# MAS3929: Bayesian Statistics and Decision Theory

Dr. Lee Fawcett

*School of Mathematics, Statistics & Physics*

**Semester 1, 2025/26**

# Contact details


| | |
|---|---|
| **My name:** | Lee Fawcett |
| **My role:** | Reader in Statistics |
| **Office:** | Room 2.07 Herschel Building |
| **Phone:** | 0191 208 7228 |
| **Email:** | `lee.fawcett@ncl.ac.uk` |
| **www:** | `www.mas.ncl.ac.uk/~nlf8` |

# What do I do?

- Teaching
  - Bayesian Statistics: MAS2903, **MAS3929**
  - Statistical Computing: MAS1608, MAS2906
  - Environmental Extremes: **MAS3921**, MAS8954, MAS8091
  - Global Opportunities: **MAS3092**, ISY3000
- Research
  - Environmental Extremes
  - Bayesian Inference
  - Statistics and Road Safety
  - Technological Innovations in Statistics Education
- Global
  - Student Mobility: Study Abroad, Field Trips
  - Student Recruitment: Europe, India, China

# Module schedule

- **Lectures** every week: Wednesdays (11am), Thursdays (12pm), both in LT2
- **Computing** every fortnight: Thursdays (3pm), Herschel PC
- Beware!
  - Teaching week 4 (w/b 13th October): Some slight schedule changes
  - Teaching week 5 (w/b 20th October): Lee away, Vianey P-R stands in
  - Computer classes: We might not use all of them!
- **Office hours**: Wednesdays (1pm), Thursdays (4pm), 2.07 Herschel

# Module assessment

- January Exam (**80%**)
  - 2 hours
  - Four equally-weighted questions, of which you must choose three
  - Each question will have some easy bits, some hard bits
  - At least two specimens!
- Written Report (**20%**)
  - Calculations, derivations interpretation/insights/discussion
  - Will not go beyond what we cover in Chapters 2 and 3
  - Can work in small groups of up to 4 students if you wish
  - Deadline for submission: **4pm, Wednesday 19th November 2025**
- Numbas assessment (**0%**)
  - After the written report has been submitted, a Numbas assessment will go live
  - "Formative" assessment – counts for 0% of your module grade
  - Closes at **4pm, Friday 5th December 2025**

# Module assessment

- Worksheets (**0%**) – see Chapters 6–8
  - Will coincide with the fortnightly computer practical sessions
  - Will consist of warm-up questions (Chapter 6), exam-style questions (Chapter 7) and computing questions (Chapter 8)
  - Are completely formative – full solutions are (already!) available

# Module resources

- All lecture materials will be available in advance on the MAS3929 Canvas page
- **Notes** provide a single, complete resource for the module
  - HTML screen-reader-friendly, accessible lecture notes
  - Traditional PDF lecture notes for printing:
    - With gaps, for students who like to write stuff down during lectures
    - With gaps filled in
- **Slides** – summarised version of notes, available by Friday each week
- **Solutions**
  - Non-assessed work – already there!
  - Written report – posted 1 week after the deadline

# Recommended textbooks

- "*Bayes' Rule: A Tutorial Introduction to Bayesian Analysis*" - James Stone
- "*Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*" - John Krushke
- "*Bayesian Statistics: An Introduction*" - Peter Lee

"Bayes' rule" is a good introduction to the main concepts in Bayesian statistics but doesn't cover everything in this course. The other books are broader references which go well beyond the contents of this course.

# Chapter 2

## Single parameter problems

## 2.1 Prior and posterior distributions

Data: $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$

Model: pdf/pf $f(\boldsymbol{x}|\theta)$ depends on a single parameter $\theta$
$\rightarrow$ Likelihood function $f(\boldsymbol{x}|\theta)$
considered as a function of $\theta$ for known $\boldsymbol{x}$

Prior beliefs: pdf/pf $\pi(\theta)$

Combine using Bayes Theorem

Posterior beliefs: pdf/pf $\pi(\theta|\boldsymbol{x})$

Posterior distribution summarises all our current knowledge about the parameter $\theta$

## Bayes Theorem

The posterior probability (density) function for $\theta$ is

$$\pi(\theta|\boldsymbol{x}) = \frac{\pi(\theta)\, f(\boldsymbol{x}|\theta)}{f(\boldsymbol{x})}$$

where

$$f(\boldsymbol{x}) = \begin{cases} \int_\Theta \pi(\theta)\, f(\boldsymbol{x}|\theta)\, d\theta & \text{if } \theta \text{ is continuous,} \\ \\ \sum_\Theta \pi(\theta)\, f(\boldsymbol{x}|\theta) & \text{if } \theta \text{ is discrete.} \end{cases}$$

As $f(\boldsymbol{x})$ is not a function of $\theta$, Bayes Theorem can be rewritten as

$$\pi(\theta|\boldsymbol{x}) \propto \pi(\theta) \times f(\boldsymbol{x}|\theta)$$

i.e. **posterior** $\propto$ **prior** $\times$ **likelihood**

## Example 2.1

- Data:

| Year | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 | 2023 | 2024 |
|---|---|---|---|---|---|---|---|---|
| Cases | 2 | 0 | 0 | 0 | 1 | 0 | 2 | 1 |

Table: Number of earthquakes along the *East Anatolian Fault*, 2017–2024.

- Assume that earthquakes occur at random at a constant rate $\theta$ in time
- This means the data are from a Poisson process and so are a random sample from a Poisson distribution with rate $\theta$
- A seismologist was consulted to construct a suitable prior distribution for the rate $\theta$

## Example 2.1

A seismologist gives their prior distribution for the rate $\theta$ as a $Ga(2, 1)$ distribution, with density

$$\pi(\theta) = \theta e^{-\theta}, \quad \theta > 0, \tag{1}$$

and mean $E(\theta) = 2$ and variance $Var(\theta) = 2$. Determine the posterior distribution for $\theta$.

## Solution

. . . Solution...

# Summary

- Model: $X_i|\theta \sim Po(\theta)$, $i = 1, 2, \ldots, 8$ (independent)
- Prior: $\theta \sim Ga(2, 1)$
- Data: in Table above
- Posterior: $\theta|\boldsymbol{x} \sim Ga(8, 9)$

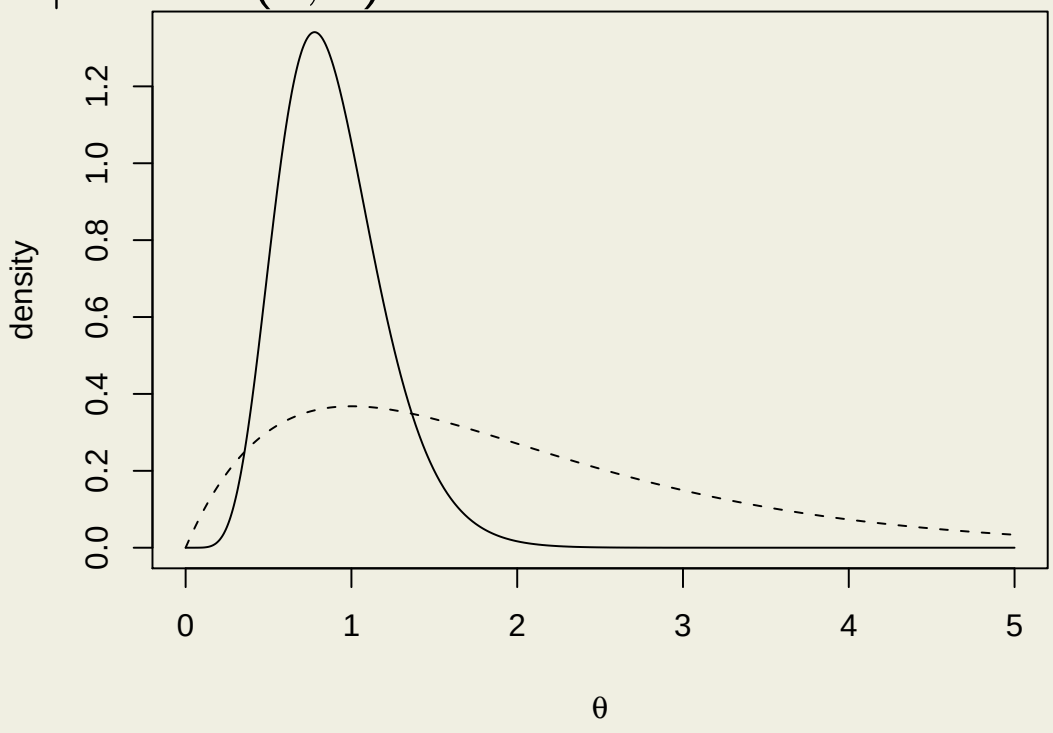



Figure: Prior (dashed) and posterior (solid) densities for $\theta$

| | Prior | Likelihood | Posterior |
|---|---|---|---|
| $Mode(\theta)$ | 1.00 | 0.75 | 0.78 |
| $E(\theta)$ | 2.00 | – | 0.89 |
| $SD(\theta)$ | 1.41 | – | 0.31 |

Table: Changes in beliefs about $\theta$

- Likelihood mode $<$ prior mode $\rightarrow$ posterior mode moves in direction of likelihood mode $\rightarrow$ posterior mode $<$ prior mode
- Reduction in variability from the prior to the posterior

## Example 2.2

We now consider the general case (of Example 2.1). Suppose $X_i|\theta \sim Po(\theta)$, $i = 1, 2, \ldots, n$ (independent) and our prior beliefs about $\theta$ are summarised by a $Ga(g, h)$ distribution (with $g$ and $h$ known), with density

$$\pi(\theta) = \frac{h^g \, \theta^{g-1} e^{-h\theta}}{\Gamma(g)}, \quad \theta > 0. \tag{4}$$

Determine the posterior distribution for $\theta$.

## Solution

...

(1.1)

(1.2)

# Summary

- Model: $X_i|\theta \sim Po(\theta)$, $i = 1, 2, \ldots, n$ (independent)
- Prior: $\theta \sim Ga(g, h)$
- Data: observe $\boldsymbol{x}$
- Posterior: $\theta|\boldsymbol{x} \sim Ga(g + n\bar{x}, h + n)$
- Taking $g \geq 1$

| | Prior (4) | Likelihood (5) | Posterior (6) |
|---|---|---|---|
| $Mode(\theta)$ | $(g-1)/h$ | $\bar{x}$ | $(g + n\bar{x} - 1)/(h + n)$ |
| $E(\theta)$ | $g/h$ | – | $(g + n\bar{x})/(h + n)$ |
| $SD(\theta)$ | $\sqrt{g}/h$ | – | $\sqrt{g + n\bar{x}}/(h + n)$ |

Table: Changes in beliefs about $\theta$

## Comments

- Posterior mean is greater than the prior mean if and only if the likelihood mode is greater than the prior mean, that is,

$$E(\theta|\boldsymbol{x}) > E(\theta) \quad \Longleftrightarrow \quad Mode_\theta\{f(\boldsymbol{x}|\theta)\} > E(\theta)$$

- Standard deviation of the posterior distribution is smaller than that of the prior distribution if and only if the sample mean is not too large, that is

$$SD(\theta|\boldsymbol{x}) < SD(\theta) \quad \Longleftrightarrow \quad Mode_\theta\{f(\boldsymbol{x}|\theta)\} < \left(2 + \frac{n}{h}\right) E(\theta),$$

  and that this will be true in large samples.

## Example 2.3

Suppose we have a random sample from a normal distribution with unknown mean $\mu$ but known precision $\tau$: $X_i|\mu \sim N(\mu, 1/\tau)$, $i = 1, 2, \ldots, n$ (independent).

Suppose our prior beliefs about $\mu$ can be summarised by a $N(b, 1/d)$ distribution, with probability density function

$$\pi(\mu) = \left(\frac{d}{2\pi}\right)^{1/2} \exp\left\{-\frac{d}{2}(\mu - b)^2\right\}. \tag{7}$$

Determine the posterior distribution for $\mu$.

Hint:

$$d(\mu - b)^2 + n\tau(\bar{x} - \mu)^2 = (d + n\tau)\left\{\mu - \left(\frac{db + n\tau\bar{x}}{d + n\tau}\right)\right\}^2 + c$$

where $c$ does not depend on $\mu$.

## Solution

$$\cdots \tag{1.3}$$

$$\tag{1.4}$$

$$\tag{1.5}$$

# Summary

- Model: $X_i|\mu \sim N(\mu, 1/\tau)$, $i = 1, 2, \ldots, n$ (independent), with $\tau$ known
- Prior: $\mu \sim N(b, 1/d)$
- Data: observe $\boldsymbol{x}$
- Posterior: $\mu|\boldsymbol{x} \sim N(B, 1/D)$, where

$$B = \frac{db + n\tau\bar{x}}{d + n\tau} \qquad \text{and} \qquad D = d + n\tau$$

| | Prior (7) | Likelihood (8) | Posterior (10) |
|---|---|---|---|
| $Mode(\mu)$ | $b$ | $\bar{x}$ | $(db + n\tau\bar{x})/(d + n\tau)$ |
| $E(\mu)$ | $b$ | – | $(db + n\tau\bar{x})/(d + n\tau)$ |
| $Precision(\mu)$ | $d$ | – | $d + n\tau$ |

Table: Changes in beliefs about $\mu$

- Posterior mean is greater than the prior mean if and only if the likelihood mode (sample mean) is greater than the prior mean, that is

$$E(\mu|\boldsymbol{x}) > E(\mu) \quad \Longleftrightarrow \quad Mode_{\mu}\{f(\boldsymbol{x}|\mu)\} > E(\mu)$$

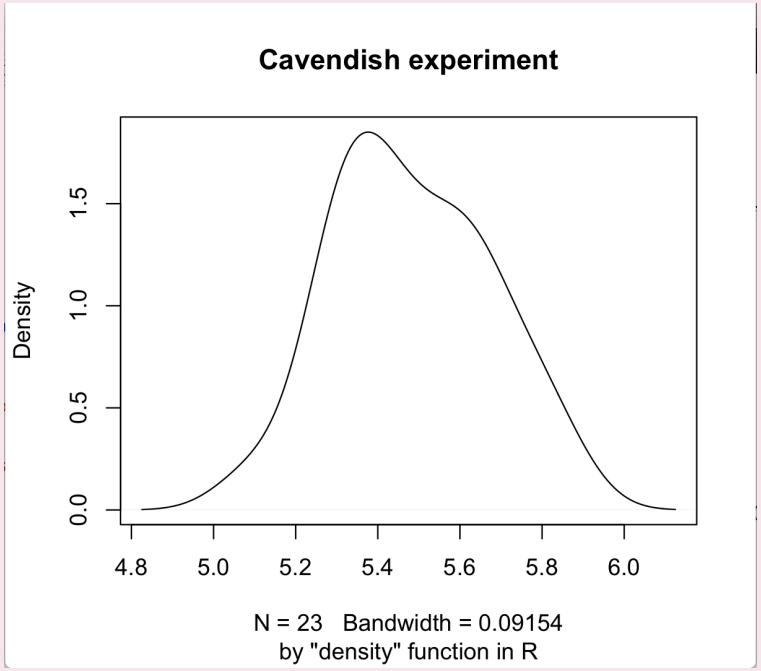
- Standard deviation of the posterior distribution is smaller than that of the prior distribution, that is

$$SD(\mu|\boldsymbol{x}) < SD(\mu)$$

## Example 2.4

The 18th century physicist Henry Cavendish made 23 experimental determinations of the earth's density, and these data (in $g/cm^3$) are shown below.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 5.36 | 5.29 | 5.58 | 5.65 | 5.57 | 5.53 | 5.62 | 5.29 |
| 5.44 | 5.34 | 5.79 | 5.10 | 5.27 | 5.39 | 5.42 | 5.47 |
| 5.63 | 5.34 | 5.46 | 5.30 | 5.78 | 5.68 | 5.85 | |



## Example 2.4

- Cavendish asserts that the error standard deviation of these measurements is $0.2\,g/cm^3$
- Assume that they are normally distributed with mean equal to the true earth density $\mu$, that is, $X_i|\mu \sim N(\mu, 0.2^2)$, $i = 1, 2, \ldots, 23$
- Use a normal prior distribution for $\mu$ with mean $5.41\,g/cm^3$ and standard deviation $0.4\,g/cm^3$
- Derive the posterior distribution for $\mu$

## Solution

...

# Summary

- Small increase in mean from prior to posterior
- Large decrease in uncertainty from prior to posterior

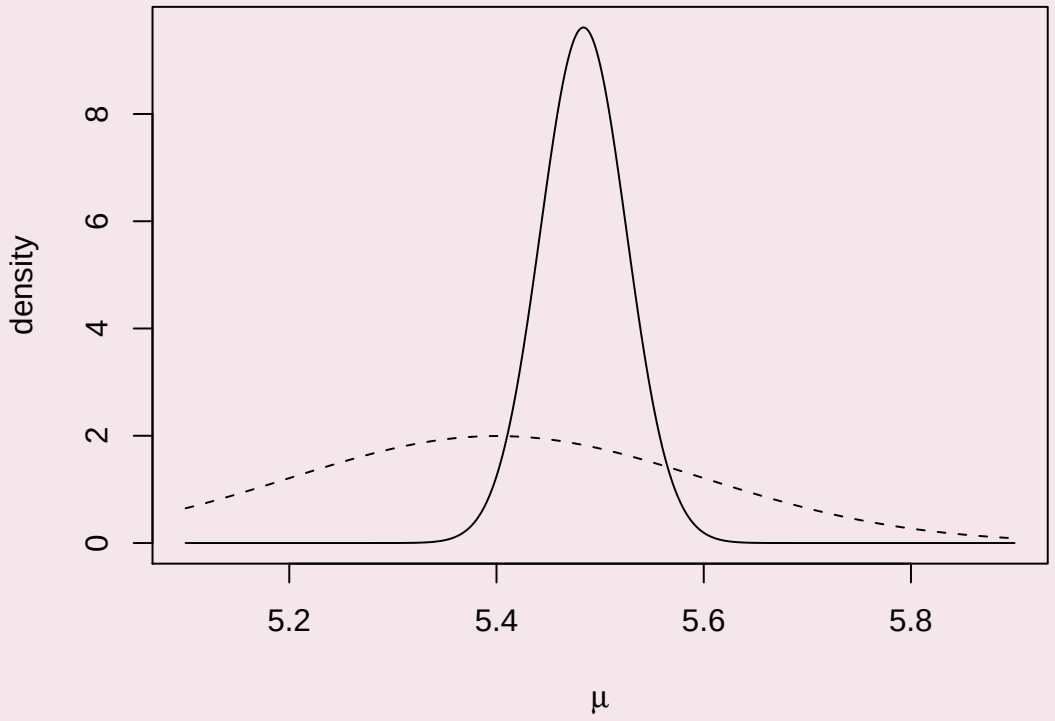



Figure: Prior (dashed) and posterior (solid) densities for the earth's density

# Summary

- The actual mean density of the earth is $5.515\,g/cm^3$ (Wikipedia)
- We can determine the (posterior) probability that the mean density is within 0.1 of this value as follows...
- The posterior distribution is $\mu|\boldsymbol{x} \sim N(5.484, 0.0415^2)$ and so

$$Pr(5.415 < \mu < 5.615|\boldsymbol{x}) = 0.9510,$$

calculated using the `R` command

```
pnorm(5.615,5.484,0.0415)-pnorm(5.415,5.484,0.0415)
```

# Summary

- Without the data, the only basis for determining the earth's density is via the prior distribution
- Here the prior distribution is $\mu \sim N(5.41, 0.4^2)$ and so the (prior) probability that the mean density is within 0.1 of the (now known) true value is

$$Pr(5.415 < \mu < 5.615) = 0.1896,$$

calculated using the R command

```
pnorm(5.615,5.41,0.4)-pnorm(5.415,5.41,0.4)
```

# 2.2 Different levels of prior knowledge

## 2.2.1 Substantial Prior Knowledge

- We have *substantial prior information* for $\theta$ when the prior distribution *dominates* the likelihood function, that is $\pi(\theta|\boldsymbol{x}) \sim \pi(\theta)$

- Difficulties:
  1. Intractability of mathematics in deriving the posterior
  2. Practical formulation of the prior distribution – coherently specifying prior beliefs in the form of a probability distribution is far from straightforward, let alone reconciling differences between experts!

