SCS 3251 – MCMC

Bayesian Statistics II

Course Roadmap

Module / Week	Title
1	Introduction to Statistics for Data Science
2	Probability
3	Distribution of Random Variables
4	Inference
5	Model Building
6	Linear Regression
7	Multiple & Logistic Regression
8	Guest Speakers & Review
9	Introduction to Bayesian Inference
10	Markov Chain Monte Carlo
11	Multi-level Models
12	Presentations
13	Final Exam

Module 10: Learning Objectives

- Exponential and normal likelihoods
- Non-informative priors
- MonteCarlo Sampling
  - Metropolis
  - Gibbs
- Applications

Bayesian Statistics Framework

1. Identify the data relevant to the research questions
2. Define a descriptive model for the relevant data.
3. Specify a prior distribution on the parameters
4. Use Bayesian inference to re-allocate credibility across parameter values.
5. Check that the posterior predictions mimic the data with reasonable accuracy (i.e., conduct a "posterior predictive check"). If not, then consider a different descriptive model.

Exponential data

- Bus comes in average every 10 minutes

$$Y \sim Exp(\lambda)$$

prior mean 1/lambda. It turns out the Gamma distribution is conjugate of the exponential likelihood. i.e. choose

$$\Gamma(100, 1000)$$

$$prior\ std.\ dev:\ 1/100$$

$$prior\ interval\ : 0.1 \pm 0.02$$

Exponential data (cont)

- Observed data Y = 12

$$
\begin{aligned}
f(\lambda|y) &\propto f(y|\lambda)f(\lambda) \\
&\propto \lambda e^{-\lambda y}\lambda^{\alpha-1}e^{-\beta\lambda} \\
&\propto \lambda^{(\alpha+1)-1}e^{-(\beta+y)\lambda}
\end{aligned}
$$

$$
\begin{aligned}
\lambda|y &\sim \Gamma(\alpha+1, \beta+y) \\
&\sim \Gamma(101, 1012)
\end{aligned}
$$

$$Posterior\ mean = 101/1012 \sim 1/10.02$$

Normal Data with known variance

$$X_i \sim N(\mu, \sigma_o^2)$$

- A normal prior has a normal conjugate posterior

$$prior\ \mu \sim N(m_0, s_o^2)$$

where the posterior

$$f(\mu|\bar{x}) \propto f(\tilde{x}|\mu)f(\mu)$$

$$\mu|\tilde{x} \sim N\left(\frac{\frac{n\bar{x}}{\sigma_0^2}+\frac{m_0}{s_0^2}}{\frac{n}{\sigma_0^2}+\frac{1}{s_0^2}}, \frac{1}{\frac{n}{\sigma_0^2}+\frac{1}{s_0^2}}\right)$$

- The mean can be rewritten as:

$$\frac{n}{n+\frac{\sigma_0^2}{s_0^2}}\bar{x} + \frac{\frac{\sigma_0^2}{s_0^2}}{n+\frac{\sigma_0^2}{s_0^2}}m_0$$

Normal Data - Example

- Suppose a chemist wants to measure the mass of a sample. Her balance has a know standard deviation of 0.2 milligrams
- By looking at the sample she thinks the mass is about 10 milligrams and based on her previous experience her uncertainty on the guess is 2 mlg
- After 5 measurements: data mean = 10.5
- Updating her posterior:
  - Mean = 10.4999
  - Std. Dev = 0.080
- The posterior mean shifted and the uncertainty dropped!

Normal Data with unknown variance

$$X_i|\mu,\sigma^2 \sim N(\mu,\sigma^2)$$

- Express conjugate prior in a hierarchical fashion

$$\text{prior} \rightarrow \mu|\sigma^2 \sim N(m_0, \sigma^2/w), \ \text{where } w = \sigma^2/\sigma_\mu^2$$

$$\text{prior} \rightarrow \sigma^2 \sim \Gamma^{-1}(\alpha,\beta)$$

so, the posterior

$$\sigma^2|\tilde{x} \sim \Gamma^{-1}(\alpha+\frac{n}{2},\beta+\frac{1}{2}\sum_{i=1}^{n}(x_o-\bar{x})^2+\frac{nw}{2(n+w)}(\bar{x}-m)^2)$$

$$\mu|\sigma^2,\tilde{x} \sim N(\frac{w}{n+w}m+\frac{n}{n+w}\bar{x},\frac{\sigma^2}{n+w})$$

In some cases, we really only care about mu. So we can marginalize sigma:

