ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2014)
76, *Part* 3, *pp.* 485–493

The deviance information criterion: 12 years on

David J. Spiegelhalter,

University of Cambridge, UK

Nicola G. Best,

Imperial College School of Public Health, London, UK

Bradley P. Carlin

University of Minnesota, Minneapolis, USA

and Angelika van der Linde

Bremen, Germany

[*Presented to* The Royal Statistical Society *at its annual conference in a session organized by the* Research Section *on Tuesday, September 3rd, 2013*, Professor G. A. Young *in the Chair*]

Summary. The essentials of our paper of 2002 are briefly summarized and compared with other criteria for model comparison. After some comments on the paper's reception and influence, we consider criticisms and proposals for improvement made by us and others.

Keywords: Bayesian; Model comparison; Prediction

1. Some background to model comparison

Suppose that we have a given set of candidate models, and we would like a criterion to assess which is 'better' in a defined sense. Assume that a model for observed data y postulates a density $p(y|\theta)$ (which may include covariates etc.), and call $D(\theta) = -2\log\{p(y|\theta)\}$ the deviance, here considered as a function of θ. Classical model choice uses hypothesis testing for comparing nested models, e.g. the deviance (likelihood ratio) test in generalized linear models. For non-nested models, alternatives include the Akaike information criterion

$$\mathrm{AIC} = -2\log\{p(y|\hat{\theta})\} + 2k$$

where $\hat{\theta}$ is the maximum likelihood estimate and k is the number of parameters in the model (dimension of Θ).

AIC is built with the aim of favouring models that are likely to make good predictions. Since we generally do not have independent validation data, we can assess which model best predicts the *observed* data by using the deviance, but if parameters have been estimated we need some penalty for this double use of the data. AIC's penalty of $2k$ has been shown to be asymptotically equivalent to leave-one-out cross-validation. However, AIC does not work in models with informative prior information, such as hierarchical models, since the prior effectively acts to

Address for correspondence: David J. Spiegelhalter, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, Cambridge, CB3 0WB, UK.
E-mail: d.spiegelhalter@statslab.cam.ac.uk

 1369–7412/12/76485

'restrict' the freedom of the model parameters, so the appropriate 'number of parameters' is generally unclear.

Schwarz's Bayesian information criterion BIC is given by

$$\text{BIC} = -2\log\{p(y|\hat{\theta})\} + k\log(n);$$

this penalizes complexity more than AIC and so will favour simpler models.

AIC and BIC represent two very different approaches to model comparison. Since differences in BIC provide an approximation to the log(Bayes factor) under specific 'unit information' prior assumptions (Kass and Raftery, 1995), BIC is essentially trying to identify the 'true' model and, if a candidate model is the genuine data-generating mechanism, BIC will be asymptotically consistent in selecting such a model. In contrast, AIC is not asymptotically consistent, as it is not seeking to select the 'true model', but rather to make a pragmatic choice that explains data and will make good short-term predictions. It is perhaps therefore rather surprising how often these two criteria produce similar rankings of candidate models.

2. Spiegelhalter *et al.* (2002)

The BUGS project started in 1989, and by 1992 we had a program that could fit arbitrarily complex Bayesian models by using Markov chain Monte Carlo methods (Thomas *et al.*, 1992). Although this was a considerable advance, it was always clear that there was a pressing need for a model comparison criterion such as those provided in standard statistical packages. In 1996 we suggested the posterior mean deviance $\mathbf{E}_{\theta|y}[D(\theta)] = \overline{D(\theta)}$ as a measure of fit (Spiegelhalter *et al.*, 1996), but we acknowledged that it was unclear what penalty for increased complexity should be applied.

A draft paper in 1997 suggested $\frac{1}{2}\mathbf{V}_{\theta|y}\{D(\theta)\}$ as a penalty (see later), but this was superseded in 1998 by the first draft of the paper on the deviance information criterion DIC that was submitted to the Royal Statistical Society (authors Spiegelhalter, Best and Carlin). After rather negative referees' reports had been received, the first author went into a massive sulk for a year, but recruiting van der Linde into the team in 2000 led to new vitality and a new draft was submitted in 2001, and finally read to the Society in 2002.

