An Introduction to Bayesian Statistics

David Carlson

July 9, 2021

The Scientific Method

- 1) Ask a question or pose a problem

The Scientific Method

- 1) Ask a question or pose a problem
- 2) Collect the existing relevant information

The Scientific Method

- 1) Ask a question or pose a problem
- 2) Collect the existing relevant information
- 3) Based on information, design an investigation or experiment to address the question

The Scientific Method

- 1) Ask a question or pose a problem
- 2) Collect the existing relevant information
- 3) Based on information, design an investigation or experiment to address the question
- 4) Carry out the investigation or experiment

The Scientific Method

- 1) Ask a question or pose a problem
- 2) Collect the existing relevant information
- 3) Based on information, design an investigation or experiment to address the question
- 4) Carry out the investigation or experiment
- 5) Use the evidence to update previously known information, then draw conclusions

The Scientific Method

- 1) Ask a question or pose a problem
- 2) Collect the existing relevant information
- 3) Based on information, design an investigation or experiment to address the question
- 4) Carry out the investigation or experiment
- 5) Use the evidence to update previously known information, then draw conclusions
- 6) Repeat 3—5 as necessary

Where Does Statistics Fit In?

- Central to steps 2, 3, and 5

Where Does Statistics Fit In?

- Central to steps 2, 3, and 5
  - Experimental design, survey sampling

Where Does Statistics Fit In?

- Central to steps 2, 3, and 5
  - Experimental design, survey sampling
  - Statistical inference

Where Does Statistics Fit In?

- Central to steps 2, 3, and 5
  - Experimental design, survey sampling
  - Statistical inference
- Bayesian statistics is particularly well-suited for step 2 and 5

Where Does Statistics Fit In?

- Central to steps 2, 3, and 5
  - Experimental design, survey sampling
  - Statistical inference
- Bayesian statistics is particularly well-suited for step 2 and 5
  - Step 2: Prior information

Where Does Statistics Fit In?

- Central to steps 2, 3, and 5
  - Experimental design, survey sampling
  - Statistical inference
- Bayesian statistics is particularly well-suited for step 2 and 5
  - Step 2: Prior information
  - Step 5: Prior information $\rightarrow$ posterior information

Economic Applications of Bayesian Statistics

Figure 5: Econometrica Containing "Bayes" or "Bayesian"

Common Applications of Bayesian Statistics

- Economics

Common Applications of Bayesian Statistics

- Economics
- Marketing

Common Applications of Bayesian Statistics

- Economics
- Marketing
- Education

Common Applications of Bayesian Statistics

- Economics
- Marketing
- Education
- Medical research

Common Applications of Bayesian Statistics

- Economics
- Marketing
- Education
- Medical research
- Genomics

Common Applications of Bayesian Statistics

- Economics
- Marketing
- Education
- Medical research
- Genomics
- Weather

Common Applications of Bayesian Statistics

- Economics
- Marketing
- Education
- Medical research
- Genomics
- Weather
- The list goes on...

Theoretical Differences to Frequentist Approaches

- Probability as the subjective experience of uncertainty

Theoretical Differences to Frequentist Approaches

- Probability as the subjective experience of uncertainty
- No notion of infinitely repeating an event of interest

Theoretical Differences to Frequentist Approaches

- Probability as the subjective experience of uncertainty
- No notion of infinitely repeating an event of interest
- Using as much prior information as possible as well as personal judgment (placing a bet)

Theoretical Differences to Frequentist Approaches

- Probability as the subjective experience of uncertainty
- No notion of infinitely repeating an event of interest
- Using as much prior information as possible as well as personal judgment (placing a bet)
- Once an outcome is revealed, prior information is updated

Table of Frequentist vs. Bayesian Interpretations

	Frequentist statistics	Bayesian statistics
Definition of the p value	The probability of observing the same or more extreme data assuming that the null hypothesis is true in the population	The probability of the (null) hypothesis
Large samples needed?	Usually, when normal theory-based methods are used	Not necessarily
Inclusion of prior knowledge possible?	No	Yes
Nature of the parameters in the model	Unknown but fixed	Unknown and therefore random
Population parameter	One true value	A distribution of values reflecting uncertainty
Uncertainty is defined by	The sampling distribution based on the idea of infinite repeated sampling	Probability distribution for the population parameter
Estimated intervals	Confidence interval: Over an infinity of samples taken from the population, 95% of these contain the true population value	Credibility interval: A 95% probability that the population value is within the limits of the interval

Bayesian vs. Frequentist

- Frequentist: Parameter is unknown, but fixed

Bayesian vs. Frequentist

- Frequentist: Parameter is unknown, but fixed
- Bayesian: Unknown parameters are treated as uncertain and described by probability distribution

Bayesian vs. Frequentist

- Frequentist: Parameter is unknown, but fixed
- Bayesian: Unknown parameters are treated as uncertain and described by probability distribution
- Frequentist: 95 of 100 replications of same experiment capture the fixed but unknown parameter

Bayesian vs. Frequentist

- Frequentist: Parameter is unknown, but fixed
- Bayesian: Unknown parameters are treated as uncertain and described by probability distribution
- Frequentist: 95 of 100 replications of same experiment capture the fixed but unknown parameter
- Bayesian: Probability that a parameter lies in the credible interval

Priors

- Understanding of parameters of interest given our current data depends on our prior knowledge about the parameters of interest weighted by the current evidence given those parameters of interest

Priors

- Understanding of parameters of interest given our current data depends on our prior knowledge about the parameters of interest weighted by the current evidence given those parameters of interest
- Priors: Equally important to quantify ignorance as to quantify cumulative understanding of a problem

Priors

- Understanding of parameters of interest given our current data depends on our prior knowledge about the parameters of interest weighted by the current evidence given those parameters of interest
- Priors: Equally important to quantify ignorance as to quantify cumulative understanding of a problem
- Objective prior: Pure ignorance

Priors

- Understanding of parameters of interest given our current data depends on our prior knowledge about the parameters of interest weighted by the current evidence given those parameters of interest
- Priors: Equally important to quantify ignorance as to quantify cumulative understanding of a problem
- Objective prior: Pure ignorance
- Prior reflects knowledge about parameters before observing current data

Priors

- Understanding of parameters of interest given our current data depends on our prior knowledge about the parameters of interest weighted by the current evidence given those parameters of interest
- Priors: Equally important to quantify ignorance as to quantify cumulative understanding of a problem
- Objective prior: Pure ignorance
- Prior reflects knowledge about parameters before observing current data
- Science can be accumulative!

Bayes' Theorem

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A) \times \mathrm{P}(A)}{\mathrm{P}(B)}$$

An Example of Bayes' Theorem

- Enzyme-Linked Immuno Sobert Assay (ELISA)

An Example of Bayes' Theorem

- Enzyme-Linked Immuno Sobert Assay (ELISA)
  - Diagnosis tool for HIV

An Example of Bayes' Theorem

- Enzyme-Linked Immuno Sobert Assay (ELISA)
  - Diagnosis tool for HIV
  - Low cost

An Example of Bayes' Theorem

- Enzyme-Linked Immuno Sobert Assay (ELISA)
  - Diagnosis tool for HIV
  - Low cost
  - Less accurate compared to more expensive tools

An Example of Bayes' Theorem

- Enzyme-Linked Immuno Sobert Assay (ELISA)
  - Diagnosis tool for HIV
  - Low cost
  - Less accurate compared to more expensive tools
- Let $A = \{\text{the patient is positive}\}$. Is it proper to use $\Pr(A)$?

An Example of Bayes' Theorem

- Enzyme-Linked Immuno Sobert Assay (ELISA)
  - Diagnosis tool for HIV
  - Low cost
  - Less accurate compared to more expensive tools
- Let $A = \{\text{the patient is positive}\}$. Is it proper to use $\Pr(A)$?
- How should we update the uncertainty after a test?

An Example of Bayes' Theorem (cont.)

- Data: Test result (positive/negative)

An Example of Bayes' Theorem (cont.)

- Data: Test result (positive/negative)
- Update the uncertainty using Bayes' theorem

$$\mathrm{P}(A|+) = \frac{\mathrm{P}(+|A)\mathrm{P}(A)}{\mathrm{P}(+|A)\mathrm{P}(A) + \mathrm{P}(+|\neg A)\mathrm{P}(\neg A)}$$

An Example of Bayes' Theorem (cont.)

- Data: Test result (positive/negative)
- Update the uncertainty using Bayes' theorem

$$\mathrm{P}(A|+) = \frac{\mathrm{P}(+|A)\mathrm{P}(A)}{\mathrm{P}(+|A)\mathrm{P}(A) + \mathrm{P}(+|\neg A)\mathrm{P}(\neg A)}$$

- Sensitivity: $\mathrm{P}(+|A) = 0.95$

An Example of Bayes' Theorem (cont.)

- Data: Test result (positive/negative)
- Update the uncertainty using Bayes' theorem

$$\mathrm{P}(A|+) = \frac{\mathrm{P}(+|A)\mathrm{P}(A)}{\mathrm{P}(+|A)\mathrm{P}(A) + \mathrm{P}(+|\neg A)\mathrm{P}(\neg A)}$$

- Sensitivity: $\mathrm{P}(+|A) = 0.95$
- Specificity: $\mathrm{P}(-|\neg A) = 1 - \mathrm{P}(+|\neg A) = 0.98$

An Example of Bayes' Theorem (cont.)

- Data: Test result (positive/negative)
- Update the uncertainty using Bayes' theorem

$$\mathrm{P}(A|+) = \frac{\mathrm{P}(+|A)\mathrm{P}(A)}{\mathrm{P}(+|A)\mathrm{P}(A) + \mathrm{P}(+|\neg A)\mathrm{P}(\neg A)}$$

- Sensitivity: $\mathrm{P}(+|A) = 0.95$
- Specificity: $\mathrm{P}(-|\neg A) = 1 - \mathrm{P}(+|\neg A) = 0.98$
- Population prevalance: $\mathrm{P}(A) = 0.01$

An Example of Bayes' Theorem (cont.)

- Data: Test result (positive/negative)
- Update the uncertainty using Bayes' theorem

$$\mathrm{P}(A|+) = \frac{\mathrm{P}(+|A)\mathrm{P}(A)}{\mathrm{P}(+|A)\mathrm{P}(A) + \mathrm{P}(+|\neg A)\mathrm{P}(\neg A)}$$

- Sensitivity: $\mathrm{P}(+|A) = 0.95$
- Specificity: $\mathrm{P}(-|\neg A) = 1 - \mathrm{P}(+|\neg A) = 0.98$
- Population prevalance: $\mathrm{P}(A) = 0.01$
- $\mathrm{P}(A|+) = 0.32$

An Example of Bayes' Theorem (cont.)

- Data: Test result (positive/negative)
- Update the uncertainty using Bayes' theorem

$$\mathrm{P}(A|+) = \frac{\mathrm{P}(+|A)\mathrm{P}(A)}{\mathrm{P}(+|A)\mathrm{P}(A) + \mathrm{P}(+|\neg A)\mathrm{P}(\neg A)}$$

- Sensitivity: $\mathrm{P}(+|A) = 0.95$
- Specificity: $\mathrm{P}(-|\neg A) = 1 - \mathrm{P}(+|\neg A) = 0.98$
- Population prevalance: $\mathrm{P}(A) = 0.01$
- $\mathrm{P}(A|+) = 0.32$
- False positive rate: $\mathrm{P}(\neg A|+) = 1 - \mathrm{P}(A|+) = 0.68$

An Example of Bayes' Theorem (cont.)

- Data: Test result (positive/negative)
- Update the uncertainty using Bayes' theorem

$$\mathrm{P}(A|+) = \frac{\mathrm{P}(+|A)\mathrm{P}(A)}{\mathrm{P}(+|A)\mathrm{P}(A) + \mathrm{P}(+|\neg A)\mathrm{P}(\neg A)}$$

- Sensitivity: $\mathrm{P}(+|A) = 0.95$
- Specificity: $\mathrm{P}(-|\neg A) = 1 - \mathrm{P}(+|\neg A) = 0.98$
- Population prevalance: $\mathrm{P}(A) = 0.01$
- $\mathrm{P}(A|+) = 0.32$
- False positive rate: $\mathrm{P}(\neg A|+) = 1 - \mathrm{P}(A|+) = 0.68$
- False negative rate: $\mathrm{P}(A|-) = 0.00052$

General Form of Bayes' Theorem

- Model: $y|\theta \sim f(y|\theta)$

General Form of Bayes' Theorem

- Model: $y|\theta \sim f(y|\theta)$
  - y: data

General Form of Bayes' Theorem

- Model: $y|\theta \sim f(y|\theta)$
  - y: data
  - θ: parameter

General Form of Bayes' Theorem

- Model: $y|\theta \sim f(y|\theta)$
  - y: data
  - θ: parameter
- Bayes' theorem

$$
\begin{aligned}
f(\theta|y) &= \frac{f(\theta, y)}{f(y)} \\
&= \frac{f(y|\theta)f(\theta)}{f(y)} \\
&\propto f(y|\theta)f(\theta)
\end{aligned}
$$

General Form of Bayes' Theorem

- Model: $y|\theta \sim f(y|\theta)$
  - y: data
  - θ: parameter
- Bayes' theorem

$$
\begin{aligned}
f(\theta|y) &= \frac{f(\theta, y)}{f(y)} \\
&= \frac{f(y|\theta)f(\theta)}{f(y)} \\
&\propto f(y|\theta)f(\theta)
\end{aligned}
$$

- Posterior $\propto$ likelihood $\times$ prior

Key Issues in Bayesian Statistics

- How to specify prior distributions?

Key Issues in Bayesian Statistics

- How to specify prior distributions?
- How to compute the posterior $f(\theta|y)$?

