Bayesian Analysis (2014) **1**, Number 1, pp. 1–14

# Deviance Information Criteria for Mixtures of Distributions

Chanmin Kim *

**Abstract.** As a Bayesian criterion for model comparison, Spiegelhalter et al. (2002) proposed the deviance information criterion (DIC) consisting of two parts: a classical estimate of fit and the effective number of parameters. This model comparison method is constructed based on generalized linear models, and thus it may be inappropriate to use the classical DIC in the cases of the mixtures of distributions because of the "label switching" problem (Marin et al. 2005). Celeux et al. (2006) proposed several modified DIC constructions and assessed their behaviors under a mixture of distributions, but they do not provide the properties of alternative criteria well enough and those alternatives still suffer from the issue of using the observation data twice in estimating the posterior distribution and in computing the posterior mean of the expected log-likelihood. Here, we study $DIC_3$, one of the variations Celeux discussed, and propose a new criterion to lessen the potential bias from using the dataset twice. We compare our proposed criterion to other model selection criteria based on two numerical examples, the Galaxy dataset (Roeder 1990) and the simulated dataset.

**Keywords:** DIC, Mixtures of Distributions, Bayesian nonparametric models

## 1 Introduction

It is difficult to set the golden rule for model assessment and model comparison, despite a number of methods proposed to evaluate the goodness of models: well-known examples include AIC (Akaike 1973), BIC (Schwarz 1978), NIC (Murata et al. 1994) and TIC (Takeuchi 1976). All look for the model that best balances "model fit" and "model complexity". For Bayesian model selection, the deviance information criterion (DIC) Spiegelhalter et al. (2002) has been widely used, primarily developed for the case of generalized linear models. In the context of mixture models, however, DIC does not behave satisfactorily due to the multimodality issue and the so-called "label switching" issue (Marin et al. 2005) that renders the mixture model invariant under permutation of the indices of the components.

In order to overcome this problem, Celeux et al. (2006) proposed several modified DIC ($DIC_2 \sim DIC_8$) constructions and assessed their behavior under mixtures of distributions. Although their new constructions are mainly for latent variable and missing data models, $DIC_2$ and $DIC_3$ are for the fully observed data model and the latter is particularly dedicated to the case of the mixture models. The authors do not spend much time discussing the properties of $DIC_3$. Additionally, $DIC_3$, as with DIC, has a

*Department of Biostatistics, Harvard School of Public Health, Boston, MA02115 ckim@hsph.harvard.edu

 

deficiency in that it uses the data twice in estimating the posterior distribution and in computing the posterior mean of the expected log-likelihood which may result in biased estimates (Robert and Titterington 2002; Ando 2007).

These issues together have motivated us to explore the properties of $\mathrm{DIC}_3$ as a model selection tool in the context of mixture models and to develop a modified criterion to lessen the issue raised by using the observation data twice. We advocate implementing $\mathrm{DIC}_3$ to handle the "label switching" problem with the leave-one-out predictive density (Geisser and Eddy 1979; Gelfand 1996) to avoid the double use of the dataset. That is, for the likelihood of the $i$-th individual, we use all observations except the $i$-th observation to fit the model and use the remaining observation to evaluate the model. Although it seems like using the same dataset several times in terms of an overall likelihood, for an individual it actually reduces the overfitting by using the dataset once. We contend that our proposed criterion outperforms its competitors under several scenarios in Section 4.

The remainder of the article is organized as follows. In Section 2, we briefly introduce DIC from Spiegelhalter et al. (2002) and its problems with mixture models. In Section 3, we illustrate the alternative criterion $\mathrm{DIC}_3$ from Celeux et al. (2006) and examine its properties. Also, we propose a modified $\mathrm{DIC}_3$ to overcome the issue of using the data twice. We then compare our criterion to other deviance information criteria based on the Galaxy dataset (Roeder 1990) and a simulated dataset in Section 4. Section 5 concludes with a discussion.

## 2 Deviance Information Criterion

For a Bayesian criterion of model selection, Spiegelhalter et al. (2002) proposed the deviance information criterion (DIC), which has been used as a model selection criterion in much of the literature. In this section, we briefly introduce DIC and describe the issue with the mixture models. For data $\mathbf{y}$ and probability model $p(\mathbf{y}|\boldsymbol{\theta})$ with parameters of interest $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the Bayesian deviance is defined as

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

where $h(\mathbf{y})$ is taken to be a fully specified standardizing term that is a function of data alone. Then, their proposal of the effective number of parameters (or effective dimensions) is defined as

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

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

$$\overline{D(\boldsymbol{\theta})} = E_{\theta|y}[-2 \log p(\mathbf{y}|\boldsymbol{\theta})] + 2 \log h(\mathbf{y}), \tag{2}$$

and $D(\overline{\boldsymbol{\theta}})$ is the deviance measured at the posterior mean of the parameters $\boldsymbol{\theta}$. In Equation (1), $p_D$ is independent of the choice of the function $h(\cdot)$ by the definition. Under the assumption of a normal approximation to the likelihood and negligible prior information, Spiegelhalter et al. (2002) show that $p_D$ is approximately the true number of parameters. Also, it can be interpreted as the degree of "overfitting" due to

incorporating $\overline{\boldsymbol{\theta}}$ to estimate parameters of interest in a Bayesian perspective. Hence, we can think of $p_D$ as the effective number of parameters in a model or its complexity so that it functions as a penalty term. It is worth mentioning that the effective number of parameters is different from the nominal number of parameters, which determines the penalty term for model complexity in AIC. The nominal number of parameters is directly computed from a model, while the effective number of parameters is obtained from the data (Shriner and Yi 2009). It is also worth noting that Ando (2007) argues $p_D$ can be thought of as the asymptotic bias to the posterior mean of the expected log likelihood. This is based on the aforementioned discussion from Robert and Titterington (2002) such that the same data were used twice in the construction of DIC. We revisit this issue in Section 3.2.

