J. R. Statist. Soc. B (2002)
64, *Part* 4, *pp.* 583–639

Bayesian measures of model complexity and fit

David J. Spiegelhalter,

Medical Research Council Biostatistics Unit, Cambridge, UK

Nicola G. Best,

Imperial College School of Medicine, London, UK

Bradley P. Carlin

University of Minnesota, Minneapolis, USA

and Angelika van der Linde

University of Bremen, Germany

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, March 13th, 2002*, Professor D. Firth *in the Chair*]

Summary. We consider the problem of comparing complex hierarchical models in which the number of parameters is not clearly defined. Using an information theoretic argument we derive a measure p_D for the effective number of parameters in a model as the difference between the posterior mean of the deviance and the deviance at the posterior means of the parameters of interest. In general p_D approximately corresponds to the trace of the product of Fisher's information and the posterior covariance, which in normal models is the trace of the 'hat' matrix projecting observations onto fitted values. Its properties in exponential families are explored. The posterior mean deviance is suggested as a Bayesian measure of fit or adequacy, and the contributions of individual observations to the fit and complexity can give rise to a diagnostic plot of deviance residuals against leverages. Adding p_D to the posterior mean deviance gives a *deviance information criterion* for comparing models, which is related to other information criteria and has an approximate decision theoretic justification. The procedure is illustrated in some examples, and comparisons are drawn with alternative Bayesian and classical proposals. Throughout it is emphasized that the quantities required are trivial to compute in a Markov chain Monte Carlo analysis.

Keywords: Bayesian model comparison; Decision theory; Deviance information criterion; Effective number of parameters; Hierarchical models; Information theory; Leverage; Markov chain Monte Carlo methods; Model dimension

1. Introduction

The development of Markov chain Monte Carlo (MCMC) methods 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). This ability naturally leads us to wish to compare alternative model formulations with the aim of identifying a class of succinct models which appear to describe the information in the data adequately: for example, we might ask whether we need to incorporate

Address for correspondence: David J. Spiegelhalter, Medical Research Council Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge, CB2 2SR, UK.
E-mail: david.spiegelhalter@mrc-bsu.cam.ac.uk

 1369–7412/02/64583

a random effect to allow for overdispersion, what distributional forms to assume for responses and random effects, and so on.

Within the classical modelling framework, model comparison generally takes place by defining a measure of *fit*, typically a deviance statistic, and *complexity*, the number of free parameters in the model. Since increasing complexity is accompanied by a better fit, models are compared by trading off these two quantities and, following early work of Akaike (1973), proposals are often formally based on minimizing a measure of expected loss on a future replicate data set: see, for example, Efron (1986), Ripley (1996) and Burnham and Anderson (1998). A model comparison using the Bayesian information criterion also requires the specification of the number of parameters in each model (Kass and Raftery, 1995), but in complex hierarchical models parameters may outnumber observations and these methods clearly cannot be directly applied (Gelfand and Dey, 1994). The most ambitious attempts to tackle this problem appear in the smoothing and neural network literature (Wahba, 1990; Moody, 1992; MacKay, 1995; Ripley, 1996). This paper suggests Bayesian measures of complexity and fit that can be combined to compare models of arbitrary structure.

In the next section we use an information theoretic argument to motivate a complexity measure p_D for the effective number of parameters in a model, as the difference between the posterior mean of the deviance and the deviance at the posterior estimates of the parameters of interest. This quantity can be trivially obtained from an MCMC analysis and algebraic forms and approximations are unnecessary for its use. We nevertheless investigate some of its formal properties in the following three sections: Section 3 shows that p_D is approximately the trace of the product of Fisher's information and the posterior covariance matrix, whereas in Section 4 we show that for normal models p_D corresponds to the trace of the 'hat' matrix projecting observations onto fitted values and we illustrate its form for various hierarchical models. Its properties in exponential families are explored in Section 5.

The posterior mean deviance $\bar{D}$ can be taken as a Bayesian measure of fit or 'adequacy', and Section 6 shows how in exponential family models an observation's contributions to $\bar{D}$ and p_D can be used as residual and leverage diagnostics respectively. In Section 7 we tentatively suggest that the adequacy $\bar{D}$ and complexity p_D may be added to form a *deviance information criterion* DIC which may be used for comparing models. We describe how this parallels the development of non-Bayesian information criteria and provide a somewhat heuristic decision theoretic justification. In Section 8 we illustrate the use of this technique on some reasonably complex examples. Finally, Section 9 draws some conclusions concerning these proposed techniques.

2. The complexity of a Bayesian model

2.1. 'Focused' full probability models

Parametric statistical modelling of data y involves the specification of a probability model $p(y|\theta)$, $\theta \in \Theta$. For a Bayesian 'full' probability model, we also specify a prior distribution $p(\theta)$ which may give rise to a marginal distribution

$$p(y) = \int_{\Theta} p(y|\theta)\, p(\theta)\, \mathrm{d}\theta. \tag{1}$$

Particular choices of $p(y|\theta)$ and $p(\theta)$ will be termed a model 'focused' on Θ. Note that we might further parameterize our prior with unknown 'hyperparameters' ψ to create a hierarchical model, so that the full probability model factorizes as

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

Then, depending on the parameters in focus, the model may compose the likelihood $p(y|\theta)$ and prior

$$p(\theta) = \int_{\Psi} p(\theta|\psi)\, p(\psi)\, \mathrm{d}\psi,$$

or the likelihood

$$p(y|\psi) = \int_{\Theta} p(y|\theta)\, p(\theta|\psi)\, \mathrm{d}\theta$$

and prior $p(\psi)$. Both these models lead to the same marginal distribution (1) but can be considered as having different numbers of parameters. A consequence is that in hierarchical modelling we cannot uniquely define a 'likelihood' or 'model complexity' without specifying the level of the hierarchy that is the focus of the modelling exercise (Gelfand and Trevisani, 2002). In fact, by focusing our models on a particular set of parameters Θ, we essentially reduce all models to non-hierarchical structures.

For example, consider an unbalanced random-effects one-way analysis of variance (ANOVA) focused on the group means:

$$y_i|\theta_i \sim N(\theta_i, \tau_i^{-1}), \qquad \theta_i \sim N(\psi, \lambda^{-1}), \qquad i = 1, \ldots, p. \tag{2}$$

This model could also be focused on the overall mean ψ to give

$$y_i|\psi \sim N(\psi, \tau_i^{-1} + \lambda^{-1}),$$

in which case it could reasonably be considered as having a different complexity.

It is natural to wish to measure the complexity of a focused model, both in its own right, say to assess the degrees of freedom of estimators, and as a contribution to model choice: for example, criteria such as BIC (Schwarz, 1978), AIC (Akaike, 1973), TIC (Takeuchi, 1976) and NIC (Murata *et al.*, 1994) all trade off model fit against a measure of the effective number of parameters in the model. However, the foregoing discussion suggests that such measures of complexity may not be unique and will depend on the number of parameters in focus. Furthermore, the inclusion of a prior distribution induces a dependence between parameters that is likely to reduce the effective dimensionality, although the degree of reduction may depend on the data that are available. Heuristically, complexity reflects the 'difficulty in estimation' and hence it seems reasonable that a measure of complexity may depend on both the prior information concerning the parameters in focus and the specific data that are observed.

2.2. *Is there a true model?*

We follow Box (1976) in believing that 'all models are wrong, but some are useful'. However, it can be useful to posit a 'true' distribution $p^{\mathrm{t}}(Y)$ of unobserved future data Y since, for any focused model, this defines a 'pseudotrue' parameter value θ^{t} (Sawa, 1978) which specifies a likelihood $p(Y|\theta^{\mathrm{t}})$ that minimizes the Kullback–Leibler distance $E^{\mathrm{t}}[\log\{p^{\mathrm{t}}(Y)\}/p(Y|\theta^{\mathrm{t}})]$ from $p^{\mathrm{t}}(Y)$. Having observed data y, under reasonably broad conditions (Berk, 1966; Bunke and Milhaud, 1998) $p(\theta|y)$ converges to θ^{t} as information on the components of θ increases. Thus Bayesian analysis implicitly relies on $p(Y|\theta^{\mathrm{t}})$ being a reasonable approximation to $p^{\mathrm{t}}(Y)$, and we shall indicate where we make use of this 'good model' assumption.

2.3. True and estimated residual information

The residual information in data y conditional on θ may be defined (up to a multiplicative constant) as $-2\log\{p(y|\theta)\}$ (Kullback and Leibler, 1951; Burnham and Anderson, 1998) and can be interpreted as a measure of 'surprise' (Good, 1956), logarithmic penalty (Bernardo, 1979) or uncertainty. Suppose that we have an estimator $\tilde{\theta}(y)$ of the pseudotrue parameter θ^{t}. Then the excess of the true over the estimated residual information will be denoted

$$d_\Theta\{y, \theta^{\mathrm{t}}, \tilde{\theta}(y)\} = -2\log\{p(y|\theta^{\mathrm{t}})\} + 2\log[p\{y|\tilde{\theta}(y)\}]. \tag{3}$$

This can be thought of as the reduction in surprise or uncertainty due to estimation, or alternatively the degree of 'overfitting' due to $\tilde{\theta}(y)$ adapting to the data y. We now argue that d_Θ may form the basis for both classical and Bayesian measures of model dimensionality, with each approach differing in how it deals with the unknown true parameters in d_Θ.

2.4. Classical measures of model dimensionality

In a non-Bayesian likelihood-based context, we may take $\tilde{\theta}(y)$ to be the maximum likelihood estimator $\hat{\theta}(y)$, expand $2\log\{p(y|\theta^{\mathrm{t}})\}$ around $2\log[p\{y|\hat{\theta}(y)\}]$, take expectations with respect to the unknown true sampling distribution $p^{\mathrm{t}}(Y)$ and hence show (Ripley, 1996) (page 34) that

$$E^{\mathrm{t}}[d_\Theta\{Y, \theta^{\mathrm{t}}, \tilde{\theta}(Y)\}] \approx p^* = \mathrm{tr}(KJ^{-1}), \tag{4}$$

where

$$\begin{aligned} J &= -E^{\mathrm{t}}\left[\frac{\partial^2 \log\{p(Y|\theta^{\mathrm{t}})\}}{\partial\theta^2}\right], \\ K &= \mathrm{var}^{\mathrm{t}}\left[\frac{\partial \log\{p(Y|\theta^{\mathrm{t}})\}}{\partial\theta}\right]. \end{aligned} \tag{5}$$

This is the measure of complexity that is used in TIC (Takeuchi, 1976). Burnham and Anderson (1998) (page 244) pointed out that

$$p^* = \mathrm{tr}(J\Sigma), \tag{6}$$

where $\Sigma = J^{-1}KJ^{-1}$ is the familiar 'sandwich' approximation to the variance–covariance matrix of the $\hat{\theta}(y)$ (Huber, 1967). If $p^{\mathrm{t}}(y) = p(y|\theta^{\mathrm{t}})$, i.e. one of the models is true, then $K = J$ and $p^* = p$, the number of independent parameters in Θ.

For example, in a fixed effect ANOVA model

$$y_i|\theta_i \sim N(\theta_i, \tau_i^{-1}), \qquad i = 1, \ldots, p,$$

with τ_i^{-1}s known,

$$d_\Theta\{y, \theta^{\mathrm{t}}, \hat{\theta}(y)\} = \sum_i \tau_i (y_i - \theta_i^{\mathrm{t}})^2,$$

whose expectation under $p^{\mathrm{t}}(Y)$ is $p^* = \Sigma_i \tau_i E^{\mathrm{t}}(Y_i - \theta^{\mathrm{t}})^2$. If the model is true, $E^{\mathrm{t}}(Y_i - \theta^{\mathrm{t}})^2 = \tau_i^{-1}$ and so $p^* = p$.

Ripley (1996) (page 140) showed how this procedure may be extended to 'regularized' models in which a specified prior term $p(\theta)$ is introduced to form a penalized log-likelihood. Replacing $\log(p)$ by $\log\{p(y|\theta)\} + \log\{p(\theta)\}$ in equations (5) yields a more general definition of p^* that

was derived by Moody (1992) and termed the 'effective number of parameters'. This is the measure of dimensionality that is used in NIC (Murata *et al.*, 1994): the estimation of p^* is generally not straightforward (Ripley, 1996).

In the random-effects ANOVA example with $\theta_i \sim N(\psi, \lambda^{-1})$, ψ and λ known, let $\rho_i = \tau_i/(\tau_i + \lambda)$ be the intraclass correlation coefficient in the ith group. We then obtain

$$p^* = \sum_i \rho_i \tau_i \, E^{\mathrm{t}}(Y_i - \theta^{\mathrm{t}})^2, \tag{7}$$

which becomes

$$p^* = \sum_i \rho_i \tag{8}$$

if the likelihood is true.

2.5. *A Bayesian measure of model complexity*

From a Bayesian perspective, the unknown θ^{t} may be replaced by a random variable θ. Then $d_\Theta\{y, \theta, \tilde{\theta}(y)\}$ can be estimated by its posterior expectation with respect to $p(\theta|y)$, denoted

$$\begin{aligned} p_D\{y, \Theta, \tilde{\theta}(y)\} &= E_{\theta|y}[d_\Theta\{y, \theta, \tilde{\theta}(y)\}] \\ &= E_{\theta|y}[-2 \log\{p(y|\theta)\}] + 2 \log[p\{y|\tilde{\theta}(y)\}]. \end{aligned} \tag{9}$$

$p_D\{y, \Theta, \tilde{\theta}(y)\}$ is our proposal as the effective number of parameters with respect to a model with focus Θ: we shall usually drop the arguments $\{y, \Theta, \tilde{\theta}(y)\}$ from the notation. In our examples we shall generally take $\tilde{\theta}(y) = E(\theta|y) = \bar{\theta}$, the posterior mean of the parameters. However, we note that it is not strictly necessary to use the posterior mean as an estimator of either d_Θ or θ, and the mode or median could be justified (Section 2.6).

Taking $f(y)$ to be some fully specified standardizing term that is a function of the data alone, p_D may be written as

$$p_D = \overline{D(\theta)} - D(\bar{\theta}) \tag{10}$$

where

$$D(\theta) = -2 \log\{p(y|\theta)\} + 2 \log\{f(y)\}.$$

We shall term $D(\theta)$ the 'Bayesian deviance' in general and, more specifically, for members of the exponential family with $E(Y) = \mu(\theta)$ we shall use the saturated deviance $D(\theta)$ obtained by setting $f(y) = p\{y|\mu(\theta) = y\}$: see Section 8.1.

Equation (10) shows that p_D can be considered as a 'mean deviance minus the deviance of the means'. A referee has pointed out the related argument used by Meng and Rubin (1992), who showed that such a difference, between the average of log-likelihood ratios and the likelihood ratio evaluated at the average (over multiple imputations) of the parameters, is the key quantity in estimating the degrees of freedom of a test.

For example, in the random-effects ANOVA (2) with ψ and λ known,

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

which is -2 log(likelihood) standardized by the term $-2\log\{f(y)\} = \Sigma_i \log(2\pi/\tau_i)$ obtained from setting $\theta_i = y_i$. Now $\theta_i|y \sim N\{\rho_i y_i + (1-\rho_i)\psi, \rho_i\tau_i^{-1}\}$ and hence it can be shown that the posterior distribution of $D(\theta)$ has the form

$$D(\theta) \sim \sum \rho_i \chi^2\{1, (y_i - \psi)^2(1-\rho_i)\lambda\},$$

where $\chi^2(a, b)$ is a non-central χ^2-distribution with mean $a + b$. Thus, since $\rho_i\lambda = (1-\rho_i)\tau_i$, we have

$$\overline{D(\theta)} = \sum \rho_i + \sum \tau_i(1-\rho_i)^2(y_i - \psi)^2,$$

$$D(\bar{\theta}) = \sum \tau_i(1-\rho_i)^2(y_i - \psi)^2,$$

and so

$$p_D = \sum_i \rho_i = \sum_i \frac{\tau_i}{\tau_i + \lambda}. \tag{11}$$

The effective number of parameters is therefore the sum of the intraclass correlation coefficients, which essentially measures the sum of the ratios of the precision in the likelihood to the precision in the posterior. This exactly matches Moody's approach (8) when the model is true.

If ψ is unknown and given a uniform hyperprior we obtain a posterior distribution $\psi \sim N\{\bar{y}, (\lambda\Sigma\rho_i)^{-1}\}$, where $\bar{y} = \Sigma\rho_i y_i/\Sigma\rho_i$. It is straightforward to show that

$$\overline{D(\theta)} = \sum \rho_i + \lambda\sum \rho_i(1-\rho_i)(y_i - \bar{y})^2 + \sum \rho_i(1-\rho_i)/\sum \rho_i,$$

$$D(\bar{\theta}) = \lambda\sum \rho_i(1-\rho_i)(y_i - \bar{y})^2,$$

and so $p_D = \Sigma\rho_i + \Sigma\rho_i(1-\rho_i)/\Sigma\rho_i$. If the groups are independent, $\lambda = 0$, $\rho_i = 1$ and $p_D = p$. If the groups all have the same mean, $\lambda \to \infty$, $\rho_i \to 0$ and $p_D \to 1$. If all group precisions are equal, $p_D = 1 + (p-1)\rho$, as obtained by Hodges and Sargent (2001).

2.6. Some observations on p_D

(a) Equation (10) may be rewritten as

$$\overline{D(\theta)} = D(\bar{\theta}) + p_D, \tag{12}$$

which can be interpreted as a classical 'plug-in' measure of fit plus a measure of complexity. Thus our Bayesian measure of fit, $\overline{D(\theta)}$, could perhaps be better considered as a measure of 'adequacy', and we shall use these terms interchangeably. However, in Section 7.3 we shall suggest that an additional penalty for complexity may be reasonable when making model comparisons.

(b) Simple use of the Bayes theorem reveals the expression

$$p_D = E_{\theta|y}\left[-2\log\left\{\frac{p(\theta|y)}{p(\theta)}\right\}\right] + 2\log\left\{\frac{p(\tilde{\theta}|y)}{p(\tilde{\theta})}\right\},$$

which can be interpreted as (minus twice) the posterior estimate of the gain in information provided by the data about θ, minus the plug-in estimate of the gain in information.

(c) It is reasonable that the effective number of parameters in a model might depend on the data, the choice of focus Θ and the prior information (Section 2.1). Less attractive, perhaps, is that p_D may also depend on the choice of estimator $\tilde{\theta}(y)$, since this can produce a lack of invariance of p_D to apparently innocuous transformations, such as making inferences on logits instead of probabilities in Bernoulli trials. Our usual choice of the posterior mean is largely based on the subsequent ability to investigate approximate forms for p_D (Section 3), and the positivity properties described below. A choice of, say, posterior medians would produce a measure of model complexity that was invariant to univariate 1–1 transformations, and we explore this possibility in Section 5.

(d) It follows from equation (10) and Jensen's inequality that, when using the posterior mean as an estimator $\tilde{\theta}(y)$, $p_D \geqslant 0$ for any likelihood that is log-concave in θ, with 0 being approached for a degenerate prior on θ. Non-log-concave likelihoods can, however, give rise to a negative p_D in certain circumstances. For example, consider a single observation from a Cauchy distribution with deviance $D(\theta) = 2 \log\{1 + (y - \theta)^2\}$, with a discrete prior assigning probability 1/11 to $\theta = 0$ and 10/11 to $\theta = 3$. If we observe $y = 0$, then the posterior probabilities are changed to 0.5 and 0.5, and so $\bar{\theta} = 1.5$. Thus $p_D = \overline{D(\theta)} - D(\bar{\theta}) = \log(10) - 2 \log(13/4) = \log(160/169) < 0$. Our experience has been that negative p_Ds indicate substantial conflict between the prior and data, or where the posterior mean is a poor estimator (such as a symmetric bimodal distribution).

(e) The posterior distribution that is used in obtaining p_D conditions on the truth of the model, and hence p_D may only be considered an appropriate measure of the complexity of a model that reasonably describes the data. This is reflected in the finding that p_D in the simple ANOVA example (11) will not necessarily be approximately equivalent to the classical p^* (7) if the assumptions of the model are substantially inaccurate. This good model assumption (Section 2.2) is further considered when we come to comparisons of models (Section 7.3).

(f) Provided that $D(\theta)$ is available in closed form, p_D may be easily calculated after an MCMC run by taking the sample mean of the simulated values of $D(\theta)$, minus the plug-in estimate of the deviance using the sample means of the simulated values of θ. No 'small sample' adjustment is necessary. This ease of computation should be contrasted with the frequent difficulty within the classical framework with deriving the functional form of the measure of dimensionality and its subsequent estimation.

(g) Since the complexity depends on the focus, a decision must be made whether nuisance parameters, e.g. variances, are to be included in Θ or integrated out before specifying the model $p(y|\theta)$. However, such a removal of nuisance parameters may create computational difficulties.

p_D has been defined and is trivially computable by using MCMC methods, and so strictly speaking there is no need to explore exact forms or approximations. However, to provide insight into the behaviour of p_D, the following three sections consider the form of p_D in different situations and draw parallels with alternative suggestions: note that we are primarily concerned with the 'preasymptotic' situation in which prior opinion is still influential and the likelihood has not overwhelmed the prior.

3. Forms for p_D based on normal approximations

In Section 2.1 we argued that focused models are essentially non-hierarchical with a likelihood $p(y|\theta)$ and prior $p(\theta)$. Before considering particular assumptions for these we examine the form

of p_D under two general conditions: approximately normal likelihoods and negligible prior information.

3.1. p_D assuming a normal approximation to the likelihood

We may expand $D(\theta)$ around $E_{\theta|y}(\theta) = \bar{\theta}$ to give, to second order,

$$D(\theta) \approx D(\bar{\theta}) + (\theta - \bar{\theta})^{\mathrm{T}} \frac{\partial D}{\partial \theta}\bigg|_{\bar{\theta}} + \frac{1}{2}(\theta - \bar{\theta})^{\mathrm{T}} \frac{\partial^2 D}{\partial \theta^2}\bigg|_{\bar{\theta}} (\theta - \bar{\theta}), \tag{13}$$

$$= D(\bar{\theta}) - 2(\theta - \bar{\theta})^{\mathrm{T}} L'_{\bar{\theta}} - (\theta - \bar{\theta})^{\mathrm{T}} L''_{\bar{\theta}} (\theta - \bar{\theta}) \tag{14}$$

where $L = \log\{p(y|\theta)\}$ and L' and L'' represent first and second derivatives with respect to θ. This corresponds to a normal approximation to the likelihood.

Taking expectations of equation (14) with respect to the posterior distribution of θ gives

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

where $V = E\{(\theta - \bar{\theta})(\theta - \bar{\theta})^{\mathrm{T}}\}$ is the posterior covariance matrix of θ, and $-L''_{\bar{\theta}}$ is the observed Fisher information evaluated at the posterior mean of θ. Thus

$$p_D \approx \mathrm{tr}(-L''_{\bar{\theta}} V), \tag{15}$$

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. We note the parallel with the classical p^* in equation (6).

We also note that

$$L''_{\bar{\theta}} = Q''_{\bar{\theta}} - P''_{\bar{\theta}}$$

where $Q'' = \partial^2 \log\{p(\theta|y)\}/\partial\theta^2$ and $P'' = \partial^2 \log\{p(\theta)\}/\partial\theta^2$, and hence approximation (15) can be written

$$p_D \approx \mathrm{tr}(-Q''_{\bar{\theta}} V) - \mathrm{tr}(-P''_{\bar{\theta}} V).$$

Under approximate posterior normality $V^{-1} \approx -Q''_{\bar{\theta}}$ and hence

$$p_D \approx p - \mathrm{tr}(-P''_{\bar{\theta}} V) \tag{16}$$

where p is the cardinality of Θ.

3.2. p_D for approximately normal likelihoods and negligible prior information

Consider a focused model in which $p(\theta)$ is assumed to be dominated by the likelihood, either because of assuming a 'flat' prior or by increasing the sample size. Assume that the approximation

$$\theta|y \sim N(\hat{\theta}, -L''_{\hat{\theta}}) \tag{17}$$

holds, where $\bar{\theta} = \hat{\theta}$ are the maximum likelihood estimates such that $L'_{\hat{\theta}} = 0$ (Bernardo and Smith (1994), section 5.3). From equation (14)

$$\begin{aligned} D(\theta) &\approx D(\hat{\theta}) - (\theta - \hat{\theta})^{\mathrm{T}} L''_{\hat{\theta}} (\theta - \hat{\theta}) \\ &\approx D(\hat{\theta}) + \chi^2_p, \end{aligned} \tag{18}$$

since, by approximation (17), $-(\theta - \hat{\theta})^{\mathrm{T}} L''_{\hat{\theta}} (\theta - \hat{\theta})$ has an approximate χ^2-distribution with p degrees of freedom.

Rearranging approximation (18) and taking expectations with respect to the posterior distribution of θ reveals that

$$p_D = E_{\theta|y}\{D(\theta)\} - D(\hat{\theta}) \approx p,$$

i.e. p_D will be approximately the true number of parameters: this approximation could also be derived by letting $P''_{\bar{\theta}} \to 0$ in approximation (16). This approximate identity is illustrated in Section 8.1.

