Deviance Information Criteria for Missing Data Models

G. Celeux[1], F. Forbes[2], C.P. Robert[3]*, and D.M. Titterington[4]†

[1]*INRIA FUTURS, Orsay,* [2]*INRIA Rhône-Alpes,*

[3]*CREST and CEREMADE, Uni. Paris Dauphine,*

and [4]*University of Glasgow*

Abstract

The deviance information criterion (DIC) introduced by Spiegelhalter et al. (2002) for model assessment and model comparison is directly inspired by linear and generalised linear models, but it is open to different possible variations in the setting of missing data models, depending in particular on whether or not the missing variables are treated as parameters. In this paper, we reassess the criterion for such models and compare different DIC constructions, testing the behaviour of these various extensions in the cases of mixtures of distributions and random effect models.

Keywords: completion, deviance, DIC, EM algorithm, MAP, model comparison, mixture model, random effect model.

1 Introduction

When developing their theory of the *deviance information criterion* (DIC) for the assessment and comparison of models, Spiegelhalter et al. (2002) mostly

*Research partially conducted while visiting the University of Glasgow in Feb. 2002, 2003, and 2005. The author wishes to thank the Department of Statistics for its continuing hospitality and warm welcome over the years!

†Research partially conducted while visiting INRIA Rhône-Alpes in the Spring and Autumn of 2002. The author wishes to thank INRIA for its hospitality and its support.

focussed on the case of generalised linear models, although they concluded their seminal paper with a discussion of the possibilities of extending this notion to models like mixtures of distributions. The ensuing discussion in the *Journal of the Royal Statistical Society* pointed out the possible difficulties of defining DIC precisely in these scenarios. In particular, DeIorio and Robert (2002) described some possible inconsistencies in the definition of a DIC for mixture models, while Richardson (2002) presented an alternative notion of DIC, again in the context of mixture models.

The fundamental versatility of the DIC criterion is that, in hierarchical models, basic notions like parameters and *deviance* may take several equally acceptable meanings, with direct consequences for the properties of the corresponding DICs. As pointed out in Spiegelhalter et al. (2002), this is not a problem *per se* when parameters of interest (or a "focus") can be identified but this is not always the case when practitioners compare models. The diversity of the numerical answers associated with the different focusses is then a real difficulty of the method. As we will see, these different choices can produce quite distinct evaluations of the effective dimension p_D that is central to the DIC criterion. (Although this is not in direct connection with our missing data set-up, nor with the DIC criterion, note that Hodges and Sargent (2001) also describes the derivation of degrees of freedom in loosely parameterised models.)

There is thus a need to evaluate and compare the properties of the most natural choices of DICs: The present paper reassesses the definition and connection of various DIC criteria for missing data models. In Section 2, we recall the notions introduced in Spiegelhalter et al. (2002). Section 3 presents a typology of the possible extensions of DIC for missing data models, while Section 4 constructs and compares these extensions in random effect models, and Section 5 does the same for mixtures of distributions. We conclude the paper with a discussion of the relevance of the various extensions in Section 6.

2 Bayesian measures of complexity and fit

For competing parametric statistical models, $f(\mathbf{y}|\theta)$, the construction of a generic model-comparison tool is a difficult problem with a long history. In particular, the issue of devising a selection criterion that works both as a measure of fit and as a measure of complexity is quite challenging. In this paper, we examine solely the criteria developed in Spiegelhalter et al. (2002), including, in connection with model complexity, their measure, p_D, of the effective number of parameters in a model. We refer the reader to this paper, the ensuing discussion and to Hodges and Sargent (2001) for further references. This quantity is based on a *deviance*, defined by

$$D(\theta) = -2\log f(\mathbf{y}|\theta) + 2\log h(\mathbf{y})\ ,$$

where $h(\mathbf{y})$ is some fully specified standardizing term which is function of the data alone. Then the *effective dimension* p_D is defined as

$$p_D = \overline{D(\theta)} - D(\tilde{\theta})\,, \tag{1}$$

where $\overline{D(\theta)}$ is the posterior mean deviance,

$$\overline{D(\theta)} = \mathbb{E}_\theta[-2\log f(\mathbf{y}|\theta)|\mathbf{y}] + 2\log h(\mathbf{y}),$$

which can be regarded as a Bayesian measure of fit, and $\tilde{\theta}$ is an estimate of θ depending on $\mathbf{y}$. The posterior mean $\overline{\theta} = \mathbb{E}[\theta|\mathbf{y}]$ is often a natural choice for $\tilde{\theta}$ but the posterior mode or median can also be justified as an alternative. Note that p_D is completely independent of the choice of the standardizing h. As explicitly pointed out in Spiegelhalter et al. (2002), the fact that p_D does depend on the choice of the estimate $\tilde{\theta}$ and more generaly on the parameterisation of the model is one of the difficulties of this approach that can only be solved when there is a clear "focus" on the parameter of interest in the model. In the event of a global model comparison where no particular parameterisation enjoys a special position, as for instance in the comparison of the number of components in a mixture of distributions, there is no intrinsic definition to the dimension p_D.

A corresponding *Deviance Information Criterion* (DIC) for model comparison is advanced by Spiegelhalter et al. (2002) from this construction:

$$\begin{aligned}
\text{DIC} &= \overline{D(\theta)} + p_D \\
&= D(\tilde{\theta}) + 2p_D \\
&= 2\overline{D(\theta)} - D(\tilde{\theta}) \\
&= -4\mathbb{E}_\theta[\log f(\mathbf{y}|\theta)|\mathbf{y}] + 2\log f(\mathbf{y}|\tilde{\theta}).
\end{aligned} \tag{2}$$

For model comparison, we need to set $h(\mathbf{y}) = 1$, for all models, so we take

$$D(\theta) \quad = \quad -2\log f(\mathbf{y}|\theta). \tag{3}$$

Provided that $D(\theta)$ is available in closed form, $\overline{D(\theta)}$ can easily be approximated using an MCMC run by taking the sample mean of the simulated values of $D(\theta)$. (If $f(\mathbf{y}|\theta)$ is *not* available in closed form, as is often the case for missing data models like (4), further simulations can provide a converging approximation or, as we will see below, can be exploited directly in alternative representations of the likelihood and of the DIC criterion.) When $\overline{\theta} = \mathbb{E}[\theta|\mathbf{y}]$ is used, $D(\overline{\theta})$ can also be approximated by plugging in the sample mean of the simulated values of θ. As pointed out by Spiegelhalter et al. (2002), this choice of $\tilde{\theta}$ ensures that p_D is positive when the density is log-concave in θ, but it is not appropriate when θ is discrete-valued since $\mathbb{E}[\theta|\mathbf{y}]$ usually fails to take one of these values. Also, the effective dimension p_D may well be negative for models outside the log-concave densities. We will discuss further the issue of choosing (or not choosing) $\tilde{\theta}$ in the following sections.

3 DICs for missing data models

In this section, we describe alternative definitions of DIC in missing data models, that is, when

$$f(\mathbf{y}|\theta) = \int f(\mathbf{y}, \mathbf{z}|\theta) d\mathbf{z}\,, \tag{4}$$

by attempting to write a typology of natural DICs in such settings. Missing data models thus involve variables $\mathbf{z}$ which are non-observed, or missing, in addition to the observed variables $\mathbf{y}$. There are numerous occurrences of such models in theoretical and practical Statistics and we refer to Little and Rubin (1987), McLachlan and Krishnan (1997) and Cappé et al. (2005) for different accounts of the topic. Whether or not the missing data $\mathbf{z}$ are truly meaningful for the problem is relevant for the construction of the DIC criterion because the focus of inference may be on the parameter θ, the pair (θ, z) or on z only, as in classification problems.

The observed data associated with this model will be denoted by $\mathbf{y} = (y_1, \ldots, y_n)$ and the corresponding *missing data* by $\mathbf{z} = (z_1, \ldots, z_n)$. Following the EM terminology, the likelihood $f(\mathbf{y}|\theta)$ is often called the *observed likelihood* while $f(\mathbf{y}, \mathbf{z}|\theta)$ is called the *complete likelihood*. We will use as illustrations of such models the special cases of random effect and mixture models in Sections 4 and 5.

3.1 Observed DICs

A first category of DICs is associated with the observed likelihood, $f(\mathbf{y}|\theta)$, under the assumption that it can be computed in closed form (which is for instance the case for mixture models but does not always hold for hidden Markov models). While, from (3),

$$\overline{D(\theta)} = -2\mathbb{E}_\theta\left[\log f(\mathbf{y}|\theta)|\mathbf{y}\right]$$

is clearly and uniquely defined, even though it may require (MCMC) simulation to be computed approximately, choosing $\tilde{\theta}$ is more delicate and the definition of the second term $D(\tilde{\theta})$ in (1) is not unique.

