12 Hierarchical models

Motivating example: consider measurements from individuals i within groups j. Assume $y_{i,j} \sim \mathrm{N}(\mu_{i,j}, \sigma^2)$ Two extreme models could be:

$$\mu_{i,j} = \mu$$

and

$$\mu_{i,j} = \mu_j$$

In the first, all observations are modeled via common mean parameter. Therefore, all data are pooled together. In the second, the mean is group specific, so that each group is estimated separately. If there are important group differences, pooling is not good. But if there are very limited data in some groups, the corresponding estimate becomes very uncertain. There is some information that could be 'borrowed' from other groups, with the hierarchical model:

$$\mu_{i,j} = \mu + e_j \quad , e_j \sim \mathrm{N}(0, \sigma_e^2)$$

Each group mean is allowed to variate around the global mean μ, and these variations between groups are modeled as $\mathrm{N}(0, \sigma_e^2)$ with σ_e^2 to be estimated (variance component).

12.1 Applications of hierarchical models

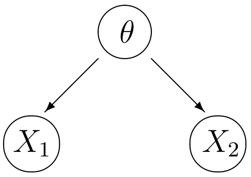
Hierarchical models are a powerful tool for making *synthesis of several sources of information.* Some examples in different contexts:

• Categorical data
(several 2 × 2 tables or a large K × K -table, sparse data, empty categories)
• Spatial analysis
(several geographical areas as categories)
• Temporal analysis
(temporal dependency, dynamics)
• Hierarchical population structures
(individual, group, population, random effects and overdispersed data)
• Latent/hidden structures
(observed disease cases of type A vs. all cases of type A vs. all cases of all types)
• Meta-analysis
(several results from literature, selection bias, publication bias)
• Model averaging
(data model, model parameters, 'model of models')

Conditionally, the distributions are still tractable and can be presented in a DAG. This structure is also exploited in MCMC algorithms, where e.g. full conditionals may be obtained for Gibbs.

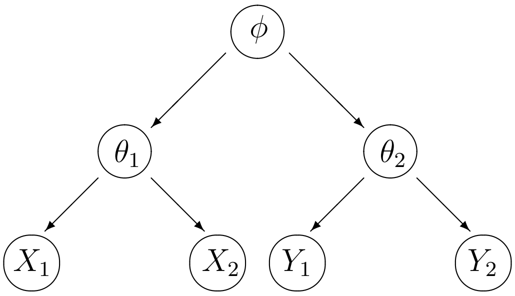
12.2 Hierarchical structures and exchangeability

As before, consider a model graphically as Directed Acyclic Graph (DAG). For example, with only two variables (X_1, X_2):



This is not really a hierarchical model yet, because it only has two parts: the prior $\pi(\theta)$, and the conditional probability of observable data $\pi(X_1, X_2 \mid \theta) = \pi(X_1 \mid \theta)\pi(X_2 \mid \theta)$. If both X_1 and X_2, or either one of them is observed, we can compute the posterior $\pi(\theta \mid X_i)$, or $\pi(\theta \mid X_1, X_2)$. And based on the posterior, we can compute posterior predictive distribution for a next variable X_3.

Recall de Finetti theorem: before observing, the variables $X_1, X_2, \ldots$ were *exchangeable*. From exchangeability it follows, that our probabilities are necessarily such that they can be constructed as a prior distribution $\pi(\theta)$ *and* the conditional distribution $\pi(X_i \mid \theta)$ so that the variables X_i are conditionally independent of each other, given θ. But exchangeability depends on our background information. If we know that the variables X_i represent e.g. measurements of animals from the same farm, then - if nothing more is known - they would be exchangeable for us. **But** if we know that the variables are measurements of animals from two different farms so that $X_1, \ldots, X_k$ are from farm A, and $Y_1, \ldots, Y_r$ are from farm B, the whole set of observations would no longer be exchangeable to us because we know that there can be important differences between farms, which could lead to very different outcomes for Y-variables compared to X. *Knowing* which measurement came from which farm makes a difference. Variables X would still be exchangeable *within* farm A, and variables Y would still be exchangeable *within* farm B. (This is called partial exchangeability). Therefore, (by de Finetti), each set of variables would be conditionally independent, given a *farm specific parameter* θ_i, and we would have a prior $\pi(\theta_i)$. Moreover, the farm specific parameters would be exchangeable - if nothing more is known about the farms: these farms are just some farms from a larger population of farms. Therefore, (by de Finetti), parameters θ_i would be conditionally independent, given some higher level parameter so that $\pi(\theta_1, \theta_2 \mid \phi) = \pi(\theta_1 \mid \phi)\pi(\theta_2 \mid \phi)$. This could be drawn as a DAG:



This is a hierarchical model, that is mathematically written as:

$\pi(\phi)$	hyper prior
$\pi(\theta_i \mid \phi)$	prior
$\pi(X_{ij} \mid \theta_i)$	data generating model

Inference: the model could be used for estimating farm specific quantities (θ_i), but it could also be used for estimating higher level quantities (ϕ) describing the larger population of farms, based on the observed results from several farms.

The posterior density would be multidimensional:

$$\pi(\phi, \theta_1, \theta_2 \mid X_{11}, \ldots, X_{1,J_1}, X_{21}, \ldots, X_{2,J_2})$$

$$\propto \pi(X_{11}, \ldots, X_{1,J_1} \mid \theta_1)\pi(X_{21}, \ldots, X_{2,J_2} \mid \theta_2)\pi(\theta_1 \mid \phi)\pi(\theta_2 \mid \phi)\pi(\phi)$$

Predictions: the model could be used for predicting a new variable within a single farm, but it could also be used for predicting a completely new farm, based on the observed results from several farms.

Gelman et al: Bayesian data analysis:

> *In practice, ignorance implies exchangeability.*
> *Generally, the less we know about a problem,*
> *the more confidently we can make claims of exchangeability.*
> *(This is not, we hasten to add, a good reason to limit our knowledge of*
> *a problem before embarking on statistical analysis!)*

Note: if we know the measurements were from different farms but if we still don't know which of them came from which farm, the measurements would still be exchangeable in our prior state of knowledge. It would also be possible to define (for each measurement) an underlying hidden variable that is an indicator of the farm. This would lead to a cluster model where the measurements would be grouped, or classified, probabilistically to different groups.

Finally: a hierarchical model needs hierarchical data! I.e. groups within groups.

12.3 Example: blood pressure

Example 9.1, BMUW, p. 308. The variability in measurements is due to between-subject variability and within-subject variability. The measurements of blood pressure were taken twice from each individual:

```
list(n=20, K=2, y=structure(.Data=c(108, 98, 91, 94, 93, 96, 104,
99, 99, 97, 95, 98, 93, 97, 99, 96, 90, 100, 92, 95, 101, 89, 97,
97, 97, 100, 96, 95, 106, 100, 100, 98, 90, 99, 88, 98, 92, 92, 100,
101), .Dim = c(20, 2) ) )
```

A simple model would be

$$Y_{ij} = \mu + a_i + \epsilon_{ij}$$

with $\epsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$ and $a_i \sim \mathrm{N}(0, \sigma_a^2)$. Or simply:

$$Y_{ij} \sim \mathrm{N}(\mu_{ij}, \sigma^2)$$

with $\mu_{ij} = \mu + a_i$ and $a_i \sim \mathrm{N}(0, \sigma_a^2)$. The 'random' effect a_i is set for each individual and it has an effect for all measurements from this individual. The random effects variance describes how much the individuals differ. (between subject variability). Often, gamma-priors are used for $1/\sigma^2 = \tau$, but e.g. Gelman has proposed to use uniform priors on σ instead. In some situations, the seemingly flat gamma-density might not be reasonably uninformative. It is only reasonably flat when we are away from zero, but has its peak at zero. If the likelihood happens to be reasonably high near zero, problems with gamma priors can emerge.

