Bayesian Data Analysis

Modeling and Computing

Yang Chen

Department of Statistics

University of Michigan

ychenang@umich.edu

July 11, 2019

Plan

Course Participants

Tell me about yourself.

Course Participants

Tell me about yourself.

And what you know about Bayesian inference.

What to expect from this tutorial?

What to expect from this tutorial?

1. Basics of Bayesian Modeling
   - Examples of Bayesian Inference

What to expect from this tutorial?

1. Basics of Bayesian Modeling
    - Examples of Bayesian Inference
2. Computational algorithms in Bayesian statistics
    - MCMC and Stan

What to expect from this tutorial?

1. Basics of Bayesian Modeling
   - Examples of Bayesian Inference
2. Computational algorithms in Bayesian statistics
   - MCMC and Stan

Want to learn more?

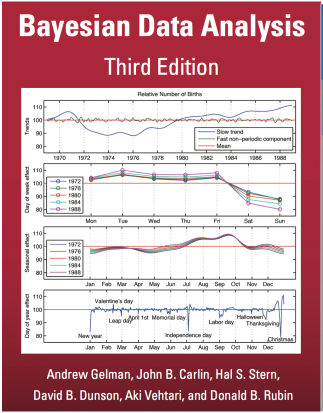
What to expect from this tutorial?

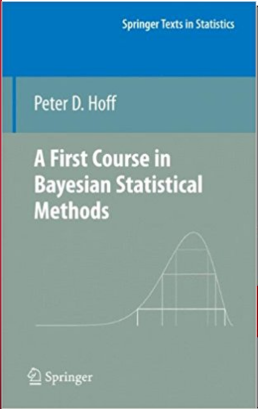
1. Basics of Bayesian Modeling
    - Examples of Bayesian Inference
2. Computational algorithms in Bayesian statistics
    - MCMC and Stan

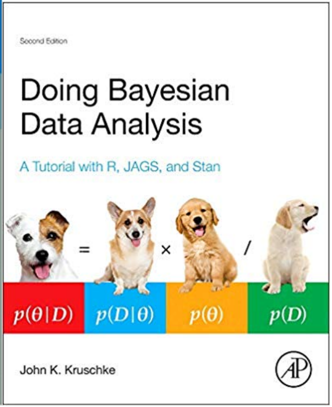
Want to learn more?

- STATS 451 (2019 Fall), STATS 551 (2020 Winter)

Major References







Plan

1. Introduction

2. Basics of Bayesian Modeling
   - What is Bayesian Data Analysis?
   - Bayesian Data Analysis Recipe
   - Bayesian Data Analysis: Examples

3. Computational Algorithms
   - Markov Chain Monte Carlo
   - Approximate Computation

Plan

Bayesian Data Analysis

Quantities we observe

Data.

Bayesian Data Analysis

Quantities we observe

Data.

Quantities we wish to learn

Parameters.

Bayesian Data Analysis

Quantities we observe	Quantities we wish to learn
Data.	Parameters.

- Statistical inference:

  estimate unknown (parameters) from known (data).

Bayesian Data Analysis

Quantities we observe	Quantities we wish to learn
Data.	Parameters.

- Statistical inference:

  estimate unknown (parameters) from known (data).

- Bayesian methods:

  quantify uncertainty in statistical inferences.

Bayesin Inference

- The process of ‘inductive thinking’ via Bayes’ rule.

Bayesin Inference

- The process of 'inductive thinking' via Bayes' rule.
- Bayesian methods provide
  - models for rational, quantitative learning
  - estimators that work for small and large sample sizes
  - methods for generating statistical procedures in complicated problems

Plan

Three Steps of Bayesian Analysis

- Setting up a full probability model
    - Joint probability distribution of all observed & unobserved quantities
- Conditioning on observed: posterior distribution
    - Conditional probability distribution of unobserved given observed
- Evaluating fitting & Interpreting posterior distributions.

Example: estimating 5-year survival probability of a new drug.

Three Steps of Bayesian Analysis

- Setting up a full probability model
  - Joint probability distribution of all observed & unobserved quantities

Three Steps of Bayesian Analysis

- Setting up a full probability model
  - Joint probability distribution of all observed & unobserved quantities

Clinical trial example

For $1 \leq i \leq n$, $y_i = 1$ if alive and 0 otherwise. θ is probability of survival.

