

Postgraduate Symposium  
December 2006

Byron Bayes

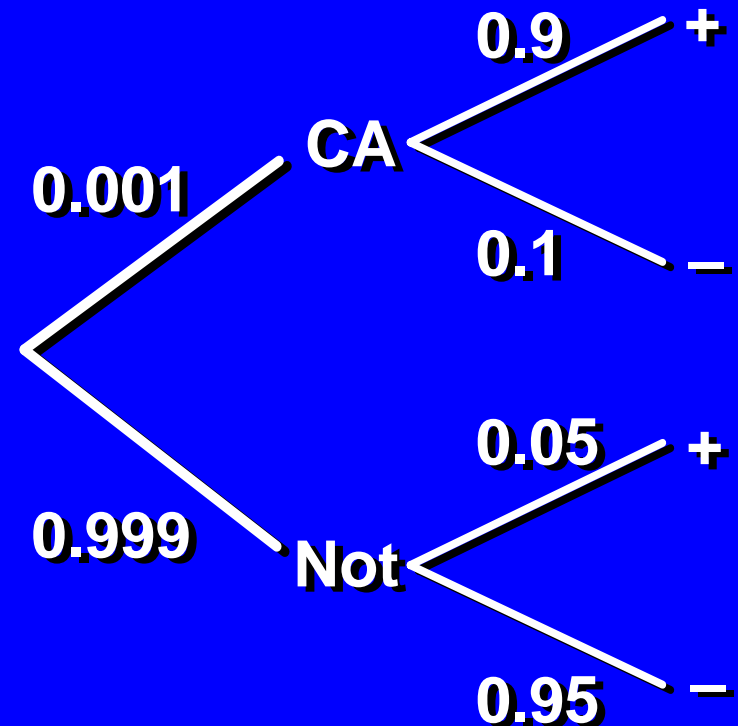
Kerrie Mengersen  
QUT

# Plan

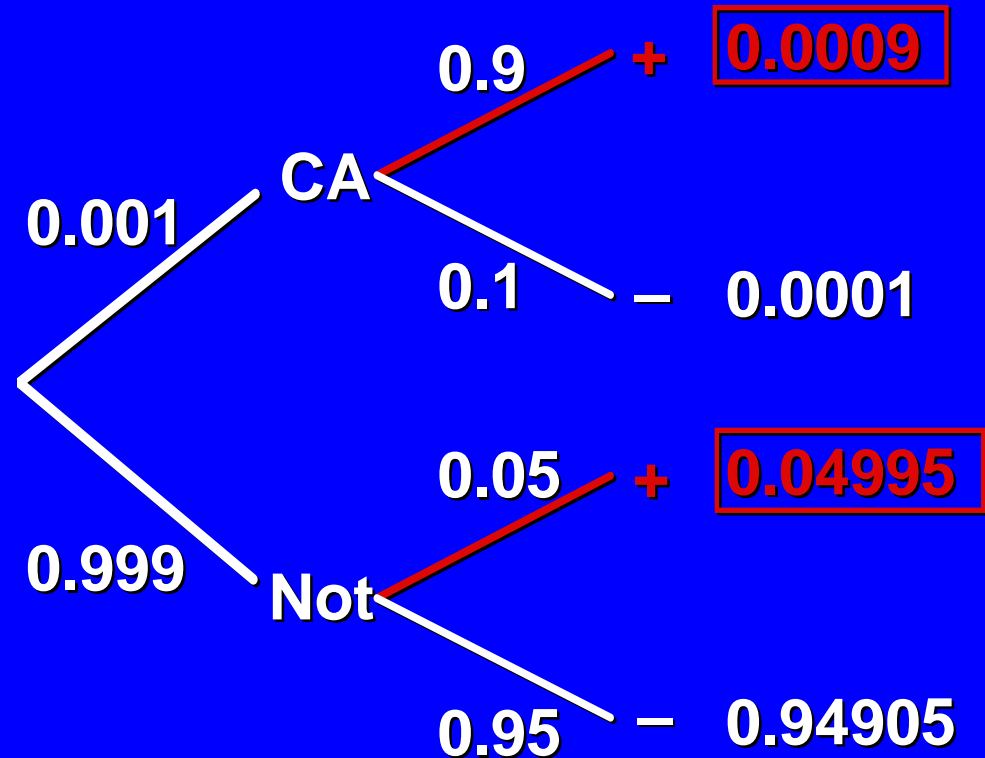
1. Basics of Bayesian Inference
2. Markov chain Monte Carlo
3. Introduction to WinBUGS
4. Case studies

# Decision trees: screening for cancer

- The probability of cancer is 0.0001
- The probability of a positive test given cancer = 0.9  
(sensitivity)
- The probability of a negative test given no cancer = 0.95  
(specificity)



*What is the chance of cancer, given that the test is positive?*



Now, instead of  
 $P(+ve\ test\ | \ cancer)$   
we want  
 $P(cancer\ | \ +ve\ test)$

$$\begin{aligned} \text{Prob.} &= \Pr(\text{cancer and +ve test}) / \Pr(+ve\ test) \\ &= 0.0009 / (0.0009 + 0.04995) \\ &= 0.0177 \end{aligned}$$

# What has this to do with modelling?

$$P(A|B) = P(A \text{ and } B)/P(B)$$

*and*  $P(B|A) = P(B \text{ and } A)/P(A)$

*so*  $P(A|B) = P(B|A)P(A) / P(B)$

**Think of:**  $A=\theta$  (unknown parameters, etc)  
 $B=y$  (known 'data')

**So**  $P(\theta|y) = P(y|\theta) p(\theta) / p(y)$

***This is Bayesian analysis!***

# Posterior $\propto$ Data $\times$ Prior

Unknown  $\theta$ : parameters, missing data, latent variables etc

What do we know  
from our data?

**likelihood**

$p(y|\theta)$  + model

What else do know  
we about  $\theta$ ?

**prior**

$p(\theta)$

Combined info. about  $\theta$  *given* data  $y$

**posterior**

$p(\theta|y)$

# Why Bayes?

Bayesian methods allow us to:

- Think differently about estimating and interpreting unknown parameters  
“what are possible values of this parameter?”
- Combine prior information with the data  
“what else do I know about this parameter and model?”
- Describe many sources of uncertainty in the model  
“how sure am I about the inputs to my model?”
- Describe complex systems using hierarchical or multi-level models

# Why Bayes?

Bayesian computational methods (such as MCMC) allow us to:

- Use non-standard distributions
- Fit very complex models
- Obtain a very wide variety of estimates
- Make a very wide range of inferences, based directly on posterior probabilities

# Example: Estimating a probability

- **Unknown:**  $\theta$ : probability of survival after surgery

- **Data:** 29 patients: 22 survive and 7 die.

- **Likelihood:**  $y/\theta \sim \text{Binomial}(n, \theta)$

$$p(y | \theta) \propto \theta^y (1-\theta)^{n-y} \quad p(y=22|\theta) \propto \theta^{22}(1-\theta)^7$$

- **Prior for  $\theta$ :** continuous distribution between 0 and 1:

$$\theta \sim \text{Beta}(a, b)$$

$$p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$$

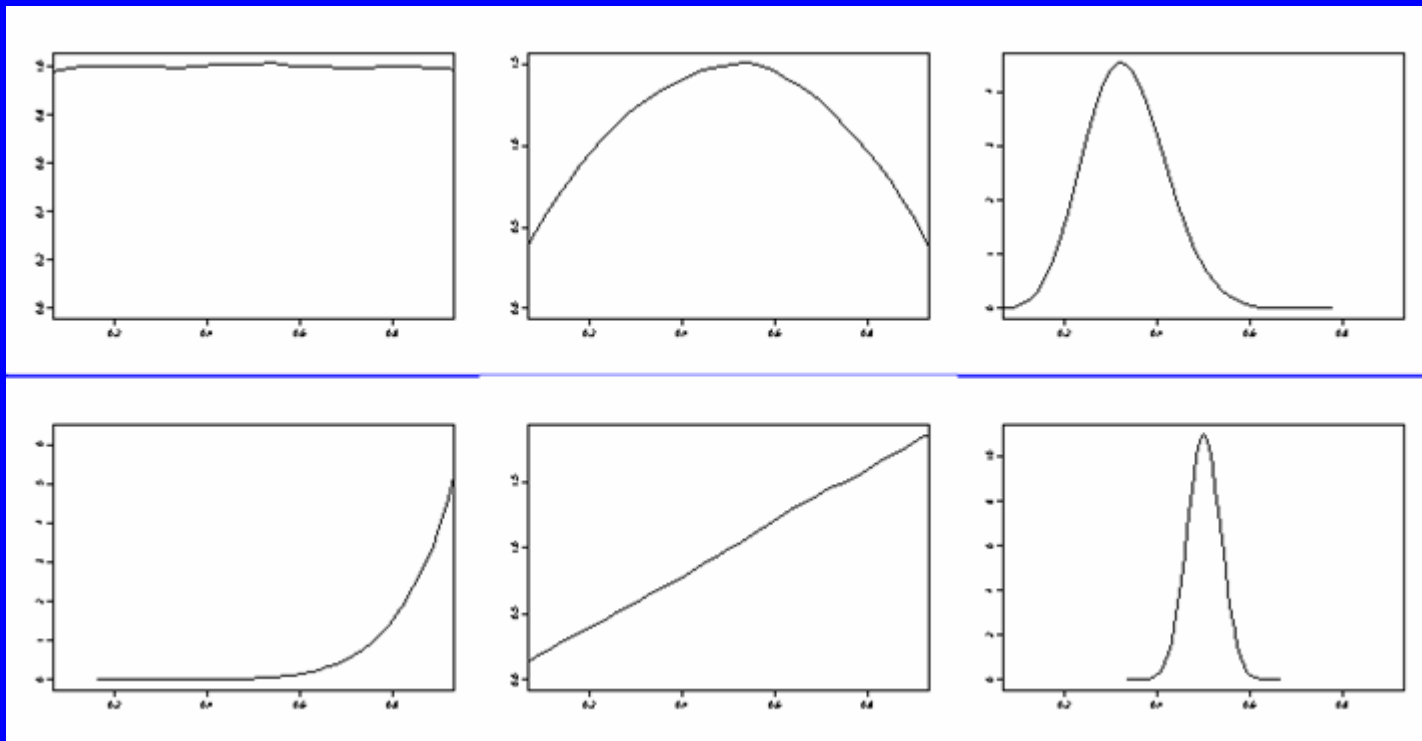
Beta(a,b):

$$\text{Mean} = E(\theta) = a/(a+b)$$

$$\text{Var}(\theta) = ab/\{(a+b)^2(a+b+1)\}$$

Match the plots to the distributions:

Beta(1,1), Beta(2,2), Beta(100,100), Beta(2,1), Beta(10,20), Beta(9,1)



# Back to estimating a proportion

- *Prior: Some alternatives*

1. Assume we ‘know nothing’ about  $\theta$ , so we set a uniform prior  $\theta \sim U(0,1)$ , equivalently,  $\theta \sim \text{Beta}(1,1)$
2. Based on past information, adopt a  $\text{Beta}(9,1)$  prior.
3. Based on expert info, assume a  $\text{Beta}(100,100)$  prior.

