Bayesian Analysis (2006) **1**, Number 4, pp. 701–706

Rejoinder

G. Celeux*, F. Forbes†, C.P. Robert‡ and D.M. Titterington§

We are grateful to all discussants for their comments and to an editor for initiating this discussion. Rather than addressing each discussion separately, we identify several themes of interest and contention among the discussants that we now develop separately.

1 Foundations of DIC

A theme common to all discussions is that DIC is so far more of a plausible measure of complexity than a well-grounded criterion. We completely agree with this perspective and even share the more radical prognosis of Meng and Vaida that DIC may simply lack a theoretical foundation. Indeed, there are deeper concerns with DIC than just that of a definition in the missing data case. In this regard, we do agree with Carlin that our "casework" analysis cannot solve the problem of defining a proper DIC for missing data and even less in general. Therefore, Carlin's point that *"authors do not refer at all to any derivation, nor to any subsequent interpretation of model complexity"* is both true and meaningless: if DIC as originally defined is a universal way of evaluating model fit or model complexity, it should also apply in the missing data setting and we showed here that it clearly does not. The main conclusion of our paper is thus that DIC lacks a natural generalisation outside exponential families or, alternatively, that it happened to work within exponential families while lacking a true theoretical foundation. Similarly, regarding Meng and Vaida's criticisms about our proposal of an almost tautological emphasis, we (obviously!) cannot agree: in the paper, we are considering models that can be *fruitfully* regarded as missing data models, that is models for which there is a many to one mapping linking the complete data and the observed data.

Some discussants attempt to provide alternatives that could establish theoretical foundations for DIC. For instance, van der Linde focusses on DIC as an approximate estimated loss, in the same way that BIC is an approximate log Bayes factor, even though she is obviously less critical of DIC in exponential families. She seems to envisage our developments as the result of various approximations. In that perspective, we could wonder what is the whole point of producing such criteria. If the approximation (of a posterior loss?) cannot be evaluated, we should then consider other models in which no approximation is required and then check the appropriateness of each approximation. Further, while using true loss functions is usually sensible (Celeux et al. 2000), it remains to be seen which loss functions correspond to each of the DIC_i's, if any. (In this regard, DIC_2 could be described in a sense as being a more robust version of the basic DIC_1.) This obviously does not relate to the hair(y) loss mentioned by Meng and Vaida!

*INRIA FUTURS, Orsay, France, mailto:gilles.celeux@inrialpes.fr
†INRIA Rhône-Alpes, France, mailto:Florence.Forbes@inrialpes.fr
‡CREST and CEREMADE, Uni. Paris Dauphine, France, mailto:xian@ceremade.dauphine.fr
§University of Glasgow, Glasgow, UK, mailto:mike@stats.gla.ac.uk

The very idea of loss function is nonetheless very central to the debate, since DIC appears as a portmanteau substitute for well-defined loss functions. While debating about DIC, we are so far forgetting a central issue, namely what we plan to do with the output of a model comparison exercise. In fact, there is a "dark history" of Bayesian model assessment waiting to be told, in that almost all attempts have stepped outside Bayesian boundaries in order to evaluate the fit of a model. These attempts include that of Robert and Rousseau (2002) and involve p-values that are not strictly Bayesian, or that are not evaluated via a Bayesian perspective. We can therefore truly wonder whether or not it is possible to compare or even to define model complexity within the Bayesian paradigm. At a naïve level, an obvious answer is that we cannot, since we cannot look at a model without standing outside this model. At another level, however, we could answer positively, since tools like Bayes factors and even BIC are already available. But this is not really a less naïve answer! Plummer's alternative is thus interesting in this respect as (a) it does not depend on parameterisation and (b) it is a quantity that can be evaluated a posteriori. Its main drawbacks are that it does not necessarily relate to the original problem, and also that it uses the replica distribution rather than the predictive distribution, which has been advocated in Bayesdom as paramount; see for example van der Linde's discussion or Robert and Rousseau (2002). Also, this only defines a particular type of complexity (or of true dimension) but it does not allow for the comparison of models.