It is well documented that the posterior mean $\overline{\boldsymbol{\theta}}$ could be replaced by the mode and median in computing $D(\overline{\boldsymbol{\theta}})$ (Celeux et al. 2006). A benefit of using the posterior mean $\overline{\boldsymbol{\theta}}$ as an estimator of parameters of interest is that $p_D \geq 0$ for a log-concave likelihood in $\boldsymbol{\theta}$ by Jensen's inequality. Yet a negative $p_D$ can be still generated by non-log-concave likelihoods as well as poor posterior mean estimates, which will be discussed in Section 2.1

Since Equation (2) measures a Bayesian model fit and Equation (1) penalizes for model complexity, Spiegelhalter et al. (2002) defines the deviance information criterion DIC as

$$\begin{aligned} \text{DIC} &= \overline{D(\boldsymbol{\theta})} + p_D \\ &= E_{\theta|y}[-2\log p(\mathbf{y}|\boldsymbol{\theta})] + E_{\theta|y}[-2\log p(\mathbf{y}|\boldsymbol{\theta})] + 2\log p(\mathbf{y}|\overline{\boldsymbol{\theta}}) \\ &= E_{\theta|y}[-4\log p(\mathbf{y}|\boldsymbol{\theta})] + 2\log p(\mathbf{y}|\overline{\boldsymbol{\theta}}), \end{aligned}$$

where $h(\mathbf{y})$ is set to 1. Markov Chain Monte Carlo (MCMC) typically makes it straightforward to compute $\overline{\boldsymbol{\theta}}$ and DIC.

As we mentioned earlier, however, there are issues with DIC, which include the issue Spiegelhalter et al. (2002) raised and a negative $p_D$ due to non-log-concave likelihoods. Besides those, there is another issue: that the choice of $\overline{\boldsymbol{\theta}}$ is difficult in distributions that are not unimodal. As is typically the case for a mixture of distributions.

## 2.1 Issue for Mixtures of Distributions

Mixtures of distributions are easy to construct and are widely used to flexibly estimate a density. Such mixture models are invariant under permutation of the indices of the components so that we cannot identify the parameter $\theta_i \in (\theta_1, \cdots, \theta_K)$ marginally (Marin et al. 2005), where $\theta_i$ denotes the parameter of the $i$-th distribution in the mixture. In terms of the likelihood, parameter $\theta_i$ is indistinguishable from parameter $\theta_j$ for $i \neq j$. This exchangeability implies $O(K!)$ modes for a $K$ components mixture model since one local maximum $(\theta_1, \cdots, \theta_K)$ gives rise to $K!$ local maxima for all permutations of $(1, \cdots, K)$. On the level of prior distributions, if we use an exchangeable prior on $(\theta_1, \cdots, \theta_K)$, marginals of $\theta_i$ for $i = 1, \cdots, K$ are identical.

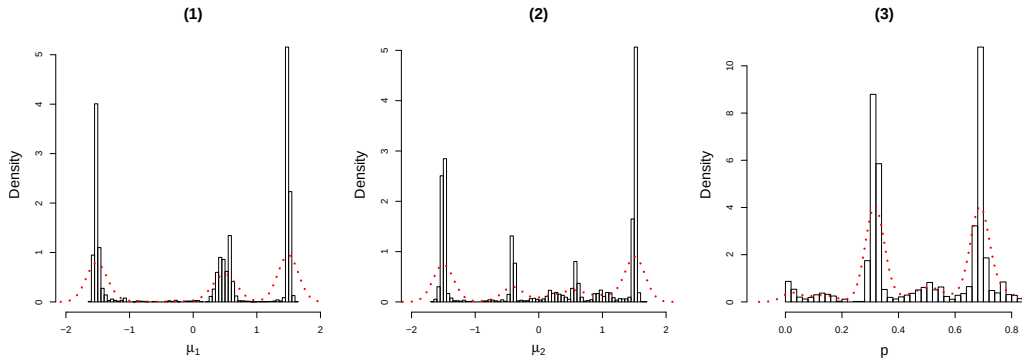


Figure 1: Posterior distributions of $\mu_1$, $\mu_2$ and $p$ from 2 chains. In each plot, it has two modes clustered around the true values (and additional modes), which implies two components of the mixture model are indistinguishable.

Figure 1 illustrates label-switching phenomenon (Marin et al. 2005) via the simple example of a two-component normal mixture

$$p \times \mathcal{N}(\mu_1, \sigma_1^2) + (1-p) \times \mathcal{N}(\mu_2, \sigma_2^2), \tag{3}$$

where the vector of parameters, $\boldsymbol{\theta} = (p, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$, is taken to be $(0.3, -1.5, 1.5, 0.5^2, 0.5^2)$ for simulating a dataset of 2000 observations from the model (3). Based on the simulated dataset, we run the model in OpenBUGS to obtain posterior samples of parameters. Figure 1 represents posterior samples of $\mu_1, \mu_2$ and $p$ from 2 chains of 15000 MCMC iterations after discarding 5000 runs as burn-in. Each posterior distribution, as we expected, has two modes clustered around the true values of parameters, $\mu_1, \mu_2, p$ and $1-p$ due to the fact that parameters under label 1 and label 2 are not distinguishable from each other and additional modes to represent a single component distribution. Here, the posterior mean estimates are $E_{\theta|y}(\mu_1) = 0.197$ and $E_{\theta|y}(\mu_2) = 0.155$ which may lead to a negative value of $p_D$. To remedy a problem, we may impose an constraint on the parameter $p$ such as $p \in [0, 0.5)$. However, it results in $E_{\theta|y}(\mu_1) = -0.392$ and $E_{\theta|y}(\mu_2) = 0.153$, which are still biased.

To obtain DIC, we need to compute the deviance measured at the posterior mean (or mode, or median) of the parameters $\boldsymbol{\theta}$, $D(\overline{\boldsymbol{\theta}})$. This plug-in estimate of the deviance is inadequate and may generate a negative value if the posterior means (or modes, or medians) of parameters are poor estimators, as we observed in the previous example in Equation (3). Hence, DIC is an improper comparison criterion in the context of the mixture of distributions, as such this original construction of DIC needs to be slightly modified.

