

Reversible jump: Application to genetic association studies

Dave Lunn

Dept. Epidemiology & Public Health, Imperial College, London

IceBUGS Workshop: February 11–12, 2006, Hanko, Finland

- Association studies
- Variable selection
- Automatic curve fitting
- Categorical data

Motivation: association studies

- Suppose we have phenotypes y_i measured on $i = 1, \dots, N$ individuals (y_i may be multivariate)
- Also observe each individual's genotype at Q SNP loci:

$$g_{ij} \in \{0, 1, 2\}, \quad i = 1, \dots, N, \quad j = 1, \dots, Q$$

- What is the 'best' set of genetic predictors for \mathbf{y} ?
- Forward/backward selection unreliable
- Instead: let set of predictors be a parameter of the model, θ , say
 \Rightarrow "variable selection model"
- We don't know (a priori) how many predictors form the 'best' model
 $\Rightarrow \theta$ has unknown dimension
- *Reversible jump* algorithm recently implemented in WinBUGS for handling certain types of "trans-dimensional" model

Variable selection

- Let X denote $N \times C$ matrix of *all* potential predictors (arranged in columns)
- For example:
 - * $X_{.j} = g_{.j}, \quad j = 1, \dots, Q \quad (C = Q)$
 - * $X_{.j} = I(g_{.j} = 0), \quad X_{.j+Q} = I(g_{.j} = 2), \quad j = 1, \dots, Q \quad (C = 2Q)$

- Suppose phenotypes measured on continuous scale:

$$y_i \sim \text{Normal}(\mu_i, \sigma^2), \quad i = 1, \dots, N$$

- Variable selection model:

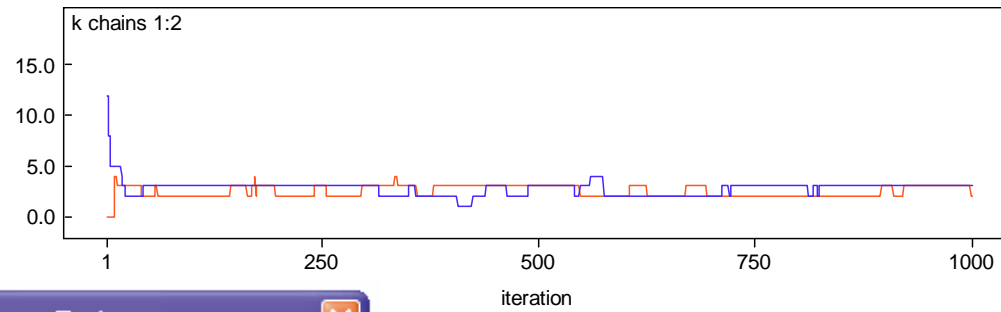
$$\mu_i = W_i \beta, \quad W_i = \left(1, X_{i\theta_1}, X_{i\theta_2}, \dots, X_{i\theta_k}\right)$$

- k = number of currently selected predictors
- θ = vector of k selected columns of X
- β = vector of $k + 1$ regression coefficients

WinBUGS code

```
model {  
  for (i in 1:N) {  
    y[i] ~ dnorm(mu[i], tau)  
    for (j in 1:Q) {  
      X[i, j] <- equals(g[i, j], 0)  
      X[i, (j + Q)] <- equals(g[i, j], 2)  
    }  
  }  
  
  mu[1:N] <- jump.lin.pred(X[1:N, 1:C], k, tau.beta)  
  id <- jump.model.id(mu[1:N])  
  k ~ dbin(0.5, C)  
  tau ~ dgamma(0.001, 0.001)  
}
```

