MAS8303 Modern Bayesian Inference
Part 2

M. Farrow
School of Mathematics and Statistics
Newcastle University

Semester 1, 2012-13

Chapter 5

Random Effects and Hierarchical Models

5.1 Random Effects

5.1.1 Fixed and random effects

Consider Example 2 of Lecture 1.3.1. The data gave the gains in weight of rats fed on four different diets. The diets differed in terms of the amount of protein ("low" or "high") and the source of the protein ("beef" or "cereal"). The population mean weight gains with each diet are considered to be parameters of the model. If we were to observe very large numbers of rats with each diet then we would gain very precise information about the values of these parameters. In the limit, we would know the values exactly. The differences in population mean weight gains between the diets are regarded as *fixed* but *unknown*.

We can write the four means in the form

$$\begin{aligned}\mu_1 &= \mu - \beta_a - \beta_s + \gamma,\\ \mu_2 &= \mu + \beta_a - \beta_s - \gamma,\\ \mu_3 &= \mu - \beta_a + \beta_s - \gamma,\\ \mu_4 &= \mu + \beta_a + \beta_s + \gamma.\end{aligned}$$

Then $\beta_a,\ \beta_s, \gamma$ are all regarded as *fixed effects*.

Now consider another example. This example comes from Davies and Goldsmith (1972). The experiment concerned testing the strength of Portland cement. The cement was divided into small samples. Each sample was then mixed with water and worked. This process is called "gauging." Each sample was then cast into a cube and allowed to set. The samples were then tested for strength. This is known as "breaking."

Three different people did the gauging and three different people did the breaking. There are thus nine combinations of gauger and breaker. In each combination there were four cubes. The data, in pounds per square inch, are given in Table 5.1.

Let $y_{i,j,k}$ be the i^{th} observation made with Gauger j and breaker k, a realisation of the random variable $Y_{i,j,k}$. Then we might write

$$Y_{i,j,k} \mid \mu_{j,k}, \tau_\varepsilon \sim N(\mu_{j,k},\ \tau_\varepsilon^{-1})$$

where

$$\mu_{j,k} = \mu + \alpha_j + \beta_k + \gamma_{j,k}.$$

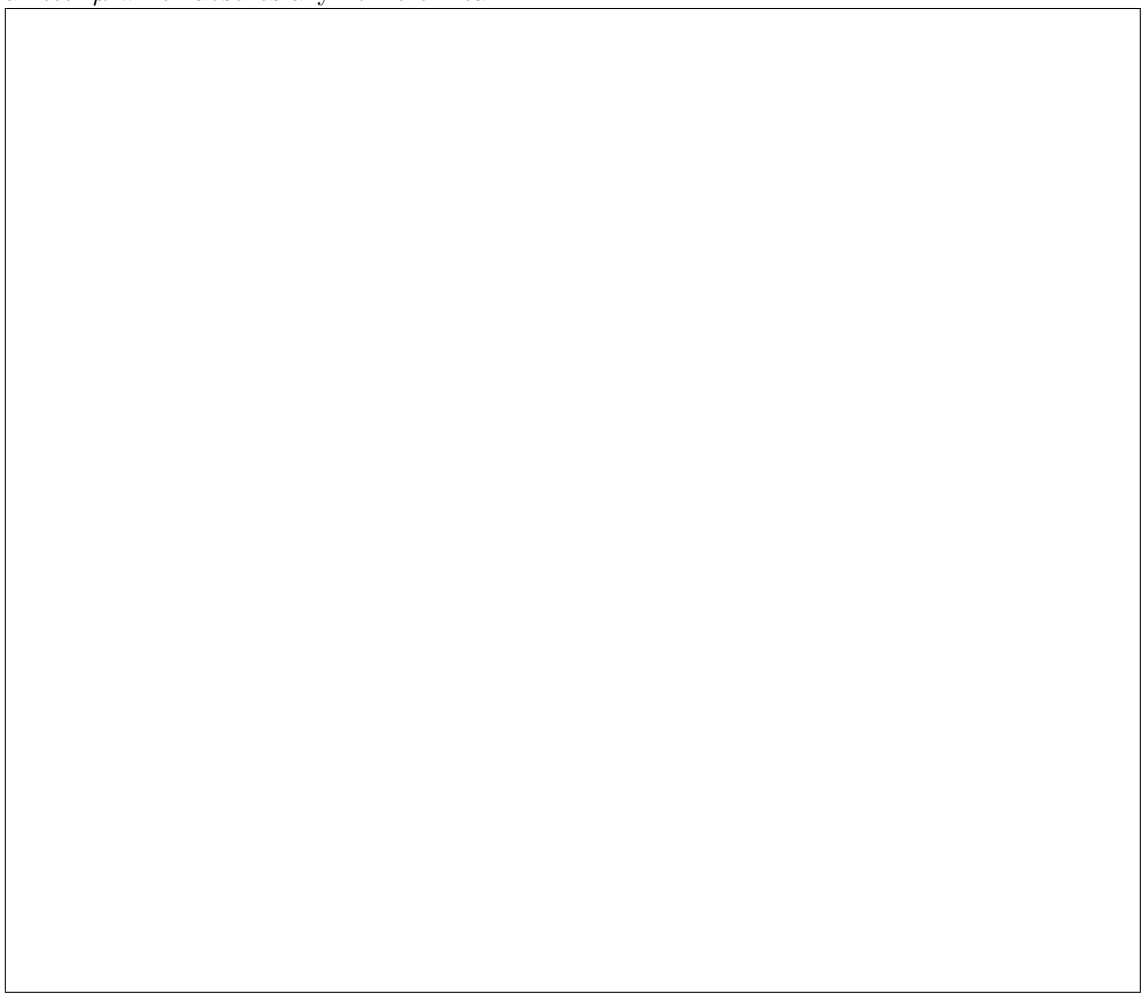
Here, just as in the rats example, we have the main effects of two factors and an interaction effect. The gauger effects are $\alpha_1, \alpha_2, \alpha_3$, the breaker effects are $\beta_1, \beta_2, \beta_3$ and the interaction effects are $\gamma_{1,1}, \ldots, \gamma_{3,3}$. However these are not regarded as *fixed* effects. Instead they are regarded as *random effects*. This is because we are not just interested in the effects of *these* gaugers and

