The BUGS Project
version 0.5 manual

Next Previous Contents
Next: Acknowledgements **Previous:** 8 Errors

9 Topics in modelling

We shall examine some general issues in Bayesian analysis and associated implementation in BUGS.

9.1 General Strategy

We are unwilling, and unable, to recommend any fixed strategy for analysing data. However, some general points are as follows.

- Start with small simple models which have been used in other software and for which you know the answers.
- Develop more complex models incrementally.
- Check final answers by starting from different initial values and a different random number seed, running for long periods, and using different parameterisations.
- Perform a few updates before plunging into long runs in order to assess timings and examine ballpark results.

For a collection of case-studies and practical advice on using MCMC methods, see Gilks *et al.* (1995).

9.2 `Non-informative priors'

9.2.1 Issues with improper priors

The appropriate specification of `non-informative' priors is an old problem in Bayesian statistics, and is particularly important when techniques are to be used in a scientific context in which an `objective' inference is required. BUGS requires that a full probability model is defined, and hence forces all priors to be proper, and the problem arises of specifying appropriate priors on `founder' nodes (those with no parents) in the graph.

It is, however, vital to distinguish parameters of primary interest from those which specify secondary structure for the model. The former will generally be location parameters, such as regression coefficients, and in many cases a normal prior with an extremely small precision (large variance) is adequate. Our examples often feature a choice $\beta \sim \mathrm{Normal}(0.0, 0.0001)$, corresponding to a standard deviation of 100, so that a 95% prior interval is around ± 200, and the prior will be locally uniform over the region supported by the likelihood. Alternatively a uniform prior on a suitably wide interval could be given , such as $\beta \sim \mathrm{Unif}(-200, 200)$

Secondary aspects of a model include, say, the precision of random effects in a hierarchical model. More care is required in setting such parameters and we recommend thinking carefully about reasonable values for such parameters in advance and so specifying fairly informative priors wherever possible - the inclusion of such

external information is unlikely to bias the scientific conclusions of a study, although some sensitivity analysis should be performed to reassure consumers.

A particular, but insufficiently recognised, problem occurs with the precision parameter in a random effects model [DuMouchel, 1990, DuMouchel and Waternaux, 1992]. Let us call this parameter τ, with inverse σ^2, the variance of the random effects. The standard `non-informative' prior would generally be considered to be $p(\sigma^2) \propto 1/\sigma^2$, equivalent to $p(\tau) \propto 1/\tau$, which arises from assuming the log of the scale parameter has a uniform distribution on $(-\infty, \infty)$.

However, the boundary value $\sigma^2 = 0$ is supported by non-negligible likelihood since it is theoretically possible that there are no subject-specific random effects and, remarkably, the improper prior $p(\sigma^2) \propto 1/\sigma^2$ goes to infinity at zero at a rate sufficient to lead to a *posterior* distribution that is also improper. Assuming a prior that is `only just' proper, such $\tau \sim \mathrm{Gamma}(\epsilon, \epsilon)$ with mean 1 and variance $1/\epsilon$, will formally avoid strict impropriety: such priors are investigated below.

9.2.2 Pareto priors for precision parameters

An alternative approach is to choose an improper prior such as $p(\sigma^2)$ being locally uniform on $(0, \infty)$ [Gelman and Rubin, 1992, Carlin, 1992], which does not lead to an improper posterior. Keeping within a general class of proper priors, we assume we would like a prior on σ^k that is uniform on a range $(0, r)$. This implies that $\tau = \sigma^{-2}$ has a *Pareto* distribution with parameters $(k/2, r^{-2/k})$, where the Pareto distribution (Section 4.4) has the form

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

Thus a uniform prior for σ or σ^2 on $(0, r)$ would respectively give rise to a Pareto $(\frac{1}{2}, r^{-2})$ or a Pareto $(1, r^{-1})$ distribution on τ.

Assuming τ is the precision parameter for a normal distribution gives a likelihood contribution for τ from *n* observations proportional to $\tau^{n/2} e^{-\tau S/2}$, where *S* is a sum of squares statistic. Thus a Pareto prior for τ is conjugate with the likelihood within the class of left-truncated gamma distributions, and leads to a posterior distribution proportional to $\tau^{n/2-\alpha-1} e^{-\tau S/2};\ \tau > c$. We note that the Pareto distribution is not log-concave in τ, and hence BUGS recognises the conjugacy and constructs the full conditional distribution: this will be log-concave provided $n \geq 2(\alpha + 1)$: hence for a locally uniform prior on σ we require at least *n=3* contributions to the precision estimation, and at least *n= 4* to assume a locally uniform distribution for σ^2.

These and other priors are explored in Figure 3.

Figure 3: Prior beliefs arising from an assumption that a precision parameter τ has various forms, shown on scales in τ and σ.

9.2.3 Assessment of proper priors for precision parameters

Instead of using an `off-the-shelf' prior, it is perhaps more appropriate to give some thought as to a reasonable proper prior for the random-effect precision parameter using knowledge of the particular context. Below we give some examples of the sort of thinking that could give rise to proper prior specification. We shall consider normal, logistic and Poisson regression, in which it is assumed that the random effect for subject *i* enters the linear

predictor additively as a variable $b_i \sim \mathbf{Normal}(0, \tau)$.

Normal

Suppose we have a regression model such that the *j*th observation on the *i*th subject has distribution

$$\begin{aligned} y_{ij} &\sim \mathbf{Normal}(\beta^T x_{ij} + b_i, \tau_\epsilon) \\ b_i &\sim \mathbf{Normal}(0, \tau) \end{aligned}$$

Plausible values for τ will depend primarily on the measurement precision τ_ϵ - it is unlikely that the between-subject precision τ is substantially smaller than the within-subject precision τ_ϵ, and plausible that it is considerably greater.

Logistic

Assume a regression model

$$\begin{aligned} r_i &\sim \mathbf{Binomial}(p_i, n_i) \\ logit(p_i) &= \beta^T x_{ij} + b_i \\ b_i &\sim \mathbf{Normal}(0, \tau) \end{aligned}$$

This model means that 95% of subjects with identical covariates will have log(odds) of the event in question occuring in a range with width $2 \times 1.96/\sqrt{\tau}$.