We note in passing that we might use MCMC output to estimate the classical deviance $D(\hat{\theta})$ of any likelihood-based model by

$$\hat{D}(\hat{\theta}) = E_{\theta|y}\{D(\theta)\} - p. \tag{19}$$

Although the 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, and so the estimator given by equation (19) may be preferable.

4. p_D for normal likelihoods

In this section we illustrate the formal behaviour of p_D for normal likelihoods by using exact and approximate identities. However, it is important to keep in mind that in practice such forms are unnecessary for computation and that p_D should automatically allow for fixed effects, random effects and unknown precisions.

4.1. The normal linear model

We consider the general hierarchical normal model described by Lindley and Smith (1972). Suppose that

$$\begin{aligned} y &\sim N(A_1\theta, C_1), \\ \theta &\sim N(A_2\psi, C_2) \end{aligned} \tag{20}$$

where all matrices and vectors are of appropriate dimension, and C_1 and C_2 are assumed known and θ is the focus: unknown precisions are considered in Section 4.5. Then the standardized deviance is $D(\theta) = (y - A_1\theta)^{\mathrm{T}} C_1^{-1} (y - A_1\theta)$, and the posterior distribution for θ is normal with

mean $\bar{\theta} = Vb$ and covariance V: V and b will be left unspecified for the moment. Expressing $y - A_1\theta$ as $y - A_1\bar{\theta} + A_1\bar{\theta} - A_1\theta$ reveals that

$$D(\theta) = D(\bar{\theta}) - 2(y - A_1\bar{\theta})^{\mathrm{T}} C_1^{-1} A_1(\theta - \bar{\theta}) + (\theta - \bar{\theta})^{\mathrm{T}} A_1^{\mathrm{T}} C_1^{-1} A_1(\theta - \bar{\theta}).$$

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

$$\bar{D} = D(\bar{\theta}) + \mathrm{tr}(A_1^{\mathrm{T}} C_1^{-1} A_1 V),$$

and thus $p_D = \mathrm{tr}(A_1^{\mathrm{T}} C_1^{-1} A_1 V)$. We note that $A_1^{\mathrm{T}} C_1^{-1} A_1$ is the Fisher information $-L''$, V is the posterior covariance matrix and hence

$$p_D = \mathrm{tr}(-L'' V): \tag{21}$$

an exact version of approximation (15). It is also clear that in this context p_D is invariant to affine transformations of θ.

If ψ is assumed known, then Lindley and Smith (1972) showed that $V^{-1} = A_1^{\mathrm{T}} C_1^{-1} A_1 + C_2^{-1}$ and hence from equation (21)

$$p_D = p - \mathrm{tr}(C_2^{-1} V) \tag{22}$$

as an exact version of approximation (16); then $0 \leqslant p_D \leqslant p$, and $p - p_D$ is the measure of the 'shrinkage' of the posterior estimates towards the prior means. If $(C_2^{-1}V)^{-1} = A_1^{\mathrm{T}} C_1^{-1} A_1 C_2 + I_p$ has eigenvalues $\lambda_i + 1, i = 1, \ldots, p$, then

$$p_D = \sum_{i=1}^{p} \frac{\lambda_i}{\lambda_i + 1}, \tag{23}$$

and hence the upper bound for p_D is approached as the eigenvalues of C_2 become large, i.e. the prior becomes flat. It can further be shown, in the case $A_1 = I_n$, that p_D is the sum of the squared canonical correlations between data Y and the 'signal' θ.

4.2. *The 'hat' matrix and leverages*

A revealing identity is found by noting that $b = A_1^{\mathrm{T}} C_1^{-1} y$ and the fitted values for the data are given by $\hat{y} = A_1\bar{\theta} = A_1 Vb = A_1 V A_1^{\mathrm{T}} C_1^{-1} y$. Thus the hat matrix that projects the data onto the fitted values is $H = A_1 V A_1^{\mathrm{T}} C_1^{-1}$, and

$$p_D = \mathrm{tr}(A_1^{\mathrm{T}} C_1^{-1} A_1 V) = \mathrm{tr}(A_1 V A_1^{\mathrm{T}} C_1^{-1}) = \mathrm{tr}(H). \tag{24}$$

This identity also holds assuming that ψ is unknown with a uniform prior, in which case Lindley and Smith (1972) showed that $V^{-1} = A_1^{\mathrm{T}} C_1^{-1} A_1 + C_2^{-1} - C_2^{-1} A_2 (A_2^{\mathrm{T}} C_2^{-1} A_2)^{-1} A_2^{\mathrm{T}} C_2^{-1}$.

The identification of the effective number of parameters with the trace of the hat matrix is a standard result in linear modelling and has been applied to smoothing (Wahba, 1990) (page 63) and generalized additive models (Hastie and Tibshirani (1990), section 3.5), and is also the conclusion of Hodges and Sargent (2001) 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: see Section 4.4 for further discussion. Note that $\mathrm{tr}(H)$ 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 return to this identity in Section 6.3.

Ye (1998) considered the independent normal model

$$y_i \sim N(\theta_i, \tau^{-1})$$

and suggested that the effective number of parameters should be $\Sigma_i h_i$, where

$$h_i(\theta) = \frac{\partial E_{y|\theta}(\tilde{\theta}_i)}{\partial \theta_i}: \tag{25}$$

the average sensitivity of an unspecified estimate $\tilde{\theta}_i$ to a small change in y_i. This is a generalization of the trace of the hat matrix discussed above. In the context of the normal linear models, it is straightforward to show that $E_{Y|\theta}(\tilde{\theta}) = H\theta$, and hence $p_D = \text{tr}(H)$ matches Ye's suggestion for model complexity. Further connections with Ye (1998) are described in Section 7.2.

4.3. *Example: Laird–Ware mixed models*

Laird and Ware (1982) specified the mixed normal model as

$$y \sim N(X\alpha + Z\beta, C_1),$$
$$\beta \sim N(0, D),$$

where the covariance matrices C_1 and D are currently assumed known. The random effects are β, and the fixed effects are α, and placing a uniform prior on α we can write this model within the general Lindley–Smith formulation (20) by setting $\theta = (\alpha, \beta)$, $A_1 = (X, Z)$, $\psi = 0$ and C_2 as a block diagonal matrix with ∞ in the top left-hand block, D in the bottom right and 0 elsewhere.

We have already shown that in these circumstances $p_D = \text{tr}\{A_1^{\mathrm{T}} C_1^{-1} A_1 (A_1^{\mathrm{T}} C_1^{-1} A_1 + C_2^{-1})^{-1}\}$, and substituting in the appropriate entries for the Laird–Ware model gives $p_D = \text{tr}(V^* V^{-1})$, where

$$V^* = \begin{pmatrix} X^{\mathrm{T}} C_1^{-1} X & X^{\mathrm{T}} C_1^{-1} Z \\ Z^{\mathrm{T}} C_1^{-1} X & Z^{\mathrm{T}} C_1^{-1} Z \end{pmatrix},$$
$$V = \begin{pmatrix} X^{\mathrm{T}} C_1^{-1} X & X^{\mathrm{T}} C_1^{-1} Z \\ Z^{\mathrm{T}} C_1^{-1} X & Z^{\mathrm{T}} C_1^{-1} Z + D^{-1} \end{pmatrix}$$

which is the precision of the parameter estimates assuming that $D^{-1} = 0$, relative to the precision assuming informative D.

4.4. *Frequentist approaches to model complexity: smoothing and normal non-linear models*

A common model in semiparametric regression is

$$y \sim N(X\alpha + \beta, \tau^{-1} C_1),$$
$$\beta \sim N(0, \lambda^{-1} D),$$

where β is a vector of length n of function values of the nonparametric part of an interpolation spline (Wahba, 1990; van der Linde, 1995) and C_1 and D are assumed known. Motivated by the need to estimate the unknown scale factors τ^{-1} and λ^{-1}, for many years the effective number of parameters has been taken to be the trace of the hat matrix (Wahba (1990), page 63) and so, for example, $\hat{\tau}^{-1}$ is the residual sum of squares divided by the 'effective degrees

of freedom' $n - \mathrm{tr}(H)$. In this class of models this measure of complexity coincides with p_D. Interest in regression diagnostics (Eubank, 1985; Eubank and Gunst, 1986) and cross-validation to determine the smoothing parameter τ/λ (Wahba (1990), section 4.2) also drew attention to the diagonal entries of the hat matrix as leverage values.

Links to partially Bayesian interpolation models have been provided by Kimeldorf and Wahba (1970) and Wahba (1978, 1983) and further work built on these ideas. For example, another large class of models can be formulated by using the following extension to the Lindley–Smith model:

$$y \sim N\{g(\theta), \tau^{-1}C_1\},$$
$$\theta \sim N(A_2\psi, \lambda^{-1}D)$$

where g is a non-linear expression as found, for example, in pharmacokinetics or neural networks: in many situations $A_2\psi$ will be 0 and C_1 and D will be identity matrices. Define

$$q(\theta) = (y - g(\theta))^{\mathrm{T}} C_1^{-1} (y - g(\theta)),$$
$$r(\theta) = (\theta - A_2\psi)^{\mathrm{T}} D^{-1} (\theta - A_2\psi)$$

as the likelihood and prior residual variation. MacKay (1992) suggested estimating τ and λ by maximizing the 'type II' likelihood $p(y|\lambda, \tau)$ derived from integrating out the unknown θ from the likelihood. Setting derivatives equal to 0 eventually reveals that

$$\hat{\tau}^{-1} = \frac{q(\bar{\theta})}{n - p_D},$$
$$\hat{\lambda}^{-1} = \frac{r(\bar{\theta})}{p_D},$$

which are the fitted likelihood and prior residual variation, divided by the appropriate effective degrees of freedom: $p_D = \mathrm{tr}(H)$ is the key quantity.

These results were derived by MacKay (1992) in the context of 'regularization' in complex interpolation models such as neural networks, in which the parameters θ are standardized and assumed to have independent normal priors with mean 0 and precision λ. Then expression (16) may be written

$$p_D \approx p - \lambda\,\mathrm{tr}(V). \tag{26}$$

However, MacKay's use of approximation (26) requires the evaluation of $\mathrm{tr}(V)$, whereas our p_D arises without any additional computation. We would also recommend including λ and τ in the general MCMC estimation procedure, rather than relying on type II maximum likelihood estimates (Ripley (1996), page 167). In this and the smoothing context a fully Bayesian analysis requires prior distributions for τ^{-1} and λ^{-1} to be specified (van der Linde, 2000), and this will both change the complexity of the model and require a choice of estimator of the precisions. We shall now illustrate the form of p_D in the restricted situation of unknown τ^{-1}.

4.5. Normal models with unknown sampling precision

Introducing unknown variances as part of the focus confronts us with the need to choose a form for the plug-in posterior estimates. We may illustrate this issue by extending the general hierarchical normal model (20) to the conjugate normal–gamma model with an unknown scale

parameter τ in both the likelihood and the prior (Bernardo and Smith (1994), section 5.2.1). Suppose that

$$\begin{aligned} y &\sim N(A_1\theta, \tau^{-1}C_1), \\ \theta &\sim N(A_2\psi, \tau^{-1}C_2), \end{aligned} \tag{27}$$

and we focus on (θ, τ). The standardized deviance is $D(\theta, \tau) = \tau\, q(\theta) - n\log(\tau)$, where

$$q(\theta) = (y - A_1\theta)^{\mathrm{T}} C_1^{-1}(y - A_1\theta)$$

is the residual variation. Then, for a currently unspecified estimator $\hat{\tau}$,

$$\begin{aligned} p_D &= E_{\theta,\tau|y}(D|\theta, \tau) - D(\bar{\theta}, \hat{\tau}) \\ &= E_{\tau|y}[E_{\theta|\tau,y}\{\tau\, q(\theta)\} - n\ \log(\tau)] - \{\hat{\tau}\, q(\bar{\theta}) - n\log(\hat{\tau})\} \\ &= \mathrm{tr}(H) + q(\bar{\theta})(\bar{\tau} - \hat{\tau}) - n\{\overline{\log(\tau)} - \log(\hat{\tau})\} \end{aligned} \tag{28}$$

where $H = A_1^{\mathrm{T}}C_1^{-1}A_1(A_1^{\mathrm{T}}C_1^{-1}A_1 + C_2^{-1})^{-1}$ is the hat matrix which does not depend on τ. Thus the additional uncertain scale parameter adds the second two terms to the complexity of the model.

A conjugate prior $\tau \sim \mathrm{gamma}(a, b)$ leads to a posterior distribution $\tau|y \sim \mathrm{gamma}(a + n/2, b + S/2)$, where

$$S = (y - A_1A_2\psi)^{\mathrm{T}}(C_1 + A_1^{\mathrm{T}}C_2A_1)^{-1}(y - A_1A_2\psi).$$

It remains to choose the estimator $\hat{\tau}$ to place in equation (28), and we shall consider two options.

Suppose that we parameterize in terms of τ and use

$$\hat{\tau} = \bar{\tau} = \frac{a + n/2}{b + S/2},$$

making the second term in equation (28) 0. Now if $X \sim \mathrm{gamma}(a, b)$, then $E\{\log(X)\} = \psi(a) - \log(b)$ where ψ is the digamma function, and so $\overline{\log(\tau)} = \psi(a + n/2) - \log(b + S/2)$. Hence the term contributing to p_D due to the unknown precision is

$$\begin{aligned} p_D - \mathrm{tr}(H) &= -n\left\{\psi\left(a + \frac{n}{2}\right) - \log\left(a + \frac{n}{2}\right)\right\} \\ &\approx 1 - \frac{2a - \frac{1}{3}}{2a + n} \end{aligned}$$

using the approximation $\psi(x) \approx \log(x) - 1/2x - 1/12x^2$. This term will tend to $1 + 1/3n$ as prior information becomes negligible and hence will be close to the 'correct' value of 1 for moderate sample sizes.

If we were to parameterize in terms of $\log(\tau)$ and to use $\hat{\tau} = \exp\{\overline{\log(\tau)}\}$, the third term in equation (28) is 0 and the second term can be shown to be $1 - O(n^{-1})$. Thus for reasonable sample sizes the choice of parameterization of the unknown precision will make little difference to the measure of complexity. However, in Section 7 we shall argue that the log-scale may be more appropriate owing to the better approximation to likelihood normality.

5. Exponential family likelihoods

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

$$\log\{p(y_{ij}|\theta_i, \phi)\} = w_i\{y_{ij}\theta_i - b(\theta_i)\}/\phi + c(y_{ij}, \phi), \tag{29}$$

where

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

and w_i is a constant. If the canonical parameterization Θ is the focus of the model, then writing $\bar{b}_i = E_{\theta_i|y}\{b(\theta_i)\}$ we easily obtain that the contribution of the ith group to the effective number of parameters is

$$p_{Di}^{\Theta} = 2n_i w_i\{\bar{b}_i - b(\bar{\theta}_i)\}/\phi. \tag{30}$$

These likelihoods highlight the issue of the lack of invariance of p_D to reparameterization, since the mean parameterization μ will give a different complexity p_{Di}^{μ}. This is first explored within simple binomial and Poisson models with conjugate priors, and then exact and approximate forms of p_D are examined for generalized linear and generalized linear mixed models.

5.1. Binomial likelihood with conjugate prior

In the notation of equation (29), $\phi = 1$, $w_i = 1$ and $\theta = \text{logit}(\mu) = \log\{\mu/(1-\mu)\}$, and the (unstandardized) deviance is

$$D(\mu_i) = -2y_i \log(\mu_i) - 2(n_i - y_i)\log(1 - \mu_i)$$

where $y_i = \Sigma_j y_{ij}$. A conjugate prior $\mu_i = \{1 + \exp(-\theta_i)\}^{-1} \sim \text{beta}(a, b)$ provides a posterior $\mu_i \sim \text{beta}(a + y_i, b + n_i - y_i)$ with mean $(a + y_i)/(a + b + n_i)$. Now, if $X \sim \text{beta}(a, b)$, then $E\{\log(X)\} = \psi(a) - \psi(a + b)$ and $E\{\log(1 - X)\} = \psi(b) - \psi(a + b)$ where ψ is the digamma function, and hence it can be shown that

$$\overline{D(\mu_i)} = \overline{D(\theta_i)} = -2y_i\,\psi(a + y_i) - 2(n_i - y_i)\,\psi(b + n_i - y_i) + 2n_i\,\psi(a + b + n_i)$$
$$D(\bar{\mu}_i) = -2y_i \log(a + y_i) - 2(n_i - y_i)\log(b + n_i - y_i) + 2n_i \log(a + b + n_i)$$
$$D(\bar{\theta}_i) = -2y_i\,\psi(a + y_i) + 2y_i\,\psi(b + n_i - y_i) + 2n_i \log[1 + \exp\{\psi(a + y_i) - \psi(b + n_i - y_i)\}],$$
$$D(\mu_i^{\text{med}}) = D(\theta_i^{\text{med}}) = -2y_i \log(\mu_i^{\text{med}}) - 2(n_i - y_i)\log(1 - \mu_i^{\text{med}})$$

where μ_i^{med} denotes the posterior median of μ_i.

Exact p_{D_i}s are obtainable by subtraction, and Fig. 1 shows how the value of p_{D_i} depends on the parameterization, the data and the prior. We may also gain further insight into the behaviour of p_{D_i} by considering approximate formulae for the mean and canonical parameterizations by using $\psi(x) \approx \log(x) - 1/2x \approx \log(x - \frac{1}{2})$. This leads to

$$p_{D_i}^{\mu} \approx \frac{y_i}{a + y_i} + \frac{n_i - y_i}{b + n_i - y_i} - \frac{n_i}{a + b + n_i}, 7$$
$$p_{D_i}^{\Theta} \approx \frac{n_i}{a + b + n_i - \frac{1}{2}}. \tag{31}$$

We make the following observations.

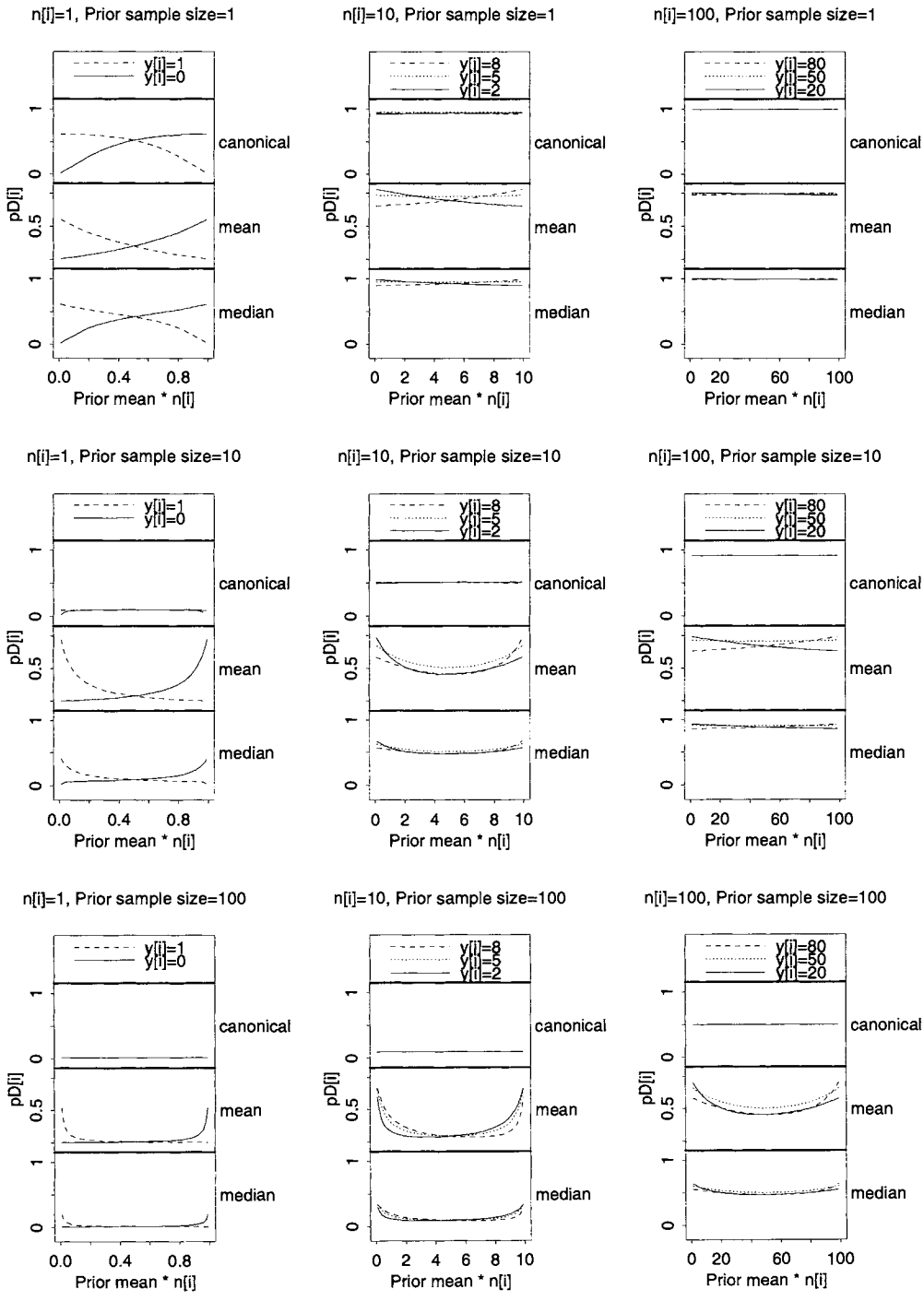

Fig. 1. Binomial likelihood—contribution of the ith group to the effective number of parameters under various parameterizations (canonical $p_{D_i}^{\Theta}$, mean $p_{D_i}^{\mu}$ and median $p_{D_i}^{\text{med}}$) as a function of the data (sample size n_i and observed proportion y_i/n_i) and prior (effective prior sample size $a+b$ and prior mean $a/(a+b)$): we are seeking agreement between alternative parameterizations with little dependence on data

5.1.1. Behaviour of p_D

For all three parameterizations, as the sample size in each group increases relative to the effective prior sample size, its contribution to p_{D_i} tends towards 1.

5.1.2. Agreement between parameterizations

The agreement between parameterizations is generally reasonable except in the situations in which the prior sample size is 10 times that of the data. While the canonical parameterization has $p_{D_i} \approx 1/11$, the mean and median give increased p_{D_i} for extreme prior means.

5.1.3. Dependence on data

With the exception of the sparse data and weak prior scenario for which the approximate formulae do not hold, the canonical $p_{D_i}^{\Theta}$ does not depend on the data observed and is approximately the ratio of the sample size to the effective posterior sample size. When the mean and median forms depend on data (say when $n_i = 1$ and $a + b = 10$), p_{D_i} is higher in situations of prior–data conflict.

5.2. Poisson likelihood with conjugate prior

In the notation of equation (29), $\phi = 1, w_i = 1$ and $\theta = \log(\mu)$, and the (unstandardized) deviance is $D(\mu_i) = -2y_i \log(\mu_i) + 2n_i\mu_i$. A conjugate prior $\mu_i = \exp(\theta_i) \sim \text{gamma}(a, b)$ gives a posterior $\mu_i \sim \text{gamma}(a + y_i, b + n_i)$ with mean $(a + y_i)/(b + n_i)$. If $X \sim \text{gamma}(a, b)$, then $E\{\log(X)\} = \psi(a) - \log(b)$ and hence we can show that

$$\overline{D(\mu_i)} = \overline{D(\theta_i)} = -2y_i\{\psi(a + y_i) - \log(b + n_i)\} + 2n_i\frac{a + y_i}{b + n_i},$$

$$D(\bar{\mu}_i) = -2y_i\{\log(a + y_i) - \log(b + n_i)\} + 2n_i\frac{a + y_i}{b + n_i},$$

$$D(\bar{\theta}_i) = -2y_i\{\psi(a + y_i) - \log(b + n_i)\} + 2n_i\frac{\exp\{\psi(a + y_i)\}}{b + n_i},$$

$$D(\mu_i^{\text{med}}) = D(\theta_i^{\text{med}}) = -2y_i \log(\mu_i^{\text{med}}) + 2n_i\mu_i^{\text{med}}.$$

Exact p_{D_i}s are obtainable by subtraction. Fig. 2 shows how the value of p_{D_i} relates to the parameterization, the data and the prior. Using the same approximation as previously, approximate p_{D_i}s for the mean and canonical parameterizations are

$$p_{D_i}^{\mu} \approx y_i/(a + y_i),$$
$$p_{D_i}^{\Theta} \approx n_i/(b + n_i).$$

5.2.1. Behaviour of p_{D_i}

For all three parameterizations, as the sample size in each group increases relative to the effective prior sample size, its contribution to p_{D_i} tends towards 1.

