Introduction to Bayesian Data Analysis and Markov Chain Monte Carlo

Jeffrey S. Morris
University of Texas M.D. Anderson Cancer Center
Department of Biostatistics
jeffmo@mdanderson.org

September 20, 2002

Abstract

The purpose of this talk is to give a brief overview of Bayesian Inference and Markov Chain Monte Carlo methods, including the Gibbs Sampler and Metropolis Hastings algorithm.

Outline

- Bayesian vs. Frequentist paradigm
- Bayesian Inference and MCMC
  - ⋆ Gibbs Sampler
  - ⋆ Metropolis-Hastings Algorithm
- Assessing Convergence of MCMC
- Hierarchical Model Example
- MCMC: Benefits and Cautions

Frequentist vs. Bayesian paradigms

- Data: $\mathbf{X}$ Parameters: $\mathbf{\Theta}$

Frequentist vs. Bayesian paradigms

- Data: $\mathbf{X}$ Parameters: $\boldsymbol{\Theta}$
- To a frequentist:
  - ⋆ The data $\mathbf{X}$ are random, and the parameters $\boldsymbol{\Theta}$ are fixed.

Frequentist vs. Bayesian paradigms

- Data: $\mathbf{X}$ Parameters: $\boldsymbol{\Theta}$
- To a frequentist:
  - ⋆ The data $\mathbf{X}$ are random, and the parameters $\boldsymbol{\Theta}$ are fixed.
  - ⋆ (ML) Inference is performed by finding $\boldsymbol{\Theta}$ such that $f(\mathbf{X}|\boldsymbol{\Theta})$ is maximized.

Frequentist vs. Bayesian paradigms

- Data: $\mathbf{X}$ Parameters: $\mathbf{\Theta}$

- To a frequentist:
  - ⋆ The data $\mathbf{X}$ are random, and the parameters $\mathbf{\Theta}$ are fixed.
  - ⋆ (ML) Inference is performed by finding $\mathbf{\Theta}$ such that $f(\mathbf{X}|\mathbf{\Theta})$ is maximized.
  - ⋆ We <u>cannot</u> make probability statements about parameters, but only can make statements about performance of estimators over repeated sampling (e.g.confidence intervals).

Frequentist vs. Bayesian paradigms

- Data: $\mathbf{X}$ Parameters: $\boldsymbol{\Theta}$

- To a frequentist:

  ⋆ The data $\mathbf{X}$ are random, and the parameters $\boldsymbol{\Theta}$ are fixed.
  ⋆ (ML) Inference is performed by finding $\boldsymbol{\Theta}$ such that $f(\mathbf{X}|\boldsymbol{\Theta})$ is maximized.
  ⋆ We <u>cannot</u> make probability statements about parameters, but only can make statements about performance of estimators over repeated sampling (e.g.confidence intervals).

- To a Bayesian:

  ⋆ The current data $\mathbf{X}$ is fixed, and the unknown parameters $\boldsymbol{\Theta}$ are random.

Frequentist vs. Bayesian paradigms

- Data: $\mathbf{X}$ Parameters: $\boldsymbol{\Theta}$
- To a frequentist:
  - ⋆ The data $\mathbf{X}$ are random, and the parameters $\boldsymbol{\Theta}$ are fixed.
  - ⋆ (ML) Inference is performed by finding $\boldsymbol{\Theta}$ such that $f(\mathbf{X}|\boldsymbol{\Theta})$ is maximized.
  - ⋆ We <u>cannot</u> make probability statements about parameters, but only can make statements about performance of estimators over repeated sampling (e.g.confidence intervals).
- To a Bayesian:
  - ⋆ The current data $\mathbf{X}$ is fixed, and the unknown parameters $\boldsymbol{\Theta}$ are random.
  - ⋆ Inference is performed via the posterior distribution $f(\boldsymbol{\Theta}|\mathbf{X})$.

Frequentist vs. Bayesian paradigms

- Data: $\mathbf{X}$ Parameters: $\boldsymbol{\Theta}$

- To a frequentist:
  - ⋆ The data $\mathbf{X}$ are random, and the parameters $\boldsymbol{\Theta}$ are fixed.
  - ⋆ (ML) Inference is performed by finding $\boldsymbol{\Theta}$ such that $f(\mathbf{X}|\boldsymbol{\Theta})$ is maximized.
  - ⋆ We cannot make probability statements about parameters, but only can make statements about performance of estimators over repeated sampling (e.g.confidence intervals).

- To a Bayesian:
  - ⋆ The current data $\mathbf{X}$ is fixed, and the unknown parameters $\boldsymbol{\Theta}$ are random.
  - ⋆ Inference is performed via the posterior distribution $f(\boldsymbol{\Theta}|\mathbf{X})$.
  - ⋆ We can make probability statements about parameters, since they are random quantities (e.g. credible intervals)

Bayes' Rule

- The posterior distribution is computed by applying **Bayes' Rule:**

$$f(\mathbf{\Theta}|\mathbf{X}) = \frac{f(\mathbf{X}|\mathbf{\Theta})f(\mathbf{\Theta})}{f(\mathbf{X})}$$

Bayes' Rule

- The posterior distribution is computed by applying **Bayes' Rule:**

$$f(\mathbf{\Theta}|\mathbf{X}) = \frac{f(\mathbf{X}|\mathbf{\Theta})f(\mathbf{\Theta})}{f(\mathbf{X})}$$

- $f(\mathbf{X}|\mathbf{\Theta})$= Likelihood

Bayes' Rule

- The posterior distribution is computed by applying **Bayes' Rule:**

$$f(\boldsymbol{\Theta}|\mathbf{X}) = \frac{f(\mathbf{X}|\boldsymbol{\Theta})f(\boldsymbol{\Theta})}{f(\mathbf{X})}$$

- $f(\mathbf{X}|\boldsymbol{\Theta})$= Likelihood
- $f(\boldsymbol{\Theta})$= Prior Distribution

Bayes' Rule

- The posterior distribution is computed by applying **Bayes' Rule:**

$$f(\mathbf{\Theta}|\mathbf{X}) = \frac{f(\mathbf{X}|\mathbf{\Theta})f(\mathbf{\Theta})}{f(\mathbf{X})}$$

- $f(\mathbf{X}|\mathbf{\Theta})$= Likelihood
- $f(\mathbf{\Theta})$= Prior Distribution
  - ⋆ Reflects prior knowledge about $\mathbf{\Theta}$

Bayes' Rule

- The posterior distribution is computed by applying **Bayes' Rule:**

$$f(\mathbf{\Theta}|\mathbf{X}) = \frac{f(\mathbf{X}|\mathbf{\Theta})f(\mathbf{\Theta})}{f(\mathbf{X})}$$

- $f(\mathbf{X}|\mathbf{\Theta})$= Likelihood
- $f(\mathbf{\Theta})$= Prior Distribution
  - ⋆ Reflects prior knowledge about $\mathbf{\Theta}$
  - ⋆ Sometimes controversial

Bayes' Rule

- The posterior distribution is computed by applying **Bayes' Rule:**

$$f(\mathbf{\Theta}|\mathbf{X}) = \frac{f(\mathbf{X}|\mathbf{\Theta})f(\mathbf{\Theta})}{f(\mathbf{X})}$$

- $f(\mathbf{X}|\mathbf{\Theta})$= Likelihood
- $f(\mathbf{\Theta})$= Prior Distribution
  - ⋆ Reflects prior knowledge about $\mathbf{\Theta}$
  - ⋆ Sometimes controversial
  - ⋆ If little information available, just use *diffuse* priors (avoid improper priors)

Bayes' Rule

- The posterior distribution is computed by applying **Bayes' Rule:**

$$f(\mathbf{\Theta}|\mathbf{X}) = \frac{f(\mathbf{X}|\mathbf{\Theta})f(\mathbf{\Theta})}{f(\mathbf{X})}$$

- $f(\mathbf{X}|\mathbf{\Theta})$= Likelihood
- $f(\mathbf{\Theta})$= Prior Distribution
  - ⋆ Reflects prior knowledge about $\mathbf{\Theta}$
  - ⋆ Sometimes controversial
  - ⋆ If little information available, just use *diffuse* priors (avoid improper priors)
- $f(\mathbf{X})$= Marginal Distribution = $\int f(\mathbf{X}|\mathbf{\Theta})f(\mathbf{\Theta})d\mathbf{\Theta}$