Key Issues in Bayesian Statistics

- How to specify prior distributions?
- How to compute the posterior $f(\theta|y)$?
  - In general we need to know the marginal distribution $f(y) = \int f(y|\theta)f(\theta)d\theta$

Key Issues in Bayesian Statistics

- How to specify prior distributions?
- How to compute the posterior $f(\theta|y)$?
  - In general we need to know the marginal distribution $f(y) = \int f(y|\theta)f(\theta)d\theta$
  - Direct computation of this integral is difficult especially when θ is high-dimensional

Key Issues in Bayesian Statistics

- How to specify prior distributions?
- How to compute the posterior $f(\theta|y)$?
  - In general we need to know the marginal distribution $f(y) = \int f(y|\theta)f(\theta)d\theta$
  - Direct computation of this integral is difficult especially when θ is high-dimensional
  - Simulation is the general solution

Key Issues in Bayesian Statistics

- How to specify prior distributions?
- How to compute the posterior $f(\theta|y)$?
  - In general we need to know the marginal distribution $f(y) = \int f(y|\theta)f(\theta)d\theta$
  - Direct computation of this integral is difficult especially when θ is high-dimensional
  - Simulation is the general solution
- How to summarize the posterior?

Key Issues in Bayesian Statistics

- How to specify prior distributions?
- How to compute the posterior $f(\theta|y)$?
  - In general we need to know the marginal distribution $f(y) = \int f(y|\theta)f(\theta)d\theta$
  - Direct computation of this integral is difficult especially when θ is high-dimensional
  - Simulation is the general solution
- How to summarize the posterior?
  - Posterior mean, median, or mode

Key Issues in Bayesian Statistics

- How to specify prior distributions?
- How to compute the posterior $f(\theta|y)$?
  - In general we need to know the marginal distribution $f(y) = \int f(y|\theta)f(\theta)d\theta$
  - Direct computation of this integral is difficult especially when θ is high-dimensional
  - Simulation is the general solution
- How to summarize the posterior?
  - Posterior mean, median, or mode
  - Credible interval

Comparison of Bayesian and Frequentist Methods

- Interval estimation

Comparison of Bayesian and Frequentist Methods

- Interval estimation
  - Frequentist: Confidence interval

Comparison of Bayesian and Frequentist Methods

- Interval estimation
  - Frequentist: Confidence interval
    - ★ The experiment should be repeatable

Comparison of Bayesian and Frequentist Methods

- Interval estimation
  - Frequentist: Confidence interval
    - ★ The experiment should be repeatable
    - ★ Not conditional on the data

Comparison of Bayesian and Frequentist Methods

- Interval estimation
  - Frequentist: Confidence interval
    - ★ The experiment should be repeatable
    - ★ Not conditional on the data
  - Bayesian: Credible interval

Comparison of Bayesian and Frequentist Methods

- Interval estimation
  - Frequentist: Confidence interval
    - The experiment should be repeatable
    - Not conditional on the data
  - Bayesian: Credible interval
- Hypothesis testing

Comparison of Bayesian and Frequentist Methods

- Interval estimation
  - Frequentist: Confidence interval
    - ★ The experiment should be repeatable
    - ★ Not conditional on the data
  - Bayesian: Credible interval
- Hypothesis testing
  - Consider testing $H_0 : \theta = 0.5$ vs. $H_1 : \theta > 0.5$

Comparison of Bayesian and Frequentist Methods

- Interval estimation
  - Frequentist: Confidence interval
    - ★ The experiment should be repeatable
    - ★ Not conditional on the data
  - Bayesian: Credible interval
- Hypothesis testing
  - Consider testing $H_0 : \theta = 0.5$ vs. $H_1 : \theta > 0.5$
  - Frequentist:

Comparison of Bayesian and Frequentist Methods

- Interval estimation
  - Frequentist: Confidence interval
    - ★ The experiment should be repeatable
    - ★ Not conditional on the data
  - Bayesian: Credible interval
- Hypothesis testing
  - Consider testing $H_0 : \theta = 0.5$ vs. $H_1 : \theta > 0.5$
  - Frequentist:
    - ★ p-value can only be computed when the paramter value is given under the null

Comparison of Bayesian and Frequentist Methods

- Interval estimation
  - Frequentist: Confidence interval
    - ★ The experiment should be repeatable
    - ★ Not conditional on the data
  - Bayesian: Credible interval
- Hypothesis testing
  - Consider testing $H_0 : \theta = 0.5$ vs. $H_1 : \theta > 0.5$
  - Frequentist:
    - ★ p-value can only be computed when the paramter value is given under the null
    - ★ If we switch the role of two hypotheses, frequentists cannot compute p-value

Comparison of Bayesian and Frequentist Methods

- Interval estimation
  - Frequentist: Confidence interval
    - ★ The experiment should be repeatable
    - ★ Not conditional on the data
  - Bayesian: Credible interval
- Hypothesis testing
  - Consider testing $H_0 : \theta = 0.5$ vs. $H_1 : \theta > 0.5$
  - Frequentist:
    - ★ p-value can only be computed when the paramter value is given under the null
    - ★ If we switch the role of two hypotheses, frequentists cannot compute p-value
  - Bayesian: No trouble

Advantages and Disadvantages of Bayesian

- Disadvantages of Bayesian

Advantages and Disadvantages of Bayesian

- Disadvantages of Bayesian
  - Subjective choice of prior

Advantages and Disadvantages of Bayesian

- Disadvantages of Bayesian
  - Subjective choice of prior
  - Computation of posterior can be difficult

Advantages and Disadvantages of Bayesian

- Disadvantages of Bayesian
  - Subjective choice of prior
  - Computation of posterior can be difficult
- Advantages of Bayesian

Advantages and Disadvantages of Bayesian

- Disadvantages of Bayesian
  - Subjective choice of prior
  - Computation of posterior can be difficult
- Advantages of Bayesian
  - Can easily incorporate prior information

Advantages and Disadvantages of Bayesian

- Disadvantages of Bayesian
  - Subjective choice of prior
  - Computation of posterior can be difficult
- Advantages of Bayesian
  - Can easily incorporate prior information
  - Inferences are conditional on the actual data

Advantages and Disadvantages of Bayesian

- Disadvantages of Bayesian
  - Subjective choice of prior
  - Computation of posterior can be difficult
- Advantages of Bayesian
  - Can easily incorporate prior information
  - Inferences are conditional on the actual data
  - Results are more easily interpretable

Advantages and Disadvantages of Bayesian

- Disadvantages of Bayesian
  - Subjective choice of prior
  - Computation of posterior can be difficult
- Advantages of Bayesian
  - Can easily incorporate prior information
  - Inferences are conditional on the actual data
  - Results are more easily interpretable
  - Inferences are based on posterior, which is conceptually simple

Advantages and Disadvantages of Bayesian

- Disadvantages of Bayesian
  - Subjective choice of prior
  - Computation of posterior can be difficult
- Advantages of Bayesian
  - Can easily incorporate prior information
  - Inferences are conditional on the actual data
  - Results are more easily interpretable
  - Inferences are based on posterior, which is conceptually simple
  - In principle, all problems can be solved

Priors

- Mixture of conjugate prior $\rightarrow$ mixture of conjugate posterior

Priors

- Mixture of conjugate prior $\rightarrow$ mixture of conjugate posterior
  - Prior $\times$ posterior takes known form

Priors

- Mixture of conjugate prior $\rightarrow$ mixture of conjugate posterior
  - Prior $\times$ posterior takes known form
  - Computationally efficient

Priors

- Mixture of conjugate prior $\rightarrow$ mixture of conjugate posterior
  - Prior $\times$ posterior takes known form
  - Computationally efficient
  - Analytically tractible

Priors

- Mixture of conjugate prior $\rightarrow$ mixture of conjugate posterior
  - Prior $\times$ posterior takes known form
  - Computationally efficient
  - Analytically tractible
- Elicited (informative) prior

Priors

- Mixture of conjugate prior $\rightarrow$ mixture of conjugate posterior
  - Prior $\times$ posterior takes known form
  - Computationally efficient
  - Analytically tractible
- Elicited (informative) prior
  - Histogram

Priors

- Mixture of conjugate prior $\rightarrow$ mixture of conjugate posterior
  - Prior $\times$ posterior takes known form
  - Computationally efficient
  - Analytically tractible
- Elicited (informative) prior
  - Histogram
  - Parameters in priors using moments or quantiles

Priors

- Mixture of conjugate prior $\rightarrow$ mixture of conjugate posterior
  - Prior $\times$ posterior takes known form
  - Computationally efficient
  - Analytically tractible
- Elicited (informative) prior
  - Histogram
  - Parameters in priors using moments or quantiles
  - Interactive computer programs

Priors

- Mixture of conjugate prior $\rightarrow$ mixture of conjugate posterior
  - Prior $\times$ posterior takes known form
  - Computationally efficient
  - Analytically tractible
- Elicited (informative) prior
  - Histogram
  - Parameters in priors using moments or quantiles
  - Interactive computer programs
  - Using a hierarchical structure is often helpful

Priors

- Mixture of conjugate prior $\rightarrow$ mixture of conjugate posterior
  - Prior $\times$ posterior takes known form
  - Computationally efficient
  - Analytically tractible
- Elicited (informative) prior
  - Histogram
  - Parameters in priors using moments or quantiles
  - Interactive computer programs
  - Using a hierarchical structure is often helpful
- Noninformative prior

Priors

- Mixture of conjugate prior $\rightarrow$ mixture of conjugate posterior
  - Prior $\times$ posterior takes known form
  - Computationally efficient
  - Analytically tractible
- Elicited (informative) prior
  - Histogram
  - Parameters in priors using moments or quantiles
  - Interactive computer programs
  - Using a hierarchical structure is often helpful
- Noninformative prior
- Empirical Bayes

Noninformative Prior

- Flat (uniform)

Noninformative Prior

- Flat (uniform)
  - Improper when the support is not finite

Noninformative Prior

- Flat (uniform)
  - Improper when the support is not finite
  - Improper prior can lead to proper posterior

Noninformative Prior

- Flat (uniform)
  - Improper when the support is not finite
  - Improper prior can lead to proper posterior
  - Usually not flat after transformation

Noninformative Prior

- Flat (uniform)
  - Improper when the support is not finite
  - Improper prior can lead to proper posterior
  - Usually not flat after transformation
- Jeffrey's prior: Invariant under one-to-one transformation

Noninformative Prior

- Flat (uniform)
  - Improper when the support is not finite
  - Improper prior can lead to proper posterior
  - Usually not flat after transformation
- Jeffrey's prior: Invariant under one-to-one transformation
- Location invariant prior for location family

Noninformative Prior

- Flat (uniform)
  - Improper when the support is not finite
  - Improper prior can lead to proper posterior
  - Usually not flat after transformation
- Jeffrey's prior: Invariant under one-to-one transformation
- Location invariant prior for location family
- Scale invariant prior for scale family

Some Basic Models

- Beta-binomial model

Some Basic Models

- Beta-binomial model
  - $y|\theta \sim \text{binomial}(n, \theta)$

Some Basic Models

- Beta-binomial model
  - $y|\theta \sim \text{binomial}(n, \theta)$
  - $\theta \sim \text{beta}(a, b)$

Some Basic Models

- Beta-binomial model
  - $y|\theta \sim \text{binomial}(n, \theta)$
  - $\theta \sim \text{beta}(a, b)$
  - Posterior: beta

Some Basic Models

- Beta-binomial model
  - $y|\theta \sim \text{binomial}(n, \theta)$
  - $\theta \sim \text{beta}(a, b)$
  - Posterior: beta
- Poisson model

Some Basic Models

- Beta-binomial model
  - $y|\theta \sim \text{binomial}(n, \theta)$
  - $\theta \sim \text{beta}(a, b)$
  - Posterior: beta
- Poisson model
  - $y_1, \ldots, y_n|\theta \sim \text{Poisson}$

Some Basic Models

- Beta-binomial model
  - $y|\theta \sim \text{binomial}(n, \theta)$
  - $\theta \sim \text{beta}(a, b)$
  - Posterior: beta
- Poisson model
  - $y_1, \ldots, y_n|\theta \sim$ Poisson
  - Prior: gamma

Some Basic Models

- Beta-binomial model
  - $y|\theta \sim \text{binomial}(n, \theta)$
  - $\theta \sim \text{beta}(a, b)$
  - Posterior: beta
- Poisson model
  - $y_1, \ldots, y_n|\theta \sim$ Poisson
  - Prior: gamma
  - Marginal model: Negative-binomial

Some Basic Models

- Beta-binomial model
  - $y|\theta \sim \text{binomial}(n, \theta)$
  - $\theta \sim \text{beta}(a, b)$
  - Posterior: beta
- Poisson model
  - $y_1, \ldots, y_n|\theta \sim$ Poisson
  - Prior: gamma
  - Marginal model: Negative-binomial
- Exponential model

Some Basic Models

- Beta-binomial model
  - $y|\theta \sim \text{binomial}(n, \theta)$
  - $\theta \sim \text{beta}(a, b)$
  - Posterior: beta
- Poisson model
  - $y_1, \ldots, y_n|\theta \sim \text{Poisson}$
  - Prior: gamma
  - Marginal model: Negative-binomial
- Exponential model
  - $y_1, \ldots, y_n|\theta \sim \text{Exp}(\theta)$

Some Basic Models

- Beta-binomial model
  - $y|\theta \sim \text{binomial}(n, \theta)$
  - $\theta \sim \text{beta}(a, b)$
  - Posterior: beta
- Poisson model
  - $y_1, \ldots, y_n|\theta \sim \text{Poisson}$
  - Prior: gamma
  - Marginal model: Negative-binomial
- Exponential model
  - $y_1, \ldots, y_n|\theta \sim \text{Exp}(\theta)$
  - Prior: gamma

