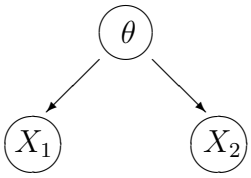
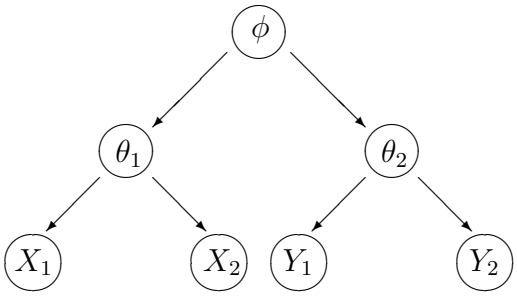
8 Hierarchical models

Any bayesian model can be visualized 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.

What was the basis for this model? 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 partial exchangeability). Therefore, (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, (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.

8.1 Example: genotypes in a family

Each individual has one of the genotypes: AA, Aa or aa and the prior probabilities of these could be based on the assumption of a stable population ('random mixing' of genotypes) so that $P(AA) = 0.25, P(Aa) = 0.5$ ja $P(aa) = 0.25$.

(1) If a child was detected to have genotype AA, then what is the posterior distribution of the parents' genotypes?
(2) What is then the posterior predictive probability that the next born child is of type AA? What is the posterior predictive probability that the child of this child is of type AA?
(3) Draw the model as a DAG.

(1) Denote the parent's genotypes as: X, Y, and the genotype of the 1st child as C_1:

X, Y	prior probability	$P(C_1 = AA \mid X, Y)$	prior $\times$ $P(C_1 = AA \mid X, Y)$	post.
AA, AA	$0.25 \times 0.25 = 1/16$	1	1/16	1/4
AA, Aa	$0.25 \times 0.5 = 2/16$	1/2	1/16	1/4
AA, aa	$0.25 \times 0.25 = 1/16$	0	0	0
Aa, AA	$0.5 \times 0.25 = 2/16$	1/2	1/16	1/4
Aa, Aa	$0.5 \times 0.5 = 4/16$	1/4	1/16	1/4
Aa, aa	$0.5 \times 0.25 = 2/16$	0	0	0
aa, AA	$0.25 \times 0.25 = 1/16$	0	0	0
aa, Aa	$0.25 \times 0.5 = 2/16$	0	0	0
aa, aa	$0.25 \times 0.25 = 1/16$	0	0	0

(2) Posterior predictive probability that the next child would be AA is:

$$P(C_2 = AA \mid C_1 = AA) = \sum_{X,Y} P(C_2 = AA \mid X, Y) P(X, Y \mid C_1 = AA)$$

$$= 1 * 1/4 + 0.5 * 1/4 + 0.5 * 1/4 + 0.25 * 1/4 = 9/4 * 1/4 = 9/16$$

Posterior predictive probabilities for all possible genotypes are

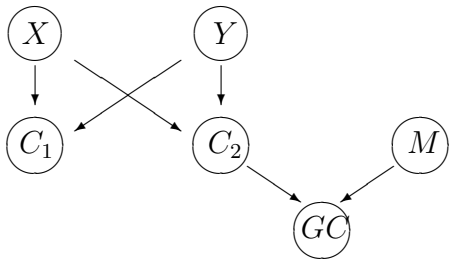
$$\begin{aligned} P(C_2 = AA \mid C_1 = AA) &= 9/16 \\ P(C_2 = Aa \mid C_1 = AA) &= 6/16 \\ P(C_2 = aa \mid C_1 = AA) &= 1/16 \end{aligned}$$

Therefore, posterior predictive probability for a grandchild, GC, (child of C_2), is based on the prediction of the child C_2 and an unknown mate M whose probabilities are described by the prior $P(AA) = 0.25, P(Aa) = 0.5, P(aa) = 0.25$:

C_2, M	$P(C_2, M \mid C_1 = AA)$	$P(GC = AA \mid C_2, M)$	$P(C_2, M \mid C_1 = AA)P(GC = AA \mid C_2, M)$
AA, AA	$(9/16) \times 0.25 = 9/64$	1	9/64
AA, Aa	$(9/16) \times 0.5 = 9/32$	1/2	9/64
AA, aa	$(9/16) \times 0.25 = 9/64$	0	0
Aa, AA	$(6/16) \times 0.25 = 6/64$	1/2	3/64
Aa, Aa	$(6/16) \times 0.5 = 6/32$	1/4	3/64
Aa, aa	$(6/16) \times 0.25 = 6/64$	0	0
aa, AA	$(1/16) \times 0.25 = 1/64$	0	0
aa, Aa	$(1/16) \times 0.5 = 1/32$	0	0
aa, aa	$(1/16) \times 0.25 = 1/64$	0	0
			$= \mathbf{24/64}$

So, the prediction is $P(GC = AA \mid C_1 = AA) = \frac{9}{64} + \frac{9}{64} + \frac{3}{64} + \frac{3}{64} = 24/64 = 3/8 = 0.375$. The prior probability for some 'random' individual would be $P(AA) = 0.25$, but after we know there was one child in this family who was of type AA, this increases the probability that another child is of type AA, which again influences our predictions of the grandchild. Observations about one child affect even the probabilities of his/her unborn brother's/sister's children. This would not happen if we knew exactly the parents' genotypes.

(3) DAG:



8.2 Information synthesis

Hierarchical models are a powerful tool for making synthesis of several sources of information. In the genetics example above, the model could be extended by incorporating data from another family, to predict genotypes of common children and grandchildren. The same principle of modeling can be applied in a variety of applications. For example:

• Categorical data
(several 2 × 2 tables or a large K × K -table, sparse data, empty categories)
• Spatial analysis
(several geographical areas as categories)
• Hierarchical population structures
(individual, group, population, random effects models, 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 selection
(data model, model parameters, 'model of models')

Hierarchical models are not the only form of information synthesis. Even calculation of simple average can be seen as a synthesis of the data points. Also fitting a simple linear regression $y = ax + b$ to describe observed set of points (x, y) is a synthesis of data points. In expert elicitation, Delphi method aims at a synthesis of expert opinions. Draper et al describes broadly the different goals of information synthesis as:

Combining Information:
1.combining within data
2.combining between data
3.combining data & judgement
4.combining judgement information

8.3 Example: smoothing mortality rates with WinBUGS

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.

