Tutorial Workshop on Parameter Estimation for Biological Models

Introduction to Bayesian Analysis

Brian Reich, NC State

July, 2018

A motivating example

- Student 1 will write down a number and then flip a coin
- **If the flip is heads**, they will honestly tell student 2 if the number is even or odd
- **If the flip is tails**, they will lie
- Student 2 will then guess if the number is odd or even
- Let θ be probability that student 2 correctly guesses whether the number is even or odd

A motivating example

Before we start,

1. What's your best guess about θ?

2. What's the probability that θ is greater than a half?

A motivating example

After collecting the data,

1. What's your best guess about θ?

2. What's the probability that θ is greater than a half?

Frequentist approach

- A **frequentist procedure** quantifies uncertainty in terms of *repeating the process that generated the data many times*
- The parameters θ are fixed and unknown
- The sample (data) Y is random
- A frequentest would **never** say $\text{Prob}(\theta > 0) = 0.60$ because θ is not a random variable
- All probability statements should be made about randomness in the data

Frequentist approach

- A **frequentist procedure** quantifies uncertainty in terms of *repeating the process that generated the data many times*
- The parameters θ are fixed and unknown
- The sample (data) Y is random
- A frequentest would **never** say $\text{Prob}(\theta > 0) = 0.60$ because θ is not a random variable
- All probability statements should be made about randomness in the data

Frequentist approach

- A **frequentist** procedure quantifies uncertainty in terms of *repeating the process that generated the data many times*
- A **statistic** $\hat{\theta}$ is a summary of the sample
- For example, the sample mean $\hat{\theta} = \bar{X}$ is a statistic, and it is an **estimator** of the population mean $\theta = \mu$
- **Sampling distribution**: the distribution of $\hat{\theta}$ that arises from repeating the process that generated the data many times
- A frequentist would **never** say "the distribution of μ is Normal(4.2,1.2)"

Frequentist approach

- A **frequentist** procedure quantifies uncertainty in terms of *repeating the process that generated the data many times*
- A **95% confidence interval** (l, u) is an interval constructed from the data that should contain the true value of the parameter 95% of the time if we repeated the process that generated the data many times and computed an interval each time
- A frequentist would **never** say "the probability that the true mean is in the interval (3.4, 4.5) is 0.95"

Frequentist approach

- A **frequentist** procedure quantifies uncertainty in terms of *repeating the process that generated the data many times*
- A common approach for testing a hypothesis is to reject the null if a test statistic exceeds a threshold
- For example, we might reject $\mathcal{H}_0 : \mu \leq 0$ in favor of the alternative $\mathcal{H}_1 : \mu > 0$ if $\bar{X} > T$
- A **p-value** is the probability of observing a test statistic at least as extreme as observed in the sample if we repeated the process that generated the data many times
- A frequentist would **never** say "the probability that the null hypothesis is true is 0.03"

How would a frequentist answer these questions?

Before we start:

1. What's your best guess about θ?
2. What's the probability that θ is greater than a half?

After we have observed some trials:

1. What's your best guess about θ now?
2. What's the probability that θ is greater than a half now?

How about a frequentist answer these questions?

Before we start:

1. What's your best guess about θ?

   I don't know

2. What's the probability that θ is greater than a half?

   This question is nonsense, θ is not a random variable

After we have observed some trials:

1. What's your best guess about θ now?

   The sample proportion

2. What's the probability that θ is greater than a half now?

   This question is nonsense, θ is not a random variable

The Bayesian approach

- Bayesians also view θ as fixed and unknown
- However, we express our uncertainty about θ using probability distributions
- Probability statements like this are intuitive (to me at least)
- The distribution before observing the data is the **prior distribution**
- Example: Prob$(\theta > 0.5) = 0.6$.
- This is **subjective** in that people may have different priors
- There is also a field called objective Bayes

The Bayesian approach

- Our uncertainty about θ is changed (hopefully reduced) after observing the data
- The uncertainty distribution of θ after observing the data is the **posterior distribution**
- **Bayes theorem** provides the rule for updating the prior

$$f(\theta|Y) = \frac{f(Y|\theta)f(\theta)}{f(Y)}$$

- In words: Posterior $\propto$ Likelihood$\cdot$prior
- The main difference between Bayesian and frequentist statistics is that all inference is conditional on the single data set we observed Y

Back to the example

- Say we observed $Y = 60$ successes in $n = 100$ trials
- The parameter $\theta \in [0, 1]$ is the true probability of success
- In most cases we would select a prior that puts probability on all values between 0 and 1
- If we have no relevant prior information we might use the prior

$$\theta \sim \text{Uniform}(0, 1)$$

  so that all values between 0 and 1 are equally likely

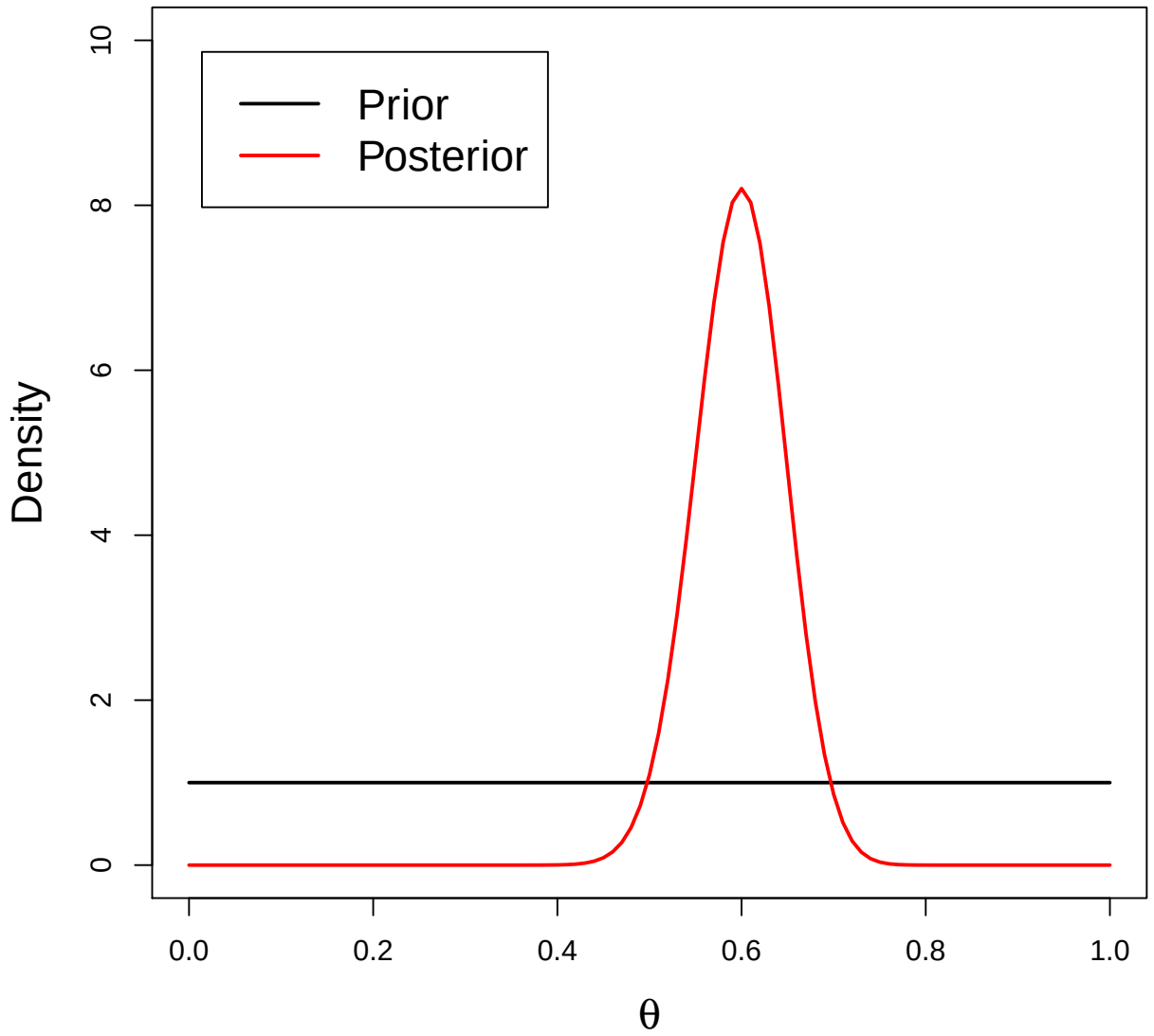
- This is an example of an **uninformative prior**

Posterior distribution

- The likelihood is $Y|\theta \sim \text{Binomial}(n, \theta)$
- The uniform prior is $\theta \sim \text{Uniform}(0, 1)$
- Then it turns out the posterior is