Some Common Conjugate Priors

Likelihood	Prior	Posterior
$X\|\theta \sim \mathcal{N}(\theta, \sigma^2)$	$\theta \sim \mathcal{N}(\mu, \tau^2)$	$\theta\|X \sim \mathcal{N}(\frac{\tau^2}{\sigma^2+\tau^2}X + \frac{\sigma^2}{\sigma^2+\tau^2}\mu, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2})$
$X\|\theta \sim \mathcal{B}(n, \theta)$	$\theta \sim \mathcal{B}e(\alpha, \beta)$	$\theta\|X \sim \mathcal{B}e(\alpha + x, n - x + \beta)$
$X_1, \ldots, X_n\|\theta \sim \mathcal{P}(\theta)$	$\theta \sim \mathcal{G}a(\alpha, \beta)$	$\theta\|X_1, \ldots, X_n \sim \mathcal{G}a(\sum_i X_i + \alpha, n + \beta)$.
$X_1, \ldots, X_n\|\theta \sim \mathcal{NB}(m, \theta)$	$\theta \sim \mathcal{B}e(\alpha, \beta)$	$\theta\|X_1, \ldots, X_n \sim \mathcal{B}e(\alpha + mn, \beta + \sum_{i=1}^n x_i)$
$X \sim \mathcal{G}(n/2, 2\theta)$	$\theta \sim \mathcal{IG}(\alpha, \beta)$	$\theta\|X \sim \mathcal{IG}(n/2 + \alpha, (x/2 + \beta^{-1})^{-1})$
$X_1, \ldots, X_n\|\theta \sim \mathcal{U}(0, \theta)$	$\theta \sim \mathcal{P}a(\theta_0, \alpha)$	$\theta\|X_1, \ldots, X_n \sim \mathcal{P}a(\max\{\theta_0, x_1, \ldots, x_n\}\alpha + n)$
$X\|\theta \sim \mathcal{N}(\mu, \theta)$	$\theta \sim \mathcal{IG}(\alpha, \beta)$	$\theta\|X \sim \mathcal{IG}(\alpha + 1/2, \beta + (\mu - X)^2/2)$
$X\|\theta \sim \mathcal{G}a(\nu, \theta)$	$\theta \sim \mathcal{G}a(\alpha, \beta)$	$\theta\|X \sim \mathcal{G}a(\alpha + \nu, \beta + x)$

Likelihood Principle

- Likelihood principle: In the inference about θ, after y is observed, all relevant experimental information is contained in the likelihood function for the observed y. Furthermore, two likelihood functions contain the same information about θ if they are proportional to each other

Likelihood Principle

- Likelihood principle: In the inference about θ, after y is observed, all relevant experimental information is contained in the likelihood function for the observed y. Furthermore, two likelihood functions contain the same information about θ if they are proportional to each other
- Consider testing the fairness of a coin:

$$H_0 : \theta = \frac{1}{2} \text{ vs. } H_1 : \theta > \frac{1}{2}$$

Likelihood Principle

- Likelihood principle: In the inference about θ, after y is observed, all relevant experimental information is contained in the likelihood function for the observed y. Furthermore, two likelihood functions contain the same information about θ if they are proportional to each other
- Consider testing the fairness of a coin:

$$H_0 : \theta = \frac{1}{2} \text{ vs. } H_1 : \theta > \frac{1}{2}$$

- Data: An experiment is conducted and 9 heads and 3 tails are observed

Two Possible Experiments

- Binomial: 12 toss in total

Two Possible Experiments

- Binomial: 12 toss in total
- Negative binomial: keep tossing until get three tails

Two Possible Experiments

- Binomial: 12 toss in total
- Negative binomial: keep tossing until get three tails
- Likelihoods are proportional

Two Possible Experiments

- Binomial: 12 toss in total
- Negative binomial: keep tossing until get three tails
- Likelihoods are proportional
- Conclusions based on p-values are contradictive $\rightarrow$ violation of likelihood principle

Two Possible Experiments

- Binomial: 12 toss in total
- Negative binomial: keep tossing until get three tails
- Likelihoods are proportional
- Conclusions based on p-values are contradictive $\rightarrow$ violation of likelihood principle
- Bayesian method has no difficulty $\rightarrow$ the same conclusion under both scenarios

Some General Facts About Bayesian Inference

- On average, posterior distribution is less variable than the prior distribution

Some General Facts About Bayesian Inference

- On average, posterior distribution is less variable than the prior distribution
  - Prior variance: $var(\theta)$

Some General Facts About Bayesian Inference

- On average, posterior distribution is less variable than the prior distribution
  - Prior variance: $var(\theta)$
  - Posterior variance: $var(\theta|y)$

Some General Facts About Bayesian Inference

- On average, posterior distribution is less variable than the prior distribution
  - Prior variance: $var(\theta)$
  - Posterior variance: $var(\theta|y)$
  - $var(\theta) = E(var(\theta|y)) + var(E(\theta|y)) \geq E(var(\theta|y))$

Some General Facts About Bayesian Inference

- On average, posterior distribution is less variable than the prior distribution
  - Prior variance: $var(\theta)$
  - Posterior variance: $var(\theta|y)$
  - $var(\theta) = E(var(\theta|y)) + var(E(\theta|y)) \geq E(var(\theta|y))$
- Sequential updates in Bayesian inference

Some General Facts About Bayesian Inference

- On average, posterior distribution is less variable than the prior distribution
  - Prior variance: $var(\theta)$
  - Posterior variance: $var(\theta|y)$
  - $var(\theta) = E(var(\theta|y)) + var(E(\theta|y)) \geq E(var(\theta|y))$
- Sequential updates in Bayesian inference
  - Prior: $p(\theta)$

Some General Facts About Bayesian Inference

- On average, posterior distribution is less variable than the prior distribution
  - Prior variance: $var(\theta)$
  - Posterior variance: $var(\theta|y)$
  - $var(\theta) = E(var(\theta|y)) + var(E(\theta|y)) \geq E(var(\theta|y))$
- Sequential updates in Bayesian inference
  - Prior: $p(\theta)$
  - After first batch of data $y_1 \rightarrow p(\theta|y_1) \propto p(y_1|\theta)p(\theta)$

Some General Facts About Bayesian Inference

- On average, posterior distribution is less variable than the prior distribution
  - Prior variance: $var(\theta)$
  - Posterior variance: $var(\theta|y)$
  - $var(\theta) = E(var(\theta|y)) + var(E(\theta|y)) \geq E(var(\theta|y))$
- Sequential updates in Bayesian inference
  - Prior: $p(\theta)$
  - After first batch of data $y_1 \rightarrow p(\theta|y_1) \propto p(y_1|\theta)p(\theta)$
  - After second batch y_2 (assume it is conditionally independent from y_1) $\rightarrow p(\theta|y_1, y_2) \propto p(y_2|\theta)p(\theta|y_1) \propto p(y_2|\theta)p(y_1\theta)p(\theta) = p(y_1, y_2|\theta)p(\theta)$

Some General Facts About Bayesian Inference

- On average, posterior distribution is less variable than the prior distribution
  - Prior variance: $var(\theta)$
  - Posterior variance: $var(\theta|y)$
  - $var(\theta) = E(var(\theta|y)) + var(E(\theta|y)) \geq E(var(\theta|y))$
- Sequential updates in Bayesian inference
  - Prior: $p(\theta)$
  - After first batch of data $y_1 \rightarrow p(\theta|y_1) \propto p(y_1|\theta)p(\theta)$
  - After second batch y_2 (assume it is conditionally independent from y_1) $\rightarrow p(\theta|y_1, y_2) \propto p(y_2|\theta)p(\theta|y_1) \propto p(y_2|\theta)p(y_1\theta)p(\theta) = p(y_1, y_2|\theta)p(\theta)$
  - This is the same as if we observed both batches together

Simulate Normal Random Variables: Box–Muller Transformation

- We require two random variables, U and V, uniformly distributed on $[0, 1]$. Set

$$R = \sqrt{-2 \log V},$$
$$\theta = 2\pi U,$$

and

$$Z_1 = R \cos \theta,$$
$$Z_2 = R \sin \theta.$$

Then they are independent standard normal variables. To obtain two standard normal variables with correlation ρ, take

$$X = Z_1$$
$$Y = \rho Z_1 + \sqrt{1 - \rho^2} Z_2$$

Multiple Linear Regression

- Response: y

Multiple Linear Regression

- Response: y
- Explanatory variables: $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$

Multiple Linear Regression

- Response: y
- Explanatory variables: $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$
- Goal: Find $y = f(\boldsymbol{X})$

Multiple Linear Regression

- Response: y
- Explanatory variables: $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$
- Goal: Find $y = f(\boldsymbol{X})$
- Data: $(\boldsymbol{x}_i, y_i), i = 1, \ldots, n$

Multiple Linear Regression

- Response: y
- Explanatory variables: $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$
- Goal: Find $y = f(\boldsymbol{X})$
- Data: $(\boldsymbol{x}_i, y_i), i = 1, \ldots, n$
- Model: $y_i = f(\boldsymbol{x}_i) + \epsilon_i$,

Multiple Linear Regression

- Response: y
- Explanatory variables: $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$
- Goal: Find $y = f(\boldsymbol{X})$
- Data: $(\boldsymbol{x}_i, y_i), i = 1, \ldots, n$
- Model: $y_i = f(\boldsymbol{x}_i) + \epsilon_i$,
- Typical assumption for normal linear model:

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) iid$$

Multiple Linear Regression

- Response: y
- Explanatory variables: $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$
- Goal: Find $y = f(\boldsymbol{X})$
- Data: $(\boldsymbol{x}_i, y_i), i = 1, \ldots, n$
- Model: $y_i = f(\boldsymbol{x}_i) + \epsilon_i$,
- Typical assumption for normal linear model:

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) iid$$

- $E(\boldsymbol{y}|\boldsymbol{X}) = f(\boldsymbol{X})$

Multiple Linear Regression

- Response: y
- Explanatory variables: $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$
- Goal: Find $y = f(\boldsymbol{X})$
- Data: $(\boldsymbol{x}_i, y_i), i = 1, \ldots, n$
- Model: $y_i = f(\boldsymbol{x}_i) + \epsilon_i$,
- Typical assumption for normal linear model:

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) iid$$

- $E(\boldsymbol{y}|\boldsymbol{X}) = f(\boldsymbol{X})$
- $\boldsymbol{y}|\boldsymbol{X} \sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 I)$

Frequentist Inference

- Ordinary least squares

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$
$$\hat{\sigma}^2 = s^2 = \frac{1}{n-k}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

Bayesian Inference

- Noninformative prior:

$$f(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{X}) \propto \sigma^{-2}$$

Bayesian Inference

- Noninformative prior:

$$f(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{X}) \propto \sigma^{-2}$$

- Posterior:

$$\begin{aligned} \boldsymbol{\beta}, \sigma^2 | \boldsymbol{y} &\sim \mathcal{N}(\hat{\boldsymbol{\beta}}, V_\beta \sigma^2) \\ \frac{(n-k)s^2}{\sigma^2} | \boldsymbol{y} &\sim \chi^2_{n-k} \\ \sigma^2 | \boldsymbol{y} &\sim Inv - \chi^2(n-k, s^2) \end{aligned}$$

Bayesian Inference

- Noninformative prior:

$$f(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{X}) \propto \sigma^{-2}$$

- Posterior:

$$\begin{aligned} \boldsymbol{\beta}, \sigma^2 | \boldsymbol{y} &\sim \mathcal{N}(\hat{\boldsymbol{\beta}}, V_\beta \sigma^2) \\ \frac{(n-k)s^2}{\sigma^2} | \boldsymbol{y} &\sim \chi^2_{n-k} \\ \sigma^2 | \boldsymbol{y} &\sim Inv - \chi^2(n-k, s^2) \end{aligned}$$

- Marginal posterior of $\boldsymbol{\beta}|\boldsymbol{y}$ is the multivariate t-distribution with $n-k$ degrees of freedom

Hierarchical Model

- Powerful technique for describing complex models. Idea is to break the model down into smaller easier understood pieces, which when put together describes the joint distribution of all data and parameters

Hierarchical Model

- Powerful technique for describing complex models. Idea is to break the model down into smaller easier understood pieces, which when put together describes the joint distribution of all data and parameters
- Why go hierarchical?