$$\mu|\tilde{x} \sim t-distributed$$

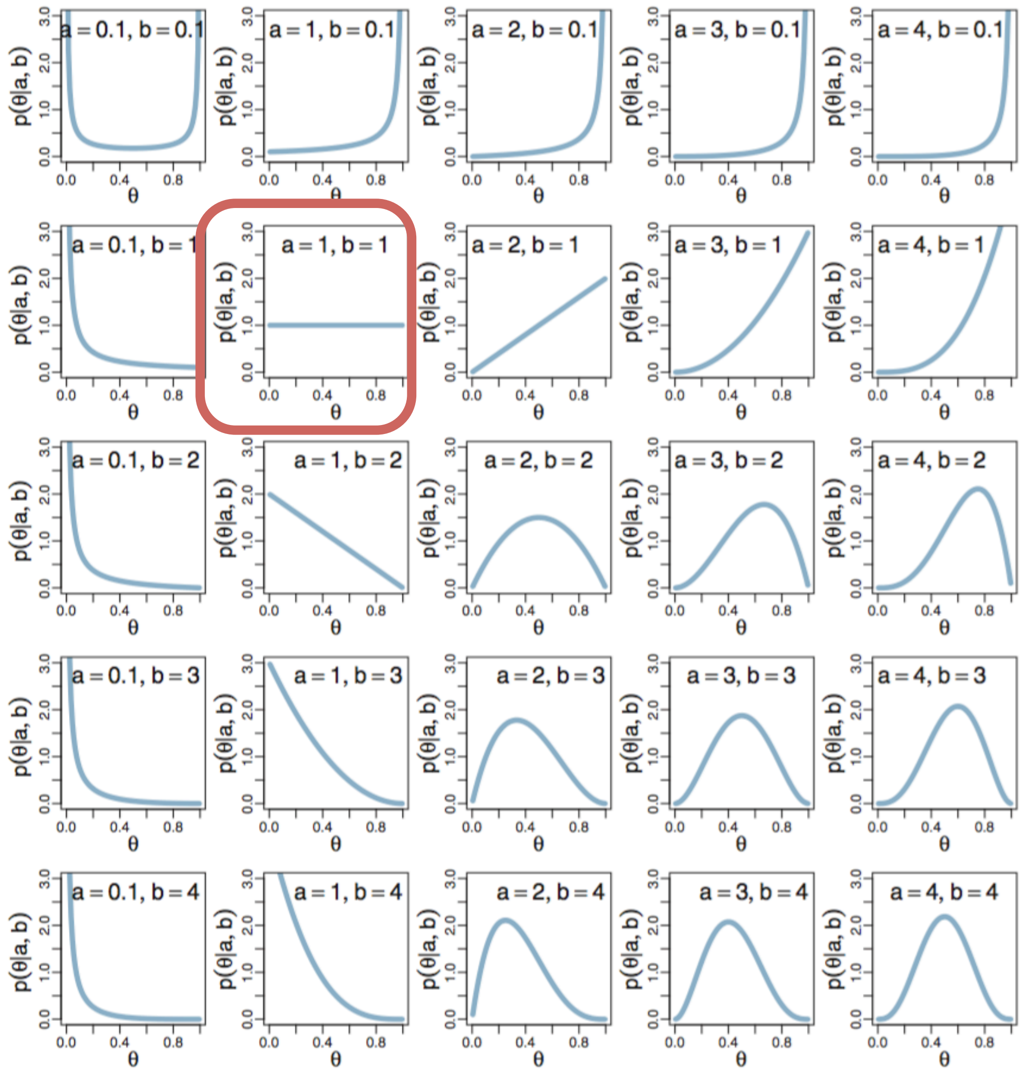
Non-informative priors

- We have seen several approaches to choosing priors. One additional approach is to be as objective as possible:
  - Minimize the amount of information that goes into the prior,
  - Data would have maximum influence on the posterior
- Example: Coin flip example: $Y_i \sim Bernoulli(\theta)$
  - Non-informative prior: $\theta \sim Uniform[0,1] = Beta(1,1)$
  - Effective sample size? Sum of the parameters = 1+1 = 2
  - How can we reduce the information present?

    $$Beta(0.5, 0.5) ... Beta(0.001, 0.001) ... \rightarrow Beta(0,0)$$

    the smaller the information in the prior, the more importance is given to the data.
  - But Beta(0,0) is not a proper prior. Why? $Beta(0,0) \propto \theta^{-1}(1-\theta)^{-1}$
  - Can we do inference with an improper prior?

Beta distribution



Non-informative priors (cont)

- If we collect data, and as long as we observe at least 1 head and 1 tail, we can get a posterior:

$$Beta(y, n-y)$$

$$\text{posterior} \quad \text{mean:} \frac{y}{n} = \hat{\theta}$$

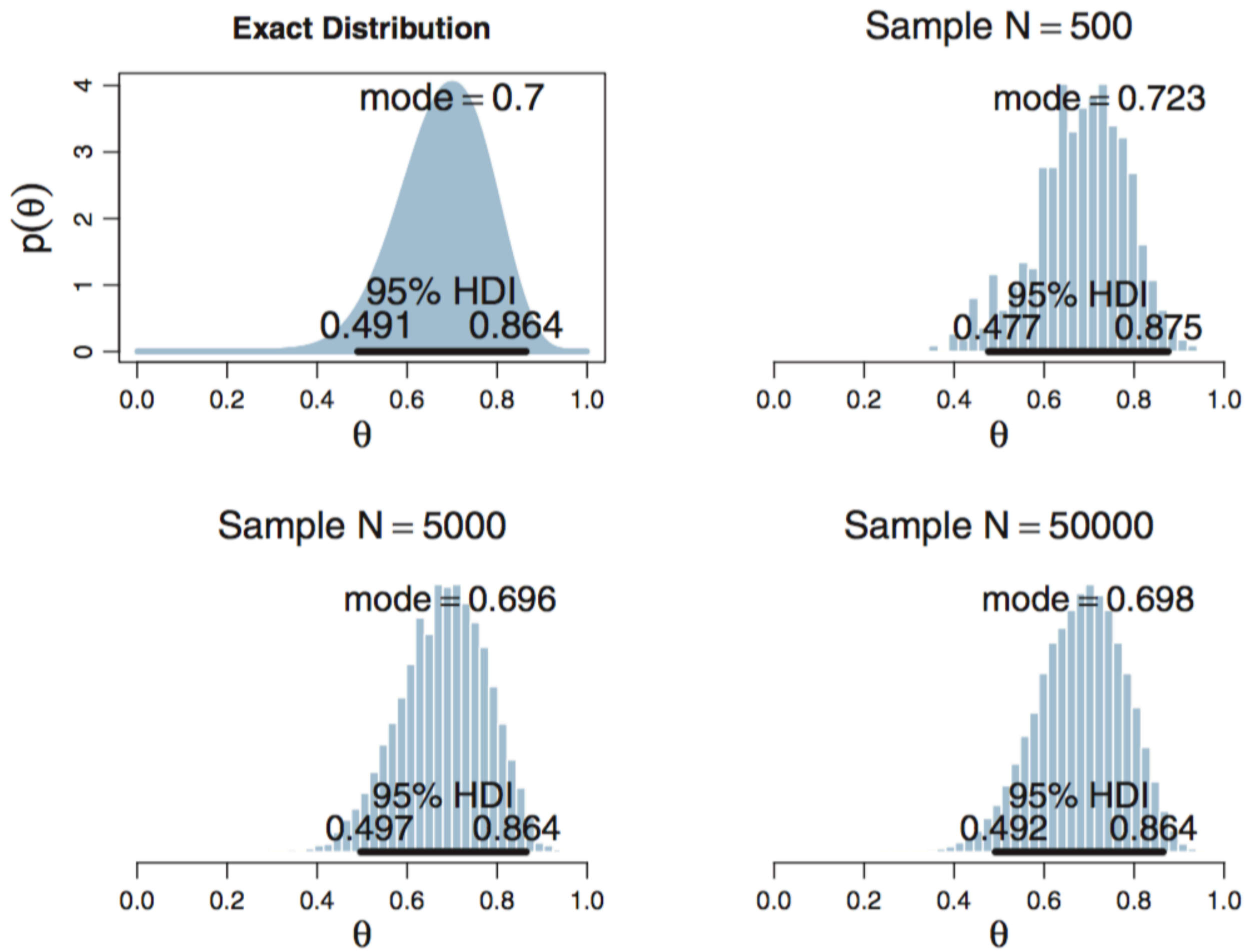
- In simple models, non-informative priors often produce posterior mean and estimates that are equivalent to the common frequentist/MLE estimates.
  - We can still use these prior as long as the posterior distribution is proper.
  - In addition to the point estimate, we can have a posterior distribution for the parameter which allows us to calculate posterior probabilities and credibility integrals

SAMPLING - MCMC

MCMC

- Methods to producing accurate approximations to Bayesian posterior distributions.
- MCMC assumes that the prior and the likelihood are specified by a function that is easily computed
- MCMC results in an approximation of the posterior distribution, $p(\theta|D)$, in the form of large number of θ values sampled from the distribution
- Estimates can be used to calculate expected value, HDI, etc.

Sampling from distribution



John K. Kruschke, Doing Bayesian Data Analysis

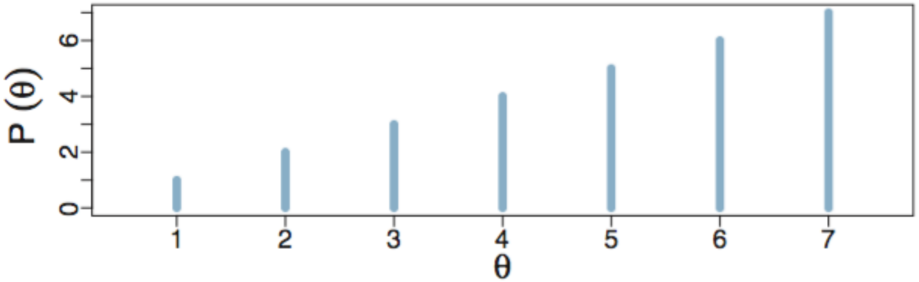
The Traveling Politician

- How can we sample from a large number of representative samples from a distribution?
- Example: Traveling politician in a long chain of islands.
  - Goal is to visit all islands proportional to their relative population.
  - Advisors know population of the current island, and the two adjacent islands.
  - Each day decide if :
    - Stays on the current island
    - moves to the island to the west
    - moves to the island to the east

Travelling Politician (cont)

- Heuristic for moving:
  - Flips a fair coin to decide East or West
    - If the proposed island has a higher population than the current island -> move.
    - If proposed island population is smaller move with probability proportional to the population ratio. $p_{move} = P_{proposed} / P_{current}$
  - In the long run, the probability that the politician is on any one of the islands matches the relative population of the island

