Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models

David J Spiegelhalter * **Nicola G Best** † **Bradley P Carlin** ‡

March 30, 1998

Abstract

We consider the problem of comparing complex hierarchical models in which the number of parameters is not clearly defined. We follow Dempster in examining the posterior distribution of the log-likelihood under each model, from which we derive measures of fit and complexity (the effective number of parameters). These may be combined into a *Deviance Information Criterion* (DIC), which is shown to have an approximate decision-theoretic justification. Analytic and asymptotic identities reveal the measure of complexity to be a generalisation of a wide range of previous suggestions, with particular reference to the neural network literature. The contributions of individual observations to fit and complexity can give rise to a diagnostic plot of deviance residuals against leverages. The procedure is illustrated in a number of examples, and throughout it is emphasised that the required quantities are trivial to compute in a Markov chain Monte Carlo analysis, and require no analytic work for new models.

1 Introduction

The development of Markov chain Monte Carlo (MCMC) has made it possible to fit increasingly large classes of models with the aim of exploring real-world complexities of data (Gilks *et al.*, 1996). Being able to fit such models naturally leads to the wish to compare alternative formulations with the aim of identifying a class of succinct plausible models: for example, we might ask whether we need to incorporate a random effect to allow for over-dispersion, whether allowing measurement error helps explain the data, what distributional forms to assume, and so on.

Within the classical modelling framework, model comparison takes place by defining a measure of *fit*, typically the deviance statistic, and *complexity*, the number of free parameters in the model. Since increasing complexity is accompanied by better fit, models are compared by trading these two quantities off using likelihood ratio tests, Akaike's information criterion, or one of a number of other suggestions (Aitkin, 1991). Bayesian model comparison using Schwarz's information criterion as a Bayes factor approximation also requires specification of the number of parameters in each model (Kass and Raftery, 1995). Unfortunately, in complex hierarchical models, in which parameters

*MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR, UK: e-mail `david.spiegelhalter@mrc-bsu.cam.ac.uk`

†Dept Epidemiology and Public Health, Imperial College School of Medicine at St Mary's, Norfolk Place, London W2 1PG, UK: e-mail `n.best@ic.ac.uk`

‡Division of Biostatistics, Box 303 Mayo Building, University of Minnesota, Minneapolis, MN 55455-0392, USA: e-mail `brad@muskie.biostat.umn.edu`

generally outnumber observations, these methods clearly cannot be directly applied (Gelfand and Dey, 1994). The most ambitious attempts to tackle this problem appear in the neural network literature (Moody, 1992; MacKay, 1995; Ripley, 1996).

In the next section we follow Dempster (1974) (recently reprinted as Dempster (1997*b*)) in basing comparisons on the posterior distributions of the deviance (-2 log-likelihood + some standardising factor) under each model. We identify 'fit' as the posterior mean of the deviance, and 'complexity' (i.e. the effective number of parameters, p_D) as the difference between the posterior mean of the deviance and the deviance based on the posterior means of the parameters. These quantities can be trivially obtained from a Markov chain Monte Carlo (MCMC) analysis. The fit and complexity are then added to form a *Deviance Information Criterion* (DIC) which may be used for model comparison. In Section 3 we illustrate the use of this simple technique on a reasonably complex example in spatial modelling with covariates.

The remainder of the paper attempts to justify the use of these measures of fit, complexity, and their combination into a single model comparison criterion. In Section 4 we use a heuristic asymptotic argument to see DIC as a generalisation of Akaike's Information Criterion (AIC) (Akaike, 1973) , and then show that DIC has an approximate decision-theoretic justification in terms of minimising expected loss in predicting a replicate dataset, and that in some situations an absolute measure of fit is obtained. Although not necessary for the computation of DIC, it is useful to examine exact and asymptotic forms within standard models (Section 5). We show that in hierarchical normal linear models p_D has a sensible closed form as the trace of of the 'hat' matrix that projects data onto the fitted values, and hence can be related to many other suggestions. Asymptotic approximations in the exponential family reveal further relationships. In Section 6 we argue that each observation's contribution to p_D can be interpreted as its leverage, and show how to obtain diagnostic plots of leverages against deviance residuals as a by-product of an MCMC analysis, regardless of the form of the model, and without any analytic effort.

Section 7 contains a set of examples to show how DIC works in practice, and in Section 8 we discuss how it fits into the general model comparison literature. Finally, Section 9 presents a critique of the method, identifying some areas for further research.

2 Measures of fit and complexity from the posterior distribution of the deviance

Suppose we are fitting a model with observed data y and unknown quantities ϕ, which may include parameters at different levels of the model, latent variables, missing data and so on. The Bayesian approach specifies a joint distribution $p(y, \phi)$, which will generally consist of a product of many terms through conditional independence assumptions. This joint distribution can be written as

$$p(y, \phi) = p(y|\theta)\ p(\theta|\psi)\ p(\psi)$$

where $\phi = (\theta, \psi)$ and y is conditionally independent of ψ given θ. Thus θ are the parameters that directly influence y (*e.g.* true means), while typically ψ are hyper-parameters that govern the form of the prior distribution for θ. Our interest will focus on the 'lowest-level' θ parameters since they directly influence the fit and predictive ability of the model.

Dempster (1974) long ago suggested direct consideration of the posterior distribution of the log-likelihood of the data, equivalent to examining the posterior distribution of

$$D(\theta) = -2 \log p(y|\theta) + 2 \log f(y),$$

where $f(y)$ is some fully specified standardising term that is a function of the data alone and hence does not affect model comparison. We shall term $D(\theta)$ the 'Bayesian deviance', and introduce two specific standardisations: first, the null standardisation $D_0(\theta) = -2\log(\text{likelihood})$ obtained by assuming $f(y)$ is the perfect predictor that gave probability 1 to each observation, and second, for members of the one parameter exponential family with $E(Y) = \mu(\theta)$, the saturated deviance $D_S(\theta)$ obtained by setting $f(y) = p(y|\mu(\theta) = y)$.

The posterior distribution of D is based on $p(\theta|y)$, where $p(\theta|y) \propto p(y|\theta)\, p(\theta)$, and $p(\theta) = \int p(\theta|\psi)\, p(\psi) d\psi$. Dempster (1974) provided some basic examples and some suggestions for comparison of distributions of the log-likelihood, and a number of papers have featured plots or summaries such as posterior means: see, for example, Raghunathan (1988), Zeger and Karim (1991), Gilks *et al.* (1993) and Richardson and Green (1997). However, these authors appear unclear about how to compare models of differing complexity, and in a discussion to a republishing of his 1974 paper, Dempster (1997*a*) gives little extra guidance, stating that "one should plot a representation, perhaps from Markov chain Monte Carlo methods, of the posterior distributions of log-likelihood under each competing model and compare. I hesitate to dictate a procedure for choosing among models."

We shall not be so hesitant, and offer a set of suggestions. First, summarise the 'fit' of a model by the posterior expectation of the deviance;

$$\overline{D} = E_{\theta|y}[D].$$

Second, measure the 'complexity' of a model by the effective number of parameters p_D, defined as the expected deviance minus the deviance evaluated at the posterior expectations;

$$\begin{aligned} p_D &= E_{\theta|y}[D] - D(E_{\theta|y}[\theta]) \\ &= \overline{D} - D(\overline{\theta}). \end{aligned}$$

Finally, models may be compared using a *Deviance Information Criterion*, defined as

$$\begin{aligned} \text{DIC} &= \overline{D} + p_D \\ &= D(\overline{\theta}) + 2p_D; \end{aligned}$$

This will be shown to be a natural generalisation of Akaike's Information Criterion.

DIC may be calculated during an MCMC run by monitoring both θ and $D(\theta)$, and at the end of the run simply taking the sample mean of the simulated values of D, minus the plug-in estimate of the deviance using the sample means of the simulated values of θ. Smaller values of DIC indicate a better-fitting model. As with many previously-proposed model comparison tools, DIC consists of two terms, one representing goodness-of-fit, and the other a penalty for increasing model complexity.

We need to emphasise that we do not recommend that DIC be used as a strict criterion for model *choice* or as a basis for model averaging (Draper, 1995). Selecting a single model is a complex procedure involving background knowledge and other factors such as the robustness of inferences to alternative models with similar support (Box and Tiao, 1973): model choice may be unnecessary in the first place and is certainly very difficult to formalise. We rather view DIC as a method for screening alternative formulations in order to produce a list of candidate models for further consideration. See Section 8 for further discussion of this issue.

3 A running example: the spatial distribution of lip cancer in Scotland

To illustrate the practical application of our suggestion, we analyse data on the rates of lip cancer in 56 counties in Scotland (Clayton and Kaldor, 1987; Breslow and Clayton, 1993). The data include observed (y_i) and expected (E_i) numbers of cases for each county i (where the expected counts are based on the age- and sex-standardised national rate applied to the population at risk in each county), a covariate (x_i) representing the percentage of the population in each county who are engaged in agriculture, fishing or forestry (used as a proxy for sunlight exposure), plus the 'location' of each county expressed as a list ($\mathcal{A}_i$) of its n_i adjacent counties. We assume the usual Poisson model for cancer incidence within each county:

$$y_i \sim \text{Poisson}(\lambda_i E_i)$$

where λ_i denotes the underlying true area-specific relative risk of lip cancer. Using the standard canonical parameterisation, $\theta_i = \log \lambda_i$ may be expressed using either a pooled model (with or without the covariate):

$$\begin{aligned} \text{Model 1:} \quad \theta_i &= \alpha_0 \\ \text{Model 2:} \quad \theta_i &= \alpha_0 + \beta x_i \end{aligned}$$

or a saturated model:

$$\text{Model 3:} \quad \theta_i = \alpha_i$$

where locally uniform priors (actually Normal with mean 0 and variance 10000) are placed on β, α_0 and α_i, $i = 1, ..., 56$. However, models 1 and 2 make no allowance for variation between the true risk ratios in each county (other than that associated with the covariate in model 2), whilst model 3 assumes independence between the county-specific risk ratios (essentially yielding the maximum likelihood estimates $\hat{\lambda}_i = \frac{y_i}{E_i}$). A more plausible assumption is that the true county-specific risk ratios lie somewhere between the independent and pooled estimates, thus motivating the use of random effects models. In the present example, specification of the random effects population distribution is complicated by the possibility that the λ_i may be spatially correlated. That is, the risk ratios in two neighbouring counties may be more similar than risk ratios in two counties further apart, possibly due to dependence on unmeasured risk factors which vary smoothly with geographic location. We thus consider comparison of the following six random and mixed effects models in addition to models 1–3 above:

$$\begin{aligned} \text{Model 4:} \quad \theta_i &= \alpha_0 + \gamma_i \\ \text{Model 5:} \quad \theta_i &= \alpha_0 + \gamma_i + \beta x_i \\ \text{Model 6:} \quad \theta_i &= \phi_i \\ \text{Model 7:} \quad \theta_i &= \phi_i + \beta x_i \\ \text{Model 8:} \quad \theta_i &= \gamma_i + \phi_i \\ \text{Model 9:} \quad \theta_i &= \gamma_i + \phi_i + \beta x_i \end{aligned}$$

where α_0 and β are as for models 1 and 2, γ_i are exchangeable random effects with a Normal prior distribution having zero mean and precision τ_γ, and ϕ_i are spatial random effects with a conditional autoregressive prior (Besag, 1974) given by

$$\phi_i | \phi_{\backslash i} \sim \text{Normal}\Big(\frac{1}{n_i} \sum_{j \in \mathcal{A}_i} \phi_j, \frac{1}{n_i \tau_\phi}\Big) \ .$$

	Model	$\overline{D}$	$D(\overline{\theta})$	p_D	DIC
1	pooled	381.7	380.7	1.0	382.7
2	cov	248.7	238.6	2.1	242.8
3	saturated	56.0	3.1	52.9	108.9
4	exch	60.6	16.8	43.8	104.4
5	exch + cov	62.2	22.5	39.7	101.9
6	spat	56.9	25.3	31.6	88.5
7	spat + cov	59.6	30.2	29.4	89.0
8	exch + spat	56.5	24.0	32.5	89.0
9	exch + spat + cov	59.0	28.7	30.3	89.3

Table 1: Deviance summaries for lip cancer data: 'cov' is a model with the covariate, 'exch' means an exchangeable random effect, 'spat' is a spatially correlated randoj effect.