The model is equivalent to

$$y_i \mid \mu, \sigma^2, \sigma_a^2 \sim \mathrm{N}\left(\begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma^2 + \sigma_a^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma^2 + \sigma_a^2 \end{bmatrix} \right)$$

Here we have the following results, conditionally on $\mu, \sigma_a^2, \sigma^2$:

$$\begin{aligned} \mathrm{E}(Y_{ij}) &= \mu \\ \mathrm{Var}(Y_{ij}) &= \mathrm{Var}(a_i) + \mathrm{Var}(\epsilon_{ij}) = \sigma_a^2 + \sigma^2 \\ \mathrm{Cov}(Y_{i1}, Y_{i2}) &= \mathrm{Cov}(\mu + a_i + \epsilon_{i1}, \mu + a_i + \epsilon_{i2}) \\ &= \mathrm{Var}(a_i) + \mathrm{Cov}(a_i, \epsilon_{i2}) + \mathrm{Cov}(\epsilon_{i1}, a_i) + \mathrm{Cov}(\epsilon_{i1}, \epsilon_{i2}) = \sigma_a^2 \end{aligned}$$

The result for covariance follows from recalling that $\mathrm{Cov}(\sum X_i, \sum Y_i) = \sum\sum \mathrm{Cov}(X_i, Y_j)$ and that $\mathrm{E}(\epsilon_{ij}) = \mathrm{E}(a_i) = 0$, and that $\epsilon_{i1} \perp \epsilon_{i2}$, $\epsilon_{ij} \perp a_i$, (conditional independence).

Therefore, the total variance of response Y_{ij} is $\sigma_a^2 + \sigma^2$ and the covariance of two measurements from the same subject i is the between-subject variability σ_a^2 (i.e. random effects variance).

The within-subject correlation is thus $\sigma_a^2/(\sigma_a^2 + \sigma^2)$.

```
 model{
 for(i in 1:n){
 for(j in 1:K){
   y[i,j] ~ dnorm(mu[i,j],tau)
   mu[i,j] <- m + a[i]
 }
 a[i] ~ dnorm(0,tau.a)
 }
 m ~ dnorm(0,0.001)
 tau ~ dgamma(0.001,0.001)
 tau.a ~ dgamma(0.001,0.001)
}
```

Compute the posterior distribution in WinBUGS. Try also the 2-dimensional model using `dmnorm`.

Computational note: the above model could also be coded as

```
 mu[i,j] ~  dnorm(m,tau.a)
```

instead of

```
 mu[i,j] <- m + a[i]
 a[i] ~ dnorm(0,tau.a)
```

This is called 'hierarchical centering'. Although the two are logically implementations of the same model, the former method leads to more efficient MCMC in BUGS.

12.3.1 Missing values

Try the model with some missing values Y_{ij}:

```
list(n=20, K=2, y=structure(.Data=c(108, NA, 91, NA, NA, NA, 104,
NA, 99, 97, 95, 98, 93, 97, 99, 96, 90, 100, 92, 95, 101, 89, 97,
97, 97, 100, 96, 95, 106, 100, 100, 98, 90, 99, 88, 98, 92, 92, 100,
101), .Dim = c(20, 2) ) )
```

Benefits from hierarchical model: for missing values, it shrinks the estimates toward overall mean by borrowing information from other individuals, but accounts for individual differences. Hence, estimates should be more reliable. Alternatives for the hierarchical model? The missing data estimates could be based on global mean, ignoring individual effects $Y_{ij} \sim \mathrm{N}(\mu, \sigma^2)$. Another alternative is $Y_{ij} \sim \mathrm{N}(\mu + a_i, \sigma^2)$ but then the problem is if all measurements for subject i are missing: effect a_i could only be drawn from the prior, i.e. the model could not learn from other individuals.

12.4 Example: cluster sampling

Hierarchical models can be applied in the analysis of data that resulted from cluster sampling. For example, assume that we collect a random sample of units (households, farms, or other groups). Within each unit, we collect a sample of individuals and observe the infectious status for each. The units have differences so that in some of them, almost all individuals are infected, but some units have very little infected individuals. (This type of problems occur also when we have repeated samples from same individuals). Assume the extreme case with either 100% or 0% infection prevalence within unit. Then, observing only one individual from the unit would be enough to determine whether there is 100% or 0% infection. Observing any additional individuals would not provide any further information. Therefore, sample size of one individual per unit would be just as informative as any other sample size. But if the within unit prevalence is 50%, then a large sample of individuals would be needed for an accurate estimate of within unit prevalence. Why?

$$X_i \sim \mathrm{Bin}(N_i, p_i)$$

$$V(X_i \mid p_i, N_i) = p_i(1 - p_i)N_i \qquad \text{with maximum at } p_i = 0.5$$

$$
\begin{aligned}
\pi(p_i \mid N_i, X_i) &= \mathrm{Beta}(X_i+1, N_i-X_i+1)\\
E(p_i \mid X_i, N_i) &= \frac{X_i+1}{N_i+2}\\
V(p_i \mid X_i, N_i) &= \frac{(X_i+1)(N_i-X_i+1)}{(N_i+2)^2(N_i+3)}
\end{aligned}
$$

So, if we look at the posterior variance of p_i as a function of X_i, it is proportional to the function $f(X_i) = -X_i^2+N_iX_i+N+1$. This function is at maximum when $f'(X_i)=0$, that is, when $X_i = N_i/2$. If the true prevalence is really $p_i = 0.5$, then the expected outcome is exactly $X_i = N_i/2$, so that we can expect the posterior variance to be high.

Each unit specific sample helps us to estimate the unit specific prevalence, but to analyze the whole data we need to account for the cluster specific information as well as the overall information from all clusters. Hence the model:

$$
\begin{aligned}
&X_i \sim \mathrm{Bin}(p_i, N_i)\\
&p_i \sim \mathrm{Beta}(\alpha, \beta)\\
&\alpha \sim \text{hyper prior}\\
&\beta \sim \text{hyper prior}
\end{aligned}
$$

Often, logit-transform is used $\theta_i = \mathrm{logit}(p_i) = \log(p_i/(1-p_i))$ so that the (normal) prior is specified for $\theta_i \in \mathbb{R}$.

In Gelman et al [?], example of tumors in rats: N_i rats, X_i rats with tumors, $i = 1, \ldots, 70$. Based on sample mean and sample variance of the observed fractions, point estimate was obtained: $\hat{\alpha}, \hat{\beta} = (1.4, 8.6)$. This is not a bayesian analysis because it is not based on any full probability model. However, part of the data could be used as historical data, for deriving a prior density, and the rest of the data could be used for computing the posterior. In the rat tumor example, beta-prior was used, but noting that a uniform improper prior on $\mathrm{logit}(\alpha/(\alpha+\beta))$ and $\log(\alpha/\beta)$ would yield an improper posterior. Also uniform improper prior for $(\alpha/(\alpha+\beta), (\alpha+\beta))$ or (α, β) would not work. In this example, they chose a uniform prior on $(\alpha/(\alpha+\beta)), (\alpha+\beta)^{-1/2}$, corresponding to the prior

$$
\pi(\alpha, \beta) \propto (\alpha+\beta)^{-5/2}
$$

which is not something that is usually seen in applications, but demonstrates an example when we wish to use beta-distribution as a model directly for p_i, and then we would need a hyper prior for the beta-parameters. It is more typical to use some link function for p_i and then apply hierarchical normal model.