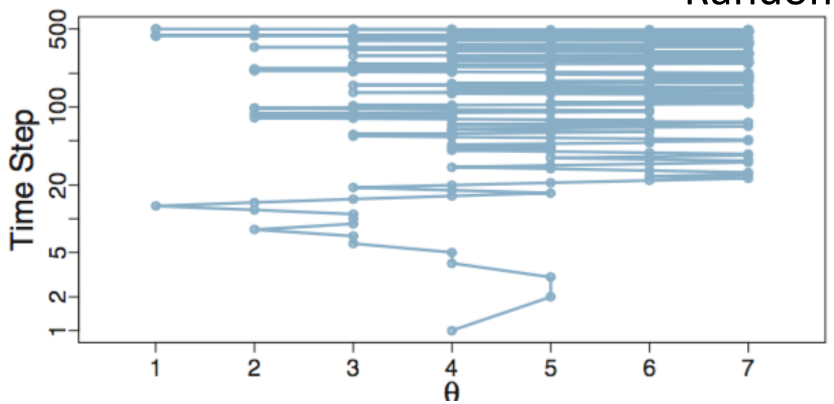
Travelling Politician



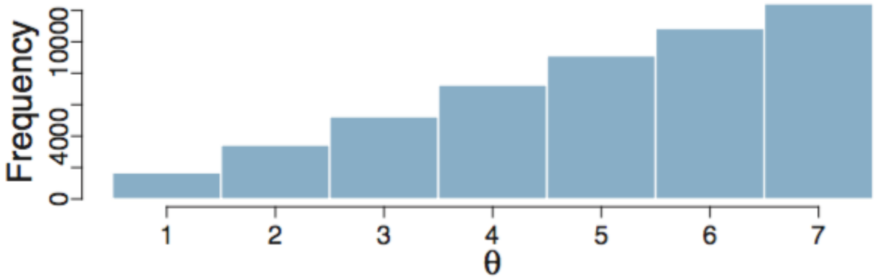
True Population

$$p_{\text{move}} = \min\left(\frac{P(\theta_{\text{proposed}})}{P(\theta_{\text{current}})}, 1\right)$$