Spiegelhalter *et al.* (2002) suggested a measure of 'effective number of parameters'

$$p_D = \mathbf{E}_{\theta|y}[-2\log\{p(y|\theta)\}] + 2\log[p\{y|\tilde{\theta}(y)\}].$$

If we take $\tilde{\theta} = \mathbf{E}[\theta|y]$, then $p_D =$ 'posterior mean deviance − deviance of posterior means'. In normal linear hierarchical models:

$$p_D = \text{tr}\{\hat{I}(\bar{\theta})\,\mathbf{V}(\theta|y)\},$$

where $\mathbf{V}(\theta|y)$ is the posterior covariance matrix, $\hat{I}(\bar{\theta})$ is the observed Fisher information evaluated at the posterior mean, and estimates the inverse covariance matrix of the maximum likelihood estimate. In general, the justification of the trace depends on the approximate normality of the posterior distribution.

If there is 'vague' prior information, $\bar{\theta} \approx \hat{\theta}$, and so $D(\theta) \approx D(\bar{\theta}) + \chi^2_k$; hence $p_D \approx \mathbf{E}[\chi^2_k] = k$, the true number of parameters.

Spiegelhalter *et al.* (2002) proposed a deviance information criterion based on the principle DIC = 'goodness of fit' + 'complexity'. DIC is defined, analogously to AIC, as

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

DIC can be easily monitored in BUGS.

3. How has Spiegelhalter *et al.* (2002) been received?

At the meeting considerable doubts were expressed about the paper: Jim Smith (University of Warwick) said

> 'I shall not address technical inaccuracies but just present four foundational problems that I have with the model selection in this paper. ... So my suggestion to a practitioner would be: if you must use a formal selection criterion do not use DIC'.

Philip Dawid (University College London) was concerned about the lack of consistency:

> '... should we not seek reassurance that a procedure performs well in those simple cases for which its performance can be readily assessed, before trusting it on more complex problems?'.

Subsequent citations suggest that Professor Smith's exhortations went unheeded. Spiegelhalter *et al.* (2002) was the third most cited paper in international mathematical sciences between 1998 and 2008, and up to November 2013 it had over 2500 citations on the *Web of Knowledge* and over 4600 on Google Scholar. The citations show no sign of abating—at the time of the presentation of this material (September 2013), around two new papers a day were reported as citing Spiegelhalter *et al.* (2002) on Google Scholar. Recent examples included applications in drug safety, tree breeding, distributions of platypus and assessing bias in the Eurovision Song Contest, reflecting the extensive use of DIC in biostatistics, ecology and general applications of hierarchical models.

4. So what is 'wrong' with DIC?

Common criticisms (including those following this paper) include, roughly in increasing order of importance the following.

(a) *p_D is not invariant to reparameterization*. For example, we would obtain a (slightly) different p_D (and hence DIC) if we parameterized in terms of σ or $\log(\sigma)$, even if the priors on each were mathematically equivalent. p_D can even be negative if the posterior of θ is very non-normal and so $\bar{\theta}$ does not provide a very good estimate of θ. See below for some ways of dealing with this issue.

(b) *Lack of consistency*: the lack of consistency concerns many critics, but the stated aim of DIC is in optimizing short-term predictions of a particular type, and not in trying to identify the 'true' model: except in rare and stylized circumstances, we contend that such an entity is an unattainable ideal.

(c) *It is not based on a proper predictive criterion*. Although, like AIC, DIC is supposedly based on a criterion for predicting replicate data Y^{rep}, it uses plug-in predictions $p(Y^{\text{rep}}|\hat{\theta})$, rather than full predictive distributions $p(Y^{\text{rep}}) = \int p(Y^{\text{rep}}|\theta)\, p(\theta|y)\mathrm{d}\theta$—the latter would also provide invariance to reparameterization.