Gamma(0.001, 0.001) and Gamma(1,1) priors are assumed for the random effects precision parameters τ_γ and τ_ϕ respectively. The prior for the latter is weakly informative in order to improve the stability and convergence properties of the model, since we note that there is considerable non-identifiability in the above parameterisations. However, this does not influence the model comparison which is based only on the fitted θ_i's.

For this Poisson model we adopt the classical deviance (McCullagh and Nelder, 1989)[p 34]

$$D_S(\theta) = 2\sum_i \left[y_i \log \frac{y_i}{e^{\theta_i} E_i} - (y_i - e^{\theta_i} E_i) \right]$$

obtained by taking $-2 \log f(y) = -2 \sum_i \log p(y_i | \theta_i = \log \frac{y_i}{E_i}) = 208.0$ as the standardising factor.

For each model we ran an MCMC sampler in BUGS (Spiegelhalter *et al.*, 1996*a*) for 5000 interations following a burn-in period of 1000 iterations. As suggested by Dempster (1974), Figure 1 shows a kernel-density smoothed plot of the resulting posterior distributions of the deviance under each competing model. Apart from revealing the clear unacceptability of Models 1 and 2, this clearly illustrates the difficulty of formally comparing posterior deviances on the basis of such plots alone.

In Table 1, we present the results of our own suggestion for summarising the deviance for competing models. For each model, $\overline{D}$ is simply the mean of the posterior samples of D, and $D(\overline{\theta})$ is calculated by plugging the posterior means of the relevant parameters (α_0, α_i, β, γ_i, ϕ_i) into the linear predictor θ_i.

Beginning with the fixed effects models 1 and 2, we note that $p_D \approx 1$ and $p_D \approx 2$ respectively, which are the 'true' number of parameters. For model 3, $p_D = 52.9$ which is slightly lower than the true 56 parameters.

The six random effects models have values of p_D ranging from about 30 to 44. Somewhat surprisingly, the effective number of parameters in models 8 and 9, which each contain 112 random effects (56 spatial and 56 exchangeable) are almost identical to the effective number of parameters in models 6 and 7, which each contain only the 56 spatial random effects. There are approximately 10 *more* effective parameters in models 4 and 5. This suggests that the exchangeable random effects do less well at explaining the between-area variability in relative risk compared to the spatial random effects, and indeed, are probably redundant once the spatial effects are also included in the model.

Turning to the comparison of DIC for each model, we first note that DIC is subject to Monte Carlo

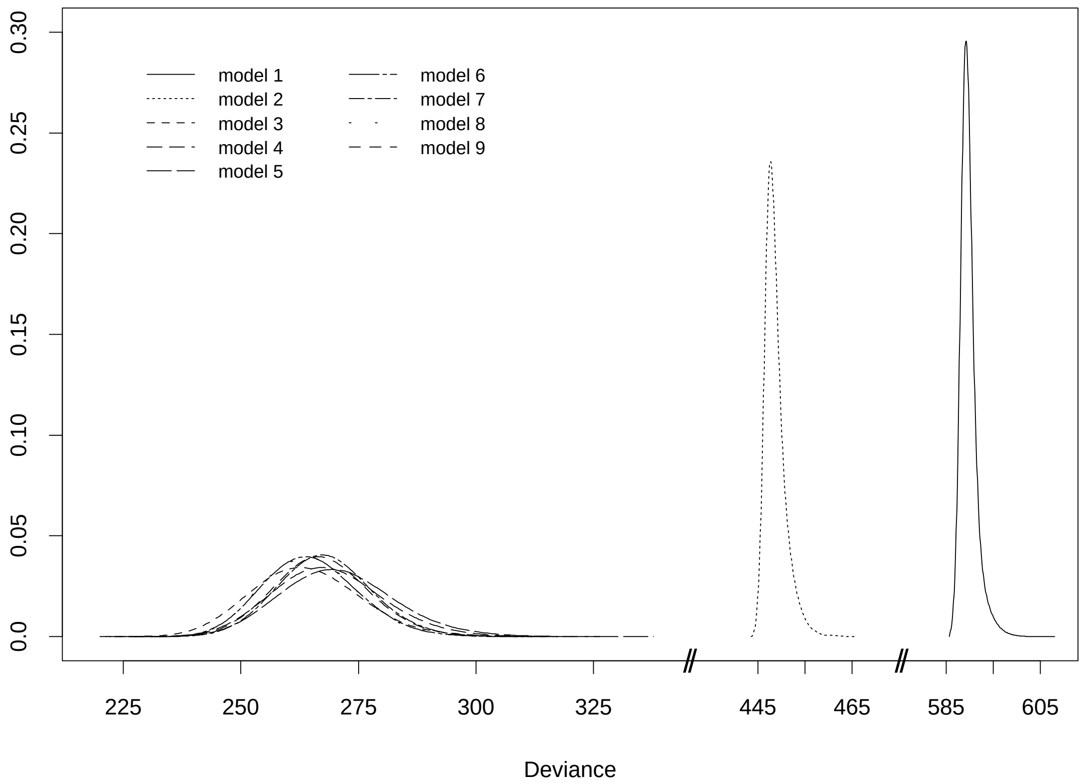

Figure 1: Posterior distributions of the deviance for each model considered in the lip cancer example

sampling error, since it is a function of stochastic quantities generated under an MCMC sampling scheme. Whilst computing the precise standard errors for our DIC values is a subject of ongoing research, the standard errors for the $\bar{D}$ values are readily obtained, and provide a good indication of the accuracy of DIC and p_D. In any case, in several runs using different initial values and random number seeds for this example, the DIC and p_D estimates obtained never varied by more than 0.5. As such, we are confident that, even allowing for Monte Carlo error, any of Models 6-9 are superior (in terms of DIC performance) to models 3-5, which are in turn superior to models 1-2. Comparison of DIC for models 6-9 suggests that the four spatial models are virtually indistiguishable in terms of overall fit. However, in Section 6 we show how to produce diagnostic plots of leverage against deviance residuals using the individual components of DIC: these suggest important differences in fit for certain aspects of models 6–9.

4 Theory and asymptotics

4.1 A heuristic derivation of p_D and DIC

We emphasise that the asymptotics described in this section reflect the situation where increasing information about individual members of $\theta = (\theta_1, ..., \theta_p)$ is received, *i.e.* where the number of observations grows with respect to the number of parameters.

We first expand $D(\theta)$ around $E_{\theta|y}[\theta] = \overline{\theta}$ to give, to second order,

$$\begin{aligned} D(\theta) &\approx D(\overline{\theta}) + (\theta - \overline{\theta})^T \left.\frac{\delta D}{\delta \theta}\right|_{\overline{\theta}} + \frac{1}{2}(\theta - \overline{\theta})^T \left.\frac{\delta^2 D}{\delta \theta^2}\right|_{\overline{\theta}} (\theta - \overline{\theta}) && (1) \\ &= D(\overline{\theta}) - 2(\theta - \overline{\theta})^T L'_{\overline{\theta}} - (\theta - \overline{\theta})^T L''_{\overline{\theta}}(\theta - \overline{\theta}) && (2) \end{aligned}$$

where $L = \log p(y|\theta) = -D/2$ and L' and L'' represent first and second derivatives with respect to θ.

Consider now a non-hierarchical prior in which $p(\theta)$ is assumed to be completely specified with no unknown parameters. It is well known that asymptotically

$$\theta|\mathbf{y} \sim \mathrm{N}\left(\hat{\theta}, -L''_{\hat{\theta}}\right) \qquad (3)$$

where $\overline{\theta} = \hat{\theta}$ are the maximum likelihood estimates such that $L'_{\hat{\theta}} = 0$. Writing $D_{\mathrm{non}}(\theta)$ to represent the deviance for a non-hierarchical model we thus obtain from (2)

$$\begin{aligned} D_{\mathrm{non}}(\theta) &\approx D(\hat{\theta}) - (\theta - \hat{\theta})^T L''_{\hat{\theta}}(\theta - \hat{\theta}) \\ &= D(\hat{\theta}) + \chi^2_p, && (4) \end{aligned}$$

since, by (3), $-(\theta - \hat{\theta})^T L''_{\hat{\theta}}(\theta - \hat{\theta})$ has an approximate chi-squared distribution with p degrees of freedom.

Rearranging (4) and taking expectations with respect to the posterior distribution of θ reveals

$$p \approx E_{\theta|y}[D_{\mathrm{non}}] - D(\hat{\theta}), \qquad (5)$$

i.e., the number of parameters is approximately the expected deviance $\overline{D} = E_{\theta|y}[D_{\mathrm{non}}]$ minus the fitted deviance. Akaike's information criterion (Akaike, 1973) is AIC $= D(\hat{\theta}) + 2p$, and hence from (5) may be written

$$\mathrm{AIC} \approx \overline{D} + p; \qquad (6)$$

the expected deviance plus the number of parameters.

Our suggestion for hierarchical models thus follows equations (5) and (6) but substituting the posterior mean $\overline{\theta}$ for the maximum likelihood estimate $\hat{\theta}$. It is a generalisation of Akaike's criterion: for non-hierarchical models, $\overline{\theta} \approx \hat{\theta}$, $p_D \approx p$ and DIC $\approx$ AIC.

4.1.1 MCMC estimation of the maximum likelihood deviance

We note in passing that since $\mathrm{V}_{\theta|y}[D_{\mathrm{non}}(\theta)] = 2p$, we can use MCMC output to estimate the classical deviance $D(\hat{\theta})$ of any non-hierarchical model by

$$\hat{D}(\hat{\theta}) = E[D] - \frac{1}{2}V[D], \tag{7}$$

using the empirical mean and variance of the sampled values for D. Although this maximum likelihood deviance is theoretically the minimum of D over all feasible values of θ, $D(\hat{\theta})$ will generally be very badly estimated by the sample minimum over an MCMC run.

The above discussion suggests that the posterior distributions for the non-hierarchical Models 1,2 and 3 in Figure 1 should be approximately shifted χ^2 distributions with 1,2 and 56 degrees of freedom respectively, and hence have variances of 2,4 and 112. In fact the respective variances of these distributions are 2.2, 3.8 and 111.0, and equation (7) provides classical deviance estimates of 380.67, 238.64 and 1.00 compared with the true values of 380.73, 238.62 and 0.00. The chi-squared approximation appears rather good, and a reasonably accurate estimate of the maximum likelihood deviance is obtained.

4.2 Asymptotic properties of p_D

Taking expectations of (2) with respect to the posterior distribution of θ gives

$$\begin{aligned} E_{\theta|y}D(\theta) &\approx D(\overline{\theta}) - \mathrm{E}\left[\mathrm{tr}\left((\theta-\overline{\theta})^T L''_{\overline{\theta}}(\theta-\overline{\theta})\right)\right] \\ &= D(\overline{\theta}) - \mathrm{E}\left[\mathrm{tr}\left(L''_{\overline{\theta}}(\theta-\overline{\theta})(\theta-\overline{\theta})^T\right)\right] \\ &= D(\overline{\theta}) - \mathrm{tr}\left(L''_{\overline{\theta}}\,\mathrm{E}\left[(\theta-\overline{\theta})(\theta-\overline{\theta})^T\right]\right) \\ &= D(\overline{\theta}) + \mathrm{tr}\left(-L''_{\overline{\theta}}\,V\right) \end{aligned}$$

where $V = \mathrm{E}\left[(\theta-\overline{\theta})(\theta-\overline{\theta})^T\right]$ is the posterior covariance matrix of θ, and $-L''_{\overline{\theta}}$ is the observed Fisher's information evaluated at the posterior mean of θ. Note that this will be the asymptotic covariance of θ had a locally uniform prior been adopted.