- *Posterior:*

$$\begin{aligned} P(\theta | y) &\propto \text{likelihood} \times \text{prior} \\ &= \theta^y (1-\theta)^{n-y} \times \theta^{a-1} (1-\theta)^{b-1} \\ &= \theta^{y+a-1} (1-\theta)^{n-y+b-1} \end{aligned}$$

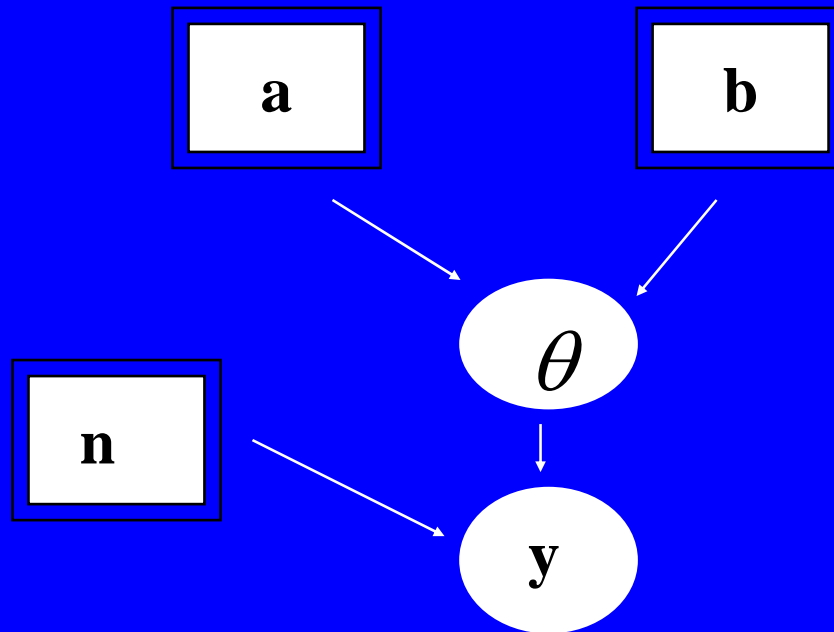
$$\theta | y \sim \text{Beta}(y+a, n-y+b)$$

# Graphical model (DAG)

- **Model**

$$y \sim \text{Binomial}(\theta, n)$$

$$\theta \sim \text{Beta}(a, b)$$



# Your turn!

- Data:
  - a) 22 successes, 7 failures
  - b) 220 successes, 70 failures
- Priors for  $\theta$ :
  - a) Beta(1,1)
  - b) Beta(9,1)
  - c) Beta(100,100)

*For each of these:*

- 1. What is the observed proportion of successes?*
- 2. What is the prior mean for  $\theta$ ?*
- 3. What is the posterior mean for  $\theta$  given the data?*
- 4. What general conclusions can you make about the influence of priors and sample size?*

# Answers:

Sample proportion =  $22/29 = 0.76$

***Beta(1,1):***

Prior mean =  $1/(1+1) = 0.5$

Posterior Beta(23,8); mean =  $23/31 = 0.74$

***Beta(9,1):***

Prior mean =  $9/(9+1) = 0.90$

Posterior Beta(31,8); mean =  $31/39 = 0.79$

***Beta(100,100):***

Prior mean =  $100/(100+100) = 0.5$

Posterior Beta(122,107); mean =  $122/229 = 0.53$

# Influences on posterior

- The posterior mean is a compromise between the prior mean and the data.
- The stronger the prior, the more weight the prior has in the posterior.
- The larger the sample size, the more weight the likelihood has in the posterior.

# Dynamic Updating

- If we obtain more data, we do not have to redo all of the analysis: our posterior from the first analysis simply becomes our prior for this next analysis.

- Binomial example:

Stage 0. Prior  $p(\theta) \sim \text{Beta}(1,1)$ ; ie  $E(\theta)=0.5$ .

Stage 1. Observe  $y=22$  ‘presences’ from 29 sites.

Likelihood:  $p(y|\theta) \sim \text{Bin}(n=29, \theta)$ ;

Posterior:  $p(\theta|y) \sim \text{Beta}(23,8)$ ; ie  $E(\theta|y) = 0.74$

Stage 2: Observe 5 more ‘presences’ from 10 sites.

Likelihood:  $p(y|\theta) \sim \text{Bin}(n=10, \theta)$ ;

Prior  $p(\theta) \sim \text{Beta}(23,8)$ ;

Posterior  $p(\theta|y) \sim \text{Beta}(28,13)$ ; ie  $E(\theta|y) = 0.68$ .

# Normal Model, unknown mean, one observation

Suppose that we have one observation from a normal distribution, known variance.

What do we learn about the population mean?

- **Likelihood:**  $y|\theta \sim \mathbf{N}(\mu, \sigma^2)$

- **Prior:**  $\mu \sim \mathbf{N}(\mu_0, \sigma_0^2)$

$\mu_0, \sigma_0^2$  can be specified values representing our “best guess” at the true mean and how certain we are of this. (Or we can put priors on these values as well.)

- Likelihood:  $y|\theta \sim \mathbf{N}(\mu, \sigma^2)$
- Prior:  $\mu \sim \mathbf{N}(\mu_0, \sigma_0^2)$

- Posterior:

$$p(\mu/y) \sim \mathbf{N}(\mu_1, \sigma_1^2)$$

Posterior mean is a weighted average of prior and data

$$\mu_1 = ( \mu_0/\sigma_0^2 + y/\sigma^2 ) / (1/\sigma_0^2 + 1/\sigma^2)$$

$$1/\sigma_1^2 = 1/\sigma_0^2 + 1/\sigma^2$$

Posterior variance also combines variances from prior and data

# Your turn!

- Suppose that we wish to estimate the mean of a normal distribution with  $\sigma^2=3$ , so  $p(y|\mu) \sim N(\mu, \sigma^2=3)$
- Assume our prior  $p(\mu)$  is  $N(\mu_0=0, \sigma_0^2=1)$ .
- We make one observation:  $y=2$ .
- What is the posterior distribution for  $\mu$  given  $y$ ?
- What if the prior is  $N(2,1)$ ?  $N(0,10)$ ?

# Answers

- Observe  $y = 2$  ;  $p(y|\mu) \sim N(\mu, \sigma^2=3)$
- If prior  $p(\mu) \sim N(\mu_0=0, \sigma_0^2=1)$ :  
posterior mean:  $\mu_1 = (0/1 + 2/3) / (1/1 + 1/3) = 0.50$   
posterior variance:  $1/\sigma_1^2 = 1/1 + 1/3 = 1.333$  so  $\sigma_1^2 = 0.75$
- If prior  $p(\mu) \sim N(\mu_0=2, \sigma_0^2=1)$   
 $\mu_1 = (2/1 + 2/3) / (1/1 + 1/3) = 2$   
 $1/\sigma_1^2 = 1/1 + 1/3 = 1.333$  so  $\sigma_1^2 = 0.75$
- If prior  $p(\mu) \sim N(\mu_0=0, \sigma_0^2=10)$   
 $\mu_1 = (0/10 + 2/3) / (1/10 + 1/3) = 1.54$   
 $1/\sigma_1^2 = 1/10 + 1/3 = 0.4433$  so  $\sigma_1^2 = 2.31$



## Case Study 1: locally-specific, globally-supported risk models

- Risk models based on local data can be unreliable (eg 5000 hospital admissions)
- We can use ‘gold standard’ models as *priors* (eg APACHE, 17000 patients from 40 US hospitals)

# Model

For each of our patients, the outcome (dead/alive) is

$$y_i | \theta_i \sim \text{Bin}(1, \theta_i)$$

$$\text{logit}(\theta_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{29} x_{i29}$$

$$\beta_i \sim \text{N}(\text{mean}, \text{var})$$

←  
APACHE III coefficient

# Results

- Inclusion of local data did not substantially change most coefficients.
- Both models had excellent discrimination based on ROC analysis.
- However, the Bayesian model was better than the APACHE-III model w.r.t calibration to the local data and prediction of deaths.
- Annual updating was easy and effective. Most coefficients were stable over time, but some changed in a consistent pattern.

# Example: Linear model

- **Explanation**

30 rats, weighed weekly for 5 weeks.

Model as random effects linear growth curve.

**Weight  $Y_{ij}$  of rat  $i$  on day  $x_j$**

$x_j =$      **8**     **15**     **22**     **29**     **36**

**Rat 1**                    151    199    246    283    320

**Rat 2**                    145    199    249    293    354

...

