Bayesian Analysis (2006) **1**, Number 4, pp. 681–686

Comment on Article by Celeux et al.

Martyn Plummer[*]

The Deviance Information Criterion (DIC) for model choice was introduced by Spiegelhalter et al. (2002) as a Bayesian analogue of the Akaike Information Criterion. The aim of DIC is not to identify the "true" probability model, but to find a parsimonious description of the data Y in terms of parameters θ. The parameters are of lower dimension than the data, either because θ is restricted to a low-dimensional subspace, or because it has a highly structured prior. A penalty function p_D measures the "effective number of parameters" of the model, and this is added to a measure of fit – the expected deviance – to give the DIC. Given a set of models, the one with the smallest DIC has the best balance between goodness of fit and model complexity.

DIC has received a mixed reception, as shown by the discussion of Spiegelhalter et al. (2002). On the one hand, it gives a pragmatic solution to the problem of model choice, and is now routinely available in the software WinBUGS (Spiegelhalter et al. 2004). On the other hand, a number of technical and conceptual difficulties with the criterion remain. Celeux et al. (2006) investigate these difficulties in the context of missing data models, and in particular mixture models. They have produced 8 variations on the theme of DIC. Some of these variations address the problem of finding a good "plug-in" estimate of θ, which is necessary for the calculation of the penalty p_D. Others are innovations that provide a way of calculating DIC in missing data models, which might otherwise be intractable. I have attempted to classify these criteria according to their level of "focus".

1 Focus

The concept of focus is fundamental to understanding DIC, since DIC is not a global evaluation of the model, but of a particular partition $f(y|\theta)f(\theta)$. In a hierarchical model, there may be a multitude of choices for θ, so this choice – the focus of DIC – must be made explicit. Figure 1 shows the directed acyclic graph defined by the mixture model considered in section 5 of Celeux et al. (2006). Any edge cut of this graph defines a partition of the model into a "likelihood" part $f(y|\theta)$ and a "prior" part $f(\theta)$. Figure 1 shows 3 of the 11 possible cuts.

At the lowest level of focus, corresponding to cut F_1, θ is empty and the likelihood term is the predictive distribution $\widehat{f}(y) = \mathrm{E}_{\theta|Y}(f(y|\theta))$. This level of focus is appropriate if the mixture components have no physical meaning, but are just a convenient semi-parametric way of describing the distribution $f(y)$. At a higher level, F_2, the focus is on the mean μ_k, standard deviation σ_k, and probability p_k of the mixture component k for $k = 1 \ldots K$. This is the focus of the "observed DICs" (DIC$_1$ and DIC$_2$) of Celeux et al. (2006). DIC$_3$ does not have an unambiguous focus, as it has elements of both F_1 and

[*] International Agency for Research on Cancer, Lyon, France, http://www-ice.iarc.fr/~martyn/

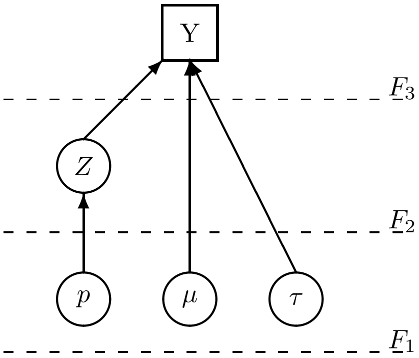

Figure 1: Directed acyclic graph of a normal mixture model, showing different levels of focus for the DIC

F_2. It will not be considered further here.

A third level of focus, F_3, uses indicator variables Z_i, which identify the group that each observation belongs to. This level of focus – which is also used by DIC$_7$ of Celeux et al. (2006) – emphasizes the ability of the model to accurately classify the observed data into groups, and not just to characterize the population they were drawn from.

