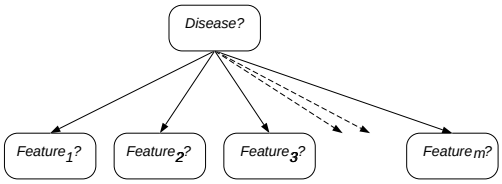
A much-used classification model

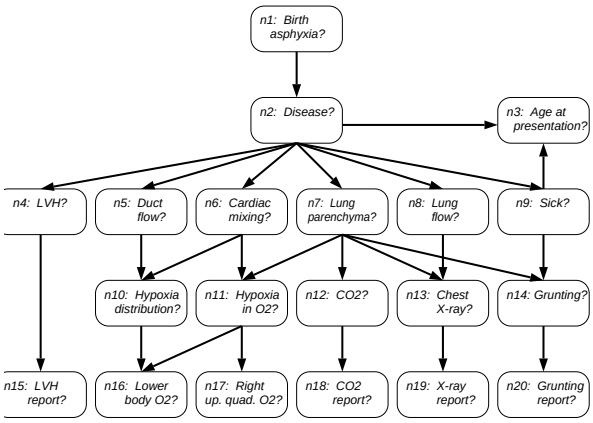
This graph expresses the view that, once the disease class is known, information about one set of feature variables is of no further relevance to predicting the values of some other disjoint set.



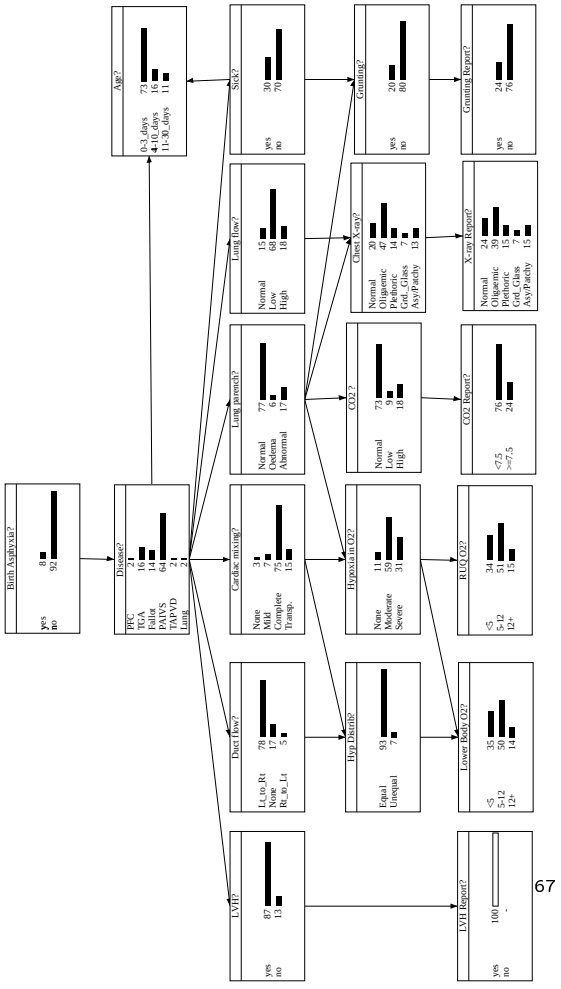
Directed graphical model representing conditional independence of feature variables within each disease class – the "idiot's Bayes" model This is equivalent to assuming

$$p(D, F_1, \ldots, F_m) = p(D) \prod_{i=1}^{m} p(F_i|D).$$

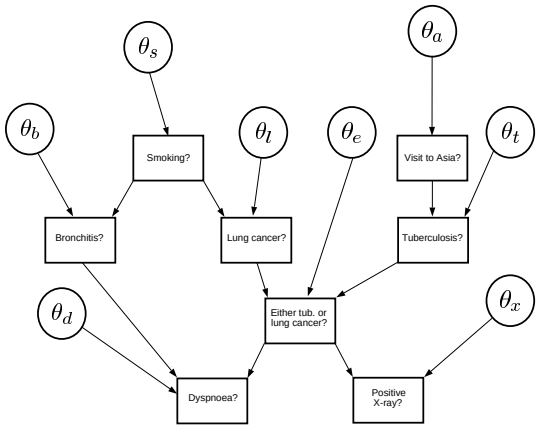
The CHILD network



Directed acyclic graph representing the incidence and presentation of six possible diseases that would lead to a "blue" baby (*LVH* = Left Ventricular Hypertrophy).



From Bayesian networks to Bayesian graphical models:



Network from Lauritzen and Spiegelhalter (1988) with supplementary 'parameter' nodes, representing marginally independent random quantities $\theta_v, v \in V$ whose realizations specify the conditional probability tables for the network.

Suppose we observe I sets of binomial data, e.g.

