INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES
Série des Documents de Travail du CREST
(Centre de Recherche en Economie et Statistique)

n° 2003-30

Deviance Information Criteria for Missing Data Models

G. CELEUX[1]
F. FORBES[2]
C. P. ROBERT[3]
D. M. TITTERINGTON[4]

[1] INRIA Rhône-Alpes.
[2] INRIA Rhône-Alpes.
[3] CREST and CEREMADE, University Paris-Dauphine.
[4] University of Glasgow.

Deviance Information Criteria for Missing Data Models

G. Celeux[♭], F. Forbes[♭], C.P. Robert[‡] , and D.M. Titterington[♯]

[♭]*INRIA Rhône-Alpes,*
[‡]*CREST-INSEE, and CEREMADE-Université Paris Dauphine,*
and [♯]*University of Glasgow*

Abstract

The deviance information criterion (DIC) introduced by Spiegelhalter et al. (2002) is directly inspired by linear and generalised linear models, but it is not so naturally defined for missing data models. In this paper, we reassess the criterion for such models, testing the behaviour of various extensions in the cases of mixture and random effect models.

Keywords: *deviance, completion, MAP, mixture model, random effect model.*

Résumé

Le critère d'information de la déviance (DIC) a é introduit par Spiegelhalter et al. (2002) dans le contexte particulier des modèles linéaires généralisés. Son extension aux modèles à variables latentes s'avère plus délicat. Cet article propose une série d'extensions à ces modèles, qui sont testées sur les modèles particuliers des mélanges et des effets aléatoires.

Mots clés : *déviance, complétion, maximum a posteriori, mélange, modèle à effets aléatoires.*

Deviance Information Criteria for Missing Data Models

G. Celeux[♭], F. Forbes[♭], C.P. Robert[♯*], and D.M. Titterington[♯†]
[♭] *INRIA Rhône-Alpes,* [♯] *CREST and CEREMADE, Uni. Paris Dauphine and* [♯] *University of Glasgow*

Abstract

The deviance information criterion (DIC) introduced by Spiegelhalter et al. (2002) is directly inspired by linear and generalised linear models, but it is not so naturally defined for missing data models. In this paper, we reassess the criterion for such models, testing the behaviour of various extensions in the cases of mixture and random effect models.
Keywords: deviance, completion, MAP, mixture model, random effect model.

1 Introduction

When developing their theory of the *deviance information criterion* (DIC), Spiegelhalter et al. (2002) did not give much consideration to models outside the scope 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 pointed out the possible difficulties of defining DIC precisely in these scenarios. In particular,

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

†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.

DeIorio and Robert (2002) described some possible inconsistencies in the definition of a DIC for mixture models, the most apparent being the occurrence of negative dimension parameters p_D, while Richardson (2002) presented an alternative notion of DIC, again in the context of mixture models. The difficulty with more general models is that basic notions like *deviance* may take several equally acceptable meanings, with direct consequences for the properties of the corresponding DICs. There is thus a need to evaluate and compare these properties for the most natural choices of DICs.

The present paper reassesses this problem in greater depth and generality 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 in 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, $P(y|\theta)$, 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. This quantity is based on a *deviance*, defined by

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

where $f(\mathbf{y})$ 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,

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

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. Note that p_D is completely independent of the choice of the standardizing f. As already pointed out in Spiegelhalter et al. (2002), the fact that p_D may depend on the choice of the estimate $\tilde{\theta}$ is one of the difficulties of this approach.

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

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

For model comparison, we can assume without loss of generality that $f(\mathbf{y}) = 1$, for all models, so we take

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

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 $D(\theta)$. 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, as revealed by DeIorio and Robert (2002), the effective dimension p_D may well be negative in the case of mixtures of distributions. 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 possible extensions of DIC in missing data models, that is, when

$$P(\mathbf{y}|\theta) = \int P(\mathbf{y}, \mathbf{z}|\theta) d\mathbf{z}\,,$$

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 $\mathbf{y}$. (Whether or not $\mathbf{z}$ is physically meaningful for the problem is not relevant here.) 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 $P(\mathbf{y}|\theta)$ is often called the *observed likelihood* while $P(\mathbf{y}, \mathbf{z}|\theta)$ 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, $P(\mathbf{y}|\theta)$, 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). Following from (3), while