Smith *et al.* (1995) derived the following prior within the context of meta-analysis. Suppose that we thought it plausible that there was roughly one order of magnitude difference between the odds on failure for subjects with identical covariates, and we interpreted this as having a prior belief that 95% of subjects had log(odds) in a range of width $\log 10 = 2.3$. This implies that $(2 \times 1.96/2.3)^2 = 2.9$ is a reasonable guess for τ. Suppose in addition that we would be rather surprised to find two orders of magnitude difference between the odds on failure between subjects, and we would interpret such a finding as requiring 95% of subjects to have log(odds) in a range of width $\log 100 = 4.6$. This gives a `low' value for τ of $(2 \times 1.96/4.6)^2 = .73$. A gamma distribution with parameters (3,1) has a mean of 3 and 96% probability of exceeding .73, and hence $\tau \sim \mathbf{Gamma}(3, 1)$ might be an appropriate proper prior in this context. Figure 3 shows that the implied distribution for σ seemsquite reasonable, although ruling out either very low values (no random effect) and very high values (independent observations).

This analysis can be adjusted to reflect alternative prior beliefs: for example a guess that 95% of log(odds) were within a two-fold range ($\hat{\tau} = 32$), with a ten-fold range being unlikely ($\tau < 2.9$), would lead to a prior $\tau \sim \mathbf{Gamma}(1.44, .045)$, with mean 32 and standard deviation 27.

Poisson

Assume a regression model

$$\begin{aligned} r_i &\sim \mathbf{Poisson}(\mu_i) \\ log(\mu_i) &= \beta^T x_{ij} + b_i \\ b_i &\sim \mathbf{Normal}(0, \tau) \end{aligned}$$

This model means that 95% of subjects with identical covariates will have log(mean) in a range with width $2 \times 1.96/\sqrt{\tau}$.

Following the analysis for the logistic model, we might think it plausible in a specific application that subjects with the same covariates had expectations that varied within one order of magnitude, but unlikely

to vary over two orders. This leads to the same assumption of a Gamma(3,1) prior for τ .

Figure 3 shows a variety of prior suggestions plotted on scales for τ (.25 to 100) and σ (.1 to 2). The selected prior distributions show a variety of behaviours on the σ scale. The `just' proper Gamma(.001,.001) prior favours small values of σ in a reasonable manner, and we have found this to generally lead to sensible conclusions. Although the crucial aspect is whether the shape of the prior is reasonable in the area supported by the likelihood, the major reason for introducing substantial prior opinion is when the likelihood is fairly flat. In this situation some external judgement of plausible values of the precision parameter is unavoidable.

9.3 Model criticism and selection

Although the emphasis in this manual and examples is firmly on drawing inferences assuming a model is an appropriate assumption, serious statistical science requires us to check these assumptions and justify our particular choice of model. We briefly outline some recent suggestions for model checking and comparison using MCMC methods - see the cited references for more details. We shall indicate which aspects can be easily included in a BUGS analysis.

We shall distinguish three objectives: model checking through examination of individual observations, comparison between two or more competitor models, and global checks of goodness-of-fit. In all cases we exploit the idea of comparing observed statistics with our expectations, were we making predictions conditional on the truth of a particular model assumption.

9.3.1 Checking through examination of individual observations.

Consider data $y_1, ..., y_I$ and parameters θ under the assumed model. Gelfand *et al*. (1992) suggest a series of `checking functions' which can be calculated for each observation, involving a comparison of the observed y_i with a predictive distribution $p(Y_i)$. Leaving the basis for this predictive distribution aside for the moment, their suggestions include

1. the residual: $y_i - \mathrm{E}(Y_i)$
2. the standardised residual: $(y_i - \mathrm{E}(Y_i)) / \sqrt{\mathrm{V}(Y_i)}$
3. the chance of getting a more extreme observation: $\min(p(Y_i < y_i), p(Y_i \geq y_i))$
4. the chance of getting a more `surprising' observation: $p(Y_i : p(Y_i) \leq p(y_i))$
5. the predictive ordinate of the observation: $p(y_i)$

The first and last require some standard against which to compare, whereas the middle three options speak for themselves.

Two situations can be identified: when the data $y_1, ..., y_I$ form a separate evaluation set, and when they were used for deriving the posterior distribution of the parameters.

1. *Separate evaluation data available.* In this case the posterior distribution is based on a `training set' $\underline{x}$. The predictive distribution is given by

$$p(Y_i|\underline{x}) = \int p(Y_i|\underline{x}, \theta) p(\theta|\underline{x}) d\theta.$$

   In many cases, unless say there is an explicit dependence on previous observations, the *y*'s are conditionally independent of the *x*'s given the total set of unknowns θ . Thus the required integral is simply

$$p(Y_i|\underline{x}) = \int p(Y_i|\theta) p(\theta|\underline{x}) d\theta.$$

This just requires adding additional random nodes Y_i in the graph with the appropriate parents, and monitoring the samples generated for Y_i.

The observed values y_i can then be compared with their predictive distributions which may be summarised using the `stats` command. As an application of the third checking function, one could then simply see how often the true observations lie within the nominal predictive intervals. Alternatively, one could calculate the first four checking functions above by explicitly including new nodes representing, for example, $\mathrm{E}(Y_i|\theta)$ and $\mathrm{V}(Y_i|\theta)$ and so obtain estimates of $\mathrm{E}(Y_i)$ and $\mathrm{V}(Y_i)$ directly by monitoring these nodes and noting the empirical mean.

2. *No separate evaluation data available.* In this case the predictive distribution of Y_i should ideally be conditional on the model and remainder of the data in order to perform ``cross-validation". Hence for each observation y_i we would require a distribution $p(Y_i|\underline{y}_{\setminus i})$, where is the rest of the data excluding y_i.

   Unfortunately this is generally difficult to do within BUGS. An exception is the final checking function, since we can explicitly calculate the predictive ordinate within BUGS. Gelfand and Dey (1994) point out the interesting fact that

$$\frac{1}{p(y_i|\underline{y}_{\setminus i})} = \int \frac{1}{p(y_i|\underline{y}_{\setminus i},\theta)} p(\theta|\underline{y})\, d\theta$$

   and hence a Monte Carlo estimate of $p(y_i|\underline{y}_{\setminus i})$ is obtained as a harmonic mean of $p(y_i|\underline{y}_{\setminus i},\theta)$. This may be accomplished in BUGS by constructing a node which takes on values $p(y_i|\underline{y}_{\setminus i},\theta)^{-1}$, and then taking the inverse of its empirical mean. However, harmonic means are notoriously unstable so care is required regarding convergence.

   An approximation to the cross-validatory method is to use the methods for a separate evaluation set, but replacing $\underline{x}$ by $\underline{y}$. Hence our predictive distribution is

$$p(Y_i|\underline{y}) = \int p(Y_i|\underline{y},\theta) p(\theta|\underline{y}) d\theta$$

   which will usually be expressible as

$$p(Y_i|\underline{y}) = \int p(Y_i|\theta) p(\theta|\underline{y}) d\theta$$

   and hence just requires adding additional random nodes Y_i in the graph with the same parents as y_i.