- I=12 Hospitals performing cardiac surgery in babies
- Number of surgical failures (deaths) per centre

	Hospital						
	A	**B**	**C**	....	**J**	**K**	**L**
No. of ops. n	47	148	119	....	97	256	360
No. of deaths r	0	18	8	....	8	29	24

Surgical example continued

We shall treat this problem using a Bayesian hierarchical model, with a random-effects logistic regression model and 'vague' priors on its hyperparameters.

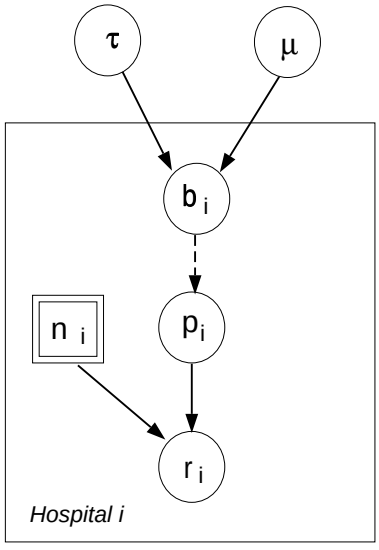
- *Likelihood:*
  - Binomial response with random effects:

$$\begin{aligned} r_i &\sim \text{Binomial}(\pi_i, n_i) \\ \text{logit}(\pi_i) &= \alpha_i \\ \alpha_i &\sim \text{Normal}(\mu, 1/\tau) \end{aligned}$$

- *Priors:*

$$\begin{aligned} \mu &\sim \text{Normal}(0.0, 1,000,000) \\ \tau &\sim \text{Gamma}(1.0\text{e-}3, 1.0\text{e-}3) \\ \sigma &= \sqrt{\tau^{-1}} \end{aligned}$$

Graph for Example 1



Recap: Principals of Bayesian Inference

- Joint posterior distribution:

$$p(\underline{\theta}|\underline{x}) \propto p(\underline{x}|\underline{\theta})p(\underline{\theta})$$

  $\underline{x}$ = observed data
  $\underline{\theta}$ = unobserved variables
  (e.g. parameters, missing data, latent variables)
- Typically, we require inference on *single* parameters, say θ_k
- This is achieved by averaging (integrating) the joint posterior over all other unknowns $\theta_1, \theta_2, ..., \theta_{k-1}, \theta_{k+1}, ...\theta_p$, i.e.

$$p(\theta_k|\underline{x}) = \int p(\underline{\theta}|\underline{x}) d\theta_1 d\theta_2 ... d\theta_{k-1} d\theta_{k+1} ... d\theta_p$$

- Analytical or numerical (e.g. Laplace) integration is typically intractable for real-life problems
- Monte Carlo integration = *simulation* from the joint posterior
- Marginal posterior inference then involves simple data summaries

Monte Carlo Integration

- Suppose we can draw samples from the joint posterior distribution for $\underline{\theta}$, *i.e.*

$$\underline{\theta}^{(1)}, \underline{\theta}^{(2)}, ..., \underline{\theta}^{(n)} \sim p(\underline{\theta}|\underline{x})$$

- Then

$$\begin{aligned} E(g(\underline{\theta})) &= \int g(\underline{\theta}) p(\underline{\theta}|\underline{x}) d\underline{\theta} \\ &\approx \frac{1}{n} \sum_{i=1}^{n} f(\underline{\theta}^{(i)}) \end{aligned}$$

this is Monte Carlo integration

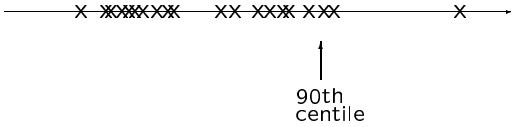
Monte Carlo integration is (sometimes) easy!

Assume we have a sample from $p(\theta_k \mid \underline{x})$

- Posterior mean

$$E(\theta_k) \approx \frac{1}{N} \sum_{i=1}^{N} \theta_k^{(i)}$$

- Quantiles



- Kernel density estimates

$$\hat{p}(\theta_k|\underline{x}) \approx \frac{1}{N} \sum_{i=1}^{N} h(\theta_k; \theta_k^{(i)})$$

How do we sample from the posterior?

- In general, we want samples from the joint posterior distribution $p(\underline{\theta}|\underline{x})$
- *Independent* sampling from $p(\underline{\theta}|\underline{x})$ may be difficult
- **BUT** *dependent* sampling from a *Markov chain* with $p(\underline{\theta}|\underline{x})$ as its stationary (equilibrium) distribution is easier
- A sequence of random variables $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, ...$ forms a Markov chain if

$$\theta^{(i+1)} \sim p(\theta|\theta^{(i)})$$

i.e. conditional on the value of $\theta^{(i)}$, $\theta^{(i+1)}$ is independent of $\theta^{(i-1)}, ..., \theta^{(0)}$

- Theorems exist which show that $\frac{1}{n}\sum_{i=1}^{n} f(\theta^{(i)}) \to E(f(\theta))$ as $n \to \infty$ when $\theta^{(1)}, ..., \theta^{(n)}$ are sampled from a suitable Markov chain

To summarize:

Bayesian posterior inference may be achieved via Monte Carlo integration using simulated values of all the unknown quantites $\underline{\theta}$ in the model generated from a Markov chain with $p(\underline{\theta}|\underline{x})$ as its stationary distribution

This is **Markov chain Monte Carlo (MCMC)**

How do we design a Markov chain with $p(\underline{\theta}|\underline{x})$ as its unique stationary distribution?