# 3 Alternative Criteria for Model Comparison

To overcome the issue raised in the mixture of distributions (and missing data models), Celeux et al. (2006) proposed 7 alternative constructions of DIC such as $\text{DIC}_2, \text{DIC}_3, \cdots, \text{DIC}_8$. Among those, the $\text{DIC}_3$ is designed for the case of mixtures of distributions, in that the inferential focus is on the density $p(\mathbf{y}|\boldsymbol{\theta})$ itself instead of parameters $\boldsymbol{\theta}$. Specifically, $D(\overline{\boldsymbol{\theta}})$ is estimated from $\hat{p}(\mathbf{y})$, an estimator of the density $p(\mathbf{y}|\boldsymbol{\theta})$, which is invariant to label switching of the component. This density estimator is easily obtained by MCMC simulations: for example, in a mixture of normal distributions

$$p(y_i \,|\, \boldsymbol{\theta}) = \sum_{j=1}^{K} \pi_j \, \mathcal{N}(y_i \,|\, \mu_j, \sigma_j^2),$$

we can obtain $\hat{p}(\mathbf{y})$ as follows

$$\hat{p}(\mathbf{y}) = \prod_{i=1}^{n} \hat{p}(y_i) = \prod_{i=1}^{n} \left\{ \frac{1}{L} \sum_{l=1}^{L} \sum_{j=1}^{K} \pi_j^{(l)} \, \mathcal{N}(y_i \,|\, \mu_j^{(l)}, \sigma_j^{2,(l)}) \right\},$$

where $\boldsymbol{\theta} = \{\mu_j, \sigma_j^2, \pi_j \,:\, j = 1, \cdots, K\}$ and $(\mu_j^{(l)}, \sigma_j^{2,(l)}, \pi_j^{(l)})$ are a set of posterior samples for the $j$-th component parameters from the $l$-th MCMC iteration. Then, this approximately holds $\hat{p}(y_i) \approx E_{\theta|y}[p(y_i|\boldsymbol{\theta})]$ which results in $\hat{p}(\mathbf{y}) = \prod_{i=1}^{n} \hat{p}(y_i) \approx \prod_{i=1}^{n} E_{\theta|y}[p(y_i|\boldsymbol{\theta})]$. Finally, Celeux et al. (2006) defined $\text{DIC}_3$ as

$$\begin{aligned} \text{DIC}_3 &= \overline{D(\boldsymbol{\theta})} + p_D^{\star} \\ &\approx E_{\theta|y}[-2 \log p(\mathbf{y}|\boldsymbol{\theta})] + \big\{ E_{\theta|y}[-2 \log p(\mathbf{y}|\boldsymbol{\theta})] + 2 \log \hat{p}(\mathbf{y}) \big\}, \end{aligned} \tag{4}$$

where $p_D^{\star}$ indicates the new effective dimension.

Celeux et al. (2006), however, do not discuss the properties of $\text{DIC}_3$ after they propose the constructive definition of it. Here, we examine a property of $\text{DIC}_3$ that was never made explicit through the new form of $p_D^{\star} = E_{\theta|y}[-2 \log p(\mathbf{y}|\boldsymbol{\theta})] + 2 \log \hat{p}(\mathbf{y})$.

## 3.1 Properties of $\text{DIC}_3$

To exploit particular properties of $\text{DIC}_3$, we need to rewrite $p_D^{\star}$ as

$$\begin{aligned} p_D^{\star} &= -2 E_{\theta|y}[\log p(\mathbf{y}|\boldsymbol{\theta})] + 2 \log \hat{p}(\mathbf{y}) \\ &\approx -2 E_{\theta|y}\big[ \log \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta}) \big] + 2 \log \prod_{i=1}^{n} E_{\theta|y}\big[ p(y_i|\boldsymbol{\theta}) \big] \\ &= -2 \sum_{i=1}^{n} E_{\theta|y}\big[ \log p(y_i|\boldsymbol{\theta}) \big] + 2 \sum_{i=1}^{n} \log E_{\theta|y}\big[ p(y_i|\boldsymbol{\theta}) \big] \\ &= -2 \sum_{i=1}^{n} \Big\{ E_{\theta|y}\big[ \log p(y_i|\boldsymbol{\theta}) \big] - \log E_{\theta|y}\big[ p(y_i|\boldsymbol{\theta}) \big] \Big\}. \end{aligned} \tag{5}$$

We now introduce a measurable function $X_i$ such that $X_i(\boldsymbol{\theta}) = p(y_i|\boldsymbol{\theta})$. Then, we expand $\log X_i$ around $E_{\theta|y}[X_i]$ to the second order:

$$\log X_i \approx \log E_{\theta|y}[X_i] + \frac{1}{E_{\theta|y}[X_i]}\big(X_i - E_{\theta|y}[X_i]\big) - \frac{1}{2\big(E_{\theta|y}[X_i]\big)^2}\big(X_i - E_{\theta|y}[X_i]\big)^2 \quad (6)$$

After taking expectation on both sides with respect to the posterior distribution of $\boldsymbol{\theta}$, Equation (6) implies:

$$\begin{aligned} -2E_{\theta|y}[\log X_i] + 2\log E_{\theta|y}[X_i] &= \frac{1}{\big(E_{\theta|y}[X_i]\big)^2} E_{\theta|y}\big[\big(X_i - E_{\theta|y}[X_i]\big)^2\big] \\ &= \frac{V_{\theta|y}\big[X_i\big]}{\big(E_{\theta|y}[X_i]\big)^2}, \end{aligned} \quad (7)$$

where the first order term in Equation (6) is canceled out by the facts that $E_{\theta|y}\big[X_i - E_{\theta|y}[X_i]\big] = 0$ and $V_{\theta|y}\big[X_i\big]$ is the variance of $X_i$ with respect to the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$. If we replace $X_i$ with $p(y_i|\boldsymbol{\theta})$ in Equation (7) and plug it into Equation (5), then $p_D^\star$ can be represented as

$$p_D^\star = \sum_{i=1}^{n} \left\{ \frac{V_{\theta|y}\big[p(y_i|\boldsymbol{\theta})\big]}{\big(E_{\theta|y}[p(y_i|\boldsymbol{\theta})]\big)^2} \right\}. \quad (8)$$

Here, Equation (8) indicates that $p_D^\star$ is nonnegative. Hence, a potential issue of a negative $p_D$ from the poor plug-in estimators, which was pointed out in Section 2.1, is less of an issue in $\mathrm{DIC}_3$. To be precise, this result is based on an asymptotic expansion for $\log p(y_i|\boldsymbol{\theta})$ so that this only explains the case that the posterior distribution of $\boldsymbol{\theta}$ concentrates values of $p(y_i|\boldsymbol{\theta})$ around $E_{\boldsymbol{\theta}|y}[p(y_i|\boldsymbol{\theta})]$. However, it can also be seen by Jensen's inequality $E_{\theta|y}\big[\log p(y_i|\boldsymbol{\theta})\big] \leq \log E_{\theta|y}\big[p(y_i|\boldsymbol{\theta})\big]$ if well behaved densities and by Equation (5).