   If we do wish to sample from the correct cross-validatory predictive distribution, this can be carried out using an additional importance sampling step to remove the effect of y_i when repredicting Y_i [Gelfand *et al.*, 1992], although this would have to be carried out external to BUGS.

9.3.2 Comparison between two or more candidate models

Bayes factor approaches.

The Bayesian theory of model comparison is based on the Bayes factor, which for two competing models M_1 and M_2 is defined as the ratio of the marginal ordinates of the observed data

$$\frac{p(\underline{y}|M_1)}{p(\underline{y}|M_2)} = \frac{\int p(\underline{y}|\theta_1, M_1)p(\theta_1|M_1)\,d\theta_1}{\int p(\underline{y}|\theta_2, M_2)p(\theta_2|M_2)\,d\theta_2}$$

where θ_i are the unobserved quantities in model *i*. There is a great deal of recent literature on the difficult issues surrounding the calculation and interpretation of Bayes factors: see for example Kass and Raftery (1995), Gelfand and Dey (1994). Carlin and Chib (1995) describe one means of calculating Bayes factors within a Monte Carlo run, and this is illustrated in our `pines` example.

Cross-validatory measures.

We note the analysis first described by Smith (1991) in which it is argued that a Bayes factor approach is only strictly appropriate if we truly believe that one, and only one, of the candidate models is true. If, as would generally be the case, the assumed models are simply considered as useful proxies for some theoretically and practically indeterminate `true' model, then model comparison based on cross-validatory measures may be more suitable.

Comparison between models can therefore be based on an accumulation over observations of the checking functions described above. Examples might include the sum of squared or absolute unstandardised residuals $y_i - \mathrm{E}(Y_i)$, or the negative cross-validatory log-likelihood

$$-\sum_{i=1}^{n} \log p(y_i|\underline{y}_{\setminus i}).$$

The latter results in a comparison between models based on the `pseudo-Bayes factor' [Geisser and Eddy, 1979, Gelfand and Dey, 1994].

We note that such accumulation needs to be carried out after a BUGS run using the estimates obtained from the monitors.

`Deviance' measures

Dempster (1974) suggested examining the posterior distribution of the log-likelihood of the observed data: this is straightforward to calculate with a BUGS run by forming a variable with the value $-2\log p(\underline{y}|\theta)$. For non-hierarchical models, the minimum feasible value of this quantity is the traditional deviance statistic, and in these circumstances a natural connection is made to classical model comparison. However, in for hierarchical models the minimum is likely to be very poorly estimated by the sample minimum, and the mean might be a more reasonable measure: this has been implemented by, for example, Gilks *et al.* (1992) and Zeger and Karim (1991) for model comparison. This quantity is illustrated in the `seeds` example for logistic models, the `salm` example for Poisson models, the `lsat` and `beetles` examples for binary data, the `alli` example for multinomial data and the `jaw` example for multivariate normal data. However, further work is required before giving recommendations on its use.

9.3.3 Global goodness-of-fit tests based on Bayesian p-values

Rubin (1984) introduced a `Bayesianly-justifiable' frequentist model checking device. Suppose we calculate a statistic $d(\underline{y})$ that we expect might be sensitive to departures of interest from the assumed model. For example, if we doubted the normal assumption for the population distribution in random-effects meta-analysis, we might calculate the range of the observed empirical log-odds ratios in the trials. Allowing design aspects of the study to be fixed, such as the sample sizes of the trials, we could then generate replicate sets of data based on our current beliefs about the parameters $p(\theta|\underline{y})$, and re-calculate for each replicate set $\underline{y}^{rep}$ the statistic $d(\underline{y}^{rep})$ which we may abbreviate to d^{rep}. Our observed statistic can then be compared with the distribution $p(d^{rep})$ generated under the assumed model. If $d(\underline{y})$ lies in the extreme tails of this distribution then evidence exists against the model.

This approach can be easily accommodated in BUGS by producing replicate data-sets and constructing nodes to calculate *d* based on the original and replicate data. The empirical distribution of d^{rep} can then be monitored using the `stats` command.

Gelman *et al.* (1996) extend this work of Rubin to allow discrepancy measures $d(\underline{y}, \theta)$ that depend both on the data and parameters, and this is straightforward to program within BUGS.

9.4 Implementation of selected model checking criteria in BUGS

We use the trivial `line` example introduced in Section 1.4 of this manual to illustrate how to implement a number of the above checking criteria in BUGS. (Note that we have changed the data point (2,3) in the `line.dat` file to (2,7) to represent an `outlier').

The first 3 checking functions listed in Section 9.3.1 may be calculated directly in BUGS. Note that in this example, $E(Y_i)$ is denoted by `mu[i]`. To compute $p(Y_i < y_i)$ (see checking function 3), we first obtain values of the random variable Y_i by generating a replicate data set `Y.rep[i]` which depends on the current values of `mu[]` and `tau` at each iteration. The `step()` function is then used to calculate the variable `p.smaller[i]` which takes the value 1 if `Y[i]-Y.rep[i]` ≥ 0 and zero otherwise. Hence the posterior mean of `p.smaller[i]` is simply the proportion of iterations for which `Y.rep[i]` < `Y[i]` (*i.e.* $p(Y_i < y_i)$). It follows that $p(Y_i \geq y_i) = 1 -$ posterior mean of `p.smaller`; the chance of observing a more *extreme* value for Y_i is thus the minimum of these two probabilities. To compute the fifth checking function listed in Section 9.3.1, we calculate the variable `p.inv[i]` (the inverse of the likelihood for observation *i*) at each iteration. Having completed a BUGS run, we then calculate the inverse of the posterior mean of `p.inv[i]` to obtain the predictive ordinate $p(y_i|\underline{y}_{\setminus i})$. In addition, these values may be used to compute the negative cross-validatory log-likelihood discussed in Section 9.3.2.

The model `deviance' may be computed directly in the BUGS code by calculating the log-likelihood contribution (`log.like[i]`) for each observation, and then summing and multiplying by *-2*. Monitoring this sum over iterations will yield a posterior distribution for the deviance.