Bayes' Rule

- The posterior distribution is computed by applying **Bayes' Rule:**

$$f(\mathbf{\Theta}|\mathbf{X}) = \frac{f(\mathbf{X}|\mathbf{\Theta})f(\mathbf{\Theta})}{f(\mathbf{X})}$$

- $f(\mathbf{X}|\mathbf{\Theta})$= Likelihood
- $f(\mathbf{\Theta})$= Prior Distribution
  - ⋆ Reflects prior knowledge about $\mathbf{\Theta}$
  - ⋆ Sometimes controversial
  - ⋆ If little information available, just use *diffuse* priors (avoid improper priors)
- $f(\mathbf{X})$= Marginal Distribution = $\int f(\mathbf{X}|\mathbf{\Theta})f(\mathbf{\Theta})d\mathbf{\Theta}$
  - ⋆ Difficult to compute (usually intractable integral)

Bayes' Rule

- The posterior distribution is computed by applying **Bayes' Rule:**

$$f(\boldsymbol{\Theta}|\mathbf{X}) = \frac{f(\mathbf{X}|\boldsymbol{\Theta})f(\boldsymbol{\Theta})}{f(\mathbf{X})}$$

- $f(\mathbf{X}|\boldsymbol{\Theta})$= Likelihood
- $f(\boldsymbol{\Theta})$= Prior Distribution
  - ⋆ Reflects prior knowledge about $\boldsymbol{\Theta}$
  - ⋆ Sometimes controversial
  - ⋆ If little information available, just use *diffuse* priors (avoid improper priors)
- $f(\mathbf{X})$= Marginal Distribution = $\int f(\mathbf{X}|\boldsymbol{\Theta})f(\boldsymbol{\Theta})d\boldsymbol{\Theta}$
  - ⋆ Difficult to compute (usually intractable integral)
  - ⋆ Often not necessary to compute.

Conjugate priors

- *Conjugate priors*: $f(\mathbf{\Theta})$ and $f(\mathbf{\Theta}|\mathbf{X})$ have same distributional form.

Conjugate priors

- *Conjugate priors*: $f(\boldsymbol{\Theta})$ and $f(\boldsymbol{\Theta}|\mathbf{X})$ have same distributional form.
- Examples: Normal-Normal, Beta-Binomial, Gamma-Poisson

Conjugate priors

- *Conjugate priors*: $f(\mathbf{\Theta})$ and $f(\mathbf{\Theta}|\mathbf{X})$ have same distributional form.
- Examples: Normal-Normal, Beta-Binomial, Gamma-Poisson
- Ex: $(X|\theta) \sim$ Binomial(n, θ); $\qquad \theta \sim$ Beta(α, β)

Conjugate priors

- *Conjugate priors*: $f(\boldsymbol{\Theta})$ and $f(\boldsymbol{\Theta}|\mathbf{X})$ have same distributional form.
- Examples: Normal-Normal, Beta-Binomial, Gamma-Poisson
- Ex: $(X|\theta) \sim$ Binomial(n, θ); $\qquad \theta \sim$ Beta(α, β)

$$f(\theta|X) \quad \propto \quad f(X|\theta)f(\theta)$$

Conjugate priors

- *Conjugate priors*: $f(\boldsymbol{\Theta})$ and $f(\boldsymbol{\Theta}|\mathbf{X})$ have same distributional form.
- Examples: Normal-Normal, Beta-Binomial, Gamma-Poisson
- Ex: $(X|\theta) \sim$ Binomial(n, θ); $\qquad \theta \sim$ Beta(α, β)

$$\begin{aligned} f(\theta|X) &\propto f(X|\theta)f(\theta) \\ &\propto \theta^X(1-\theta)^{n-X}\theta^\alpha(1-\theta)^\beta \end{aligned}$$

Conjugate priors

- *Conjugate priors*: $f(\boldsymbol{\Theta})$ and $f(\boldsymbol{\Theta}|\mathbf{X})$ have same distributional form.
- Examples: Normal-Normal, Beta-Binomial, Gamma-Poisson
- Ex: $(X|\theta) \sim$ Binomial(n, θ); $\qquad \theta \sim$ Beta(α, β)

$$
\begin{aligned}
f(\theta|X) &\propto f(X|\theta)f(\theta) \\
&\propto \theta^X(1-\theta)^{n-X}\theta^\alpha(1-\theta)^\beta \\
&= \theta^{\alpha+X}(1-\theta)^{\beta+n-X}
\end{aligned}
$$

Conjugate priors

- *Conjugate priors*: $f(\mathbf{\Theta})$ and $f(\mathbf{\Theta}|\mathbf{X})$ have same distributional form.
- Examples: Normal-Normal, Beta-Binomial, Gamma-Poisson
- Ex: $(X|\theta) \sim$ Binomial(n, θ); $\theta \sim$ Beta(α, β)

$$\begin{aligned} f(\theta|X) &\propto f(X|\theta)f(\theta) \\ &\propto \theta^X(1-\theta)^{n-X}\theta^\alpha(1-\theta)^\beta \\ &= \theta^{\alpha+X}(1-\theta)^{\beta+n-X} \\ &= \text{kernel of Beta}(\alpha+X, \beta+n-X) \end{aligned}$$

Conjugate priors

- *Conjugate priors*: $f(\boldsymbol{\Theta})$ and $f(\boldsymbol{\Theta}|\mathbf{X})$ have same distributional form.
- Examples: Normal-Normal, Beta-Binomial, Gamma-Poisson
- Ex: $(X|\theta) \sim$ Binomial(n, θ); $\theta \sim$ Beta(α, β)

$$\begin{aligned} f(\theta|X) &\propto f(X|\theta)f(\theta) \\ &\propto \theta^X(1-\theta)^{n-X}\theta^\alpha(1-\theta)^\beta \\ &= \theta^{\alpha+X}(1-\theta)^{\beta+n-X} \\ &= \text{kernel of Beta}(\alpha+X, \beta+n-X) \end{aligned}$$

- For single parameter problem: conjugate priors allow closed form posterior distributions.

Conjugate priors

- *Conjugate priors*: $f(\mathbf{\Theta})$ and $f(\mathbf{\Theta}|\mathbf{X})$ have same distributional form.
- Examples: Normal-Normal, Beta-Binomial, Gamma-Poisson
- Ex: $(X|\theta) \sim$ Binomial(n, θ); $\qquad \theta \sim$ Beta(α, β)

$$\begin{aligned} f(\theta|X) &\propto f(X|\theta)f(\theta) \\ &\propto \theta^X(1-\theta)^{n-X}\theta^\alpha(1-\theta)^\beta \\ &= \theta^{\alpha+X}(1-\theta)^{\beta+n-X} \\ &= \text{kernel of Beta}(\alpha+X, \beta+n-X) \end{aligned}$$

- For single parameter problem: conjugate priors allow closed form posterior distributions.
- What if we don't want to use conjugate priors?

Conjugate priors

- *Conjugate priors*: $f(\mathbf{\Theta})$ and $f(\mathbf{\Theta}|\mathbf{X})$ have same distributional form.
- Examples: Normal-Normal, Beta-Binomial, Gamma-Poisson
- Ex: $(X|\theta) \sim$ Binomial(n, θ); $\qquad \theta \sim$ Beta(α, β)

$$\begin{aligned} f(\theta|X) &\propto f(X|\theta)f(\theta) \\ &\propto \theta^X(1-\theta)^{n-X}\theta^\alpha(1-\theta)^\beta \\ &= \theta^{\alpha+X}(1-\theta)^{\beta+n-X} \\ &= \text{kernel of Beta}(\alpha+X, \beta+n-X) \end{aligned}$$

- For single parameter problem: conjugate priors allow closed form posterior distributions.
- What if we don't want to use conjugate priors?

  What if we have multiple parameters?