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

is clearly and uniquely defined, even though it may require (MCMC) simulation to be computed approximately, choosing $\tilde{\theta}$ is more delicate so that 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 5.5, 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 priors and likelihood are invariant with respect to the labels of the components, all posterior means are identical and the mixture collapses to a single-component mixture (Celeux et al., 2000). As a result,

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

is often not a good choice. For instance, in the mixture case, DIC_1 almost always leads to a negative value for p_D (DeIorio and Robert, 2002).

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

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

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

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

Recall for instance that, for the K-component mixture problem, there exist a multiple of $K!$ marginal posterior modes. Note that, when the prior on

θ is uniform, so that $\widehat{\theta}(\mathbf{y})$ is also the maximum likelihood estimator, the corresponding p_D,

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

is necessarily positive.

In the mixture case, where nonidentifiability is endemic, and for all models where this is the case, a more natural choice for $D(\tilde{\theta})$ is to select an estimator $\widehat{P}(\mathbf{y})$ of the density $P(\mathbf{y}|\theta)$, since this function is invariant under permutation of the component labels. For instance, one can use the posterior expectation $\mathbb{E}_\theta\left[P(\mathbf{y}|\theta)|\mathbf{y}\right]$. (Note that this is also independent of the representation (3).) In terms of functional estimation, this approach provides stable evaluations that are superior to the plug-in estimates $P(\mathbf{y}|\widehat{\theta})$; the density estimator is easily approximated by an MCMC evaluation. For instance, for a Gaussian mixture with density

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

we have

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

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 $p = (p_1,\ldots,p_K)^t$, and m denotes the number of MCMC simulations. This is also the MCMC predictive density, and this leads to another criterion

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

where $\widehat{P}(\mathbf{y}) = \prod_{i=1}^{n}\widehat{P}(y_i)$. 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 also that the relative values of $\widehat{P}(\mathbf{y})$, for different models, constitute the posterior Bayes factors of Aitkin (1991).

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

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

In addition to the difficulty of choosing $\tilde{\theta}$ as in the previous section, we have the problem of defining the fixed point deviance, $D(\tilde{\theta})$. On the basis of an interpretation of the DIC as the expectation of the joint DIC over the missing data, a natural idea is to take

$$\begin{aligned}D(\tilde{\theta}|\mathbf{y}, \mathbf{z}) &= -2\log P(\mathbf{y}, \mathbf{z}|\mathbb{E}_\theta[\theta|\mathbf{y}, \mathbf{z}]) \\ \text{DIC}(\mathbf{y}, \mathbf{z}) &= -4\mathbb{E}_\theta\left[\log P(\mathbf{y}, \mathbf{z}|\theta)|\mathbf{y}, \mathbf{z}\right] + 2\log P(\mathbf{y}, \mathbf{z}|\mathbb{E}_\theta[\theta|\mathbf{y}, \mathbf{z}]),\end{aligned}$$

and then define

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

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 is to consider $\mathbf{Z}$ as an additional parameter 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,

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

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). The only relevant estimator $(\widehat{\mathbf{Z}}(\mathbf{y}), \widehat{\theta}(\mathbf{y}))$ in this setting thus seems to be the *joint* 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).

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

The DIC corresponding to this analysis is then

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

which, barring a poor MCMC approximation to the MAP estimate, should lead to a positive effective dimension p_{D5}, given that, under a flat prior, the second part of DIC_5 is the maximum (modulo 2) of the function integrated in the first part against θ and $\mathbf{Z}$. Note however, that DIC_5 is somewhat inconsistent in the way it takes $\mathbf{z}$ into account. The posterior deviance incorporates $\mathbf{z}$ as missing variables while $D(\tilde{\theta})$ and therefore p_{D5} regards $\mathbf{z}$ as an additional parameter (see Section 4 for an illustration).

Another interpretation, and thus another DIC, can be derived from the EM analysis of the missing data model. Recall (Dempster et al., 1977; Robert and Casella, 1999, §5.3.3) that the core function of the EM algorithm is

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

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

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

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

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

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

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

rather than of

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

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.