Thus

$$p_D \approx \mathrm{tr}\left(-L''_{\overline{\theta}}V\right), \tag{8}$$

which can be thought of as a measure of the ratio of the information in the likelihood about the parameters as a fraction of the total information in the likelihood and the prior. Under asymptotic posterior normality we have that

$$V^{-1} \approx -L''_{\overline{\theta}} - P''_{\overline{\theta}}$$

where $P'' = \delta^2 \log p(\theta)/\delta\theta^2$, and hence (8) can be written

$$\begin{aligned} p_D &\approx \operatorname{tr}\left((V^{-1} + P''_{\bar{\theta}})V\right) \\ &= p - \operatorname{tr}\left(-P''_{\bar{\theta}}V\right). \end{aligned} \tag{9}$$

In particular, consider the problem of 'regularisation' in complex interpolation models such as neural networks, in which the parameters θ are standardised and assumed to have independent normal priors with precision α. Then expression (9) may be written

$$p_D \approx p - \alpha \operatorname{tr}(V), \tag{10}$$

as obtained by MacKay (1992).

In many conditionally independent hierarchical models θ will be the same length as y and $-L''_{\bar{\theta}}$ will be a diagonal matrix with ith entry $-L''_i = -\frac{\delta^2 \log p(y_i|\theta)}{\delta\theta^2}\Big]_{\bar{\theta_i}}$, and so

$$p_D = \sum_{i=i}^{p} -L''_i \; V(\theta_i|y) \;, \tag{11}$$

showing that each parameter contributes the ratio of the information in the likelihood $-L''_i$ to its posterior precision $V^{-1}(\theta_i|y)$.

4.3 An approximate decision-theoretic justification for DIC

Suppose we wish to make predictions on a replicate dataset Y_{rep} which has an identical design to the observed data y, as in the framework described by Gelfand and Ghosh (1998). Assume the 'true' model is $p(Y_{rep}|\theta)$, and the loss in using an estimate $\tilde{\theta}$ is given by

$$L(\theta, \tilde{\theta}) = \mathrm{E}_{Y_{rep}|\theta}[-2 \log p(Y_{rep}|\tilde{\theta})],$$

the predicted loss using a proper logarithmic scoring rule (Bernardo, 1979).. Denote $-2 \log p(Y_{rep}|\tilde{\theta})$ by $D_{rep}(\tilde{\theta})$. Then following the approach of Ripley (1996)[p33], this loss can be broken down into

$$L(\theta, \tilde{\theta}) = \mathrm{E}_{Y_{rep}|\theta}[D_{rep}(\tilde{\theta}) - D_{rep}(\theta)] + \mathrm{E}_{Y_{rep}|\theta}[D_{rep}(\theta) - D(\theta)] + [D(\theta) - D(\tilde{\theta})] + D(\tilde{\theta}). \tag{12}$$

We shall denote the first two terms by L_1 and L_2 respectively.

Expanding the first term to second order gives

$$L_1(\theta, \tilde{\theta}) \approx \mathrm{E}_{Y_{rep}|\theta}[-2(\tilde{\theta} - \theta)^T L'_{rep,\theta} - (\tilde{\theta} - \theta)^T L''_{rep,\theta}(\tilde{\theta} - \theta)]$$

where $L_{rep,\theta} = \log p(Y_{rep}|\theta)$. Since $\mathrm{E}_{Y_{rep}|\theta}[L'_{rep,\theta}] = 0$, we obtain after some rearrangement

$$L_1(\theta, \tilde{\theta}) \approx \operatorname{tr}\left(I_\theta(\tilde{\theta} - \theta)(\tilde{\theta} - \theta)^T\right)$$

where $I_\theta = \mathrm{E}_{Y_{rep}|\theta}[-L''_{rep,\theta}]$ is Fisher's information in Y_{rep}, and hence also in y. This might reasonably be approximated by the observed information at the estimated parameters, so that

$$L_1(\theta, \tilde{\theta}) \approx \operatorname{tr}\left(-L''_{\tilde{\theta}}(\tilde{\theta} - \theta)(\tilde{\theta} - \theta)^T\right). \tag{13}$$

The second term in (12) may be written as $L_2 = \mathrm{E}_{Y_{rep}|\theta}[-2\log p(Y_{rep}|\theta)] + 2\log p(y|\theta)$: this depends only on the observed data and the true parameter θ, and for any value of θ has sampling expectation zero.

Suppose that under a particular model assumption we obtain a posterior distribution $p(\theta|y)$. Then from (12) and (13) our posterior expected loss when adopting the estimator $\tilde{\theta}$ is

$$\mathrm{E}_{\theta|y}L(\theta,\tilde{\theta}) \approx \mathrm{tr}\left(-L''_{\tilde{\theta}}\ \mathrm{E}_{\theta|y}(\theta-\tilde{\theta})(\theta-\tilde{\theta})^T\right) + \mathrm{E}_{\theta|y}L_2(\theta) + \mathrm{E}_{\theta|y}[D(\theta) - D(\tilde{\theta})] + D(\tilde{\theta}).$$

Ideally we would choose $\tilde{\theta}$ to minimise this function, but this would be extremely complex. In practice, we may approximate the true Bayes estimate by the posterior mean $\overline{\theta}$, making the expected loss

$$\mathrm{E}_{\theta|y}L(\theta,\overline{\theta}) \approx \mathrm{tr}\left(-L''_{\overline{\theta}}V\right) + \mathrm{E}_{\theta|y}L_2(\theta) + p_D + D(\overline{\theta}),$$

where V has been previously defined as the posterior covariance of θ, and $p_D = \overline{D} - D(\overline{\theta})$. Since we have already shown in (8) that $p_D \approx \mathrm{tr}\left(-L''_{\overline{\theta}}V\right)$, we finally obtain the attractive result that the expected posterior loss when adopting a particular posterior distribution is approximately $DIC = 2p_D + D(\overline{\theta})$, plus a term with expectation zero whichever model is true.

Kass and Raftery (1995) criticise Akaike (1973) for using a plug-in predictive distribution as we have done above, rather than the full predictive distribution obtained by integrating out the unknown parameters. The above justification must therefore be taken as fairly heuristic.

4.4 Asymptotic sampling theory properties of the posterior expected deviance

Suppose that all aspects of the assumed model are true. Then before observing y our expectation of the posterior expected deviance is

$$\begin{aligned} E_y(\overline{D}) &= E_y\left[E_{\theta|y}D(\theta)\right] \\ &= E_\theta\left[E_{y|\theta}\left[-2\log p(y|\theta) + 2\log f(y)\right]\right] \end{aligned}$$

by reversing the conditioning between y and θ. Suppose θ is of dimension p and let $f(y) = p(y|\theta'(y))$ where $\theta'(y)$ are the standard maximum likelihood estimates. Then

$$E_{y|\theta}\left[-2\log\frac{p(y|\theta)}{p(y|\theta'(y))}\right]$$

is simply the expected likelihood ratio statistic for the fitted values $\theta'(y)$ with respect to the true null model θ, and hence under the standard conditions is asymptotically p. Hence we expect, if the model is true, the posterior expected deviance (standardised by the maximised log-likelihood) to be p, the dimension of θ.

In particular, consider the one-parameter exponential family where $p = n$, the total sample size. The likelihood is maximised by substituting y_i for the mean of y_i, and the posterior mean of the deviance has approximate sampling expectation of n if the model is true. This might be appropriate for checking the overall goodness-of-fit of the model. This will be exact for normal models with known variance, but in general will only be reliable if each observation provides considerable information about its mean (McCullagh and Nelder, 1989, p. 36). Note that comparing $\overline{D}$ with n is precisely the same as comparing $D(\overline{\theta})$ with $n - p_D$, the effective degrees of freedom.

In Table 1 we might therefore compare the column $\overline{D}$ with the sample size $n = 56$. This suggests that all models 3 to 9 provide an adequate overall fit to the data, and that the comparison is based on their complexity alone.

5 Results for some common model classes

5.1 Normal models

5.1.1 One-way ANOVA

Consider the one-way analysis of variance with known variance components, i.e.

$$y_i \sim N(\theta_i, \lambda_i^{-1}),\ i = 1, ..., p$$

giving

$$D(\theta) = \sum_i \lambda_i (y_i - \theta_i)^2,$$

which is -2 log(likelihood) standardised by the term $-2 \log f(y) = \sum_i \log \frac{2\pi}{\lambda_i}$ obtained from setting $\theta_i = y_i$.

Saturated model. We assume the θ_i's have independent locally uniform priors, so that $\theta_i|y \sim N(y_i, \lambda_i^{-1})$, and $D_{\text{sat}}(\theta) = \chi_p^2$. Thus

$$\overline{D}_{\text{sat}} = p, \quad D_{\text{sat}}(\overline{\theta}) = 0,$$

and so $p_D = p$, the true number of parameters.

Pooled model. We assume $\theta_i = \theta$ for all i, where θ has a locally uniform prior. Then $\theta|y \sim N(\overline{y}, 1/(\sum \lambda_i)),\ \overline{y} = \sum \lambda_i y_i / \sum \lambda_i$, and so $D_{\text{pool}}(\theta) = \sum_i \lambda_i (y_i - \overline{y})^2 + \chi_1^2$. Thus

$$\overline{D}_{\text{pool}} = \sum_i \lambda_i (y_i - \overline{y})^2 + 1, \quad D_{\text{pool}}(\overline{\theta}) = \sum_i \lambda_i (y_i - \overline{y})^2,$$

and so $p_D = 1$, the true number of parameters.

Exchangeable model, known normal prior. We assume a prior $\theta_i \sim N(\mu, \alpha^{-1})$ where μ, α are assumed known. Then $\theta_i|y \sim N(\rho_i y_i + (1 - \rho_i)\mu, \rho_i \lambda_i^{-1})$ where $\rho_i = \lambda_i/(\lambda_i + \alpha)$ is the likelihood precision as a fraction of the posterior precision: in the language of traditional mixed modelling this is equal to the intra-class correlation coefficient. It can be easily shown that

$$D_{\text{exch}}(\theta) = \sum \rho_i \chi^2(1, (y_i - \mu)^2 (1 - \rho_i)\alpha),$$

where $\chi^2(a, b)$ is a non-central chi-square with mean $a + b$ and variance $2(a + 2b)$. Thus, since $\rho_i \alpha = (1 - \rho_i)\lambda_i$, we have

$$\overline{D}_{\text{exch}}(\theta|\mu) = \sum \rho_i + \sum \lambda_i (1 - \rho_i)^2 (y_i - \mu)^2, \quad D_{\text{exch}}(\overline{\theta}|\mu) = \sum \lambda_i (1 - \rho_i)^2 (y_i - \mu)^2,$$

and so

$$p_D = \sum_i \rho_i = \sum_i \frac{\lambda_i}{\lambda_i + \alpha}. \tag{14}$$

The effective number of parameters is therefore the sum of the intra-class correlation coefficients, which essentially measures the sum of the ratios of the precision in the likelihood to the precision in the posterior, as described in equation (11) above.

Exchangeable model, unknown normal prior mean. Giving μ a locally uniform prior, we obtain a posterior distribution $\mu \sim N(\overline{y}, (\alpha \sum \rho_i)^{-1})$, where $\overline{y} = \sum \rho_i y_i / \sum \rho_i$. It is straightforward to show that

$$\begin{aligned} \overline{D}_{\text{exch}}(\theta) &= \sum \rho_i + \alpha \sum \rho_i (1-\rho_i)(y_i - \overline{y})^2 + \sum \rho_i (1-\rho_i) / \sum \rho_i \\ D_{\text{exch}}(\overline{\theta}) &= \alpha \sum \rho_i (1-\rho_i)(y_i - \overline{y})^2, \end{aligned}$$

and so $p_D = \sum \rho_i + \sum \rho_i (1-\rho_i)/\sum \rho_i$.

If all group precisions are equal, $p_D = 1 + (p-1)\rho$, as obtained by Hodges and Sargent (1998).