Random Walk



Frequency Distribution



John K. Kruschke, Doing Bayesian Data Analysis

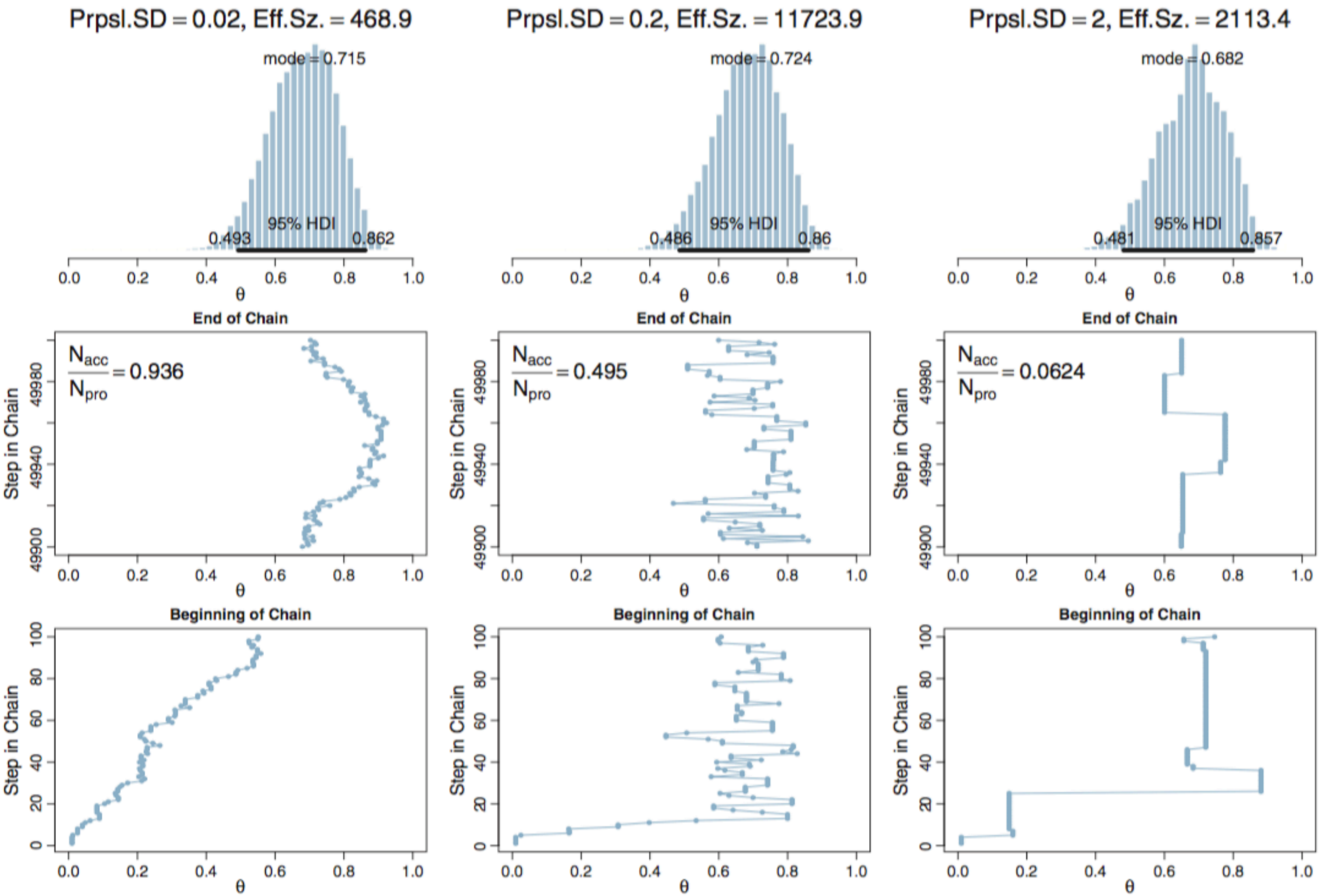
Metropolis Algorithm

- We have some target distribution $P(\theta)$, over a multidimensional continuous parameter space from which we would like to generate representative sample values.
- We must be able to compute the value of $P(\theta)$ for any candidate value of θ.
- The distribution, $P(\theta)$, does not have to be normalized, however. It merely needs to be non-negative.
- In typical applications, $P(\theta)$ is the unnormalized posterior distribution on θ, which is to say, it is the product of the likelihood and the prior.

Metropolis Algorithm

- Start at an arbitrary initial value of θ (in the valid range)
- Randomly generate a proposed jump
- Compute the probability of moving to the proposed value $p_{\text{move}} = \min\left(1, \frac{P(\theta_{\text{pro}})}{P(\theta_{\text{cur}})}\right)$
  - Accept the proposed parameter value if a random value sampled from a [0,1] uniform distribution is less than p_{move}
  - Otherwise reject the proposed parameter value and tally the current value again
- Repeat the above steps until it is judged that a sufficient representative sample has been generated

Bernoulli x=14, N = 20



John K. Kruschke, Doing Bayesian Data Analysis

Two-dimentional prior, likelihood and posterior

- To estimate the parameters θ$_2$ and θ$_1$ we must specify what we believe about them. And because they form a probability

$$\iint \mathrm{d}\theta_1 \mathrm{d}\theta_2 p(\theta_1, \theta_2) = 1$$

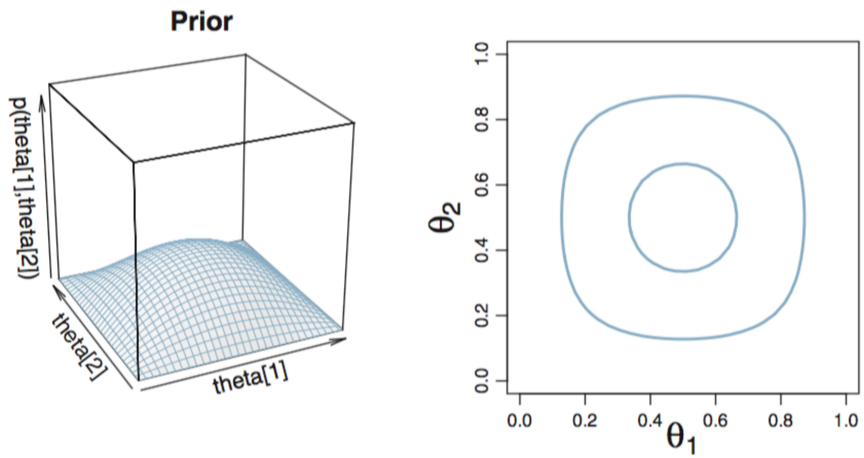
- Additionally, we also have some observed data. Assuming that θ$_2$ and θ$_1$ are independent

$$p(y_1|\theta_1, \theta_2) = p(y_1|\theta_1) \text{ and } p(y_2|\theta_1, \theta_2) = p(y_2|\theta_2)$$

- With posterior distribution:

$$\begin{aligned} p(\theta_1, \theta_2|D) &= p(D|\theta_1, \theta_2)p(\theta_1, \theta_2) \,/\, p(D) \\ &= p(D|\theta_1, \theta_2)p(\theta_1, \theta_2) \Big/ \iint \mathrm{d}\theta_1 \mathrm{d}\theta_2 p(D|\theta_1, \theta_2)p(\theta_1, \theta_2) \end{aligned}$$

2D-Bernoulli



Prior Beta(2,2)

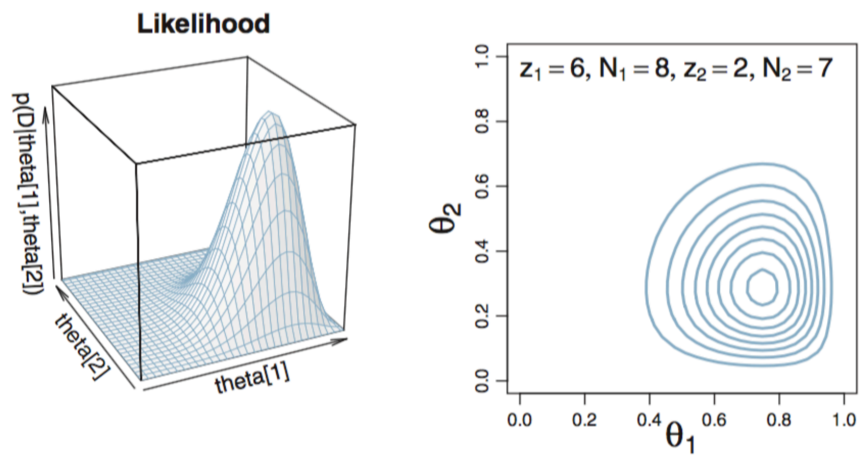
Observed: $\{z_1 = 6, N_1 = 8,\}\{z_2 = 2, N_2 = 7\}$

