Biostatistics (2008), **9**, 3, *pp.* 523–539
doi:10.1093/biostatistics/kxm049
Advance Access publication on January 21, 2008

Penalized loss functions for Bayesian model comparison

MARTYN PLUMMER

International Agency for Research on Cancer,
150 Cours Albert Thomas, 69372 Lyon Cedex 08, France
plummer@iarc.fr

SUMMARY

The deviance information criterion (DIC) is widely used for Bayesian model comparison, despite the lack of a clear theoretical foundation. DIC is shown to be an approximation to a penalized loss function based on the deviance, with a penalty derived from a cross-validation argument. This approximation is valid only when the effective number of parameters in the model is much smaller than the number of independent observations. In disease mapping, a typical application of DIC, this assumption does not hold and DIC under-penalizes more complex models. Another deviance-based loss function, derived from the same decision-theoretic framework, is applied to mixture models, which have previously been considered an unsuitable application for DIC.

Keywords: Bayesian model comparison; Deviance information criterion; Disease mapping; Markov chain Monte Carlo methods; Mixture models.

1. INTRODUCTION

Model choice is a fundamental part of data analysis, and the various attempts to formalize this activity have created a large and diverse literature. A considerable diversity of approaches can be found even within Bayesian discussions of model choice. The Bayes factor (Kass and Raftery, 1995), which quantifies the weight of evidence in favor of one model over another, is widely recognized as a formally correct solution to the model-choice problem. However, Bayes factors have some practical limitations: they are undefined when the model parameters are given improper prior distributions and are numerically unstable when proper, but diffuse, reference priors are used. Although some consider this as evidence against the use of reference priors, modifications to the Bayes factor have been proposed to overcome it (O'Hagan, 1995; Berger and Pericchi, 1996) by sacrificing a small fraction of the data for estimation of the model parameters and using the remainder for the calculation of the Bayes factor.

The notion of splitting the data between parameter estimation and assessment of model adequacy is used more explicitly in cross-validation (Geisser and Eddy, 1979). This is a utilitarian approach to model choice, in which a model is considered "useful" if, given a set of data, it makes accurate out-of-sample predictions. This goal may be contrasted with the search for the "true" model that motivates the use of Bayes factors.

A third approach to Bayesian model choice is based on hypothetical replicates from the same process that generated the data. In this posterior predictive approach, replicate data sets are simulated from the posterior distribution of the model parameters, and the adequacy of the model is assessed by the faithfulness of these replicates to the original data. The posterior predictive approach is recommended as a general framework for model criticism by Gelman *and others* (2002). Model-choice criteria in this framework have been proposed by Laud and Ibrahim (1995) and Gelfand and Ghosh (1998).

A further, more recent addition to the collection of Bayesian model-choice methods is the deviance information criterion (DIC) (Spiegelhalter *and others*, 2002), a Bayesian analogue of classical model-choice criteria, such as the Akaike information criterion (AIC). DIC combines a measure of model fit—the expected deviance—with a measure of model complexity—the effective number of parameters. DIC is simple to calculate using Markov chain Monte Carlo (MCMC) simulation and is routinely implemented in the WinBUGS software package (Spiegelhalter *and others*, 2004). It is now described in textbooks on Bayesian data analysis gelman:etal:2002,banerjee:etal:2004 and is widely used in papers on applied Bayesian statistics, particularly in biomedical applications. The Institute for Scientific Information web of knowledge currently lists 534 citations of Spiegelhalter *and others* (2002) and its precursor technical report (Spiegelhalter *and others*, 1998). Typical applications include lung cancer prediction (Clements *and others*, 2005), monitoring of depression (Elliott *and others*, 2005), monitoring health care quality (Daniels and Normand, 2006), and spatial analysis of cancer mortality (Liu *and others*, 2005).

Despite the practical advantages of DIC, its theoretical foundations remain controversial. Spiegelhalter *and others* (2002) provided a heuristic justification of DIC, which was further explored by van der Linde (2004, 2005). However, Celeux *and others* (2006b) have suggested that DIC lacks a natural generalization outside of exponential families. In addition, various *ad hoc* extensions or modifications to DIC have been proposed (Gelman *and others*, 2002; Plummer, 2002; Celeux *and others*, 2006a), none of them more convincing than DIC itself.

Given the conflict between applied and theoretical statistical perspectives on DIC, anyone wishing to use it as a model-choice tool faces a choice between hedging the use of DIC with a discussion of its potential limitations (see, e.g. Elliott *and others*, 2005) or trusting the expert judgment that "experience with DIC to date suggests that it works remarkably well" (Banerjee *and others*, 2004, p. 108).

The purpose of the current paper is to provide a more formal justification for DIC by demonstrating the link between DIC and cross-validation. This justification suggests some limitations on situations where DIC may be used, as well as providing improved asymptotic approximations. The plan of the paper is as follows. Section 2 sets up model comparison as a decision-theoretic problem and derives penalized loss functions using a cross-validation argument. In Sections 3 and 4, the theoretical properties of these loss functions are examined in normal linear models and exponential family models, respectively. Section 5 discusses the application of a particular penalized loss function in general models and relates the penalty to the "effective number of parameters." Sections 6 and 7 apply the penalized loss functions to examples in mixture modeling and disease mapping. Section 8 ends with a discussion.

2. Loss functions for model choice

Consider an idealized situation in which 2 independent data sets are available: a set of training data $\mathbf{Z}$ and a set of test data $\mathbf{Y} = \{Y_1, \dots, Y_n\}$. Suppose that model adequacy is assessed by a loss function $L(\mathbf{Y}, \mathbf{Z})$, which measures the ability of the model to make accurate predictions of $\mathbf{Y}$ from $\mathbf{Z}$. Given a set of candidate probability models for $\mathbf{Y}$ and $\mathbf{Z}$, we choose the one with the smallest loss. More pragmatically, we may choose a subset of models with losses close to the minimum for further consideration.

Suitable loss functions can be derived from scoring rules. A scoring rule measures the utility of a probabilistic forecast of $\mathbf{Y}$, represented by the probability density function $p^{\dagger}(\mathbf{y})$. A scoring rule is called "proper" if its expected value is maximized when $p^{\dagger}(\cdot)$ is the true probability density of $\mathbf{Y}$ and is "strictly proper" if this is the unique maximum. Strictly proper scoring rules are reviewed by Gneiting and Raftery (2007). Bernardo (1979) added the additional assumption of "locality"—the condition that the scoring rule depends only on the value of Y that actually obtains—and showed that the unique local, strictly proper scoring rule is the log-scoring rule

$$A\log\{p^{\dagger}(y)\} + B(y),$$

where A is a constant and $B(y)$ is an arbitrary function of y.

In this paper, we shall not consider the comparison of arbitrary models, but consider a more limited situation where all candidate models share a common vector of parameters $\boldsymbol{\theta}$—called the "focus" of inference by Spiegelhalter *and others* (2002)—and differ only in the prior structure imposed on $\boldsymbol{\theta}$. We also assume that $\mathbf{Y}$ and $\mathbf{Z}$ are conditionally independent given $\boldsymbol{\theta}$ so that $p(\mathbf{Y}|\boldsymbol{\theta}, \mathbf{Z}) = p(\mathbf{Y}|\boldsymbol{\theta})$. In this context, the log-scoring rule for Y becomes the log-likelihood of $\boldsymbol{\theta}$. Equivalently, when considered as a loss function rather than a utility, it becomes the deviance function $D(\boldsymbol{\theta}) = -2\log\{p(\mathbf{Y} \mid \boldsymbol{\theta})\}$.

