Session 2. Bayesian inference: conjugate models and Markov chain Monte Carlo

Outline

Conjugate Bayesian inference

- Bayes Theorem and Bayesian inference
- Bayesian inference for a proportion
- Bayesian inference for the mean of normal, with known variance

Practical exercises 1-3

Markov chain Monte Carlo

- Gibbs sampling
- Convergence diagnostics

Practical exercises 4-6

Conjugate models

Bayesian inference

The Bayesian analyst needs to explicitly state

- a reasonable opinion concerning the plausibility of different values of the parameters *excluding* the evidence from the current study (the **prior distribution**)
- the support for different values of the parameters based *solely* on data from the trial (the **likelihood/sampling model**[1]),

and to combine these two sources to produce

- a final opinion about the parameters (the **posterior distribution**)

One can view the Bayesian approach as a formalisation of the process of learning from experience

[1] We use the terms "likelihood" and "sampling model" interchangeably

Bayes theorem

Let $A_1, \ldots, A_J$ be a set of mutually exclusive events with $p(\bigcup_j A_j) = \sum_{j=1}^{J} p(A_j) = 1$

Let B be another event, then

$$p(A_i \mid B) = \frac{p(B \mid A_i)p(A_i)}{\sum_{j=1}^{J} p(B \mid A_j)p(A_j)}$$

Example: Bayes theorem in diagnostic testing

D-dimer is a test used in hospital for deep vein thrombosis (DVT).

- Laboratory-based D-dimer tests[2] for DVT have sensitivity 0.93 and specificity 0.48
- Suppose we use such D-dimer tests in a population with DVT prevalence of $1/100 = 0.01$
- What is the chance that a patient with a positive D-dimer test result actually has a DVT?

[2] NICE Evidence Review NG158, 2020

Example: Bayes theorem in diagnostic testing (cont)

Let A be the event that the patient truly has a DVT

Let $\overline{A}$ be the event that the patient truly does not have a DVT

Let B be the event that they have a positive D-dimer test result

We want $p(A \mid B)$.

"93% sensitivity" means that $p(B \mid A) = 0.93$.

"48% specificity" means that $p(B \mid \overline{A}) = 1 - 0.48 = 0.52$.

By Bayes theorem

$$p(A \mid B) = \frac{p(B \mid A)p(A)}{p(B \mid A)p(A) + p(B \mid \overline{A})p(\overline{A})}$$
$$= \frac{0.93 \times 0.01}{0.93 \times 0.01 + 0.52 \times 0.99} = 0.018$$

Thus, *in this population*, 98.2% of those with a positive D-dimer test result will *not*, in fact, have a DVT.

Example: Bayes theorem in diagnostic testing (cont)

Easier to explain as expected outcomes in a large number of cases

	No DVT ($\overline{A}$)	Has DVT (A)	Total
D-dimer $-$ ($\overline{B}$)	47,520	70	47,590
D-dimer $+$ (B)	51,480	930	52,410
Total	99,000	1,000	100,000

$$p(A \mid B) = 930/52410 = 0.018$$

Example: Bayes theorem in diagnostic testing (cont)

Bayes Theorem tells us how the test result changes our beliefs — often our intution is poor when processing probabilistic evidence

- The disease prevalence (here $p = 0.01$) is like a prior probability
- Observing a positive D-dimer result causes us to modify this probability to $p = 0.018$, which is our posterior probability that the patient has a DVT.

Here Bayes theorem was used for combining probabilities that are assumed known.

Bayesian inference takes this idea further: it is the use of Bayes theorem in general statistical analyses, when parameters are the unknown quantities and their prior distributions need to be specified.

Bayesian inference

Bayesian methods distinguish:

1. Observable quantities y: the data
   - Bayesian inference treats these as fixed (after observing them)
2. All unknown quantities θ: statistical parameters, missing data, mismeasured data ...
   - Bayesian inference treats all of these as random variables

Contrast with the frequentist framework where:

- parameters are fixed, non-random quantities
- probability statements concern the data

Bayesian inference (cont)

As with any statistical analysis, we start by positing the **likelihood** $p(y \mid \theta)$, which relates all variables into a **'full probability model'**.

- θ is unknown so should have a **probability distribution** reflecting our uncertainty about it before seeing the data
  $\rightarrow$ need to specify a **prior distribution** $p(\theta)$

- After observing data y, it is known so we condition on it, and use Bayes theorem to obtain the conditional probability distribution for unobserved quantities of interest given the data y:

$$p(\theta \mid y) = \frac{p(\theta)\, p(y \mid \theta)}{\int p(\theta)\, p(y \mid \theta)\, d\theta} \propto p(\theta)\, p(y \mid \theta)$$

  This is the **posterior distribution**

Inference for proportions

Suppose we observe r positive responses out of n patients.

Assuming patients are independent, with common unknown response rate θ, leads to a binomial likelihood

$$p(r \mid n, \theta) = \binom{n}{r} \theta^r (1-\theta)^{n-r} \propto \theta^r (1-\theta)^{n-r}$$

As discussed in Session 1, often use a Beta(a, b) prior distribution for θ

$$p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1} \propto \theta^{a-1} (1-\theta)^{b-1}$$

Need to calculate the posterior distribution $p(\theta \mid r, n)$

Inference for proportions: mathematical derivation

Combining binomial likelihood and beta prior gives

$$\begin{aligned} p(\theta \mid r, n) &\propto p(r \mid n, \theta) \times p(\theta) \\ &\propto \theta^r (1-\theta)^{n-r} \times \theta^{a-1}(1-\theta)^{b-1} \\ &= \theta^{r+a-1}(1-\theta)^{n-r+b-1} \end{aligned}$$

We can recognise this as proportional to the pdf of a Beta$(r+a, n-r+b)$ distribution

So the posterior distribution $p(\theta \mid r, n)$ is Beta$(r+a, n-r+b)$

Posterior distribution is in the same family as the prior, so we say the Beta distribution is a conjugate prior for the binomial likelihood

With fixed a and b, as r and n increase, $E(\theta \mid r, n) \to r/n$ (the MLE), and the variance tends to zero

Inference for proportions: predictive distributions

In Session 1, we looked at the prior predictive distribution

$$p(r) = \int p(r \mid n, \theta) \times p(\theta) d\theta$$

which we found was a Beta-Binomial distribution, with parameters (a, b, n)

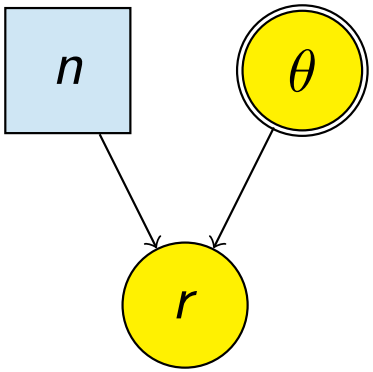
After observing r, we can provide a revised predictive distribution, the posterior predictive distribution for the number of successes out of m in the future:

$$p(r_{\text{pred}}) = \int p(r_{\text{pred}} \mid m, \theta) \times p(\theta \mid r, n) d\theta$$

This is again a Beta-Binomial distribution, but now with parameters $(r + a, n - r + b, m)$

Inference for proportions: DAG representation

Before observing r...