5.2.2. Agreement between parameterizations

The agreement between parameterizations is best when there is no conflict between the prior expectation and the data, but it can be substantial when such conflict is extreme. The median

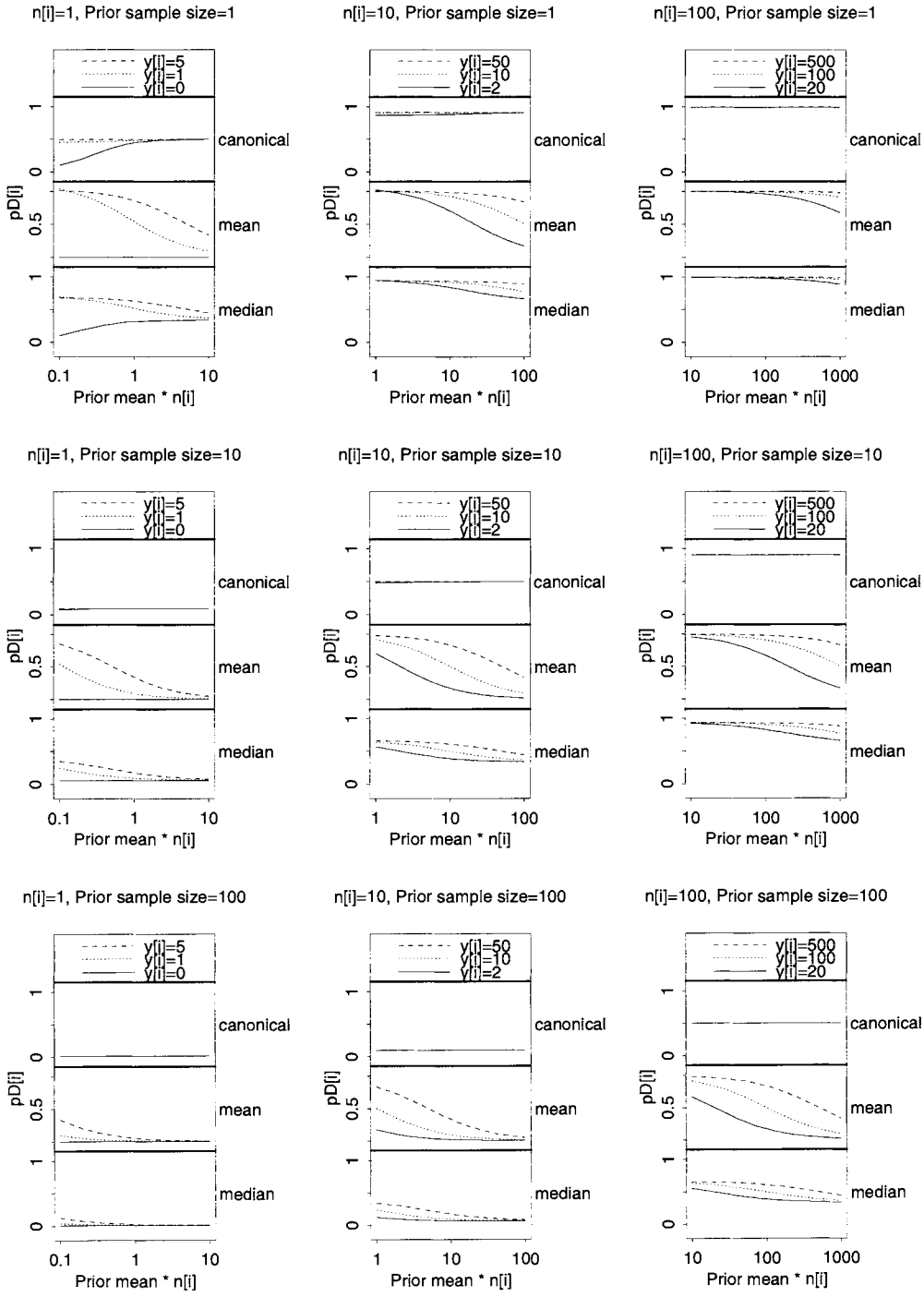

Fig. 2. Poisson likelihood—contribution of the ith group to the effective number of parameters under various parameterizations (canonical $p_{D_i}^{\Theta}$, mean $p_{D_i}^{\mu}$ and median $p_{D_i}^{\text{med}}$) as a function of the data (sample size n_i and observed total y_i) and prior (mean $n_i a/b$ and 'sample size' b)

estimator leads to a p_{D_t} that is intermediate between those derived from the canonical and mean parameterizations.

5.2.3. *Dependence on data*

Except in the situation of a single $y_i = 0$ with weak prior information, the approximation for the canonical $p^{\Theta}_{D_t}$ is very accurate and so $p^{\Theta}_{D_t}$ does not depend on the data observed. There can be a substantial dependence for the mean parameterization, with $p^{\mu}_{D_t}$ being higher when the prior mean underestimates the data.

5.2.4. *Conclusion*

In conclusion, for both binomial and Poisson data there is reasonable agreement between the different p_{D_t}s provided that the model provides a reasonable fit to the data, i.e. there is not strong conflict between the prior and data. The canonical parameterization appears preferable, both for its lack of dependence on the data and for its generally close approximation to the invariant p_{D_t} based on a median estimator. Thus we would not normally expect the choice of parameterization to have a strong effect, although in Section 8.3 we present an example of a Bernoulli model where this choice does prove to be important.

5.3. Generalized linear models with canonical link functions

Here we shall focus on the canonical parameterization in terms of θ_i, both for the reasons outlined above and because its likelihood should better fulfil a normal approximation (Slate, 1994): related identities are available for the mean parameterization in terms of $\mu_i = \mu(\theta_i)$. We emphasize again that the approximate identities that are derived in this and the following section are only for understanding the behaviour of p_D in idealized circumstances (i.e. known precision parameters) and are not required for computation in practical situations.

Following McCullagh and Nelder (1989) we assume that the mean μ_i of y_{ij} is related to a set of covariates x_i through a link function $g(\mu_i) = x_i^{\mathrm{T}}\alpha$, and that g is the canonical link $\theta(\mu)$. The second-order Taylor series expansion of $D(\theta_i)$ around $D(\bar{\theta}_i)$ yields an approximate normal distribution for working observations and hence derivations of Section 3 apply. We eventually obtain

$$p_D \approx \mathrm{tr}\{X^{\mathrm{T}}WX\, V(\alpha|y)\}$$

where W is diagonal with entries

$$W_i = \frac{w_i}{\phi} n_i\, b''(\bar{\theta}_i),$$

the generalized linear model iterated weights (McCullagh and Nelder (1989), page 40): ϕ is assumed known.

Under an $N(\alpha_0, C_2)$ prior on α, the prior contribution to the negative Hessian matrix at the mode is just C_2^{-1}, so under the canonical link the approximate normal posterior has variance

$$V(\alpha|y) = (C_2^{-1} + X^{\mathrm{T}}WX)^{-1},$$

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

5.4. Generalized linear mixed models

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

Using the same argument as for generalized linear models (Section 5.3), we find that

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

where

$$V^* = \begin{pmatrix} X^{\mathrm{T}} W^{-1} X & X^{\mathrm{T}} W^{-1} Z \\ Z^{\mathrm{T}} W^{-1} X & Z^{\mathrm{T}} W^{-1} Z \end{pmatrix},$$

$$V = \begin{pmatrix} X^{\mathrm{T}} W^{-1} X & X^{\mathrm{T}} W^{-1} Z \\ Z^{\mathrm{T}} W^{-1} X & Z^{\mathrm{T}} W^{-1} Z + D^{-1} \end{pmatrix}.$$

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. Diagnostics for fit and influence

6.1. Posterior expected deviance as a Bayesian measure of fit or 'adequacy'

The posterior mean of the deviance $E_{\theta|y}\{D(\theta)\} = \overline{D(\theta)}$ has often been used to compare models informally: see, for example, Dempster (1974) (reprinted as Dempster (1997a)), Raghunathan (1988), Zeger and Karim (1991), Gilks *et al.* (1993) and Richardson and Green (1997). These researchers have, however, not been explicit about whether, or how much, such a measure might be traded off against increasing complexity of a model: Dempster (1997b) suggested plotting log-likelihoods from MCMC runs but hesitated to dictate a model choice procedure. We shall discuss this further in Section 7.3. In Section 2.6 we argued that $\overline{D(\theta)}$ already incorporates some penalty for complexity and hence we use the term 'adequacy' and 'Bayesian fit' interchangeably.

6.2. Sampling theory diagnostics for lack of Bayesian fit

Suppose that all aspects of the model were assumed true. Then before observing data Y our expectation of the posterior expected deviance is

$$\begin{aligned} E_Y(\bar{D}) &= E_Y[E_{\theta|y}\{D(\theta)\}] \\ &= E_\theta(E_{Y|\theta}[-2\log\{p(Y|\theta)\} + 2\log\{f(Y)\}]) \end{aligned} \tag{32}$$

by reversing the conditioning between Y and θ. If $f(Y) = p\{Y|\hat{\theta}(Y)\}$ where $\hat{\theta}(Y)$ is the standard maximum likelihood estimate, then

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

is simply the expected likelihood ratio statistic for the fitted values $\hat{\theta}(Y)$ with respect to the true null model θ and hence under standard conditions is approximately $E(\chi^2_p) = p$, the dimensionality of θ. From equation (32) we therefore expect, if the model is true, the posterior expected deviance (standardized by the maximized log-likelihood) to be $E_Y(\bar{D}) \approx E_\theta(p) = p$, the number of free parameters in θ. This might be appropriate for checking the overall goodness of fit of the model.

In particular, consider the one-parameter exponential family where $p = n$, the total sample size. The likelihood is maximized by substituting y_i for the mean of y_i, and the posterior mean of the standardized deviance has approximate sampling expectation n if the model is true. This will be exact for normal models with known variance, but in general it will only be reliable if each observation provides considerable information about its mean (McCullagh and Nelder (1989),

page 36). Note that comparing $\bar{D}$ with n is precisely the same as comparing the 'classical' fit $D(\bar{\theta})$ with $n - p_D$, the effective degrees of freedom.

It is then natural to consider the contribution D_i of each observation i to the overall mean deviance, so that

$$\bar{D} = \sum_i \bar{D}_i = \sum_i dr_i^2$$

where $dr_i = \pm\sqrt{\bar{D}_i}$ (with the sign given by the sign of $y_i - E(y_i|\bar{\theta})$) termed the Bayesian deviance residual, defined analogously to McCullagh and Nelder (1989), page 39. See Section 8.1 for an application of this procedure.

6.3. Leverage diagnostics

In Section 4.1 we noted that in normal linear models the contribution p_{Di} of each observation i to p_D turned out to be its leverage, defined as the relative influence that each observation has on its own fitted value. For y_i conditionally independent given θ, it can be shown that

$$p_{Di} = -2\left(E_{\theta|y}\left[\log\left\{ \frac{p(\theta|y_i)}{p(\theta)} \right\} \right] - \log\left\{ \frac{p(\bar{\theta}|y_i)}{p(\bar{\theta})} \right\} \right)$$

which reflects its interpretation as the difficulty in estimating θ with y_i.

It may be possible to exploit this interpretation in general model fitting, and as a by-product of MCMC estimation to obtain estimates of leverage for each observation. Such diagnostics are illustrated in Section 8.1.

7. A model comparison criterion

7.1. Model 'selection'

There has been a long and continuing debate about whether the issue of selecting a model as a basis for inferences is amenable to a strict mathematical analysis using, for example, a decision theoretic paradigm: see, for example, Key *et al.* (1999). Our approach here can be considered to be semiformal. Although we believe that it is useful to have measures of fit and complexity, and to combine them into overall criteria that have some theoretical justification, we also feel that an overformal approach to model 'selection' is inappropriate since so many other features of a model should be taken into account before using it as a basis for reporting inferences, e.g. the robustness of its conclusions and its inherent plausibility. In addition, in many contexts it may not be appropriate to 'choose' a single model. Our development closely follows that of Section 2.

A characteristic that is common to both Bayesian and classical approaches is the concept of an independent replicate data set Y_{rep}, derived from the same data-generating mechanism as gave rise to the observed data. Suppose that the loss in assigning to a set of data Y a probability $p(Y|\tilde{\theta})$ is $\mathcal{L}(Y, \tilde{\theta})$. We assume that we shall favour models $p(Y|\tilde{\theta})$ for which $\mathcal{L}(Y, \tilde{\theta})$ is expected to be small, and thus a criterion can be based on an estimate of $E_{Y_{\text{rep}}|\theta^t}\{\mathcal{L}(Y_{\text{rep}}, \tilde{\theta})\}$.

A natural, but optimistic, estimate of this quantity is the 'apparent' loss $\mathcal{L}\{y, \tilde{\theta}(y)\}$ that is suffered on repredicting the observed y that gave rise to $\tilde{\theta}(y)$. We follow Efron (1986) in defining the 'optimism' that is associated with this estimator as c_Θ, where

$$E_{Y_{\text{rep}}|\theta^t}[\mathcal{L}\{Y_{\text{rep}}, \tilde{\theta}(y)\}] = \mathcal{L}\{y, \tilde{\theta}(y)\} + c_\Theta\{y, \theta^t, \tilde{\theta}(y)\}. \tag{33}$$

Both classical and Bayesian approaches to estimating the optimism c_Θ will now be examined when assuming a logarithmic loss function $\mathcal{L}(Y, \tilde{\theta}) = -2\log\{p(Y|\tilde{\theta})\}$: as in Section 2, the classical approach attempts to estimate the sampling expectation of c_Θ, whereas the Bayesian approach is based on a direct calculation of the posterior expectation of c_Θ.

7.2. *Classical criteria for model comparison*

From the previous discussion, approximate forms for the expected optimism

$$\pi(\theta^t) = E_{Y|\theta^t}[c_\Theta\{Y, \theta^t, \tilde{\theta}(Y)\}]$$

will, from equation (33), yield criteria for a comparison of models that are based on minimizing

$$\hat{E}_{Y_{\text{rep}}|\theta^t}[\mathcal{L}\{Y_{\text{rep}}, \tilde{\theta}(y)\}] = \mathcal{L}\{y, \tilde{\theta}(y)\} + \hat{\pi}(\theta^t). \tag{34}$$

Efron (1986) derived the expression for $\pi(\theta^t)$ for exponential families and for general loss functions. In particular, for the logarithmic loss function, Efron showed that

$$\pi_E(\theta^t) = 2\sum_i \text{cov}^t(\hat{Y}_i, Y_i), \tag{35}$$

where $\hat{Y}_i$ is the fitted value arising from the estimator $\tilde{\theta}$: if $\tilde{\theta}$ corresponds to maximum likelihood estimation based on a linear predictor with p parameters, then $\pi_E(\theta^t) \approx 2p$. Hence Efron's result can be thought of as generalizing Akaike (1973), who sought to minimize the expected Kullback–Leibler distance between the true and estimated predictive distribution and showed under broad conditions that $\pi(\theta^t) \approx 2p$.

This in turn suggests that $\pi_E/2$, derived from equation (35), may be adopted as a measure of complexity in more complex modelling situations. Ye and Wong (1998) extended the work mentioned in Section 4.2 to show that $\pi_E/2$ for exponential families can be expressed as a sum of the average sensitivity of the fitted values $\hat{y}_i$ to a small change in y_i: this quantity is termed by Ye and Wong the 'generalized degrees of freedom' when using a general estimation procedure. In normal models with linear estimators $\hat{y}_i = \tilde{\theta}_i(y) = \Sigma_j h_{ij} y_j$, and so $\pi(\theta^t) = 2\,\text{tr}(H)$. Finally, Ripley (1996) extended the analysis described in Section 2.4 to show that if the model assumed is not true then $\pi(\theta^t) \approx 2p^*$, where p^* is defined in equation (4). See Burnham and Anderson (1998) for a full and detailed review of all aspects of estimation of $\pi(\theta^t)$.

These classical criteria for general model comparison are thus all based on equation (34) and can all be considered as corresponding to a plug-in estimate of fit, plus twice the effective number of parameters in the model. We shall now adapt this structure to a Bayesian context.

7.3. *Bayesian criteria for model comparison*

Gelfand and Ghosh (1998) and Laud and Ibrahim (1995) both attempted strict decision theoretic approaches to model choice based on expected losses on replicate data sets. Our approach is more informal, in aiming to identify models that best explain the observed data, but with the expectation that they are likely to minimize uncertainty about observations generated in the same way. Thus, by analogy with the classical results described above, we propose a *deviance information criterion* DIC, defined as a classical estimate of fit, plus twice the effective number of parameters, to give

$$\text{DIC} = D(\bar{\theta}) + 2p_D \tag{36}$$

$$= \bar{D} + p_D \tag{37}$$

by definition of p_D (10): equation (37) shows that DIC can also be considered as a Bayesian measure of fit or adequacy, penalized by an additional complexity term p_D. From the results in Section 3.2, we immediately see that in models with negligible prior information DIC will be approximately equivalent to Akaike's criterion.

An approximate decision theoretic justification for DIC can be obtained by mimicking the development of Ripley (1996) (page 33) and Burnham and Anderson (1998) (chapter 6). Using the logarithmic loss function in equation (33), we obtain

$$c_\Theta\{y, \theta^t, \tilde{\theta}(y)\} = E_{Y_{\text{rep}}|\theta^t}\{D_{\text{rep}}(\tilde{\theta})\} - D(\tilde{\theta})$$

where $-2\log[p\{Y_{\text{rep}}|\tilde{\theta}(y)\}]$ is denoted $D_{\text{rep}}(\tilde{\theta})$ and so on: note in this section that D is an unstandardized deviance ($f(\cdot) = 1$). It is convenient to expand c_Θ into the three terms

$$c_\Theta = E_{Y_{\text{rep}}|\theta^t}\{D_{\text{rep}}(\tilde{\theta}) - D_{\text{rep}}(\theta^t)\} + E_{Y_{\text{rep}}|\theta^t}\{D_{\text{rep}}(\theta^t) - D(\theta^t)\} + \{D(\theta^t) - D(\tilde{\theta})\}; \qquad (38)$$

we shall denote the first two terms by $\mathcal{L}_1$ and $\mathcal{L}_2$ respectively and, since we are taking a Bayesian perspective, replace the true θ^t by a random quantity θ.

Expanding the first term to second order gives

$$\mathcal{L}_1(\theta, \tilde{\theta}) \approx E_{Y_{\text{rep}}|\theta}\{-2(\tilde{\theta} - \theta)^{\mathrm{T}} L'_{\text{rep},\theta} - (\tilde{\theta} - \theta)^{\mathrm{T}} L''_{\text{rep},\theta}(\tilde{\theta} - \theta)\}$$

where $L_{\text{rep},\theta} = \log\{p(Y_{\text{rep}}|\theta)\}$. Since $E_{Y_{\text{rep}}|\theta}(L'_{\text{rep},\theta}) = 0$ from standard results for score statistics, we obtain after some rearrangement

$$\mathcal{L}_1(\theta, \tilde{\theta}) \approx \text{tr}\{I_\theta(\tilde{\theta} - \theta)(\tilde{\theta} - \theta)^{\mathrm{T}}\}$$

where $I_\theta = E_{Y_{\text{rep}}|\theta}(-L''_{\text{rep},\theta})$ is the assumed Fisher information in Y_{rep}, and hence also in y. Making the good model assumption (Section 2.2), this might reasonably be approximated by the observed information at the estimated parameters, so

$$\mathcal{L}_1(\theta, \tilde{\theta}) \approx \text{tr}\{-L''_{\tilde{\theta}}(\tilde{\theta} - \theta)(\tilde{\theta} - \theta)^{\mathrm{T}}\}. \qquad (39)$$

Suppose that under a particular model assumption we obtain a posterior distribution $p(\theta|y)$. Then from approximations (38) and (39) our posterior expected optimism when adopting this model and the estimator $\tilde{\theta}$ is

$$E_{\theta|y}(c_\Theta) \approx \text{tr}[-L''_{\tilde{\theta}}\, E_{\theta|y}\{(\theta - \tilde{\theta})(\theta - \tilde{\theta})^{\mathrm{T}}\}] + E_{\theta|y}\{\mathcal{L}_2(y, \theta)\} + E_{\theta|y}\{D(\theta) - D(\tilde{\theta})\}.$$

Using the posterior mean $\bar{\theta}$ as our estimator makes the expected optimism

$$E_{\theta|y}(c_\Theta) \approx \text{tr}(-L''_{\bar{\theta}} V) + E_{\theta|y}\{\mathcal{L}_2(y, \theta)\} + p_D, \qquad (40)$$

where V again is defined as the posterior covariance of θ, and $p_D = \bar{D} - D(\bar{\theta})$. Now

$$\mathcal{L}_2(y, \theta) = E_{Y_{\text{rep}}|\theta}[-2\log\{p(Y_{\text{rep}}|\theta)\}] + 2\log\{p(y|\theta)\},$$

and so $E_Y[E_{\theta|Y}\{\mathcal{L}_2(Y, \theta)\}] = E_\theta[E_{Y|\theta}\{\mathcal{L}_2(Y, \theta)\}] = 0$. We have already shown in approximation (15) that $p_D \approx \text{tr}(-L''_{\bar{\theta}} V)$, and hence from expressions (33) and (40) the expected posterior loss when adopting a particular model is

$$D(\bar{\theta}) + E_{\theta|y}(c_\Theta) \approx D(\bar{\theta}) + 2p_D = \text{DIC},$$

neglecting a term $E_{\theta|y}\{\mathcal{L}_2(y, \theta)\}$ which is expected to be 0. This derivation has assumed that

D is an unstandardized deviance: common standardization across models will leave unchanged the property that differences in DIC are estimates of differences in expected loss in prediction.

We make the following observations concerning this admittedly heuristic justification of DIC. First, for the general normal linear model (20), it is straightforward to show that $\mathcal{L}_2(y, \theta) = p - (y - A_1\theta)^{\mathrm{T}} C_1^{-1} (y - A_1\theta)$ where p is the dimensionality of θ, and hence for true θ has sampling distribution $p - \chi_p^2$ with mean 0 and variance $2p$. This parallels the classical development in which Ripley (1996) (page 34) pointed out that the equivalent term is $O(\sqrt{n})$: we would hope that this factor will tend to cancel when assessing differences in DIC, but this requires further investigation.

Second, this development draws heavily on the approximations in Section 3 and hence encourages parameterizations in which likelihood normality is more plausible.

Third, we are attempting to evaluate the consequences of assuming a particular model, using an analysis that is based on that very assumption. This use of the good model assumption (Section 2.2) argues for the use of DIC in comparing models that have already been shown to be adequate candidates for explaining the observations.

8. Examples

p_D and DIC have already been applied by other researchers in a variety of contexts, such as alternative models for diagnostic probabilities in screening studies (Erkanli *et al.*, 1999), longitudinal binary data using Markov regression models (Erkanli *et al.*, 2001), spline models with Bernoulli responses (Biller and Fahrmeir, 2001), multistage models for treatment usage which combine to form a total DIC (Gelfand *et al.*, 2000), complex spatial models for Poisson counts (Green and Richardson, 2000), pharmacokinetic modelling (Rahman *et al.*, 1999) and structures of Bayesian neural networks (Vehtari and Lampinen, 1999). The following examples illustrate the use of p_D and DIC to compare alternative prior and likelihood structures.

8.1. The spatial distribution of lip cancer in Scotland

We consider data on the rates of lip cancer in 56 districts 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-standardized national rate applied to the population at risk in each county) plus the 'location' of each county expressed as a list ($\mathcal{A}_i$) of its n_i adjacent counties. We assume that the cancer counts within each county y_i follow a Poisson distribution with mean $\exp(\theta_i)E_i$ where $\exp(\theta_i)$ denotes the underlying true area-specific relative risk of lip cancer. We then consider the following set of candidate models for θ_i, reflecting different assumptions about the between-county variation in (log-) relative risk of lip cancer: model 1,

$$\theta_i = \alpha_0;$$

model 2,

$$\theta_i = \alpha_0 + \gamma_i;$$

model 3,

$$\theta_i = \alpha_0 + \delta_i;$$

model 4,

$$\theta_i = \alpha_0 + \gamma_i + \delta_i;$$

model 5,

$$\theta_i = \alpha_i.$$

An improper uniform prior is placed on α_0, independent (proper) normal priors with large variance are specified for each α_i $(i = 1, \ldots, 56)$, γ_i are exchangeable random effects with a normal prior distribution having zero mean and precision λ_γ, and δ_i are spatial random effects with a conditional autoregressive prior (Besag, 1974) given by

$$\delta_i | \delta_{\setminus i} \sim \text{normal}\left(\frac{1}{n_i} \sum_{j \in \mathcal{A}_i} \delta_j, \frac{1}{n_i \lambda_\delta}\right).$$

A sum-to-zero constraint is imposed on the $\{\delta_i\}$ for identifiability, and weakly informative gamma(0.5,0.0005) priors are assumed for the random effects precision parameters λ_γ and λ_δ. These five models cover the spectrum between the pooled model 1 that makes no allowance for variation between the true risk ratios in each county and the saturated model 5 that assumes independence between the county-specific risk ratios (essentially yielding the maximum likelihood estimates $\hat{\theta}_i = \log(y_i/E_i)$). The random-effects models 2–4 allow the county-specific relative risks to be similar but not identical, with the autoregressive term allowing for the possibility of spatially correlated variation.

We use the saturated deviance (McCullagh and Nelder (1989), page 34)

$$D(\theta) = 2 \sum_i [y_i \log\{y_i / \exp(\theta_i) E_i\} - \{y_i - \exp(\theta_i) E_i\}]$$

obtained by taking $-2 \log\{f(y)\} = -2\Sigma_i \log\{p(y_i|\hat{\theta}_i)\} = 208.0$ as the standardizing factor (see Section 2.5). This allows calculation of absolute measures of fit (see Section 6.2). For model comparisons, however, it is sufficient to take the standardizing factor as $f(y) = 1$. For each model we ran two independent chains of an MCMC sampler in WinBUGS (Spiegelhalter *et al.*, 2000) for 15000 iterations each, following a burn-in period of 5000 iterations. As suggested by Dempster (1997b), Fig. 3 shows a kernel density smoothed plot of the resulting posterior distributions of the deviance under each competing model. Apart from revealing the obvious unacceptability of model 1, this clearly illustrates the difficulty of formally comparing posterior deviances on the basis of such plots alone.