Moreover, Equation (8) can be represented as a function of a coefficient of variation (CV)

$$p_D^\star = \sum_{i=1}^{n} \left\{ \frac{\sqrt{V_{\theta|y}\big[p(y_i|\boldsymbol{\theta})\big]}}{\big(E_{\theta|y}[p(y_i|\boldsymbol{\theta})]\big)} \right\}^2 = \sum_{i=1}^{n} CV_i^2, \quad (9)$$

where $CV_i$ is the ratio of the standard deviation $\sqrt{V_{\theta|y}\big[p(y_i|\boldsymbol{\theta})\big]}$ to the mean $E_{\theta|y}[p(y_i|\boldsymbol{\theta})]$ for $i = 1, \cdots, n$. In the CV for the model fit, generally, we consider the numerator of the CV as the root mean squared deviation and regard a lower CV as being suggestive of a good model fit due to the smaller residuals relative to the predicted value. In our case, the mean $E_{\theta|y}[p(y_i|\boldsymbol{\theta})]$ is the posterior predictive distribution for a data point $y_i$ and the variance $V_{\theta|y}\big[p(y_i|\boldsymbol{\theta})\big] = E_{\theta|y}[(p(y_i|\boldsymbol{\theta}) - E_{\theta|y}[p(y_i|\boldsymbol{\theta})])^2]$ evaluates the closeness of the estimated predictive distribution to the distributions under various values of $\theta$ so that $p_D^\star$ in (9) measures how well predictive distributions, $\hat{p}(y_i) = E_{\theta|y}[p(y_i|\boldsymbol{\theta})]$ for $i = 1, \cdots, n$, are estimated with less variability. In this setting, the model with the

smaller $p_D^\star$ is indicative of a good model fit. Hence, the application of $p_D^\star$ is consistent with that of the previous $p_D$ in Section 2 in terms of preferring smaller values.

It is worth noting that the widely applicable information criterion (WAIC) (Watanabe 2010) resembles the $\text{DIC}_3$ in that WAIC uses the slightly different functional variance form

$$p_{WAIC} = \sum_{i=1}^{n} V_{\theta|y}\big[\log p(y_i|\boldsymbol{\theta})\big]$$

to measure the effective number of parameters and in that it uses $\log \hat{p}(y_i)$ instead of $\log p(y_i|\overline{\boldsymbol{\theta}})$ to evaluate the deviance. Although both criteria can be used in the case of the mixture models to guarantee a positive effective number of parameters, $\text{DIC}_3$ stems directly from the DIC which means both that it is comparable to DIC in its interpretation and that preference of smaller $p_D^\star$ is advocated by the coefficient of variation for the model fit.

## 3.2 New Criterion for Model Comparison

As we mentioned earlier, $\text{DIC}_3$ shares with DIC that it uses the data $\mathbf{y}$ twice in their construction (Robert and Titterington 2002). First, the data is used to estimate the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$, it is used the second time to compute the posterior expectation of $p(\mathbf{y}|\boldsymbol{\theta})$ or the posterior predictive distribution $\hat{p}(\mathbf{y})$. As a result, the model chosen by $\text{DIC}_3$(DIC) may be too complex due to overfitting (Ando 2007).

Here, we propose a modified $\text{DIC}_3$ motivated by the leave-one-out cross validation predictive density (Geisser and Eddy 1979), also known as the conditional predictive ordinate (CPO) (Gelfand 1996), to avoid the double use of the data. Let $\mathbf{y}_{(i)} = (y_1, \cdots, y_{i-1}, y_{i+1}, \cdots, y_n)$ denote the set $\mathbf{y}$ with the $i$-th observation $y_i$ set aside. For the likelihood of the $i$-th observation, $y_{(i)}$ is used to fit the model and estimate the posterior distribution and $y_i$ is used to evaluate the model. Then the leave-one-out posterior mean of deviance (hereafter referred to as $\overline{D(\boldsymbol{\theta})}_{loo}$) is defined as

$$\overline{D(\boldsymbol{\theta})}_{loo} = -2\sum_{i=1}^{n}\left\{\int \log p(y_i \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \,|\, \mathbf{y}_{(i)})\, d\boldsymbol{\theta}\right\}, \tag{10}$$

and the leave-one-out posterior predictive distribution (hereafter referred to as $\hat{p}(\mathbf{y})_{loo}$) is specified as

$$\hat{p}(\mathbf{y})_{loo} = \prod_{i=1}^{n} \hat{p}(y_i)_{loo} = \prod_{i=1}^{n}\left\{\int p(y_i \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \,|\, \mathbf{y}_{(i)})\, d\boldsymbol{\theta}\right\}. \tag{11}$$

With Equations (10) and (11), we now propose a leave-one-out version of $\text{DIC}_3$ (hereafter

referred to as $\text{DIC}_{loo}$) as follows:

$$\begin{aligned}
&\text{DIC}_{loo} \\
&= \overline{D(\boldsymbol{\theta})}_{loo} + p_{loo} \\
&= \overline{D(\boldsymbol{\theta})}_{loo} + \left\{\overline{D(\boldsymbol{\theta})}_{loo} + 2\log \hat{p}(\mathbf{y})_{loo}\right\} \\
&= \sum_{i=1}^{n} E_{\theta|\mathbf{y}_{(i)}}[-2\log p(y_i \,|\, \boldsymbol{\theta})] + \left\{\sum_{i=1}^{n} E_{\theta|\mathbf{y}_{(i)}}[-2\log p(y_i \,|\, \boldsymbol{\theta})] + 2\sum_{i=1}^{n} \log E_{\theta|\mathbf{y}_{(i)}}[p(y_i \,|\, \boldsymbol{\theta})]\right\} \\
&= -4\sum_{i=1}^{n}\left\{\int \log p(y_i \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \,|\, \mathbf{y}_{(i)})\, d\boldsymbol{\theta}\right\} + 2\sum_{i=1}^{n}\left\{\log \int p(y_i \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \,|\, \mathbf{y}_{(i)})\, d\boldsymbol{\theta}\right\}. \qquad (12)
\end{aligned}$$

Hence, $\text{DIC}_{loo}$ can avoid a potential overfitting problem by using the data once both in constructing a posterior distribution and in computing a predictive distribution for each $i$-th observation $i = 1, \cdots, n$.

