

MAS8303 Modern Bayesian Inference
Part 2

M. Farrow
School of Mathematics and Statistics
Newcastle University

Semester 1, 2012-13

Chapter 3

Missing Data and Data Augmentation

3.1 Introduction to Graphical Models

In the rest of the module it will sometimes be useful to use graphical representations of models. We will look at a particular type of graph called a *directed acyclic graph* or *dag*.

Suppose that we are going to observe a number of animals of the same species. Each animal might or might not have a particular gene. Suppose that all animals in the population are considered to be exchangeable with respect to having this gene. Let $T_i = 1$ if animal i has the gene. Otherwise $T_i = 0$. Because of the exchangeability, we can represent the relationships in our beliefs about T_1, T_2, \dots by introducing θ to represent the unknown overall proportion of animals in the population which have the gene. The graph for two animals is shown in Figure 3.1.

With three unknowns, A, B, C , we can always write the joint probability as

$$\Pr(A, B, C) = \Pr(A) \Pr(B|A) \Pr(C|A, B).$$

In the example this might have led us to write

$$\Pr(\theta, T_1, T_2) = \Pr(\theta) \Pr(T_1|\theta) \Pr(T_2|\theta, T_1).$$

In fact the last term is just $\Pr(T_2|\theta)$ since T_2 is *conditionally independent* of T_1 given θ . In other words we do not draw an arrow (or *arc*) from T_1 to T_2 in figure 3.1. We can build up the joint probability as a product of one marginal probability and a sequence of conditional probabilities. The direction of the arcs denotes the order in which we are doing this and the arcs leading into a node indicate on which other unknowns we need to condition at each step. In a way, the important feature, therefore, is which possible arcs are missing. Note that we can not have a directed cycle in such a graph. The graphs are sometimes called *directed acyclic graphs* or DAGs. They are also sometimes called *influence diagrams*.

Figure 3.2 gives another example of a DAG. Here A and D are independent, B and E are conditionally independent given C and each of B, E is conditionally independent of each of A, D given C . The joint probability can be written

$$\Pr(A, B, C, D, E) = \Pr(A) \Pr(D) \Pr(C|A, D) \Pr(B|C) \Pr(E|C). \quad (3.1)$$

There is no unique graph for a group of random variables.

Suppose, in the example of figure 3.2, we wanted A to be the only node with no parents. This means reversing the direction of the arc between D and C . It is not quite as simple as this though. The joint probability is given by (3.1). By Bayes theorem,

