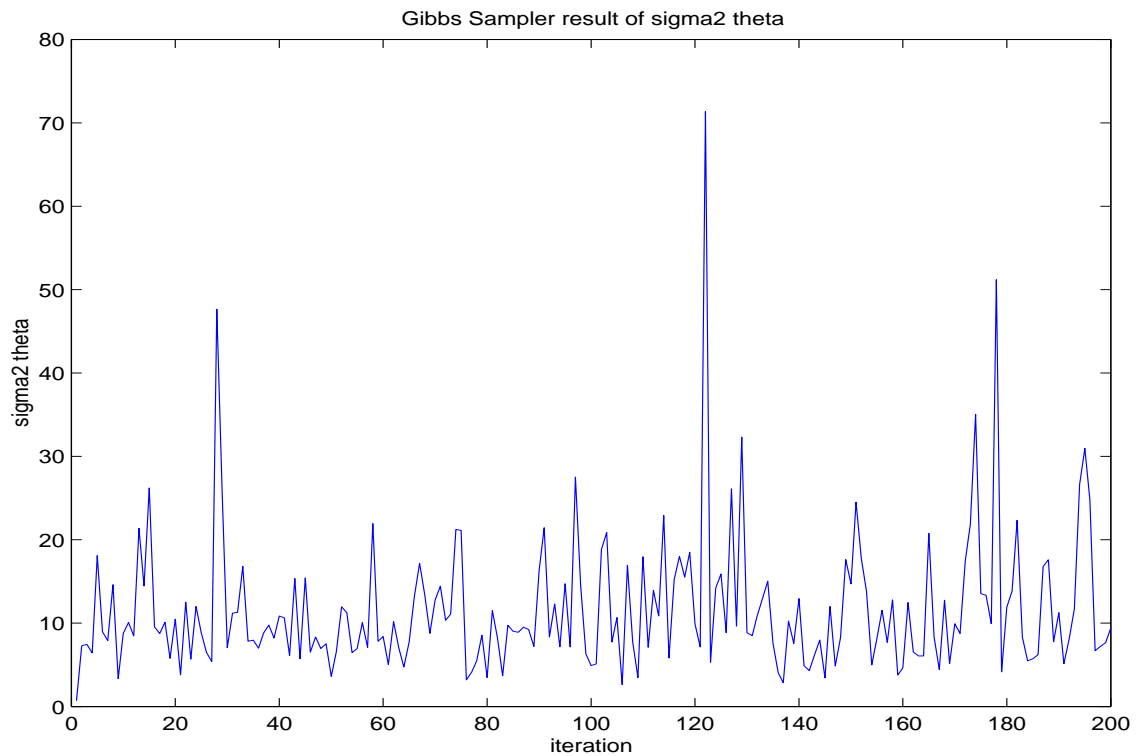
## EXAMPLE 5: TRACEPLOT FOR $\mu$



	True	Simulation
Value	-0.0380	0.7658 (standard error = 0.2613)

- Starting values are (0.86, -0.17, 1.22, 0.69, 5.45).
- Table entries are based on  $N = 4,000$  iterations of the MC with a burn-in period of  $B = 1,000$ .

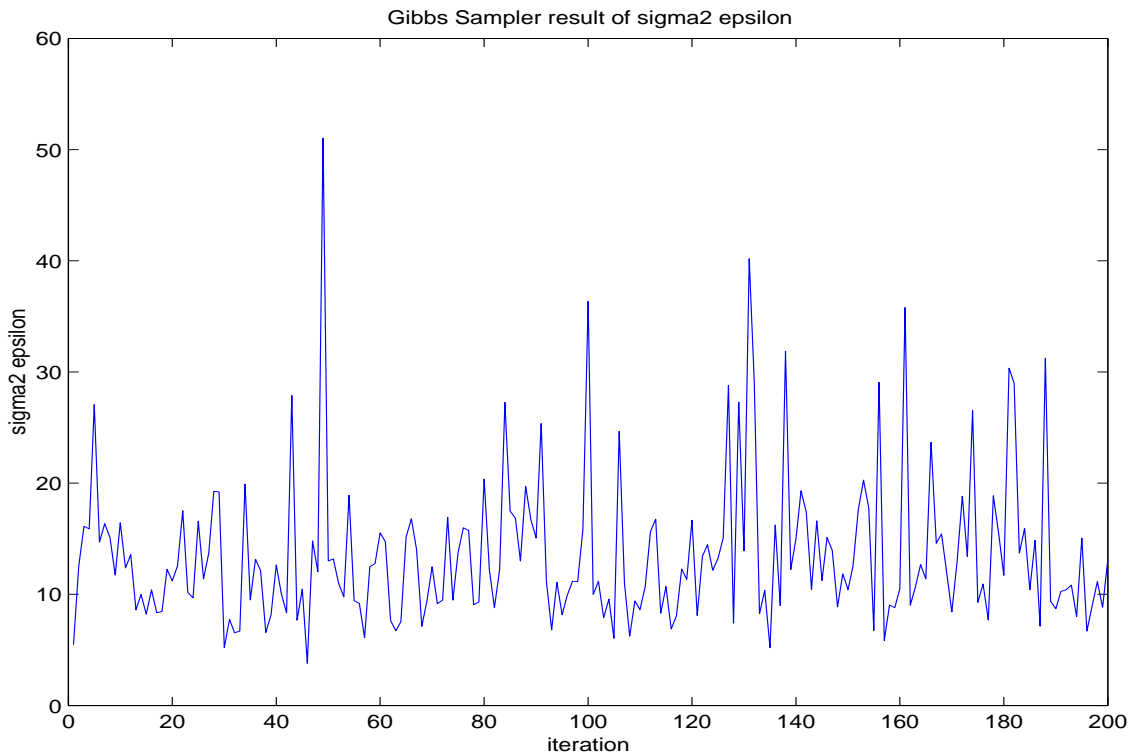
## EXAMPLE 5: TRACEPLOT FOR $\sigma_\theta^2$



	True	Simulation
Value	14.7876	14.0902 (standard error = 2.5243)

- Starting values are (0.86, -0.17, 1.22, 0.69, 5.45).
- Table entries are based on  $N = 4,000$  iterations of the MC with a burn-in period of  $B = 1,000$ .