We consider 2 loss functions based on the deviance: the "plug-in deviance"

$$L^{\mathrm{p}}(\mathbf{Y}, \mathbf{Z}) = -2\log[p\{\mathbf{Y} \mid \bar{\boldsymbol{\theta}}(\mathbf{Z})\}],$$

where $\bar{\boldsymbol{\theta}}(\mathbf{Z}) = E(\boldsymbol{\theta} \mid \mathbf{Z})$, and the "expected deviance"

$$L^{\mathrm{e}}(\mathbf{Y}, \mathbf{Z}) = -2\int \mathrm{d}\boldsymbol{\theta}\, p(\boldsymbol{\theta} \mid \mathbf{Z})\log\{p(\mathbf{Y} \mid \boldsymbol{\theta})\},$$

where the expectation is taken over the posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{Z}$, and the test data $\mathbf{Y}$ are considered fixed. Since the deviance is defined only up to an additive function of the data, both these loss functions are relative losses: their absolute value has no meaning, and only the difference in loss between 2 candidate models can be interpreted.

Although L^{p} and L^{e} are both derived from the deviance function, there are some important differences between them. The plug-in deviance is sensitive to reparameterization. Changing the coordinates of $\boldsymbol{\theta}$ changes the definition of the posterior expectation $\overline{\boldsymbol{\theta}}(\mathbf{Z})$, and hence the loss function L^{p}. The expected deviance, on the other hand, is coordinate free. The plug-in deviance gives equal loss to all models that yield the same posterior expectation of $\boldsymbol{\theta}$, regardless of the precision of this estimate. The expected deviance is a function of the full posterior of $\boldsymbol{\theta}$ given $\mathbf{Z}$, and therefore takes precision of the estimates into account.

2.1 *Penalized losses*

The problem is how to proceed when there are no training data $\mathbf{Z}$, and the test data $\mathbf{Y}$ must be used both to estimate $\boldsymbol{\theta}$ and to assess the adequacy of the model. The simplest solution is to reuse the observed data, creating a loss function $L(\mathbf{Y}, \mathbf{Y})$. We refer to this as the "exact replicate" form of the loss function. For example, the posterior Bayes factor proposed by Aitkin (1991) is equivalent to using the exact replicate form of the loss $L(\mathbf{Y}, \mathbf{Z}) = -\log\{p(\mathbf{Y} \mid \mathbf{Z})\}$. However, the posterior Bayes factor leads to inconsistent inference (Lindley, 1991).

In general, $L(\mathbf{Y}, \mathbf{Y})$ gives an optimistic assessment of model adequacy as it uses the same data twice: once for calculating the posterior distribution of the model parameters and again in place of new test data. The degree of optimism can be estimated for loss functions that decompose into the sum of contributions from each individual Y_i:

$$L(\mathbf{Y}, \mathbf{Z}) = \sum_{i=1}^{n} L(Y_i, \mathbf{Z}). \tag{2.1}$$

This condition is satisfied by deviance-based loss functions when $Y_1, \ldots, Y_n$ are conditionally independent given $\boldsymbol{\theta}$. For such loss functions, the extent to which $L(Y_i, \mathbf{Y})$ overstates the model adequacy can be assessed by comparing it with the cross-validation loss $L(Y_i, \mathbf{Y}_{-i})$, where $\mathbf{Y}_{-i}$ denotes the set of observations $\{Y_1, \ldots, Y_n\}$ with Y_i removed. The expected decrease in loss from using $L(Y_i, \mathbf{Y})$ in place of

$L(Y_i, \mathbf{Y}_{-i})$ is

$$p_{\text{opt}_i} = E\{L(Y_i, \mathbf{Y}_{-i}) - L(Y_i, \mathbf{Y}) \mid \mathbf{Y}_{-i}\}. \tag{2.2}$$

Following the terminology of Efron (1983), we refer to p_{opt_i} as the "optimism" of $L(Y_i, \mathbf{Y})$. The penalized loss function

$$L(Y_i, \mathbf{Y}) + p_{\text{opt}_i}$$

has the same expectation, given $\mathbf{Y}_{-i}$, as the cross-validation loss $L(Y_i, \mathbf{Y}_{-i})$. The 2 loss functions are therefore equivalent to an observer who has not seen Y_i.

The same argument can be applied to each observation Y_i in turn. We propose to use the sum of the penalized loss functions to assess model adequacy. This gives $L(\mathbf{Y}, \mathbf{Y}) + p_{\text{opt}}$ where the total optimism $p_{\text{opt}} = \sum_i p_{\text{opt}_i}$ is a rational cost that must be paid for using the data $\mathbf{Y}$ twice.

2.2 *The DIC*

The DIC is defined by Spiegelhalter *and others* (2002) as

$$\text{DIC} = \overline{D} + p_D,$$

where $\overline{D} = E(D \mid \mathbf{Y})$ is considered to be a measure of model fit, and p_D is the "effective number of parameters," a measure of model complexity, defined by

$$p_D = \overline{D} - D\{\bar{\theta}(\mathbf{Y})\}.$$

In the loss function notation of this paper, $\overline{D} = L^{\text{e}}(\mathbf{Y}, \mathbf{Y})$ and $D\{\bar{\theta}(\mathbf{Y})\} = L^{\text{p}}(\mathbf{Y}, \mathbf{Y})$. Hence, p_D can be rewritten as

$$p_D = L^{\text{e}}(\mathbf{Y}, \mathbf{Y}) - L^{\text{p}}(\mathbf{Y}, \mathbf{Y}).$$

Moreover, p_D can be decomposed into the sum of individual contributions $p_D = \sum_i p_{D_i}$, where

$$p_{D_i} = L^{\text{e}}(Y_i, \mathbf{Y}) - L^{\text{p}}(Y_i, \mathbf{Y}).$$

3. Normal linear models

Normal linear models provide an opportunity to express the penalties p_D and p_{opt} in closed form and so gain a heuristic understanding of the behavior of the penalized loss functions and their relationship to DIC. Following Spiegelhalter *and others* (2002, Section 4.1), we consider the hierarchical linear model of Lindley and Smith (1972)

$$\mathbf{Y} \mid \boldsymbol{\theta} \sim N(A_1\boldsymbol{\theta}, C_1),$$

$$\boldsymbol{\theta} \mid \boldsymbol{\psi} \sim N(A_2\boldsymbol{\psi}, C_2),$$

where, in this context, $\mathbf{Y}$ is a vector of observations, the matrices A_1, A_2, C_1, C_2 are known, and $\boldsymbol{\psi}$ is a vector of hyperparameters which may be either fixed or unknown with an improper flat hyperprior.