Model specification

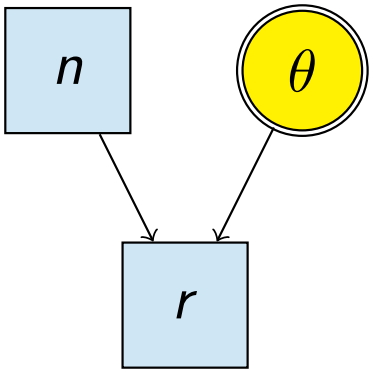
$$\theta \sim \text{Beta}(a, b)$$
$$r \mid (n, \theta) \sim \text{Binomial}(r \mid n, \theta)$$

Prior predictive distribution

$$r \sim \text{Beta-Binomial}(a, b, n)$$

Inference for proportions: DAG representation (cont)

After observing r...



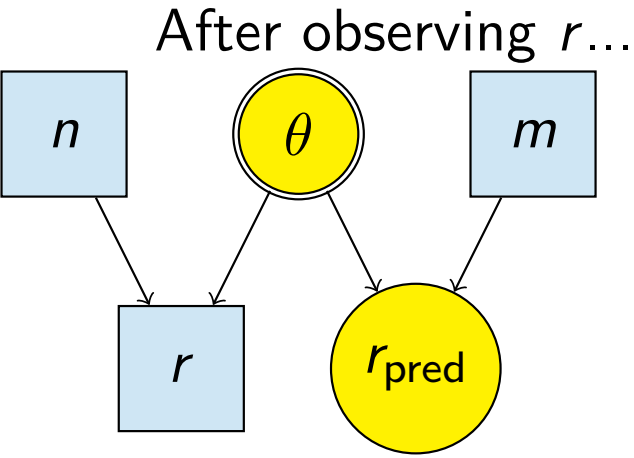
Model specification

$$\theta \sim \text{Beta}(a, b)$$
$$r \mid (n, \theta) \sim \text{Binomial}(r \mid n, \theta)$$

Posterior distribution

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

Inference for proportions: DAG representation (cont)



Posterior predictive distribution in m trials

$$r_{\text{pred}} \mid (r, n, m) \sim \text{Beta-Binomial}(r + a, n - r + b, m)$$

Inference for proportions: Drug example

- Recall example from session 1, where we consider early investigation of a new drug
- Experience with similar compounds has suggested that response rates between 0.2 and 0.6 could be feasible
- We interpreted this as a distribution with mean $= 0.4$, standard deviation 0.1 and showed that a Beta$(9.2, 13.8)$ distribution has these properties
- Suppose we now treat $n = 20$ volunteers with the compound and observe $r = 15$ positive responses

Inference for proportions: Drug example (cont)

Model specification: fix number of trials n; and values of prior hyperparameters a and b

$$\theta \sim \text{Beta}(a = 9.2, b = 13.8)$$
$$r \mid (n = 20, \theta) \sim \text{Binomial}(r \mid n = 20, \theta)$$

Prior predictive distribution

$$r \sim \text{Beta-Binomial}(a = 9.2, b = 13.8, n = 20)$$

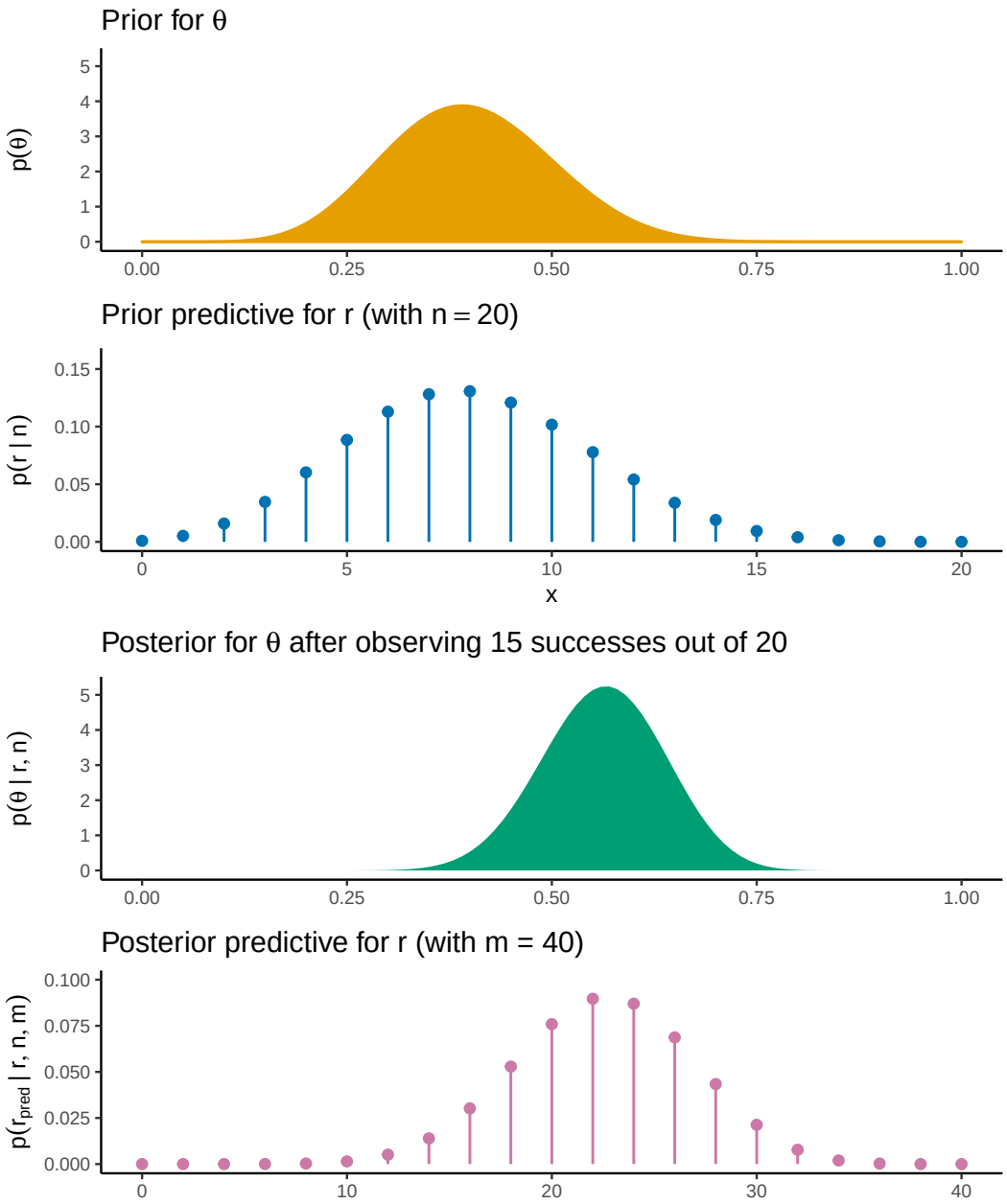
Posterior distribution: observe r

$$\theta \mid (r = 15, n = 20) \sim \text{Beta}(15 + 9.2, 20 - 15 + 13.8)$$

Posterior predictive distribution for $m = 40$ future trials

$$r_{\text{pred}} \mid (r = 15, n = 20, m = 40) \sim \text{Beta-Binomial}(24.2, 18.8, 40)$$