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

Bayesian learning: $Y = 60$ and $n = 100$



Beta prior

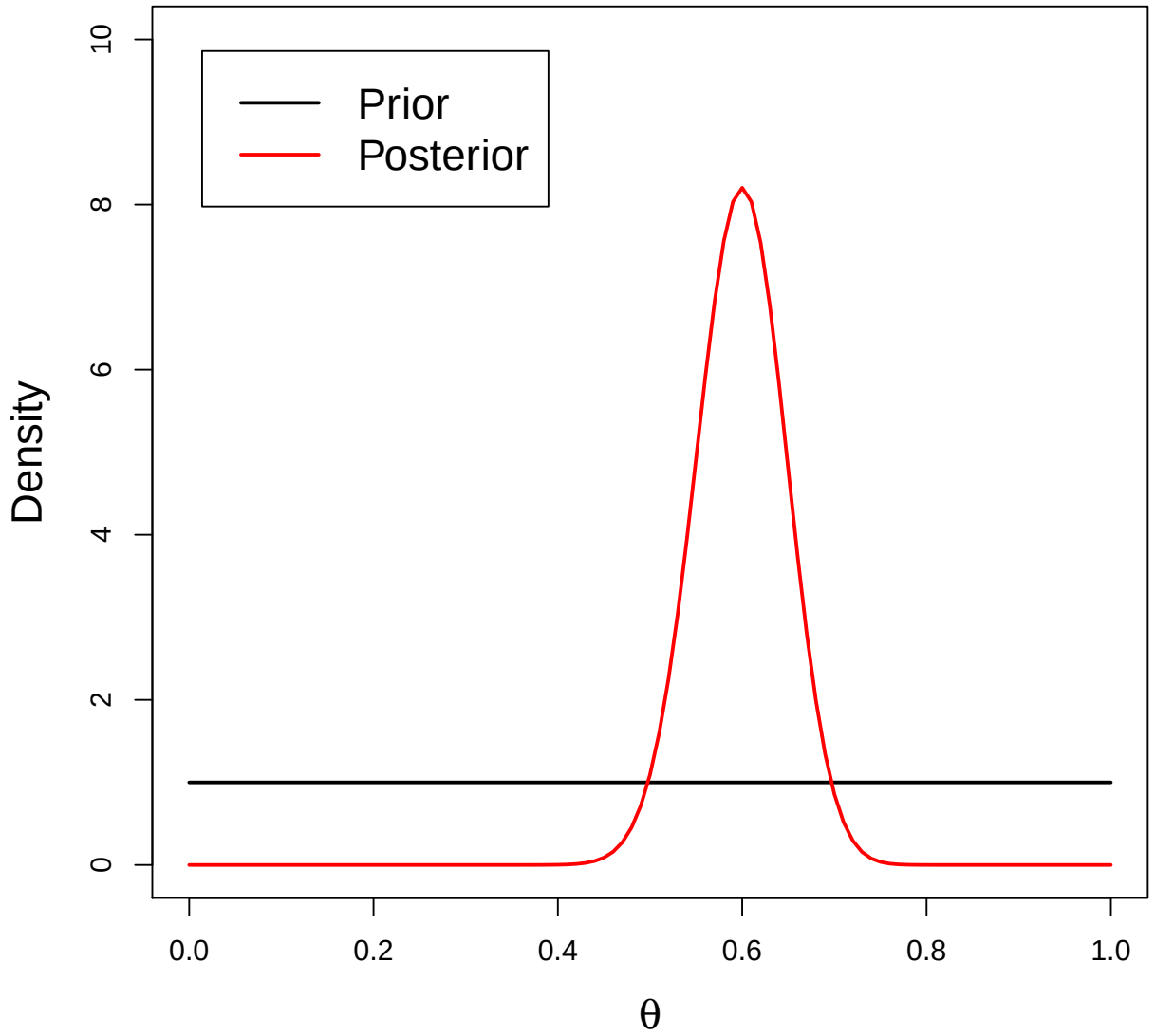
- The uniform prior represents prior ignorance
- To encode prior information we need a more general prior
- The beta distribution is a common prior for a parameter that is bounded between 0 and 1
- If $\theta \sim \text{Beta}(a, b)$ then the posterior is

$$\theta|Y \sim \text{Beta}(Y + a, n - Y + b)$$

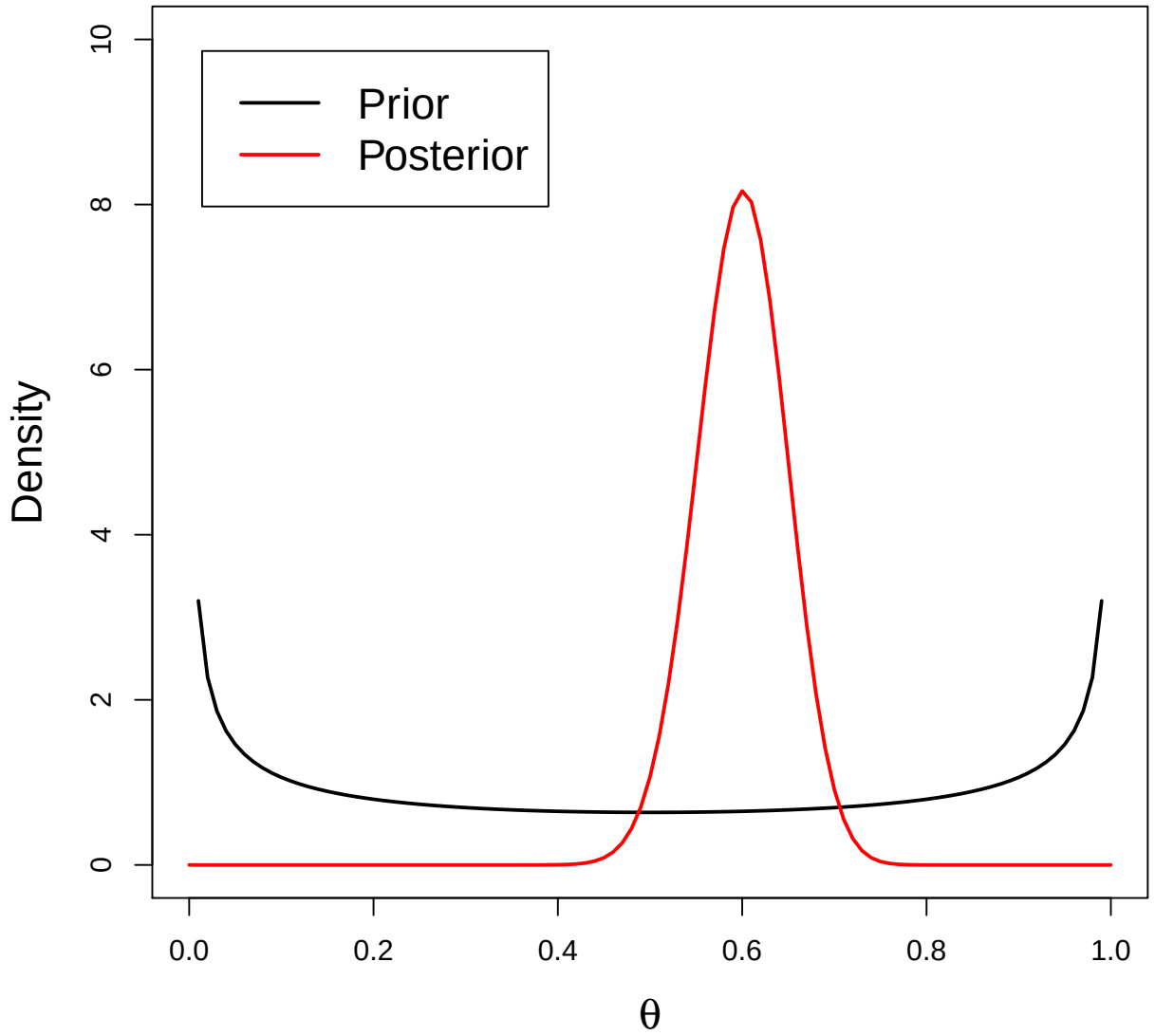
- The posterior mean and variance are

$$\text{E}(\theta|Y) = \frac{Y + a}{n + a + b} \quad \text{and} \quad \text{V}(\theta|Y) = \frac{(Y + a)(n - Y + b)}{(n + a + b)^2(n + a + b + 1)}$$

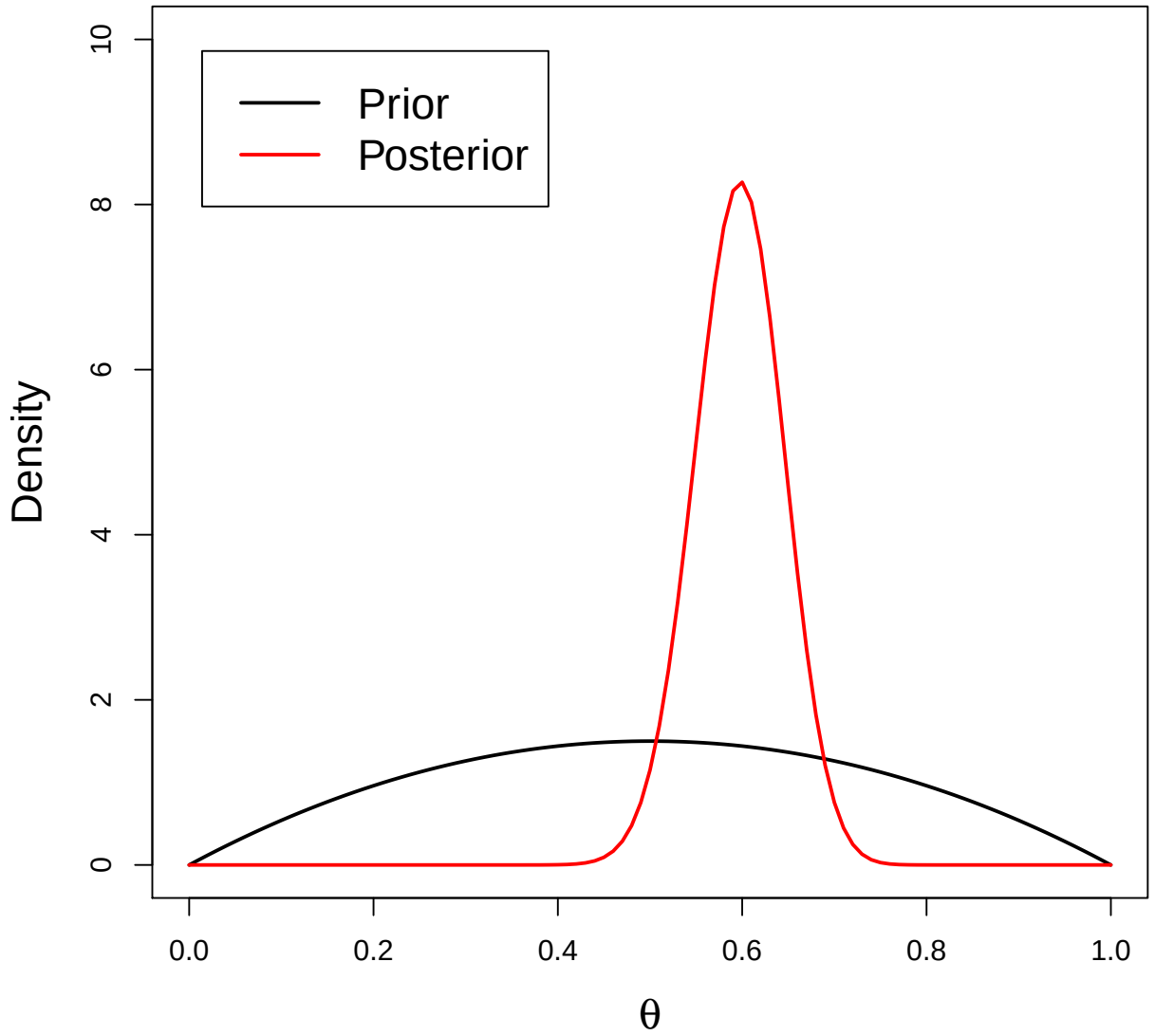
Prior 1: $\theta \sim \text{Beta}(1, 1)$