The replicate data set `Y.rep[]` mentioned earlier is also used to compute the Bayesian p-values described in Section 9.3.3. Here we consider 2 different statistics for $d(\underline{y})$ which may be sensitive to outlying obervations in a Normal model. These are the coefficients of skewness and kurtosis, which are estimated by the average value of the third and fourth moments about the mean divided by the third and fourth powers of the standard deviation respectively. For samples from a Normal distribution, the skewness coefficient has expectation zero, whilst the kurtosis coefficient has expectation 3. Positive skewness indicates an extended right tail, whilst negative skewness implies an extended left tail. A kurtosis coefficient > 3 indicates an excess of values near the mean and far from it, with a corresponding depletion elsewhere: this is the manner in which the *t*-distribution departs from the Normal. Values < 3 result from distributions with a flatter top than Normal. For large sample sizes ($n \approx 150$ for skewness and $n \approx 1000$ for kurtosis), the sampling distributions of the coefficients are approximately Normal with known variances. For smaller sample sizes, these sampling distributions are not appropriate for testing whether the observed coefficients depart significantly from their expected values. The problem of an unkown sampling distribution may be overcome by implementing the ideas of Gelman *et al.* (1996), who suggest computing the *empirical* distribution of the relevant statistic $d(\underline{y}, \theta)$ conditional on current beliefs about the `true' model parameters. In the `line` example, this is achieved by computing the average of the standardised third (`skew.rep`) and fourth (`kurtosis.rep`) moments using the replicate data set `Y.rep[]` at each iteration. These values are then compared to the relevant average moment (`skew.obs` or `kurtosis.obs`) computed from the observed data. The proportion of iterations for which `skew.obs` < `skew.rep` and `kurtosis.obs` <

kurtosis.rep correspond to the Bayesian p-values for testing significant departures from the values of skewness and kurtosis expected under Normality.

The BUGS code below shows the declarations needed to implement each of the above checking criteria for the line example.

```
model line;
const
 N = 5,  # number of observations
 PI = 3.141593;

var
 x[N],Y[N],mu[N],alpha,beta,tau,sigma,x.bar,resid[N],sresid[N],
 m3[N],m4[N],skew.obs,kutosis.obs,p.skew,p.kurtosis, Y.rep[N],
 resid.rep[N],sresid.rep[N],m3.rep[N],m4.rep[N],skew.rep,kutosis.rep,
 p.smaller,like[N],p.inv[N],log.like[N],deviance;

data in  "line.dat";
inits in "line.in";

{
 # Model

 for (i in 1:N) {
   mu[i] <- alpha + beta*(x[i] - x.bar);
   Y[i] ~ dnorm(mu[i],tau);
 }
 x.bar <- mean(x[]);
 alpha ~ dnorm(0.0,1.0E-4);  beta ~ dnorm(0.0,1.0E-4);
 tau ~ dgamma(1.0E-3,1.0E-3);  sigma <- 1.0/sqrt(tau);

 # Model checking

 for (i in 1:N) {
 # Residuals and moments for observed data
   resid[i] <- Y[i] - mu[i];             # Checking fn. 1
   sresid[i] <- resid[i]*sqrt(tau);      # Checking fn. 2

   m3[i] <- pow(sresid[i],3);
   m4[i] <- pow(sresid[i],4);

 # Replicate data set
   Y.rep[i]  ~ dnorm(mu[i],tau);
   p.smaller[i] <- step(Y[i]-Y.rep[i]);  # Checking fn. 3

 # Residuals and moments for replicate data
   resid.rep[i] <- Y.rep[i] - mu[i];
   sresid.rep[i] <- resid.rep[i]*sqrt(tau);

   m3.rep[i] <- pow(sresid.rep[i],3);
   m4.rep[i] <- pow(sresid.rep[i],4);

 # Likelihood for each observed Y[i]
   like[i] <- sqrt(tau/(2*PI))*exp(-0.5*pow(sresid[i],2));
   p.inv[i] <- 1/like[i];         # For checking fn. 5
   log.like[i] <- log(like[i]);   # Log-likelihood for deviance
 }

 # Bayesian p-values
```

```
    skew.obs <- sum(m3[])/N;     skew.rep <- sum(m3.rep[])/N;
    p.skew <- step(skew.rep-skew.obs);

    kurtosis.obs <- sum(m4[])/N;     kurtosis.rep <- sum(m4.rep[])/N;
    p.kurtosis <- step(kurtosis.rep-kurtosis.obs);

  # Deviance distribution
    deviance <- -2*sum(log.like[]);
}
```

Table 6 gives the results for each of these model checks based on a 10000 iteration BUGS run.

Table 6: Results of selected model diagnostics for the `line` *example*

Residuals $y_i - \mathrm{E}(Y_i)$		
resid[1]	-2.05	*95% interval:* (−8.09, 4.03)
resid[2]	3.57	*95% interval:* (−0.70, 7.74)
resid[3]	−0.82	*95% interval:* (−4.41, 2.64))
resid[4]	-1.20	*95% interval:* (−5.54, 3.12)
resid[5]	0.41	*95% interval:* (−5.69, 6.52)
Standardized Residuals $(y_i - \mathrm{E}(Y_i))/\sqrt{\mathrm{V}(Y_i)}$		
sresid[1]	-0.74	*95% interval:* (−2.44, 0.89)
sresid[2]	1.31	*95% interval:* (−0.14, 2.91)
sreisd[3]	−0.30	*95% interval:* (−1.21, 0.60)
sreisd[4]	−0.44	*95% interval:* (−1.60, 0.67)
sresid[5]	0.15	*95% interval:* (−1.36, 1.65)
Probability of more extreme observation $\min(p(Y_i \le y_i), p(Y_i \ge y_i))$		
$\min(\overline{\texttt{p.smaller[1]}}, 1 - \overline{\texttt{p.smaller[1]}})$	0.29	
$\min(\overline{\texttt{p.smaller[2]}}, 1 - \overline{\texttt{p.smaller[2]}})$	0.15	
$\min(\overline{\texttt{p.smaller[3]}}, 1 - \overline{\texttt{p.smaller[3]}})$	0.40	
$\min(\overline{\texttt{p.smaller[4]}}, 1 - \overline{\texttt{p.smaller[4]}})$	0.35	
$\min(\overline{\texttt{p.smaller[5]}}, 1 - \overline{\texttt{p.smaller[5]}})$	0.45	
Predictive ordinate $p(y_i \mid \underline{y}_{\setminus i})$		
$1/\ \overline{\texttt{p.inv[1]}}$	0.039	
$1/\ \overline{\texttt{p.inv[2]}}$	0.008	
$1/\ \overline{\texttt{p.inv[3]}}$	0.010	
$1/\ \overline{\texttt{p.inv[4]}}$	0.086	
$1/\ \overline{\texttt{p.inv[5]}}$	0.069	
Negative cross-validatory log-likelihood $-\sum_{i=1}^{n} \log p(y_i \mid \underline{y}_{\setminus i})$		
$-\sum_{i=1}^{5} \log(1/\overline{\texttt{p.inv[i]}})$	15.47	
Bayesian p-values $p(d < d^{rep})$		
$d =$ **Skewness coefficient:**		
skew.obs	0.45	*95% interval:* (−3.06, 4.90)
skew.rep	−0.08	*95% interval:* (−23.42,21.49)
p.skew	0.44	
$d =$ **Kurtosis coefficient:**		
kurtosis.obs	3.46	*95% interval:* (0.06, 17.29)
kurtosis.rep	39.87	*95% interval:* (0.06, 228.90)
p.kurtosis	0.70	
Deviance $-2 \log p(\underline{y} \mid \theta)$		
deviance	25.27	*Minimum:* 20.92