12.5 Random farm effects for pigs

Previously, we had a model where the first measurements (at farms) were modeled to depend on the farm type. But there could be significant farm-to-farm variation. In the extreme situation, there might be farms for which all pigs would be negative, while at other farms they would all be positive. The individual measurements would then be highly dependent within farms. Consequently, sampling only one pig would be (nearly) sufficient to describe all pigs at the same farm. If we allow random farm effects, then we assume that there may be possibly large deviations between individual farms of the same type. In this way, we also adjust for the fact that individual pig measurements are dependent within the same farm. (They are conditionally independent, given the farm effect). The random effect could be incorporated by using the logit-link:

$$\begin{array}{l}\mathrm{logit}(q_i) = \sum_{k=1}^{3} u_k I_k(i) + \epsilon_i \\ \epsilon_i \sim \mathrm{N}(0, \sigma_\epsilon^2) \\ u_k \sim \mathrm{N}(0, 0.001) \\ \tau_\epsilon \sim \Gamma(0.01, 0.01) \quad , \text{ or } \sigma \sim \mathrm{U}(0, 100)\end{array}$$

Then, the probability to be positive at farm of type k with $\epsilon_i = 0$ is $Q_k = \exp(u_k)/(1+\exp(u_k))$. This corresponds to the prevalence for farm type k.

Observe that only one small conventional farm (out of five) had any positive animals ($2/22 \approx 0.09$). Without farm effects, the prevalence for small conventional farms was estimated as a common parameter ($\approx 0.02 \approx 2/(25+25+25+22+25)$). With farm effects, the single positive farm is allowed to be explained by this random farm effect. Hence, the overall estimate for small conventional farms is not so much affected by this single farm which seems to be an exception to the pattern. The prevalence estimate for small conventional farms becomes now lower, ≈ 0.007. We could use the same logit-model either with our without random effects to compare results. Model comparison with DIC indicates that the random effects model is better.

12.6 Normal hierarchical models: priors for variance components

Assume we have sample data y_{ij}, representing measurements from individuals i, $(i = 1, \ldots, n_j)$, from j different groups. We denote the sample mean of the jth group by $\bar{y}_{(.,j)}$. If we assume the variance σ^2 is the same in each group, we obtain the following hierarchical model:

$$\begin{array}{ll}\text{Level 1:} & \mathrm{N}(y_{ij} \mid \theta_j, \sigma^2) \quad , \text{ that is: } \mathrm{N}(\bar{y}_{(.,j)} \mid \theta_j, \sigma_j^2) \text{ where } \sigma_j^2 = \sigma^2/n_j. \\ \text{Level 2:} & \mathrm{N}(\theta_j \mid \mu, \sigma_\theta^2),\end{array}$$

Here, a *hyper prior* (Level 3) is set for parameters μ and σ_θ, assuming that σ_j^2 is known. For the hyper priors we choose:

$$\pi(\mu, \sigma_\theta) = \pi(\mu \mid \sigma_\theta)\pi(\sigma_\theta) \propto \pi(\sigma_\theta),$$

which is improper density for μ, $U(-\infty, \infty)$, but assumes some proper density for σ_θ.
The joint posterior is then of the form:

$$\pi(\theta, \mu, \sigma_\theta \mid y) \propto \pi(\mu, \sigma_\theta) \prod_{j=1}^{J} \mathrm{N}(\theta_j \mid \mu, \sigma_\theta^2) \prod_{j=1}^{J} \mathrm{N}(\bar{y}_{(\cdot,j)} \mid \theta_j, \sigma_j^2),$$

which depends on data only via group means $\bar{y}_{(\cdot,j)}$.

After some mathematical manipulation, the marginal posterior of σ_θ is found to be of the form

$$\pi(\sigma_\theta \mid y) \propto \pi(\sigma_\theta) V_\mu^{1/2} \prod_{j=1}^{J} (\sigma_j^2 + \sigma_\theta^2)^{-1/2} \exp\Big(- \frac{(\bar{y}_{(\cdot,j)} - \hat{\mu})^2}{2(\sigma_j^2 + \sigma_\theta^2)} \Big).$$

$$\hat{\mu} = \frac{\sum_{j=1}^{J} \frac{\bar{y}_{(\cdot,j)}}{\sigma_j^2+\sigma_\theta^2}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\sigma_\theta^2}} \qquad \text{and} \qquad V_\mu^{-1} = \sum_{j=1}^{J} \frac{1}{\sigma_j^2+\sigma_\theta^2}.$$

Note: as an uninformative prior, we can use the improper prior $\pi(\sigma_\theta) \propto 1$, BUT the improper prior $\pi(\log(\sigma_\theta)) \propto 1$ leads to improper posterior!

Recommended: $\pi(\sigma_\theta) = \mathrm{U}(0, L)$. See the paper by Gelman, 2006. With some simulated data y, and some choice of prior $\pi(\sigma_\theta)$, the above marginal posterior density could be studied by plotting it in R, to check empirically the phenomenon reported in Gelman paper. Below is some R-code for this. In the R-code, the marginal density function (apart from normalizing constant) is computed for σ_θ for which the prior density $\pi(\sigma_\theta)$ needs to be expressed in the function. If the prior is uniform, then we simply have a constant term (e.g. 1). But if the prior is defined for σ_θ^2 as Inv-Gamma(a, b), then this is the same as Gamma(a, b) for $\tau_\theta = 1/\sigma_\theta^2$, and gamma-densities are computed numerically in R by using `dgamma()`. But we need to compute the corresponding prior density function for σ_θ, which we can get from variable transformation theorem as: Gamma$(1/\sigma_\theta^2 \mid a, b) \mid \frac{\mathrm{d}\sigma_\theta^{-2}}{\mathrm{d}\sigma_\theta} \mid$. The code below assumes equal number (I) of observations in each group, and J groups.

```
sigbetween <- 1; sigwithin <- 2;
J <- 3; I <- 5;
a <- 0.001; b<- 0.001; L <- 100
m <- numeric()
sigw2 <- numeric()
y <- matrix(NA,J,I)
for(j in 1:J){
sigw2[j] <- sigwithin^2/I # "sig2_j", n_j=I
m[j] <- rnorm(1,0,sigbetween)
for(i in 1:I){
y[j,i] <- rnorm(1,m[j],sigwithin)
}
}
ybar <- numeric()
for(j in 1:J){ybar[j] <- mean(y[j,])}

muhat <- function(sigb2)
sum(ybar/(sigw2+sigb2))/sum(1/(sigw2+sigb2))
Vmu <- function(sigb2)1/sum(1/(sigw2+sigb2))

sigb <- seq(0.02,5,by=0.01)
sigb2 <- sigb^2

marginalg<-numeric()
marginalu<-numeric()
marginallike<-numeric()
for(u in 1:length(sigb)){
```

```
marginalg[u]<-
(2/sigb[u]^3)*dgamma(1/sigb2[u],a,b)*
sqrt(Vmu(sigb2[u]))*
prod((1/sqrt(sigw2+sigb2[u]))*
exp(-0.5*((ybar-muhat(sigb2[u]))^2)/
(sigw2+sigb2[u])))
marginalu[u]<-
(sigb[u]<L)*(sigb[u]>0)/L*
sqrt(Vmu(sigb2[u]))*
prod((1/sqrt(sigw2+sigb2[u]))*
exp(-0.5*((ybar-muhat(sigb2[u]))^2)/
(sigw2+sigb2[u])))
marginallike[u]<-
sqrt(Vmu(sigb2[u]))*
prod((1/sqrt(sigw2+sigb2[u]))*
exp(-0.5*((ybar-muhat(sigb2[u]))^2)/
(sigw2+sigb2[u])))
}
par(mfcol=c(3,1))
plot(sigb,marginalg,type='l',
main=expression(paste("Margin. post.,
with inv-gamma-prior for ",sigma[b]^2)),xlab=expression(sigma[b]))
plot(sigb,marginalu,type='l',
main=expression(paste("Margin. post.,
with U-prior for ",sigma[b])),xlab=expression(sigma[b]))
plot(sigb,marginallike,type='l',
main=expression(paste("Margin. likelihood for ",
sigma[b])),xlab=expression(sigma[b]))
```

12.6.1 Schools example

This is from the Gelman et al book and also discussed in Gelman 2006 paper. In USA, school students are tested with SAT (Scholastic Aptitude Test). In 8 schools, there was a coaching program, and we are interested in the effect of the coaching on the SAT scores. The data represent group means (mean effects on SAT score) and variances (no results of single students given) from 8 schools.