2 An alternative penalty function

In the discussion of Spiegelhalter et al. (2002), I suggested an alternative definition of p_D that does not require a plug-in estimate (Plummer 2002). This is based on the Kullback-Leibler information divergence between between the predictive distributions at two different values of θ

$$I(\theta^{(0)}, \theta^{(1)}) = \mathrm{E}_{[Y_{rep} | \theta^{(0)}]} \left\{ \log \left(\frac{f(Y_{rep} \mid \theta^{(0)})}{f(Y_{rep} \mid \theta^{(1)})} \right) \right\},$$

The penalty p_D can be defined as the expected value of $I(\theta^{(0)}, \theta^{(1)})$ when $\theta^{(0)}$, $\theta^{(1)}$ are independent samples from the posterior distribution of θ. For linear mixed model, this expression is identical to the p_D of Spiegelhalter et al. (2002). It also generalizes easily to other models for which p_D is not easily calculated. Although this definition of p_D lacks a formal derivation, it is interesting to consider in the current context as one of many possible generalizations of DIC for mixture models. It has some useful properties: it is always non-negative and is independent of the coordinates of θ. It also provides an interpretation of p_D as a penalty for inconsistency. A model is penalized if it gives high posterior probabilities to two different values of θ that give inconsistent predictions for Y.

To my knowledge, this definition of p_D has never been used in practice. I took the opportunity to examine its empirical behaviour. Table 1 shows results for the galaxy data set, for the three levels of focus considered in figure 1. The DICs have been standardized so that the DIC for the best fitting model is zero. For each model, the default prior of Richardson and Green (1997) was used. The penalty p_D was estimated using two parallel MCMC chains.

With focus F_1, the penalty term is zero for all models. Although the model with K=7 groups is, in fact, more complex than the model with K=2, this complexity appears in the likelihood, and is ignored by DIC, which only takes account of complexity in the prior. Unsurprisingly, the model with the largest number of components gives the best fit to the data, although there are clearly diminishing returns from adding extra components.

When the focus is at level F_2, the penalty p_D increases quite rapidly with the number of components. Consequently, the model with $K = 3$ is strongly favoured. This DIC behaves similarly to DIC$_2$ of Celeux et al. (2006) up to $K = 5$, but penalizes models $K = 6$ and $K = 7$ more strongly.

When the focus is at level F_3, the penalty p_D is not monotonic in K. It drops dramatically for $K = 3$, increases to a peak at $K = 5$ and then diminishes again slightly. The small value of p_D for $K = 3$ may appear surprising, but it is consistent with the notion of p_D as a penalty for inconsistent prediction. Inspection of figure 1 of Celeux et al. (2006), which shows a histogram of the galaxy data, shows that the galaxies can be divided by eye into three groups, a central mass and two small outlying groups: one around $10 \times 10^6 m/s$ and the other around $32 \times 10^6 m/s$. The three-component model unambiguously classifies the observed galaxies into these three groups, as a result of which the predictions for Y_i given Z_i are quite consistent between different draws from the posterior distribution of Z. The decrease in DIC for $K \geq 5$ is somewhat harder to explain. In fact, the DIC for an 8-component model is smaller than the DIC for the 3-component model (data not shown). One is left with a choice between a simple model with poor fit, and a complex model with good fit. Of course, this is exactly the choice that DIC was designed to resolve, but the differences in this case are extreme. The measures of fit and complexity both change by about 100, and it is not clear that they are both commensurable over such a wide range. Although DIC$_7$ has the same focus, it behaves quite differently.

3 DIC with missing data

The remaining DICs considered by Celeux et al. (2006) are extensions of the original concept of DIC. They are designed for missing data problems, in which the likelihood $f(y|\theta)$ may not be available in closed form and hence for which DIC cannot be calculated. Again, these can be distinguished by their level of focus: the "complete DICs" with focus F_2 and the "conditional DICs" with focus F_3. I believe there is an ambiguity in the definition of the complete DICs, due to the fact that the log likelihood $\log(f(y, z|\theta))$ is only defined up to an arbitrary function of the data. When Y, Z are both observed, this is not a problem, as *differences* between DICs are well defined, even if the absolute values depend on the underlying measure used to define the density f. However, when Z is missing, this is no longer true. The difference between values of $\mathrm{E}_{Z|Y}(\log(f(y,z|\theta)))$ for two different models will, in general, be sensitive to the underlying measure, since the predictive distributions $f(Z|Y)$ will be different. This may not be an insurmountable problem, but is an extra complication not shared by other DICs.