**Rat 30**                   153    200    244    286    324

# Model 1 for Rats

- **Model**

$$y_{ij} \sim \text{Normal} ( \alpha_i + \beta_i (x_j - \bar{x}), \sigma_C^2 )$$

- **Priors**

$$\alpha_i \sim \text{Normal} ( \alpha_C, \sigma_\alpha^2 )$$

$$\beta_i \sim \text{Normal} ( \beta_C, \sigma_\beta^2 )$$

$$\alpha_C \sim \text{Normal} ( 0, \text{large var.} )$$

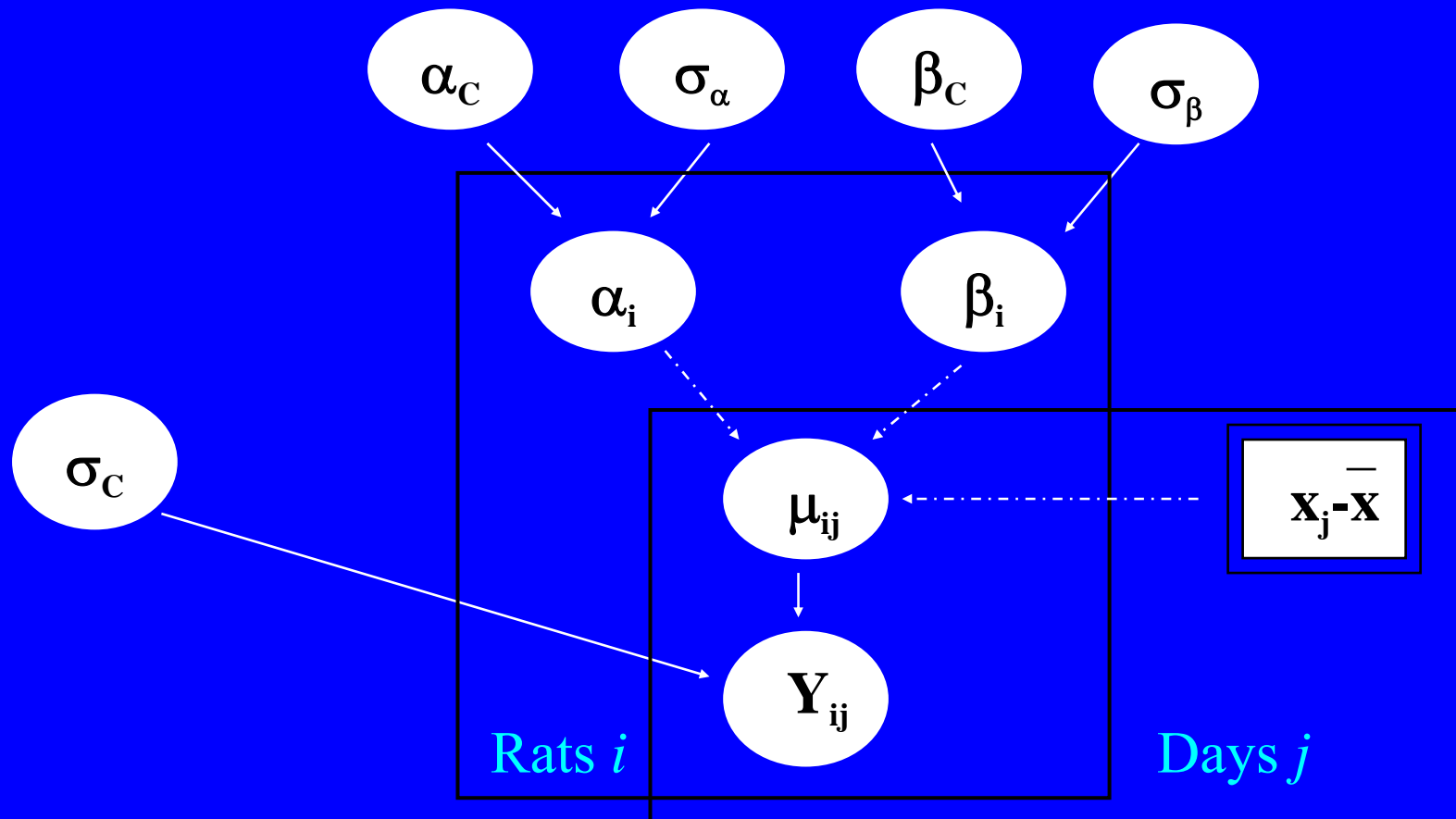
$$\beta_C \sim \text{Normal} ( 0, \text{large var.} )$$

$$\sigma_C \sim \text{Uniform}(0, 100 )$$

$$\sigma_\alpha \sim \text{Uniform}(0, 100 )$$

$$\sigma_\beta \sim \text{Uniform}(0, 100 )$$

# DAG for rats model



# Markov chain Monte Carlo

- Typically our posterior distribution is complex, with many parameters  $(\theta_1, \theta_2, \theta_3, \dots)$ .
- “Decompose” into a sequence of conditional distributions:  
 $(\theta_1 | \theta_2, \theta_3, \dots) \propto \dots$   
 $(\theta_2 | \theta_1, \theta_3, \dots) \propto \dots$
- Simulate from each conditional distribution in turn.
- Advantages:
  1. The conditional distributions are simpler.
  2. We only need to know them as ‘proportional to’.
  3. We can use Markov chain theory to make statements about behaviour and convergence of the chain

# Example

- Consider a single observation  $(y_1, y_2)$  from a bivariate normal population with unknown mean  $(\theta_1, \theta_2)$  and known covariance matrix

$$\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

- Assume uniform priors for  $\theta_1$  and  $\theta_2$ .
- Then  $p(\theta_1, \theta_2 | y)$  is multidimensional, but

$$\theta_1 | \theta_2 \sim \text{Univariate normal}$$

$$\theta_2 | \theta_1 \sim \text{Univariate normal}$$

# Gibbs sampler

Sample from

$$\theta_1 \mid \theta_2, y \sim \text{N}(y_1 + \rho(\theta_2 - y_2), 1 - \rho^2)$$

Then sample from

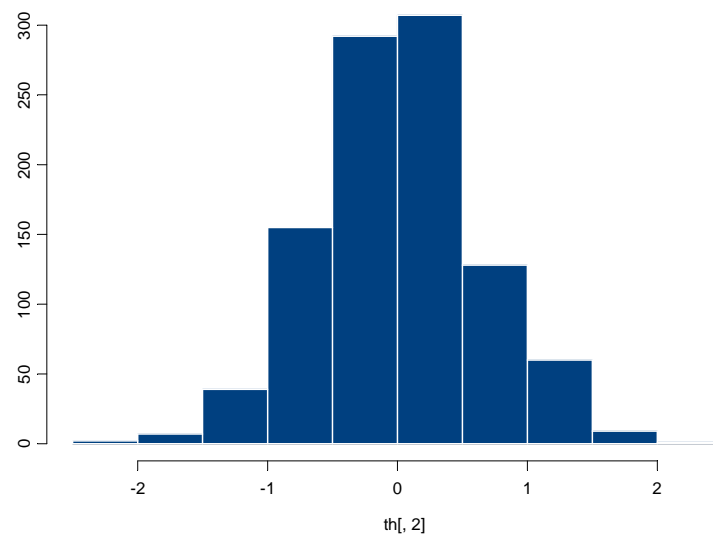
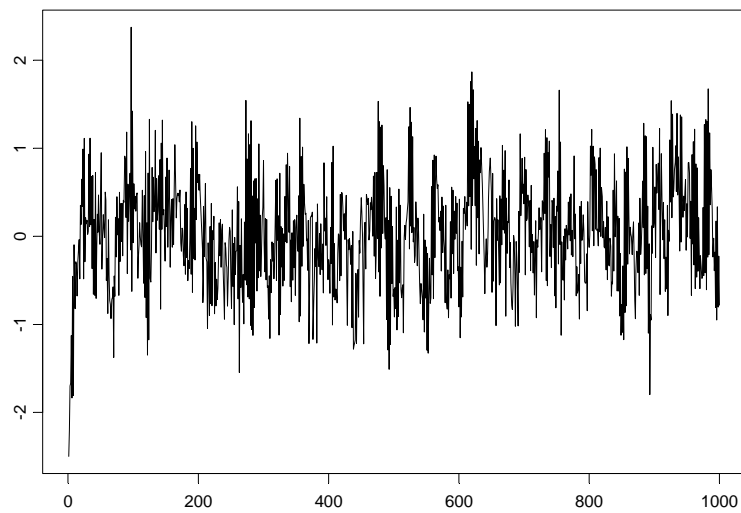
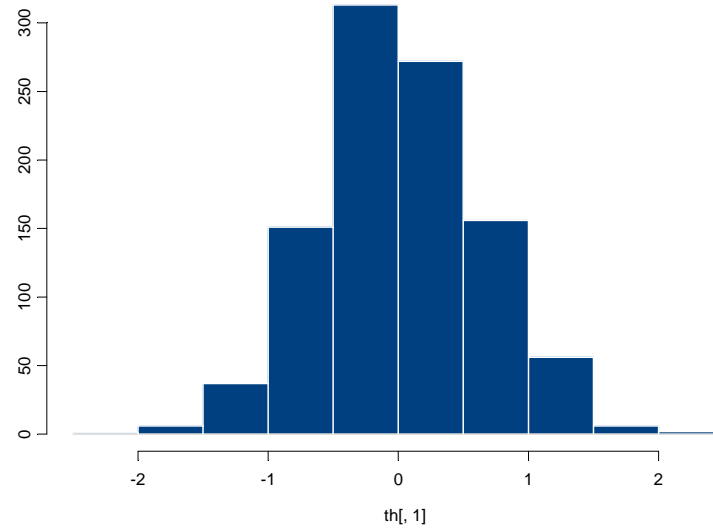
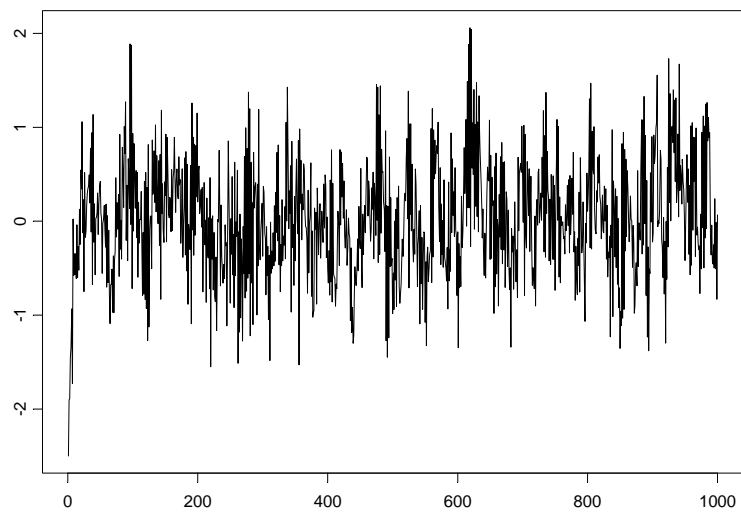
$$\theta_2 \mid \theta_1, y \sim \text{N}(y_2 + \rho(\theta_1 - y_1), 1 - \rho^2)$$