Hierarchical Model

- Powerful technique for describing complex models. Idea is to break the model down into smaller easier understood pieces, which when put together describes the joint distribution of all data and parameters
- Why go hierarchical?
  - Non-hierarchical models with few parameters generally don't fit the data well

Hierarchical Model

- Powerful technique for describing complex models. Idea is to break the model down into smaller easier understood pieces, which when put together describes the joint distribution of all data and parameters
- Why go hierarchical?
  - Non-hierarchical models with few parameters generally don't fit the data well
  - Non-hierarchical models with many parameters tend to fit the data well, but have poor predictive ability (overfitting)

Hierarchical Model

- Powerful technique for describing complex models. Idea is to break the model down into smaller easier understood pieces, which when put together describes the joint distribution of all data and parameters
- Why go hierarchical?
  - Non-hierarchical models with few parameters generally don't fit the data well
  - Non-hierarchical models with many parameters tend to fit the data well, but have poor predictive ability (overfitting)
  - Hierarchical models can often fit data with a small number of parameters but can also do well in prediction

A Simple Example

- Observed datum is $\boldsymbol{X}$

$$\boldsymbol{X}|\theta \sim \mathcal{N}(\theta, 1)$$

A Simple Example

- Observed datum is $\boldsymbol{X}$

$$\boldsymbol{X}|\theta \sim \mathcal{N}(\theta, 1)$$

- First-stage prior: $\theta|\tau^2 \sim \mathcal{N}(0, \tau^2)$

A Simple Example

- Observed datum is $\boldsymbol{X}$

$$\boldsymbol{X}|\theta \sim \mathcal{N}(\theta, 1)$$

- First-stage prior: $\theta|\tau^2 \sim \mathcal{N}(0, \tau^2)$
- Second-stage prior: $\tau^2 \sim \pi$ where π is completely specified

A Simple Example

- Observed datum is $\boldsymbol{X}$

$$\boldsymbol{X}|\theta \sim \mathcal{N}(\theta, 1)$$

- First-stage prior: $\theta|\tau^2 \sim \mathcal{N}(0, \tau^2)$
- Second-stage prior: $\tau^2 \sim \pi$ where π is completely specified
- Possibly more levels

A Simple Example

- Observed datum is $\boldsymbol{X}$

$$\boldsymbol{X}|\theta \sim \mathcal{N}(\theta, 1)$$

- First-stage prior: $\theta|\tau^2 \sim \mathcal{N}(0, \tau^2)$
- Second-stage prior: $\tau^2 \sim \pi$ where π is completely specified
- Possibly more levels
  - Second-stage prior: $\tau^2|\alpha \sim gamma(\alpha, 1)$

A Simple Example

- Observed datum is $\boldsymbol{X}$

$$\boldsymbol{X}|\theta \sim \mathcal{N}(\theta, 1)$$

- First-stage prior: $\theta|\tau^2 \sim \mathcal{N}(0, \tau^2)$
- Second-stage prior: $\tau^2 \sim \pi$ where π is completely specified
- Possibly more levels
  - Second-stage prior: $\tau^2|\alpha \sim gamma(\alpha, 1)$
  - Third-stage prior: $\alpha \sim Exp(1)$

Hierarchical Linear Model

$$
\begin{aligned}
Y|X, \beta, \Sigma &\sim \mathcal{N}(X\beta, \Sigma) \\
\beta|X_\beta, \alpha, \Sigma_\beta &\sim \mathcal{N}(X_\beta\alpha, \Sigma_\beta) \\
\alpha|\alpha_0, \Sigma_\alpha &\sim \mathcal{N}(\alpha_0, \Sigma_\alpha)
\end{aligned}
$$

Simple Random Effects Model

- J groups

Simple Random Effects Model

- J groups
- Data in group j: $Y_{1j}, \ldots, Y_{n_j j}$

Simple Random Effects Model

- J groups
- Data in group j: $Y_{1j}, \ldots, Y_{n_j j}$
- $Y_{ij}|\beta_j, \sigma^2 \sim \mathcal{N}(\beta_j, \sigma^2)$ independent, $j = 1, \ldots, J$, $i = 1, \ldots, n_j$

Simple Random Effects Model

- J groups
- Data in group j: $Y_{1j}, \ldots, Y_{n_j j}$
- $Y_{ij}|\beta_j, \sigma^2 \sim \mathcal{N}(\beta_j, \sigma^2)$ independent, $j = 1, \ldots, J$, $i = 1, \ldots, n_j$
- Random effects: $\beta_j|\alpha, \sigma_\beta^2 \sim \mathcal{N}(\alpha, \sigma_\beta^2)$ iid

Simple Random Effects Model

- J groups
- Data in group j: $Y_{1j}, \ldots, Y_{n_j j}$
- $Y_{ij}|\beta_j, \sigma^2 \sim \mathcal{N}(\beta_j, \sigma^2)$ independent, $j = 1, \ldots, J$, $i = 1, \ldots, n_j$
- Random effects: $\beta_j|\alpha, \sigma_\beta^2 \sim \mathcal{N}(\alpha, \sigma_\beta^2)$ iid
- $\alpha \sim \mathcal{N}(\alpha_0, \sigma_\alpha^2)$

Posterior Inference by Simulation

- Approach 1: Independence sampling

Posterior Inference by Simulation

- Approach 1: Independence sampling
  - Simulate independent samples from the posterior distributions

Posterior Inference by Simulation

- Approach 1: Independence sampling
  - Simulate independent samples from the posterior distributions
  - Draw $\theta^{(1)}, \ldots, \theta^{(M)}$ iid from the posterior distribution $f(\theta|y)$

Posterior Inference by Simulation

- Approach 1: Independence sampling
  - Simulate independent samples from the posterior distributions
  - Draw $\theta^{(1)}, \ldots, \theta^{(M)}$ iid from the posterior distribution $f(\theta|y)$
  - This is all we have discussed so far

Posterior Inference by Simulation

- Approach 1: Independence sampling
  - Simulate independent samples from the posterior distributions
  - Draw $\theta^{(1)}, \ldots, \theta^{(M)}$ iid from the posterior distribution $f(\theta|y)$
  - This is all we have discussed so far
- Approach 2: Markov Chain Monte Carlo (MCMC)

Posterior Inference by Simulation

- Approach 1: Independence sampling
  - Simulate independent samples from the posterior distributions
  - Draw $\theta^{(1)}, \ldots, \theta^{(M)}$ iid from the posterior distribution $f(\theta|y)$
  - This is all we have discussed so far
- Approach 2: Markov Chain Monte Carlo (MCMC)
  - Draw $\theta^{(i+1)}$ from $g(\theta^{(i+1)}|\theta^{(i)}$ such that

$$f(\theta^{(i+1)}) \to f(\theta|y)$$

Posterior Inference by Simulation

- Approach 1: Independence sampling
  - Simulate independent samples from the posterior distributions
  - Draw $\theta^{(1)}, \ldots, \theta^{(M)}$ iid from the posterior distribution $f(\theta|y)$
  - This is all we have discussed so far
- Approach 2: Markov Chain Monte Carlo (MCMC)
  - Draw $\theta^{(i+1)}$ from $g(\theta^{(i+1)}|\theta^{(i)}$ such that

$$f(\theta^{(i+1)}) \to f(\theta|y)$$

  - Gibbs sampling

Posterior Inference by Simulation

- Approach 1: Independence sampling
  - Simulate independent samples from the posterior distributions
  - Draw $\theta^{(1)}, \ldots, \theta^{(M)}$ iid from the posterior distribution $f(\theta|y)$
  - This is all we have discussed so far
- Approach 2: Markov Chain Monte Carlo (MCMC)
  - Draw $\theta^{(i+1)}$ from $g(\theta^{(i+1)}|\theta^{(i)}$ such that

$$f(\theta^{(i+1)}) \rightarrow f(\theta|y)$$

  - Gibbs sampling
  - Metropolis-Hastings

Posterior Inference by Simulation

- Approach 1: Independence sampling
  - Simulate independent samples from the posterior distributions
  - Draw $\theta^{(1)}, \ldots, \theta^{(M)}$ iid from the posterior distribution $f(\theta|y)$
  - This is all we have discussed so far
- Approach 2: Markov Chain Monte Carlo (MCMC)
  - Draw $\theta^{(i+1)}$ from $g(\theta^{(i+1)}|\theta^{(i)}$ such that

$$f(\theta^{(i+1)}) \to f(\theta|y)$$

  - Gibbs sampling
  - Metropolis-Hastings
  - Hamiltonian dynamics

Gibbs Sampler

- Used for multiparameter models

Gibbs Sampler

- Used for multiparameter models
- Parameter: $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$

Gibbs Sampler

- Used for multiparameter models
- Parameter: $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$
- An iterative algorithm

Gibbs Sampler

- Used for multiparameter models
- Parameter: $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$
- An iterative algorithm
  - Draw θ_1 from $p(\theta_1 | \theta_2, \theta_3, \ldots, \theta_k, y)$

Gibbs Sampler

- Used for multiparameter models
- Parameter: $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$
- An iterative algorithm
  - Draw θ_1 from $p(\theta_1 | \theta_2, \theta_3, \ldots, \theta_k, y)$
  - Draw θ_2 from $p(\theta_2 | \theta_1, \theta_3, \ldots, \theta_k, y)$

Gibbs Sampler

- Used for multiparameter models
- Parameter: $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$
- An iterative algorithm
  - Draw θ_1 from $p(\theta_1 | \theta_2, \theta_3, \ldots, \theta_k, y)$
  - Draw θ_2 from $p(\theta_2 | \theta_1, \theta_3, \ldots, \theta_k, y)$
  - $\ldots$

Gibbs Sampler

- Used for multiparameter models
- Parameter: $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$
- An iterative algorithm
  - Draw θ_1 from $p(\theta_1 | \theta_2, \theta_3, \ldots, \theta_k, y)$
  - Draw θ_2 from $p(\theta_2 | \theta_1, \theta_3, \ldots, \theta_k, y)$
  - $\ldots$
  - Draw θ_k from $p(\theta_k | \theta_1, \theta_2, \ldots, \theta_{k-1}, y)$

Gibbs Sampler (cont.)

- Full conditional distribution $p(\theta_j|\theta_{-j}, y)$, where $\theta_{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_k)$

Gibbs Sampler (cont.)

- Full conditional distribution $p(\theta_j|\theta_{-j}, y)$, where $\theta_{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_k)$
- In iteration t, draw $\theta_j^t = p(\theta_j|\theta_{-j}^t, y)$, where $(\theta_1^t, \ldots, \theta_{j-1}^t, \theta_{j+1}^{t-1}, \ldots, \theta_k^{t-1})$

Gibbs Sampler (cont.)

- Full conditional distribution $p(\theta_j|\theta_{-j}, y)$, where $\theta_{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_k)$
- In iteration t, draw $\theta_j^t = p(\theta_j|\theta_{-j}^t, y)$, where $(\theta_1^t, \ldots, \theta_{j-1}^t, \theta_{j+1}^{t-1}, \ldots, \theta_k^{t-1})$
- Each θ_j is updated conditional on the latest values of θ

Example: Simulate from a Bivariate Normal Distribution

S

- Joint distribution

$$\boldsymbol{Z} = (X, Y)' \sim \mathcal{N}(0, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix})$$

Python Code

```
import numpy as np
x = [0]
y = [0]
rho = .9
c = np.sqrt(1 - rho**2)
for i in range (6000):
    x.append(rho*y[i - 1] + c*np.random.normal(0, 1, 1))
    y.append(rho*x[i] + c*np.random.normal(0, 1, 1))

plt.style.use('classic')
plt.figure()
plt.subplot(3, 1, 1)
plt.plot(x[0:14], y[0:14], '-')
plt.subplot(3, 1, 2)
plt.plot(x[0:49], y[0:49], '-')
plt.subplot(3, 1, 3)
plt.plot(x, y, '-');
```

The Resulting Posterior

General Property of Gibbs Sampler

- Output: A dependent sequence

$$\theta^{(1)} = \{\theta_1^{(1)}, \ldots, \theta_p^{(1)}\}$$
$$\theta^{(2)} = \{\theta_1^{(2)}, \ldots, \theta_p^{(2)}\}$$
$$\vdots$$
$$\theta^{(S)} = \{\theta_1^{(S)}, \ldots, \theta_p^{(S)}\}$$

General Property of Gibbs Sampler

- Output: A dependent sequence

$$\theta^{(1)} = \{\theta_1^{(1)}, \ldots, \theta_p^{(1)}\}$$
$$\theta^{(2)} = \{\theta_1^{(2)}, \ldots, \theta_p^{(2)}\}$$
$$\vdots$$
$$\theta^{(S)} = \{\theta_1^{(S)}, \ldots, \theta_p^{(S)}\}$$

- $\theta^{(S)}$ depends on $\theta^{(0)}, \ldots, \theta^{(S-1)}$ only through $\theta^{(S-1)}$

General Property of Gibbs Sampler

- Output: A dependent sequence

$$\theta^{(1)} = \{\theta_1^{(1)}, \ldots, \theta_p^{(1)}\}$$
$$\theta^{(2)} = \{\theta_1^{(2)}, \ldots, \theta_p^{(2)}\}$$
$$\vdots$$
$$\theta^{(S)} = \{\theta_1^{(S)}, \ldots, \theta_p^{(S)}\}$$

- $\theta^{(S)}$ depends on $\theta^{(0)}, \ldots, \theta^{(S-1)}$ only through $\theta^{(S-1)}$
- $\theta^{(S)}$ is conditionally independent of $\theta^{(0)}, \ldots, \theta^{(S-2)}$ given $\theta^{(S-1)}$

General Property of Gibbs Sampler

- Output: A dependent sequence

$$\theta^{(1)} = \{\theta_1^{(1)}, \ldots, \theta_p^{(1)}\}$$
$$\theta^{(2)} = \{\theta_1^{(2)}, \ldots, \theta_p^{(2)}\}$$
$$\vdots$$
$$\theta^{(S)} = \{\theta_1^{(S)}, \ldots, \theta_p^{(S)}\}$$

- $\theta^{(S)}$ depends on $\theta^{(0)}, \ldots, \theta^{(S-1)}$ only through $\theta^{(S-1)}$
- $\theta^{(S)}$ is conditionally independent of $\theta^{(0)}, \ldots, \theta^{(S-2)}$ given $\theta^{(S-1)}$
- This is called a Markov property, and the sequence a Markov chain

General Property of Gibbs Sampler

- Output: A dependent sequence

$$\theta^{(1)} = \{\theta_1^{(1)}, \ldots, \theta_p^{(1)}\}$$
$$\theta^{(2)} = \{\theta_1^{(2)}, \ldots, \theta_p^{(2)}\}$$
$$\vdots$$
$$\theta^{(S)} = \{\theta_1^{(S)}, \ldots, \theta_p^{(S)}\}$$

- $\theta^{(S)}$ depends on $\theta^{(0)}, \ldots, \theta^{(S-1)}$ only through $\theta^{(S-1)}$
- $\theta^{(S)}$ is conditionally independent of $\theta^{(0)}, \ldots, \theta^{(S-2)}$ given $\theta^{(S-1)}$
- This is called a Markov property, and the sequence a Markov chain
- For the models in this class, the sampling distribution of $\theta^{(S)}$ approaches the target distribution as $S \to \infty$, regardless of starting value

$$Pr(\theta^{(S)} \in A \to \int_A p(\theta) d\theta \text{ as } S \to \infty$$

General Property of Gibbs Sampler (cont.)