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

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

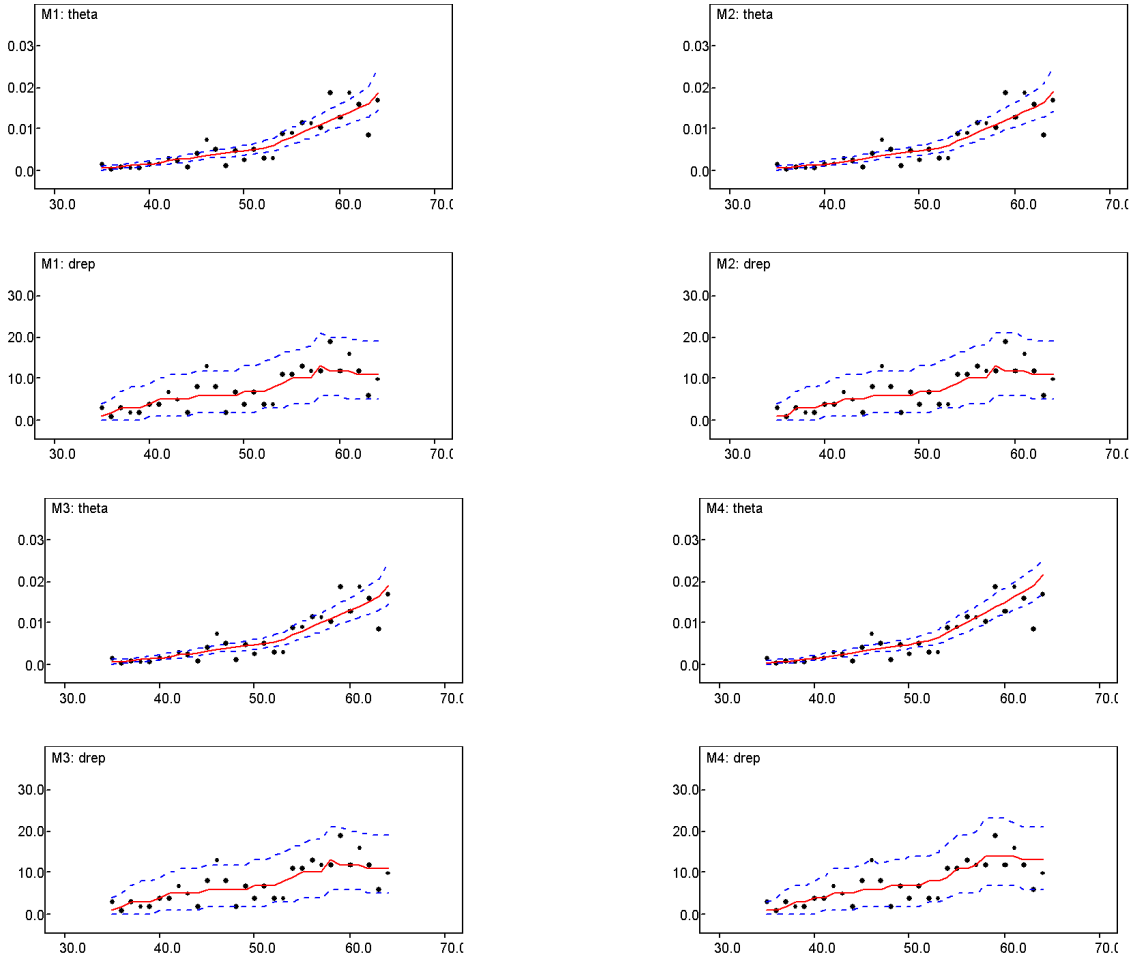


Figure 1: Mortality estimation: WinBUGS outputs.

```
 # 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 downloadable package for such models: **GeoBUGS**.

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:

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

The prior $\tau = 1/\sigma^2 \sim \mathrm{Gamma}(0.01, 0.01)$ gives results somewhat more similar to the earlier ones.

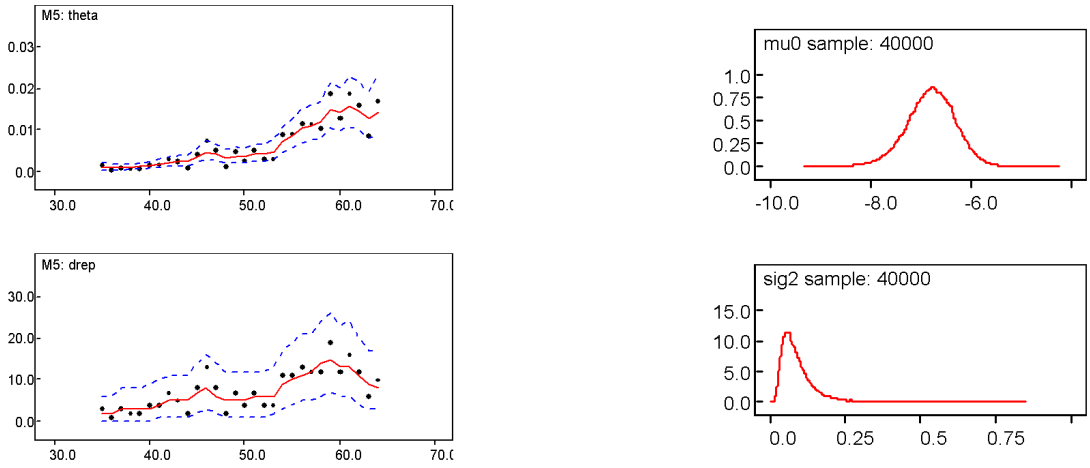


Figure 2: Mortality estimation: WinBUGS outputs.

8.4 Hierarchical normal model

Gelman et al: Bayesian data analysis.

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

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

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 assume some proper density for σ_θ. (We return to this in the end).

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 $\bar{y}_{(\cdot,j)}$.

Marginal full conditional density (could be used in Gibbs-sampler) for each θ_j is:

$$\pi(\theta_j \mid \mu, \sigma_\theta, y) \sim N(\hat{\theta}_j, V_j)$$

$$\hat{\theta}_j = \frac{\frac{1}{\sigma_j^2}\bar{y}_{(\cdot,j)} + \frac{1}{\sigma_\theta^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\sigma_\theta^2}} \qquad V_j = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\sigma_\theta^2}}.$$

This is obtained by taking the part of the above joint posterior density that involves only θ_j, wiping out remaining terms that are constant with respect to θ_j. That is: $\mathrm{N}(\bar{y}_{(\cdot,j)} \mid \theta_j, \sigma_j^2) \times \mathrm{N}(\theta_j \mid \mu, \sigma_\theta^2)$, and then completing the square for $(\theta_j - \ldots)^2$.

The posterior of hyper parameters could be written (in principle) as:

$$\pi(\mu, \sigma_\theta \mid y) \propto \pi(\mu, \sigma_\theta) \underbrace{\pi(y \mid \mu, \sigma_\theta)}_{(\star)}.$$

In this expression, the conditional density $(\star)$ is a distribution of data, given hyper parameters, and it is usually difficult to evaluate (θ-parameters need to be integrated out), but in the case of normal density this is easy, because we know that:

$$\pi(\bar{y}_{(\cdot,j)} \mid \mu, \sigma_\theta, \sigma_j) = \mathrm{N}(\mu, \sigma_j^2 + \sigma_\theta^2),$$

{This is based on the following general result of normal densities: if $X \sim \mathrm{N}(\mu, \sigma_1^2)$ and $\mu \sim \mathrm{N}(0, \sigma_2^2)$, then the joint distribution of (X, μ) is bivariate normal (due to the quadratic function of X, μ), and so the marginal distribution of X is also normal. The marginal mean and variance is obtained applying the results about conditional means and conditional variances in the preliminary notes. Thus the variance $\sigma_1^2 + \sigma_2^2$. }

Therefore:

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