Non-conjugate Case

- Suppose we are interested in the posterior mean:

$$E(\boldsymbol{\Theta}|\mathbf{X}) = \int \boldsymbol{\Theta} f(\Theta|\mathbf{X}) d\boldsymbol{\Theta}$$

Non-conjugate Case

- Suppose we are interested in the posterior mean:

$$
\begin{aligned}
E(\mathbf{\Theta}|\mathbf{X}) &= \int \mathbf{\Theta} f(\Theta|\mathbf{X}) d\mathbf{\Theta} \\
&= \frac{\int \mathbf{\Theta} f(\mathbf{X}|\mathbf{\Theta}) f(\mathbf{\Theta}) d\mathbf{\Theta}}{\int f(\mathbf{X}|\mathbf{\Theta}) f(\mathbf{\Theta}) d\mathbf{\Theta}}
\end{aligned}
$$

Non-conjugate Case

- Suppose we are interested in the posterior mean:

$$
\begin{aligned}
E(\boldsymbol{\Theta}|\mathbf{X}) &= \int \boldsymbol{\Theta} f(\Theta|\mathbf{X}) d\boldsymbol{\Theta} \\
&= \frac{\int \boldsymbol{\Theta} f(\mathbf{X}|\boldsymbol{\Theta}) f(\boldsymbol{\Theta}) d\boldsymbol{\Theta}}{\int f(\mathbf{X}|\boldsymbol{\Theta}) f(\boldsymbol{\Theta}) d\boldsymbol{\Theta}}
\end{aligned}
$$

- How do we compute this integral if it is intractable?

Non-conjugate Case

- Suppose we are interested in the posterior mean:

$$
\begin{aligned}
E(\boldsymbol{\Theta}|\mathbf{X}) &= \int \boldsymbol{\Theta} f(\Theta|\mathbf{X}) d\boldsymbol{\Theta} \\
&= \frac{\int \boldsymbol{\Theta} f(\mathbf{X}|\boldsymbol{\Theta}) f(\boldsymbol{\Theta}) d\boldsymbol{\Theta}}{\int f(\mathbf{X}|\boldsymbol{\Theta}) f(\boldsymbol{\Theta}) d\boldsymbol{\Theta}}
\end{aligned}
$$

- How do we compute this integral if it is intractable?
  - ⋆ Numerical Integration (Quadrature)

Non-conjugate Case

- Suppose we are interested in the posterior mean:

$$
\begin{aligned}
E(\boldsymbol{\Theta}|\mathbf{X}) &= \int \boldsymbol{\Theta} f(\Theta|\mathbf{X}) d\boldsymbol{\Theta} \\
&= \frac{\int \boldsymbol{\Theta} f(\mathbf{X}|\boldsymbol{\Theta}) f(\boldsymbol{\Theta}) d\boldsymbol{\Theta}}{\int f(\mathbf{X}|\boldsymbol{\Theta}) f(\boldsymbol{\Theta}) d\boldsymbol{\Theta}}
\end{aligned}
$$

- How do we compute this integral if it is intractable?
  - ⋆ Numerical Integration (Quadrature)
    May not work if there are many parameters.

Non-conjugate Case

- Suppose we are interested in the posterior mean:

$$
\begin{aligned}
E(\boldsymbol{\Theta}|\mathbf{X}) &= \int \boldsymbol{\Theta} f(\Theta|\mathbf{X}) d\boldsymbol{\Theta} \\
&= \frac{\int \boldsymbol{\Theta} f(\mathbf{X}|\boldsymbol{\Theta}) f(\boldsymbol{\Theta}) d\boldsymbol{\Theta}}{\int f(\mathbf{X}|\boldsymbol{\Theta}) f(\boldsymbol{\Theta}) d\boldsymbol{\Theta}}
\end{aligned}
$$

- How do we compute this integral if it is intractable?
  - ⋆ Numerical Integration (Quadrature)
    May not work if there are many parameters.
  - ⋆ Monte Carlo integration

Markov Chain Monte Carlo: Monte Carlo Integration

- *Monte Carlo* integration:

  Estimate integrals by randomly drawing samples from the required distribution.

Markov Chain Monte Carlo: Monte Carlo Integration

- *Monte Carlo* integration:

  Estimate integrals by randomly drawing samples from the required distribution.

$$E(\mathbf{\Theta}|\mathbf{X}) = \int \mathbf{\Theta} f(\mathbf{\Theta}|\mathbf{X}) d\mathbf{\Theta}$$

Markov Chain Monte Carlo: Monte Carlo Integration

- *Monte Carlo* integration:

  Estimate integrals by randomly drawing samples from the required distribution.

$$
\begin{aligned}
E(\boldsymbol{\Theta}|\mathbf{X}) &= \int \boldsymbol{\Theta} f(\boldsymbol{\Theta}|\mathbf{X}) d\boldsymbol{\Theta} \\
&\approx \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{\Theta}_t,
\end{aligned}
$$

  where $\boldsymbol{\Theta}_t \sim f(\boldsymbol{\Theta}|\mathbf{X})$

Markov Chain Monte Carlo: Monte Carlo Integration

- *Monte Carlo* integration:

  Estimate integrals by randomly drawing samples from the required distribution.

$$
\begin{aligned}
E(\boldsymbol{\Theta}|\mathbf{X}) &= \int \boldsymbol{\Theta} f(\boldsymbol{\Theta}|\mathbf{X}) d\boldsymbol{\Theta} \\
&\approx \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{\Theta}_t,
\end{aligned}
$$

  where $\boldsymbol{\Theta}_t \sim f(\boldsymbol{\Theta}|\mathbf{X})$

- We still need a method for drawing samples from the posterior distribution:

Markov Chain Monte Carlo: Monte Carlo Integration

- *Monte Carlo* integration:

  Estimate integrals by randomly drawing samples from the required distribution.

$$\begin{aligned} E(\boldsymbol{\Theta}|\mathbf{X}) &= \int \boldsymbol{\Theta} f(\boldsymbol{\Theta}|\mathbf{X}) d\boldsymbol{\Theta} \\ &\approx \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{\Theta}_t, \end{aligned}$$

  where $\boldsymbol{\Theta}_t \sim f(\boldsymbol{\Theta}|\mathbf{X})$

- We still need a method for drawing samples from the posterior distribution:
  - ⋆ Rejection Sampling

Markov Chain Monte Carlo: Monte Carlo Integration

- *Monte Carlo* integration:

  Estimate integrals by randomly drawing samples from the required distribution.

$$\begin{aligned} E(\mathbf{\Theta}|\mathbf{X}) &= \int \mathbf{\Theta} f(\mathbf{\Theta}|\mathbf{X}) d\mathbf{\Theta} \\ &\approx \frac{1}{n} \sum_{t=1}^{n} \mathbf{\Theta}_t, \end{aligned}$$

  where $\mathbf{\Theta}_t \sim f(\mathbf{\Theta}|\mathbf{X})$

- We still need a method for drawing samples from the posterior distribution:
  - ⋆ Rejection Sampling
  - ⋆ Importance Sampling

Markov Chain Monte Carlo: Monte Carlo Integration

- *Monte Carlo* integration:

  Estimate integrals by randomly drawing samples from the required distribution.

$$\begin{aligned} E(\mathbf{\Theta}|\mathbf{X}) &= \int \mathbf{\Theta} f(\mathbf{\Theta}|\mathbf{X}) d\mathbf{\Theta} \\ &\approx \frac{1}{n}\sum_{t=1}^{n} \mathbf{\Theta}_t, \end{aligned}$$

  where $\mathbf{\Theta}_t \sim f(\mathbf{\Theta}|\mathbf{X})$

- We still need a method for drawing samples from the posterior distribution:
  - ⋆ Rejection Sampling
  - ⋆ Importance Sampling
  - ⋆ Markov Chain

Markov Chain Monte Carlo: Markov Chains

- *Markov Chain* : Method to draw samples from a desired stationary distribution.

Markov Chain Monte Carlo: Markov Chains

- *Markov Chain* : Method to draw samples from a desired stationary distribution.

- Steps:

1. Obtain starting values Θ_0

Markov Chain Monte Carlo: Markov Chains

- *Markov Chain* : Method to draw samples from a desired stationary distribution.

- Steps:

1. Obtain starting values $\mathbf{\Theta}_0$
2. Sample $\mathbf{\Theta}_1$ from suitably chosen transition kernel $P(\mathbf{\Theta}_1|\mathbf{\Theta}_0)$

Markov Chain Monte Carlo: Markov Chains

- *Markov Chain* : Method to draw samples from a desired stationary distribution.

- Steps:

1. Obtain starting values $\mathbf{\Theta}_0$
2. Sample $\mathbf{\Theta}_1$ from suitably chosen transition kernel $P(\mathbf{\Theta}_1|\mathbf{\Theta}_0)$
3. Repeat second step n times to obtain chain $\{\mathbf{\Theta}_0, \mathbf{\Theta}_1, \ldots, \mathbf{\Theta}_n\}$.