- More importantly, for most functions g of interest,

$$\frac{1}{S}\sum_{s=1}^{S} g(\theta^{(s)}) \to E[g(\theta)] = \int g(\theta)p(\theta)d\theta \text{ as } S \to \infty$$

General Property of Gibbs Sampler (cont.)

- More importantly, for most functions g of interest,

$$\frac{1}{S}\sum_{s=1}^{S} g(\theta^{(s)}) \to E[g(\theta)] = \int g(\theta)p(\theta)d\theta \text{ as } S \to \infty$$

- One can approximate $E[g(\theta)]$ with sample average of $\{g(\theta^{(1)}), \ldots, g(\theta^{(S)})\}$. This is the Monte Carlo part

General Property of Gibbs Sampler (cont.)

- More importantly, for most functions g of interest,

$$\frac{1}{S}\sum_{s=1}^{S} g(\theta^{(s)}) \to E[g(\theta)] = \int g(\theta)p(\theta)d\theta \text{ as } S \to \infty$$

- One can approximate $E[g(\theta)]$ with sample average of $\{g(\theta^{(1)}), \ldots, g(\theta^{(S)})\}$. This is the Monte Carlo part
- Hence, we call this method Markov chain Monte Carlo (MCMC)

Distinguishing Parameter Estimation from Posterior Approximation

- Bayesian data analysis using Monte Carlo methods

Distinguishing Parameter Estimation from Posterior Approximation

- Bayesian data analysis using Monte Carlo methods
  - Data analysis: The statistical part

Distinguishing Parameter Estimation from Posterior Approximation

- Bayesian data analysis using Monte Carlo methods
  - Data analysis: The statistical part
  - Numerical approximation: The Monte Carlo part

Distinguishing Parameter Estimation from Posterior Approximation

- Bayesian data analysis using Monte Carlo methods
  - Data analysis: The statistical part
  - Numerical approximation: The Monte Carlo part
- Ingredients of Bayesian data analysis

Distinguishing Parameter Estimation from Posterior Approximation

- Bayesian data analysis using Monte Carlo methods
  - Data analysis: The statistical part
  - Numerical approximation: The Monte Carlo part
- Ingredients of Bayesian data analysis
  - Model specification

Distinguishing Parameter Estimation from Posterior Approximation

- Bayesian data analysis using Monte Carlo methods
  - Data analysis: The statistical part
  - Numerical approximation: The Monte Carlo part
- Ingredients of Bayesian data analysis
  - Model specification
  - Prior specification

Distinguishing Parameter Estimation from Posterior Approximation

- Bayesian data analysis using Monte Carlo methods
  - Data analysis: The statistical part
  - Numerical approximation: The Monte Carlo part
- Ingredients of Bayesian data analysis
  - Model specification
  - Prior specification
  - Posterior summary

Distinguishing Parameter Estimation from Posterior Approximation

- Bayesian data analysis using Monte Carlo methods
  - Data analysis: The statistical part
  - Numerical approximation: The Monte Carlo part
- Ingredients of Bayesian data analysis
  - Model specification
  - Prior specification
  - Posterior summary
- When the posterior distribution is complicated, we can "look at" the posterior by studying Monte Carlo samples from the posterior

Distinguishing Parameter Estimation from Posterior Approximation

- Monte Carlo and MCMC sampling algorithms

Distinguishing Parameter Estimation from Posterior Approximation

- Monte Carlo and MCMC sampling algorithms
  - Are not models

Distinguishing Parameter Estimation from Posterior Approximation

- Monte Carlo and MCMC sampling algorithms
  - Are not models
  - They do not generate more information than is in y and $p(\theta)$

Distinguishing Parameter Estimation from Posterior Approximation

- Monte Carlo and MCMC sampling algorithms
  - Are not models
  - They do not generate more information than is in y and $p(\theta)$
  - They are simply ways of looking at $p(\theta|y)$

Distinguishing Parameter Estimation from Posterior Approximation

- Monte Carlo and MCMC sampling algorithms
  - Are not models
  - They do not generate more information than is in y and $p(\theta)$
  - They are simply ways of looking at $p(\theta|y)$
- Estimation: How we use $p(\theta|y)$ to make inferences about θ

Distinguishing Parameter Estimation from Posterior Approximation

- Monte Carlo and MCMC sampling algorithms
  - Are not models
  - They do not generate more information than is in y and $p(\theta)$
  - They are simply ways of looking at $p(\theta|y)$
- Estimation: How we use $p(\theta|y)$ to make inferences about θ
- Approximation: The use of Monte Carlo procedures to approximate integrals

Additional Issues

- Length of a chain

Additional Issues

- Length of a chain
- Multiple chains (in parallel)

Additional Issues

- Length of a chain
- Multiple chains (in parallel)
- Burn-in

Additional Issues

- Length of a chain
- Multiple chains (in parallel)
- Burn-in
- Reduce autocorrelation by thinning

Additional Issues

- Length of a chain
- Multiple chains (in parallel)
- Burn-in
- Reduce autocorrelation by thinning
- Convergence assessment

Additional Issues

- Length of a chain
- Multiple chains (in parallel)
- Burn-in
- Reduce autocorrelation by thinning
- Convergence assessment
- We will deal with these later

A Motivating Example for Generalized Linear Models

- A sample from of a population of 52 female song sparrows was studied over the course of a summer and their reproductive activities were recorded

A Motivating Example for Generalized Linear Models

- A sample from of a population of 52 female song sparrows was studied over the course of a summer and their reproductive activities were recorded
  - The age and number of new offspring were recorded for each sparrow

A Motivating Example for Generalized Linear Models

- A sample from of a population of 52 female song sparrows was studied over the course of a summer and their reproductive activities were recorded
  - The age and number of new offspring were recorded for each sparrow
- Goal

A Motivating Example for Generalized Linear Models

- A sample from of a population of 52 female song sparrows was studied over the course of a summer and their reproductive activities were recorded
  - The age and number of new offspring were recorded for each sparrow
- Goal
  - Understand the relationship between age and reproductive success

A Motivating Example for Generalized Linear Models

- A sample from of a population of 52 female song sparrows was studied over the course of a summer and their reproductive activities were recorded
  - The age and number of new offspring were recorded for each sparrow
- Goal
  - Understand the relationship between age and reproductive success
  - Make population forecasts for this group of birds

The Data

Poisson Regression

- Let y = number of offspring, x = age

Poisson Regression

- Let y = number of offspring, x = age
- $y|x \sim Poisson(\theta_x), \theta_x = E(y|x)$

Poisson Regression

- Let y = number of offspring, x = age
- $y|x \sim Poisson(\theta_x), \theta_x = E(y|x)$
- Separate estimate of θ_x for each age group may be imprecise when the number of birds of the same age is small

Poisson Regression

- Let y = number of offspring, x = age
- $y|x \sim Poisson(\theta_x), \theta_x = E(y|x)$
- Separate estimate of θ_x for each age group may be imprecise when the number of birds of the same age is small
- We may assume $\theta_x = \beta_1 + \beta_2 x + \beta_3 x^2$

Poisson Regression

- Let y = number of offspring, x = age
- $y|x \sim Poisson(\theta_x), \theta_x = E(y|x)$
- Separate estimate of θ_x for each age group may be imprecise when the number of birds of the same age is small
- We may assume $\theta_x = \beta_1 + \beta_2 x + \beta_3 x^2$
  - But this may produce negative values of θ_x

Poisson Regression

- Let y = number of offspring, x = age
- $y|x \sim Poisson(\theta_x), \theta_x = E(y|x)$
- Separate estimate of θ_x for each age group may be imprecise when the number of birds of the same age is small
- We may assume $\theta_x = \beta_1 + \beta_2 x + \beta_3 x^2$
  - But this may produce negative values of θ_x
- A better model: $\log \theta_x = \beta_1 + \beta_2 x + \beta_3 x^2$

Poisson Regression

- Let y = number of offspring, x = age
- $y|x \sim Poisson(\theta_x), \theta_x = E(y|x)$
- Separate estimate of θ_x for each age group may be imprecise when the number of birds of the same age is small
- We may assume $\theta_x = \beta_1 + \beta_2 x + \beta_3 x^2$
  - But this may produce negative values of θ_x
- A better model: $\log \theta_x = \beta_1 + \beta_2 x + \beta_3 x^2$
  - i.e. $\theta_x = \exp(\beta_1 + \beta_2 x + \beta_3 x^2)$

Poisson Regression

- Let y = number of offspring, x = age
- $y|x \sim Poisson(\theta_x), \theta_x = E(y|x)$
- Separate estimate of θ_x for each age group may be imprecise when the number of birds of the same age is small
- We may assume $\theta_x = \beta_1 + \beta_2 x + \beta_3 x^2$
  - But this may produce negative values of θ_x
- A better model: $\log \theta_x = \beta_1 + \beta_2 x + \beta_3 x^2$
  - i.e. $\theta_x = \exp(\beta_1 + \beta_2 x + \beta_3 x^2)$
- Poisson regression: $y|x \sim Poisson(\exp(\boldsymbol{\beta}^T \boldsymbol{x}))$

Generalized Linear Model

- Logistic regression

Generalized Linear Model

- Logistic regression
  - $y|x \sim Bernoulli(\theta_x), \theta_x = P(y = 1|x) = E(y|x)$

Generalized Linear Model

- Logistic regression
  - $y|x \sim Bernoulli(\theta_x), \theta_x = P(y = 1|x) = E(y|x)$
  - $\text{logit}(\theta_x) = \log \frac{\theta_x}{1-\theta_x} = \boldsymbol{\beta}^T \boldsymbol{x}$

Generalized Linear Model

- Logistic regression
  - $y|x \sim Bernoulli(\theta_x), \theta_x = P(y = 1|x) = E(y|x)$
  - $\text{logit}(\theta_x) = \log \frac{\theta_x}{1-\theta_x} = \boldsymbol{\beta}^T \boldsymbol{x}$
  - Link function: logit

Generalized Linear Model

- Logistic regression
  - $y|x \sim Bernoulli(\theta_x), \theta_x = P(y = 1|x) = E(y|x)$
  - $\text{logit}(\theta_x) = \log \frac{\theta_x}{1-\theta_x} = \boldsymbol{\beta}^T \boldsymbol{x}$
  - Link function: logit
- In general, $g(E(y|x)) = \boldsymbol{\beta}^T \boldsymbol{x}$

Generalized Linear Model

- Logistic regression
  - $y|x \sim Bernoulli(\theta_x), \theta_x = P(y = 1|x) = E(y|x)$
  - $\text{logit}(\theta_x) = \log \frac{\theta_x}{1-\theta_x} = \boldsymbol{\beta}^T \boldsymbol{x}$
  - Link function: logit
- In general, $g(E(y|x)) = \boldsymbol{\beta}^T \boldsymbol{x}$
  - $g()$: Link function

Generalized Linear Model

- Logistic regression
  - $y|x \sim Bernoulli(\theta_x), \theta_x = P(y = 1|x) = E(y|x)$
  - $\text{logit}(\theta_x) = \log \frac{\theta_x}{1-\theta_x} = \boldsymbol{\beta}^T \boldsymbol{x}$
  - Link function: logit
- In general, $g(E(y|x)) = \boldsymbol{\beta}^T \boldsymbol{x}$
  - $g()$: Link function
- No conjugate prior for GLM except identity link $\rightarrow$ we need a more general MCMC algorithm than Gibbs sampling

The Metropolis Algorithm

- When using simulation to describe the posterior $p(\theta|y)$, the goal is to construct a large collection of θ-values, $\{\theta^{(1)}, \ldots, \theta^{S}\}$, whose empirical distribution approximates $p(\theta|y)$

The Metropolis Algorithm

- When using simulation to describe the posterior $p(\theta|y)$, the goal is to construct a large collection of θ-values, $\{\theta^{(1)}, \ldots, \theta^{S}\}$, whose empirical distribution approximates $p(\theta|y)$
  - We need

$$\frac{\#\{\theta^{(s)}\text{'s in the collection } = \theta_a\}}{\#\{\theta^{(s)}\text{'s in the collection } = \theta_b\}} \approx \frac{p(\theta_a|y)}{p(\theta_b|y)}$$

The Metropolis Algorithm

- When using simulation to describe the posterior $p(\theta|y)$, the goal is to construct a large collection of θ-values, $\{\theta^{(1)}, \ldots, \theta^{S}\}$, whose empirical distribution approximates $p(\theta|y)$
  - We need

$$\frac{\#\{\theta^{(s)}\text{'s in the collection } = \theta_a\}}{\#\{\theta^{(s)}\text{'s in the collection } = \theta_b\}} \approx \frac{p(\theta_a|y)}{p(\theta_b|y)}$$

- How to construct $\{\theta^{(1)}, \ldots, \theta^{S}\}$

The Metropolis Algorithm

- When using simulation to describe the posterior $p(\theta|y)$, the goal is to construct a large collection of θ-values, $\{\theta^{(1)}, \ldots, \theta^{S}\}$, whose empirical distribution approximates $p(\theta|y)$
  - We need

$$\frac{\#\{\theta^{(s)}\text{'s in the collection } = \theta_a\}}{\#\{\theta^{(s)}\text{'s in the collection } = \theta_b\}} \approx \frac{p(\theta_a|y)}{p(\theta_b|y)}$$

- How to construct $\{\theta^{(1)}, \ldots, \theta^{S}\}$
  - Suppose we have $\{\theta^{(1)}, \ldots, \theta^{S}\}$ and want to add new $\theta^{(S+1)}$

The Metropolis Algorithm

- When using simulation to describe the posterior $p(\theta|y)$, the goal is to construct a large collection of θ-values, $\{\theta^{(1)}, \ldots, \theta^{S}\}$, whose empirical distribution approximates $p(\theta|y)$
  - We need

$$\frac{\#\{\theta^{(s)}\text{'s in the collection } = \theta_a\}}{\#\{\theta^{(s)}\text{'s in the collection } = \theta_b\}} \approx \frac{p(\theta_a|y)}{p(\theta_b|y)}$$

- How to construct $\{\theta^{(1)}, \ldots, \theta^{S}\}$
  - Suppose we have $\{\theta^{(1)}, \ldots, \theta^{S}\}$ and want to add new $\theta^{(S+1)}$
  - We may choose a candidate θ^* near $\theta^{(S)}$

The Metropolis Algorithm

- When using simulation to describe the posterior $p(\theta|y)$, the goal is to construct a large collection of θ-values, $\{\theta^{(1)}, \ldots, \theta^{S}\}$, whose empirical distribution approximates $p(\theta|y)$
  - We need

$$\frac{\#\{\theta^{(s)}\text{'s in the collection } = \theta_a\}}{\#\{\theta^{(s)}\text{'s in the collection } = \theta_b\}} \approx \frac{p(\theta_a|y)}{p(\theta_b|y)}$$

- How to construct $\{\theta^{(1)}, \ldots, \theta^{S}\}$
  - Suppose we have $\{\theta^{(1)}, \ldots, \theta^{S}\}$ and want to add new $\theta^{(S+1)}$
  - We may choose a candidate θ^* near $\theta^{(S)}$
  - Should we include θ^* into the collection?