In fact, within missing data models like the mixture model of Section 5, the parameters θ are not always *identifiable* and the posterior mean $\overline{\theta} = \mathbb{E}_\theta[\theta|\mathbf{y}]$ can then be a very poor estimator. For instance, in the mixture case, if both prior and likelihood are invariant with respect to the labels of the components, all marginals (in the components) are the same, all posterior means are identical, and the plug-in mixture then collapses to a single-component mixture (Celeux et al., 2000). As a result,

$$\mathrm{DIC}_1 = -4\mathbb{E}_\theta\left[\log f(\mathbf{y}|\theta)|\mathbf{y}\right] + 2\log f(\mathbf{y}|\mathbb{E}_\theta\left[\theta|\mathbf{y}\right])$$

is often not a good choice. For instance, in the mixture case, DIC_1 quite often leads to a negative value for p_D. (The reason for this is that, even under an identifiability constraint, the posterior mean borrows from several modal regions of the posterior density and ends up with a value that is located between modes, see also the discussion in Marin et al., 2005.)

A more relevant choice for $\tilde{\theta}$ is the posterior mode or modes,

$$\widehat{\theta}(\mathbf{y}) = \arg\max_\theta f(\theta|\mathbf{y}),$$

since this depends on the posterior distribution of the whole vector θ, rather than on the marginal posterior distributions of its elements as in the mixture case. This leads to the alternative "observed" DIC

$$\mathrm{DIC}_2 = -4\mathbb{E}_\theta\left[\log f(\mathbf{y}|\theta)|\mathbf{y}\right] + 2\log f(\mathbf{y}|\widehat{\theta}(\mathbf{y})).$$

Recall that, for the K-component mixture problem, there exist a multiple of $K!$ marginal posterior modes. Note also that, when the prior on θ is uniform, so that $\widehat{\theta}(\mathbf{y})$ is also the maximum likelihood estimator, which can be computed by the EM algorithm, the corresponding p_D,

$$p_D = -2\mathbb{E}_\theta \left[\log f(\mathbf{y}|\theta)|\mathbf{y}\right] + 2\log f(\mathbf{y}|\widehat{\theta}(\mathbf{y}))\,,$$

is necessarily positive. However, positivity does not always hold for other prior distributions, even though it is asymptotically true when the Bayes estimator is convergent.

When non-identifiability is endemic, as in mixture models, the parameterisation by θ of the model $f(\mathbf{y}|\theta)$ is often not relevant and the inferential focus is mostly on the density itself. In this setting, a more natural choice for $D(\tilde{\theta})$ is to select an estimator $\widehat{f}(\mathbf{y})$ of the density $f(\mathbf{y}|\theta)$, since this function is invariant under permutation of the component labels. For instance, one can use the posterior expectation $\mathbb{E}_\theta\left[f(\mathbf{y}|\theta)|\mathbf{y}\right]$. (Note that this is also independent of the representation (4).) In terms of functional estimation, this approach provides stable evaluations that are superior to the plug-in estimates $f(\mathbf{y}|\widehat{\theta})$; furthermore, the density estimator is easily approximated by an MCMC evaluation. For instance, for a Gaussian mixture with density

$$f(y|\theta) = \sum_{i=1}^{K} p_i \phi(y|\mu_i, \sigma_i^2),$$

we have

$$\widehat{f}(y) = \frac{1}{m}\sum_{l=1}^{m}\sum_{i=1}^{K} p_i^{(l)}\phi(y|\mu_i^{(l)}, \sigma_i^{2(l)}) \approx \mathbb{E}_\theta[f(y|\theta)|\mathbf{y}]\,,$$

where $\phi(y|\mu_i, \sigma_i^2)$ denotes the density of the normal $\mathcal{N}(\mu_i, \sigma_i^2)$ distribution, $\theta = \{p, \mu, \sigma^2\}$ with $\mu = (\mu_1, \ldots, \mu_K)^t$, $\sigma^2 = (\sigma_1^2, \ldots, \sigma_K^2)^t$ and $p = (p_1, \ldots, p_K)^t$, m denotes the number of MCMC simulations and $(p_i^{(l)}, \mu_i^{(l)}, \sigma_i^{2(l)})_{1\le i\le m}$ is the result of the l-th MCMC iteration. This is also the MCMC predictive density, and this leads to another criterion,

$$\mathrm{DIC}_3 = -4\mathbb{E}_\theta\left[\log f(\mathbf{y}|\theta)|\mathbf{y}\right] + 2\log \widehat{f}(\mathbf{y})\,,$$

where $\widehat{f}(\mathbf{y}) = \prod_{i=1}^n \widehat{f}(y_i)$. Note that this is also the proposal of Richardson (2002) in her discussion of Spiegelhalter et al. (2002). This is quite a sensible alternative, since the predictive distribution is quite central to Bayesian inference. (See, for instance, the notion of Bayes factors, which are ratios of predictives, Robert, 2001.) Note however that the relative values of $\widehat{f}(\mathbf{y})$, for different models, also constitute the "posterior Bayes factors" of Aitkin (1991) which came under strong criticism in the ensuing discussion.

3.2 Complete DICs

The missing data structure makes available many alternative representations of the DIC, by reallocating the positions of the log and of the various expectations. This is not simply a formal exercise: missing data models offer a wide variety of interpretations depending on the chosen representation for the missing data structure.

We can first note that, using the complete likelihood $f(\mathbf{y}, \mathbf{z}|\theta)$, we can set $\overline{D(\theta)}$ as the posterior expected value (over the missing data) of the joint deviance,

$$\begin{aligned}\overline{D(\theta)} &= -2\mathbb{E}_\theta \left\{ \mathbb{E}_\mathbf{Z} \left[\log f(\mathbf{y}, \mathbf{Z}|\theta) | \mathbf{y}, \theta \right] | \mathbf{y} \right\} \\ &= -2\mathbb{E}_\mathbf{Z} \left\{ \mathbb{E}_\theta \left[\log f(\mathbf{y}, \mathbf{Z}|\theta) | \mathbf{y}, \mathbf{Z} \right] | \mathbf{y} \right\} \\ &= -2\mathbb{E}_{\theta,\mathbf{Z}} \left[\log f(\mathbf{y}, \mathbf{Z}|\theta) | \mathbf{y} \right].\end{aligned}$$

In addition to the difficulty of choosing $\tilde{\theta}$, already encountered in the previous section, we now have the problem of defining the fixed point deviance, $D(\tilde{\theta})$, in connection with the missing data structure. Using the same motivations as for the EM algorithm (McLachlan and Krishnan, 1997), we can first define a complete data DIC, by defining the complete data estimator $\mathbb{E}_\theta[\theta|\mathbf{y}, \mathbf{z}]$, which does not suffer from identifiability problems since the components are identified by $\mathbf{z}$, and then obtain DIC for the complete model as

$$\mathrm{DIC}(\mathbf{y}, \mathbf{z}) = -4\mathbb{E}_\theta \left[\log f(\mathbf{y}, \mathbf{z}|\theta) | \mathbf{y}, \mathbf{z} \right] + 2 \log f(\mathbf{y}, \mathbf{z}|\mathbb{E}_\theta[\theta|\mathbf{y}, \mathbf{z}]).$$

As in the EM algorithm, we can then integrate this quantity to define

$$\begin{aligned}\mathrm{DIC}_4 &= \mathbb{E}_\mathbf{Z} \left[\mathrm{DIC}(\mathbf{y}, \mathbf{Z}) | \mathbf{y} \right] \\ &= -4\mathbb{E}_{\theta,\mathbf{Z}} \left[\log f(\mathbf{y}, \mathbf{Z}|\theta) | \mathbf{y} \right] + 2\mathbb{E}_\mathbf{Z} \left[\log f(\mathbf{y}, \mathbf{Z}|\mathbb{E}_\theta[\theta|\mathbf{y}, \mathbf{Z}]) | \mathbf{y} \right].\end{aligned}$$

This requires the computation of a posterior expectation for each value of Z, but this is usually not difficult as the complete model is often chosen for its simplicity.

A second solution that integrates the notion of "focus" defended in Section 2.1 of Spiegelhalter et al. (2002) is to consider $\mathbf{Z}$ as an additional parameter (of interest) rather than as a missing variable and to use a pivotal quantity $D(\tilde{\theta})$ based on estimates of both $\mathbf{z}$ and θ; that is, informally,

$$D(\tilde{\theta}) = -2 \log f(\mathbf{y}, \widehat{\mathbf{z}}(\mathbf{y}) | \widehat{\theta}(\mathbf{y})).$$

Once again, we must stress that, in missing data problems like the mixture model, the choices for these estimators are quite delicate as the expectations of $\mathbf{Z}$, given $\mathbf{y}$, are poor estimators, being for instance all identical under exchangeable priors (see Section 5.2) and, besides, most often taking values outside the support of $\mathbf{Z}$, as in the mixture case. For this purpose, the only relevant estimator $(\widehat{\mathbf{z}}(\mathbf{y}), \widehat{\theta}(\mathbf{y}))$ in this setting seems thus to be the *joint* maximum a posteriori (MAP) estimator of the pair $(\mathbf{z}, \theta)$, given $\mathbf{y}$, unless one is ready to define a proper loss function as in Celeux et al. (2000) which somehow contradicts the initial purpose of DIC since a loss function should also integrate the model choice aspects in that case. Note that, in the event of non-identifiability or simply multimodality, the MAP estimates are not unique but they are all equivalent. In the case of K-component mixtures, choosing one MAP estimate is then equivalent to selecting one of the $K!$ possible component orderings.