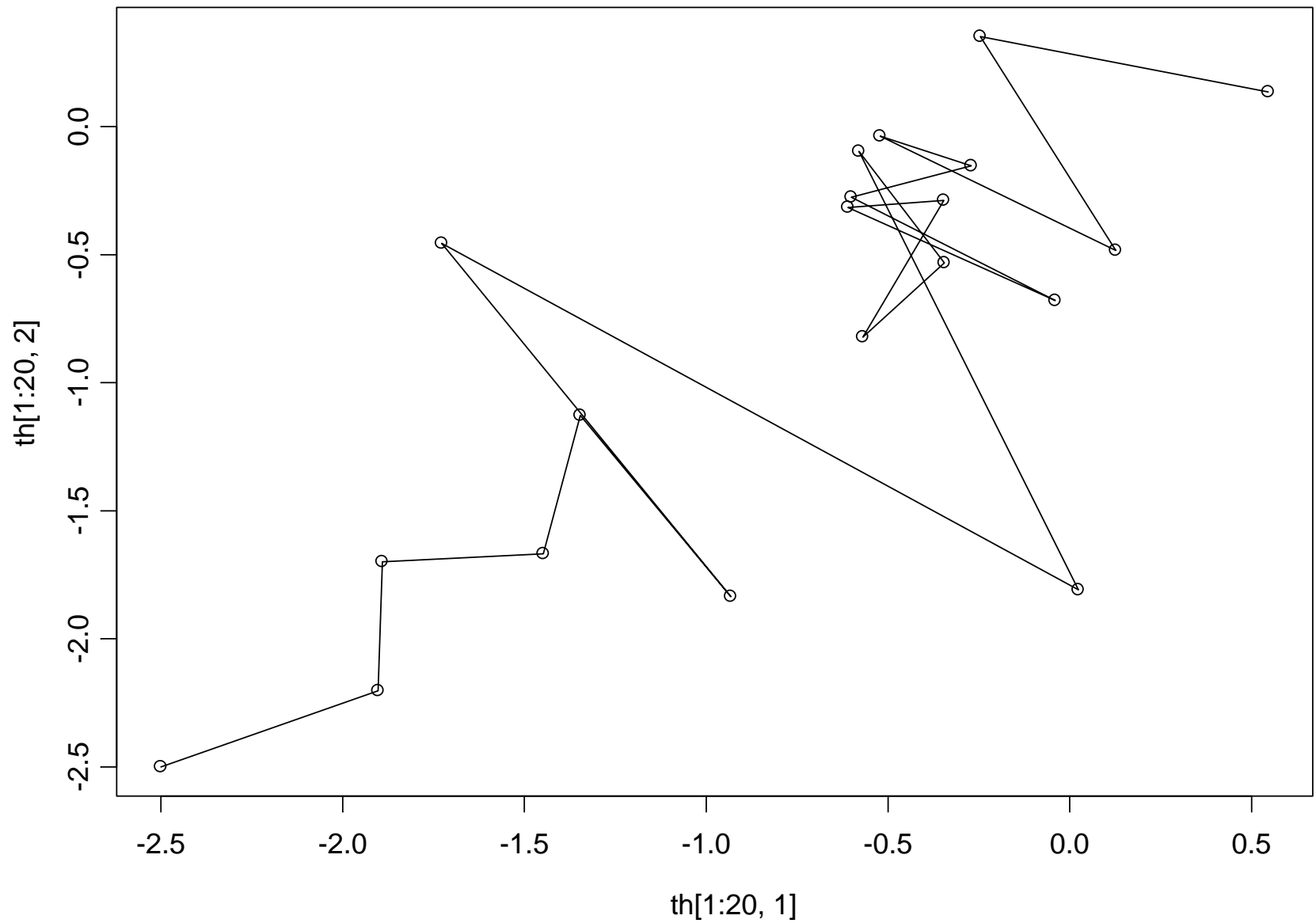
# Example

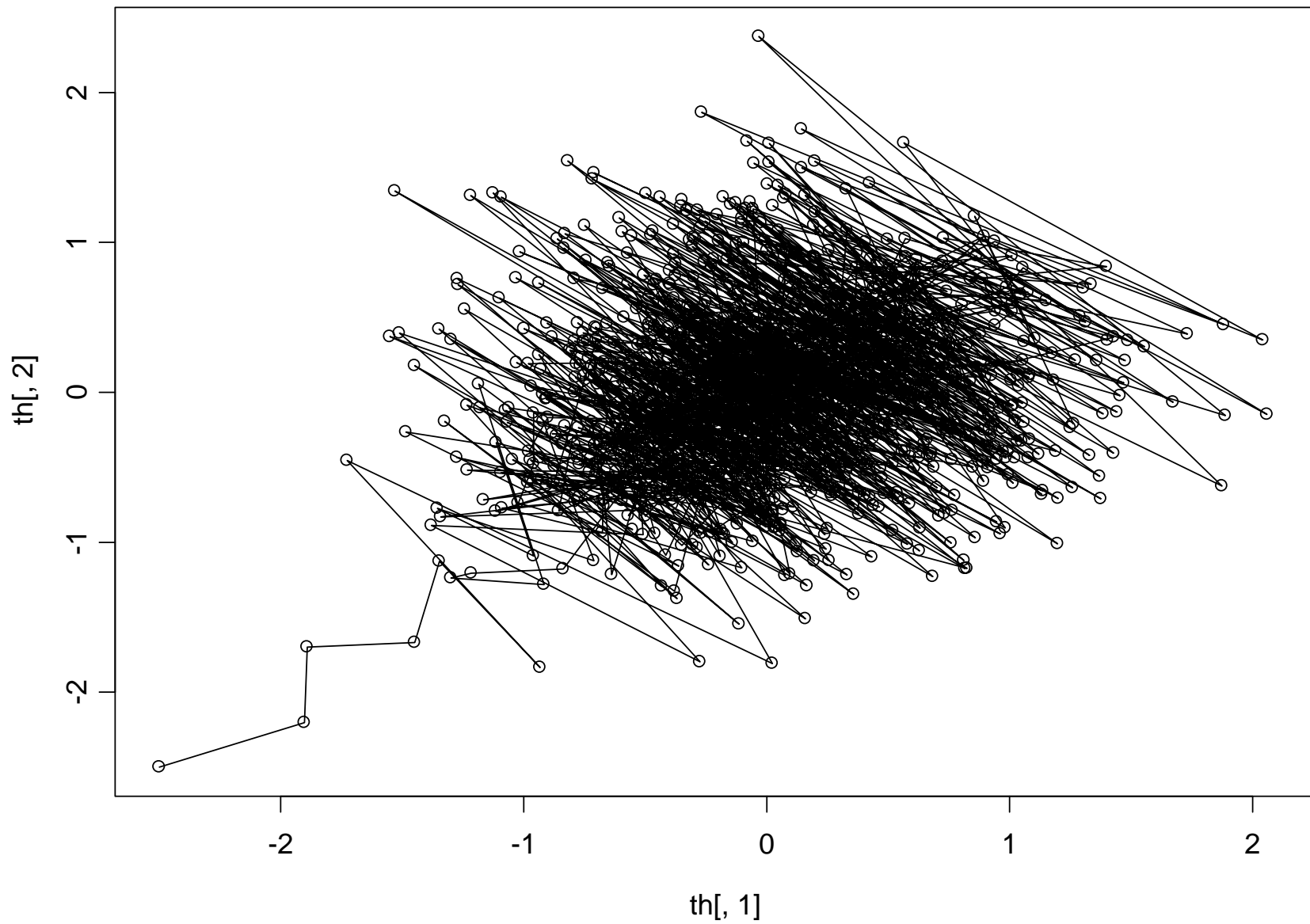
$\rho = 0.8$

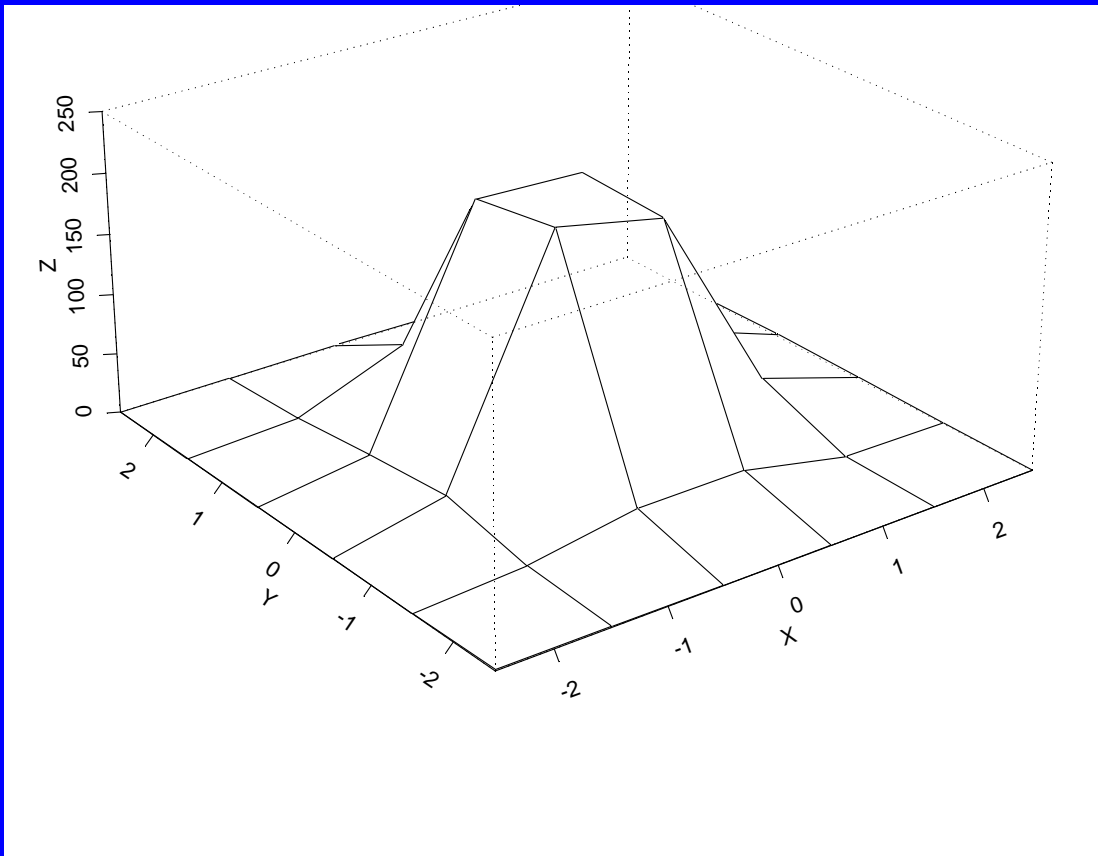
Data:  $(y_1, y_2) = (0,0)$

|       | theta1      | theta2      |
|-------|-------------|-------------|
| [1,]  | -2.50000000 | -2.50000000 |
| [2,]  | -1.90294512 | -2.20308841 |
| [3,]  | -1.89128340 | -1.69893746 |
| [4,]  | -1.44797468 | -1.66859001 |
| [5,]  | -1.34664264 | -1.12758074 |
| [6,]  | -0.93261816 | -1.83510560 |
| [7,]  | -0.72743540 | -0.45567164 |
| [8,]  | 0.02365003  | -1.80804471 |
| [9,]  | -0.57995614 | -0.09685846 |
| [10,] | -0.34490984 | -0.53282284 |









### Posterior estimates:

| parameter  | mean   | s.d.  |
|------------|--------|-------|
| $\theta_1$ | 0.002  | 0.003 |
| $\theta_2$ | -0.001 | 0.002 |

# Your turn!

- Implement this Gibbs sampler
  - By hand
  - Using Excel

# WinBUGS

## Bayesian inference Using Gibbs Sampling

- Can be used directly
- Can be used with other packages, eg R, Matlab
- Can be developed using OpenBUGS

## Alternatives:

- Other software packages eg MLWin, SAS
- Own code (in R, Matlab, C etc)

# Recall the rats model

- **Model**