Inference for proportions: Drug example (cont)



Inference for proportions: using JAGS

- In JAGS there is no need to explicitly specify posterior
- Only need to provide the model specification and data
- JAGS contains algorithms to evaluate the posterior given (almost) arbitrary specification of prior and likelihood
  - posterior doesn't need to be closed form
  - may exploit conjugacy when it exists
- Learn about the posterior via samples drawn from it – "Monte Carlo"

JAGS will use all information you provide: the output will reflect all the information you provide

Inference for proportions: Drug example using JAGS

The model can be written in JAGS:

```
model {
  theta ~ dbeta(9.2, 13.8)
  r ~ dbin(theta, 20)
}
```

- If we provide no observations of r, then JAGS's output for
  - `theta` will represent its prior distribution
  - `r` will represent its prior predictive distribution
- If we provide an observation $r = 15$ as data `list(r = 15)`, then JAGS's output for
  - `theta` will represent its posterior (given $r = 15$).
  - `r` will be fixed at its observed value

Inference for proportions: Drug example using JAGS (cont)

Equivalently we can leave the fixed values n, a and b as variables in the model:

```
model {
  theta ~ dbeta(a, b)
  r ~ dbin(theta, n)
}
```

and provide the fixed values as extra "data" points, along with the observed data $r = 15$:

```
list(a = 9.2, b = 13.8, n = 20, r = 15)
```

Both will provide identical results.

Inference for proportions: Drug example using JAGS (cont)

We can obtain predictions for the number of successes out of m trials simply by adding an extra copy of the "likelihood":

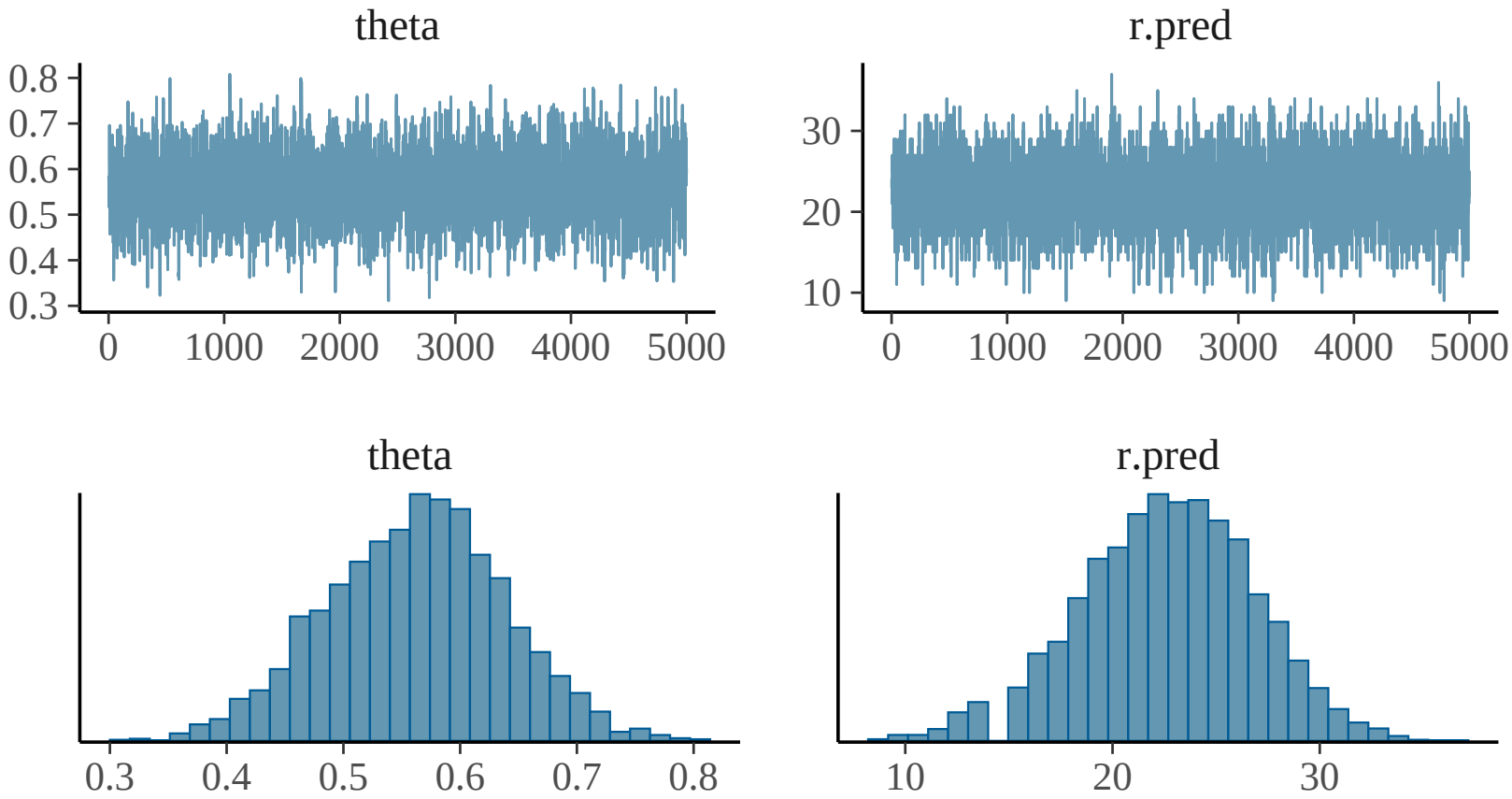
```
model {
  theta ~ dbeta(a, b)
  r ~ dbin(theta, n)
  r.pred ~ dbin(theta, m)
}
```

with a fixed value for m specified as data (as per n, a and b previously).

Inference for proportions: Drug example using JAGS (cont)

- If we provide no observations of r, then JAGS's output for
  - `theta` will represent its prior distribution
  - `r` will represent its prior predictive distribution, with n trials
  - `r.pred` will represent its prior predictive distribution, with m trials
- If we provide an observation $r = 15$ as `list(r = 15)`, then JAGS's output for
  - `theta` will represent its posterior (given $r = 15$).
  - `r` will be fixed at its observed value
  - `r.pred` will represent the posterior predictive distribution, with m trials

Inference for proportions: Drug example results



Top row: traceplots, showing the drawn sample against Monte Carlo sample number (iteration) Bottom row: histograms of the drawn samples

Inference for proportions: Drug example results (cont)

```
# A tibble: 2 × 10
  variable   mean median     sd    mad     q5    q95  rhat ess_bulk ess_tail
  <chr>     <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl> <dbl>    <dbl>    <dbl>
1 theta     0.562  0.566 0.0764 0.0760  0.434  0.686  1.00    2780.    3013.
2 r.pred   22.5   23     4.27   4.45   15     29      1.00    3582.    3769.
```