Exchangeable model, general known prior. Suppose θ_i has a specified prior $p(\theta)$, and the posterior distribution of θ_i has mean $\overline{\theta}_i$ and precision τ_i. Then this 'general' deviance has the form

$$D(\theta) = \sum_i \lambda_i (y_i - \overline{\theta}_i)^2 + \sum_i \lambda_i 2 (y_i - \overline{\theta}_i)(\overline{\theta}_i - \theta_i) + \sum_i \lambda_i (\overline{\theta}_i - \theta_i)^2$$

and so

$$\overline{D}_{\text{gen}} = \sum_i \lambda_i (y_i - \overline{\theta}_i)^2 + \sum_i \lambda_i / \tau_i, \quad D_{\text{gen}}(\overline{\theta}) = \sum_i \lambda_i (y_i - \overline{\theta}_i)^2,$$

and so $p_D = \sum \lambda_i / \tau_i$, so that each observation contributes the ratio of its posterior precision based on the likelihood alone, to its posterior precision based on the full model.

Suppose further that the posterior distribution of θ_i given $y_{\backslash i}$, *i.e.* all the data except y_i, is approximately normal with mean $\overline{\theta}'_i$. Then from standard normal prior/posterior analysis, $\overline{\theta}_i = w_i y_i + (1 - w_i)\overline{\theta}'_i$, where $w_i = \lambda_i / \tau_i$. Thus the contribution to the total number of parameters is the relative weight given to the observation in estimating its mean.

We note in passing that Ye (1998) suggests that, for general normal models, the contribution of each θ_i to the effective number of parameters should be

$$h_i(\theta) = \frac{\delta E_{y|\theta}[\hat{\theta}_i]}{\delta \theta_i}.$$

In our Bayesian framework we equate $\hat{\theta}_i$ to $\overline{\theta}_i$, and hence $E_{y|\theta}[\hat{\theta}_i] = w_i \theta_i + (1 - w_i) E_{y|\theta}[\overline{\theta}'_i]$. Since the second term does not depend on θ_i, we thus obtain

$$h_i(\theta) = w_i,$$

and hence in this context Ye's suggestion is a special case of our general deviance-based measure of complexity.

5.1.2 The general normal linear model

We consider the general hierarchical normal model using the notation of Lindley and Smith (1972). Suppose

$$\begin{aligned} y &\sim N(A_1 \theta_1, C_1) \\ \theta_1 &\sim N(A_2 \theta_2, C_2) \end{aligned}$$

where all matrices and vectors are of appropriate dimension. Then the standardised deviance is $D(\theta_1) = (y - A_1 \theta_1)^T C_1^{-1} (y - A_1 \theta_1)$. Assume the posterior distribution for θ_1 is normal with mean

$\overline{\theta}_1 = Bb$ and covariance B: B and b will be left unspecified for the moment. Then expressing $y - A_1\theta_1$ as $y - A_1\overline{\theta}_1 + A_1\overline{\theta}_1 - A_1\theta_1$ reveals that

$$D(\theta_1) = D(\overline{\theta}_1) + 2(y - A_1\overline{\theta}_1)^T C_1^{-1}(\theta_1 - \overline{\theta}_1) + (\theta_1 - \overline{\theta}_1)^T A_1^T C_1^{-1} A_1(\theta_1 - \overline{\theta}_1).$$

Taking expectations with respect to the posterior distribution of θ_1 eliminates the middle term and gives

$$\overline{D} = D(\overline{\theta}_1) + \mathrm{tr}(A_1^T C_1^{-1} A_1 B),$$

and thus $p_D = \mathrm{tr}(A_1^T C_1^{-1} A_1 B)$.

If θ_2 is assumed known, then Lindley and Smith show that $B^{-1} = A_1^T C_1^{-1} A_1 + C_2^{-1}$ and hence $p_D = \mathrm{tr}[A_1^T C_1^{-1} A_1 (A_1^T C_1^{-1} A_1 + C_2^{-1})^{-1}] = p - \mathrm{tr}[C_2^{-1}(A_1^T C_1^{-1} A_1)^{-1}]$.

A more revealing identity is found by assuming θ_2 is unknown with a locally uniform prior. Then Lindley and Smith show that $B^{-1} = A_1^T C_1^{-1} A_1 + C_2^{-1} - C_2^{-1} A_2 (A_2^T C_2^{-1} A_2)^{-1} A_2^T C_2^{-1}$ and $b = A_1^T C_1^{-1} y$. The fitted values for the data are given by $\hat{y} = A_1\overline{\theta}_1 = A_1 B b = A_1 B A_1^T C_1^{-1} y$, and so the 'hat' matrix that projects the data onto the fitted values is $H = A_1 B A_1^T C_1^{-1}$. Now $p_D = \mathrm{tr}(A_1^T C_1^{-1} A_1 B) = \mathrm{tr}(A_1 B A_1^T C_1^{-1}) = tr(H)$.

This identification of the effective number of parameters with the trace of the 'hat' matrix is a standard result in linear modelling, and extends to the general class of smoothing and generalised additive models (Hastie and Tibshirani, 1990)[Sec 3.5], and is also the conclusion of Hodges and Sargent (1998) in the context of general linear models. The advantage of using the deviance formulation for specifying p_D is that all matrix manipulation and asymptotic approximation is avoided. Note that $tr(H)$ is the sum of terms which in regression diagnostics are identified as the individual *leverages*, the influence of each observation on its fitted value: we shall exploit this identity in Section 6.

5.1.3 The general normal non-linear model

A large class of models can be formulated using the following extension to the Lindley-Smith model discussed above:

$$\begin{aligned} y &\sim N(g(\theta_1), \tau^{-1} D_1) \\ \theta_1 &\sim N(A_2\theta_2, \alpha^{-1} D_2) \end{aligned}$$

where g is a non-linear expression as found, for example, in pharmacokinetics or neural networks, and τ and α are likelihood and prior precisions respectively (presumed known for the present): in many situations $A_2\theta_2$ will be 0 and D_1, D_2 will be identity matrices. Define

$$\begin{aligned} q(\theta_1) &= (y - g(\theta_1)^T D_1^{-1}(y - g(\theta_1)) \\ r(\theta_1) &= (\theta_1 - A_2\theta_2)^T D_2^{-1}(\theta_1 - A_2\theta_2) \end{aligned}$$

as the likelihood and prior residual variation.

Assuming asymptotic posterior normality, we have $\theta_1 | y \sim N(\overline{\theta_1}, V)$ where

$$\begin{aligned} V^{-1} &\approx -L''_{\overline{\theta_1}} - P''_{\overline{\theta_1}} \\ &= \frac{\tau}{2} q''(\overline{\theta_1}) + \frac{\alpha}{2} D_2^{-1} \end{aligned}$$

which from equation (9) leads to

$$p_D \approx p - \mathrm{tr}\left(\left[\frac{\tau}{\alpha} q''(\overline{\theta_1}) D_2 + I_p \right]^{-1} \right).$$

Let $q''(\overline{\theta_1})D_2$ have eigenvalues $\lambda_i, i = 1, ..., p$. Then $\frac{\tau}{\alpha} q''(\overline{\theta_1})D_2 + I_p$ has eigenvalues $\lambda_i \tau/\alpha + 1$, and hence

$$p_D = \sum \frac{\lambda_i \tau}{(\lambda_i \tau + \alpha)}.$$

Setting $\tau = 1$ gives the result (14) found in the one-way analysis of variance example in Section 5.1.1, but this more general expression was described by MacKay (1992).

MacKay (1992) shows a consequence of this formulation can be the use of 'effective degrees of freedom' in estimating variance paramteers. Assume τ and α are unknown and to be estimated by maximising the 'Type II' likelihood $p(y|\alpha, \tau)$ derived from integrating out the unknown θ_1 from the likelihood. From a standard Laplace approximation (Kass and Raftery, 1995), this is is equivalent to minimising

$$\begin{aligned} -2 \log p(y|\alpha, \tau) &= -2 \log p(y|\overline{\theta_1}, \alpha, \tau) - 2 \log p(\overline{\theta_1}|\alpha, \tau) + 2 \log p(\overline{\theta_1}|y, \alpha, \tau) \\ &\approx \tau q(\overline{\theta_1}) + n \log 2\pi + \log |D_1/\tau| \\ &\quad + \alpha r(\overline{\theta_1}) + p \log 2\pi + \log |D_2/\alpha| \\ &\quad - p \log 2\pi - \log |V|. \end{aligned}$$

Now

$$\begin{aligned} \log |D_2/\alpha| - \log |V| &= \log |\alpha^{-1} D_2 V^{-1}| \\ &= \log |\frac{\tau}{\alpha} q''(\overline{\theta_1}) D_2 + I_p| \\ &= \sum \log \frac{\lambda_i \tau + \alpha}{\alpha} \end{aligned}$$

by the previous derivation. Thus we seek to minimise

$$-2 \log p(y|\alpha, \tau) \approx \tau q(\overline{\theta_1}) - N \log \tau + \alpha\, r(\overline{\theta_1}) - p \log \alpha + \sum \log(\lambda_i \tau + \alpha).$$

Setting derivatives equal to zero reveals that

$$\begin{aligned} \hat{\tau}^{-1} &= \frac{q(\overline{\theta_1})}{n - p_D} \\ \hat{\alpha}^{-1} &= \frac{r(\overline{\theta_1})}{p_D} \end{aligned}$$

which are the fitted likelihood and prior residual variation, divided by their effective degrees of freedom derived from p_D.

These results were derived by MacKay (1992) using the form of p_D given in equation (10), although we emphasise two differences in our result. First, while Mackay needs to specifically evaluate $\mathrm{tr}(V)$, our p_D arises without any additional computation. Second, we would recommend including α and τ in the general MCMC estimation procedure, rather than relying on Type II maximum likelihood estimates (Ripley, 1996)[p. 167].

5.2 General one-parameter exponential family

We assume that we have p groups of observations, where each of the n_i observations in group i has the same distribution. Following McCullagh and Nelder (1989), we define a one-parameter exponential family for the jth observation in the ith group as

$$\log p(y_{ij}|\theta_i, \phi) = w_i(y_{ij}\theta_i - b(\theta_i))/\phi + c(y_{ij}, \phi),$$

where standard results are that

$$\mu_i = E(Y_{ij}|\theta_i, \phi) = b'(\theta_i), \quad V(Y_{ij}|\theta_i, \phi) = b''(\theta_i)\phi/w_i.$$

Both canonical and mean parameterisations may be of interest. Here we shall focus on the canonical parameterisation in terms of θ_i, as its posterior distribution should better fulfill the asymptotic normal approximation underlying the optimality criterion described in Section 4.3: related identities are of course available for the mean parameterisation in terms of $\mu_i = \mu(\theta_i)$. Often the choice of parameterisation has little impact on the resulting conclusions. However, in Section 7.3 we present an example of a Bernoulli model where this choice does prove to be important, and in Section 9 we further discuss parameterisation invariance issues.

Writing $\overline{b_i} = E_{\theta_i|y}[b(\theta_i)]$, we easily obtain that the contribution of the ith group to the effective number of parameters is $p_{Di} = 2n_i w_i(\overline{b_i} - b(\overline{\theta}_i))/\phi$.

5.2.1 Poisson likelihood with conjugate prior

In this case $\phi = 1, w_i = 1, b(\theta) = e^\theta$ and hence $p_{Di} = 2n_i(E_{\theta_i|y}[e^{\theta_i}] - e^{\overline{\theta}_i})$.

Let us assume a conjugate prior $\mu_i = e^{\theta_i} \sim \Gamma(\alpha, \beta)$. Now if $X \sim \Gamma(a, b)$, then $\mathrm{E}(\log X) = \psi(a) - log(b)$, where $\psi(t) = \delta \log \Gamma(t)/\delta t$ is the digamma function. It follows that $p_{Di} = 2n_i(\alpha + y_i - e^{\psi(\alpha+y_i)})/(\beta + n_i)$, where $y_i = \sum_j y_{ij}$. Using Stirling's approximation, $\psi(x) = \log x - 1/(2x) + O(x^{-2})$ (see e.g. Abramowitz and Stegun (1970)[p.259], and a one-term Maclaurin expansion of the exponential function, we obtain that

$$p_D \approx \sum_i n_i/(\beta + n_i).$$

We note that this does not depend on the data observed, and each group contributes the fraction of its sample size to the effective sample size underlying the posterior. As the sample size in each group increases, its contribution tends towards 1.