$\text{DIC}_{loo}$ inherits the aforementioned properties in Section 3.1 Specifically, Equation (8) can be easily restated with respect to the leave-one-out effective dimension (hereafter $p_{loo}$) such that

$$p_{loo} = \sum_{i=1}^{n}\left\{\frac{V_{\theta|y_{(i)}}[p(y_i|\boldsymbol{\theta})]}{(E_{\theta|y_{(i)}}[p(y_i|\boldsymbol{\theta})])^2}\right\}, \qquad (13)$$

where $E_{\theta|y_{(i)}}(\cdot)$ and $V_{\theta|y_{(i)}}(\cdot)$ are the mean and variance functions, respectively, with respect to the leave-one-out posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y}_{(i)})$ for the $i$-th observation. This implies that effective dimension is positive and that the smaller value is better in terms of the coefficient of variation representation.

We can obtain another benefit from $\text{DIC}_{loo}$ through the second term of Equation (12), also known as the CPO, in that we can check a model fit for an individual point $y_i$. If the CPO for the $i$-th observation $\text{CPO}_i = \int p(y_i|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y}_{(i)})d\boldsymbol{\theta}$ is low, then we consider the $i$-th observation as an outlier, given the current model fit, so that we can determine the model adequacy in each observation level as well as in population level. Congdon (2005) recommends using scaled CPOs which are obtained after dividing each $\text{CPO}_i$ by its maximum and considering observations with scaled CPOs under 0.01 as outliers. The CPO is readily available from the process of computing $\text{DIC}_{loo}$ so that we can use both DIC and CPO for evaluating the model fit.

Another approach comparable to $\text{DIC}_{loo}$ is the leave-one-out cross validation whose effective dimension (hereafter referred to as $p_{loo-cv}$) is defined Gelman et al. (2013) as

$$p_{loo-cv} = 2\sum_{i=1}^{n} \log E_{\theta|y}[p(y_i|\boldsymbol{\theta})] - 2\sum_{i=1}^{n} \log E_{\theta|y_{(i)}}[p(y_i|\boldsymbol{\theta})], \qquad (14)$$

where the first term is deviance (multiplied by -1) and the second term is the logarithm of the cross validated pointwise probability ($\text{lppd}_{loo-cv}$) (Gelman et al. 2013). The discrepancy between this criterion and ours is that $p_{loo}$ gives a smaller penalty compared to $p_{loo-cv}$ when the CPO approximates the posterior predictive distribution well. Thus,

our approach ($\text{DIC}_{loo}$) is more likely to select the model that explains the data with less variability (i.e., the CPO approximates the posterior predictive distribution well for $i = 1, \cdots, n$). This can be seen by re-expressing Equations (13) and (14) as

$$p_{loo} = \sum_{i=1}^{n} \left\{ \frac{p(y_i|\mathbf{y})}{p(y_i|\mathbf{y}_{(i)})} - 1 \right\},$$

and

$$p_{loo-cv} = 2\sum_{i=1}^{n} \log \frac{p(y_i|\mathbf{y})}{p(y_i|\mathbf{y}_{(i)})},$$

where $p(y_i|\mathbf{y}) = \int p(y_i|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$ and $p(y_i|\mathbf{y}_{(i)}) = \int p(y_i|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y}_{(i)})d\boldsymbol{\theta}$. A detailed derivation is in the supplementary material. Hence, when there is a large discrepancy between the CPO and the posterior predictive distribution overall, $p_{loo}$ gives a larger penalty than $p_{loo-cv}$ does by the fact that $2\log x < x - 1$ for $x > 3.51$(approximately) (e.g., $\frac{p(y_i|\mathbf{y})}{p(y_i|\mathbf{y}_{(i)})} > 3.51$). In contrast, $p_{loo}$ gives a smaller penalty than $p_{loo-cv}$ does when there is a small discrepancy between the CPO and the posterior predictive distribution by the fact that $2\log x \geq x - 1$ for $1 \leq x \leq 3.51$(approximately) (e.g., $1 \leq \frac{p(y_i|\mathbf{y})}{p(y_i|\mathbf{y})_{(i)}} \leq 3.51$).

### Posterior Computation

The quantities in Equations (10) and (11) can be estimated using MCMC such that, for example in the case of normal mixtures from Section 3, the second term of Equation (12) can be approximated by

$$2\log \hat{p}(\mathbf{y})_{loo} = 2\sum_{i=1}^{n} \log \hat{p}(y_i)_{loo} \approx 2\sum_{i=1}^{n} \log \left\{ \frac{1}{L}\sum_{l=1}^{L}\sum_{j=1}^{K} \pi_{j,(i)}^{(l)}\, \mathcal{N}(y_i \,|\, \mu_{j,(i)}^{(l)}, \sigma_{j,(i)}^{2,(l)}) \right\}, \tag{15}$$

where $(\mu_{j,(i)}^{(l)}, \sigma_{j,(i)}^{2,(l)}, \pi_{j,(i)}^{(l)})$ are a set of the $l$-th posterior samples for the $j$-th component parameters from the $i$-th leave-one-out posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y}_{(i)})$. Similarly, the first term of Equation (12) can be approximated.

However, this naive approximation is computationally expensive in that, for each observation $y_i$, it requires a separate MCMC run to estimate the leave-one-out posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y}_{(i)})$. That is, for a large number of observations $n$, it is computationally burdensome to run all $n$ separate MCMC algorithms. To overcome this issue, Gelfand (1996) shows that the leave-one-out predictive distribution can be estimated by a Monte Carlo integration with respect to the posterior predictive distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$ so that this approach just needs to run a single MCMC algorithm. Specifically, we can calculate the MC integration approximation of the first term of Equation (12) as

$$-4\sum_{i=1}^{n} \left\{ \int \log p(y_i \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \,|\, \mathbf{y}_{(i)})\, d\boldsymbol{\theta} \right\} \quad \approx \quad -4\sum_{i=1}^{n} \left\{ \frac{1}{H}\sum_{h=1}^{H} b_h^i \log p(y_i \,|\, \boldsymbol{\theta}_h) \right\},$$

where $\boldsymbol{\theta}_h$ is a set of posterior samples from the $h$-th MCMC iteration of $\pi(\boldsymbol{\theta}|\mathbf{y})$ and $b_h^i$ is a weight for importance sampling such that

$$b_h^i = \frac{\pi(\boldsymbol{\theta}_h \,|\, \mathbf{y}_{(i)})}{\pi(\boldsymbol{\theta}_h \,|\, \mathbf{y})} \approx \frac{1}{\left\{ \frac{1}{J} \sum_{j=1}^{J} \frac{1}{p(y_i \,|\, \boldsymbol{\theta}_j)} \right\} p(y_i \,|\, \boldsymbol{\theta}_h)}, \tag{16}$$

under assuming the conditional independence of $y_i$ given $\theta$ (Gelfand 1996). A detailed derivation of Equation (16) is in the supplementary material. The computation of the second term of Equation (12) follows similarly. Then, the $\mathrm{DIC}_{loo}$ can be estimated by a single MCMC run:

$$\mathrm{DIC}_{loo} \;=\; \sum_{i=1}^{n} \left[ -\frac{4}{H} \sum_{h=1}^{H} b_h^i \log p(y_i \,|\, \boldsymbol{\theta}_h) + 2 \log \left\{ \frac{1}{H} \sum_{h=1}^{H} b_h^i \, p(y_i \,|\, \boldsymbol{\theta}_h) \right\} \right],$$

where $b_h^i$ is specified in Equation (16) and $\boldsymbol{\theta}_h$ is a set of the $h$-th posterior samples drawn from $\pi(\boldsymbol{\theta}|\mathbf{y})$. Hence, a single MCMC run suffices to calculate $\mathrm{DIC}_{loo}$.