And the probability of at least 25 successes in 40 future trials:

```
> mean(r.pred >= 25)
[1] 0.3304
```

Exact answers from conjugate analysis

- θ: mean 0.563 and standard deviation 0.075
- r_{pred}: mean 22.51 and standard deviation 4.31.
- Probability of at least 25: 0.329

MCMC results are within Monte Carlo error of the true values

Normal distribution

Suppose we have n independent observations from a normal distribution with unknown mean μ but known variance σ^2

$$y_i \sim \mathsf{N}(\mu, \sigma^2) \qquad\qquad i = 1, \ldots, n$$

Suppose we choose a normal prior distribution for μ, with fixed mean γ and fixed variance ω^2

$$\mu \sim \mathsf{N}(\gamma, \omega^2)$$

Normal distribution – mathematical derivation

Combining normal likelihood and normal prior gives the posterior:

$$p(\mu \mid y_1, \ldots, y_n) \propto p(\mu) \prod_{i=1}^{n} p(y_i \mid \mu)$$

$$= \exp\left[-\frac{1}{2}\left\{\frac{(\mu-\gamma)^2}{\omega^2}\right\}\right] \exp\left[-\frac{1}{2}\left\{\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{\sigma^2}\right\}\right]$$

After some algebra , we can recognise this as proportional to the PDF of a normal:

$$\mu \mid (y_1, \ldots, y_n) \sim \mathrm{N}\left(\frac{n_0\gamma + n\overline{y}}{n_0+n}, \frac{\sigma^2}{n_0+n}\right)$$

where we have rewritten prior equivalently as $\mu \sim \mathrm{N}(\gamma, \sigma^2/n_0)$
where $n_0 = \sigma^2/\omega^2$, and where $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Normal distribution (cont)

- As n_0 tends to 0, the prior variance becomes larger and the distribution becomes 'flatter', and in the limit the prior distribution becomes essentially uniform over $-\infty, \infty$
- Posterior mean $(n_0\mu + n\overline{y})/(n_0 + n)$ is a weighted average of the prior mean μ and parameter estimate $\overline{y}$, weighted by their precisions (relative 'sample sizes'), and so is always a compromise between the two
- Posterior variance is based on an implicit sample size equivalent to the sum of the prior 'sample size' n_0 and the sample size of the data n
- As $n \to \infty$, $p(\mu \mid y) \to \mathrm{N}(\overline{y}, \sigma^2/n)$ *i.e.* not depending on prior
- Compare with frequentist setting, the MLE is $\hat{\mu} = \bar{y}$ with $\mathrm{SE}(\hat{\mu}) = \sigma/\sqrt{n}$, and sampling distribution

$$p(\hat{\mu} \mid \mu) = p(\bar{y} \mid \mu) = \mathrm{N}(\mu, \sigma^2/n)$$

Normal distribution: JAGS implementation

Important

JAGS uses precision ($=$ variance^{-1} $=$ stddev^{-2}) rather than variance to describe the dispersion of a normal distribution.

Write precision $\tau = \sigma^{-2}$ and precision $\phi = \omega^{-2}$.

For a single observation (ie $n = 1$), the model specification is

$$y \sim \mathrm{N}(\mu, \text{variance} = \tau^{-1})$$
$$\mu \sim \mathrm{N}(\gamma, \text{variance} = \phi^{-1})$$

Then in JAGS the model is

```
model {
  mu ~ dnorm(gamma, phi)
  y ~ dnorm(mu, tau)
}
```

where we must provide fixed values for γ, τ and ϕ as data (and y).

Normal distribution: JAGS implementation

When we have n observations, the model specification is

$$y_i \sim \mathrm{N}(\mu, \text{variance} = \tau^{-1}) \qquad i = 1, \ldots, n$$
$$\mu \sim \mathrm{N}(\gamma, \text{variance} = \phi^{-1})$$

Then in JAGS the model is

```
model {
  mu ~ dnorm(gamma, phi)
  for (i in 1:n){
    y[i] ~ dnorm(mu, tau)
  }
}
```

where we must provide JAGS with

- A vector of observations $y = (y_1, \ldots, y_n)$
- fixed values for n, γ, τ and ϕ as data

Functions of parameters

Suppose we have a parameter θ in our model, but we are really interested in $g(\theta)$

Easy using Monte Carlo: just calculate required function $g(\theta)$ for each Monte Carlo sample and summarise posterior samples of $g(\theta)$

For example, suppose we are interested in the variance $\sigma^2 = 1/\tau$. Two equivalent ways:

1. Add an extra line into the `model` that you provide to JAGS

   ```
   sigma.squared <- 1/tau
   ```

   and monitor the quantity `sigma.squared` using JAGS.

2. Or monitor `tau` using JAGS, then in R calculate

   ```
   sigma.squared <- 1/tau
   ```

JAGS: declarative model specification

Model specification in JAGS is declarative not procedural

- not providing a "procedure" for inference, as you typically would in most programming languages such as R
- instead we are describing ("declaring") the model specification to JAGS, which then works out what to do

Means you can specify model components in any order.

e.g. the following two models are identical:

```
model {
  mu ~ dnorm(gamma, phi)
  for (i in 1:n){
    y[i] ~ dnorm(mu, tau)
  }
  sigma.squared <- 1/tau
}
```

```
model {
  sigma.squared <- 1/tau
  for (i in 1:n){
    y[i] ~ dnorm(mu, tau)
  }
  mu ~ dnorm(gamma, phi)
}
```

Normal distribution example: Assay

- Assays used in clinical settings are not perfect – they all come with measurement error
- Suppose three independent measurements for the same specimen are taken:

$$y_1 = 119.04, \quad y_2 = 144.08, \quad y_3 = 116.57$$

- Suppose we know that the assay measurement error has a standard deviation $\sigma = 5$ IU/ml
- What should we estimate the concentration is in this specimen?

Let the mean be denoted μ.

A standard analysis estimates μ by the sample mean $\overline{y} = 126.56$ IU/ml, with standard error $\sigma/\sqrt{n} = 5/\sqrt{3} = 2.88$ IU/ml, and 95% confidence interval $\overline{y} \pm 1.96 \times \sigma/\sqrt{n}$, *i.e.* 120.91–132.22 IU/ml

Normal distribution example: Assay (cont)

Suppose another laboratory has previously assayed this specimen, and estimated that the mean concentration was 108 IU/ml with standard deviation 5.5 IU/ml