	Breaker 1		Breaker 2		Breaker 3	
Gauger 1	5280	5520	4340	4400	4160	5180
	4760	5800	5020	6200	5320	4600
Gauger 2	4420	5280	5340	4880	4180	4800
	5580	4900	4960	6200	4600	4480
Gauger 3	5360	6160	5720	4760	4460	4930
	5680	5500	5620	5560	4680	5600

Table 5.1: Breaking strengths (pounds per square inch) of cement samples.

breakers but in how much variation there is between gaugers generally and between breakers generally. We regard these gaugers as a sample from the population of gaugers and these breakers as a sample from the population of breakers.

We do not constrain the effects to sum to zero, or fix one of them to be zero. Instead we regard them as samples from a distribution with zero mean. The mean is zero because we include the parameter μ which absorbs any nonzero mean.



Notice two difference between this random effects model and the fixed effects model which we used for the rats example.

1. We suppose that we might observe a new gauger or a new breaker in the future. No matter how many observations we make with the gaugers and breakers in our sample, we will never be able to predict exactly the mean for a new gauger-breaker combination which have not yet observed because this will involve new realisations from the random effects distributions.

2. Suppose that, instead of giving $\tau_\alpha,\ \tau_\beta,\ \tau_\gamma$ prior distributions, we simply chose values for them. Then the model would be very similar to a fixed effects model. The only differences would be point 1, which refers to how we interpret the results in terms of future observations, and the fact that we do not constrain the effects to sum to zero. This latter point would mean that the individual model effects would not be identifiable but the nine means for combinations of gauger and breaker would still be identifiable. However, in fact, we do not choose values for these precisions (i.e. for the variance components) but regard them as unknown and learn about them from the data. This means that we use the data to tell us how similar we can expect future gauger-breaker combinations to be to those which we have already seen.

5.1.2 Evaluation of posterior distribution

Given a model such as the cement-testing example, we can easily use MCMC with data augmentation to sample from the posterior distribution. We regard the random effects as auxiliary variables. I will illustrate the method in terms of the cement example. The auxiliary data are $\alpha_1, \ldots, \alpha_3,\ \beta_1, \ldots, \beta_3, \gamma_{1,1}, \ldots, \gamma_{3,3}$.

A possible MCMC scheme is as follows. Sketching a DAG might help to see how this works.

1 Sample τ_ε : Given values for the fixed effect μ and the random effects we have

$$Y_{i,j,k} - \mu - \alpha_j - \beta_k - \gamma_{j,k} \sim N(0,\ \tau_\varepsilon^{-1})$$

With a gamma prior for τ_ε we get a gamma fcd for τ_ε and it is easy to sample from this.

2 Sample μ : Given values for the error precision τ_ε and the random effects we have

$$Y_{i,j,k} - \alpha_j - \beta_k - \gamma_{j,k} \sim N(\mu,\ \tau_\varepsilon^{-1})$$

With a normal prior for μ we get a normal fcd for μ and it is easy to sample from this.

3 Sample τ_α : Given τ_α we have $\alpha_j \sim N(0,\ \tau_\alpha^{-1})$. So, given values for $\alpha_1, \ldots, \alpha_3$ and a gamma prior for τ_α, the fcd for τ_α is a gamma distribution and it is easy to sample from this.

4 Sample τ_β : Given τ_β we have $\beta_k \sim N(0,\ \tau_\beta^{-1})$. So, given values for $\beta_1, \ldots, \beta_3$ and a gamma prior for τ_β, the fcd for τ_β is a gamma distribution and it is easy to sample from this.

5 Sample τ_γ : Given τ_γ we have $\gamma_{j,k} \sim N(0,\ \tau_\gamma^{-1})$. So, given values for $\gamma_{1,1}, \ldots, \gamma_{3,3}$ and a gamma prior for τ_γ, the fcd for τ_γ is a gamma distribution and it is easy to sample from this.

6 Sample $\alpha_1, \ldots, \alpha_3$: Given values for the fixed effect μ, for τ_α and for the other random effects we have

$$Y_{i,j,k} - \mu - \beta_k - \gamma_{j,k} \sim N(\alpha_j,\ \tau_\varepsilon^{-1})$$

The "prior" for α_j here is the conditional distribution of α_j given τ_α which is $\alpha_j \mid \tau_\alpha \sim N(0,\ \tau_\alpha^{-1})$. The resulting fcd is normal and it is easy to sample from this. The fcd for α_j just involves the data through $y_{1,j,1}, \ldots, y_{4,j,3}$.

7 Sample $\beta_1, \ldots, \beta_3$: Given values for the fixed effect μ, for τ_β and for the other random effects we have

$$Y_{i,j,k} - \mu - \alpha_j - \gamma_{j,k} \sim N(\beta_k,\ \tau_\varepsilon^{-1})$$

The "prior" for β_k here is the conditional distribution of β_k given τ_β which is $\beta_k \mid \tau_\beta \sim N(0,\ \tau_\beta^{-1})$. The resulting fcd is normal and it is easy to sample from this. The fcd for β_k just involves the data through $y_{1,1,k}, \ldots, y_{4,3,k}$.

8 Sample $\gamma_{1,1}, \ldots, \gamma_{3,3}$: Given values for the fixed effect μ, for τ_γ and for the other random effects we have

$$Y_{i,j,k} - \mu - \alpha_j - \beta_k \sim N(\gamma_{j,k},\ \tau_\varepsilon^{-1})$$

The "prior" for $\gamma_{j,k}$ here is the conditional distribution of $\gamma_{j,k}$ given τ_γ which is $\gamma_{j,k} \mid \tau_\gamma \sim N(0,\ \tau_\gamma^{-1})$. The resulting fcd is normal and it is easy to sample from this. The fcd for $\gamma_{j,k}$ just involves the data through $y_{1,j,k}, \ldots, y_{4,j,k}$.