$$\Pr(C | A, D) = \frac{\Pr(C | A) \Pr(D | A, C)}{\Pr(D | A)}$$

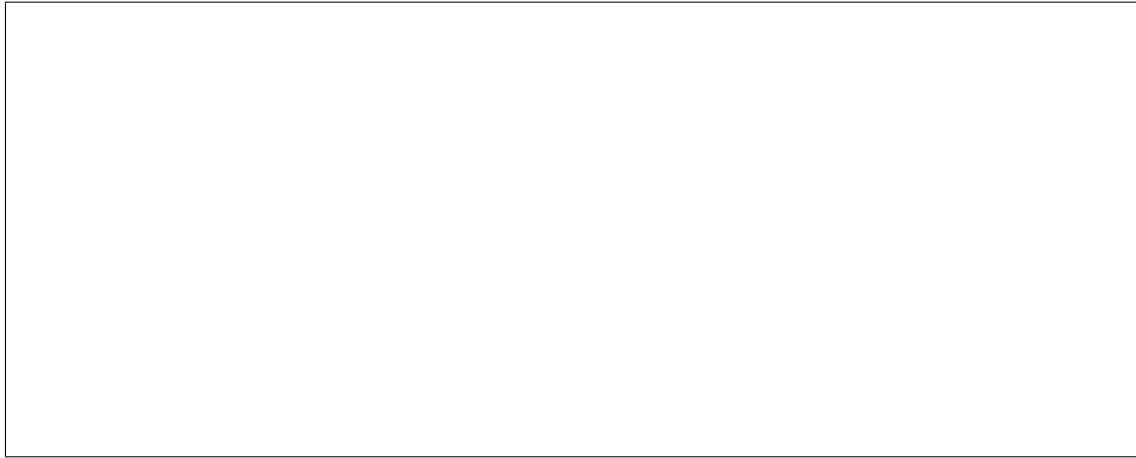


Figure 3.1: Graphical model for animals example

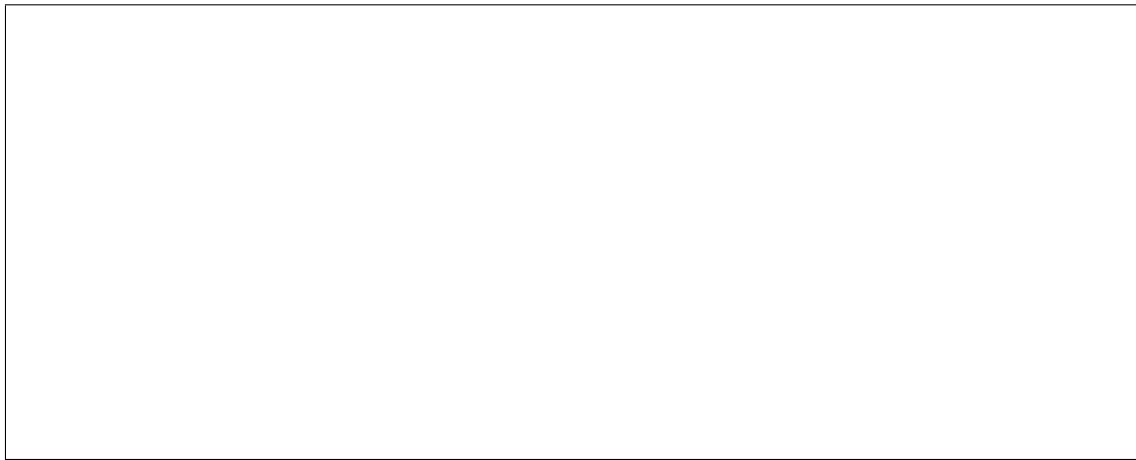


Figure 3.2: Directed acyclic graph

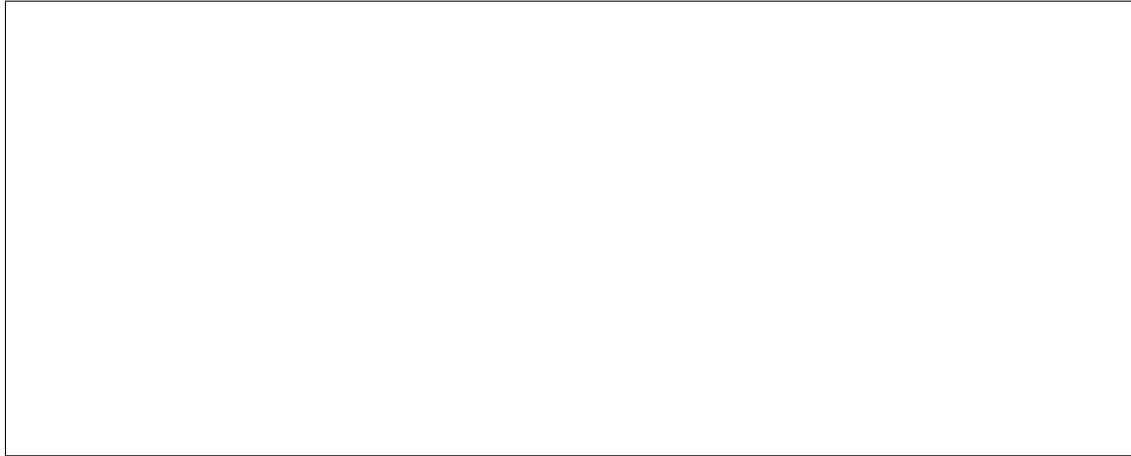


Figure 3.3: Arc reversal

but $\Pr(D | A) = \Pr(D)$ so

$$\Pr(D) \Pr(C | A, D) = \Pr(C | A) \Pr(D | A, C).$$

Hence we replace (3.1) with

$$\Pr(A, B, C, D, E) = \Pr(A) \Pr(C|A) \Pr(D|A, C) \Pr(B|C) \Pr(E|C).$$

So (unless D is conditionally independent of A given C , which would not be true in general) we need to add an arc from A to D , as in figure 3.3. The general rule is that we can reverse the direction of an arc between two nodes, N_1 and N_2 , provided that

1. we do not create a directed cycle by doing so and
2. any node which is a parent of either N_1 or N_2 is made a parent of both.

See Figure 3.3.

Going back to figure 3.2, suppose we wished to eliminate C . Then we could replace (3.1) with

$$\Pr(A, B, D, E) = \Pr(A) \Pr(D) \Pr(B|A, D) \Pr(E|A, B, D)$$

which gives the diagram in figure 3.4. To see this, first note that we can *always* remove a node which has no children without having to make any other changes. This is obvious since we are just dropping a term from the end of the joint probability factorisation. So, we can arrange for the node which we want to delete to have no children by suitable arc reversals. For example, starting with figure 3.2, we could reverse the arc between B and C and then reverse the arc between C and E . This would leave us with figure 3.4. If we did the reversals in the other order we would get a slightly different result.

The procedure using arc reversals to eliminate a node will work but we might end with a graph which has more arcs than are necessary. There is a general rule which can be used which avoids this problem. (You need not memorise this rule). The general rule is as follows. If a node N is eliminated then:

- every child of N inherits all parents of N ,
- every pair of the children of N is connected by an arc,
- every child that receives an arrow from another child inherits all parents of the latter.

See Pearl, Geiger and Verma (1990) p82.

Figure 3.5 shows a simple repeated measures model with three observations on each of two individuals. The model is as follows.

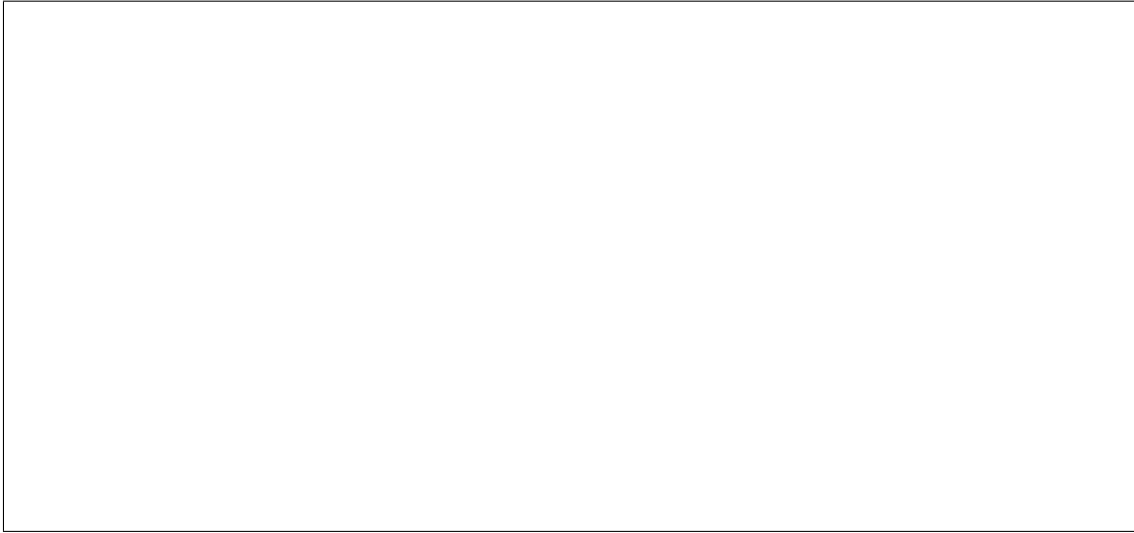


Figure 3.4: Node deletion

Suppose we have k samples of observations and observation j in sample i is

$$Y_{ij} = \theta_i + \varepsilon_{ij}$$

for $j = 1, \dots, J$, where $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ and $\theta_i \sim N(\mu, \sigma_\theta^2)$ (all independent). Suppose we have independent priors for the three parameters:

$$\begin{aligned} \mu &\sim N(\mu_0, \sigma_0^2) \\ \sigma_\theta^2 &\sim \text{IG}(a_1, b_1) \\ \sigma_\varepsilon^2 &\sim \text{IG}(a_2, b_2) \end{aligned}$$

where IG stands for “inverse gamma” (i.e. $(\sigma^2)^{-1}$ has a gamma distribution).

The diagram shows the model with $k = 2$ and $J = 3$.

We will see similar models in a later lecture.

Figure 3.6 shows an example of how we might represent relationships in our prior beliefs about quantities. Here M_1, M_2, M_3 are all related because they share a common parent, U_3 . Thus, if we learn something about M_1 this will affect our beliefs about M_2 and M_3 . Furthermore M_1 is more strongly related to M_2 than to M_3 because M_1 and M_2 share a common parent, U_1 , which is not a parent of M_3 .