- Metropolis *et al.* (1953) showed how to do this
- This method was generalized by Hastings (1970)
- **Gibbs Sampling** (see Geman and Geman (1984), Gelfand and Smith (1990), Casella and George (1992)) is a special case of the Metropolis-Hastings algorithm which generates a Markov chain by sampling from **full conditional distributions**
- See Gilks, Richardson and Spiegelhalter (1996) for a full introduction and many worked examples.

Gibbs sampling

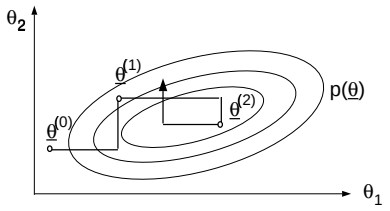
Let our vector of unknowns $\underline{\theta}$ consist of k sub-components $\underline{\theta} = (\theta_1, \theta_2, ..., \theta_k)$
e.g. for the HB model, $\underline{\theta} = (\alpha, \beta, \gamma, \sigma^2)$

1) Choose starting values $\theta_1^{(0)}$, $\theta_2^{(0)}$, ..., , $\theta_k^{(0)}$

2) Sample $\theta_1^{(1)}$ from $p(\theta_1|\theta_2^{(0)}, \theta_3^{(0)}, ..., , \theta_k^{(0)}, \underline{x})$
   Sample $\theta_2^{(1)}$ from $p(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, ..., , \theta_k^{(0)}, \underline{x})$
   .....
   Sample $\theta_k^{(1)}$ from $p(\theta_k|\theta_1^{(1)}, \theta_2^{(1)}, ..., , \theta_{k-1}^{(1)}, \underline{x})$

3) Repeat step 2 many 1000s of times
   – eventually obtain sample from $p(\underline{\theta}|\underline{x})$

Gibbs sampling ctd.



- Sample $\theta_1^{(1)}$ from $p(\theta_1|\theta_2^{(0)}, \underline{x})$
- Sample $\theta_2^{(1)}$ from $p(\theta_2|\theta_1^{(1)}, \underline{x})$
- Sample $\theta_1^{(2)}$ from $p(\theta_1|\theta_2^{(1)}, \underline{x})$
- .....

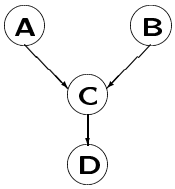
$\underline{\theta}^{(n)}$ forms a Markov chain with (*eventually*) a stationary distribution $p(\underline{\theta}, \underline{x})$.

Sampling from full conditional distributions

- To do Gibbs sampling, we need to sample from distributions of the form
  $p(\theta_j|\theta_1, ..., \theta_{j-1}, \theta_{j+1}, ...\theta_k, \underline{x})$
- In general, these full conditional density functions, f, are
  – one-dimensional (*i.e.* θ_j is a scalar)
  – of complex algebraic form that doesn't simplify to a known distribution
  – log-concave *i.e.* Hessian matrix (matrix of 2nd derivatives) for $\log f$ is negative semi-definite
- For one-dimensional, log-concave full conditional distributions we can use *Adaptive Rejection Sampling* (Gilks and Wild, 1992)
- GLMs with canonical link always give log-concave full conditionals

Constructing full conditionals for a model

- Recall the conditional indepence graph from Lecture 2



- *Directed Local Markov property*

$$\begin{aligned} v &\perp\!\!\!\perp \text{non-descendants}[v] \mid \text{parents}[v] \\ D &\perp\!\!\!\perp A, B \mid C \end{aligned}$$

- *Factorisation of joint distribution*

$$\begin{aligned} p(V) &= \prod_{v \epsilon V} p(v \mid \text{parents}[v]) \\ &= p(A)\, p(B)\, p(C \mid A, B)\, p(D \mid C) \end{aligned}$$

- *Full conditional for* A

$$\begin{aligned} p(A \mid .) &\propto \text{terms on RHS containing } A \\ &\propto p(A)\, p(C \mid A, B) \end{aligned}$$

The BUGS program

Bayesian inference Using Gibbs Sampling

- Language for specifying complex Bayesian models
- Parsing and pre-processing
- Constructs object-oriented internal representation of the model graph by identifying parents and children
- Simulation from full conditionals using Gibbs sampling
- 'Classic' BUGS is now superceded by WinBUGS, which incorporates the DoodleBUGS graphical model editor, and on-line viewing of the simulations.