Note that this is by no means the *only* way to evaluate the posterior distribution. In fact this algorithm may be subject to poor mixing. However it is simple to implement.

5.1.3 More general models

I have explained random effects models in terms of a simple example with two factors, each with three levels, and an interaction. Of course we could have much more complicated models with more factors and interactions. The principles remain the same though.

We could also have models which contain non-normal distributions We will see an example of this later.

5.1.4 Mixed models

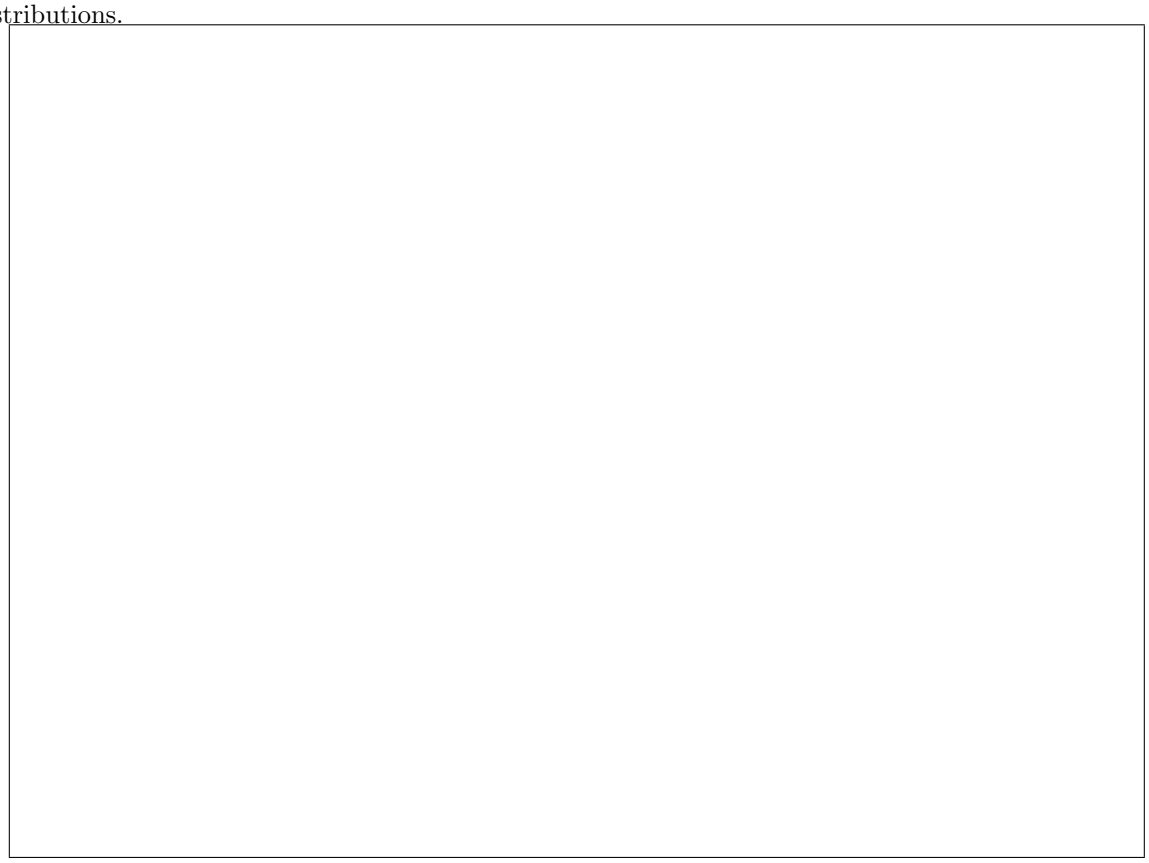
We can also have models in which some effects are fixed and some random. For example, in testing two drugs for the control of high blood pressure, each patient might provide a number of blood pressure measurements while being treated with each of the drugs (e.g. in a crossover trial). We would normally regard the drug effects as fixed but the patient effects, and any patient-drug interaction, as random effects. At step 2 in the algorithm above we would sample all of the fixed effects.

A model containing both fixed and random effects is called a *mixed model*. A mixed model where the random effects distributions are normal and the error distribution is normal and the means are linear functions of the effects is a *linear mixed model*. We could also have, for example, a *generalised linear mixed model* in which the error distribution might be, for example, Poisson or binomial, the means are related to linear predictors by a link function and the linear predictors are linear functions of fixed and random effects.

5.2 Hierarchical Models

5.2.1 Hierarchical structures

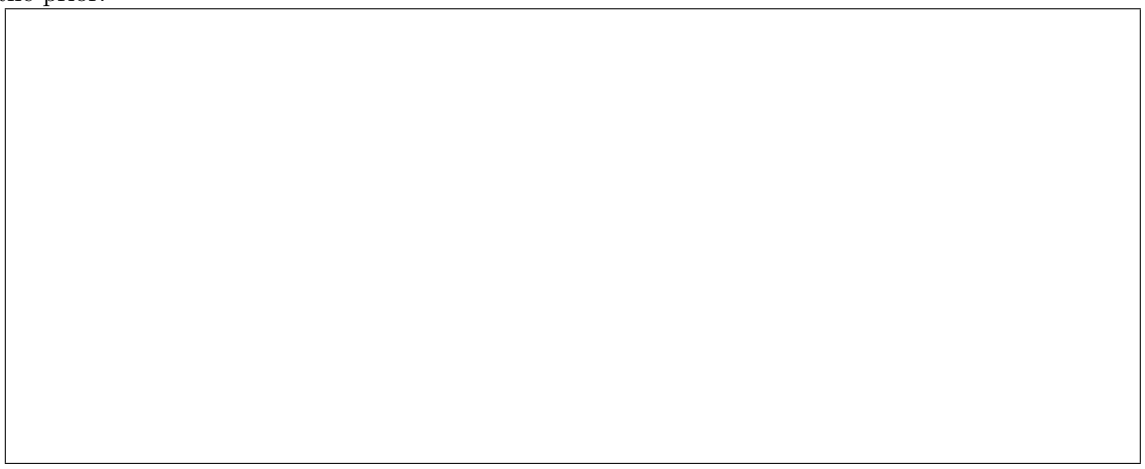
We are going to look at models and priors where we have two or more "levels" of conditional distributions.

