Bayesian Analysis (2006) **1**, Number 4, pp. 677–680

# Comments on Article by Celeux et al.

Ming-Hui Chen*

I would like to congratulate the authors for developing a major extension of the deviance information criterion (DIC) introduced by Spiegelhalter et al. (2002) in the setting of missing data models. Recently, DIC is becoming increasingly popular for model assessment and model comparison. One of the main reasons for this is that DIC is well defined under improper priors as long as the resulting posteriors are proper and it is generally easy to compute.

Missing data models are routinely encountered in practice. There are several challenges posed by missing data. First, it is very difficult to reconstruct missing data. In most cases, the lost information due to missing data is not easy to recover. Second, it is more challenging to develop a measure of model complexity, which is a key issue in developing a model comparison criterion. Computation is another obstacle in dealing with such models. I am glad to see that the authors tackle this difficult problem and propose several natural extensions of DIC for these models.

Other Bayesian criterion based tools for model assessment and model comparison are available but not mentioned in the article. The Conditional Predictive Ordinate (CPO) statistic has been widely used in the statistical literature under various contexts. A detailed discussion of the CPO statistic and its applications to model assessment can be found in Geisser (1993), Gelfand and Dey (1994), and Gelfand et al. (1992). As shown in Gelfand and Dey (1994), asymptotically the CPO statistic has a similar dimensional penalty as AIC. In this perspective, the CPO statistic may be similar to DIC. The L measure criterion is another useful tool for model comparison. The L measure is constructed from the posterior predictive distribution of the data, and can be written as a sum of two components, one involving the means of the posterior predictive distribution and the other involving the variances. The L measure was introduced by Ibrahim and Laud (1994) for normal linear models and Gelfand and Ghosh (1998) for generalized linear models. The theoretical properties were examined in detail by Ibrahim et al. (2001). Chen et al. (2004) proposed the weighted L measure, which is a natural extension of the L measure. Both the CPO statistic and the L measures are well defined under improper priors. Thus, these criteria are similar to the DIC in this sense.

To examine performance of various DICs, I consider a small simulation study using a binary regression model with probit link. Suppose $y_i$ takes values 0 or 1 with probability

$$p_i = P(y_i = 1|\beta, x_i) = \Phi(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}), \tag{1}$$

where $\Phi$ is the standard normal cumulative distribution function, $x_i = (1, x_{i1}, \ldots, x_{i4})'$ is a $5 \times 1$ vector of covariates, which includes an intercept, and $\beta = (\beta_0, \beta_1, \ldots, \beta_4)$. The

*Department of Statistics, University of Connecticut, Storrs, CT, http://www.stat.uconn.edu/~mhchen

 

model (1) can be rewritten via the latent variable approach of Albert and Chib (1993). Specifically, we define

$$y_i = \begin{cases} 1 & \text{if } z_i \geq 0, \\ 0 & \text{if } z_i < 0, \end{cases} \tag{2}$$

and

$$z_i = x_i'\beta + \epsilon_i, \quad \epsilon_i \sim N(0,1). \tag{3}$$

Then, the joint distribution of $y_i$ and $z_i$ is given by

$$f(y_i, z_i|\beta, x_i) = \left[y_i 1_{\{z_i \geq 0\}} + (1-y_i)1_{\{z_i<0\}}\right] \times \frac{1}{\sqrt{2\pi}} \exp\{-\frac{(z_i - x_i\beta)^2}{2}\}. \tag{4}$$

Since the $z_i$'s are not observed, (2) and (3) can be viewed as a missing data model.

For each simulated dataset, $n$ independent binary responses ($y_i$'s) are generated with success probability $p_i$ given in (1) for $i = 1, 2, \ldots, n$, where $(x_{i1}, x_{i2}, x_{i3}, x_{i4})$ are independently and identically distributed random vectors from a multivariate normal distribution with means $(0, 0, 0, 0)$, variances $(16, 9, 0.3, 3)$, and a correlation matrix

$$\begin{pmatrix} 1 & 0.6 & 0 & 0 \\ 0.6 & 1 & 0.8 & 0 \\ 0 & 0.8 & 1 & 0.6 \\ 0 & 0 & 0.6 & 1 \end{pmatrix}.$$

I consider $\beta = (-1.0, 3.0, 0, -1.5, 0)'$, $\beta = (-1.0, 3.0, 2.0, -1.5, 0)'$ and $\beta = (-1.0, 3.0, 2.0, -1.5, 1)'$, which correspond to the true models $(x_1, x_3)$, $(x_1, x_2, x_3)$, and $(x_1, x_2, x_3, x_4)$, respectively. Sample size is taken to be $n = 200$ and for each $\beta$, I generate 200 independent datasets. For each simulated dataset, I fit $2^4 - 1 = 15$ models, so that each model includes an intercept. For each model, an improper uniform prior for $\beta$, i.e., $\pi(\beta) \propto 1$, is used. In this simulation study, I compare the following criteria: $\text{DIC}_1$, $\text{DIC}_4$, $\text{DIC}_5$, $\text{DIC}_7$, AIC, BIC, and L measure ($L(\nu)$). $\text{DIC}_1$ is constructed from the distribution $f(y_i|\beta, x_i)$ in (1) while $\text{DIC}_4$, $\text{DIC}_5$, and $\text{DIC}_7$ are constructed from the latent variable model defined by (2) and (3). Also, AIC and BIC are given by