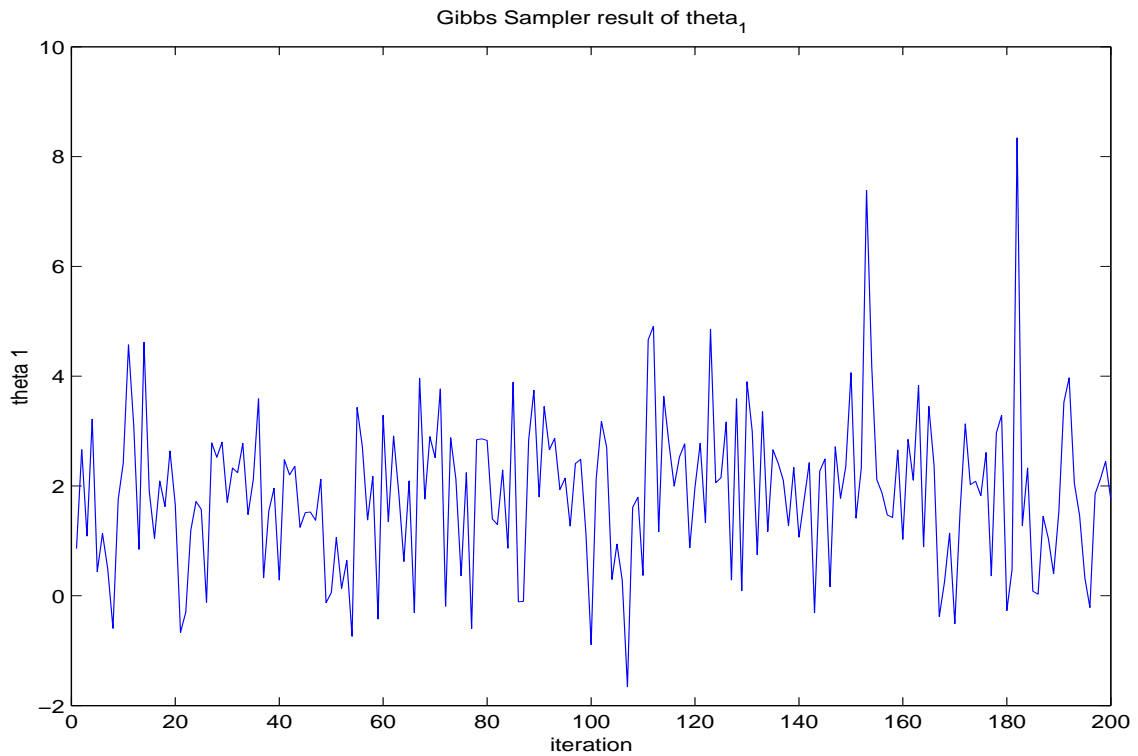
## EXAMPLE 5: TRACEPLOT FOR $\sigma_{\epsilon}^2$



	True	Simulation
Value	14.7876	14.0902 (standard error = 2.5243)

- Starting values are (0.86, -0.17, 1.22, 0.69, 5.45).
- Table entries are based on  $N = 4,000$  iterations of the MC with a burn-in period of  $B = 1,000$ .

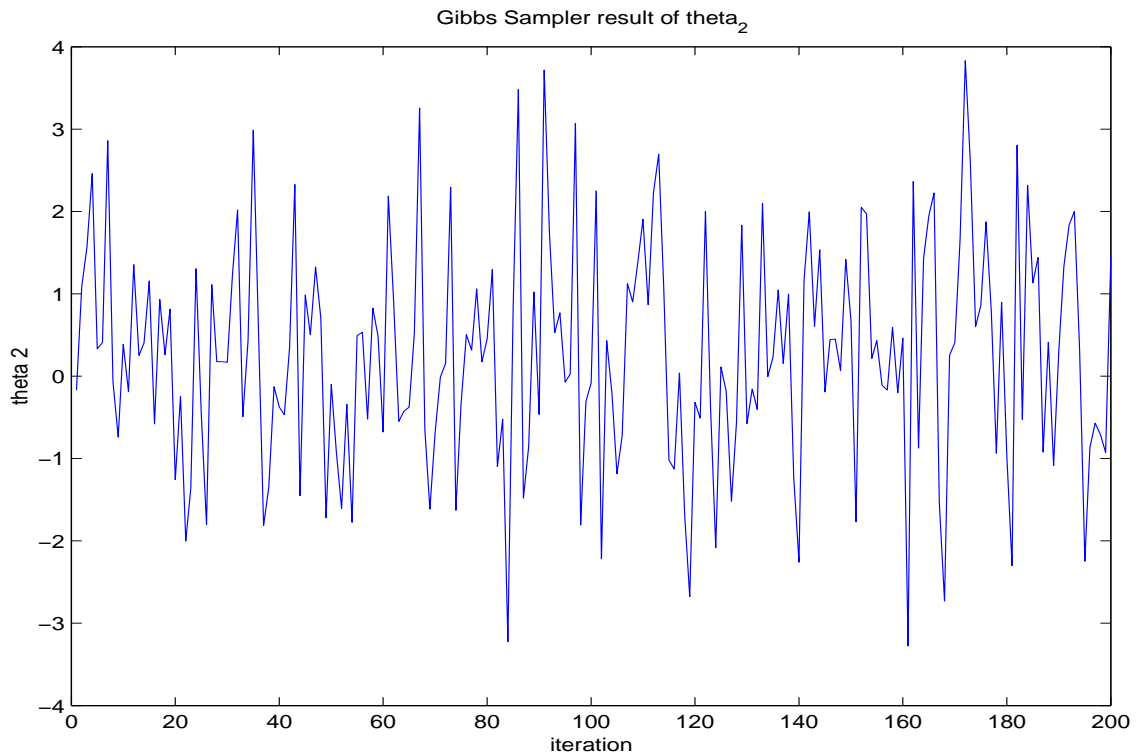
## EXAMPLE 5: TRACEPLOT FOR $\theta_1$



	Value	Standard error
Posterior Mean	1.8162	0.1518

- Starting values are (0.86, -0.17, 1.22, 0.69, 5.45).
- Table entries are based on  $N = 4,000$  iterations of the MC with a burn-in period of  $B = 1,000$ .

## EXAMPLE 5: TRACEPLOT FOR $\theta_2$



	Value	Standard error
Posterior Mean	0.4072	0.1489

- Starting values are (0.86, -0.17, 1.22, 0.69, 5.45).
- Table entries are based on  $N = 4,000$  iterations of the MC with a burn-in period of  $B = 1,000$ .

## IMPLEMENTATION ISSUES

How many parallel Markov chains should be run ?