(d) *It has a weak theoretical justification.* As Celeux *et al.* (2006) showed well in the context of mixture models, DIC is not based on a universal principle that could lead to a procedure that was both computationally practical and generically applicable. DIC, as with other prediction-based criteria, starts with a particular posterior predictive target based on replicates conditional on certain elements of the design remaining fixed—when we assume a parametric model this replication structure defines the 'focus' of the analysis. As implemented in BUGS, DIC takes the lowest level parameters as the focus, but this is only to make the technique computationally feasible.

The predictive target is usually approximated by, for example, Taylor or entropy expansions and the approximate terms estimated from the observed data. These steps result

in penalty terms for model complexity—van der Linde (2012) argued that this can be thought of as expressing the dependence between data and parameters in a fully specified Bayesian joint model. Unfortunately neither the approximation nor the estimation steps are universally valid but depend on the class of models being considered. We also note that a derivation of both AIC and DIC may make use of the concept of a 'true' distribution, even though we have just denied that, in general, any of our models can be considered true. This may appear contradictory but, although we regard a 'true' model as a useful mathematical concept, it is practically unattainable, perhaps like ∞.

5. What has been done to try to improve on DIC?

5.1. Patch up p_D

WinBUGS currently uses the posterior mean of stochastic parents of θ, i.e., if there are stochastic nodes ϕ such that $\theta = f(\phi)$, then $D(\tilde{\theta}) = D\{f(\bar{\phi})\}$. It would be better if BUGS used the posterior mean of an appropriate function of the 'direct parameters' (e.g. those that appear in the BUGS distribution syntax) to give the plug-in deviance, rather than the posterior means of the 'stochastic parents'. This would make DIC invariant to reparameterization.

The BUGS software architecture meant that this was not a straightforward change but is currently being implemented in OpenBUGS (better late than never).

5.2. p_V: an alternative measure of complexity

Suppose that we have a non-hierarchical model with a weak prior, so that $D(\theta) \approx D(\bar{\theta}) + \chi^2_k$. Then $\mathbf{E}[D(\theta)] \approx D(\bar{\theta}) + k$, so $p_D \approx k$ and $\mathbf{V}\{D(\theta)\} \approx 2k$. Thus, with negligible prior information, half the variance of the deviance is an estimate of the number of free parameters in the model. This estimate generally turns out to be remarkably robust and accurate, and this has suggested the use of $p_V = \mathbf{V}(D)/2$ as an estimate of the effective number of parameters in a model in more general situations with informative prior information. This was originally suggested in a working paper in 1997 and has since been proposed by Gelman *et al.* (2004). The posterior distribution of the deviance is not affected by equivalent reparameterizations, and so p_V will be invariant.

5.3. Allowing for prediction

van der Linde (2005, 2012), Plummer (2008) and Ando (2012) have all concluded that DIC overfits because it uses a plug-in estimate in the predictive target rather than the proper predictive distribution, and all have concluded that a better method would be to increase the penalty:

$$\begin{aligned} \mathrm{DIC}^* &= D(\bar{\theta}) + 3p_D \\ &= \bar{D} + 2p_D, \end{aligned}$$

so that choosing an 'average target' (with an expectation) rather than a 'representative target' with an estimate 'costs' one p_D. So far the argument has been theoretical rather than practical, although experiences with AIC as a special case point in the same direction. More studies are needed to evaluate the differences in ranking models with proper priors resulting from the use of DIC or DIC*.

6. The Watanabe–Akaike criterion: a worthy successor?

Spiegelhalter *et al.* (2002) always assumed that their preliminary efforts would be superseded by others. There is now a bewildering alphabet of 'information criteria', including AIC (in multiple

versions), BIC, BPIC, CIC, DIC, EIC, FIC, NIC, TIC. . ., although DIC is still holding ground—for a discussion addressing information theoretical issues and embedding DIC, see van der Linde (2012).