## 2.2.2 Limited prior information

Pragmatic approach:

- Choose a distribution which makes the Bayes updating from prior to posterior mathematically straightforward
- Use what prior information is available to determine the parameters of this distribution

Previous examples:

1. Poisson random sample, Gamma prior distribution $\longrightarrow$ Gamma posterior distribution
2. Normal random sample (known variance), Normal prior distribution $\longrightarrow$ Normal posterior distribution

In these examples, the prior distribution and the posterior distribution come from the same family

## Definition: Conjugacy

Suppose that data $\boldsymbol{x}$ are to be observed with distribution $f(\boldsymbol{x}|\theta)$. A family $\mathfrak{F}$ of prior distributions for $\theta$ is said to be *conjugate* to $f(\boldsymbol{x}|\theta)$ if for every prior distribution $\pi(\theta) \in \mathfrak{F}$, the posterior distribution $\pi(\theta|\boldsymbol{x})$ is also in $\mathfrak{F}$.

## Comment

The conjugate family depends crucially on the model chosen for the data $\boldsymbol{x}$.

For example, the only family conjugate to the model "random sample from a Poisson distribution" is the Gamma family. Here, the likelihood is $f(\boldsymbol{x}|\theta) \propto \theta^{n\bar{x}} e^{-n\theta}$, $\theta > 0$. Therefore we need a family with density $f(\theta|\boldsymbol{a})$ and parameters $\boldsymbol{a}$ such that

$$f(\theta|\boldsymbol{A}) \propto f(\theta|\boldsymbol{a}) \times \theta^{n\bar{x}} e^{-n\theta}, \quad \theta > 0$$

$$\implies \quad f(\theta|\boldsymbol{a}) \propto \theta^{a_1} e^{-a_2\theta}, \quad \theta > 0$$

that is, the Gamma family of distributions

## 2.2.3 Vague prior knowledge/prior ignorance

- Have very little or no prior information about $\theta$
- Still must choose a prior distribution
- Sensible to choose a prior distribution which is not concentrated about any particular value, that is, one with a very large variance
- Most of the information about $\theta$ will be passed through to the posterior distribution via the data, and so we have $\pi(\theta|\boldsymbol{x}) \sim f(\boldsymbol{x}|\theta)$
- Vague prior: Represent vague prior knowledge by using a prior distribution which is conjugate to the model for $\boldsymbol{x}$ and which has as large a variance as possible
- Prior ignorance: Uniform prior? Jeffreys prior?

$$I(\theta) = E_{\boldsymbol{X}|\theta}\left[-\frac{\partial^2}{\partial\theta^2}\log f(\boldsymbol{X}|\theta)\right],$$

and then:

$$\pi(\theta) \propto \sqrt{I(\theta)}.$$

### Example 2.5

Suppose we have a random sample from a $N(\mu, 1/\tau)$ distribution (with $\tau$ known). Determine the posterior distribution assuming a vague prior for $\mu$.

### Solution

...

## Example 2.6

Suppose we have a random sample from an Poisson distribution, that is, $X_i|\theta \sim Po(\theta)$, $i = 1, 2, \ldots, n$ (independent). Determine the posterior distribution assuming a vague prior for $\theta$.

## Solution

- The conjugate prior distribution is a Gamma distribution
- The $Ga(g, h)$ distribution has mean $m = g/h$ and variance $v = g/h^2$
- Rearranging gives $g = m^2/v$ and $h = m/v$
- Clearly $g \to 0$ and $h \to 0$ as $v \to \infty$ (for fixed $m$)
- We have seen that, for this model, using a $Ga(g, h)$ prior distribution results in a $Ga(g + n\bar{x}, h + n)$ posterior distribution
- Therefore, taking a vague prior distribution will give a $Ga(n\bar{x}, n)$ posterior distribution

Note that the posterior mean is $\bar{x}$ (the likelihood mode) and that the posterior variance $\bar{x}/n \to 0$ and $n \to \infty$.

## Extra example

Suppose we have a random sample from an exponential distribution, that is, $X_i|\theta \sim Exp(\theta)$, $i = 1, 2, \ldots, n$ (independent). Determine the Jeffreys prior for this model.

## Solution

$\ldots$

# What have we done so far?

1. **Posterior $\propto$ Prior $\times$ Likelihood**
   - **Example: Poisson distribution** with gamma prior (food poisoning, and generic/algebraic example)
   - **Example: Normal distribution** (known precision) with Normal prior for the mean (generic/algebraic example, and Cavendish experiment). (Memorise the results from the algebraic example!)
   - **Example: Binomial distribution** with beta prior (homework)
   - **Example: Exponential distribution** with Normal prior (homework)
   - Make sure you're familiar with simple functions in `R` for computing probabilities (e.g. `pnorm`), and make sure you can sketch/plot distributions!

# What have we done so far?

2. **Different levels of prior information**
   - **Substantial** prior information: $\pi(\theta|\boldsymbol{x}) \sim \pi(\theta)$, but often difficult to construct
   - **Limited** prior information: Use conjugate prior with parameters determined by the information provided by the expert
   - **Vague** prior: Use conjugate prior with maximised variance
   - **Ignorance** prior: Use a Jeffreys prior, or perhaps a $U(0,1)$?
   - For vague/ignorance priors, $\pi(\theta|\boldsymbol{x}) \sim f(\boldsymbol{x}|\theta)$

3. **Posterior distribution when the sample is very large**

4. **Bayesian inference – confidence intervals and prediction**

# 2.3 Asymptotic Posterior Distribution

## Background

- There are many asymptotic results in Statistics
- The **Central Limit Theorem** is a statement about the asymptotic distribution of $\bar{X}_n$ as the sample size $n \to \infty$, where the $X_i$ are i.i.d. with known mean $\mu$ and known (finite) variance $\sigma^2$
- Under different distributions of $X_i$, $\bar{X}_n$ has the same moments, $E(\bar{X}_n) = \mu$ and $Var(\bar{X}_n) = \sigma^2/n$, but its distribution varies
- The CLT says that as $n \to \infty$, regardless of the distribution of $X_i$,

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{\mathcal{D}} N(0,1)$$

## Theorem

Suppose we have a statistical model $f(\boldsymbol{x}|\theta)$ for data $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$, together with a prior distribution $\pi(\theta)$ for $\theta$. Then

$$\sqrt{J(\hat{\theta})}\,(\theta - \hat{\theta})|\boldsymbol{x} \xrightarrow{\mathcal{D}} N(0,1) \qquad \text{as } n \to \infty,$$

where $\hat{\theta}$ is the likelihood mode and $J(\theta)$ is the observed information

$$J(\theta) = -\frac{\partial^2}{\partial\theta^2} \log f(\boldsymbol{x}|\theta)$$

## Usage

In large samples,

$$\theta|\boldsymbol{x} \sim N\left(\hat{\theta}, J(\hat{\theta})^{-1}\right), \qquad \text{approximately.}$$

- In large samples ($n$ large), how the prior distribution is specified does not matter.

## Example 2.7

Suppose we have a random sample from a $N(\mu, 1/\tau)$ distribution (with $\tau$ known). Determine the asymptotic posterior distribution for $\mu$.
Recall that

$$f(\boldsymbol{x}|\mu) = \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left\{-\frac{\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right\},$$

and therefore

$$\log f(\boldsymbol{x}|\mu) = \frac{n}{2}\log\tau - \frac{n}{2}\log(2\pi) - \frac{\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\Rightarrow \quad \frac{\partial}{\partial\mu}\log f(\boldsymbol{x}|\mu) = -\frac{\tau}{2}\times\sum_{i=1}^{n}-2(x_i-\mu) = \tau\sum_{i=1}^{n}(x_i-\mu) = n\tau(\bar{x}-\mu)$$

$$\Rightarrow \quad \frac{\partial^2}{\partial\mu^2}\log f(\boldsymbol{x}|\mu) = -n\tau \qquad \Rightarrow \quad J(\mu) = -\frac{\partial^2}{\partial\mu^2}\log f(\boldsymbol{x}|\mu) = n\tau.$$

## Solution

...

# 2.4 Bayesian inference

- The posterior distribution $\pi(\theta|\boldsymbol{x})$ summarises all our information about $\theta$ to date
- It can answer the questions: How to estimate the value of $\theta$, and what is the uncertainty of the estimator?

## 2.4.1 Estimation

### Point estimates

Many useful summaries, such as

- the mean: $E(\theta|\boldsymbol{x})$
- the mode: $Mode(\theta|\boldsymbol{x})$
- the median: $Median(\theta|\boldsymbol{x})$

## Interval Estimates

- A more useful summary of the posterior distribution is one which also reflects its variation
- A $100(1-\alpha)\%$ *Bayesian confidence interval* for $\theta$ is any region $C_\alpha$ that satisfies $Pr(\theta \in C_\alpha | \boldsymbol{x}) = 1 - \alpha$
- If $\theta$ is continuous with posterior density $\pi(\theta|\boldsymbol{x})$ then

$$\int_{C_\alpha} \pi(\theta|\boldsymbol{x})\, d\theta = 1 - \alpha$$

- The usual correction is made for discrete $\theta$, that is, we take the largest region $C_\alpha$ such that $Pr(\theta \in C_\alpha | \boldsymbol{x}) \leq 1 - \alpha$
- Bayesian confidence intervals are sometimes called *credible regions* or *plausible regions*
- Clearly these intervals are not unique, since there will be many intervals with the correct probability coverage for a given posterior distribution

## Highest density intervals

- A $100(1-\alpha)\%$ *highest density interval* (HDI) for $\theta$ is the region

$$C_\alpha = \{\theta : \ \pi(\theta|\boldsymbol{x}) \geq \gamma\}$$

  where $\gamma$ is chosen so that $Pr(\theta \in C_\alpha|\boldsymbol{x}) = 1-\alpha$
- It is a $100(1-\alpha)\%$ Bayesian confidence interval but only includes the most likely values of $\theta$
- This region is sometimes called a *most plausible Bayesian confidence interval*
- If the posterior distribution has many modes then it is possible that the HDI will be the union of several disjoint regions

## Highest density intervals

- If the posterior distribution is unimodal (has one mode) and symmetric about its mean then the HDI is an equi-tailed interval, that is, takes the form $C_\alpha = (a, b)$, where

$$Pr(\theta < a|\boldsymbol{x}) = Pr(\theta > b|\boldsymbol{x}) = \alpha/2$$

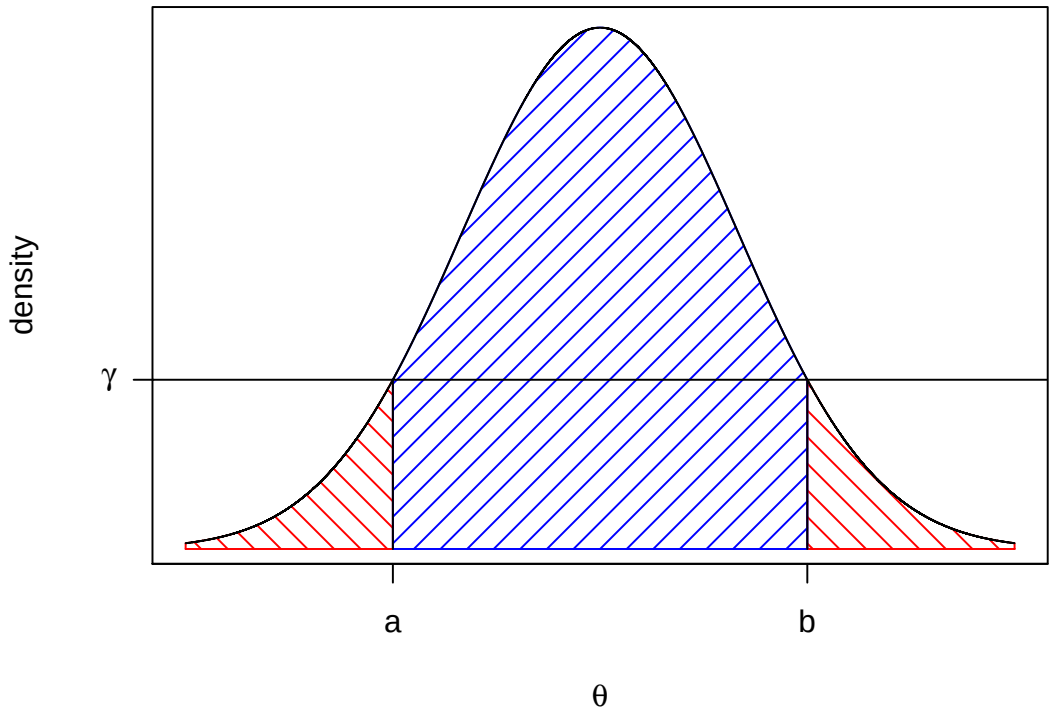



Figure: HDI for $\theta$

## Example 2.8

Suppose we have a random sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ from a $N(\mu, 1/\tau)$ distribution (where $\tau$ is known). We have seen that, assuming vague prior knowledge, the posterior distribution is $\mu|\boldsymbol{x} \sim N\{\bar{x}, 1/(n\tau)\}$. Determine the $100(1-\alpha)\%$ HDI for $\mu$.

## Solution

$\ldots$

## Comment

Note that this interval is numerically identical to the 95% frequentist confidence interval for the (population) mean of a normal random sample with known variance. However, the interpretation is very different.

## Interpretation of confidence intervals

- $C_B$ is a 95% Bayesian confidence interval for $\theta$
- $C_F$ is a 95% frequentist confidence interval for $\theta$

These intervals do not have the same interpretation:

- the probability that $C_B$ contains $\theta$ is 0.95
- the probability that $C_F$ contains $\theta$ is either 0 or 1
- the interval $C_F$ covers the true value $\theta$ on 95% of occasions — in repeated applications of the formula

## Example 2.9

Recall Example 2.1 on the number of earthquakes along the *East Anatolian Fault*. The data were modelled as a random sample from a Poisson distribution with mean $\theta$. Using a $Ga(2, 1)$ prior distribution, we found the posterior distribution to be $\theta|\boldsymbol{x} \sim Ga(8, 9)$, with density

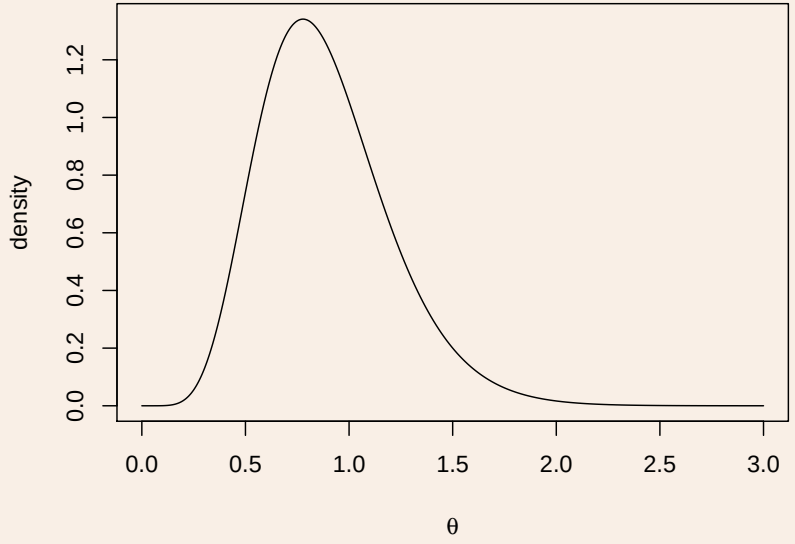



Figure: Posterior density for $\theta$

Determine the 95% HDI for $\theta$.

## Solution

. . .

## Simple way to calculate the HDI

- Use the `R` function `hdiGamma` in the package `nclbayes`
- It calculates the HDI for any Gamma distribution
- Here we use

```
library(nclbayes)
hdiGamma(p=0.95,a=8,b=9)
```

- The package also has functions
  - `hdiBeta` for the Beta distribution
  - `hdiInvchi` for the Inv-Chi distribution (introduced in Chapter 3)

# 2.4.2 Prediction

- Much of statistical inference (both Frequentist and Bayesian) is aimed towards making statements about a parameter $\theta$
- Often the inferences are used as a yardstick for similar future experiments
- For example, we may want to predict the outcome when the experiment is performed again

## Predicting the future

- There will be uncertainty about the future outcome of an experiment
- Suppose this future outcome $Y$ is described by a probability (density) function $f(y|\theta)$
- If $\theta$ were known, say $\theta_0$, then any prediction can do no better than one based on $f(y|\theta = \theta_0)$
- What if $\theta$ is unknown?

## Frequentist solution

- Get estimate $\hat{\theta}$ and use $f(y|\theta = \hat{\theta})$ but this ignores uncertainty on $\hat{\theta}$
- Better: use $E_{\hat{\theta}}\{f(y|\theta = \hat{\theta})\}$ to average over uncertainty on $\hat{\theta}$

## Bayesian solution

- Use the *predictive distribution*, with density

$$f(y|\boldsymbol{x}) = \int_{\Theta} f(y|\theta)\,\pi(\theta|\boldsymbol{x})\,d\theta$$

  when $\theta$ is a continuous quantity
- Notice this could be rewritten as

$$f(y|\boldsymbol{x}) = E_{\theta|\boldsymbol{x}}\{f(y|\theta)\}$$

- Uses $f(y|\theta)$ but weights each $\theta$ by our posterior beliefs

## Prediction interval

- Useful range of plausible values for the outcome of a future experiment
- Similar to a HDI interval
- A $100(1-\alpha)\%$ *prediction interval* for $Y$ is the region

$$C_\alpha = \{y : \ f(y|\boldsymbol{x}) \geq \gamma\}$$

where $\gamma$ is chosen so that $Pr(Y \in C_\alpha|\boldsymbol{x}) = 1-\alpha$

## Example 2.10

Recall Example 2.1 on the number of earthquakes along the *East Anatolian Fault*.

The data for 2017–2024 were modelled as a random sample from a Poisson distribution with mean $\theta$.

Using a $Ga(2, 1)$ prior distribution, we found the posterior distribution to be $\theta|\boldsymbol{x} \sim Ga(8, 9)$.

Determine the predictive distribution for the number of earthquakes for the following year (2025).

## Solution

$\cdots$

## Comments

- This predictive probability function is related to that of a negative binomial distribution
- If $Z \sim NegBin(r, p)$ then

$$Pr(Z = z) = \binom{z-1}{r-1} p^r (1-p)^{z-r}, \quad z = r, r+1, \ldots$$

and so $W = Z - r$ has probability function

$$Pr(W = w) = Pr(Z = w+r) = \binom{w+r-1}{r-1} p^r (1-p)^w, \quad w = 0, 1, \ldots$$

- This is the same probability function as our predictive probability function, with $r = 8$ and $p = 0.9$
- Therefore $Y|\boldsymbol{x} \sim NegBin(8, 0.9) - 8$
- Note that, unfortunately, `R` also calls the distribution of $W$ a negative binomial distribution with parameters $r$ and $p$: `dnbinom(y,r,p)`

- To distinguish between this distribution and the $NegBin(r, p)$ distribution used above, we shall denote the distribution of $W$ as a $NegBin_{\mathrm{R}}(r, p)$ distribution – it has mean $r(1-p)/p$ and variance $r(1-p)/p^2$
- Thus $Y|\boldsymbol{x} \sim NegBin_{\mathrm{R}}(8, 0.9)$

## Comparison between predictive and naive predictive distributions

- Can compare the predictive distribution $Y|\boldsymbol{x}$ with a naive predictive $Y|\theta = \hat{\theta} \sim Po(0.75)$ where

$$f(y|\theta = \hat{\theta}) = \frac{0.75^y \, e^{-0.75}}{y!}, \quad y = 0, 1, \ldots \; .$$

- Probability functions:

| $y$<br>`y = 0:5` | correct<br>$f(y\|\boldsymbol{x})$<br>`dnbinom(y,8,0.9)` | naive<br>$f(y\|\theta = \hat{\theta})$<br>`dpois(y,0.75)` |
|---|---|---|
| 0 | 0.430 | 0.472 |
| 1 | 0.344 | 0.354 |
| 2 | 0.155 | 0.133 |
| 3 | 0.052 | 0.033 |
| 4 | 0.014 | 0.006 |
| 5 | 0.003 | 0.001 |
| $\geq 6$ | 0.005 | 0.002 |

- The naive predictive distribution is a predictive distribution which uses a degenerate posterior distribution $\pi^*(\theta|\boldsymbol{x})$
- Here $Pr_{\pi^*}(\theta = 0.75|\boldsymbol{x}) = 1$ and standard deviation $SD_{\pi^*}(\theta|\boldsymbol{x}) = 0$
- The correct posterior standard deviation of $\theta$ is $SD_{\pi}(\theta|\boldsymbol{x}) = \sqrt{8}/9 = 0.314$
- Using a degenerate posterior distribution results in the naive predictive distribution having too small a standard deviation:

$$SD(Y|\boldsymbol{x}) = \begin{cases} 0.994 & \text{using the correct } \pi(\theta|\boldsymbol{x}) \\ 0.866 & \text{using the naive } \pi^*(\theta|\boldsymbol{x}), \end{cases}$$

  these values being calculated from $NegBin_{\mathrm{R}}(8, 0.9)$ and $Po(0.75)$ distributions
- $\{0, 1, 2\}$ is a 92.9% prediction set/interval
- $\{0, 1, 2\}$ is a 95.9% prediction set/interval using the the more "optimistic" naive predictive distribution

## Predictive distribution (general case)

- Generally requires calculation of a non-trivial integral (or sum)

$$f(y|\boldsymbol{x}) = \int_{\Theta} f(y|\theta)\,\pi(\theta|\boldsymbol{x})\,d\theta$$

- Easier method available when using a conjugate prior distribution
- Suppose $\theta$ is a continuous quantity and $\boldsymbol{X}$ and $Y$ are independent given $\theta$
- Using Bayes Theorem, the posterior distribution for $\theta$ given $\boldsymbol{x}$ and $y$ is

$$\begin{aligned}
\pi(\theta|\boldsymbol{x}, y) &= \frac{\pi(\theta) f(\boldsymbol{x}, y|\theta)}{f(\boldsymbol{x}, y)} \\
&= \frac{\pi(\theta) f(\boldsymbol{x}|\theta) f(y|\theta)}{f(\boldsymbol{x}) f(y|\boldsymbol{x})} \qquad \text{since } \boldsymbol{X} \text{ and } Y \text{ are indep given } \theta \\
&= \frac{\pi(\theta|\boldsymbol{x})\, f(y|\theta)}{f(y|\boldsymbol{x})}.
\end{aligned}$$

Rearranging, we obtain ...