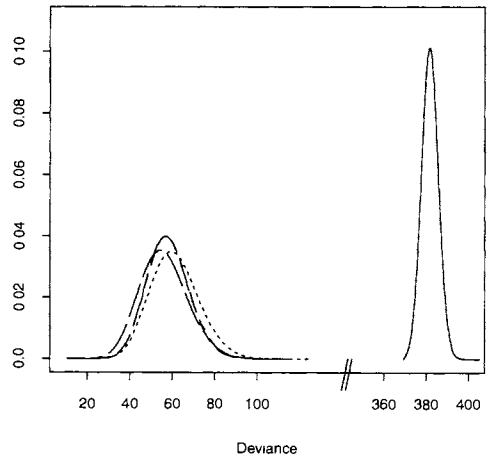

Fig. 3. Posterior distributions of the deviance for each model considered in the lip cancer example: ———, model 1; · · · · · ·, model 2; - - - - - - -, model 3; – – –, model 4; — —, model 5

Table 1. Deviance summaries for the lip cancer data using three alternative parameterizations (mean, canonical and median) for the plug-in deviance†

Model	$\bar{D}$	$D(\bar{\mu})$	p_D^{μ}	DIC^{μ}	$D(\bar{\theta})$	p_D^{θ}	DIC^{θ}	$D(med)$	p_D^{med}	DIC^{med}
1, pooled	381.7	380.7	1.0	382.7	380.7	1.0	382.7	380.7	1.0	382.7
2, exchangeable	61.1	18.2	42.9	104.0	17.7	43.4	104.5	17.6	43.5	104.6
3, spatial	58.3	26.6	31.7	89.9	27.1	31.2	89.5	27.2	31.1	89.3
4, exchangeable + spatial	57.9	26.1	31.8	89.7	26.5	31.4	89.3	26.6	31.3	89.2
5, saturated	55.9	0.0	55.9	111.7	3.1	52.8	108.6	1.4	54.5	110.4

†Exchangeable means an exchangeable random effect; spatial is a spatially correlated random effect.

The deviance summaries proposed in this paper are shown for the lip cancer data in Table 1: $\bar{D}$ is simply the mean of the posterior samples of the saturated deviance; $D(\bar{\mu})$ is calculated by plugging the posterior mean of $\mu_i = \exp(\theta_i)E_i$ into the saturated deviance; $D(\bar{\theta})$ is calculated by plugging the posterior means of the relevant parameters (α_0, α_i, γ_i and/or δ_i) into the linear predictor θ_i and then evaluating the saturated deviance; $D(\text{med})$ is calculated by plugging the posterior median of θ_i (or, equivalently, of μ_i) into the saturated deviance. The results are remarkably similar for the three alternative parameterizations of the plug-in deviance. For fixed effects models we would expect from Section 3.2 that p_D should be approximately the true number of independent parameters. For the pooled model 1, $p_D = 1.0$ as expected, whereas, for the saturated model 5, p_D ranges from 52.8 to 55.9 depending on the parameterization that is used, which is close to the true value of 56 parameters. The models containing spatial random effects (either with or without additional exchangeable effects) both have around 31 effective parameters, whereas the model with only exchangeable random effects has about 12 additional effective parameters. On the basis of the results of Section 5.2 comparing p_D for Poisson likelihoods with different priors, this suggests that the spatial model provides stronger prior information than does the exchangeable model for these data.

Turning to the comparison of DIC for each model, we first note that DIC is subject to Monte Carlo sampling error, since it is a function of stochastic quantities generated under an MCMC sampling scheme. Whereas computing the precise standard errors for our DIC values is a subject of on-going 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, either of models 3 or 4 is superior (in terms of DIC performance) to models 2 or 5, which are in turn superior to model 1. A comparison of DIC for models 3 and 4 suggests that the two spatial models are virtually indistinguishable in terms of the overall fit: pragmatically, we might prefer reporting model 3 since its DIC is only marginally greater than the more complex model 4.

Considering now the absolute measure of fit suggested in Section 6.2, we compare the values of $\bar{D}$ in Table 1 with the sample size $n = 56$. This suggests that all models except the pooled model 1 provide an adequate overall fit to the data, and that the comparison is essentially based on their complexity alone.

Following the discussion in Section 6, Fig. 4 shows a plot of deviance residuals dr_i against leverages p_{D_i} for each of the five models considered. The broken curves 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 the overall DIC for that model. For models 2–5, parabolas are marked at values of $c = 1, 2, 5$, and any data point whose contribution DIC_i is greater than 2 is labelled by its

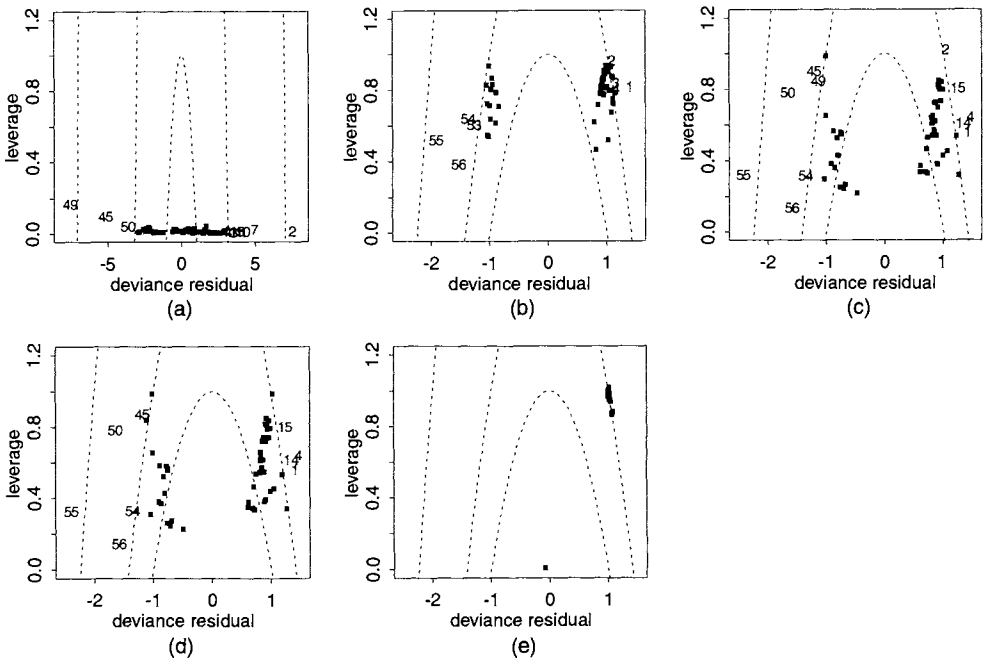

Fig. 4. Diagnostics for the lip cancer example—residuals *versus* leverages (the parabolas indicate contributions of 1, 2 or 5 to the total DIC (apart from model 1): (a) model 1; (b) model 2; (c) model 3; (d) model 4; (e) model 5

observation number. For model 1, parabolas are marked at $c = 1, 10, 50$, since the size of the deviance residuals and individual contributions to DIC are much larger and, for clarity, only points for which DIC_i is greater than 10 are marked by their observation number. Observations 55 and 56, the only districts with $y_i = 0$, are clearly identified as potential outliers under each of the random-effects models 2–4, as is observation 1 (the district with the highest observed risk ratio y_i/E_i). A few other observations (2, 3, 4, 53 and 54) have contributions DIC_i that are just larger than 2 under model 2: with the exception of the three districts already discussed, these five districts have the most extreme observed risk ratios and so their estimates tend to be shrunk furthest under the exchangeable model. Observations 14, 15, 45 and 50 appear to be outliers in models 3 and 4 which have a spatial effect, but not in the remaining models. A further investigation reveals that the observed risk ratios in these districts are extreme compared with those in each of their neighbouring districts. For example district 50 has only six cases compared with 19.6 expected, whereas each of its three neighbouring districts have high observed counts (17, 16 and 16) relative to those expected (7.8, 10.5 and 14.4). The spatial prior in models 3 and 4 causes the estimated rate in district 50 to be smoothed towards the mean of its neighbours' rates, thus leading to the discrepancy between observed and fitted values, and since the observation still exercises considerable weight on its fitted value the leverage is high as well. However, overall we might not consider that there is sufficient evidence to cast doubt on any particular observations.

8.2. Robust regression using the stack loss data

Spiegelhalter *et al.* (1996) (pages 27–29) considered a variety of error structures for the oft-analysed 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 $z_{ij} = (x_{ij} - \bar{x}_{\cdot j})/\mathrm{sd}(x_{\cdot j})$, the standardized covariates, the presence of a few prominent outliers among the $n = 21$ cases motivates a comparison of the following four error distributions: model 1,

$$y_i \sim \text{normal}(\mu_i, \tau^{-1});$$

model 2,

$$y_i \sim \text{DE}(\mu_i, \tau^{-1});$$

model 3,

$$y_i \sim \text{logistic}(\mu_i, \tau^{-1});$$

model 4,

$$y_i \sim t_d(\mu_i, \tau^{-1})$$

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

A well-known alternative to the direct fitting of many symmetric but non-normal error distributions is through scale mixtures of normals (Andrews and Mallows, 1974). From page 210 of Carlin and Louis (2000), we have the alternate t_d-formulation model 5,

$$y_i \sim \text{normal}\left(\mu_i, \frac{1}{w_i \tau}\right),$$
$$w_i \sim \frac{1}{d}\chi^2_d = \text{gamma}\left(\frac{d}{2}, \frac{d}{2}\right).$$

Unlike our other examples the form of the likelihood changes with each model, so we must use the full normalizing constants when computing $-2\log\{p(y|\mu, \tau)\}$.

Following Spiegelhalter *et al.* (1996) we set $d = 4$, and for each model we placed essentially flat priors on the β_j (actually normal with mean 0 and precision 0.00001) and $\log(\tau)$ (actually gamma(0.001,0.001) 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 2 (note that the mean parameterization and the canonical parameterization 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 models. 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 2–2.5 parameters. However, the model's smaller DIC-value implies that the extra mixing parameters are

Table 2. Deviance results for the stack loss data

Model	$\bar{D}$	$D(\bar{\theta})$	p_D	*DIC*
1, normal	110.1	105.0	5.1	115.2
2, double exponential	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

worthwhile in an overall quality-of-fit sense. We emphasize that the results from models 4 and 5 need not be equal since, although they lead to the same marginal likelihood for the y_i, they correspond to different prediction problems.

Finally, plots of deviance residuals *versus* leverages (which are not shown) clearly identify the observations determined to be 'outlying' by several previous researchers who analysed this data set.

8.3. Longitudinal binary observations: the six-cities study

To illustrate how the mean and canonical parameterizations (introduced in Section 5 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, $i = 1, \ldots, I$, $j = 1, \ldots, J$, for each of $I = 537$ children living in Stuebenville, Ohio, at $J = 4$ time points. 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

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

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 $N(0, \lambda^{-1})$ specification, which allow for dependence between the longitudinal responses for child i. The model choice issue here is to determine the most appropriate link function $g(\cdot)$ among three candidates, namely the logit, the probit and the complementary log–log-links. More formally, our three models are model 1,

$$g(p_{ij}) = \text{logit}(p_{ij}) = \log\{p_{ij}/(1 - p_{ij})\},$$

model 2,

$$g(p_{ij}) = \text{probit}(p_{ij}) = \Phi^{-1}(p_{ij}),$$

and model 3,

$$g(p_{ij}) = \text{cloglog}(p_{ij}) = \log\{-\log(1 - p_{ij})\}.$$

Table 3. Results for both parameterizations of the Bernoulli panel data

Model	$\bar{D}$	*Results for the canonical parameterization*			*Results for the mean parameterization*		
		$D(\bar{\theta})$	p_D	*DIC*	$D(\bar{\theta})$	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, complementary log–log	1180.9	956.5	224.4	1405.3	1013.7	167.2	1348.1

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

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

Placing flat priors on the β_k and a gamma(0.001,0.001) prior on λ, and running the Gibbs sampler for 5000 iterations following a burn-in period of 1000 iterations produces the deviance summaries in Table 3 for the canonical and mean parameterizations: the canonical parameterization constructs $\bar{\theta}$ as the mean of the linear predictors β and b_i, and then uses the appropriate linking transformation (logit, probit or complementary log–log) to obtain the imputed means for the p_{ij}. The mean parameterization simply uses the means of the p_{ij} themselves when computing $D(\bar{\theta})$. Natarajan and Kass (2000) have pointed out potential problems with the gamma(0.001,0.001) prior on λ, but in this context the 537 random effects ensure that these findings are robust to the choice of prior for λ.

The posterior standard deviation $\sqrt{\lambda^{-1}}$ of the random effects is estimated to be 2.2 (standard deviation 0.2), which indicates extremely high unexplained overdispersion and hence considerable prior–data conflict: this should warn us of a potential lack of robustness in our procedure. We have a sample size of $n_i = 4$ for each of $I = 537$ individuals, and an average p_{D_i} for the canonical parameterization of around 0.4–0.5. From approximation (31), this indicates a prior sample size $a + b$ of around 4–6. Referring to the evidence in Fig. 1 concerning low prior and observation sample sizes ($n_i = 1$; $a + b = 1$), we might expect the mean parameterization to display decreased complexity compared with the canonical, and this is borne out in the results. DIC prefers the complementary log–log-link under the canonical parameterization, but the probit link under the mean parameterization. We repeat that we prefer the canonical results because of the improved normality of the likelihoods and their lack of dependence on observed data: however, none of the models explain the data very well, and the lack of consensus suggests caution in using any of the models.

9. Discussion

Here we briefly discuss relationships to other suggestions and give some guidance on the practical use of the techniques described in this paper.

9.1. Relationship of p_D and DIC to other suggestions

9.1.1. Cross-validation

Stone (1977) showed the asymptotic equivalence of model comparison based on cross-validation

and AIC, whereas Wahba (1990) (page 52) showed how a generalized cross-validation criterion leads to the use of $n - \mathrm{tr}(H)$ as a denominator in the estimation of residual mean-squared error. We would expect our measure of model complexity p_D to be strongly related to cross-validatory assessment, but this requires further investigation.

9.1.2. *Other predictive loss functions*

Kass and Raftery (1995) criticized Akaike (1973) for using a plug-in predictive distribution as we have done in Section 7.3, rather than the full predictive distribution obtained by integrating out the unknown parameters. A criterion based on this predictive distribution is also invariant to reparameterizations. Laud and Ibrahim (1995) and Gelfand and Ghosh (1998) suggested minimizing a predictive 'discrepancy measure' $E\{d(Y_{\mathrm{new}}, y)|y\}$, where Y_{new} is a draw from the posterior predictive distribution $p(Y_{\mathrm{new}}|y)$, and we might for instance take $d(Y_{\mathrm{new}}, y) = (Y_{\mathrm{new}} - y)^{\mathrm{T}}(Y_{\mathrm{new}} - y)$. They showed that their measures also have attractive interpretations as weighted sums of 'goodness of fit' and 'predictive variability penalty' terms. However, a proper choice of the criterion requires fairly involved analytic work, as well as several subjective choices about the utility function that is appropriate for the problem at hand. Furthermore, the one-way ANOVA model in Section 2.5 gives rise to a fit term equivalent to $D(\bar{\theta})$, and a predictive variability term equal to $p_D + p$. Thus their suggestion is equivalent in this context to the comparison by our Bayesian measure of fit $\bar{D}$ which, although invariant to parameterization, does not seem to penalize complexity sufficiently.

In general the use of a plug-in estimate appears to 'cost' an extra penalty of p_D.

9.1.3. *Bayes factors*

Bayes factors are criteria based on a comparison of the marginal likelihoods (1) (Kass and Raftery, 1995), and a common approximation is the Bayesian (or Schwarz) information criterion (Schwarz, 1978), which for a model with p parameters and n observations is given by

$$\mathrm{BIC} = -2\log\{p(y|\hat{\theta})\} + p\log(n).$$

Bernardo and Smith (1994) (chapter 6) argued that this formulation may only be appropriate in circumstances where it was really believed that one and only one of the competing models was in fact true, and the crucial issue was to choose this correct model, and that in other circumstances criteria based on short-term prediction, such as cross-validation, may be more appropriate. We support this view and refer to Han and Carlin (2001) for a review of some of the computational and conceptual difficulties in using Bayes factors to compare complex hierarchical models. Whether DIC can be justified as a basis for model averaging remains open for investigation.

9.2. Practical issues in using DIC

9.2.1. *Invariance*

p_D may be only approximately invariant to the chosen parameterization, since different fitted deviances $D(\bar{\theta})$ may arise from substituting posterior means of alternative choices of θ. The example in Section 8.3 shows that this choice could be important with Bernoulli data.

In Section 5 we explored the use of the posterior median as an estimator leading to an invariant p_D. This has two possible disadvantages: we do not have a proof that p_D will be positive and some additional computational difficulty in that the full sample needs to be retained. In addition the approximate properties based on Taylor series expansions in Section 3 may not hold, although

this may be only of theoretical interest. Currently we recommend calculation of DIC on the basis of several different estimators, with a preference for posterior means based on parameterizations obeying approximate likelihood normality.

9.2.2. Focus of analysis

As we saw in the stack loss example of Section 8.2, there may be sensitivity to apparently innocuous restructuring of the model: this is to be expected since by making such changes we are altering the definition of a replicate data set, and hence one would expect DIC to change. For example, consider a model comprising a mixture of normal distributions. If this assumption was solely to obtain a flexible functional form, then the appropriate likelihood would comprise the mixture. If, however, we were interested in the membership of individual observations, then the likelihoods would be normal and the membership variables would contribute to the complexity of the model. Thus the parameters in the focus of a model should ideally depend on the purpose of the investigation, although in practice it is likely that the focus may be chosen on computational grounds as providing likelihoods that are available in closed form.

9.2.3. Nuisance parameters

Strictly speaking, nuisance parameters should first be integrated out to leave a likelihood depending solely on parameters in focus. In practice, however, parameters such as variances are likely to be included in the focus and add to the estimated complexity: we would recommend posterior means of log-variances as estimators.

9.2.4. What is an important difference in DIC?

Burnham and Anderson (1998) suggested models receiving AIC within 1–2 of the 'best' deserve consideration, and 3–7 have considerably less support: these rules of thumb appear to work reasonably well for DIC. Certainly we would like to ensure that differences are not due to Monte Carlo error: although this is straightforward for $\bar{D}$, Zhu and Carlin (2000) have explored the difficulty of assessing the Monte Carlo error on DIC.

9.2.5. Asymptotic consistency

As with AIC, DIC will not consistently select the true model from a fixed set with increasing sample sizes. We are not greatly concerned about this: we neither believe in a true model nor would expect the list of models being considered to remain static as the sample size increased.

9.3. Conclusion

In conclusion, our suggestions have a similar 'information theoretic' background to frequentist measures of model complexity and criteria for model comparison but are based on expectations with respect to parameters in place of sampling expectations. DIC can thus be viewed as a Bayesian analogue of AIC, with a similar justification but wider applicability. It is also applicable to any class of model, involves negligible additional analytic work or Monte Carlo sampling and appears to perform reasonably across a range of examples. We feel that p_D and DIC deserve further investigation as tools for model assessment and comparison.

Acknowledgements

We are very grateful for the generous discussion and criticism of the participants in the pro-

gramme on neural networks and machine learning that was held at the Isaac Newton Institute for Mathematical Sciences in 1997, and to Andrew Thomas for so quickly implementing our changing ideas into WinBUGS. NGB received partial support from Medical Research Council grant G9803841, and BPC received partial support from National Institute of Allergy and Infectious Diseases grant 1-R01-AI41966.

References

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Information Theory* (eds B. N. Petrov and F. Csáki), pp. 267–281. Budapest: Akadémiai Kiadó.

Andrews, D. F. and Mallows, C. L. (1974) Scale mixtures of normal distributions. *J. R. Statist. Soc.* B, **36**, 99–102.

Berk, R. H. (1966) Limiting behaviour of posterior distributions when the model is incorrect. *Ann. Math. Statist.*, **37**, 51–58.

Bernardo, J. M. (1979) Expected information as expected utility. *Ann. Statist.*, **7**, 686–690.

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

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Statist. Soc.* B, **36**, 192–236.

Biller, C. and Fahrmeir, L. (2001) Bayesian varying-coefficient models using adaptive regression splines. *Statist. Modlng*, **1**, 195–211.

Box, G. E. P. (1976) Science and statistics. *J. Am. Statist. Ass.*, **71**, 791–799.

Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalized linear mixed models. *J. Am. Statist. Ass.*, **88**, 9–25.

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

Bunke, O. and Milhaud, X. (1998) Asymptotic behaviour of Bayes estimates under possibly incorrect models. *Ann. Statist.*, **26**, 617–644.

Burnham, K. P. and Anderson, D. R. (1998) *Model Selection and Inference.* New York: Springer.

Carlin, B. P. and Louis, T. A. (2000) *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd edn. Boca Raton: Chapman and Hall–CRC Press.

Chib, S. and Greenberg, E. (1998) Analysis of multivariate probit models. *Biometrika*, **85**, 347–361.

Clayton, D. G. and Kaldor, J. (1987) Empirical Bayes estimates of age-standardised relative risks for use in disease mapping. *Biometrics*, **43**, 671–681.

Dempster, A. P. (1974) The direct use of likelihood for significance testing. In *Proc. Conf. Foundational Questions in Statistical Inference* (eds O. Barndorff-Nielsen, P. Blaesild and G. Schou), pp. 335–352. Aarhus: University of Aarhus.

——— (1997a) The direct use of likelihood for significance testing. *Statist. Comput.*, **7**, 247–252.

——— (1997b) Commentary on the paper by Murray Aitkin, and on discussion by Mervyn Stone. *Statist. Comput.*, **7**, 265–269.

Efron, B. (1986) How biased is the apparent error rate of a prediction rule? *J. Am. Statist. Ass.*, **81**, 461–470.

Erkanli, A., Soyer, R. and Angold, A. (2001) Bayesian analyses of longitudinal binary data using markov regression models of unknown order. *Statist. Med.*, **20**, 755–770.

Erkanli, A., Soyer, R. and Costello, E. (1999) Bayesian inference for prevalence in longitudinal two-phase studies. *Biometrics*, **55**, 1145–1150.

Eubank, R. L. (1985) Diagnostics for smoothing splines. *J. R. Statist. Soc.* B, **47**, 332–341.

Eubank, R. and Gunst, R. (1986) Diagnostics for penalized least-squares estimators. *Statist. Probab. Lett.*, **4**, 265–272.

Fitzmaurice, G. and Laird, N. (1993) A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, **80**, 141–151.

Gelfand, A. E. and Dey, D. K. (1994) Bayesian model choice: asymptotics and exact calculations. *J. R. Statist. Soc.* B, **56**, 501–514.

Gelfand, A. E., Ecker, M. D., Christiansen, C., McLaughlin, T. J. and Soumerai, S. B. (2000) Conditional categorical response models with application to treatment of acute myocardial infarction. *Appl. Statist.*, **49**, 171–186.

Gelfand, A. and Ghosh, S. (1998) Model choice: a minimum posterior predictive loss approach. *Biometrika*, **85**, 1–11.

Gelfand, A. E. and Trevisani, M. (2002) Inequalities between expected marginal log likelihoods with implications for likelihood-based model comparison. *Technical Report.* Department of Statistics, University of Connecticut, Storrs.

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

Gilks, W. R., Wang, C. C., Coursaget, P. and Yvonnet, B. (1993) Random-effects models for longitudinal data using Gibbs sampling. *Biometrics*, **49**, 441–453.

Good, I. J. (1956) The surprise index for the multivariate normal distribution. *Ann. Math. Statist.*, **27**, 1130–1135.