Perhaps the most promising innovation may be the Watanabe–Akaike (or 'widely applicable') information criterion WAIC (Watanabe, 2010). Assume that the y_is are conditionally independent given θ, and we aim to minimize the predictive criterion

$-2\{$expected log(pointwise predictive density) for a replicate data set$\}$

$$= -2 \sum_i \mathbf{E}_f[\log\{p(Y_i^{\text{rep}}|y)\}],$$

where f is the 'true' density. The measure of 'fit' equals

$$\begin{aligned} -2 \log(\text{pointwise predictive density}) &= -2 \log\left\{ \prod_i p(y_i|y) \right\} \\ &= -2 \sum_i \log\left\{ \int p(y_i|\theta)\, p(\theta|y)\, \mathrm{d}\theta \right\}. \end{aligned}$$

This is computed in Markov chain Monte Carlo algorithms as

$$-2 \sum_i \log\left\{ \frac{1}{S} \sum_{s=1}^{S} p(y_i|\theta^s) \right\},$$

where θ^s is simulated from the distribution of θ: this is a proper predictive density rather than a plug-in.

Analogously to p_D and p_V, there are two WAIC complexity penalties:

$$p_{\text{WAIC1}} = \sum_i \mathbf{E}_{\theta|y}[-2 \log\{p(y_i|\theta)\}] + 2 \sum_i \log\{p(y_i|y)\};$$

$$p_{\text{WAIC2}} = \tfrac{1}{4} \sum_i \mathbf{V}_{\theta|y}[-2 \log\{p(y_i|\theta)\}].$$

Model comparison by WAIC has been shown to be asymptotically equal to using Bayesian leave-one-out cross-validation (Watanabe, 2010), is invariant to reparameterization and will deal with mixture models. See Gelman *et al.* (2013) for a full discussion of WAIC and comparison with DIC and other criteria.

The success of DIC has rested largely on its ease of computation and availability in standard software. WAIC involves Monte Carlo estimation of predictive densities, rather than log(densities), and as such could be considerably more tricky to implement robustly.

7. Conclusions

The authors of the 2002 DIC-paper still feel that DIC was a good idea but have always been aware of its limitations. We are somewhat surprised that it has lasted so well, but in spite of its problems we are confident that it has benefited a large number of researchers in a range of substantive fields. In our personal experience, when it has come out with idiotic results such as a negative p_D, then this has acted as an appropriate warning of issues in the modelling that need to be addressed.

Nevertheless there remain reasonable criticisms of its derivation and use. DIC has stimulated rich developments, and we eagerly await routine implementation of worthy alternatives.

References

Ando, T. (2012) Predictive Bayesian model selection. *Am. J. Math. Mangmnt Sci.*, **31**, 13–38.

Celeux, G., Forbes, F., Robert, C. and Titterington, D. (2006) Deviance information criteria for missing data models. *Baysn Anal.*, **1**, 651–706.
Gelman, A., Carlin, J. C., Stern, H. and Rubin, D. B. (2004) *Bayesian Data Analysis*, 2nd edn. New York: Chapman and Hall.
Gelman, A., Hwang, J. and Vehtari, A. (2013) Understanding predictive information criteria for Bayesian models. *Statist. Comput.*, to be published.
Kass, R. and Raftery, A. (1995) Bayes factors and model uncertainty. *J. Am. Statist. Ass.*, **90**, 773–795.
van der Linde, A. (2005) DIC in variable selection. *Statist. Neerland.*, **59**, 45–56.
van der Linde, A. (2012) A Bayesian view of model complexity. *Statist. Neerland.*, **66**, 253–271.
Plummer, M. (2008) Penalized loss functions for Bayesian model comparison. *Biostatistics*, **9**, 523–539.
Thomas, A., Spiegelhalter, D. J. and Gilks, W. G. (2002) BUGS: a program to perform Bayesian inference using Gibbs sampling. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 837–842. Oxford: Oxford University Press.
Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *J. R. Statist. Soc.* B, **64**, 583–639.
Spiegelhalter, D., Thomas, A., Best, N. and Gilks, W. (1996) *BUGS 0.5: Bayesian Inference using Gibbs Sampling*, version ii. Cambridge: Medical Research Council Biostatistics Unit.
Watanabe, S. (2010) Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.*, **11**, 3571–3594.