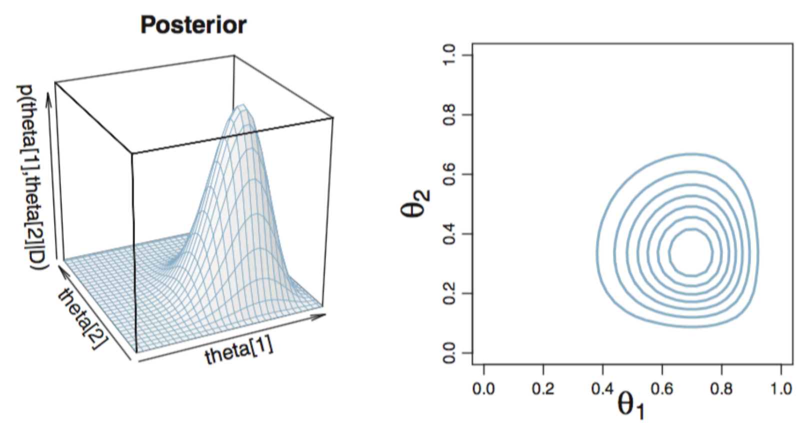
Prior: $\text{Beta}(\theta_1 | a_1, b_1) \cdot \text{Beta}(\theta_2 | a_2, b_2)$,

Posterior :

$\text{Beta}(\theta_1 | z_1 + a_1, N_1 - z_1 + b_1) \cdot$

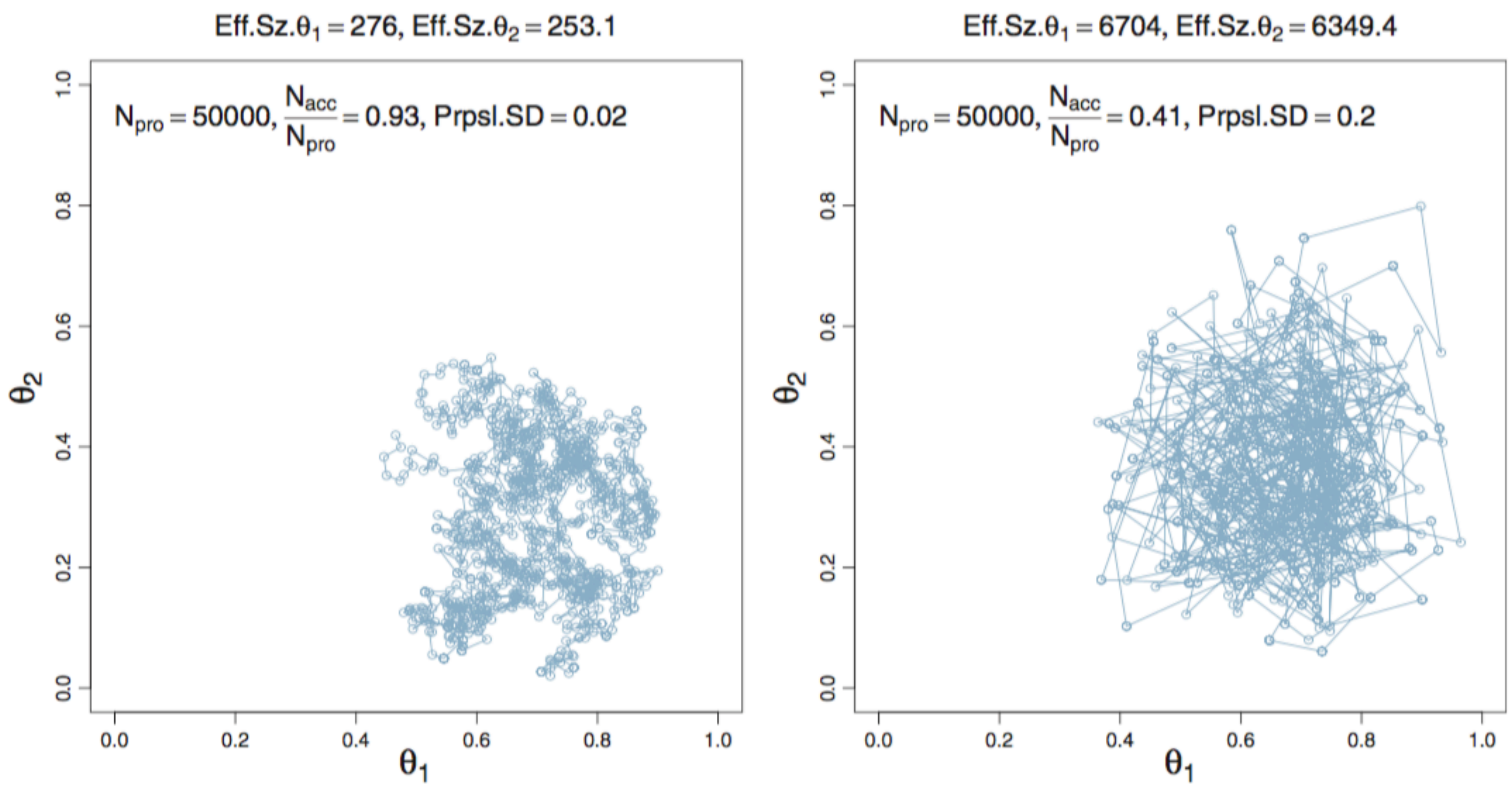
$\text{Beta}(\theta_2 | z_2 + a_2, N_2 - z2 + b2)$