3.3 Conditional DICs

A third approach to DICs in the context of missing variable models is to use the conditional likelihood, $P(\mathbf{y}|\mathbf{z}, \theta)$, where $\mathbf{z}$ is then considered as an additional parameter. This approach has obvious asymptotic and coherency difficulties, as discussed in the past 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 $\mathbf{z}$.)

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

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

where again the pair $(\mathbf{z}, \theta)$ is estimated by the joint MAP estimator, 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.

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

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

where $\widehat{\theta}(\mathbf{y}, \mathbf{z})$ is an estimator of θ based on $P(\mathbf{y}, \mathbf{z}|\theta)$, 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 they 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. Then

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

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

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

In this section, we use a flat prior for θ.

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

$$\begin{aligned} \widehat{P}(\mathbf{y}) &= \mathbb{E}_\theta[P(\mathbf{y}|\theta)|\mathbf{y}] \\ &= 2^{-1/2}\ P(\mathbf{y}|\hat{\theta}(\mathbf{y})), \end{aligned} \tag{4}$$

so that

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

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

Surprisingly, the relationship (4) is valid even though both $P(\cdot|\hat{\theta}(\mathbf{y}))$ and $\hat{P}(\cdot)$ are densities. Indeed, this identity only holds for the particular value $\mathbf{y}$ corresponding to the observations. For other values of z, $\hat{P}(z)$ is not equal to $P(z|\hat{\theta}(\mathbf{y}))/\sqrt{2}$. Note also that it makes sense that p_{D3} is smaller than p_{D2} in that the predictive 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,

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

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

After further detailed calculations we obtain

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

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 treated as missing variables.

4.3 Conditional DICs and further remarks

DIC_7 and DIC_8 involve $P(\mathbf{y}|\mathbf{z}, \theta)$. 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.

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 mentioned in Section 3.2.

5 Mixtures of distributions

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

mixture in which

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

with notation as defined in Section 3.1. The observed likelihood is

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

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

$$P(\mathbf{y}, \mathbf{z}|\theta) \;=\; \prod_{i=1}^{n} \prod_{j=1}^{K} \left\{p_j \phi(y_i|\mu_j, \sigma_j^2)\right\}^{z_{ij}}\,. \tag{5}$$

5.1 Observed DICs

Since $P(\mathbf{y}|\theta)$ 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) 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

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

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 (Stephens, 2000), and even then they often lead to negative p_D's. In view of this considerable drawback, the DIC_1 criterion is not to be considered further for this problem.

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 $\mathbb{E}_{\theta,\mathbf{Z}}\left[\log P(\mathbf{y},\mathbf{Z}|\theta)|\mathbf{y}\right]$, except that we also need to simulate the $\mathbf{z}$'s.

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

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

with

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

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

$$\frac{1}{m}\sum_{l=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{K} t_{ij}(\mathbf{y},\theta^{(l)})\log\{p_j^{(l)}\phi(y_i|\mu_j^{(l)},\sigma_j^{2(l)})\}\,. \tag{6}$$

Then the second term in DIC_4,

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

can be approximated by

$$\frac{2}{m}\sum_{l=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{K} z_{ij}^{(l)}\log\{\overline{p}_j^{(l)}\phi(y_i|\overline{\mu}_j^{(l)},\overline{\sigma}_j^{2(l)})\} \tag{7}$$

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,

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

where $P(p)$ is a Dirichlet density $\mathcal{D}(.\,|\alpha_1, \ldots, \alpha_K)$, $P(\mu_j|\sigma_j)$ is a normal density $\phi(.\,|\xi_j, \sigma_j^2/n_j)$ and $P(\sigma_j^2)$ 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 (6), 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 $P(y|z,\theta)$ 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, $D(\overline{z},\overline{\theta})$, is readily obtained, while the computations for DIC_8 of

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

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}.