Comments on the presentation

Elías Moreno (*Universidad de Granada, Spain*) **and Francisco-José Vázquez-Polo** (*Universidad de Las Palmas de Gran Canaria, Spain*)

Spiegelhalter *et al.* (2002) is an interesting paper, in which a new dimension correction to penalize overfitted models is presented. It has given rise to considerable discussion; here, we focus on the DIC model selection procedure that is defined in it.

11 years later, model selection for complex models remains an open problem. The weak link of the Bayesian model selection approach is the elicitation of the prior over models and over the model parameters to be used in the procedure. Several priors have been proposed for interesting model selection problems, such as variable selection in high dimensional regression, clustering, change points and classification, but none of them satisfy all reasonable requirements. Thus, we fully agree with the authors' claim in justifying DIC that '*full elicitation of informative priors and utilities is simply not feasible in most situations*'. However, this does not imply that in model selection we can avoid the use of priors in a coherent way (Berger and Pericchi, 2001).

Does the deviance information criterion have a justification from a decision theory viewpoint?

In model selection we have a sample $\mathbf{y}_n$ of size n, a discrete class of k competing sampling models $\mathfrak{M}$, the sampling density of model M_i, $f(\mathbf{y}_n|\theta_i, M_i)$, and a prior for models and model parameters $\pi(\theta_i, M_i) = \pi(\theta_i|M_i)\,\pi(M_i)$, where $\theta_i \in \Theta_i$. The parameter spaces are typically continuous.

In model selection the quantity of interest is the model, and therefore the decision space is $\mathfrak{D} = \{d_j, j = 1, \ldots, k\}$, where d_j is the decision to choose model M_j, and the states of nature is the class of models $\mathfrak{M}$. Given a loss function $\mathfrak{L}(d_i, M_j)$, $\mathfrak{L}: \mathfrak{D} \times \mathfrak{M} \to \mathbb{R}^+$, the optimal Bayesian decision is to choose the model M^π such that

$$M^\pi = \arg\min_{i=1,\ldots,k} \sum_{j=1}^{k} \mathfrak{L}(d_i, M_j)\,\pi(M_j|\mathbf{y}_n),$$

where

$$\pi(M_j|\mathbf{y}_n) = \frac{m_j(\mathbf{y}_n)\,\pi(M_j)}{\sum_{j=1}^{k} m_j(\mathbf{y}_n)\,\pi(M_j)},$$

and the marginal $m_i(\mathbf{y}_n) = \int_{\Theta_i} f(\mathbf{y}_n|\theta_i, M_i)\,\pi(\theta_i|M_i)\,\mathrm{d}\theta_i$ is the likelihood of model M_i, $i = 1, \ldots, k$. This means that, whatever loss function $\mathfrak{L}(d_i, M_j)$ we use, the optimal decision depends on the posterior model probabilities, i.e. the decision formulation takes into account the uncertainty of the model. However, DIC does not depend on $\pi(M_j|\mathbf{y}_n)$, $j = 1, \ldots, k$.

Does the deviance information criterion correspond to a Bayesian procedure?