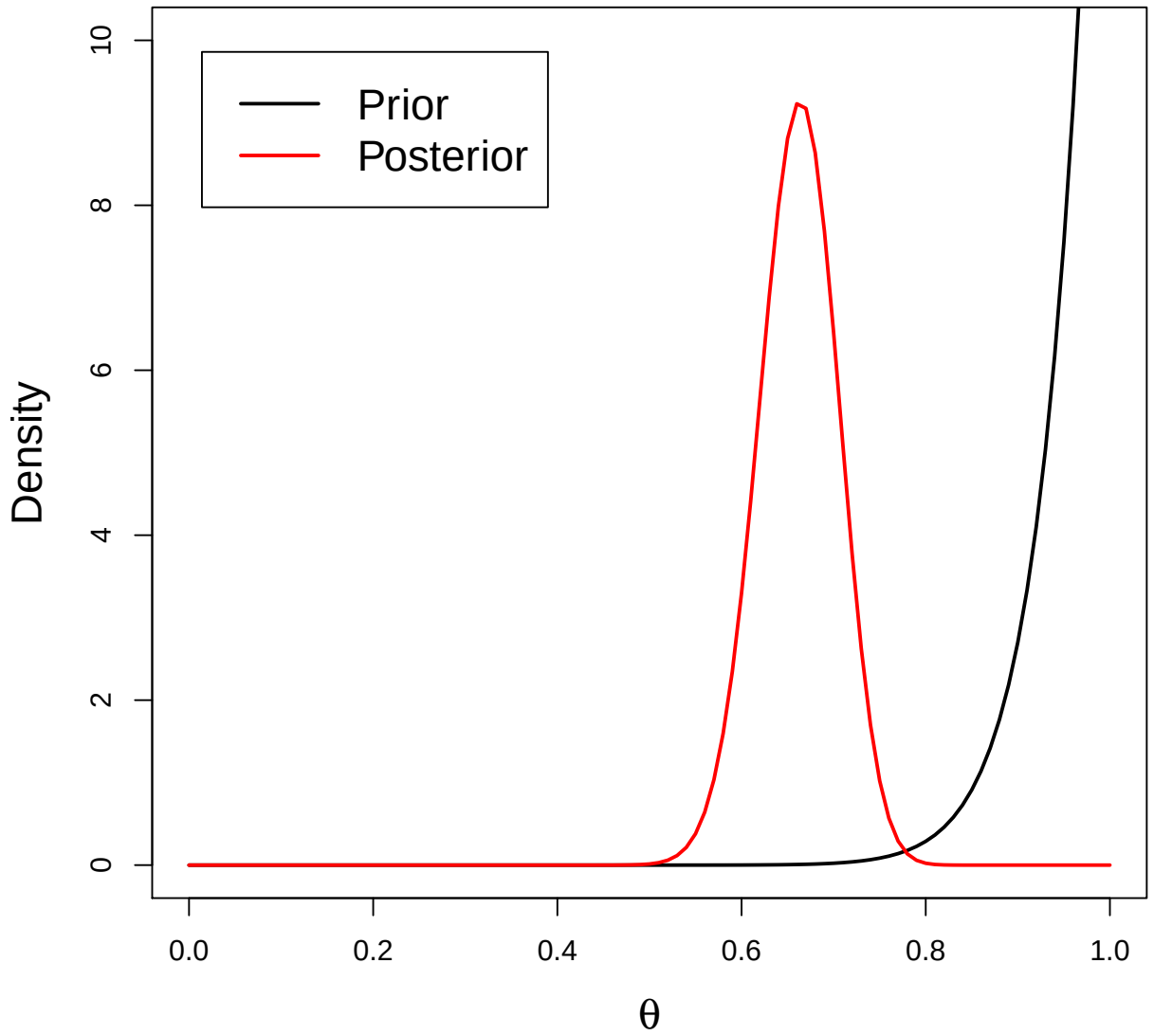
Prior 2: $\theta \sim \text{Beta}(0.5, 0.5)$



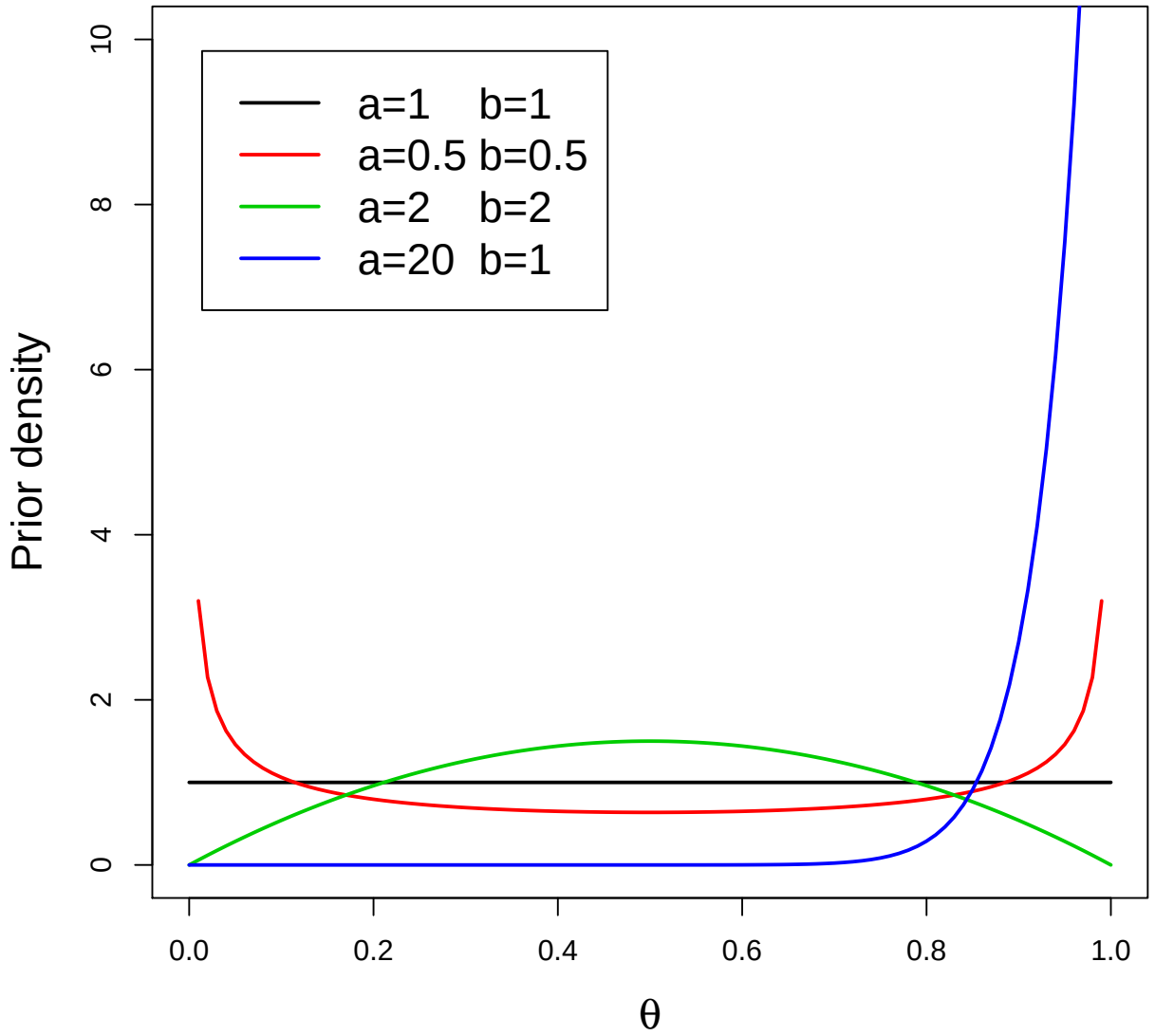
Prior 3: $\theta \sim \text{Beta}(2, 2)$