John K. Kruschke, Doing Bayesian Data Analysis

2D Bernoulli - Metropolis

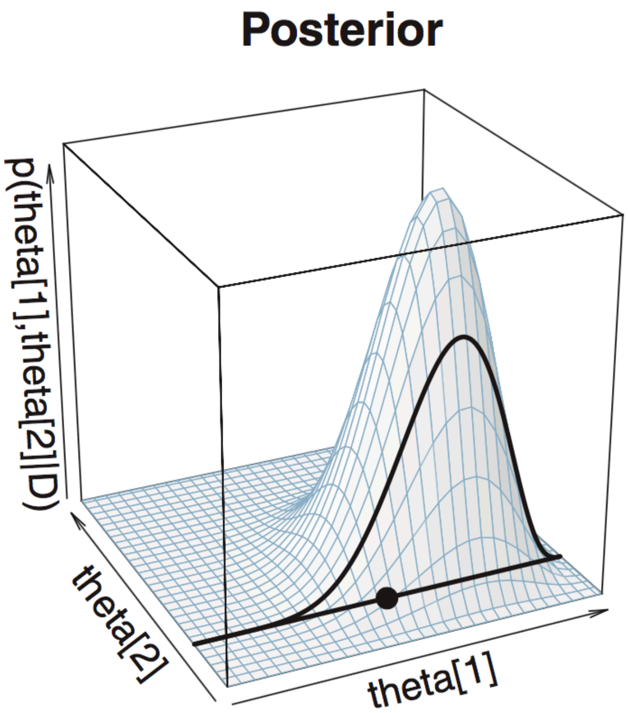


- In the limit of infinite random walks, the Metropolis algorithm yields arbitrarily accurate representations of the underlying posterior distribution.