Three Steps of Bayesian Analysis

- Setting up a full probability model
  - Joint probability distribution of all observed & unobserved quantities

Clinical trial example

For $1 \leq i \leq n$, $y_i = 1$ if alive and 0 otherwise. θ is probability of survival.

Mathematically

$$y_i \overset{i.i.d.}{\sim} p(\cdot|\theta), 1 \leq i \leq n; \quad \theta \sim p(\cdot).$$

$p(\cdot|\theta)$: conditional probability density (distribution); $p(\cdot)$: marginal distribution. Same notation for continuous & discrete densities.

Three Steps of Bayesian Analysis

Numerical formulation of joint beliefs about y and θ:

Three Steps of Bayesian Analysis

Numerical formulation of joint beliefs about y and θ:

- For each θ, our prior distribution $p(\theta)$ describes our belief that θ represents the true population characteristics.

Three Steps of Bayesian Analysis

Numerical formulation of joint beliefs about y and θ:

- For each θ, our prior distribution $p(\theta)$ describes our belief that θ represents the true population characteristics.
- For each θ and y, our sampling model $p(y|\theta)$ describes our belief that y would be the outcome of our study if we knew θ to be true.

Three Steps of Bayesian Analysis

Numerical formulation of joint beliefs about y and θ:

- For each θ, our prior distribution $p(\theta)$ describes our belief that θ represents the true population characteristics.
- For each θ and y, our sampling model $p(y|\theta)$ describes our belief that y would be the outcome of our study if we knew θ to be true.

Once we obtain data y, the last step is to update our beliefs about θ:

$$\dots\dots$$

Three Steps of Bayesian Analysis

- Conditioning on observed: posterior distribution
  - Conditional probability distribution of unobserved given observed

Three Steps of Bayesian Analysis

- Conditioning on observed: posterior distribution
    - Conditional probability distribution of unobserved given observed

Bayes Rule

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)},$$

where $p(y) = \sum_\theta p(\theta)p(y|\theta)$, $y = \{y_1, y_2, \cdots\}$.

Three Steps of Bayesian Analysis

- Conditioning on observed: posterior distribution
  - Conditional probability distribution of unobserved given observed

Bayes Rule

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)},$$

where $p(y) = \sum_\theta p(\theta)p(y|\theta)$, $y = \{y_1, y_2, \cdots\}$.

Unnormalized posterior density

$$p(\theta|y) \propto p(\theta)p(y|\theta).$$

Three Steps of Bayesian Analysis

- Conditioning on observed: posterior distribution
  - Conditional probability distribution of unobserved given observed

Bayes Rule

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)},$$

where $p(y) = \sum_\theta p(\theta)p(y|\theta)$, $y = \{y_1, y_2, \cdots\}$.

Unnormalized posterior density

$$p(\theta|y) \propto p(\theta)p(y|\theta).$$

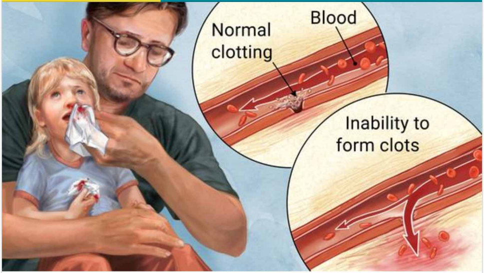
normalizing constant: $p(y)$ — y is observed.

Plan

Example: inference about a genetic status

Hemophilia

ABOUT   SYMPTOMS   TREATMENTS

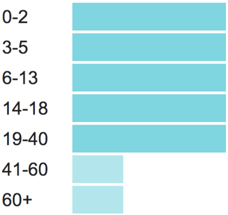


A disorder in which blood doesn't clot normally.