Green, P. and Richardson, S. (2002) Hidden Markov models and disease mapping. *J. Am. Statist. Ass.*, to be published.
Han, C. and Carlin, B. (2001) MCMC methods for computing Bayes factors: a comparative review. *J. Am. Statist. Ass.*, **96**, 1122–1132.
Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models.* London: Chapman and Hall.
Hodges, J. and Sargent, D. (2001) Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika*, **88**, 367–379.
Huber, P. J. (1967) The behaviour of maximum likelihood estimates under non-standard conditions. In *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability* (eds L. M. LeCam and J. Neyman), vol. 1, pp. 221–233. Berkeley: University of California Press.
Kass, R. and Raftery, A. (1995) Bayes factors and model uncertainty. *J. Am. Statist. Ass.*, **90**, 773–795.
Key, J. T., Pericchi, L. R. and Smith, A. F. M. (1999) Bayesian model choice: what and why? In *Bayesian Statistics 6* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 343–370. Oxford: Oxford University Press.
Kimeldorf, G. and Wahba, G. (1970) A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist.*, **41**, 495–502.
Kullback, S. and Leibler, R. A. (1951) On information and sufficiency. *Ann. Math. Statist.*, **22**, 79–86.
Laird, N. M. and Ware, J. H. (1982) Random effects models for longitudinal data. *Biometrics*, **38**, 963–974.
Laud, P. W. and Ibrahim, J. G. (1995) Predictive model selection. *J. R. Statist. Soc.* B, **57**, 247–262.
Lee, Y. and Nelder, J. A. (1996) Hierarchical generalized linear models (with discussion). *J. R. Statist. Soc.* B, **58**, 619–678.
van der Linde, A. (1995) Splines from a Bayesian point of view. *Test*, **4**, 63–81.
——(2000) Reference priors for shrinkage and smoothing parameters. *J. Statist. Planng Inf.*, **90**, 245–274.
Lindley, D. V. and Smith, A. F. M. (1972) Bayes estimates for the linear model (with discussion). *J. R. Statist. Soc.* B, **34**, 1–44.
MacKay, D. J. C. (1992) Bayesian interpolation. *Neur. Computn*, **4**, 415–447.
——(1995) Probable networks and plausible predictions—a review of practical Bayesian methods for supervised neural networks. *Netwrk Computn Neur. Syst.*, **6**, 469–505.
McCullagh, P. and Nelder, J. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.
Meng, X.-L. and Rubin, D. B. (1992) Performing likelihood ratio tests with multiply imputed data sets. *Biometrika*, **79**, 103–112.
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* (eds J. E. Moody, S. J. Hanson and R. P. Lippmann), pp. 847–854. San Mateo: Morgan Kaufmann.
Murata, N., Yoshizawa, S. and Amari, S. (1994) Network information criterion—determining the number of hidden units for artificial neural network models. *IEEE Trans. Neur. Netwrks*, **5**, 865–872.
Natarajan, R. and Kass, R. E. (2000) Reference Bayesian methods for generalised linear mixed models. *J. Am. Statist. Ass.*, **95**, 227–237.
Raghunathan, T. E. (1988) A Bayesian model selection criterion. *Technical Report.* University of Washington, Seattle.
Rahman, N. J., Wakefield, J. C., Stephens, D. A. and Falcoz, C. (1999) The Bayesian analysis of a pivotal pharmacokinetic study. *Statist. Meth. Med. Res.*, **8**, 195–216.
Richardson, S. and Green, P. J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. R. Statist. Soc.* B, **59**, 731–792.
Ripley, B. D. (1996) *Pattern Recognition and Neural Networks.* Cambridge: Cambridge University Press.
Sawa, T. (1978) Information criteria for choice of regression models: a comment. *Econometrica*, **46**, 1273–1291.
Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–466.
Slate, E. (1994) Parameterizations for natural exponential-families with quadratic variance functions. *J. Am. Statist. Ass.*, **89**, 1471–1482.
Spiegelhalter, D. J., Thomas, A. and Best, N. G. (2000) *WinBUGS Version 1.3 User Manual.* Cambridge: Medical Research Council Biostatistics Unit. (Available from `http://www.mrc-bsu.cam.ac.uk/bugs`.)
Spiegelhalter, D. J., Thomas, A., Best, N. G. and Gilks, W. R. (1996) *BUGS Examples Volume 1, Version 0.5 (Version ii).* Cambridge: Medical Research Council Biostatistics Unit.
Stone, M. (1977) An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. R. Statist. Soc.* B, **39**, 44–47.
Takeuchi, K. (1976) Distribution of informational statistics and a criterion for model fitting (in Japanese). *Suri-Kagaku*, **153**, 12–18.
Vehtari, A. and Lampinen, J. (1999) Bayesian neural networks with correlated residuals. In *IJCNN'99: Proc. 1999 Int. Joint Conf. Neural Networks.* New York: Institute of Electrical and Electronic Engineers.
Wahba, G. (1978) Improper priors, spline smoothing and the problem of guarding against model errors in regressions. *J. R. Statist. Soc.* B, **40**, 364–372.
——(1983) Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. R. Statist. Soc.* B, **45**, 133–150.

———(1990) *Spline Models for Observational Data.* Philadelphia: Society for Industrial and Applied Mathematics.
Ye, J. (1998) On measuring and correcting the effects of data mining and model selection. *J. Am. Statist. Ass.*, **93**, 120–131.
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, Chicago.
Zeger, S. L. and Karim, M. R. (1991) Generalised linear models with random effects; a Gibbs sampling approach. *J. Am. Statist. Ass.*, **86**, 79–86.
Zhu, L. and Carlin, B. (2000) Comparing hierarchical models for spatio-temporally misaligned data using the deviance information criterion. *Statist. Med.*, **19**, 2265–2278.

Discussion on the paper by Spiegelhalter, Best, Carlin and van der Linde

S. P. Brooks (*University of Cambridge*)

This is a wonderful paper containing a wide array of interesting ideas. It seems to me very much like a first step (and in the right direction) and I am sure that it will be seen as both a focus and a source of inspiration for future developments in this area.

As the authors point out, their p_D and the deviance information criterion (DIC) statistics have already been widely used within the Bayesian literature. Given this history and in the previous absence of a published source for these ideas, it is easy to misunderstand what p_D actually does. Certainly, before reading this paper, but having read several others which use the DIC, I thought that the p_D-statistic was a clever way of avoiding the problem that Bayesians have when it comes to calculating the number of parameters in any hierarchical model. Essentially the problem is one of deciding which variables in the posterior are model parameters and which are hyperparameters arising from the prior. However, p_D does not help us here and that is why we have Section 2.1 explaining that this choice is up to the reader. The authors refer to this as choosing the 'focus' for the analysis. Sadly, in many cases the calculation of p_D will be impossible for the focus of primary interest since the deviance will not be available in closed from (this includes random effects and state space models, for example), so this remains an open problem.

What p_D *does* do is to tell you, once you have chosen your focus, how many parameters you lose (or even gain?) by being Bayesian. The number of degrees of freedom (or parameters) in a model is clear from the (focused) likelihood. However, by combining the likelihood with the prior we almost always impose additional restrictions on the parameter space, effectively reducing the degrees of freedom of our model. Take the authors' saturated model of Section 8.1, in which parameters $\alpha_1, \ldots, \alpha_{56}$ are given a prior with some unknown mean μ and fixed variance σ^2. Clearly, in the limit as σ^2 goes to 0, we essentially remove the 56 individual parameters α_i and effectively replace them with a single parameter μ. I guess that this is fairly obvious with hindsight as is the case with many great ideas. None-the-less it is a credit to the authors firstly for seeing it and, more importantly, for actually deriving a procedure for dealing with it.

This prior-induced parameter reduction can be clearly observed in Fig. 5 in which we plot the value of p_D^θ against $\log(\sigma^2)$ both for a hyperprior $\mu \sim N(0, 1000)$ and for $\mu = 0$ (the authors are unclear about which, if either, they actually use in Section 8.1). We can see that, as σ^2 decreases, the effective number of parameters decreases to either 1 or 0 depending on whether or not μ itself is a parameter, i.e. which prior is chosen. It is interesting to note the rapid decline in p_D for variances between 1 and 0.01, but what is particularly interesting about this plot is that, as σ^2 increases, p_D converges to a fixed maximum well below 56, the number of parameters in the likelihood. As an experiment, if we take $\sigma^2 = 10^{30}$ or even the Jeffreys prior for the μ_i, a value for p_D exceeding 53.1 is never obtained (modulo Monte Carlo error). This suggests that we automatically lose three parameters just by being Bayesian, even if we are as vague as we could possibly be with our prior. Quoting Bernardo and Smith (1994), page 298, 'every prior specification has *some* informative posterior or predictive implications .... There is no "objective" prior that represents ignorance.' Of course, the authors' Table 1 suggests that if we took the median as the basis for the calculation of p_D then we might obtain different results; indeed we seem to regain several parameters this way! Unfortunately, analytic investigation of the p_D-statistic is essentially limited to the case where we take $\tilde{\theta}(y)$ to be the posterior mean, so we have little idea of the extent and nature of the variability across parameterizations. This choice is likely to have a significant effect on any inference based on the corresponding p_D-statistic and further (no doubt simulation-based) investigation along these lines would certainly be very helpful.

As well as the construction of the p_D-statistic, the paper also derives a new criterion for model comparison labelled the DIC. The authors provide a heuristic justification for the DIC, but there are clearly several alternatives. One obvious extension of the usual Akaike information criterion (AIC) statistic to

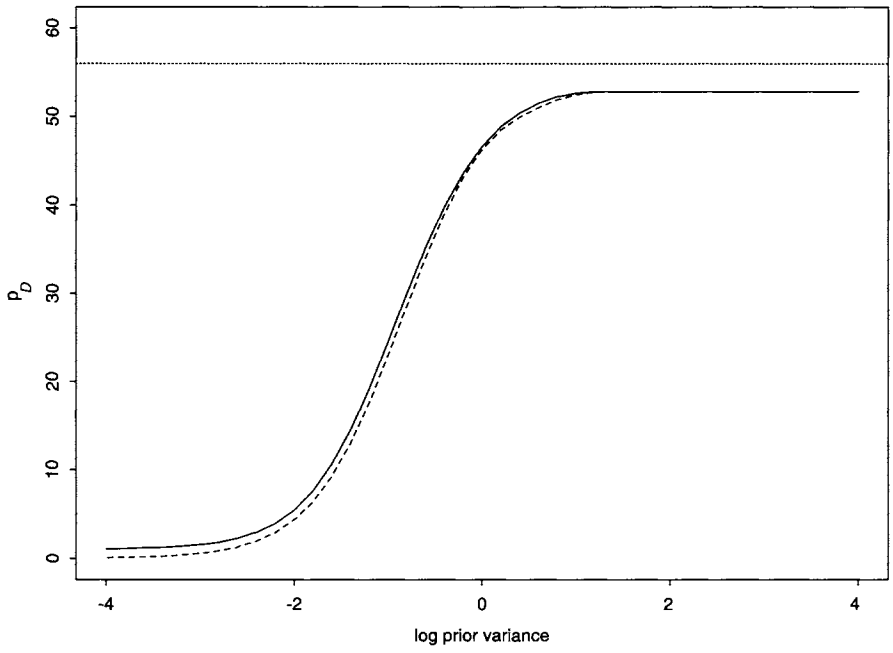

Fig. 5. Plot of p_D^{θ} for the saturated model of Section 8.1 demonstrating its dependence on the prior variance for the random effects: ———, p_D-statistic with an $N(0, 1000)$ hyperprior for μ: - - - - -, corresponding value when we fix $\mu = 0$; ·····, number of parameters in the likelihood

the Bayesian context is to calculate its posterior expectation, EAIC $= \overline{D(\theta)} + 2p$ (rather than evaluating it at the posterior mode under a flat prior), or to take the deviance calculated at the posterior mean, i.e. taking $D(\bar{\theta}) + 2p$. Of course, as with the DIC, posterior medians, modes etc. could also be taken and similar extensions could be applied to the corrected AIC statistic and the Bayesian information criterion for example. Further, the number of parameters in each of these expressions might be replaced by p_D to gain even more potential criteria. Table 4 gives the posterior model probabilities and posterior-averaged information criteria (based on p, rather than p_D), including DIC, for autoregressive models of various orders fitted to the well-known lynx data (Priestley (1981), section 5.5). We note the broad agreement between the DIC, EAIC and EAIC_c (as is common in my own experience and, I think, expected by the authors), but that EBIC locates an entirely different model. We note also that the posterior model probabilities correctly identify the fact that two models appear to describe the data well and it is the only criterion to identify correctly the existence of two distinct modes in the posterior.

Given the number of approximations and assumptions that are required to obtain the DIC it can only really be used as a broad brush technique for discriminating between obviously disparate models, in much the same way as any of the alternative information criteria suggested above might be used. However, in many realistic applications there may be two or more models with sufficiently similar DIC that it is impossible to choose between the two. The only sensible choice in this circumstance is to model-average (see Section 9.1.3). Burnham and Anderson (1998), section 4.2, suggested the use of AIC weights and these are also given in Table 4 together with the corresponding weights for the other criteria. Essentially, these are obtained by subtracting from each AIC the value associated with the 'best' model and then setting

$$w_k \propto \exp\{-\Delta\text{AIC}(k)/2\}$$

where $\Delta\text{AIC}(k)$ denotes the transformed AIC-value for model k. These weights are then normalized to sum to 1 over the models under consideration.

Note the distinct differences between the weights and the posterior model probabilities given in Table 4, suggesting that only one or the other can really make any sense. We note here that similar comparisons have been made in the context of other examples. In the context of a log-linear contingency

Table 4. Effective number of parameters, values of DIC and the posterior expectation of various information criteria for fitting an autoregressive model of order k (with $k + 1$ parameters including the error variance) to the lynx data†

k	p_D	*DIC*	*EAIC*	*EBIC*	$EAIC_c$	$\pi(K = k)$	w_k^{DIC}	w_k^{EAIC}	w_k^{EBIC}	w^{EAIC_c}
1	1.88	206.66	206.78	209.51	206.81	0.000	0.000	0.000	0.000	0.000
2	2.85	126.58	127.72	**133.19**	127.83	**0.243**	0.000	0.003	**0.858**	0.011
3	3.78	127.06	129.27	*137.48*	129.50	0.016	0.000	0.001	**0.101**	0.005
4	4.76	125.52	128.75	139.70	129.12	0.007	0.000	0.002	0.033	0.006
5	5.70	125.23	129.52	143.20	130.08	0.002	0.000	0.001	0.006	0.004
6	6.62	126.30	131.68	148.09	132.46	0.001	0.000	0.004	0.000	0.001
7	7.60	122.34	128.72	147.88	129.78	0.002	0.000	0.002	0.001	0.004
8	8.61	121.81	129.19	151.08	130.56	0.002	0.000	0.001	0.000	0.003
9	9.58	122.75	131.16	155.79	132.89	0.001	0.000	0.001	0.000	0.001
10	10.54	118.94	128.40	155.76	130.53	0.002	0.001	0.002	0.000	0.003
11	11.33	**106.51**	**117.16**	147.26	**119.75**	0.154	**0.431**	**0.566**	0.001	**0.624**
12	12.61	*106.89*	*118.27*	151.10	*121.36*	**0.268**	**0.356**	**0.325**	0.000	**0.280**
13	13.56	*108.74*	*121.17*	156.74	*124.81*	0.135	0.142	0.076	0.000	0.050
14	14.46	110.77	124.30	162.61	128.54	0.067	0.051	0.016	0.000	0.008
15	15.37	112.896	127.42	168.47	132.32	0.000	0.019	0.003	0.000	0.001

†Criterion entries in bold indicate the model minimizing the relevant criterion, whereas those in italics denote alternative plausible models under the rules of thumb discussed in Section 9.2.4. Probabilities π or weights w in bold denote the top two models in each case. Here, $EAIC_c$ denotes the posterior mean of the corrected EAIC (Burnham and Anderson, 1998), $\pi(K = k)$ the corresponding posterior model probability under a flat prior across models and the w_k^X the corresponding Akaike weights (or equivalent). The posterior model probabilities were kindly provided by Ricardo Ehlers.

table analysis, King (2001), Table 2.5, found that two models have posterior probability 0.557 and 0.057 but corresponding DIC weights of 0.062 and 0.682 respectively. Similar examples in which the DIC and posterior model probabilities give wildly different results are provided by King and Brooks (2001). Do the authors have any feel for why these two approaches might give such different results? Which would they recommend be used and do they have any suggestions for alternative DIC-based weights for model averaging which might lead to more sensible results? Surely, the only sensible approach is to calculate posterior model probabilities via transdimensional Markov chain Monte Carlo methods. When, then, do the authors suggest that the DIC might be used? What, in practical terms is the question that the DIC is answering as opposed to the posterior model probabilities?

The incorporation of the DIC-statistic into WinBUGS 1.4 ensures its ultimate success, but I have grave misgivings concerning the blind application of a 'default' DIC-statistic for model determination problems particularly given its heuristic derivation and the series of essentially arbitrary assumptions and approximations on which it is based. The authors 'recommend calculation of DIC on the basis of several different estimators'. The option to choose different parameterizations is not available in the beta version of WinBUGS 1.4; will it be added to later versions? What about options for the all-important choice of focus? What do the authors suggest we do when the same parameterization is not calculable for all models being compared? Could not the choice of parameterization for each model adversely influence the results, particularly for models with large numbers of parameters (where a small percentage change in p_D might mean a large absolute change in the corresponding DIC)?

The paper, like any good discussion paper, leaves various other open questions. For example: why take $\mathbb{E}_{\theta|y}[d_\Theta]$ in equation (9) and not the mode or median; how should we decide when to take $\hat{\theta}$ to be the mean, median, mode etc. as this will surely lead to different comparative results for the DIC; when is p_D negative and why; in an entirely practical sense, how does model comparison with the DIC compare with that via posterior model probabilities and why do they differ—can both be 'correct' in any meaningful way? On page 613, the authors write 'p_D and DIC deserve *further investigation* as tools for model assessment and comparison' and I would certainly agree that they do. I have very much enjoyed thinking about some of these ideas over the past few weeks and I am very grateful to the authors for the opportunity and motivation to do so. It therefore gives me great pleasure to propose the vote of thanks.

Jim Smith (*University of Warwick, Coventry*)

I shall not address technical inaccuracies but just present four foundational problems that I have with the model selection in this paper.

(a) Bayesian models are designed to make plausible predictive statements about future observables. The predictive implications of all the prior settings on variances in the worked examples in Section 8 are unbelievable. They do not represent carefully elicited expert judgments but the views of a vacuous software user. Early in Section 1 the authors state that they want to identify succinct models 'which appear to describe the information [about wrong "true" parameter values (see Section 2.2)?] in the *data* accurately'. But in a Bayesian analysis a separation between information in the data and in the prior is artificial and inappropriate. For example where do I input extraneous data used as the basis of my prior? When do I stop calling this data (and so include it in $D(\cdot)$) and instead call it prior information? This forces the authors to use default priors.

A Bayesian analysis on behalf of a remote auditing expert (Smith, 1996) might require the selection of a prior that is robust within a *class* of belief of different experts (e.g. Pericchi and Walley (1991)). Default priors can sometimes be justified for simple models. Even then, models within a selection class need to have compatible parameterizations: see Moreno *et al.* (1998). However, in examples where 'the number of parameters outnumbers observations'—they claim their approach addresses—default priors are unlikely to exhibit any robustness. In particular, outside the domain of vague location estimation or separating variance estimation (discussed in Section 4), apparently default priors can have strong influence on model implications and hence selection.

(b) Suppose that we need to select models whose predictive implications we do not believe. Surely we should try to ensure that prior information in each model corresponds to predictive statements that are comparable. Such issues, not addressed here, are considered by Madigan and Raftery (1991) for simple discrete Bayesian models. But outside linear models with known variances this is a difficult problem. Furthermore it is well known that calibration is a fast function (Cooke, 1991). In particular apparently inconsequential deviations from the features of a model 'not in focus' tend to dominate $D(\theta)$ and $\overline{D(\theta)}$. A trivial example of this occurs when we plan to forecast X_2 having observed an independent identically distributed $X_1 = 0.01$ which under models M1 and M2 have respective Gaussian distributions $N(100, 10000)$ and $N(0, 0.001)$. Then, for most priors, model M1 is strongly preferred although its predictions about X_2 are less 'useful' (Section 2.2). The authors' premise that all the models they entertain are 'wrong' allows these calibration issues to bite theoretically even in the limit, unlike their asymptotically consistent rivals. The authors, however, do no more than to acknowledge the existence of this core difficulty after the example in Section 8.3.

(c) Suppose that problems (a) and (b) do not bite. Then the 'vector of parameters of focus' (POF) will have a critical influence on any ensuing inference. How in practice do we specify this? The authors state without elaboration that this 'should depend on the purpose of the investigation' (Section 9.2.2). But it appears that in practice the POF is calculated on 'computational grounds', their software capability driving their inference.

The high influence of the choice of the POF is illustrated in the example in Section 8.2. Here models 4 and 5 are predictively identical but model 5 has a significantly smaller deviance information criterion DIC than model 4. The authors conclude that 'the extra mixing parameters are worthwhile': why? In what practical sense is this helpful? This example illustrates that the unguided choice of the POF will often be inferentially critical. Incidentally in this example the order of DIC is not (as stated) consistent with the thickness of tails of the sample distribution, the thickest-tailed distribution being model 4.

(d) But ignoring all these difficulties there still remains the acknowledged choice of (re)parameterization governing the choice of $\bar{\theta}$ which initially we shall assume to be the mean. Consider the case when the POF θ is one dimensional with strictly increasing posterior distribution function $F(\theta|y)$, and G_μ is a distribution function of a random variable with mean μ. Then the reparameterization of θ to $\phi_\mu = G_\mu^{-1}\{F(\theta|y)\}$ has $E(\phi_\mu) = \mu$. Thus $D(\bar{\theta})$ (or $D(\bar{\phi})$) is arbitrary within the range of $D(\cdot)$. Thus, contrary to Section (5.1.4), the choice of parameterization of θ with non-degenerate posterior will always be critical. But no *general* selection guidance is given here. In observation (c) of Section 2.6 the authors suggest the use of the posterior median instead of the mean if this can be calculated easily from their output: not a solution when the POF is more than one dimensional. Even familiar transforms of marginal medians to contrasts and means or means and variances to means and coefficients of variation will not exhibit the required sorts of invariance.

There may be theoretical reasons to use DIC but I do not believe that this paper gives them. So my suggestion to a practitioner would be: if you must use a formal selection criterion do not use DIC. I second the vote of thanks.

The vote of thanks was passed by acclamation.

Aki Vehtari (*Helsinki University of Technology*)
The authors mention that the deviance information criterion DIC estimates the expected loss, with deviance as the loss function. This connection should be emphasized more. It should be remembered that the estimation of the expected deviance was Akaike's motivation for deriving the very first information criterion AIC (Akaike, 1973). In prediction and decision problems, it is natural to assess the predictive ability of the model by estimating the expected utilities, as the principle of rational decisions is based on maximizing the expected utility (Good, 1952) and the maximization of expected likelihood maximizes the information gained (Bernardo, 1979). It is often useful to use other than likelihood-based utilities. For example, in classification problems it is much more meaningful for the application expert to know the expected classification accuracy than just the expected deviance value (Vehtari, 2001). Given an arbitrary utility function u, it is possible to use Monte Carlo samples to estimate $E_\theta[\bar{u}(\theta)]$ and $\bar{u}(E_\theta[\theta])$, and then to compute an expected utility estimate as

$$\bar{u}_{\mathrm{DIC}} = \bar{u}(E_\theta[\theta]) + 2\{E_\theta[\bar{u}(\theta)] - \bar{u}(E_\theta[\theta])\},$$

which is a generalization of DIC (Vehtari, 2001).

The authors also mention the known asymptotic relationship of AIC to cross-validation (CV). Equally important is to note that the same asymptotic relationship holds also for NIC (Stone (1977), equation (4.5)). The asymptotic relationship is not surprising, as it is known that CV can also be used to estimate expected utilities with Bayesian justification (Bernardo and Smith (1994), chapter 6, Vehtari (2001) and Vehtari and Lampinen (2002a)). Below some main differences between CV and DIC are listed. See Vehtari (2001) and Vehtari and Lampinen (2002b) for full discussion and empirical comparisons. CV can use full predictive distributions. In the CV approach, there are no parameterization problems, as it deals directly with predictive distributions. CV estimates the expected utility directly, but it can also be used to estimate the effective number of parameters if desired. In the CV approach, it is easy to estimate the distributions of the expected utility estimates, which can for example be used to determine automatically whether the difference between two models is 'important'. Importance sampling leave-one-out CV (Gelfand *et al.*, 1992; Gelfand, 1996) is computationally as light as DIC, but it seems to be numerically more unstable. k-fold CV is very stable and reliable, but it requires k times more computation time to use. k-fold CV can also handle finite range dependences in the data. For example, in the six-cities study, the wheezing statuses of a single child at different ages are not independent. DIC, which assumes independence, underestimates the expected deviance. In k-fold CV it is possible to group the dependent data and to handle independent groups and thus to obtain better estimates (Vehtari, 2001; Vehtari and Lampinen, 2002b).

Martyn Plummer (*International Agency for Research on Cancer, Lyon*)
I congratulate the authors on their thought-provoking paper. I would like to offer one constructive suggestion and one criticism.

Firstly, I have a proposal for a modified definition of the effective number of parameters p_D. Starting from the Kullback–Leibler information divergence between the predictive distributions at two different values of θ

$$I(\theta^0, \theta^1) = E_{Y_{\mathrm{rep}}|\theta^0}\left[\log\left\{\frac{p(Y_{\mathrm{rep}}|\theta^0)}{p(Y_{\mathrm{rep}}|\theta^1)}\right\}\right],$$

I suggest that p_D be defined as the expected value of $I(\theta^0, \theta^1)$ when θ^0 and θ^1 are independent samples from the posterior distribution of θ. This modified definition yields exactly the same expression for p_D in the normal linear model with known variance. In general, it should give a similar estimate of p_D when θ has an asymptotic normal distribution. This version of p_D can also be decomposed into influence diagnostics when the likelihood factorizes as in Section 6.3. It has the theoretical advantages of being non-negative and co-ordinate free. A practical advantage is that p_D can be estimated via Markov chain Monte Carlo sampling using two parallel chains by taking the sample average of

$$\log\left\{\frac{p(Y^0_{\mathrm{rep}}|\theta^0)}{p(Y^0_{\mathrm{rep}}|\theta^1)}\right\}$$

where the superscript denotes the chain to which each quantity belongs. The Monte Carlo error of this estimate is easily calculated and the difficulties discussed by Zhu and Carlin (2000) can thus be avoided.