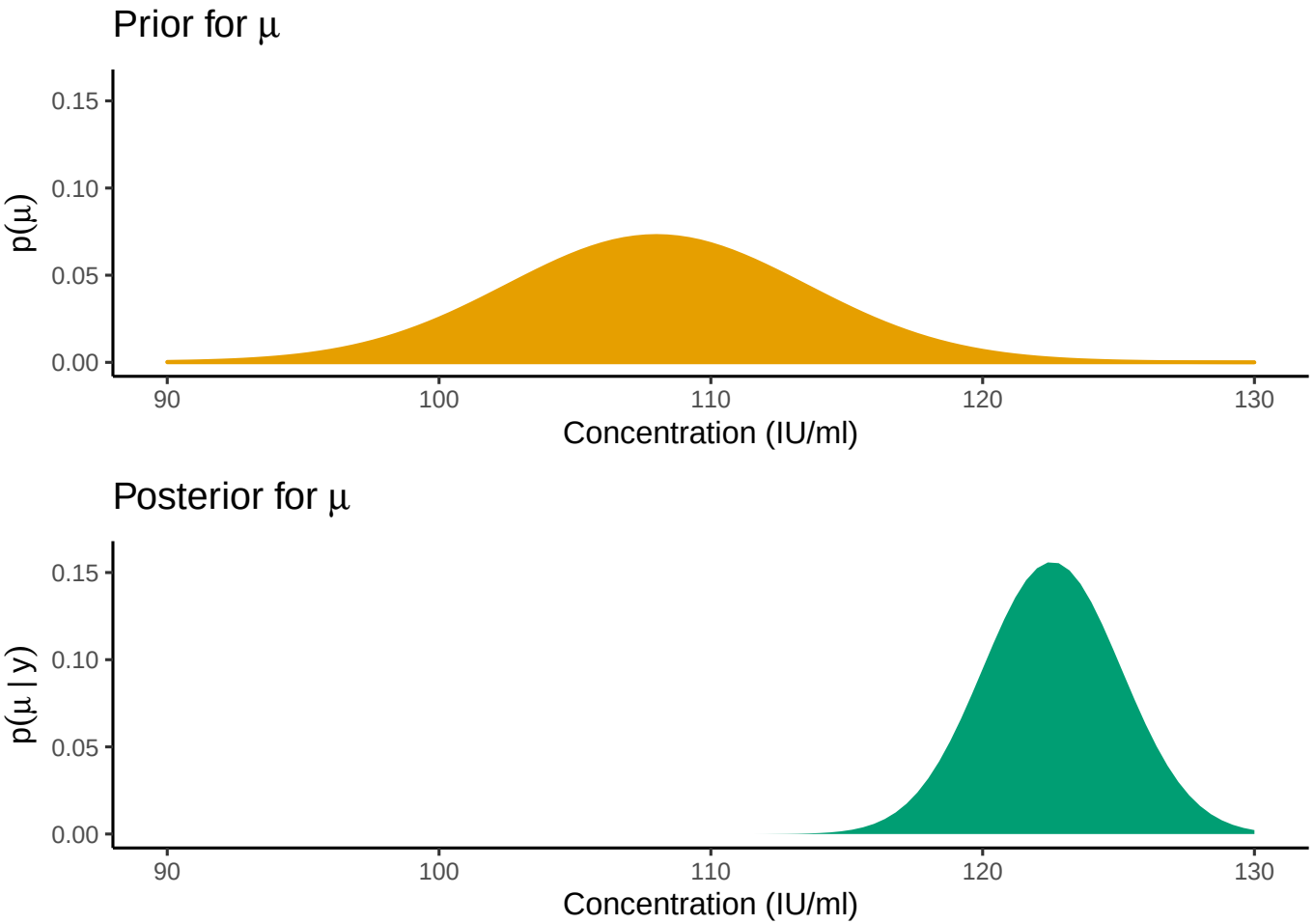
- suggests $\mathrm{N}(108, 5.5^2)$ prior for μ
- if we express the prior standard deviation in terms of $\sigma/\sqrt{n_0}$ (where $\sigma = 5$), we can solve to find $n_0 = (5/5.5)^2 = 0.83$
- so our prior can be written as $\mu \sim \mathrm{N}(108, \sigma^2/0.83)$

Posterior for μ is then

$$p(\mu \mid y) = \mathrm{N}\left(\frac{0.93 \times 108 + 3 \times 126.56}{0.83 + 3}, \frac{5^2}{0.83 + 3}\right)$$
$$= \mathrm{N}(122.55, 2.56^2)$$

giving 95% interval for μ of 117.54 to 127.56 IU/ml

Normal distribution example: Assay (cont)



Normal distribution: prediction

Denoting the posterior mean and variance as $\mu_{\text{post}} = (n_0\mu + n\bar{y})/(n_0 + n)$ and $\sigma^2_{\text{post}} = \sigma^2/(n_0 + n)$, the *predictive distribution* for a new observation $\tilde{y}$ is

$$p(\tilde{y} \mid y) = \int p(\tilde{y} \mid y, \mu)p(\mu \mid y)d\mu$$

which generally simplifies to

$$p(\tilde{y} \mid y) = \int p(\tilde{y} \mid \mu)p(\mu \mid y)d\mu$$

which can be shown to give

$$p(\tilde{y} \mid y) \sim \text{N}\left(\mu_{\text{post}}, \sigma^2_{\text{post}} + \sigma^2\right)$$

So the predictive distribution is centred around the posterior mean with variance equal to sum of the posterior variance and the sample variance of $\tilde{y}$

Summary

For all these examples, we see that

- the posterior mean is a compromise between the prior mean and the MLE
- the posterior s.d. is less than each of the prior s.d. and the s.e.(MLE)
  'A Bayesian is one who, vaguely expecting a horse and catching a glimpse of a donkey, strongly concludes he has seen a mule'

As $n \to \infty$,

- the posterior mean $\to$ the MLE
- the posterior s.d. $\to$ the s.e.(MLE)
- the posterior does not depend on the prior.

These observations are generally true, when the MLE exists and is unique.

Conjugate prior families

When the posterior is in the same family as the prior then we have what is known as *conjugacy*. This has the advantage that prior parameters can usually be interpreted as a *prior sample*. Examples include:

Likelihood	Parameter	Prior	Posterior
Normal	mean	Normal	Normal
Normal	precision	Gamma	Gamma
Binomial	success prob.	Beta	Beta
Poisson	rate or mean	Gamma	Gamma

- Conjugate prior distributions are mathematically convenient, but do not exist for all likelihoods, and can be restrictive
- Computations for non-conjugate priors are harder, but possible using MCMC (see next lecture)

Key points so far

1. With data y and unknown quantities θ we need:
   - **Prior distribution** $p(\theta)$ – we *choose* this to represent our opinion of the plausibility of unknown quantities θ *before seeing the data*
   - **Likelihood** $p(y \mid \theta)$ – this describes how data depend on the parameter values θ
   - **Posterior distribution** $p(\theta \mid y)$ – this describes our uncertainty about θ *after seeeing the data*. We calculate this using Bayes Theorem.
2. In JAGS, we simply provide the model specification: the prior and likelihood.
3. When prior and likelihood are from the same family, they are known as **conjugate**. This simplifies the mathematics of Bayes Theorem – but JAGS can handle non-conjugate models as well (coming next)

Practical 1

1. Conjugate inference in R
   - Doing conjugate Bayesian inference exactly and using Monte Carlo in R.
2. Basic JAGS model
   - A demonstration of conjugate Bayesian inference in JAGS
3. Implement a JAGS model for yourself
   - Ensure you understand the role of all involved quantities
   - Prediction in JAGS

Markov chain Monte Carlo

Bayesian computation

Suppose we have a vector of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ in our model.