Markov Chain Monte Carlo: Markov Chains

- *Markov Chain* : Method to draw samples from a desired stationary distribution.
- Steps:
  1. Obtain starting values $\boldsymbol{\Theta}_0$
  2. Sample $\boldsymbol{\Theta}_1$ from suitably chosen transition kernel $P(\boldsymbol{\Theta}_1|\boldsymbol{\Theta}_0)$
  3. Repeat second step n times to obtain chain $\{\boldsymbol{\Theta}_0, \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_n\}$.
- Theorems show that, under certain regularity conditions, the chain will converge to a particular *stationary distribution* after suitable *burn-in* period.

Markov Chain Monte Carlo: Markov Chains

- *Markov Chain* : Method to draw samples from a desired stationary distribution.

- Steps:

1. Obtain starting values $\mathbf{\Theta}_0$
2. Sample $\mathbf{\Theta}_1$ from suitably chosen transition kernel $P(\mathbf{\Theta}_1|\mathbf{\Theta}_0)$
3. Repeat second step n times to obtain chain $\{\mathbf{\Theta}_0, \mathbf{\Theta}_1, \ldots, \mathbf{\Theta}_n\}$.

- Theorems show that, under certain regularity conditions, the chain will converge to a particular *stationary distribution* after suitable *burn-in* period.

- End result: A (correlated) sample from the stationary distribution.

Markov Chain Monte Carlo

- Given Markov Chain $\{\mathbf{\Theta}_0, \mathbf{\Theta}_1, \ldots, \mathbf{\Theta}_n\}$ with stationary distribution $f(\mathbf{\Theta}|\mathbf{X})$ with burn-in m, we can estimate the posterior mean using Monte Carlo integration:

Markov Chain Monte Carlo

- Given Markov Chain $\{\mathbf{\Theta}_0, \mathbf{\Theta}_1, \ldots, \mathbf{\Theta}_n\}$ with stationary distribution $f(\mathbf{\Theta}|\mathbf{X})$ with burn-in m, we can estimate the posterior mean using Monte Carlo integration:

$$E(\mathbf{\Theta}|\mathbf{X}) \approx \frac{1}{n-m} \sum_{t=m+1}^{n} \mathbf{\Theta}_t.$$

Markov Chain Monte Carlo

- Given Markov Chain $\{\mathbf{\Theta}_0, \mathbf{\Theta}_1, \ldots, \mathbf{\Theta}_n\}$ with stationary distribution $f(\mathbf{\Theta}|\mathbf{X})$ with burn-in m, we can estimate the posterior mean using Monte Carlo integration:

$$E(\mathbf{\Theta}|\mathbf{X}) \approx \frac{1}{n-m} \sum_{t=m+1}^{n} \mathbf{\Theta}_t.$$

- Other quantities can also be computed from Markov Chain:

Markov Chain Monte Carlo

- Given Markov Chain $\{\mathbf{\Theta}_0, \mathbf{\Theta}_1, \ldots, \mathbf{\Theta}_n\}$ with stationary distribution $f(\mathbf{\Theta}|\mathbf{X})$ with burn-in m, we can estimate the posterior mean using Monte Carlo integration:

$$E(\mathbf{\Theta}|\mathbf{X}) \approx \frac{1}{n-m} \sum_{t=m+1}^{n} \mathbf{\Theta}_t.$$

- Other quantities can also be computed from Markov Chain:
  - ⋆ Standard errors
  - ⋆ Quantiles
  - ⋆ Density estimates

Markov Chain Monte Carlo

- Given Markov Chain $\{\boldsymbol{\Theta}_0, \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_n\}$ with stationary distribution $f(\boldsymbol{\Theta}|\mathbf{X})$ with burn-in m, we can estimate the posterior mean using Monte Carlo integration:

$$E(\boldsymbol{\Theta}|\mathbf{X}) \approx \frac{1}{n-m} \sum_{t=m+1}^{n} \boldsymbol{\Theta}_t.$$

- Other quantities can also be computed from Markov Chain:
  - ⋆ Standard errors
  - ⋆ Quantiles
  - ⋆ Density estimates
- Samples can be used to perform any Bayesian inference of interest.
- How do we generate the Markov Chain?

Gibbs Sampler

- *Gibbs Sampler*(Geman and Geman, 1984):

  Markov transition kernel consists of drawing from *full conditional* distributions.

Gibbs Sampler

- *Gibbs Sampler*(Geman and Geman, 1984):

  Markov transition kernel consists of drawing from *full conditional* distributions.

- Suppose $\boldsymbol{\Theta} = \{\theta_1, \theta_2, \ldots, \theta_p\}^T$.

Gibbs Sampler

- *Gibbs Sampler*(Geman and Geman, 1984):

  Markov transition kernel consists of drawing from *full conditional* distributions.

- Suppose $\mathbf{\Theta} = \{\theta_1, \theta_2, \ldots, \theta_p\}^T$.

  Full conditional distribution for parameter i: $f(\theta_i | \mathbf{X}, \mathbf{\Theta}_{-i})$

Gibbs Sampler

- *Gibbs Sampler*(Geman and Geman, 1984):

  Markov transition kernel consists of drawing from *full conditional* distributions.

- Suppose $\mathbf{\Theta} = \{\theta_1, \theta_2, \ldots, \theta_p\}^T$.

  Full conditional distribution for parameter i: $f(\theta_i|\mathbf{X}, \mathbf{\Theta}_{-i})$

  Conditions on:

  ⋆ The data $\mathbf{X}$
  ⋆ The values for all other parameters $\mathbf{\Theta}_{-i}$.

Gibbs Sampler

- Steps of Gibbs sampler:

Gibbs Sampler

- Steps of Gibbs sampler:

1. Choose a set of starting values $\mathbf{\Theta}^{(0)}$.

Gibbs Sampler

- Steps of Gibbs sampler:

1. Choose a set of starting values $\mathbf{\Theta}^{(0)}$.
2. Generate $(\mathbf{\Theta}^{(1)}|\mathbf{\Theta}^{(0)})$ by sampling:

$$\begin{array}{l} \theta_1^{(1)} \text{ from } f(\theta_1^{(1)}|\mathbf{X}, \mathbf{\Theta}_{-1}^{(0)}) \\ \theta_2^{(1)} \text{ from } f(\theta_2^{(1)}|\mathbf{X}, \mathbf{\Theta}_{-2}^{(0)}) \\ \quad\vdots \\ \theta_p^{(1)} \text{ from } f(\theta_p^{(1)}|\mathbf{X}, \mathbf{\Theta}_{-p}^{(0)}) \end{array}$$

Gibbs Sampler

- Steps of Gibbs sampler:

1. Choose a set of starting values $\mathbf{\Theta}^{(0)}$.
2. Generate $(\mathbf{\Theta}^{(1)}|\mathbf{\Theta}^{(0)})$ by sampling:

$\theta_1^{(1)}$ from $f(\theta_1^{(1)}|\mathbf{X}, \mathbf{\Theta}_{-1}^{(0)})$

$\theta_2^{(1)}$ from $f(\theta_2^{(1)}|\mathbf{X}, \mathbf{\Theta}_{-2}^{(0)})$

$\vdots$

$\theta_p^{(1)}$ from $f(\theta_p^{(1)}|\mathbf{X}, \mathbf{\Theta}_{-p}^{(0)})$

3. Repeat step two to get chain of length n: $\{\mathbf{\Theta}^{(0)}, \mathbf{\Theta}^{(1)}, \ldots \mathbf{\Theta}^{(n)}\}$.

Gibbs Sampler

- Steps of Gibbs sampler:

1. Choose a set of starting values $\mathbf{\Theta}^{(0)}$.
2. Generate $(\mathbf{\Theta}^{(1)}|\mathbf{\Theta}^{(0)})$ by sampling:

   $\theta_1^{(1)}$ from $f(\theta_1^{(1)}|\mathbf{X}, \mathbf{\Theta}_{-1}^{(0)})$
   $\theta_2^{(1)}$ from $f(\theta_2^{(1)}|\mathbf{X}, \mathbf{\Theta}_{-2}^{(0)})$
   $\vdots$
   $\theta_p^{(1)}$ from $f(\theta_p^{(1)}|\mathbf{X}, \mathbf{\Theta}_{-p}^{(0)})$
3. Repeat step two to get chain of length n: $\{\mathbf{\Theta}^{(0)}, \mathbf{\Theta}^{(1)}, \ldots \mathbf{\Theta}^{(n)}\}$.
4. Assuming convergence by iteration m, compute posterior mean, quantiles, etc. using samples m through n.