Given that this estimator is not available in closed form, we can choose to estimate it by using the best–in terms of the values of the posterior distribution proportional to $f(\mathbf{y}, \mathbf{z}|\theta)f(\theta)$–pair that arose during the MCMC iterations. Note that the missing data structure is usually chosen so that the joint distribution $f(\mathbf{y}, \mathbf{z}|\theta)$ is available in closed form. Thus, even if the MAP estimate cannot be derived analytically, the values of $f(\mathbf{y}, \mathbf{z}|\theta)f(\theta)$ at the simulated pairs $(\mathbf{z}, \theta)$ can be computed.

The DIC corresponding to this analysis is then

$$\mathrm{DIC}_5 = -4\mathbb{E}_{\theta,\mathbf{Z}}\left[\log f(\mathbf{y}, \mathbf{Z}|\theta)|\mathbf{y}\right] + 2\log f(\mathbf{y}, \widehat{\mathbf{z}}(\mathbf{y})|\widehat{\theta}(\mathbf{y}))\,,$$

which, barring a poor MCMC approximation to the MAP estimate, should lead to a positive effective dimension,

$$p_{D5} = -2\mathbb{E}_{\theta,\mathbf{Z}}\left[\log f(\mathbf{y}, \mathbf{Z}|\theta)|\mathbf{y}\right] + 2\log f(\mathbf{y}, \widehat{\mathbf{z}}(\mathbf{y})|\widehat{\theta}(\mathbf{y}))\,,$$

given that, under a flat prior, the second part of DIC_5 is the maximum of the function integrated in the first part over θ and $\mathbf{Z}$. Note however that DIC_5 is somewhat inconsistent in the way it takes $\mathbf{z}$ into account. The posterior deviance, that is, the first part of DIC_5, incorporates $\mathbf{z}$ as missing variables while $D(\tilde{\theta})$ and therefore p_{D5} regard $\mathbf{z}$ as an additional parameter (see Sections 4 and 5 for illustrations).

Another interpretation of the posterior deviance and a corresponding DIC can be directly derived from the EM analysis of the missing data model.

Recall (Dempster et al., 1977; McLachlan and Krishnan, 1997; Robert and Casella, 2001, §5.3.3) that the core function of the EM algorithm is

$$Q(\theta|\mathbf{y},\theta_0) = \mathbb{E}_{\mathbf{Z}}\left[\log f(\mathbf{y},\mathbf{Z}|\theta)|\mathbf{y},\theta_0\right] ,$$

where θ_0 represents the "current" value of θ, and $Q(\theta|\mathbf{y},\theta_0)$ is maximised over θ in the "M" step of the algorithm, to provide the following "current" value θ_1. The function Q is usually easily computable, as for instance in the mixture case. Therefore, another natural choice for $D(\tilde{\theta})$ is to take

$$D(\tilde{\theta}) = -2Q(\widehat{\theta}(\mathbf{y})|\mathbf{y},\widehat{\theta}(\mathbf{y})) = -2\mathbb{E}_{\mathbf{Z}}[\log f(\mathbf{y},\mathbf{Z}|\widehat{\theta}(\mathbf{y}))|\mathbf{y},\widehat{\theta}(\mathbf{y})] ,$$

where $\widehat{\theta}(\mathbf{y})$ is an estimator of θ based on $f(\theta|\mathbf{y})$, such as the marginal MAP estimator, or, maybe more naturally, a fixed point of Q, such as an EM maximum likelihood estimate. This choice leads to a corresponding DIC

$$\mathrm{DIC}_6 = -4\mathbb{E}_{\theta,\mathbf{Z}}\left[\log f(\mathbf{y},\mathbf{Z}|\theta)|\mathbf{y}\right] + 2\mathbb{E}_{\mathbf{Z}}[\log f(\mathbf{y},\mathbf{Z}|\widehat{\theta}(\mathbf{y}))|\mathbf{y},\widehat{\theta}(\mathbf{y})] .$$

As for DIC_4, this strategy is consistent in the way it regards $\mathbf{Z}$ as missing information rather than as an extra parameter, but it is not guaranteed to lead to a *positive* effective dimension p_{D6}, as the maximum likelihood estimator gives the maximum of

$$\log \mathbb{E}_{\mathbf{Z}}\left[f(\mathbf{y},\mathbf{Z}|\theta)|\mathbf{y},\theta\right]$$

rather than of

$$\mathbb{E}_{\mathbf{Z}}\left[\log f(\mathbf{y},\mathbf{Z}|\theta)|\mathbf{y},\theta\right]$$

the latter of which is smaller since log is a concave function. An alternative to the maximum likelihood estimator would be to choose $\widehat{\theta}(\mathbf{y})$ to maximise $Q(\theta|\mathbf{y},\theta)$, which represents a more challenging problem, not addressed by EM unfortunately.

3.3 Conditional DICs

A third category of constructions of DICs in the context of missing variable models is to adopt a different inferential focus and consider $\mathbf{z}$ as an additional parameter. The DIC can then be based on the conditional likelihood, $f(\mathbf{y}|\mathbf{z},\theta)$. This approach has obvious asymptotic and coherency difficulties, as discussed in previous literature (Marriott, 1975; Bryant and Williamson, 1978; Little and Rubin, 1983), but it is computationally feasible and can thus be compared with the other solutions above. (Note in addition that DIC_5 is situated between the complete and the conditional approaches in that it uses the complete likelihood but similarly estimates $\mathbf{z}$. As we will see below, it appears quite naturaly as an extension of DIC_7.)

A natural solution in this framework is then to apply the original definition of DIC to the conditional distribution, which leads to

$$\mathrm{DIC}_7 = -4\mathbb{E}_{\theta,\mathbf{Z}}\left[\log f(\mathbf{y}|\mathbf{Z},\theta)|\mathbf{y}\right] + 2\log f(\mathbf{y}|\widehat{\mathbf{z}}(\mathbf{y}),\widehat{\theta}(\mathbf{y}))\,,$$

where again the pair $(\mathbf{z},\theta)$ is estimated by the joint MAP estimator, $(\widehat{\mathbf{z}}(\mathbf{y}),\widehat{\theta}(\mathbf{y}))$ approximated by MCMC. This approach leads to a positive effective dimension p_{D7}, under a flat prior for $\mathbf{z}$ and θ, for the same reasons as for DIC_5.

Note that there is a strong connection between DIC_5 and DIC_7 in that

$$\mathrm{DIC}_5 = \mathrm{DIC}_7 + \left\{-4\mathbb{E}_{\theta,\mathbf{Z}}\left[\log f(\mathbf{Z}|\mathbf{y},\theta)|\mathbf{y}\right] + 2\log f(\widehat{\mathbf{z}}(\mathbf{y})|\mathbf{y},\widehat{\theta}(\mathbf{y}))\,,\right\}$$

the additional term being similar to the difference between $Q(\theta|\mathbf{y},\theta_0)$ and the observed log-likelihood in the EM algorithm. The difference between DIC_5 and DIC_7 is not necessarily positive even though it appears as a DIC on the conditional distribution, given that $D(\tilde{\theta})$ is evaluated at the joint MAP estimate.

An alternative solution is to separate θ from $\mathbf{Z}$, taking once more the missing data perspective, as in DIC_4; that is, to condition first on $\mathbf{Z}$ and then integrate over $\mathbf{Z}$ conditional on $\mathbf{y}$, giving

$$\mathrm{DIC}_8 = -4\mathbb{E}_{\theta,\mathbf{Z}}\left[\log f(\mathbf{y}|\mathbf{Z},\theta)|\mathbf{y}\right] + 2\mathbb{E}_{\mathbf{Z}}\left[\log f(\mathbf{y}|\mathbf{Z},\widehat{\theta}(\mathbf{y},\mathbf{Z}))|\mathbf{y}\right]\,,$$

where $\widehat{\theta}(\mathbf{y},\mathbf{z})$ is an estimator of θ based on $f(\mathbf{y},\mathbf{z}|\theta)$, such as the posterior mean (which is now a correct estimator because it is based on the joint distribution) or the MAP estimator of θ (conditional on both $\mathbf{y}$ and $\mathbf{z}$). Here $\mathbf{Z}$ is dealt with as missing variables rather than as an additional parameter. The simulations in Section 5.5 illustrate that DIC_8 actually behaves differently from DIC_7 when estimating the complexity through p_D.

4 Random effect models

In this section we list the various DICs in the context of a simple random effect model. Some of the details of the calculations are not given here but are available from the authors. The model was discussed in Spiegelhalter et al. (2002), but here we set it up as a missing data problem, with the random effects regarded as missing values, because computations are feasible in closed form for this model and allow for a better comparison of the different DICs. More specifically, we focus on the way these criteria account for complexity, i.e. on the p_Ds, since there is no real model-selection issue in this setting.

