Bayesian measures of model complexity and fit

David J Spiegelhalter [*] **Nicola G Best** [†] **Bradley P Carlin** [‡]
Angelika van der Linde [§]

August 6, 2001

Abstract

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 measure of fit, and the contributions of individual observations to 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* (DIC) for comparing models, which is related to other information criteria and has an approximate decision-theoretic justification. The procedure is illustrated in a number of examples, and comparisons drawn with alternative Bayesian and classical proposals. Throughout it is emphasised that the required quantities are trivial to compute in a Markov chain Monte Carlo analysis.

1 Introduction

The development of Markov chain Monte Carlo (MCMC) has made it possible to fit increasingly large classes of models with the aim of exploring real-world complexities of data (Gilks *et al.*, 1996). This ability naturally leads us to wish to compare alternative formulations with the aim of identifying a class of succinct models which appear to describe the data adequately: for example, we might ask whether we need to incorporate a random effect to allow for over-dispersion, 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

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

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

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

[§] Dept. of Mathematics, Institute of Statistics, University of Bremen, PO Box 330 440, 28334 Bremen, Germany : e-mail `a.vdl@t-online.de`

model. Since increasing complexity is accompanied by better fit, models are compared by trading off these two quantities and, following early work of Akaike (1973), proposals are often formally based on minimising a measure of expected loss on a future replicate dataset: see, for example, Efron (1986), Ripley (1996), and Burnham and Anderson (1998). Bayesian model comparison using Schwarz's information criterion as a Bayes factor approximation also requires specification of the number of parameters in each model (Kass and Raftery, 1995), 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 a measure of complexity and fit that can be combined in order to compare models of arbitrary structure.

In the next section we use an information-theoretic argument to suggest 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 a Markov chain Monte Carlo (MCMC) analysis, and Section 3 shows that p_D is approximately the trace of the product of Fisher's information and the posterior covariance matrix. 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 illustrate its form for different hierarchical models, while its properties in exponential families are explored in Section 5. The posterior mean deviance $\overline{D}$ can be taken as a measure of fit, and Section 6 shows how in exponential family models an observation's contributions to $\overline{D}$ and p_D can be used as residual and leverage diagnostics respectively.

In Section 7 we tentatively suggest that the fit $\overline{D}$ and complexity p_D may be added to form a *Deviance Information Criterion* (DIC) which may be used for model comparison. We describe how this parallels development of non-Bayesian information criteria, and has an approximate decision-theoretic justification. In Section 8 we illustrate the use of this technique on a number of reasonably complex examples. Finally, Section 9 draws some conclusions concerning these proposed techniques.

2 The complexity of a Bayesian model

2.1 'Focussed' full probability models

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

$$p(y) = \int_{\Theta} p(y|\theta)p(\theta)d\theta. \tag{0.1}$$

Particular choices of $p(y|\theta)$ and $p(\theta)$ will be termed a model 'focussed' on Θ. Note that we might further parameterise our prior with unknown 'hyper-parameters' ψ to create a hierarchical model, so that the full probability model factorises as $p(y, \theta, \psi) = p(y, \theta)p(\theta|\psi)p(\psi)$. Then depending on the parameters in focus, the model may compose likelihood $p(y|\theta)$ and prior $p(\theta) = \int_{\Psi} p(\theta|\psi)p(\psi)d\psi$, or likelihood $p(y|\psi) = \int_{\Theta} p(y|\theta)p(\theta|\psi)d\theta$ and prior $p(\psi)$. Both these models lead to the same marginal distribution (0.1), 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, 2000). In fact by focussing 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 focussed on the group means:

$$y_i|\theta_i \sim N(\theta_i, \tau_i^{-1}), \qquad \theta_i \sim N(\psi, \lambda^{-1}),\ i = 1, ..., p. \tag{0.2}$$

This model could also be focussed 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 focussed model, both in its own right, say to assess 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 dependency between parameters that is likely to reduce the effective dimensionality, although the degree of reduction may depend on the data available. Heuristically, complexity reflects 'difficult in estimation' and hence it seems reasonable that a measure of complexity may depend both on the prior information concerning the parameters in focus, and the specific data observed.

2.2 Is there a true model?

We follow Box (1976) in believing "all models are wrong, but some are useful". However, it can be useful to posit a 'true' distribution $p^T(Y)$ of unobserved future data Y conditional on the focussed model considered, since this defines a 'pseudo-true' parameter value θ^T (Sawa, 1978) as that which specifies a likelihood $p(Y|\theta^T)$ that minimises the Kullback-Leibler distance $E^T[\log p^T(Y)/p(Y|\theta^T)]$ from $p^T(Y)$. 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.

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 we have an estimator $\tilde{\theta}(y)$ of the pseudo-true parameter θ^T. Then the excess of the true over the estimated residual information will be denoted

$$d_\Theta(y, \theta^T, \tilde{\theta}(y)) = -2\log p(y|\theta^T) + 2\log p(y|\tilde{\theta}(y)). \tag{0.3}$$

This can be thought of as the reduction in surprise or uncertainty due to estimation, or alternatively the degree of 'over-fitting' 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 $2\log p(y|\theta^T)$ around $2\log p(y|\hat{\theta}(y))$, take expectations with respect to the unknown true sampling distribution $p^T(Y)$, and hence show (Ripley, 1996)[p 34]

$$\mathrm{E}^T\left[d_\Theta(Y,\theta^T,\tilde{\theta}(Y))\right] \approx p^* = \mathrm{tr}(KJ^{-1}), \tag{0.4}$$

where

$$J = -\mathrm{E}^T\left[\frac{\delta^2 \log p(Y,\theta^T)}{\delta\theta^2}\right], \quad K = \mathrm{Var}^T\left[\frac{\delta \log p(Y,\theta^T)}{\delta\theta}\right]. \tag{0.5}$$

This is the measure of complexity used in TIC (Takeuchi, 1976). Burnham and Anderson (1998)[p. 244] point out that

$$p^* = \mathrm{tr}(J\Sigma), \tag{0.6}$$

where $\Sigma = J^{-1}KJ^{-1}$ is the familiar 'sandwich' approximation to the variance-covariance matrix of the $\hat{\theta}(y)$ (Huber, 1967). Note that if $p^T(y) = p(y|\theta^T)$, *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

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

with τ_i^{-1}'s known,

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

whose expectation under $p^T(Y)$ is $p^* = \sum_i \tau_i V^T(Y_i)$. If the model is true, $V^T(Y_i) = \tau_i^{-1}$ and so $p^* = p$.

Ripley (1996)[p 140] shows how this procedure may be extended to 'regularised' models in which a specified prior term $p(\theta)$ is introduced to form a penalized log-likelihood $L_1 = \log p(y|\theta) + \log p(\theta)$. Replacing $\log p$ by L_1 in (0.5), and defining θ^T as obeying $\delta L_1(Y,\theta^T)/\delta\theta = 0$, 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 used in NIC (Murata *et al.*, 1994): estimation of p^* is generally not straightforward (Ripley, 1996).

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

$$p^* = \sum_i \rho_i \tau_i V^T(Y_i), \tag{0.7}$$

which becomes

$$p^* = \sum_i \rho_i \tag{0.8}$$

if the likelihood is true.

2.5 A Bayesian measure of model complexity

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

$$\begin{aligned} p_D(y,\Theta,\tilde{\theta}(y)) &= E_{\theta|y}[d_\Theta(y,\theta,\tilde{\theta}(y))] \\ &= E_{\theta|y}[-2\log p(y|\theta)] + 2\log p(y|\tilde{\theta}(y)). \end{aligned} \qquad (0.9)$$

$p_D(y,\Theta,\tilde{\theta}(y))$ is our proposal as the 'effective number of parameters' with respect to a model with focus Θ: we will usually drop the arguments $(y,\Theta,\tilde{\theta}(y))$ from the notation. In our examples we will generally take $\tilde{\theta}(y) = E[\theta|y] = \overline{\theta}$, 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: see Section 2.6.