Gibbs Sampler

- Steps of Gibbs sampler:

  1. Choose a set of starting values $\mathbf{\Theta}^{(0)}$.
  2. Generate $(\mathbf{\Theta}^{(1)}|\mathbf{\Theta}^{(0)})$ by sampling:

     $\theta_1^{(1)}$ from $f(\theta_1^{(1)}|\mathbf{X}, \mathbf{\Theta}_{-1}^{(0)})$

     $\theta_2^{(1)}$ from $f(\theta_2^{(1)}|\mathbf{X}, \mathbf{\Theta}_{-2}^{(0)})$

     $\vdots$

     $\theta_p^{(1)}$ from $f(\theta_p^{(1)}|\mathbf{X}, \mathbf{\Theta}_{-p}^{(0)})$
  3. Repeat step two to get chain of length n: $\{\mathbf{\Theta}^{(0)}, \mathbf{\Theta}^{(1)}, \ldots \mathbf{\Theta}^{(n)}\}$.
  4. Assuming convergence by iteration m, compute posterior mean, quantiles, etc. using samples m through n.

- Many variations possible:

Gibbs Sampler

- Steps of Gibbs sampler:

1. Choose a set of starting values $\mathbf{\Theta}^{(0)}$.
2. Generate $(\mathbf{\Theta}^{(1)}|\mathbf{\Theta}^{(0)})$ by sampling:

$$\begin{array}{l} \theta_1^{(1)} \text{ from } f(\theta_1^{(1)}|\mathbf{X}, \mathbf{\Theta}_{-1}^{(0)}) \\ \theta_2^{(1)} \text{ from } f(\theta_2^{(1)}|\mathbf{X}, \mathbf{\Theta}_{-2}^{(0)}) \\ \quad \vdots \\ \theta_p^{(1)} \text{ from } f(\theta_p^{(1)}|\mathbf{X}, \mathbf{\Theta}_{-p}^{(0)}) \end{array}$$

3. Repeat step two to get chain of length n: $\{\mathbf{\Theta}^{(0)}, \mathbf{\Theta}^{(1)}, \ldots \mathbf{\Theta}^{(n)}\}$.
4. Assuming convergence by iteration m, compute posterior mean, quantiles, etc. using samples m through n.

- Many variations possible:

  ⋆ Parameters to update each iteration, order of updating

Gibbs Sampler

- Steps of Gibbs sampler:

1. Choose a set of starting values $\boldsymbol{\Theta}^{(0)}$.
2. Generate $(\boldsymbol{\Theta}^{(1)}|\boldsymbol{\Theta}^{(0)})$ by sampling:

$$\begin{array}{l} \theta_1^{(1)} \text{ from } f(\theta_1^{(1)}|\mathbf{X}, \boldsymbol{\Theta}_{-1}^{(0)}) \\ \theta_2^{(1)} \text{ from } f(\theta_2^{(1)}|\mathbf{X}, \boldsymbol{\Theta}_{-2}^{(0)}) \\ \quad\quad \vdots \\ \theta_p^{(1)} \text{ from } f(\theta_p^{(1)}|\mathbf{X}, \boldsymbol{\Theta}_{-p}^{(0)}) \end{array}$$

3. Repeat step two to get chain of length n: $\{\boldsymbol{\Theta}^{(0)}, \boldsymbol{\Theta}^{(1)}, \ldots \boldsymbol{\Theta}^{(n)}\}$.
4. Assuming convergence by iteration m, compute posterior mean, quantiles, etc. using samples m through n.

- Many variations possible:
  - ⋆ Parameters to update each iteration, order of updating
  - ⋆ 'Blocking' parameters together, working with marginalized distributions

Gibbs Sampler

- Steps of Gibbs sampler:

  1. Choose a set of starting values $\mathbf{\Theta}^{(0)}$.
  2. Generate $(\mathbf{\Theta}^{(1)}|\mathbf{\Theta}^{(0)})$ by sampling:

     $\theta_1^{(1)}$ from $f(\theta_1^{(1)}|\mathbf{X}, \mathbf{\Theta}_{-1}^{(0)})$
     $\theta_2^{(1)}$ from $f(\theta_2^{(1)}|\mathbf{X}, \mathbf{\Theta}_{-2}^{(0)})$
     $\vdots$
     $\theta_p^{(1)}$ from $f(\theta_p^{(1)}|\mathbf{X}, \mathbf{\Theta}_{-p}^{(0)})$
  3. Repeat step two to get chain of length n: $\{\mathbf{\Theta}^{(0)}, \mathbf{\Theta}^{(1)}, \ldots \mathbf{\Theta}^{(n)}\}$.
  4. Assuming convergence by iteration m, compute posterior mean, quantiles, etc. using samples m through n.

- Many variations possible:

  ⋆ Parameters to update each iteration, order of updating
  ⋆ 'Blocking' parameters together, working with marginalized distributions

- If conjugate priors used for all parameters, full conditionals in closed form.

Gibbs Sampler

- Steps of Gibbs sampler:
  1. Choose a set of starting values $\mathbf{\Theta}^{(0)}$.
  2. Generate $(\mathbf{\Theta}^{(1)}|\mathbf{\Theta}^{(0)})$ by sampling:

     $\theta_1^{(1)}$ from $f(\theta_1^{(1)}|\mathbf{X}, \mathbf{\Theta}_{-1}^{(0)})$

     $\theta_2^{(1)}$ from $f(\theta_2^{(1)}|\mathbf{X}, \mathbf{\Theta}_{-2}^{(0)})$

     $\vdots$

     $\theta_p^{(1)}$ from $f(\theta_p^{(1)}|\mathbf{X}, \mathbf{\Theta}_{-p}^{(0)})$
  3. Repeat step two to get chain of length n: $\{\mathbf{\Theta}^{(0)}, \mathbf{\Theta}^{(1)}, \ldots \mathbf{\Theta}^{(n)}\}$.
  4. Assuming convergence by iteration m, compute posterior mean, quantiles, etc. using samples m through n.
- Many variations possible:
  - ⋆ Parameters to update each iteration, order of updating
  - ⋆ 'Blocking' parameters together, working with marginalized distributions
- If conjugate priors used for all parameters, full conditionals in closed form.
- What if we don't have closed form distributions for full conditionals?

Metropolis-Hastings Algorithm

- *Metropolis-Hastings algorithm* (Metropolis et al. 1953, Hastings 1970):

  Method to construct a Markov Chain for θ, even if closed form expression for distribution is not available.

Metropolis-Hastings Algorithm

- *Metropolis-Hastings algorithm* (Metropolis et al. 1953, Hastings 1970):

  Method to construct a Markov Chain for θ, even if closed form expression for distribution is not available.

  $\pi(\theta)$: kernel of distribution of interest for θ, $f(\theta_i^{(t)}|\mathbf{X}, \mathbf{\Theta}_{-i}^{(t-1)})$.

Metropolis-Hastings Algorithm

- *Metropolis-Hastings algorithm* (Metropolis et al. 1953, Hastings 1970):

  Method to construct a Markov Chain for θ, even if closed form expression for distribution is not available.

  $\pi(\theta)$: kernel of distribution of interest for θ, $f(\theta_i^{(t)} | \mathbf{X}, \mathbf{\Theta}_{-i}^{(t-1)})$.
- Steps:

Metropolis-Hastings Algorithm

- *Metropolis-Hastings algorithm* (Metropolis et al. 1953, Hastings 1970):

  Method to construct a Markov Chain for θ, even if closed form expression for distribution is not available.

  $\pi(\theta)$: kernel of distribution of interest for θ, $f(\theta_i^{(t)}|\mathbf{X}, \mathbf{\Theta}_{-i}^{(t-1)})$.
- Steps:

1. Get $\theta^{(0)}$= starting value for θ.

Metropolis-Hastings Algorithm

- *Metropolis-Hastings algorithm* (Metropolis et al. 1953, Hastings 1970):

  Method to construct a Markov Chain for θ, even if closed form expression for distribution is not available.

  $\pi(\theta)$: kernel of distribution of interest for θ, $f(\theta_i^{(t)}|\mathbf{X}, \mathbf{\Theta}_{-i}^{(t-1)})$.
- Steps:

1. Get $\theta^{(0)}$= starting value for θ.
2. Get θ^*=proposed value for $\theta^{(1)}$, by sampling from *proposal density* $q(\theta|\mathbf{X}, \theta^{(0)})$.