Suppose that $\mathbf{Y}$ can be broken down into subvectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ that are conditionally independent given $\boldsymbol{\theta}$. The covariance matrix C_1 is thus block diagonal with blocks $C_{11}, \ldots, C_{1n}$, and the matrix A_1 can be decomposed into corresponding row blocks $A_{11}, \ldots, A_{1n}$. Let $V = \text{Var}(\boldsymbol{\theta} \mid \mathbf{Y})$ and $V_{-i} = \text{Var}(\boldsymbol{\theta} \mid \mathbf{Y}_{-i})$. The 2 variances are related by

$$V^{-1} = V_{-i}^{-1} + A_{1i}^T C_{1i}^{-1} A_{1i}. \tag{3.1}$$

Using the plug-in deviance L^{p} as a loss function, the optimism for observation i is

$$p_{\mathrm{opt}_i} = \mathrm{Tr}(C_{1i}^{-1} A_{1i} V A_{1i}^T) + \mathrm{Tr}(C_{1i}^{-1} A_{1i} V_{-i} A_{1i}^T).$$

This expression can be rewritten in terms of the "hat" matrix $H = C_1^{-1} A_1 V A_1^T$, which projects the data $\mathbf{Y}$ onto the fitted values. The importance of the hat matrix was highlighted by Spiegelhalter *and others* (2002), who showed that $p_D = \mathrm{Tr}(H)$. With the additional assumption of independence of $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$, the contribution of Y_i to p_D can be written as $p_{D_i} = \mathrm{Tr}(H_i)$, where $H_i = C_{1i}^{-1} A_{1i} V A_{1i}^T$. Using the relation (3.1) between V and V_{-i}, the optimism for observation i can be rewritten as

$$p_{\mathrm{opt}_i} = \mathrm{Tr}(H_i) + \mathrm{Tr}\{(I - H_i)^{-1} H_i\}.$$

Summing over i, the penalized plug-in deviance is

$$D\{\bar{\boldsymbol{\theta}}(\mathbf{Y})\} + p_D + \sum_{i=1}^{n} \mathrm{Tr}\{(I - H_i)^{-1} H_i\},$$

or equivalently

$$\overline{D} + \sum_{i=1}^{n} \mathrm{Tr}\{(I - H_i)^{-1} H_i\}.$$

A similar argument shows that the penalized expected deviance can be written as

$$\overline{D} + 2\sum_{i=1}^{n} \mathrm{Tr}\{(I - H_i)^{-1} H_i\}.$$

Both penalized loss functions can therefore be written in a form that can be compared directly with DIC.

3.1 *Large sample behavior*

When $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are all scalar, the penalty term $\mathrm{Tr}\{(I - H_i)^{-1} H_i\}$ simplifies to $p_{D_i}/(1 - p_{D_i})$. The large sample behavior of this penalty depends on the dimension of $\boldsymbol{\theta}$. For regular models, in which the dimension of $\boldsymbol{\theta}$ is fixed, p_{D_i} is $\mathrm{O}(n^{-1})$. The penalized losses can then be written as

$$\overline{D} + k p_D + \mathrm{O}(n^{-1}),$$

where $k = 1$ for the plug-in deviance and $k = 2$ for the expected deviance. The penalized plug-in deviance is therefore asymptotically equivalent to DIC in regular linear models with a scalar outcome.

In random-effect models, it is quite common for the dimension of $\boldsymbol{\theta}$ to increase with n. In this case, the behavior of the penalized plug-in deviance may be very different from DIC. This can be illustrated with a 1-way analysis of variance (ANOVA) model

$$Y_i \mid \theta_i \sim N(\theta_i, \tau_i^{-1}),$$

$$\theta_i \mid \psi \sim N(\psi, \lambda^{-1}),$$

where the precision parameters $\tau_1, \ldots, \tau_n$ are fixed, and ψ is given a flat hyperprior. Consider a situation in which the candidate models are indexed by λ, and the deviance is defined as

$$D(\boldsymbol{\theta}) = \sum_{i=1}^{n} (y_i - \theta_i)^2 \tau_i.$$

As shown by Spiegelhalter *and others* (2002), the contribution to the effective number of parameters from observation i is

$$p_{D_i} = \rho_i + \frac{\rho_i(1-\rho_i)}{\sum_{j=1}^{n} \rho_j},$$

where $\rho_i = \tau_i/(\lambda + \tau_i)$ is the intraclass correlation coefficient.

There are 2 limiting situations of interest. In the limit $\lambda \to \infty$, the ANOVA model tends to a pooled model, in which all observations have the same prior mean ψ. In this limit, $p_D = 1$, and both DIC and the penalized plug-in deviance are equal to $\sum_i \tau_i(Y_i - \overline{Y})^2 + 2$, where $\overline{Y} = \sum_i \tau_i Y_i \big/ \sum_j \tau_j$. In the opposite limit $\lambda \to 0$, the ANOVA model tends to a saturated fixed-effects model, in which $\mathbf{Y}_{-i}$ contains no information about the mean of Y_i. In this limit, $p_D = n$ and DIC $= 2n$, but the penalized plug-in deviance tends to infinity.

For linear models with scalar outcomes, therefore, DIC is a good approximation to the penalized plug-in deviance whenever $p_{D_i} \ll 1$ for all i. A necessary condition for this condition to hold is $p_D \ll n$. The ratio p_D/n may therefore be used as an indicator of the validity of DIC in such models.

4. Exponential family models

In an exponential family model, the log probability density takes the form

$$\log\{p(y_i \mid \theta_i)\} = \{y_i\theta_i - b(\theta_i)\}\varphi^{-1} - c(y_i, \varphi).$$

The density can also be written in terms of the mean value parameter $\mu_i = E(Y_i \mid \theta_i)$.

Using the plug-in deviance as a loss function, and assuming that the scale parameter φ is known, the optimism for observation i is

$$p_{\text{opt}_i} = 2\varphi^{-1}\text{Cov}(\theta_i, \mu_i \mid \mathbf{Y}_{-i}) - p_{D_i}(\mathbf{Y}_{-i}) + E\left\{p_{D_i}(\mathbf{Y}) \mid \mathbf{Y}_{-i}\right\}. \tag{4.1}$$

Here, the dependency of p_{D_i} on the data, which is usually ignored, is made explicit by writing it as $p_{D_i}(\mathbf{Y})$, and $p_{D_i}(\mathbf{Y}_{-i})$ denotes the effective number of parameters evaluated using the posterior distribution of $\boldsymbol{\theta}$ given the partial data $\mathbf{Y}_{-i}$ instead of the full data $\mathbf{Y}$.

Further simplification is possible noting that the observed value of $p_{D_i}(\mathbf{Y})$ is an unbiased estimate of the last term in (4.1). Making this substitution, the penalized plug-in deviance can be expressed approximately as

$$\overline{D} + 2\varphi^{-1}\sum_{i=1}^{n}\text{Cov}(\theta, \mu_i \mid \mathbf{Y}_{-i}) - p_{D_i}(\mathbf{Y}_{-i}). \tag{4.2}$$

Calculations for the expected deviance L^{e} are simpler and do not require any approximations. The penalized expected deviance is

$$\overline{D} + 2\varphi^{-1}\sum_{i=1}^{n}\text{Cov}(\theta_i, \mu_i \mid \mathbf{Y}_{-i}). \tag{4.3}$$

By taking a quadratic expansion of $b(\cdot)$ around $\theta_i = E(\theta_i \mid \mathbf{Y}_{-i})$, it can be shown that

$$p_{D_i}(\mathbf{Y}_{-i}) \approx \varphi^{-1}\text{Cov}(\theta_i, \mu_i \mid \mathbf{Y}_{-i}).$$

Hence, when the penalized loss functions are written in terms of $\overline{D}$, the penalty is approximately twice the size in (4.3) as in (4.2).