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

$$p_D = \overline{D(\theta)} - D(\overline{\theta}) \qquad (0.10)$$

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 5.

(0.10) 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 show 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 with ψ,λ known,

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

which is -2 log(likelihood) standardised by the term $-2\log f(y) = \sum_i \log\frac{2\pi}{\tau_i}$ 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 chi-square distribution with mean $a+b$. Thus, since $\rho_i\lambda = (1-\rho_i)\tau_i$, we have

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

and so

$$p_D = \sum_i \rho_i = \sum_i \frac{\tau_i}{\tau_i+\lambda}. \qquad (0.11)$$

The effective number of parameters is therefore the sum of the intra-class correlation coefficients, which essentially measures the sum of the ratios of the precision in the likelihood to the precision in the posterior. This exactly matches Moody's approach (0.8) when the model is true.

Giving ψ a uniform hyper-prior we obtain a posterior distribution $\psi \sim N(\overline{y}, (\lambda \sum \rho_i)^{-1})$, where $\overline{y} = \sum \rho_i y_i / \sum \rho_i$. It is straightforward to show that

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

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

1. Simple use of Bayes theorem reveals the expression

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

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

2. It is reasonable that the effective number of parameters in a model might depend both 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. 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.

3. It follows from (0.10) and Jensen's inequality that, when using the posterior mean as an estimator $\tilde{\theta}(y)$, $p_D \geq 0$ 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 .5 and .5, and so $\overline{\theta} = 1.5$. Thus $p_D = \overline{D(\theta)} - D(\overline{\theta}) = \log 10 - 2\log 13/4 = \log 160/169 < 0$. Our experience has been that negative p_D's indicate substantial conflict between prior and data.

4. The posterior distribution used in obtaining p_D assumes the model is valid, and hence p_D may only be considered an appropriate measure of the complexity of a model that reasonably describes the data. Thus our simple ANOVA example shows that p_D (0.11) will not necessarily be approximately equivalent to the classical p^* (0.7) if the assumptions of the model are substantially inaccurate. This suggests that, for models under serious consideration, we should be willing to assume that $p(Y|\theta^T)$ is a reasonable approximation to $p^T(Y)$.

5. Since the complexity depends on the focus, a decision must be made whether nuisance parameters, for example variances, are to be included in Θ or integrated out before specifying the likelihood $p(y|\theta)$.

6. 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. 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.

p_D has been defined and is trivially computable using MCMC, and so strictly speaking there is no need to explore exact forms or approximations. However, in order 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 'pre-asymptotic' 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 focussed 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 $E_{\theta|y}[\theta] = \overline{\theta}$ to give, to second order,

$$\begin{aligned} D(\theta) &\approx D(\overline{\theta}) + (\theta - \overline{\theta})^T \left.\frac{\delta D}{\delta \theta}\right|_{\overline{\theta}} + \frac{1}{2}(\theta - \overline{\theta})^T \left.\frac{\delta^2 D}{\delta \theta^2}\right|_{\overline{\theta}} (\theta - \overline{\theta}) & (0.12) \\ &= D(\overline{\theta}) - 2(\theta - \overline{\theta})^T L'_{\overline{\theta}} - (\theta - \overline{\theta})^T L''_{\overline{\theta}}(\theta - \overline{\theta}) & (0.13) \end{aligned}$$

where $L = \log p(y|\theta) = -D/2$ 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 (0.13) with respect to the posterior distribution of θ gives

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

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

Thus

$$p_D \approx \mathrm{tr}\left(-L''_{\overline{\theta}} V\right), \qquad (0.14)$$

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 (0.6).

We also note that

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

where $Q'' = \delta^2 \log p(\theta|y)/\delta\theta^2$ and $P'' = \delta^2 \log p(\theta)/\delta\theta^2$, and hence (0.14) can be written

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

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

$$p_D \approx p - \mathrm{tr}\left(-P''_{\overline{\theta}}V\right) \tag{0.15}$$

where p is the cardinality of Θ.

3.2 p_D for approximately normal likelihoods and negligible prior information

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

$$\theta|y \sim \mathrm{N}\left(\hat{\theta}, -L''_{\hat{\theta}}\right) \tag{0.16}$$

holds, where $\overline{\theta} = \hat{\theta}$ are the maximum likelihood estimates such that $L'_{\hat{\theta}} = 0$ (Bernardo and Smith, 1994)[Ch 5.3]. From (0.13)

$$\begin{aligned} D(\theta) &\approx D(\hat{\theta}) - (\theta - \hat{\theta})^T L''_{\hat{\theta}}(\theta - \hat{\theta}) \\ &\approx D(\hat{\theta}) + \chi^2_p, \end{aligned} \tag{0.17}$$

since, by (0.16), $-(\theta - \hat{\theta})^T L''_{\hat{\theta}}(\theta - \hat{\theta})$ has an approximate chi-squared distribution with p degrees of freedom.

Rearranging (0.17) and taking expectations with respect to the posterior distribution of θ reveals

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

i.e., p_D will be approximately the true number of parameters. This approximate identity is illustrated in Section 8.

We note in passing that in this context $\mathrm{V}_{\theta|y}[D(\theta)] \approx 2p$, so we can use MCMC output to estimate the classical deviance $D(\hat{\theta})$ of any likelihood-based model by

$$\hat{D}(\hat{\theta}) = E_{\theta|y}[D(\theta)] - \frac{1}{2}V_{\theta|y}[D(\theta)], \tag{0.18}$$

using the empirical mean and variance of the sampled values for D. Although this maximum likelihood deviance is theoretically the minimum of D over all feasible values of θ, $D(\hat{\theta})$ will generally be very badly estimated by the sample minimum over an MCMC run, and so the estimator given by (0.18) may be preferable.

4 p_D for normal likelihoods

4.1 The normal linear model

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

$$y \sim N(A_1\theta, C_1), \qquad \theta \sim N(A_2\psi, C_2) \tag{0.19}$$

where all matrices and vectors are of appropriate dimension, and C_1, C_2 are assumed known. Then the standardised deviance is $D(\theta) = (y - A_1\theta)^T C_1^{-1}(y - A_1\theta)$. Assume the posterior distribution for θ is normal with mean $\overline{\theta} = Vb$ and covariance V: V and b will be left unspecified for the moment. Then expressing $y - A_1\theta$ as $y - A_1\overline{\theta} + A_1\overline{\theta} - A_1\theta$ reveals that

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

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

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

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

$$p_D = \mathrm{tr}\left(-L'' V\right) : \tag{0.20}$$

an exact version of (0.14). 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 show that $V^{-1} = A_1^T C_1^{-1} A_1 + C_2^{-1}$ and hence from (0.20)

$$p_D = p - \mathrm{tr}\left(C_2^{-1} V\right) \tag{0.21}$$

as an exact version of (0.15), and so $0 \leq p_D \leq p$, and $p - p_D$ is the measure of the 'shrinkage' of the posterior estimates towards the prior means. If $(C_2^{-1}V)^{-1} = A_1^T C_1^{-1} A_1 C_2 + I_p$ has eigenvalues $\lambda_i + 1, i = 1, ..., p$, then

$$p_D = \sum_{i=1}^{p} \frac{\lambda_i}{\lambda_i + 1}, \tag{0.22}$$

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 $b = A_1^T C_1^{-1} y$ and the fitted values for the data are given by $\hat{y} = A_1\overline{\theta} = A_1 V b = A_1 V A_1^T C_1^{-1} y$. Thus the 'hat' matrix that projects the data onto the fitted values is $H = A_1 V A_1^T C_1^{-1}$, and

$$p_D = \mathrm{tr}(A_1^T C_1^{-1} A_1 V) = \mathrm{tr}(A_1 V A_1^T C_1^{-1}) = \mathrm{tr}(H). \tag{0.23}$$

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

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)[p.63] and generalised additive models (Hastie and Tibshirani, 1990)[Sec 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) considers the independent normal model

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

and suggests that the effective number of parameters should be $\sum_i h_i$, where

$$h_i(\theta) \quad = \quad \frac{\delta E_{y|\theta}[\tilde{\theta}_i]}{\delta \theta_i} : \tag{0.24}$$

the average sensitivity of an unspecified estimate $\tilde{\theta}_i$ to a small change in y_i. This is a generalisation of the trace of the 'hat' matrix discussed above. In the context of the normal linear models, it is straightforward to show that $\mathrm{E}_{Y|\theta}(\overline{\theta}) = H\theta$, and hence $p_D = \mathrm{tr}(H)$ 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

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

where the covariance matrices C_1 and D are currently assumed known. The random effects are β, fixed effects are α, and placing a uniform prior on α we can write this model within the general Lindley-Smith formulation (0.19) by setting $\theta = (\alpha, \beta), A_1 = (X, Z), \psi = 0$ and C_2 as a block-diagonal matrix with infinities in the top-left block, D in the bottom right, and zeros elsewhere.