- Several different (short) runs with different initial values (Gelman and Rubin, 1992)

- (a) Gives indication of convergence

- (b) Gives a sense of statistical security

- Run one very long chain (Geyer, 1992)

- (a) Reaches parts of the posterior distribution that other schemes do not

- Experiment yourself, try one or the other, or both.

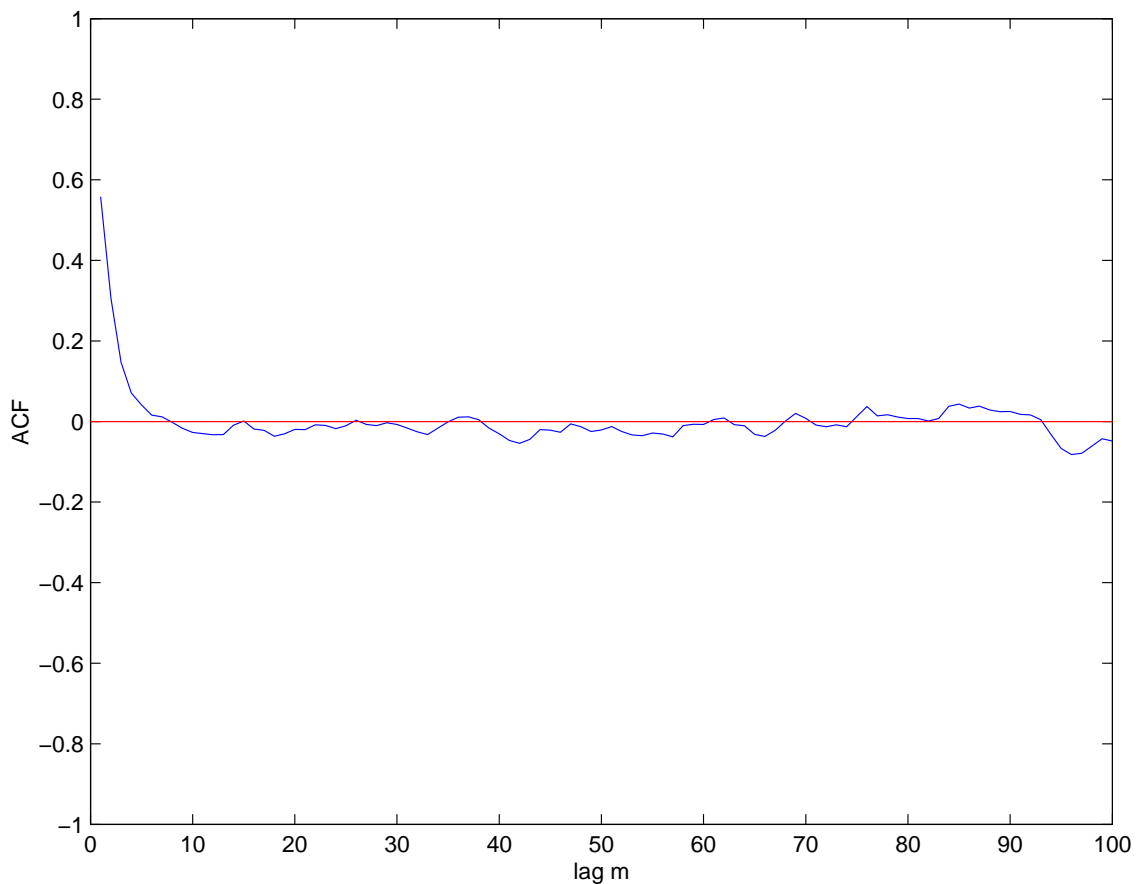
## CONVERGENCE DIAGNOSTICS

You must do:

- Traceplots of each component of the parameter  $\Theta$ .
- Plot of the autocorrelation functions. If correlations do not die down to zero, check your codes, debug!

# CONVERGENCE DIAGNOSTICS

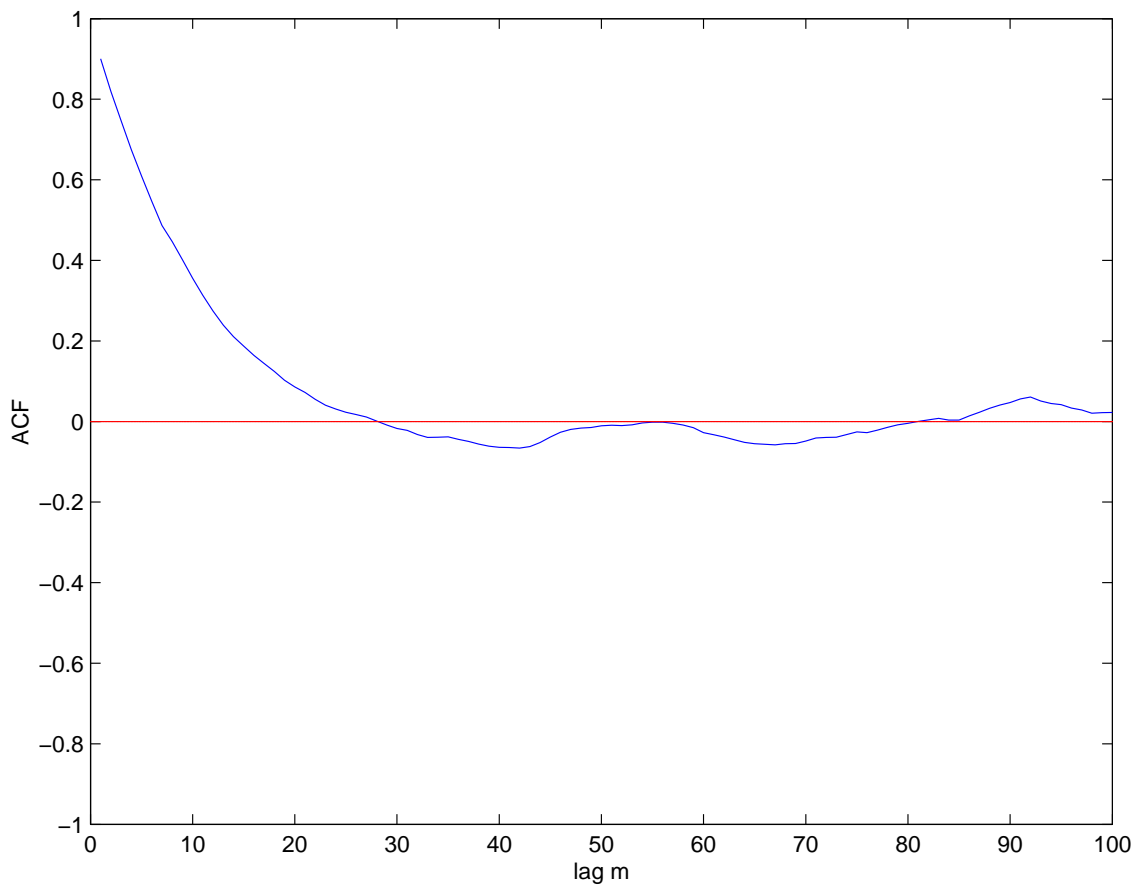
## EXAMPLE 3: INDEPENDENCE SAMPLER WITH CAUCHY(0,0.5)



Autocorrelation functions were calculated based on  $N = 4,000$  with a burn-in period of  $B = 1,000$ .