Efron (1986) considered the error of prediction rules in exponential family models from a repeated sampling perspective. He derived an expression for the optimism of the plug-in deviance:

$$2\varphi^{-1}E\left\{\sum_{i=1}^{n}(y_i-\mu_i)\overline{\theta}_i \mid \boldsymbol{\theta}\right\},$$

which may be seen as an empirical estimate of $2\varphi^{-1}\sum_i \mathrm{Cov}(Y_i, \overline{\theta}_i \mid \boldsymbol{\theta})$, and is thus a close frequentist analogue of p_{opt}.

5. The effective number of parameters

Outside of the normal linear and exponential family models considered in Sections 3 and 4, the optimism of the plug-in deviance is hard to calculate. It is based on the distribution of the plug-in estimator $\overline{\boldsymbol{\theta}}(\mathbf{Y})$ given the partial data $\mathbf{Y}_{-i}$ which is, in general, intractable. However, the optimism of the expected deviance can be estimated in general models using MCMC methods. It also has some useful information-theoretic properties.

Given 2 density functions $p(x)$ and $q(x)$ for a random variable X, let

$$I(p,q)=\int \mathrm{d}x p(x)\log\left\{\frac{p(x)}{q(x)}\right\}$$

denote the Kullback–Leibler information divergence between p and q, and let

$$J(p,q)=I(p,q)+I(q,p)$$

denote the undirected divergence. The difference in loss between the cross-validation and exact replicate forms of L^{e} can be expressed in terms of J.

$$L^{\mathrm{e}}(Y_i,\mathbf{Y}_{-i})-L^{\mathrm{e}}(Y_i,\mathbf{Y})=J\{p(\theta \mid \mathbf{Y}_{-i}),\, p(\theta \mid \mathbf{Y})\}.$$

An immediate corollary is that $L^{\mathrm{e}}(Y_i,\mathbf{Y})\leqslant L^{\mathrm{e}}(Y_i,\mathbf{Y}_{-i})$. Thus, the exact replicate form of the expected deviance is always optimistic in its assessment of model adequacy.

An alternative representation of p_{opt} uses the predictive distribution $p(y_i^{\text{rep}} \mid \boldsymbol{\theta})$ where y_i^{rep} is a hypothetical replicate of observation y_i for fixed $\boldsymbol{\theta}$. Let

$$J_i(\boldsymbol{\theta},\boldsymbol{\theta}')=J\{p(y_i^{\text{rep}} \mid \boldsymbol{\theta}),\, p(y_i^{\text{rep}} \mid \boldsymbol{\theta}')\}.$$

Then, the optimism of L^{e} can be written as

$$p_{\text{opt}_i}=\int \mathrm{d}\boldsymbol{\theta}\int \mathrm{d}\boldsymbol{\theta}' p(\boldsymbol{\theta} \mid \mathbf{Y}_{-i})p(\boldsymbol{\theta}' \mid \mathbf{Y}_{-i})J_i(\boldsymbol{\theta},\boldsymbol{\theta}'). \tag{5.1}$$

This suggests a way of estimating p_{opt} by MCMC methods. Consider 2 parallel chains, both with stationary distribution $p(\boldsymbol{\theta} \mid \mathbf{Y}_{-i})$. Let $\boldsymbol{\theta}^c$ be a sample of $\boldsymbol{\theta}$ from Markov chain c, and let Y_i^{repc} be a replicate of Y_i simulated from the density $p(y_i \mid \boldsymbol{\theta}^c)$. Then, p_{opt_i} can be estimated by taking the sample mean of

$$\widehat{p}_{\text{opt}_i}=\log\left\{\frac{p(y^{\text{rep1}} \mid \boldsymbol{\theta}^1)}{p(y^{\text{rep1}} \mid \boldsymbol{\theta}^2)}\right\}+\log\left\{\frac{p(y^{\text{rep2}} \mid \boldsymbol{\theta}^2)}{p(y^{\text{rep2}} \mid \boldsymbol{\theta}^1)}\right\}$$

over parallel iterations of the 2 chains.

This method of estimation has the disadvantage of requiring n separate MCMC runs, with a single observation deleted in each run. Since MCMC methods are typically time consuming, this may be infeasible. An alternative is to use importance sampling, taking samples from the full posterior of $\boldsymbol{\theta}$ given $\mathbf{Y}$ and calculating the weighted mean of $\widehat{p}_{\text{opt}_i}$ with relative weights

$$w(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2) = \frac{1}{p(y_i \mid \boldsymbol{\theta}^1)} \frac{1}{p(y_i \mid \boldsymbol{\theta}^2)}.$$

If $p(\boldsymbol{\theta} \mid \mathbf{y})$ differs substantially from $p(\boldsymbol{\theta} \mid \mathbf{y}_{-i})$, then the importance weights are unstable and the variance of this estimator is inflated. It may even have infinite variance when there are highly influential observations in the data set (Peruggia, 1997). Conversely, importance sampling is efficient when Y_i does not have a strong influence on the posterior of $\boldsymbol{\theta}$. For regular models, the influence becomes vanishingly small as $n \rightarrow \infty$. In this case, the sample variance of the weights $w(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2)$ tends to 0 and p_{opt_i} can be estimated from the full posterior without reweighting the samples. This approximation gives $p_{\text{opt}} \approx 2p_D^*$, where p_D^* is an alternative definition of the "effective number of parameters" proposed by Plummer (2002).

This approximation assumes that each observation has negligible influence on the posterior of $\boldsymbol{\theta}$ and cannot be used outside of this asymptotic situation. For example, Plummer (2006) applied p_D^* as an *ad hoc* penalty in mixture models and found that p_D^* may increase or decrease when more components are added to the mixture. This counterintuitive result may be explained by the fact that $2p_D^*$ is a poor approximation to p_{opt} in this context.

6. Application to mixture models

Finite mixture models present a severe challenge to DIC as a model-choice criterion, as noted by several contributors to the discussion of Spiegelhalter *and others* (2002). In such models, the posterior expectation is not a suitable plug-in estimate for the model parameters since it lies in between multiple modes of the posterior density, and alternative plug-in estimators are hard to define. Celeux *and others* (2006a) applied no less than eight variations of DIC to this problem, but finally were unable to recommend any of them, concluding that DIC was neither a well-defined criterion nor a solution to a well-defined optimization problem.