The Bayesian procedures automatically penalize model complexity without any adjustment (Dawid, 2002),

and this is a good reason to require a model selection procedure to be Bayesian. Another reason is that the competing models can be averaged, with the weights being the model posterior probabilities, whereas for Schwarz's Bayesian information criterion BIC, to compare model M_i with M_j,

$$-2\log\{\mathrm{BIC}_{ij}(\mathbf{y}_n)\} = -2\log\left[\frac{f\{\mathbf{y}_n|\hat{\theta}_i(\mathbf{y}_n), M_i\}}{f\{\mathbf{y}_n|\hat{\theta}_j(\mathbf{y}_n), M_j\}}\right] + (d_i - d_j)\log(n),$$

where d_i and d_j are the dimensions of Θ_i and Θ_j, there is a Bayes factor B_{ij} such that $|-2\log(B_{ij}) - 2\log(\mathrm{BIC}_{ij})| = O_P(n^{-1/2})$ (Kass and Wasserman, 1995) and thus BIC asymptotically corresponds to a Bayes factor; we do not see that a similar correspondence can be established with the

$$\mathrm{DIC}_{ij}(\mathbf{y}_n) = -2\log\left[\frac{f\{\mathbf{y}_n|\bar{\theta}_i(\mathbf{y}_n), M_i\}}{f\{\mathbf{y}_n|\bar{\theta}_j(\mathbf{y}_n), M_j\}}\right] + \mathrm{correction}_{ij}$$

where $\bar{\theta}_i(\mathbf{y}_n) = E_{\theta_i|\mathbf{y}_n}(\theta_i)$, and

$$\mathrm{correction}_{ij} = 4(E_{\theta_i|\mathbf{y}_n}[\log\{f(\mathbf{y}_n|\theta_i, M_i)\} - E_{\theta_j|\mathbf{y}_n}[\log\{f(\mathbf{y}_n|\theta_j, M_i)\}]) + 4\log\left[\frac{f\{\mathbf{y}_n|\tilde{\theta}_i(\mathbf{y}_n), M_i\}}{f\{\mathbf{y}_n|\tilde{\theta}_j(\mathbf{y}_n), M_j\}}\right].$$

We note that under mild conditions $|\hat{\theta}(\mathbf{y}_n) - \bar{\theta}(\mathbf{y}_n)| = O_P(n^{-1})$, and hence the main difference between BIC and DIC comes from the correction term. As a result of this term, DIC does not correspond to a Bayesian procedure.

Asymptotics

DIC is not a consistent model selection procedure and, although it is a negative property of the procedure, this does not seem to worry the authors, who argued that '*we neither believe in a true model nor would expect the list of models being considered*'. This implies that the probability of a model has no meaning, as no model space is considered. However, the point is that, if we applied DIC to a case in which the class of models were known, we would have consistency.

Some statisticians (for instance Fraser (2011)) have suggested that the sampling properties of the Bayesian methods should be studied. In this respect, Wasserman (2011) asserted that 'we must be vigilant and pay careful attention to the sampling properties of procedures'. We agree with both these views and believe that consistency is a very useful sampling property that allows us to compare the behaviour of alternative Bayesian model selection procedures for complex models.

Consistency in a model selection procedure for a given class of models $\mathfrak{M}$ means that, when sampling from a model in $\mathfrak{M}$, the posterior probability of this model tends to 1 as the sample size tends to ∞. Bayesian procedures for model selection are typically consistent when the dimension of the models is small compared with the sample size (Dawid, 1992; Casella *et al.*, 2009). Furthermore, when the model from which we are sampling is not in the class $\mathfrak{M}$, the Bayesian procedure asymptotically chooses a model in $\mathfrak{M}$ that is as close as possible to the true model, in the Kullback–Leibler distance.

However, consistent Bayesian procedures for low dimensional models are not necessarily consistent for high dimensional models. For example:

(a) Schwarz's approximation to the Bayes factor BIC is not necessarily consistent in high dimensional settings (Berger, 2003; Moreno *et al.*, 2010);
(b) when the number of models increases with the sample size, as occurs in clustering, change point or classification problems, consistency of the Bayesian model selection procedure depends not only on the prior over the model parameters but also on the prior over the models (in fact, default priors that are commonly used for discrete spaces may give an inconsistent Bayesian model selection procedure, as occurs in clustering when using the uniform prior over the models (Casella *et al.*, 2014));
(c) in variable selection in regression when the number of regressors p increases with the sample size, i.e. $p = O(n^b)$, $0 \leqslant b \leqslant 1$, some priors that are commonly used over the model parameters and over the model space make the Bayesian procedures inconsistent (for instance, the g-priors (Zellner, 1986) with $g = n$ produce an inconsistent Bayesian procedure).