## Candidate's formula

- The predictive p(d)f is

$$f(y|\boldsymbol{x}) = \frac{f(y|\theta)\pi(\theta|\boldsymbol{x})}{\pi(\theta|\boldsymbol{x}, y)}$$

- The RHS looks as if it depends on $\theta$ but it doesn't: all terms in $\theta$ cancel
- For this formula to be useful, we have to be able to work out $\theta|\boldsymbol{x}$ and $\theta|\boldsymbol{x}, y$ fairly easily
- This is the case when using conjugate priors

## Example 2.11

Rework Example 2.10 using Candidate's formula to determine the number of cases in 2006.

## Solution

...

# 2.5 Learning objectives

By the end of this chapter, you should be able to:

1. Determine the likelihood function using a random sample from **any** distribution
2. Combine this likelihood function with **any** prior distribution to obtain the posterior distribution
3. Name the posterior distribution if it is a "standard" distribution listed in these notes or on the exam paper – this list may well include distributions that are standard within the subject but which you have not met before. If the posterior is not a "standard" distribution then it is okay just to give its density (or probability function) up to a constant.
4. Do all the above for a particular data set or for a general case with random sample $x_1, \ldots, x_n$

5. Describe the different levels of prior information; determine and use conjugate priors and vague priors
6. Determine the asymptotic posterior distribution
7. Determine the predictive distribution, particularly when having a random sample from any distribution and a conjugate prior via Candidate's formula
8. Describe and calculate the confidence intervals, HDIs and prediction intervals

# Chapter 3

## Inference for a normal population

## 3.1 Bayes Theorem for many parameters

Data: $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$

Model: depends on many parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^T$
pdf/pf $f(\boldsymbol{x}|\boldsymbol{\theta}) \rightarrow$ Likelihood function $f(\boldsymbol{x}|\boldsymbol{\theta})$

Prior beliefs: pdf/pf $\pi(\boldsymbol{\theta})$

Combine using Bayes Theorem

Posterior beliefs: pdf/pf $\pi(\boldsymbol{\theta}|\boldsymbol{x})$

Posterior distribution summarises all our current knowledge about the parameter $\boldsymbol{\theta}$

## Bayes Theorem

The posterior probability (density) function for $\boldsymbol{\theta}$ is

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{\pi(\boldsymbol{\theta})\, f(\boldsymbol{x}|\boldsymbol{\theta})}{f(\boldsymbol{x})}$$

where

$$f(\boldsymbol{x}) = \begin{cases} \int_{\Theta} \pi(\boldsymbol{\theta})\, f(\boldsymbol{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta} & \text{if } \boldsymbol{\theta} \text{ is continuous,} \\ \\ \sum_{\Theta} \pi(\boldsymbol{\theta})\, f(\boldsymbol{x}|\boldsymbol{\theta}) & \text{if } \boldsymbol{\theta} \text{ is discrete.} \end{cases}$$

As before, this can be rewritten as

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) \propto \pi(\boldsymbol{\theta}) \times f(\boldsymbol{x}|\boldsymbol{\theta})$$

$$i.e. \;\; \text{posterior} \propto \text{prior} \times \text{likelihood}$$

## Generalised $t$ distribution: $X \sim t_a(b, c)$

- Density for $x \in \mathbb{R}$

$$f(x|a, b, c) = \frac{\Gamma\left(\frac{a+1}{2}\right)}{\sqrt{ac\pi}\,\Gamma\left(\frac{a}{2}\right)} \left\{1 + \frac{(x-b)^2}{ac}\right\}^{-\frac{a+1}{2}}$$

- Parameters: $a > 0$, $b \in \mathbb{R}$, $c > 0$
- $E(X) = Mode(X) = b$ and $Var(X) = \dfrac{ac}{a-2}$, if $a \geq 2$
- Generalisation of the standard $t$-distribution since $(X - b)/\sqrt{c} \sim t_a$
- $t_a(0, 1) \equiv t_a$
- $\lim_{a\to\infty} t_a(b, c) = N(b, c)$

## Example 3.1

- If $X \sim t_a(b, c)$ then show that $Y = (X - b)/\sqrt{c} \sim t_a$, with density

$$f(y) = \frac{\Gamma\left(\frac{a+1}{2}\right)}{\sqrt{a\pi}\,\Gamma\left(\frac{a}{2}\right)} \left\{1 + \frac{y^2}{a}\right\}^{-\frac{a+1}{2}}, \qquad -\infty < y < \infty,$$

Recall the general result: if $X$ is a random variable with probability density function $f_X(x)$ and $g$ is a bijective (1–1) function then the random variable $Y = g(X)$ has probability density function

$$f_Y(y) = f_X\left\{g^{-1}(y)\right\} \left| \frac{d}{dy} g^{-1}(y) \right|. \tag{11}$$

## Solution

...

## Comments

- `R` functions `pgt` and `dgt` in the package `nclbayes` give values for $F_X(x)$ and $f_X(x)$ when $X \sim t_a(b, c)$
- Relationship between the generalised $t$ distribution and the standard $t$ distribution is similar to that of the normal distribution and the standard normal distribution:

$$X \sim N(b, c) \quad \Longrightarrow \quad \frac{X - b}{\sqrt{c}} \sim N(0, 1)$$

$$X \sim t_a(b, c) \quad \Longrightarrow \quad \frac{X - b}{\sqrt{c}} \sim t_a$$

## Inverse Chi distribution: $Y \sim \mathit{Inv\text{-}Chi}(a, b)$

- Density

$$f(y|a,b) = \frac{2b^a y^{-2a-1} e^{-b/y^2}}{\Gamma(a)}, \quad y > 0$$

- Parameters: $a > 0,\ b > 0$
- $E(Y) = \dfrac{\sqrt{b}\,\Gamma(a - 1/2)}{\Gamma(a)}$
- $Var(Y) = \dfrac{b}{a-1} - E(Y)^2$, if $a \geq 1$
- The name of the distribution comes from the fact that $1/Y^2 \sim Ga(a, b) \equiv \chi^2_{2a}/(2b)$

## 3.2 Prior to posterior analysis

- Data: $X_i|\mu,\tau \sim N(\mu, 1/\tau)$, $i = 1, 2, \ldots, n$ (indep)
- Prior: $\mu|\tau \sim N\left(b, \frac{1}{c\tau}\right)$, $\tau \sim Ga(g, h)$, with joint density, for $\mu \in \mathbb{R}$, $\tau > 0$

$$\begin{aligned}\pi(\mu,\tau) &= \pi(\mu|\tau)\pi(\tau) \\ &= \left(\frac{c\tau}{2\pi}\right)^{1/2} \exp\left\{-\frac{c\tau}{2}(\mu-b)^2\right\} \times \frac{h^g \tau^{g-1} e^{-h\tau}}{\Gamma(g)} \\ &\propto \tau^{g-\frac{1}{2}} \exp\left\{-\frac{\tau}{2}\left[c(\mu-b)^2 + 2h\right]\right\} \qquad (12)\end{aligned}$$

- We write $\begin{pmatrix}\mu \\ \tau\end{pmatrix} \sim NGa(b, c, g, h)$
- Determine the posterior distribution for $\begin{pmatrix}\mu \\ \tau\end{pmatrix}$

  Hint:

$$c(\mu-b)^2 + n(\bar{x}-\mu)^2 = (c+n)\left\{\mu - \left(\frac{cb+n\bar{x}}{c+n}\right)\right\}^2 + \frac{nc(\bar{x}-b)^2}{c+n}$$

## Solution

$$\cdots \tag{2.1}$$

## 3.2.1 Marginal distributions

- If $\begin{pmatrix}\mu\\ \tau\end{pmatrix} \sim NGa(b,c,g,h)$ then $\tau \sim Ga(g,h)$
- The (marginal) density for $\mu$ is, for $\mu \in \mathbb{R}$

$$
\begin{aligned}
\pi(\mu) &= \int_0^\infty \pi(\mu,\tau)\,d\tau \\
&\propto \int_0^\infty \tau^{g-\frac{1}{2}} \exp\left\{-\frac{\tau}{2}\left[c(\mu-b)^2+2h\right]\right\} d\tau \\
&\propto \int_0^\infty \tau^{g+\frac{1}{2}-1} \exp\left\{-\frac{\tau}{2}\left[c(\mu-b)^2+2h\right]\right\} d\tau \\
&\propto \frac{\Gamma\left(g+\frac{1}{2}\right)}{[\{c(\mu-b)^2+2h\}/2\}]^{g+\frac{1}{2}}} \quad \text{using } \int_0^\infty \theta^{a-1}e^{-b\theta}\,d\theta = \frac{\Gamma(a)}{b^a} \\
&\propto h^{-g-1/2}\left\{1+\frac{c(\mu-b)^2}{2h}\right\}^{-g-1/2} \\
&\propto \left\{1+\frac{c(\mu-b)^2}{2h}\right\}^{-\frac{2g+1}{2}} \\
&\qquad \text{i.e.} \quad \mu \sim t_{2g}\left(b,\frac{h}{gc}\right)
\end{aligned}
\tag{14}
$$

## Summary of marginal distributions

- The data: $X_i|\mu,\tau \sim N(\mu, 1/\tau)$
- Priors: $\mu|\tau \sim N(b, 1/c\tau)$, $\tau \sim Ga(g,h)$
- The joint prior $\begin{pmatrix}\mu\\ \tau\end{pmatrix} \sim NGa(b,c,g,h)$ has marginal distributions
  - $\mu \sim t_{2g}\left(b, \frac{h}{gc}\right)$ (Proven in today's lecture)
  - $\tau \sim Ga(g,h)$

  Also $\sigma = 1/\sqrt{\tau} \sim Inv\text{-}Chi(g,h)$ (One for you in the assignment!)
- The posterior $\left.\begin{pmatrix}\mu\\ \tau\end{pmatrix}\right| \boldsymbol{x} \sim NGa(B,C,G,H)$ has marginal distributions
  - $\mu|\boldsymbol{x} \sim t_{2G}\left(B, \frac{H}{GC}\right)$ (Exactly the same proof as for $\mu \sim t_{2g}$)
  - $\tau|\boldsymbol{x} \sim Ga(G,H)$

  Also $\sigma|\boldsymbol{x} \sim Inv\text{-}Chi(G,H)$ (One for you in the assignment!)

## Example 3.2

- The 18th century physicist Henry Cavendish made 23 experimental determinations of the earth's density, and these data (in $g/cm^3$) are

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 5.36 | 5.29 | 5.58 | 5.65 | 5.57 | 5.53 | 5.62 | 5.29 |
| 5.44 | 5.34 | 5.79 | 5.10 | 5.27 | 5.39 | 5.42 | 5.47 |
| 5.63 | 5.34 | 5.46 | 5.30 | 5.78 | 5.68 | 5.85 | |

  with sufficient statistics $n = 23$, $\bar{x} = 5.4848$, $s = 0.1882$
- We consider $X|\mu, \tau \sim N(\mu, 1/\tau)$, with $\tau$ unknown
- In Example 2.4 we assumed $X_i|\mu \sim N(\mu, 0.2^2)$ with $\mu \sim N(5.41, 0.4^2)$ and $\tau$ known with $\tau = 1/0.2^2 = 25$.
- Must specify the parameters in the $NGa(b, c, g, h)$ prior distribution for $(\mu, \tau)^T$. Suppose we choose $Var(\tau) = 250$.
- Choice of $b$ and $c$: the $NGa$ prior distribution has $\mu|\tau \sim N\{b, 1/(c\tau)\}$. Matching the prior for $\mu|\tau = 25$ gives $b = 5.41$ and $c = 0.25$

  Choice of $g$ and $h$: the $NGa$ prior distribution has $\tau \sim Ga(g, h)$. Choose $E(\tau) = 25$, giving $g = 2.5$ and $h = 0.1$

- Summary: take prior

$$\begin{pmatrix} \mu \\ \tau \end{pmatrix} \sim NGa(b = 5.41, c = 0.25, g = 2.5, h = 0.1)$$

- Is the marginal prior distribution for $\mu$ close to what we used in Example 2.4? Yes, see plot below

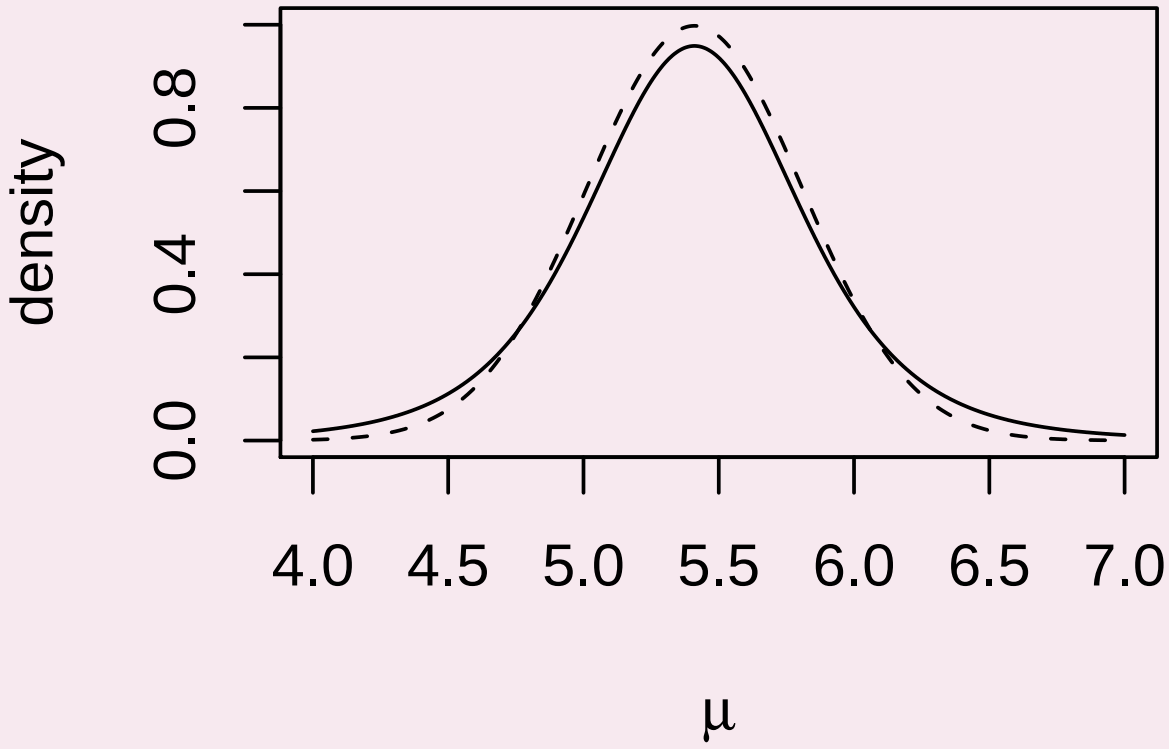



Figure: Marginal prior density for $\mu$: new version (solid) and previous version (dashed)

- Determine the posterior distribution for $(\mu, \tau)^T$. Also determine the marginal prior distribution for $\tau$ and for $\sigma$, and the marginal posterior distribution for each of $\mu$, $\tau$ and $\sigma$.

**Solution**

...

# Comparison of prior and posterior marginal distributions

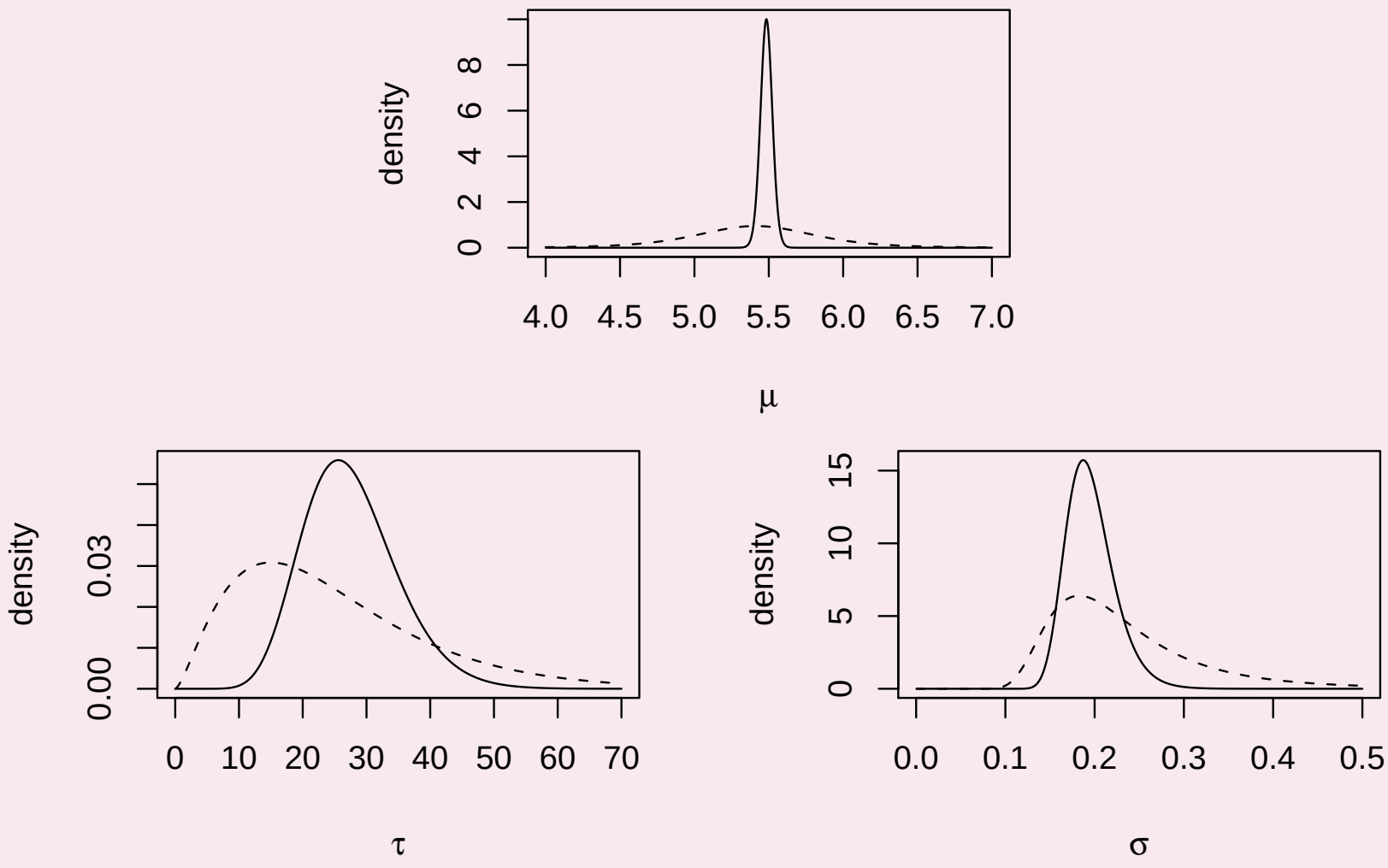



Figure: Prior (dashed) and posterior (solid) densities for $\mu$, $\tau$ and $\sigma$

# Review of contour plots for bivariate distribution

- Bivariate normal: $\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}\right\}$, with density $\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - 2\rho\frac{xy}{\sigma_x\sigma_y}\right)\right\}$.