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

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

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

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

A common model in semiparametric regression is

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

where β is a vector of length n of function values of the non-parametric part of an interpolation spline (Wahba, 1990; van der Linde, 1995), and C_1, 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)[p 63] and so, for example, $\hat{\tau}^{-1}$ is the residual sum of squares divided by the 'effective degrees of freedom' $n - \text{tr}(H)$. 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)[Ch 4.2] also drew attention to the diagonal entries of the hat matrix as leverage values.

A link to partially Bayesian interpolation models has been provided by G and G (1970); Wahba (9878); Wahba (1983) and further work built on these ideas. For example, another large class of models can be formulated using the following extension to the Lindley-Smith model:

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

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, D will be identity matrices. Define

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

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

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

which are the fitted likelihood and prior residual variation, divided by the effective degrees of freedom which turns out to be equivalent to $p_D = \text{tr}(H)$.

These results were derived by MacKay (1992) in the context of 'regularisation' in complex interpolation models such as neural networks, in which the parameters θ are standardised and assumed to have independent normal priors with mean 0 and precision λ. Then expression (0.15) may be written

$$p_D \approx p - \lambda\ \text{tr}(V). \tag{0.25}$$

However, Mackay's use of (0.25) requires evaluation of $\text{tr}(V)$, while 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)[p. 167]. In this and the smoothing context a fully Bayesian analysis requires prior distributions for τ^{-1}, 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 (0.19) to the conjugate normal-gamma model with an unknown scale parameter τ in both likelihood and prior (Bernardo and Smith, 1994)[Ch 5.2.1]. Suppose

$$y \sim N(A_1\theta, \tau^{-1}C_1), \qquad \theta \sim N(A_2\psi, \tau^{-1}C_2), \tag{0.26}$$

and we focus on (θ, τ). The standardised deviance is $D(\theta,\tau) = \tau q(\theta) - n \log \tau$, where $q(\theta) = (y - A_1\theta)^T C_1^{-1}(y - A_1\theta)$ is the residual variation. Then, for a currently unspecified estimator $\hat{\tau}$,

$$\begin{aligned} p_D &= \mathrm{E}_{\theta,\tau|y}[D|\theta,\tau] - D(\overline{\theta},\hat{\tau}) \\ &= \mathrm{E}_{\tau|y}\left[\mathrm{E}_{\theta|\tau,y}[\tau q] - n\log\tau\right] - \left[\hat{\tau}q(\overline{\theta}) - n\log\hat{\tau}\right] \\ &= \mathrm{tr}(H) + q(\overline{\theta})(\overline{\tau} - \hat{\tau}) - n(\overline{\log\tau} - \log\hat{\tau}) \qquad (0.27) \end{aligned}$$

where $H = A_1^T C_1^{-1} A_1 (A_1^T C_1^{-1} A_1 + C_2^{-1})^{-1}$ 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 $\tau|y \sim \mathrm{Gamma}(a + \frac{n}{2}, b + \frac{S}{2})$, where $S = (y - A_1A_2\psi)^T(C_1 + A_1^T C_2 A_1)^{-1}(y - A_1A_2\psi)$. It remains to choose the estimator $\hat{\tau}$ to place in (0.27), and we shall consider two options.

Suppose we parametrise in terms of τ and use $\hat{\tau} = \overline{\tau} = (a + \frac{n}{2})/(b + \frac{S}{2})$, making the second term in (0.27) zero. Now if $X \sim \mathrm{Gamma}(a,b)$, then $\mathrm{E}(\log X) = \psi(a) - \log(b)$ where ψ is the digamma function, and so $\overline{\log\tau} = \psi(a + \frac{n}{2}) - \log(b + \frac{S}{2})$. Hence the term contributing to p_D due to the unknown precision is

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

using the approximation $\psi(x) \approx \log x - \frac{1}{2x} - \frac{1}{12x^2}$. This term will tend to $1 + 1/3n$ as prior information becomes negligible.

Were we to parameterise in terms of $\log\tau$ and use $\hat{\tau} = exp(\overline{\log\tau})$, the third term in (0.27) is 0 and the second term can be shown to be $1 - O(n^{-1})$. Thus for reasonable sample sizes the choice of parameterisation 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 due 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

$$\log p(y_{ij}|\theta_i,\phi) = w_i(y_{ij}\theta_i - b(\theta_i))/\phi + c(y_{ij},\phi), \qquad (0.28)$$

where

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

If the canonical parameterisation Θ is the focus of the model, then writing $\overline{b_i} = E_{\theta_i|y}[b(\theta_i)]$, we easily obtain that the contribution of the ith group to the effective number of parameters is

$$p_{Di}^{\Theta} = 2n_i w_i(\overline{b_i} - b(\overline{\theta}_i))/\phi. \qquad (0.29)$$

These likelihoods highlight the issue of the lack of invariance of p_D to re-parameterisation, since the mean parameterisation 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 generalised linear and generalised linear mixed models.

5.1 Binomial likelihood with conjugate prior

In the notation of (0.28), $\phi = 1, w_i = 1$, and $\theta = \text{logit}(\mu) = \log[\mu/(1-\mu)]$, and the (unstandardised) deviance is

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

where $y_i = \sum_j y_{ij}$. A conjugate prior $\mu_i = (1+e^{-\theta_i})^{-1} \sim \text{Beta}(a,b)$ 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 $E(\log X) = \psi(a) - \psi(a+b), E[\log(1-X)] = \psi(b) - \psi(a+b)$ where ψ is the digamma function, and hence it can be shown that:

$$\begin{aligned}
\overline{D(\mu_i)} = \overline{D(\theta_i)} &= -2y_i\psi(a+y_i) - 2(n_i-y_i)\psi(b+n_i-y_i) + 2n_i\psi(a+b+n_i) \\
D(\overline{\mu_i}) &= -2y_i\log(a+y_i) - 2(n_i-y_i)\log(b+n_i-y_i) + 2n_i\log(a+b+n_i) \\
D(\overline{\theta_i}) &= -2y_i\psi(a+y_i) + 2y_i\psi(b+n_i-y_i) + 2n_i\log(1+e^{\psi(a+y_i)-\psi(b+n_i-y_i)}) \\
D(\mu_i^{med}) = D(\theta_i^{med}) &= -2y_i\log\mu_i^{med} - 2(n_i-y_i)\log(1-\mu_i^{med})
\end{aligned}$$

Exact p_{D_i}'s are obtainable by subtraction, and Figure 1 shows how the value of p_{D_i} depends on the parameterisation, 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 parameterisations by using $\psi(x) \approx \log x - \frac{1}{2x} \approx \log(x - \frac{1}{2})$. This leads to

$$\begin{aligned}
p_{D_i}^{\mu} &\approx \frac{y_i}{a+y_i} + \frac{n_i-y_i}{b+n_i-y_i} - \frac{n_i}{a+b+n_i} \\
p_{D_i}^{\Theta} &\approx \frac{n_i}{a+b+n_i-\frac{1}{2}}
\end{aligned}$$

We make the following observations. *Behaviour of p_D:* for all three parameterisations, as the sample size in each group increases relative to the effective prior sample size, its contribution to p_{D_i} tends towards 1. *Agreement between parameterisations:* this is generally reasonable except in the situations in which the prior sample size is 10 times that of the data. While the canonical parameterisation has $p_{D_i} \approx 1/11$, the mean and median give increased p_{D_i} for extreme prior means. *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 sample size to effective posterior sample size. When there is dependence on data, p_{D_i} is higher in situations of prior/data conflict.