The "conditional DIC", DIC_8, based on $f(y|z,\theta)$ appears particularly useful for models that can be factorized as

$$f(Y|Z,\theta)f(\theta)f(Z|\varphi)f(\varphi)$$

and where the focus of interest is on θ. When Z is observed, inference on θ is independent of φ. But when Z is missing, φ is a nuisance parameter that must be eliminated. DIC_8 eliminates φ using a two-step procedure: first calculate what the value that DIC would take if Z were observed, using the sub-model $f(Y|Z,\theta)f(\theta)$, then calculate the expected value of this DIC using the posterior distribution of Z.

Table 2 shows results for the galaxy data set using this approach, and the definition of p_D based on the information divergence. The mixture model factorizes, as above, with $\theta = \{\mu, \tau\}$ and $\varphi = p$. The estimates were calculated by drawing 2000 samples from the posterior distribution of Z, and then calculating DIC for each sampled value pretending that it was observed.

The expected deviance is the same as for the standard DIC, with focus F_3. Small differences in $\overline{D}$ between table 1 and 2 are due to sampling error. The penalty term is, however, much lower in table 2, and so this DIC apparently favours a model with more components. However, this may not be a valid comparison. If we could, in fact, observe Z, then we would know exactly how many components were represented in the data. There would then be no need to compare models with different numbers of components, as in table 2.

I would suggest that this approach to DIC should be limited to situations that satisfy two criteria. Firstly, that Z should be observable in principle, so there is no ambiguity over whether Z is a construct of the model or a platonic true score. Secondly, all models under comparison should use the same sub-model $f(z|\varphi)f(\varphi)$. Neither assumption holds for the current example.

In summary Celeux et al. (2006) provide a useful extension of DIC for missing data models. If used with caution, this could extend the application of DIC to situations in which it could not, otherwise, be calculated. The diverse behaviour of different generalizations of DIC is notable. Ultimately, there is a limit to our ability to judge DIC by its empirical behaviour, and these difficulties underscore the lack of a solid theoretical foundation for DIC and its derived measures. More work is required in this direction.

K	Focus								
	F_1			F_2 $\{\mu, \sigma, p\}$			F_3 $\{\mu, \sigma, Z\}$		
	$\overline{D}$	p_D	DIC	$\overline{D}$	p_D	DIC	$\overline{D}$	p_D	DIC
2	442.1	0	46.2	445.8	9.2	16.5	405.5	67.5	116.2
3	412.5	0	16.6	418.0	20.5	0.0	343.0	13.8	0.0
4	403.2	0	7.3	412.2	35.2	9.0	306.1	106.7	56.0
5	398.2	0	2.3	408.4	42.4	12.4	271.9	132.2	47.3
6	396.5	0	0.6	406.9	72.6	41.0	250.8	128.1	22.1
7	395.9	0	0.0	406.4	97.5	65.4	236.9	123.8	3.9

Table 1: Results for the galaxy data set at different levels of focus: expected deviance $\overline{D}$, penalty p_D, and DIC relative to best fitting model.

K	$\mathrm{E}_{Z\mid Y}(\overline{D})$	$\mathrm{E}_{Z\mid Y}(p_D)$	$\mathrm{E}_{Z\mid Y}(DIC)$
2	405.7	4.1	159.2
3	343.1	8.1	100.6
4	305.4	9.2	64.1
5	272.0	10.9	32.4
6	251.1	12.5	13.1
7	236.4	14.1	0.0

Table 2: Expected DIC for the galaxy data set if Z were observed, with focus $F_3(\mu, \sigma, Z)$.

References

Celeux, G., Forbes, F., Robert, C. P., and Titterington, D. M. (2006). "Deviance information criteria for missing data models." *Bayesian Analysis (in press)*. 681, 682, 683, 684

Plummer, M. (2002). "Discussion of Speigelhalter et al." *Journal of the Royal Statistical Society - Series B*, 64: 620. 682

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. 683

Spiegelhalter, D. J., Best, N. G., Carlin, B. R., and van der Linde, A. (2002). "Bayesian measures of complexity and fit." *Journal of the Royal Statistical Society - Series B*, 64: 583–616. 681, 682

Spiegelhalter, D. J., Thomas, A., Best, N., and Lunn, D. (2004). *WinBUGS User Manual, Version 2.0*. 681