Rare

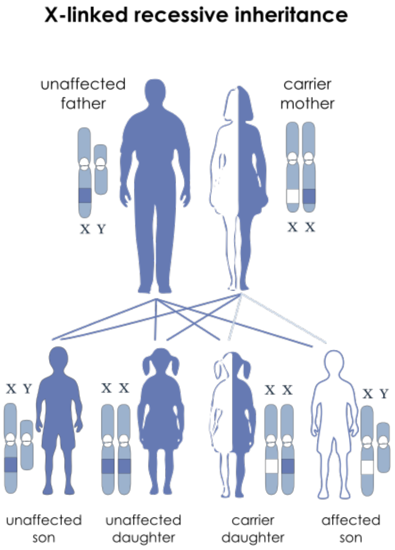
Fewer than 200,000 US cases per year

Ages affected



Genders affected

Example: inference about a genetic status



Humans

Male: XY chromosome.

Female: XX chromosome.

Hemophilia

- Male with the disease-causing gene on X: affected.
- Female with the disease-causing gene on one of two X: not affected.
- Female with the disease-causing gene on both two X: affected.

Example: inference about a genetic status

What is the genetic status of an unaffected woman with an affected brother, an unaffected father and an unaffected mother?

Example: inference about a genetic status

What is the genetic status of an unaffected woman with an affected brother, an unaffected father and an unaffected mother?

Only two possibilities: a carrier of the gene ($\theta = 1$) or not ($\theta = 0$).

Example: inference about a genetic status

What is the genetic status of an unaffected woman with an affected brother, an unaffected father and an unaffected mother?

Only two possibilities: a carrier of the gene ($\theta = 1$) or not ($\theta = 0$).

Prior Distribution

- Affected brother + Unaffected mother $\Rightarrow$ Mother carries one 'good' and one 'bad' gene.

Example: inference about a genetic status

What is the genetic status of an unaffected woman with an affected brother, an unaffected father and an unaffected mother?

Only two possibilities: a carrier of the gene ($\theta = 1$) or not ($\theta = 0$).

Prior Distribution

- Affected brother + Unaffected mother $\Rightarrow$
  Mother carries one 'good' and one 'bad' gene.
- + Unaffected father $\Rightarrow$
  She has fifth-fifty chance of having the gene.

Example: inference about a genetic status

What is the genetic status of an unaffected woman with an affected brother, an unaffected father and an unaffected mother?

Only two possibilities: a carrier of the gene ($\theta = 1$) or not ($\theta = 0$).

Prior Distribution

- Affected brother + Unaffected mother $\Rightarrow$
  Mother carries one 'good' and one 'bad' gene.
- + Unaffected father $\Rightarrow$
  She has fifth-fifty chance of having the gene.
- Prior distribution for θ:

$$P(\theta = 1) = P(\theta = 0) = \frac{1}{2}.$$

Example: inference about a genetic status

Data

Neither of her two sons is affected ($y_1 = y_2 = 0$).

- Two sons: independent and not identical twins.

Example: inference about a genetic status

Data

Neither of her two sons is affected ($y_1 = y_2 = 0$).

- Two sons: independent and not identical twins.

Likelihood

Example: inference about a genetic status

Data

Neither of her two sons is affected ($y_1 = y_2 = 0$).

- Two sons: independent and not identical twins.

Likelihood

$$P(y_1 = y_2 = 0|\theta = 1) = 0.5 \times 0.5 = 0.25,$$
$$P(y_1 = y_2 = 0|\theta = 0) = 1 \times 1 = 1.$$

Example: inference about a genetic status

Posteiror Distribution

Let $y = (y_1, y_2)$, then Bayes' rule gives

Example: inference about a genetic status

Posteiror Distribution

Let $y = (y_1, y_2)$, then Bayes' rule gives

$$P(\theta = 1|y) = \frac{P(y|\theta = 1)P(\theta = 1)}{P(y|\theta = 1)P(\theta = 1) + P(y|\theta = 0)P(\theta = 0)}$$
$$= \frac{0.25 * 0.5}{0.25 * 0.5 + 1 * 0.5} = 0.2,$$

which is smaller than 0.5 (given by the prior).

Example: inference about a genetic status

Posteiror Distribution

Let $y = (y_1, y_2)$, then Bayes' rule gives

$$P(\theta = 1|y) = \frac{P(y|\theta = 1)P(\theta = 1)}{P(y|\theta = 1)P(\theta = 1) + P(y|\theta = 0)P(\theta = 0)}$$
$$= \frac{0.25 * 0.5}{0.25 * 0.5 + 1 * 0.5} = 0.2,$$

which is smaller than 0.5 (given by the prior).