# CONVERGENCE DIAGNOSTICS

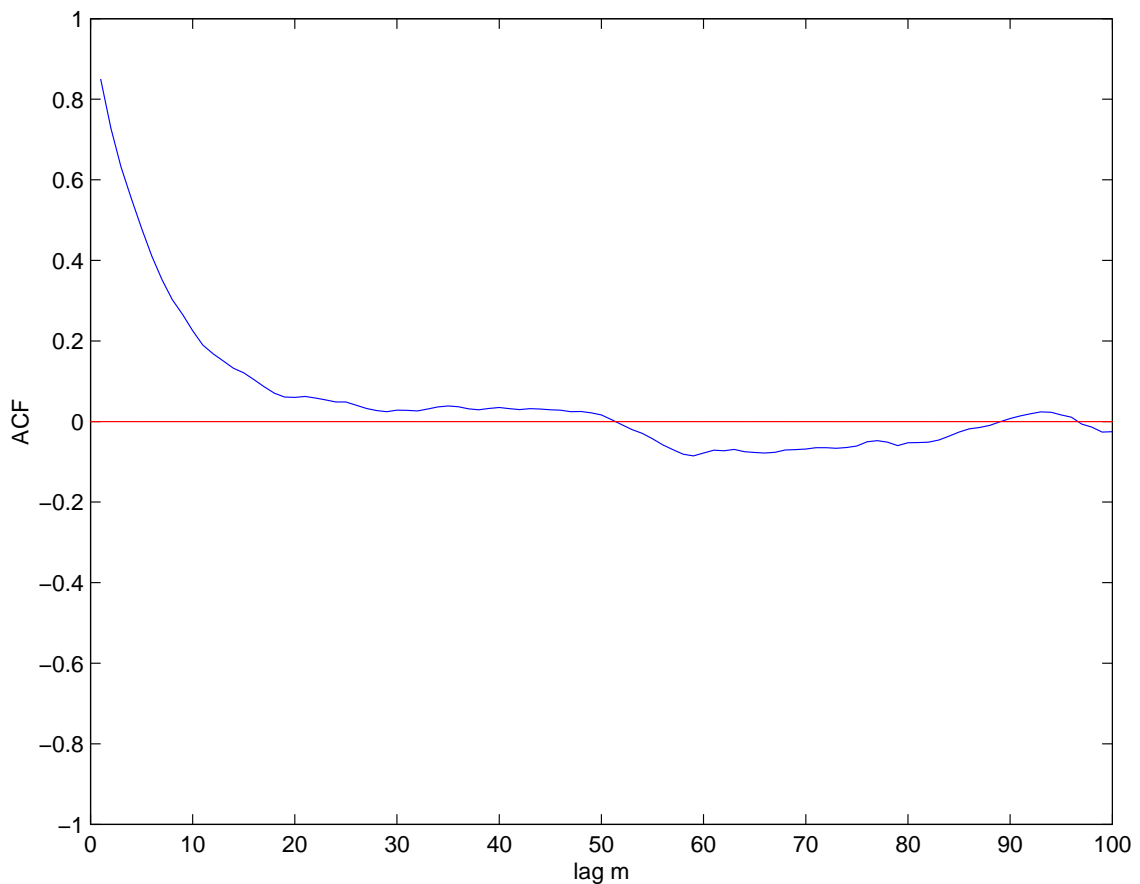
## EXAMPLE 3: INDEPENDENCE SAMPLER WITH CAUCHY(-0.1,0.5)



Autocorrelation functions were calculated based on  $N = 4,000$  with a burn-in period of  $B = 1,000$ .

# CONVERGENCE DIAGNOSTICS

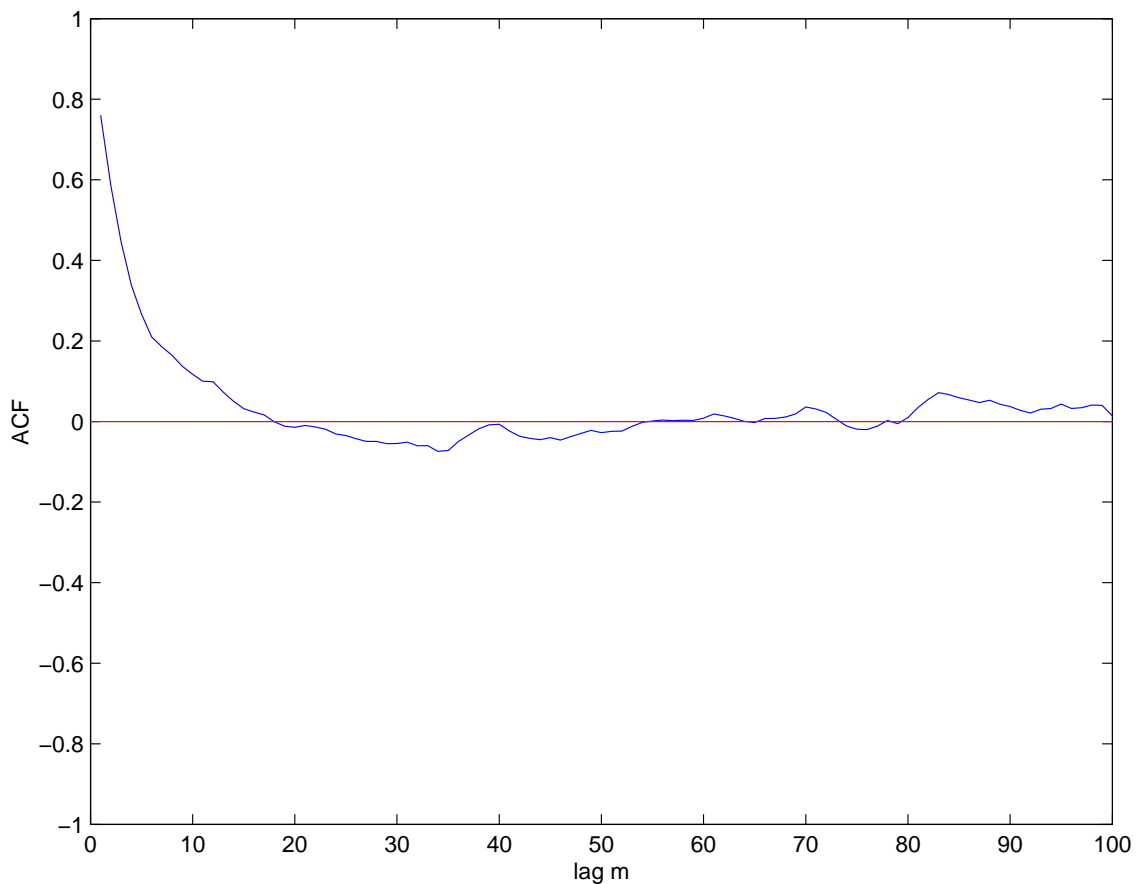
## EXAMPLE 3: RANDOM WALK SAMPLER WITH CAUCHY(0,2)