Clearly, the difficulties in defining a plug-in estimate rule out the use of L^{p} as a loss function. We therefore consider the penalized expected deviance L^{e} as a loss function in this problem, using an example considered by Richardson and Green (1997). This example concerns an experiment by Bechtel *and others* (1993) on 245 unrelated subjects to determine their acetylator phenotype by measuring the ratio of 2 urinary metabolites after oral administration of caffeine. The acetylator phenotype is determined by sequence variants in the NAT2 gene, which has been associated with cancer risk at several sites in the body (Hein *and others*, 2000). Figure 1 shows a histogram of the ratio of the 2 urinary metabolites (AFMU:1X) on a log scale. The bimodal distribution suggests that the subjects can be classified into 2 subpopulations of "fast" and "slow" acetylators.

These data can be represented by a normal mixture model in which Y_i is the value of $\log(\text{AFMU}/1\text{X})$ for subject i. In a normal mixture model with G components, the population is composed of G groups with means $\mu_1, \ldots, \mu_G$ and variances $\sigma_1^2, \ldots, \sigma_G^2$, respectively. The observation Y_i is drawn from group g with prior probability π_g. This model can be expressed as

$$p(y_i \mid x_i, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \phi\left(\frac{y_i - \mu_{x_i}}{\sigma_{x_i}}\right),$$

$$P(X_i = x \mid \boldsymbol{\pi}) = \pi_x,$$

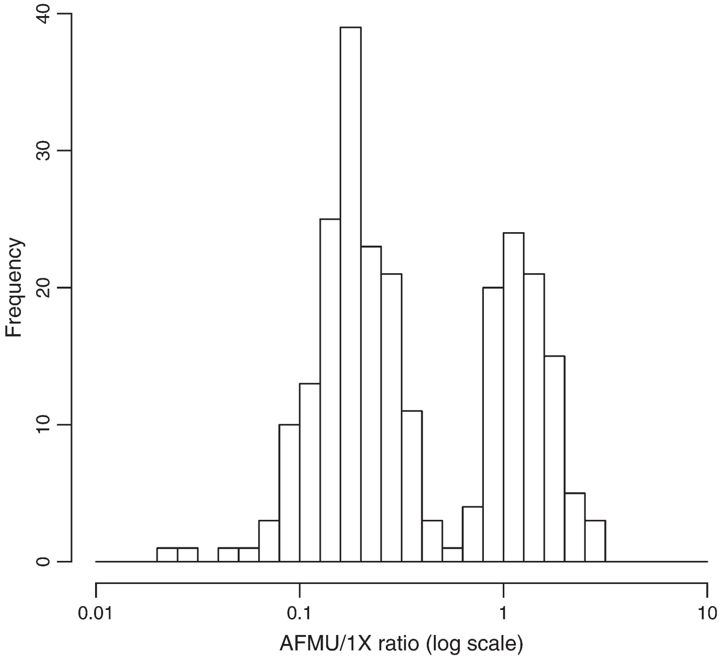

Fig. 1. Histogram of the ratio of 2 urinary metabolites in 245 subjects after oral administration of caffeine, from an assay designed to distinguish fast and slow acetylator phenotypes.

where X_i is a latent indicator variable for the group to which individual i belongs, and $\phi(\cdot)$ is the density function of a standard normal distribution.

When using MCMC methods, it is more efficient to eliminate X_i by marginalization and define the mixture model as

$$p(y_i \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \sum_{g=1}^{G} \pi_g \phi\left(\frac{Y_i - \mu_g}{\sigma_g}\right). \tag{6.1}$$

Richardson and Green (1997) considered the number of components G to be an unknown parameter in the model and determined the posterior distribution of G using reversible jump MCMC. In the current setting, there are several candidate models, each with a different fixed value of G. The aim is to choose the value (or values) of G required to accurately describe the distribution of Y.

There are 2 possible levels of focus in mixture models. The parameter in focus $\boldsymbol{\theta}$ can be defined either as $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}\}$ or as $\boldsymbol{\theta} = \{\mathbf{X}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}\}$. In both cases, the condition that $Y_1, \ldots, Y_n$ are conditionally independent given $\boldsymbol{\theta}$ is satisfied. In the latter case, interest lies not only in determining the population-level parameters but also in accurately classifying the group to which each individual belongs.

Mixture models with 2–5 components were fitted to the acetylator phenotype data using JAGS, an alternative MCMC engine for the BUGS language. The "mix" module of JAGS allows the normal mixture model (6.1) to be defined in the BUGS language and uses a random-walk Metropolis–Hastings sampler with tempered transitions to jump between the multiple modes of the posterior density (Neal, 1996; Celeux *and others*, 2000). Samples were drawn from 20 000 iterations after a burn-in period of 20 000 iterations.

The weakly informative, data-dependent prior of Richardson and Green (1997) was used with one modification. The prior distribution of $\boldsymbol{\pi}$ was defined as a Dirichlet($\boldsymbol{\delta}$) distribution with $\delta_g = 5$ for $g = 1, \ldots, G$. Compared with the value $\delta_g = 1$ used by Richardson and Green (1997), this prior encodes the same information as directly observing X_i on $4G$ individuals and finding 4 of them in each group. This prior improves the numerical stability of the MCMC estimates by preventing π_g from becoming stuck at a value close to 0, a phenomenon that effectively collapses the mixture model into one with fewer than G components and confounds the distinction between the different candidate models. Although this prior is clearly informative, the maximum number of pseudo-observations it encodes is 20, which is substantially smaller than the sample size of 245.

The optimism of L^{e} was calculated by importance sampling as detailed in Section 5. When $\mathbf{X}$ is in focus, the posterior density of $\boldsymbol{\theta} = \{\mathbf{X}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}\}$, given the partial data $\mathbf{Y}_{-i}$, factorizes

$$p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}, X_i \mid \mathbf{Y}_{-i}) = p(X_i \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}) p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma} \mid \mathbf{Y}_{-i}).$$

This factorization was exploited, when calculating the estimator $\widehat{p}_{\mathrm{opt}_i}$, by drawing samples of X_i from its prior density $p(X_i \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$. Using this *post hoc* simulation of X_i, the importance weights are the same as when focus is on $\{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}\}$.

Table 1 shows the expected deviance $\overline{D}$, the optimism p_{opt}, and the penalized expected deviance $\overline{D} + p_{\mathrm{opt}}$ for models with 2–5 components. When the focus did not include $\mathbf{X}$, the expected deviance was identical (within Markov chain error) for models with 3–5 components. This reflects a masking phenomenon: when additional components beyond 3 were added to the model, they adopted a similar mean and standard deviation to one of the existing 3 components. The model with 3 components had the smallest penalized loss. Despite the obvious bimodal distribution in Figure 1, the addition of a third component improved the model adequacy by accounting for the small number of outliers with low values of log(AFMU/IX). Such outliers are a typical artifact of the log transformation.

When the focus included $\mathbf{X}$, the expected deviance decreased with increasing number of components, due to the decrease in residual variation in Y as the number of components is increased. The optimism of $L^{\mathrm{e}}(\mathbf{Y}, \mathbf{Y})$ was very large in all models, and the model with 2 components was strongly preferred. When $\mathbf{X}$ was in focus, the addition of redundant mixture components was not harmless, as it greatly increased the classification error, leading to a steep increase in the optimism.