Metropolis-Hastings Algorithm

- *Metropolis-Hastings algorithm* (Metropolis et al. 1953, Hastings 1970):

  Method to construct a Markov Chain for θ, even if closed form expression for distribution is not available.

  $\pi(\theta)$: kernel of distribution of interest for θ, $f(\theta_i^{(t)}|\mathbf{X}, \mathbf{\Theta}_{-i}^{(t-1)})$.
- Steps:

1. Get $\theta^{(0)}$= starting value for θ.
2. Get θ^{*}=proposed value for $\theta^{(1)}$, by sampling from *proposal density* $q(\theta|\mathbf{X}, \theta^{(0)})$.
3. Compute $\alpha(\theta^{(0)}, \theta^{*})$=min$\left(1, \frac{\pi(\theta^{*})q(\theta^{(0)}|\theta^{*})}{\pi(\theta^{(0)}q(\theta^{*}|\theta^{(0)}))}\right)$.

Metropolis-Hastings Algorithm

- *Metropolis-Hastings algorithm* (Metropolis et al. 1953, Hastings 1970):

  Method to construct a Markov Chain for θ, even if closed form expression for distribution is not available.

  $\pi(\theta)$: kernel of distribution of interest for θ, $f(\theta_i^{(t)}|\mathbf{X}, \mathbf{\Theta}_{-i}^{(t-1)})$.

- Steps:

  1. Get $\theta^{(0)}$= starting value for θ.
  2. Get θ^*=proposed value for $\theta^{(1)}$, by sampling from *proposal density* $q(\theta|\mathbf{X}, \theta^{(0)})$.
  3. Compute $\alpha(\theta^{(0)}, \theta^*)$=min$\left(1, \frac{\pi(\theta^*)q(\theta^{(0)}|\theta^*)}{\pi(\theta^{(0)}q(\theta^*|\theta^{(0)})}\right)$.
  4. Generate u $\sim$Uniform(0,1).

Metropolis-Hastings Algorithm

- *Metropolis-Hastings algorithm* (Metropolis et al. 1953, Hastings 1970):

  Method to construct a Markov Chain for θ, even if closed form expression for distribution is not available.

  $\pi(\theta)$: kernel of distribution of interest for θ, $f(\theta_i^{(t)}|\mathbf{X}, \mathbf{\Theta}_{-i}^{(t-1)})$.
- Steps:

1. Get $\theta^{(0)}$= starting value for θ.
2. Get θ^*=proposed value for $\theta^{(1)}$, by sampling from *proposal density* $q(\theta|\mathbf{X}, \theta^{(0)})$.
3. Compute $\alpha(\theta^{(0)}, \theta^*)$=min$\left(1, \dfrac{\pi(\theta^*)q(\theta^{(0)}|\theta^*)}{\pi(\theta^{(0)}q(\theta^*|\theta^{(0)})}\right)$.
4. Generate u ∼Uniform(0,1).
   If $u < \alpha \Rightarrow$ let $\theta^{(1)} = \theta^*$, else let $\theta^{(1)} = \theta^{(0)}$.

Metropolis-Hastings Algorithm

- *Metropolis-Hastings algorithm* (Metropolis et al. 1953, Hastings 1970):

  Method to construct a Markov Chain for θ, even if closed form expression for distribution is not available.

  $\pi(\theta)$: kernel of distribution of interest for θ, $f(\theta_i^{(t)}|\mathbf{X}, \mathbf{\Theta}_{-i}^{(t-1)})$.

- Steps:

  1. Get $\theta^{(0)}$= starting value for θ.
  2. Get θ^*=proposed value for $\theta^{(1)}$, by sampling from *proposal density* $q(\theta|\mathbf{X}, \theta^{(0)})$.
  3. Compute $\alpha(\theta^{(0)}, \theta^*)$=min$\left(1, \dfrac{\pi(\theta^*)q(\theta^{(0)}|\theta^*)}{\pi(\theta^{(0)}q(\theta^*|\theta^{(0)})}\right)$.
  4. Generate u ~Uniform(0,1).
     If $u < \alpha \Rightarrow$ let $\theta^{(1)} = \theta^*$, else let $\theta^{(1)} = \theta^{(0)}$.

- Types of proposals: Random Walk, Independence, Symmetric

Assessing Convergence of Markov Chains

- The Markov Chain is known to converge to the stationary distribution of interest, but how do I know when convergence has been achieved?

Assessing Convergence of Markov Chains

- The Markov Chain is known to converge to the stationary distribution of interest, but how do I know when convergence has been achieved?

  i.e. How do I decide how long the burn-in should be?

Assessing Convergence of Markov Chains

- The Markov Chain is known to converge to the stationary distribution of interest, but how do I know when convergence has been achieved?

  i.e. How do I decide how long the burn-in should be?

1. Look at time series plots for the parameters.

Assessing Convergence of Markov Chains

- The Markov Chain is known to converge to the stationary distribution of interest, but how do I know when convergence has been achieved?

  i.e. How do I decide how long the burn-in should be?

1. Look at time series plots for the parameters.
2. Run multiple chains with divergent starting values.

Assessing Convergence of Markov Chains

- The Markov Chain is known to converge to the stationary distribution of interest, but how do I know when convergence has been achieved?

  i.e. How do I decide how long the burn-in should be?

1. Look at time series plots for the parameters.
2. Run multiple chains with divergent starting values.
3. Run formal diagnostics (Gelman and Rubin 1992, Geweke 1992)

Assessing Convergence of Markov Chains

- The Markov Chain is known to converge to the stationary distribution of interest, but how do I know when convergence has been achieved?

  i.e. How do I decide how long the burn-in should be?

  1. Look at time series plots for the parameters.
  2. Run multiple chains with divergent starting values.
  3. Run formal diagnostics (Gelman and Rubin 1992, Geweke 1992)

- Other issues:

  ⋆ Length of chain

Assessing Convergence of Markov Chains

- The Markov Chain is known to converge to the stationary distribution of interest, but how do I know when convergence has been achieved?

  i.e. How do I decide how long the burn-in should be?

  1. Look at time series plots for the parameters.
  2. Run multiple chains with divergent starting values.
  3. Run formal diagnostics (Gelman and Rubin 1992, Geweke 1992)

- Other issues:

  ⋆ Length of chain
  ⋆ Thinning to decrease autocorrelation

Assessing Convergence of Markov Chains

- The Markov Chain is known to converge to the stationary distribution of interest, but how do I know when convergence has been achieved?

  i.e. How do I decide how long the burn-in should be?

1. Look at time series plots for the parameters.
2. Run multiple chains with divergent starting values.
3. Run formal diagnostics (Gelman and Rubin 1992, Geweke 1992)