Autocorrelation functions were calculated based on  $N = 4,000$  with a burn-in period of  $B = 1,000$ .

# CONVERGENCE DIAGNOSTICS

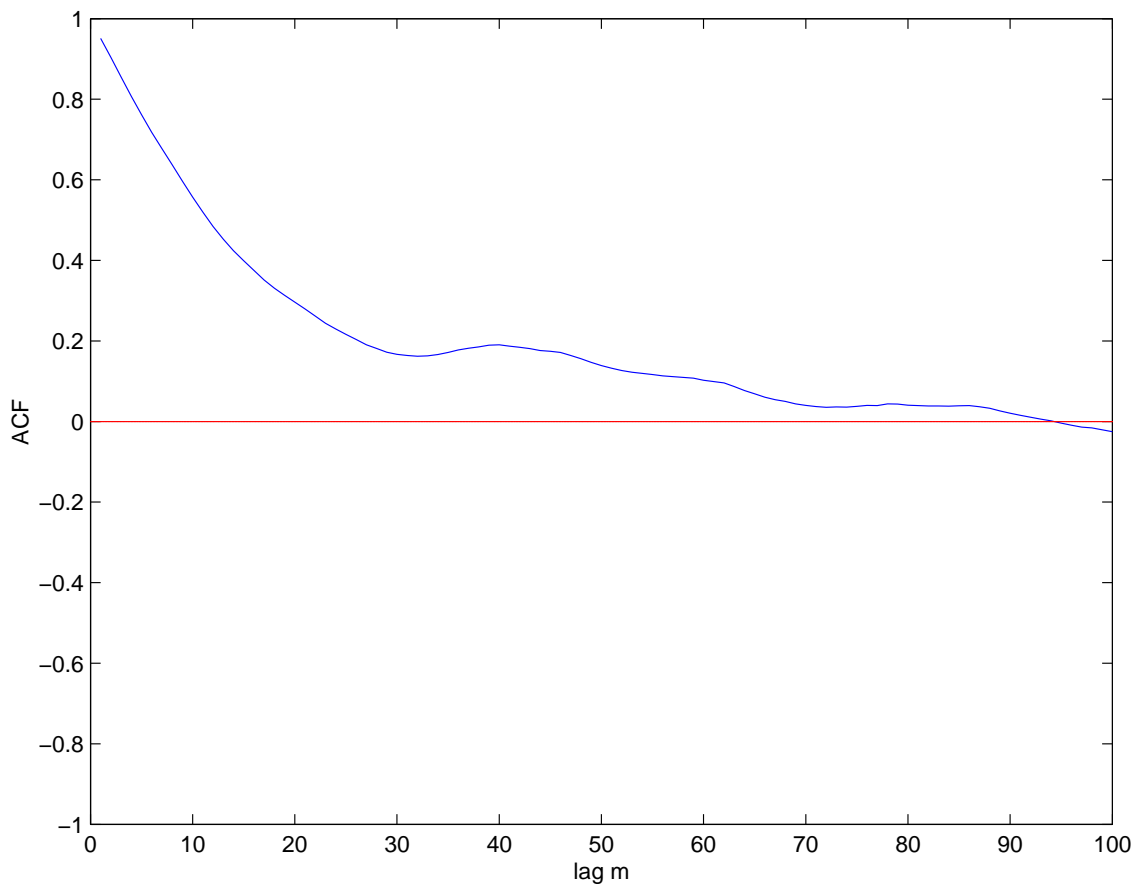
## EXAMPLE 3: RANDOM WALK SAMPLER WITH CAUCHY(0,0.2)



Autocorrelation functions were calculated based on  $N = 4,000$  with a burn-in period of  $B = 1,000$ .

# CONVERGENCE DIAGNOSTICS

## EXAMPLE 3: RANDOM WALK SAMPLER WITH CAUCHY(0,0.02)



Autocorrelation functions were calculated based on  $N = 4,000$  with a burn-in period of  $B = 1,000$ .

## GELMAN AND RUBIN (1992)

Based on the idea that given convergence has taken place, different chains will have the same distribution. Can be checked based on a suitable metric.

### ALGORITHM:

(a) Use  $K$  initial values. Iterate  $B$  steps for burn-in and  $(N-B)$  additional steps for monitoring.

(b) Calculate the following statistics:

Within chain variance,  $\mathcal{W}$

$$= \frac{1}{K(N-B-1)} \sum_{j=1}^K \sum_{t=B+1}^N (h(X_j^{(t)}) - \bar{h}_{BN,j})^2$$

Between chain variance,  $\mathcal{B}$

$$= \frac{N-B}{K-1} \sum_{j=1}^K (\bar{h}_{BN,j} - \bar{h}_{BN,\cdot})^2$$

where

$$\bar{h}_{BN,j} = \frac{1}{N-B} \sum_{t=B+1}^N h(X_j^{(t)}), \text{ and}$$

$$\bar{h}_{BN,\cdot} = \frac{1}{K} \sum_{j=1}^K \bar{h}_{BN,j}$$

- The pooled posterior variance estimate is

$$\hat{\mathcal{V}} = \left(1 - \frac{1}{N-B}\right) \mathcal{W} + \left(1 + \frac{1}{K}\right) \frac{1}{N-B} \mathcal{B}$$

- The Gelman-Rubin statistic is

$$\sqrt{R} = \sqrt{\frac{\hat{\mathcal{V}}}{\mathcal{W}}}$$

- Intuition:

(a) Before convergence,  $\mathcal{W}$  underestimates total posterior variance because it has not fully explored the target distribution.

