Reply to the discussion

David Spiegelhalter, Nicky Best, Brad Carlin and Angelika van der Linde
Revised May 3rd, 2002

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

1 Model focus and definition of deviance

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

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

model assessment and selection is an issue in hierarchical modelling and not specifically Bayesian.

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

2 Features of p_D

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

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

Replying to those discussants who were concerned about observing $p_D < n$ under 'flat' priors, we re-emphasize that $p_D = n$ was obtained theoretically only in the normal case or under normal approximations. There is no proof that $p_D = n$ for general distributions. In the case of S. Brooks' illustration using the Scottish lip cancer data, in which he shows that p_D appears to 'lose' 2 to 3 (modulo Monte Carlo error) parameters under such priors, we point out that 2 of the 56 observations in this dataset are zero with small expected values, and so contribute negligibly to the Poisson deviance. We have replicated his analysis

replacing these 2 observations by non-zero counts, and found that p_D increases by about 2 to around 55.5.

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

There is evidence (Daniels and Kass, 1999, 2001) that, in the absence of missing data, the use of default priors for variance components typically has little effect on the posteriors for the main effects in a model. Still, J. Smith and J. Bernardo observe that the flat priors that may maximize p_D are not necessarily weakly informative, and we agree. Reference priors that are least informative in an information-theoretical sense can be easily studied in some of our examples. For example, Figure 1 displays the performance of the Beta$(\frac{1}{2}, \frac{1}{2})$ reference prior (corresponding to a prior sample size of $n_i = a + b = 1$) for the binomial likelihood, and the approximation (31) indicates that $p^{\Theta}_{D_i}$ based on the reference prior is greater than $p^{\Theta}_{D_i}$ based on the uniform Beta$(1, 1)$ prior (which has prior sample size $n_i = 2$). Similarly for a Poisson likelihood the reference prior $\pi(\mu_i) \propto \sqrt{\mu_i}$ yields a $\Gamma(y_i + \frac{1}{2}, n_i)$ posterior distribution corresponding to $a = \frac{1}{2}$, $b \to 0$. Hence $p^{\mu}_{D_i} \approx \frac{y_i}{y_i + \frac{1}{2}}$ and $p^{\Theta}_{D_i} \approx \frac{n_i}{n_i} = 1$ might be compared to the values shown in Figure 2.

3 Properties of DIC

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

Though we agree with S. Brooks that due to its heuristic derivation DIC may be considered as a "broad brush technique," we do not regard it to be as arbitrary as the alternatives he suggests. In particular we do not feel that

terms of "fit" and "complexity" can be arbitrarily combined, but reemphasize that a measure of model complexity results from correcting overfit due to an approximation of the expected loss that "uses the observations twice." Similarly we would like to see a justification of A. Vehtari's estimates of expected utilities as valid approximations generalizing DIC.

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

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

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

4 Applications

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

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

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

5 Alternatives to DIC

Several discussants (S. Brooks, A.P. Dawid, S. Sahu) feel that DIC suffers in comparison to more traditional Bayesian model selection criteria based on posterior model probabilities and Bayes factors. Here we can only repeat that our deliberate intention was to offer an alternative to Bayes factors, which are most suitable when the entire collection of candidate models can be specified ahead of time (the "$\mathcal{M}$-closed" case of Bernardo and Smith, 1994). In our practical experience, the model building, criticism, and rebuilding process is typically an iterative, "$\mathcal{M}$-open" one in which the ultimate model collection is rarely known ahead of time, and here DIC may well emerge as more appropriate. Moreover,

Bayes factors address how well the prior has predicted the observed data; this prior predictive emphasis ultimately leads to the Lindley paradox. DIC instead addresses how well the *posterior* might predict future data generated by the same mechanism that gave rise to the observed data; this posterior predictive outlook might be considered intuitively more appealing in many practical contexts. We emphasise that these techniques are intended to answer different questions and cannot be expected to give the same conclusions: in any case, posterior model probabilities may be highly dependent on within- and between-model priors, so their comparison with DIC is not straightforward. On a related point, several discussants (S. Brooks, K. Burnham, D. Draper) mention the possible alternative of model averaging. We do not, however, see any justification for transforming DIC values to relative probabilities, and in any case the prior on the model space may be difficult to develop, and might even reasonably be related to model complexity!

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

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

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

References

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

[2] Daniels, M. J. and Kass, R. E. (2001). Shrinkage estimators for covariance matrices. *Biometrics*, **57**, 1173–1184

[3] Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika*, **83**, 875–890

[4] Holmes, C. and Denison, D. (1999). Bayesian Wavelet analysis with a model complexity prior. In *Bayesian Statistics 6*, (Eds.: J.M.Bernardo et al.) Oxford University Press, 769–776.

[5] Volinsky, C. T. and Raftery, A. E. (2000). Bayesian information criterion for censored survival models. *Biometrics*, **56**, 256–262.