The response variable y_j is the estimated coaching effect on 'SAT-Score'. Also, the sampling variance σ_j^2 of these estimates are reported for each school. These play the same role as $\bar{y}_{\cdot,j}$ and σ_j^2 in the previous theory.

School	Est. treatm. effect y_j	S.E. of estimate, σ_j^2
A	28	15
B	8	10
C	-3	16
D	7	11
E	-1	9
F	1	11
G	18	10
H	12	18

Consider separate estimates: if each school is analyzed separately, assuming the normal model, $y_j \sim \mathrm{N}(\theta_j, \sigma_j^2)$, we get 95% CIs for θ_js of the form $y_j \pm 1.96\sigma_j$, and they all overlap substantially. It is difficult to distinguish between any of the experiments.

Consider pooled estimate: the overlap in the separate posterior intervals suggests that all experiments might be estimating the same quantity. If we then make the hypothesis that all experiments have the same mean effect and produce independent estimates of this common effect, the observations could be modeled as normally distributed with known variances. $y_j \sim \mathrm{N}(\mu, \sigma_j^2)$. With uninformative prior, the posterior of the common μ is

$$\mathrm{N}\left(\frac{\sum \frac{1}{\sigma^2} \bar{y}_{(\cdot,j)}}{\sum \frac{1}{\sigma_j^2}}, (\sum \frac{1}{\sigma_j^2})^{-1} \right)$$

The pooled estimate (posterior mean) is 7.9 and the posterior variance is 17.4. From this we get the 95% CI $7.9 \pm 1.96\sqrt{17.4} = [-0.3, 16.0]$.

Problem: based on separate estimates, for school A we would have 50% probability that the effect is *larger* than 28. Based on pooled estimates, for school A we would have 50% probability that the effect is *smaller* than 7.9. **Both results seem unrealistic**.

→ We would like a compromise that combines information from all eight experiments without assuming all the θ_j are equal. The bayesian analysis with the hierarchical model provides this.
The hierarchical normal model, with constant σ_j^2, is

$$\begin{aligned} y_j &\sim \mathrm{N}(\theta_j, \sigma_j^2) \\ \theta_j &\sim \mathrm{N}(\mu_\theta, \sigma_\theta^2) \\ \pi(\mu_\theta, \sigma_\theta) &\propto 1 \end{aligned}$$

Below the BUGS code for comparing different priors on σ_θ^2.

```
model{
for(i in 1:8){
y[1,i] <- estimate[i]
y[2,i] <- estimate[i]
y[1,i] ~ dnorm(theta[1,i],tau.y[i])
y[2,i] ~ dnorm(theta[2,i],tau.y[i])
tau.y[i] <- pow(sd[i],-2)
theta[1,i] ~ dnorm(m[1],tau.theta[1])
theta[2,i] ~ dnorm(m[2],tau.theta[2])
}
m[1] ~ dnorm(0,0.001)
m[2] ~ dnorm(0,0.001)
tau.theta[1] ~ dgamma(0.01,0.01)
tau.theta[2] <- pow(sig,-2); sig ~ dunif(0,1000)
sig.theta[1] <- pow(tau.theta[1],-1/2)
sig.theta[2] <- pow(tau.theta[2],-1/2)
}
list(estimate=c(28,8,-3,7,-1,1,18,12),sd=c(15,10,16,11,9,11,10,18))
```

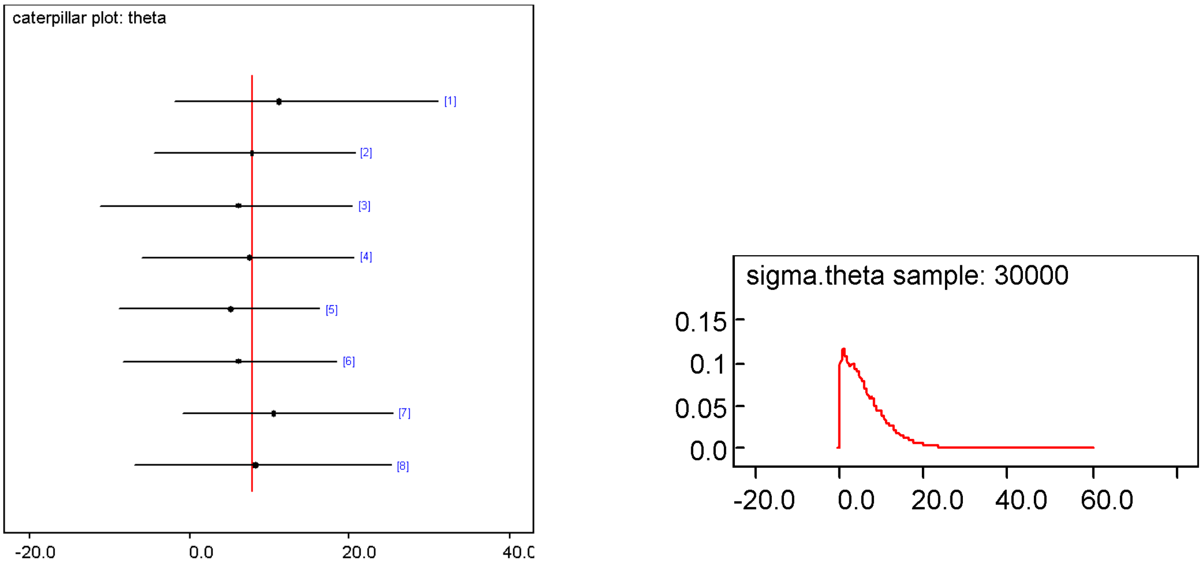


Figure 1: School specific estimates (θ_j) from hierarchical model (left) and the marginal posterior density of σ_θ (right).