Recall that we edited observation y_2 to create an outlier. Each of the checking functions described in Section 9.3.1 indicate that y_2 is indeed the most outlying value: y_2 has the largest residuals, the smallest probability of observing a more extreme value and the smallest predictive ordinate. However, the Bayesian p-values suggest that the observed skewness and kurtosis coefficients are consistent with the values expected from a random sample of 5 points from a Normal distribution. That is, there is no evidence against the assumption of Normal errors for this data. Notice the wide 95% credible intervals for the empirical distribution of `skew.rep` and `kurtosis.rep`,

suggesting large sampling variation for these coefficients when based on small sample sizes.

The deviance and negative cross-validatory log-likelihood are intended for comparing 2 or more alternative models. To illustrate this, we fitted a second model to the `line` data, in which the observations were assumed to follow a Student's *t* distribution on 3 degrees of freedom. This yielded a posterior mean deviance of 24.56 (minimum = 19.71) and a negative cross-validatory log-likelihood of 15.44, all of which are slightly smaller than the corresponding values for the Normal model. This suggests that a t-distributed error structure on 3 degrees of freedom provides a marginally better fit to the edited `line` data than does a Normal error structure.

9.5 Ranking

Besag *et al.* (1995) emphasise how MCMC methods are particularly suitable for deriving posterior probabilities of complex functions of multiple parameters, simply by counting the proportion of iterations for which the specific condition obtains. A special but useful example is when inferring the rank order of a set of parameters, for example the mortality rates in a set of hospitals. This is described in detail in the `surgical` example, in which it is shown how to establish the rank order at each iteration by means of the `step` function.

9.6 Measurement error

There has been increasing acknowledgement of the importance of measurement error in epidemiology, and this has led to a large literature [Gail, 1990] on alternative methods for adjusting inferences based on the standard assumption of that measured covariates represent the true `exposure' to whatever risk factor is of interest. Richardson and Gilks (1993, 1994) discuss the Bayesian approach to measurement error using graphical models and Gibbs sampling: here we shall indicate how alternative models for measurement error may be accommodated using BUGS, and also issue a warning about the lack of robustness that can result from a full Bayesian analysis.

Suppose we measure a covariate z_i on an individual *i*, but we feel this is only an approximation for the `true' covariate value which we shall denote $x_i$ . We consider two possible reasons for this approximation: first, that there is `measurement error' either due to the measurement instrument used or due to random fluctuations around a long-term mean, and second that there is `rounding error' due simply to recording with insufficient inaccuracy: an example of the latter is the notorious preference for numbers ending in 0 or 5 in records of blood pressure. Of course, it is possible for both of these causes to exist together.

For each of these causes, we can consider two possible structures relating the observed z_i to the true underlying x_i : first, in the `classical error' model $z_i$ depends on $x_i$ , and we explicitly assume an `exposure model', *i.e.* a probability distribution for x_i in the population; second, we can adopt the `Berkson' model and consider the true x_i as depending on z_i - while this may seem odd at first it is essentially equivalent to assuming a uniform exposure distribution.

Figure 4: Alternative graphical models for measurement and rounding error.

The four alternatives are briefly discussed below, with their possible representation in BUGS, and shown graphically in Figure 4.

Classical measurement error

Assume z_i is normal with mean x_i and precision τ_ϵ , and a normal exposure distribution with mean θ and precision ψ . τ_ϵ will need to be assumed known or a relevant calibration sample available (see the `cervix` example). Alternatively repeated measures of x_i may be used to provide information on τ_ϵ .

There may be external prior information on θ and ψ, or there may be additional relevant sets of data to include in the model, or we may adopt `vague' priors.

```
z[i] ~ dnorm(x[i], tau.e);
x[i] ~ dnorm(theta, psi);
```

Berkson measurement error

Assume x_i is normal with mean z_i and precision τ_ϵ, with the same need for external information on τ_ϵ as above (see the `air` example).

```
x[i] ~ dnorm(z[i], tau.e);
```

Classical rounding error

Assume z_i is x_i rounded to a simple number, so that the real information we have about x_i is that it lies in an interval centred on z_i with width 2*c*, where *2c* is the precision of measurement. For instance, if x_i is only recorded to the nearest whole number ending in 0 or 5 (as in the blood pressure example),

We may again assume a normal exposure distribution with mean θ and precision ψ.

```
z[i] ~ dnorm(x[i], tau.e);
x[i] ~ dnorm(theta, psi) I(lower[i], upper[i]);
lower[i] <- z[i] - c;
upper[i] <- z[i] + c;
```

The restriction in the range of the conditional distribution is dealt with by BUGS using the `I(,)` construct, and represented in the graph by dashed undirected links.

Berkson rounding error

We again assume that z_i is x_i rounded to a simple number, but now interpret this as meaning that x_i could lie anywhere in a range $z_i - c$ to $z_i + c$.

```
x[i] ~ dunif(lower[i], upper[i]);
lower[i] <- z[i] - c;
upper[i] <- z[i] + c;
```

As a simple example of these methods, suppose we have an observed measurement vector $z = (1,2,3,4,5)$. We shall assume $\tau_\epsilon = 4$, which gives a standard deviation of .5 and hence a 95% interval of around ± 1, and $c = .5$ which says observations have been rounded to the nearest whole number. Fairly vague priors are assumed for θ (mean 0, standard deviation 100) and ψ a gamma(.001,.001) distribution). We obtain the estimates shown in Table 7 based on 5000 iterations. We note that the classical approach has pulled the estimated *x*'s towards each other, while the Berkson model produces no change.