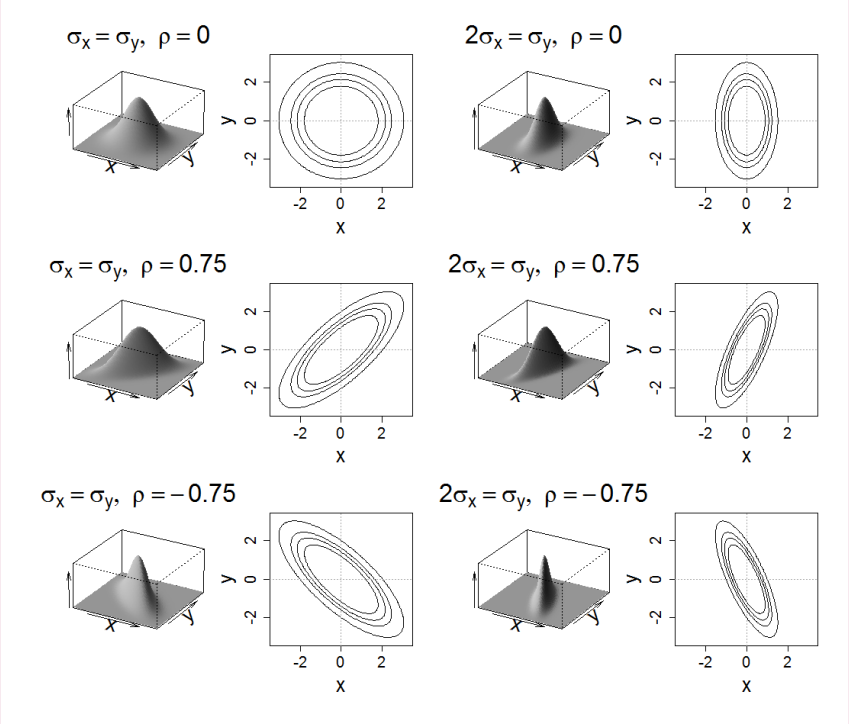



Figure: Contour plots for different bivariate normal densities

# Comparison of prior and posterior distributions

- Can plot their contours using R command NGacontour:

```
mu=seq(4.5,6.5,len=1000); tau=seq(0,71,len=1000)
NGacontour(mu,tau,b,c,g,h,lty=3); NGacontour(mu,tau,B,C,G,H,add=TRUE)
```

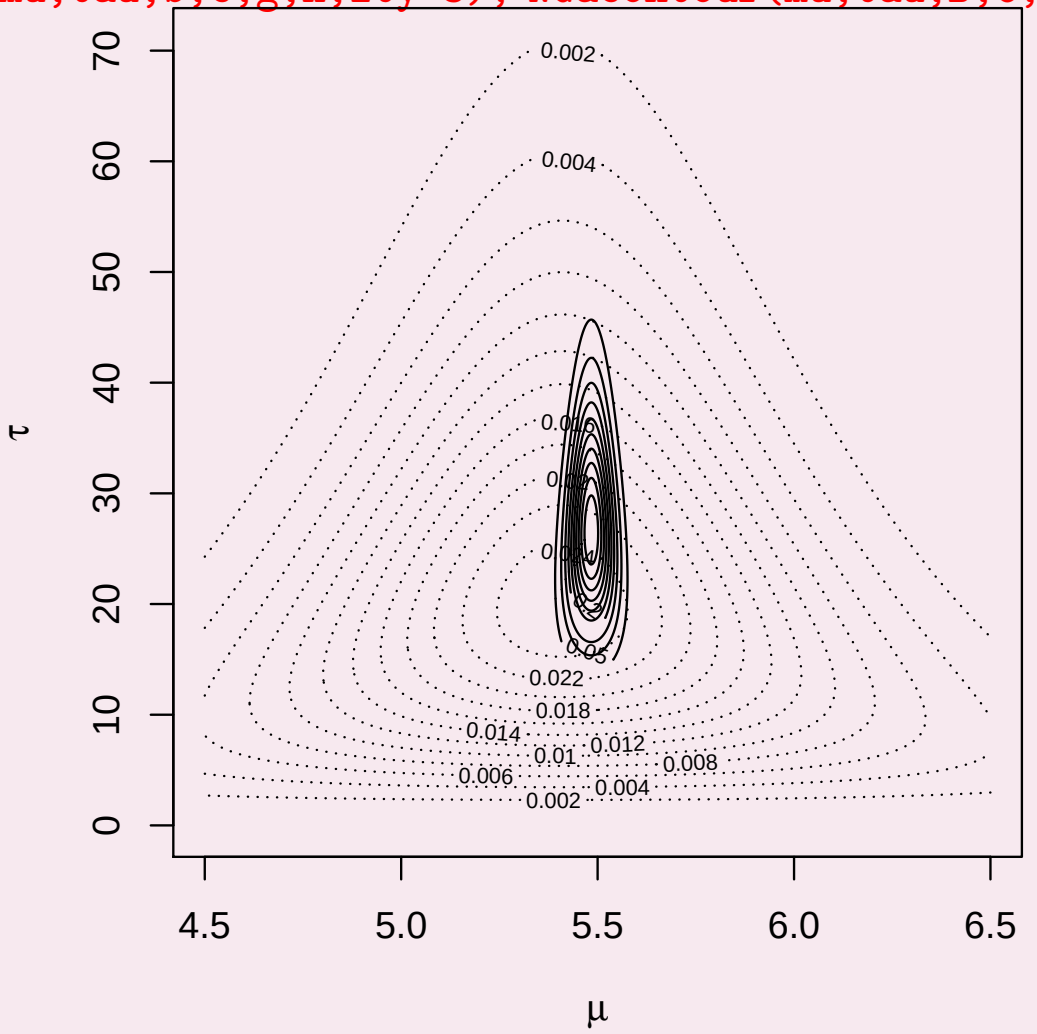



Figure: Contour plot of the prior (dashed) and posterior (solid) densities for $(\mu, \tau)^T$.

## Comments

- Wikipedia tells us that the actual mean density of the earth is $5.515\, g/cm^3$
- What is the (posterior) probability that the mean density is within 0.1 of this value?

$$\mu|\boldsymbol{x} \sim t_{28}(5.484, 0.001561) \quad \Rightarrow \quad Pr(5.415 < \mu < 5.615|\boldsymbol{x}) = 0.9529$$

using
`pgt(5.615,28,5.484,0.001561)-pgt(5.415,28,5.484,0.001561)`

- Without the data, the only basis for determining the earth's density is via the prior distribution

$$\mu \sim t_5(5.41, 0.16) \quad \Rightarrow \quad Pr(5.415 < \mu < 5.615) = 0.1802$$

using `pgt(5.615,5,5.41,0.16)-pgt(5.415,5,5.41,0.16)`

- These probability calculations demonstrate that the data have been very informative and changed our beliefs about the earth's density

# Confidence intervals and regions

## Example 3.3

Determine the $100(1-\alpha)\%$ highest density interval (HDI) for the population mean $\mu$ in terms of quantiles of the standard $t$-distribution

## Solution

$\cdots$

## Calculating 95% posterior confidence intervals using R

- $\mu|\boldsymbol{x} \sim t_{2G}\{B, H/(GC)\}$
  Symmetric distribution $\rightarrow$ HDI and equi-tailed intervals are the same
  `c(qgt(0.025,2*G,B,H/(G*C)),qgt(0.975,2*G,B,H/(G*C)))`

- $\tau|\boldsymbol{x} \sim Ga(G, H)$
  Skewed distribution $\rightarrow$ HDI and equi-tailed intervals are different
  `hdiGamma(0.95,G,H)`
  `c(qgamma(0.025,G,H),qgamma(0.975,G,H))`

- $\sigma|\boldsymbol{x} \sim \textit{Inv-Chi}(G, H)$
  Skewed distribution $\rightarrow$ HDI and equi-tailed intervals are different
  `hdiInvchi(0.95,G,H)`
  `c(qinvchi(0.025,G,H),qinvchi(0.975,G,H))`

## Results from this data analysis . . .

| | Prior | Posterior | |
|---|---|---|---|
| $\mu$: | $(4.3818, 6.4382)$ | $(5.4031, 5.5649)$ | |
| $\tau$: | $(1.4812, 55.9573)$ | $(14.0193, 42.2530)$ | $\leftarrow$ HDI |
| | $(4.1561, 64.1625)$ | $(15.0674, 43.7625)$ | |
| $\sigma$: | $(0.1062, 0.4246)$ | $(0.1466, 0.2505)$ | $\leftarrow$ HDI |
| | $(0.1248, 0.4905)$ | $(0.1512, 0.2576)$ | |

- Posterior HDI and equi-tailed intervals for $\tau$ are fairly similar but prior intervals are not
- Ditto for $\sigma$
- Why?

  The prior distributions are quite skewed but the posterior distributions are fairly symmetric

## Confidence regions

- We have looked at (marginal) HDIs
- Can be useful to also look at (joint) confidence regions

## Example 3.4

Determine a joint $\alpha$ confidence region for $(\mu, \tau)^T$ by calculating $k$ satisfying $P(\pi(\mu, \tau) > k) = \alpha$ or $P(\pi(\mu, \tau) > k \mid \boldsymbol{x}) = \alpha$.

## Solution

...

- Can plot these regions using `NGacontour` by using the appropriate value for the contour level

```
mu=seq(3.5,7.5,len=1000)
tau=seq(0,80,len=1000)
NGacontour(mu,tau,b,c,g,h,p=c(0.95,0.9,0.8),lty=3)
NGacontour(mu,tau,B,C,G,H,p=c(0.95,0.9,0.8),add=TRUE)
```

# Confidence regions for $(\mu, \tau)^T$

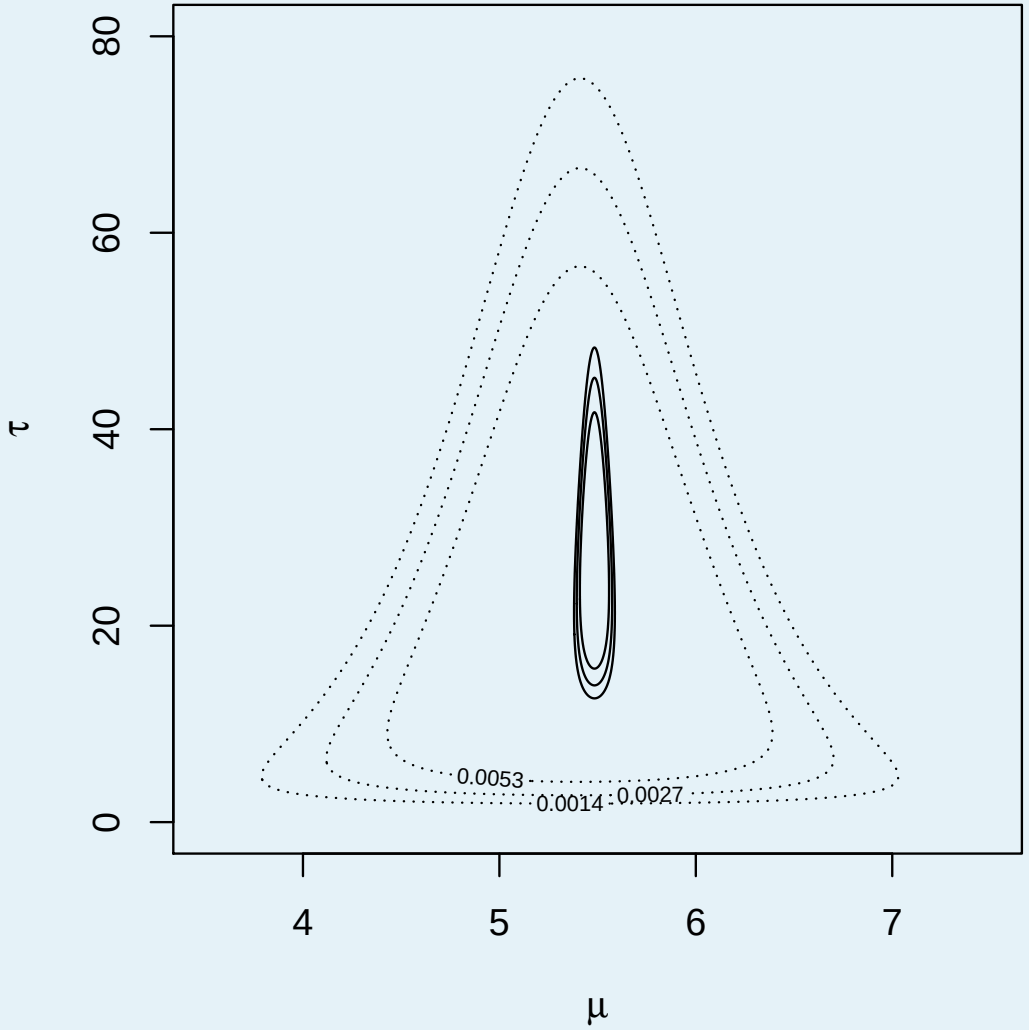



Figure: 95%, 90% and 80% prior (dashed) and posterior (solid) confidence regions for $(\mu, \tau)^T$; 95% (outer), 80% (inner).

# Focusing on central part of plot . . .

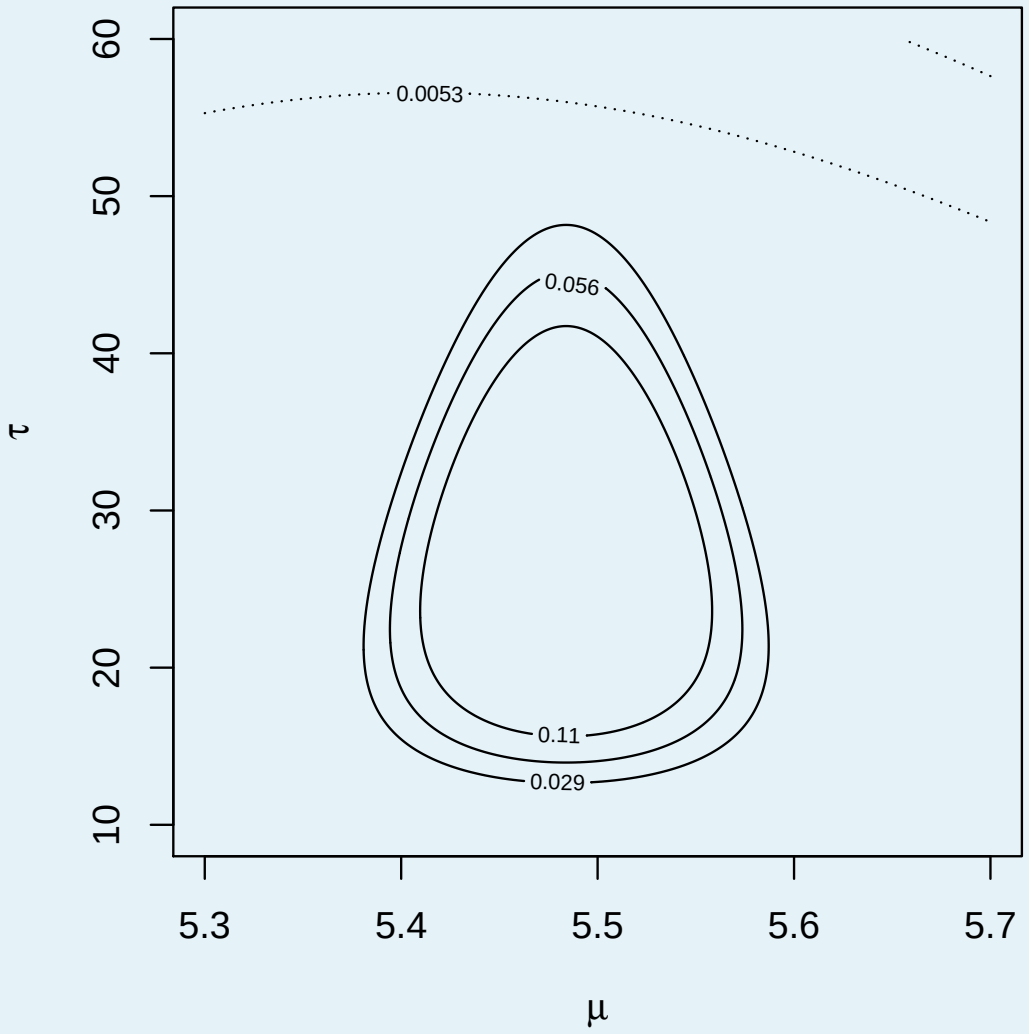



Figure: 95%, 90% and 80% prior (dashed) and posterior (solid) confidence regions for $(\mu, \tau)^T$; 95% (outer), 80% (inner).

## 3.4 Predictive distributions

- In this model we can determine the predictive distribution via

$$f(y|\boldsymbol{x}) = \int f(y|\mu,\tau)\,\pi(\mu,\tau|\boldsymbol{x})\,d\mu\,d\tau$$

  or by using Candidate's formula (as this is a conjugate analysis)
- But, for this model, there is a more straightforward way

. . .

- These predictive intervals can be calculated easily using the R function qgt
- In Example 2.2, the prior and posterior predictive HDIs for a new value $Y$ from the population are $(4.2604, 6.5596)$ and $(5.0855, 5.8825)$ respectively, calculated using

```
c(qgt(0.025,2*g,b,h*(c+1)/(g*c)),qgt(0.975,2*g,b,h*(c+1)/(g*c)))
c(qgt(0.025,2*G,B,H*(C+1)/(G*C)),qgt(0.975,2*G,B,H*(C+1)/(G*C)))
```

## 3.5 Summary

Suppose we have a normal random sample with $X_i|\mu,\tau \sim N(\mu, 1/\tau)$, $i = 1, 2, \ldots, n$ (independent)

- $(\mu,\tau)^T \sim NGa(b, c, g, h)$ is a conjugate prior distribution
- The posterior distribution is $(\mu,\tau)^T|\boldsymbol{x} \sim NGa(B, C, G, H)$ where the posterior parameters are given in the proof in Section 2.2.
- The marginal prior distributions are $\mu \sim t_{2g}\{b, h/(gc)\}$, $\tau \sim Ga(g, h)$, $\sigma = 1/\sqrt{\tau} \sim \textit{Inv-Chi}(g, h)$
- The marginal posterior distributions are $\mu|\boldsymbol{x} \sim t_{2G}\{B, H/(GC)\}$, $\tau|\boldsymbol{x} \sim Ga(G, H)$, $\sigma|\boldsymbol{x} \sim \textit{Inv-Chi}(G, H)$
- Prior and posterior means and standard deviations for $\mu$, $\tau$ and $\sigma$ can be calculated from the properties of the $t$, *Gamma* and *Inv-Chi* distributions
- Prior and posterior probabilities and densities for $\mu$, $\tau$ and $\sigma$ can be calculated using the `R` functions `pgt, dgt, pgamma, dgamma, pinvchi, dinvchi`

Suppose we have a normal random sample with $X_i|\mu, \tau \sim N(\mu, 1/\tau)$, $i = 1, 2, \ldots, n$ (independent)

- HDIs or equi-tailed CIs for $\mu$, $\tau$ and $\sigma$ can be calculated using `qgt`, `hdiGamma`, `hdiInvchi`, `qgamma`, `qinvchi`
- Contour plots of the prior and posterior densities for $(\mu, \tau)^T$ can be plotted using the `NGacontour` function
- Prior and posterior confidence regions for $(\mu, \tau)^T$ can be plotted using the `NGacontour` function
- The predictive distribution for a new observation $Y$ from the population is $Y|\boldsymbol{x} \sim t_{2G}\{B, H(C+1)/(GC)\}$ and its HDI can be calculated using the `qgt` function

# 3.6 Why do we have so many different distributions?

- So far we have used many distributions, some you will have met before and some will be new
- After a while the variety and sheer number of different distributions can become overwhelming
- Why do we need so many distributions and why do we name so many of them?