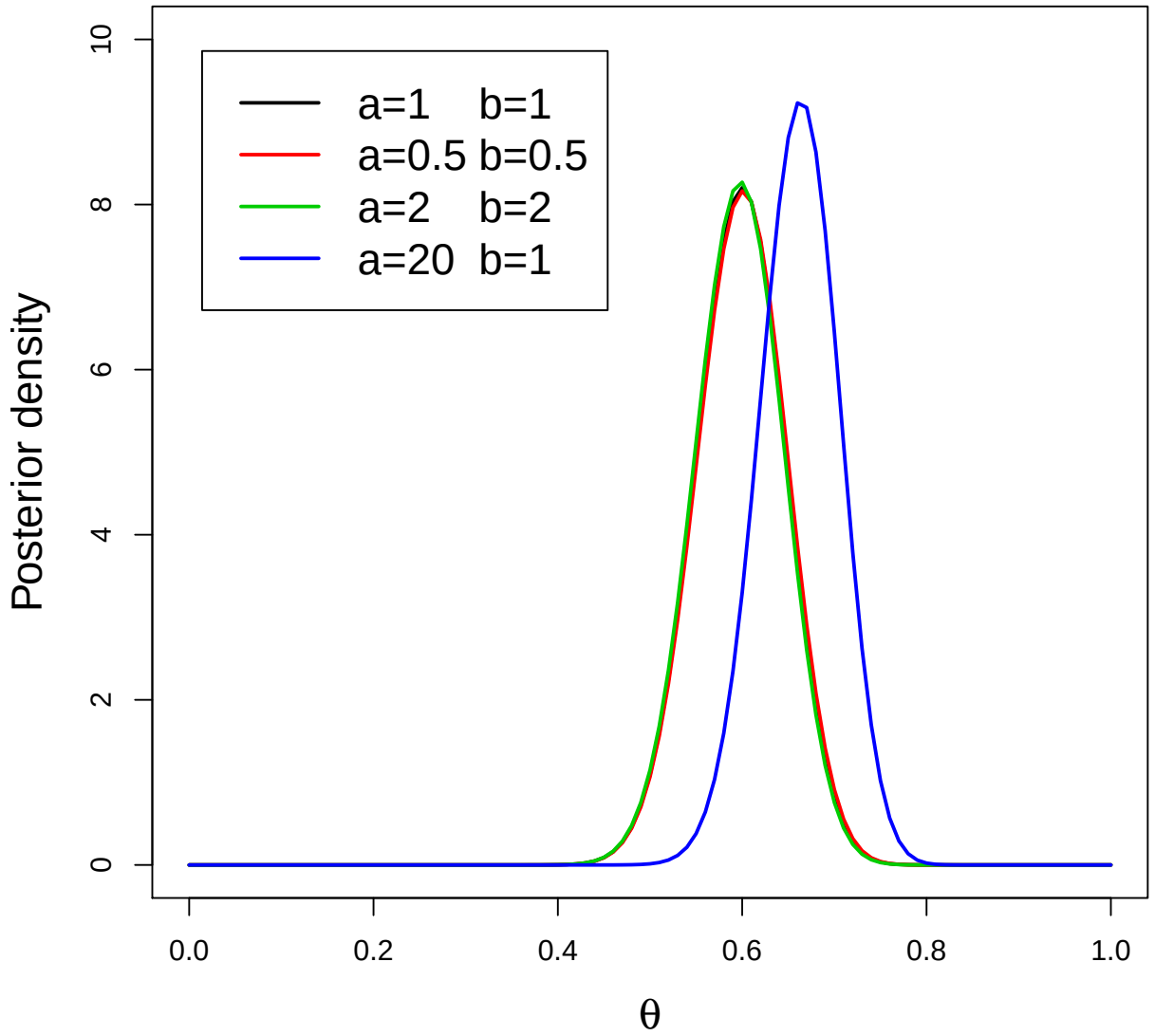
Prior 4: $\theta \sim \text{Beta}(20, 1)$



Plot of different beta priors



Plots of the corresponding posteriors



Senstivity to the prior

		Prior			Posterior		
a	b	Mean	SD	P$>$0.5	Mean	SD	P$>$0.5
1	1	0.50	0.29	0.50	0.60	0.05	0.98
0.5	0.5	0.50	0.50	0.50	0.60	0.05	0.98
2	2	0.50	0.22	0.50	0.60	0.05	0.98
20	1	0.95	0.05	1.00	0.66	0.04	1.00

Summary

- The first three priors give essentially the same results
- Say the objective is to test $\mathcal{H}_o : \theta \leq 0.5$ versus $\mathcal{H}_A : \theta > 0.5$
- In these three cases we can say that after observing the data the probability of the null is only 0.02 and the alternative is 50 times more likely than the null
- The final prior strongly favored large θ and gave different results
- How would we argue this analysis is useful?

Advantages of the Bayesian approach

- Bayesian concepts (posterior prob of the null) are arguably easier to interpret than frequentist ideas (p-value)
- We can incorporate scientific knowledge via the prior
- Excellent at quantifying uncertainty in complex problems
- In some cases the computing is easier
- Provides a framework to incorporate data/information from multiple sources

Disadvantages of Bayesian methods

- Picking a prior is subjective
- Procedures with frequentist properties are desirable
- Computing can be slow or unstable for hard problems
- Less common/familiar
- Nonparametric methods are challenging

Models with more than one parameter

- Thus far we have studied single-parameter models, but most analyses have several parameters
- For example, consider the normal model: $Y_i \sim \text{N}(\mu, \sigma^2)$ with priors $\mu \sim \text{N}(\mu_0, \sigma_0^2)$ and $\sigma^2 \sim \text{InvGamma}(a, b)$
- We want to study the joint posterior distribution $p(\mu, \sigma^2|\mathbf{Y})$
- As another example, consider the simple linear regression model

$$Y_i \sim \text{N}(\beta_0 + X_{1i}\beta_1, \sigma^2)$$

- We want to study the joint posterior $f(\beta_0, \beta_1, \sigma^2|\mathbf{Y})$

Models with more than one parameter

- Thus far we have studied single-parameter models, but most analyses have several parameters
- For example, consider the normal model: $Y_i \sim \text{N}(\mu, \sigma^2)$ with priors $\mu \sim \text{N}(\mu_0, \sigma_0^2)$ and $\sigma^2 \sim \text{InvGamma}(a, b)$
- We want to study the joint posterior distribution $p(\mu, \sigma^2|\mathbf{Y})$
- As another example, consider the simple linear regression model

$$Y_i \sim \text{N}(\beta_0 + X_{1i}\beta_1, \sigma^2)$$

- We want to study the joint posterior $f(\beta_0, \beta_1, \sigma^2|\mathbf{Y})$

Models with more than one parameter

- How to compute high-dimensional (many parameters) posterior distributions?
- How to visualize the posterior?
- How to summarize them concisely?

Models with more than one parameter

- How to compute high-dimensional (many parameters) posterior distributions?
- How to visualize the posterior?
- How to summarize them concisely?

Bayesian one-sample t-test

- In this section we will study the one-sample t-test in depth
- Likelihood: $Y_i|\mu,\sigma \sim \text{N}(\mu,\sigma^2)$ independent over $i = 1,...,n$
- Priors: $\mu \sim \text{N}(\mu_0,\sigma_0^2)$ independent of $\sigma^2 \sim \text{InvGamma}(a,b)$
- The joint (bivariate PDF) of (μ,σ^2) is proportional to

$$\left\{\sigma^n \exp\left[-\frac{\sum_{i=1}^n (Y_i-\mu)^2}{2\sigma^2}\right]\right\} \exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right] (\sigma^2)^{a-1} \exp(-\frac{b}{\sigma^2})$$

- How to summarize this complicated function?

Bayesian one-sample t-test

- In this section we will study the one-sample t-test in depth
- Likelihood: $Y_i|\mu,\sigma \sim \text{N}(\mu,\sigma^2)$ independent over $i = 1,...,n$
- Priors: $\mu \sim \text{N}(\mu_0,\sigma_0^2)$ independent of $\sigma^2 \sim \text{InvGamma}(a,b)$
- The joint (bivariate PDF) of (μ,σ^2) is proportional to

$$\left\{\sigma^n \exp\left[-\frac{\sum_{i=1}^n (Y_i-\mu)^2}{2\sigma^2}\right]\right\} \exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right] (\sigma^2)^{a-1} \exp(-\frac{b}{\sigma^2})$$

- How to summarize this complicated function?

Plotting the posterior on a grid

- For models with only a few parameters we could simply plot the posterior on a grid
- That is, we compute $p(\mu, \sigma^2 | Y_1, ..., Y_n)$ for all combinations of m values of μ and m values of σ^2
- The number of grid points is m^p where p is the number of parameters in the model
- See `http://www4.stat.ncsu.edu/~reich/ABA/code/NN`

Summarizing the results in a table

- Typically we are interested in the marginal posterior

$$f(\mu|\mathbf{Y}) = \int_0^\infty p(\mu, \sigma^2|\mathbf{Y}) d\sigma^2$$

  where $\mathbf{Y} = (Y_1, ..., Y_n)$
- This accounts for our uncertainty about σ^2
- We could also report the marginal posterior of σ^2
- Results are usually given in a table with marginal mean, SD, and 95% interval for all parameters of interest
- The marginal posteriors can be computed using numerical integration
- See `http://www4.stat.ncsu.edu/~reich/ABA/code/NN`