The Metropolis Algorithm (cont.)

- If $p(\theta^*|y) > p(\theta^{(S)}|y)$ we clearly want θ^* in the collection

The Metropolis Algorithm (cont.)

- If $p(\theta^*|y) > p(\theta^{(S)}|y)$ we clearly want θ^* in the collection
- If $p(\theta^*|y) < p(\theta^{(S)}|y)$ we need to decide to keep it or not

The Metropolis Algorithm (cont.)

- If $p(\theta^*|y) > p(\theta^{(S)}|y)$ we clearly want θ^* in the collection
- If $p(\theta^*|y) < p(\theta^{(S)}|y)$ we need to decide to keep it or not
- Decision depends on ratio $r = \frac{p(\theta^*|y)}{p(\theta^{(S)}|y)}$

The Metropolis Algorithm (cont.)

- If $p(\theta^*|y) > p(\theta^{(S)}|y)$ we clearly want θ^* in the collection
- If $p(\theta^*|y) < p(\theta^{(S)}|y)$ we need to decide to keep it or not
- Decision depends on ratio $r = \frac{p(\theta^*|y)}{p(\theta^{(S)}|y)}$
- We can always calculate this ratio easily

$$r = \frac{p(\theta^*|y)}{p(\theta^{(S)}|y)} = \frac{p(y|\theta^*)p(\theta^*)}{p(y)} \frac{p(y)}{p(y|\theta^{(S)})p(\theta^{(S)})} = \frac{p(y|\theta^*)p(\theta^*)}{p(y|\theta^{(S)})p(\theta^{(S)})}$$

The Metropolis Algorithm (cont.)

- How to choose θ^*?

The Metropolis Algorithm (cont.)

- How to choose θ^*?
  - Sample θ^* from a proposal distribution $J(\theta^* | \theta^{(S)})$

The Metropolis Algorithm (cont.)

- How to choose θ^*?
  - Sample θ^* from a proposal distribution $J(\theta^*|\theta^{(S)})$
  - The Metropolis algorithm uses a symmetric proposal distribution $(J(\theta_a|\theta_b) = J(\theta_b|\theta_a))$

The Metropolis Algorithm (cont.)

- How to choose θ^*?
  - Sample θ^* from a proposal distribution $J(\theta^*|\theta^{(S)})$
  - The Metropolis algorithm uses a symmetric proposal distribution $(J(\theta_a|\theta_b) = J(\theta_b|\theta_a))$
  - Example: $\theta^*|\theta^{(S)} \sim U(\theta^{(S)} - \delta, \theta^{(S)} + \delta)$ or $\theta^*|\theta^{(S)} \sim \mathcal{N}(\theta^{(S)}, \delta^2)$

The Metropolis Algorithm (cont.)

- 1) Sample $\theta^* \sim J(\theta|\theta^{(s)})$

The Metropolis Algorithm (cont.)

- 1) Sample $\theta^* \sim J(\theta|\theta^{(s)})$
- 2) Compute the acceptance ratio

$$r = \frac{p(\theta^*|y)}{p(\theta^{(s)}|y)} = \frac{p(y|\theta^*)p(\theta^*)}{p(y|\theta^{(s)}p(\theta^{(s)})}.$$

The Metropolis Algorithm (cont.)

- 1) Sample $\theta^* \sim J(\theta|\theta^{(s)})$
- 2) Compute the acceptance ratio

$$r = \frac{p(\theta^*|y)}{p(\theta^{(s)}|y)} = \frac{p(y|\theta^*)p(\theta^*)}{p(y|\theta^{(s)}p(\theta^{(s)})}.$$

- 3) Let

$$\theta^{(s+1)} = \begin{cases} \theta^* & \text{with probability } \min(r,1) \\ \theta^{(s)} & \text{with probability } 1-\min(r,1) \end{cases}$$

The Metropolis Algorithm (cont.)

- 1) Sample $\theta^* \sim J(\theta|\theta^{(s)})$
- 2) Compute the acceptance ratio

$$r = \frac{p(\theta^*|y)}{p(\theta^{(s)}|y)} = \frac{p(y|\theta^*)p(\theta^*)}{p(y|\theta^{(s)}p(\theta^{(s)})}.$$

- 3) Let

$$\theta^{(s+1)} = \begin{cases} \theta^* & \text{with probability } \min(r,1) \\ \theta^{(s)} & \text{with probability } 1-\min(r,1) \end{cases}$$

- Step 3 can be accomplished by sampling $u \sim \text{uniform}(0,1)$ and setting $\theta^{(s+1)} = \theta^*$ if $u < r$ and setting $\theta^{(s+1)} = \theta^{(s)}$ otherwise

The Metropolis Algorithm (cont.)

- 1) Sample $\theta^* \sim J(\theta|\theta^{(s)})$
- 2) Compute the acceptance ratio

$$r = \frac{p(\theta^*|y)}{p(\theta^{(s)}|y)} = \frac{p(y|\theta^*)p(\theta^*)}{p(y|\theta^{(s)}p(\theta^{(s)})}.$$

- 3) Let

$$\theta^{(s+1)} = \begin{cases} \theta^* & \text{with probability } \min(r,1) \\ \theta^{(s)} & \text{with probability } 1-\min(r,1) \end{cases}$$

- Step 3 can be accomplished by sampling $u \sim \text{uniform}(0,1)$ and setting $\theta^{(s+1)} = \theta^*$ if $u < r$ and setting $\theta^{(s+1)} = \theta^{(s)}$ otherwise
- Output: A Markov chain $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$

Example: Normal Model with Known Variance

- Data: $y_1, \ldots, y_n | \theta \sim \mathcal{N}(\theta, \sigma^2)$

Example: Normal Model with Known Variance

- Data: $y_1, \ldots, y_n | \theta \sim \mathcal{N}(\theta, \sigma^2)$
- Prior: $\theta \sim \mathcal{N}(\mu, \tau^2)$

Example: Normal Model with Known Variance

- Data: $y_1, \ldots, y_n | \theta \sim \mathcal{N}(\theta, \sigma^2)$
- Prior: $\theta \sim \mathcal{N}(\mu, \tau^2)$
- Suppose that
  $\sigma^2 = 1, \tau^2 = 10, \mu = 5, n = 5, y = (9.37, 10.18, 9.16, 11.60, 10.33)$

Example: Normal Model with Known Variance

- Data: $y_1, \ldots, y_n | \theta \sim \mathcal{N}(\theta, \sigma^2)$
- Prior: $\theta \sim \mathcal{N}(\mu, \tau^2)$
- Suppose that
  $\sigma^2 = 1, \tau^2 = 10, \mu = 5, n = 5, y = (9.37, 10.18, 9.16, 11.60, 10.33)$
- From what we learned before, we can show that $\theta | y \sim \mathcal{N}(10.03, 0.44)$

Example: Normal Model with Known Variance

- Data: $y_1, \ldots, y_n | \theta \sim \mathcal{N}(\theta, \sigma^2)$
- Prior: $\theta \sim \mathcal{N}(\mu, \tau^2)$
- Suppose that
  $\sigma^2 = 1, \tau^2 = 10, \mu = 5, n = 5, y = (9.37, 10.18, 9.16, 11.60, 10.33)$
- From what we learned before, we can show that $\theta | y \sim \mathcal{N}(10.03, 0.44)$
- Let's sample from the posterior distribution using Metropolis algorithm and compare to exact solution

Example: Normal Model with Known Variance (cont.)

- The acceptance ratio is

$$r = \frac{p(\theta^*|\boldsymbol{y})}{p(\theta^{(s)}|\boldsymbol{y}} = \left(\frac{\prod_{i=1}^n \Phi(y_i, \theta^*, \sigma)}{\prod_{i=1}^n \Phi(y_i, \theta^{(s)}, \sigma)}\right) \times \left(\frac{\Phi(\theta^*, \mu, \tau)}{\Phi(\theta^{(s)}, \mu, \tau)}\right)$$

Example: Normal Model with Known Variance (cont.)

- The acceptance ratio is

$$r = \frac{p(\theta^*|\boldsymbol{y})}{p(\theta^{(s)}|\boldsymbol{y}} = \left(\frac{\prod_{i=1}^n \Phi(y_i, \theta^*, \sigma)}{\prod_{i=1}^n \Phi(y_i, \theta^{(s)}, \sigma)}\right) \times \left(\frac{\Phi(\theta^*, \mu, \tau)}{\Phi(\theta^{(s)}, \mu, \tau)}\right)$$

- Often, the above calculation is numerically unstable and may compute logarithm instead

$$\log r = \sum_{i=1}^n [\log \Phi(y_i, \theta^*, \sigma) - \log \Phi(y_i, \theta^{(s)}, \sigma)] + [\log \Phi(\theta^*, \mu, \tau) - \log \Phi(\theta^{(s)}, \mu, \tau)].$$

Example: Normal Model with Known Variance (cont.)

- The acceptance ratio is

$$r = \frac{p(\theta^*|\boldsymbol{y})}{p(\theta^{(s)}|\boldsymbol{y}} = \left(\frac{\prod_{i=1}^n \Phi(y_i, \theta^*, \sigma)}{\prod_{i=1}^n \Phi(y_i, \theta^{(s)}, \sigma)}\right) \times \left(\frac{\Phi(\theta^*, \mu, \tau)}{\Phi(\theta^{(s)}, \mu, \tau)}\right)$$

- Often, the above calculation is numerically unstable and may compute logarithm instead

$$\log r = \sum_{i=1}^n [\log \Phi(y_i, \theta^*, \sigma) - \log \Phi(y_i, \theta^{(s)}, \sigma)] + [\log \Phi(\theta^*, \mu, \tau) - \log \Phi(\theta^{(s)}, \mu, \tau)].$$

- On the log scale, proposal is accepted if $\log U(0,1) < \log r$

Python Code

```
import numpy as np
import scipy.stats as stats
s2 = 1
t2 = 10
mu = 5
y = np.array([9.37, 10.18, 9.16, 11.60, 10.33])
theta = 0
delta2 = 2
S = 10000
THETA = []
```

Python Code (cont.)

```
for s in range(S):
    theta_star = np.random.normal(theta, np.sqrt(delta2), 1)
    log_r =
     (np.sum(np.log(stats.norm(theta_star, np.sqrt(s2)).pdf(y))) -
      np.sum(np.log(stats.norm(theta, np.sqrt(s2)).pdf(y)))) +
      (np.log(stats.norm(mu, np.sqrt(t2)).pdf(theta_star)) -
       np.log(stats.norm(mu, np.sqrt(t2)).pdf(theta)))
    if np.log(np.random.random(1)) < log_r:
        theta = theta_star

    THETA.append(theta)

import matplotlib as mpl
import matplotlib.pyplot as plt
plt.style.use('classic')
plt.plot(range(S), THETA)

import seaborn as sns
sns.distplot(THETA, hist = True)
```

Results

12
10
8
6
4
2
0
0
2000
4000
6000
8000
100

1.0
0.8
0.6
0.4
0.2
0.0
−2
0
2
4
6
8
10
12
14

General Property of the Metropolis Algorithm

- Under some mild conditions, the marginal distribution of $\theta^{(S)}$ approximates the posterior distribution $p(\theta|y)$ for large S.