Adding more data

Another unaffected son.

Example: inference about a genetic status

Posteiror Distribution

Let $y = (y_1, y_2)$, then Bayes' rule gives

$$P(\theta = 1|y) = \frac{P(y|\theta = 1)P(\theta = 1)}{P(y|\theta = 1)P(\theta = 1) + P(y|\theta = 0)P(\theta = 0)}$$
$$= \frac{0.25 * 0.5}{0.25 * 0.5 + 1 * 0.5} = 0.2,$$

which is smaller than 0.5 (given by the prior).

Adding more data

Another unaffected son.

$$P(\theta = 1|y_1, y_2, y_3) = \frac{0.5 * 0.2}{0.5 * 0.2 + 1 * 0.8} = 0.111.$$

Example: Estimating a probability from binomial data

- Data: a sequence of 'Bernoulli trials', either 0 or 1.

Example: Estimating a probability from binomial data

- Data: a sequence of 'Bernoulli trials', either 0 or 1.

- Binomial sampling model

$$p(y|\theta) = \left(\begin{array}{c} n \\ y \end{array} \right) \theta^y (1-\theta)^{n-y}.$$

Example: Estimating a probability from binomial data

- Data: a sequence of ‘Bernoulli trials’, either 0 or 1.
- Binomial sampling model

$$p(y|\theta) = \left(\begin{array}{c} n \\ y \end{array} \right) \theta^y (1-\theta)^{n-y}.$$

- Example: estimating probability of female birth.

Example: Estimating a probability from binomial data

- Data: a sequence of 'Bernoulli trials', either 0 or 1.
- Binomial sampling model

$$p(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}.$$

- Example: estimating probability of female birth.
- Prior for θ: uniform on $(0, 1)$.

Example: Estimating a probability from binomial data

- Data: a sequence of 'Bernoulli trials', either 0 or 1.
- Binomial sampling model

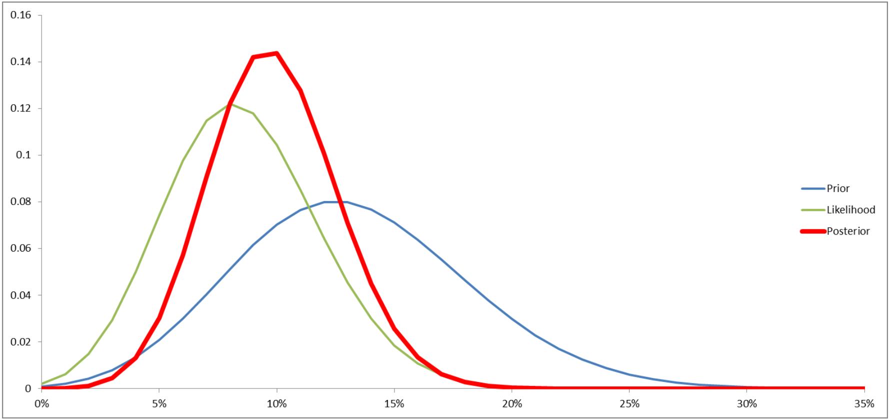
$$p(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}.$$

- Example: estimating probability of female birth.
- Prior for θ: uniform on $(0, 1)$.
- Posterior for θ:

$$\theta|y \sim \text{Beta}(y+1, n-y+1).$$

Example: estimating percentage of Dunkin' lovers

$$\text{Prior Knowledge} + \text{Data} = \text{Current Knowledge}$$



Freq vs. Bayes: Binomial Example

	Frequentist Inference	Bayesian Inference
Estimator	$\hat{\theta} = \frac{y}{n}$ (MLE)	$\frac{y+1}{n+2}$ (Posterior mean)
Variability	$\frac{\hat{\theta}(1-\hat{\theta})}{n}$ (Asymptotically)	$\frac{(y+1)(n-y+1)}{(n+2)^2(n+3)}$ (Posterior variance)
Interval	$[\hat{\theta} \pm 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}]$ $\approx$ Confidence Interval	[a, b] Posterior Interval

Remark: $[a, b]$, s.t. $\int_a^b \frac{\theta^y(1-\theta)^{n-y}}{B(y+1, n-y+1)} = 0.95$.