Frequentist analysis of a normal mean

- In frequentist statistics the estimate of the mean is $\bar{Y}$
- If σ is known the 95% interval is

$$\bar{Y} \pm z_{0.975} \frac{\sigma}{\sqrt{n}}$$

where z is the quantile of a normal distribution

- If σ is unknown the 95% interval is

$$\bar{Y} \pm t_{0.975,n-1} \frac{s}{\sqrt{n}}$$

where t is the quantile of a t-distribution

Bayesian analysis of a normal mean

- The Bayesian estimate of μ is its marginal posterior mean
- The interval estimate is the 95% posterior interval
- If σ is known the posterior of $\mu|\mathbf{Y}$ is Gaussian and the 95% interval is

$$\mathrm{E}(\mu|\mathbf{Y}) \pm z_{0.975}\mathrm{SD}(\mu|\mathbf{Y})$$

- If σ is unknown the marginal (over σ^2) posterior of μ is t with $\nu = n + 2a$ degrees of freedom.
- Therefore the 95% interval is

$$\mathrm{E}(\mu|\mathbf{Y}) \pm t_{0.975,\nu}\mathrm{SD}(\mu|\mathbf{Y})$$

- See "Marginal posterior of μ" on `http://www4.stat.ncsu.edu/~reich/ABA/derivations5.pdf`

Bayesian analysis of a normal mean

- The Bayesian estimate of μ is its marginal posterior mean
- The interval estimate is the 95% posterior interval
- If σ is known the posterior of $\mu|\mathbf{Y}$ is Gaussian and the 95% interval is

$$\mathrm{E}(\mu|\mathbf{Y}) \pm z_{0.975}\mathrm{SD}(\mu|\mathbf{Y})$$

- If σ is unknown the marginal (over σ^2) posterior of μ is t with $\nu = n + 2a$ degrees of freedom.
- Therefore the 95% interval is

$$\mathrm{E}(\mu|\mathbf{Y}) \pm t_{0.975,\nu}\mathrm{SD}(\mu|\mathbf{Y})$$

- See "Marginal posterior of μ" on `http://www4.stat.ncsu.edu/~reich/ABA/derivations5.pdf`

Bayesian analysis of a normal mean

- The Bayesian estimate of μ is its marginal posterior mean
- The interval estimate is the 95% posterior interval
- If σ is known the posterior of $\mu|\mathbf{Y}$ is Gaussian and the 95% interval is

$$\mathrm{E}(\mu|\mathbf{Y}) \pm z_{0.975}\mathrm{SD}(\mu|\mathbf{Y})$$

- If σ is unknown the marginal (over σ^2) posterior of μ is t with $\nu = n + 2a$ degrees of freedom.
- Therefore the 95% interval is

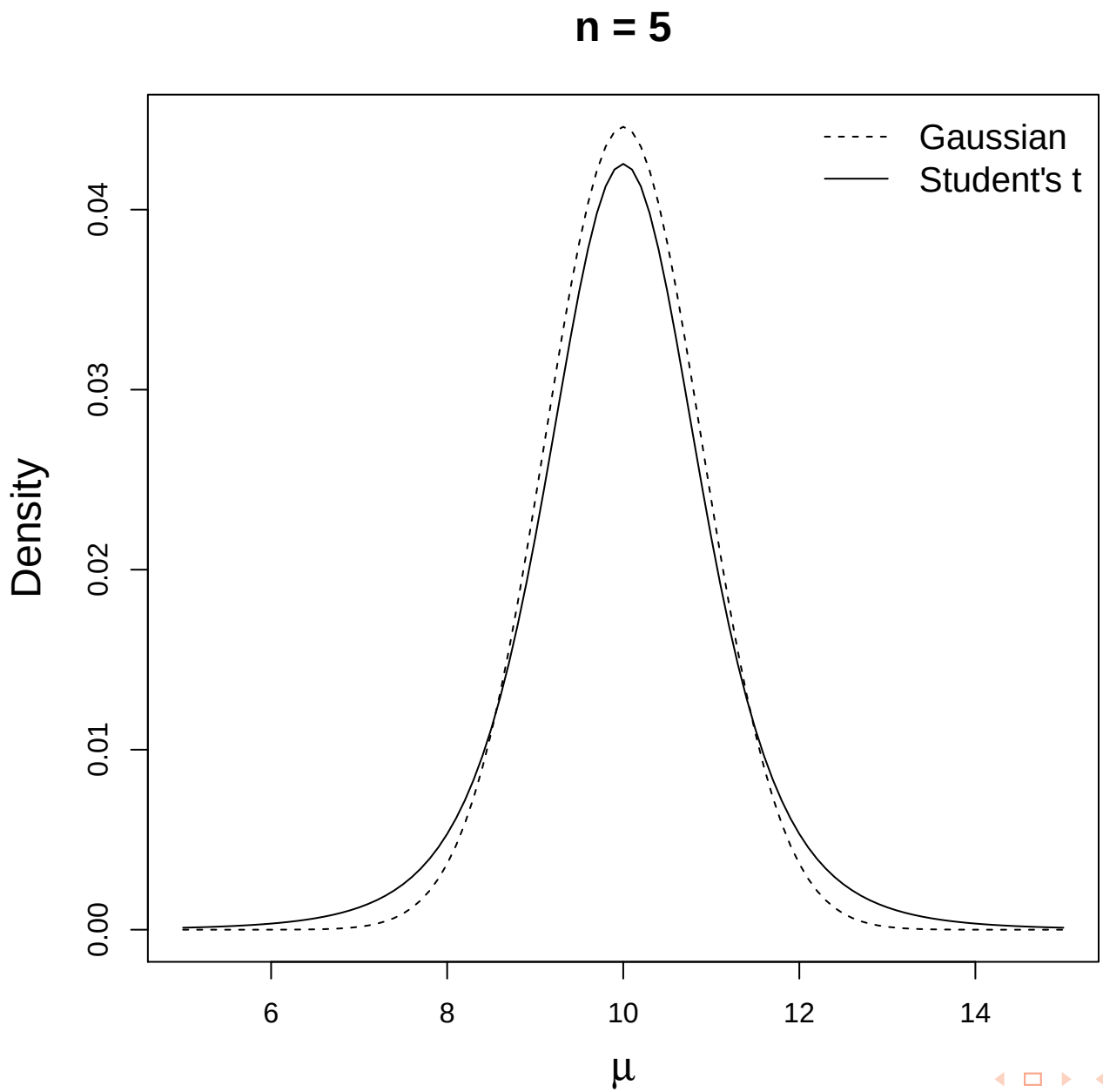
$$\mathrm{E}(\mu|\mathbf{Y}) \pm t_{0.975,\nu}\mathrm{SD}(\mu|\mathbf{Y})$$

- See "Marginal posterior of μ" on `http://www4.stat.ncsu.edu/~reich/ABA/derivations5.pdf`

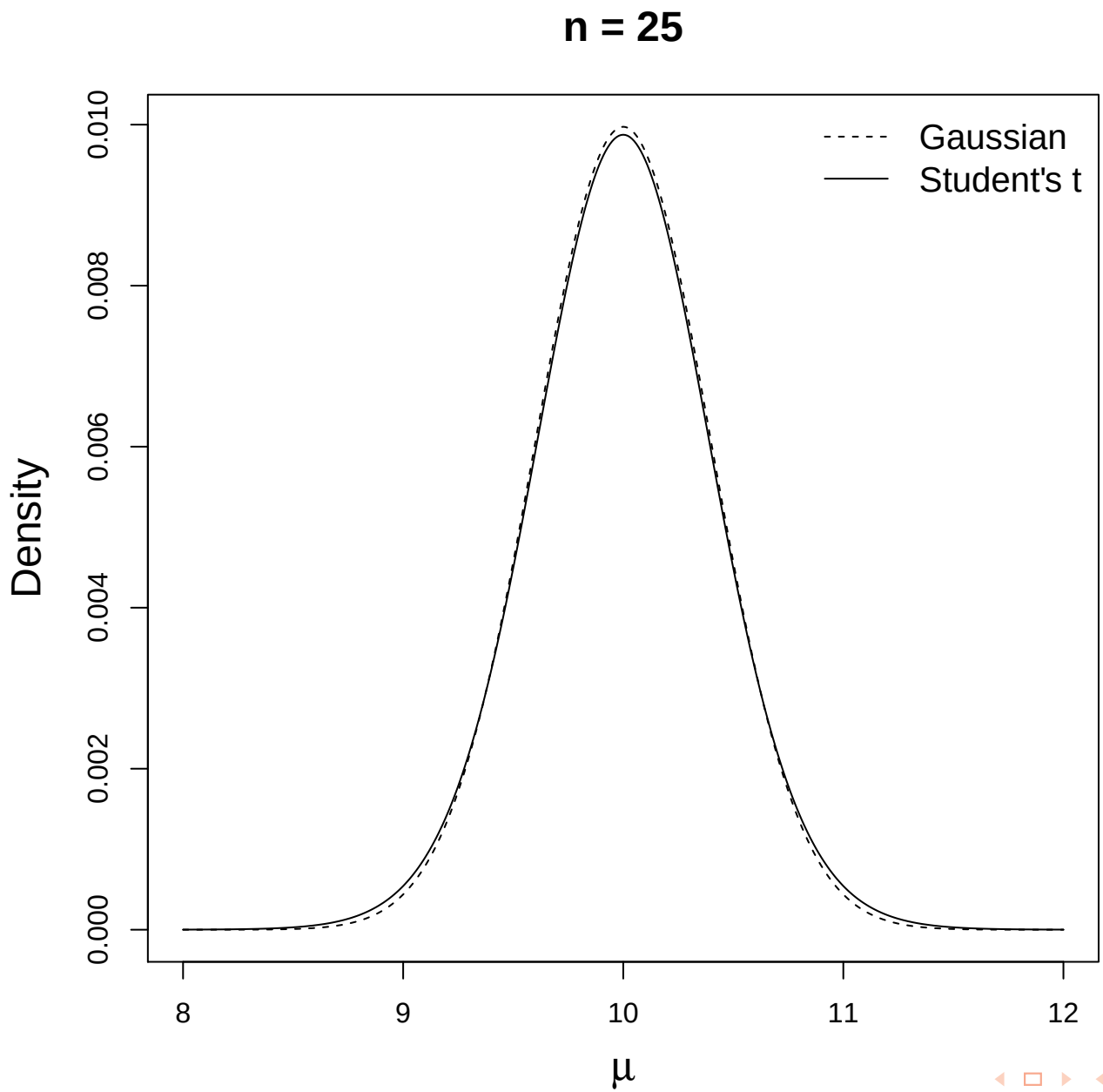
Bayesian analysis of a normal mean

- The following two slides give the posterior of μ for a data set with sample mean 10 and sample variance 4
- The Gaussian analysis assumes $\sigma^2 = 4$ is known
- The t analysis integrates over uncertainty in σ^2
- As expected, the latter interval is a bit wider