12.6.2 Simulated data in R, computation in WinBUGS, outputs in R

Next, similar experiments with hierarchical normal model in BUGS, but running it from R: set the working directory to where you have the following files: `hier_model_normal.bug` and `hier_model_normal2.bug`

```
model{   # hier_model_normal.bug
for(j in 1:J){  # groups
for(i in 1:I){  # individuals within groups
y[j,i] ~ dnorm(theta[j],tau.y)
}
theta[j] ~ dnorm(mu.theta,tau.theta)
}
```

```
mu.theta ~ dnorm(0,1.0E-6)
tau.theta ~  dgamma(a,b)
sigma.theta <- 1/sqrt(tau.theta)
ICC <- (1/tau.theta)/(1/tau.y +1/tau.theta)
}

model{  # hier_model_normal2.bug
for(j in 1:J){  # groups
for(i in 1:I){  # individuals within groups
y[j,i] ~ dnorm(theta[j],tau.y)
}
theta[j] ~ dnorm(mu.theta,tau.theta)
}
mu.theta ~ dnorm(0,1.0E-6)
tau.theta <- pow(sigma.theta,-2)
sigma.theta  ~ dunif(0,1000)
ICC <- (1/tau.theta)/(1/tau.y +1/tau.theta)
}



library(R2WinBUGS)
sigbetween <- 1; sigwithin <- 2;
J <- 3; I <- 5; # J groups, I individuals per group
m <- numeric()
y <- matrix(NA,J,I)
for(j in 1:J){
m[j] <- rnorm(1,0,sigbetween)
for(i in 1:I){
y[j,i] <- rnorm(1,m[j],sigwithin)
}
}
par(mfcol=c(2,1))
tau.y <- 1/(sigwithin^2)
a <- 0.0001; b <- 0.0001 # gamma-prior parameters
data <- list("y","J","I","a","b","tau.y")
parameters <- c("theta","mu.theta","tau.theta","ICC")
inits<-function(){list(theta=rnorm(J,0,0.1),tau.theta=1,mu.theta=0)}
res.sim <-bugs(data,inits,parameters,"hier_model_normal.bug",n.chains=1,n.iter=7500,
               n.burnin=500,n.thin=1);
attach.bugs(res.sim)
plot(density(1/sqrt(tau.theta)),xlim=c(0,10))

data2 <- list("y","J","I","tau.y")
parameters <- c("theta","mu.theta","tau.theta","ICC")
inits2<-function(){list(theta=rnorm(J,0,0.1),sigma.theta=1,mu.theta=0)}
res.sim2<-bugs(data2,inits2,parameters,"hier_model_normal2.bug",n.chains=1,n.iter=7500,
                n.burnin=500,n.thin=1);
attach.bugs(res.sim2)
plot(density(1/sqrt(tau.theta)),xlim=c(0,10))

res <- matrix(NA,J,4)
for(j in 1:J){
res[j,1]<-mean(y[j,]);
res[j,2]<-quantile(theta[,j],0.025,names=FALSE)
res[j,3]<-mean(theta[,j])
res[j,4]<-quantile(theta[,j],0.975,names=FALSE)
}
plot(res[,1],res[,3],xlim=c(-1.5,1.5),ylim=c(-1.5,1.5),
     xlab=expression(bar(y)[j]),
     ylab=expression(paste("E(",theta[j],"|y), and 95% CI")),
     main=expression(sigma[within]==2))
for(j in 1:J){points(c(res[j,1],res[j,1]),c(res[j,2],res[j,4]),'l')}
points(c(-2,2),c(-2,2),'l')
points(c(-2,2),c(mean(mu.theta),mean(mu.theta)),'l')
points(mean(y[]),mean(mu.theta),pch=15)
```

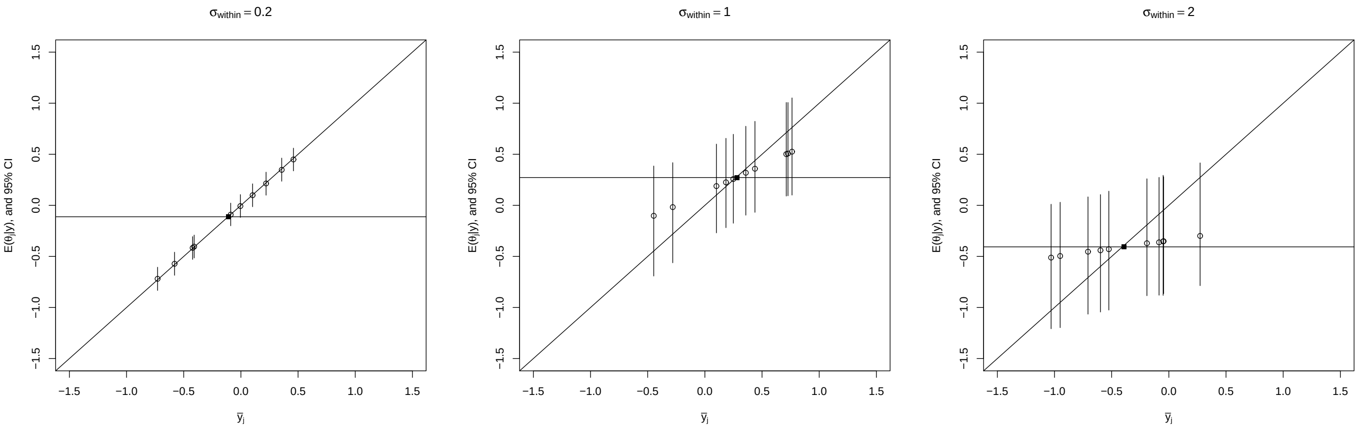


Figure 2: Shrinkage of group means θ_j towards global average. Results from different data generated from $\sigma_\theta = 1/2$ and $\sigma_{\text{within}} \in \{1/5, 1, 2\}$. In all analyses, σ_{within} is assumed known, leaving $\sigma_\theta, \theta_1, \ldots, \theta_J, \mu$ to be estimated. (Prior: $\pi(\sigma_\theta) = \text{U}(0, 1000)$). Horizontal line represents posterior mean of the global parameter $E(\mu \mid y)$. When the within group variance is large, we have more uncertainty about θ_j, hence they become more drawn towards the global estimate. The same phenomenon happens in e.g. Poisson models of disease incidence over geographical areas: in areas with small population, we are more uncertain about the area specific mean 'θ_j'. The hierarchical model would then drag the estimate towards local (neighborhood) or global estimate (depending on model structure). Only when the group level information is strong, we have $E(\theta_j \mid y) \approx \bar{y}_j$. When it's weak, we get $E(\theta_j \mid y) \approx \bar{y}$.

12.7 Example: Rats, hierarchical linear model

See WinBUGS examples Vol 1. These data represent the weights (Y) of 30 rats at 5 different times X. (All rats measured at the same time). The simple approach assumes a linear growth model that has individual parameters for each rat

$$Y_{ij} = \alpha_i + \beta_i X_j + \epsilon_{ij}$$

so that

$$Y_{ij} \sim \mathrm{N}(\alpha_i + \beta_i X_j, \sigma^2)$$

With standardized explanatory variables:

$$Y_{ij} \sim \mathrm{N}(\alpha_i + \beta_i (X_j - \bar{X}), \sigma^2)$$

The rat specific growth parameters α_i, β_i have priors with hyper parameters.

$\tau_c = 1/\sigma^2$	$\sim$	$\mathrm{Gamma}(0.001, 0.001)$
α_i	$\sim$	$\mathrm{N}(\alpha_c, \sigma^2_\alpha)$
α_c	$\sim$	$\mathrm{N}(0, 10^6)$
$\tau_\alpha = 1/\sigma^2_\alpha$	$\sim$	$\mathrm{Gamma}(0.001, 0.001)$
β_i	$\sim$	$\mathrm{N}(\beta_c, \sigma^2_\beta)$
β_c	$\sim$	$\mathrm{N}(0, 10^6)$
$\tau_\beta = 1/\sigma^2_\beta$	$\sim$	$\mathrm{Gamma}(0.001, 0.001)$

From the model, predictions can be done for missing values in different levels: (1) predicting missing values of the same rat, (2) predicting missing rats.