Since it is equal to using draws of $\boldsymbol{\theta}$ from the full posterior distribution, $\pi(\boldsymbol{\theta}|\boldsymbol{y})$, with weight proportional to $1/p(y_i \,|\, \boldsymbol{\theta}_h)$, there may be an issue that the importance-weighted estimate can be unstable due to the unbounded weights, $1/p(y_i \,|\, \boldsymbol{\theta}_h)$. In that case, Vehtari and Gelman (2014) suggest using truncated importance resampling which is replacing $w_h = 1/p(y_i \,|\, \boldsymbol{\theta}_h)$ by $\min(w_h, \sqrt{H}\bar{w})$ where $w = \sum w_h/H$ to stabilize the weights.

# 4 Numerical Examples

In this section, we compare our proposed method ($\mathrm{DIC}_{loo}$) with other DICs based on the Galaxy dataset (Roeder 1990) and the simulated dataset that Celeux et al. (2006) used for their numerical comparisons. The former contains the velocities of 82 galaxies and can be found in the `MASS::galaxies` library of `R`. The latter includes 164 observations simulated from the 4 components mixture of normals

$$0.288\mathcal{N}(0, 0.2^2) + 0.260\mathcal{N}(-1.5, 0.5^2) + 0.171\mathcal{N}(2.2, 3.4^2) + 0.281\mathcal{N}(3.3, 0.5^2). \tag{17}$$

In Table 1, we obtain the results after fitting $K$-component mixtures of normals for the Galaxy dataset. As expected, the classical DIC is not appropriate for the case due to generating negative values of effective dimension ($p_D$). It gives negative values of $p_D$ after $K = 4$ which may indicate that the label switching issue becomes substantial in the plug-in estimate as the number of component increases. The remaining penalties behave well in terms of being positive and non-decreasing with the number of mixture components ($p_D^\star, p_{loo}, p_{loo-cv}$). However, their magnitudes are different such that the effective dimension of $\mathrm{DIC}_3$ ($p_D^\star$) is always smaller than those of two competing criteria and the effective dimension of $\mathrm{DIC}_{loo-cv}$ ($p_{loo-cv}$) is the largest. For a given $K$-component mixture of independent normals, we would anticipate the effective number of parameters should be around $3K - 1$ ($K$ mean parameters, $K$ variance parameters, $K - 1$ weight

parameters). As one can see, our proposed method ($\text{DIC}_{loo}$) generates an effective dimension closest to the expected value for the given $K$ component mixture model among three criteria $\text{DIC}_3$, $\text{DIC}_{loo}$ and $\text{DIC}_{loo-cv}$. When we examine the results for $K = 7$, for example,$\text{DIC}_3$ underestimates the appropriate effective dimension by $7.811(20 - 12.189)$ and $\text{DIC}_{loo-cv}$ overestimate it by $7.499(27.499 - 20)$, whereas $\text{DIC}_{loo}$ gives 19.316 which is the closest to 20. Henceforth, $\text{DIC}_{loo}$ behaves properly with respect to the effective dimension.

When it comes to the values of DICs, $\text{DIC}_3$ and $\text{DIC}_{loo-cv}$ agree as both identify the $K = 6$ component mixture of normals as the best model. On the contrary, $\text{DIC}_{loo}$ chooses the $K = 3$ model. Here, $\text{DIC}_{loo}$ is distinct from the other two criteria in that $\text{DIC}_{loo}$ prefers a parsimonious model for the number of components in mixture models. In Figure 2, posterior predicted densities are plotted under $K = 3$ (left) and $K = 6$ (right) scenarios along with the density of the observed data from the Galaxy dataset. It is worth noting that the most substantial difference between two predictive densities is the model under $K = 6$, which captures additional two posterior modes at the cost of incorporating $3 \times 3 - 1 = 8$ more parameters.

| K | DIC ($p_D$) | | $\text{DIC}_3$ ($p_D^\star$) | | $\text{DIC}_{loo}$ ($p_{loo}$) | | $\text{DIC}_{loo-cv}$ ($p_{loo-cv}$) | |
|---|---|---|---|---|---|---|---|---|
| 2 | 564.19 | (4.93) | 563.11 | (3.85) | 572.96 | (4.68) | 563.61 | (8.21) |
| 3 | 534.14 | (6.21) | 533.79 | (5.86) | 557.38 | (10.62) | 536.14 | (14.08) |
| 4 | 506.32 | (-14.39) | 530.08 | (9.37) | 562.43 | (14.78) | 532.86 | (21.51) |
| 5 | 346.78 | (-172.23) | 529.99 | (10.97) | 565.92 | (16.59) | 532.75 | (23.70) |
| 6 | 218.48 | (-298.84) | 528.62 | (11.31) | 569.64 | (18.78) | 532.08 | (26.08) |
| 7 | 266.01 | (-252.03) | 530.23 | (12.19) | 571.98 | (19.32) | 533.35 | (27.49) |

Table 1: Results from various model selection criteria with effective dimensions (in parentheses) for the Galaxy dataset. After running 20000 MCMC interations, discard the first 10000 interation as burn-in. K represents the number of normal components.