5.4 A relationship between DIC_2 and DIC_4

We have

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

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

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

We can write

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

Then

$$\mathbb{E}_{\theta,\mathbf{Z}}\left[\log P(\mathbf{y},\mathbf{Z}|\theta)|\mathbf{y}\right] = \mathbb{E}_{\theta}\left[\log P(\mathbf{y}|\theta)|\mathbf{y}\right] - \mathbb{E}_{\theta}\left[\mathrm{Ent}\left\{P(\mathbf{Z}|\mathbf{y},\theta)\right\}|\mathbf{y}\right], \quad (8)$$

where the entropy,

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

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

$$\begin{aligned} DIC_4 &= DIC_2 + 4\mathbb{E}_{\theta}\left[\mathrm{Ent}\left\{P(\mathbf{Z}|\mathbf{y},\theta)\right\}|\mathbf{y}\right] \\ &\quad +2\mathbb{E}_{\mathbf{Z}}\left[\log P(\mathbf{y},\mathbf{Z}|\mathbb{E}_{\theta}[\theta|\mathbf{y},\mathbf{Z}])|\mathbf{y}\right] - 2\log P(\mathbf{y}|\widehat{\theta}(\mathbf{y}), \end{aligned}$$

Now, assuming that

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

This should be valid provided that

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

where

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

The last two terms can be further written as

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

Then

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

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

We therefore have

$$\mathrm{DIC}_4 \approx \mathrm{DIC}_2 + 2\mathbb{E}_\theta \left[\mathrm{Ent}\left\{P(\mathbf{Z}|\mathbf{y},\theta)\right\}|\mathbf{y}\right], \tag{9}$$

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.

5.5 A numerical comparison

When calculating the various DICs for the Galaxy dataset 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. 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 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 $P(\mathbf{y}|\hat{\theta}(\mathbf{y}))$. The adequacy of the plug-in estimates is shown in Figures 1 and 2.

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 $20,000$ iterations for $K=2$. For this number of components, the differences between the estimates are negligible. The

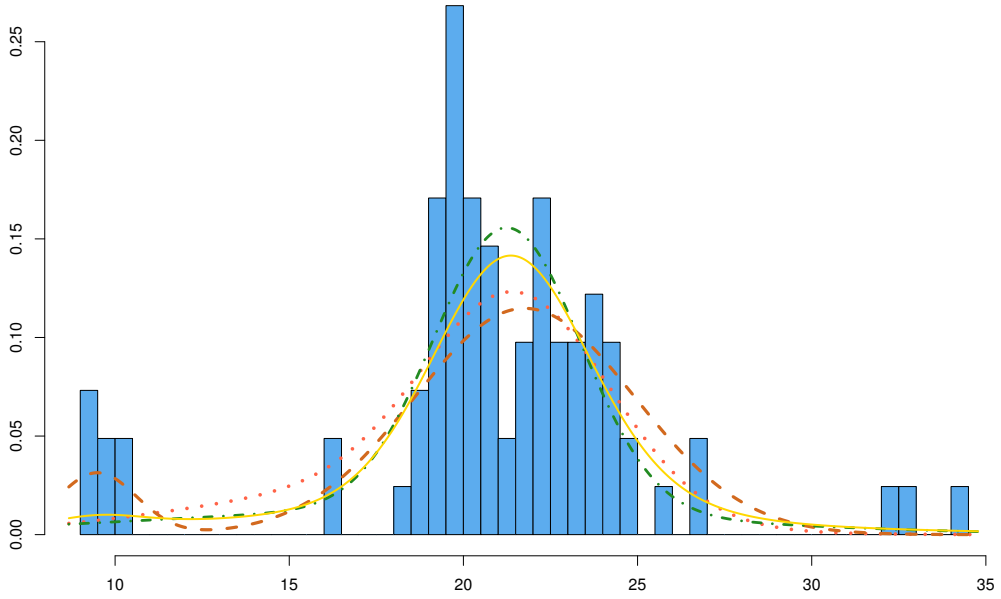

Figure 1: Galaxy dataset of 82 observations with $K = 2$ components fitted: average density (gold and full), plug-in density with average parameters (tomato and dots), plug-in density with marginal MAP parameters (forest green and dots-and-dashes), and plug-in density with joint MAP parameters (chocolate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