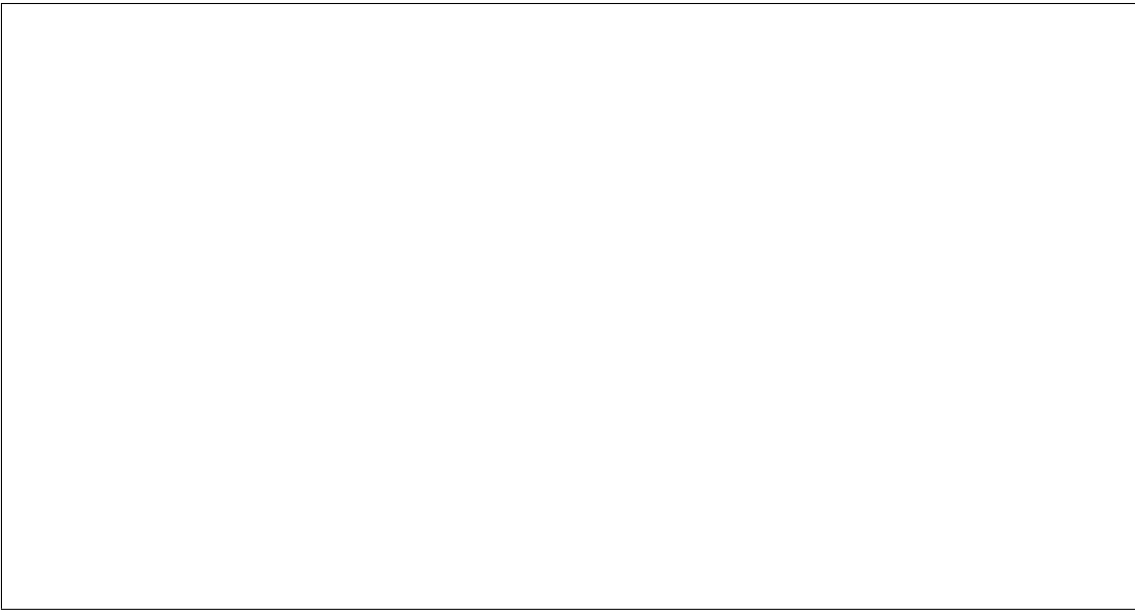


Figure 3.5: Repeated measures model



Figure 3.6: Three related quantities

3.2 Missing Data

3.2.1 Motivation

Consider a regression model with a dependent variable Y and explanatory variables X_1, \dots, X_p . This could be a linear model, a generalised linear model or some other kind of model such as a survival model. In some cases the values of X_1, \dots, X_p will be deliberately chosen in a designed experiment. In other cases the data will arise from an *observational study* in which we observe the variables for each of a sample of individuals from some population. In the latter case it is possible that, for some reason, the values of one or more of the explanatory variables are missing. In fact this is not particularly unusual. Regression is not the only situation where this might be a problem. It might apply in other cases when we make multivariate observations.

What can we do? We can not simply make inferences about the model parameters in the ordinary way when observations on some variables are missing. One possibility would be simply to delete any observation where there are missing data. This is not satisfactory for two reasons as follow.

1. We would be losing information which could be used.
2. We might make misleading inferences as a result. It could be that the cases where data are missing are also different in some other way from the complete cases.

From the Bayesian viewpoint the missing values are simply “unknowns” in the same way as other unknowns such as model parameters. We can therefore extend our model to include a model for the variables which may be missing. If a variable X is sometimes, but not always, missing then we can (subject to certain assumptions – see below) use the cases where it is present to learn about its relationship with other variables. Therefore we can obtain a posterior distribution for the missing values. In fact we obtain a joint posterior distribution for the model parameters and the missing values. We can then “integrate out” the missing values to obtain the marginal posterior distribution of the model parameters. In effect, our inferences about model parameters are “averaged” over the distribution of possible values of the missing data. MCMC methods are well suited to handling problems of this type and it is often quite straightforward to handle missing-data problems using software such as BUGS.

We need to do two things:

1. We need to consider the nature of the “missingness” to see what it makes sense to do.
2. If it does make sense to proceed then we need a “missing data model”. That is a model which shows how the variables which are sometimes missing are related to other variables.

3.2.2 Different kinds of missingness

Suppose that we make a multivariate observation \underline{Y} on each of a number of individuals. In the regression example above, \underline{Y} would contain both the dependent variable Y and the explanatory variable X_1, \dots, X_p . For an individual i the vector of values of the variables is $\underline{y}_i = (y_{i,1}, \dots, y_{i,J})^T$. However some of the values $y_{i,1}, \dots, y_{i,J}$ might be missing and therefore not observed. We introduce the *inclusion indicator* $I_i = (I_{i,1}, \dots, I_{i,J})^T$ where $I_{i,j} = 1$ if $y_{i,j}$ is observed and $I_{i,j} = 0$ if $y_{i,j}$ is missing.

We introduce (vector) parameters θ, ϕ such that, given θ and ϕ , we can write the joint probability (density) of \underline{y}_i, I_i as

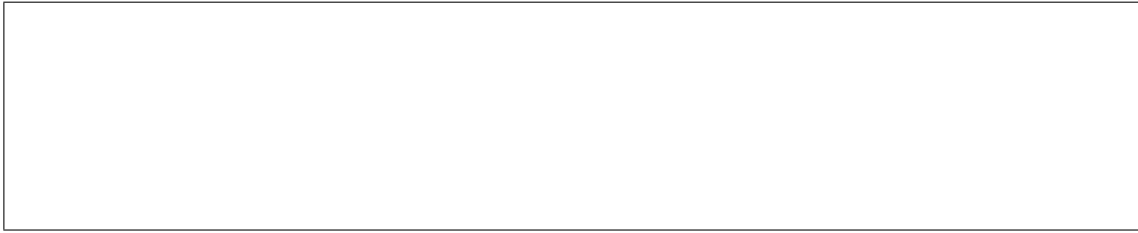
$$f_{y,I}(\underline{y}_i, I_i | \theta, \phi) = f_y(\underline{y}_i | \theta) p_I(I_i | \underline{y}_i, \phi).$$

Let us consider all of the data (for all individuals) together. We write \underline{y} for the complete data on all individuals, I for the inclusion indicator for all individuals, which is now a matrix, and so on. Then we write


$$f_{y,I}(\underline{y}, I | \theta, \phi) = f_y(\underline{y} | \theta) p_I(I | \underline{y}, \phi).$$

(The meaning of f and p has, of course, changed here).

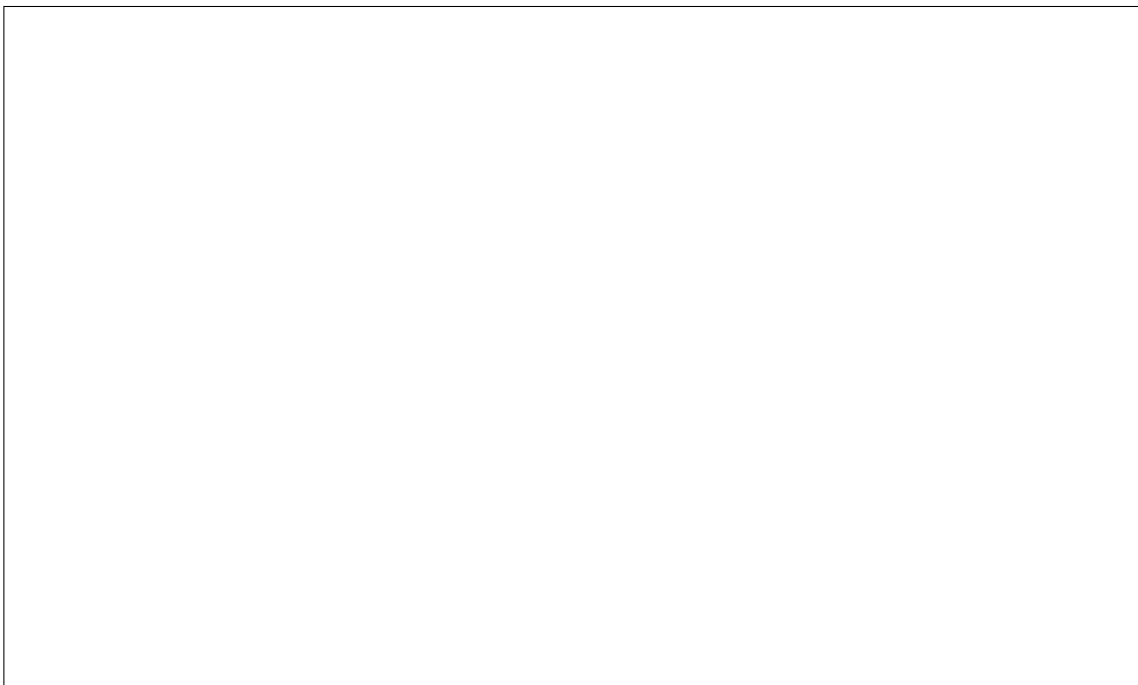
Missing data mechanism :

A large, empty rectangular box with a thin black border, intended for a response to the question 'Missing data mechanism :'. It occupies the upper middle portion of the page.