## Justification

- Statistics studies the random variation in experiments, samples and processes
- The variety of applications leads to their randomness being described by many different distributions
- In many applications, bespoke distributions need to be formulated
- However, some distributions come up time and time again for modelling random variation in data and for describing prior beliefs

## Justification (continued)

- It is helpful for us to be able to refer to these distributions – and so we give each one a name – and be able to quote known results for these distributions such as their mean and variance
- For example, in this chapter you have been introduced to
  - a generalisation of the $t$-distribution
  - the inverse chi distribution
  - The Normal-Gamma distribution
- We have been able to use results for their mean and variance to study prior and posterior distributions and have been able to plot these distributions using functions in the `R` package
- You will meet several other new distributions in the remainder of the module

## Justification (continued)

- Obviously it's useful to have a working knowledge of each of these distributions but not vital to remember all their properties
- The exam paper will contain a list of all the distributions used in the exam, together with their density (or probability function) and any useful proprieties such as their mean and variance (as needed for the exam)

# 3.7 Learning objectives

By the end of this chapter, you should be able to:

1. Determine the posterior distribution for $(\mu, \tau)^T$
2. Determine and use the univariate prior and posterior distributions
3. Determine confidence intervals, HDIs and confidence regions
4. Determine the predictive distribution of another value from the population, and its predictive interval
5. Determine the predictive distribution of the mean of another random sample from the population

both in general and for a particular prior and data set.

Also you should be able to:

6. Appreciate the benefit of naming distributions and for having lists of properties for these distributions

# Chapter 4

## Non-conjugate multi-parameter problems

# Chapter 4. Non-conjugate multi-parameter problems

- Here we will study some multi-parameter problems in which the prior distribution does not have to be conjugate
- Inferences are made by using techniques which simulate realisations from the posterior distribution
- These methods are generally referred to as Markov Chain Monte Carlo techniques, and often abbreviated to MCMC
- Two main MCMC techniques:
  1. the Gibbs sampler – at the forefront of the recent MCMC revolution
  2. Metropolis-Hastings sampling
- MCMC schemes based on the combination of these two fundamental techniques are still at the forefront of MCMC research

# 4.1 Why is non-conjugate inference not straightforward?

## Example 4.1

- Consider again the problem in section 3.2

  $X_i|\mu,\tau \sim N(\mu, 1/\tau)$, $i = 1, 2, \ldots, n$ (independent)
- Here we showed that a $NGa$ prior for $(\mu,\tau)^T$ was conjugate
- But what if a $NGa(b, c, g, h)$ prior distribution does not adequately represent our prior beliefs?
- Suppose instead that our prior beliefs are represented by independent priors for the parameters

$$\mu \sim N\left(b, \frac{1}{c}\right) \qquad \text{and} \qquad \tau \sim Ga(g, h)$$

What is the posterior distribution for $(\mu,\tau)^T$?

## Solution

...

## In this case ...

- What is the posterior mean of $\mu$ and of $\tau$?
- What are their marginal distributions?
- How can we calculate the moments $E(\mu^{m_1}\tau^{m_2}|\boldsymbol{x})$ of this posterior distribution?

## Marginal posterior density for $\mu$

$$\pi(\mu|\boldsymbol{x}) = \int_0^\infty \pi(\mu,\tau|\boldsymbol{x})\,d\tau$$

$$= \frac{\displaystyle\int_0^\infty \tau^{g+\frac{n}{2}-1}\exp\left\{-\frac{c}{2}(\mu-b)^2 - h\tau - \frac{n\tau}{2}\left[s^2+(\bar{x}-\mu)^2\right]\right\}d\tau}{\displaystyle\int_{-\infty}^\infty\int_0^\infty \tau^{g+\frac{n}{2}-1}\exp\left\{-\frac{c}{2}(\mu-b)^2 - h\tau - \frac{n\tau}{2}\left[s^2+(\bar{x}-\mu)^2\right]\right\}d\tau\,d\mu}$$

## Marginal posterior density for $\tau$

$$\pi(\tau|\boldsymbol{x}) = \int_{-\infty}^{\infty} \pi(\mu, \tau|\boldsymbol{x})\, d\mu$$

$$= \frac{\displaystyle\int_{-\infty}^{\infty} \tau^{g+\frac{n}{2}-1} \exp\left\{-\frac{c}{2}(\mu-b)^2 - h\tau - \frac{n\tau}{2}\left[s^2 + (\bar{x}-\mu)^2\right]\right\} d\mu}{\displaystyle\int_{-\infty}^{\infty}\int_{0}^{\infty} \tau^{g+\frac{n}{2}-1} \exp\left\{-\frac{c}{2}(\mu-b)^2 - h\tau - \frac{n\tau}{2}\left[s^2 + (\bar{x}-\mu)^2\right]\right\} d\tau\, d\mu}$$

## General moments

$$E(\mu^{m_1}\tau^{m_2}|\boldsymbol{x})$$

$$= \int_{-\infty}^{\infty}\int_{0}^{\infty} \mu^{m_1}\tau^{m_2}\pi(\mu, \tau|\boldsymbol{x})\, d\tau\, d\mu$$

$$= \frac{\displaystyle\int_{-\infty}^{\infty}\int_{0}^{\infty} \mu^{m_1}\tau^{m_2} \times \tau^{g+\frac{n}{2}-1} \exp\left\{-\frac{c}{2}(\mu-b)^2 - h\tau - \frac{n\tau}{2}\left[s^2 + (\bar{x}-\mu)^2\right]\right\} d\tau\, d\mu}{\displaystyle\int_{-\infty}^{\infty}\int_{0}^{\infty} \tau^{g+\frac{n}{2}-1} \exp\left\{-\frac{c}{2}(\mu-b)^2 - h\tau - \frac{n\tau}{2}\left[s^2 + (\bar{x}-\mu)^2\right]\right\} d\tau\, d\mu}$$

## Comments

- These integrals cannot be determined analytically
- It is possible to use numerical integration methods or approximations (for large $n$)
- However, in general, the accuracy of the numerical approximation to the integral deteriorates as the dimension of the integral increases
- How do we analyse models with a large number of parameters?

## Comment

- Not using a conjugate prior distribution can cause many basic problems such as difficulty in plotting the marginal posterior density or determining posterior moments
- But having to use conjugate priors is far too restrictive for many real data analyses:
  (i) our prior beliefs may not be captured using a conjugate prior
  (ii) most models for complex data do not have conjugate priors
- Until relatively recently, practical Bayesian inference for real complex problems was either not feasible or only undertaken by the dedicated few prepared to develop bespoke computer code to numerically evaluate all the integrals etc.

# Examples

- Poisson distribution for our data, $X_i|\lambda \sim Po(\lambda)$:

$$\begin{aligned} \lambda \sim Ga(g,h) \quad &\longrightarrow \quad \lambda|\boldsymbol{x} \sim Ga(g+n\bar{x}, h+n) \\ \lambda \sim Weibull(a,b) \quad &\longrightarrow \quad \lambda|\boldsymbol{x} \sim \textbf{??} \end{aligned}$$

- Normal distribution for our data, $X_i|\mu,\tau \sim N(\mu, 1/\tau)$:

$$\begin{aligned} &\mu|\tau \sim N\left(b, \frac{1}{c\tau}\right), \tau \sim Ga(g,h) \\ &\quad \longrightarrow (\mu,\tau)^T \sim NGa(b,c,g,h) \quad \longrightarrow \quad (\mu,\tau)^T|\boldsymbol{x} \sim NGa(B,C,G,H) \\ &\mu \sim N\left(b, \frac{1}{c}\right); \quad \tau \sim Ga(g,h) \quad \longrightarrow \quad (\mu,\tau)^T|\boldsymbol{x} \sim \textbf{??} \end{aligned}$$

# Examples

- Poisson distribution for our data, $X_i|\lambda \sim Po(\lambda)$:

$$\lambda \sim Ga(g,h) \quad \longrightarrow \quad \lambda|\boldsymbol{x} \sim Ga(g+n\bar{x}, h+n)$$
$$\lambda \sim Weibull(a,b) \quad \longrightarrow \quad \lambda|\boldsymbol{x} \sim \textbf{??}$$

- Normal distribution for our data, $X_i|\mu,\tau \sim N(\mu, 1/\tau)$:

$$\mu|\tau \sim N\left(b, \frac{1}{c\tau}\right), \tau \sim Ga(g,h)$$
$$\longrightarrow (\mu,\tau)^T \sim NGa(b,c,g,h) \quad \longrightarrow \quad (\mu,\tau)^T|\boldsymbol{x} \sim NGa(B,C,G,H)$$
$$\mu \sim N\left(b, \frac{1}{c}\right); \quad \tau \sim Ga(g,h) \quad \longrightarrow \quad (\mu,\tau)^T|\boldsymbol{x} \sim \textbf{??}$$

**All hell breaks loose.**

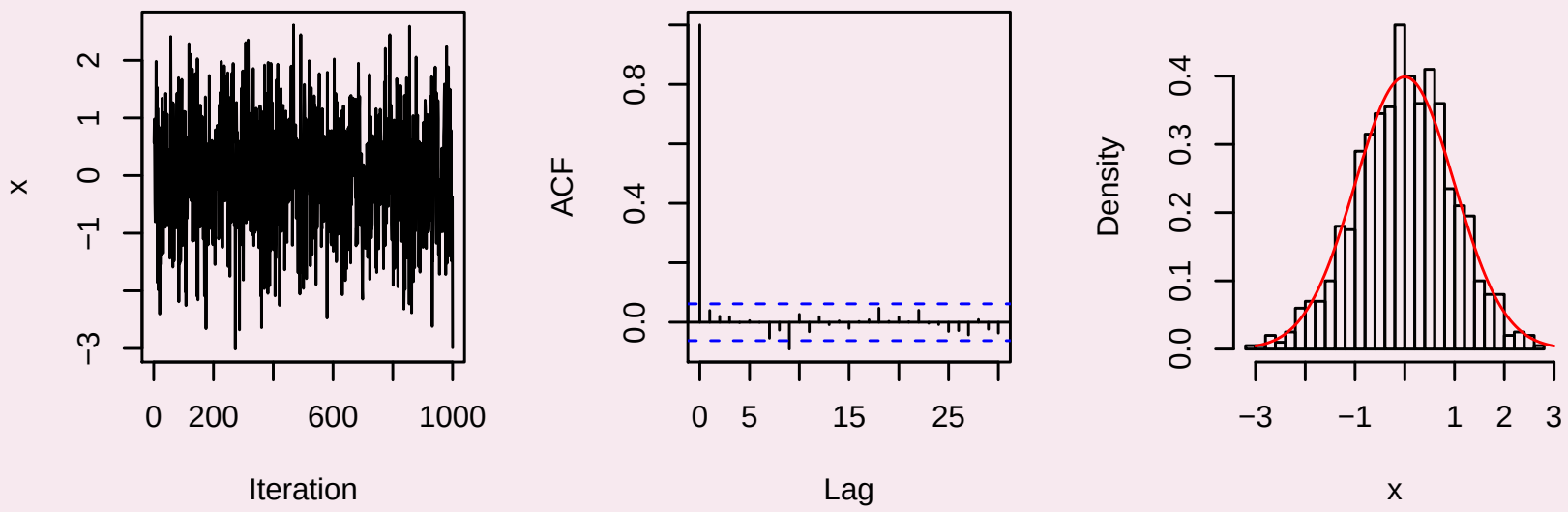
# 4.2 Simulation-based inference

- Can get around the problem of having to work out integrals
- Base inferences on simulated realisations from the posterior distribution
- This is the fundamental idea behind **Markov Chain Monte Carlo** (MCMC) methods
- If we could simulate from the posterior distribution then we could use a very large sample of realisations to determine posterior means, standard deviations, correlations, joint densities, marginal densities etc.

**Black box simulator demo in R**

*(Ignore the upcoming slides if you'd rather try out the R code!)*

- Imagine you wanted to know about the standard normal distribution – its shape, its mean, its standard deviation
- But didn't know any mathematics so that you couldn't derive say the distribution's zero mean and unit variance
- However you've been given a "black box" which can simulate realisations from this distribution
- Here we'll use the `R` function `rnorm()` as the black box simulator
- If you decide to generate 1K realisations the output might look something like



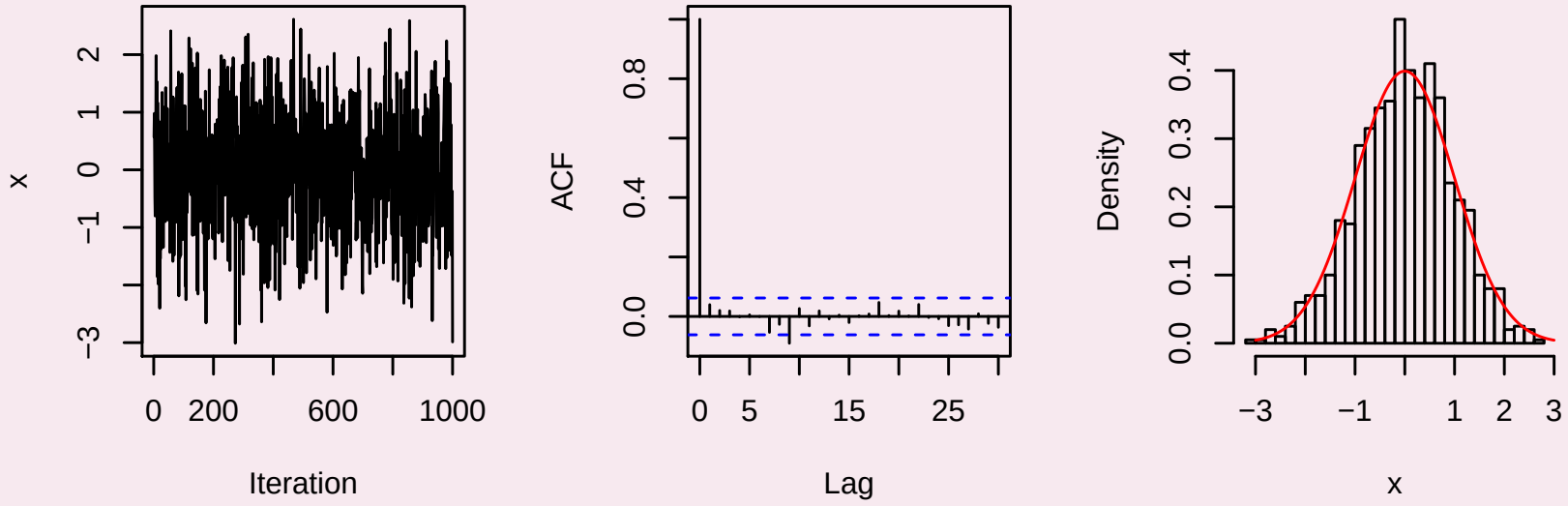
## In these plots

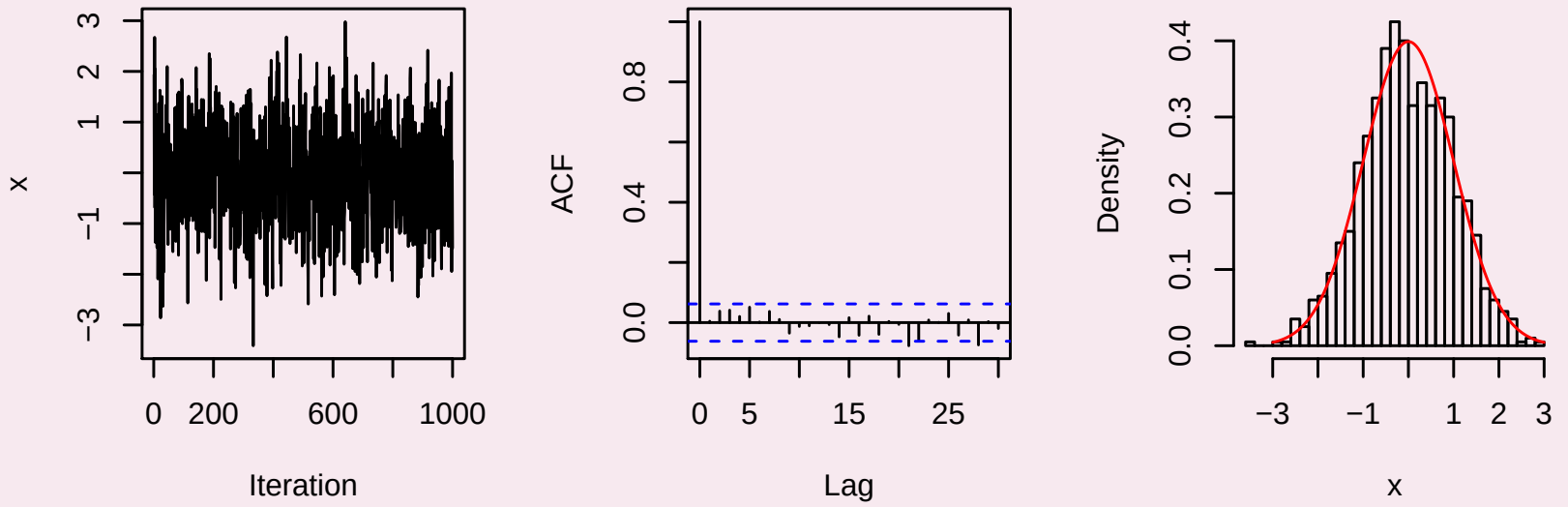
Left: the trace plot of the output – the realisations from the black box sampler in the order they are produced

Mid: the autocorrelation (ACF) plot – shows the correlation between $\{x_1, \ldots, x_{N-k}\}$ and $\{x_{k+1}, \ldots, x_N\}$ for $k \geq 0$. Call $k$ as the lag. It shows how correlated the realisations are at different lags. The lag 0 autocorrelation $corr(x_i, x_i)$ must be one (by definition). Here the (sample) correlation between say consecutive values $corr(x_i, x_{i+1})$ will be almost zero, which shows that the simulator `rnorm()` produces independent realisations. This is also the case for correlations at all positive lags. This sample ACF plot tells that the simulator generates independent sample.

Right: the density histogram of the realisations. This tells that what the standard normal distribution is like.

- If you simulate another 1K realisations then the output you get will be a different set of realisations but from the same distribution
- Both sets of realisations might look like



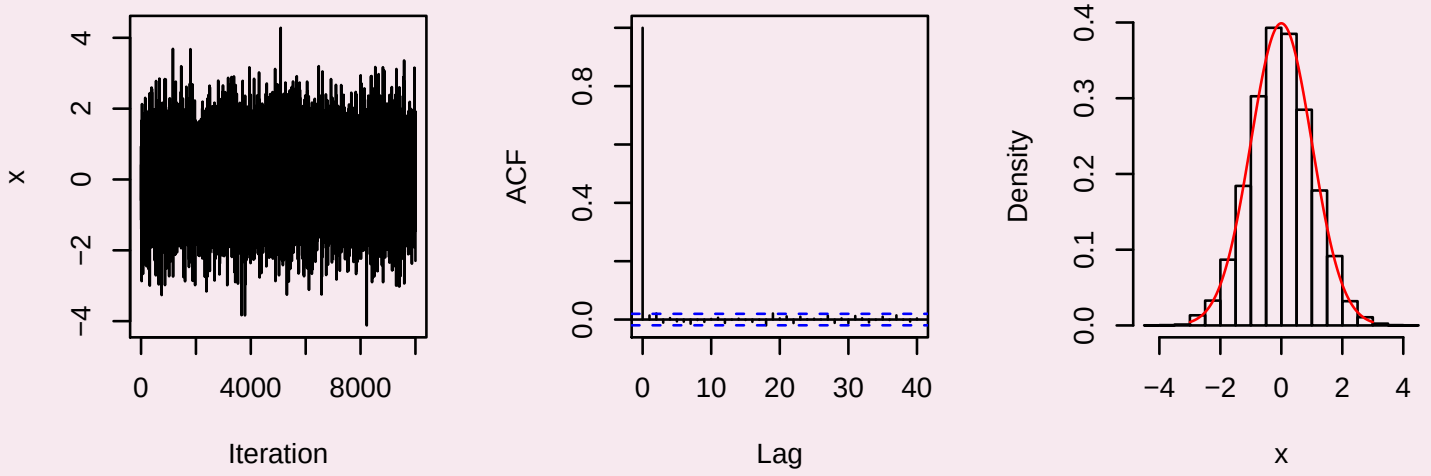
- We can also estimate various quantities of the standard normal distribution from these simulated datasets of 1K realisations