(b)  $\hat{\mathcal{V}}$ , on the other hand, overestimates the variance because the starting points are over-dispersed relative to the target.

- $R$  is called the PSRF, or, the potential scale reduction factor:  $R$  close to 1 indicates convergence and vice versa.

## PSRFs FOR EXAMPLE 3

- IS Cauchy(0,0.5):  $R = 1.0000$
  - IS Cauchy(-1,0.5):  $R = 1.0115$
  - RWS Cauchy(0,2):  $R = 1.0006$
  - RWS Cauchy(0,0.2):  $R = 1.0029$
  - RWS Cauchy(0,0.02):  $R = 1.0054$
- 
- Five different starting values are chosen:  $-2, -1, 0, 1$  and  $2$ .
  - Table entries are based on  $N = 4,000$  iterations of the MC with a burn-in period of  $B = 1,000$ .

## PSRFs FOR HMM EXAMPLE

- Recall that the data realized were

$$Y = (0.21, 2.01, -0.36, -2.46, -2.61)$$

- Three different sets of starting values were chosen:  
(i)  $Y$ ,  $Y + 0.2$  and  $Y - 0.2$ .

- $N = 4,000$  iterations of the MC with a burn-in period of  $B = 1,000$  was run.

- $S_0$ :  $R = 1.0000$
- $S_1$ :  $R = 1.0002$
- $S_2$ :  $R = 1.0001$
- $S_3$ :  $R = 1.0000$
- $S_4$ :  $R = 0.9999$

- Check out the histograms and the estimates of the posterior means, variances and standard errors.

## PSRFs FOR VARIANCE COMPONENT MODEL

- Recall that the data was

$$Y = (1.70, 1.30, 3.53, 1.14, 3.15, \\ -2.25, 2.29, 1.77 - 3.80, 3.36)$$

- Three different sets of starting values of  $(\theta_1, \theta_2, \mu, \sigma_\theta^2, \sigma_\epsilon^2)$  were chosen:

(i) (-3.7201, 4.8943, -0.0379, 12.1644, 14.7669)

(ii) (-2.9594, 1.8781, -0.0395, 12.6582, 15.3662)

(iii) (0.7465, -3.3410, -0.0390, 12.5008, 15.1752)

- $N = 4,000$  iterations of the MC with a burn-in period of  $B = 1,000$  was run.

- $\theta_1$ : R = 1.0001
- $\theta_2$ : R = 0.9999
- $\mu$ : R = 1.0000
- $\sigma_\theta^2$ : R = 0.9999
- $\sigma_\epsilon^2$ : R = 0.9999

- Check out the histograms and the estimates of the posterior means, variances and standard errors.

## BAYESIAN MODEL DIAGNOSTICS

- Let  $Y_1, Y_2, \dots, Y_n$  be iid from  $f(y | \theta)$ . The unknown parameter is denoted by  $\Theta = \{\theta\}$ .
- We want to examine the influence of  $y_j$  to the fit.
- Do this by cross validation using the predictive distribution of  $y_j$ .

$$p(y_j | y_{-j}) = \int_{\Theta} f(y_j | y_{-j}, \theta) \cdot p(\theta | y_{-j}) d\theta$$

- This is called the conditional predictive ordinate (CPO).
- Can estimate the residual by

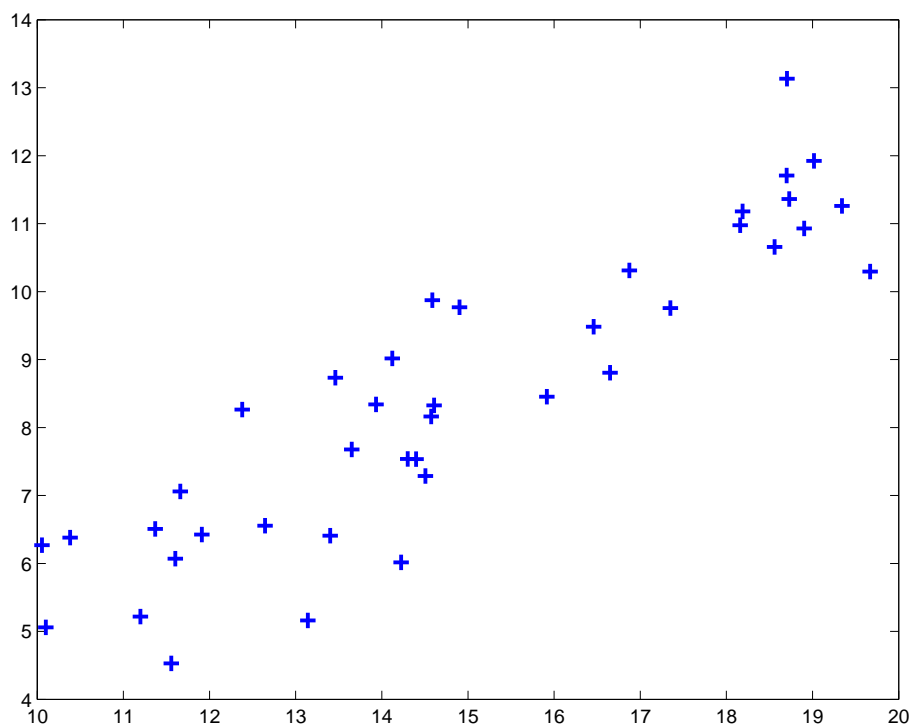
$$y_j - E(y_j | y_{-j})$$

## EXAMPLE 6: SIMPLE LINEAR REGRESSION

$$M_1: y_i = \beta_0 + \epsilon_i, \epsilon_i \stackrel{i.i.d}{\sim} N(0, 1)$$

$$M_2: y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \epsilon_i \stackrel{i.i.d}{\sim} N(0, 1)$$

DATA:



Sample size:  $n = 40$ .