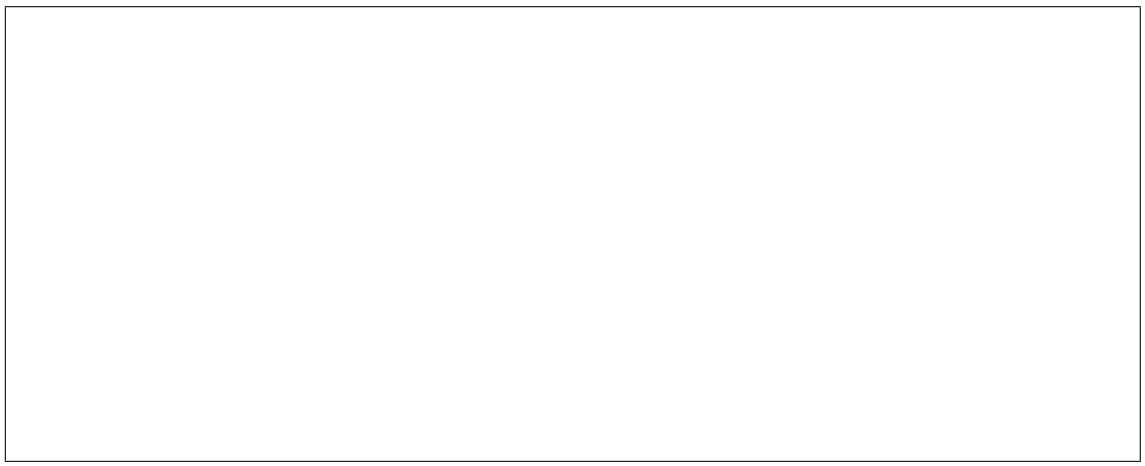


Notice that

- there are several *levels* in this structure and
- the structure is *nested* or *hierarchical*.

Here $Y_{i,j,k}$ is the i^{th} obervation within sub-group j of group k. Two observations within the same subgroup are more strongly correlated with each other than two observations within different subgroups. Two observations in different subgroups within the same group are more strongly correlated than two observations in different groups. The group means are themselves correlated in the prior.





Therefore, in the example above:

When we looked at mixture models we said that there were two different reasons why we might use a mixture model, depending on whether or not we supposed that there really were subpopulations. The distinction between a "hierarchical prior" and a "multilevel model" or "hierarchical model" is of a similar nature.

In some cases we are really only interested in one level of unit, such as the sub-groups indexed j, k above, and other levels, e.g. the groups indexed k above, are introduced simply to give a covariance structure to the prior. In this case we would regard this as a "hierarchical prior."

In other cases the levels might have "physical" interpretations. For example, $Y_{i,j,k}$ could be the score obtained in a test by pupil i in school j of education authority k. Then the values of the education-authority effects b_k and the school effect $a_{j,k}$ might be of interest in themselves.

5.2.2 Hierarchical priors and "borrowing strength"

We have seen hierarchical priors already. We might make observations on members of a number of groups, e.g. weight gains of rats given different diets. So $Y_{i,j}$ is the i^{th} observation in Group j. Then, given μ_j and σ_Y^2, we have

$$Y_{i,j} \mid \mu_j, \sigma_Y^2 \sim N(\mu_j,\ \sigma_Y^2).$$

We need a prior for $\mu_1, \ldots, \mu_J$ but, if we are measuring the same thing in these groups, e.g. weight gain, then it seems reasonable that these means will be positively correlated in our prior. So, given μ_0 and σ_μ^2, we write

$$\mu_j \mid \mu_0, \sigma_\mu^2 \sim N(\mu_0,\ \sigma_\mu^2).$$

Then we give a prior to μ_0 with

$$\mu_0 \sim N(m_0,\ v_0).$$

Thus, in our prior, given σ_μ^2, each of μ_j and $\mu_{j'}$ has mean m_0 and variance $v_0 + \sigma_\mu^2$ but they also have covariance v_0 when $j \neq j'$.

Typically we would also give a prior to σ_Y^2. We might simply choose a value for σ_μ^2 but we might choose to give it a prior as well. Choosing to give σ_μ^2 a distribution has two effects.

- Because of the covariance structure, the posterior means of $\mu_1, \ldots, \mu_J$ will tend to be closer to their common overall mean than the sample means of the data are. This is a similar effect to the posterior mean being closer to the prior mean than the sample mean is when we have a single sample. This effect is called *shrinkage*. The degree of shrinkage depends, in part, on the relative sizes of the variances. If we choose the value of σ_μ^2 then we are (almost) choosing the degree of shrinkage. (The degree of shrinkage also depends on σ_Y^2 and we allow this to be unknown). If we allow σ_μ^2 to be unknown and give it a prior then we give the data more influence over the degree of shrinkage.

- If we expect to observe other related groups in the future then, learning about σ_μ^2 from the data allows us to change our minds about how close we expect these future group means to be to the means for groups which we have seen. We would have to believe that, in some sense, the future groups would be drawn from "the same population." (Usually this means that we would believe that groups were exchangeable).

Borrowing strength

The shrinkage effect noted above has an important benefit. Consider the following (very simplified) example.