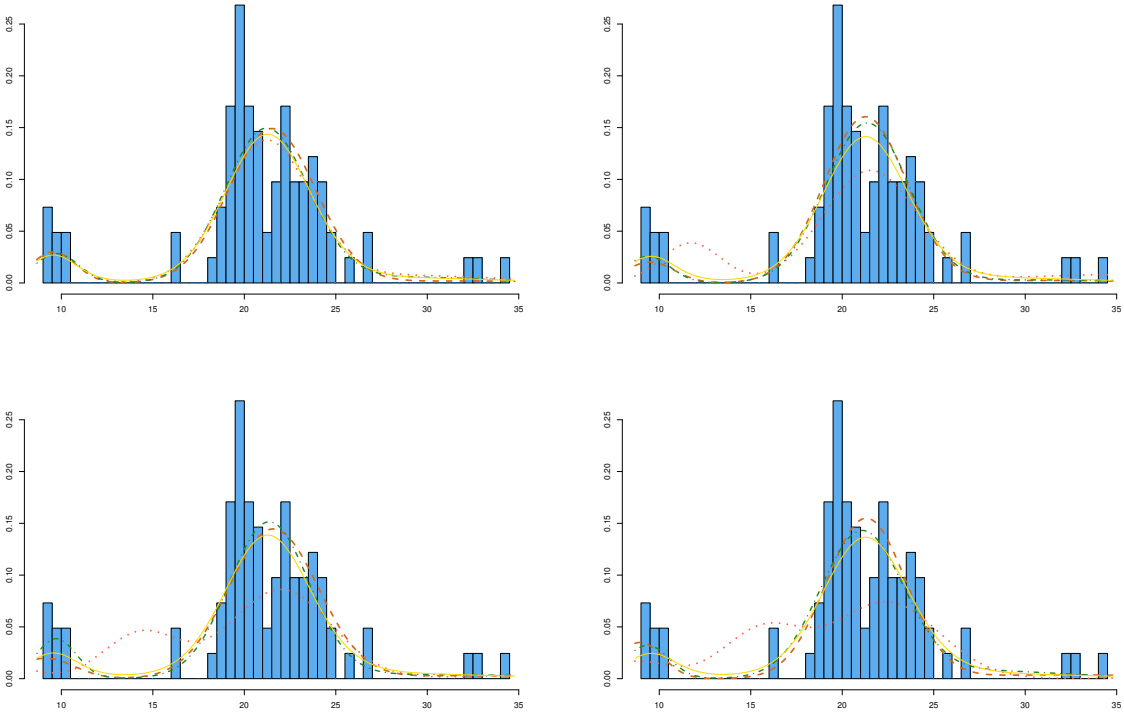

Figure 2: Galaxy dataset of 82 observations with $K = 3, 4, 5, 6$ components fitted: average density (gold and full), plug-in density with average parameters (tomato and dots), plug-in density with marginal MAP parameters (forest green and dots-and-dashes), and plug-in density with joint MAP parameters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