The posterior distribution is

$$p(\boldsymbol{\theta} \mid y) = \frac{p(\boldsymbol{\theta})\, p(y \mid \boldsymbol{\theta})}{\int p(\boldsymbol{\theta})\, p(y \mid \boldsymbol{\theta})\, d\boldsymbol{\theta}}$$

What do we want to know about a posterior distribution?

- marginal posteriors for particular components θ_i

$$p(\theta_i \mid y) = \int \int \cdots \int p(\boldsymbol{\theta} \mid y)\; d\boldsymbol{\theta}_{(-i)}$$

  where $\boldsymbol{\theta}_{(-i)}$ denotes the vector $\boldsymbol{\theta}$ excluding θ_i

- Properties of the marginal posteriors $p(\theta_i \mid y)$ such as
  - the mean $= E(\theta_i) = \int \theta_i p(\theta_i \mid y) d\theta_i$
  - tail areas $= p(\theta_i > A) = \int_A^\infty p(\theta_i \mid y) d\theta_i$ etc.

Bayesian computation

For conjugate models, we wrote

$$p(\boldsymbol{\theta} \mid y) \propto p(\boldsymbol{\theta})\, p(y \mid \boldsymbol{\theta})$$

and then recognised this was proportional to the pdf of a family of known distributions. This made it straightforward to

- plot the pdf
  - using the functions in R etc
  - by Monte Carlo sampling from the posterior
- to calculate properties of the posterior
  - by calculating these on pen-and-paper (or looking it up)
  - by Monte Carlo sampling from the posterior

Bayesian computation

For non-conjugate models, we can still write down

$$p(\boldsymbol{\theta} \mid y) \propto p(\boldsymbol{\theta})\, p(y \mid \boldsymbol{\theta})$$

but this won't have a recognisable form as a 'standard' distribution. So the expression it is not directly useful.

We can also write down the full equation

$$p(\boldsymbol{\theta} \mid y) = \frac{p(\boldsymbol{\theta})\, p(y \mid \boldsymbol{\theta})}{\int p(\boldsymbol{\theta})\, p(y \mid \boldsymbol{\theta})\, d\boldsymbol{\theta}}$$

but the integral in the denominator can't usually be solved, so this expression is also not directly useful either.

Can't solve on pen-and-paper $\rightarrow$ use Monte Carlo again?

Monte Carlo integration

Suppose we could draw T samples from the joint posterior distribution for $\boldsymbol{\theta}$

$$\left(\theta_1^{(1)}, \ldots, \theta_k^{(1)}\right), \left(\theta_1^{(2)}, \ldots, \theta_k^{(2)}\right), \ldots, \left(\theta_1^{(T)}, \ldots, \theta_k^{(T)}\right) \sim p(\boldsymbol{\theta} \mid y)$$

Then,

1. $\theta_1^{(1)}, \ldots, \theta_1^{(T)}$ are a sample from the marginal posterior $p(\theta_1 \mid y)$
2. And we can do *Monte Carlo integration*

$$E(g(\theta_1)) = \int g(\theta_1) p(\theta_1 \mid y) d\theta_1 \approx \frac{1}{T} \sum_{t=1}^{T} g(\theta_1^{(i)})$$

Theorems exist which prove convergence in limit as $t \to \infty$

Monte Carlo with dependent samples

So far we have drawn *independent* samples for use in Monte Carlo

$$\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(T)} \sim p(\boldsymbol{\theta} \mid y)$$

Not possible in general — but can draw *dependent* samples.

- A particular form of dependence turns out to be convenient, in which each sample is drawn depending on the previous sample.
- More precisely, we draw samples from a Markov chain

  $$\boldsymbol{\theta}^{(1)} \sim p(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(0)}), \quad \boldsymbol{\theta}^{(2)} \sim p(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(1)}), \quad \ldots$$

  so that, given $\boldsymbol{\theta}^{(t)}$, $\boldsymbol{\theta}^{(t+1)}$ is independent of $\boldsymbol{\theta}^{(0)}, \ldots, \boldsymbol{\theta}^{(t-1)}$
- We design the chain such that the stationary distribution of the Markov chain is $p(\boldsymbol{\theta} \mid y)$
- Even with dependent samples, as $t \to \infty$, can prove Monte Carlo integration still works

MCMC algorithms

How to design and draw samples from such a Markov chain?

- Several standard 'recipes' available for designing Markov chains with required stationary distribution $p(\boldsymbol{\theta} \mid y)$
  - Gibbs sampling
  - Slice sampling: samples uniformly under the density function
  - Metropolis-Hastings: a very general formulation allowing "proposed moves" that are accepted or rejected
  - Hamiltonian Monte Carlo: exploits gradient information. Implemented in *e.g.* Stan (https://mc-stan.org) and Turing (https://turing.ml).
- JAGS automatically an MCMC algorithm (from Gibbs sampling, slice sampling, and random walk Metropolis-Hastings).

(Bayesian inference without MCMC is possible using the Laplace Approximation. Very fast but approximates only marginal posteriors. Implemented in R-INLA (https://www.r-inla.org))

Two-stage Gibbs sampling for a bivariate normal

Consider random variables (θ_1, θ_2) with a bivariate normal distrib.

$$(\theta_1, \theta_2) \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

So the conditional distributions are:

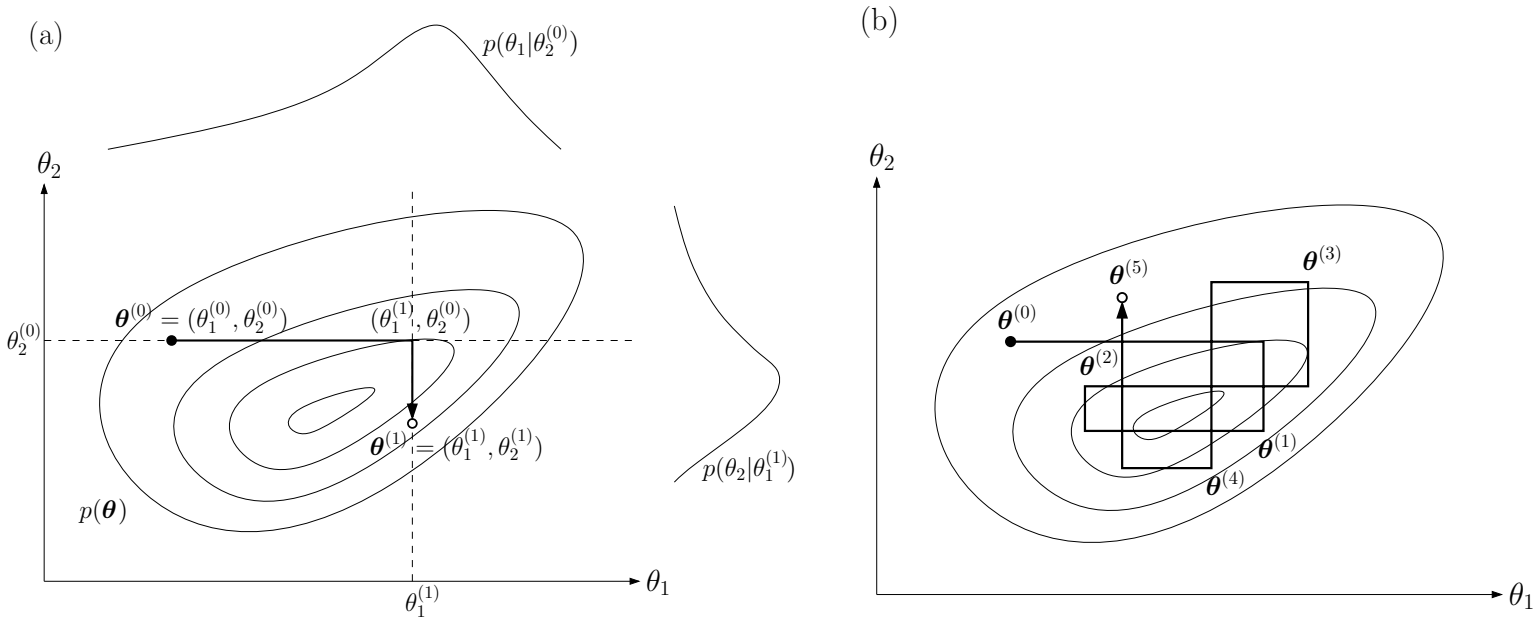
$$\theta_1 \mid \theta_2 \sim \mathrm{N}(\rho\,\theta_2, 1-\rho^2) \qquad \theta_2 \mid \theta_1 \sim \mathrm{N}(\rho\,\theta_1, 1-\rho^2)$$