We are interested in the rates of a disease in different areas of the country. In area j the population at risk is n_j. (In reality we would usually also take into account age groups etc.). Our model says that the number of cases in area j is Y_j which, conditional on a rate parameter θ_j, has a Poisson distribution

$$Y_j \mid \theta_j \sim \text{Po}(n_j\theta_j).$$

Now the mean of this distribution $n_j\theta_j$ might be a small number (e.g. 10) so that the standard deviation of the Poisson distribution is quite large compared to its mean. If we try to make inferences about the individual rates λ_j treating them independently then there is little information in the data about each. On the other hand, if we pool all of the data and assume that $\lambda_1 = \lambda_2 = \cdots = \lambda_J$, then we lose any possibility of detecting unusual rates in particular places. Instead we compromise and use a hierarchical prior. Given $a,\ b$ we give θ_j a gamma distribution

$$\theta_j \mid a, b \sim \text{Ga}(a,\ b).$$

we can then give a prior to $a,\ b$.

In this way the posterior distribution for θ_j uses information not only from Y_j but also from the observations in other areas. For example, the posterior means in cases with unusually large Y values will be "shrunk" somewhat towards the overall average. This is called "borrowing strength."

In *spatial statistics*, more complicated models are used in which the parameter in an area is more strongly correlated with the parameters in neighbouring areas.

5.2.3 Data augmentation and MCMC

Clearly, just as in Lecture 5.1 on random effects, hierarchical structures such as those discussed here give rise to a straightforward application of MCMC with data augmentation, regarding the different levels of random effects as auxiliary data. So, for example, in the first, normal, example above we could regard $\{A_{j,k}\}$ and $\{B_k\}$ as auxiliary data. For fixed values of these the likelihood is simple and sampling values for the parameters is simple. When the values of the parameters and one set of auxiliary variables is fixed, it is simple to sample values for the other set of auxiliary variables.

5.2.4 Multilevel models

As noted above, in some cases we are interested in the random effects themselves, rather than either just the population parameters ($\mu,\ \sigma_B^2,\ \sigma_A^2,\ \sigma_Y^2$) or just the first-level parameters ($\{A_{j,k}\}$).

5.3 Repeated Measures

5.3.1 Introduction

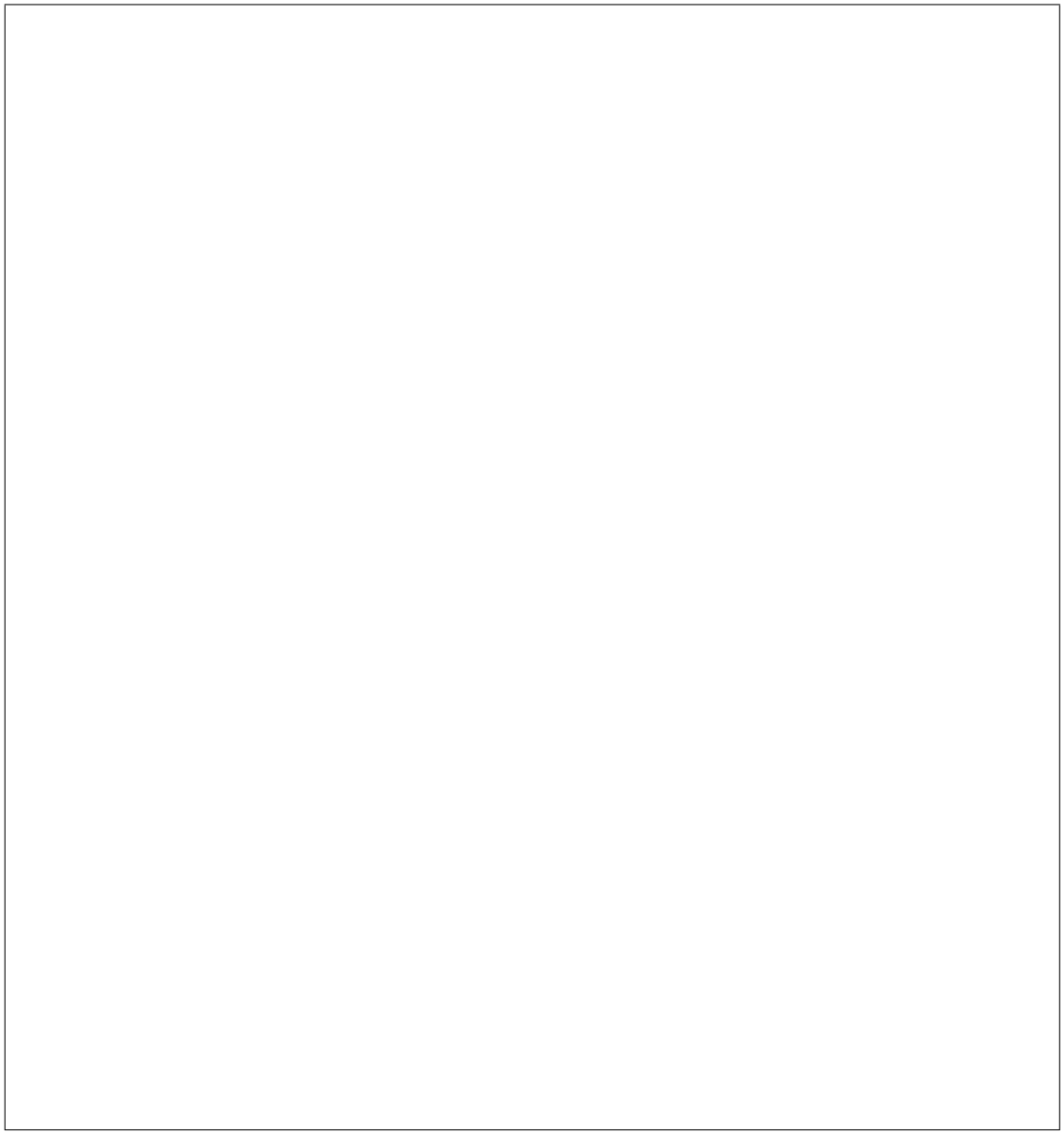
Among the types of problem where random effects are used are *repeated measures* models, where several observations are made on the same individual, and *longitudinal data*, where we are particularly interested in how repeated measurements taken on individuals change over time.