## LET'S DEVELOP THE METHODOLOGY

- Index model  $M_k$  by  $k$ ,  $k = 1, 2$ . We can write both models in the matrix form

$$\mathbf{Y}_{n \times 1} = X_{n \times k}^{(k)} \beta_{k \times 1}^{(k)} + \epsilon_{n \times 1}$$

where

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \beta^{(k)} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{(k-1)} \end{pmatrix},$$

$$X^{(1)} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \quad \text{and} \quad X^{(2)} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

- Likelihood:

$$L^{(k)}(\mathbf{Y} | \beta^{(k)}) = \frac{1}{(2\pi)^{n/2}} \times \exp \left\{ -\frac{1}{2} (\mathbf{Y} - X^{(k)} \beta^{(k)})^T (\mathbf{Y} - X^{(k)} \beta^{(k)}) \right\}$$

- Prior on  $\beta^{(k)}$  is  $N(0, I/c)$ :

$$\pi_0(\beta^{(k)}) = \prod_{j=0}^{(k-1)} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{c}{2}\beta^T \beta\right\}$$

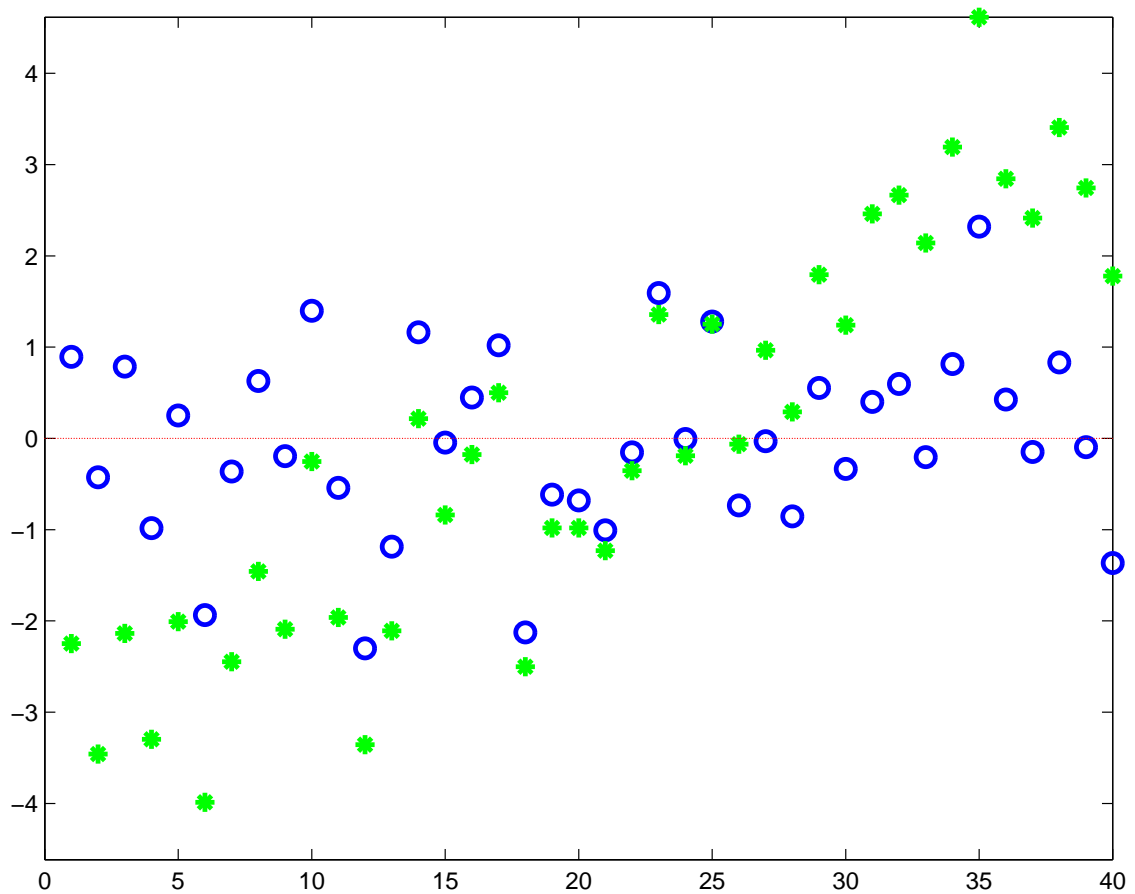
- STEP 1: Calculate the posterior of  $\beta^{(k)}$
- STEP 2: Reduce  $n$  to  $n - 1$  for the cross validation procedure
- STEP 3: Calculate  $E(y_j | y_{-j})$
- STEP 4: Calculate  $y_j - E(y_j | y_{-j})$

## SIMPLE LINEAR REGRESSION (CONT.)

Graph of residuals based on:

(a)  $M_1$  (in green)

(b)  $M_2$  (in blue)



## MODEL SELECTION

We can do model selection based on the pseudo-Bayes factor given by

$$\text{PBF} = \frac{\prod_{j=1}^n f(y_j | y_{-j}, M_1)}{\prod_{j=1}^n f(y_j | y_{-j}, M_2)}$$

This is a variant of the Bayes factor

$$\text{BF} = \frac{\text{Marginal likelihood under } M_1}{\text{Marginal likelihood under } M_2}$$

FOR OUR EXAMPLE 6: PBF is  $1.3581 \times 10^{-31}$ .

(Observations came from  $M_2$  with  $\beta_0 = -0.4, \beta_1 = 0.6$ )

## WHAT IF CLOSED FORMS ARE NOT AVAILABLE

- In the example we used, the predictive density  $f(y_j | y_{-j})$  and the expected value  $E(y_j | y_{-j})$  could be calculated in a closed form.

- However, this is not always the case.