This joint density can further be decomposed as a product of marginal densities:

$$\pi(\mu, \sigma_\theta \mid y) = \underbrace{\pi(\mu \mid \sigma_\theta, y)}_{(\dagger)} \underbrace{\pi(\sigma_\theta \mid y)}_{(\dagger\dagger)},$$

where (†) is:

$$\mathrm{N}(\hat{\mu}, V_\mu),$$

This normal density (with expressions for mean and variance) appears when we notice that the expression (♣) can be written in the form (wiping out constant terms with respect to μ):

$$(\clubsuit) \propto \underbrace{\pi(\mu \mid \sigma_\theta)}_{=1} \prod_{j=1}^{J} \frac{1}{\sqrt{\sigma_j^2 + \sigma_\theta^2}} \exp\Big(-0.5(\mu - \bar{y}_{(\cdot,j)})^2/(\sigma_j^2 + \sigma_\theta^2)\Big)$$

$$\propto \exp\Big(-0.5 \sum_{j=1}^{J} \frac{(\mu - \bar{y}_{(\cdot,j)})^2}{\sigma_j^2 + \sigma_\theta^2}\Big)$$

$$= \exp\Big(-0.5 \sum_{j=1}^{J} \Big[\frac{\mu^2}{\sigma_j^2 + \sigma_\theta^2} - \frac{2\mu\bar{y}_{(\cdot,j)}}{\sigma_j^2 + \sigma_\theta^2} + \frac{\bar{y}_{(\cdot,j)}^2}{\sigma_j^2 + \sigma_\theta^2}\Big]\Big)$$

$$\propto \exp\Big(-0.5 \Big[\mu^2 \sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \sigma_\theta^2} - 2\mu \sum_{j=1}^{J} \frac{\bar{y}_{(\cdot,j)}}{\sigma_j^2 + \sigma_\theta^2}\Big]\Big)$$

$$= \exp\Bigg(-0.5 \sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \sigma_\theta^2} \Bigg[\mu^2 - 2\mu \frac{\sum_{j=1}^{J} \frac{\bar{y}_{(\cdot,j)}}{\sigma_j^2 + \sigma_\theta^2}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \sigma_\theta^2}}\Bigg]\Bigg)$$

$$\propto \exp\Bigg(-0.5 \Bigg[\mu - \frac{\sum \frac{\bar{y}_{(\cdot,j)}}{\sigma_j^2 + \sigma_\theta^2}}{\sum \frac{1}{\sigma_j^2 + \sigma_\theta^2}}\Bigg]^2 \Big/ \Big(\sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \sigma_\theta^2}\Big)^{-1}\Bigg),$$

And this shows the result:

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

The other (††) marginal density is:

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

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!

What happens to $E(\theta_j \mid \sigma_\theta, y)$ and $V(\theta_j \mid \sigma_\theta, y)$ when σ_θ grows bigger?

Answer:

$$\lim_{\sigma_\theta \to \infty} E(\theta_j \mid \sigma_\theta, y) = \bar{y}_{(\cdot,j)}$$

and

$$\lim_{\sigma_\theta \to \infty} V(\theta_j \mid \sigma_\theta, y) = \sigma_j^2.$$

8.5 Hierarchical normal model with WinBUGS: Schools example

Schools example (Gelman et al): 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) from 8 schools.

school	estimated effect y_j	SD (σ_j)
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

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

This is a special case of the earlier solution, with $\sigma_\theta = 0$. (Note that here y_j is the same as $\bar{y}_{(\cdot,j)}$). 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

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

In WinBUGS, this could be coded as:

```
model{
for(j in 1:J){
y[j] ~ dnorm(theta[j],tau.y[j])
theta[j] ~ dnorm(mu.theta,tau.theta)
tau.y[j] <- pow(sigma.y[j],-2)
}
mu.theta ~ dnorm(0,1.0E-6)
tau.theta <- pow(sigma.theta,-2)
sigma.theta ~ dunif(0,1000)
}
list(J=8,y=c(28,8,-3,7,-1,1,18,12),sigma.y=c(15,10,16,11,9,11,10,18))
```

In the theoretical examples, the prior of μ was improper $U(-\infty, \infty)$. In WinBUGS, this is usually replaced by some flat normal density ~`dnorm(0,1.0E-6)`, but it would also be possible to define ~`dflat()` which corresponds to $U(-\infty, \infty)$. (Initial value must be assigned, it cannot be generated from `dflat()`).

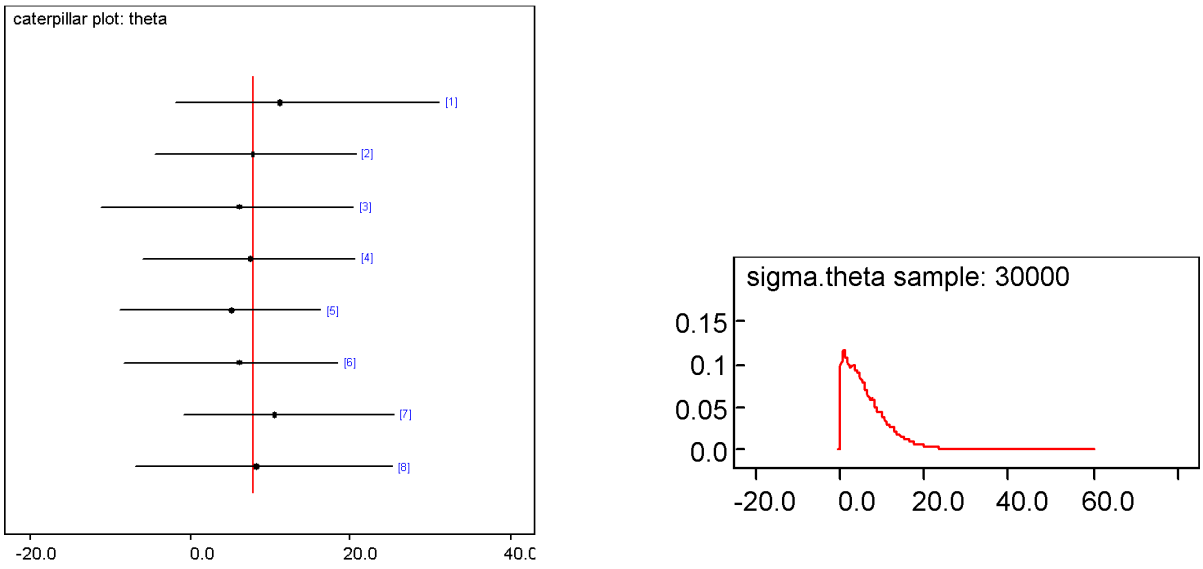


Figure 3: School specific estimates (θ_j) from hierarchical model (left) and the marginal posterior density of σ_θ (right).

8.6 Example: dyes, variance components

This is WinBUGS example, from Examples Vol 1. Box and Tiao (1973) analyse data first presented by Davies (1967) concerning batch to batch variation in yields of dyestuff (väriaine). The data (shown below) arise from a balanced experiment whereby the total product yield was determined for 5 samples from each of 6 randomly chosen batches of raw material.