Observed data and likelihood :

A large, empty rectangular box with a thin black border, intended for a response to the question 'Observed data and likelihood :'. It occupies the middle portion of the page.

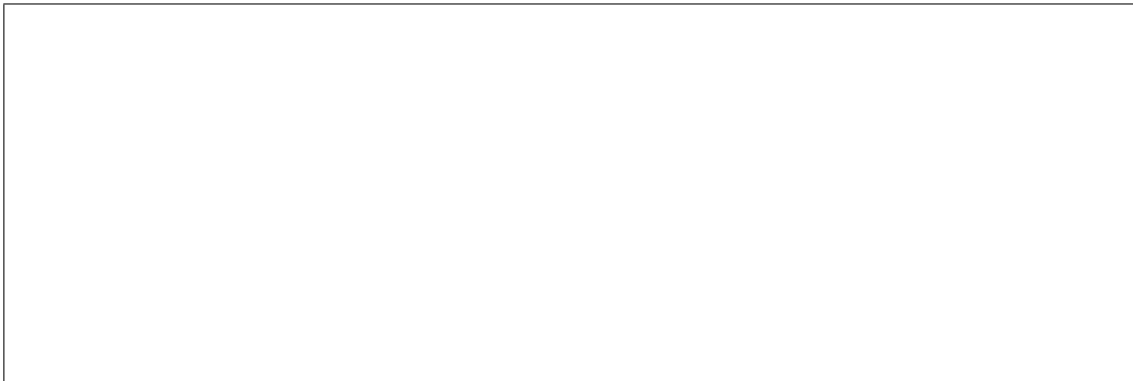
Missingness at random :

A large, empty rectangular box with a thin black border, intended for a response to the question 'Missingness at random :'. It occupies the lower portion of the page.

Ignorable missing-data mechanism :



Missingness completely at random :



Notice that it is not usually necessary to assume MCAR for Bayesian inference. It is usually sufficient to have MAR - ignorable.

The MAR assumption is more plausible when we observe a large number of variables since the observed values are then more likely to convey enough information to make missingness conditionally independent of the missing values.

3.2.3 Missing data models

Consider the abrasion loss example in Practical 1 (Section 1.4.1). Suppose that some of the hardness (X_1) and tensile strength (X_2) measurements are missing.

We need a model for the joint distribution of X_1 and X_2 . The existing regression model is just a model for the conditional distribution of Y given X_1 and X_2 . For example, we could say

$$\begin{aligned}
 X_1 \mid \mu_1, \tau_1 &\sim N(\mu_1, \tau_1^{-1}) \\
 \mu_1 &\sim N(60, 400) \\
 \tau_1 &\sim \text{Ga}(1, 100) \\
 X_2 \mid x_1, \delta_1, \delta_2, \tau_2 &\sim N(\delta_1 + \delta_2[x_1 - 60], \tau_2^{-1})
 \end{aligned}$$

$$\begin{aligned}\delta_1 &\sim N(200, 2500) \\ \delta_2 &\sim N(0, 1) \\ \tau_2 &\sim \text{Ga}(1, 2000)\end{aligned}$$

Notice that we have done this by giving X_1 a distribution and then giving X_2 a conditional distribution given X_1 . (Of course we could have done it the other way round). There are many possibilities for the way we build a “missing data model” depending on what the variables are. For example, we might have a binary variable which we could relate to a continuous variable through a logistic regression (or we could give the continuous variable two different conditional distributions depending on the value of the binary variable).

Figure 3.7 shows a BUGS model specification in the abrasion loss example. (Note that we are not using the fully conjugate prior here). The data file would simply contain NA where a value is missing.

```
model abrasion

{for (i in 1:30)
  {loss[i]~dnorm(lossmean[i],tau)
   lossmean[i]<=-alpha+beta[1]*(hard[i]-60)+beta[2]*(tens[i]-200)
   hard[i]~dnorm(muhard,tau.hard)
   tens[i]~dnorm(tensmean[i],tau.tens)
   tensmean[i]<=-delta[1]+delta[2]*(hard[i]-60)
  }

  alpha~dnorm(150,0.000625)
  beta[1]~dnorm(0,0.0025)
  beta[2]~dnorm(0,0.0025)
  beta0<=-alpha+60*beta[1]+200*beta[2]
  muhard~dnorm(60,0.0025)
  delta[1]~dnorm(200,0.0004)
  delta[2]~dnorm(0,1)

  tau.tens~dgamma(1,2000)
  tau.hard~dgamma(1,100)
  tau~dgamma(2,3200)
}
```

Figure 3.7: BUGS model specification for abrasion loss example with missing data

3.3 Data augmentation

3.3.1 Introduction

Some models have rather complicated likelihood functions which, if handled directly, would lead to difficult calculations. Sometimes it is possible to make things much simpler by introducing extra variables, known as *auxiliary variables*, which are not observed but, which, if they were observed, would make the likelihood simpler. These auxiliary variables are then treated as if they were missing data. This is known as *data augmentation*. MCMC methods are well suited to this approach.

3.3.2 Example 1: Mixture models

In MAS3301 we looked at the use of mixture distributions as priors. We can also use mixtures as sampling distributions. The likelihood can be complicated and difficult to calculate but we can make things much simpler by introducing a group-membership variable which is unobserved. We will look at the case of mixtures in Chapter 4.

3.3.3 Example 2: Student t -model

In a normal linear model we have

$$Y_i | \mu_i, \sigma^2 \sim N(\mu_i, \sigma^2)$$

$$\text{where } \mu_i = \sum_{j=1}^J \beta_j x_{i,j}$$

Suppose instead we want to use a Student's t -distribution for the errors so

$$\frac{Y_i - \mu_i}{\sigma} | \mu_i, \tau \sim t_d,$$

where t_d represents the Student's t distribution on d degrees of freedom. We assume that d is chosen. A small value of d makes the error distribution “heavy-tailed.”

The likelihood in this model is such that sampling for a Gibbs sampler would be difficult. However we can overcome this problem by introducing auxiliary variables X_i where

$$d\sigma^2 X_i \sim \chi_d^2.$$

Then we let

$$Y_i | \mu_i, X_i \sim N(\mu_i, X_i^{-1}).$$

Now we get the following properties.

- Y_i has the required error distribution.