12.8 Example: smoothing mortality rates

Broffit (1988) described a model for bayesian graduation (smoothing) of mortality rates, subject to the restriction that the mortality is increasing function of age, over 35-64 years. Based on insurance records, e_i was the number of people insured ('exposure') in the ith age group, d_i was the number of insured who died. We can compare non-hierarchical and hierarchical models, but in all models we constrain the true mortality rates to be monotonically increasing with age: $\theta_{35} < \theta_{36} < \ldots < \theta_{64}$. This is a *strong assumption*, based on assumed biological effects of ageing. The rates may not be monotonic over the whole life span. For example, if age groups 19-20 were included, mortality might be non-monotonic due to driving accidents of young drivers.

$$d_i \sim \mathrm{Poisson}(\theta_i e_i)$$

$\mathbf{M_1}$:	informative prior	$\mathbf{M_2}$:	'global' hierarchical
θ_i	$\sim \mathrm{Gamma}(\alpha_i, \beta_i) I_{\{\theta_i \in (\theta_{i-1}, \theta_{i+1})\}}$	θ_i	$\sim \mathrm{Gamma}(\alpha, \beta) I_{\{\theta_i \in (\theta_{i-1}, \theta_{i+1})\}}$
α_i	$= 1.5079$	α	$\sim \mathrm{Exp}(0.01)$
β_i	$= 229.3094$	β	$\sim \mathrm{Exp}(0.01)$
$\mathbf{M_3}$:	'global' hierarchical with prior data θ_i^P	$\mathbf{M_4}$:	hierarchical with prior data θ_i^P
θ_i	$\sim \mathrm{Gamma}(\alpha, \beta) I_{\{\theta_i \in (\theta_{i-1}, \theta_{i+1})\}}$	θ_i	$\sim \mathrm{Gamma}(\alpha_i, \beta_i) I_{\{\theta_i \in (\theta_{i-1}, \theta_{i+1})\}}$
α	$\sim \mathrm{Exp}(0.01)$	α_i	$\sim \mathrm{Exp}(0.01)$
β	$\sim \mathrm{Exp}(0.01)$	β_i	$\sim \mathrm{Exp}(0.01)$
θ_i^P	$\sim \mathrm{Gamma}(\alpha, \beta)$	θ_i^P	$\sim \mathrm{Gamma}(\alpha_i, \beta_i)$

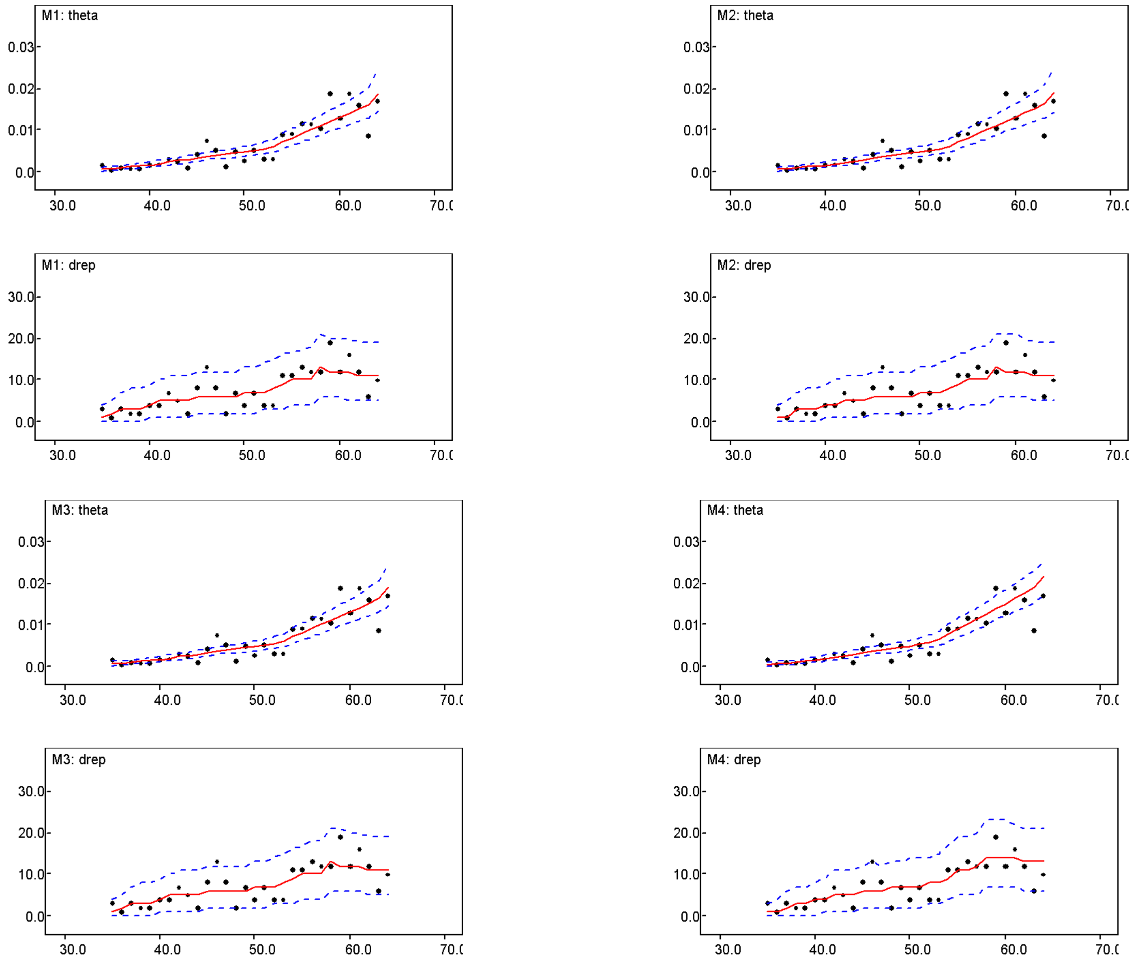


Figure 3: Mortality estimation: WinBUGS outputs.

```
 # Note: set Options -> Updater options -> iterations 1000000
 model{
 # Model M4 structure.
 for( i in 1:k ){
  age[i] <- i+34
  d[i] ~ dpois( lambda[i] ); lambda[i]<-e[i]*theta[i]; thetae[i]<-d[i]/e[i]
  drep[i] ~ dpois(lambda[i])
  theta[1] ~ dgamma(alpha[1],beta[1])I(,theta[2])
  for(i in 2:(k-1)){theta[i] ~ dgamma(alpha[i],beta[i])I(theta[i-1],theta[i+1])}
  theta[k] ~ dgamma(alpha[k],beta[k])I(theta[k-1],B)
  for(i in 1:k){
  thetap[i] ~ dgamma( alpha[i], beta[i] ); # prior data !!
  alpha[i] ~ dexp(0.01); beta[i] ~ dexp(0.01) } }
#################################
 list(B=0.025,k=30, d = c( 3, 1, 3, 2, 2,4, 4, 7, 5, 2, 8, 13,
 8, 2, 7,4, 7, 4, 4, 11,11, 13, 12, 12, 19, 12, 16, 12, 6, 10),
 e = c( 1771.5, 2126.5, 2743.5, 2766.0, 2463.0, 2368.0, 2310.0,
 2306.5, 2059.5, 1917.0, 1931.0, 1746.5, 1580.0, 1580.0, 1467.5,
 1516.0, 1371.5, 1343.0, 1304.0, 1232.5, 1204.5, 1113.5, 1048.0,
 1155.0, 1018.5, 945.0, 853.0, 750.0, 693.0, 594.0 ),
 thetap = c( 0.0012308, 0.0012808, 0.0013609, 0.0014811, 0.0016213,
 0.0017816, 0.0019519, 0.0021423, 0.0023628, 0.0026134, 0.0028942,
 0.0031951, 0.0035362, 0.0039377, 0.0044097, 0.0049422, 0.0054850,
 0.0060382, 0.0066017, 0.0072663, 0.0080523, 0.0090710, 0.0101210,
 0.0111823, 0.0122548, 0.0133386, 0.0145047, 0.0158753, 0.0174514, 0.0192848))
####################################
 list(theta = c(0.0004702563, 0.0007230658, 0.0008120179, 0.0010432968,
 0.0010934937, 0.0012658228, 0.0016891892, 0.0016934801, 0.0017316017,
 0.0024277737, 0.0026385224, 0.0029784066, 0.0030349014, 0.0030674847,
 0.0041429311, 0.0047700170, 0.0050632911, 0.0051039008, 0.0074434583,
 0.0086580087, 0.0089249493, 0.0091324201, 0.0103896104, 0.0114503817,
 0.0116748990, 0.0126984127, 0.0160000000, 0.0168350168, 0.0186548846,
 0.0187573271)
 ,alpha = c( 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
 1, 1, 1, 1, 1, 1, 1, 1, 1, 1 )
 ,beta = c( 100, 100, 100, 100, 100, 100, 100, 100, 100, 100,
 100, 100, 100, 100, 100, 100, 100, 100, 100, 100,
 100, 100, 100, 100, 100, 100, 100, 100, 100, 100 )
```