5.2 Poisson likelihood with conjugate prior

In the notation of (0.28), $\phi = 1, w_i = 1, \theta = \log\mu$, and the (unstandardised) deviance is $D(\mu_i) = -2y_i\log\mu_i + 2n_i\mu_i$. A conjugate prior $\mu_i = e^{\theta_i} \sim \text{Gamma}(a,b)$ 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

$$\begin{aligned}
\overline{D(\mu_i)} = \overline{D(\theta_i)} &= -2y_i(\psi(a+y_i) - \log(b+n_i)) + 2n_i\frac{(a+y_i)}{(b+n_i)} \\
D(\overline{\mu_i}) &= -2y_i(\log(a+y_i) - \log(b+n_i)) + 2n_i\frac{(a+y_i)}{(b+n_i)} \\
D(\overline{\theta_i}) &= -2y_i(\psi(a+y_i) - \log(b+n_i)) + 2n_i\frac{e^{\psi(a+y_i)}}{(b+n_i)} \\
D(\mu_i^{med}) = D(\theta_i^{med}) &= -2y_i\log\mu_i^{med} + 2n_i\mu_i^{med}
\end{aligned}$$

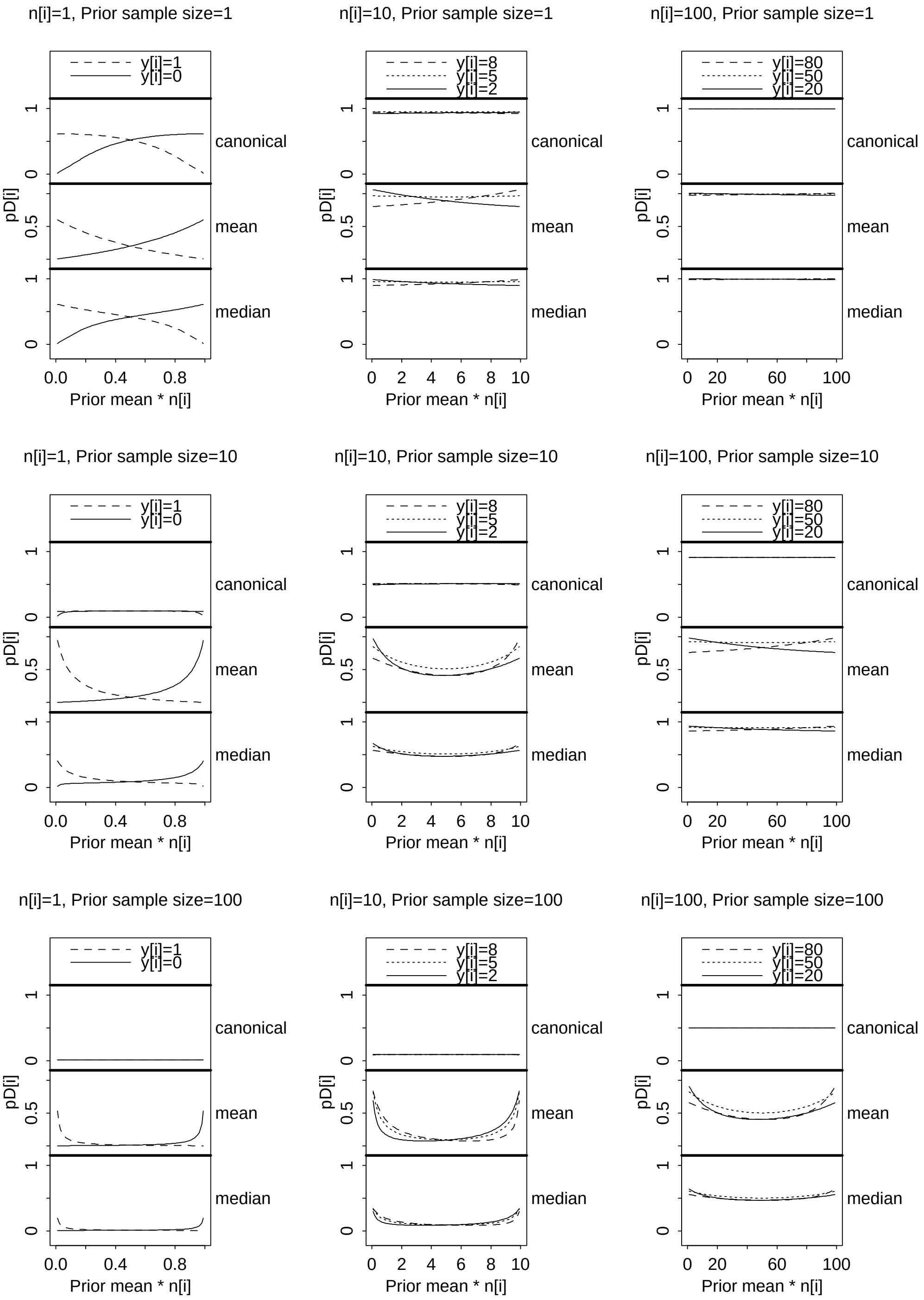

Figure 1: Binomial likelihood: contribution of the ith group to the effective number of parameters under different parameterisations (canonical $p_{D_i}^{\Theta}$, mean $p_{D_i}^{\mu}$ and median $p_{D_i}^{med}$) as a function of the data (sample size, n_i, and observed proportion, y_i/n_i) and prior (effective prior sample size, a+b, and the prior mean, $a/(a+b)$).

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

$$\begin{aligned} p^{\mu}_{D_i} &\approx \frac{y_i}{a+y_i} \\ p^{\Theta}_{D_i} &\approx \frac{n_i}{b+n_i} \end{aligned}$$

Behaviour of p_{D_i}*:* for all three parameterisations, as the sample size in each group increases relative to the effective prior sample size, its contribution to p_{D_i} tends towards 1. *Agreement between parameterisations:* this is best when there is no conflict between the prior expectation and the data, but can be substantial when such conflict is extreme. The median estimator leads to a p_{D_i} intermediate between those derived from the canonical and mean parameterisations. *Dependence on data:* except in the situation of a single $y_i = 0$ with weak prior information, the approximation for the canonical $p^{\Theta}_{D_i}$ is very accurate and so p_{D_i} does not depend on the data observed. There can be substantial dependence for the mean parameterisation, with p_{D_i} being higher when the prior mean underestimates the data.

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

5.3 Generalised linear models with canonical link functions

Here we shall focus on the canonical parameterisation in terms of θ_i, both for the reasons outlined above and because its likelihood should better fulfill the normal approximation underlying the optimality criterion described in Section 7.3: related identities are available for the mean parameterisation in terms of $\mu_i = \mu(\theta_i)$. See SLATE (1994) for more refined analysis of likelihood normality in this context.

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

$$p_D \approx \mathrm{tr}\left[X^T W X \quad V(\alpha|y)\right].$$

where W is diagonal with entries

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

the GLM iterated weights (McCullagh and Nelder, 1989, p. 40).