A Gibbs sampler for this distribution, starts from initial values ${\theta_1}^{(0)}, {\theta_2}^{(0)}$. Then for $t = 1, 2, \ldots, T$

1. Draw ${\theta_1}^{(t)} \sim p(\theta_1 \mid {\theta_2}^{(t-1)})$
2. Draw ${\theta_2}^{(t)} \sim p(\theta_2 \mid {\theta_1}^{(t)})$

This draws T iterations or T MCMC samples or T samples: these terms are used interchangeably (note "samples" here does not refer to the observations/data you are using in your analysis!)

Two-stage Gibbs sampling for a bivariate normal (cont)



- Sample ${\theta_1}^{(1)}$ from $p(\theta_1 \mid {\theta_2}^{(0)})$
- Sample ${\theta_2}^{(1)}$ from $p(\theta_2 \mid {\theta_1}^{(1)})$
- Sample ${\theta_1}^{(2)}$ from $p(\theta_1 \mid {\theta_2}^{(1)})$
- .....

Note the way the Gibbs sampler always move parallel to an axis: this means it can performs poorly when correlation is high.

Gibbs sampling for general posterior distributions

To draw samples from $p(\boldsymbol{\theta} \mid y)$, with parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$, given observations y.

We use the "full conditional distributions" which condition on all other parameters.

1. Choose starting values ${\theta_1}^{(0)}$, ${\theta_2}^{(0)}$, ..., ${\theta_k}^{(0)}$
2. Sample ${\theta_1}^{(1)}$ from $p(\theta_1 \mid {\theta_2}^{(0)}, \theta_3^{(0)}, \ldots, {\theta_k}^{(0)}, y)$
   Sample ${\theta_2}^{(1)}$ from $p(\theta_2 \mid {\theta_1}^{(1)}, \theta_3^{(0)}, \ldots, {\theta_k}^{(0)}, y)$
   ...
   Sample ${\theta_k}^{(1)}$ from $p(\theta_k \mid {\theta_1}^{(1)}, {\theta_2}^{(1)}, \ldots, \theta_{k-1}^{(1)}, y)$

   } "full conditional distributions"
3. Repeat step 2 many 1000s of times

JAGS can calculate the form of these full conditionals automatically

Convergence: basics

As the number of iterations $t \to \infty$, eventually we know the samples will be from the posterior.

However, we can't draw infinite samples or wait forever...

- We must choose a finite set of samples to use for our Monte Carlo estimates
- These samples must be from the stationary distribution of the Markov chain: that is the chain must have converged.

Convergence here means to a distribution (the required posterior), not to a single value.

- Not like optimisation algorithms that seek *e.g.* the maximum

Convergence: basics (cont)

In practice the following simple approach is used.

Draw T iterations using MCMC, then

- Declare that the chain converges after T^* iterations
- Samples 1 to $T^* - 1$ are thus pre-convergence
  - Discard these samples. We call these burn-in samples
- Samples T^* to T are thus from the stationary distribution
  - Use these samples to estimate all results.

We need to choose T and T^*

There is no formula or simple foolproof rule to choose these – this is the main practical challenge of using MCMC

Convergence

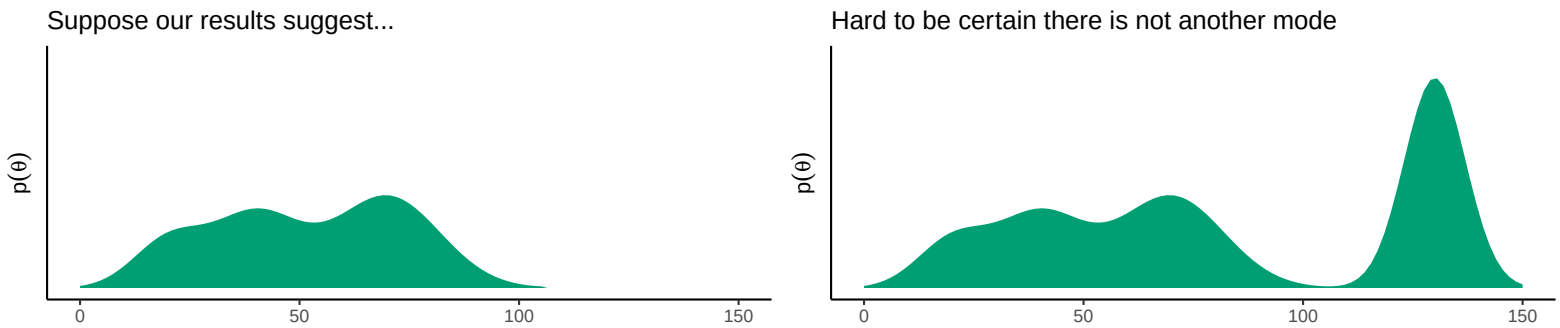
The following workflow is often used:

1. Choose initial T and T^*
   - Usually chosen fairly arbitrarily, *e.g.* $T = 4000$ and $T^* = T/2$
   - May be informed by experience with similar models/data
2. Run T MCMC iterations
3. Assess whether iterations T^* to T are "converged"
   - Convergence diagnosis
4. If so, check whether Monte Carlo estimates resulting from iterations T^* to T are stable enough
   - Monte Carlo standard error and effective sample size
5. Otherwise, increase T and/or T^*, and repeat...