$$y_{ij} \sim \text{Normal} ( \alpha_i + \beta_i (x_j - \bar{x}), \sigma_C^2 )$$

- **Priors**

$$\alpha_i \sim \text{Normal} ( \alpha_C, \sigma_\alpha^2 )$$

$$\beta_i \sim \text{Normal} ( \beta_C, \sigma_\beta^2 )$$

$$\alpha_C \sim \text{Normal} ( 0, \text{large var.} )$$

$$\beta_C \sim \text{Normal} ( 0, \text{large var.} )$$

$$\sigma_C \sim \text{Uniform}(0, 100 )$$

$$\sigma_\alpha \sim \text{Uniform}(0, 100 )$$

$$\sigma_\beta \sim \text{Uniform}(0, 100 )$$

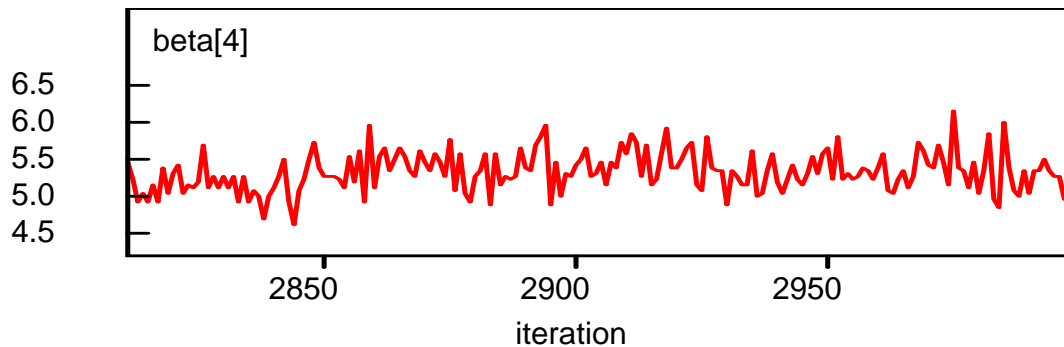
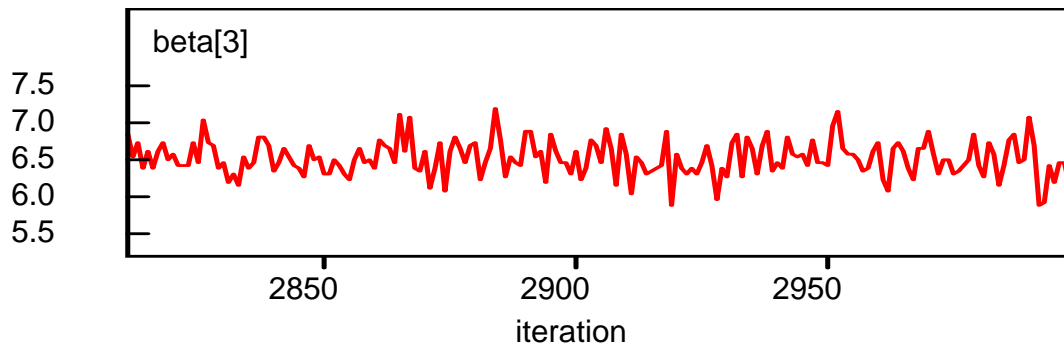
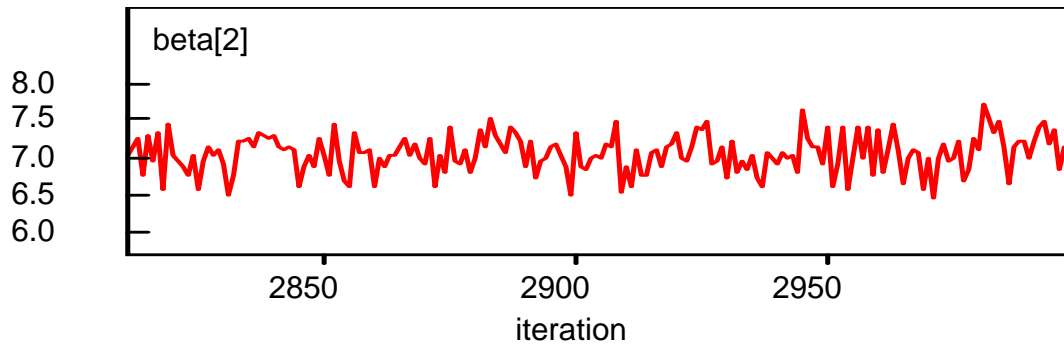
# WinBUGS code for rats

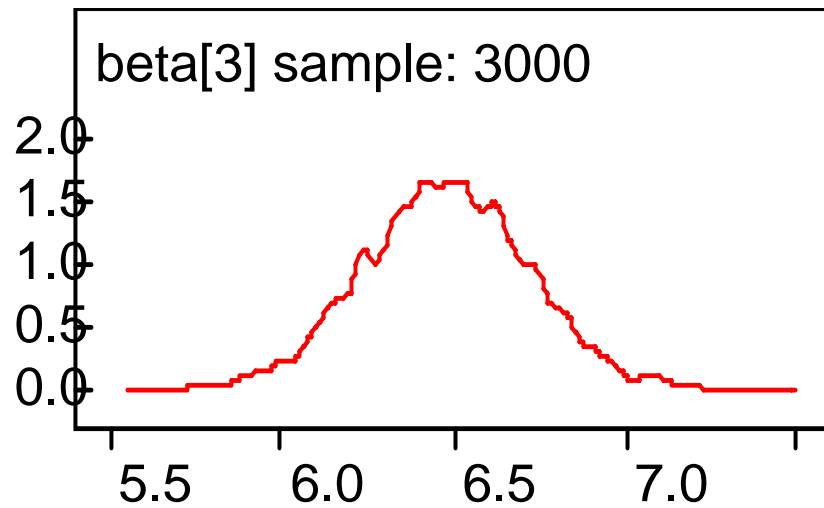
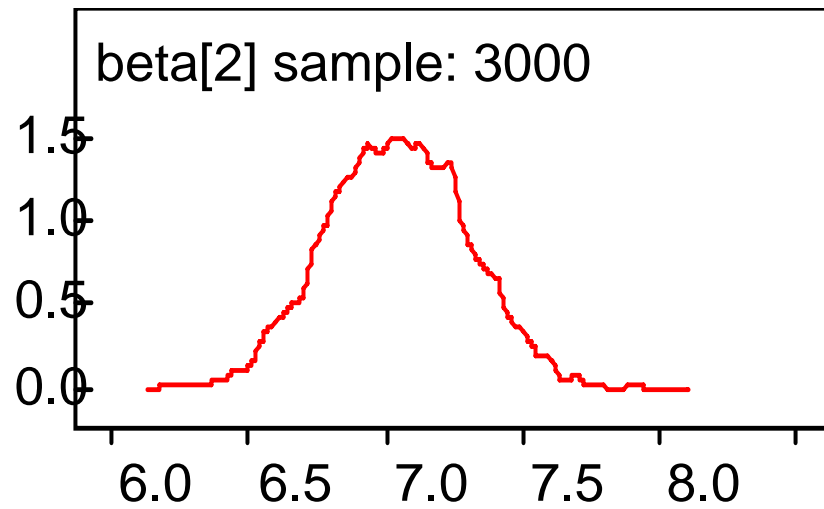
N = no. rats, T = no. time periods

```
model{
  for (i in 1:N) {
    for (j in 1:T) {
      mu[i,j] <- alpha[i] + beta[i] * (x[j] - x.bar)
      Y[i,j] ~ dnorm(mu[i,j], tau.y)
    }
    alpha[i] ~ dnorm(mu.a, tau.a)
    beta[i] ~ dnorm(mu.b, tau.b)}
  mu.a ~ dnorm(0, 1.0E-4)
  mu.b ~ dnorm(0, 1.0E-4)
  tau.y <- 1/(sig.y*sig.y)
  sig.y ~ dunif(0,100)
  tau.a <- 1/(sig.a*sig.a)
  sig.a ~ dunif(0,100)
  tau.b <- 1/(sig.b*sig.b)
  sig.b ~ dunif(0,100)
  x.bar <- mean( x[] )
}
```