Recall the surgical example

- *Likelihood:*
  - Binomial response with random effects:

$$\begin{aligned} r_i &\sim \text{Binomial}(\pi_i, n_i) \\ \text{logit}(\pi_i) &= \alpha_i \\ \alpha_i &\sim \text{Normal}(\mu, \tau) \end{aligned}$$

- *Priors:*

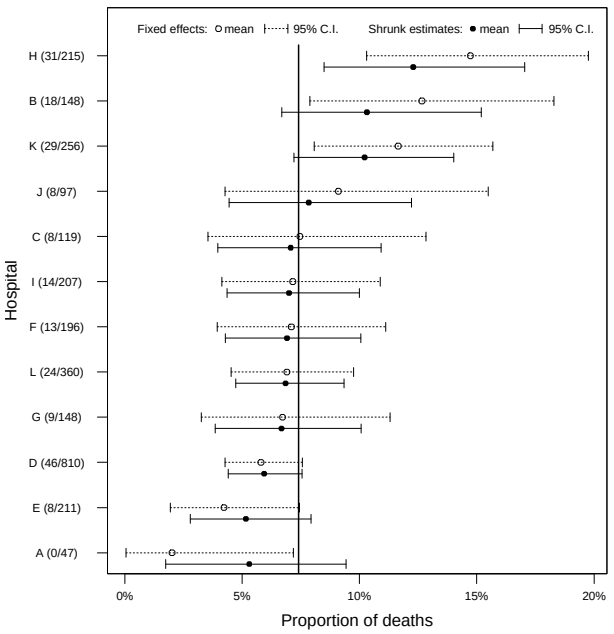
$$\begin{aligned} \mu &\sim \text{Normal}(0.0, 1.0\text{e-}6) \\ \tau &\sim \text{Gamma}(1.0\text{e-}3, 1.0\text{e-}3) \\ \sigma &= \sqrt{\tau^{-1}} \end{aligned}$$

Classic BUGS code for surgical example

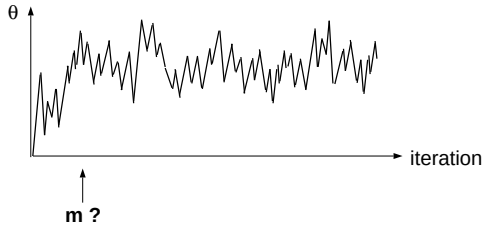
```
model surgical;
const
   N = 12; # number of hospitals
var
   r[N],pi[N],n[N],alpha[N],
   mu,tau,sigma,pop.mean;

data in ''surgical.dat'';
inits in ''surgical.in'';
{
    # Likelihood:
    for (i in 1:N) {
        r[i] ~ dbin(p[i], n[i]);
        logit(pi[i]) <- alpha[i];
        alpha[i] ~ dnorm(mu, tau);
    }
    # Priors:
    mu ~ dnorm(0.0,1.0E-6);
    tau ~ dgamma(1.0E-3, 1.0E-3);
    sigma <- 1.0/sqrt(tau);
    pop.mean <- exp(mu/(1+mu));
}
```

Note that the normal distribution is parameterised in terms of its *precision* (1/variance).

Example 1: Fixed and Smoothed estimates of Surgical Mortality Rates



Convergence diagnosis

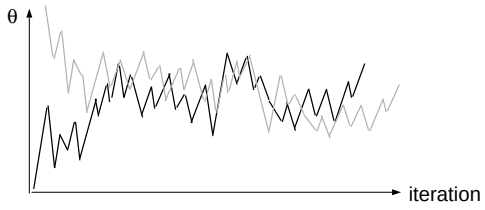


- Strictly speaking, convergence is only achieved for $n = \infty$
  $\Rightarrow$ concept of 'burn-in' is not well-defined theoretically
- **BUT**, we only need Markov chain to be 'approaching' convergence for Monte Carlo integration to yield a consistent estimate of the true expectation
- How do we determine m, the number of 'burn-in' iterations?
- Several 'convergence diagnostics' exist
- Focus on Gelman & Rubin (1992)

Convergence diagnosis ctd.