5.2.2 Bernoulli likelihood with conjugate prior

In this case $\phi = 1, w_i = 1, \theta = logit(\mu) = \log[\mu/(1-\mu)], b(\theta) = \log(1 + e^\theta)$ and hence $p_{Di} = 2n_i\{E_{\theta_i|y}[\log(1 + e^{\theta_i})] - \log(1 + e^{\overline{\theta}_i})\}$.

Let us assume a conjugate prior $\mu_i = (1 + e^{-\theta_i})^{-1} \sim Beta(\alpha, \beta)$. Now if $X \sim Beta(a, b)$, then $E(\log X) = \psi(a) - \psi(a + b), E[\log(1 - X)] = \psi(b) - \psi(a + b)$. Hence it can be shown that

$$p_{Di} = 2n_i\{\psi(\alpha + \beta + n_i) - \log(e^{\psi(\beta+n_i-y_i)} + e^{\psi(\alpha+y_i)})\}.$$

A similar use of Stirling's approximation to that in the previous subsection shows that $p_D \approx \sum_i n_i/(\alpha+\beta+n_i)$. We note that this does not depend on the data observed, and again is the ratio of sample size to effective posterior sample size. As the sample size in each group increases, its contribution tends towards 1.

5.2.3 Asymptotic form

A second order Taylor expansion of $D(\theta_i)$ around $D(\overline{\theta}_i)$ yields

$$D(\theta_i) \approx D(\overline{\theta}_i) - 2(\theta_i - \overline{\theta}_i)w_i(y_i - n_i b'(\overline{\theta}_i))/\phi + (\theta_i - \overline{\theta}_i)^2 w_i n_i b''(\overline{\theta}_i)/\phi$$

and hence as each n_i increases the contribution to p_D tends to

$$p_{Di} \approx \frac{w_i}{\phi} n_i b''(\overline{\theta}_i)\ V(\theta_i|y),$$

the precision in the likelihood divided by the posterior precision.

It can be shown that for Poisson and Bernoulli likelihoods and conjugate priors this approximation yields precisely the results obtained above by using Stirling's approximation on the exact contribution.

5.3 Generalised linear and mixed models with canonical link functions

Following McCullagh and Nelder (1989) we assume the mean μ_i of y_{ij} is related to a set of covariates x_i through a link function $g(\mu_i) = x_i^T\alpha$, and that g is the canonical link $\theta(\mu)$.

5.3.1 Canonical parameterisation

Using the asymptotic result above, we have

$$p_{Di} \approx \frac{w_i}{\phi} n_i b''(\overline{\theta}_i)\ V(x_i^T\alpha|y) = \frac{w_i}{\phi} n_i b''(\overline{\theta}_i)\ x_i^T V(\alpha|y) x_i.$$

Since $b''(\theta_i) = \mu'(\theta_i) = 1/g'(\theta_i)$, and $V(y|\theta_i) = b''(\theta_i)\phi/w_i$,

$$\frac{w_i}{\phi} n_i b''(\overline{\theta}_i) = \frac{n_i}{g'^2(\overline{\theta}_i)\, V(y|\overline{\theta}_i)} = W_i,$$

where W_i are the GLM iterated weights (McCullagh and Nelder, 1989, p. 40). Hence

$$p_D = tr\left[X^T W X \quad V(\alpha|y)\right];$$

Under a $N(\alpha_0, V)$ prior on α, the prior contribution to the negative Hessian matrix at the mode is just V^{-1}, so under the canonical link the approximate normal posterior has variance

$$V(\alpha|y) = [V^{-1} + X^T W X]^{-1},$$

again producing the p_D as a measure of the ratio of likelihood to posterior information.

5.3.2 Laird-Ware normal models

Laird and Ware (1982) specified the mixed normal model

$$y \sim N(X\alpha + Z\beta, R), \quad b \sim N(0, D),$$

where R and D are currently assumed known. The random effects are β, fixed effects are α, and placing a uniform prior on α we can write this model within the general Lindley-Smith formulation by setting $R = C_1, (\alpha, \beta) = \theta_1, (X, Z) = A_1, 0 = \theta_2$ and C_2 as a block-diagonal matrix with infinities in the top-left block, D is the bottom right, and zeros elsewhere.

We have already shown that in these circumstances

$$p_D = \mathrm{tr}[A_1^T C_1^{-1} A_1 (A_1^T C_1^{-1} A_1 + C_2^{-1})^{-1}],$$

and substituting in the appropriate entries for the Laird-Ware model gives $p_D = tr(V^* V^{-1})$, where

$$V^* = \begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z \end{bmatrix}, \quad V = \begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + D^{-1} \end{bmatrix}$$

which is the precision of the parameter estimates assuming $D = \infty$, relative to the precision assuming D.

5.3.3 Generalised linear mixed models

We now consider the class of generalized linear mixed models with canonical link, in which $g(\mu_i) = x_i^T \alpha + z_i^T \beta$, where $\beta \sim N(0, D)$ (Breslow and Clayton, 1993).

Using the same argument as for GLMs, we find that

$$p_D = tr\left[(X, Z)^T W (X, Z) V((\alpha, \beta)|y)\right] = p_D = tr(V^* V^{-1}) \ ,$$

where

$$V^* = \begin{bmatrix} X^T W^{-1} X & X^T W^{-1} Z \\ Z^T W^{-1} X & Z^T W^{-1} Z \end{bmatrix}, \quad V = \begin{bmatrix} X^T W^{-1} X & X^T W^{-1} Z \\ Z^T W^{-1} X & Z^T W^{-1} Z + D^{-1} \end{bmatrix}.$$

This matches the proposal of Lee and Nelder (1996) except their D^{-1} is a diagonal matrix of the second derivatives of the prior likelihood for each random effect.

6 Using DIC for model diagnostics

In Section 5.1.2 we noted that in general linear models the contribution of each observation to p_D turned out to be its leverage, defined as the relative influence each observation has on its own fitted value, and we suggest that this interpretation may be taken in general model fitting. Thus as a by-product of MCMC estimation we may obtain deviance residuals and estimates of leverage for each observation. Suppose we are working with the saturated deviance D_S, where the contribution of an individual observation i to DIC is

$$\begin{aligned} \mathrm{DIC}_i &= \overline{D}_{S_i} + p_{Di} \\ &= {dr_i}^2 + p_{Di} \end{aligned}$$

where $dr_i = \pm\sqrt{D}_{S_i}$ (with sign given by the sign of $(y_i - E(y_i|\overline{\theta}))$) is the deviance residual, defined analagously to McCullagh and Nelder (1989) p 39.

A wide range of diagnostic plots of these quantities are possible, following the structure established in non-hierarchical models. For example, Figure 2 shows a plot of dr_i against p_{Di} for each of the 9 models considered for the lip cancer example introduced Section 3. The dashed lines marked on each plot are of the form $x^2 + y = c$ and points lying along such a parabola will each contribute an amount $\text{DIC}_i = c$ to DIC for that model. For models 3–9, parabolas are marked at values of $c = 1$, 2 and 5, and any data point whose contribution $\text{DIC}_i > 2$ is labelled by its observation number. For models 1 and 2, parabolas are marked at $c = 1$, 10 and 50, since the size of the deviance resiudals and individual contributions to DIC are much larger. For clarity, only points for which $\text{DIC}_i > 10$ are marked by their observation number.

Figure 2 identifies observations 55 and 56, the only counties with $y_i = 0$, as outliers under each of the random effects models 4–9. Observation 50 appears to be an outlier in models 6–9 which have a spatial effect, but not in the remaining models. Further investigation reveals that county 50 has only 6 cases compared to 19.6 expected, whilst each of its 3 neighbouring counties have high observed counts (17, 16, 16) relative to expected (7.8, 10.5, 14.4). The spatial prior in models 6–9 causes the estimated rate in county 50 to be smoothed towards the mean of its neighbours' rates, thus leading to the discrepancy between observed and fitted values. However since the observation still exercises considerable weight on its fitted value the leverage is high as well.

7 Further numerical examples

7.1 Seeds: random effects logistic regression with alternative priors

Crowder (1978) presents an analysis of the proportion of seeds that germinate on each of 21 plates arranged according to a 2×2 factorial layout depending on binary variables seed type (x_1) and root extract (x_2). The number of seeds r_i which germinate out of the n_i seeds on plate i are assumed to follow a binomial distribution: $r_i \sim \text{Binomial}(p_i, n_i)$. We compare the following logistic regression models for $\theta_i = \text{logit}(p_i)$:

$$
\begin{array}{lrcl}
\text{Model 1:} & \theta_i & = & \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_{1i} x_{2i} \\
\text{Model 2:} & \theta_i & = & \alpha_i \\
\text{Model 3:} & \theta_i & = & \beta_0 + b_i \\
\text{Model 4:} & \theta_i & = & \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_{1i} x_{2i} + b_i
\end{array}
$$

where locally uniform priors (actually Normal with mean 0 and precision 0.00001) are placed on β_0, β_1 , β_2 , β_{12} and α_i, $i = 1, ..., 21$, and b_i are exchangeable random effects with a Normal prior distribution having zero mean and precision τ. Four alternative prior specifications were considered for the random effects precision:

$$
\begin{array}{lcl}
\text{Prior A:} \quad \tau & \sim & \text{Gamma}(0.001, 0.001) \\
\text{Prior B:} \quad \tau & \sim & \text{Gamma}(3, 1) \\
\text{Prior C:} \quad \tau & \sim & \text{Pareto}(0.5, 1) \\
\text{Prior D:} \quad \tau & \sim & \text{Pareto}(0.5, 4)
\end{array}
$$

Prior A is 'just' proper but diffuse, having mean 1 and variance 1000. Prior B is proper and has mean 3 — see Smith *et al.* (1995) for an argument in favour of this prior. Priors C and D are