John K. Kruschke, Doing Bayesian Data Analysis

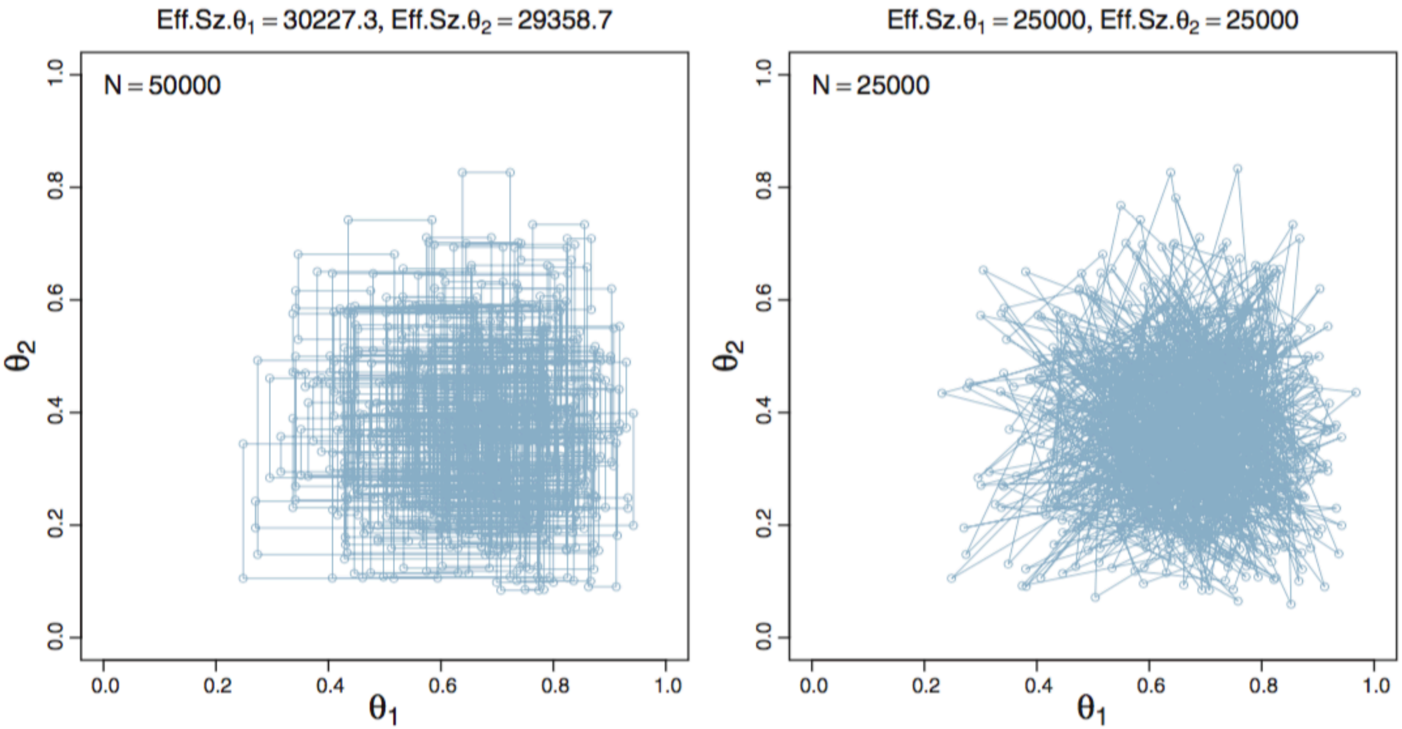
Gibbs Sampling

- Similar to Metropolis. Difference is in how the steps are taken:
  - At each poin in the walk on of the components parameters is selected. Gibbs sampling then chooses a new value for that parameter only
  - The new value for θ_i combined with the unchanged values ofr all other θ_j becomes the new position in the random walk



Gibbs Sampling

Gibbs sampling vs Metropolis sampling



John K. Kruschke, Doing Bayesian Data Analysis

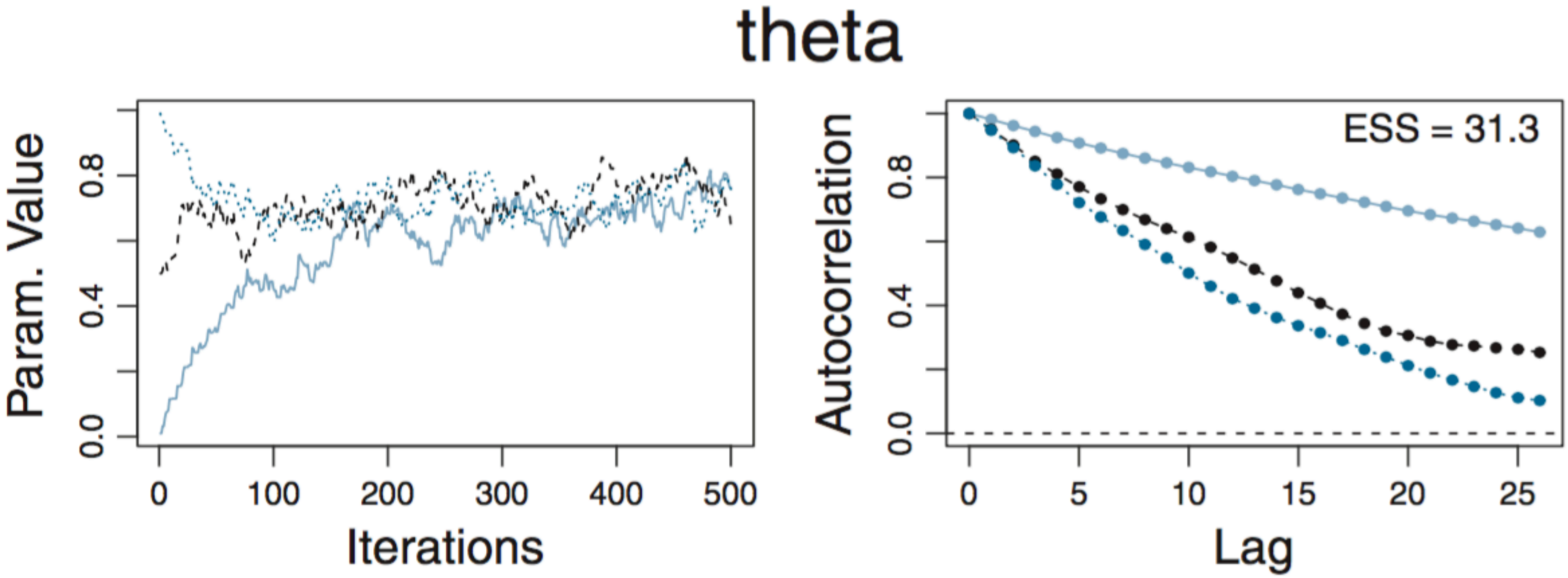
Gibbs sampling (cont)

- Advantages:
  - There is no need to tune a proposal distribution and no inefficiency of rejected proposals.
- Restrictions:
  - The conditional probabilities of each parameter on the others must be calculated
- Disadvantage:
  - It progress can be stalled by highly correlated parameters

MCMC Goals

- Values in the chain:
  - Must be representative of the posterior distribution
  - Should be of sufficient size so that estimates are accurate and stable
- Chain should be generated efficiently, whit as few as possible steps

MCMC Goals



$$\mathrm{ESS} = N \Bigg/ \left(1 + 2\sum_{k=1}^{\infty} \mathrm{ACF}(k)\right)$$

$$\mathrm{MCSE} = \mathrm{SD}/\sqrt{\mathrm{ESS}}$$

Questions?

Homework

- See notebook in the blackboard for the course