Suppose therefore that, for $i = 1, \ldots, p$,

$$y_i = z_i + \epsilon_i,$$

where $z_i \sim \mathcal{N}(\theta, \lambda^{-1})$ and $\epsilon_i \sim \mathcal{N}(0, \tau_i^{-1})$, with all random variables independent and with λ and the τ_i's known. We use a flat prior for θ. Then

$$\begin{aligned}
\log f(\mathbf{y}, \mathbf{z}|\theta) &= \log f(\mathbf{y}|\mathbf{z}) + \log f(\mathbf{z}|\theta) \\
&= -p \log 2\pi + \frac{1}{2}\sum_i \log (\lambda\tau_i) - \frac{1}{2}\sum_i \tau_i(y_i - z_i)^2 \\
&\qquad -\frac{1}{2}\lambda \sum_i (z_i - \theta)^2.
\end{aligned}$$

Marginally, $y_i \sim \mathcal{N}(\theta, \tau_i^{-1} + \lambda^{-1}) \sim \mathcal{N}(\theta, 1/(\lambda\rho_i))$, where $\rho_i = \tau_i/(\lambda + \tau_i)$. Thus

$$\log f(\mathbf{y}|\theta) = -\frac{p}{2}\log 2\pi + \frac{1}{2}\sum_i \log (\lambda\rho_i) - \frac{\lambda}{2}\sum_i \rho_i(y_i - \theta)^2.$$

4.1 Observed DICs

For this example

$$\theta|\mathbf{y} \sim \mathcal{N}\left(\frac{\sum_i \rho_i y_i}{\sum_i \rho_i}, \frac{1}{\lambda \sum_i \rho_i}\right),$$

and therefore the posterior mean and mode of θ, given $\mathbf{y}$, are both equal to $\hat{\theta}(\mathbf{y}) = \sum_i \rho_i y_i / \sum_i \rho_i$. Thus

$$\mathrm{DIC}_1 = \mathrm{DIC}_2 = p \log 2\pi - \sum_i \log(\lambda\rho_i) + \lambda \sum_i \rho_i (y_i - \hat{\theta}(\mathbf{y}))^2 + 2.$$

Furthermore, $p_{D1} = p_{D2} = 1$.

For DIC_3 it turns out that

$$\begin{aligned}
\widehat{f}(\mathbf{y}) &= \mathbb{E}_\theta[f(\mathbf{y}|\theta)|\mathbf{y}] \\
&= 2^{-1/2} f(\mathbf{y}|\hat{\theta}(\mathbf{y})), \qquad (5)
\end{aligned}$$

so that

$$\mathrm{DIC}_3 = \mathrm{DIC}_1 - \log 2$$

and $p_{D3} = 1 - \log 2$.

Surprisingly, the relationship (5) is valid even though both $f(\cdot|\hat{\theta}(\mathbf{y}))$ and $\hat{f}(\cdot)$ are densities. Indeed, this identity only holds for the particular value $\mathbf{y}$ corresponding to the observations. For other values of z, $\hat{f}(z)$ is not equal to $f(z|\hat{\theta}(\mathbf{y}))/\sqrt{2}$. Note also that it makes sense that p_{D3} is smaller than p_{D2} in that the predictive distribution is not necessarily of the same complexity as the sum of the dimensions of its parameters.

4.2 Complete DICs

For the random effect model,

$$\begin{aligned}
\overline{D(\theta)} &= -2\mathbb{E}_{\theta,\mathbf{Z}}\left[\log f(\mathbf{y},\mathbf{Z}|\theta)|\mathbf{y}\right] \\
&= -2\mathbb{E}_{\mathbf{Z}}\left\{\mathbb{E}_\theta\left[\log f(\mathbf{y},\mathbf{Z}|\theta)|\mathbf{y},\mathbf{Z}\right]|\mathbf{y}\right\} \\
&= 2p\log 2\pi - \sum_i \log(\lambda\tau_i) \\
&\qquad +\mathbb{E}_{\mathbf{Z}}\left[\sum_i \tau_i(y_i - z_i)^2 + \lambda\sum_i (z_i - \bar{z})^2 + 1|\mathbf{y}\right] \\
&= -2\mathbb{E}_{\mathbf{Z}}\left[\log f(\mathbf{y},\mathbf{Z}|\mathbb{E}_\theta[\theta|\mathbf{y},\mathbf{Z}])|\mathbf{y}\right] + 1,
\end{aligned}$$

since $\theta|\mathbf{y},\mathbf{z} \sim \mathcal{N}(\bar{z}, \frac{1}{\lambda p})$. As a result, $p_{D4} = 1$.

After further detailed calculations we obtain

$$\begin{aligned}
\mathrm{DIC}_4 &= 2p\log 2\pi - \sum_i \log(\lambda\tau_i) + \sum_i \lambda\rho_i(1-\rho_i)(y_i - \hat{\theta}(\mathbf{y}))^2 \\
&\qquad +\lambda\sum_i \hat{z}_i^2 - \lambda p\,\hat{\theta}(\mathbf{y})^2 + 2 + p \\
&= \mathrm{DIC}_2 + p\log 2\pi + p + \sum_i \log\frac{\rho_i}{\tau_i}.
\end{aligned}$$

We also obtain $p_{D5} = 1 + p$, $p_{D6} = 1$,

$$\mathrm{DIC}_5 = \mathrm{DIC}_4 + p,$$

and

$$\mathrm{DIC}_6 = \mathrm{DIC}_5 - p = \mathrm{DIC}_4.$$

The value of p_{D5} is not surprising since, in DIC_5, $\mathbf{z}$ is regarded as an extra parameter of dimension p. This is not the case in DIC_6 since, in the computation of $D(\tilde{\theta})$, $\mathbf{z}$ is then treated as missing variables.

4.3 Conditional DICs and further remarks

DIC_7 and DIC_8 involve $f(\mathbf{y}|\mathbf{z},\theta)$. In the random effect model, this quantity does not depend on parameter θ so that computing the p_Ds and therefore the DICs does not really make sense. For instance, p_{D8} would be 0 and p_{D7}, although different from 0, because $\mathbf{z}$ is considered as an additional parameter, would not take θ into account either (unless indirectly through $\mathbf{z}$).

Note however that

$$\begin{aligned}\mathrm{DIC}_7 &= p \log 2\pi - \sum_i \log \tau_i + \lambda \sum_i \rho_i(1-\rho_i)(y_i - \hat{\theta}(\mathbf{y}))^2 \\ &\quad +2[\sum_r \rho_r + \{\sum_r \rho_r(1-\rho_r)\}/(\sum_r \rho_r)],\end{aligned}$$

which appears at the end of Section 2.5 of Spiegelhalter *et al.* (2002), corresponding to a 'change of focus'.

The DICs ($\mathrm{DIC}_{1,2,4,6}$) leading to the same measure of complexity through $p_D = 1$ but to different posterior deviances show how the latter can incorporate an additional penalty by measuring the amount of missing information, corresponding to $\mathbf{z}$, in DIC_4 and DIC_6. DIC_5 incorporates the missing information in the posterior deviance while p_{D5} regards $\mathbf{z}$ as an extra p-dimensional parameter ($p_{D5} = 1 + p$). This illustrates the unsatisfactory inconsistency in the way DIC_5 takes $\mathbf{z}$ into account, as pointed out in Section 3.2.

5 Mixtures of distributions

As suggested in Spiegelhalter et al. (2002) and detailed in the ensuing discussion, an archetypical example of a missing data model is the K-component normal mixture in which

$$f(y|\theta) = \sum_{j=1}^{K} p_j \phi(y|\mu_j, \sigma_j^2), \qquad \sum_{j=1}^{K} p_j = 1\,,$$

with notation as defined in Section 3.1. Note at this point that, while mixtures can be easily handled by winBUGS 1.4, the use of DIC for these models is not possible in the current version (see Spiegelhalter et al., 2004).

The observed likelihood of a mixture is

$$f(\mathbf{y}|\theta) = \prod_{i=1}^{n} \sum_{j=1}^{K} p_j \phi(y_i|\mu_j, \sigma_j^2)\,.$$

This model can be interpreted as a missing data model problem if we introduce the membership variables $\mathbf{z} = (z_1, \ldots, z_n)$, a set of K-dimensional indicator vectors, denoted by $z_i = \{z_{i1}, \ldots, z_{iK}\} \in \{0,1\}^K$, so that $z_{ij} = 1$ if and only if y_i is generated from the normal distribution $\phi(.|\mu_j, \sigma_j^2)$, conditional on z_i, and $P(Z_{ij} = 1) = p_j$. The corresponding complete likelihood is then

$$f(\mathbf{y}, \mathbf{z}|\theta) \;=\; \prod_{i=1}^{n} \prod_{j=1}^{K} \left\{p_j \phi(y_i|\mu_j, \sigma_j^2)\right\}^{z_{ij}}\,. \tag{6}$$

5.1 Observed DICs

Since $f(\mathbf{y}|\theta)$ is available in closed form, the missing data $\mathbf{z}$ can be ignored and the expressions (2) and (3) for the deviance and DIC can be computed using m simulated values $\theta^{(1)}, \ldots, \theta^{(m)}$ from an MCMC run. (We refer the reader to Celeux et al. (2000) and Marin et al. (2005) for details of the now-standard implementation of an MCMC algorithm in mixture settings.) The first term of DIC_1, DIC_2 and DIC_3 is therefore approximated by an MCMC algorithm as

$$\begin{aligned}
\overline{D(\theta)} &\approx -\frac{2}{m}\sum_{l=1}^{m} \log f(y|\theta^{(l)}) \\
&= -\frac{2}{m}\sum_{l=1}^{m}\sum_{i=1}^{n} \log\left\{\sum_{j=1}^{K} p_j^{(l)} \phi(y_i|\mu_j^{(l)}, \sigma_j^{2(l)})\right\}\,,
\end{aligned}$$

where m is the number of iterations and $(p_j^{(l)}, \mu_j^{(l)}, \sigma_j^{2(l)})_{1\le j\le k}$ are the simulated values of the parameters.