A benefit we can obtain from DICs for mixture models is that those criteria can be used in a comparison of Bayesian nonparametric models such as

$$\begin{aligned} Y &\sim \mathcal{N}(\mu_i, \Sigma_i), \quad i = 1, \cdots, n, \\ (\mu_i, \Sigma_i) &\sim G, \quad i = 1, \cdots, n, \\ G &\sim DP(\alpha\, G_0), \end{aligned}$$

where $G_0$ is the base distribution and $\alpha$ is the mass parameter. This is possible by making use of the stick-breaking construction (Sethuraman 1994), which essentially implies a mixture model as follows:

$$f(y) = \sum_{k=1}^{\infty} \beta_k, \mathcal{N}(y; \mu_k, \Sigma_k),$$

where $\beta_k = \beta'_k \prod_{j<k}(1 - \beta'_j)$, $\beta'_j \sim \text{Beta}(1, \alpha)$ and $(\mu_k, \Sigma_k) \sim G_0$. Although it is an infinite mixture of normals, for computations, it is approximated by setting $\beta'_C \equiv 1$

with a fixed $C$ (Ishwaran and James 2001)

$$f(y) = \sum_{k=1}^{C} \beta_k, \mathcal{N}(y; \mu_k, \Sigma_k),$$

which is now a finite mixture model. Under this Bayesian nonparametric model with the maximum number of clusters $C$ being 30, we obtain the results of the aforementioned DIC's based on the Galaxy dataset in Table 2. BUGS code for the Bayesian nonparametric model is in the supplementary material. Like the previous results under the parametric models, classical DIC gives a negative value for effective dimension $p_D$. The effective dimension is the smallest under $\text{DIC}_3$ and the largest when $\text{DIC}_{loo-cv}$ is considered. When $\text{DIC}_3$ is used for a model selection criterion, we prefer a parametric model only when $K = 5, 6$. Under $\text{DIC}_{loo}$, we prefer a parametric model when $K = 3, 4, 5, 6$. For $\text{DIC}_{loo-cv}$, we prefer a parametric model when $K = 4, 5, 6$. It is worth noting that the DIC values under the Bayesian nonparametric model are similar to those under the parametric model of $K = 7$. This indicates that the effective number of clusters being used in the Bayesian nonparametric model is approximately 7.

| | DIC ($p_D$) | | $\text{DIC}_3$ ($p_D^\star$) | | $\text{DIC}_{loo}$ ($p_{loo}$) | | $\text{DIC}_{loo-cv}$ ($p_{loo-cv}$) | |
|---|---|---|---|---|---|---|---|---|
| BNP | 182.65 | (-335.11) | 530.02 | (12.25) | 570.43 | (18.58) | 533.28 | (27.77) |

Table 2: Results for the Galaxy dataset under a Bayesian nonparametric (BNP) model with setting $C = 30$ which allows the maximum number of clusters to be 30.

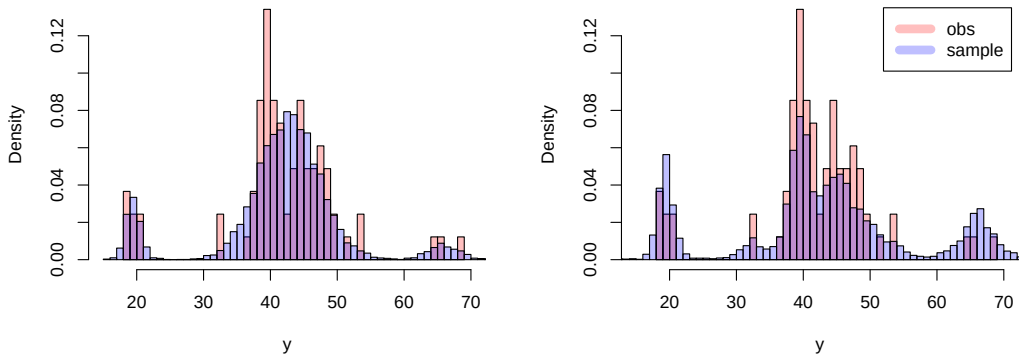


Figure 2: Density plots of 82 observations for the Galaxy dataset (obs; red) and 10000 samples from $K$ component mixture models (sample; blue): $K = 3$ (left) and $K = 6$ (right).

In addition to the Galaxy dataset, we analyzed the simulated dataset of 164 observations comprised of 4 component mixture of normal distributions in (17). In Table 3,

the results from the DICs are presented. Similar to the previous result, classical DIC behaves unsatisfactorily in that it has negative effective dimensions ($p_D$) after $K = 5$. Interestingly, $\text{DIC}_3$ and $\text{DIC}_{loo-cv}$ do not behave well in terms of effective dimensions ($p_D^\star$ and $p_{loo-cv}$) since their values are fluctuating after $K = 5$, which may not be proper. On the other hand, $\text{DIC}_{loo}$ behaves satisfactorily in terms of non-decreasing $p_{loo}$. Also, its value is close to $3K - 1$ for a given $K$ component model.

Although two criteria, $\text{DIC}_3$ and $\text{DIC}_{loo-cv}$, are not satisfactory in the values of effective dimensions, $\text{DIC}_3$ selects the $K = 4$ component model as the best model, which is an appropriate choice under our simulation and $\text{DIC}_{loo-cv}$ chooses the model with an additional component ($K = 5$). It is worth noting, however, that the differences of deviance information criteria between $K = 4$ and $K = 5$ are subtle in the cases of $\text{DIC}_3$ and $\text{DIC}_{loo-cv}$, while our proposed $\text{DIC}_{loo}$ chooses the $K = 4$ component model by a considerable margin. In Figure 3, the density plots of 164 observations and 10,000 predictive posterior samples under the $K = 4$ components normal model are depicted.

| K | DIC ($p_D$) | | $\text{DIC}_3$ ($p_D^\star$) | | $\text{DIC}_{loo}$ ($p_{loo}$) | | $\text{DIC}_{loo-cv}$ ($p_{loo-cv}$) | |
|---|---|---|---|---|---|---|---|---|
| 2 | 673.57 | (4.42) | 672.80 | (3.66) | 680.54 | (3.2) | 672.91 | (7.42) |
| 3 | 623.57 | (7.16) | 623.33 | (6.92) | 639.05 | (7.64) | 623.77 | (14.28) |
| 4 | 587.64 | (10.05) | 587.17 | (9.58) | 611.92 | (11.79) | 588.35 | (20.34) |
| 5 | -3382.37 | (-3960.5) | 588.46 | (10.34) | 617.07 | (13.59) | 587.91 | (22.12) |
| 6 | 442.58 | (-157.42) | 609.71 | (9.70) | 643.67 | (15.81) | 612.04 | (21.75) |
| 7 | 496.48 | (-103.59) | 609.87 | (9.79) | 646.71 | (17.09) | 612.51 | (22.22) |

Table 3: Results from various model selection criteria with effective dimensions (in parentheses) for the simulated data set with 200 observations. After running 25000 MCMC interations, discard the first 15000 interation as burn-in. $K$ represents the number of normal components.