```
list(batches = 6, samples = 5,
       y = structure(
```

```
.Data = c(1545, 1440, 1440, 1520, 1580,
 1540, 1555, 1490, 1560, 1495,
 1595, 1550, 1605, 1510, 1560,
 1445, 1440, 1595, 1465, 1545,
 1595, 1630, 1515, 1635, 1625,
 1520, 1455, 1450, 1480, 1445), .Dim = c(6, 5)))
```

The model is

$$\begin{aligned} y_{ij} &\sim \mathrm{N}(\mu_i, \sigma^2_{\text{within}}) \\ \mu_i &\sim \mathrm{N}(\theta, \sigma^2_{\text{between}}) \end{aligned}$$

In the WinBUGS example, standard non-informative priors for θ and precision τ_{within} were used.

Three alternative priors for the between-batch variance: prior 1 is a uniform prior on the between batch standard deviation σ_{between}, prior 2 is a uniform prior on the intra-class correlation coefficient $\sigma^2_{\text{between}}/(\sigma^2_{\text{between}} + \sigma^2_{\text{within}})$, and prior 3 is a gamma(0.001, 0.001) prior on the between batch precision τ_{between}.

8.7 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?

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

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

$$\pi(p_i \mid N_i, X_i) = \mathrm{Beta}(X_i + 1, N_i - X_i + 1)$$

$$E(p_i \mid X_i, N_i) = \frac{X_i + 1}{N_i + 2}$$

$$V(p_i \mid X_i, N_i) = \frac{(X_i + 1)(N_i - X_i + 1)}{(N_i + 2)^2(N_i + 3)}$$

So, if we look at the posterior variance of p_i as a function of X_i, it is proportional to the function $f(X_i) = -X_i^2 + N_i X_i + N + 1$. 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:

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

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

In Gelman et al [3], 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 $\text{logit}(\alpha/(\alpha+\beta))$ 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

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

8.8 Mixture 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, 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

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

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 $\pi_i = \mathrm{N}(\theta_i, \sigma^2)$:

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

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

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, in a situation where we don't know which observations came from which group.

8.8.1 Example: kangaroo skulls

A set of measurements were made from female (F) and male (M) kangaroo skulls, [8]. First, knowing the sex ('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:

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