Bayesian analysis of a normal mean



Bayesian analysis of a normal mean



Bayesian one sample t-test

- The one-sided test of $H_1 : \mu \leq 0$ versus $H_2 : \mu > 0$ is conducted by computing the posterior probability of each hypothesis
- This is done with the `pt` function in `R`
- The two-sided test of $H_1 : \mu = 0$ versus $H_2 : \mu \neq 0$ is conducted by either
  - Determining if 0 is in the 95% posterior interval
  - Bayes factor (later)

Bayesian one sample t-test

- The one-sided test of $H_1 : \mu \leq 0$ versus $H_2 : \mu > 0$ is conducted by computing the posterior probability of each hypothesis
- This is done with the `pt` function in `R`
- The two-sided test of $H_1 : \mu = 0$ versus $H_2 : \mu \neq 0$ is conducted by either
  - Determining if 0 is in the 95% posterior interval
  - Bayes factor (later)

Methods for dealing with multiple parameters

- In this case, we were able to compute the marginal posterior in closed form (a t distribution)
- We were also able to compute the posterior on a grid
- For most analyses the marginal posteriors will not be a nice distributions, and a grid is impossible if there are many parameters
- We need new tools!

Methods for dealing with multiple parameters

- In this case, we were able to compute the marginal posterior in closed form (a t distribution)
- We were also able to compute the posterior on a grid
- For most analyses the marginal posteriors will not be a nice distributions, and a grid is impossible if there are many parameters
- We need new tools!

Methods for dealing with multiple parameters

Some approaches to dealing with complicated joint posteriors:

- Just use a point estimate, ignore uncertainty
- Approximate the posterior as normal
- Numerical integration
- Monte Carlo sampling

Summarizing a posterior

- Given the data and prior the posterior is determined
- Summarizing the posterior gives parameter estimates, intervals, and hypothesis tests
- Most of these computations are integrals over the posterior
- For simple problems we can do these integrals with pencil and paper, or google
- For medium problems can be solved with numerical integration
- For hard problems we usually use MCMC

Monte Carlo sampling

- Monte Carlo (MC) sampling is the predominant method of Bayesian inference because it can be used for high-dimensional models (i.e., with many parameters)

- The main idea is to approximate posterior summaries by drawing samples from the posterior distribution, and then using these samples to approximation posterior summaries of interest

Monte Carlo sampling

- Notation: Let $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$ be the collection of all parameters in the model
- Notation: Let $\mathbf{Y} = (Y_1, ..., Y_n)$ be entire dataset
- The posterior $f(\boldsymbol{\theta}|\mathbf{Y})$ is a distribution
- If $\theta^{(1)}, ..., \theta^{(S)}$ are samples from $f(\boldsymbol{\theta}|\mathbf{Y})$, then the mean of the S samples approximates the posterior mean
- This only provides approximations of the posterior summaries of interest.
- But how to draw samples from some arbitrary distribution $p(\boldsymbol{\theta}|\mathbf{Y})$?

Software options

- There are now many software options for performing MC sampling
- There are SAS procs and R functions for particular analyses (e.g., the function `BLR` for linear regression)
- There are also all-purpose programs that work for virtually any user-specified model: OpenBUGS; JAGS; Proc MCMC; STAN; INLA (not MC)

MCMC

We will study the algorithms behind these programs, which is important because it helps:

- Select models and priors conducive to MC sampling
- Anticipate bottlenecks
- Understand error messages and output
- Design your own sampler if these off-the-shelf programs are too slow

The most common algorithms are **Gibbs** and **Metropolis** sampling

Gibbs sampling

- Gibbs sampling was proposed in the early 1990s (Geman and Geman, 1984; Gelfand and Smith, 1990) and fundamentally changed Bayesian computing
- Gibbs sampling is attractive because it can sample from high-dimensional posteriors
- The main idea is to break the problem of sampling from the high-dimensional joint distribution into a series of samples from low-dimensional conditional distributions
- Updates can also be done in blocks (groups of parameters)
- Because the low-dimensional updates are done in a loop, samples are not independent
- The dependence turns out to be a Markov distribution, leading to the name Markov chain Monte Carlo (MCMC)

Gibbs sampling for the Gaussian model

- Likelihood: $Y_i|\mu,\sigma \sim \text{N}(\mu,\sigma^2)$ independent over $i = 1,...,n$
- Priors: $\mu \sim \text{N}(\mu_0,\sigma_0^2)$ independent of $\sigma^2 \sim \text{InvGamma}(a,b)$
- The full conditional (FC) distribution is the distribution of one parameter taking all other as fixed and known
- FC1: $\mu|\sigma^2,\mathbf{Y} \sim \text{Normal}\left[\frac{n\bar{Y}\sigma^{-2}+\mu_0\sigma_0^{-2}}{n\sigma^{-2}+\sigma_0^{-2}}, \frac{1}{n\sigma^{-2}+\sigma_0^{-2}}\right]$
- FC2: $\sigma^2|\mu,\mathbf{Y} \sim \text{InvGamma}\left[\frac{n}{2}+a, \frac{1}{2}\sum_{i-1}^{n}(Y_i-\mu)^2+b\right]$

Gibbs sampling for the Gaussian model

- Likelihood: $Y_i|\mu,\sigma \sim \text{N}(\mu,\sigma^2)$ independent over $i = 1,...,n$
- Priors: $\mu \sim \text{N}(\mu_0,\sigma_0^2)$ independent of $\sigma^2 \sim \text{InvGamma}(a,b)$
- The full conditional (FC) distribution is the distribution of one parameter taking all other as fixed and known
- FC1: $\mu|\sigma^2,\mathbf{Y} \sim \text{Normal}\left[\frac{n\bar{Y}\sigma^{-2}+\mu_0\sigma_0^{-2}}{n\sigma^{-2}+\sigma_0^{-2}}, \frac{1}{n\sigma^{-2}+\sigma_0^{-2}}\right]$
- FC2: $\sigma^2|\mu,\mathbf{Y} \sim \text{InvGamma}\left[\frac{n}{2}+a, \frac{1}{2}\sum_{i-1}^{n}(Y_i-\mu)^2+b\right]$

Gibbs sampling

- In the Gaussian model $\boldsymbol{\theta} = (\mu, \sigma^2)$ so $\theta_1 = \mu$ and $\theta_2 = \sigma^2$
- The algorithm begins by setting initial values for all parameters, $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, ..., \theta_p^{(0)})$.
- Variables are then sampled one at a time from their full conditional distributions,

$$p(\theta_j | \theta_1, ..., \theta_{j-1}, \theta_{j+1}, ..., \theta_p, \mathbf{Y})$$

- Rather than 1 p-dimensional joint sample, we make p 1-dimensional samples.
- The process is repeated until the required number of samples have been generated.

Gibbs sampling

A Set initial value $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, ..., \theta_p^{(0)})$

B For iteration t,

FC1 Draw $\theta_1^{(t)} | \theta_2^{(t-1)}, ..., \theta_p^{(t-1)}, \mathbf{Y}$

FC2 Draw $\theta_2^{(t)} | \theta_1^{(t)}, \theta_3^{(t-1)}, ..., \theta_p^{(t-1)}, \mathbf{Y}$

...

FCp Draw $\theta_p^{(t)} | \theta_1^{(t)}, ..., \theta_{p-1}^{(t)}, \mathbf{Y}$

We repeat step B S times giving posterior draws

$$\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(S)}$$

Why does this work?

- $\theta^{(0)}$ isn't a sample from the posterior, it is an arbitrarily chosen initial value
- $\theta^{(1)}$ likely isn't from the posterior either. Its distribution depends on $\theta^{(0)}$
- $\theta^{(2)}$ likely isn't from the posterior either. Its distribution depends on $\theta^{(0)}$ and $\theta^{(1)}$
- **Theorem**: For any initial values, the chain will eventually converge to the posterior
- **Theorem**: If $\theta^{(s)}$ is a sample from the posterior, then $\theta^{(s+1)}$ is too

Convergence

- We need to decide:
  1. When has it converged?
  2. When have we taken enough samples to approximate the posterior?
- Once we decide the chain has converged at iteration T, we discard the first T samples as "burn-in"
- We use the remaining $S - T$ to approximate the posterior
- For example, the posterior mean (marginal over all other parameters) of θ_j is

$$\mathrm{E}(\theta_j|\mathbf{Y}) \approx \frac{1}{S-T} \sum_{s=S-T+1}^{S} \theta_j^{(s)}$$

Tuning the MCMC algoritm

- MCMC is beautiful because it can handle virtually any statistical model and it is usually pretty easy to write functional code
- However, for hard problems great care must be taken to ensure that the algorithm has converged
- There are three main decisions:
  - Selecting the initial values
  - Determining if/when the chain(s) has converged
  - Selecting the number of samples needed to approximate the posterior