For DIC_1, the posterior means of the parameters are computed as the MCMC sample means of the simulated values of θ, but, as mentioned in

Section 3.1 and further discussed in Celeux et al. (2000), these estimators are not meaningful if no identifiability constraint is imposed on the model (see also Stephens, 2000), and even then they often lead to negative p_D's. For instance, for the Galaxy dataset (Roeder, 1990) discussed below in Section 5.5, we get p_{D1} equal to $-1.8, 5.0, -77.7, -69.1, -82.9, -65.5$, for $K = 2, \ldots, 7$, respectively, using 5000 iterations of the Gibbs sampler. In view of this considerable drawback, the DIC_1 criterion is not to be considered further for this model.

5.2 Complete DICs

The first terms of DIC_4, DIC_5 and DIC_6 are all identical. In view of this, we can use the same MCMC algorithm as before to come up with an approximation to $\mathbb{E}_{\theta,\mathbf{Z}}\left[\log f(\mathbf{y}, \mathbf{Z}|\theta)|\mathbf{y}\right]$, except that we also need to simulate the $\mathbf{z}$'s.

Recall that $\mathbb{E}_{\theta,\mathbf{Z}}\left[\log f(\mathbf{y}, \mathbf{Z}|\theta)|\mathbf{y}\right] = \mathbb{E}_\theta\left\{\mathbb{E}_Z\left[\log f(\mathbf{y}, \mathbf{Z}|\theta)|\mathbf{y}, \theta\right]|\mathbf{y}\right\}$ (Section 3.2). Given that, for mixture models, $\mathbb{E}_\mathbf{Z}[\log f(\mathbf{y}, \mathbf{Z}|\theta)|\mathbf{y}, \theta]$ is available in closed form as

$$\mathbb{E}_\mathbf{Z}[\log f(\mathbf{y}, \mathbf{Z}|\theta)|\mathbf{y}, \theta] = \sum_{i=1}^{n}\sum_{j=1}^{K} P(Z_{ij} = 1|\theta, \mathbf{y}) \log(p_j \phi(y_i|\mu_j, \sigma_j^2)),$$

with

$$P(Z_{ij} = 1|\theta, \mathbf{y}) = \frac{p_j \phi(y_i|\mu_j, \sigma_j^2)}{\sum_{k=1}^{K} p_k \phi(y_i|\mu_k, \sigma_k^2)} \stackrel{\text{def}}{=} t_{ij}(\mathbf{y}, \theta),$$

this approximation is obtained from the MCMC output $\theta^{(1)}, \ldots, \theta^{(m)}$ as

$$\frac{1}{m}\sum_{l=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{K} t_{ij}(\mathbf{y}, \theta^{(l)}) \log\{p_j^{(l)} \phi(y_i|\mu_j^{(l)}, \sigma_j^{2(l)})\}. \tag{7}$$

Then the second term in DIC_4,

$$2\mathbb{E}_\mathbf{Z}[\log f(\mathbf{y}, \mathbf{Z}|\mathbb{E}_\theta[\theta|\mathbf{y}, \mathbf{z}]|\mathbf{y}],$$

can be approximated by

$$\frac{2}{m}\sum_{l=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{K} z_{ij}^{(l)} \log\{\overline{p}_j^{(l)} \phi(y_i|\overline{\mu}_j^{(l)}, \overline{\sigma}_j^{2(l)})\} \tag{8}$$

where $\overline{\theta}^{(l)} = \overline{\theta}(\mathbf{y}, z^{(l)}) = \mathbb{E}[\theta|\mathbf{y}, z^{(l)}]$, which can be computed exactly, as shown below, using standard results in Bayesian analysis (Robert, 2001). The prior on θ is assumed to be a product of conjugate densities,

$$f(\theta) = f(p) \prod_{j=1}^{K} f(\mu_j, \sigma_j)\,,$$

where $f(p)$ is a Dirichlet density $\mathcal{D}(.\,|\alpha_1, \ldots, \alpha_K)$, $f(\mu_j|\sigma_j)$ is a normal density $\phi(.\,|\xi_j, \sigma_j^2/n_j)$ and $f(\sigma_j^2)$ is an inverse gamma density $\mathcal{IG}(.\,|\nu_j/2, s_j^2/2)$. The quantities $\alpha_1, \ldots, \alpha_K$, ξ_j, n_j, ν_j and s_j^2 are fixed hyperparameters. It follows that

$$\begin{aligned}
\bar{p}_j^{(l)} &= \mathbb{E}_\theta[p_j|\mathbf{y}, \mathbf{z}^{(l)}] = \left\{\alpha_j + m_j^{(l)}\right\} \Big/ \sum_{k=1}^{K} \alpha_k + n \\
\bar{\mu}_j^{(l)} &= \mathbb{E}_\theta[\mu_j|\mathbf{y}, \mathbf{z}^{(l)}] = \left\{n_j\xi_j + m_j^{(l)}\widehat{\mu}_j^{(l)}\right\} \Big/ (n_j + m_j^{(l)}) \\
\bar{\sigma}_j^{2(l)} &= \mathbb{E}_\theta[\sigma_j^2|\mathbf{y}, \mathbf{z}^{(l)}] = \left\{s_j^2 + \widehat{s}_j^{2(l)} + \frac{n_j m_j^{(l)}}{n_j + m_j^{(l)}}(\widehat{\mu}_j^{(l)} - \xi_j)^2\right\} \Big/ (\nu_j + m_j^{(l)} - 2),
\end{aligned}$$

with

$$m_j^{(l)} = \sum_{i=1}^{n} z_{ij}^{(l)}, \quad \widehat{\mu}_j^{(l)} = \frac{1}{m_j^{(l)}} \sum_{i=1}^{n} z_{ij}^{(l)} y_i\,, \quad \widehat{s}_j^{2(l)} = \sum_{i=1}^{n} z_{ij}^{(l)} (y_i - \widehat{\mu}_j^{(l)})^2\,,$$

and with the $\mathbf{z}^{(l)} = (z_1^{(l)}, \ldots, z_n^{(l)})$ simulated at the ℓth iteration of the MCMC algorithm.

If we use approximation (7), the DIC criterion is then

$$\begin{aligned}
\mathrm{DIC}_4 \;\approx\; & -\frac{4}{m} \sum_{l=1}^{m} \sum_{i=1}^{n} \sum_{j=1}^{K} t_{ij}(y, \theta^{(l)}) \log\{p_j^{(l)} \phi(y_i|\mu_j^{(l)}, \sigma_j^{2(l)})\} \\
& + \frac{2}{m} \sum_{l=1}^{m} \sum_{i=1}^{n} \sum_{j=1}^{K} z_{ij}^{(l)} \log\{\overline{p}_j^{(l)} \phi(y_i|\overline{\mu}_j^{(l)}, \overline{\sigma}_j^{2(l)})\}\,.
\end{aligned}$$

Similar formulae apply for DIC_5, with the central deviance $D(\overline{\theta})$ being based instead on the (joint) MAP estimator.

In the case of DIC_6, the central deviance also requires less computation, since it is based on an estimate of θ that does not depend on $\mathbf{z}$. We then obtain

$$-\frac{4}{m}\sum_{l=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{K} t_{ij}(\mathbf{y},\theta^{(l)})\log(p_j^{(l)}\phi(y_i \mid \mu_j^{(l)},\sigma_j^{2(l)}))$$
$$+\,2\sum_{i=1}^{n}\sum_{j=1}^{K} t_{ij}(\mathbf{y},\bar\theta)\log(\bar p_j\phi(y_i \mid \bar\mu_j,\bar\sigma_j^2)),$$

as an approximation to DIC_6.

5.3 Conditional DICs

Since the conditional likelihood $f(y|z,\theta)$ is available, we can also use criteria DIC_7 and DIC_8. The first term can be approximated in a similar fashion to the previous section, namely as

$$\overline{D(Z,\theta)} \approx -\frac{2}{m}\sum_{l=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{K} t_{ij}(\mathbf{y},\theta^{(l)})\log\phi(y_i|\mu_j^{(l)},\sigma_j^{2(l)})\,.$$

The second term of DIC_7, $D(\overline{\mathbf{z}},\overline{\theta})$, is readily obtained, while the computations for DIC_8 of

$$\mathbb{E}_{\mathbf{Z}}[\log f(\mathbf{y}|\mathbf{Z},\widehat{\theta}(\mathbf{y},\mathbf{Z}))|\mathbf{y}]$$

are very similar to those proposed above for the approximation of DIC_6.