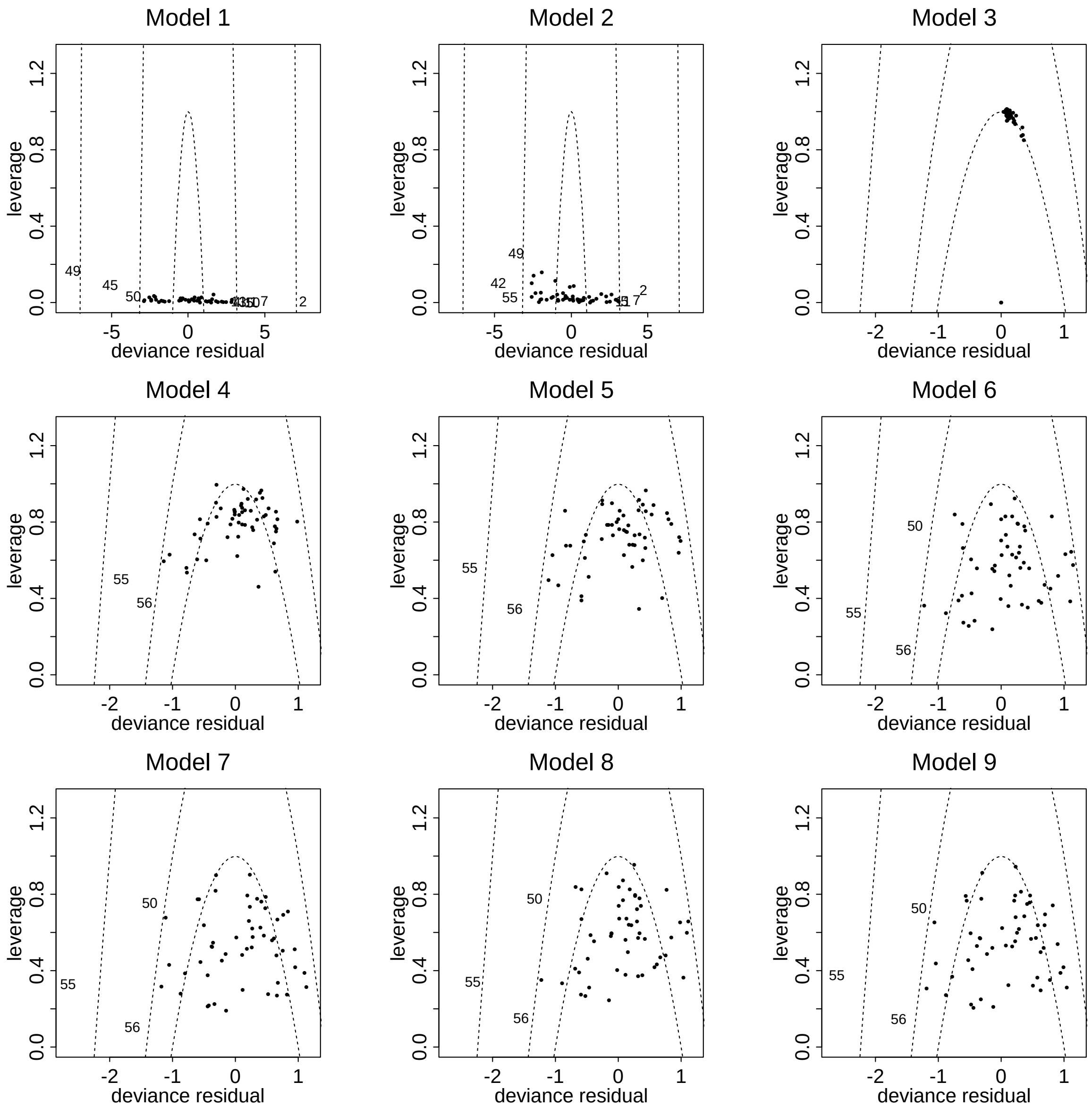

Figure 2: Posterior distributions of the deviance for each model considered in the lip cancer example

Model	$\overline{D_S}$	$D_S(\overline{\theta})$	p_D	DIC	σ	(95% interval)
1	37.4	33.3	4.1	41.5		
2	20.8	.3	20.5	41.3		
3A	23.9	8.1	15.8	39.7	0.68	(0.41, 1.09)
3B	24.0	8.4	15.6	39.6	0.65	(0.45, 0.92)
3C	23.6	7.2	16.4	40.0	0.70	(0.44, 0.97)
3D	27.9	14.9	13.0	40.9	0.45	(0.35, 0.50)
4A	25.9	14.9	11.0	36.9	0.29	(0.09, 0.55)
4B	20.9	6.7	14.2	35.1	0.48	(0.33, 0.69)
4C	23.7	11.7	12.0	35.7	0.34	(0.07, 0.65)
4D	24.0	13.0	11.0	35.0	0.31	(0.08, 0.49)

Table 2: Deviance and posterior summaries for the seed germination data.

equivalent to a uniform prior on (0, 1) or (0, 0.5) respectively for $\sigma = \sqrt{\frac{1}{\tau}}$, the standard deviation of the random effects. We emphasise that such a list of priors must be made strictly external to the information in the data, otherwise the model comparison criterion will be unreasonably favourable.

For this binomial model we adopt the saturated deviance (McCullagh and Nelder, 1989)[p. 34]

$$D_S(\theta) = 2\sum_i \left[r_i \log \frac{r_i}{n_i p_i} + (r_i - n_i) \log \frac{1 - r_i/n_i}{1 - p_i} \right]$$

obtained by taking $p_i = \frac{e^{\theta_i}}{1+e^{\theta_i}}$ and $-2\log f(r) = -2\sum_i \log p(r_i|\theta_i = \text{logit}\frac{r_i}{n_i}) = 76.6$ as the standardising factor.

For each model and prior we ran an MCMC sampler in BUGS for 5000 interations following a burn-in period of 1000 iterations. The resulting deviance summaries are given in Table 2.

The estimates of p_D for the two non-hierarchical models (4.1 for model 1 and 20.5 for model 2) closely approximate the actual number of parameters in each model (4 and 21 respectively). The estimates of p_D for model 3 imply that the 21 random effects contribute approximately 12-15 effective parameters depending on the prior (plus 1 parameter for β_0). However, including the three covariate effects in model 4 results in a net *decrease* in p_D: the 21 random effects now contribute only 7-10 effective parameters, since the covariates explain a substantial amount of the between plate variation. This conclusion is supported by examination of DIC, which favours model 4 (under any of the four priors) over the other models considered.

Turning to the comparison of priors for τ, we reach different conclusions depending on the linear predictor. For model 3 (which includes only an intercept term plus the random effects), DIC makes little distinction between priors A, B and C, but prefers each over prior D. Examination of the posterior distributions for $\sigma = \sqrt{\frac{1}{\tau}}$ explains why: priors A, B and C yield similar estimates of σ, with a 95% interval of approximately 0.4 – 1.0. However, prior D does not support values of $\sigma > 0.5$, thus constraining the random effects to be less variable than the data suggest. Including the covariate and interaction effects in the linear predictor of model 4 explains a considerable amount of the between plate variation, with a corresponding reduction in the estimated size of σ under all four priors. Prior D now provides support across the plausible range of values for σ, and hence is no longer penalised by DIC. By contrast, the greater uncertainty associated with the 'just proper' prior A results in a higher value of DIC compared to the remaining 3 priors.

From the perspective of absolute fit, comparing $\overline{D}_S$ with $n = 21$ shows the hierarchical models are reasonable with 4B being particularly favoured. These data are also used by Lee and Nelder (1996) to illustrate their method for estimating the scaled deviance and degrees of freedom of hierarchical generalised linear models, where our model 4A corresponds to the GLMM model shown in their Table 2. Their estimate of 10.0 degrees of freedom is identical to our value of $n - p_D = 21 - 11.0 = 10.0$, reinforcing the analysis of Section 5.3.3. Their estimated scaled deviance of 12.0 contrasts with our $D_S(\overline{\theta}) = 14.9$ (however, their scaled deviance is based on direct estimates of the mean p, rather than on estimates of the canonical parameters, and if we use the equivalent mean parameterisation we obtain $D_S(\overline{\theta}) = 12.9$.)

From the perspective of absolute fit, comparing $\overline{D}_S$ with n shows the hierarchical models are reasonable with 4A being particularly favoured. These data are also used by Lee and Nelder (1996) to illustrate their method for estimating the scaled deviance and degrees of freedom of hierarchical generalised linear models, where our model 4A corresponds to the GLMM model shown in their Table 2. Their estimate of 10.0 degrees of freedom is identical to our value of $n - p_D = 21 - 11.0 = 10.0$, reinforcing the analysis of Section 5.3.3. Their estimated scaled deviance of 12.0 contrasts with our $D_S(\overline{\theta}) = 14.9$ (however, their scaled deviance is based on estimates of the mean μ of y, and if we use the mean parameterisation we obtain $D_S(\overline{\theta}) = 12.9$.)

7.2 Stacks: robust regression

Spiegelhalter *et al.* (1996*b*)[pp.27–29] consider a variety of different error structures for the oft-analyzed stack loss data of Brownlee (1965). Here the response variable (y), the amount of stack loss (escaping ammonia in an industrial application), is regresssed on three predictor variables: air flow (x_1), temperature (x_2), and acid concentration (x_3). Assuming the usual linear regression structure

$$\mu_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3}$$

where $z_{ij} = (x_{ij} - \bar{x}_{.j})/sd(x_{.j})$, the standardized covariates, the presence of a few prominent outliers amongst the $n = 21$ cases motivates comparison of the following four error distributions:

$$\begin{array}{ll} \text{Model 1:} & y_i \sim \text{Normal}(\mu_i, \tau^{-1}) \\ \text{Model 2:} & y_i \sim \text{DE}(\mu_i, \tau^{-1}) \\ \text{Model 3:} & y_i \sim \text{Logistic}(\mu_i, \tau^{-1}) \\ \text{Model 4:} & y_i \sim \text{t}_d(\mu_i, \tau^{-1}) \end{array}$$

where DE denotes the double exponential (Laplace) distribution, and t_d denotes the Student's t distribution with d degrees of freedom.

Unlike our other examples the form of the likelihood changes with each model, so we must take care with normalizing constants when computing null deviances (- 2 log(likelihoods)). These emerge as follows:

$$\begin{array}{l} D_{01} = \sum_{i=1}^{n} \left[\tau(y_i - \mu_i)^2 - \log(\frac{\tau}{2\pi})\right] \\ D_{02} = \sum_{i=1}^{n} \left[2\tau|y_i - \mu_i| - 2\log(\frac{\tau}{2})\right] \\ D_{03} = \sum_{i=1}^{n} \left[4\log(1 + e^{\tau(y_i-\mu_i)}) - 2\log\tau - 2\tau(y_i - \mu_i)\right] \\ D_{04} = \sum_{i=1}^{n} \left\{(d+1)\log\left[1 + \frac{\tau}{d}(y_i - \mu_i)^2\right] - \log(\frac{\tau}{d\pi}) - 2\log\Gamma(\frac{d+1}{2}) + 2\log\Gamma(\frac{d}{2})\right\} \end{array}$$

A well-known alternative to direct fitting of many symmetric but nonnormal error distributions is through scale mixtures of normals (Andrews and Mallows, 1974). From p.210 of Carlin and Louis

(1996), we have the alternate t_d formulation

$$\text{Model 5:} \quad y_i \sim \text{Normal}(\mu_i, \tfrac{1}{w_i \tau}), \; w_i \sim \tfrac{1}{d}\chi^2_d = \text{Gamma}(\tfrac{d}{2}, \tfrac{d}{2}) \;,$$

and corresponding null deviance expression

$$D_{05} = \sum_{i=1}^{n} \left[w_i \tau (y_i - \mu_i)^2 - \log\left(\frac{w_i \tau}{2\pi}\right)\right] \;,$$

a simple modification of the Gaussian expression.

Model	$\overline{D}_0$	$D_0(\overline{\theta})$	p_D	DIC
1 Normal	110.1	105.0	5.1	115.2
2 DE	107.9	102.3	5.6	113.5
3 Logistic	109.5	104.2	5.3	114.8
4 t_4	108.7	103.2	5.5	114.2
5 t_4 as scale mixture	102.1	94.5	7.6	109.7

Table 3: Deviance results for stack loss data.

Following Spiegelhalter *et al.* (1996*b*) we set $d = 4$, and for each model we placed essentially flat priors (actually normal with mean 0 and precision 0.00001) on the β_j, a vague Gamma(0.001, 0.001) prior on τ, and ran the Gibbs sampler in BUGS for 5000 iterations following a burn-in period of 1000 iterations.

Replacing τ and w_i by their posterior means where necessary for the $D(\bar{\theta})$ calculation, the resulting deviance summaries are shown in Table 3 (note that the mean parametrization and the canonical parametrization are equivalent here, since the mean μ_i is a linear function of the canonical β parameters). Beginning with a comparison of the first four models, the estimates of p_D are all just over 5, the correct number of parameters for this example. The DIC values imply that Model 2 (double exponential) is best, followed by the t_4, the logistic, and finally the normal. Clearly this order is consistent with the models' respective abilities to accommodate outliers.

Turning to the normal scale mixture representation for the t_4 likelihood (Model 5), the p_D value is 7.6, suggesting that the w_i random effects contribute only an extra 2 to 2.5 parameters. However, the model's smaller DIC value implies that the extra mixing parameters are "worth it" in an overall quality of fit sense. We emphasize that the results from Models 4 and 5 need not be equal since, while they lead to the same marginal likelihood for the y_i, they correspond to different prediction problems.

Finally, plots of deviance residuals versus leverages (not shown) clearly identify the observations determined to be 'outlying' by several previous authors analysing this dataset.