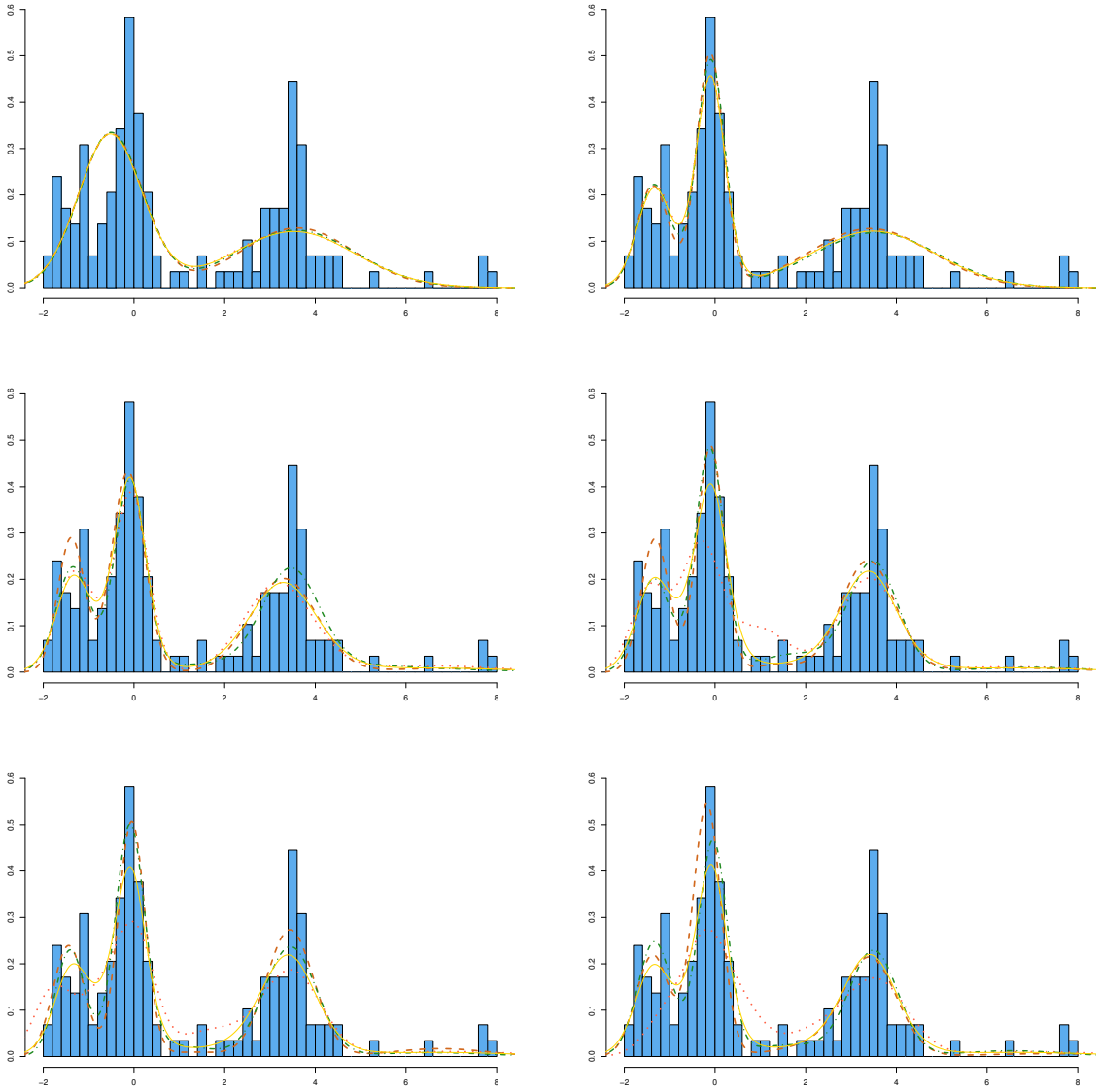

Figure 3: Simulated dataset of 146 observations with $K = 2, 3, 4, 5, 6, 7$ components fitted: average density (gold and full), plug-in density with average parameters (tomato and dots), plug-in density with marginal MAP parameters (forest green and dots-and-dashes), and plug-in density with joint MAP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

K	DIC_2	DIC_3	DIC_4	DIC_5	DIC_6	DIC_7	DIC_8
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.

same applies to Figure 3, for $K = 3$. The differences start to appear 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.

DIC_2 and DIC_3 behave similarly as for the galaxy dataset. DIC_5 and DIC_6 are not satisfactory producing negative p_D's, which for DIC_5 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 p_D's. Only DIC_4 and DIC_8 gives reasonable p_D's with DIC_4 selecting $K = 4$ while DIC_8 is selecting $K = 7$. Note that DIC_5 and DIC_6 select the right number of components! For 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 $\mathbf{Z}$ (while 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, DIC_4 does not necessarily lead to the most suitable model. For instance, in the mixture case, it chooses the mixture model with the

K	DIC_2	DIC_3	DIC_4	DIC_5	DIC_6	DIC_7	DIC_8
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).

cluster structure with 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 produces increasing p_Ds for increasing complexity is not very surprising, but it points out a drawback with this kind of criteria, 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.

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.

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.

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.

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

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, 1999.

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.

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