2 Complexity and focus

As noted in Plummer's discussion, an interesting point in Spiegelhalter et al. (2002) is the concept of *focus*. Missing data models clearly give rise to different types of focus, as stressed by both van der Linde and Meng and Vaida (Sections 4 and 5). This feature makes a big difference with ordinary models since possible focusses for missing data models are multifaceted and (much) more numerous than those of standard models, assuming that we do not introduce an articial level of completion!

We thus appreciate the different focusses proposed by Plummer, although they only apply in simple problems: as the hierarchy becomes more and more complex, the number of possible focusses simply explodes. They highlight the complex nature of the notion of complexity rather than truly solving the problem. Indeed, Plummer's empirical results are rather unhelpful, seeming;y not behaving satisfactorily as K increases. For instance, in Plummer's Figure 1, we could introduce a fourth focus where (μ, τ) would come down at the level of Z, even if this may be a completely artificial represention.

In the case of mixtures, this has the interesting effect of reminding us of the very different nature of p compared with both other parameters. As already stated in Celeux et al. (2000), some natural loss functions for mixture estimation simply omit the parameter p if for instance allocation is taken into account. There is therefore something delicate and indefinite about p. Note that in Table 2 of Plummer the expected p_D is strikingly close to $2K$ (excluding p then), except for $K = 3, 4$. The last column of Table 1 in Plummer's discussion is also intriguing: p_D and DIC move in such

a non-monotonic way that the argument about a simple-and-good-enough model vs. a complex-but-better-fitting model is far from convincing.

To answer van der Linde's question, the complexity of a predictive density is for us the complexity of the underlying model, since the degree(s) of complexity (in the posterior distribution) has been integrated out in the calculation of the predictive. (Think for instance of model averaging which is a *proper* Bayes solution: the weighted sum of predictive densities of different complexities has no well-defined complexity.) We also fail to see how DIC has brought a *"quantification of the reduction of model complexity due to the information in a prior"*, although this would suggest using instead Meng and Vaida's posterior version.

A question raised when reading the discussion is whether or not the nuisance parameters in a model are appropriately treated by DIC. In a sense, this is another type of problem where the definitions of p_D and DIC are unclear, the missing data taking the place of the nuisance parameters. Section 5 of Meng and Vaida's discussion as well as Plummer take alternative positions on that problem, and there are possibly many more others.

3 Plug-in estimates

Without going so far as to agree fully with Dawid's complete dismissal of DIC in his discussion of Spiegelhalter et al. (2002), we concede that using a *plug-in* estimate disqualifies the technique from being properly Bayesian. In the case of mixture models, the problem runs deeper since there is not even a clear-cut estimate without an associated loss function. (This difficulty with DIC is stressed both by Meng and Vaida and by Plummer.) If we want to keep using DIC, it seems that the Bayes estimate of the density is more appropriate for reasons stated in the original paper. If instead we use the predictive then another term should replace the plug-in.

Carlin's suggestion of replacing a plug-in degree of freedom by its posterior distribution is obviously most appealing from a Bayesian point of view, even though the implementation of this principle in a unified methodology may also be *"a few years away"*.

The way Plummer defines p_D is also sensible and the numerical illustrations for the galaxy benchmark dataset are of interest. However, for focus F3, the decrease in DIC for $K \geq 5$ is hard to explain: it could be related to numerical imprecision when deriving its p_D proposal. (We take the opportunity to address here Carlin's last comment about MCMC convergence. While we completely agree that non-identifiable settings are usually welcomed in terms of MCMC convergence, we are rather confident that our sampler has converged within the number of simulations we ran and thus that the exotic behaviour of some DIC_i's is not the result of lack of convergence.)