The last column of Table 1 shows the difference Δ in penalized loss with respect to the optimal model (model 2 or 3 depending on the focus). This difference Δ can be decomposed into individual contributions $\Delta = \sum_i \Delta_i$. The standard error of Δ under repeated sampling can be estimated as $\sqrt{n}$ times the sample standard deviation of $\Delta_1, \ldots, \Delta_n$. This standard error is shown in brackets.

When focus is on $\{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}\}$, the difference $\Delta = 6.1$ between the models with 2 and 3 components is less than its standard error (9.0) under repeated sampling. Thus, either model 2 or 3 might appear to

Table 1. *Expected deviance $\overline{D}$, optimism p_{opt}, and penalized expected deviance for the acetylator phenotype data. Δ is the difference in penalized deviance with respect to the optimal model and SE(Δ) is its standard error*

Level of focus	Number of components	$\overline{D}$	p_{opt}	$\overline{D} + p_{\mathrm{opt}}$	Δ	SE(Δ)
$\{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}\}$	2	594.1	9.5	603.6	6.2	9.0
	3	583.5	14.1	597.5	0.0	
	4	583.3	16.3	599.5	2.0	0.7
	5	583.4	18.0	601.3	3.8	1.2
$\{\mathbf{X}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}\}$	2	278.0	2995.0	3273.0	0.0	
	3	245.7	3328.9	3574.6	301.6	17.8
	4	231.8	3544.3	3776.1	503.1	14.5
	5	220.1	3869.5	4089.6	816.7	15.3

be optimal in a hypothetical repetition of the experiment in the same population. In all other cases, Δ is much larger than its standard error, suggesting stability of Δ under repeated sampling.

7. Application to disease mapping

A disease map shows the spatial distribution of some summary measure of a given disease. When mapping the incidence of rare diseases in small geographic areas, the incidence-rate estimates displayed in the map may be unstable due to small case counts within each area. Clayton and Kaldor (1987) proposed a solution to this instability using a model in which the underlying rates are spatially smooth. They applied this approach to lip cancer incidence among males in the 56 counties of Scotland in the years 1975–1980 (Kemp *and others*, 1985). The example was revisited by Breslow and Clayton (1993) and was considered as a model-choice example by Spiegelhalter *and others* (2002).

Spatial models have proven a popular application of DIC. For example, Banerjee *and others* (2004) make extensive use of DIC for model selection. It is therefore useful to compare the behavior of DIC with a penalized loss function. Since the purpose of disease mapping is to make accurate point predictions of the disease rates in each area, it is natural to use the plug-in deviance, which depends only on the point estimates, as a model-choice criterion.

In the lip cancer example, Y_i represents the number of cases observed in county i. A generalized linear mixed model (GLMM) is used for Y_i with Poisson family, log link, and linear predictor

$$\log(\mu_i) = \alpha_0 + \gamma_i + \delta_i + \log(E_i), \tag{7.1}$$

where α_0 is a fixed effect with an improper flat prior; $\gamma_1, \dots, \gamma_n$ are uncorrelated random effects; $\delta_1, \dots, \delta_n$ are spatially correlated random effects with a conditional autoregressive prior (Besag *and others*, 1991), and constraint $\sum_i \delta_i = 0$ for identifiability; and $E_1, \dots, E_n$ are the expected numbers of cases given the age structure in each county. It is customary to treat $E_1, \dots, E_n$ as if they were constants derived from external data, even though they are derived from the national cancer rates of Scotland, so that $\sum_i Y_i \approx \sum_i E_i$.

Four variations on model (7.1) were considered as candidate models. Each model included the fixed effect α_0 and offset $\log(E_i)$, but the models varied by the presence or absence of random effects:

1. A “pooled” model with no random effects.
2. An “exchangeable” model with only uncorrelated random effects (γ).
3. A “spatial” model with only autocorrelated random effects (δ).
4. The “full” model with both exchangeable and spatial random effects (γ, δ).

All calculations were carried out using WinBUGS (Spiegelhalter *and others*, 2004), using samples drawn on 15 000 iterations, following a burn-in of 5000 iterations. The optimism p_{opt} was calculated from $n = 56$ runs with each value of Y_i deleted in turn. The intercept parameter α_0 was given an improper flat prior, and the precision parameters for the random effects were given gamma(0.5, 0.0005) priors.

Table 2 shows the effective number of parameters p_D in each of the 4 models. To improve comparability with p_D, the optimism p_{opt} is expressed as the “residual optimism” r_{opt}, defined by

$$L^{\text{p}}(\mathbf{Y}, \mathbf{Y}) + p_{\text{opt}} = \overline{D} + r_{\text{opt}}.$$

Table 2 shows r_{opt} calculated with the approximate formula (4.2). Although r_{opt} and p_D are both close to 1 for the pooled model, their concordance is poor for the other models, with the residual penalty r_{opt} being much larger than the effective number of parameters p_D.

This example is sufficiently small ($n = 56$) for r_{opt} to be estimated from n MCMC runs with each observation left out in turn. But such detailed calculation is not feasible in general. Therefore, Table 2 also

Table 2. *Penalties for the Scottish lip cancer incidence data. These are added to $\overline{D}$ to obtain, respectively, DIC (p_D), the penalized plug-in deviance (r_{opt}), and approximate penalized plug-in deviance ($A1$, $A2$)*

Model	p_D	r_{opt}	$A1$	$A2$
Pooled	1.0	1.1	1.1	
Exchangeable	43.5	570.5	305.1	574.6
Spatial	31.0	163.9	119.8	159.3
Exchangeable + spatial	31.6	166.4	124.5	162.9

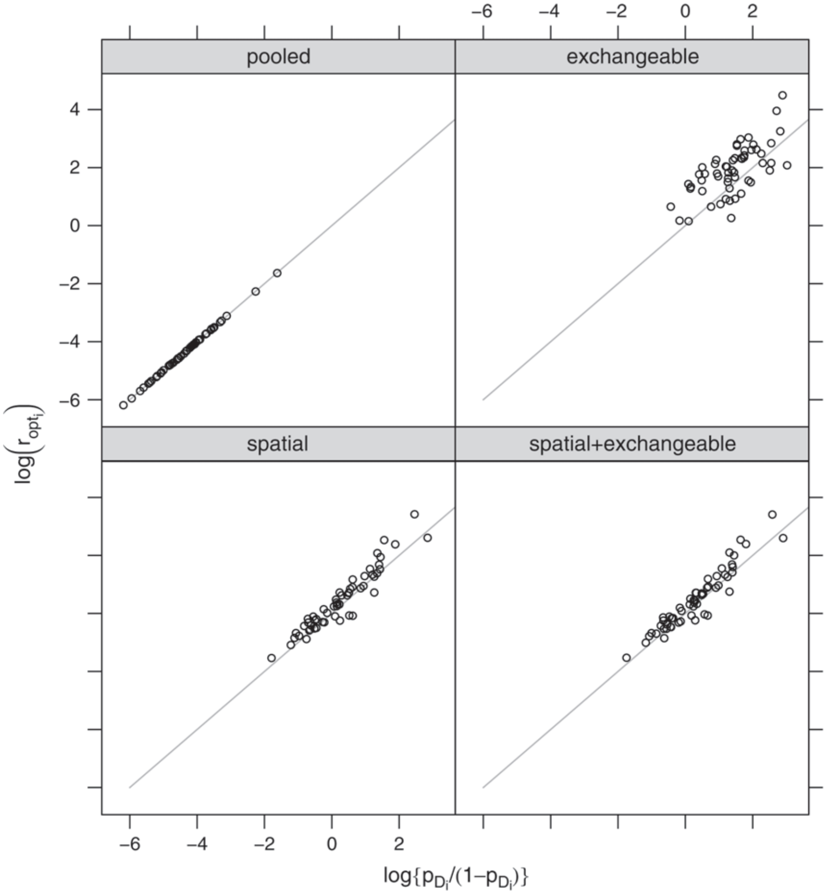

Fig. 2. Correlation between contributions to the residual optimism r_{opt} and the effective number of parameters p_D from each county in 4 separate models applied to the Scottish lip cancer data. The diagonal line marks equality of the 2 axes.