$$\text{AIC} = -2\log L(\hat{\beta}|D) + 2p, \quad \text{BIC} = -2\log L(\hat{\beta}|D) + p\log(n),$$

where $p$ is the dimension of $\beta$, $L(\hat{\beta}|D)$ is the likelihood function evaluated at the maximum likelihood estimate $\hat{\beta}$, and $D = ((y_i, x_i), i = 1, 2, \ldots, n)$ denotes the observed data. For the model (1), the L measure with a quadratic loss introduced by Gelfand and Ghosh (1998) takes the form

$$L(\nu) = \sum_{i=1}^{n} \mu_i(1-\mu_i) + \nu \sum_{i=1}^{n} (\mu_i - y_i)^2,$$

where $\mu_i = E[\Phi(x_i'\beta)|D)]$ and $0 < \nu \leq 1$. The performance evaluation criterion is a 0-1 loss function, the loss being 0 if the true model is selected and 1 otherwise. Table 1 shows the results.

| True model | $DIC_1$ | $DIC_4$ | $DIC_5$ | $DIC_7$ | AIC | BIC | $L(0.5)$ |
|---|---|---|---|---|---|---|---|
| $(x_1, x_2, x_3, x_4)$ | 96 | 1 | 0 | 19 | 83 | 23 | 142 |
| $(x_1, x_2, x_3)$ | 78 | 2 | 0 | 30 | 74 | 32 | 60 |
| $(x_1, x_3)$ | 89 | 25 | 0 | 25 | 101 | 108 | 26 |

Table 1: Frequencies of Ranking the True Model as Best Based on 200 Datasets

From Table 1, we saw that no single measure is dominant in all three cases. The L measure performed better when the true model becomes more complex and BIC performed better when the true model is more parsimonious. In all three cases, $DIC_1$ and AIC performed reasonably well. However, $DIC_4$, $DIC_5$, and $DIC_7$ had much worse performance than $DIC_1$. This result is not surprising at all as little information is available in the observed data $D$ regarding latent variables $z_i$'s. In fact, the binary response $y_i$ is determined only by sign of $z_i$ while not by actual value of $z_i$. Thus, a variation of DIC constructed from the distribution of missing data would have a little power in selecting the true model.

Throughout the article, most attention has been paid to computational applicability or simplicity while less effort has been put on examining power of the measure being proposed. The above small simulation study suggests that the best DIC may be $DIC_1$, which is constructed from the distribution of observed data. My first question to the authors is: When or for which missing data models would the proposed variations of DIC constructed from the joint distribution of observed and missing data perform better or at least not much worse than $DIC_1$? My second question: Would DIC be more preferable over other criterion based measures such as CPO or L measure in the presence of missing data? My third question is: For certain missing data models, would it be better to construct DIC from either the marginal distribution $\int f(y|z,\theta)dz$ or the conditional distribution $f(y|z,\theta)$ while treating both $z$ and $\theta$ as parameters similar to the one proposed by Huang et al. (2005) in the setting of missing covariates data models?

## References

Albert, J. H. and Chib, S. (1993). "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association*, 88: 669–679. 678

Chen, M.-H., Dey, D. K., and Ibrahim, J. G. (2004). "Bayesian Criterion Based Model Assessment for Categorical Data." *Biometrika*, 91: 45–63. 677

Geisser, S. (1993). *Predictive Inference: An Introduction*. London: Chapman and Hall. 677

Gelfand, A. E. and Dey, D. K. (1994). "Bayesian Model Choice: Asymptotics and Exact Calculations." *Journal of the Royal Statistical Society, Series B*, 56: 501–514. 677

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). "Model Determinating Using Predictive Distributions with Implementation via Sampling-based Methods (with Discussion)." In Bernado, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (eds.), *Bayesian Statistics* **4**, 147–167. Oxford: Oxford University Press. 677

Gelfand, A. E. and Ghosh, S. K. (1998). "Model choice: A Minimum Posterior Predictive Loss Approach." *Biometrika*, 85: 1–13. 677, 678

Huang, L., Chen, M.-H., and Ibrahim, J. G. (2005). "Bayesian Analysis for Generalized Linear Models with Nonignorably Missing Covariates." *Biometrics*, 61: 767–780. 679

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). "Criterion Based Methods for Bayesian Model Assessment." *Statistica Sinica*, 11: 419–443. 677

Ibrahim, J. G. and Laud, P. W. (1994). "A Predictive Approach to the Analysis of Designed Experiments." *Journal of the American Statistical Association*, 89: 309–319. 677