Proof: Since $d\sigma^2 X_i \sim \chi_d^2$ we have $X_i \sim \text{Ga}(d/2, d\sigma^2/2)$. Therefore the joint density of X_i and Y_i given μ_i and σ^2 is

$$\begin{aligned} f_{X,Y}(x_i, y_i | \mu_i, \sigma^2) &= (2\pi)^{-1/2} x_i^{1/2} \exp\left\{-\frac{x_i}{2}(y_i - \mu_i)^2\right\} \frac{(d\sigma^2/2)^{d/2} x_i^{d/2-1} e^{-d\sigma^2 x_i/2}}{\Gamma(d/2)} \\ &= (2\pi)^{-1/2} \frac{(d\sigma^2/2)^{d/2}}{\Gamma(d/2)} x_i^{(d+1)/2-1} e^{-x_i[d\sigma^2+(y_i-\mu_i)^2]/2} \\ &= \frac{(2\pi)^{-1/2} (d\sigma^2/2)^{d/2} \Gamma([d+1]/2)}{\Gamma(d/2) ([d\sigma^2+(y_i-\mu_i)^2]/2)^{(d+1)/2}} \\ &\quad \times \frac{([d\sigma^2+(y_i-\mu_i)^2]/2)^{(d+1)/2}}{\Gamma([d+1]/2)} x_i^{(d+1)/2-1} e^{-x_i[d\sigma^2+(y_i-\mu_i)^2]/2} \end{aligned}$$

Now integrate with respect to x_i and the second term, which is a gamma density, integrates to 1. So the density of Y_i is

$$\begin{aligned} f_Y(Y_i) &= \frac{(2\pi)^{-1/2}(d\sigma^2/2)^{d/2}\Gamma([d+1]/2)}{\Gamma(d/2)([d\sigma^2 + (y_i - \mu_i)^2]/2)^{(d+1)/2}} \\ &= \pi^{-1/2}d^{d/2}\sigma^d \frac{\Gamma([d+1]/2)}{\Gamma(d/2)} [d\sigma^2 + (y_i - \mu_i)^2]^{-(d+1)/2} \\ &= (\pi d)^{-1/2}\sigma^{-1} \frac{\Gamma([d+1]/2)}{\Gamma(d/2)} \left[1 + d^{-1} \left(\frac{y_i - \mu_i}{\sigma} \right)^2 \right]^{-(d+1)/2} \end{aligned}$$

Now let $T_i = (Y_i - \mu_i)/\sigma$. Then $dt/dy = 1/\sigma$ so the density of T_i is

$$f_T(t_i) = (\pi d)^{-1/2} \frac{\Gamma([d+1]/2)}{\Gamma(d/2)} \left[1 + \frac{t_i^2}{d} \right]^{-(d+1)/2}.$$

This is the density of a t_d distribution, as required.

- Given a multivariate normal prior for β_1, \dots, β_J , the full conditional distribution of β_1, \dots, β_J (conditioning on values for X_1, \dots, X_J) is also multivariate normal and therefore easy to sample.
- Given a value for σ^2 and values for μ_1, \dots, μ_n , the full conditional distribution for each X_i is a gamma distribution and therefore easy to sample.
- Given values for X_1, \dots, X_n and a gamma prior for σ^2 (Note: σ^2 in this case, not $\tau = \sigma^{-2}$), the full conditional distribution of σ^2 is also a gamma distribution and easy to sample.

3.3.4 Example 3: Integrated moving average processes

Integrated moving average processes are commonly used as models for nonstationary time series. The integrated first order moving average process, denoted IMA (0,1,1), is especially useful.

Let the observation at time t be y_t . Let $z_t = y_t - y_{t-1}$. Then we model z_t using a first order moving average, MA(1), process

$$z_t = \varepsilon_t + \theta\varepsilon_{t-1}$$

where $\dots, \varepsilon_{t-2}, \varepsilon_{t-1}, \varepsilon_t, \varepsilon_{t+1}, \dots$ are independent and each has distribution $\varepsilon_j \sim N(0, \sigma^2)$, given σ^2 .

This is a stationary process. For reasons of identifiability we restrict θ to $-1 \leq \theta \leq 1$. Given the parameters, the moments of the process are as follows.

$$\begin{aligned} E(z_t) &= 0, \\ \gamma_0 = \text{var}(z_t) &= \sigma^2(1 + \theta^2), \\ \gamma_1 = \text{covar}(z_t, z_{t-1}) &= \sigma^2\theta, \\ \gamma_k = \text{covar}(z_t, z_{t-k}) &= 0 \quad (k > 1). \end{aligned}$$

(Note that the mean is zero because we have not added a nonzero “drift” into the model).

If we observe $\underline{y} = (y_1, \dots, y_n)^T$ then, we can transform this to the equivalent observation y_1, z_2, \dots, z_n where z_2, \dots, z_n is a realisation of a MA(1) process. Given the parameters, this is an observation from a multivariate normal distribution in which the variance matrix is a function of the unknown parameters σ^2 and θ .

Things take on a simpler form if we observe that, conditional on ε_{t-1} and the model parameters, the distribution of z_t is normal with mean

$$E(z_t \mid \varepsilon_{t-1}, \underline{\theta}) = \theta\varepsilon_{t-1}$$

and variance σ^2 . This now looks like a straightforward normal linear regression (with no intercept because we are not fitting a nonzero mean). However, we need to know the values of ε_{t-1} . Apart

from one catch, this problem is easily solved since, if we know ε_{t-1} and z_t and the parameters, we can calculate

$$\varepsilon_t = z_t - \theta\varepsilon_{t-1}.$$

Therefore we can calculate the ε values recursively through the time series. The catch is that we need a starting value at the beginning of the series and we do not have it. That is, we do not have ε_1 which we need to calculate ε_2 . The solution is to *augment* the data by including ε_1 as an auxiliary variable. Since y_1 and ε_1 are not independent and we have the observation y_1 , we really ought to use this information. One way to do this is to write $y_1 = \mu + \varepsilon_1$ and specify a (normal) prior for μ .