Jump output

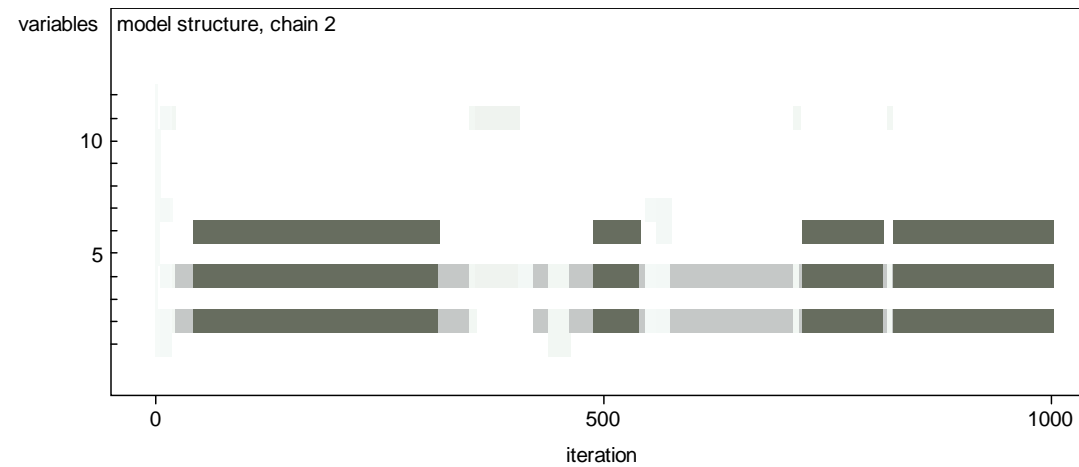
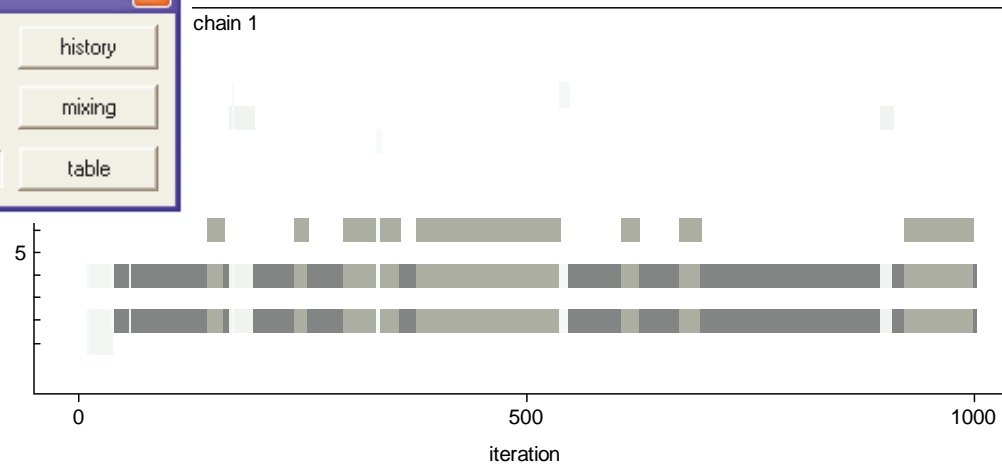


Jump Summary Tool

id node:

beg:

end: razor:



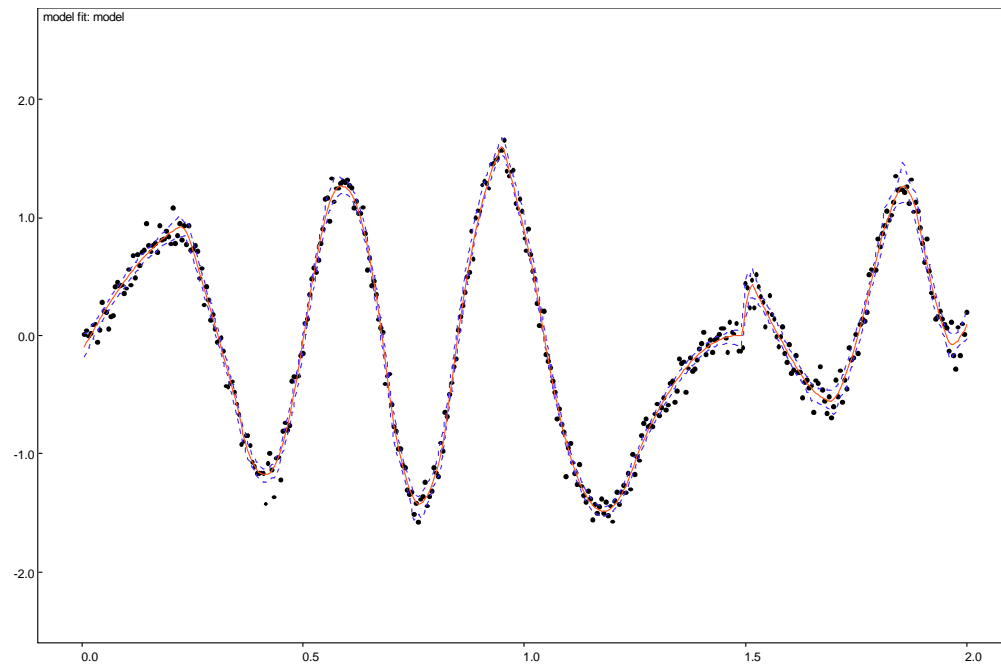
Jump output continued

model structure	posterior prob.	cumulative prob.
010101000000	0.4825	0.4825
010100000000	0.3905	0.873
010100000010	0.0315	0.9045
110100000000	0.0245	0.929
000100000010	0.0245	0.9535

variable no.	marginal prob.
1	0.034
2	0.9625
3	0.001
4	0.994
5	0.002
6	0.4945
7	0.0245
8	0.0025
9	0.0025
10	0.0045
11	0.0655
12	0.0065

Additional covariates

- What if we have additional covariates?
 - Can incorporate into variable selection model
 - Can incorporate into additional variable selection model
 - Can incorporate into fixed part of model
 - * Can control for arbitrary relationships using *automatic splines*:



Automatic curve fitting

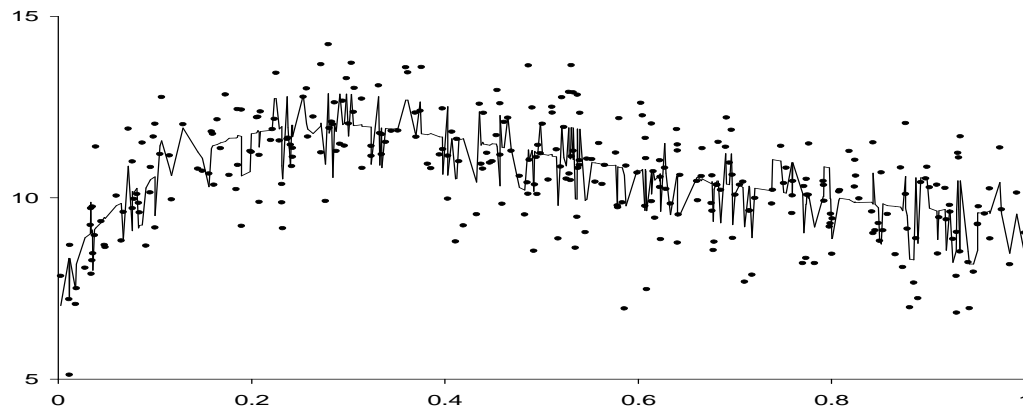
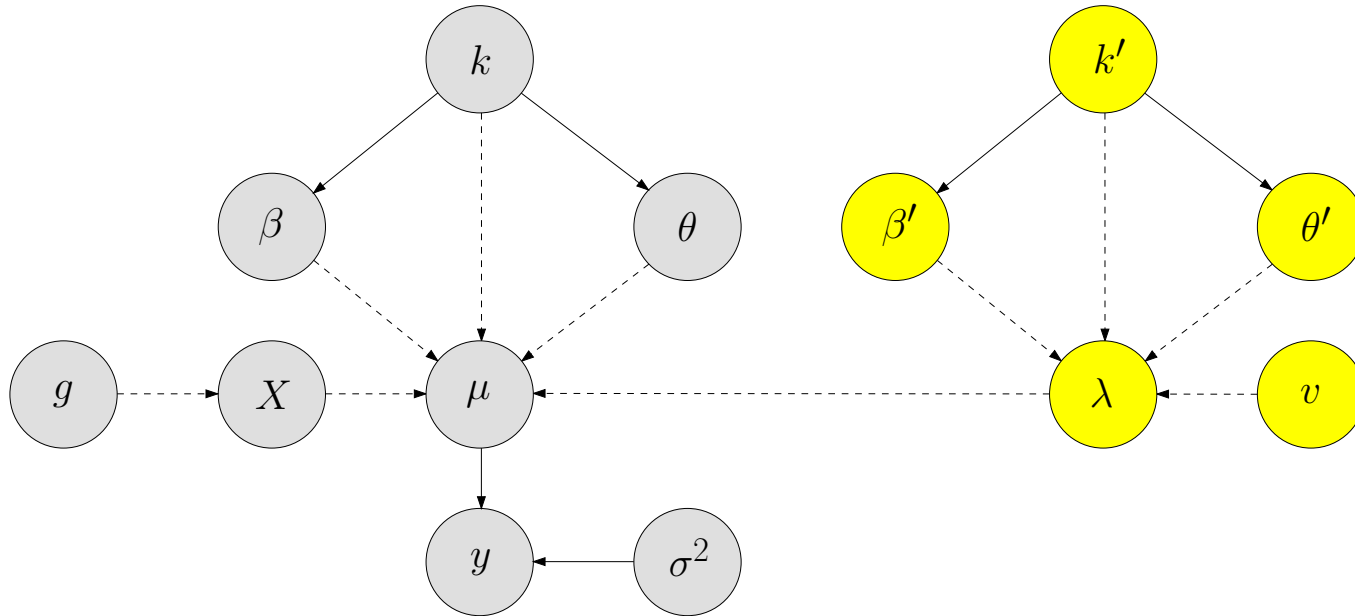
- Use quadratic spline, for example

$$\mu_i = W_i\beta + \lambda_i$$

$$\lambda_i = \beta'_1 + \beta'_2(v_i - v_0) + \beta'_3(v_i - v_0)^2 + \sum_{n=1}^{k'} \beta'_{n+3}(v_i - \theta'_n)^2$$

```
for (i in 1:N) {mu[i] <- select[i] + lambda[i]}
select[1:N] <- jump.lin.pred(X[1:N, 1:C], k, tau.beta)
lambda[1:N] <- jump.pw.poly.c.quad(v[1:N], k.prime,
                                tau.beta, v0, max)
```

Combining models



Binary and other data types

- Natural model for binary data:

$$y_i \sim \text{Bernoulli}(p_i), \quad \ell(p_i) = \mu_i, \quad i = 1, \dots, N$$

- But, current implementation of reversible jump requires conjugacy for regression coefficients $\beta \Rightarrow$ likelihood must be normal or Student-t
- Consider

$$y_i \sim \text{Bernoulli}(p_i), \quad p_i = I(z_i \geq 0), \quad z_i \sim \text{Normal}(\mu_i, 1), \quad i = 1, \dots, N$$

$$\Rightarrow \Pr(y_i = 1) = \int_0^\infty p(z_i | \mu_i) dz_i = \Phi(\mu_i) = \text{probit}^{-1}(\mu_i)$$

- \therefore equivalent but the latter provides a normal likelihood for β
- Specify via `y[i] ~ dbern.aux(z[i])`
- Also 'dcat.aux(.)' and 'dgene.aux(.)'