Under a $N(\alpha_0, C_2)$ 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

$$V(\alpha|y) = [C_2^{-1} + X^T W X]^{-1} ,$$

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

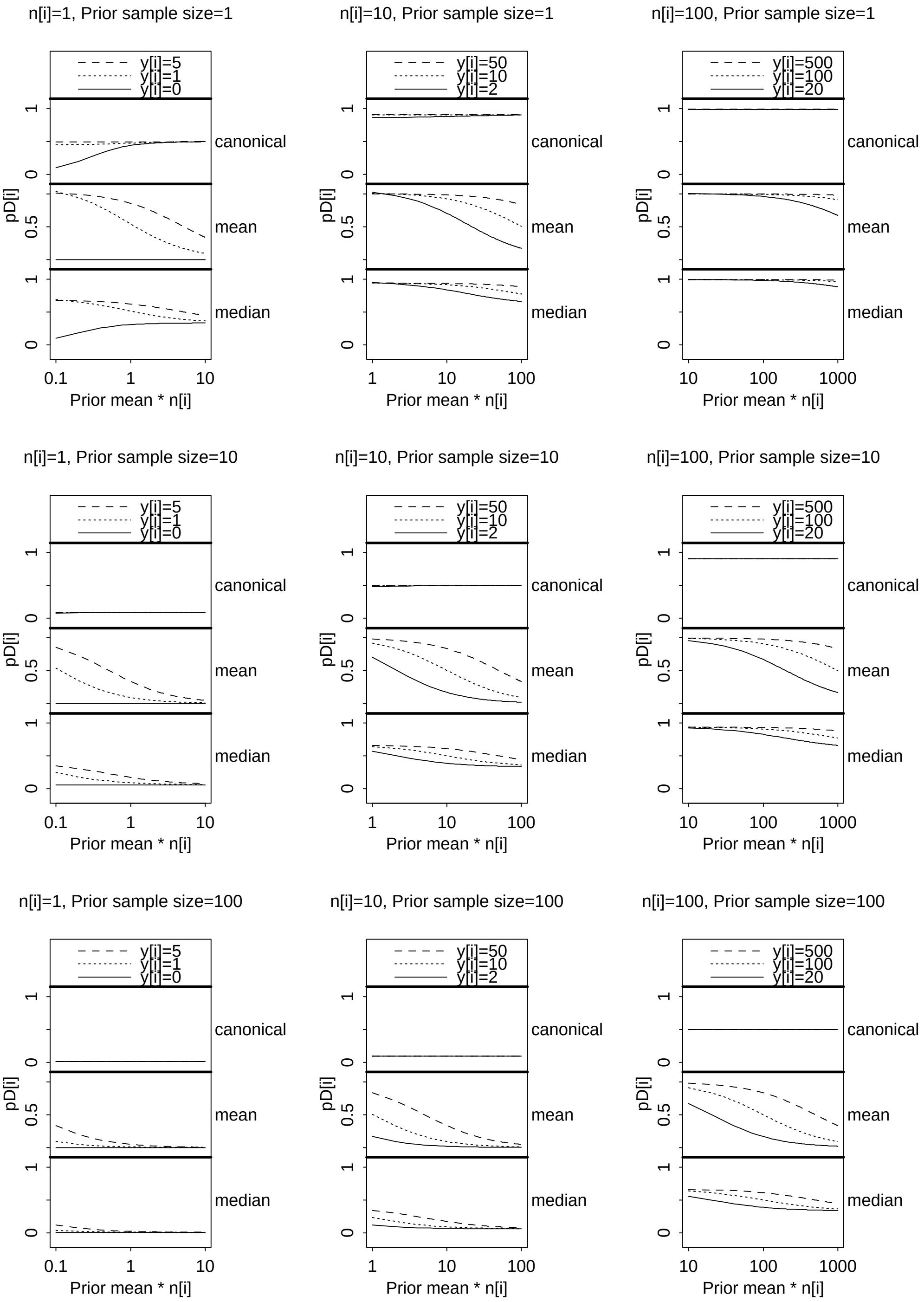

Figure 2: Poisson likelihood: contribution of the ith group to the effective number of parameters under different parameterisations (canonical $p_{D_i}^{\Theta}$, mean $p_{D_i}^{\mu}$ and median $p_{D_i}^{med}$) as a function of the data (sample size n_i and observed total y_i) and prior (expectation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

5.4 Generalised linear mixed models

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

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

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

where

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

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

6 Diagnostics for fit and influence

6.1 Posterior expected deviance as a measure of fit

The posterior mean of the deviance $E_{\theta|y}D(\theta)$ has been used as a measure of model fit by a number of authors: see, for example, Dempster (1974) (reprinted as Dempster (1997*b*)), Raghunathan (1988), Zeger and Karim (1991), Gilks *et al.* (1993) and Richardson and Green (1997). These authors have, however, not been explicit about how such a measure might be traded off against increasing complexity of a model: Dempster (1997*a*) suggests plotting log-likelihoods from MCMC runs but hesitates to dictate a model choice procedure. This issue is considered in Section 7.

6.2 Sampling-theory diagnostics for lack-of-fit

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

$$\begin{aligned} E_Y(\overline{D}) &= E_Y\left[E_{\theta|y}D(\theta)\right] && (0.30) \\ &= E_\theta\left[E_{Y|\theta}\left[-2\log p(Y|\theta) + 2\log f(Y)\right]\right] \end{aligned}$$

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

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

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 p, the dimensionality of θ. From (0.31) we therefore expect, if the model is true, the posterior expected deviance (standardised by the maximised log-likelihood) to be $E_Y(\overline{D}) \approx E_\theta[p] = p$, 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 maximised by substituting y_i for the mean of y_i, and the posterior mean of the standardised deviance has approximate sampling expectation n if the model is true. This will be exact for normal models with known variance, but in general will only be reliable if each observation provides considerable information about its mean (McCullagh and Nelder, 1989, p. 36). Note that comparing $\overline{D}$ with n is precisely the same as comparing $D(\overline{\theta})$ with $n - p_D$, the effective degrees of freedom.

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

$$\overline{D} = \sum_i D_i = \sum_i dr_i^2$$

where $dr_i = \pm\sqrt{D_i}$ (with sign given by the sign of $(y_i - E(y_i|\overline{\theta}))$) termed the Bayesian deviance residual, defined analagously to McCullagh and Nelder (1989) p 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 each observation has on its own fitted value. In general it can be shown, for y_i conditionally independent given θ, that

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

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 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 strict statistical analysis using, for example, a decision-theoretic paradigm: see, for example, (Key *et al.*, 1999). Our approach here can be considered semi-formal. While we believe 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 over-formal 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: for example the robustness of its conclusions, its inherent plausibility and so on. In addition, in many contexts it may not be appropriate to 'choose' a single model. Our development closely follows that of Section 2.

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

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

$$c_\Theta(y, \theta^T, \tilde{\theta}(y)) \;=\; \mathrm{E}_{Y_{rep}|\theta^T}\left[L(Y_{rep}, \tilde{\theta}(y))\right] - L(y, \tilde{\theta}(y)) \qquad (0.31)$$

and assume a logarithmic loss function $L(Y, \tilde{\theta}) = -2\log p(Y|\tilde{\theta})$.

Both classical and Bayesian approaches to estimating the optimism will now be examined: as in Section 2, the classical approach attempts to estimate the sampling expectation of c_Θ, while the Bayesian approach is based on 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

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

will, from (0.31), yield criteria for model comparison based on minimising

$$\hat{E}_{Y_{rep}|\theta^T}\left[L(Y_{rep}, \tilde{\theta}(y))\right] = L(y, \tilde{\theta}(y)) + \hat{\pi}(\theta^T). \qquad (0.32)$$

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

$$\pi_E(\theta^T) = 2\sum_i \mathrm{Cov}^T(\hat{Y}_i, Y_i), \qquad (0.33)$$

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 minimise 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 (0.33), may be adopted as a measure of complexity in more complex modelling situations. Ye and Wong (1998) extend 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 'generalised degrees of freedom' when using a general estimation procedure. In normal models with linear estimators $\hat{y}_i = \tilde{\theta}_i(y) = \sum_j h_{ij}y_j$, and so $\pi(\theta^T) = 2\mathrm{tr}(H)$. Finally, Ripley (1996) extends the analysis described in Section 2.4 to show that if the assumed model is not true then $\pi(\theta^T) \approx 2p^*$, where p^* is defined in (0.4). 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 (0.32), 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 adapt this structure to a Bayesian context.