Table 7: Estimated **x***'s using different measurement error models*

	Estimated x_i			
Observed z_i	Classical measurement error	Classical rounding error	Berkson measurement error	Berkson rounding error
1	1.19	1.05	1.00	1.00
2	2.10	2.02	2.00	2.00
3	3.00	2.99	3.00	3.00
4	3.90	3.98	4.00	4.00
5	4.80	4.94	5.00	5.00

Suppose we follow the `line` example of Section 1.4 and now consider 5 observed (x, y) pairs $(1,1), (2,3), (3,3), (4,3), (5,5)$. We shall fit a simple linear regression of *y* on *x*, using the notation

$$\begin{aligned} Y_i &\sim \text{Normal}(\mu_i, \tau) \\ \mu_i &= \alpha + \beta(x_i - 3) \end{aligned}$$

where the *x*'s have been standardised around the mean of the *z*'s.

The results in Table 8 are based on 5000 iterations.

Table 8: Estimated **x***'s using different measurement error models when embedded in regression problem.*

		Estimated x_i			
Observed z_i	no error	Classical measurement error	Classical rounding error	Berkson measurement error	Berkson rounding error
1	1	1.14	1.04	0.87	0.96
2	2	2.61	2.13	2.46	2.11
3	3	3.02	3.00	2.98	3.00
4	4	3.40	3.86	3.50	3.89
5	5	4.89	4.96	5.09	5.03
β	.79 (.37)	1.00 (.39)	.84 (.38)	.86 (.36)	.81 (.40)
σ	1.00 (.67)	.57 (.61)	.94 (.65)	.66 (.85)	.86 (.72)

In contrast to the analysis before considering the *y*'s, in which only the classical approach pulled all estimates pulled towards the middle, we see that there is substantial adjustment with all methods. The *x*'s are essentially pulled towards values that provide a better fit to the linear model, so that the classical approach makes a substantial adjustment to the second and fourth values, and the Berkson model estimates x_1 and x_5 as being even more extreme than the surrogate *z*. The classical measurement error approach also shows a dramatic increase in the estimate of β.

Is this appropriate behaviour? MacMahon *et al*. (1990) have provided a highly influential use of measurement error adjustment, in which the *x*'s were all adjusted *before* consideration of the regression analysis: they were all pulled towards the middle on the basis of an external sample. Simultaneous adjustment for measurement error and estimation of regression coefficients clearly leads to adjustments towards improving the fit of the assumed regression model, which of course is perfectly coherent *given the assumption that the regression model is true*. However, rarely do we feel so confident in a regression model.

9.7 Multinomial-Poisson transformation

A widely applicable statistical trick for handling structured multinomial probabilities is to pretend the counts in the categories are actually independent Poisson counts: see Baker (1995) for a long list of possible applications. We show three possible applications of this technique, which may be used to speed sampling in a variety of models.

9.7.1 Multinomial-logistic models

We assume there are *K* categories, with the probability of the *k*th category given covariates x_i given by

$$p(k|x_i) = \frac{e^{\beta_k' x_i}}{\sum_{j=1}^{K} e^{\beta_j' x_i}}$$

where $\beta_1 = 0$ for identifiability. Equivalently

$$\log \frac{p(k|x_i)}{p(1|x_i)} = \beta_k' x_i.$$

For the *i*th covariate pattern, let the observed counts be $Y_{i1}, ..., Y_{iK}, \sum_{j=1}^{K} Y_{ij} = n_i$. Then the likelihood is

$$\prod_{i=1}^{I} \frac{e^{\sum_{j=1}^{K} Y_{ij} \beta_j' x_i}}{\left[\sum_{j=1}^{K} e^{\beta_j' x_i}\right]^{n_i}}$$

This likelihood can be handled directly within BUGS (see Section 4.8 and the `alli` and `endo` examples). However, in some circumstances the following trick may be more efficient.

Suppose we assume the data in fact were generated by

$$\begin{aligned} Y_{ij} &\sim \text{Poisson}(\mu_{ij}) \\ \mu_{ij} &= \mu_{i1} e^{\beta_j' x_i} \end{aligned}$$

Then the likelihood is

$$\prod_{i=1}^{I} \prod_{j=1}^{K} \mu_{ij}^{Y_{ij}} e^{-\mu_{ij}}$$

$$= \prod_{i=1}^{I} \mu_{i1}^{n_i} e^{\sum_{j=1}^{K} Y_{ij} \beta_j' x_i} \; e^{-\mu_{i1} \sum_{j=1}^{K} e^{\beta_j' x_i}}$$

Let the μ_{i1}'s have independent gamma(*a,b*) priors: integrating them out gives a marginal likelihood of β_k's

$$\prod_{i=1}^{I} e^{\sum_{j=1}^{K} Y_{ij} \beta_j' x_i} \int \mu_{i1}^{n_i} e^{-\mu_{i1} \sum_{j=1}^{K} e^{\beta_j' x_i}} \; \mu_{i1}^{a-1} e^{-b\mu_{i1}} d\mu_{i1}$$

$$\propto \prod_{i=1}^{I} \frac{e^{\sum_{j=1}^{K} Y_{ij} \beta_j' x_i}}{\left[\sum_{j=1}^{K} e^{\beta_j' x_i} + b\right]^{n_i + a}}$$

so as $a, b \to 0$, the correct likelihood is obtained.

Since a gamma(0,0) on μ_{i1} is formally equivalent to a uniform prior on $\log \mu_{i1}$, we could also formulate the problem as

$$\log \mu_{ij} = \lambda_i + \beta_j' x_i$$

and give λ_i a locally uniform prior.

The use of this technique is illustrated in the `alli` example.

9.7.2 Conditional likelihoods in case-control studies

Conditional likelihoods in case-control studies provide a further application of the multinomial-Poisson transformation. Chapter 8 of Breslow and Day (1980) describe a structure with *I* case-control strata, each containing one case with covariates x_{i0} , and m_i controls with covariates $x_{i1}, \ldots, x_{im_i}$. The conditional likelihood for relative risk parameters β is

$$\prod_{i=1}^{I} \frac{e^{\beta' x_{i0}}}{\sum_{j=0}^{m_i} e^{\beta' x_{ij}}}$$

Assume the data are given by the disease status indicator $Y_{i0} = 1, Y_{i1} = \cdots = Y_{im_i} = 0$, and were in fact generated by

$$\begin{aligned} Y_{ij} &\sim \text{Poisson}(\mu_{ij}) \quad j = 0, 1, \ldots, m_i \\ \log \mu_{ij} &= \lambda_i + \beta' x_{ij} \end{aligned}$$

Then the likelihood is

$$\prod_{i=1}^{I} \exp(\lambda_i + \beta' x_{i0})\, e^{-\lambda_i + \sum_{j=0}^{m_i} \exp(\beta' x_{ij})}$$

and giving each λ_i a locally uniform prior leads to a marginal likelihood of identical form to the conditional likelihood. So in the basic example of a single case and control and a single exposure variable $x_{ij} = 1$ if exposed, 0 otherwise, this would lead to likelihood contributions of $e^{\beta}/(1 + e^{\beta})$ when the case is exposed and the control not, $1/(1 + e^{\beta})$ when the control is exposed and the case not, and is uninformative if the case and control have matching exposure status. See the `endo` example.