C.I.: If confidence intervals are constructed using a given confidence level in an infinite number of independent experiments, the proportion of those intervals that contain the true value of the parameter will match the confidence level.

Plan

Bayesian Computation

- Computation of posterior distribution

$$p(\theta|y) = c \cdot p(y|\theta)p(\theta) \propto p(y|\theta)p(\theta),$$

where $c = \int p(y|\theta)p(\theta)d\theta$.

  - Posterior inference: quantities which are functions of θ.

Bayesian Computation

- Computation of posterior distribution

$$p(\theta|y) = c \cdot p(y|\theta)p(\theta) \propto p(y|\theta)p(\theta),$$

where $c = \int p(y|\theta)p(\theta)d\theta$.

  - Posterior inference: quantities which are functions of θ.

- Computation of posterior predictive distribution

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta.$$

Bayesian computation – sampling from unnormalized densities, $p(\theta|y)$.

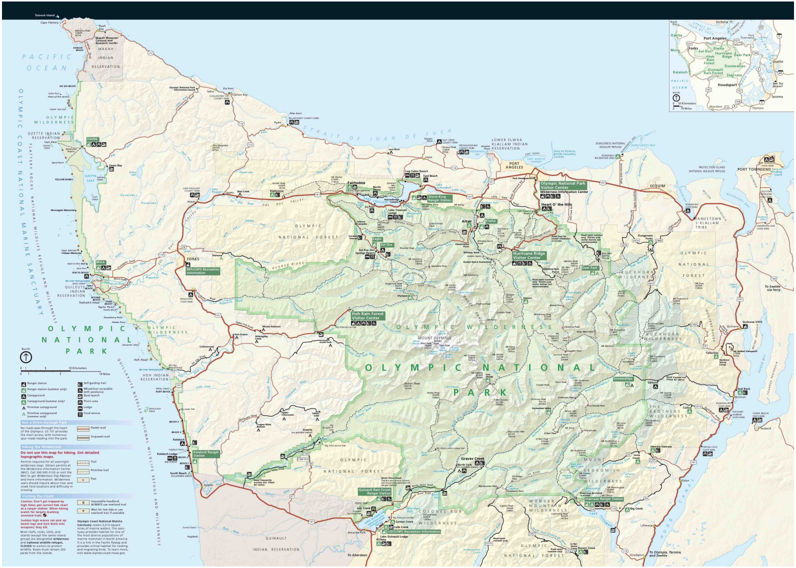
Bayesian Computation

Where is the mountain?

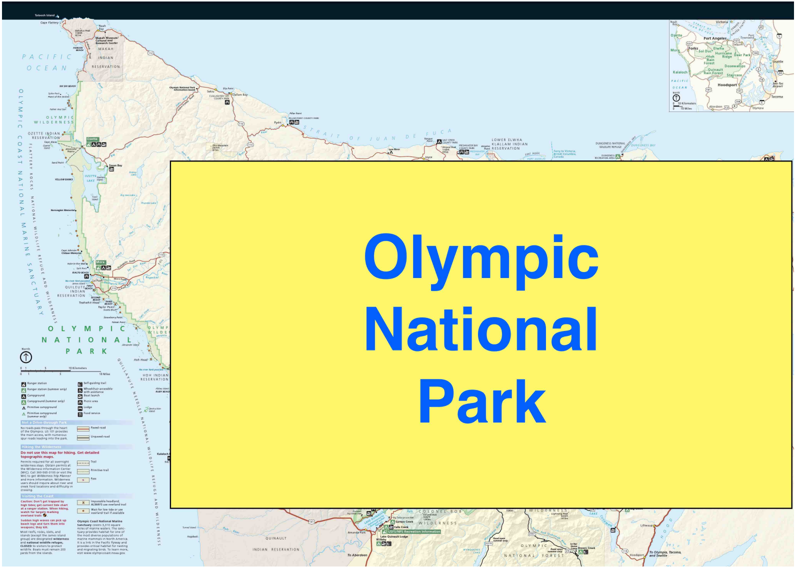
Bayesian Computation

Explore the mountain.

If you have a map – Importance Sampling.



If you DON't have a map.



But you have Time and Patience.