A puzzling part of Meng and Vaida's discussion is their Section 6, where they happily start mixing even further Bayesian and frequentist tools and objects! The fact that the (more convincing) posterior equivalent of p_D is not working as well is indeed quite

intriguing although Plummer somehow gives the hint of an answer in his first paragraph, namely that there are many ways of decomposing a joint distribution into $f(y|\theta)f(\theta)$, just as the number of missing data representations are infinite. First note that using the posterior instead of the likelihood in DIC is nominally Bayesian but not truly Bayesian as the concept is still frequentist. (The fact that p_D^B is constant in the example is not a difficulty per se: after all this really is a one-parameter problem and it is difficult to look at it otherwise.) There is also the issue that incorporating the prior into the complexity measure confounds the complexity due to the model with the complexity due to the prior and this is very confusing when different models are being compared because we need to use one prior for each model. The final part of Meng and Vaida's Section 6 also makes limited sense (to us at least) because of its systematic interweaving of Bayes and non-Bayes rules and concepts. The only conclusion we could derive from this part is that *ad hoc* criteria can breed even more criteria with seemingly the same validity, which is not necessarily the conclusion expected by the authors...

4 Missing data specifics

For missing data models and in particular for the mixture model, several discussants (Carlin, Meng and Vaida, Plummer) seem to prefer DIC_7 when the focus emphasizes the ability of the model to classify the observed data accurately into groups because, as noted by Carlin, this criterion treats $\mathbf{Z}$ and θ symmetrically. However, a potential default of DIC_7 is that it treats the missing data as parameters. Thus, the number of parameters to be estimated grows to infinity with the sample size for many models including the mixture model. Moreover, it can be remarked that in full Bayesian approaches of the mixture model (see Marin et al. 2005, for a recent survey) the $\mathbf{Z}$ are not treated as parameters (with a prior distribution) but as missing data. In this context, our favorite criterion remains DIC_4 even though this criterion is not invariant to the choice of $\mathbf{Z}$, as noted in the paper and as stressed by Plummer. In our opinion, this problem is essentially formal: when the focus is on imputing values for the missing data, the choice of $\mathbf{Z}$ does not suffer from any ambiguity from a practical point of view.

The point of the last section of Plummer's discussion about the missing or arbitrary function of the data Y was altogether missed by us, although it also replicates a statement in Spiegelhalter et al. (2002). We indeed have trouble in understanding why $f(y, z|\theta)$ is not defined exactly. Is this problem deeper than a mere measure-theoretic subtlety? We would also take issue with the last paragraph of this discussion in that we are not completely convinced that we should use *any* of the DIC's we examined!

5 Conclusion

It seems to us that, if DIC is to 'work' in general then the basic approach, in other words DIC_1 (or arguably DIC_2), should produce satisfactory results, since this is Spiegelhalter *et al.*'s (2002) criterion. In this paper, we have highlighted in some detail the problems in applying DIC beyond the exponential family case. Our goal was not to find a 'cure-

all', so that the existence of a generally-applicable measure remains an open question. In other words, the definition of a deviance information criterion, albeit immensely desirable, remains *ad hoc* at this stage and is not even close to being a well-defined ideal criterion or the solution of a well-defined optimisation problem. There is thus a need to reappraise its properties or to start afresh with a new deviance information criterion based on decision theoretic grounds.

References

Celeux, G., Hurn, M., and Robert, C. P. (2000). "Computational and inferential difficulties with mixture posterior distributions." *Journal of the American Statistical Association*, 95(3): 957–979. 701, 702

Marin, J., Mengersen, K., and Robert, C. (2005). "Bayesian Modelling and Inference on Mixtures of Distributions." In Rao, C. and Dey, D. (eds.), *Handbook of Statistics*, volume 25 (to appear). Springer-Verlag, New York. 704

Robert, C. P. and Rousseau, J. (2002). "A mixture approach to Bayesian goodness of fit." Technical Report 2002-9, Université Paris Dauphine. 702

Spiegelhalter, D. J., Best, N., B.P., C., and van der Linde, A. (2002). "Bayesian measures of model complexity and fit." *Journal of the Royal Statistical Society, Series B*, 64: 583–640. 702, 703, 704