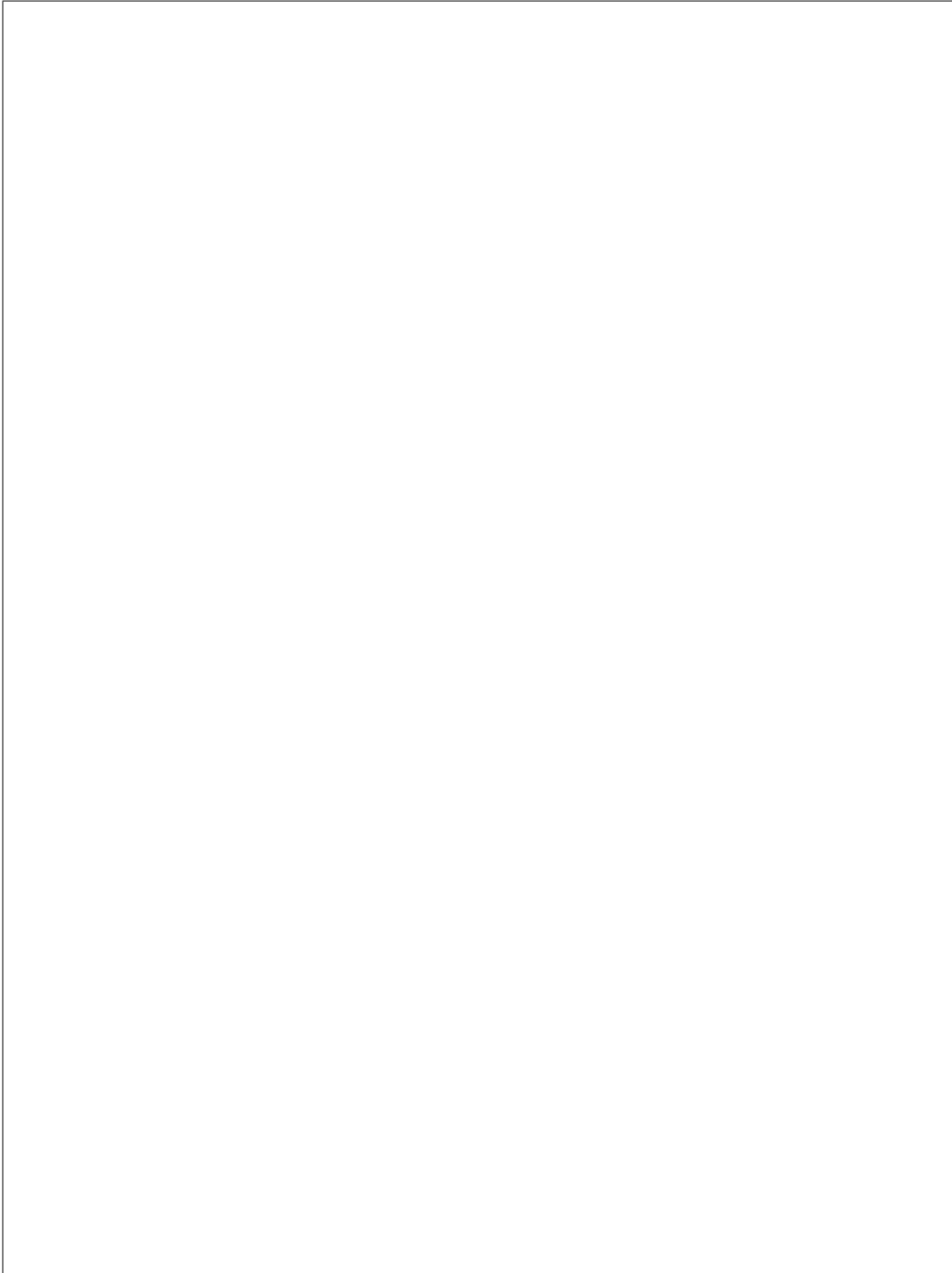


Figure 3.8 shows a possible BUGS model specification using this second approach. We can regard x_1, \dots, x_n as auxiliary data. In fact, in this model specification I have also introduced x_0 to help to construct the prior. There are different ways to construct priors for models of this sort but a method like that shown here is often used if we wish to use the model for forecasting.

```

model ima
{
  for (i in 1:n)
    {y[i]~dnorm(x[i],tau.y)
    }

  x0~dnorm(400,0.0001)
  x[1]~dnorm(x0,tau.x)

  for (i in 2:n)
    {x[i]~dnorm(x[i-1],tau.x)
    }

  tau.x~dgamma(1,10)
  tau.y~dgamma(1,10)

  k<- -(tau.y/tau.x+2)
  theta1<-k+sqrt(pow(k,2)-4)
  theta2<-k-sqrt(pow(k,2)-4)
  theta<-max(theta1,theta2)
  sigmasq<- 1/(theta*tau.y)
}

```

Figure 3.8: BUGS model specification for integrated moving average IMA (0,1,1) process

3.4 Practical 3

3.4.1 Abrasion Loss

Consider the abrasion loss example in Practical 1 (Section 1.4.1). Suppose that some of the hardness (X_1) and tensile strength (X_2) measurements are missing.

1. Obtain a copy of the data file `abrasion.txt` used in Practical 1. Edit the file by changing the first two hardness values to `NA` and the third and fourth tensile strength values to `NA`. Add headings for the columns as shown below. The first few lines of the file should now look like this.

```

loss[] hard[] tens[]
372 NA 162
206 NA 233
175 61 NA
154 66 NA
136 71 231
112 71 237

```

We now have four missing values in the file.

Add the word `END` at the end of the file. The end of the file should now look like this.

```

283 65 148
267 74 144
215 81 134
148 86 127
END

```

Save the file with the name `abmiss.txt` .

2. Obtain a copy of the BUGS model file from the module Web page.
3. In R, check the syntax of the model.

```
modelCheck("abmissbug.txt")
```

4. Load the data.

```
modelData("abmiss.txt")
```

5. Compile the model.

```
modelCompile()
```

6. Generate initial values.

```
modelGenInits()
```

7. Run the sampler for a burn-in period.

```
modelUpdate(5000)
```

8. Start recording samples. You do not have to record everything, of course, but, in this case, it might be interesting to see the sampled values of the missing data as well as the model parameters.

```
samplesSet(c("beta0", "beta", "tau", "muhard", "tau.hard", "delta", "tau.tens", "hard", "tens"))
```

9. Run the sampler for some more iterations.

```
modelUpdate(10000)
```

10. Look at the results in the various ways which you have seen and compare the posterior summaries with those obtained in Practical 1 when no observations were missing.
11. Try running this model again but this time with two chains to check convergence. To see what happens we will not use a burn-in.

```
modelCheck("abmissbug.txt")
modelData("abmiss.txt")
modelCompile(2)
modelGenInits()
samplesSet(c("beta0", "beta", "tau", "muhard", "tau.hard", "delta", "tau.tens", "hard", "tens"))
modelUpdate(10000)
samplesHistory("beta0")
```

Etc.

```

model abrasion

{for (i in 1:30)
  {loss[i]~dnorm(lossmean[i],x[i])
   lossmean[i]<-alpha+beta[1]*(hard[i]-60)+beta[2]*(tens[i]-200)
   x[i]~dgamma(5,w)
  }

alpha~dnorm(150,0.000625)
beta[1]~dnorm(0,0.0025)
beta[2]~dnorm(0,0.0025)
beta0<-alpha+60*beta[1]+200*beta[2]

w<-5*v
v~dgamma(2,0.00125)

}

```

Figure 3.9: BUGS model specification for abrasion loss example with Student t errors.

3.4.2 Abrasion loss with t errors

Let us analyse the abrasion loss data again, this time with no missing data but with Student t errors.

1. Create a new model file, perhaps by editing `abmissbug.txt`. Save the file as `abtbody.txt`. The file should look like figure 3.9. We set the degrees of freedom to 10. We will use data augmentation (although, in fact, BRugs is probably clever enough to handle the problem even without this).
2. Create a data file, called `abrasion.txt`, like `abmiss.txt` but with none of the observations missing.
3. Try the analysis.

```

modelCheck("abtbody.txt")
modelData("abrasion.txt")
modelCompile()
modelGenInits()
modelUpdate(5000)
samplesSet(c("beta0","beta","v"))
modelUpdate(10000)
samplesHistory("beta0")

```

Etc.

4. Compare the results with the results in Practical 1. What do you think is the effect of using t rather than normal errors?

3.4.3 IBM Stock Prices

Table 3.1 shows 160 consecutive daily IBM common stock closing prices. The data may be obtained from the Module Web Page. They were obtained from Box and Jenkins (1976). Box and Jenkins suggest fitting an IMA (0,1,1) model to these data.