- Other issues:
  - ⋆ Length of chain
  - ⋆ Thinning to decrease autocorrelation

Example: Hierarchical Models

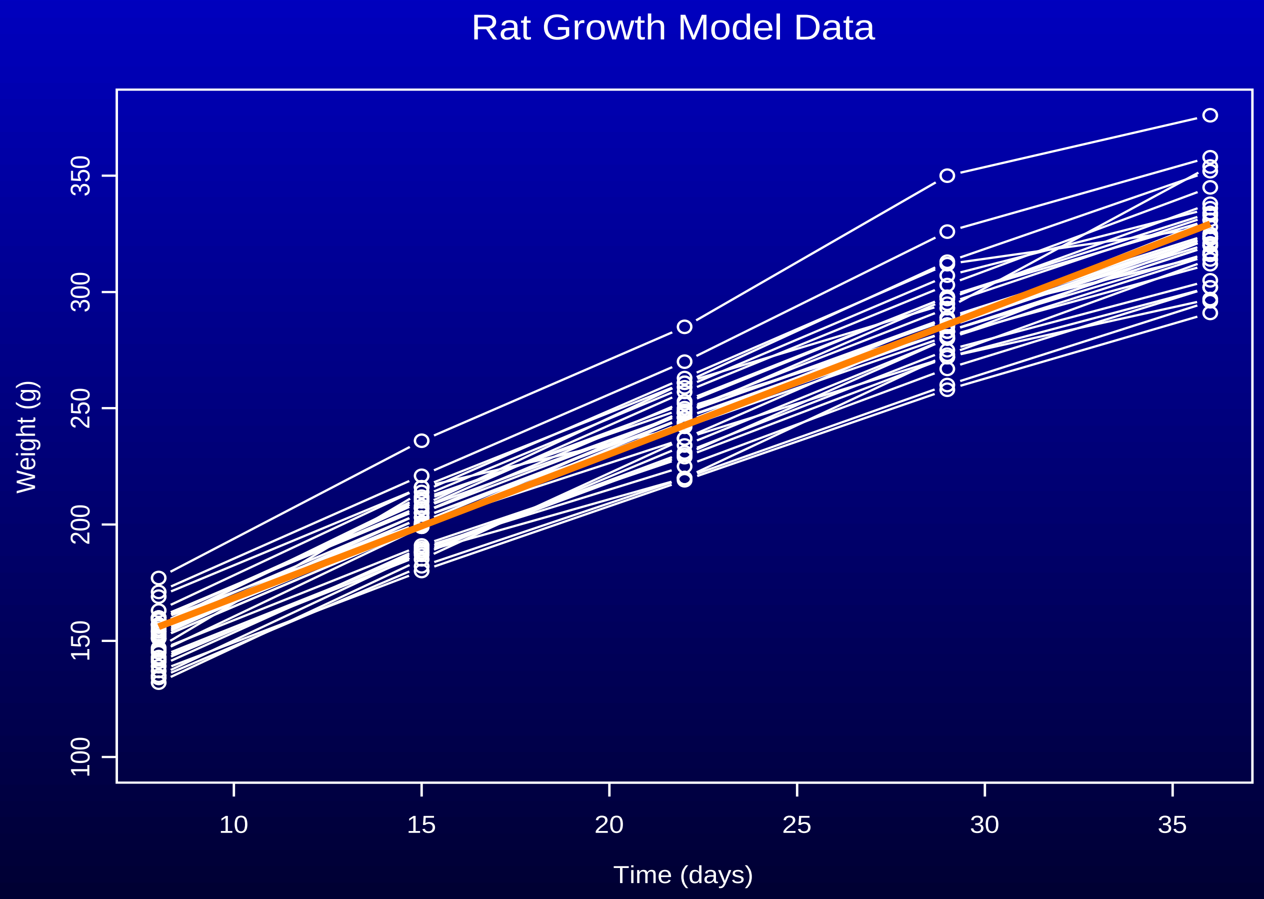
- Example: Growth curves for rats.

Example: Hierarchical Models

- Example: Growth curves for rats.
- Data Y_{ij} consists of weights for 30 rats over 5 weeks.

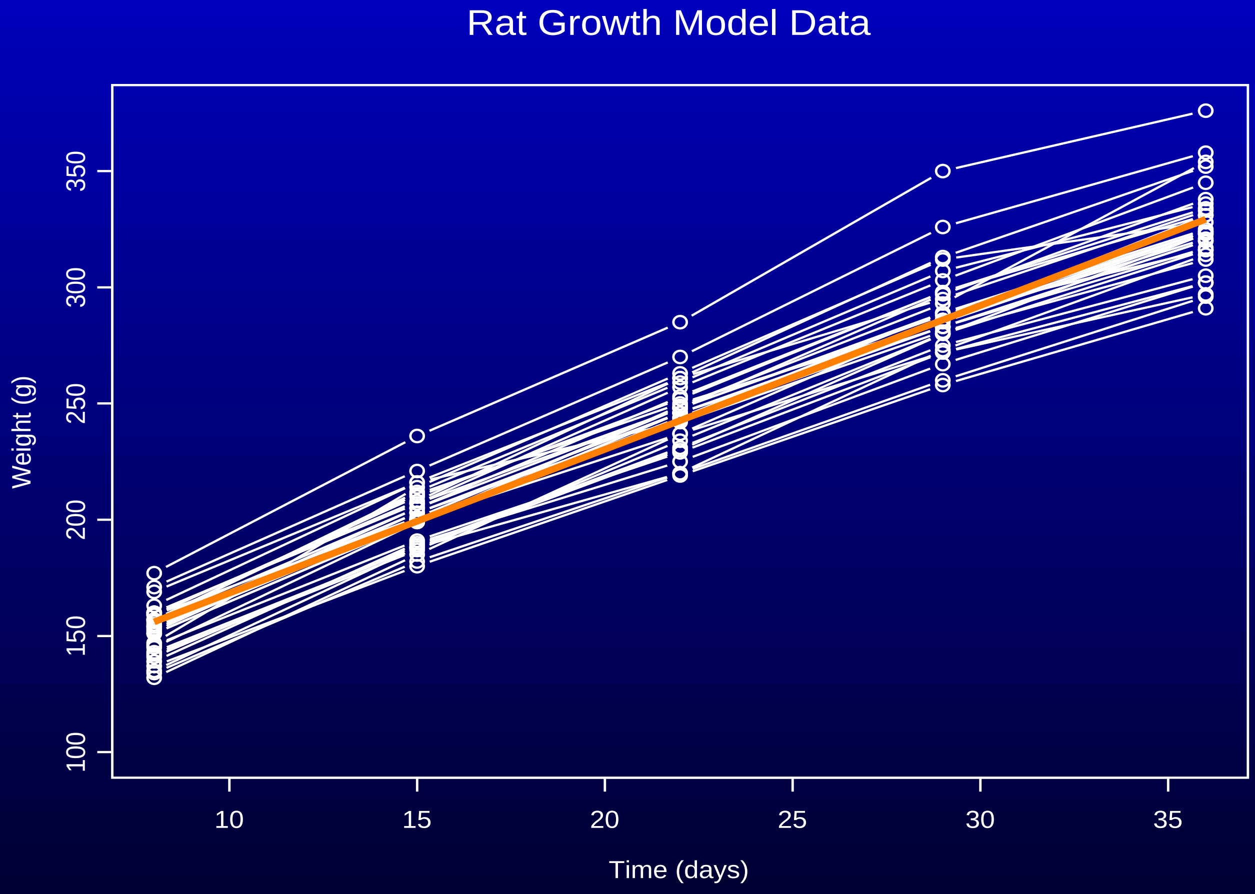
Example: Hierarchical Models

- Example: Growth curves for rats.
- Data Y_{ij} consists of weights for 30 rats over 5 weeks.



Example: Hierarchical Models

- Example: Growth curves for rats.
- Data Y_{ij} consists of weights for 30 rats over 5 weeks.



- Can estimate mean growth curve by linear regression, but growth curve models necessary to get standard errors right.

Example: Hierarchical Models

- Model: $$Y_{ij} \sim \text{Normal}(\mu_{ij}, \tau_c)$$

Example: Hierarchical Models

- Model:

$$\begin{aligned} Y_{ij} &\sim \text{Normal}(\mu_{ij}, \tau_c) \\ \mu_{ij} &= \alpha_i + \beta_i(x_j - \overline{x}) \end{aligned}$$

Example: Hierarchical Models

- Model:

$$\begin{aligned} Y_{ij} &\sim \text{Normal}(\mu_{ij}, \tau_c) \\ \mu_{ij} &= \alpha_i + \beta_i(x_j - \overline{x}) \\ \alpha_i &\sim \text{Normal}(\alpha_c, \tau_\alpha) \end{aligned}$$

Example: Hierarchical Models

- Model:

$$
\begin{aligned}
Y_{ij} &\sim \text{Normal}(\mu_{ij}, \tau_c) \\
\mu_{ij} &= \alpha_i + \beta_i(x_j - \overline{x}) \\
\alpha_i &\sim \text{Normal}(\alpha_c, \tau_\alpha) \\
\beta_i &\sim \text{Normal}(\beta_c, \tau_\beta)
\end{aligned}
$$

Example: Hierarchical Models

- Model:

$$\begin{aligned} Y_{ij} &\sim \text{Normal}(\mu_{ij}, \tau_c) \\ \mu_{ij} &= \alpha_i + \beta_i(x_j - \overline{x}) \\ \alpha_i &\sim \text{Normal}(\alpha_c, \tau_\alpha) \\ \beta_i &\sim \text{Normal}(\beta_c, \tau_\beta) \end{aligned}$$

- Model could be fit using linear mixed model or Bayesian hierarchical model.