shows 2 approximations to the residual penalty r_{opt} that might be more generally applicable. Approximation $A1$ is based on the observation that the Wald approximation to the log-likelihood for a GLMM with canonical link has the same structure as the linear model of Section 3 (Spiegelhalter *and others*, 2002) and therefore the relation $r_{\text{opt}} \approx \sum_i p_{D_i}/(1 - p_{D_i})$, derived from the linear model, may hold asymptotically. The accuracy of this approximation is poor, although it does give values closer to the true penalties than p_D.

Figure 2 examines approximation $A1$ in more detail, showing the contributions r_{opt_i} and $p_{D_i}/(1-p_{D_i})$ for $i = 1, \ldots, n$ in all 4 models. Although the relation $\log(r_{\text{opt}_i}) \sim \log\{p_{D_i}/(1-p_{D_i})\}$ holds broadly, it is only precise for the pooled model, and is particularly poor for the exchangeable model.

Approximation $A2$ uses replication of the random effects to obtain an approximation to the partial posterior of θ_i given $\mathbf{Y}_{-i}$ from a single MCMC run as suggested by Marshall and Spiegelhalter (2003). The last column of Table 2 shows that approximation $A2$ works well in this example.

8. Discussion

If DIC is regarded as an approximation to the penalized plug-in deviance, then this perspective imposes some important restrictions on its use. A necessary assumption for the penalized plug-in deviance, and hence DIC, is that the data can be broken down into components that are conditionally independent given the parameters in focus. A second important assumption is that the effective number of parameters p_D must be small in relation to the sample size n. When this assumption does not hold, $2p_D$ is a poor approximation to the optimism of the plug-in deviance, and DIC under-penalizes complex models. The disease-mapping example shows that it is easy to construct a mixed model for which p_D is the same order as n.

These observations are not new. The asymptotic justification of DIC using a cross-validation argument is a Bayesian analogue of the result of Stone (1977) on the equivalence of cross-validation and the AIC. Indeed, in the discussion of Spiegelhalter *and others* (2002), Stone (2002) emphasized the importance of the independence assumption for DIC. The poor empirical performance of DIC compared to cross-validation when the assumption $p_D \ll n$ does not hold was highlighted by Vehtari and Lampinen (2002). From a frequentist perspective, AIC is also known to under-penalize models in small samples. Burnham and Anderson (2000) recommended the routine use of AIC_{c}, a small sample correction to AIC proposed by Hurvich and Tsai (1989), which imposes a higher penalty than AIC unless the number of parameters is much smaller than the sample size.

In view of these limitations, some practical advice is required on what to do when DIC cannot be justified as approximation to the penalized plug-in deviance. One solution would be to define DIC$_\text{c}$, the corrected DIC, as

$$\text{DIC}_\text{c} = \overline{D} + \sum_{i=1}^{n} p_{D_i}/(1 - p_{D_i}).$$

DIC$_\text{c}$ is exactly equal to the penalized plug-in deviance for linear models with known variance and may be a useful approximation for GLMMs. However, this approximation gives poor results in the Scottish lip cancer data, for which $p_D/n \approx 0.55$. This may be a typical situation in random-effect models. In general, for accurate estimation of the penalized plug-in deviance, special techniques, such as resampling of random effects, may be necessary. This undermines one of the attractions of DIC, which is at its ease of calculation.

Using the expected deviance as a loss function, the penalized loss function resembles DIC, but with a penalty approximately twice the size of p_D in regular exponential family models (van der Linde, 2005). The penalized expected deviance can be calculated in situations where DIC cannot easily be defined, due to the lack of a suitable plug-in estimate. In particular, it provides a solution to the model-choice problem for mixture models, which was previously considered intractable. It is tempting to suggest the penalized expected deviance as a general model-choice criterion in place of DIC. Even though it leads to a different compromise between model fit and complexity, the use of the expected deviance, instead of the plug-in deviance, does not necessarily lead to a different ranking of candidate models. Practical experience with DIC shows that much of the variation between models comes from $\overline{D}$ rather than p_D. In such cases, the conclusions about the best model may be robust to substantial misspecification of the penalty.

Although this paper has concentrated on loss functions based on the deviance, other penalized loss functions can be derived from the same framework. For example, one possibility is to use the predictive log-density $L^{\mathrm{pr}}(Y_i, \mathbf{Z}) = -2\log\{p(Y_i \mid \mathbf{Z})\}$. The use of L^{pr} as a loss function does not require the candidate models to have a common focus. The cross-validation form of this loss function $\sum_i L^{\mathrm{pr}}(Y_i \mid \mathbf{Y}_{-i})$ was proposed by Geisser and Eddy (1979). In regular linear models, it can be shown that the penalized predictive log-density is asymptotically equal to $\overline{D}$, in agreement with the asymptotic calculations of Gelfand and Dey (1994). $\overline{D}$ is generally considered not to give a sufficient penalty to complex models, and therefore L^{pr} may not be as useful as L^{e} or L^{p} in situations when all 3 loss functions can be defined.

The question of what constitutes a noteworthy difference in DIC between 2 models has not yet received a satisfactory answer. Whereas calibration scales have been proposed for Bayes factors (Kass and Raftery, 1995), no credible scale has been proposed for the difference in DIC between 2 models. Indeed, such a scale is unlikely be useful. Ripley (1996) shows that the sampling error of the difference in AIC between 2 models is $\mathrm{O}_p(1)$ when the models are nested, and the smaller model is true, but the error may be $\mathrm{O}_p(\sqrt{n})$ for nonnested models. No absolute scale for interpretation of AIC could be valid in both situations. DIC inherits this behavior since it includes AIC as a special case.

In the absence of an absolute scale for interpreting differences in L^{e} and L^{p}, some probabilistic recalibration is required. In Section 6, a simple approach was used, comparing the difference Δ in the penalized loss between 2 models with an empirical estimate of its standard error under repeated sampling. In the mixture model example, this calculation showed that an apparently large difference (6.1) in favor of a model with 3 components compared with 2 components was not reproducible under repeated sampling, and therefore exclusion of the model with 2 components would lead to poorly reproducible inference. Other, more computationally intensive, approaches are possible, such as calculating the posterior predictive distribution of Δ under the chosen optimal model or using the Bayesian bootstrap (Vehtari, 2001).