For exponential family models, $I(\theta^0, \theta^1)$ can be expressed in closed form and there is no need to simulate replicate observations Y_{rep}. When the scale parameter ϕ is known, the expression for p_{D_i} simplifies to

$$p_{D_i} = n_i w_i \operatorname{cov}\{\theta_i, \mu(\theta_i)|Y\} / \phi.$$

This gives a surprising resolution to the problem of whether to use the canonical or mean parameterization to estimate p_D.

On a more negative note, I am not convinced by the heuristic derivation of the deviance information criterion DIC in Section 7.3. I followed this derivation for the linear model of Section 4.1, for which it is not necessary to make any approximations. The term with expectation 0, neglected in the final expression, is $p - p_D - D(\tilde{\theta})$. Adding this to DIC gives an expected loss of $p + p_D$ which is not useful as a model choice criterion. I am not suggesting that the use of DIC is wrong, but a formal derivation is lacking.

Mervyn Stone (*University College London*)

The paper is rather economical with the 'truth'. The *truth* of $p^{\text{t}}(Y)$ corresponds fixedly to the *conditions* of the experimental or observational set-up that ensures independent future replication Y_{rep} or internal independence of $y = \mathbf{y} = (y_1, \ldots, y_n)$ (not excluding an implicit concomitant x). For $p^{\text{t}}(Y) \approx p(Y|\theta^{\text{t}})$, θ must parameterize a scientifically plausible family of alternative distributions of Y under those conditions and is therefore a *necessary* 'focus' if the 'good [true] model' idea is to be invoked: think of tossing a bent coin. Changing focus is not an option.

Any connection of p_D with cross-validatory assessment would need truth as $p^{\text{t}}(\mathbf{y}) = p^{\text{t}}(y_1) \ldots p^{\text{t}}(y_n)$. If $l = \log(p)$ is an acceptable measure of predictive success, $A = \Sigma_i\, l(y_i|\tilde{\theta}_{-i})$ is a one-out estimate of $E_{p^{\text{t}}(\mathbf{Y})}[\Sigma_i\, l\{Y_i|\tilde{\theta}(\mathbf{y})\}]$. Multiplied by -2, this connects with equation (33) only when the θ-model is true with $Y_1, \ldots, Y_n$ independent.

Extending Stone (1977) to the posterior mode for prior $p(\theta)$, with n large, $A \approx L_{\tilde{\theta}}(\mathbf{y}) - \Pi(\mathbf{y})$ where

$$\Pi(\mathbf{y}) = -\operatorname{tr}\{L''_{\tilde{\theta}} + l''(\tilde{\theta})\}^{-1} \sum_i l'_{\tilde{\theta}}(y_i) l'_{\tilde{\theta}}(y_i)^{\text{T}}$$

and $l(\theta) = \log\{p(\theta)\}$. If $l''(\tilde{\theta})$ is negative definite, the typically non-negative penalty $\Pi(\mathbf{y})$ is smaller for the posterior mode than for the maximum likelihood estimate. For the maximum likelihood estimate, $l''(\tilde{\theta}) = \mathbf{O}$ gives $\Pi(\mathbf{y})$ estimating p^*, but the general form probably gives Ripley's p^*.

If Section 7.3 could be rigorously developed (the use of $E_{\mathbf{Y}}$ does look suspicious!), another connection (via equation (33)) might be that DIC $\approx -2A$. But, since Section 7.3 invokes the 'good model' assumption and small $|\tilde{\theta} - \theta|$ for the Taylor series expansion (i.e. large n), such a connection would be as contrived as that of A with the Akaike information criterion: why not stick with the pristine (nowadays calculable) form of A—which does not need large n or truth, and which accommodates estimation of θ at the independence level of a hierarchical Bayesian model? If sensitivity of the logarithm to negligible probabilities is objectionable, Bayesians should be happy to substitute a subjectively preferable measure of predictive success.

Christian P. Robert (*Université Paris Dauphine*) **and D. M. Titterington** (*University of Glasgow*)

A question that arises regarding this thought challenging paper was actually raised in the discussion of Aitkin (1991), namely that the data seem to be used *twice* in the construction of p_D. Indeed, y is used the first time to produce the posterior distribution $\pi(\theta|y)$ and the associated estimate $\tilde{\theta}(y)$. The (Bayesian) deviance criterion then computes the posterior expectation of the *observed* likelihood $p(y|\theta)$,

$$\int \log\{p(y|\theta)\}\, \pi(\mathrm{d}\theta|y) \propto \int \log\{p(y|\theta)\}\, p(y|\theta)\, \pi(\mathrm{d}\theta),$$

and thus uses y again, similarly to Aitkin's posterior Bayes factor

$$\int p(y|\theta)\, \pi(\mathrm{d}\theta|y).$$

This repeated use of y would appear to be a potential factor for overfitting.

It thus seems more pertinent (within the Bayesian paradigm) to follow an integrated approach along the lines of the posterior *expected* deviance of Section 6.2,

$$\int E_{Y|\theta}[-2\log\{p(Y|\theta)\} + 2\log\{f(Y)\}]\pi(\mathrm{d}\theta|y)$$

because this quantity would be strongly related to the posterior *expected* loss defined by the logarithmic deviance,

$$d(\theta, \tilde{\theta}) = E_{Y|\theta}[\log\{p(Y|\theta)\} - \log\{p(Y|\tilde{\theta})\}],$$

advocated in Robert (1996) and Dupuis and Robert (2002) as an intrinsic loss adequate for model fitting. In fact, the connection between p_D, the deviance information criterion and the logarithmic deviance would suggest the use of this loss $d(\theta, \tilde{\theta})$ to compute the estimate plugged in p_D as the intrinsic Bayes estimator

$$\begin{aligned}\theta^{\pi}(y) &= \arg\min_{\tilde{\theta}}\{E_{\theta|y}(E_{Y|\theta}[\log\{p(Y|\theta)\} - \log\{p(Y|\tilde{\theta})\}])\}\\ &= \arg\max[E_{Y|y}\{p(Y|\tilde{\theta})\}]\end{aligned}$$

where the last expectation is computed under the predictive distribution. Not only does this make sense because of the aforementioned connection, but it also provides an estimator that is completely invariant to reparameterization and thus avoids the possibly difficult choice of the parameterization of the problem. (See Celeux *et al.* (2000) for an illustration in the set-up of mixtures.)

J. A. Nelder (*Imperial College of Science, Technology and Medicine, London*)
My colleague Professor Lee has made some general points connecting the subject of this paper to our work on likelihood-based hierarchical generalized linear models. I want to make one specific point and two general ones.

(a) Professor Dodge has shown that, of the 21 observations in the stack loss data set, only five have not been declared to be outliers by someone! Yet there is a simple model in which no observation appears as an outlier. It is a generalized linear model with gamma distribution, log-link and linear predictor $x_2 + \log(x_1) * \log(x_3)$. This gives the following entries for Table 2 in the paper

98.3 92.1 6.2 104.5

(I am indebted to Dr Best for calculating these). It is clearly better than the existing models used in Table 2.

(b) This example illustrates my first general point. I believe that the time has passed when it was enough to assume an identity link for models while allowing the distribution only to change. We should take as our base-line set of models at least the generalized linear model class defined by distribution, link and linear predictor, with choice of scales for the covariates in the last named.

(c) My second general point is that there is, for me, not nearly enough model checking in the paper (I am assuming that the use of such techniques is not against the Bayesian rules). For example, if a set of random effects is sufficiently large in number and the model postulates that they are normally distributed, their estimates should be graphed to see whether they look like a sample from such a distribution. If they look, for example, strongly bimodal, then the model must be revised.

Anthony Atkinson (*London School of Economics and Political Science*)
This is an interesting paper which tackles important problems. In my comments I concentrate on regression models: the points extend to the more complicated models at the centre of the authors' presentation.

It is stressed in Section 7.1 that information criteria assume a replication of the observations; in regression this would be with the same X-matrix. But, the simulations of Atkinson (1980) showed that, to predict over a different region, higher values of the penalty coefficient than two in equation (36) are needed. Do the authors know of any analytical results in this area?

Information criteria for model selection are based on aggregate statistics. Fig. 4 shows an alternative and more informative breakdown of one criterion into the contributions of individual observations than that given by Weisberg (1981). However, it does not show the effect of the deletion of observations on model choice. Atkinson and Riani (2000) used the forward search to analyse the stack loss data, for which symmetrical error distributions were considered in Section 8.2. Their Fig. 4.28 shows that the square-root transformation is the only one supported by all the data. The forward plot of residuals, Fig. 3.27, is stable, with observations 4 and 21 outlying. This diagnostic technique complements the choice of a model using information criteria calculated over a set of models that is too narrow.

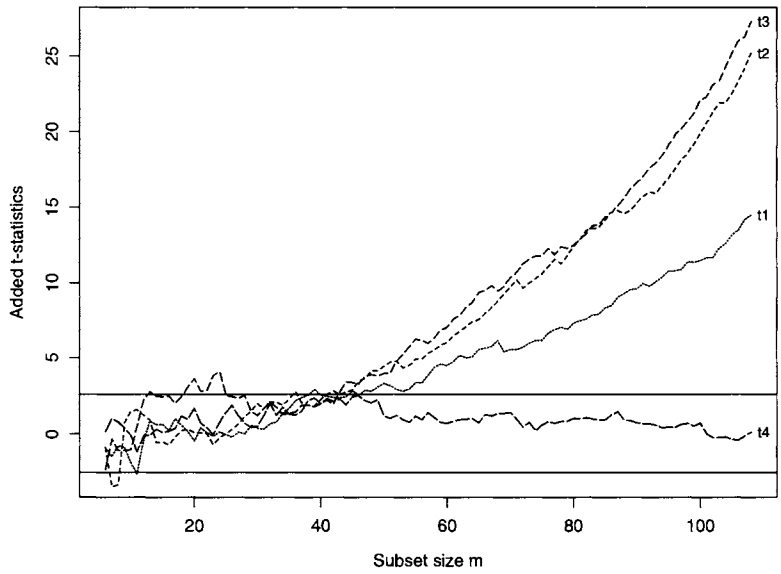

Fig. 6. Transformed surgical unit data: forward plot of the four added variable t-statistics: three variables are needed in the model—x_4 is not significant

An example of model choice potentially confounded by the presence of several outliers is provided by 108 observations on the survival of patients following liver surgery from Neter *et al.* (1996), pages 334 and 438. There are four explanatory variables. Fig. 6 shows the evolution of the added variable t-tests for the variables during the forward search with log(survival time) as the response: the evidence for the importance of all variables except x_4 increases steadily during the search. Atkinson and Riani (2002) modify the data to produce two different effects. The forward plots of the t-tests in Fig. 7(a) show that now x_1

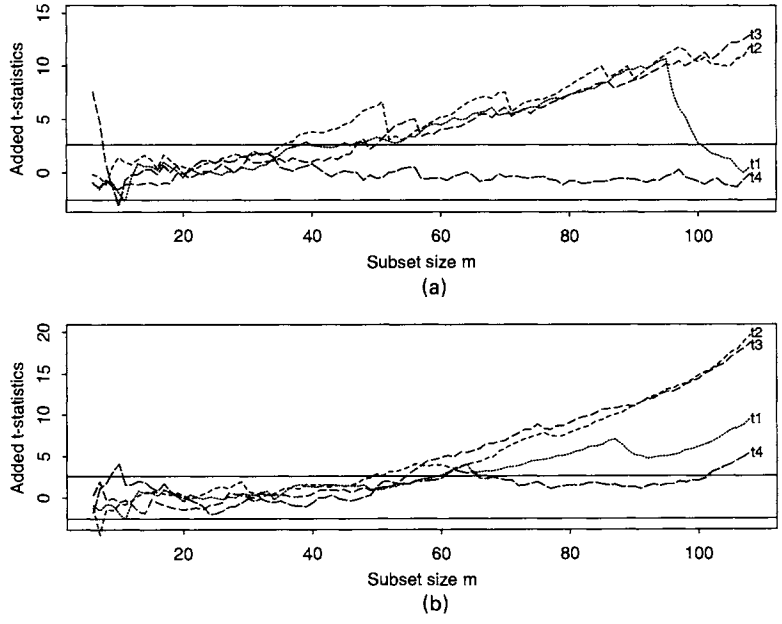

Fig. 7. Modified transformed surgical unit data: (a) outliers render x_1 non-significant; (b) now the outliers make x_4 significant (both (a) and (b) show forward plots of added variable t-statistics)

is non-significant at the end of the search. The plot identifies the group of modified observations which have this effect on the t-test for x_1. Fig. 7(b) shows the effect of a different contamination, which makes x_4 significant at the end of the search.

The use of information criteria in the selection of models is a first step, which needs to be complemented by diagnostic tests and plots. These examples show that the forward search is an extremely powerful tool for this purpose. It also requires many fits of the model to subsets of the data. Can it be combined with the appreciable computations of the authors' Markov chain Monte Carlo methods?

A. P. Dawid (*University College London*)

This paper should have been titled 'Measures of Bayesian model complexity and fit', for it is the models, not the measures, that are Bayesian. Once the ingredients of a problem have been specified, any relevant question has a unique Bayesian answer. Bayesian methodology should focus on specification issues or on ways of calculating or approximating the answer. Nothing else is required.

Classical criteria overfit complex models, necessitating some form of penalization, and this paper lies firmly in that tradition. But with Bayesian techniques (Kass and Raftery, 1995) overfitting is not a problem: the marginal likelihood automatically penalizes model complexity without any need for further adjustment. In particular, Bayesian model choice is consistent in the 'good model' case (Dawid, 1992a). In Section 9.2.5 the authors brush aside the failure of their deviance information criterion procedure to share this consistency property; but should we not seek reassurance that a procedure performs well in those simple cases for which its performance can be readily assessed, before trusting it on more complex problems?

I contest the view (Section 9.1.3) that likelihood is relevant only under the good model assumption: from a decision theoretic perspective, we can always regard the 'log-loss' scoring rule $S(p, y) := -\log\{p(y)\}$ as a measure of the inadequacy of an assessed density $p(\cdot)$ in the light of empirical data y (Dawid, 1986). Moreover, when y is a sequence $y^n = (y_1, \ldots, y_n)$ of not necessarily independent or identically distributed variables, we have

$$-\log\{p(y^n)\} = \sum_{i=1}^{n} -\log\{p(y_i|y^{i-1})\}, \tag{41}$$

the ith term measuring the performance of the Bayesian probability forecast for y_i on the basis of analysis of earlier data only (Cowell *et al.* (1999), chapters 10 and 11). This representation clearly demonstrates why unadjusted marginal likelihood offers a valid measure of model fit: each 'test' observation y_i is always entirely disjoint from the associated 'training' data y^{i-1}. If desired, we can generalize this prequential formulation of marginal likelihood by inserting other loss functions (Dawid, 1992b) or using other model fitting methods (Skouras and Dawid, 1999). Such procedures exhibit a natural consistency property even under model misspecification (Dawid, 1991; Skouras and Dawid, 2000).

One place where a Bayesian might want a measure of model complexity is as a substitute for p in the Bayes information criterion approximation to marginal likelihood, e.g. for hierarchical models. But in such cases the definition of the sample size n can be just as problematic as that of the model dimension p. What we need is a better substitute for the whole term $p \log(n)$.

Andrew Lawson and Allan Clark (*University of Aberdeen*)

We would like to make several comments on this excellent paper.

Our prime concern here is the fact that the deviance information criterion DIC is not designed to provide a sensible measure of model complexity when the parameters in the model take the form of locations in some $\mathcal{R}$-dimensional space. In the spatial context, this could mean the locations of cluster centres or, more generally, the components of a mixture. Clearly the averaging of parameters in these contexts is nonsensical but is a fundamental ingredient of DIC's penalty term $D(\bar{\theta})$. Even if an alternative measure of central tendency is used it remains inappropriate to average over configurations where locations in the chosen space are parameters (e.g. cluster detection modelling in spatial epidemiology (McKeague and Loiseaux, 2002; Gangnon and Clayton, 2002). In the case of the Bayes information criterion, however, it might be possible to replace the penalty $p \ln(n)$ by an average number of parameters (in a reversible jump context) such as $\bar{p} \ln(n)$, where p is the number of parameters and n the sample size. This would at least approximately accommodate the varying dimension but would not require the averaging of parameters (as compared with DIC). This was suggested in Lawson (2000).

The second point of concern is the relationship of the goodness of fit to convergence of the Markov chain Monte Carlo samplers for which DIC is designed. If posterior marginal distributions are multimodal then

the conventional convergence diagnostic will fail (as they will usually find too much variability in individual chains), and also DIC will average over the modes.

We are also somewhat concerned and puzzled by the results for the Scottish lip cancer data set. In Table 1, excepting the saturated model, the largest penalty terms are for the exchangeable model and not those with either spatial or spatial and exchangeable components. We also note that it is not strictly appropriate to fit a spatial-only model without the exchangeable component.

Finally we note that alternative approaches have recently been proposed (Plummer, 2002).

José M. Bernardo (*Universitat de València*)

This interesting paper discusses rather polemic issues and offers some reasonable suggestions. I shall limit my comments to some points which could benefit from further analysis.

(a) The authors point out that their proposal is not invariant under reparameterization and show that differences may be large. The use of the median would make the result invariant in one dimension, but it is not trivial to extend this to many dimensions. An attractive, general invariant estimator is the *intrinsic* estimator obtained by minimizing the *reference* posterior expectation of the intrinsic loss $\delta(\hat{\theta}, \theta)$ (Bernardo and Suarez, 2002) defined as the *minimum* logarithmic divergence between $p(x|\hat{\theta})$ and $p(x|\theta)$. Under regularity conditions and moderate or large samples, this is well approximated by $(E[\theta|\mathbf{x}] + M[\theta|\mathbf{x}])/2$, the average between the reference posterior mean and mode. Other invariant estimators may be obtained by minimizing the posterior expectation of $\delta(\hat{\theta}, \theta)$ obtained from either a proper subjective prior or an improper prior which, as the reference prior, is obtained from an algorithm which is invariant under reparameterization.

(b) The authors use 'essentially flat' or 'weakly informative' priors, i.e. conjugate-like priors with very small parameter values. This is dangerous and is *not* recommended. There is no reason to believe that those priors are weakly informative on the parameters of interest. Indeed, these limiting proper priors can have hidden undesirable features such as strong biases (cf. the Stein paradox). Moreover, they may approximate a prior function which would result in an improper posterior and using a 'vague' proper prior in that case does not solve the problem; the answer will then typically be extremely sensitive to the hyperparameters chosen for the vague proper prior and, since the Markov chain Monte Carlo algorithm will converge because the posteriors are guaranteed to be proper, one might not notice anything wrong. If full, credible, subjective elicitation is not possible then one should use formal methods to derive an appropriate reference prior.

(c) The authors' brief comment (in Section 9.2.4) on the calibration of the deviance information criterion DIC is too short to offer guidance. With Bayes factors, we have a direct interpretation of the numbers obtained. The Bayesian reference criterion (Bernardo, 1999) is defined in terms of natural information units (and may also be described in terms of log-odds). Is there a natural interpretation for DIC?

(d) The important particular case of nested models is not discussed in the paper. Would the authors comment on the behaviour on DIC in that case (and hence on their implication on precise hypothesis testing)? For instance, what is DIC's recommendation for the simple canonical problem of testing a value for a normal mean? It seems to me that, like Akaike's information criterion or the Bayesian reference criterion (but not the Bayes information criterion or Bayes factors), DIC would avoid Lindley's paradox. Is this so?

Sujit K. Sahu (*University of Southampton*)

This impressive paper shows how the very complicated business of model complexity can be assessed easily by using Markov chain Monte Carlo methods. My comments mostly concern the foundational aspects of the methods proposed and the interrelationship of the deviance information criterion DIC and other Bayesian model selection criteria.

The paper provides a long list of models and the associated p_D, the effective number of parameters. In each of these cases p_D is interpreted nicely in terms of model quantities. However, there is an unappealing feature of p_D that I would like to point out in the discussion below.

Consider the set-up leading to equation (23). Assume further that $A_1 = \mathbf{1}$, $C_1 = 1$ and $C_2 = \tau^2$. Thus the likelihood is $N(\theta, 1)$ and the prior is $N(0, \tau^2)$. Then equation (23) yields that

$$p_D = \frac{1}{1 + 1/n\tau^2}.$$

Assuming τ^2 to be finite it is seen that p_D increases to 1 as $n \to \infty$. The unappealing point is that the

effective number of parameters is larger for larger sample sizes; conventional intuition suggests otherwise. The number of unknowns (i.e. the effective number of parameters) should decrease as more data are obtained under this very simple static model. In spite of the authors' views on asymptotics or consistency, this point deserves further explanation as it is valid even when small sample sizes are considered.

In Section 9.1 the relationship between DIC and other well-known Bayesian model selection criteria including the Bayes factor is discussed. Although DIC is not to be viewed as a formal model choice criterion (according to the authors), it is often (and it will be) used to perform model selection; see for example the references cited by the authors. In this regard a more precise statement about the relationship between the Bayes factor and DIC can be made. I illustrate this with the above simple example taken from the paper.

Assume that the observation model is $N(\theta, 1)$ and the prior for θ is $N(0, \tau^2)$. Suppose that model 0 specifies that $H_0 : \theta = 0$ and model 1 says that $H_1 : \theta \neq 0$. I assume that both n and τ^2 are finite and thus avoid the problems with interpretation of the Bayes factor and Lindley's paradox. Using the Bayes factor, model 0 will be selected if

$$n\bar{y}^2 < (1 + n\tau^2)\frac{\log(1 + n\tau^2)}{n\tau^2}.$$

In contrast, DIC selects model 0 if

$$n\bar{y}^2 < (1 + n\tau^2)\frac{2}{2 + n\tau^2}.$$

Clearly, if DIC selects model 0 then the Bayes factor will also select model 0. It is also observed that the Bayes factor allows for higher $|\bar{y}|$-values without rejecting the simpler model. In effect DIC is seen to have the much discussed poor behaviour of a conventional significance test which criticizes the simpler null hypothesis too much and often rejects it when it should not.

Sylvia Richardson (*Imperial College School of Medicine, London*)

I restrict my comments on this far-reaching paper to the use of the deviance information criterion DIC for choosing within a family of models and the behaviour of p_D as a penalization.

My first remark concerns the spatial example of Section 8. The DIC-values for the 'spatial' and the 'spatial plus exchangeable' models are nearly identical. Thus, the authors resort to external pragmatic considerations for preferring the simpler model, while the more complex one is not penalized.

Table 5. Performance of DIC for mixture models with different numbers of components

	Results for the following values of k:				
	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
Bimod ($n = 200$)					
DIC(k)	566.7	567.7	568.5	569.2	570.0
$E(D\|y, k)$	563.4	563.7	564.1	564.5	565.0
p_D	3.3	4	4.4	4.7	5
Skew ($n = 200$)					
DIC(k)	545.5	535.9	535.5	535.7	535.8
$E(D\|y, k)$	540.3	530.1	530.0	530.2	530.4
p_D	5.2	5.8	5.5	5.5	5.4
North–south ($n = 94$)					
DIC(k)	110.5	110.9	110.9	110.5	110.8
$E(D\|y, k)$	94.2	91.9	89.6	87.7	86.2
p_D	16.3	19.0	21.3	22.8	24.6

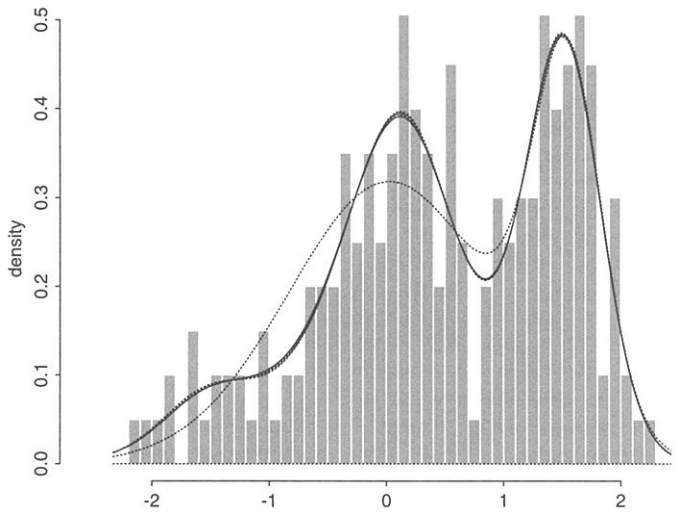

Fig. 8. Predictive densities for the skew data set: ·····, $k = 2$; ———, unconditional (results for $k = 3, 4, 5$ are superimposed)

Turning to mixture models and the comparison between models with different numbers of components, I discuss two situations. The first concerns simple Gaussian mixtures with an unknown number of components; $y_i \sim \Sigma_{j=1}^{k} w_j f(\cdot|\theta_j)$, $i = 1, \ldots, n$, where $f(\cdot|\theta_j)$ is Gaussian. To calculate DIC in this setting, let us focus on mixtures as flexible distributions and use the conditional density for a new observation y^* : $g(y^*) = p(y^*|y, w, \theta, k)$ to calculate the deviance $D(g) = -2\,\Sigma_{i=1}^{n} \log\{g(y_i)\}$ and take its expectation over the Markov chain Monte Carlo run, conditional on k. We have $p_D(k) = E\{D(g)\} - D(\hat{g}_k)$, where $\hat{g}_k = p(y^*|y, k)$.