Note however that the weights p_j are no longer part of the DIC factor, except through the posterior weights t_{ij}, since $f(y|z,\theta)$ does not depend on p.

5.4 A relationship between DIC_2 and DIC_4

We have

$$\mathrm{DIC}_2 = -4\mathbb{E}_\theta\left[\log f(\mathbf{y}|\theta)|\mathbf{y}\right] + 2\log f(\mathbf{y}|\widehat{\theta}(\mathbf{y}))\,,$$

where $\widehat{\theta}(\mathbf{y})$ denotes a posterior mode of θ, and

$$\mathrm{DIC}_4 = -4\mathbb{E}_{\theta,\mathbf{Z}}\left[\log f(\mathbf{y},\mathbf{Z}|\theta)|\mathbf{y}\right] + 2\mathbb{E}_{\mathbf{Z}}\left[\log f(\mathbf{y},\mathbf{Z}|\mathbb{E}_\theta[\theta|\mathbf{y},\mathbf{Z}])|\mathbf{y}\right].$$

We can write

$$\begin{aligned}\mathbb{E}_{\theta,\mathbf{Z}}\left[\log f(\mathbf{y},\mathbf{Z}|\theta)|\mathbf{y}\right] &= \mathbb{E}_{\theta}\left[\mathbb{E}_{\mathbf{Z}}\left\{\log f(\mathbf{y},\mathbf{Z}|\theta)|\mathbf{y},\theta\right\}|\mathbf{y}\right]\\ &= \mathbb{E}_{\theta}\left[\mathbb{E}_{\mathbf{Z}}\left\{\log f(\mathbf{y}|\theta)|\mathbf{y},\theta\right\}|\mathbf{y}\right]\\ &\quad +\mathbb{E}_{\theta}\left[\mathbb{E}_{\mathbf{Z}}\left\{\log f(\mathbf{Z}|\mathbf{y},\theta)|\mathbf{y},\theta\right\}|\mathbf{y}\right].\end{aligned}$$

Then

$$\mathbb{E}_{\theta,\mathbf{Z}}\left[\log f(\mathbf{y},\mathbf{Z}|\theta)|\mathbf{y}\right] = \mathbb{E}_{\theta}\left[\log f(\mathbf{y}|\theta)|\mathbf{y})\right] - \mathbb{E}_{\theta}\left[\mathrm{Ent}\left\{f(\mathbf{Z}|\mathbf{y},\theta)\right\}|\mathbf{y}\right], \tag{9}$$

where the entropy,

$$\mathrm{Ent}\left\{f(\mathbf{Z}|\mathbf{y},\theta)\right\} = -\mathbb{E}_{\mathbf{Z}}\left\{\log f(\mathbf{Z}|\mathbf{y},\theta)|\mathbf{y},\theta\right\} > 0$$

is a measure of the mixture overlap. When the mixture components are well separated this entropy term is near 0 and it is far from 0 when the mixture components are poorly separated.

It follows that

$$\begin{aligned}\mathrm{DIC}_4 &= \mathrm{DIC}_2 + 4\mathbb{E}_{\theta}\left[\mathrm{Ent}\left\{f(\mathbf{Z}|\mathbf{y},\theta)\right\}|\mathbf{y}\right]\\ &\quad +2\mathbb{E}_{\mathbf{Z}}\left[\log f(\mathbf{y},\mathbf{Z}|\mathbb{E}_{\theta}[\theta|\mathbf{y},\mathbf{Z}])|\mathbf{y}\right] - 2\log f(\mathbf{y}|\widehat{\theta}(\mathbf{y}),\end{aligned}$$

Now, we assume that

$$\mathbb{E}_{\mathbf{Z}}\left[\log f(\mathbf{y},\mathbf{Z}|\mathbb{E}[\theta|\mathbf{y},\mathbf{Z}])|\mathbf{y}\right] \approx \mathbb{E}_{\mathbf{Z}}\left[\log f(\mathbf{y},\mathbf{Z}|\widehat{\theta}(\mathbf{y}))|\mathbf{y}\right].$$

This approximation should be valid provided that

$$\mathbb{E}[\theta|\mathbf{y},\widehat{\mathbf{z}}(\mathbf{y})] = \widehat{\theta}(\mathbf{y})$$

where

$$\widehat{\mathbf{z}}(\mathbf{y}) = \arg\max_{z} f(\mathbf{z}|\mathbf{y}).$$

The last two terms can be further written as

$$\begin{aligned}2\mathbb{E}_{\mathbf{Z}}\left[\log f(\mathbf{y},\mathbf{Z}|\mathbb{E}_{\theta}[\theta|\mathbf{y},\mathbf{Z}])|\mathbf{y}\right] - 2\log f(\mathbf{y}|\widehat{\theta}(\mathbf{y}) \approx\\ 2\mathbb{E}_{\mathbf{Z}}\left[\log f(\mathbf{Z}|\mathbf{y},\widehat{\theta}(\mathbf{y}))|\mathbf{y}\right].\end{aligned}$$

Then

$$\begin{aligned}
\mathbb{E}_\theta\left[\mathrm{Ent}\left\{f(\mathbf{Z}|\mathbf{y},\theta)\right\}|\mathbf{y}\right] &= -\mathbb{E}_\theta\left[\mathbb{E}_{\mathbf{Z}}\left\{\log f(\mathbf{Z}|\mathbf{y},\theta)|\mathbf{y},\theta\right\}|\mathbf{y}\right] \\
&= -\mathbb{E}_{\theta,\mathbf{Z}}\left[\log f(\mathbf{Z}|\mathbf{y},\theta)|\mathbf{y}\right] \\
&\approx -\mathbb{E}_{\mathbf{Z}}\left[\log f(\mathbf{Z}|\mathbf{y},\widehat{\theta}(\mathbf{y}))|\mathbf{y}\right].
\end{aligned}$$

The last approximation should be reasonable when the posterior for θ given $\mathbf{y}$ is sufficiently concentrated around its mode.

We therefore have

$$\mathrm{DIC}_4 \approx \mathrm{DIC}_2 + 2\mathbb{E}_\theta\left[\mathrm{Ent}\left\{f(\mathbf{Z}|\mathbf{y},\theta)\right\}|\mathbf{y}\right], \tag{10}$$

from which it follows that $\mathrm{DIC}_4 > \mathrm{DIC}_2$ and that the difference between the two criteria is twice the posterior mean of the mixture entropy. This inequality can be verified in the experiments that follow. The important point to note about this inequality is that DIC_4 and DIC_2 are of different natures, with DIC_4 penalizing poorly separated mixtures. Note also that van der Linde (2004) provides an alternative entry to DIC via entropy and information representations, although her paper is focussed on variable selection. (Using the Fisher information in the context of mixtures is also quite challenging (McLachlan and Krishnan, 1997).)

5.5 A numerical comparison

When calculating the various DICs for the Galaxy dataset (Roeder, 1990), now used in most papers on mixture estimation, we obtain the results presented in Table 1. As one can see, DIC_5 and DIC_6 do not behave satisfactorily: the former leads to excessively large and non-increasing p_D's, presumably because of its inconsistency in dealing with $\mathbf{Z}$ and to poor MCMC approximations to the MAP estimates. The results from DIC_6 are not reliable because of computational problems in the computation of the marginal MAP estimates. DIC_7 leads to larger p_D's too, presumably as a side effect of incorporating $\mathbf{Z}$ as a parameter, whereas DIC_8 behaves satisfactorily with respect to p_D, considering that for a K-component mixture the number of parameters is $3K-1$. Finally, note that all DICs indicate $K = 3$ as the appropriate number of components. In addition, the effective dimension for DIC_3 stabilises after $K = 3$, indicating that the extra components do not greatly contribute to the deviance of the model, which may not be so appropriate. The same is observed for DIC_4 to a lesser extent. DIC_2 gives reasonable results until $K = 4$. The subsequent degradation for $K = 5$ and $K = 6$ may come from the instability in the plug-in estimate $f(\mathbf{y}|\hat{\theta}(\mathbf{y}))$. The adequacy of the plug-in estimates is shown in Figures 1 and 2.