Gelman & Rubin (1992)

- Many long runs
- Widely differing starting points



- Convergence assessed by quantifying whether sequences are much further apart than expected based on their internal variability
- Diagnostic uses components of variance of the multiple sequences

`CODA`

- Output processor for `BUGS`
- Can be used with any MCMC program
- `http://mrc-bsu.cam.ac.uk/bugs`
- Menu-driven set of S-Plus functions (now also available in freeware R)

A) Convergence Diagnosis:
- Geweke (1992)
- Gelman & Rubin (1992)
- Raftery & Lewis (1992)
- Heidelberger & Welch (1983)
- Autocorrelations
- Cross-correlations

B) Summary statistics:
- Empirical means, sd's and quantiles
- Standard error of the mean

C) Graphical:
- Sample trace for each variable
- Kernel density
- Plots of the autocorrelation function
- Plots of the cross-correlations
- Plots of Geweke's diagnostic
- Plots of Gelman & Rubin's diagnostic

Some practical issues:

- Initial values
- Parameterization
- Priors

Initial values

- *All* unknown variables must be given initial values. These are either
  - specified by the user in the `.in` file (Classic BUGS) or in a document (WinBUGS)

  or
  - generated automatically in `BUGS` using forward sampling from prior
- If variable given values in data *and* initial values files ⇒ error!
- For
  - fixed effect regression coefficients
  - population hyperparameters
  - parameters with vague priors

  always specify initial values in the `.in` file
  - avoids generation of inappropriate values

    e.g. `tau ~ dgamma(0.001, 0.001)`
    - failing to specify and initial value gives:

```
-- error -- Invalid value for node tau
            - expected positive value.
```


- Poor choice of inital values may cause convergence problems e.g.

```
for(i in 1:N) {
  for(j in 1:T) {
    Y[i,j] ~ dnorm(mu[i,j],tau.c);
    mu[i,j] <- alpha[i] + beta[i]*(x[j]-x.bar);
  }
  alpha[i] ~ dnorm(alpha.c,tau.alpha);
  beta[i] ~ dnorm(beta.c,tau.beta);
}
alpha.c ~ dnorm(0,1.0E-4);
beta.c ~ dnorm(0,1.0E-4);
tau.alpha ~ dgamma(1.0E-3,1.0E-3);
tau.beta ~ dgamma(1.0E-3,1.0E-3);
tau.c ~ dgamma(1.0E-3,1.0E-3);
```

Data (`Y`):

```
 151 199 246 283 320 145 199 249
 293 354 147 214 263 312 328......
```

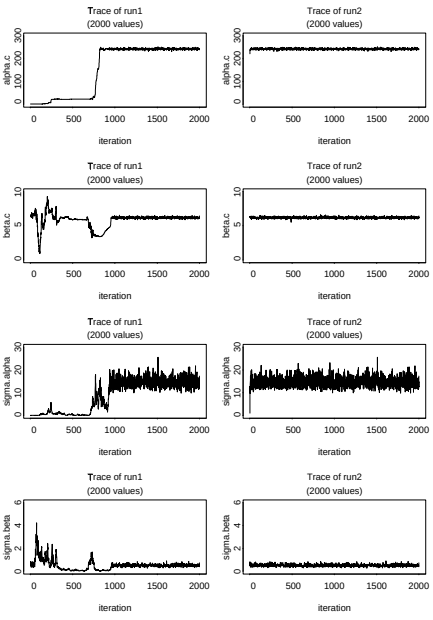
Inital values (run 1):

```
list(alpha.c=0.1, beta.c=0, tau.alpha=1000,
     tau.beta=1, tau.c=1, seed=489225436)
```

Inital values (run 2):

```
list(alpha.c=100, beta.c=0, tau.alpha=1,
     tau.beta=1, tau.c=1, seed=643828390)
```

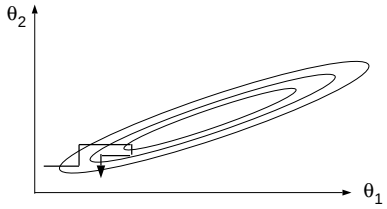

Effect of bad initial values on convergence



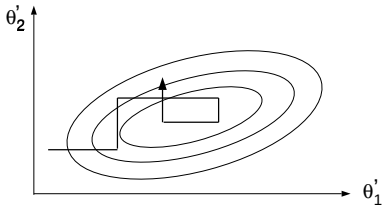
Parameterization

- MCMC samplers often show poor mixing
  - i.e. sampler does not move rapidly throughout the support of the target distribution
- Slows convergence and increases Monte Carlo error variance
- Chains tend to be highly autocorrelated
- Often caused by high posterior correlations between model parameters
- 'Ordered over-relaxation' (Neal, 1998) deliberately produces negatively correlated samples, and is implemented in WinBUGS.

- Slow Mixing



- Fast Mixing



Options for reparameterizing a model

1. Regression models

- Consider simple regression model

$$\begin{aligned} y_i &\sim \text{Normal}(\mu_i, \sigma^2) \\ \mu_i &= \alpha + \beta x_i \\ \alpha, \beta &\sim \text{flat priors} \end{aligned}$$

- Posterior correlation between α and β is

$$\rho_{\alpha\beta} = -\frac{\overline{x}}{\sqrt{\overline{x}^2 + \frac{1}{n}\sum_i (x_i - \overline{x})^2}}$$

- If $\overline{x} >>$ sd(x) $\Rightarrow \rho_{\alpha\beta} \rightarrow \pm 1$
- Remedy: standardize x_i about the sample mean $\overline{x}$:

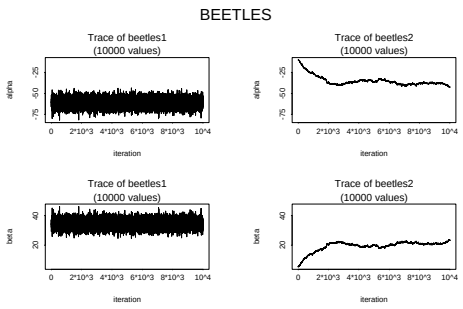
$$\mu_i = \alpha' + \beta'(x_i - \overline{x})$$

- Posterior correlation $\rho_{\alpha'\beta'} = 0$

- e.g. `beetles` example:

Model 1: $\text{logit} p_i = \alpha' + \beta(x_i - \overline{x})$

Model 2: $\text{logit} p_i = \alpha + \beta x_i$



- For multiple regression models, rescaling covariates to roughly equalise their sample standard deviations often helps to reduce correlation

Non-informative priors

- May not want priors to be influential
- Distinguish
  - *primary* parameters of interest
  - *secondary* structure used for smoothing *etc.*

1. *Location parameters e.g.* regression coefficients:

$$\alpha \sim \text{Normal}(0.0, 0.0001)$$

   $\Rightarrow$ standard deviation of 100
   $\Rightarrow$ 95% prior interval ± 200
   $\Rightarrow$ prior will be locally uniform over the region supported by the likelihood

2. *Scale parameters e.g.* precision of random effects:
   - Standard 'reference' prior

$$p(\tau) \propto \frac{1}{\tau}$$
$$p(\sigma^2) \propto \frac{1}{\sigma^2}$$

   $\Rightarrow$ at second level of hierarchy this will give an **improper** posterior distribution since $\sigma^2 = 0, \tau = \infty$ is supported by non-negligible likelihood.

- Options:

A) 'Just proper': $\tau \sim \text{Gamma}(\epsilon, \epsilon)$

B) $\sigma \sim$ uniform $(0, r) \iff \tau \sim \text{Pareto}(\frac{1}{2}, r^{-2})$ where

$$\tau \sim \text{Pareto}(\alpha, c) \iff p(\tau) = \alpha c^{\alpha} \tau^{-(\alpha+1)} \quad \tau > c$$

C) Think about proper prior

Example of a Bayesian GLMM: Hepatitis B Immunisation (Spiegelhalter et al, 1996)

Background
- Hepatitis B (HB) is endemic in Africa
- National program of childhood vaccination against HB introduced in Gambia
- Program effectiveness depends on duration of immunity afforded by vaccination

Data
- 106 children immunized against HB
- For each child: anti-HB titre measured at time of vaccination (baseline) and on 2 or 3 follow-up occasions

Study objective
- To obtain a model useful for predicting an individual child's protection against HB after vaccination

Related studies
- Similar study in Senegal found:

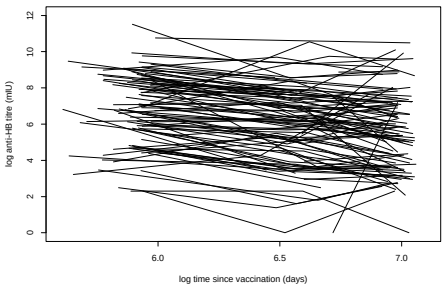
$$\text{anti-HB titre} \propto \frac{1}{T}$$

where T = time since HB vaccination

1. Probability dist^n (likelihood) for responses:

$$y_{ij} \sim \text{Normal}(\mu_{ij}, \sigma^2)$$

   where y_{ij} = *log* of the jth anti-HB titre measurement for child i
2. Linear predictor?
   - Assume that response (log anti-HB titre) may depend on log-time t_{ij} and on log-baseline titre $y0_i$.



- Allow separate intercept and slope for each child:

$$\begin{aligned} y_{ij} &\sim \text{Normal}(\mu_{ij}, \sigma^2) \\ \mu_{ij} &= \alpha_i + \beta_i(t_{ij} - \bar{t}) + \gamma(y0_i - \overline{y0}) \end{aligned}$$

- What prior distributions should we chose for the α_i's and β_i's?
- Assume that all the α_i's follow a *common population* prior distribution, and likewise for the β_i's *e.g.*