The mixture of g-priors with respect to the inverse gamma$(g|1/2, n/2)$, or the intrinsic priors (Moreno *et al.*, 1998) over the model parameters when combined with the independent Bernoulli prior on the model space (George and McCulloch, 1997; Raftery *et al.*, 1997), may also provide an inconsistent Bayesian procedure.

These results show that consistency can be a very useful property for the difficult task of selecting priors for model selection in complex models.

(This research was partially funded by grants MTM2011-28945 and ECO2009-14152 (Ministerio de Ciencia e Innovación, Spain).)

Christian P. Robert (*University of Warwick, Coventry, and Université Paris-Dauphine*)

The main issue with DIC undoubtedly is the question of its worth for (or within) Bayesian decision analysis (since I doubt whether there are many proponents of DIC outside the Bayesian community). The appeal of DIC is, I presume, to deliver a *single* summary per model for all models under comparison and to allow therefore for a complete ranking of those models. I, however, object to the worth of simplicity for simplicity's sake: models are complex (albeit less than reality) and their usages are complex as well. To consider that model A is to be preferred over model B just because $\mathrm{DIC(A)} = 1228 < \mathrm{DIC(B)} = 1237$ is a travesty of the complex mechanisms at play behind model choice, especially given the wealth of information that is provided by a Bayesian framework. (Non-Bayesian paradigms may be more familiar with procedures based on a single estimator value.) And to abstain from accounting for the significance of the difference between DIC(A) and DIC(B) clearly makes matters worse.

This is not even discussing the stylized setting where one model is considered as 'true' and where procedures are compared by their ability to recover the 'truth'. David Spiegelhalter repeatedly mentioned during his talk that he was not interested in this. This stance inevitably brings another objection, though, namely that models—as tools instead of approximations to reality—can only be compared against their predictive abilities, which DIC seems unable to capture. Once again, what is needed in this approach to model comparison is a multifactor and all-encompassing criterion that evaluates the predictive models in terms of their recovery of some features of the phenomenon under study, or of the process being conducted. (Even stooping to a one-dimensional loss function that is supposed to summarize the purpose of the model comparison does not produce anything close to the DIC-function, unless one agrees to massive approximations.)

Obviously, considering that asymptotic consistency is of no importance whatsoever (as repeated by David Spiegelhalter in his presentation) easily avoids some embarrassing questions, except the (still embarrassing) question about the true purpose of statistical models and procedures. How can those be compared if no model is true and if accumulating data from a given model is not meaningful? How can simulation be conducted in such a barren landscape? I find this minimalist attitude the more difficult to accept that models are truly used as if they were or could be true, at several stages in the process. It also prevents the study of the criterion under model misspecification, which would clearly be of interest.

Another point worth discussing, which has already been exposed in Celeux *et al.* (2006), is that there is no unique driving principle for constructing DICs. In Celeux *et al.* (2006) inspired from the discussion by DeIorio and Robert (2002), we examined eight different and natural versions of DIC for mixture models, resulting in highly diverging values for DIC and the effective dimension of the parameter; I believe that such a lack of focus is bound to reappear in any multimodal setting and I fear that the answer about (eight) different focus on what matters in the model is too cursory and lacks direction for the hapless practitioner.

My final and critical remark about DIC is that the criterion shares very much the same perspective as Murray Aitkin's integrated likelihood, as already stressed in Robert and Titterington (2002). Both Aitkin (1991, 2010) and Spiegelhalter *et al.* (2002) considered a posterior distribution on the likelihood function, taken as a function of the parameter but omitting the delicate fact that it also depends on the observable and hence does not exist *a priori*. See Gelman *et al.* (2013) for a detailed review of Aitkin (2010), since most of the criticisms therein equally apply to DIC, and I shall not reproduce them here, except to point out that DIC escapes the Bayesian framework (and thus requires even more its own justifications).