1. Use the model file shown in Figure 3.8 to analyse these data. You can obtain the model file from the Module Web Page.

460	457	452	459	462	459	463	479	493	490	492	498	499	497	496	490
489	478	487	491	487	482	479	478	479	477	479	475	479	476	476	478
479	477	476	475	475	473	474	474	474	465	466	467	471	471	467	473
481	488	490	489	489	485	491	492	494	499	498	500	497	494	495	500
504	513	511	514	510	509	515	519	523	519	523	531	547	551	547	541
545	549	545	549	547	543	540	539	532	517	527	540	542	538	541	541
547	553	559	557	557	560	571	571	569	575	580	584	585	590	599	603
599	596	585	587	585	581	583	592	592	596	596	595	598	598	595	595
592	588	582	576	578	589	585	580	579	584	581	581	577	577	578	580
586	583	581	576	571	575	575	573	577	582	584	579	572	577	571	560

Table 3.1: One hundred and sixty consecutive daily IBM common stock closing prices. The data are to be read along the rows.

2. We can calculate forecasts in a straightforward way. Edit the data file. Change $n=160$ to $n=164$. Add NA four times at the end of the list of y values, separated by commas. Repeat the analysis but this time you can record the values of the “missing” Y values, by monitoring y , and thus obtain a forecast distribution.

3.5 Exercise

Table 3.2 shows the numbers of patients undergoing surgery and the numbers who died in the hospital following surgery in two areas of the USA, broken down by age-group and sex. The data are taken from Mosteller and Tukey (1977).

We propose the following model. There are four area-sex groups:

Group 1 : Males in Area 1.

Group 2 : Females in Area 1.

Group 3 : Males in Area 2.

Group 4 : Females in Area 2.

Given the model parameters, the number of deaths in Area-Sex Group j and Age-group k has a binomial $\text{Bin}(n_{j,k}, p_{j,k})$ distribution where $n_{j,k}$ is the number of patients undergoing surgery and

$$\log\left(\frac{p_{j,k}}{1-p_{j,k}}\right) = \alpha_j + \beta_j(x_k - 50)$$

where x_k is the mid-point of the age-range for age-group k .

We need to make inferences about the eight model parameters, $\alpha_1, \dots, \alpha_4, \beta_1, \dots, \beta_4$.

1. Suppose that we consider “typical” patients aged 50. Suppose that for such patients, the probability p_0 of death is α_0 and we give α_0 a normal prior distribution. Suppose that, in our prior beliefs, $\Pr(p_0 < 0.02) = \Pr(p_0 > 0.10) = 0.025$. Find the mean and variance of our normal prior distribution for α_0 .
2. Our joint prior distribution for $\alpha_1, \dots, \alpha_4$ can be represented as follows. We write

$$\begin{aligned}\alpha_j | \bar{\alpha} &\sim N(\bar{\alpha}, V_{\alpha,1}) \quad \text{for } j = 1, \dots, 4, \\ \bar{\alpha} &\sim N(m_\alpha, V_{\alpha,0}).\end{aligned}$$

Here $\alpha_1, \dots, \alpha_4$ are conditionally independent given $\bar{\alpha}$. We choose to make $V_{\alpha,0} = V_{\alpha,1}$ and $V_{\alpha,0} + V_{\alpha,1}$ gives the prior variance of α_0 . Find the values of $V_{\alpha,0}$ and $V_{\alpha,1}$.

Age	Area 1				Area 2			
	Total undergoing surgery		Number dying		Total undergoing surgery		Number dying	
	Males	Females	Males	Females	Males	Females	Males	Females
5-14	4272	3911	9	11	1739	1758	5	2
15-24	2835	2989	23	5	1233	1244	14	1
25-34	2785	2606	19	8	989	1004	8	3
35-44	1930	1886	16	15	897	922	9	13
45-54	1497	1524	59	40	921	961	28	15
55-64	960	1013	101	52	686	739	68	37
65-75	652	855	185	118	611	784	159	73
76-83	186	287	97	108	189	290	86	88

Table 3.2: Deaths following surgery in two areas of the USA

3. We propose a matching structure for β_1, \dots, β_4 with β_1, \dots, β_4 independent of $\alpha_1, \dots, \alpha_4$ in the prior.

$$\begin{aligned}\beta_j | \bar{\beta} &\sim N(\bar{\beta}, V_{\beta,1}) \quad \text{for } j = 1, \dots, 4. \\ \bar{\beta} &\sim N(m_\beta, V_{\alpha,0}).\end{aligned}$$

Here β_1, \dots, β_4 are conditionally independent given $\bar{\beta}$. We choose to make $V_{\beta,0} = V_{\beta,1}$ and $V_{\beta,0} + V_{\beta,1} = 0.0004$. Find the values of $V_{\beta,0}$ and $V_{\beta,1}$. The value of m_β is 0.0.

4. Construct a suitable BRugs model file. Hint: You can use a construction such as `alpha[group[i]]` to denote α_j where observation i belongs to group j .
5. The data are available from the Module Web Page in a file called `surgicaldata.txt`. The data have been arranged into four columns as follows.
- **group**: the area-sex group number as above.
 - **age**: the midpoint of the age range for the age-group.
 - **patients**: the number of patients undergoing surgery.
 - **deaths**: the number of deaths.

Use BRugs to find the posterior distribution of the model parameters. Check convergence of the sampler.

6. Present summaries of the inference, including posterior means and standard deviations of the parameters.
7. Find the posterior mean and standard deviation of $\log(p_1^*/p_3^*)$ where p_1^* is the probability of death for a fifty-year-old male in area 1 and p_3^* is the probability of death for a fifty-year-old male in area 2. Plot the posterior probability density function of this quantity and comment.

3.6 Problems 4

Solutions to all questions are to be submitted in the Homework Letterbox no later than 4.00pm on Wednesday December 12th. Please note that you should give some attention to the presentation of your work. Describe the data, model, prior etc. and explain what you have done. Comment on your conclusions. A listing of the output from a R session with one or two things written on it will not get a very good mark on its own.

In questions 2 and 3, each student is given different data. For this purpose each student is given a reference number according to the table below. Please use the correct data and write your reference number on your work.

Reference numbers

Browning,	Bethany Megan	11
Bulmer,	Rebecca Louise	12
Chaffey,	Adam John	13
Cherlin,	Svetlana	14
Clawson,	Rebecca	15
Consul,	Juliana Iworikumo	16
Dickens,	Jordan Mark	17
Goodall,	Elizabeth Adeline	18
Halliwell,	James William	19
Jones,	Dean Robert Matthew	20
Moffatt,	Joseph Michael	21
Mossop,	Helen	22
Sofro,	A'Yunin	23
Sutherland,	Fiona	24
Varey,	Emma Catherine	25
Wong,	Goldie Sin Man	26

Problems

1. Full conditional distribution

Certain components are manufactured in batches. Each batch contains n components. The components in N batches are then tested and some are found to be defective. Let the number of defective components in batch i be x_i . We suppose that, given the value of π_i , where $0 < \pi_i < 1$, the value of x_i is an observation from a binomial distribution $X_i \sim \text{Bin}(n, \pi_i)$ and X_i and X_j are independent given π_i and π_j when $i \neq j$. Let $\eta_i = \log_e \{\pi_i / (1 - \pi_i)\}$. We suppose that, given the values of μ and τ , η_i is an observation from the normal $N(\mu, \tau^{-1})$ distribution and η_i and η_j are independent, when $i \neq j$, given the values of μ and τ . Finally we have independent prior distributions for μ and τ with μ having a normal prior, $\mu \sim N(m, v)$, and τ having a gamma prior, $\tau \sim \text{Ga}(a, b)$.