General problem: analyzing rare disease incidence. Assume a population is stratified into groups (age groups, geographical areas, or other groups). Disease incidence rates are observed in each group (usually as number of cases per 100000 per year). In groups with small population counts, the observed rates easily show very high or very low values, more than in groups with large population. If the disease is relatively rare, then small populations will typically show zero cases. Does it mean that the risk there is zero? How should the (positive, not zero) disease rate there be estimated, considering that we have observations from *all* population groups? Likewise, if there happens to be one or two disease cases in a small group, the observed rate would be very high. Does it mean that the risk is extremely high there? How should we down-weight the estimate, considering observations from all groups?

• Weighting of local point estimate d_i/e_i and global mean μ that results from all data. Small population estimates are shrunk more towards the global mean value than large population estimates.

• Weighting of local point estimates d_i/e_i and local mean μ_{S_i} that results from local data (adjacent geographical regions, adjacent age groups, etc). Small population estimates are shrunk more towards the neighborhood mean value than large population estimates.

Both approaches can be done explicitly by defining a corresponding hierarchical model. In WinBUGS, there is a special package for such models: **GeoBUGS**, see GeoBUGS manual.

Note: if the prior is defined only locally, depending on the 'nearest neighbor' parameters, the resulting prior is not necessarily a proper distribution. In Poisson models, especially in spatial models, priors are often only locally defined (relative to other parameters) and improper unless the parameters (or some of them) are fixed in terms of absolute values. In typical applications, the data (according to Poisson model) is sufficient to make sure that the posterior exists even though the prior is improper and only locally defined. But then we cannot use the prior predictive distribution.

As an example of local smoothing with proper prior, consider the previous example with the priors:

$$\begin{array}{ll} \log\theta_1 & \sim \mathrm{N}(\mu_0, 1000) \\ \mu_0 & \sim \mathrm{N}(0, 1000) \\ \log\theta_{i+1} & \sim \mathrm{N}(\log\theta_i, \sigma^2) \\ \sigma^2 & \sim \mathrm{Gamma}(0.01, 0.01) \end{array}$$

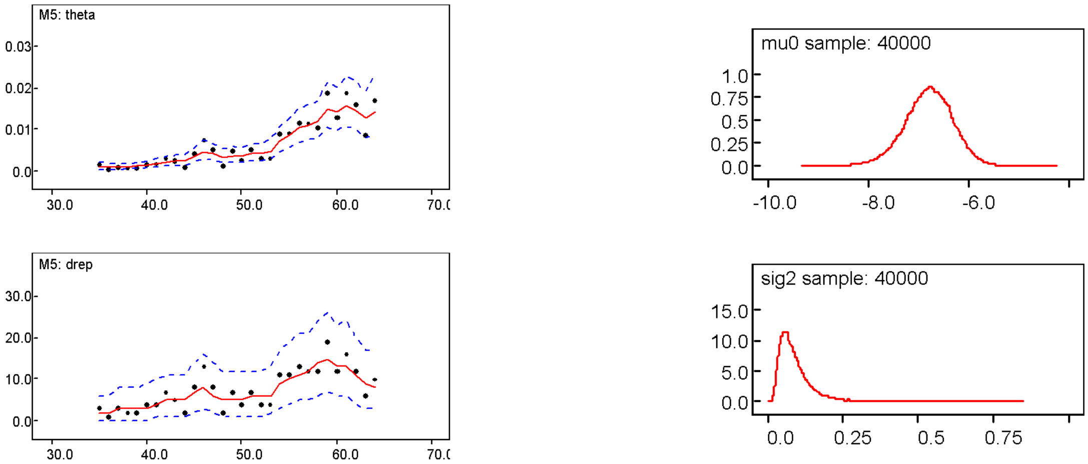


Figure 4: Mortality estimation: WinBUGS outputs.

12.9 Mixture models

Mixture models can also be constructed as hierarchical models. A mixture model is a mixture of probability distributions so that

$$\pi(X \mid \theta_1, \ldots, \theta_k, w_1, \ldots, w_k) = \sum_{i=1}^{k} w_i \pi_i(X \mid \theta_i) \qquad \text{where } \sum w_i = 1.$$

For example: X could be the optically measured length of fish. (Or a function of several measurements). Assuming that the fish can be either salmon or sea bass, its length could be modeled by two conditional distributions with different parameters:

$$\pi(X \mid \text{salmon}) = \pi_1(X \mid \theta_1) \qquad \text{and} \qquad \pi(X \mid \text{sea bass}) = \pi_2(X \mid \theta_2).$$

If w is the proportion of salmon (in the population of salmon and sea bass), the model for X would be

$$\pi(X \mid \theta_1, \theta_2, w) = w\pi_1(X \mid \theta_1) + (1 - w)\pi_2(X \mid \theta_2).$$

The model can also be written by using a *latent* variable Z_i, that is an (unobserved) indicator variable (zero or one) of whether the fish is salmon or not, so that

$$\begin{array}{lcl} Z_i & \sim & \text{Bernoulli}(w) \\ X_i \mid Z_i & \sim & \pi(X_i \mid Z_i) \\ \theta_1 & \sim & \text{prior} \\ \theta_2 & \sim & \text{prior} \\ w & \sim & \text{prior} \end{array}$$

where the distribution $\pi(X_i \mid Z_i)$ is either π_1 or π_2, depending on Z_i.

The model could be used for classification problems. As a result, we would obtain the posterior probability $P(Z_i = 1 \mid \text{data})$ describing the probability that the ith fish is salmon. The full posterior would be computed for all unknowns w, θ_1, θ_2, and Z. Note: some parameters need to be constrained, e.g. $\theta_1 > \theta_2$, because otherwise the full set of parameters would not be identifiable. Parameters are said to be *unidentifiable* when the probability of data ('likelihood function') is equal for different parameter values:

$$\pi(X \mid \psi) = \pi(X \mid \psi') \qquad \text{for some } \psi \neq \psi'.$$

In this example, if the mixture components are normal densities with different unknown means, so that $\pi_i = \text{N}(\theta_i, \sigma^2)$:

$$w\pi_1(X \mid \theta_1) + (1-w)\pi_2(X \mid \theta_2) = (1-w)\pi_1(X \mid \theta_2) + w\pi_2(X \mid \theta_1).$$

This type of unidentifiability is called '*label switching problem*', or '*aliasing*'.