- Note that we are interested in the quantity

$$E(y_j | y_{-j}) = \int_{\Theta} E(y_j | \theta) \pi(\theta | y_{-j}) d\theta$$

- Material is from Gelfand and Dey (1994)

CASE I:  $E(y_j | \theta) = a_j(\theta)$

- Dependence on  $M_k$  is suppressed.
- We want to estimate the quantity

$$E(y_j | y_{-j}) = \frac{\int_{\Theta} a_j(\theta) \pi(\theta | y_{-j}) d\theta}{\int_{\Theta} \pi(\theta | y_{-j}) d\theta}$$

- Recall importance sampling: If we have  $\theta_1^*, \theta_2^*, \dots, \theta_N^*$  i.i.d. samples from  $g(\cdot)$ , then

$$\begin{aligned} \int_{\Theta} a_j(\theta) \pi(\theta | y_{-j}) d\theta &= \int_{\Theta} a_j(\theta) \frac{\pi(\theta | y_{-j})}{g(\theta)} g(\theta) d\theta \\ &\approx \frac{1}{N} \sum_{i=1}^N a_j(\theta_i^*) \frac{\pi(\theta_i^* | y_{-j})}{g(\theta_i^*)} \end{aligned}$$

and

$$\begin{aligned} \int_{\Theta} \pi(\theta | y_{-j}) d\theta &= \int_{\Theta} \frac{\pi(\theta | y_{-j})}{g(\theta)} g(\theta) d\theta \\ &\approx \frac{1}{N} \sum_{i=1}^N \frac{\pi(\theta_i^* | y_{-j})}{g(\theta_i^*)} \end{aligned}$$

- Essential that  $g(\theta)$  closely resembles  $\pi(\theta | y_{-j})$ .
- Thus, a good choice of  $g$  is the complete posterior density,  $\pi(\theta | \mathbf{y})$ .
- Then,

$$\frac{1}{N} \sum_{i=1}^N a_j(\theta_i^*) \frac{\pi(\theta_i^* | y_{-j})}{g(\theta_i^*)} = \frac{m(\mathbf{y})}{m(y_{-j})} \frac{1}{N} \sum_{i=1}^N \frac{a_j(\theta_i^*)}{L(y_j | \theta_i^*)}$$

$$\frac{1}{N} \sum_{i=1}^N \frac{\pi(\theta_i^* | y_{-j})}{g(\theta_i^*)} = \frac{m(\mathbf{y})}{m(y_{-j})} \frac{1}{N} \sum_{i=1}^N \frac{1}{L(y_j | \theta_i^*)}$$

- So, we have

$$E(y_j | y_{-j}) \approx \frac{\frac{1}{N} \sum_{i=1}^N \frac{a_j(\theta_i^*)}{L(y_j | \theta_i^*)}}{\frac{1}{N} \sum_{i=1}^N \frac{1}{L(y_j | \theta_i^*)}}$$

## BACK TO THE REGRESSION EXAMPLE

- For  $M_1$ , we have

$$E(y_j | \beta_0) = \beta_0$$

and

$$L(y_j | \beta_0) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(y_j - \beta_0)^2\right\}$$

- The posterior  $\pi(\beta_0 | \mathbf{y})$  is given by

$$\pi(\beta_0 | \mathbf{y}) = N\left(\frac{n}{n+c}\bar{y}, \frac{1}{(n+c)}\right).$$

- So, plug-in the above expressions into the general formula of  $E(y_j | y_{-j})$  to get the explicit expression for this example.

## BACK TO THE REGRESSION EXAMPLE (CONT.)

- For  $M_2$ , we have

$$E(y_j | \beta_0, \beta_1) = \beta_0 + \beta_1 x_j$$

and

$$L(y_j | \beta_0, \beta_1) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(y_j - \beta_0 - \beta_1 x_j)^2\right\}$$

- The posterior  $\pi(\beta_0, \beta_1 | \mathbf{y})$  is given by

$$\pi(\beta_0, \beta_1 | \mathbf{y}) = N(\hat{\beta}_c, A)$$

where

$$\hat{\beta}_c = (X'X + cI)^{-1} X'Y$$

and

$$A = (X'X + cI)^{-1}$$

- So, plug-in the above expressions into the general formula of  $E(y_j | y_{-j})$  to get the explicit expression for this example.

## CALCULATE PBF FOR THE REGRESSION EXAMPLE

- To obtain  $f(y_j | y_{-j})$ , replace  $a_j(\theta)$  by 1.
- In the regression example, the PBF is  $5.4367 \times 10^{-29}$ .

## CASE II: $E(y_j | \theta)$ NOT IN CLOSED FORM

- Recall that

$$E(y_j | y_{-j}) = \int_{y_j} y_j \pi(y_j | y_{-j}) dy_j$$

- Use importance sampling once more:

Let  $y_{j,1}^*, y_{j,2}^*, \dots, y_{j,N}^*$  be samples from  $\pi(y_j | y_{-j})$ .  
Then,

$$E(y_j | y_{-j}) \approx \frac{1}{N} \sum_{i=1}^N y_{j,i}^*$$

- How to generate samples from  $\pi(y_j | y_{-j})$ ?
  - (i) First generate  $\theta_{j,i}^*$  from  $\pi(\theta | y_{-j})$ , and
  - (ii) Then, generate  $y_{j,i}^*$  from  $L(y_j | \theta_{j,i}^*)$ .

- SIR (Sampling Importance Resampling) is one way to convert samples from  $\pi(\theta | y)$  to samples of  $\pi(\theta | y_{-j})$
- General set-up: Suppose we have  $N$  samples from  $g(\cdot)$ ,  $\theta_1, \theta_2, \dots, \theta_N$ . The goal is to obtain a sample of  $M$  observations from  $f(\cdot)$ .
- Idea: Assign a sampling weight  $w_i = w(\theta_i)$  to the sample  $\theta_i$ . If  $\theta^*$  is a draw from  $\theta_1, \theta_2, \dots, \theta_N$  with selection probabilities  $w_1, w_2, \dots, w_N$ , then

$$\begin{aligned}
 P(\theta^* \in \mathcal{B}) &= \frac{\sum_{i=1}^N w_i I\{\theta_i \in \mathcal{B}\}}{\sum_{i=1}^N w_i} \\
 &\rightarrow \frac{\int_{\theta \in \mathcal{B}} w(\theta) g(\theta) d\theta}{\int_{\theta \in \mathcal{R}} w(\theta) g(\theta) d\theta} \\
 &= \frac{\int_{\theta \in \mathcal{B}} f(\theta) d\theta}{\int_{\theta \in \mathcal{R}} f(\theta) d\theta}
 \end{aligned}$$

if the weights are chosen as

$$w_i = w(\theta_i) \propto f(\theta_i)/g(\theta_i)$$

- Normalizing by  $\sum_{i=1}^N w_i$  in the denominator helps remove unwanted constants.

## EXERCISE

- Assume  $Y_1, Y_2, \dots, Y_n$  are iid  $N(\mu, \sigma^2)$ .

- Data is

6.1	7.6	7.5	4.2	5.7
4.3	5.6	8.4	5.3	6.0
6.7	6.2	6.6	6.2	7.0
4.2	5.4	5.4	1.2	5.2

- Obtain estimates of  $\mu$  and  $\sigma^2$  using MCMC techniques and appropriate prior distributions.
- Is there evidence that the data does not come from the normal distribution?