Two cases of Gaussian mixtures were simulated (one replication): a well-separated bimodal mixture (bimod), $0.5\ N(-1.5, 0.5) + 0.5\ N(1.5, 0.5)$, and an overlapping skewed bimodal mixture (skew): $0.75\ N(0, 1) + 0.25\ N(1.5, 0.33)$, each with 200 data points.

In the clear-cut bimod case, DIC(k) is lower for $k = 2$, with a small incremental increase in both $E(D|y, k)$ and p_D as extra components are being fitted (Table 5). In the more challenging skew case, the pattern of DIC-values shows that this data set requires more than two components to be adequately fitted, but the values of DIC and p_D stay surprisingly flat between three and six components. Note that the predictive density plots conditional on $k = 3, 4, 5$ are completely superimposed (Fig. 8), indicating that more than three components can be considered as overfitting the data, in the sense that they give alternative explanations that are no better but involve increasing numbers of parameters.

The second situation is that of spatial mixture models proposed in Green and Richardson (2002) in the context of disease mapping. DIC was calculated by focusing on area-specific risk. Referring, for example, to the simple north–south (two-component) contrast defined in that paper, we find that DIC stays stable as k increases, decreasing $E(D|y, k)$ values being compensated by increasing p_D. On the basis of a mean-square error criterion between the estimated and the underlying risk surface, a deterioration of the fit would be seen with values of 0.14, 0.15 and 0.16 for $k = 2, 3, 4$ respectively.

Thus p_D acts as a sufficient penalization only in the simplest case. In other cases, DIC does not distinguish between alternative fits with increasing number of parameters.

Peter Green (*University of Bristol*)

I have two rather simple comments on this interesting, important and long-awaited paper.

The first concerns using basic distribution theory to give a surprising new perspective on p_D in the normal case, perhaps identifying a missed opportunity in exposition.

Consider first a decomposition of data as focus plus noise:

$$Y = X + Z$$

where X and Z are independent n-vectors, normally distributed with fixed means and variances, and var(Z)

is non-singular. The deviance is

$$D(X) = (Y - X)^{\mathrm{T}} \operatorname{var}(Z)^{-1}(Y - X)$$

and so

$$p_D = E[D(X)|Y] - D(E[X|Y]) = \operatorname{tr}\{\operatorname{var}(Z)^{-1}\operatorname{var}(Z|Y)\}, \tag{42}$$

using the standard expression for the expectation of a quadratic form. Several results in the paper have this form, possibly in disguise. However,

$$\begin{aligned}\operatorname{var}(Z|Y) &= \operatorname{var}(Z) - \operatorname{cov}(Z, Y)\operatorname{var}(Y)^{-1}\operatorname{cov}(Y, Z)\\ &= \operatorname{var}(Z) - \operatorname{var}(Z)\operatorname{var}(Y)^{-1}\operatorname{var}(Z)\\ &= \operatorname{var}(Z)\operatorname{var}(Y)^{-1}\{\operatorname{var}(Y) - \operatorname{var}(Z)\},\end{aligned}$$

yielding the much more easily interpretable

$$p_D = \operatorname{tr}\{\operatorname{var}(Y)^{-1}\operatorname{var}(X)\}. \tag{43}$$

This allows a very clean derivation of examples in Sections 2.5 and 4.1–4.3. For example, in the Lindley and Smith model we have $\operatorname{var}(Z) = C_1$ and $\operatorname{var}(X) = A_1 C_2 A_1^{\mathrm{T}}$, and so

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

as in equation (21) of the paper.

Turning now to hierarchical models, consider a decomposition into k independent terms

$$Y = Z_1 + Z_2 + \ldots + Z_k,$$

where all Z_i are normal, and $\operatorname{var}(Z_k)$ is non-singular. These represent all the various terms of the model: fixed effects with priors, random effects with different structures, errors at various levels; again all means and variances are fixed. Then for any level $l = 1, 2, \ldots, k - 1$ we may take the sum of the first l terms as the focus and the rest as noise.

Version (42) of p_D above is then not very promising:

$$p_D(l) = \operatorname{tr}\left\{\operatorname{var}\left(\sum_{i=l+1}^{k} Z_i\right)^{-1} \operatorname{var}\left(\sum_{i=l+1}^{k} Z_i \middle| Y\right)\right\},$$

but expression (43) gives the more compelling

$$p_D(l) = \operatorname{tr}\left\{\operatorname{var}(Y)^{-1} \operatorname{var}\left(\sum_{i=1}^{l} Z_i\right)\right\}. \tag{44}$$

Thus p_D has generated a decomposition of the overall degrees of freedom $n = \Sigma_l \operatorname{tr}\{\operatorname{var}(Y)^{-1}\operatorname{var}(Z_l)\}$ into non-negative terms attributable to the levels $l = 1, 2, \ldots, k$, just as in frequentist nested model analysis of variance. (We must take care with improper priors in using expression (44), and terms should be treated as limits as precisions go to 0.) Of course, expressions (43) and (44) fail to hold with unknown variances or with non-normal models, but the observations above do provide further motivation for accepting p_D as a measure of complexity, and suggest exploring more thoroughly its role in hierarchical models.

My second point notes that the paper has no examples with discrete 'parameters'. Conditional distributions in hierarchical models with purely categorical variables can be computed by using probability propagation methods (Lauritzen and Spiegelhalter, 1988), avoiding Markov chain Monte Carlo methods, so that p_D is again a cheap local computation. Presumably marginal posterior modes would be used for $\tilde{\theta}$. Certainly this is a context where p_D can be negative. Can connections be drawn with existing model criticism criteria in probabilistic expert systems?

The following contributions were received in writing after the meeting.

Kenneth P. Burnham (*US Geological Survey and Colorado State University, Fort Collins*)

This paper is an impressive contribution to the literature and I congratulate the authors on their achievements therein. My comments focus on the model selection aspect of the deviance information criterion DIC. My perspectives on model selection are given in Burnham and Anderson (2002), which has a focus on the Akaike information criterion AIC as derived from Kullback–Leibler information theory. A lesson that we learned was that, if the sample size n is small or the number of estimated parameters p is large relative to n, a modified AIC should be used, such as $\mathrm{AIC}_c = \mathrm{AIC} + 2p(p+1)/(n-p-1)$. I wonder whether DIC needs such a modification or if it really automatically adjusts for a small sample size or large p, relative to n. This would be a useful issue for the authors to explore in detail.

At a deeper level I maintain that model selection should be multimodel inference rather than just inference based on a single best model. Thus, model selection to me has become the computation of a set of model weights (probabilities in a Bayesian approach), based on the data and the set of models, that sum to 1. Given these weights and the fitted models (or posterior distributions), model selection uncertainty can be assessed and model-averaged inferences made. The authors clearly have this issue in mind as demonstrated by the last sentence of Section 9.1.3. I urge them to pursue this much more general implementation of model selection and to seek a theoretical or empirical basis for it with DIC.

There is a matter that I am confused about. The authors say '... we essentially reduce all models to non-hierarchical structures' (third page), and 'Strictly speaking, nuisance parameters should first be integrated out ...' (Section 9.2.3). Does this mean that we cannot make full inferences about models with random effects? Can DIC be applied to random-effects models? It seems so on the basis of their lip cancer example (Section 8.1). Can I have a model with fixed effects τ, random effects $\phi_1, \ldots, \phi_k$, with postulated distribution $g(\phi|\theta)$, θ as fixed effects (plus priors on all fixed effects) and have my focus be all of τ, ϕ and θ? Thus, I obtain shrinkage-type inferences about the ϕ_i; I do not integrate out the ϕ (AIC has been adapted to this usage).

The authors make a point (page 612) that I wish to make more strongly. It will usually not be appropriate to 'choose' a single model. Unfortunately, standard statistical model selection has been to select a single model and to ignore any selection uncertainty in the subsequent inferences.

Maria DeIorio (*University of Oxford*) **and Christian P. Robert** (*Université Paris Dauphine*)

Amidst the wide scope of possible extensions of their paper, the authors mention the case of mixtures

$$\sum_{j=1}^{k} p_j \, f(x|\theta_j),$$

which is quite interesting, as it illustrates the versatility of the deviance information criterion DIC under different representations of the same model.

In this set-up, if the p_js are known, the associated *completed* likelihood is

$$L\{\boldsymbol{\theta}|(x_1, z_1), \ldots, (x_n, z_n)\} \propto \prod_{i=1}^{n} f(x_i|\theta_{z_i}) = \prod_{j=1}^{k} \prod_{i:z_i=j} f(x_i|\theta_j). \tag{45}$$

Therefore, conditional on the latent variables $\mathbf{z} = (z_1, \ldots, z_n)$, and setting the saturated deviance $f(x)$ to 1, define

$$[\mathrm{DIC}|\mathbf{z}] = \sum_{j=1}^{k} \sum_{i:z_i=j} (-4E[\log\{f(x_i|\theta_j)\}|\mathbf{x}, \mathbf{z}\} + 2 \log\{f(x_i|\hat{\theta}_j)\}])$$

where $\hat{\theta}_j = E(\theta_j|\mathbf{x}, \mathbf{z})$ (under proper identifiability constraints; see Celeux *et al.* (2000)). The *integrated* DIC is then

$$\mathrm{DIC}_1 = \sum_{\mathbf{z} \in Z} [\mathrm{DIC}|\mathbf{z}] \Pr(\mathbf{z}|\mathbf{x}),$$

where $\Pr(\mathbf{z}|\mathbf{x})$ can be approximated (Casella *et al.*, 1999).

A second possibility is the *observed* DIC, DIC_2, based on the observed likelihood, which does not use the latent variables $\mathbf{z}$. (We note the strong dependence of DIC on the choice of the saturated function f and the corresponding lack of clear guidance outside exponential families. For instance, if $f(x_i)$ goes from the marginal density to the extreme alternative where both θ_1 and θ_2 are set equal to x_i, DIC_2 goes from -31.71 to 166.6 in the following example.)

Table 6. Comparison of the three different criteria DIC_1, DIC_2 and DIC_3 for a simulated sample of 100 observations from $0.5\mathcal{N}(5, 1.5) + 0.5\mathcal{N}(7.5, 8)$ with a conjugate prior $\theta_1 \sim \mathcal{N}(4, 5)$ and $\theta_2 \sim \mathcal{N}(8, 5)$, and of DIC based on the true complete sample (**x**, **z**) and DIC for the single-component normal model (with an $\mathcal{N}(6, 5)$ prior and a variance set of 6.07)

	Results for the following models:				
	Normal ($k = 1$)	*Complete,* [$DIC \mid \mathbf{z}$]	*Integrated,* DIC_1	*Observed,* DIC_2	*Full,* DIC_3
DIC	465.1	413.5	462.6	457.6	447.4
ΔDIC	—	−51.6	−2.5	−7.5	−17.6
p_D	0.99	1.96	2.27	1.98	28.06

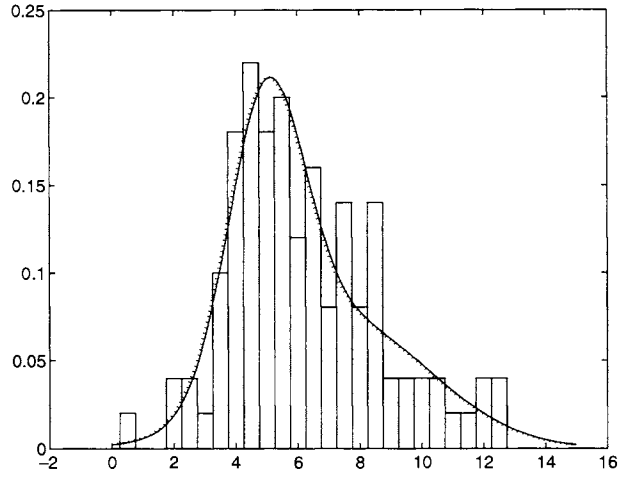

Fig. 9. Histogram of the simulated data set and true density

A third possibility is the *full* DIC, DIC_3, based on the completed likelihood (45) when it incorporates **z** as an additional parameter, in which case the saturated deviance could be the normal standardized deviance, although we still use $f(x) = 1$ for comparison.

The three possibilities above lead to rather different figures, as shown by Table 6 for the simulated data set in Fig. 9; Table 6 exhibits in addition a lack of clear domination of the mixture ($k = 2$) *versus* the normal distribution ($k = 1$) (second column), except when **z** is set to its true value (third column) or estimated (last column). Note that, for the full DIC, p_D is far from 102; this may be because, for some combinations of **z**, the likelihood is the same. (This also relates to the fact that **z** is not a parameter in the classical sense.)

David Draper (*University of California, Santa Cruz*)

The authors of this interesting paper talk about Bayesian model assessment, comparison and fit, but—if their work is to be put seriously to practical use—the real point of the paper is Bayesian model choice: we are encouraged to pick the model with the smallest deviance information criterion DIC among the class of 'good' models (those which are 'adequate candidates for explaining the observations'). (It is implicit that somehow this class has been previously specified by means that are not addressed here—would the

authors comment on how this set of models is to be identified in general?) However, in the case of model selection it would seem self-evident that *to choose a model you have to say to what purpose the model will be put*, for how else will you know whether your model is sufficiently good? We can, perhaps, use DIC to say that model 2 is better than model 1, and we can, perhaps, compare $\bar{D}$ with 'the number of free parameters in θ' to 'check the overall goodness of fit' of model 2, but we cannot use the authors' methods to say whether model 2 is sufficiently good, because the real world definition of this concept has not been incorporated into their methods. It seems hard to escape the fact that specifying the purpose to which a model will be put demands a decision theoretic basis for model choice; thus (Draper, 1999) I am firmly in the camp of Key *et al.* (1999).

See Draper and Fouskakis (2000) and Fouskakis and Draper (2002) for an example from health policy that puts this approach into practice, as follows. Most attempts at variable selection in generalized linear models conduct what might be termed a benefit-only analysis, in which a subset of the available predictors is chosen solely on the basis of predictive accuracy. However, if the purpose of the modelling is to create a scale that will be used—in an environment of constrained costs, which is frequently the case—to make predictions of outcome values for future observations, then the model selection process must seek a subset of predictors which trades off predictive accuracy against data collection cost. We use stochastic optimization methods to maximize the expected utility in a decision theoretic framework in the space of all 2^p possible subsets (for p of the order of 100), and because our predictors vary widely in how much they cost to collect (which will also often be true in practice) we obtain subsets which are sharply different from (and much better than) those identified by benefit-only methods for performing 'optimal' variable selection in regression, including DIC.

Alan E. Gelfand (*Duke University, Durham*) **and Matilde Trevisani** (*University of Trieste*)

The authors' generally informal approach motivates several remarks which we can only briefly develop here. First, in Section 2.1, we think that better terminology would be 'focused on $p(y|\theta)$' with 'interest in the models for θ', as in, for example, the example in Section 8.1 where there is no θ in the likelihood for any of the given models. Even the example in Section 8.2, where θ does not change across models, emphasizes the focus on $p(y|\theta)$ since $f(y)$ depends on the choice of p. So, here, a relative comparison of the models depends on the choices made for the fs. Without a clear prescription for f (once we leave the exponential family), the opportunity exists to fiddle the support for a model.

Though the functional form of the Bayesian deviance does not depend on $p(\theta)$, DIC and p_D will. With the authors' hierarchical specification,

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

the effective degrees of freedom will depend on $p(\psi)$. But, also, under this specification, rather than $p(y|\theta)$, we can put a different distribution, $p(y|\psi)$, in focus. Again, it seems preferable not to speak in terms of 'parameters in focus'.

Moreover, since $p(y|\theta)$ and $p(y|\psi)$ have the same marginal distribution $p(y)$, a coherent model choice criterion must provide the same value under either focus. Otherwise, a particular hierarchical specification could be given more or less support according to which distribution we focus on. But let DIC_1, p_{D_1} and $f_1(y)$ be associated with $p(y|\theta)$ and DIC_2, p_{D_2} and $f_2(y)$ with $p(y|\psi)$. To have $\mathrm{DIC}_1 = \mathrm{DIC}_2$ requires, after some algebra, that

$$\ln\{f_2(y)\} - \ln\{f_1(y)\} = p_{D_1} - p_{D_2} + E[\ln\{p(y|\psi)|y\}] - E[\ln\{p(y|\theta)|y\}].$$

Just as the functional form of $f_1(y)$ depends only on the form of $p(y|\theta)$, the form for $f_2(y)$ should depend only on $p(y|\psi)$. Evidently this is not so. For instance, under the authors' example in expression (2), $f_1(y) = 0$. The above expression yields the non-intuitive choice

$$\ln\{f_2(y)\} = \sum w_i + \frac{1}{2}\sum \ln(1 - w_i) - \lambda\,\mathrm{var}(\psi|y)\ \sum w_i^2 - \frac{\lambda}{2}\ \sum w_i^2\{y_i - E(\psi|y)\}^2$$

where $w_i = \tau_i/(\tau_i + \lambda)$. This issue is discussed further in Gelfand and Trevisani (2002).

Jim Hodges (*University of Minnesota, Minneapolis*)

This is a most interesting paper, presenting a method of tremendous generality and, as a bonus, a fine survey of related methods. I can think of a dozen models for which I would like to see p_D, but I shall ask for just one: a balanced one-way random-effects model with unknown between-group precision, in which each group has its own unknown error precision, these latter precisions being modelled as draws from, say, a common gamma distribution with unknown parameters. Thus the precisions will be shrunk as well

as the means, and presumably the two kinds of shrinkage will affect each other. The focus could be either the means or the precisions, or preferably both at once.

One thing is troubling: the possibility of a negative measure of complexity (Section 2.6, comment (d)). Hodges and Sargent (2001) is linked (shackled?) to linear model theory, in which complexity is defined as the dimension of the subspace of $\Re^n$ in which the fitted values lie. In our generalization, the fitted values may be restricted to 'using' only part of a basis vector's dimension, because they are stochastically constrained by higher levels of the model's hierarchy. (Basing complexity on fitted values may remove the need to specify a focus, although, if true, this is not obvious.) In this context, zero complexity makes sense: the fitted values lie in a space of dimension 0 specified entirely by a degenerate prior. Negative complexity, however, is uninterpretable in these terms. The authors attribute negative complexity to a poor model fit, which suggests that p_D describes something more than the fitted values' complexity *per se*. Perhaps the authors could comment further on this.

Youngjo Lee (*Seoul National University*)

It is very interesting to see the Bayesian view of Section 4.2 of Lee and Nelder (1996), which used extended or h-likelihood and in which we introduced various test statistics. For a lack of fit of the model we proposed using the scaled deviance

$$D_r = -2(\log\{p(y|\tilde{\theta}^1)\} - \log[p\{y|\mu(\theta) = y\}])$$

with degrees of freedom $E(D_r)$, estimated by $n - \mathrm{tr}(-L''_{\tilde{\theta}}V)$ where $-L''_{\tilde{\theta}} = V^*$ as in Sections 4.3 and 5.4 of this paper. We considered a wider class of models, which we called hierarchical generalized linear models (HGLMs) (see also Lee and Nelder (2001a, b)), but some of our proofs hold more widely than this, so that, for example, Section 3.1 of this paper is summarized in our Appendix D, etc. For model complexity the authors define in equation (9) the scaled deviance

$$D_m = -2[\log\{p(y|\theta)\} - \log\{p(y|\tilde{\theta}^1)\}].$$

D_r and D_m are the scaled deviances for the residual and model respectively, whose degrees of freedom add up to the sample size n. We are very glad that the authors have pointed out the importance of the parameterization of θ in forming deviances. We extended the canonical parameters of Section 5 to arbitrary links by defining the h-likelihood on a particular scale of the random parameters, namely one in which they occur linearly in the linear predictor. In HGLMs the degrees of freedom for fixed effects are integers whereas those for random effects are fractions. Thus, a GLM has integer degrees of freedom $p_m = \mathrm{rank}(X)$ because $C_2^{-1}\delta$ is 0 in Section 5, whereas the estimated degrees of freedom of D_m in HGLMs are fractions. Lee and Nelder (1996) introduced the adjusted profile h-likelihood eliminating θ, and this can be used to test various structures of the dispersion parameters λ discussed in the examples of Section 8: see the model checking plots for the lip cancer data in Lee and Nelder (2001b). Lee and Nelder (2001a) justified the simultaneous elimination of fixed and random nuisance parameters. It will be interesting to have the Bayesian view of the adjusted profile h-likelihood.

Xavier de Luna (*Umeå University*)

This interesting paper presents Bayesian measures of model complexity and fit which are useful at different stages of a data analysis. My comments will focus on their use for model selection. In this respect, one of the noticeable contributions of the paper is to propose a Bayesian analogue, the deviance information criterion DIC, to the Akaike information criterion AIC and TIC. Both DIC and TIC are generalizations of AIC. The former may be useful in a Bayesian data analysis, whereas the frequentist criterion TIC has the advantage of not requiring the 'good model' assumption discussed by the authors.

Such 'information-based' criteria use measures of model complexity (denoted p^* or p_D in the paper). It should, however, be emphasized that models can be compared without having to define and compute their complexity. Instead, out-of-sample validation methods, such as cross-validation (Stone, 1974) or prequential tests (Dawid, 1984) can be used in wide generality. Moreover, to use an estimate of p^* in a model selection criterion, some characteristics of the data-generating mechanism (DGM)—'true model' in the paper—must be known. For instance, depending on the DGM either AIC-type or Bayes information type criteria are asymptotically optimal (see Shao (1997) for a formal treatment of linear models). Thus, when little is known about the DGM, out-of-sample validation provides a *formal* and general framework to perform model selection as was presented in de Luna and Skouras (2003), in which accumulated prediction errors (defined with a loss function chosen in accordance with the purpose of the data analysis)

were advocated to compare and choose between different model selection strategies. When many models are under scrutiny, out-of-sample validation may be computationally prohibitive and generally yields high variability in the selection of a model. In such cases, different model selection strategies based on p^* (making—implicitly or explicitly—diverse DGM assumptions) can be applied to reduce the dimension of the selection problem. Accumulated prediction errors can then be used to identify the best strategy while making very few assumptions on the DGM.

Xiao-Li Meng (*Harvard University, Cambridge, and University of Chicago*)
The summary made me smile, for the 'mean of the deviance – deviance of the mean' theme once injected a small dose of excitement into my student life. I was rather intrigued by the 'cuteness' of expressions (3.4) and (3.8) of Meng and Rubin (1992), and seeing a Bayesian analogue of our likelihood ratio version certainly brought back fond memories. My excitement back then was short lived as I quickly realized that all I was deriving was just a masked version of a well-known variance formula. Let $D(x, \mu) = (x - \mu)^2$ be the deviance, a case of *realized discrepancy* of Gelman *et al.* (1996); then

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \overline{D(x_i, \mu)} - D(\bar{x}, \mu). \tag{46}$$

Although equation (46) is typically mentioned (with μ set to 0) for computational convenience, it is the back-bone of the theme under quadratic or normal approximations, or more generally with log-concave likelihoods, beyond which assumptions become much harder to justify or derive. (Obviously, equation (46) is applicable for posterior or likelihood averaging by switching x and μ.)

Section 1 contained a small puzzle. I wondered why Ye (1998) was omitted from the list of 'the most ambitious attempts', because Ye's 'data derivative' perspective goes far beyond the independent normal model cited in Section 4.2 (for example, it addresses data mining). It also provides a more original and insightful justification than normal approximations, especially considering that Markov chain Monte Carlo sampling is most needed in cases where such approximations are deemed unacceptable.

Section 2.1 presented a bigger puzzle. The authors undoubtedly would agree that a statement like 'In hierarchical modelling we cannot uniquely define a "posterior" or "model complexity" without specifying the level of the hierarchy that is the focus of the modelling exercise' is tautological. Surely the 'posterior' and thus the corresponding 'model complexity' depend on the level or parameter(s) of interest. So why does the statement become a meaningful motivation when the word posterior is replaced by 'likelihood'? There is even some irony here, because hierarchical models are models where there are unambiguous and uncontroversial *marginal* likelihoods—both $L(\theta|y) = p(y|\theta)$ and $L(\phi|y) = p(y|\phi)$ in Section 2.1 are *likelihoods* in the original sense.

Although limitations on space prevent me from describing my reactions when reading the rest, I do wish that DIC would stick out in the dazzling AIC—TIC alphabet contest, so we would all be less compelled to look for UIC (*unified or useful information criterion*?) ....

The **authors** replied later, in writing, as follows.

We thank all the contributors for their wide-ranging and provocative discussion. Our reply is organized according to a number of recurring themes, but constraints on space mean that it is impossible to address all the points raised. Echoing Brooks's opening remarks, our hope is that discussants and readers will be sufficiently inspired to pursue the ideas proposed in this paper and to address some of the unresolved issues highlighted in the discussion.