The results in this paper help to put DIC on a more formal basis. However, the model-choice criteria derived herein do not have a completely rigorous justification. The expectation (2.2) used to calculate the optimism is calculated differently for each candidate model. Although this may be acceptable with a “good model” assumption (Spiegelhalter *and others*, 2002; van der Linde, 2005), when at least one candidate model is considered to fit the data well, it may be hard to justify when comparing 2 ill-fitting models. A more rigorous approach, adopting the $\mathcal{M}$-completed perspective of Bernardo and Smith (2000), would be to calculate this expectation under a common “reference” model, representing the prior opinion of an observer who must choose between candidate models. Whether or not this makes a practical impact on the chosen model is a subject for further research.

ACKNOWLEDGMENTS

I would like to thank Jon Wakefield and David Spiegelhalter for their valuable advice and feedback. *Conflict of Interest:* None declared.

REFERENCES

AITKIN, M. (1991). Posterior Bayes factors (with discussion). *Journal of the Royal Statistical Society, Series B* **53**, 111–142.

BANERJEE, S., CARLIN, B. AND GELFAND, A. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, FL: Chapman & Hall/CRC.

BECHTEL, Y., BONAITI-PELLIE, C., POISSON, N., MAGNETTE, J. AND BECHTEL, P. (1993). A population and family study of n-acetyltransferase using caffeine urinary metabolites. *Clinical Pharmacology and Therapeutics* **54**, 134–141.

BERGER, J. AND PERICCHI, L. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* **91**, 109–122.

BERNARDO, J. (1979). Expected information as expected utility. *The Annals of Statistics* **7**, 686–690.

BERNARDO, J. AND SMITH, A. (2000). *Bayesian Theory*. Chichister, UK: Wiley.

BESAG, J., YORK, J. AND MOLLIÉ, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* **43**, 1–59.

BRESLOW, N. AND CLAYTON, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–24.

BURNHAM, K. AND ANDERSON, D. (2000). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd edition. New York: Springer.

CELEUX, G., FORBES, F., ROBERT, C. AND TITTERINGTON, D. (2006a). Deviance information criteria for missing data models. *Bayesian Analysis* **1**, 651–706.

CELEUX, G., FORBES, F., ROBERT, C. AND TITTERINGTON, D. (2006b). Rejoinder to discussion of deviance information criteria for missing data models. *Bayesian Analysis* **1**, 701–706.

CELEUX, G., HURN, M. AND ROBERT, C. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association* **95**, 957–970.

CLAYTON, D. AND KALDOR, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* **43**, 671–681.

CLEMENTS, M., ARMSTRONG, B. AND MOOLGAVKAR, S. (2005). Lung cancer rate predictions using generalized additive models. *Biostatistics* **6**, 576–589.

DANIELS, M. AND NORMAND, S. (2006). Longitudinal profiling of health care units based on continuous and discrete patient outcomes. *Biostatistics* **7**, 1–15.

EFRON, B. (1983). Estimating the error rate of a prediction rule: improvements on cross-validation. *Journal of the American Statistical Association* **78**, 316–331.

EFRON, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association* **81**, 461–470.

ELLIOTT, M., GALLO, J., HAVE, T. T., BOGNER, H. AND KATZ, I. (2005). Using a Bayesian latent growth curve model to identify trajectories of positive affect and negative events following myocardial infarction. *Biostatistics* **6**, 119–143.

GEISSER, S. AND EDDY, W. (1979). A predictive approach to model selection. *Journal of the American Statistical Association* **74**, 153–160.

GELFAND, A. AND DEY, D. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B* **56**, 501–514.

GELFAND, A. AND GHOSH, S. (1998). Model choice: a minimum posterior predictive loss approach. *Biometrika* **85**, 1–11.

GELMAN, A., CARLIN, J., STERN, H. AND RUBIN, D. (2002). *Bayesian Data Analysis*, 2nd edition. Boca Raton, FL: Chapman & Hall/CRC.

GNEITING, T. AND RAFTERY, A. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**, 359–378.

Hein, D., Doll, M., Fretland, A., Leff, M., Webb, S., Xiao, G., Devanaboyina, U., Nangju, N. and Feng, Y. (2000). Molecular genetics and epidemiology of the NAT1 and NAT2 acetylation polymorphisms. *Cancer Epidemiology, Biomarkers and Prevention* **9**, 29–42.

Hurvich, C. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297–307.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

Kemp, I., Boyle, P., Smans, M. and Muir, C. (1985). *Atlas of Cancer Incidence in Scotland, 1975–80. Incidence and Epidemiological Perspective*. IARC Scientific Publications, Number 72. Lyon, France: International Agency for Research on Cancer.

Laud, P. and Ibrahim, J. (1995). Predictive model selection. *Journal of the Royal Statistical Society, Series B* **57**, 247–262.

Lindley, D. (1991). Discussion of the paper by Aitkin. *Journal of the Royal Statistical Society, Series B* **53**, 130–131.

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

Liu, X., Wall, M. and Hodges, J. (2005). Generalized spatial structural equation models. *Biostatistics* **6**, 539–557.

Marshall, E. and Spiegelhalter, D. (2003). Approximate cross-validatory predictive checks in disease mapping models. *Statistics in Medicine* **22**, 1649–1660.

Neal, R. (1996). Sampling from multimodal distributions using tempered transitions. *Statistics and Computing* **4**, 353–366.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *Journal of the Royal Statistical Society, Series B* **57**, 99–138.

Peruggia, M. (1997). On the variability of case-deletion importance sampling weights in the Bayesian linear model. *Journal of the American Statistical Association* **92**, 199–207.

Plummer, M. (2002). Discussion of the paper by Spiegelhalter et al. *Journal of the Royal Statistical Society, Series B* **64**, 620.

Plummer, M. (2006). Comment on article by Celeux et al. *Bayesian Analysis* **1**, 681–686.

Richardson, S. and Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, Series B* **59**, 731–758.

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

Spiegelhalter, D., Best, N. and Carlin, B. (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. *Technical Report 98-009*. Division of Biostatistics, University of Minnesota. Cambridge, UK: Medical Research Council Biostatistics Unit.

Spiegelhalter, D., Best, N., Carlin, B. and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B* **64**, 583–639.

Spiegelhalter, D. J., Thomas, A., Best, N. and Lunn, D. (2004). *WinBUGS User Manual, Version 2.0*.

Stone, M. (1977). An asymptotic equivalence of choice of model cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B* **36**, 44–47.

Stone, M. (2002). Discussion of the paper by Spiegelhalter et al. *Journal of the Royal Statistical Society, Series B* **64**, 621.

van der Linde, A. (2004). On the association between a random parameter and an observation. *Test* **13**, 85–111.

VAN DER LINDE, A. (2005). DIC in variable selection. *Statistica Neerlandica* **59**, 45–56.

VEHTARI, A. (2001). Bayesian model assessment and selection using expected utilities, [PhD. Thesis]. Helsinki, Finland: Helsinki University of Technology.

VEHTARI, A. AND LAMPINEN, J. (2002). Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Computation* **14**, 2439–2468.

[*Received June 15, 2007; revised November 12, 2007; accepted for publication November 20, 2007*]