Plan

Introduction to MCMC

- Markov chain simulation (MCMC)
  - sampling from an arbitrary distribution
  - sequentially draw samples (Markov chain)

Introduction to MCMC

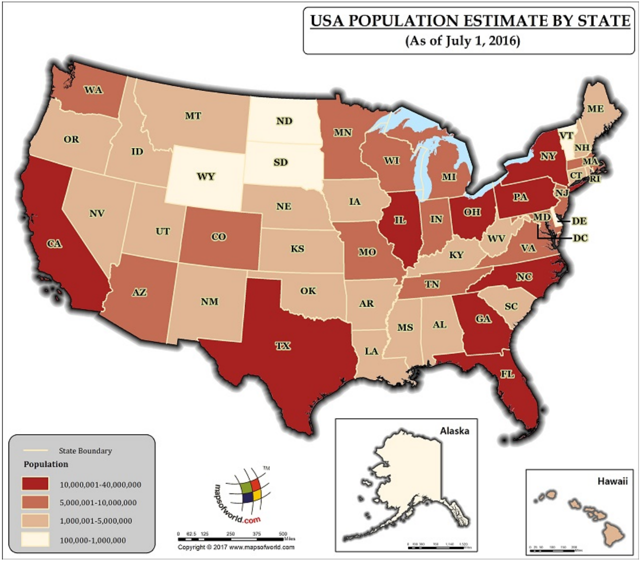
- Markov chain simulation (MCMC)
    - sampling from an arbitrary distribution
    - sequentially draw samples (Markov chain)
- Use when: not possible/efficient to sample from $p(\theta|y)$ directly.

Introduction to MCMC

- Markov chain simulation (MCMC)
  - sampling from an arbitrary distribution
  - sequentially draw samples (Markov chain)
- Use when: not possible/efficient to sample from $p(\theta|y)$ directly.
- Key: Markov process with stationary distribution $p(\theta|y)$.

Example of MCMC with Election Campaign

Example of MCMC with Election Campaign



Now that we are hiking, one step at a time...

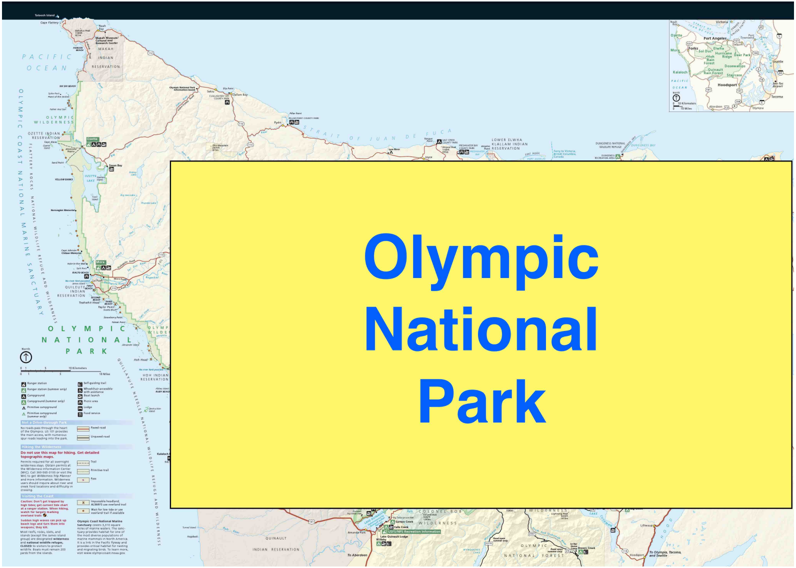
Markov Chains: Why Walk When You Can Flow?

Abstract: *If you are still using a Gibbs sampler, you are working too hard for too little result. Newer, better algorithms trade random walks for frictionless flow.*

In 1989, Depeche Mode was popular, the first version of Microsoft Office was released, large demonstrations brought down the wall separating East and West Germany, and a group of statisticians in the United Kingdom dreamed of Markov chains on the desktop. In 1997, they succeeded, with the first public release of a Windows implementation of **Bayesian inference Using Gibbs Sampling**, **BUGS**.

Click Here to Show Animation.

If you DON't have a map.



And you DON'T have Time or Patience.

And you still want to ROUGHLY SEE the mountain.