5.3.2 Example: Repeated measurements in two groups

A drug for lowering blood pressure is tested. A sample of patients with high blood pressure is divided randomly into two groups. Patients in Group 1 are given the drug. Patients in Group 2 are given a placebo. After a suitable period a sequence of five blood pressure measurements, at intervals, is made on each patient. (In this example we assume that there is no time trend).

This is really just a mixed-effects model. There is a fixed treatment effect and there are random patient effects.

Let $Y_{i,g,t}$ be the observation on patient i of group g at time t for $i = 1, \ldots, n_g,\ g = 1, 2,\ t = 1, \ldots 5$.



```
model bloodpressure

{for (i in 1:N)
   {for (t in 1:5)
      {y[i,t]~dnorm(p_[i],tau.y)
       }
    p[i]~dnorm(mu[group[i]],tau.p)
    }

 for (g in 1:2)
   {mu[g]~dnorm(mu0,0.001)
    }

 mu0~dnorm(150,0.0005)
 tau.y~dgamma(2,100)
 tau.p~dgamma(2,200)
 }
```

Figure 5.1: BUGS code for blood pressure example.

Here the variance of $\mu_1 - \mu_2$ is $\text{var}(2\delta) = 4\tilde{v}_g$. In the first form of prior we had $\text{var}(\mu_1 - \mu_2) = 2v_g$. Hence, if we set $\tilde{v}_g = v_g/2$ we get the same variances and covariances. The introduction of d_0 allows us to have a nonzero prior mean for the treatment effect.

Figure 5.1 shows some suitable BUGS code. It is assumed that the data file contains six columns. The first column contains the number of the group to which the patient belongs. The remaining five columns contain the five blood pressure measurements, in order. (With BRugs it would be necessary to load the overall sample size `N` from another file). The first form of the prior is used.

5.3.3 Autocorrelation

In the example in 5.3.2 we have made no use of the time-ordering of the observations. The five observation on a particular patient are treated as exchangeable. We might believe that neighbouring observations are likely to be more strongly correlated than observations further apart. We could allow for this by allowing autocorrelation of the observations. This could be done, for example, using an autoregressive process or a moving average process. For illustration we will use a first-order moving average process.

The model as it stands can be written

$$Y_{i,g,t} = P_{i,g} + \varepsilon_{i,g,t}$$

where $\varepsilon_{i,g,t} \sim N(0,\ \sigma^2_Y)$.

Let us replace this with

$$Y_{i,g,t} = P_{i,g} + \varepsilon_{i,g,t} + \eta_{i,g,t} + \eta_{i,g,t+1}$$

where $\varepsilon_{i,g,t} \sim N(0, \sigma^2_\varepsilon)$ and $\eta_{i,g,t} \sim N(0,\ \sigma^2_\eta)$. The conditional variance of $Y_{i,g,t}$ is now $\sigma^2_\varepsilon + 2\sigma^2_\eta$ so we would want this variance to correspond to the old σ^2_Y. The conditional covariance between $Y_{i,g,t}$ and $Y_{i,g,t'}$ is now zero for $|t - t'| > 1$ but σ^2_η for $|t - t'| = 1$.

Figure 5.2 shows modified BUGS code. Note that we have to allow for an extra $\eta_{i,g,6}$.

5.3.4 Example: growth curves

Growth curves are a special kind of longitudinal-data problem. We are often interested in how, for example, individual children or young animals grow over time.

Here is a simple example taken from Gelfand *et al.* (1990). It can also be found as an example on the BUGS Website.

```
model bloodpressure

{for (i in 1:N)
   {for (t in 1:5)
      {y[i,t]~dnorm(ymean[i,t],tau.eps)
       ymean[i,t]<-p[i]+eta[i,t]+eta[i,t+1]
       }
    for (t in 1:6)
      {eta[i,t]~dnorm(0,tau.eta)
       }
    p[i]~dnorm(mu[group[i]],tau.p)
    }

 for (g in 1:2)
   {mu[g]~dnorm(mu0,0.001)
    }

 mu0~dnorm(150,0.0005)
 tau.eps~dgamma(1,30)
 tau.eta~dgamma(1,10)
 tau.p~dgamma(2,200)
 }
```

Figure 5.2: BUGS code for blood pressure example with moving average errors.

The weights of thirty young rats are measured at weekly intervals for five weeks. A straight-line model is used to relate weight to time. (We might well want to consider a more complicated form of curve and possibly allow autocorrelation of deviations from the curve but, for this example, we will stick to a straight line with independent "errors"). However the intercept and gradient of the line are allowed to vary as random effects between rats.

The five times, in days, at which the weights are measured are $t_1 = 8,\ t_2 = 15,\ t_3 = 22,\ t_4 = 29,\ t_5 = 36$. The weight of rat i on day t_j is

$$Y_{i,j} \mid \alpha_i, \beta_i, \tau_Y \sim N(\alpha_i + \beta_i[t_j - 22], \tau_Y^{-1}).$$

Now we need a model for how α_i, β_i vary between rats. We could simply write

$$\begin{aligned} \alpha_i \mid \mu_\alpha, \tau_\alpha &\sim N(\mu_\alpha,\ \tau_\alpha^{-1}) \\ \beta_i \mid \mu_\beta, \tau_\beta &\sim N(\mu_\beta,\ \tau_\beta^{-1}) \end{aligned} \tag{5.1}$$

with α_i, β_i independent given the parameters. However it might be more realistic to allow them to have a nonzero correlation. One way to do this (though not the way that it is done on the BUGS Website) is to specify the conditional distribution of β_i given α_i. So, instead of (5.1) we write

$$\beta_i \mid \mu_\beta, \tau_\beta, \alpha_i, \gamma \sim N(\mu_\beta + \gamma[\alpha_i - \mu_\alpha],\ \tau_\beta^{-1}).$$