```
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)))
```

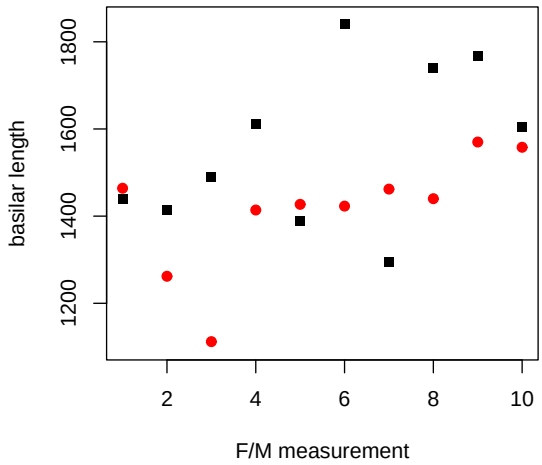


Figure 4: Basilar length of 10 F (red dot) and 10 M (black square) kangaroo skulls.

```
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])
}
```

8.8.2 Example: no training data

If no training data are available, we only have some measurements X for which we could apply a mixture model. In this situation, some constraints for parameters are needed to achieve identifiability. Without training data the problem is harder. If all component distributions have unknown means

and variances to be estimated, poor identifiability can still result and the role of priors can become more essential. It may also happen that none of the data points are assigned to a specific component distribution at some iterations, so that the parameters of that component would be solely dependent on priors. The example below shows some difficulties... try different priors.

```
model{
p[1] ~ dunif(0,1); p[2]<- 1-p[1]
k[3] <- 2; k[17] <-1 #smallest & largest values assigned to groups
for(i in 1:20){
k[i] ~ dcat(p[]);
g1[i] <- equals(k[i],1); g2[i] <- equals(k[i],2)
x[i] ~ dnorm(mu[k[i]],tau[k[i]])
}

s[1] <- sum(g1[]) # size of group 1
s[2] <- sum(g2[]) # size of group 2
mu[1] ~ dnorm(0,0.1);
mu[2] ~ dnorm(0,0.1)I(mu[1],)
tau[1] ~ dgamma(2,1); tau[2] ~ dgamma(2,1)
}
# data generated from dnorm(0,1) and dnorm(3,2):
list(x=c(2.547, 3.285, 3.533, 2.964, 2.946, 3.039, 1.597, 2.838,
3.261, 2.163,  1.332, -0.1041, -0.5081, -0.5214, 1.23,  2.08,
-1.226, 0.166, 0.222, -0.7234))
# initials:
list(tau=c(1,2),mu=c(0,3))
```

References

[1] Berger J: The Case for Objective Bayesian Analysis. Bayesian Analysis, 2006, Vol 1, 3, 385-402.

[2] Goldstein M: Subjective Bayesian Analysis: Principles and Practice. Bayesian Analysis, 2006, Vol 1, 3, 403-420.

[3] Gelman A, Carlin J B, Stern H S, Rubin D B: Bayesian data analysis, 2nd edition. Chapman & Hall/CRC. 2004.

[4] Gelman A: Prior distributions for variance parameters in hierarchical models. Bayesian Analysis, 1, No 3, pp. 515Ű533. 2006.

[5] Jaynes E T: Probability theory: the logic of science. Cambridge university press. 2003.

[6] Sivia D S: Data Analysis, a Bayesian tutorial, 2nd edition. Oxford university press. 2006.

[7] Robert C P, Casella G: Monte Carlo Statistical Methods. Springer 1999.

[8] Congdon P: Bayesian Statistical Modelling. John Wiley & Sons, Ltd. 2001.

[9] Congdon P: Applied Bayesian Modelling. John Wiley & Sons, Ltd. 2003.

[10] Bernardo J M, Smith A F M: Bayesian Theory. John Wiley & Sons, Ltd. 2000.

[11] Lee P M: Bayesian Statistics, an introduction. 3rd ed. Hodder Education. 2004.