General Property of the Metropolis Algorithm

- Under some mild conditions, the marginal distribution of $\theta^{(S)}$ approximates the posterior distribution $p(\theta|y)$ for large S.
  - For any given value of θ_a,

$$\lim_{S\to\infty} \frac{\#\{\theta\text{'s in the sequence } < \theta_a\}}{S} = p(\theta < \theta_a|y)$$

Tune the Proposal Distribution

- By choosing a proper proposal variance δ^2, we can decrease the correlation in the Markov chain

Tune the Proposal Distribution

- By choosing a proper proposal variance δ^2, we can decrease the correlation in the Markov chain
  - Faster convergence

Tune the Proposal Distribution

- By choosing a proper proposal variance δ^2, we can decrease the correlation in the Markov chain
  - Faster convergence
  - Better mixing

Tune the Proposal Distribution

- By choosing a proper proposal variance δ^2, we can decrease the correlation in the Markov chain
  - Faster convergence
  - Better mixing
  - An increase in the effective sample size

Tune the Proposal Distribution

- By choosing a proper proposal variance δ^2, we can decrease the correlation in the Markov chain
  - Faster convergence
  - Better mixing
  - An increase in the effective sample size
- Optimal proposal variance is neither too large nor too small

Tune the Proposal Distribution

- By choosing a proper proposal variance δ^2, we can decrease the correlation in the Markov chain
  - Faster convergence
  - Better mixing
  - An increase in the effective sample size
- Optimal proposal variance is neither too large nor too small
  - Small δ^2 : θ^* is too close to current value $\theta^{(s)}$. No matter accepted or not, $\theta^{(s+1)}$ is similar to $\theta^{(s)}$. Takes a long time to move away from $\theta^{(s)}$

Tune the Proposal Distribution

- By choosing a proper proposal variance δ^2, we can decrease the correlation in the Markov chain
  - Faster convergence
  - Better mixing
  - An increase in the effective sample size
- Optimal proposal variance is neither too large nor too small
  - Small δ^2 : θ^* is too close to current value $\theta^{(s)}$. No matter accepted or not, $\theta^{(s+1)}$ is similar to $\theta^{(s)}$. Takes a long time to move away from $\theta^{(s)}$
  - Large δ^2: If $\theta^{(s)}$ is close to posterior mode, θ^* is often very far from $\theta^{(s)}$ and leads to rejection. Then the chain gets stuck, because in most iterations $\theta^{(s+1)} = \theta^{(s)}$

Different Proposal Variances

- $\delta^2 = \left\{ \frac{1}{32}, \frac{1}{2}, 2, 32, 64 \right\}$

Different Proposal Variances

- $\delta^2 = \left\{\frac{1}{32}, \frac{1}{2}, 2, 32, 64\right\}$
- Corresponding autocorrelation: $(0.98, 0.77, \mathbf{0.69}, 0.84, 0.86)$

Metropolis Algorithm for Poisson Regression

- Model: $\log E(y_i|x_i) - \beta_1 + \beta_2 x_i + \beta_3 x_i^2$

Metropolis Algorithm for Poisson Regression

- Model: $\log E(y_i|x_i) - \beta_1 + \beta_2 x_i + \beta_3 x_i^2$
- Let $\boldsymbol{x}_i = (1, x_i, x_i^2)$, then $\log E(y_i|x_i) = \boldsymbol{\beta}' \boldsymbol{x}_i$

Metropolis Algorithm for Poisson Regression

- Model: $\log E(y_i|x_i) - \beta_1 + \beta_2 x_i + \beta_3 x_i^2$
- Let $\boldsymbol{x}_i = (1, x_i, x_i^2)$, then $\log E(y_i|x_i) = \boldsymbol{\beta}'\boldsymbol{x}_i$
- Prior distribution: $\boldsymbol{\beta} \sim \mathcal{N}(0, 100\boldsymbol{I})$

Metropolis Algorithm for Poisson Regression

- Model: $\log E(y_i|x_i) - \beta_1 + \beta_2 x_i + \beta_3 x_i^2$
- Let $\boldsymbol{x}_i = (1, x_i, x_i^2)$, then $\log E(y_i|x_i) = \boldsymbol{\beta}'\boldsymbol{x}_i$
- Prior distribution: $\boldsymbol{\beta} \sim \mathcal{N}(0, 100\boldsymbol{I})$
- Acceptance rate:

$$
\begin{aligned}
r &= \frac{p(\boldsymbol{\beta}^*|\boldsymbol{X}, \boldsymbol{y})}{p(\boldsymbol{\beta}^{(s)}|\boldsymbol{X}, \boldsymbol{y})} \\
&= \frac{\prod_{i=1}^n \mathcal{P}(y_i, \boldsymbol{x}_i'\boldsymbol{\beta}^*)}{\prod_{i=1}^n \mathcal{P}(y_i, \boldsymbol{x}_i'\boldsymbol{\beta}^{(s)})} \times \frac{\prod_{j=1}^3 \Phi(\beta_j^*, 0, 10)}{\prod_{j=1}^3 \Phi(\beta_j^{(s)}, 0, 10)}
\end{aligned}
$$

Metropolis Algorithm for Poisson Regression

- Model: $\log E(y_i|x_i) - \beta_1 + \beta_2 x_i + \beta_3 x_i^2$
- Let $\boldsymbol{x}_i = (1, x_i, x_i^2)$, then $\log E(y_i|x_i) = \boldsymbol{\beta}'\boldsymbol{x}_i$
- Prior distribution: $\boldsymbol{\beta} \sim \mathcal{N}(0, 100\boldsymbol{I})$
- Acceptance rate:

$$
\begin{aligned}
r &= \frac{p(\boldsymbol{\beta}^*|\boldsymbol{X}, \boldsymbol{y})}{p(\boldsymbol{\beta}^{(s)}|\boldsymbol{X}, \boldsymbol{y})} \\
&= \frac{\prod_{i=1}^n \mathcal{P}(y_i, \boldsymbol{x}_i'\boldsymbol{\beta}^*)}{\prod_{i=1}^n \mathcal{P}(y_i, \boldsymbol{x}_i'\boldsymbol{\beta}^{(s)})} \times \frac{\prod_{j=1}^3 \Phi(\beta_j^*, 0, 10)}{\prod_{j=1}^3 \Phi(\beta_j^{(s)}, 0, 10)}
\end{aligned}
$$

- Proposal distribution: Multivariate normal

Metropolis Algorithm for Poisson Regression

- Model: $\log E(y_i|x_i) - \beta_1 + \beta_2 x_i + \beta_3 x_i^2$
- Let $\boldsymbol{x}_i = (1, x_i, x_i^2)$, then $\log E(y_i|x_i) = \boldsymbol{\beta}'\boldsymbol{x}_i$
- Prior distribution: $\boldsymbol{\beta} \sim \mathcal{N}(0, 100\boldsymbol{I})$
- Acceptance rate:

$$
\begin{aligned}
r &= \frac{p(\boldsymbol{\beta}^*|\boldsymbol{X}, \boldsymbol{y})}{p(\boldsymbol{\beta}^{(s)}|\boldsymbol{X}, \boldsymbol{y})} \\
&= \frac{\prod_{i=1}^n \mathcal{P}(y_i, \boldsymbol{x}_i'\boldsymbol{\beta}^*)}{\prod_{i=1}^n \mathcal{P}(y_i, \boldsymbol{x}_i'\boldsymbol{\beta}^{(s)})} \times \frac{\prod_{j=1}^3 \Phi(\beta_j^*, 0, 10)}{\prod_{j=1}^3 \Phi(\beta_j^{(s)}, 0, 10)}
\end{aligned}
$$

- Proposal distribution: Multivariate normal
  - Choose the proposal variance similar to posterior variance

Metropolis Algorithm for Poisson Regression

- Model: $\log E(y_i|x_i) - \beta_1 + \beta_2 x_i + \beta_3 x_i^2$
- Let $\boldsymbol{x}_i = (1, x_i, x_i^2)$, then $\log E(y_i|x_i) = \boldsymbol{\beta}'\boldsymbol{x}_i$
- Prior distribution: $\boldsymbol{\beta} \sim \mathcal{N}(0, 100\boldsymbol{I})$
- Acceptance rate:

$$
\begin{aligned}
r &= \frac{p(\boldsymbol{\beta}^*|\boldsymbol{X}, \boldsymbol{y})}{p(\boldsymbol{\beta}^{(s)}|\boldsymbol{X}, \boldsymbol{y})} \\
&= \frac{\prod_{i=1}^n \mathcal{P}(y_i, \boldsymbol{x}_i'\boldsymbol{\beta}^*)}{\prod_{i=1}^n \mathcal{P}(y_i, \boldsymbol{x}_i'\boldsymbol{\beta}^{(s)})} \times \frac{\prod_{j=1}^3 \Phi(\beta_j^*, 0, 10)}{\prod_{j=1}^3 \Phi(\beta_j^{(s)}, 0, 10)}
\end{aligned}
$$

- Proposal distribution: Multivariate normal
  - Choose the proposal variance similar to posterior variance
  - From OLS, we know that $var(\hat{\beta}) = \sigma^2(X'X)^{-1}$

Metropolis Algorithm for Poisson Regression

- Model: $\log E(y_i|x_i) - \beta_1 + \beta_2 x_i + \beta_3 x_i^2$
- Let $\boldsymbol{x}_i = (1, x_i, x_i^2)$, then $\log E(y_i|x_i) = \boldsymbol{\beta}'\boldsymbol{x}_i$
- Prior distribution: $\boldsymbol{\beta} \sim \mathcal{N}(0, 100\boldsymbol{I})$
- Acceptance rate:

$$
\begin{aligned}
r &= \frac{p(\boldsymbol{\beta}^*|\boldsymbol{X}, \boldsymbol{y})}{p(\boldsymbol{\beta}^{(s)}|\boldsymbol{X}, \boldsymbol{y})} \\
&= \frac{\prod_{i=1}^n \mathcal{P}(y_i, \boldsymbol{x}_i'\boldsymbol{\beta}^*)}{\prod_{i=1}^n \mathcal{P}(y_i, \boldsymbol{x}_i'\boldsymbol{\beta}^{(s)})} \times \frac{\prod_{j=1}^3 \Phi(\beta_j^*, 0, 10)}{\prod_{j=1}^3 \Phi(\beta_j^{(s)}, 0, 10)}
\end{aligned}
$$

- Proposal distribution: Multivariate normal
  - Choose the proposal variance similar to posterior variance
  - From OLS, we know that $var(\hat{\beta}) = \sigma^2(X'X)^{-1}$
  - Construct $\hat{\sigma}^2$ using the sample variance of $\{\log(y_1 + \frac{1}{2}, \ldots, \log(y_n + \frac{1}{2})\}$

Result

- Acceptance rate: 43%

After thinning (every 10th observation)

Result

p(β2|y)
0.0
0.4
0.8
1.2
−1
0
1
2
β2
p(β3|y)
0
1
2
3
4
5
6
7
−0.6
−0.2
0.2
0.4
β3
number of offspring
0
1
2
3
1
2
3
4
5
6
age

The Metropolis-Hastings (M-H) Algorithm

- Consider to sample from a bivariate distribution $p_0(u, v)$
- Gibbs sampler

1. update U: sample $u^{(s+1)} \sim p_0(u|v^{(s)})$;
2. update V: sample $v^{(s+1)} \sim p_0(v|u^{(s+1)})$.

1. Proposal: full conditional distribution
2. Always accept

- The Metropolis algorithm

1. update U:
   a) sample $u^* \sim J_u(u|u^{(s)})$;
   b) compute $r = p_0(u^*, v^{(s)})/p_0(u^{(s)}, v^{(s)})$;
   c) set $u^{(s+1)}$ to u^* or $u^{(s)}$ with probability $\min(1, r)$ and $\max(0, 1 - r)$.
2. update V:
   a) sample $v^* \sim J_v(v|v^{(s)})$;
   b) compute $r = p_0(u^{(s+1)}, v^*)/p_0(u^{(s+1)}, v^{(s)})$;
   c) set $v^{(s+1)}$ to v^* or $v^{(s)}$ with probability $\min(1, r)$ and $\max(0, 1 - r)$.

Symmetric proposal distribution

The Metropolis-Hastings (M-H) Algorithm (cont.)

- Arbitrary proposal

1. update U:
   a) sample $u^* \sim J_u(u|u^{(s)}, v^{(s)})$;
   b) compute the acceptance ratio

$$r = \frac{p_0(u^*, v^{(s)})}{p_0(u^{(s)}, v^{(s)})} \times \frac{J_u(u^{(s)}|u^*, v^{(s)})}{J_u(u^*|u^{(s)}, v^{(s)})};$$

   c) set $u^{(s+1)}$ to u^* or $u^{(s)}$ with probability $\min(1, r)$ and $\max(0, 1-r)$.
2. update V:
   a) sample $v^* \sim J_v(v|u^{(s+1)}, v^{(s)})$;
   b) compute the acceptance ratio

$$r = \frac{p_0(u^{(s+1)}, v^*)}{p_0(u^{(s+1)}, v^{(s)})} \times \frac{J_v(v^{(s)}|u^{(s+1)}, v^*)}{J_v(v^*|u^{(s+1)}, v^{(s)})};$$

   c) set $v^{(s+1)}$ to v^* or $v^{(s)}$ with probability $\min(1, r)$ and $\max(0, 1-r)$.