WinBUGS codes  
normal distributions as  
(mean, precision)

# Trace plots for rats





# Demonstration of WinBUGS

# Example: Meta-analysis

22 trials of beta blockers to prevent mortality after heart attack

| Study | Mortality: deaths / total |                 |
|-------|---------------------------|-----------------|
|       | <i>Treated</i>            | <i>Control</i>  |
|       | $r_i^T / n_i^T$           | $r_i^C / n_i^C$ |
| 1     | 3/38                      | 3/39            |
| 2     | 7/114                     | 14/116          |
| 3     | 5/69                      | 11/93           |
| 4     | 102/1533                  | 127/1520        |
| ..... |                           |                 |
| 20    | 32/209                    | 40/218          |
| 21    | 27/391                    | 43/364          |
| 22    | 22/680                    | 39/674          |

We want to:

- Estimate the true effect (on a log-odds scale) in each trial:  $\delta_i$
- Estimate the overall effect:  $d$
- Predict the effect in a new trial  $\delta.\text{new}$

**Model:**

$$r_i^C \sim \text{Binomial}(p_i^C, n_i^C)$$

$$r_i^T \sim \text{Binomial}(p_i^T, n_i^T)$$

$$\text{logit}(p_i^C) = \mu_i$$

$$\text{logit}(p_i^T) = \mu_i + \delta_i$$

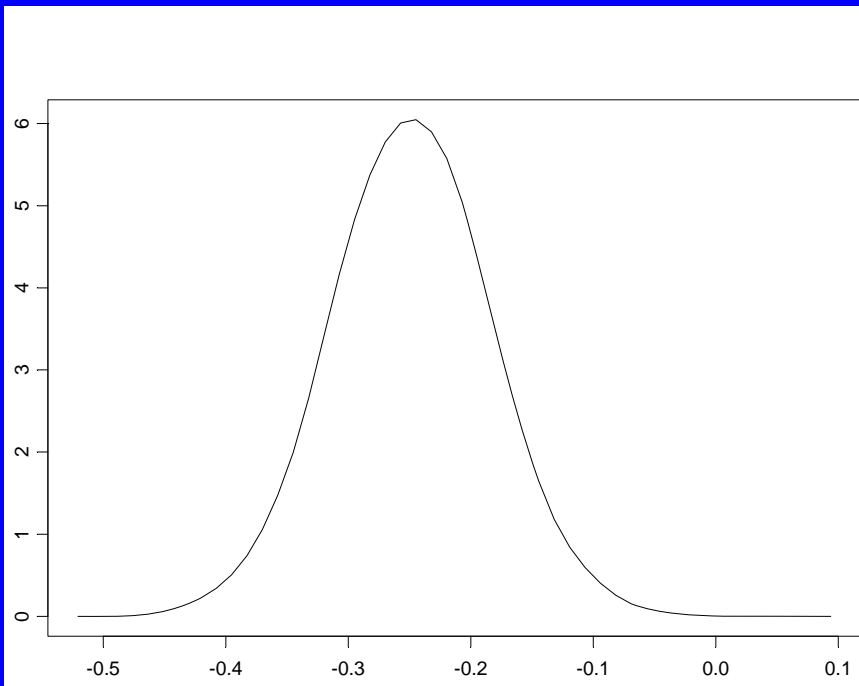
$$\delta_i \sim \text{Normal}(d, \text{tau})$$

$$\delta.\text{new} \sim \text{Normal}(d, \text{tau})$$

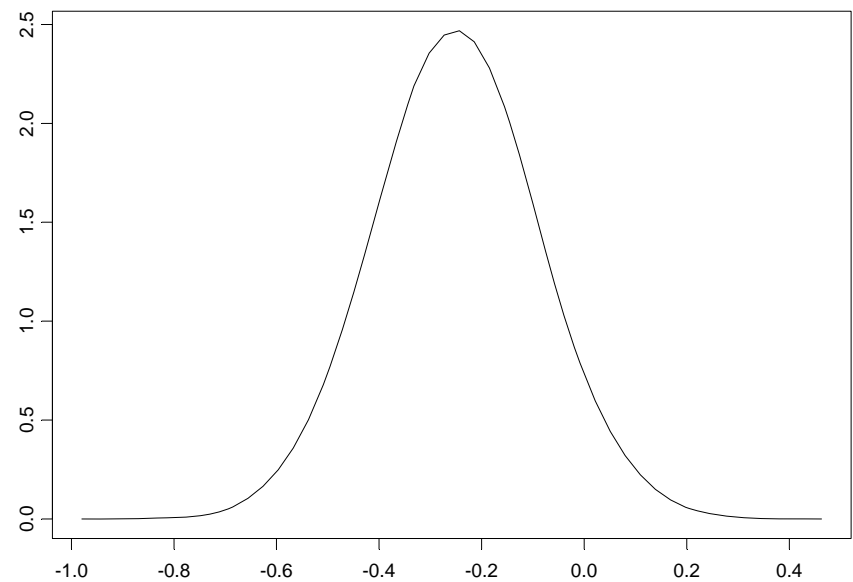
# WinBUGS code

```
model
{
  for( i in 1 : Num ) {
    rc[i] ~ dbin(pc[i], nc[i])
    rt[i] ~ dbin(pt[i], nt[i])
    logit(pc[i]) <- mu[i]
    logit(pt[i]) <- mu[i] + delta[i]
    mu[i] ~ dnorm(0.0, 1.0E-5)
    delta[i] ~ dnorm(d, tau)
  }
  d ~ dnorm(0.0, 1.0E-6)
  tau <- 1/(sigma*sigma)
  sigma ~ dunif(0, 10)
  delta.new ~ dnorm(d, tau)
}
```

| <b>node</b> | <b>mean</b> | <b>sd</b> | <b>MC error</b> | <b>2.5%</b> | <b>median</b> | <b>97.5%</b> |
|-------------|-------------|-----------|-----------------|-------------|---------------|--------------|
| d           | -0.2489     | 0.06282   | 0.002297        | -0.3734     | -0.248        | -0.1239      |
| delta.new   | -0.2496     | 0.1576    | 0.002582        | -0.5773     | -0.2514       | 0.07974      |
| sigma       | 0.1243      | 0.06834   | 0.002835        | 0.02878     | 0.1142        | 0.2796       |