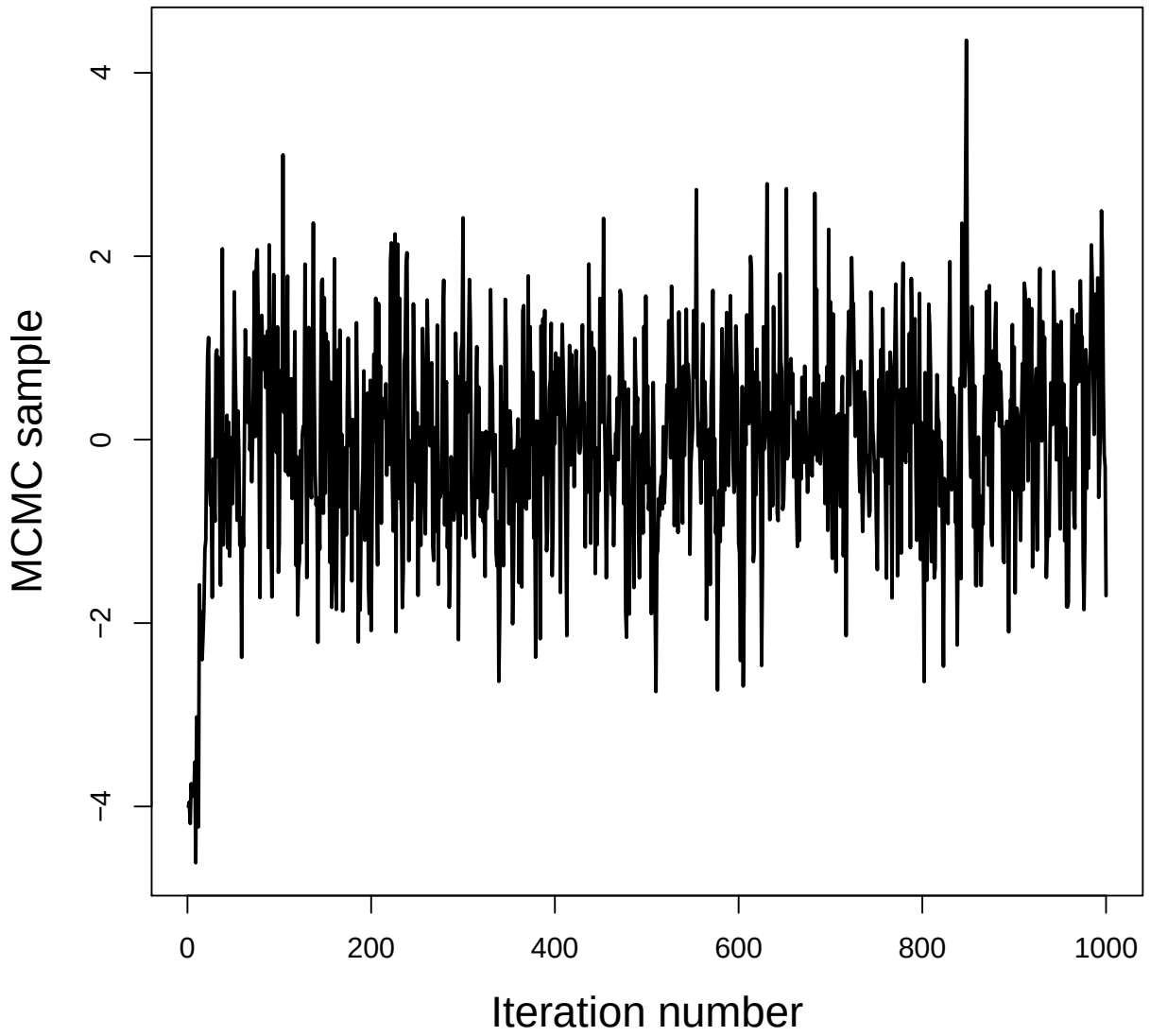
Initial values

- The algorithm will eventually converge no matter what initial values you select
- However taking time to select good initial values will speed up convergence
- It is important to try a few initial values to verify they all give the same result
- Usually 3-5 separate chains is sufficient
- **Option 1**: Select good initial values using method of moments or MLE
- **Option 2**: Purposely pick bad but different initial values for each chain to check convergence

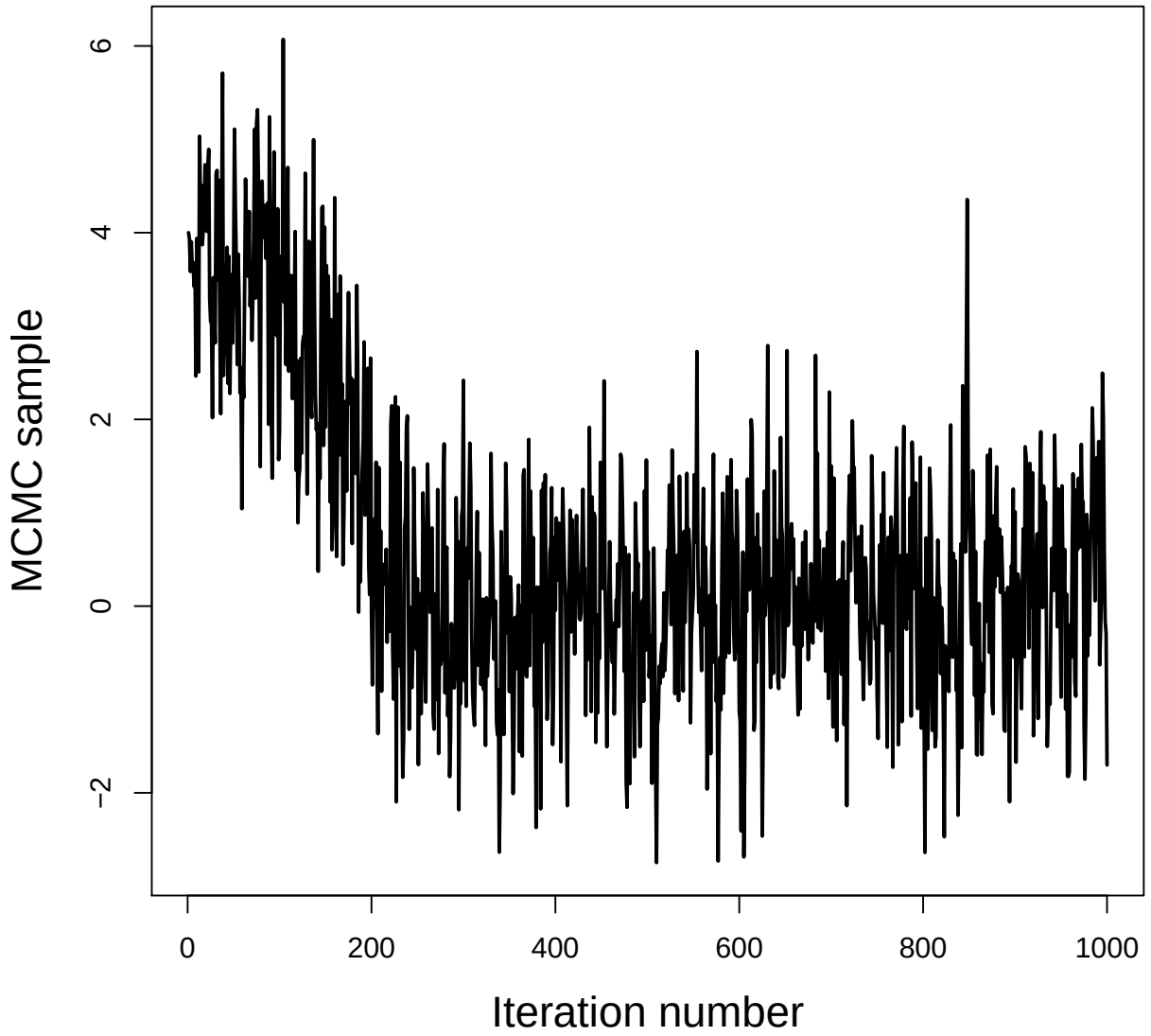
Convergence

- The first few samples are probably not draws from the posterior distribution
- It can take a few dozen, hundreds or even thousands of iterations to move from the initial values to the posterior
- When the sampler reaches the posterior this is called convergence
- Samples before convergence are discard as **burn-in**
- After convergence the samples should not converge to a single point!
- They should be draws from the posterior, and ideally look like a caterpillar or bar code