To summarize: mixture models, with latent group indicators are hierarchical models. In the top level, we have parameters of the indicators, then in the next level, given the indicator for group i we have group specific parameters for the observations in the group. This accomplishes a model where the observations within group are more correlated than observations between groups. But now the problem is more complicated because we don't know which observations came from which group. Two possible situations in classification problems: (1) we do not know the true membership for any observation, (2) we know the true membership for some observations. In the latter case, we have 'training data' which helps to estimate the parameters for each group, and then this information can be used when we estimate the unknown membership for the rest of observations.

12.9.1 Example: kangaroo skulls

A set of measurements were made from female (F) and male (M) kangaroo skulls, [**?**]. First, knowing the gender ('training data'), parameters for both conditional models could be estimated. Then, based on just the measurements, we should make a probabilistic classification of female and male skulls. Assume the following model:

$$\begin{array}{lcl} x_i \mid S_i = \text{M} & \sim & \text{N}(\mu_{\text{M}}, \sigma_M^2) \\ x_i \mid S_i = \text{F} & \sim & \text{N}(\mu_{\text{F}}, \sigma_F^2) \\ S_i & \sim & \text{Bern}(p) \\ p & \sim & \text{Beta}(1,1) \\ \mu_{(\cdot)} & \sim & \text{N}(0, 10^6) \\ \sigma_{(\cdot)} & \sim & \text{U}(0, 1000) \end{array}$$

```
list(x=structure(.Data=c(
 1439, 1,
 1413, 1,
 1490, 1,
 1612, 1,
 1388, 1,
 1840, 1,
 1294, 1,
 1740, 1,
 1768, 1,
 1604, 1,
 1464, 2,
 1262, 2,
 1112, 2,
 1414, 2,
 1427, 2,
 1423, 2,
 1462, 2,
 1440, 2,
 1570, 2,
 1558, 2).Dim=c(20,2)))
```

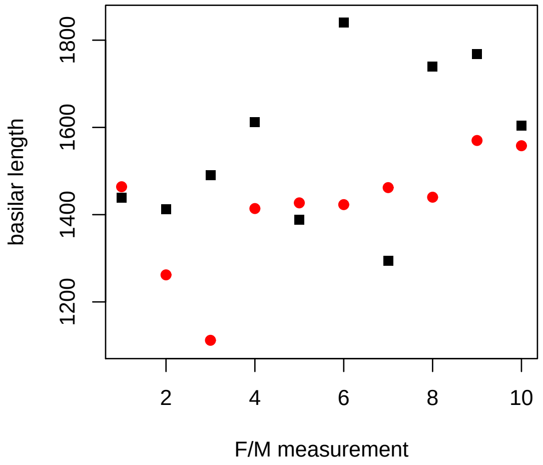


Figure 5: Basilar length of 10 F (red dot) and 10 M (black square) kangaroo skulls.

```
model{
p[1] ~ dunif(0,1); p[2]<- 1-p[1]
for(i in 1:20){
x[i,2] ~ dcat(p[1:2]);
```

```
 x[i,1] ~ dnorm(mu[x[i,2]],tau[x[i,2]]) }
 mu[1] ~ dnorm(0,0.0000001)
 mu[2] ~ dnorm(0,0.0000001)
 s[1] ~ dunif(0,1000); tau[1] <-1/(s[1]*s[1])
 s[2] ~ dunif(0,1000); tau[2] <-1/(s[2]*s[2])
 }
```

With this model and some training data, we can experiment how the classification works when we add some observations which only report the measurement, with missing gender variable M/F. Likewise, we could add some observation which only has known gender, with missing measurement. This would produce the posterior predictive distribution of the measurement for each gender. However, note that if such artificial data are added, the model will interpret these values as if they were real (only with partially missing values). Consequently, posterior of model parameters will be updated based on the whole data. If the artificial data correspond to male measurements, then the mixture probability p will be estimated towards males, etc. If we want to make posterior inference from the actual real data only, and yet produce estimates for the artificial data simultaneously, we could use the cut-function for cutting feedback from the artificial data to the actual model parameters which should be learned from the real data.

Similar classification problem in 2D-measurements is given below. The code provides classification for 2D measurement of (1800,1000) with unknown gender.

```
model{
p[1] ~ dunif(0,1); p[2]<- 1-p[1]
for(i in 1:20){
    x[i,3] ~ dcat(p[1:2])
    x[i,1:2] ~ dmnorm(mu[x[i,3],],Q[x[i,3],,])
}

z ~dcat(p[1:2])
xpred[1:2]~dmnorm(mu[z,],Q[z,,])
pmale <- 2-z  #Indicator for new observation being male.

#means
for (i in 1:2) {
    for (j in 1:2){
        mu[i,j] ~ dflat()
    }
}

#covariances
for (i in 1:2) {
    for (j in 1:2) {
        s[i,j] ~ dunif(0,1000)
        S[i,j,j] <-pow(s[i,j],2)
    }
    rho[i] ~dunif(-1,1)
    S[i,1,2] <- rho[i]*s[i,1]*s[i,2]
    S[i,2,1] <- rho[i]*s[i,1]*s[i,2]
    Q[i,1:2,1:2] <- inverse(S[i, , ])
}}
list(rho = c(0,0), s = structure(.Data=c(1,1,1,1),.Dim=c(2,2)),
mu = structure(.Data = c(1000,1000,1000,1000),.Dim=c(2,2)),z=1)
list(x=structure(.Data=c(
 1439, 824, 1,
 1413, 823, 1,
 1490, 897, 1,
 1612, 921, 1,
 1388, 805, 1,
 1840, 984, 1,
 1294, 780, 1,
 1740, 977, 1,
 1768, 968, 1,
 1604, 880, 1,
 1464, 848, 2,
 1262, 760, 2,
 1112, 702, 2,
 1414, 853, 2,
 1427, 823, 2,
 1423, 839, 2,
 1462, 873, 2,
 1440, 832, 2,
 1570, 894, 2,
 1558, 908, 2),.Dim=c(20,3)),
xpred = c(1800,1000)))
```

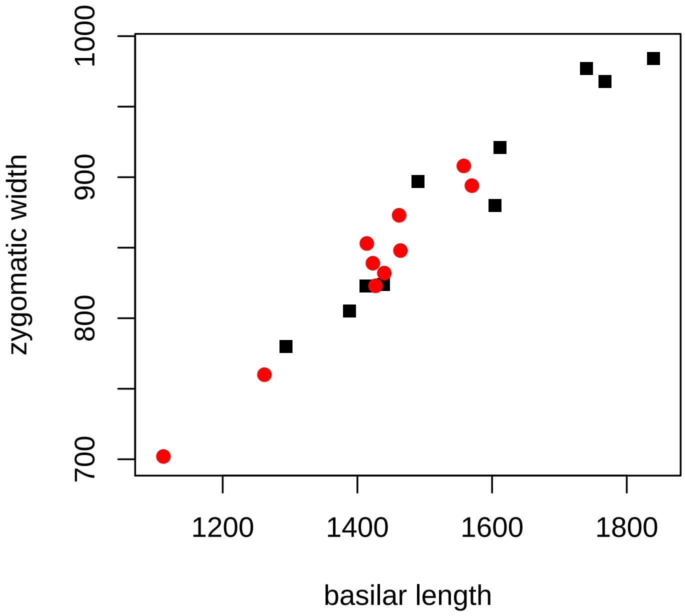


Figure 6: Kangaroo skull measurements. Red dots = female, black squares = male.