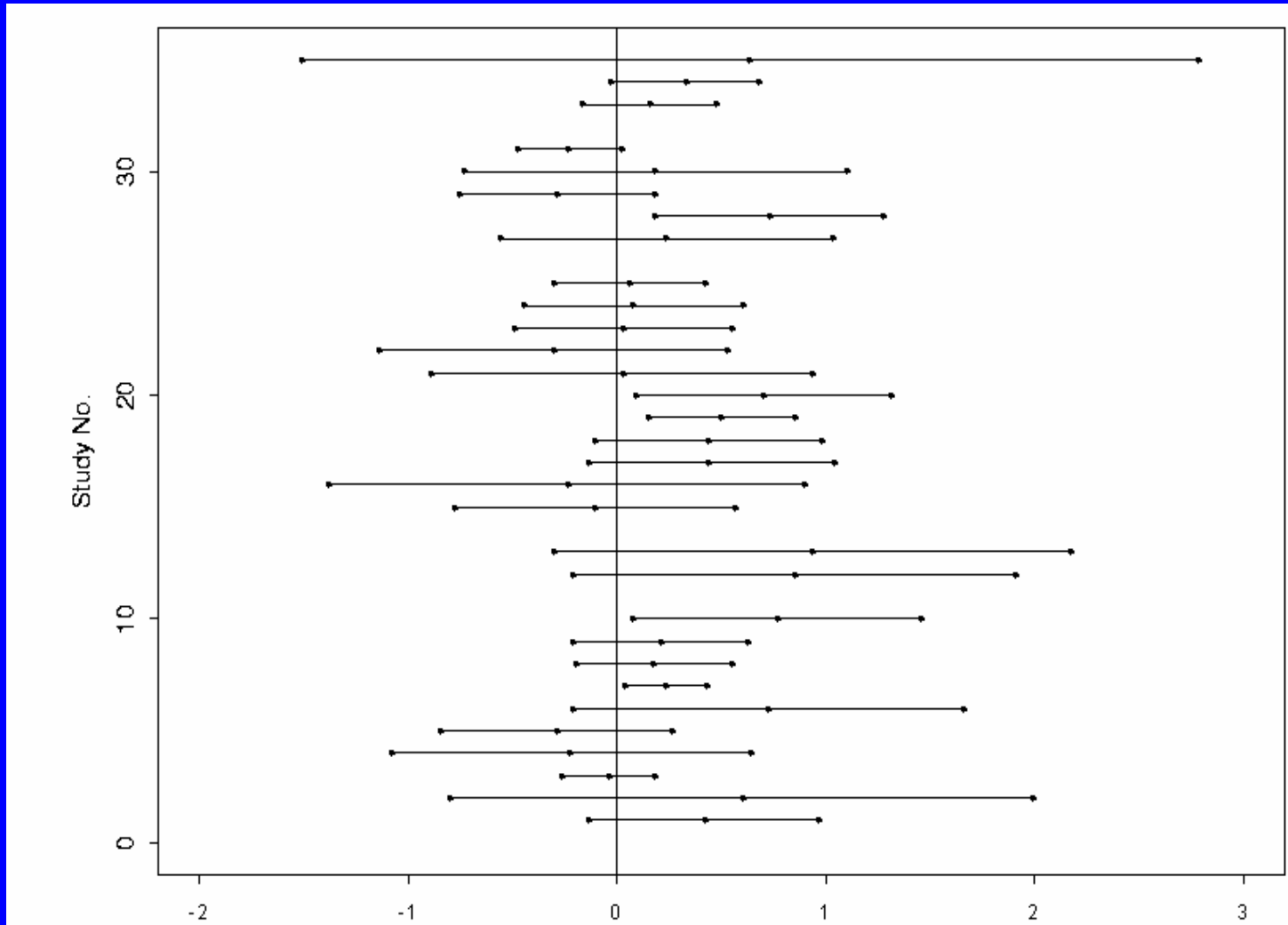


Posterior density for  $d$



Posterior density for  $\delta$ .new

# Case study 2: Quality-adjusted meta-analysis



Work  
with  
Robert  
Wolpert

log RR of lung cancer associated with passive smoking

# Adjusting for study quality

**We want: overall and study-specific true RR**

**We observe: study-specific data**

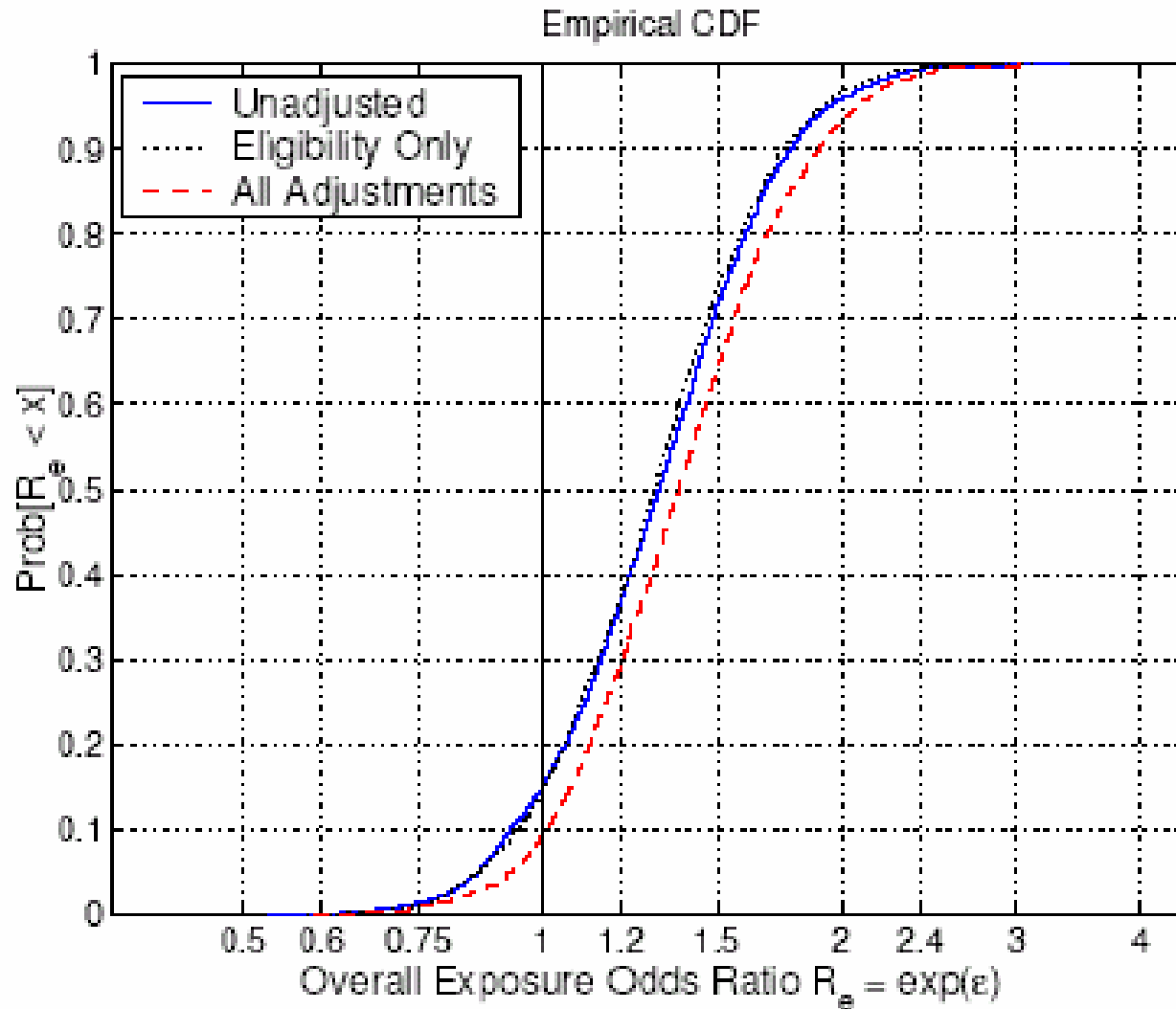
**We estimate: apparent classification probabilities**

**We need: true classification probabilities**

# Prior information about misclassification and bias

1. **Misclassification of active smoking**  
About 5% lie, 5% 'forget', 10% poorly asked
2. **Misclassification of exposure**  
'47% of currently nonsmoking wives have <1 hr/day exposure at home'  
'40-50% women with nonsmoking spouses have significant ETS exposure outside the home'
3. **Misclassification of lung cancer**  
30-40% of lung cancers seen at autopsy are missed clinically

# Overall effect of adjustment



# GeoBUGS

GeoBUGS is an add-on module to WinBUGS which provides an interface for:

- producing maps of the output from disease mapping and other spatial models
- creating and manipulating adjacency matrices that are required as input for the conditional autoregressive models (CAR) available in WinBUGS for carrying out spatial smoothing.

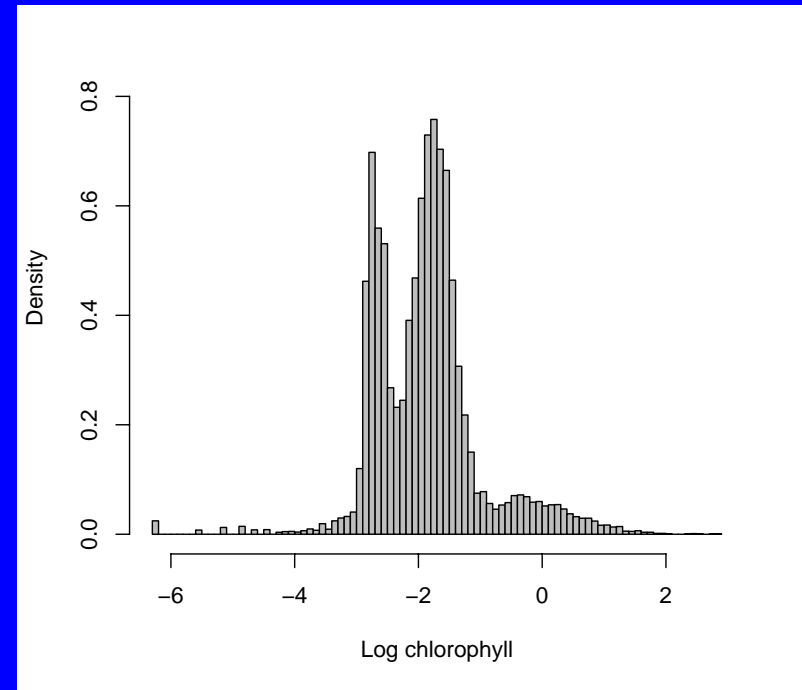
# Case Study 3: Monitoring water quality