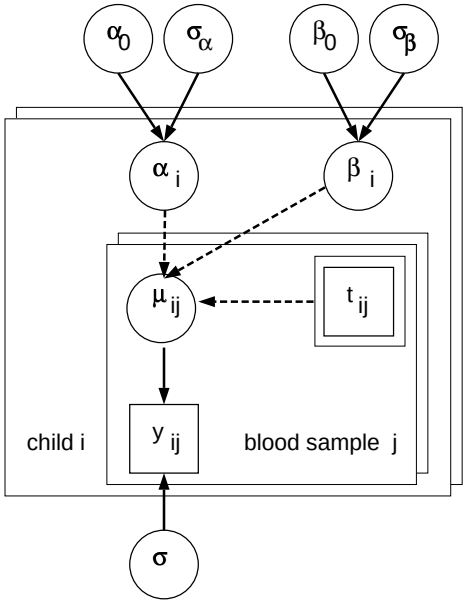
$$\begin{aligned} \alpha_i &\sim \text{Normal}(\mu_\alpha, {\sigma^2}_\alpha) \quad i = 1, ..., 106 \\ \beta_i &\sim \text{Normal}(\mu_\beta, {\sigma^2}_\beta) \quad i = 1, ..., 106 \end{aligned}$$

- We may then assume vague priors for the *hyperparameters* of the population distribution:

$$\begin{aligned} \mu_\beta, \mu_\alpha &\sim \text{Normal}(0, 10000) \\ \tau_\alpha, \tau_\beta = \frac{1}{{\sigma^2}_\alpha} &\sim \text{Gamma}(0.001, 0.001) \end{aligned}$$

- This is an example of a *Hierarchical GLM* or *Generalized Linear Mixed Model (GLMM)* or *Random Coefficients* model

Graph of a GLMM for the HB data



BUGS code for random effects growth curve.

```
{
for(j in 1:N){
    y[j]     ~ dnorm(mu[j],tau);
    mu[j] <- alpha[child[j]]
              + beta[child[j]] * log.time[j];
    }
for(i in 1:M){
    beta[i]  ~ dnorm(beta0, tau.beta);
    alpha[i] ~ dnorm(alpha0, tau.alpha);
    }
alpha0    ~ dnorm(0,0.0001);      #   priors
beta0     ~ dnorm(0,0.0001);
tau.beta  ~ dgamma(0.01,0.01);
tau.alpha ~ dgamma(0.01,0.01);
tau       ~ dgamma(0.01,0.01);

sigma       <- 1/sqrt(tau);
sigma.beta  <- 1/sqrt(tau.beta);
sigma.alpha <- 1/sqrt(tau.alpha);
```


Adjustment for observed baseline measure

y_{i0} = log(titre) on ith child at baseline.

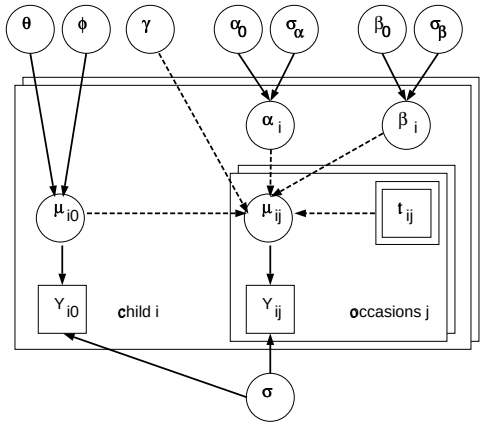
$$\begin{aligned} \mu_{ij} &= \alpha_i + \gamma y_{i0} + \beta_i t_{ij} \\ \alpha_i &\sim \text{Normal}(\alpha_0, \sigma^2_\alpha) \\ \beta_i &\sim \text{Normal}(\beta_0, \sigma^2_\beta) \end{aligned}$$

Adjustment for 'true' baseline measure

μ_{i0} = 'true' log(titre) on ith child at baseline.

$$\begin{aligned} \mu_{ij} &= \alpha_i + \gamma \mu_{i0} + \beta_i t_{ij} \\ y_{i0} &\sim \text{Normal}(\mu_{i0}, \sigma^2) \\ \alpha_i &\sim \text{Normal}(\alpha_0, \sigma^2_\alpha) \\ \beta_i &\sim \text{Normal}(\beta_0, \sigma^2_\beta) \end{aligned}$$

Allowing measurement error in the baseline measurement for the HB data



	Fixed baseline	Errors in baseline
β_0	-1.06	-1.08
	(-1.32, -0.80)	(-1.35, -0.81)
σ_β	.31	.24
	(.07, .76)	(.07, .62)
γ	.68	1.04
	(.51, .85)	(.76, 1.42)

- Use of single baseline measure grossly underestimates association: *regression dilution bias* - must allow for measurement error.
- Assuming $\beta \approx -1, \gamma \approx 1$ gives

$$\frac{\text{titre at time } t}{\text{titre at time 0}} \propto \frac{1}{t}$$

Modelling and Computational Issues

- Univariate Gibbs sampling can be very inefficient
- Many cleverer alternative sampling methods exist (WinBUGS is planning to use some of these)
- Bayesian graphical modelling naturally incorpoartes
  - Measurement error
  - Spatial correlation
  - Informative missing data
  - etc
- Important extensions include non-parametric and semi-parametric prior structures
- Wide range of applications
- Growing connection with image analysis, genetics, neural networks and other complex stochastic systems