Find a function proportional to the density of the full conditional distribution (fcd) of η_i , that is the distribution of η_i given x_i and values for μ and τ .

(10 marks)

2. Piston rings

Four compressors are located in the same building. Each has three “legs”. The compressors are of the same design and are oriented the same way. The three legs of each are labelled “North”, “Centre” and “South.” Over a certain period of time the number of failures of piston rings in each leg of each compressor is counted. These numbers are your data.

The model is as follows. Let the number of failures in leg i of compressor j be $y_{i,j}$ (where $i = 1$ for North, $i = 2$ for Centre and $i = 3$ for South). Given the value of a quantity $\lambda_{i,j} > 0$, we assume that $y_{i,j}$ is an observation from a Poisson distribution $Y_{i,j} \sim \text{Po}(\lambda_{i,j})$, with $Y_{i,j}$ independent of $Y_{i',j'}$ unless $(i, j) = (i', j')$, given the values of $\lambda_{i,j}$ and $\lambda_{i',j'}$.

The prior distribution is as follows. Let $\eta_{i,j} = \log_e(\lambda_{i,j})$. Then

$$\eta_{i,j} = \mu + \alpha_i + \beta_j + \gamma_{i,j}$$

where, $\alpha_1, \dots, \alpha_3, \beta_1, \dots, \beta_4, \gamma_{1,1}, \dots, \gamma_{3,4}$ and μ are mutually independent and

$$\begin{aligned} \mu &\sim N(3, 4), \\ \alpha_i &\sim N(0, 1), & i = 1, \dots, 3, \\ \beta_j &\sim N(0, 1), & j = 1, \dots, 4, \\ \gamma_{i,j} &\sim N(0, 0.25), & i = 1, \dots, 3, j = 1, \dots, 4. \end{aligned}$$

Data.

To obtain your data, first install the R function `pistonread`. This function may be obtained from the Module Web Page, under “Data”, or directly from

<http://www.mas.ncl.ac.uk/~nmf16/teaching/mas8303/pistonreadR.txt>

or simply by typing the following into R.

```
pistonread<-function(refno)
{data<-read.table("http://www.mas.ncl.ac.uk/~nmf16/teaching/mas8303/pistondata.txt")
  out<-cbind(data[,1],data[,2],data[,refno])
  write.table(format(out),row.names=FALSE,col.names=FALSE,quote=FALSE,file="mypistondata.txt")
}
```

Then, use the function with your reference number as the argument. For example, if your reference number is 20, type

```
pistonread(20)
```

You will then have a file (in your working directory) called `mypistondata.txt`. There will be three columns of data, as follows.

- The leg numbers (1 for North *etc*) are in column 1.
- The compressor numbers are in column 2.
- Your failure numbers are in column 3.

You will need to edit the file as follows.

- Insert a line at the top as follows.


```
leg[] comp[] y[]
```
- Insert END after the last line of data.

The file will then be ready to use as a data file with BRugs.

- Use MCMC to take samples from the posterior distribution of the unknowns in the model.

(5 marks)

- Display your results appropriately.

(4 marks)

- Explain your method and show your BUGS model specification and the commands which you have used.

(4 marks)

- Show how you have checked convergence.

(3 marks)

- Give summaries of the posterior distributions of the model unknowns. In particular, compare the failure rates in the twelve legs using the posterior distribution. What can you conclude?

(4 marks)

3. Fraud

Banks and credit card companies attempt to detect fraud by looking for unusual observations in the withdrawal data for customers. This potentially involves quite complicated models. The model in this question is a somewhat simplified version but the principal is the same.

You will each be supplied with data for five customers. For each of these customers you will be given the total withdrawals from the customer's account for each of twenty weeks. The value for customer j in week i is $y_{i,j}$.

For each customer in each week there is a small probability π that a fraud takes place. We therefore use a mixture model with two components. The component indicator for customer j in week i is $c_{i,j}$.

If $c_{i,j} = 1$ then a fraud against customer j takes place in week i .

If $c_{i,j} = 2$ then no fraud takes place against customer j in week i .

We assume that, given the model parameters, $c_{i,j}$ is independent of $c_{i',j'}$ for $(i,j) \neq (i',j')$. Given π , we have $\Pr(c_{i,j} = 1) = \pi$. Our prior distribution for π is $\text{Beta}(1, 99)$.

If $c_{i,j} = 1$ then $y_{i,j} \sim \text{Ga}(2, 0.0002)$. If $c_{i,j} = 2$ then, given α , β_j , we have $y_{i,j} \sim \text{Ga}(\alpha, \beta_j)$. We assume that $y_{i,j}$ is independent of $y_{i',j'}$ for $(i,j) \neq (i',j')$, given α , β_j and $\beta_{j'}$. Our prior distribution for α is $\text{Ga}(2, 0.5)$.

Let $\beta_j = \alpha/\lambda_j$ and $\lambda_j = \exp(\mu_j)$. Given μ_0 , τ , we have $\mu_j \sim N(\mu_0, \tau^{-1})$ with μ_j independent of $\mu_{j'}$ for $j \neq j'$. Our prior distribution for μ_0 is $\mu_0 \sim N(5.3, 1.4)$. Our prior distribution for τ is $\text{Ga}(3, 4)$.

Unless otherwise stated, the prior distributions are independent.

Data. To obtain your data, first install the R function `fraudread`. This function may be obtained from the Module Web Page, under "Data", or directly from

<http://www.mas.ncl.ac.uk/~nmf16/teaching/mas8303/fraudreadR.txt>

or simply by typing the following into R.

```
fraudread<-function(refno)
{data<-read.table("http://www.mas.ncl.ac.uk/~nmf16/teaching/mas8303/frauddata.txt")
no1<-5*(refno-11)+1
no5<-no1+4
out<-data[,no1:no5]
write.table(format(out),row.names=FALSE,col.names=FALSE,quote=FALSE,file="myfrauddata.txt")
}
```

Then, use the function with your reference number as the argument. For example, if your reference number is 20, type

```
fraudread(20)
```

You will then have a file (in your working directory) called `myfrauddata.txt`. There will be five columns of data, one for each customer, and twenty rows, one for each week. You will need to edit the file as follows.

- Insert a line at the top as follows.

y[,1] y[,2] y[,3] y[,4] y[,5]

- Insert **END** after the last line of data.

The file will then be ready to use as a data file with BRugs.

- Use MCMC to find the posterior means for $p_{i,j} = 2 - c_{i,j}$ and hence find the posterior probabilities of fraud for each customer in each week and identify any cases where fraud is likely to have occurred.

(5 marks)

- Display your results appropriately.

(4 marks)

- Explain your method and show your BUGS model specification and the commands which you have used.

(4 marks)

- Show how you have checked convergence.

(3 marks)

- Give summaries of the posterior distributions of the model parameters.

(4 marks)