7.3 Panel: Longitudinal binary observations

To illustrate how the mean and canonical parameterisations (introduced in Section5.2 and further discussed in Section 9) can sometimes lead to different conclusions, our next example considers a subset of data from the Six Cities study, a longitudinal study of the health effects of air pollution (see Fitzmaurice and Laird (1993) for the data and a likelihood-based analysis). The data consist of repeated binary measurements y_{ij} of the wheezing status (1=yes, 0=no) of child i at time j,

Model	$\overline{D}_0$	Canonical parametrization $D_0(\overline{\theta})$	p_D	DIC	Mean parametrization $D_0(\overline{\theta})$	p_D	DIC
1 logit	1166.4	917.7	248.7	1415.1	997.5	168.9	1335.3
2 probit	1148.6	885.9	262.7	1411.3	989.9	158.7	1307.3
3 cloglog	1180.9	956.5	224.4	1405.3	1013.7	167.2	1348.1

Table 4: Results for both parameterizations of Bernoulli panel data

$i = 1, \ldots, I$, $j = 1, \ldots J$, for each of $I = 537$ children living in Stuebenville, Ohio at $J = 4$ timepoints. We are given two predictor variables: a_{ij}, the age of child i in years at measurement point j (7, 8, 9, or 10 years), and s_i, the smoking status of child i's mother (1=yes, 0=no). Following the Bayesian analysis of Chib and Greenberg (1998), we adopt the conditional response model

$$\begin{aligned} Y_{ij} &\sim \text{Bernoulli}(p_{ij}) \\ p_{ij} &\equiv Pr(Y_{ij} = 1) \;=\; g^{-1}(\mu_{ij}) \\ \mu_{ij} &= \beta_0 + \beta_1 z_{ij1} + \beta_2 z_{ij2} + \beta_3 z_{ij3} + b_i \ , \end{aligned}$$

where $z_{ijk} = (x_{ijk} - \bar{x}_{..k})$, $k = 1, 2, 3$, and $x_{ij1} = a_{ij}$, $x_{ij2} = s_i$, and $x_{ij3} = a_{ij}s_i$, a smoking-age interaction term. The b_i are individual-specific random effects, initially given an exchangeable $N(0, \tau)$ specification, which allow for dependence among the longitudinal responses for child i. The model choice issue here is to determine the most appropriate link function $g(.)$ among three candidates, namely the logit, the probit, and the complementary log-log. More formally, our three models are

$$\begin{aligned} &\text{Model 1:} \quad g(p_{ij}) = \text{logit}(p_{ij}) = \log[p_{ij}/(1 - p_{ij})] \\ &\text{Model 2:} \quad g(p_{ij}) = \text{probit}(p_{ij}) = \Phi^{-1}(p_{ij}) \\ &\text{Model 3:} \quad g(p_{ij}) = \text{cloglog}(p_{ij}) = \log[-\log(1 - p_{ij})] \end{aligned}$$

Since the Bernoulli likelihood is unaffected by this choice, in all cases the null deviance takes the simple form

$$D_0 = -2 \sum_{i,j} [y_{ij} \log(p_{ij}) + (1 - y_{ij}) \log(1 - p_{ij})] \ .$$

Placing flat priors on the β_k, a vague Gamma(0.001, 0.001) prior on τ, and running the Gibbs sampler for 5000 iterations following a burn-in period of 1000 iterations produces the deviance summaries in Table 4 for the canonical and mean parametrizations, respectively, where the canonical parametrization constructs $\overline{\theta}$ as the mean of the linear predictors β and b_i, and then uses the appropriate linking transformation (logit, probit, or cloglog) to obtain the imputed means for the p_{ij}. The mean parametrization simply uses the means of the p_{ij} themselves when computing $D(\overline{\theta})$. DIC prefers the cloglog link under the canonical parametrization, but the probit link under the mean parametrization: such disagreement is unfortunate but is quite feasible with Bernoulli distributions in which the posterior distributions of the fitted means will be highly non-normal. We repeat that we prefer the canonical results due to the improved normality of the posterior distributions.

7.4 CAMCOG: Informative missing data

Best *et al.* (1996) present an analysis of longitudinal data from a study of dementia and cognitive decline in the elderly. 365 women aged 70–79 were interviewed at the start of the study, and were assessed for cognitive function using the CAMCOG test (a neuropsychological assesment scale taking values between 0 and 106). A repeat interview was conducted 5 years later, but only 237 (65%) of the original cohort were re-assessed. The remaining women had either died, moved away or refused to particpate. Best *et al.* (1996) argue that the drop-out mechanism may be non-ignorable and must be modelled explicitly. Following their published analysis, we consider a linear regression model for cognitive function at the second interview:

$$\begin{aligned} y_{2i} &\sim \text{Normal}(\mu_i, \tau^{-1}) \\ \mu_i &= \alpha + \beta y_{1i} + \theta x_i \end{aligned}$$

where y_{ki} is the CAMCOG score for subject i at interview k, and x_i represents age at second interview. We assume vague but proper priors for the regression coefficients α, β and θ (actually independent Normal with mean 0 and precision 0.00001)) and the inverse measurement variance τ (actually Gamma(0.001, 0.001)). A logistic selection model was specified for the non-response mechanism:

$$\begin{aligned} m_i &\sim \text{Bernoulli}(p_i) \\ \text{logit}(p_i) &= \psi + \phi(y_{2i} - 67.5) \end{aligned}$$

where m_i is a binary variable indicating if the CAMCOG score for subject i was missing at interview 2. We compare four alternative prior assumptions for the parameters of this model:

$$\begin{array}{lccl} \text{Prior 1:} & \psi & \sim & \text{Normal}(0.0, 10000) \\ & \phi & = & 0 \\ \text{Prior 2:} & \psi & = & -0.3 \\ & \phi & = & -0.035 \\ \text{Prior 3:} & \psi & \sim & \text{Normal}(-0.3, 0.64) \\ & \phi & \sim & \text{Normal}(-0.035, 0.0003) \\ \text{Prior 4:} & \psi & \sim & \text{Normal}(0.0, 1.0) \\ & \phi & \sim & \text{Normal}(0.0, 1.0) \end{array}$$

Prior 1 corresponds to a non-informative dropout mechanism; priors 2 and 3 represent informative prior distributions elicited from an expert psychiatrist (obtained by fitting a logistic curve to the expert's best guess at the proportion of women she would expect to drop out given different true CAMCOG scores); prior 4 is a diffuse prior on the unknown parameters of the informative dropout model.

Since it was necessary to augment the observed dataset with the missing value indicator $\mathbf{m}$ in order to explicitly model the dropout mechanism, we must take account of this additional 'data' when calculating the model null deviance, i.e.

$$D_0 = \sum_{i=1}^{237} \log 2\pi - \log\tau + \tau(y_{2i}^{obs} - \mu_i)^2 + \sum_{i=1}^{365} m_i \log p_i + (1 - m_i)\log(1 - p_i)$$

The resulting deviance summaries for each prior are shown in Table 5.

Considering first the deviance contribution from the missing data sub-model, we see that p_D closely approximates the true number of parameters for all 4 priors, and that DIC clearly rejects the non-informative dropout model (prior 1) in favour of the three informative missing data models (priors 2–4). Prior 2, in which the coefficients θ and ϕ are fixed at the expert's prior values, is slightly prefered over priors 3 and 4, in which the coefficients are assumed to be unknown. In fact, the

Model	$\overline{D}$	$D(\overline{\theta})$	p_D	DIC
Deviance contribution from y_{2i}^{obs}				
1	1639.1	1635.4	3.7	1642.8
2	1639.1	1635.7	3.4	1642.5
3	1639.1	1635.5	3.6	1642.7
4	1639.0	1635.6	3.4	1642.4
Deviance contribution from m_i				
1	455.7	454.7	1.0	456.7
2	444.4	444.3	0.1	444.5
3	445.0	443.2	1.8	446.8
4	445.1	443.2	1.9	446.9
Total deviance				
1	2094.8	2090.1	4.7	2099.5
2	2083.5	2080.0	3.5	2087.0
3	2084.1	2078.7	5.4	2089.5
4	2084.1	2078.8	5.3	2089.4

Table 5: Deviance summaries, CAMCOG data

posterior distributions for θ and ϕ under priors 3 and 4 are close to the expert's prior values, implying that her prior guess at the non-response rates as a function of true CAMCOG score was very accurate. Consequently, the additional complexity imposed by treating θ and ϕ as unknown in priors 3 and 4 is uneccessary, and is penalised by DIC.

The deviance contributions from the observed response data $y_{2i}, i = 1, ...237$ result in nearly identical values of DIC and p_D, irrespective of the prior mechanism assumed for the missing data. This seems reasonable, since the likelihood terms for this part of the model are identical under all four priors. Consequently, the differences in total DIC and p_D for each model simply reflect the differences associated with the missing data sub-models.

8 Methods for model comparison

From an applied viewpoint, our aim lies not in simply determining the effective dimension of a model, but in using the deviance to help compare competing (and perhaps nonnested) models of varying types. We are therefore adding to the long list of criteria that have been suggested for comparing or choosing between models: following the nomenclature of Dempster (1997*a*), these can be broadly distinguished into *predictive* criteria that compare the ability of prior and model assumptions to predict the currently observed data, and *postdictive* criteria that assess assumptions conditional on the observed data.

Predictive criteria: The basis for such criteria can be thought of as a sequential series of predictive statements (Dawid, 1984) which, if a full probability model is being assessed, becomes the marginal likelihood $p(y) = \int p(y|\phi)p(\phi)d\phi$. The resulting Bayes factors (Kass and Raftery, 1995) may be used to obtain posterior probabilities of competing models. However, in the discussion of Aitkin (1991), Smith pointed out that this formulation may only be appropriate in circumstances where it was really believed one and only one of the competing models were in fact 'true', and the crucial issue was to choose this correct model: this discussion is elaborated in Bernardo and Smith

(1994)[chapter 6]. We would argue that this is not the situation for the kind of examples discussed in this paper; we neither believe any of the models is actually true, nor do we wish to formulate a decision problem of strict model choice.

For a non-hierarchical model with p parameters and n observations, the Bayes (or Schwarz) information criterion (Schwarz, 1978) given by BIC $= -2\log p(y|\hat{\theta}) + p\log n$ has been widely promoted, but its implementation for hierarchical models has been controversial. This is due to the uncertainty concerning the proper values of both p and n in such situations. For instance, in our basic one-parameter exponential family hierarchical model from Section 5.2, should n be chosen as $\sum_{i=1}^{I} n_i$, the total number of observations, or I, the number of groups? If the observations within each group are independent, then the former choice appears more sensible, whereas if they are perfectly correlated, the latter is more appropriate. In practice, the true level of within-group correlation will be somewhere in between, so the choice of n should be as well. Unfortunately, asymptotic theory seems of little help here; for instance, in the context of normal linear hierarchical models, Pauler (1998) shows that even the two extreme choices for n above are defensible asymptotically. In this same context, SAS `Proc MIXED` currently employs the smaller choice, in the interest of conservatism (i.e. retaining predictors in the final model; the larger choice may often "go too far" in penalizing larger models). Working instead in the Cox survival model setting, Volinsky (1997), (unpublished Univ. of Washington PhD dissertation) obtains a similar ambiguous conclusion, where now the choice for n is between the number of patients in the study and the number of events (deaths) observed – though he suggests use of the latter, based on simulation work and analytic results from a simpler, exponential survival model. It seems possible that p_D might have a role to play in adapting BIC to hierarchical models.

Postdictive criteria: There have been a number of recent suggestions for comparing models in the light of observed data, usually based on estimates of their predictive ability on a replicate dataset.

Aitkin (1991) suggested using the posterior mean of the *likelihood*, and contrasts the resulting posterior Bayes factors (PDF) with AIC and BIC (Aitkin, 1997). We feel happier with our use of the *log*-likelihood in that the log probability ordinate is a proper scoring rule for evaluating predictions (Dawid, 1986), in contrast to the ordinate itself. In addition, the resulting penalty for complexity in the PDF appears insufficient.