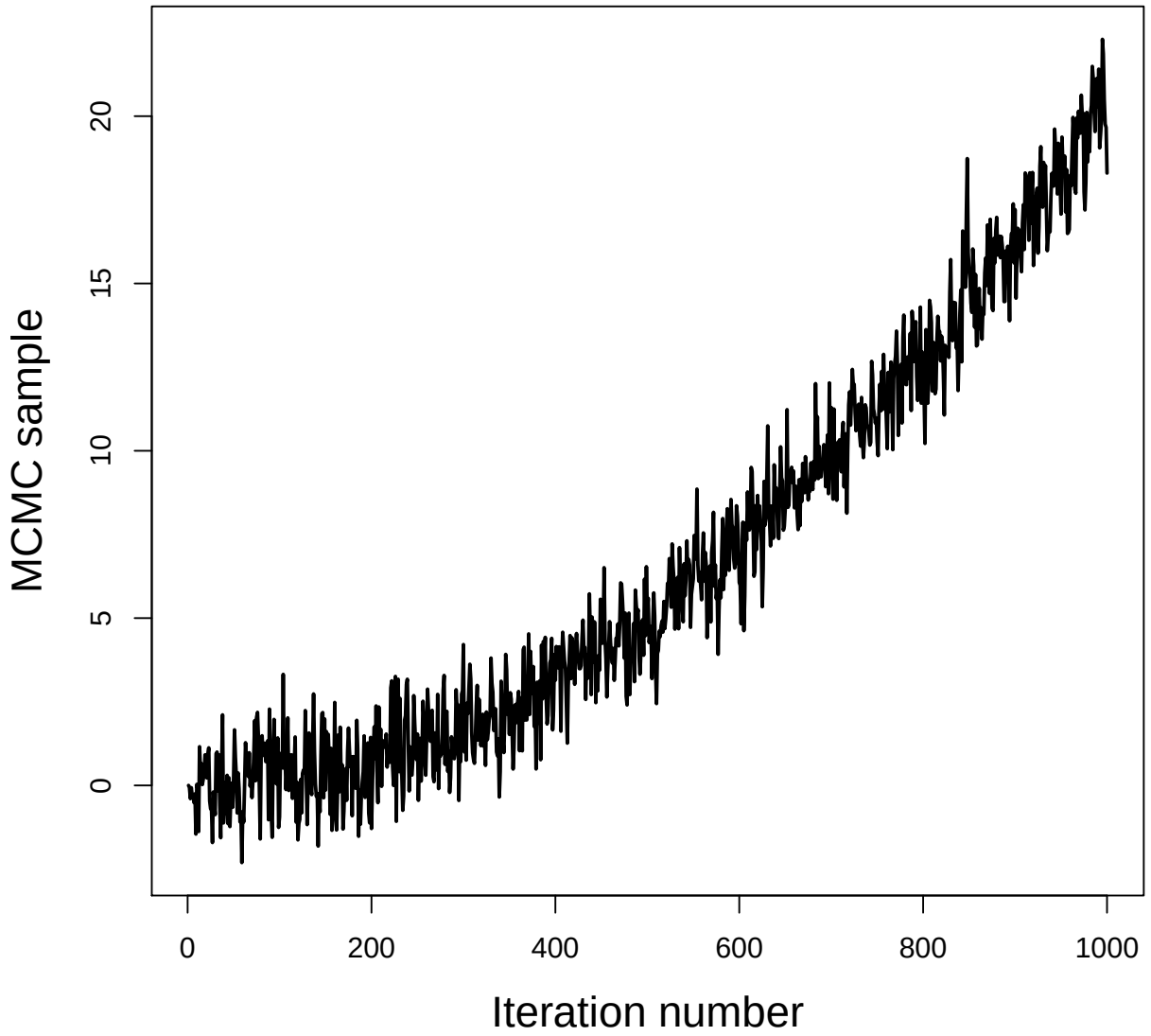
Convergence in a few iterations



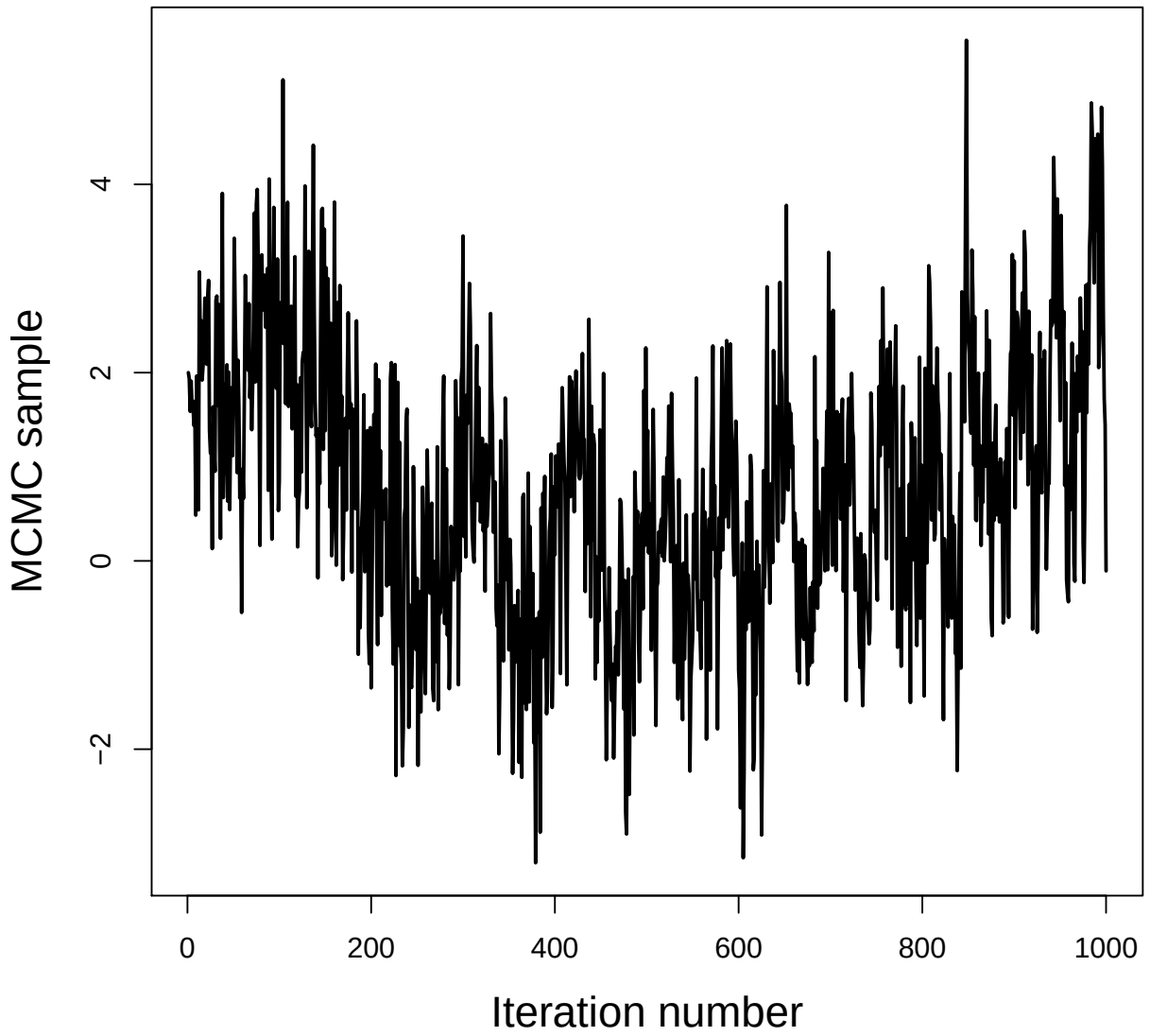
Convergence in a few hundred iterations



This one never converged



Convergence is questionable



Convergence diagnostics

- So far we have visually inspected the chains for convergence
- There are many formal diagnostics (Geweke, Gelman-Rubin, effective sample size, etc.)
- The `CODA` package in `R` has dozens of diagnostics
- Most give a measure of convergence for each parameter
- Checking convergence using these one-number summaries is more efficient and objective than visual inspection

What to do if the chains haven't converged?

- Run it longer
- Pick better initial values
- Pick tighter priors to add stability
- Update highly correlated parameters as a block
- Use a simpler model (e.g., remove collinear predictors)
- Try a different approach, e.g., STAN or write your own code

Metropolis sampling

- In Gibbs sampling each parameter is updated by sampling from its full conditional distribution
- This is possible with conjugate priors
- However, if the prior is not conjugate it is not obvious how to make a draw from the full conditional
- For example, if $Y \sim \text{Normal}(\mu, 1)$ and $\mu \sim \text{Beta}(a, b)$ then

$$p(\mu|Y) \propto \exp\left[-\frac{1}{2}(Y-\mu)^2\right] \mu^{(a-1)}(1-\mu)^{b-1}$$

- For some likelihoods there is no known conjugate prior, e.g., logistic regression
- In these cases we use Metropolis sampling

Metropolis sampling

- Metropolis sampling is a version of rejection sampling
- Let θ_j^* be the current value of the parameter being updated and $\theta_{(j)}$ be the current value of all other parameters
- You propose a random candidate based on the current value, e.g.,

$$\theta_j^c \sim \text{Normal}(\theta_j^*, s_j^2)$$

- The candidate is accepted with probability

$$R = \min\left\{1, \frac{p(\theta_j^c|\theta_{(j)}, \mathbf{Y})}{p(\theta_j^*|\theta_{(j)}, \mathbf{Y})}\right\}$$

- If the candidate is not accepted then you simply retain the previous value and move to the next step

Metropolis sampling

- The candidate standard deviation s_j is a tuning parameter
- Ideally s_j is tuned to give acceptance probability around 0.3-0.4
- If s_j is too small all candidates will be accepted but chain will move slowly
- If s_j is too large candidates will rarely be accepted and chain will get stuck
- Off-the-shelf programs have default values, and many allow you to change the value if the results are unsatisfactory

Examples

- Illustration: `http://www4.stat.ncsu.edu/~reich/ABA/code/MH.R`

- Complete example: `http://www4.stat.ncsu.edu/~reich/ABA/code/logit`

Variants

- You can combine Gibbs and Metropolis in the obvious way, sampling directly from full conditional when possible and Metropolis otherwise
- Adaptive MCMC varies the candidate distribution throughout the chain
- Hamiltonian MCMC uses the gradient of the posterior in the candidate distribution and is used in STAN

Blocked Gibbs/Metropolis

- If a group of parameters are highly correlated convergence can be slow
- One way to improve Gibbs sampling is a block update
- For example, in linear regression might iterate between sampling the block $(\beta_1, ..., \beta_p)$ and σ^2
- Blocked Metropolis is possible too
- For example, the candidate for $(\beta_1, ..., \beta_p)$ could be a multivariate normal

Summary

- With the combination of Gibbs and Metropolis-Hastings sampling we can fit virtually any model
- In some cases Bayesian computing is actually preferable to maximum likelihood analysis
- In most cases Bayesian computing is slower
- However, in the opinion of many it is worth the wait for improved uncertainty quantification and interpretability
- In all cases it is important to carefully monitor convergence