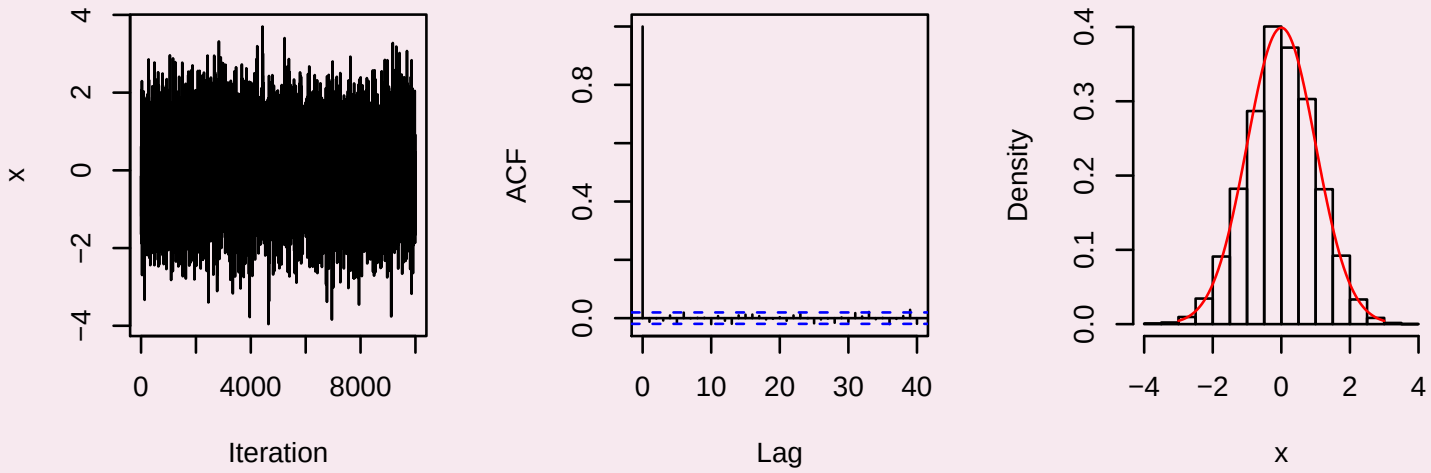
| | 1st Qu. | Median | Mean | 3rd Qu. | St.Dev. |
|---|---|---|---|---|---|
| Dataset 1 | -0.65240 | -0.00130 | -0.00192 | 0.64810 | 0.96049 |
| Dataset 2 | -0.69880 | -0.09637 | -0.03274 | 0.67330 | 0.99599 |

- The numerical summaries are slightly different but essentially tell the same story
- In each Dataset, the mean and median are around zero and the standard deviation is around one.
- We know from previous modules that there is sample variability in estimates of means from random samples

- Now look at two datasets with 10K simulated realisations:

| | 1st Qu. | Median | Mean | 3rd Qu. | St.Dev. |
|---|---|---|---|---|---|
| Dataset 3 | -0.66820 | -0.00048 | 0.00370 | 0.68070 | 0.99593 |
| Dataset 4 | -0.67130 | 0.01354 | 0.01008 | 0.67920 | 1.00691 |



- Now estimates have much less sampling variability due to the larger sample size
- In fact we can estimate any "population" quantity to any required accuracy simply by simulating a large enough collection of realisations

- These analyses show how we can make inferences, calculate means, variances, densities etc by using realisations from a distribution. In the rest of this chapter, we will look into how we can construct algorithms for simulating from (complex) posterior distributions, from which we can then make inferences

# 4.3 Motivation for MCMC methods

- We consider a generic case where we want to simulate realisations of two random variables $X$ and $Y$ with joint density $f(x, y)$
- This joint density can be factorised as

$$f(x, y) = f(x)\, f(y|x)$$

  and so we can simulate from $f(x, y)$ by first simulating $X = x$ from $f(x)$, and then simulating $Y = y$ from $f(y|x)$
- On the other hand, we can write

$$f(x, y) = f(y)\, f(x|y)$$

  and so simulate $Y = y$ from $f(y)$ and then $X = x$ from $f(x|y)$
- Assume that simulating from $f(y|x)$ and $f(x|y)$ is straightforward

# Example from MAS2906

Suppose $X \sim U(0.1, 0.5)$ and

$$f_{Y|X}(Y|X = x) = \begin{cases} xe^{-xy}, & \text{when } y \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Write an R function to generate a $2 \times n$ matrix of samples from this distribution, and produce a scatter plot.
Note that

$$Y|X = x \sim Exp(x).$$

```
example.sim = function(n) {
    output = matrix(0, 2, n)
    x = runif(n, 0.1, 0.5)
    y = rexp(n, rate = x)
    output[1, ] = x
    output[2, ] = y
    output
}
```

# Another example from MAS2906

Suppose $(X, Y)^T \sim N_2(\boldsymbol{\mu}, \Sigma)$;

- $\boldsymbol{\mu} = (\mu_x, \mu_y)^T$
- $\Sigma = \begin{pmatrix} \sigma_x^2 & \sigma_{xy}^2 \\ \sigma_{xy}^2 & \sigma_y^2 \end{pmatrix}$

What does the theory tell us about $X$, $Y$, $X|Y$ and $Y|X$?

1. $X \sim N(\mu_x, \sigma_x^2)$, $Y \sim N(\mu_y, \sigma_y^2)$
2. $X|Y = y \sim N\left(\mu_x + \rho\frac{\sigma_x}{\sigma_y}(y - \mu_y), \sigma_x^2(1 - \rho^2)\right)$ (same for $Y|X = x$)

*Recall that the correlation, $\rho$, is*

$$\rho = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y} \longrightarrow \sigma_{xy}^2 = \rho\sigma_x\sigma_y.$$

# Another example from MAS2906

- Suppose we want to investigate the marginal distributions of $X$ and $Y$
- Suppose further that we did not know much theory about the bivariate Normal; we simply knew the results in (2) telling us about the conditional distributions
- Could use a **Gibbs sampler**, in place of statistical theory, to investigate the marginals for $X$ and $Y$!
- Example:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2\left(\begin{bmatrix} 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right).$$

- We see that $\mu_x = 2, \mu_y = 3; \sigma_x = 1, \sigma_y = 1, \sigma_{xy}^2 = 0.5 \rightarrow \rho = 0.5$
- From result 2, we can write down that

$$\begin{aligned} X|Y = y &\sim N(2 + 0.5(y-3), 0.75) \quad \text{and} \\ Y|X = x &\sim N(3 + 0.5(x-2), 0.75). \end{aligned}$$

- What about the marginals for $X$ and $Y$?

*Recall that we know from the theory that $X \sim N(2,1)$ and $Y \sim N(3,1)$.*

# Another example from MAS2906

**Bivariate Normal demo in R**

# 4.4 The Gibbs sampler: Formal set-up

- Suppose we want to generate realisations from the posterior density $\pi(\boldsymbol{\theta}|\boldsymbol{x})$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^T$, and that we can simulate from the full conditional distributions (FCDs)

$$\pi(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_p, \boldsymbol{x}) = \pi(\theta_i|\cdot), \qquad i = 1, 2, \ldots, p$$

- The Gibbs sampler follows the following algorithm:
  1. Initialise the iteration counter to $j = 1$.
     Initialise the state of the chain to $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_p^{(0)})^T$.
  2. Obtain a new value $\boldsymbol{\theta}^{(j)}$ from $\boldsymbol{\theta}^{(j-1)}$ by successive generation of values

$$\begin{aligned}
\theta_1^{(j)} &\sim \pi(\theta_1|\theta_2^{(j-1)}, \theta_3^{(j-1)}, \ldots, \theta_p^{(j-1)}, \boldsymbol{x}) \\
\theta_2^{(j)} &\sim \pi(\theta_2|\theta_1^{(j)}, \theta_3^{(j-1)}, \ldots, \theta_p^{(j-1)}, \boldsymbol{x}) \\
\vdots \qquad & \qquad \vdots \qquad\qquad \vdots \\
\theta_p^{(j)} &\sim \pi(\theta_p|\theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_{p-1}^{(j)}, \boldsymbol{x})
\end{aligned}$$

  3. Change counter $j$ to $j + 1$, and return to step 2.

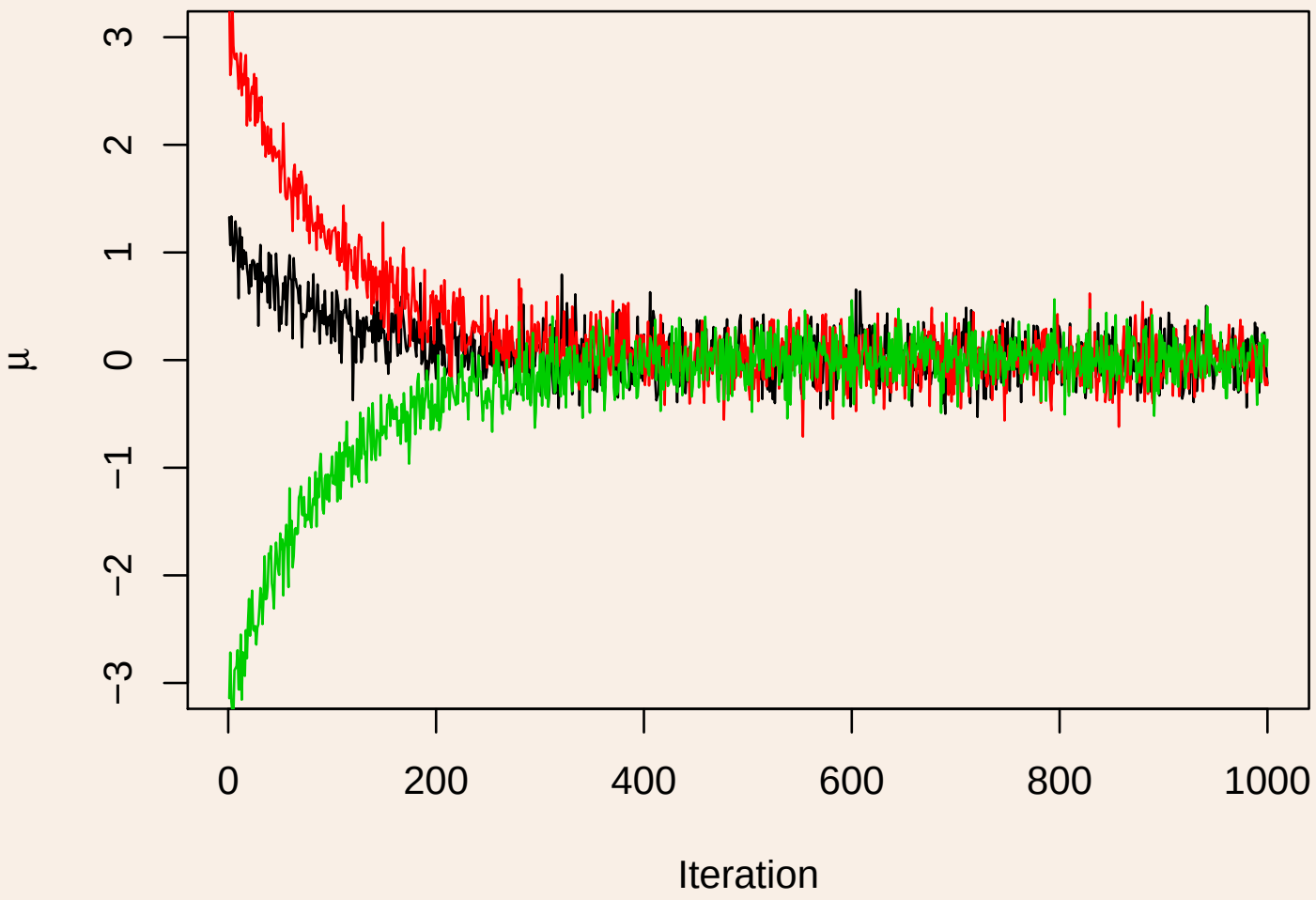
## Comments

- This algorithm defines a homogeneous Markov chain as each simulated value depends only on the previous simulated value and not on any other previous values or the iteration counter $j$
- It can be shown that $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ is the stationary distribution of this chain and so if we simulate realisations by using a Gibbs sampler, eventually the Markov chain will converge to the required posterior distribution

## More comments

- This is an iterative scheme
- It needs a period to get to convergence: the *burn-in* period
- As it is a Markov chain, successive iterates are not independent
- We can get accurate values for $E(\boldsymbol{\theta}|\boldsymbol{x})$, $SD(\boldsymbol{\theta}|\boldsymbol{x})$ or even $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ by running the sampler for a long time

# Illustration of burn-in




- Different starting points
- But distributions of the realisations after iteration 500 are the "same"

# Illustration of burn-in in two dimensions

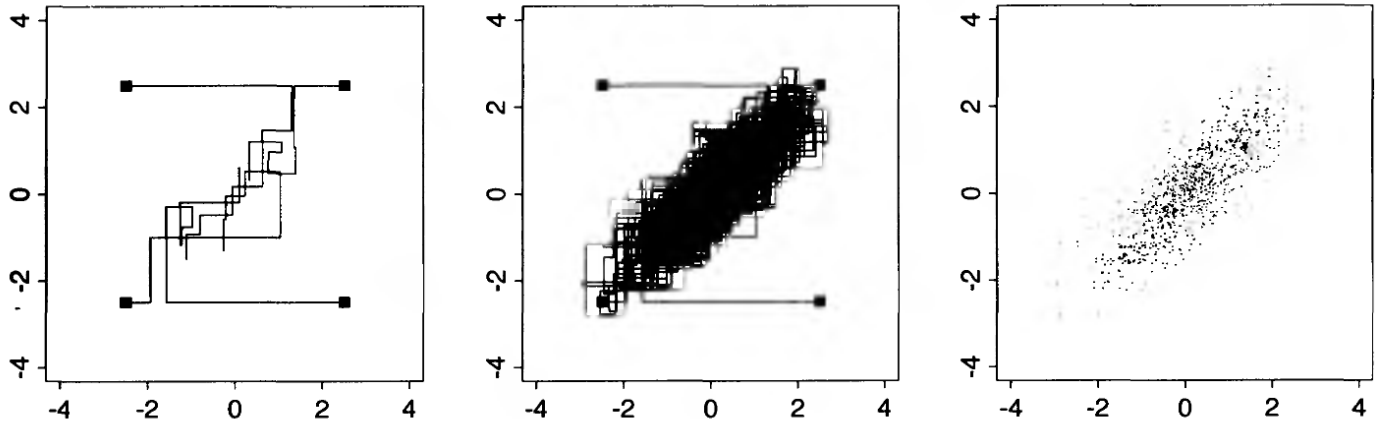



Figure 11.3 *Four independent sequences of the Gibbs sampler for a bivariate normal distribution with correlation $\rho = 0.8$, with overdispersed starting points indicated by solid squares. (a) First 10 iterations, showing the component-by-component updating of the Gibbs iterations. (b) After 500 iterations, the sequences have reached approximate convergence. Figure (c) shows the iterates from the second halves of the sequences.*

- Different starting points
- Chains converge to the "same" distribution

## 4.4.1 Processing output from a Gibbs sampler: Burn-in

- Consider a $p = 2$ parameter problem, with $\boldsymbol{\theta} = (\mu, \tau)^T$
- Run Gibbs sampler for $N$ iterations after burn-in, giving

$$\{(\mu^{(1)}, \tau^{(1)}), (\mu^{(2)}, \tau^{(2)}), \ldots, (\mu^{(N)}, \tau^{(N)})\}$$

- Can calculate features of the posterior distribution using their sample equivalents: $\bar{\mu}$, $\bar{\tau}$, $s_\mu^2$, $s_\tau^2$ and $r_{\mu\tau}$

## How accurate are these "estimates"?

- Difficult to determine exactly as they are not a *random sample*
- If it was a random sample then (for large $N$) approximate 95% CI for $\mu$ is

$$\bar{\mu} \pm \frac{1.96 s_\mu}{\sqrt{N}}$$

- But we don't have a random sample! What to do?

## 4.4.1 Processing output from a Gibbs sampler: Autocorrelation

- **Method one:** Assume the output follows some Markov chain model, and use CI of this model to approximate the true CI.
- E.g. the autoregressive model with order 1, AR(1), is

$$\mu^{(t)} - \mu = \theta(\mu^{(t-1)} - \mu) + \varepsilon_t, \text{ where } \varepsilon_t \sim N(0,1).$$

  The approximate 95% CI of $\mu$ depends on the sample autocorrelation $r(1) = Corr(\mu^{(t)}, \mu^{(t+1)})$, and is

$$\bar{\mu} \pm \frac{1.96 s_\mu}{\sqrt{N\{1 - r(1)\}^2}}.$$

  Effective sample size: $N_{eff} = N\{1 - r(1)\}^2$.
- In general, MCMC output with positive autocorrelations has $N_{eff} < N$.

## How to deal with autocorrelation

- **Method two (more popular):** `thin` the output – don't take every iterate
- Sample autocorrelation function with lag $k$ for $\mu$ is

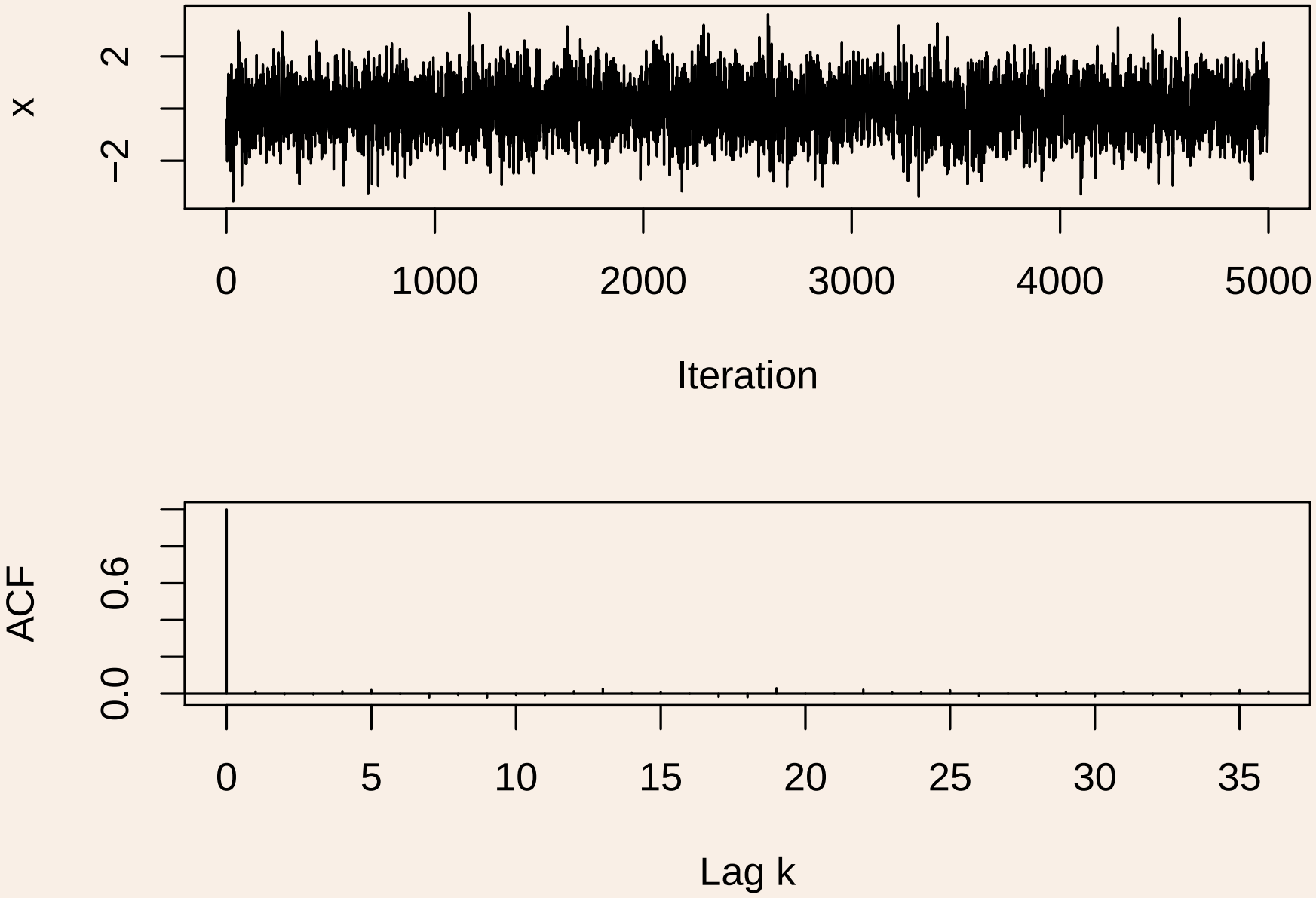
$$r(k) = Corr(\mu^{(j)}, \mu^{(j+k)}).$$

- For example, if $r(k) \simeq 0$, $k \geq 2$ then taking every 2nd iterate would give a posterior sample

$$\{(\mu^{(1)}, \tau^{(1)}), (\mu^{(3)}, \tau^{(3)}), \ldots, (\mu^{(2j+1)}, \tau^{(2j+1)}, \ldots)\}$$