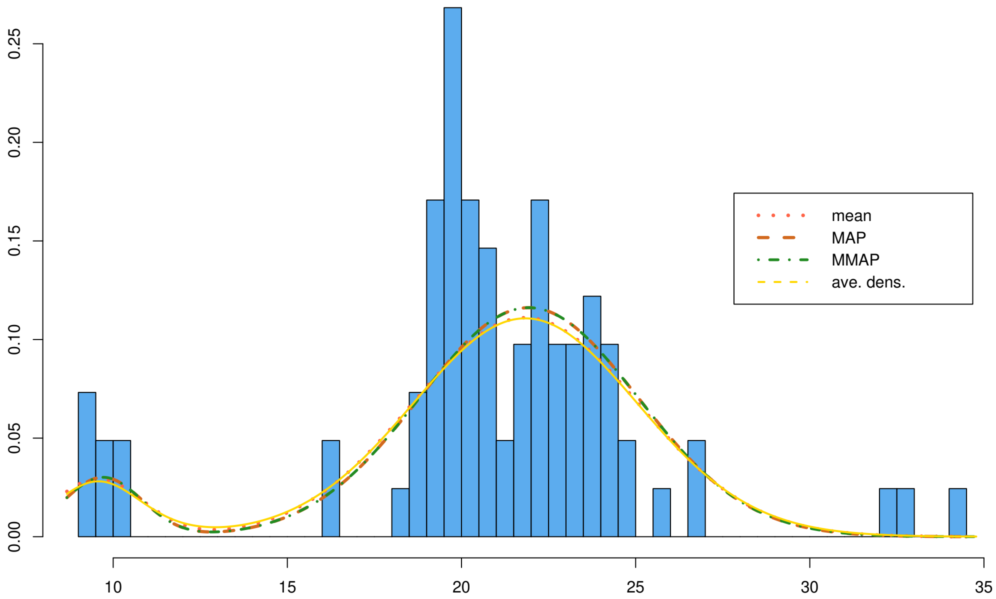

Figure 1: Galaxy dataset of 82 observations with $K = 2$ components fitted: average density, plug-in density with mean parameters, plug-in density with marginal MAP parameters, and plug-in density with joint MAP parameters, The number of iterations is 10,000 (burn-in) plus 10,000 (main).

We also analysed a dataset of 146 observations simulated from the normal mixture

$$0.288\mathcal{N}(0, .2^2) + 0.260\mathcal{N}(-1.5, .5^2) + 0.171\mathcal{N}(2.2, 3.4^2) + 0.281\mathcal{N}(3.3, .5^2)\,.$$

The simulation results are available in Table 3. Figure 3 represents the corresponding estimates after $20,000$ MCMC iterations for $K = 2$. For this number of components, the differences between the estimates are negligible. The same applies to Figure 3, for $K = 3$. The differences start to appear

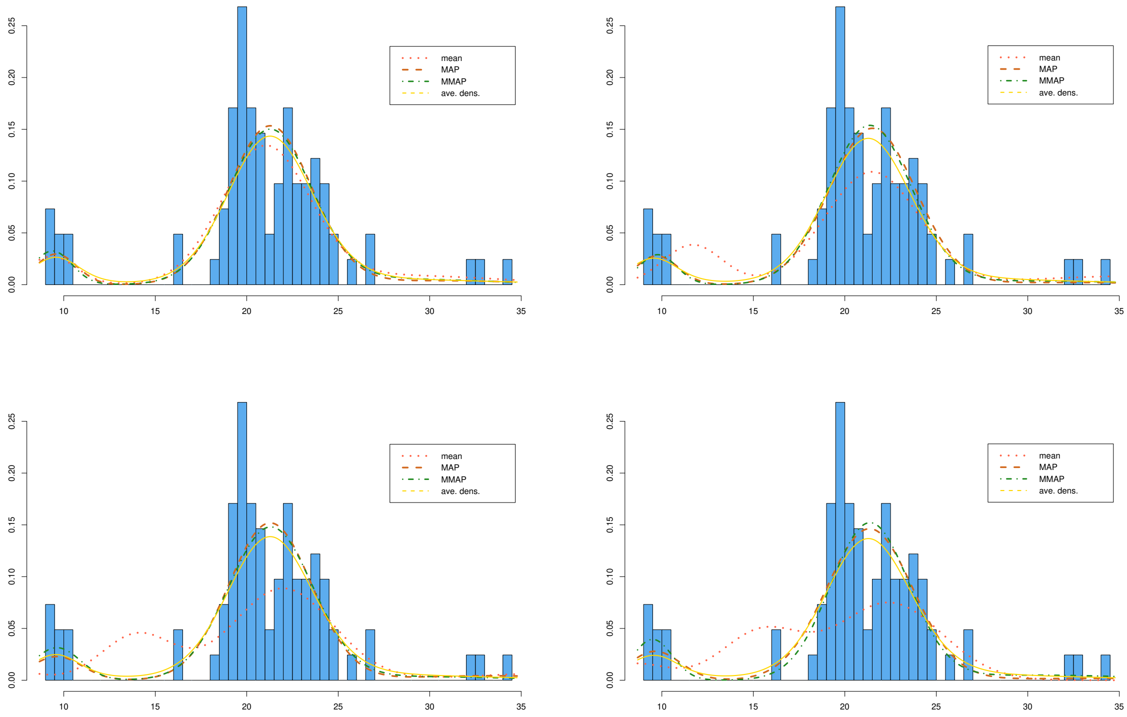

Figure 2: Galaxy dataset of 82 observations with $K = 3, 4, 5, 6$ components fitted: average density, plug-in density with average parameters, plug-in density with marginal MAP parameters, and plug-in density with joint MAP parameters. The number of iterations is 10,000 (burn-in) plus 10,000 (main).

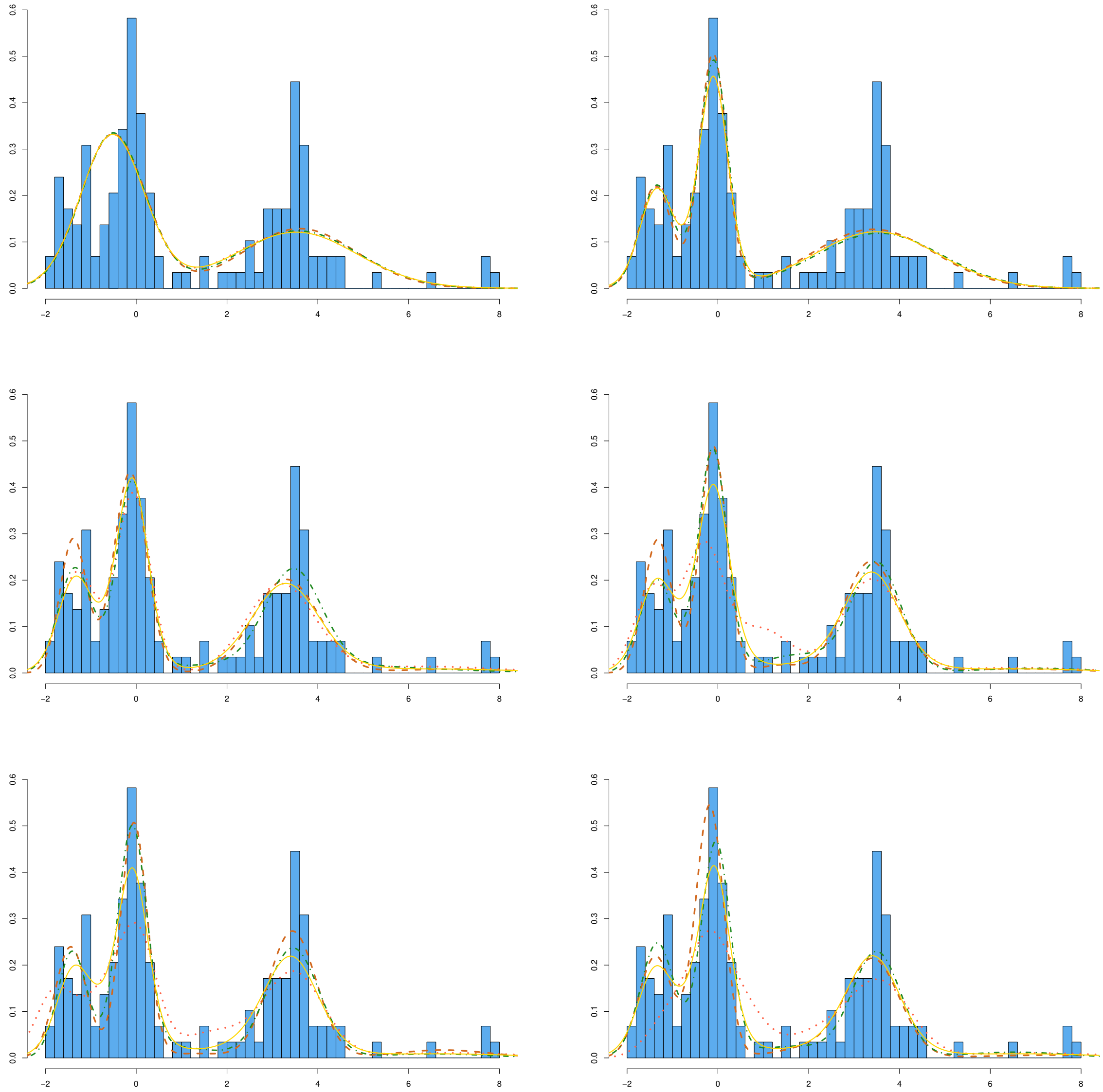

Figure 3: Simulated dataset of 146 observations with $K = 2, 3, 4, 5, 6, 7$ components fitted: average density (gold and full), plug-in density with average parameters (tomato and dots), plug-in density with marginal MAP parameters (forest green and dots-and-dashes), and plug-in density with joint MAP parameters (chocolate and dashes). The number of iterations is 10,000 (burn-in) plus 10,000 (main).