Convergence assessment

Is hard assess whether iterations T^* to T are "converged"

1. Conceptually, convergence to a distribution is less clearcut than convergence to *e.g.* a maximum
2. Hard to be sure we haven't "missed" the global mode



No general way to prove convergence. However, convergence diagnostics can identify *lack of convergence*. This can reassure us results are unlikely to be artifactual

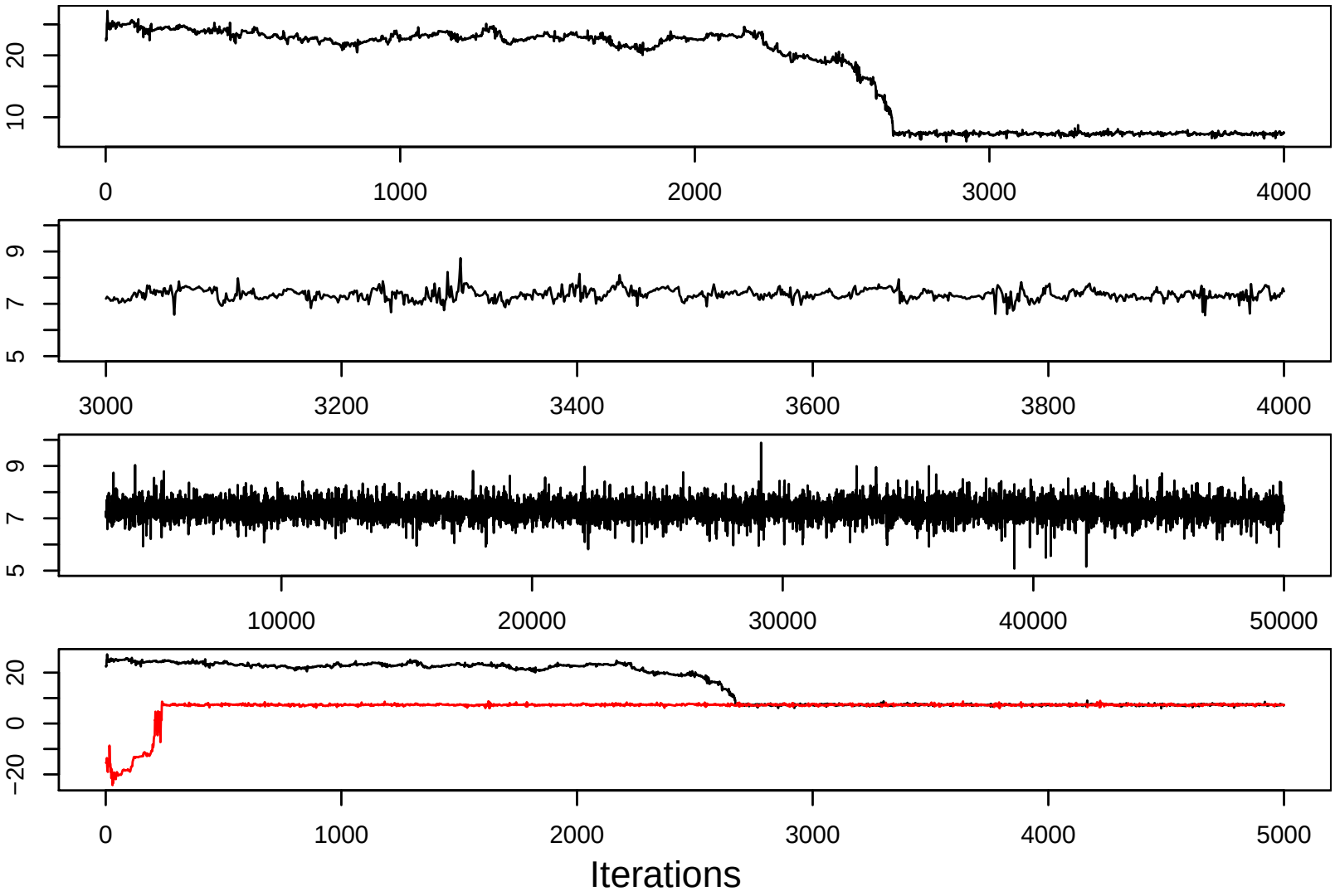
Convergence assessment

Several convergence diagnostic tools are widely accepted...

Traceplots:

- This is the simplest convergence diagnostic, but is often very effective
- Samples should look like a random scatter about a stable mean value
- Look for “fat hairy caterpillars”

Convergence: basics (cont)



Convergence: Brooks-Gelman-Rubin diagnostic

Running multiple separate MCMC chains can be very helpful

- Reassuring if all chains converge to the same distribution, and thus give the same results
- Particularly reassuring if the chains were started from fairly different initial values
- Essentially a sensitivity analysis for MCMC

When running multiple chains, select initial values to be

- different for different chains
- not so extreme, in the context of the prior distribution and the data, that the sampler will get stuck and fail to find the true posterior.

(beyond this, the exact values used are not important)

Gelman-Rubin statistic

The Gelman-Rubin statistic[3] aims to quantify (lack of) convergence using multiple chains. (*It is not foolproof!*)

- Quantifies whether chains are much further apart than expected based on their internal variability
- Several formulations are in use: in loose terms, idea is to look at

$$\widehat{R} = \sqrt{\frac{\text{across chain variance}}{\text{within chain variance}}}$$

- If all chains completely agree $= 1$, so will get nearer 1 as more iterations are drawn
- No definitive threshold to declare "convergence" but 1.1 has been proposed as a rule of thumb

[3] Also called Brooks-Gelman-Rubin statistic, Potential Scale Reduction Factor (PSRF) and Rhat (or $\widehat{R}$)

Summary of convergence diagnostics

A reasonable and widely-accepted set of convergence diagnostics is

1. Trace plot
2. Multiple chains, started from different initial values
3. Gelman-Rubin statistic

Numerous other methods have been proposed:

- Formal diagnostics for a single chain *e.g.* Geweke's diagnostic
- Trace plots using ranks rather than values
- Autocorrelation plots
- "Pairs plots" to examine correlation between parameters

Efficiency/Accuracy

After convergence, further iterations are needed to obtain samples for posterior inference.

In Session 1, with T independent samples, Monte Carlo standard error was:

$$sd(\theta)/\sqrt{T}$$

For the dependent samples produced using MCMC, the effective sample size (ESS) is usually smaller than the number of iterations/MCMC samples: highly correlated samples tell you less information. Use ESS in place of T in the formula.

If only interested in the posterior mean then could run till

$$\text{mean} \pm 2 \times \text{MC error}$$

to agree for a particular number of significant figures.

Key points about MCMC

1. MCMC is a form of Monte Carlo simulation that produces **dependent** samples.

2. Gibbs sampling is one form of MCMC that makes moves **parallel to an axis**

3. JAGS can't check that the MCMC chain it produces has converged – **you must do this**.

4. Running multiple chains from **different starting points** helps to check for convergence.

5. More iterations = more accurate posterior estimates. We can measure accuracy with **effective sample size** and **MC error**.

Practical 2

4. Convergence assessment in JAGS
   - Looking at traceplots, multiple chains and Gelman-Rubin statistics
5. Gibbs sampler for a bivariate normal in R
   - Mechanics of a Gibbs sampler
   - Example of when a Gibbs sampler works poorly
6. Gibbs sampler for a posterior (optional)
   - Mechanics of a Gibbs sampler for a posterior