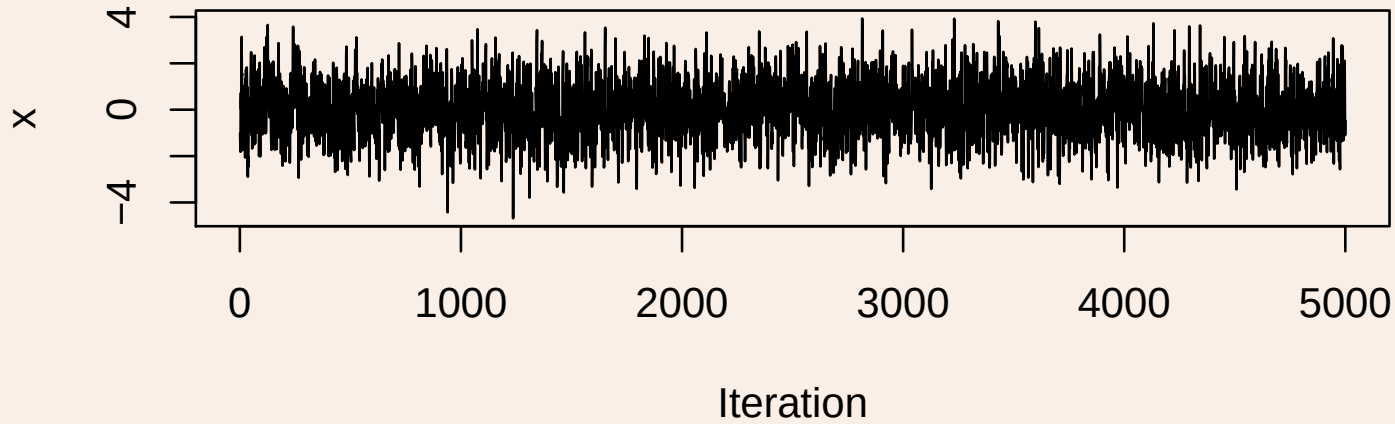
  with autocorrelation function $r^*(k) = 0$, $k \geq 1$, that is, this output has very low autocorrelation
- Generally we look at the ACF plot and choose the level of thinning needed to give low autocorrelations
- We will see how this works in some examples

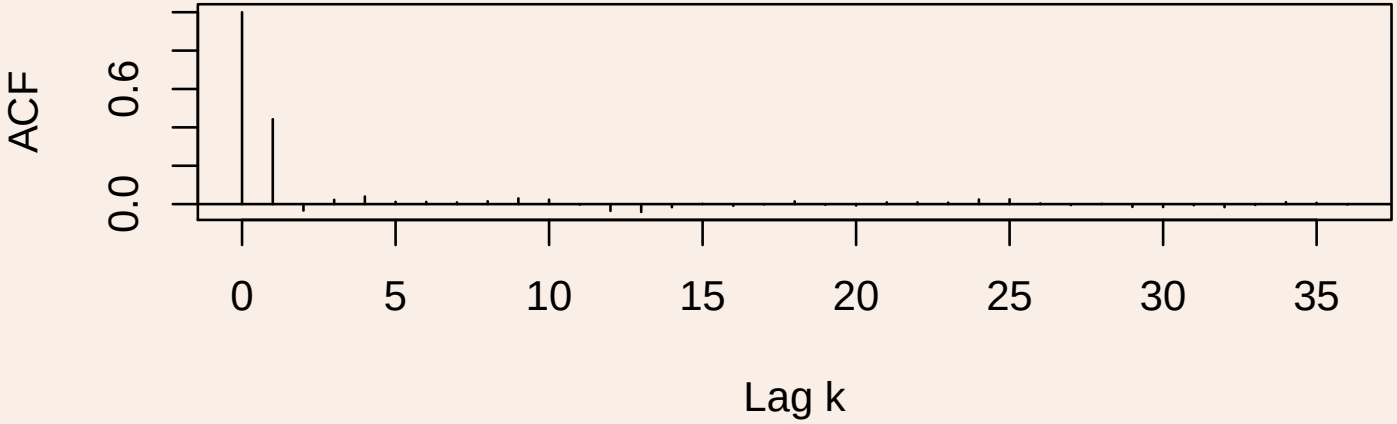
## $N(0,1)$ random sample



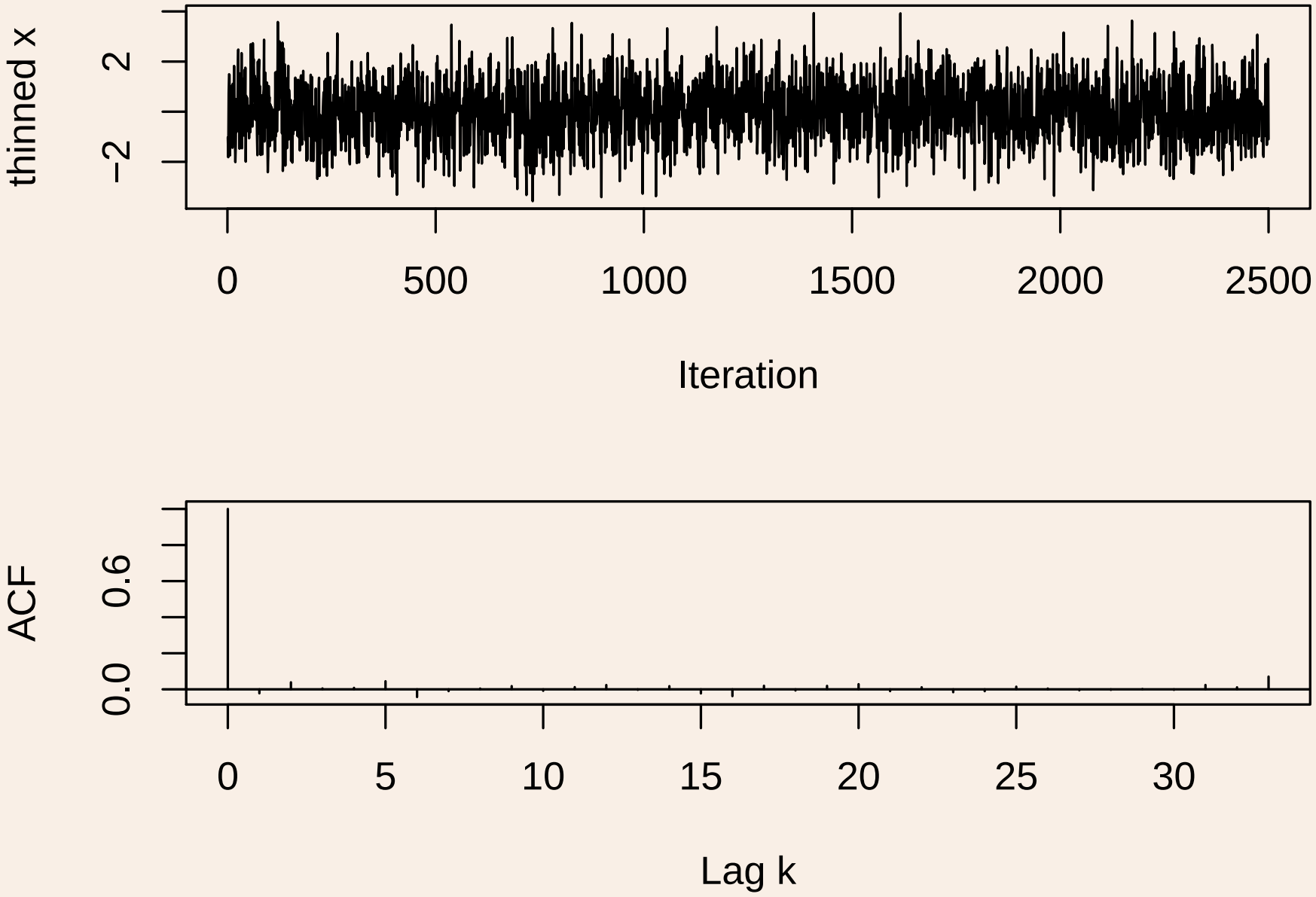
# Sample from a $MA(1)$ process

$$x_t = \varepsilon_t + \theta\varepsilon_{t-1}, \text{ where } \varepsilon_t \sim N(0,1), \text{and } r(1) = \theta, r(k) = 0, k \geq 2.$$



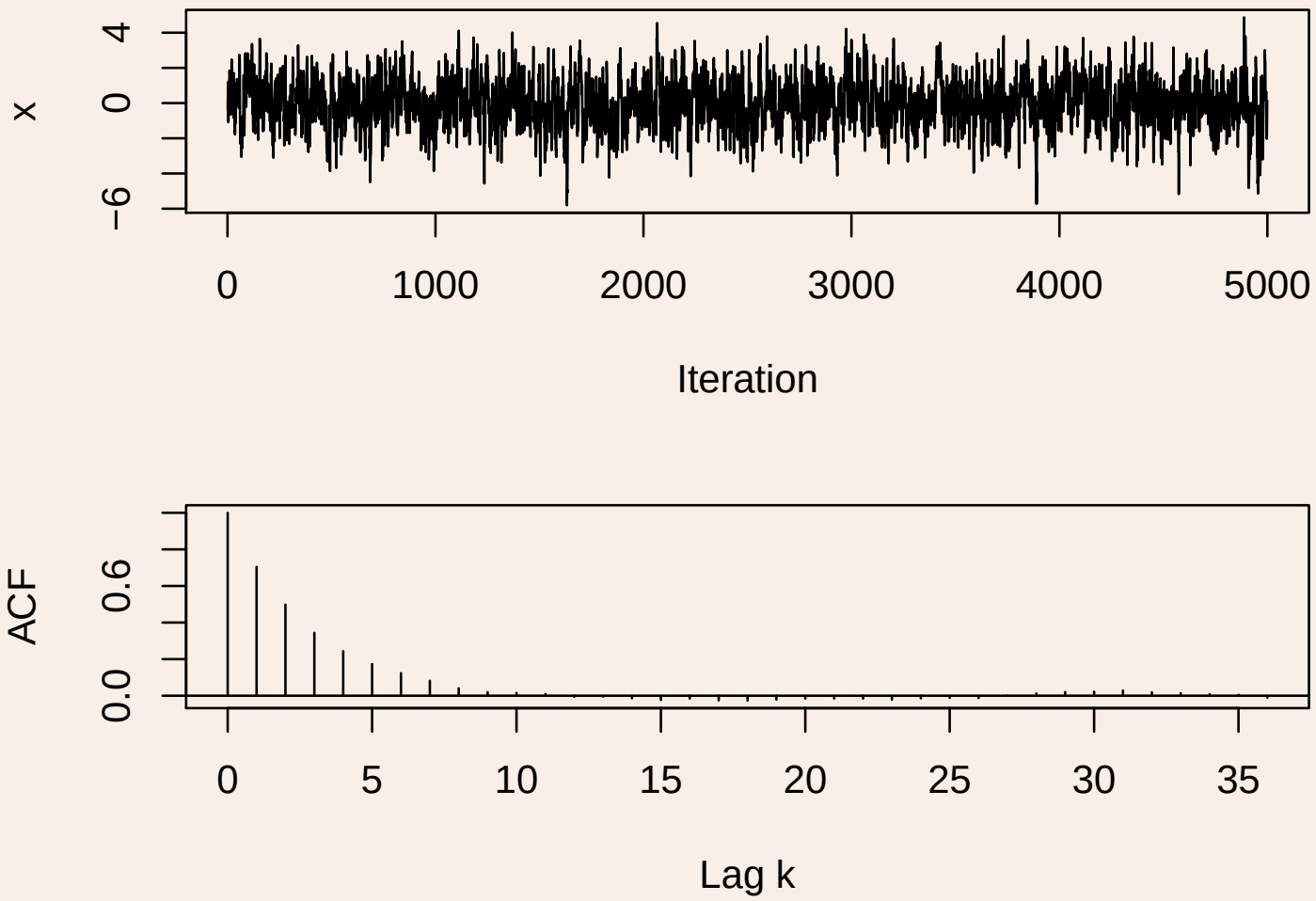



# Thinned sample from a $MA(1)$ process ($\texttt{thin} = 2$)

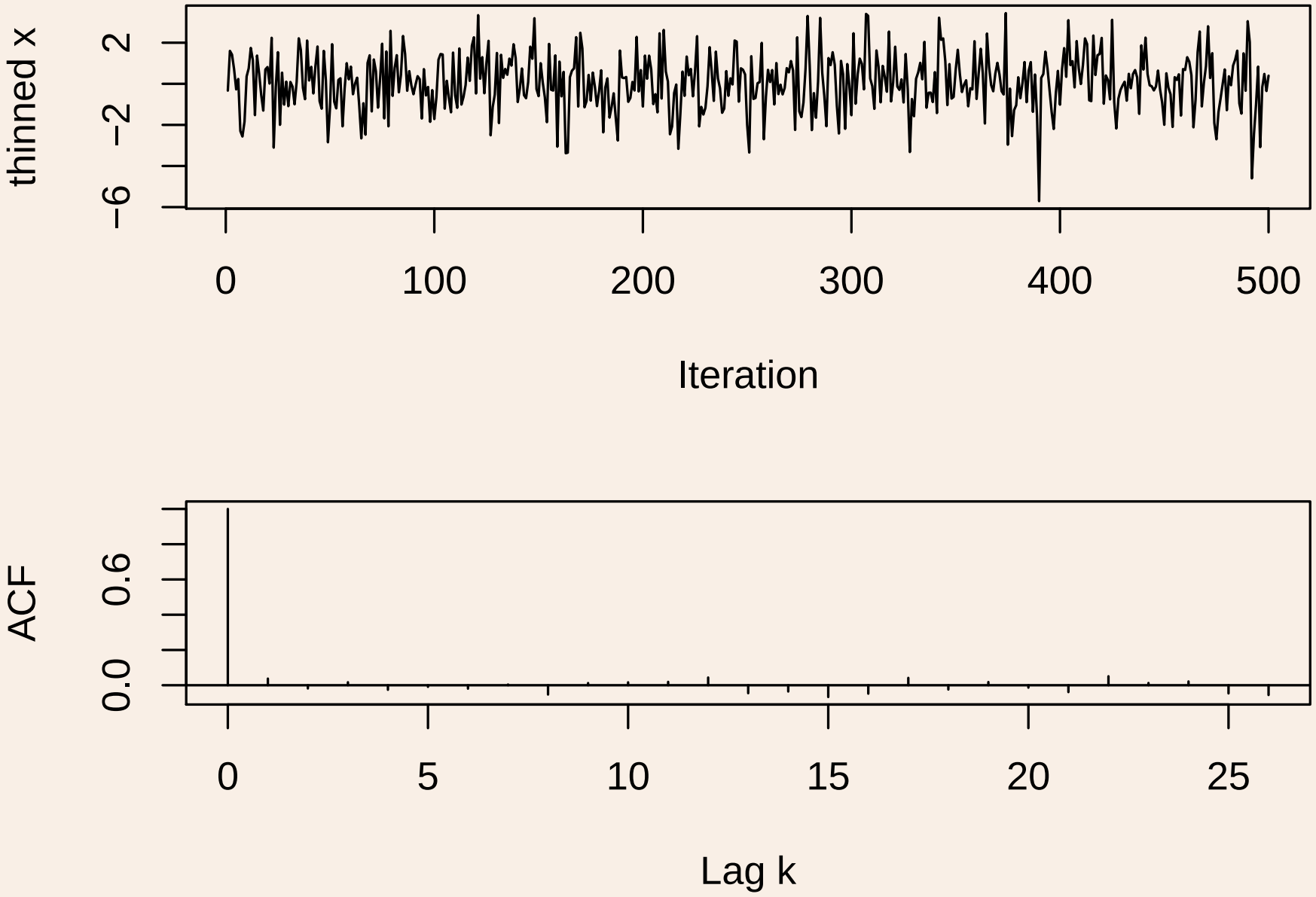


# Sample from a $AR(1)$ process

$$x_t = \theta x_{t-1} + \varepsilon_t, \text{ where } \varepsilon_t \sim N(0,1), \text{and } r(k) \simeq r(1)^k, k \geq 2.$$



# Thinned sample from a $AR(1)$ process ($\texttt{thin} = 10$)



## Strategy

1. Determine the *burn-in* period, after which the Gibbs sampler has reached its stationary distribution
   This may involve thinning the posterior sample as slowly snaking trace plots may be due to high autocorrelations rather than a lack of convergence
2. After this, determine the level of thinning to obtain a posterior sample whose autocorrelations are roughly zero
3. Repeat steps 1 and 2 several times using different initial values to make sure that the sample really is from the posterior distribution

# 4.4.2 Bayesian inference using a Gibbs sampler

- Consider again the problem in Chapter 3:
  $X_i|\mu, \tau \sim N(\mu, 1/\tau)$, $i = 1, 2, \ldots, n$ (independent)
- Here we showed that a $NGa$ prior for $(\mu, \tau)^T$ was conjugate
- But what if a $NGa(b, c, g, h)$ prior distribution does not adequately represent our prior beliefs?
- Suppose instead that our prior beliefs are represented by independent priors for the parameters

$$\mu \sim N\left(b, \frac{1}{c}\right) \qquad \text{and} \qquad \tau \sim Ga(g, h)$$

What is the posterior distribution for $(\mu, \tau)^T$?

# 4.4.2 Bayesian inference using a Gibbs sampler

## Solution

. . .

- No longer conjugate – lots of horrible integrals to work out if we want to plot the posterior, or work out posterior summaries
- Until the mid-1990s this made Bayesian inference in the non-conjugate case almost impossible

# 4.4.2 Bayesian inference using a Gibbs sampler

## Example 4.2

Construct a Gibbs sampler for the posterior distribution in Example 4.1:

- Data: $X_i|\mu, \tau \sim N(\mu, 1/\tau)$, $i = 1, 2, \ldots, n$ (independent)
- Prior:

$$\mu \sim N\left(b, \frac{1}{c}\right) \qquad \text{and} \qquad \tau \sim Ga(g, h), \qquad \text{independent}$$

## Solution

...

## Example 4.3

- Data: $X_i|\alpha, \lambda \sim Ga(\alpha, \lambda)$, $i = 1, 2, \ldots, n$ (independent), where the index $\alpha$ and scale parameter $\lambda$ are unknown
- Prior: $\alpha \sim Ga(a, b)$ and $\lambda \sim Ga(c, d)$, independent
- Determine the posterior density for $(\alpha, \lambda)^T$ and hence the posterior conditional densities for $\alpha|\lambda$ and $\lambda|\alpha$

## Solution

...

## Problem

- Although the FCD for $\lambda$ is a standard distribution and easy to simulate from, the FCD for $\alpha$ is NOT!
- Therefore we can't use a Gibbs sampler for this analysis
- Fortunately there are other methods we can use ...

# 4.5 Metropolis–Hastings sampling

- The Gibbs sampler is a very powerful tool
- Only useful if the full conditional distributions (FCDs) are standard distributions (which are easy to simulate from)
- Fortunately there is a class of methods which can be used when the FCDs are non-standard
- These methods are known as Metropolis-Hastings schemes

- Want to simulate realisations from the posterior density $\pi(\boldsymbol{\theta}|\boldsymbol{x})$
- All of the FCDs are non-standard
- Choose a proposal distribution with density $q(\boldsymbol{\theta}^*|\boldsymbol{\theta})$, which is easy to simulate from
- This distribution gives us a way of proposing new values $\boldsymbol{\theta}^*$ from the current value $\boldsymbol{\theta}$

## Metropolis–Hastings algorithm

1. Initialise the iteration counter to $j = 1$, and initialise the chain to $\boldsymbol{\theta}^{(0)}$
2. Generate a proposed value $\boldsymbol{\theta}^*$ using the proposal distribution $q(\boldsymbol{\theta}^* | \boldsymbol{\theta}^{(j-1)})$
3. Evaluate the acceptance probability $\alpha(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}^*)$ of the proposed move, where

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \min\left\{1, \frac{\pi(\boldsymbol{\theta}^* | \boldsymbol{x})\, q(\boldsymbol{\theta} | \boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta} | \boldsymbol{x})\, q(\boldsymbol{\theta}^* | \boldsymbol{\theta})}\right\}$$

4. Set $\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}^*$ with probability $\alpha(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}^*)$, and set $\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}^{(j-1)}$ otherwise; that is:

| $\boldsymbol{\theta}^{(j)}$ | $\boldsymbol{\theta}^*$ | $\boldsymbol{\theta}^{(j-1)}$ |
|---|---|---|
| Probability | $\alpha$ | $1 - \alpha$ |

5. Change the counter from $j$ to $j + 1$ and return to step 2

## Comments

- At each stage, a new value is generated from the proposal distribution
- This value is either accepted, in which case the chain moves, or rejected, in which case the chain stays where it is
- Whether or not the move is accepted or rejected depends on an acceptance probability which itself depends on the relationship between the density of interest and the proposal distribution
- The posterior density $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ only enters into the acceptance probability as a ratio, and so the method can be used when the posterior density is only known up to a scaling constant, that is,

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \min\left\{1, \frac{\pi(\boldsymbol{\theta}^*)\, f(\boldsymbol{x}|\boldsymbol{\theta}^*)\, q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta})\, f(\boldsymbol{x}|\boldsymbol{\theta})\, q(\boldsymbol{\theta}^*|\boldsymbol{\theta})}\right\}$$

- This algorithm defines a Markov chain with $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ as its stationary distribution
- It holds for all possible proposal distributions (subject to them generating realisations from the full parameter space)

## Possible proposal distributions

- Are some choices of proposal distribution better than others?
- We now discuss some commonly used proposal distributions

## Symmetric proposals

- Use a symmetric proposal distribution: $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}) = q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)$, $\forall\ \boldsymbol{\theta}, \boldsymbol{\theta}^*$
- The acceptance probability simplifies to

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \min\left\{1, \frac{\pi(\boldsymbol{\theta}^*|\boldsymbol{x})}{\pi(\boldsymbol{\theta}|\boldsymbol{x})}\right\} = \min\left\{1, \frac{\pi(\boldsymbol{\theta}^*)\, f(\boldsymbol{x}|\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta})\, f(\boldsymbol{x}|\boldsymbol{\theta})}\right\}$$

  and hence does not involve the proposal density at all
- Proposed moves which will take the chain to a region of higher posterior density are always accepted
- Moves which take the chain to a region of lower posterior density are accepted with probability proportional to the ratio of the two densities
- Moves which will take the chain to a region of very low density will be accepted with very low probability

## Random walk proposals

- Consider the random walk proposal in which the proposed value $\boldsymbol{\theta}^*$ at stage $j$ is

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}^{(j-1)} + \boldsymbol{w}_j$$

  where the $\boldsymbol{w}_j$ are independent and identically distributed random $p \times 1$ vectors (completely independent of the state of the chain)
- Suppose that the $\boldsymbol{w}_j$ have density $f(\cdot)$, which is easy to simulate from, has mean $\mathbf{0}$ and is symmetric about its mean
- We can then simulate an innovation $\boldsymbol{w}_j$, and set the proposal value to $\boldsymbol{\theta}^* = \boldsymbol{\theta}^{(j-1)} + \boldsymbol{w}_j$
- What distribution should we use for $f(\cdot)$?