Plan

Variational Inference: Introduction

- a method from machine learning

Variational Inference: Introduction

- a method from machine learning
- approximates probability densities through optimization

Variational Inference: Introduction

- a method from machine learning
- approximates probability densities through optimization
- faster than classical methods, such as MCMC

Variational Inference: Introduction

- a method from machine learning
- approximates probability densities through optimization
- faster than classical methods, such as MCMC
- Idea:

Variational Inference: Introduction

- a method from machine learning
- approximates probability densities through optimization
- faster than classical methods, such as MCMC
- Idea:
  - posit a family of densities

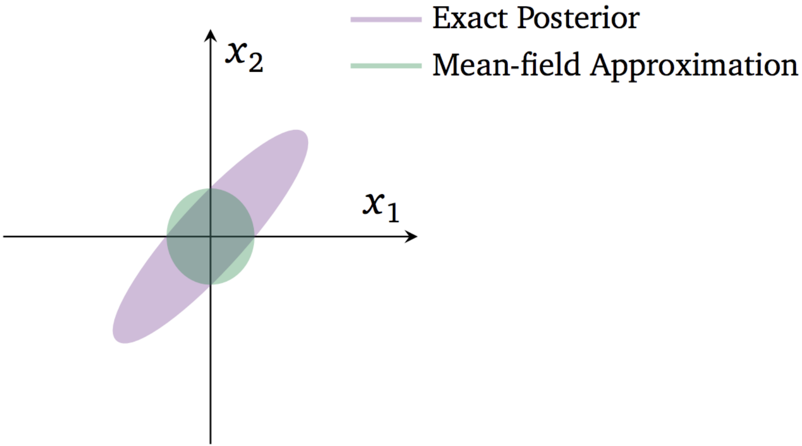
Variational Inference: Introduction

- a method from machine learning
- approximates probability densities through optimization
- faster than classical methods, such as MCMC
- Idea:
    - posit a family of densities
    - find the member of that family closest to the target

Variational Inference: Introduction

- a method from machine learning
- approximates probability densities through optimization
- faster than classical methods, such as MCMC
- Idea:
  - posit a family of densities
  - find the member of that family closest to the target
  - closeness measured by Kullback-Leibler divergence

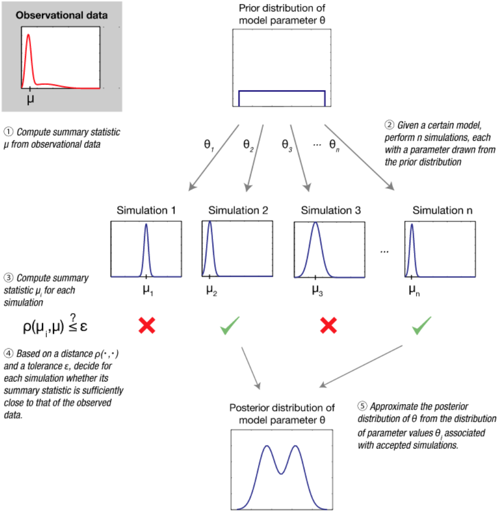
Example: Gaussian mixtures



Approximate Bayesian Computation

- Previous methods
  - Likelihood-based inference
- Complex models
  - an analytical formula might be elusive
  - the likelihood function costly to evaluate
- Simulate data – match observations

Approximate Bayesian Computation



- Bayesian Data Analysis, Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin, Third Edition, CRC Press, 2013 by Chapman and Hall/CRC.
- Hoff, Peter D. A first course in Bayesian statistical methods. Springer Science & Business Media, 2009.
- Berger, James O. Statistical decision theory and Bayesian analysis. Springer Science & Business Media, 2013.
- Robert, Christian, and George Casella. Monte Carlo statistical methods. Springer Science & Business Media, 2013.
- Liu, Jun S. Monte Carlo strategies in scientific computing. Springer Science & Business Media, 2008.
- Brooks, Steve, et al., eds. Handbook of Markov chain Monte Carlo. CRC press, 2011.
- Jackman, Simon. Bayesian analysis for the social sciences. Vol. 846. John Wiley & Sons, 2009.

- For undergrads: STATS 451.
- For graduate students: STATS 551.