# 5 Discussion

In this paper, we examined the properties of the deviance information criteria devised for mixtures of distributions as well as the classical one from Spiegelhalter et al. (2002) and proposed a modified criterion $\text{DIC}_{loo}$. As we expected, the one from Spiegelhalter et al. (2002) didn't behave well in the context of the mixture model due to the label-switching issue. When we compared our criterion to others using the Galaxy dataset, ours selected a different mixture model than the other criteria chose: ours was more parsimonious with respect to the number of components in the mixtures. The different results may lead one to consider a model from $\text{DIC}_3$ or $\text{DIC}_{loo-cv}$ as a preferred mixture model since that model is able to describe shape of observations more closely, as Figure 2 shows. Among various DICs, however, our proposed criterion outperformed in terms of capturing the right values of the penalties (*i.e.,* effective dimensions), so we may consider the model selected from $\text{DIC}_{loo}$ as the preferred model. The same phenomenon occurred in the simulated dataset such that $\text{DIC}_{loo-cv}$ chose the model having an additional component

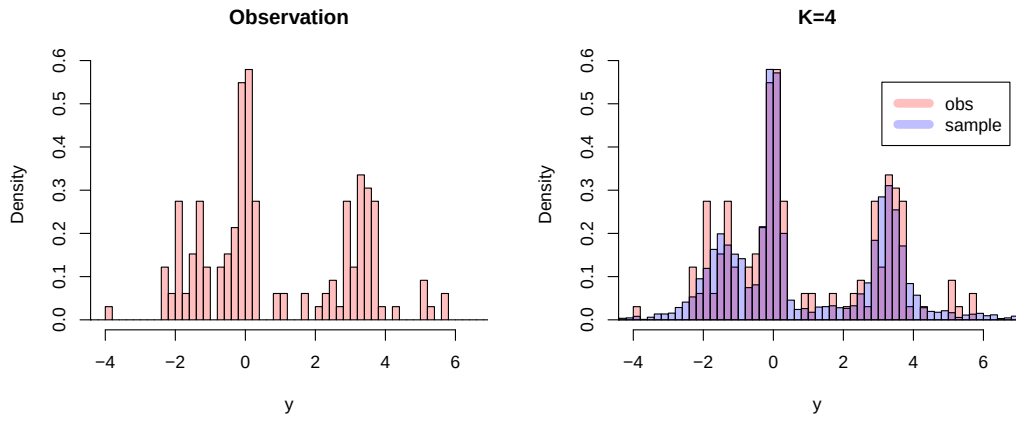


Figure 3: Density plots of 164 observations for simulated dataset (left) and 10000 samples from the $K = 4$ components mixture of normals (right).

($K = 5$) and $\text{DIC}_3$ had a little difference between $K = 4$ and $K = 5$, whereas our criterion selected $K = 4$ which is the right number of components by a large difference from other mixture cases.

Future work will examine asymptotic behavior of DICs, since we suspect that the criteria $\text{DIC}_3$, $\text{DIC}_{loo}$, and $\text{DIC}_{loo-cv}$ converge to the same criterion. This expectation is from the fact that the CPO converges to the posterior predictive distribution as the number of observations increases, provided that there is no huge influential point. Once we analytically examine the asymptotic behavior of DICs, we can also conduct numerical tests based on a large dataset such as the Shapley galaxy dataset (Drinkwater et al. 2004) which contains 4,251 observations with rich galaxy clusters, as well as some simulated datasets.

## References

Akaike, H. (1973). "Information theory and an extension of the maximum likelihood principle." In *Second international symposium on information theory*, 267–281. Akademinai Kiado.

Ando, T. (2007). "Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models." *Biometrika*, 94(2): 443–458.

Celeux, G., Forbes, F., Robert, C. P., Titterington, D. M., et al. (2006). "Deviance information criteria for missing data models." *Bayesian Analysis*, 1(4): 651–673.

Congdon, P. (2005). *Bayesian models for categorical data*. John Wiley & Sons.

Drinkwater, M. J., Gregg, M. D., Couch, W. J., Ferguson, H. C., Hilker, M., Jones, J. B., Karick, A., and Phillipps, S. (2004). "Ultra-compact dwarf galaxies in galaxy clusters." *Publications of the Astronomical Society of Australia*, 21(4): 375–378.

Geisser, S. and Eddy, W. F. (1979). "A predictive approach to model selection." *Journal of the American Statistical Association*, 74(365): 153–160.

Gelfand, A. E. (1996). "Model determination using sampling-based methods." In *Markov chain Monte Carlo in practice*, 145–161. Springer.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Ishwaran, H. and James, L. F. (2001). "Gibbs sampling methods for stick-breaking priors." *Journal of the American Statistical Association*, 96(453).

Marin, J. M., Mengersen, K., and Robert, C. P. (2005). "Bayesian modelling and inference on mixtures of distributions." *Handbook of statistics*, 25: 459–507.

Murata, N., Yoshizawa, S., and Amari, S. I. (1994). "Network information criterion-determining the number of hidden units for an artificial neural network model." *Neural Networks, IEEE Transactions on*, 5(6): 865–872.

Robert, C. P. and Titterington, D. M. (2002). "Discussion of a paper by D. J. Spiegelhalter et al." *J. R. Statist. Soc. B*, 64: 621–622.

Roeder, K. (1990). "Density estimation with confidence sets exemplified by superclusters and voids in the galaxies." *Journal of the American Statistical Association*, 85(411): 617–624.

Schwarz, G. (1978). "Estimating the dimension of a model." *The annals of statistics*, 6(2): 461–464.

Sethuraman, J. (1994). "A CONSTRUCTIVE DEFINITION OF DIRICHLET PRIORS." *Statistica Sinica*, 4: 639–650.

Shriner, D. and Yi, N. (2009). "Deviance information criterion (DIC) in Bayesian multiple QTL mapping." *Computational statistics & data analysis*, 53(5): 1850–1860.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). "Bayesian measures of model complexity and fit." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4): 583–639.

Takeuchi, K. (1976). "Distribution of information statistics and criteria for adequacy of models." *Math. Sci*, 153: 12–18.

Vehtari, A. and Gelman, A. (2014). "WAIC and cross-validation in Stan." *Manuscrip*.

Watanabe, S. (2010). "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory." *The Journal of Machine Learning Research*, 9999: 3571–3594.