This technique allows conditional inference in case-control studies to take advantage of the Bayesian modelling of proper prior information, hierarchical structure, measurement error and so on.

9.7.3 Partial likelihoods in Cox regression models

It is possible to directly model in BUGS the partial likelihood used in Cox regression models, although in practice we have found it easier to use the Poisson trick: see the `leuk` example.

9.8 Logical constraints

The `I(,)` construct described in Section 4.6 allows bounds on the range of stochastic nodes to be specified. We consider the generic statement

```
y ~ ddist( theta ) I( l, u )
```

where `ddist` represents a specific distribution with parameter or parameters θ , and the bounds *l, u* may themselves depend on θ . A number of different circumstances can apply.

***y, l ,u* all observed, θ not observed:** in this case the constraints are automatically fulfilled and the appropriate likelihood for θ will be generated.

***l ,u* observed (or *l,u* unobserved but *not* dependent on θ), y, θ not observed:** this is the case for censored survival times, where sampling can proceed accordingly.

***y, l ,u* all unobserved, θ observed:** this is handled appropriately, as illustrated in the `marsbars` example.

y, l, u, θ all unobserved, and *l,u* dependent on θ : this must be handled with great care. Essentially, the children of θ are not conditionally independent and so the total likelihood for θ is not a simple product of terms for *y*, *l* and *u*. In this situation the overall joint likelihood contribution from *y, l ,u* must be known externally and the full conditional for θ expressed in BUGS - see the `lips` example and Section 9.10 for a general discussion of how this is dealt with in a context without logical links.

9.9 Parameterisation

Appropriate parameterisation is important in MCMC work in order to improve convergence and stability of the samples. This is a large and complex topic and we merely provide some comments in relation to the worked examples.

- In regression problems it will often be helpful to standardise covariates around their mean in order to make the estimated parameters more orthogonal: see the `rats` and `epil` examples. One should also seek to transform covariates wherever possible to increase orthogonality.
- With random effects, the distributions are generally set to have mean 0: see the `epil` example for two levels of nesting. Note this changes slightly the interpretation of the fixed effects: see the `seeds` example. However, it may be more efficient to use `hierarchical centering' [Gelfand *et al.*, 1995] in which the random effect distribution is centered around the linear predictor, rather than zero. See the `dyes` example. This may be particularly true with multiple random effects, for example spatial models in which both exchangeable and auto-correlated terms are fitted.
- For fixed effects (`non-informative' priors) it is possible to use `corner constraints' in which the baseline category is set to zero: see α_1 in the `ice` example.
- With intrinsic priors the mean, and possibly other functions, of the prior distribution may not be defined: see `ice` and `lips` examples. The problem arises of how to impose sufficient constraints to ensure identifiability. Possibilities include
    1. Using corner constraints at the cost of strong induced dependencies between parameters.
    2. Constraining parameters to sum to zero as in the `seeds` example, at the cost of increased computation time.
    3. In some situations it may be possible to eliminate the grand mean and place constraints on fixed effects in order to allow freely parameterised random effects: contrasts may be calculated after sampling (see the `logRR` parameters in the `ice` example.)
    4. It is possible, but potentially dangerous, to deliberately over-parameterise and allow the sampled values to wander freely, and then display identifiable contrasts (see the β parameters in the `mice` example). While this may help with stability of estimates, the possibility of numerical problems are considerable.

9.10 Undirected links, chain graphs and neighbourhood models

Up to now the discussion has been in terms of Directed Acyclic Graphs (DAGs), which provide a direct mapping into the model specification language. The only deviation from this has been the imposition of bounds on random quantities in Section 4.6, which essentially imposes an *associational* or *undirected* link (see the `marsbars` example). We now consider the general problem of handling *chain graphs*: these also can be considered within the framework of `neighbourhood models': see Besag *et al.* (1995) for a general discusion of MCMC techniques in such models.

Chain graphs. These graphs combine directed and undirected links in such a way that there are no cycles in the graph that contain a directed link, and have been used primarily in modelling of substantive research hypotheses in the social sciences [Wermuth and Lauritzen, 1990, Cox and Wermuth, 1993]. The graphs can always be arranged in a sequence of blocks containing nodes Δ_t, where nodes within each block are connected by undirected links, and links between blocks are directed with the order constraint that a node in Δ_i can only be the parent of a node in Δ_j, where $j > i$. Thus a directed acyclic graph is a chain graph where each block only contains a single node, and an undirected graph is a chain graph with only one block: chain graphs are also known as *block recursive* graphs.

Chain graphs may represent the *chain local Markov property* [Frydenberg, 1990], which states that for a node v in block j, conditional on its parents in preceding blocks and neighbours in the same block, v is independent of all other nodes that are not descendants of v: descendants of a node are considered to be all nodes that are reachable by a path containing a directed link.

Frydenberg (1990) shows that, provided the joint distribution is everywhere positive, this property is equivalent to a joint distribution

$$p(V) = \prod_t p(\Delta_t | \mathbf{parents}[\Delta_t]). \qquad (4)$$

where $u \in \mathbf{parents}[\Delta_t]$ if there is a directed link from u to a member of Δ_t; this is as if each block has been taken as a single `node' in a directed acyclic graph. Further, each of the terms in (4) factorises into

$$p(\Delta_t | \mathbf{parents}[\Delta_t]) = \prod_{C \in \mathcal{C}_t} \phi_C(v_C). \qquad (5)$$

where $\mathcal{C}_t$ are the cliques of the undirected graph with nodes $(\Delta_t, \mathbf{parents}[\Delta_t])$ and undirected links comprising those between members of Δ_t, the links between $\mathbf{parents}[\Delta_t]$ and Δ_t with the directed arrows removed, and a complete set of links between all members of $\mathbf{parents}[\Delta_t]$.