C. Alston QUT, B. Farthing EPA, A. Steven CSIRO



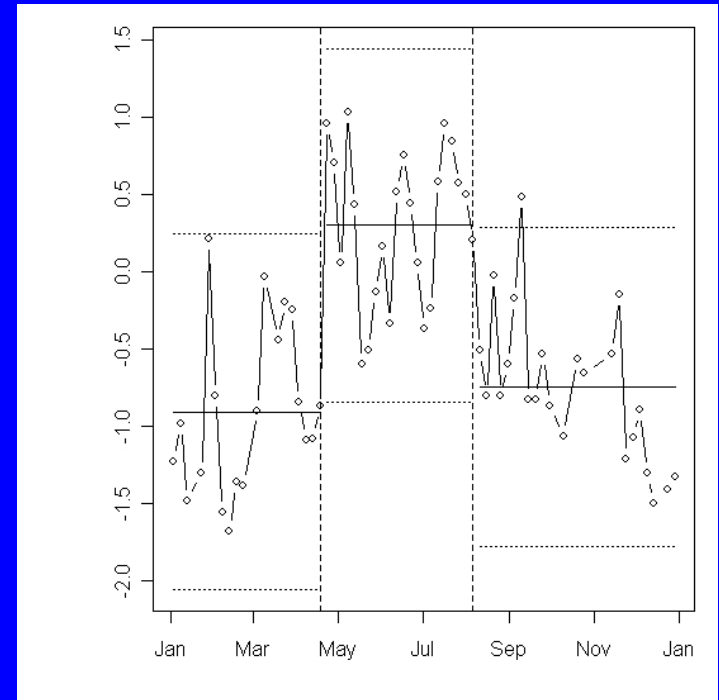
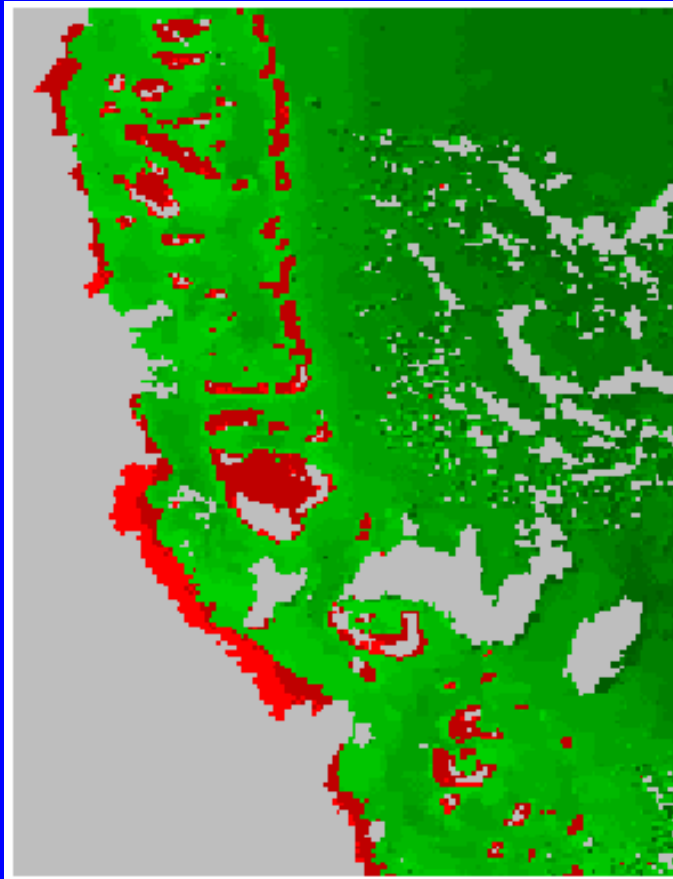
Grey = land & cloud cover

Green = chlorophyll level  
lightest = highest reading



Gaussian spatial  
mixture model to  
represent homogeneous  
regions

# From mixtures to monitoring



Tailored CUSUM and Shewart charts:

- Mixtures give regions
- Changepoint model gives season (wet/dry)

$$Y = \mu + \beta_{1,2} * \text{day}$$

# MCMC Convergence

- Theoretical results
- Empirical diagnostics
  - A wealth of options!
  - Some embedded in software (WinBUGS, CODA)

# Constructing Prior Distributions

- ? Non-informative (“let the data talk”)
- ? Vague (wide variances)
- ? Vaguely informative (“reasonable” bounds)
- ? Subjective (from other studies, experts, etc)

# Take care with ‘noninformative’ priors

- **Example:** regression model with 4 variables.

Specifying a uniform prior on the model:

$[], [1], [2], [3], [4], [12], [13], [14], [23], [24], [34], [123], [124], [134], [234], [1234]$

assigns prior probability  $6/16$  to 2-variable models  
and prior probability  $4/16$  to 3-variable models

- **Jeffreys prior** corresponds to the expected Fisher Information. All parametrizations lead to the same prior.

Example:  $\text{Beta}(\frac{1}{2}, \frac{1}{2})$  prior for binomial parameter.

# Subjective Priors

- On what information are they based?
- If based on experts, how is information ‘elicited’?
- What does the posterior estimate ‘mean’?
- How influential is the prior?

# ELICITOR (Mary Kynn)

*Interactive Software for Eliciting Informative Priors  
for a Bayesian Logistic Regression Model*

## Quick Links

[About](#)

[How it works](#)

[Download](#)

[Research](#)

[Links](#)

[Contacts](#)

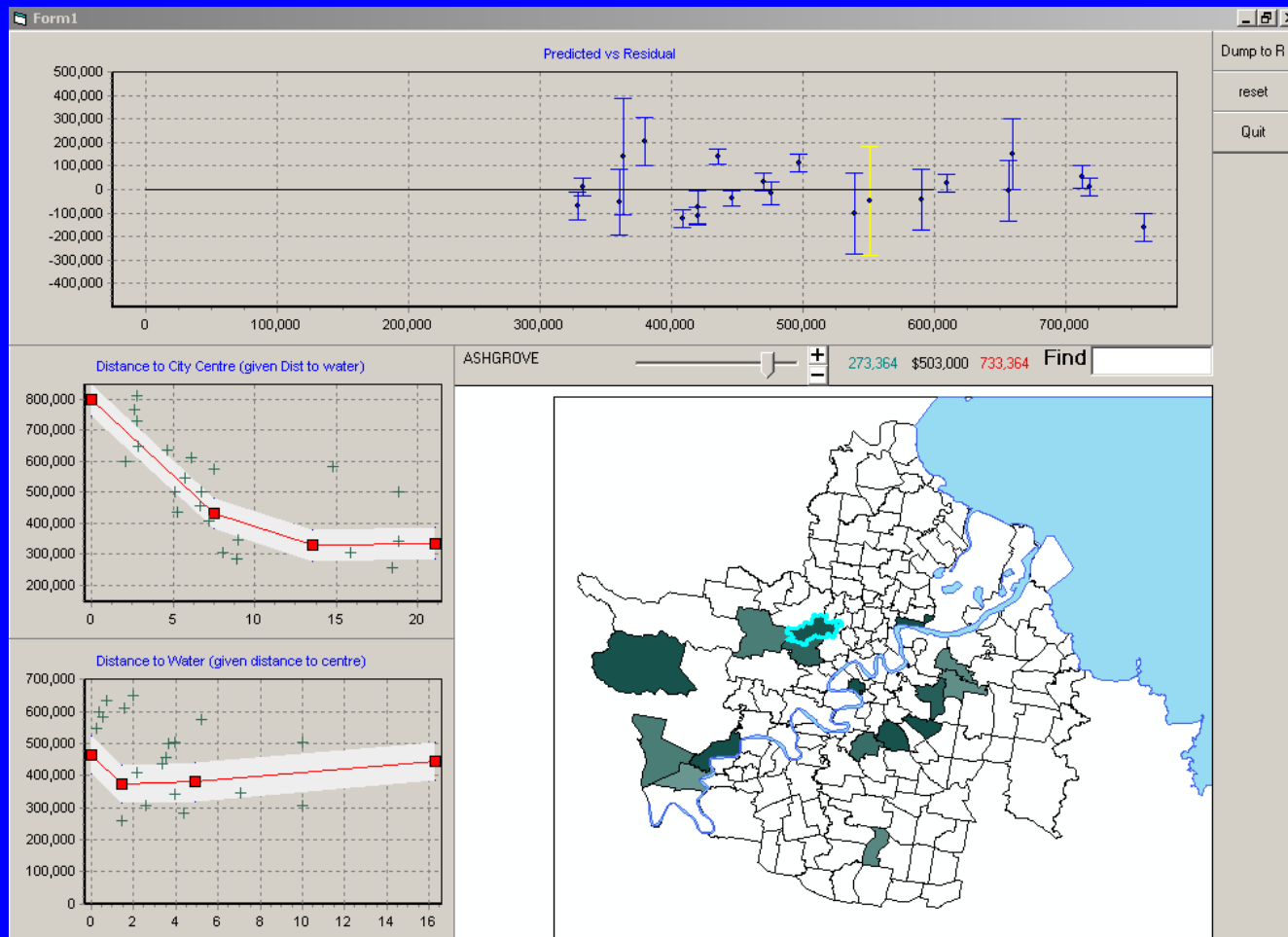


The screenshot shows the Elicitor software interface. On the left, there is a sidebar with a 'Quick Links' menu containing 'About', 'How it works', 'Download', 'Research', 'Links', and 'Contacts'. The main area displays a network diagram with nodes and edges, overlaid on a background of a water droplet on a green leaf. The word 'ELICITOR' is prominently displayed in large, bold, black letters across the center of the interface. In the top right corner, there is a table with numerical values: -0.3, 0.55, 1.66, and 2.77. Below the table, the text '= + FLOCK' is visible.

# ELICITOR

# Geographically assisted elicitation

(Robert Denham, Sama Low Choy)



# Model checking

- We are not interested in “Is the model true?” but “Do model deficiencies affect substantive inferences?”
  - Compare observed statistics with values predicted under the model
  - Compare observed data with replicated data  
If the model is adequate, replicated data generated under the model should look similar to the observed data

# Model Comparison

- Bayes factors, posterior odds, BIC, DIC
- Reversible jump MCMC,  
Birth and death MCMC
- Model averaging

# Advantages of Bayesian Analysis

- Allows modelling of ‘the real problem’
- Modular: complex systems composed of simpler parts
- Incorporates uncertainty
- Allows dynamic updating
- Uses all information
  - Outcomes based on data and priors
  - Allows broad engagement
- Provides probabilistic information about the parameters themselves

# Caveats

- Bayesian modelling is *not* a panacea for crappy data: ‘garbage in, garbage out’ still holds.
- Bayesian models are *not* built overnight. They require care with planning, inputs, sensitivity, outputs.
- Bayesian models *do* depend on prior information: this is good and bad. (Be careful that they are not simply self-fulfilling or a replacement for quality data.)
- It is important to understand what the models (and software) do and can do!

# Finally ...

*A wonderful thing is a Bayesian  
Who can model in more than one wayesian  
She can alter your prior  
to suit your desire  
And thereby meet your persuasian*