7.3 Bayesian criteria for model comparison

Gelfand and Ghosh (1998) and Laud and Ibrahim (1995) both attempt strict decision-theoretic approaches to model choice based on expected losses on replicate datasets. 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 minimise uncertainty about observations generated in the same way. Thus, in analogy to the classical results described above, we propose a *Deviance Information Criterion* (DIC), defined as a 'plug-in' estimate of fit, plus twice the effective number of parameters, to give

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

by definition of p_D (0.10). 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 mimicing the development of Ripley (1996)[p33] and Burnham and Anderson (1998)[Ch 6]. Using the logarithmic loss function in (0.31), we obtain

$$c_\Theta(y, \theta^T, \tilde{\theta}(y)) = \mathrm{E}_{Y_{rep}|\theta^T}[D_{rep}(\tilde{\theta})] - D(\tilde{\theta})$$

where $-2\log p(Y_{rep}|\tilde{\theta}(y))$ is denoted $D_{rep}(\tilde{\theta})$ and so on. c_Θ can be broken down into

$$c_\Theta = \mathrm{E}_{Y_{rep}|\theta^T}[D_{rep}(\tilde{\theta}) - D_{rep}(\theta^T)] + \mathrm{E}_{Y_{rep}|\theta^T}[D_{rep}(\theta^T) - D(\theta^T)] + [D(\theta^T) - D(\tilde{\theta})]. \qquad (0.34)$$

We shall denote the first two terms by L_1 and 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

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

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

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

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

$$L_1(\theta, \tilde{\theta}) \approx \mathrm{tr}\left(-L''_{\tilde{\theta}}(\tilde{\theta} - \theta)(\tilde{\theta} - \theta)^T\right). \qquad (0.35)$$

Suppose that under a particular model assumption we obtain a posterior distribution $p(\theta|y)$. Then from (0.34) and (0.35) our posterior expected optimism when adopting this model and the estimator $\tilde{\theta}$ is

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

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

$$\mathrm{E}_{\theta|y}[c_\Theta] \approx \mathrm{tr}\left(-L''_{\overline{\theta}}V\right) + \mathrm{E}_{\theta|y}L_2(y, \theta) + p_D, \qquad (0.36)$$

where V again is defined as the posterior covariance of θ, and $p_D = \overline{D} - D(\overline{\theta})$. Now $L_2(y,\theta) = \mathrm{E}_{Y_{rep}|\theta}[-2\log p(Y_{rep}|\theta)] + 2\log p(y|\theta)$, and so $\mathrm{E}_Y \mathrm{E}_{\theta|Y} L_2(Y,\theta) = \mathrm{E}_\theta \mathrm{E}_{Y|\theta} L_2(Y,\theta) = 0$. We have already shown in (0.14) that $p_D \approx \mathrm{tr}\left(-L''_{\overline{\theta}} V\right)$, and hence from (0.31) and (0.36) the expected posterior loss when adopting a particular model is

$$\mathrm{E}_{\theta|y}[c_\Theta] + D(\overline{\theta}) \approx 2p_D + D(\overline{\theta}) = \mathrm{DIC}$$

plus a term expected to be zero.

We make the following observations concerning this somewhat heuristic justification of DIC. First, for the general normal linear model (0.19), it is straightforward to show that $L_2(y,\theta) = p - (y - A_1\theta)^T C_1^{-1}(y - A_1\theta)$ where p is the dimensionality of θ, and hence for true θ has sampling distribution $p - \chi^2_p$ with mean 0 and variance $2p$. This parallels the classical development, in which Ripley (1996)[p 34] points out that the equivalent term is $O(\sqrt{n})$.

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

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, 2000), multi-stage 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 counties in Scotland (Clayton and Kaldor, 1987; Breslow and Clayton, 1993). The data include observed (y_i) and expected (E_i) numbers of cases for each county i (where the expected counts are based on the age- and sex-standardised national rate applied to the population at risk in each county) 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 $e^{\theta_i}E_i$ where e^{θ_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:

$$\begin{array}{lrcl}
\text{Model 1:} & \theta_i & = & \alpha_0 \\
\text{Model 2:} & \theta_i & = & \alpha_0 + \gamma_i \\
\text{Model 3:} & \theta_i & = & \alpha_0 + \delta_i \\
\text{Model 4:} & \theta_i & = & \alpha_0 + \gamma_i + \delta_i \\
\text{Model 5:} & \theta_i & = & \alpha_i
\end{array}$$

An (improper) uniform prior is placed on α_0, independent (proper) Normal priors with large variance are specified for each $\alpha_i (i = 1, ..., 56)$, γ_i are exchangeable random effects with a Normal prior

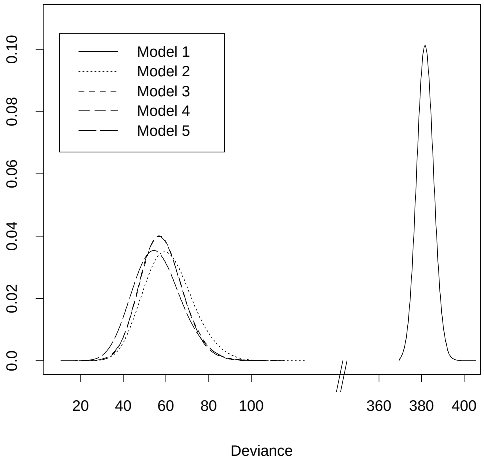

Figure 3: Posterior distributions of the deviance for each model considered in the lip cancer example

distribution having zero mean and precision λ_γ, and δ_i are spatial random effects with a conditional autoregressive prior (Besag, 1974) given by

$$\delta_i|\delta_{\backslash i} \quad \sim \quad \text{Normal}(\frac{1}{n_i}\sum_{j\in\mathcal{A}_i}\delta_j, \frac{1}{n_i\lambda_\delta}) \quad .$$

A sum-to-zero constraint is imposed on the $\{\delta_i\}$ for identifiability purposes, and weakly informative Gamma(0.5, 0.0005) priors are assumed for the random effects precision parameters λ_γ and λ_δ respectively. These five models cover the spectrum between a pooled model (1) that makes no allowance for variation between the true risk ratios in each county, to 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)[p 34] $D(\theta) = 2\sum_i \left[y_i \log(y_i/e^{\theta_i}E_i) - (y_i - e^{\theta_i}E_i)\right]$ obtained by taking $-2\log f(y) = -2\sum_i \log p(y_i|\hat{\theta}_i) =$ 208.0 as the standardising factor (see Section 2.5). 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 (1997*a*), Figure 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.

The deviance summaries proposed in this paper are shown for the lip cancer data in Table 1: $\overline{D}$ is simply the mean of the posterior samples of the saturated deviance; $D(\overline{\mu})$ is calculated by plugging the posterior mean of $\mu_i = e^{\theta_i}E_i$ into the saturated deviance; $D(\overline{\theta})$ 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; and $D(med)$ is calculated by plugging the posterior median of θ_i

	Model	$\overline{D}$	$D(\overline{\mu})$	p_D^μ	DIC$^\mu$	$D(\overline{\theta})$	p_D^θ	DIC$^\theta$	$D(med)$	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	exch	61.1	18.2	42.9	104.0	17.7	43.4	104.5	17.6	43.5	104.6
3	spat	58.3	26.6	31.7	89.9	27.1	31.2	89.5	27.2	31.1	89.3
4	exch+spat	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

Table 1: Deviance summaries for lip cancer data using three alternative parameterisations (mean, canonical and median) for the 'plug-in' deviance: 'exch' means an exchangeable random effect, 'spat' is a spatially correlated random effect.