References

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

Aitkin, M. (2010) *Statistical Inference: a Bayesian/Likelihood Approach.* Boca Raton: Chapman and Hall–CRC.

Berger, J. O. (2003) Could Fisher, Jeffreys and Neyman have agreed on testing (with discussion)? *Statist. Sci.*, **18**, 1–32.

Berger, J. O. and Pericchi, L. (2001) Objective Bayesian methods for model selection: introduction and comparison (with discussion). In *IMS Lecture Notes-Monograph Series*, vol. 38, pp. 135–203. Amsterdam: North-Holland.

Casella, G., Girón, F. J., Martinez, M. L. and Moreno, E. (2009) Consistency of Bayesian procedures for variable selection. *Ann. Statist.*, **37**, 1207–1228.

Casella, G., Moreno, E. and Girón, F. J. (2014) Cluster analysis, model selection, and prior distributions on models. *Baysn Anal.*, to be published.
Celeux, G., Forbes, F., Robert, C. P. and Titterington, D. M. (2006) Deviance information criteria for missing data models (with discussion). *Baysn Anal.*, **1**, 651–674.
Dawid, A. P. (1992) Prequential analysis, stochastic complexity and Bayesian inference. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 109–125. Oxford: Oxford University Press.
Dawid, A. P. (2002) Discussion on 'Bayesian measures of model complexity and fit' (by D. J. Spiegelhalter, N. G. Best, B. P. Carlin and A. van der Linde). *J. R. Statist. Soc.* B, **64**, 624.
DeIorio, M. and Robert, C. P. (2002) Discussion on 'Bayesian measures of model complexity and fit' (by D. J. Spiegelhalter, N. G. Best, B. P. Carlin and A. van der Linde). *J. R. Statist. Soc.* B, **64**, 629–630.
Fraser, D. A. S. (2011) Is Bayes posterior just quick and dirty confidence? *Statist. Sci.*, **26**, 299–316.
Gelman, A., Robert, C. P. and Rousseau, J. (2013) Inherent difficulties of non-Bayesian likelihood-based inference, as revealed by an examination of a recent book by Aitkin (with a reply from the author). *Statist. Risk Modlng*, **30**, 1001–1016.
George, E. I. and McCulloch, R. E. (1997) Approaches for Bayesian variable selection. *Statist. Sin.*, **7**, 339–374.
Kass, R. E. and Wasserman, L. (1995) A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Am. Statist. Ass.*, **90**, 928–934.
Moreno, E., Bertolino, F. and Racugno, W. (1998) An intrinsic limiting procedure for model selection and hypothesis testing. *J. Am. Statist. Ass.*, **93**, 1451–1460.
Moreno, E., Girón, F. J. and Casella, G. (2010) Consistency of objective Bayes factors as the model dimension grows. *Ann. Statist.*, **38**, 1937–1952.
Raftery, A., Madigan, D. and Hoeting, J. (1997) Bayesian model averaging for linear regression models. *J. Am. Statist. Ass.*, **92**, 179–191.
Robert, C. P. and Titterington, D. M. (2002) Discussion on 'Bayesian measures of model complexity and fit' (by D. J. Spiegelhalter, N. G. Best, B. P. Carlin and A. van der Linde). *J. R. Statist. Soc.* B, **64**, 621–622.
Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *J. R. Statist. Soc.* B, **64**, 583–639.
Wasserman, L. (2011) Frasian inference (discussion on the paper by Fraser, 2011). *Statist. Sci.*, **26**, 322–325.
Zellner, A. (1986) On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (eds P. K. Goel and A. Zellner), pp. 233–243. Amsterdam: North-Holland.