Software

- `First Bayes`: free software for teaching Bayesian analysis, runs under Windows.
  `http://www.maths.nott.ac.uk/personal/aoh/`
- `The Association for Uncertainty in Artificial Intelligence` provides links to software for Bayesian networks:
  `http://www.auai.org`
- Radford Neal's MCMC software for flexible Bayesian modelling for neural networks, mixture models etc is on
  `http://www.cs.utoronto.ca/ radford`
- `BUGS`: software for Gibbs sampling in complex models. Freely available with manual and examples.
  `http://www.mrc-bsu.cam.ac.uk/bugs`

Barnett, V. (1982). *Comparative Statistical Inference (Second Edition)*. J Wiley and Sons, Chichester, UK.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Phil Trans Roy Soc*, **53**, 418.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. John Wiley and Sons, Chichester, England.

Berry, D. A. and Stangl, D. K. (ed.) (1996). *Bayesian Biostatistics*. Dekker.

Breslow, N. (1990). Biostatistics and Bayes. *Statistical Science*, **5**, (3), 269–84.

Buntine, W. (1994). Operations for learning with graphical models. *Journal of Artificial Intelligence Research*, **2**, 159–255.

Buntine, W. (1996). A guide to the literature on learning probabilistic networks from data. *IEEE Transactions on Knowledge and Data Engineering*, **8**, 195–210.

Carlin, B. P. and Louis, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall, London, U.K.

Chaloner, K. (1996). Elicitation of prior distributions. In (Berry and Stangl, 1996), chapter 4, pp. 141–56.

Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion). *J Roy Statist Soc B*, **41**, 1–31.

Dawid, A. P. (1986). Probability forecasting. In *Encyclopaedia of Statistical Sciences, Vol 7*, (ed. S. Kotz and N. L. Johnson), pp. 210–8. John Wiley and Sons, New York.

Dixon, D. O. and Simon, R. (1992). Bayesian subset analysis in a colorectal cancer clinical trial. *Statistics Medicine*, **11**, 13–22.

Fleming, T. R. and Watelet, L. F. (1989). Approaches to monitoring clinical trials. *J Natl Cancer Inst*, **81**, 188–93.

Frey, B. J. (1998). *Bayesian Networks for Pattern Classification, Data Compression, and Channel Coding*. MIT Press, Cambridge MA.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman and Hall, New York.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–72.

Genest, C. and Zidek, J. (1986). Combining probability distributions: a critique and an annotated bibliography (with discussion). *Statistical Science*, **1**, 114–48.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo Methods in Practice*. Chapman and Hall, New York.

Jordan, M. (1998). *Learning in graphical models*. Kluwer Academic Publishers, Dordrecht.

Kadane, J. B. (1995). Prime time for Bayes. *Controlled Clinical Trials*, **16**, 313–8.

Kass, R. E. and Greenhouse, J. B. (1989). Comments on 'Investigating Therapies of potentialy great benefit:ECMO' by J H Ware. *Statistical Science*, **4**, 310–7.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *J Roy Statist Soc B*, **50**, 157–224.

Lee, P. M. (1989). *Bayesian Statistics: an Introduction*. Edward Arnold, London.

Lilford, R. J. and Braunholtz, D. (1996). For debate - the statistical basis of public policy - a paradigm shift is overdue. *British Medical J.*, **313**, 603–7.

Lindley, D. V. ((1985)). *Making decisions*, (second edn). John Wiley and Sons.

Louis, T. A. (1991). Using empirical Bayes methods in biopharmaceutical research. *Statistics Medicine*, **10**, 811–29.

Morris, C. N. (1983). Parametric empirical bayes inference: Theory and applications. *Journal of the American Statistical Association*, **78**, 47–55.

Neal, R. (1996). *Bayesian Learning for Neural Networks*. Springer-Verlag, New York.

Neal, R. (1998). Suppressing random walks in markov chain monte carlo using ordered overrelaxation. In *Learning in graphical models*, (ed. M. I. Jordan), pp. 205–30. Kluwer Academic Publishers, Dordrecht.

Spiegelhalter, D., Myles, J., Jones, D., and Abrams, K. (1999). Bayesian methods in health technology assessment. *Health Technology Assessment*, **to appear**.

Spiegelhalter, D. J., Best, N. G., Gilks, W. R., and Inskip, H. (1996). Hepatitis: a case study in MCMC methods. In *Markov Chain Monte Carlo Methods in Practice*, (ed. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter), pp. 21–44. Chapman and Hall, New York.

Spiegelhalter, D. J., Freedman, L. S., and Parmar, M. K. B. (1994). Bayesian approaches to randomized trials. *J. Royal Statistical Soc. Series A-Statistics Society*, **157**, 357–87.

Tversky, A. (1974). Assessing uncertainty (with discussion). *J Roy Statist Soc B*, **36**, 148–59.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Analysis*. John Wiley and Sons, Chichester.