(or equivalently, of μ_i) into the saturated deviance. The results are remarkably similar for the three alternative parameterisations 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, while for the saturated model (5), p_D ranges from 52.8 to 55.9 depending on the parameterisation 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, while the model with only exchangeable random effects has about 12 additional effective parameters. Based on 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 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. Whilst computing the precise standard errors for our DIC values is a subject of ongoing research, the standard errors for the $\overline{D}$ 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 are superior (in terms of DIC performance) to models 2 or 5, which are in turn superior to model 1. Comparison of DIC for models 3 and 4 suggests that the two spatial models are virtually indistinguishable in terms of overall fit: pragmatically, we might prefer reporting Model 3 (spatial) 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 $\overline{D}$ 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 based on their complexity alone.

Following the discussion in Section 6, Figure 4 shows a plot of deviance residuals dr_i against leverages p_{Di} for each of the five models considered. The dashed lines marked on each plot are of the form $x^2 + y = c$ and points lying along such a parabola will each contribute an amount $\text{DIC}_i = c$ to the overall DIC for that model. For models 2–5, parabolas are marked at values of $c = 1$, 2 and 5, and any data point whose contribution $\text{DIC}_i > 2$ is labelled by its observation number. For model 1, parabolas are marked at $c = 1$, 10 and 50, since the size of the deviance resiudals and individual contributions to DIC are much larger. For clarity, only points for which $\text{DIC}_i > 10$ are marked by their observation number. Observations 55 and 56, the only counties with $y_i = 0$, are clearly identified as potential outliers under each of the random effects models 2–4, as

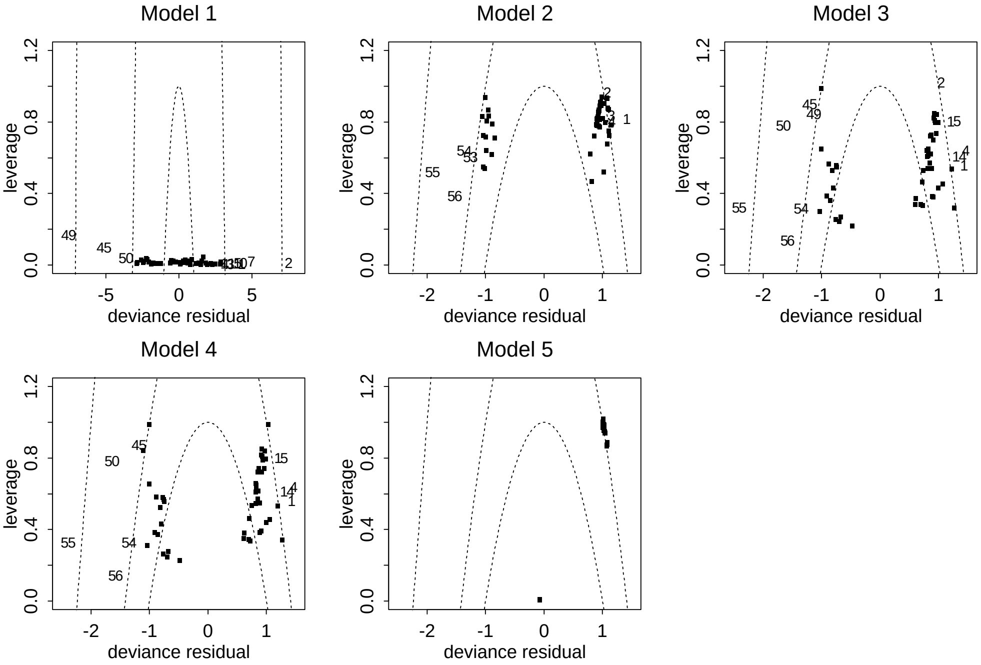

Figure 4: Diagnostics for lip cancer example: residuals vs leverages.

is observation 1 (the county with the highest observed risk ratio y_i/E_i). A few other observations (2, 3, 4, 53, 54) have contributions DIC$_i$ just larger than 2 under model 2: with the exception of the three counties already discussed, these five counties 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. Further investigation reveals that the observed risk ratios in these counties are extreme compared to those in each of their neighbouring counties. For example county 50 has only 6 cases compared to 19.6 expected, whilst each of its three neighbouring counties have high observed counts (17, 16, 16) relative to expected (7.8, 10.5, 14.4). The spatial prior in models 3 and 4 causes the estimated rate in county 50 to be smoothed towards the mean of its neighbours' rates, thus leading to the discrepancy between observed and fitted values. However since the observation still exercises considerable weight on its fitted value the leverage is high as well. Overall, we might not consider there is sufficient evidence to cast doubt on any particular observations.

8.2 Robust regression using the stack loss data

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

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

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

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

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

A well-known alternative to direct fitting of many symmetric but nonnormal error distributions is through scale mixtures of normals (Andrews and Mallows, 1974). From p.210 of Carlin and Louis (1996), we have the alternate t_d formulation

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

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

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

Table 2: Deviance results for stack loss data.

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 parameterisation and the canonical parameterisation are equivalent here, since the mean μ_i is a linear function of the canonical β parameters). Beginning with a comparison of the first four models, the estimates of p_D are all just over 5, the correct number of parameters for this example. The DIC values imply that Model 2 (double exponential) is best, followed by the t_4, the logistic, and finally the normal. Clearly this order is consistent with the models' respective abilities to accommodate outliers.

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

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

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

Table 3: Results for both parameterisations of Bernoulli panel data

8.3 Longitudinal binary observations: the Six Cities study

To illustrate how the mean and canonical parameterisations (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 $J = 4$ timepoints. We are given two predictor variables: a_{ij}, the age of child i in years at measurement point j (7, 8, 9, or 10 years), and s_i, the smoking status of child i's mother (1=yes, 0=no). Following the Bayesian analysis of Chib and Greenberg (1998), we adopt the conditional response model

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

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

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

Since the Bernoulli likelihood is unaffected by this choice, in all cases the 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, 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 parameterisations, respectively: the canonical parameterisation constructs $\overline{\theta}$ as the mean of the linear predictors β and b_i, and then uses the appropriate linking transformation (logit, probit, or cloglog) to obtain the imputed means for the p_{ij}. The mean parameterisation simply uses the means of the p_{ij} themselves when computing $D(\overline{\theta})$. 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 standard deviation of the random effects $\sqrt{\lambda^{-1}}$ is estimated to be 2.2 (standard deviation .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 4 for each of $I = 537$ individuals, and an average p_{D_i} for the canonical parameterisation of around .4 to .5. This indicates a prior sample size of around 4 to 6. Referring to the evidence in Figure 1 concerning low sample sizes, we would expect the mean parameterisation to display decreased complexity compared with the canonical, and this is borne out in the results. DIC prefers the cloglog link under the canonical parameterisation, but the probit link under the mean parameterisation. We repeat that we prefer the canonical results due to the improved normality of the likelihoods and their lack of dependence on observed data: however none of the models appear to 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 as to practical use of the techniques described in this paper.

9.1 Relationship of p_D and DIC to other suggestions

1. **Cross-validation:** Stone (1977) shows the asymptotic equivalence of model comparison based on cross-validation and AIC, while Wahba (1990)[p 52] shows how a generalised cross-validation criteria leads to the use of $n - \mathrm{tr}(H)$ as a denominator in 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.

2. **Other predictive loss functions:** Kass and Raftery (1995) criticise 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. Criteria based on this predictive distribution is also invariant to re-parameterisations. Laud and Ibrahim (1995) and Gelfand and Ghosh (1998) suggest minimising a predictive "discrepancy measure" $E[d(Y_{new}, y)|y]$, where Y_{new} is a draw from the posterior predictive distribution $p(Y_{new}|y)$, and we might for instance take $d(Y_{new}, y) = (Y_{new} - y)^T(Y_{new} - y)$. These authors show their measures also have attractive interpretations as weighted sums of "goodness of fit" and "predictive variability penalty" terms. However, proper choice of the criterion requires fairly involved analytic work, as well as several subjective choices about the utility function appropriate for the problem at hand. Furthermore, the one-way ANOVA model in Section 2.5 gives rise to a fit term equivalent to $D(\overline{\theta})$, and a predictive variability term equal to $p_D + p$. Thus their suggestion is equivalent in this context to comparison by $\overline{D}$ which, although invariant to parameterisation, does not seem to sufficiently penalize complexity.