Example: Hierarchical Models

- Model:

$$
\begin{aligned}
Y_{ij} &\sim \text{Normal}(\mu_{ij}, \tau_c) \\
\mu_{ij} &= \alpha_i + \beta_i(x_j - \overline{x}) \\
\alpha_i &\sim \text{Normal}(\alpha_c, \tau_\alpha) \\
\beta_i &\sim \text{Normal}(\beta_c, \tau_\beta)
\end{aligned}
$$

- Model could be fit using linear mixed model or Bayesian hierarchical model.
- Priors (conjugate and vague):

$$\alpha_c, \beta_c \sim \text{Normal}(0, 10^{-6})$$

Example: Hierarchical Models

- Model:

$$\begin{aligned} Y_{ij} &\sim \text{Normal}(\mu_{ij}, \tau_c) \\ \mu_{ij} &= \alpha_i + \beta_i(x_j - \overline{x}) \\ \alpha_i &\sim \text{Normal}(\alpha_c, \tau_\alpha) \\ \beta_i &\sim \text{Normal}(\beta_c, \tau_\beta) \end{aligned}$$

- Model could be fit using linear mixed model or Bayesian hierarchical model.
- Priors (conjugate and vague):

$$\begin{aligned} \alpha_c, \beta_c &\sim \text{Normal}(0, 10^{-6}) \\ \tau_c, \tau_\alpha, \tau_\beta &\sim \text{Gamma}(0.001, 0.001) \end{aligned}$$

Example: Hierarchical Models

- Model:

$$
\begin{aligned}
Y_{ij} &\sim \text{Normal}(\mu_{ij}, \tau_c) \\
\mu_{ij} &= \alpha_i + \beta_i(x_j - \overline{x}) \\
\alpha_i &\sim \text{Normal}(\alpha_c, \tau_\alpha) \\
\beta_i &\sim \text{Normal}(\beta_c, \tau_\beta)
\end{aligned}
$$

- Model could be fit using linear mixed model or Bayesian hierarchical model.
- Priors (conjugate and vague):

$$
\begin{aligned}
\alpha_c, \beta_c &\sim \text{Normal}(0, 10^{-6}) \\
\tau_c, \tau_\alpha, \tau_\beta &\sim \text{Gamma}(0.001, 0.001)
\end{aligned}
$$

- Gibbs sampler:

  Since conjugate priors were used, the full conditionals are all available in closed form and can be derived using some algebra.

Example: Hierarchical Models

- Model:

$$\begin{aligned} Y_{ij} &\sim \text{Normal}(\mu_{ij}, \tau_c) \\ \mu_{ij} &= \alpha_i + \beta_i(x_j - \overline{x}) \\ \alpha_i &\sim \text{Normal}(\alpha_c, \tau_\alpha) \\ \beta_i &\sim \text{Normal}(\beta_c, \tau_\beta) \end{aligned}$$

- Model could be fit using linear mixed model or Bayesian hierarchical model.
- Priors (conjugate and vague):

$$\begin{aligned} \alpha_c, \beta_c &\sim \text{Normal}(0, 10^{-6}) \\ \tau_c, \tau_\alpha, \tau_\beta &\sim \text{Gamma}(0.001, 0.001) \end{aligned}$$

- Gibbs sampler:

  Since conjugate priors were used, the full conditionals are all available in closed form and can be derived using some algebra.

- WinBUGS: Statistical software to perform MCMC in general problems.

Example: Hierarchical Models

- Model:

$$\begin{aligned} Y_{ij} &\sim \text{Normal}(\mu_{ij}, \tau_c) \\ \mu_{ij} &= \alpha_i + \beta_i(x_j - \overline{x}) \\ \alpha_i &\sim \text{Normal}(\alpha_c, \tau_\alpha) \\ \beta_i &\sim \text{Normal}(\beta_c, \tau_\beta) \end{aligned}$$

- Model could be fit using linear mixed model or Bayesian hierarchical model.
- Priors (conjugate and vague):

$$\begin{aligned} \alpha_c, \beta_c &\sim \text{Normal}(0, 10^{-6}) \\ \tau_c, \tau_\alpha, \tau_\beta &\sim \text{Gamma}(0.001, 0.001) \end{aligned}$$

- Gibbs sampler:

  Since conjugate priors were used, the full conditionals are all available in closed form and can be derived using some algebra.

- WinBUGS: Statistical software to perform MCMC in general problems.

Conclusions

- Why use MCMC?
  - ⋆ Flexible computing tool with ability to fit complex models.

Conclusions

- Why use MCMC?
  - ⋆ Flexible computing tool with ability to fit complex models.
  - ⋆ No need to make simplified modeling assumptions out of convenience.

Conclusions

- Why use MCMC?
  - ⋆ Flexible computing tool with ability to fit complex models.
  - ⋆ No need to make simplified modeling assumptions out of convenience.
  - ⋆ Given posterior samples, can get all benefits of Bayesian inference.

Conclusions

- Why use MCMC?
  - ⋆ Flexible computing tool with ability to fit complex models.
  - ⋆ No need to make simplified modeling assumptions out of convenience.
  - ⋆ Given posterior samples, can get all benefits of Bayesian inference.
- Words of Caution:

Conclusions

- Why use MCMC?
  - ⋆ Flexible computing tool with ability to fit complex models.
  - ⋆ No need to make simplified modeling assumptions out of convenience.
  - ⋆ Given posterior samples, can get all benefits of Bayesian inference.
- Words of Caution:
  - ⋆ Monitor convergence!

Conclusions

- Why use MCMC?
  - ⋆ Flexible computing tool with ability to fit complex models.
  - ⋆ No need to make simplified modeling assumptions out of convenience.
  - ⋆ Given posterior samples, can get all benefits of Bayesian inference.

- Words of Caution:
  - ⋆ Monitor convergence!
    - ∗ Unfortunately, the most complex models tend to converge very slowly.

Conclusions

- Why use MCMC?
  - Flexible computing tool with ability to fit complex models.
  - No need to make simplified modeling assumptions out of convenience.
  - Given posterior samples, can get all benefits of Bayesian inference.

- Words of Caution:
  - Monitor convergence!
    - Unfortunately, the most complex models tend to converge very slowly.
    - Can try blocking and marginalization to decrease correlation of model parameters in MCMC.

Conclusions

- Why use MCMC?
  - ⋆ Flexible computing tool with ability to fit complex models.
  - ⋆ No need to make simplified modeling assumptions out of convenience.
  - ⋆ Given posterior samples, can get all benefits of Bayesian inference.

- Words of Caution:
  - ⋆ Monitor convergence!
    - ∗ Unfortunately, the most complex models tend to converge very slowly.
    - ∗ Can try blocking and marginalization to decrease correlation of model parameters in MCMC.
  - ⋆ Check if your answers make sense - compare with plots and simple methods

Conclusions

- Why use MCMC?
  - ⋆ Flexible computing tool with ability to fit complex models.
  - ⋆ No need to make simplified modeling assumptions out of convenience.
  - ⋆ Given posterior samples, can get all benefits of Bayesian inference.

- Words of Caution:
  - ⋆ Monitor convergence!
    - ∗ Unfortunately, the most complex models tend to converge very slowly.
    - ∗ Can try blocking and marginalization to decrease correlation of model parameters in MCMC.
  - ⋆ Check if your answers make sense - compare with plots and simple methods
  - ⋆ Perform sensitivity analysis on priors.

Conclusions

- Why use MCMC?
  - ⋆ Flexible computing tool with ability to fit complex models.
  - ⋆ No need to make simplified modeling assumptions out of convenience.
  - ⋆ Given posterior samples, can get all benefits of Bayesian inference.
- Words of Caution:
  - ⋆ Monitor convergence!
    - ∗ Unfortunately, the most complex models tend to converge very slowly.
    - ∗ Can try blocking and marginalization to decrease correlation of model parameters in MCMC.
  - ⋆ Check if your answers make sense - compare with plots and simple methods
  - ⋆ Perform sensitivity analysis on priors.
- Other book: Gelman, Carlin, Stern, & Rubin (1995) *Bayesian Data Analysis*

References

Gelman A and Rubin DB (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457 75472.

Geman S and Geman D (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattn. Anal. Mach. Intel.* **6**, 721 75741.

Geweke J (1992). Evaluation of accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics 4*(ed. JM Bernardo, J Berger, AP Dawid and AFM Smith), pp. 169 75193. Oxford University Press.

Gilks WR, Richardson S, and Spiegelhalter DJ (1996). *Markov Chain Monte Carlo in Practice,* Chapman and Hall.

Hastings WK (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97 75109.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH and Teller E (1953). Equations of state calculations by fast computing machine. *J. Chem. Phys.* **21**, 1087 751091.