Model focus and definition of deviance
Our notion of the 'focus' of a model and its relationship to the prediction problem of interest provoked some controversy. The crucial role of the model focus is to define the (parameterization of the) likelihood, and we appreciate Gelfand and Trevisani's suggestion of the term 'focus on $p(y|\theta)$', with interest in the structure of θ, rather than models 'focused on θ'. In all our examples the likelihood has been taken to be $p(y|\theta)$ (using the notation of Section 2.1) leading to models with a closed form likelihood but an unknown number of effective parameters that we propose to estimate by p_D. However, as Brooks points out, if the focus is on $p(y|\psi)$ (i.e. integrating over the random effects θ), then in general the likelihood will no longer be available in closed form, and other methods must be sought to evaluate $p(y|\psi)$: in this circumstance the number of parameters will be the dimension of ψ or less, depending on the strength of the prior information on ψ.

Smith and others ask how the model focus should be chosen in practice. We argue that the focus is operationalized by the prediction problem of interest. For example, if the random effects θ in a hierarchical model relate to observation units such as schools or hospitals or geographical areas, where we might reasonably want to make future predictions for those same units, then taking $p(y|\theta)$ as the focus is sensible. The prediction problem is then to predict a new $Y_{i,\mathrm{rep}}$ conditional on the posterior estimate of θ_i for that unit. However, if the random effects relate to individual people, say, then we are often interested in population-average inference rather than subject-specific inference, so we may want to predict responses for a new or 'typical' individual rather than an individual who is already in the data set. In this case, it is appropriate to integrate over the θs and to predict Y_{rep} for a new individual conditional on ψ, leading to a model focused on $p(y|\psi)$. A crucial insight is that a predictive probability statement such as $p(Y_{\mathrm{rep}}|y)$ is not uniquely defined without specifying the level of the hierarchy that is kept fixed in the prediction—this defines the focus of the model. In summary, we feel that the issue of focus with respect to predictive model assessment and selection is an issue in hierarchical modelling and not specifically Bayesian.

When the forms of the likelihoods differ between models being compared, it is clearly vital to be careful that any standardizing terms that are used in the deviance are common. As observed by Smith, a comparison of models with focus at different levels of the hierarchy may not be meaningful as they correspond to different prediction problems.

Features of p_D

Several discussants questioned the definition or performance of p_D. As to the definition we maintain our claim (in spite of Dawid's comment) that it is in our models that there is a genuine Bayesian interest in quantifying the interaction between Y and Θ in probabilistic terms. One can indeed often think of p_D in terms of dimensionality as Hodges suggests, but in general we prefer to think of it as a feature of the *joint* distribution of Y and Θ. This frees it from the shackles imposed by normal linear model theory. Such a measure of interaction or model complexity may, for example, be used to reparameterize hyperparameters ψ to facilitate an intuitively interpretable specification of model priors on ψ (Holmes and Denison, 1999). Still, as suggested by Brooks, p_D may turn out to be only a step towards a (better) definition of model complexity such as that suggested by Plummer: we feel that the quantity that he proposes is intuitively intriguing and that it may be particularly appropriate in exponential families, but we wonder about its general validation and justification.

Our uncertainty about whether to recommend p_D as a definition or as an estimate of a quantity still to be defined makes it difficult to judge proposals for an 'improvement'. For example, using an invariant estimator such as that proposed by Robert and Titterington or Bernardo instead of $\bar{\theta}$ is tempting as part of a definition, but it takes into account only one feature of p_D while destroying others such as the trace approximation. Similarly the occurrence of a negative value of p_D, typically observed if the model fits poorly, might resemble a negative estimate for a positive parameter. We take a pragmatic point of view and look forward to theoretical progress that provides insight into why p_D generally appears to work well. Green provides a valuable insight into the interpretation of p_D in the normal case, using an attractive decomposition of the total predictive variance of the observables.

Replying to those discussants who were concerned about observing $p_D < n$ under 'flat' priors, we re-emphasize that $p_D = n$ was obtained theoretically only in the normal case or under normal approximations. There is no proof that $p_D = n$ for general distributions. In the case of Brooks's illustration using the Scottish lip cancer data, in which he shows that p_D appears to 'lose' two or three (modulo Monte Carlo error) parameters under such priors, we point out that two of the 56 observations in this data set are 0 with small expected values and so contribute negligibly to the Poisson deviance. We have replicated his analysis replacing these two observations by non-zero counts, and we found that p_D increases by about 2 to around 55.5.

We certainly do not recommend the unthinking use of default priors, a concern of Smith and Bernardo: on the contrary, one of our main aims is to demonstrate how an informative prior reduces model complexity. Typically a large number of parameters p relative to a small sample size n is compensated by using an informative prior, and the deviance information criterion DIC and p_D adjust accordingly without any need for additional adjustment for small sample size (see Burnham, and Lawson and Clark's comment on the example in Section 8.1).

There is evidence (Daniels and Kass, 1999, 2001) that, in the absence of missing data, the use of default priors for variance components typically has little effect on the posteriors for the main effects in a model. Still, Smith and Bernardo observe that the flat priors that may maximize p_D are not necessarily weakly informative, and we agree. Reference priors that are least informative in an information theoretical sense can

be easily studied in some of our examples. For example, Fig. 1 displays the performance of the beta($\frac{1}{2}$, $\frac{1}{2}$) reference prior (corresponding to a prior sample size of $n_i = a + b = 1$) for the binomial likelihood, and the approximation (31) indicates that $p_{D_i}^{\Theta}$ based on the reference prior is greater than $p_{D_i}^{\Theta}$ based on the uniform beta(1, 1) prior (which has prior sample size $n_i = 2$). Similarly for a Poisson likelihood the reference prior $\pi(\mu_i) \propto \sqrt{\mu_i}$ yields a $\Gamma(y_i + \frac{1}{2}, n_i)$ posterior distribution corresponding to $a = \frac{1}{2}, b \to 0$. Hence $p_{D_i}^{\mu} \approx y_i/(y_i + \frac{1}{2})$ and $p_{D_i}^{\Theta} \approx n_i/n_i = 1$ might be compared with the values shown in Fig. 2.

Properties of DIC

Another main part of the discussion focused on the properties and performance of DIC. Plummer doubted the usefulness of the expected loss that DIC approximates, but he has included a standardizing constant in the loss function which should not be present (we have made this clearer in the paper). The expected loss in the (independent) normal linear case is then $p + p_D + n \log(2\pi\sigma^2)$: this says that when comparing 'good' models with the same σ^2s the expected loss is minimized with a degenerate prior in which no parameters are estimated. This seems entirely reasonable, as all the models have equivalent fit, and so distinction is based on complexity alone. Of course in practice either σ^2 will be estimated or σ^2 will vary between models, and hence the appropriate trade-off between fit and complexity will naturally arise. A practical aspect, related to the need for 'good' models in the derivation of DIC, is that the term $\mathcal{L}_2$ ignored by DIC will tend to be negative with poorly fitting models and hence to inflate DIC: the approximation of DIC to expected loss will thus tend automatically to penalize models that are not 'good'.

Though we agree with Brooks that owing to its heuristic derivation DIC may be considered as a 'broad brush technique', we do not regard it to be as arbitrary as the alternatives that he suggests. In particular we do not feel that terms of 'fit' and 'complexity' can be arbitrarily combined, but we re-emphasize that a measure of model complexity results from correcting overfit due to an approximation of the expected loss that 'uses the observations twice'. Similarly we would like to see a justification of Vehtari's estimates of expected utilities as valid approximations generalizing DIC.

Bernardo asks for the application of DIC to nested models and hypothesis testing, in particular the occurrence of Lindley's paradox. This is an interesting question partially answered by the example discussed in Section 8.1 where some of the competing models are nested. The key point is that DIC is designed to take into account priors that are concentrated on parameters which are specified in a model, thus effectively assigning prior probability 0 to hypothetically omitted parameters (if there are remaining parameters). Let us consider Lindley's paradox in the following version: when comparing using the Bayes factor $\bar{X} \sim N(\mu_0, \sigma^2/n)$ with $\bar{X} \sim N(\mu, \sigma^2/n)$ where $\mu \sim N(\mu_1, \tau^2)$, evidence in favour of $H_0 : \mu = \mu_0$ becomes overwhelming as $\tau^2 \to \infty$ even if $\bar{x}$ would cause the rejection of H_0 at any arbitrary significance level. If σ^2 is known μ is the only parameter in the model. To apply DIC we compare the model $\bar{X} \sim N(\mu, \sigma^2/n)$ with prior $\mu \sim N(\mu_0, \tau^2)$, $\tau^2 \to 0$, corresponding to H_0 with the model with the same likelihood but prior $\mu \sim N(\mu_1, \tau^2)$, $\tau^2 \to \infty$. Then $D(\mu) = n(\bar{x} - \mu)^2/\sigma^2$, $\overline{D(\mu)} = (n/\sigma^2)\{D(\bar{\mu}) + \mathrm{var}(\mu|\bar{x})\}$ and $p_D = n/\sigma^2 \,\mathrm{var}(\mu|\bar{x})$. For $\tau^2 \to 0$, $p_D \to 0$, $\bar{\mu} \to \mu_0$ and DIC $\to D(\mu_0)$. Similarly, for $\tau^2 \to \infty$, $p_D \to 1$, $\bar{\mu} \to \bar{x}$ and DIC $\to D(\bar{x}) + 2 = 2$. Hence the model with the flat prior—the 'alternative hypothesis'—is favoured if $D(\mu_0) > 2$ or $|\sqrt{n}(\bar{x} - \mu_0)/\sigma| > 1.414$ which corresponds to a rejection of H_0 at a significance level $\alpha \approx 0.16$—exactly the behaviour of the Akaike information criterion. Thus Lindley's paradox is not observed. Similarly Sahu contrasts the prior concentrated on $\mu_0 = 0$ with an informative prior $N(0, \tau^2)$ which is centered at μ_0, also. Thus it is reasonable to reject H_0 using DIC if the data are suitably compatible with the 'alternative' prior. However, we do not accept an assessment of DIC that uses Bayes factors as a 'gold standard', since they are dealing with different prediction problems (see below).

Several discussants (Brooks, Bernardo, Burnham and Smith) were concerned with the lack of calibration of DIC. However, unlike the Bayesian reference criterion (Bernardo, 1999), which is based on a Kullback–Leibler distance and therefore a relative measure, DIC is an approximation to an absolute expected loss, and we cannot calibrate it (externally). Correspondingly, 'coherence' of model choice cannot be required in terms of equal DIC-values as Gelfand and Trevisani or Smith claim but can only be discussed in terms of model ranking by DIC. Note, by the way, that Plummer's alternative measure of model complexity, as well as our p_D, are defined relatively, indicating that these measures might be calibrated.

Finally, we certainly do not claim that applying DIC is an exhaustive tool for model assessment. Although we feel that our Fig. 4 is a step in the right direction, additional techniques such as those discussed by Nelder and Atkinson are certainly needed for refined analyses.

Applications

There were various comments on the interpretation of p_D in the Scottish lip cancer analysis (Lawson and

Clark, and Richardson) and in mixture models (Richardson, and DeIorio and Robert). Here we tend to think of p_D as the estimable dimension of the parameter space or, alternatively, as the size of the parameter space that is identifiable by the data. We repeat that the spatial model 3 in the lip cancer example (Section 8.1) provides stronger prior information than the exchangeable model 2 leading to a smaller p_D. Only the sum of the spatial and exchangeable random effects is uniquely identifiable in model 4 and so p_D remains virtually unchanged compared with the spatial-only model 3, thus justifying the lack of an additional 'penalty' for the apparently more complex model. The same is true for mixture models, where increasing the number of components does not necessarily increase the identifiable parameter space. We do appreciate the discussion of DIC in mixture models introduced by DeIorio and Robert, and by Richardson (though Richardson does not appear to have calculated DIC as we have defined it, but a different criterion based on predictive deviances). DeIorio and Robert's example nicely illustrates a range of possibilities for defining DIC in this case, although we re-emphasize that a comparison of models with different focus (e.g. their DIC_2 *versus* DIC_3) may not be meaningful, and we further note that their integrated DIC (DIC_1) does not correspond to our definition of DIC.

In response to Lawson and Clark's query about averaging 'location' parameters, we point to Green's comment concerning the calculation of p_D and DIC for models with discrete parameters, and his suggestion that marginal posterior modes could be used for $\bar{\theta}$ in this case.

We thank Nelder and Atkinson for their refinements to the analysis of the stack loss data (Section 8.2). We disagree with Smith that our models 4 and 5 for these data are predictively identical since, as already discussed, the prediction problem addressed by model 4 integrates over the random effects and corresponds to predicting stack loss for a new chimney, whereas model 5 conditions on the random effects and corresponds to predicting future stack loss for the 21 chimneys in the data set.

Alternatives to DIC

Several discussants (Brooks, Dawid and Sahu) feel that DIC suffers in comparison with more traditional Bayesian model selection criteria based on posterior model probabilities and Bayes factors. Here we can only repeat that our deliberate intention was to offer an alternative to Bayes factors, which are most suitable when the entire collection of candidate models can be specified ahead of time (the '$\mathcal{M}$ closed' case of Bernardo and Smith (1994)). In our practical experience, the model-building, criticism and rebuilding process is typically an iterative '$\mathcal{M}$ open' one in which the ultimate model collection is rarely known ahead of time, and here DIC may emerge as more appropriate. Moreover, Bayes factors address how well the prior has predicted the observed data; this prior predictive emphasis ultimately leads to the Lindley paradox. DIC instead addresses how well the *posterior* might predict future data generated by the same mechanism that gave rise to the observed data; this posterior predictive outlook might be considered intuitively more appealing in many practical contexts. We emphasize that these techniques are intended to answer different questions and cannot be expected to give the same conclusions: in any case, posterior model probabilities may be highly dependent on within- and between-model priors, so their comparison with DIC is not straightforward. On a related point, several discussants (Brooks, Burnham and Draper) mention the possible alternative of model averaging. We do not, however, see any justification for transforming DIC-values to relative probabilities, and in any case the prior on the model space may be difficult to develop, and might even reasonably be related to model complexity!

Dawid wishes for a better definition of $p \log(n)$ (instead of just p) for use in the Bayesian information criterion (BIC) but previous work has shown that many such definitions are justifiable asymptotically (e.g. Volinsky and Raftery (2000)), so this line of research does not appear promising. Regarding the suggestion by Lawson and Clark of using $\bar{p} \log(n)$ as a penalty for the BIC, this of course assumes that the number of parameters p is a suitable measure of model complexity. But most spatial models of the type that they refer to will involve random effects, where such use of the raw parameter count p would be inappropriate; indeed, this is precisely the situation that p_D was designed to address.

Vehtari and de Luna argue persuasively on behalf of cross-validation as an alternative to our posterior predictive approach that avoids a definition of complexity. Whereas no knowledge of the data-generating mechanism is required for cross-validation, the data-generating mechanism *is* necessary in a fully Bayesian analysis. Still, cross-validation as an alternative estimation method was also used to estimate model complexity by Efron (1986). We certainly acknowledge the potential of this approach, particularly in comparisons of different model selection strategies. We agree with Stone concerning further investigation of model assessment procedures in which the model is not assumed to be correct, and we refer to Konishi and Kitagawa (1996) (whose GIC adds yet further to the alphabet).

In conclusion, it is clear that several of the discussants feel that our pragmatic aims are muddying otherwise pure Bayesian waters. We feel, however, that the huge increase in the use of Bayesian methods in complex practical problems means that full elicitation of informative priors and utilities is simply not feasible in most situations, and that reasonably simple and robust methods for prior specification, model criticism and model comparison are necessary. We hope that we have made a positive contribution to the final concern.

References in the discussion

Aitkin, M. (1991) Posterior Bayes factors (with discussion). *J. R. Statist. Soc.* B, **53**, 111–142.

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Information Theory* (eds B. N. Petrov and F. Csáki), pp. 267–281. Budapest: Akadémiai Kiadó.

Atkinson, A. C. (1980) A note on the generalized information criterion for choice of a model. *Biometrika*, **67**, 413–418.

Atkinson, A. C. and Riani, M. (2000) *Robust Diagnostic Regression Analysis*. New York: Springer.

———(2002) Forward search added variable t tests and the effect of masked outliers on model selection and transformation. *Technical Report LSERR73*. London School of Economics and Political Science, London.

Bernardo, J. M. (1979) Expected information as expected utility. *Ann. Statist.*, **7**, 686–690.

———(1999) Nested hypothesis testing: the Bayesian reference criterion (with discussion). In *Bayesian Statistics 6* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 101–130. Oxford: Oxford University Press.

Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory*. New York: Wiley.

Bernardo, J. M. and Suarez, M. (2002) Intrinsic estimation. *7th Valencia Int. Meet. Bayesian Statistics, Tenerife, June*.

Burnham, K. P. and Anderson, D. R. (1998) *Model Selection and Inference: a Practical Information-theoretic Approach*. New York: Springer.

———(2002) *Model Selection and Multimodel Inference: a Practical Information-theoretical Approach*, 2nd edn. New York: Springer.

Casella, G., Robert, C. P. and Wells, M. T. (2000) Mixture models, latent variables and partitioned importance sampling. *Technical Report*. Paris.

Celeux, G., Hurn, M. and Robert, C. P. (2000) Computational and inferential difficulties with mixtures posterior distribution. *J. Am. Statist. Ass.*, **95**, 957–979.

Cooke, R. M. (1991) *Experts in Uncertainty*. Oxford: Oxford University Press.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L. and Spiegelhalter, D. J. (1999) *Probabilistic Networks and Expert Systems*. New York: Springer.

Daniels, M. J. and Kass, R. E. (1999) Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *J. Am. Statist. Ass.*, **94**, 1254–1263.

———(2001) Shrinkage estimators for covariance matrices. *Biometrics*, **57**, 1173–1184.

Dawid, A. P. (1984) Statistical theory: the prequential approach. *J. R. Statist. Soc.* A, **147**, 278–292.

———(1986) Probability forecasting. In *Encyclopedia of Statistical Sciences*, vol. 7 (eds S. Kotz, N. L. Johnson and C. B. Read), pp. 210–218. New York: Wiley-Interscience.

———(1991) Fisherian inference in likelihood and prequential frames of reference (with discussion). *J. R. Statist. Soc.* B, **53**, 79–109.

———(1992a) Prequential analysis, stochastic complexity and Bayesian inference (with discussion). In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 109–125. Oxford: Oxford University Press.

———(1992b) Prequential data analysis. In *Current Issues in Statistical Inference: Essays in Honor of D. Basu* (eds M. Ghosh and P. K. Pathak), pp. 113–126. Hayward: Institute of Mathematical Statistics.

Draper, D. (1999) Discussion on 'Decision models in screening for breast cancer' (by G. Parmigiani). In *Bayesian Statistics 6* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 541–543. Oxford: Oxford University Press.

Draper, D. and Fouskakis, D. (2000) A case study of stochastic optimization in health policy: problem formulation and preliminary results. *J. Global Optimzn*, **18**, 399–416.

Dupuis, J. and Robert, C. P. (2002) Model choice in qualitative regression models. *J. Statist. Planng Inf.*, to be published.

Efron, B. (1986) How biased is the apparent error rate of a prediction rule? *J. Am. Statist. Ass.*, **81**, 461–470.

Fouskakis, D. and Draper, D. (2002) Stochastic optimization: a review. *Int. Statist. Rev.*, to be published.

Gangnon, R. and Clayton, M. (2002) Cluster modelling for disease rate mapping. In *Spatial Cluster Modelling* (eds A. B. Lawson and D. Denison), ch. 8. New York: CRC Press.

Gelfand, A. E. (1996) Model determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), pp. 145–162. London: Chapman and Hall.

Gelfand, A. E., Dey, D. K. and Chang, H. (1992) Model determination using predictive distributions with implementation via sampling-based methods (with discussion). In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 147–167. Oxford: Oxford University Press.
Gelman, A., Meng, X.-L. and Stern, H. (1996) Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statist. Sin.*, **6**, 733–807.
Good, I. J. (1952) Rational decisions. *J. R. Statist. Soc.* B, **14**, 107–114.
Green, P. and Richardson, S. (2002) Hidden Markov models and disease mapping. *J. Am. Statist. Ass.*, to be published.
Hodges, J. and Sargent, D. (2001) Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika*, **88**, 367–379.
Holmes, C. and Denison, D. (1999) Bayesian wavelet analysis with a model complexity prior. In *Bayesian Statistics 6* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 769–776. Oxford: Oxford University Press.
Kass, R. and Raftery, A. (1995) Bayes factors and model uncertainty. *J. Am. Statist. Ass.*, **90**, 773–795.
Key, J. T., Pericchi, L. R. and Smith, A. F. M. (1999) Bayesian model choice: what and why? In *Bayesian Statistics 6* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 343–370. Oxford: Oxford University Press.
King, R. (2001) Bayesian model discrimination in the analysis of capture-recapture and related data. *PhD Thesis*. School of Mathematics, University of Bristol, Bristol.
King, R. and Brooks, S. P. (2001) Bayesian estimation of census undercount. *Biometrika*, **88**, 317–336.
Konishi, S. and Kitagawa, G. (1996) Generalised information criteria in model selection. *Biometrika*, **83**, 875–890.
Lauritzen, S. L. and Spiegelhalter, D. J. (1988) Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *J. R. Statist. Soc.* B, **50**, 157–224.
Lawson, A. B. (2000) Cluster modelling of disease incidence via rjmcmc methods: a comparative evaluation. *Statist. Med.*, **19**, 2361–2376.
Lee, Y. and Nelder, J. A. (1996) Hierarchical generalized linear models (with discussion). *J. R. Statist. Soc.* B, **58**, 619–678.
———(2001a) Hierarchical generalized linear models: a synthesis of generalized linear models, random effect models and structured dispersions. *Biometrika*, **88**, 987–1006.
———(2001b) Modelling and analysing correlated non-normal data. *Statist. Modlng*, **1**, 3–16.
de Luna, X. and Skouras, K. (2003) Choosing a model selection strategy. *Scand. J. Statist.*, to be published.
Madigan, D. and Raftery, A. E. (1991) Model selection and accounting for model uncertainty in graphical models using Occam's window. *Technical Report 213*. Department of Statistics, University of Washington, Seattle.
McKeague, I. and Loiseaux, M. (2002) Perfect sampling for point process cluster modelling. In *Spatial Cluster Modelling* (eds A. B. Lawson and D. Denison), ch. 5. New York: CRC Press.
Meng, X.-L. and Rubin, D. B. (1992) Performing likelihood ratio tests with multiply imputed data sets. *Biometrika*, **79**, 103–112.
Moreno, E., Pericchi, L. R. and Kadane, J. (1998) A robust Bayesian look at the theory of precise measurement. In *Decision Research from Bayesian Approaches to Normative Systems* (eds J. Shantan *et al.*). Boston: Kluwer.
Neter, J., Kutner, M. H., Nachtsheim, C. J. and Wasserman, W. (1996) *Applied Linear Statistical Models*, 4th edn. New York: McGraw-Hill.
Pericchi, L. R. and Walley, P. (1991) Robust Bayesian credible intervals and prior ignorance. *Int. Statist. Rev.*, **58**, 1–23.
Plummer, M. (2002) Some criteria for Bayesian model choice. *Preprint*. (Available from `http://calvin.iarc.fr/martyn/papers/`.)
Priestley, M. B. (1981) *Spectral Analysis and Time Series*. London: Academic Press.
Robert, C. P. (1996) Intrinsic loss functions. *Theory Decsn*, **40**, 191–214.
Shao, J. (1997) An asymptotic theory for linear model selection. *Statist. Sin.*, **7**, 221–264.
Skouras, K. and Dawid, A. P. (1999) On efficient probability forecasting systems. *Biometrika*, **86**, 765–784.
———(2000) Consistency in misspecified models. *Research Report 218*. Department of Statistical Science, University College London, London. (Available from: `http://www.ucl.ac.uk/Stats/research/abs00.html#218`.)
Smith, J. Q. (1996) Plausible Bayesian games. In *Bayesian Statistics 5* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 387–406. Oxford: Oxford University Press.
Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions (with discussion). *J. R. Statist. Soc.* B, **36**, 111–147.
———(1977) An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. R. Statist. Soc.* B, **36**, 44–47.
Vehtari, A. (2001) Bayesian model assessment and selection using expected utilities. *DSc Dissertation*. Helsinki University of Technology, Helsinki. (Available from `http://lib.hut.fi/Diss/2001/isbn9512257653/`.)
Vehtari, A. and Lampinen, J. (2002a) Bayesian model assessment and comparison using cross-validation predictive densities. *Neur. Computn*, **14**, in the press.

———(2002b) Cross-validation, information criteria, expected utilities and the effective number of parameters. To be published.
Volinsky, C. T. and Raftery, A. E. (2000) Bayesian information criterion for censored survival models. *Biometrics*, **56**, 256–262.
Weisberg, S. (1981) A statistic for allocating C_p to individual cases. *Technometrics*, **23**, 27–31.
Ye, J. (1998) On measuring and correcting the effects of data mining and model selection. *J. Am. Statist. Ass.*, **93**, 120–131.
Zhu, L. and Carlin, B. (2000) Comparing hierarchical models for spatio-temporally misaligned data using the deviance information criterion. *Statist. Med.*, **19**, 2265–2278.