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

3. **Bayes factors:** These are criteria based on comparison of the marginal likelihoods (0.1) (Kass and Raftery, 1995), and a common approximation is the Bayes (or Schwarz) information criterion(Schwarz, 1978) , which for a model with p parameters and n observations is given by BIC $= -2\log p(y|\hat{\theta}) + p\log n$. Bernardo and Smith (1994)[Chapter 6] argue that this formulation may only be appropriate in circumstances where it was really believed one and only one of the competing models were in fact 'true', and the crucial issue was to choose this

correct model, 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.

9.2 Practical issues in using DIC

1. **Invariance.** p_D may be only approximately invariant to the chosen parameterisation, since different fitted deviances $D(\overline{\theta})$ may arise from substituting posterior means of alternative choices of θ. The example in Section 8.3 shows 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 computational difficulty. In addition the approximate properties based on Taylor expansions in Section 3 may not hold, although this may be only of theoretical interest. Currently we recommend calculation of DIC based on a number of different estimators, with a preference for posterior means based on parameterisations obeying approximate likelihood normality.

2. **Focus of analyis.** As we see in the stacks example of Section 8.2, there may be sensitivity to apparent innocuous re-structuring of the model: this is to be expected since by making such changes one is altering the definition of a replicate dataset, 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, one 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 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 available in closed form.

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.

4. **What is an important difference in DIC?** Burnham and Anderson (1998) suggest 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 one would like to ensure that differences are not due to Monte Carlo error: while this is straightforward for $\overline{D}$, Zhu and Carlin (2000) explore the difficulty of assessing the Monte Carlo error on DIC.

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.

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 to 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 comparison.

Acknowledgements

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

References

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

Andrews, D. and Mallows, C. (1974). Scale mixtures of normality. *J Roy Statist Soc B*, **36**, 99–102.

Berk, R. H. (1966). Limiting behaviour of posterior distributions when the model is incorrect. *Annals of Mathematical Statistics*, **37**, 51–8.

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

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

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

Biller, C. and Fahrmeir, L. (2000). Bayesian varying-coefficient models using adaptive regression splines. Technical report, Institute of Statistics, Ludwig Maximilians University Munich.

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, **71**, 791–9.

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

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

Bunke, O. and Milhaud, X. (1998). Asymptotic behaviour of Bayes estimates under possibly incorrect models. *Annals Statistics*, **26**, 617–44.

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

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

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, **85**, 347–61.

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

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

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

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

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, **81**, 461–70.

Erkanli, A., Soyer, R., and Angold, A. (2001). Bayesian analyses of longitudinal binary data using markov regression models of unknown order. *Stat Med*, **20**, 755–70.

Erkanli, A., Soyer, R., and Costello, E. (1999). Bayesian inference for prevalence in longitudinal two-phase studies. *Biometrics*, **55**, 1145–50.

Eubank, R. (1985). Diagnostics for smoothing splines. *J Roy Statist Soc B*, **47**, 332–41.

Eubank, R. and Gunst, R. (1986). Diagnostics for penalized least-squares estimators. *Statistics Probability Letters*, **4**, 265–72.

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

G, K. and G, W. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann Math Statist*, **41**, 495–502.

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

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *J Roy Statist Soc B*, **56**, 501–14.

Gelfand, A. E., Ecker, M. D., Christiansen, C., MacLaughlkin, T. J., and Soumerai, S. B. (2000). Conditional categorical response models with application to treatment of acute myocardial infarction. *Applied Statistics*, **49**, 171–86.

Gelfand, A. E. and Trevisani, M. (2000). Inequalities between expected marginal log-likelihoods with implications for model comparison. Technical report, Department of Statistics, University of Connecticut, USA.

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

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

Good, I. J. (1956). The surprise index for the multivariate normal distribution. *Annals Mathematical Statistics*, **27**, 1130–5.

Green, P. and Richardson, S. (2000). Spatially correlated allocation models for count data. Technical report, Department of Mathematics, University of Bristol.

Han, C. and Carlin, B. (2001). Mcmc methods for computing bayes factors: A comparative review. *J Amer Statist Assn*, **96**, 000–000.

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

Hodges, J. and Sargent, D. (2001). Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika*, **88**, 367–79.

Huber, P. J. (1967). The behaviour of maximum likelihood estimates under non-standard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* **1**, (ed. L. M. LeCam and J. Neyman), pp. 221–33. University of California Press, Berkeley.

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

Key, J. T., Pericchi, L. R., and Smith, A. F. M. (1999). Bayesian model choice: what and why? In *Bayesian statistics 6*, (ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith), pp. 343–70. Oxford University Press, Oxford, UK.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, **22**, 79–86.

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

Laud, P. and Ibrahim, J. (1995). Predictive model selection. *J Roy Statist Soc B*, **57**, 247–62.

Lee, Y. and Nelder, J. (1996). Hierarchical generalised linear models (with discussion). *J Roy Statist Soc B*, **58**, 619–78.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *J Roy Statist Soc B*, **34**, 1–44.

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

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

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

Meng, X.-L. and Rubin, D. B. (1992). Performing likelihood ratio tests with multiply imputed data sets. *Biometrika*, **79**, 103–12.

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

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

Natarajan, R. and Kass, R. E. (2000). Reference bayesian methods for generalised linear mixed models. *J Amer Statist Assoc*, **95**, 227–37.

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

Rahman, N. J., Wakefield, J. C., Stephens, D. A., and Falcoz, C. (1999). The Bayesian analysis of a pivotal pharmacokinetic study. *Stat Methods Med Res*, **8**, 195–216.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with unknown number of components (with discussion). *J Roy Statist Soc B*, **59**, 731–92.

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

Sawa, T. (1978). Information criteria for choice of regression models: a comment. *Econometrika*, **46**, 1273–91.

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

SLATE, E. (1994). Parameterizations for natural exponential-families with quadratic variance functions. *"J Amer Statist Assoc"*, **89**, 1471–82.

Spiegelhalter, D. J., Thomas, A., and Best, N. G. (2000). *WinBUGS Version 1.3 User Manual.* MRC Biostatistics Unit, Cambridge, Available from `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).* MRC Biostatistics Unit, Cambridge.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *J Roy Statist Soc B*, **39**, 44–7.

Takeuchi, K. (1976). Distribution of informational statistics and a criterion for model fitting (in Japanese). *Suri-Kagaku (Mathematic Sciences)*, **153**, 12–8.

van der Linde, A. (1995). Splines from a Bayesian point of view. *Test*, **4**, 63–81.

van der Linde, A. (2000). Reference priors for shrinkage and smoothing parameters. *J Stat Planning Inf*, **90**, 245–74.

Vehtari, A. and Lampinen, J. (1999). Bayesian neural networks with correlating residuals. In *IJCNN'99: Proceedings of the 1999 International Joint Conference on Neural Networks.* IEEE.

Wahba, G. (1983). Bayesian confidence-intervals for the cross-validated smoothing spline. *J Roy Statist Soc B*, **45**, 133–50.

Wahba, G. (19878). Improper prirs, spline smoothing and the problem of guarding against model errors in regressions. *J Roy Statist Soc B*, **40**, 364–72.

Wahba, G. (1990). *Spline Models for Observational data.* SIAM, Philadelphia.

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

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

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

Zhu, L. and Carlin, B. (2000). Comparing hierarchical models for spatio-temporally misaligned data using the deviance information criterion. *Stat Med* , **19**, 2265–78.