Finally we give prior distributions to the model parameters. The priors given here are based (loosely) on those used in the example on the BUGS Website. They are meant to be "noninformative."

$$\begin{aligned} \mu_\alpha &\sim N(0,\ 10000) \\ \mu_\beta &\sim N(0,\ 10000) \\ \gamma &\sim N(0,\ 4000) \\ \tau_Y &\sim \text{Ga}(0.001,\ 0.001) \\ \tau_\alpha &\sim \text{Ga}(0.001,\ 0.001) \\ \tau_\beta &\sim \text{Ga}(0.001,\ 0.001) \end{aligned}$$

```
model rats

{for (i in 1:N)
   {for j in 1:5)
      {mean[i,j]<-alpha[i]+beta[i]*(t[j]-22)
       y[i,j] ~ dnorm(mean[i,j],tau.y)
       }
    alpha[i] ~ dnorm(mu.alpha,tau.alpha)
    betamean[i]<-mu.beta+gamma*(alpha[i]-mu.alpha)
    beta[i] ~ dnorm(betamean[i],tau.beta)
    }

 mu.alpha ~ dnorm(0,0.0001)
 mu.beta ~ dnorm(0,0.0001)
 gamma ~ dnorm(0,0.00025)
 tau.y ~ dgamma(0.001,0.001)
 tau.alpha ~ dgamma(0.001,0.001)
 tau.beta ~ dgamma(0.001,0.001)
 }
```

Figure 5.3: BUGS model specification for rats growth curves example.

Figure 5.3 shows suitable BUGS code.

Hospital	n_i	r_i	Hospital	n_i	r_i	Hospital	n_i	r_i
1	47	0	5	211	8	9	207	14
2	148	18	6	196	13	10	97	8
3	119	8	7	148	9	11	256	29
4	810	46	8	215	31	12	360	25

Table 5.2: Mortality in twelve hospitals performing cardiac surgery on babies. n_i : number of operations at hospital i. r_i : number of deaths at hospital i.

5.4 Practical 5

5.4.1 Hospital ranking

This example is taken from the BUGS Website. It concerns the mortality rates in twelve hospitals performing cardiac surgery in babies. The data are shown in table 5.2.

Crude methods of comparing hospitals might be misleading. For example, the variance of the observed proportions of deaths is large if the number of operations is smaller. Therefore a small hospital could appear to have a very bad rate simply because of a small number of cases. Using a random-effects model helps to smooth out such effects.

We suppose that, associated with hospital i there is a rate p_i which, if it were known, would be the probability of death at that hospital. We suppose that the number of deaths r_i out of n_i operations at hospital i has a binomial distribution

$$r_i \sim \text{Bin}(n_i,\ p_i).$$

Then we write

$$b_i = \log\left(\frac{p_i}{1-p_i}\right)$$

and

$$b_i \mid \mu, \tau \sim N(\mu,\ \tau).$$

We then gives priors to the parameters. These are the priors used on the BUGS Website. They are so-called "noninformative" priors.

$$\begin{aligned} \mu &\sim N(0,\ 10^6) \\ \tau &\sim \text{Ga}(0.001,\ 0.001) \end{aligned}$$

1. Type the following model specification into a file called `hospitalbug.txt` .

```
model hospital

{for (i in 1:N)
   {r[i]~dbin(p[i],n[i])
    logit(p[i])<-b[i]
    b[i]~dnorm(mu,tau)
    }

 mu~dnorm(0.0,1.0E-6)
 tau~dgamma(0.001,0.001)
 }
```

   Note that `1.0E-6` means 1.0×10^{-6}.

2. Type the data into a file called `hospitaldata.txt` as follows.

```
list(N=12, n=c(47,148,119,810,211,196,148,215,207,97,256,360),
           r=c(0,18,8,46,8,13,9,31,14,8,29,24))
```

3. Use BRugs to evaluate the posterior distribution. Monitor $b_1, \ldots, b_{12}$ and compare the posterior 95% intervals for these. Does any hospital stand out from the rest?

   You will need to set some initial values. For example, to run two chains, create one file called

   `hospitalinits1.txt`

   containing the following

   ```
   list(mu=-2.0, tau=2.0)
   ```

   and another file called

   `hospitalinits2.txt`

   containing the following.

   ```
   list(mu=-2.0, tau=20.0)
   ```

   You would then need to issue commands as follows.

   ```
   modelCheck("hospitalbug.txt")
   modelData("hospitaldata.txt")
   modelCompile(2)
   modelInits("hospitalinits1.txt")
   modelInits("hospitalinits2.txt")
   modelGenInits()
   ```

   You would then be ready to start updating (with or without setting a monitor). The final `modelGenInits()` is necessary because our initial value files do not specify initial values for *all* of the unknowns.

5.4.2 Rat growth

This is the "rats" example of section 5.3.4. The BUGS code is available on the Web page as `ratsbug.txt` and there are two data files, also on the Web page, called `ratsxdata.txt` and `ratsydata.txt` . Because there are two data files, you will need to start like this.

```
modelCheck("ratsbug.txt")
modelData("ratsxdata.txt")
modelData("ratsydata.txt")
```

Use BRugs to evaluate the posterior distribution. Monitor μ_α, μ_β, γ, τ_y, τ_α and τ_β. You could also monitor the regression coefficients of individual rats, α_i, β_i, if you wish.

You will need to supply initial values for some of the unknowns. I suggest that you use two chains and initialise them as follows.

```
modelCheck("ratsbug.txt")
modelData("ratsxdata.txt")
modelData("ratsydata.txt")
modelCompile(2)
modelInits("ratsinits1.txt")
modelInits("ratsinits2.txt")
modelGenInits()
```

Here is a suggestion for the contents of `ratsinits1.txt`

```
list(tau.y=1.0E-5,tau.alpha=1.0,tau.beta=10.0)
```

and here is a suggestion for the contents of `ratsinits2.txt`

```
list(tau.y=5.0E-6,tau.alpha=10.0,tau.beta=100.0)
```

Set the sample monitors and do, say, 5000 updates. Then look at the results using `samplesHistory` . You might be surprised at how poor the convergence is in this example. This is probably because the parameters are poorly identified.