Gibbs Sampler as the M-H Algorithm

- If the proposal distribution is $J_u(u^*|u^{(s)}, v^{(s)}) = p_0(u^*|v^{(s)})$ in the M-H algorithm, the acceptance ratio is

$$
\begin{aligned}
r &= \frac{p_0(u^*, v^{(s)})}{p_0 u^{(s)}, v^{(s)}} \times \frac{J_u(u^{(s)}|u^*, v^{(s)})}{J_u(u^*|u^{(s)}, v^{(s)})} \\
&= \frac{p_0(u^*, v^{(s)})}{p_0 u^{(s)}, v^{(s)}} \frac{p_0(u^{(s)}|v^{(s)})}{p_0(u^*|v^{(s)})} \\
&= \frac{p_0(u^*|v^{(s)}) p_0(v^{(s)})}{p_0(u^{(s)}|v^{(s)}) p_0(v^{(s)})} \frac{p_0(u^{(s)}|v^{(s)})}{p_0(u^*|v^{(s)})} \\
&= \frac{p_0(v^{(s)})}{p_0(v^{(s)})} = 1
\end{aligned}
$$

Requirement for the Proposal Distribution

- The proposal distribution does not depend on values in the sequence previous to $x^{(s)} \rightarrow$ Markov property

Requirement for the Proposal Distribution

- The proposal distribution does not depend on values in the sequence previous to $x^{(s)} \rightarrow$ Markov property
- Regardelss where the chain started, every value x with $p_0(x) > 0$ will eventually be proposed $\rightarrow$ irreducible

Requirement for the Proposal Distribution

- The proposal distribution does not depend on values in the sequence previous to $x^{(s)} \rightarrow$ Markov property
- Regardelss where the chain started, every value x with $p_0(x) > 0$ will eventually be proposed $\rightarrow$ irreducible
- Aperiodic: A Markov chain lacking any periodic states is called aperiodic

Requirement for the Proposal Distribution

- The proposal distribution does not depend on values in the sequence previous to $x^{(s)} \rightarrow$ Markov property
- Regardelss where the chain started, every value x with $p_0(x) > 0$ will eventually be proposed $\rightarrow$ irreducible
- Aperiodic: A Markov chain lacking any periodic states is called aperiodic
  - A value x is periodic with period $k > 1$ in a Markov chain if it can only be visited every kth iteration

Requirement for the Proposal Distribution

- The proposal distribution does not depend on values in the sequence previous to $x^{(s)} \rightarrow$ Markov property
- Regardelss where the chain started, every value x with $p_0(x) > 0$ will eventually be proposed $\rightarrow$ irreducible
- Aperiodic: A Markov chain lacking any periodic states is called aperiodic
  - A value x is periodic with period $k > 1$ in a Markov chain if it can only be visited every kth iteration
- Recurrent: A value x is said to be recurrent if, when we continue to run the Markov chain from x, we are guaranteed to eventually return to x. And we want all of the possible values of x to be recurrent in our Markov chain

Ergodic Theorem

- If $\{x^{(1)}, x^{(2)}, \ldots\}$ is an irreducible, aperiodic, and recurrent Markov chain, then there is a unique probability distribution π such that as $s \to \infty$,

$$Pr(x^{(s)} \in A) \to \pi(A) \text{ for any set } A;$$
$$\frac{1}{S}\sum g(x^{(s)}) \to \int g(x)\pi(x)dx.$$

Ergodic Theorem

- If $\{x^{(1)}, x^{(2)}, \ldots\}$ is an irreducible, aperiodic, and recurrent Markov chain, then there is a unique probability distribution π such that as $s \to \infty$,

$$Pr(x^{(s)} \in A) \to \pi(A) \text{ for any set } A;$$
$$\frac{1}{S}\sum g(x^{(s)}) \to \int g(x)\pi(x)dx.$$

- The distribution π is called the stationary distribution because it has the following property: If $x^{(s)} \sim \pi$, and $x^{(s+1)}$ is generated from the Markov chain starting at $x^{(s)}$, then $Pr(x^{(s+1)} \in A) = \pi(A)$.

Ergodic Theorem

- If $\{x^{(1)}, x^{(2)}, \ldots\}$ is an irreducible, aperiodic, and recurrent Markov chain, then there is a unique probability distribution π such that as $s \to \infty$,

$$Pr(x^{(s)} \in A) \to \pi(A) \text{ for any set } A;$$
$$\frac{1}{S}\sum g(x^{(s)}) \to \int g(x)\pi(x)dx.$$

- The distribution π is called the stationary distribution because it has the following property: If $x^{(s)} \sim \pi$, and $x^{(s+1)}$ is generated from the Markov chain starting at $x^{(s)}$, then $Pr(x^{(s+1)} \in A) = \pi(A)$.
- $\to$ Once you are sampling from the stationary distribution, you are always sampling from the stationary distribution

Combining the Metropolis and Gibbs Algorithms

- In complex models it is often the case that conditional distributions are available for some parameters but not others

Combining the Metropolis and Gibbs Algorithms

- In complex models it is often the case that conditional distributions are available for some parameters but not others
- In these situations we can combine Gibbs and Metropolis-type proposal distributions to generate a Markov chain to approximate the joint posterior distribution of all the parameters

A Regression Model with Correlated Errors

- Ordinary linear regression

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \sim \text{multivariate normal}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}).$$

- Introducing AR(1) structure

$$\Sigma = \sigma^2\mathbf{C}_\rho = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & & \\ \vdots & \vdots & & \ddots & \\ \rho^{n-1} & \rho^{n-2} & & & 1 \end{pmatrix}$$

MCMC

- Prior: $\beta \sim N(\beta_0, \Sigma_0)$ and $\sigma^{-2} \sim gamma(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2})$
- If ρ is known, a Gibbs sampler is available

$$\{\boldsymbol{\beta}|\mathbf{X}, \boldsymbol{y}, \sigma^2, \rho\} \sim \text{multivariate normal}(\boldsymbol{\beta}_n, \Sigma_n) \ , \ \text{where}$$
$$\Sigma_n = (\mathbf{X}^T \mathbf{C}_\rho^{-1} \mathbf{X}/\sigma^2 + \Sigma_0^{-1})^{-1}$$
$$\boldsymbol{\beta}_n = \Sigma_n (\mathbf{X}^T \mathbf{C}_\rho^{-1} \boldsymbol{y}/\sigma^2 + \Sigma_0^{-1} \boldsymbol{\beta}_0) \ , \ \text{and}$$

$$\{\sigma^2|\mathbf{X}, \boldsymbol{y}, \boldsymbol{\beta}, \rho\} \sim \text{inverse-gamma}([\nu_0 + n]/2, [\nu_0 \sigma_0^2 + \text{SSR}_\rho]/2) \ , \ \text{where}$$
$$\text{SSR}_\rho = (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{C}_\rho^{-1} (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}).$$

- But ρ is unknown and the full conditional distribution of ρ is not so simple

MCMC (cont.)

1. Update $\boldsymbol{\beta}$: Sample $\boldsymbol{\beta}^{(s+1)} \sim$ multivariate normal$(\boldsymbol{\beta}_n, \Sigma_n)$, where $\boldsymbol{\beta}_n$ and Σ_n depend on $\sigma^{2(s)}$ and $\rho^{(s)}$.
2. Update σ^2: Sample $\sigma^{2(s+1)} \sim$ inverse-gamma$([\nu_0+n]/2, [\nu_0\sigma_0^2+\text{SSR}_\rho]/2)$, where SSR_ρ depends on $\boldsymbol{\beta}^{(s+1)}$ and $\rho^{(s)}$.
3. Update ρ:
   a) Propose $\rho^* \sim$ uniform$(\rho^{(s)} - \delta, \rho^{(s)} + \delta)$. If $\rho^* < 0$ then reassign it to be $|\rho^*|$. If $\rho^* > 1$ reassign it to be $2 - \rho^*$.
   b) Compute the acceptance ratio

$$r = \frac{p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{\beta}^{(s+1)}, \sigma^{2(s+1)}, \rho^*)p(\rho^*)}{p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{\beta}^{(s+1)}, \sigma^{2(s+1)}, \rho^{(s)})p(\rho^{(s)})}$$

   and sample $u \sim$ uniform(0,1). If $u < r$ set $\rho^{(s+1)} = \rho^*$, otherwise set $\rho^{(s+1)} = \rho^{(s)}$.

This proposal distribution is called reflecting random walk, which ensures $0 < \rho < 1$

- It is symmetric

Example: Historical CO2 and Temperature Data

- "diffuse" prior:
  - $\beta \sim N(\beta_0, \Sigma_0)$ and $\sigma^{-2} \sim gamma(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2})$ with $\beta_0 = 0, \Sigma_0 = diag(1000), \nu_0 = 1, \sigma_0^2 = 1$
  - $\rho \sim U(0,1)$

Before Thinning

After Thinning (every 25th value)

Posterior

posterior marginal density
0
10
20
30
0.00
0.02
0.04
0.06
β_2

GLS estimate
OLS estimate
temperature
−8
−6
−4
−2
0
2
180
200
220
240
260
280
CO_2

Process the MCMC Output

- Burn-in

Process the MCMC Output

- Burn-in
- Thinning

Process the MCMC Output

- Burn-in
- Thinning
- Monitoring convergence

Process the MCMC Output

- Burn-in
- Thinning
- Monitoring convergence
- Tools

Process the MCMC Output

- Burn-in
- Thinning
- Monitoring convergence
- Tools
  - Traceplot

Process the MCMC Output

- Burn-in
- Thinning
- Monitoring convergence
- Tools
  - Traceplot
  - ACF plot

Process the MCMC Output

- Burn-in
- Thinning
- Monitoring convergence
- Tools
  - Traceplot
  - ACF plot
  - Acceptance (or rejection) rate for M-H algorithm

Process the MCMC Output

- Burn-in
- Thinning
- Monitoring convergence
- Tools
  - Traceplot
  - ACF plot
  - Acceptance (or rejection) rate for M-H algorithm
  - Quantitative procedures

Burn-In

Need for a burn-in (the "pigtail")

Mixing

- MCMC may have poor mixing when
  - autocorrelation is high
  - it takes a long time between jumps to different parts of the parameter space
  - Looking at the traceplot

- Poor mixing

- Well mixing

Gelman and Rubin (1992): Potential Scale Reduction Factor

- Based on normal theory approximation to exact Bayesian posterior inference

Gelman and Rubin (1992): Potential Scale Reduction Factor

- Based on normal theory approximation to exact Bayesian posterior inference
- Create an overdispersed estimate of the target distribution and use it to start several independent sequences

Gelman and Rubin (1992): Potential Scale Reduction Factor

- Based on normal theory approximation to exact Bayesian posterior inference
- Create an overdispersed estimate of the target distribution and use it to start several independent sequences
- For each scalar quantity of interest, after running the sampler chains estimate the factor by which the scale parameter might shrink if sampling were continued indefinitely

Gelman and Rubin (1992): Potential Scale Reduction Factor

- Based on normal theory approximation to exact Bayesian posterior inference
- Create an overdispersed estimate of the target distribution and use it to start several independent sequences
- For each scalar quantity of interest, after running the sampler chains estimate the factor by which the scale parameter might shrink if sampling were continued indefinitely
- If close to one, within-chain variance dominates between-chain variance

Gelman and Rubin (1992): Potential Scale Reduction Factor

- Based on normal theory approximation to exact Bayesian posterior inference
- Create an overdispersed estimate of the target distribution and use it to start several independent sequences
- For each scalar quantity of interest, after running the sampler chains estimate the factor by which the scale parameter might shrink if sampling were continued indefinitely
- If close to one, within-chain variance dominates between-chain variance
- This is called $\hat{R}$

About Convergence Diagnostics

- There are a lot of them

About Convergence Diagnostics

- There are a lot of them
- Can not prove the chain converges

About Convergence Diagnostics

- There are a lot of them
- Can not prove the chain converges
- Be cautious when using these diagnostics

About Convergence Diagnostics

- There are a lot of them
- Can not prove the chain converges
- Be cautious when using these diagnostics
- Use a variety of diagnosis tools rather than any single one

About Convergence Diagnostics

- There are a lot of them
- Can not prove the chain converges
- Be cautious when using these diagnostics
- Use a variety of diagnosis tools rather than any single one
- Learn as much as possible about the target density before applying the MCMC algorithm

Hamiltonian MC

- Physical analogy to Hamiltonian MC: Imagine a hockey pluck sliding over a surface without friction, being stopped at some point in time and then kicked again in a random direction

Hamiltonian MC

- Physical analogy to Hamiltonian MC: Imagine a hockey pluck sliding over a surface without friction, being stopped at some point in time and then kicked again in a random direction
- Hamiltonian MC employs the trick developed by nature. Velocity v is added to the parameters describing the system. Energy of the system consists of potential and kinetic parts

Hamiltonian MC

- Physical analogy to Hamiltonian MC: Imagine a hockey pluck sliding over a surface without friction, being stopped at some point in time and then kicked again in a random direction
- Hamiltonian MC employs the trick developed by nature. Velocity v is added to the parameters describing the system. Energy of the system consists of potential and kinetic parts
- velocities v and positions x have independent canonical distributions

Hamiltonian MC

- Physical analogy to Hamiltonian MC: Imagine a hockey pluck sliding over a surface without friction, being stopped at some point in time and then kicked again in a random direction
- Hamiltonian MC employs the trick developed by nature. Velocity v is added to the parameters describing the system. Energy of the system consists of potential and kinetic parts
- velocities v and positions x have independent canonical distributions
- So once we can sample from joint distribution $p(x, v)$, we also can sample x by ignoring computed velocities v

Pros and Cons of Hamiltonian MC

- Distances between successive generated points are typically large, so we need less iterations to get representative sampling

Pros and Cons of Hamiltonian MC

- Distances between successive generated points are typically large, so we need less iterations to get representative sampling
- price of a single iteration is higher, but HMC is still significantly more efficient

Pros and Cons of Hamiltonian MC

- Distances between successive generated points are typically large, so we need less iterations to get representative sampling
- price of a single iteration is higher, but HMC is still significantly more efficient
- Hamiltonian MC in most cases accepts new states

Pros and Cons of Hamiltonian MC

- Distances between successive generated points are typically large, so we need less iterations to get representative sampling
- price of a single iteration is higher, but HMC is still significantly more efficient
- Hamiltonian MC in most cases accepts new states
- HMC has problems with sampling from distributions with isolated local minimums puck doesn't have enough energy to jump from the first minimum to the second over the energy barrier