Laud and Ibrahim (1995) and Gelfand and Ghosh (1998) suggest minimising a predictive "discrepancy measure" $E[d(\mathbf{y}_{new}, \mathbf{y}_{obs})|\mathbf{y}_{obs}]$, where $\mathbf{y}_{new}$ is a draw from the posterior predictive distribution $p(\mathbf{y}_{new}|\mathbf{y}_{obs})$, and we might for instance take $d(\mathbf{y}_{new}, \mathbf{y}_{obs}) = (\mathbf{y}_{new} - \mathbf{y}_{obs})^T(\mathbf{y}_{new} - \mathbf{y}_{obs})$. These authors show their measures also have attractive interpretations as weighted sums of "goodness of fit" and "predictive variability penalty" terms. However, proper choice of the criterion requires fairly involved analytic work, as well as several subjective choices about the utility function appropriate for the problem at hand. Furthermore, the one-way ANOVA model in Section 5.1.1 gives rise to a fit term equivalent to $D(\overline{\theta})$, and a predictive variability term equal to $p_D + p$. Thus their suggestion is equivalent in this context to comparison by $\overline{D}$ which, although invariant to parameterisation, again does not seem to sufficiently penalize complexity.

Ye (1998) and Ye and Wong (1998) extend Akaike by arguing that predictive error on a replicate dataset is minimised by penalizing –2 log-likelihood by 2 GOF, where GOF is the 'generalised degrees of freedom' measuring the expected sensitivity of the fitted values to their corresponding observed values: essentially the expected leverage. This is extremely similar to our proposal, although they directly identify measures of leverage as the necessary components of the effective number of parameters, rather than as a consequence of a more general formulation. We should therefore expect our and Ye's proposals to give similar conclusions, although we note that his computational methods require direct estimation of leverage through a series of simulated or systematic perturbations of the data followed by repeated model fitting.

Ripley (1996)[pp 32-35, 140-141] discusses the general problem of model comparison with particular reference to the structure of neural networks, and shows that previous suggestions for the effective number of parameters made by Murata *et al.* (1994), Moody (1992) and others are all essentially equivalent to the following definition of the 'effective number of parameters' p^*. Let $p(y|\theta_0)$ be the 'closest' distribution in the assumed model to the unobservable true model $p(y)$, where closest is in terms of minimising Kullback-Leibler distance, and let $L_1(y, \theta)$ be some function of the data and parameters that we seek to minimise. Then

$$p^* = \mathrm{tr}(KJ^{-1}),$$

where

$$J = \mathrm{E}_Y \left[\frac{\delta^2 L_1(Y,\theta)}{\delta\theta^2} \right]_{\theta_0}, \quad K = \mathrm{Var}_Y \left[\frac{\delta L_1(Y,\theta)}{\delta\theta} \right]_{\theta_0}$$

where we emphasise that the expectation and variance are now with respect to the unknown true sampling distribution. If we seek to maximise the joint likelihood of the data and parameters, i.e. $-L_1 = \log p(y|\theta) + \log p(\theta)$, and are willing to assume the true model is a member of $p(y|\theta)$ (which may be reasonable in a sufficiently parameterised model), then it is straightforward to show that $V^{-1} \to nJ, -L''_{\overline{\theta}} \to nK$, and hence $p_D \approx p^*$.

9 Conclusion: a brief appraisal of issues arising with DIC

We now attempt a brief summary of the disadvantages and advantages of using p_D and DIC in routine statistical analysis.

A major problem in using p_D is that it is only asymptotically invariant to the chosen parameterisation, since different fitted deviances $D(\overline{\theta})$ may arise from substituting posterior means of alternative choices of θ. For example,

- In the one-parameter exponential family, different results will be obtained by using the mean or canonical parameterisation. The decision-theoretic justification and the asymptotic identities for DIC are all improved by approximate posterior normality, which will be better achieved by using parameters defined on the whole real line. This is the basis for recommending in Section 5.2 that the canonical parameterisation is the most approriate choice, although the panel example in Section 7.3 showed this choice could be important with Bernoulli data.

- In models with unknown scale parameters, there will be some dependence on whether to base $D(\overline{\theta})$ on the posterior means of the standard deviations, variances, precisions, log-precisions, or some other choice. In theory one could achieve invariance to the choice of nuisance parameters ψ by re-running the sampler conditional on the posterior means of the primary parameters θ, and using the resulting mean deviance $E_{\psi|\overline{\theta},y} D(\overline{\theta}, \psi)$ for $D(\overline{\theta})$ in the calculation of p_D; this would now be the effective number of primary parameters.

  Since approximate posterior normality is desirable, we would argue that log-precision is the most appropriate scale, but fortunately the choice seems to make little difference in model comparison. For example, suppose we have a normal sample $y_i \sim \mathrm{N}(\theta, \sigma^2), i = 1, ..., n,$

with standard locally uniform priors on θ and $\log \sigma$. It can be shown, up to $O(n^{-2})$, that compared to the correct answer of 2, parameterising in terms of $\log \sigma, \sigma^{-2}$ and σ^2 gives p_D's of $2 - \frac{1}{6n}, 2 + \frac{4}{3n}$ and $2 - \frac{8}{3n}$ respectively. Hence the parameterisation $\log \sigma$ is preferable, but by a rather small margin.

- As we have seen in the stacks example of Section 7.2, there may be sensitivity to apparent innocuous re-structuring of the model: this is to be expected since DIC is not a function of the marginal likelihood of the data, and hence changes that might not change the Bayes factor *do* change DIC. By making such changes one is altering the definition of a replicate dataset, and hence one would expect DIC to change.

We have shown that our suggestion is strongly related to a range of previous proposals for postdictive model choice, but has the additional advantages that it

- is completely general to any class of model
- involves little additional analytic work
- involves no extra Monte Carlo sampling
- offers an effective model screening technique that seems natural for a broad class of models and performs reasonably across a range of examples
- is already implemented in the test version of WinBUGS, the most general Bayesian software package to date, and could be easily added to any other MCMC-based package
- represents an attractive compromise between very informal model comparison methods (*e.g.* plotting posterior distributions of log-likelihoods), which are hard to use and interpret, and overly formal, decision theoretic methods (*e.g.* Bayes factors, or posterior predictive discrepancy measures), which are predicated on a choice of a single model and require substantial analytical work and subjective choices, such as the selection of an appropriate utility function.

In conclusion, we feel that p_D and DIC deserve further investigation as tools for model comparison.

Acknowledgements

We are very grateful for the generous discussion and criticism of the participants in the programme on Neural Networks and Machine Learning held at the Isaac Newton Institute for Mathematical Sciences in 1997, and to Andrew Thomas for so quickly implementing our changing ideas into WinBUGS. BPC received partial support from National Institute of Allergy and Infectious Diseases (NIAID) Grant 1-R01-AI41966.

References

Abramowitz, M. and Stegun, I. (1970). *Handbook of Mathematical Functions, 9th ed.* Dover, New York.

Aitkin, M. (1991). Posterior Bayes factors (with discussion). *Journal of the Royal Statistical Society, Series B*, **53**, 111–43.

Aitkin, M. (1997). The calibration of P-values, posterior Bayes factors and the AIC from the posterior distribution of the likelihood. *Statistics and Computing*, **7**, 253–62.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd Intl. Symp. on Information Theory*, (ed. B. Petrov and F. Csáki). Akadémiai Kiadó, Budapest.

Andrews, D. and Mallows, C. (1974). Scale mixtures of normality. *Journal of the Royal Statistical Society, Series B*, **36**, 99–102.

Bernardo, J. M. (1979). Expected information as expected utility. *Annals of Statistics*, **7**, 686–90.

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

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B*, **36**, 192–236.

Best, N. G., Spiegelhalter, D. J., Thomas, A., and Brayne, C. E. G. (1996). Bayesian analysis of realistically complex models. *Journal of the Royal Statistical Society, Series A*, **159**, 323–42.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Addison–Wesley.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal American Statistical Association*, **88**, 9–25.

Brownlee, K. A. (1965). *Statistical Theory and Methodology in Science and Engineering.* John Wiley and Sons, New York.

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

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, **85**, (to appear).

Clayton, D. G. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardised relative risks for use in disease mapping. *Biometrics*, **43**, 671–81.

Crowder, M. J. (1978). Beta-binomial Anova for proportons. *Applied Statistics*, **27**, 34–7.

Dawid, A. P. (1984). Statistical theory - the prequential approach. *Journal of the Royal Statistical Society, Series A*, **147**, 277–305.

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.

Dempster, A. P. (1974). The direct use of likelihood for significance testing. In *Proceedings of Conference on Foundational Questions in Statistical Inference*, (ed. O. Barndorff-Nielsen, P. Blaesild, and G. Schou), pp. 335–52. Department of Theoretical Statistics: University of Aarhus.

Dempster, A. P. (1997*a*). Commentary on the paper by Murray Aitkin, and on discussion by Mervyn Stone. *Statistics and Computing*, **7**, 265–9.

Dempster, A. P. (1997*b*). The direct use of likelihood for significance testing. *Statistics and Computing*, **7**, 247–52.

Draper, D. (1995). Assessment and propogation of model uncertainty. *Journal of the Royal Statistical Society, series B*, **57**, 45–97.

Fitzmaurice, G. and Laird, N. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, **80**, 141–51.

Gelfand, A. and Ghosh, S. (1998). Model choice: a minimum posterior predictive loss approach. *Biometrika.*

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B*, **56**, 501–14.

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

Gilks, W. R., Wang, C. C., Coursaget, P., and Yvonnet, B. (1993). Random-effects models for longitudinal data using gibbs sampling. *Biometrics*, **49**, 441–53.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman and Hall, London.

Hodges, J. and Sargent, D. (1998). Counting degrees of freedom in hierarchical and other richly-parameterised models. Technical report, Division of Biostatistics, University of Minnesota, USA.

Kass, R. and Raftery, A. (1995). Bayes factors and model uncertainty. *Journal of the American Statistical Association*, **90**, 773–95.

Laird, N. M. and Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics*, **38**, 963–74.

Laud, P. and Ibrahim, J. (1995). Predictive model selection. *Journal of the Royal Statistical Society, Series B*, **57**, 247–62.

Lee, Y. and Nelder, J. (1996). Hierarchical generalised linear models (with discussion). *Journal of the Royal Statistical Society, Series B*, **58**, 619–78.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society, Series B*, **34**, 1–44.

MacKay, D. J. C. (1992). Bayesian interpolation. *Neural Computation*, **4**, (3), 415–47.

MacKay, D. J. C. (1995). Probable networks and plausible predictions - a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, **6**, 469–505.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, 2nd edition.* Chapman and Hall, London.

Moody, J. E. (1992). The *effective* number of parameters: An analysis of generalization and regularization in nonlinear learning systems. In *Advances in Neural Information Processing Systems 4*, (ed. J. E. Moody, S. J. Hanson, and R. P. Lippmann), pp. 847–54. Morgan Kaufmann, San Mateo, California.

Murata, N., Yoshizawa, S., and Amari, S. (1994). Network information criterion - determining the number of hidden units for artificial neural network models. *IEEE Transactions on Neural Networks*, **5**, 865–72.

Pauler, D. (1998). The Schwarz criterion and related methods for normal linear models. *Biometrika*, **85**, to appear.

Raghunathan, T. E. (1988). A Bayesian model selection criterion. Technical report, University of Washington.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with unknown number of components (with discussion). *Journal of the Royal Statistical Society, Series B*, **59**, 731–92.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge, UK.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–4.

Smith, T. C., Spiegelhalter, D. J., and Thomas, A. (1995). Bayesian graphical modelling applied to random effects meta-analysis. *Statistics in Medicine*, **14**, 2685–99.

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Gilks, W. R. (1996*a*). *BUGS: Bayesian inference Using Gibbs Sampling, Version 0.5, (version ii)* . MRC Biostatistics Unit, Cambridge.

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Gilks, W. R. (1996*b*). *BUGS Examples Volume 1, Version 0.5, (version ii).* MRC Biostatistics Unit, Cambridge.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, **93**, 120–31.

Ye, J. and Wong, W. (1998). Evaluation of highly complex modeling procedures with binomial and poisson data. Technical report, Graduate School of Business, University of Chicago.

Zeger, S. L. and Karim, M. R. (1991). Generalised linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association*, **86**, 79–86.