## Choice of innovation distribution

- A distribution which is simple and easy to simulate from is always a good idea; for example, the uniform or normal $\cdots\cdots$
  Normal is generally better, but is a bit more expensive to simulate
- What variance should we choose for the innovation distribution?
- This choice will affect the acceptance probability, and hence the overall proportion of accepted moves
- If the variance of the innovation is too low, then most proposed values will be accepted, but the chain will move very slowly around the space — the chain is said to be too "cold"
- If the variance of the innovation is too large, very few proposed values will be accepted, but when they are, they will often correspond to quite large moves — the chain is said to be too "hot"
- Theoretically it has been shown that the optimal acceptance rate is around 0.234 — this is an asymptotic result (for large samples of data) — but experience suggests that an acceptance rate of around 20–30% is okay
- Thus, the variance of the innovation distribution should be "tuned" to get an acceptance rate of around this level

# Example: Normal random walk proposals

- Asked by the UK Met Office to analyse annual maximum wind speed observations (mph) at 206 locations across the UK
- Suppose $X_i, i = 1, ..., 50$ represent these annual maximum wind observations for a particular location in the Peak District
- An appropriate model is the *Generalised Extreme Value* distribution, that is, $X_i|\mu, \sigma, \xi \sim GEV(\mu, \sigma, \xi)$
- After discussions with Simon Brown, a climate scientist at the Met Office, we use:

$$\begin{aligned} \mu &\sim N(55, 2) \\ \sigma &\sim Ga(20, 10) \quad \text{and} \\ \xi &\sim N(0, 0.5) \end{aligned}$$

# Example: Normal random walk proposals

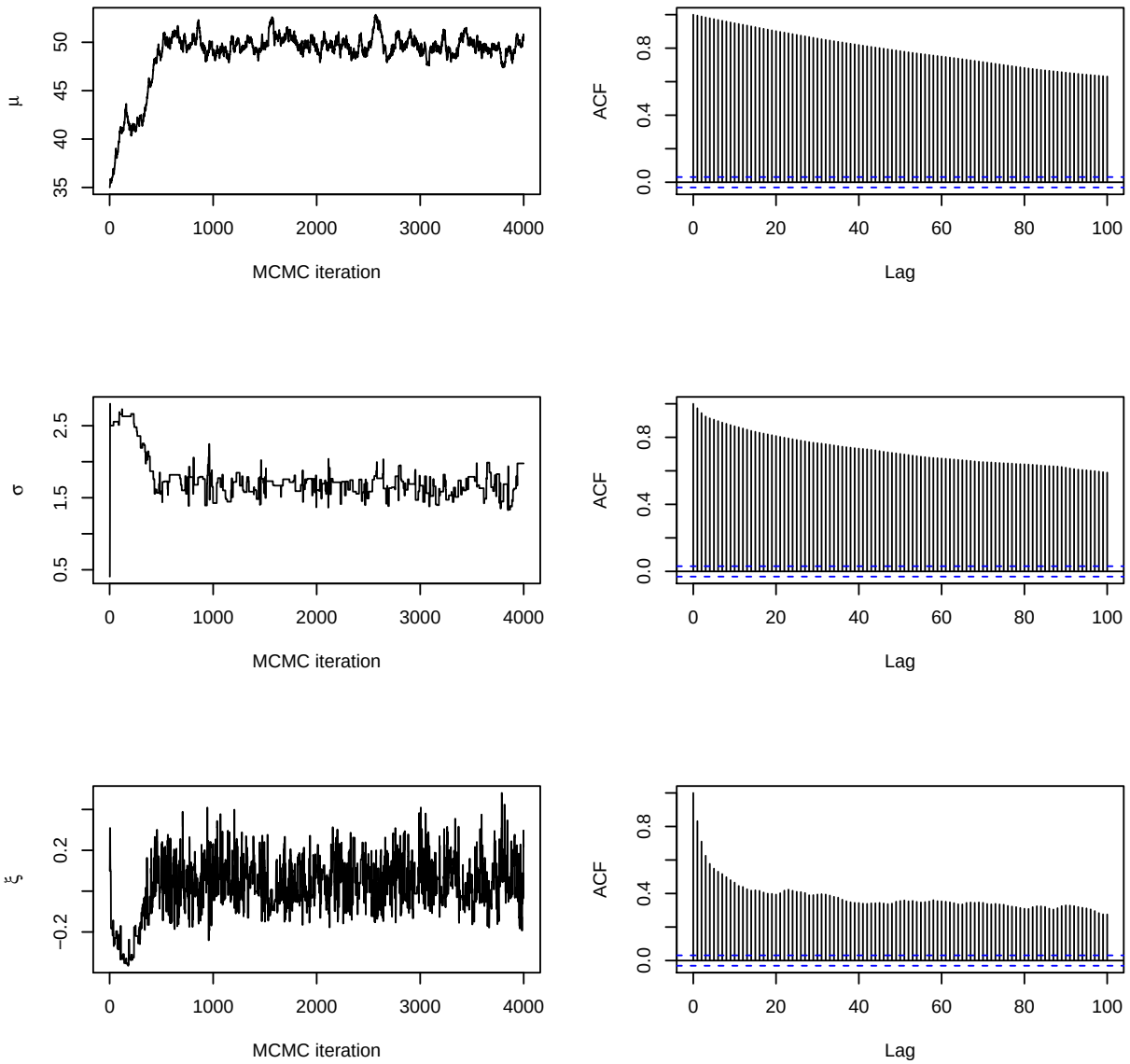
- No conjugate prior distributions exist for this model, and the full conditionals are not of standard form
- We need to use Metropolis-Hastings; we use Normal random walk updates:

$$\begin{aligned} \mu^* &= \mu^{(j-1)} + w_{\mu_j}, \\ \log \sigma^* &= \log \sigma^{(j-1)} + w_{\sigma_j} \quad \text{and} \\ \xi^* &= \xi^{(j-1)} + w_{\xi_j}, \end{aligned}$$

where

$$\begin{aligned} w_{\mu_j} &\sim N(0, \epsilon_\mu), \\ w_{\sigma_j} &\sim N(0, \epsilon_\sigma) \quad \text{and} \\ w_\xi &\sim N(0, \epsilon_\xi) \end{aligned}$$

# Example: Normal random walk proposals



# 4.6 Hybrid methods

## The algorithm

1. Initialise the iteration counter to $j = 1$
   Initialise the state of the chain to $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_p^{(0)})^T$
2. Let $\boldsymbol{\theta}_{-i}^{(j)} = \left(\theta_1^{(j)}, \ldots, \theta_{i-1}^{(j)}, \theta_{i+1}^{(j-1)}, \ldots, \theta_p^{(j-1)}\right)^T$, $i = 1, 2, \ldots, p$
   Obtain a new value $\boldsymbol{\theta}^{(j)}$ from $\boldsymbol{\theta}^{(j-1)}$ by successive generation of values
   - $\theta_1^{(j)} \sim \pi(\theta_1 | \boldsymbol{\theta}_{-1}^{(j)}, \boldsymbol{x})$ by direct simulation where the FCD is available; otherwise, use a M-H step with proposal distribution $q_1(\theta_1 | \theta_1^{(j-1)}, \boldsymbol{\theta}_{-1}^{(j)})$
   - $\theta_2^{(j)} \sim \pi(\theta_2 | \boldsymbol{\theta}_{-2}^{(j)}, \boldsymbol{x})$ by direct simulation where the FCD is available; otherwise, use a M–H step with proposal distribution $q_2(\theta_2 | \theta_2^{(j-1)}, \boldsymbol{\theta}_{-2}^{(j)})$

     $\vdots$
   - $\theta_p^{(j)} \sim \pi(\theta_p | \boldsymbol{\theta}_{-p}^{(j)}, \boldsymbol{x})$ by direct simulation where the FCD is available; otherwise, use a M-H step with proposal distribution $q_p(\theta_p | \theta_p^{(j-1)}, \boldsymbol{\theta}_{-p}^{(j)})$
3. Change counter $j$ to $j + 1$, and return to step 2

## Example 4.4

- Construct an MCMC scheme for the problem in Example 4.3, where we had a random sample of size $n$ from a gamma $Ga(\alpha, \lambda)$ distribution and independent gamma $Ga(a, b)$ and $Ga(c, d)$ prior distributions for $\alpha$ and $\lambda$ respectively
- Recall that the FCDs were

$$\pi(\alpha|\lambda, \boldsymbol{x}) \propto \frac{\alpha^{a-1} e^{(-b+n\log\bar{x}_g+n\log\lambda)\alpha}}{\Gamma(\alpha)^n}, \quad \alpha > 0$$

and

$$\pi(\lambda|\alpha, \boldsymbol{x}) \propto \lambda^{c+n\alpha-1}\, e^{-(d+n\bar{x})\lambda}, \quad \lambda > 0$$

## Solution

...

# Analysis of MCMC output (all realisations)

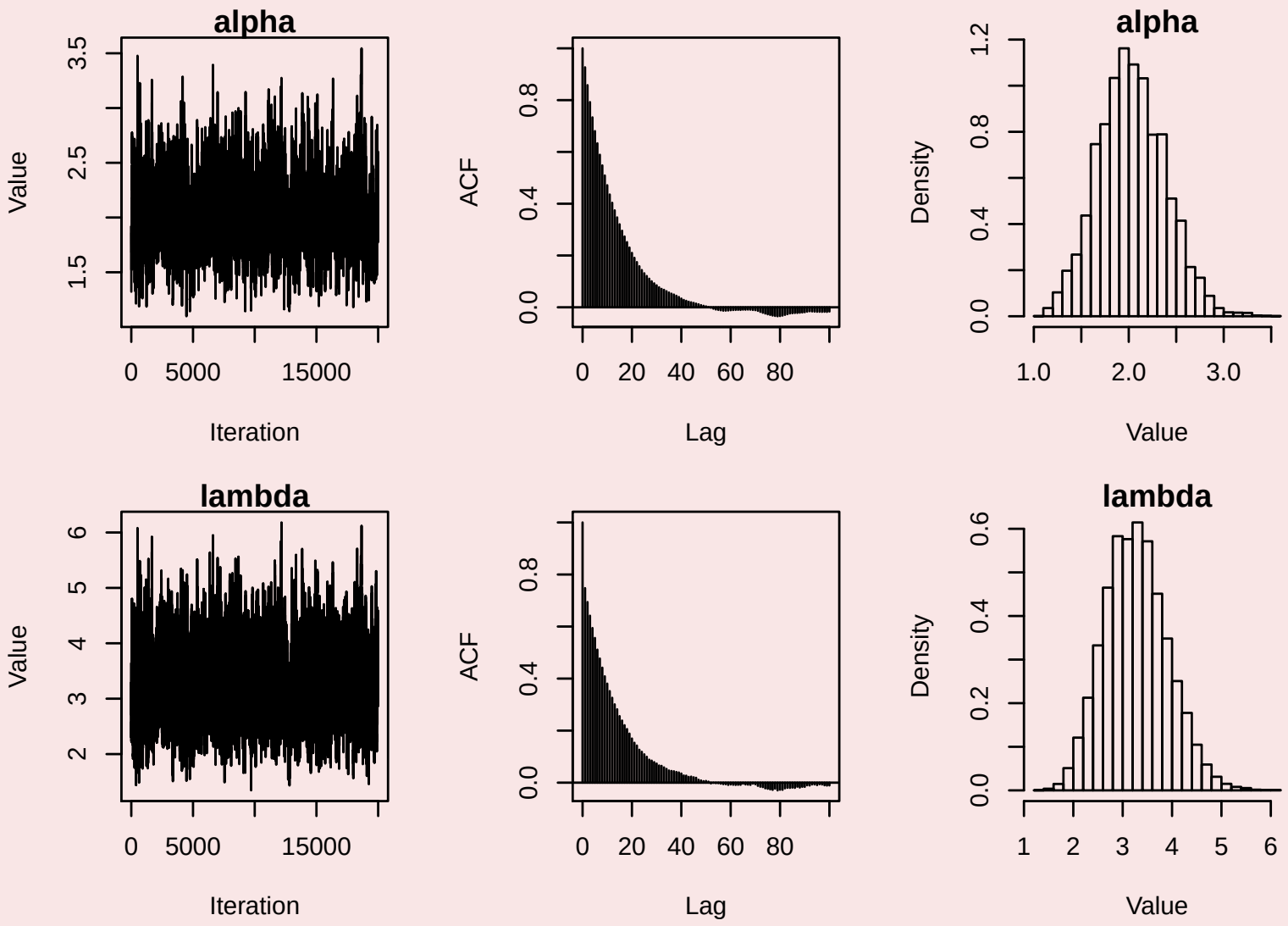



Figure: Trace plots, autocorrelation plots and histograms of the Metropolis within Gibbs sampler output

# Analysis of MCMC output (with burnin = 10, thin = 20)

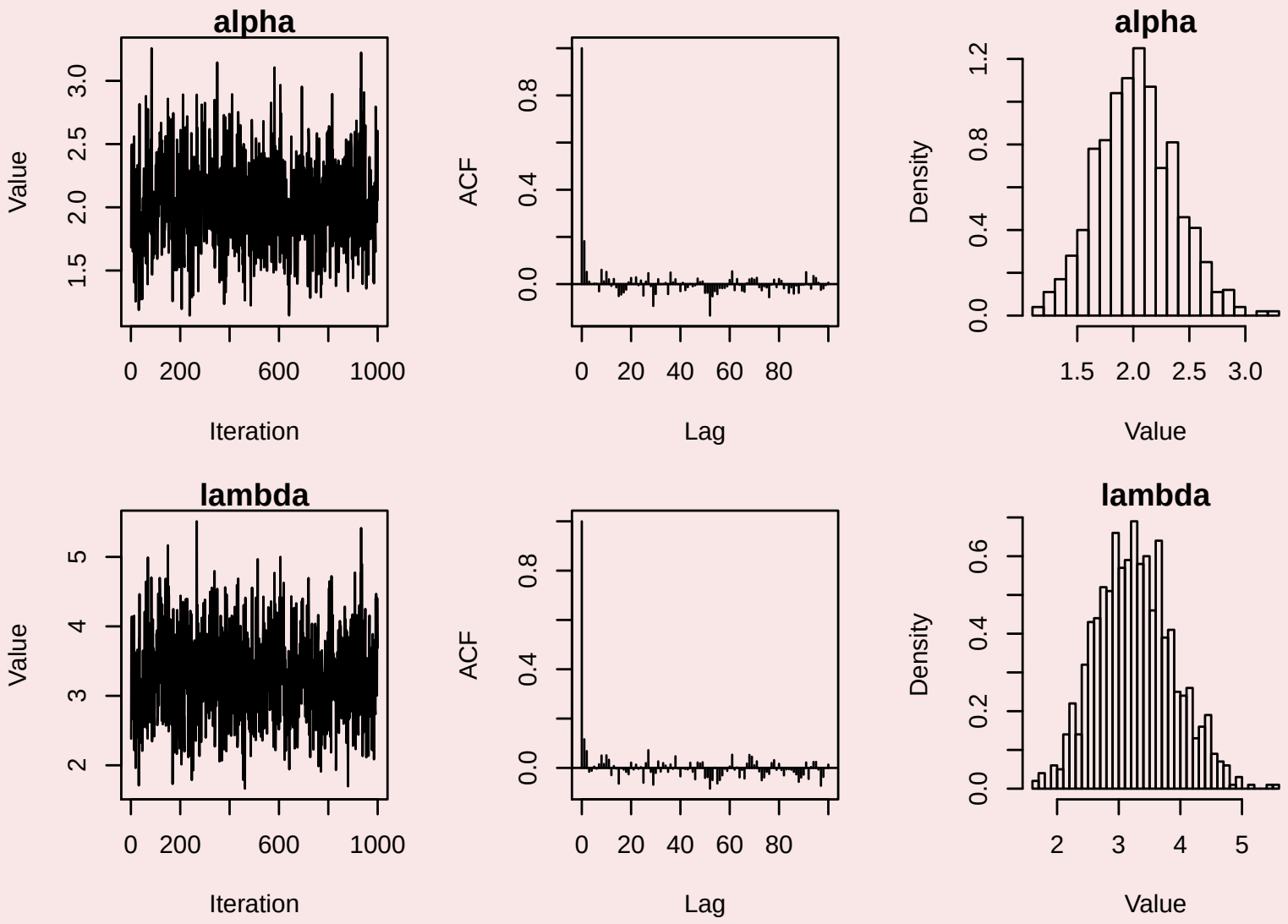



Figure: Trace plots, autocorrelation plots and histograms of the Metropolis within Gibbs sampler output

# 4.7 Summary

(i) Bayesian inference can be complicated when not using a conjugate prior distribution

(ii) One solution is to use Markov chain Monte Carlo (MCMC) methods

(iii) These work by producing realisations from the posterior distribution by constructing a Markov chain which has the posterior distribution as its stationary distribution

(iv) The MCMC methods we have studied are the Gibbs sampler, Metropolis–Hastings algorithm and the Metropolis-within-Gibbs algorithm

(v) When obtaining output from these algorithms, we need to assess whether there needs to be a burn-in period and whether the output needs to be thinned (by looking at traceplots and autocorrelation plots)

(vi) The (converged and thinned) MCMC output are realisations from the posterior distribution. It can be used to

- obtain the posterior distribution for any (joint) functions of the parameters (such as $\sigma = 1/\sqrt{\tau}$ or $(\theta_1 = \mu - \tau, \theta_2 = e^{\mu+\tau/2})^T$)
- look at bivariate posterior distributions via scatter plots
- look at univariate marginal posterior distributions via histograms or boxplots
- obtain numerical summaries such as the mean, standard deviation and confidence intervals for single variables and correlations between variables

# 4.8 Learning objectives

By the end of this chapter, you should be able to:

1. Explain why not using a conjugate prior generally causes problems in determining the posterior distribution
2. Describe the Gibbs sampler, explain why it is a Markov chain and give an outline as to why its stationary distribution is the posterior distribution
3. Describe the issues of processing MCMC output (burn-in, autocorrelation, thinning etc.) and interpret numerical/graphical output
4. Derive the full conditional densities for any posterior distribution and name these distributions if they are "standard" distributions given in the notes or on the exam paper

5. Describe a Metropolis-Hastings algorithm in general terms and when using symmetric random walk proposals or independence proposals or general proposals
6. Describe the hybrid methods *Metropolis within Gibbs*
7. provide a detailed description of **any** of the MCMC algorithms as they apply to generating realisations from **any** posterior distribution