Suppose we now require the full conditional for a node $v \in \Delta_j$. Then

$$\begin{aligned} p(v | V \setminus v) &\propto \text{terms in } p(V) \text{ containing } v \\ &= p(\Delta_j | \mathbf{parents}[\Delta_j]) \prod_k p(\Delta_k | \mathbf{parents}[\Delta_k]) \end{aligned} \qquad (6)$$

where k indexes blocks containing children of Δ_j. By conditional probability and the chain local Markov property we have for the the first term

$$p(\Delta_j | \mathbf{parents}[\Delta_j]) \propto p(v | \mathbf{neighbours}[v], \mathbf{parents}[v]).$$

In each of the likelihood terms in the product in (6), the non-children of v can be integrated out since (5) implies there will be no terms containing both v and non-children of v. Let $\Delta_{k:\mathbf{child}[v]}$ represent the members of Δ_k that are children of v. Then we obtain a full conditional distribution

$$p(v | V \setminus v) \propto p(v | \mathbf{neighbours}[v], \mathbf{parents}[v]) \prod_k p(\Delta_{k:\mathbf{child}[v]} | \mathbf{parents}[\Delta_k]) \qquad (7)$$

Gibbs sampling on chain graphs. Three main approaches for carrying out Gibbs sampling in chain graphs may be identified: treating each block as a single multivariate node, using a factorisation of univariate conditional distributions, and directly specifying the full conditional distributions.

1. Direct use of expression (6) may be feasible if joint distributions within blocks are directly given in the model specification and a procedure exists for picking out the relevant terms: this may be feasible using multivariate normal distributions in future versions of BUGS.

2. Factorisation (7) may be used if each child of *v* is in a separate block containing only one member; this occurs when sampling, for example, the random effects b_i in the `lip` example, each of which has only O_i as a child (ignoring intervening deterministic nodes). Each $\Delta_{k:\text{child}[v]}$ then has only a single member, say *w*, so that the full conditional (7) is then of the form

$$p(v|V \setminus v) \propto p(v|\text{neighbours}[v], \text{parents}[v]) \prod_{w=\text{child}[v]} p(w|\text{parents}[w])$$

   The likelihood terms may be then handled by BUGS in the usual way, but we also need to specify the prior term $p(v|\text{neighbours}[v], \text{parents}[v])$.

   Two issues arise in specifying this prior term. First, we need to ensure that the prior is derived from the correct joint distribution $p(\Delta_j|\text{parents}\Delta_j)$. Second, it should be remembered that we will also be specifying a prior term for each neighbour of *v* in Δ_j, which will be conditioning on *v* and hence appear to have *v* as a parent; BUGS would, unless told otherwise, include each such conditional distribution as a likelihood term for *v* and hence `double-count' the evidence from each neighbour. To stop this inappropriate use of a likelihood term, BUGS therefore has a precedence relation: when constructing the full conditional for a node *v*, if a node *w* is used in calculating the prior term, *w* should not also appear as a child of *v*, *i.e.* any likelihood contribution that makes use of *w* should be ignored.

3. The above procedure will fail if *v* has multiple children in a block, which occurs, for example, when sampling the hyperparameter τ that determines the degree of smoothing in the lip cancer model (see `lips` example), which has all the random effects b_i as children. We now have the additional problem that the likelihood term does not factorise into univariate terms, and so it is necessary to derive the full conditional distribution $p(v|V \setminus v)$ by algebraic means and place it directly in the model specification. The precedence relation then ensures that inappropriate likelihood terms are not added.

We note that the latter technique, over-riding the automatic construction of full conditionals, means that BUGS may be used for handling fully undirected models. However, we emphasise that **when using anything other than a DAG it is the user's responsibility to ensure they are defining a correct joint distribution**.

While BUGS is capable of handling a variety of types of neighbourhood model, its rather simplistic approach to sampling will mean it may be very inefficent compared with the specially tuned methods described, for example, by Besag *et al.* (1995).

9.11 Multivariate normal nodes

9.11.1 Use of Multivariate normal observations

The `jaw` example illustrates the use of a multivariate normal observation vector **x**, with unknown mean μ and precision **T**. This is represented in BUGS by

```
  x[] ~ dmnorm(mu[],T[,])
T[,] ~ dwish(R[,],k)
```

This will carry out multivariate updating of μ. Note our warnings in the `birats` example about instability when adopting multivariate normal observations *and* multivariate normal parmaters in the same model.

9.11.2 Use of Multivariate normal parameters

A set of regression parameters in a generalised linear model may be declared as multivariate normal (see `birats` example). However, unless the response variable is normally distributed, updating of μ will be element-wise rather than multivariate.

9.11.3 Use of the Wishart distribution

The parameterisation used for the Wishart distribution of a $p \times p$ symmetric positive definite matrix **x** is

$$f(\mathbf{x}|\mathbf{R},k) \propto |\mathbf{R}|^{\frac{k}{2}}|\mathbf{x}|^{\frac{k-p-1}{2}}e^{-\frac{1}{2}tr(\mathbf{R}\mathbf{x})}$$

is that of De Groot (1970) and not that of, say, Bernardo and Smith (1994): when reading studies it is important to check which form is being used. Here $k \geq p$ is the degrees of freedom, and **R** is a $p \times p$ symmetric non-singular matrix: our parameterisation treats the Wishart as a multivariate extension of a chi-squared distribution.

Since E(**x**) = $(\mathbf{R}/k)^{-1}$ and $\mathbf{x}^{-1}$ is a covariance matrix, we might roughly think of $\mathbf{R}/k$ as a prior estimate of the covariance based on k observations (in fact, since E($\mathbf{x}^{-1}$) = $\mathbf{R}/(k-p-1)$ and k will generally be chosen to be near or equal to p, it is perhaps best to think of **R** as an assessment of the order of magnitude of the covariance matrix.

9.11.4 Avoiding multivariate nodes for parameters

The use of multivariate normal nodes for *parameters* can be avoided by assuming ``founder'' parameters (with no parents) are *a priori* independent. If we are interested in marginal distributions of parameters this restriction to prior independence should still lead to reasonable conclusions, since dependence introduced by likelihood terms *is* appropriately handled.

Possible aids in avoiding the need for multivariate analyses include

1. Defining parameters so as to make prior independence plausible. Thus, for example, covariates should be standardised about their mean: this will also lead to reduced posterior correlation and hence better convergence.
2. Introducing additional levels of the hierarchical model to explain correlations.

Thus for *random-intercept* models with a single random effect term, it will generally be unnecessary to involve multivariate distributions. However, if there are additional random coefficients, as in teh `birats` and `schools` examples, then it is preferable to use multivariate normal models for the parameters.

Next Previous Contents
Next: Acknowledgements **Previous:** 8 Errors