K	DIC$_2$ (p_{D2})	DIC$_3$ (p_{D3})	DIC$_4$ (p_{D4})	DIC$_5$ (p_{D5})	DIC$_6$ (p_{D6})	DIC$_7$ (p_{D7})	DIC$_8$ (p_{D8})
2	453	451	502	705	501	417	410
	(5.56)	(3.66)	(5.50)	(207.88)	(4.48)	(11.07)	(4.09)
3	440	436	461	622	471	378	372
	(9.23)	(4.94)	(6.40)	(167.28)	(15.80)	(13.59)	(7.43)
4	446	439	473	649	482	388	382
	(11.58)	(5.41)	(7.52)	(183.48)	(16.51)	(17.47)	(11.37)
5	447	442	485	658	511	395	390
	(10.80)	(5.48)	(7.58)	(180.73)	(33.29)	(20.00)	(15.15)
6	449	444	494	676	532	407	398
	(11.26)	(5.49)	(8.49)	(191.10)	(46.83)	(28.23)	(19.34)
7	460	446	508	700	571	425	409
	(19.26)	(5.83)	(8.93)	(200.35)	(71.26)	(40.51)	(24.57)

Table 1: Results for the Galaxy dataset and 20,000 MCMC simulations: observed, complete and conditional DICs and corresponding effective dimensions p_D.

for $K = 4$ in Figure 3. Since the correct number of components is indeed 4, we compare the various estimates with the true values in Table 2. Figure 3 shows larger differences for $K = 5$, $K = 6$ and $K = 7$. Note that, after $K = 4$, the predictive density hardly changes. The same phenomenon occurs in Figure 2 for the Galaxy dataset.

Turning to Table 3, we see that DIC$_2$ and DIC$_3$ behave similarly as for the galaxy dataset, except that p_{D3} is decreasing from $K = 3$ and p_{D2} from $K = 5$. DIC$_5$ and DIC$_6$ are not satisfactory, since they are producing negative p_D's. (For DIC$_5$, this is not inconsistent with the remark on its positivity in Section 3.2 since for the mixture example the prior is not flat.) DIC$_7$ produces non-increasing and highly fluctuating p_D's. Only DIC$_4$ and DIC$_8$ give reasonable p_D's close to $3K - 1$ with DIC$_4$ selecting $K = 3$ (with $K = 4$ being a close second-best choice) while DIC$_8$ is selecting $K = 7$. Note that DIC$_5$ and DIC$_6$ select the right number of components! After $10,000$ more MCMC iterations, we observed that DIC$_5$ and DIC$_6$ were still selecting $K =$ 4 with negative p_D's, DIC$_4$ was selecting $K = 4$ too and the others were selecting $K = 7$.

	p_1	p_2	p_3	p_4	μ_1	μ_2	μ_3	μ_4	σ_1^2	σ_2^2	σ_3^2	σ_4^2
True	.26	.29	.17	.28	-1.5	0	2.2	3.3	.25	.04	11.6	.25
Map	.22	.39	.37	.027	-1.38	-.13	3.30	7.02	.09	.13	.53	1.64
Mean	.23	.35	.35	.061	-1.27	-.02	3.19	6.33	.17	.14	.56	2.99
Mmap	.21	.34	.14	.31	-1.35	-.08	3.12	3.46	.13	.11	7.04	.38

Table 2: Estimation results for the simulated dataset with 146 observations, $K = 4$ and $10,000$ MCMC simulations.

6 Conclusion

This paper has shown that the deviance information criterion of Spiegelhalter et al. (2002) and the corresponding effective dimension allow for a wide range of interpretations and extensions outside exponential families, as was already apparent from the published discussion of the paper. What we have found in addition through theoretical and experimental studies is that some of these extensions, while as "natural" as the others, are simply not adequate for evaluating the complexity and fit of a model, either because they give *negative* effective dimensions or because they exhibit too much variability from one model to the next. While Spiegelhalter et al. (2002) argue that negative p_D's are indicative of a possibly poor fit between the model and the data, there is no explanation of that kind in our cases: for the same data and the same model, some DICs are associated with positive p_Ds and others are not.

Among the various criteria, DIC_3 and DIC_4 stand out as being the most reliable of the DICs we studied: they are more resistant to poor estimates in that DIC_3 does not depend on estimates (in the classical sense) and DIC_4 relies on a conditional estimate that gets averaged over iterations. However, the behaviour of DIC_3 in terms of the corresponding p_D is questionable. If one of these two DICs needs to be picked out as *the* DIC for missing data models, it is undoubtedly DIC_4, as it builds on the missing data structure rather naturally, starting from the complete DIC and integrating over the missing variables. However, DIC_4 is not invariant to the choice of $\mathbf{Z}$, whereas DIC_3 is. This DIC takes into account the missing data structure but it favors models minimizing the missing information (as shown in Section 5.4). While a sensible choice and focussing on the missing data structure, DIC_4 does not necessarily lead to the most suitable model. For instance, in the mixture case,

K	DIC$_2$ (p_{D2})	DIC$_3$ (p_{D3})	DIC$_4$ (p_{D4})	DIC$_5$ (p_{D5})	DIC$_6$ (p_{D6})	DIC$_7$ (p_{D7})	DIC$_8$ (p_{D8})
2	581	582	598	579	602	409	398
	(5.10)	(6.25)	(5.12)	(-13.48)	(9.20)	(15.73)	(4.13)
3	554	557	569	481	584	317	319
	(11.44)	(15.08)	(6.76)	(-81.67)	(21.56)	(7.23)	(8.42)
4	539	534	572	393	541	260	228
	(17.0)	(11.4)	(9.1)	(-170.2)	(-21.8)	(42.6)	(10.0)
5	540	529	610	432	657	280	219
	(21.6)	(11.1)	(12.0)	(-165.7)	(59.3)	(74.7)	(13.4)
6	537	527	653	486	730	251	215
	(19.6)	(10.3)	(16.4)	(-150.9)	(93.0)	(52.8)	(16.7)
7	534	526	687	550	739	248	210
	(17.86)	(9.84)	(20.73)	(-116.62)	(72.32)	(58.54)	(20.12)

Table 3: Results for the simulated dataset with 146 observations and $20,000$ MCMC simulations: observed, complete and conditional DICs (first line) and corresponding effective dimensions p_D (second line).

it chooses the mixture model with the cluster structure for which there is the greatest evidence, and this model can be different from the most relevant model. Nonetheless, DICs can be seen as a Bayesian version of AIC and, as pointed out by several discussants in Spiegelhalter et al. (2002), they may underpenalize model complexity: DIC$_4$ can therefore be expected to reduce this tendency in a sensible way.

The fact that DIC$_7$ may produce increasing p_Ds for increasing complexity is not very surprising, but it points out a drawback with this kind of criterion, because considering $\mathbf{Z}$ as an additional parameter makes the (conditional) model too adaptive to be well-discriminating. Similarly, DIC$_8$ is not very discriminating but it may warrant further investigation: it is rather stable for varying Ks and it leads to p_D values close to the number of parameters in the model, at least in the case of the mixture model.

Acknowledgements

We are grateful to the editorial team of *Bayesian Analysis* for their suggestions that helped us to prepare this version of the paper.

References

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

P. Bryant and J. A. Williamson. Asymptotic behaviour of lassification maximum likelihood estimates. *Biometrika*, 65:273–281, 1978.

O. Cappé, E. Moulines, and T. Rydén. *Hidden Markov Models*. Springer-Verlag, New York, 2005.

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

M. DeIorio and C. P. Robert. Discussion of Spiegelhalter et al. *Journal of the Royal Statistical Society, Series B*, 64:629–630, 2002.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

J.S. Hodges and D.J. Sargent. Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika*, 88:367–379, 2001.

R. J .A. Little and D. B. Rubin. On jointly estimating parameters and missing data by maximizing the complete-data likelihood. *American Statistician*, 37:218–220, 1983.

R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data*. J. Wiley, New York, 1987.

J.M. Marin, K.L. Mengersen, and C.P. Robert. Bayesian modelling and inference on mixtures of distributions. In C.R. Rao and D. Dey, editors, *Handbook of Statistics*, volume 25 (to appear). Springer-Verlag, New York, 2005.

F. H. C. Marriott. Separating mixtures of mixture distributions. *Biometrics*, 31:767–769, 1975.

G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. J. Wiley, New York, 1997.

S. Richardson. Discussion of Spiegelhalter et al. *Journal of the Royal Statistical Society, Series B*, 64:631, 2002.

C. P. Robert. *The Bayesian Choice (second edition)*. Springer Verlag, 2001.

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer Verlag, New York, second edition, 2001.

K. Roeder. Density estimation with confidence sets exemplified by superclusters and voids in galaxies. *Journal of the American Statistical Association*, 85:617–624, 1990.

D. J. Spiegelhalter, N.G. Best, Carlin B.P., and A. van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, 64:583–640, 2002.

D.J. Spiegelhalter, A. Thomas, N.G. Best, and D. Lunn. WinBUGS 1.4 user manual. Technical report, MRC Biostatistics Unit, Cambridge, 2004. URL `http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/manual14.pdf`.

M. Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B*, 62:795–809, 2000.

A. van der Linde. DIC in variable selection. Technical report, Institute of Statistics, University of Bremen, 2004. URL `http://www.math.uni-bremen.de/~avdl/download/papers/varsel2.pdf`.