In fact things behave much better if we assume that α_i and β_i are conditionally independent given $\mu_\alpha, \mu_\beta, \tau_\alpha, \tau_\beta$. Try this. Replace the lines

```
betamean[i]<-mu.beta+gamma*(alpha[i]-mu.alpha)
beta[i] ~ dnorm(betamean[i],tau.beta)
```

with the single line

```
beta[i] ~ dnorm(mu.beta,tau.beta)
```

in the model file and try fitting the model again.

Further investigation shows that the posteriors for τ_Y, τ_α and τ_β are sensitive to the choice of priors for these parameters suggesting that the parameters are not well identified. Nevertheless, in this case, the "noninformative" priors seem to give sensible results.

5.4.3 Vertebral fractures in older women.

Here is one for you to do yourselves.

The data come from Cooper *et al.* (1991). Osteoporosis is a problem for many post-menopausal women. It can lead to bone fractures. Women were screened for evidence of vertebral fractures according to a certain criterion. A subset of the data were as follows.

i	Age group	Total number	Number with fracture
1	50-54	17	1
2	55-59	282	12
3	60-64	244	17
4	65-69	218	23
5	70-74	120	9
6	75-79	105	11
7	80-	18	5

Let the lower age limit of age-group i be x_i. Let the number of women screened in this group be n_i and let the number classified as having vertebral fractures be y_i. Then we assume

$$y_i \mid n_i, p_i \sim \text{Bin}(n_i,\ p_i)$$

with

$$\log\left(\frac{p_i}{1-p_i}\right) = \eta_i = \beta_0 + \beta_1 x_i + \delta_i.$$

Here β_0 and β_1 are parameters about which we wish to learn. We adopt the following independent prior distributions.

$$\begin{aligned} \beta_0 &\sim N(-3,\ 5), \\ \beta_1 &\sim N(0,\ 1). \end{aligned}$$

Because the relationship between age and logit of fracture rate might not really be a straight line we allow some deviation by adding a random variable δ_i with

$$\delta_i \sim N(0, 0.001).$$

Use BRugs to evaluate the posterior distribution of $\beta_0,\ \beta_1,\ \eta_1, \ldots, \eta_7$.

Convergence and mixing are poor. You will need a long burn-in. It might help if you use two initial value files such as the following.

```
list(beta0=-4.5, beta1=0.01)

list(beta0=-5.5, beta1=0.06)
```

Try also monitoring $p_1, \ldots, p_7$. You can do this using

```
samplesSet("p")
```

The results are quite interesting.

References

Azzalini, A. and Bowman, A.W., 1990. A look at some data on the Old Faithful geyser. *Journal of the Royal Statistical Society, Series C*, **39**, 357-366.

Breslow, N.E. and Clayton, D.G., 1993. Approximate inference in generalised linear mixed models. *Journal of the American Statistical Association*, **88**, 9-25.

Cooper, C., Shah, S., Hand, D.J., Adams, J., Compston, J., Davie, M. and Woolf,A., 1991. Screening for vertebral osteoporosis using individual risk factors. *Osteoporosis International*, **2**, 48-53.

Cowburn, G.J., 2003. *Bayesian mixture modelling with application to road traffic flow.* Ph.D. thesis, University of Sunderland.

Cowburn, G.J. and Farrow, M., 2007. Mixtures and diagnostic plots in modelling road traffic vehicle headways. Statistical Modelling, 7, 73-89.

Davies, O.L. and Goldsmith, P.L. (eds), 1972. *Statistical Methods in Research and Production*, 4th edition. Edinburgh: Oliver and Boyd.

Freeman, D.H., 1987. *Applied Categorical Data Analysis.* New York: Marcel Dekker.

Gelfand, A.E., Hills, S.E., Racine-Poon, A. and Smith, A.F.M., 1990. Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, **85**, 972-85.

Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J. and Ostrowski, E., 1994. *A Handbook of Small Data Sets.* Chapman and Hall, London.

Krzanowski, W.J., 1988. *Principles of Multivariate Analysis.* Oxford: Oxford University Press.

Lunn, D.J., Thomas, A., Best, N.G., and Spiegelhalter, D.,J., 2000. WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, **10**, 325-337.

Open University, 1983. *MDST242 Statistics in Society, Unit C3: Is my child normal?.* Milton Keynes: The Open University.

Mosteller, F. and Tukey, J.W., 1977. *Data analysis and regression.* Reading, Massachusetts: Addison-Wesley.

Pearl,J., Geiger,D. and Verma,T., 1990, The logic of influence diagrams, In *Influence Diagrams, Belief Nets and Decision Analysis* (R.M.Oliver, J.Q.Smith eds.) New York: Wiley.

Phillips, D.P., 1978. Airplane accident fatalities increase just after newspaper stories about murder and suicide. *Science*, **201**, 748-750.

Scott, S.L., 2002. Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association*, **97**, 337-351.

Snedecor, G.W. and Cochran, G.C., 1967. *Statistical Methods*, 6th edition. Iowa State University Press, Ames, Iowa.

Sokal, R.R. and Rohlf, F.J., 1981. *Biometry*, 2nd edition. San Francisco: W.H. Freeman.

Spiegelhalter, D.J., Thomas, A., Best, N.G. and Gilks, W.R., 1995. *BUGS: Bayesian inference Using Gibbs Sampling, Version 0.50.* MRC Biostatistics Unit, Cambridge.

Thall, P.F. and Vail, S.C., 1990. Some covariance models for longitudinal count data with overdispersion. Biometrics, **46**, 657-71.

Till, R., 1974. *Statistical Methods for the Earth Scientist.* London: MacMillan.