

## Applying deviance information criteria (DIC) to models of pre-symptomatic transmission of SARS-CoV-2

A comprehensive guide to deviance information criteria (DIC) for missing data models exists in the paper by Celeux et al; here, we introduce only those definitions and concepts applicable to our models. The first part is devoted to DIC in general, while the second part shows how DIC is calculated for our models in particular.

### Part I: Theory

#### *Introduction*

Let  $f(\mathbf{y}|\theta)$  be the likelihood of observing some data  $\mathbf{y}$ , given a model  $f$  with parameter(s)  $\theta$ .

The *deviance* of model  $f$  with parameters  $\theta$  is defined as

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

where  $h(\mathbf{y})$  is a function of the data alone, i.e. it does not depend on  $\theta$ .

The deviance information criterion, or DIC, is defined by the following expression:

$$\text{DIC} = \overline{D(\theta)} + p_D$$

$\overline{D(\theta)}$  is the posterior mean deviance, a Bayesian measure of model fit, which is found by taking the expectation of  $D(\theta)$  over the posterior distribution of  $\theta$ :

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

$p_D$  is the effective number of parameters, a measure of model complexity, and is given by

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

where  $\tilde{\theta}$  is an estimate for  $\theta$  given  $\mathbf{y}$ , such as the posterior mean  $E_{\theta}[\theta|\mathbf{y}]$ .

We can therefore rewrite DIC as follows:

$$\begin{aligned} \text{DIC} &= 2\overline{D(\theta)} - D(\tilde{\theta}) \\ &= -4E_{\theta}[\log f(\mathbf{y}|\theta)|\mathbf{y}] + 2\log f(\mathbf{y}|\tilde{\theta}) + 2\log h(\mathbf{y}) \end{aligned}$$

Since  $h(\mathbf{y})$  does not depend on the model or parameters, it is irrelevant for model comparison; we can therefore set  $h(\mathbf{y}) = 1$  which makes  $2\log h(\mathbf{y}) = 0$ , simplifying the expression for DIC to

$$\text{DIC} = -4E_{\theta}[\log f(\mathbf{y}|\theta)|\mathbf{y}] + 2\log f(\mathbf{y}|\tilde{\theta})$$

#### *Estimating DIC from Markov chain Monte Carlo (MCMC)*

DIC can be readily applied to model fitting by MCMC because the posterior distribution of  $\theta$  is approximated by the post-convergence Markov chain. As long as the likelihood function  $f(\mathbf{y}|\theta)$  is available in closed form, the posterior mean deviance,  $\overline{D(\theta)}$ , can be estimated by averaging  $D(\theta)$  over all of the steps in the chain, and  $D(\hat{\theta})$  simply requires an estimator of  $\theta$ , such as the posterior mean.

### *Extension of DIC to missing data models*

With data-augmented models, the likelihood function  $f(\mathbf{y}|\theta)$  is often not available in closed form, because it depends on missing data  $\mathbf{z}$  as well as the observed data  $\mathbf{y}$ . In this case,  $f(\mathbf{y}|\theta)$  is called as the *observed* likelihood, while  $f(\mathbf{y}, \mathbf{z}|\theta)$  is termed the *complete* likelihood.

The DIC for a missing data model can therefore be re-written in terms of the complete likelihood, as follows:

$$\text{DIC} = -4E_{\theta}[\log f(\mathbf{y}, \mathbf{z}|\theta)|\mathbf{y}, \mathbf{z}] + 2\log f(\mathbf{y}, \mathbf{z}|E_{\theta}[\theta|\mathbf{y}, \mathbf{z}])$$

Since the data  $\mathbf{z}$  are, by definition, missing, this quantity can not be computed directly; however, if the distribution of  $\mathbf{Z}$  is known or can be approximated (e.g. using a data augmentation MCMC algorithm), it is sufficient to take the expectation of DIC with respect to  $\mathbf{Z}$ :

$$\begin{aligned} \text{DIC} &= E_{\mathbf{Z}}[\text{DIC}(\mathbf{y}, \mathbf{Z})|\mathbf{y}] \\ &= -4E_{\theta, \mathbf{Z}}[\log f(\mathbf{y}, \mathbf{Z}|\theta)|\mathbf{y}] + 2E_{\mathbf{Z}}[\log f(\mathbf{y}, \mathbf{Z}|E_{\theta}[\theta|\mathbf{y}, \mathbf{Z}])|\mathbf{y}] \end{aligned}$$

The first term in this expression can be estimated using the posterior distributions of  $\theta$  and  $\mathbf{Z}$  from a data augmentation MCMC algorithm, but the second term requires calculation of the posterior mean  $E_{\theta}[\theta|\mathbf{y}, \mathbf{Z}]$  for each value of  $\mathbf{Z}$ , which is inconvenient. However, this term can be reformulated to make estimation more straightforward.

Recall that the formula for DIC is as follows:

$$\text{DIC} = 2\overline{D(\theta)} - D(\hat{\theta})$$

And recall that, if we set  $h(\mathbf{y}) = 1$ , the formula for deviance simplifies to

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

and therefore the second term in the DIC expression can be rewritten as

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

Expectation-maximization (EM) algorithms suggest a way to approximate the log-likelihood in the context of missing data:

$$\log f(\mathbf{y}|\theta) = E_{\mathbf{Z}}[\log f(\mathbf{y}, \mathbf{Z}|\hat{\theta}(\mathbf{y}))|\mathbf{y}, \hat{\theta}(\mathbf{y})]$$

where  $\hat{\theta}(\mathbf{y})$  is an estimator of  $\theta$  based on the observed data  $\mathbf{y}$ .

Thus, an alternative way of writing DIC is as follows:

$$\text{DIC} = -4E_{\theta, \mathbf{Z}}[\log f(\mathbf{y}, \mathbf{Z}|\theta)|\mathbf{y}] + 2E_{\mathbf{Z}}[\log f(\mathbf{y}, \mathbf{Z}|\hat{\theta}(\mathbf{y}))|\mathbf{y}, \hat{\theta}(\mathbf{y})]$$

The second term in this expression can be estimated by running a second MCMC with the parameters  $\theta$  fixed at  $\hat{\theta}(\mathbf{y})$  and taking the expectation of the log-likelihood with respect to the posterior distribution of the missing data  $\mathbf{Z}$ .

### Part II: Application

We now show how DIC is calculated for the models of pre-symptomatic transmission described in the Materials & Methods. Recall that the generation interval  $\tau$  is assumed to follow a gamma distribution with shape parameter  $\alpha$  and rate parameter  $\beta$  (incubation-independent model) or  $\beta/\theta_1$  (incubation-dependent model). The incubation periods of the infector and infectee are denoted  $\theta_1$  and  $\theta_2$ , respectively, while  $\delta$  refers to the serial interval.

We used an MCMC algorithm with data augmentation to estimate the parameters  $\alpha$  and  $\beta$ , as well as the missing data  $\theta_1$  and  $\theta_2$ , by fitting to the serial interval data  $\delta$ . Thus, in the notation of Part 1, the serial interval data  $\delta$  are the observed data  $\mathbf{y}$ , the incubation periods  $\theta_1$  and  $\theta_2$  comprise the missing data  $\mathbf{z}$ , and the parameters  $\alpha$  and  $\beta$  are the model parameters  $\theta$ .

At the end of Part 1, we arrived at the following expression for DIC for a missing data model:

$$\text{DIC} = -4E_{\theta, \mathbf{Z}}[\log f(\mathbf{y}, \mathbf{Z}|\theta)|\mathbf{y}] + 2E_{\mathbf{Z}}[\log f(\mathbf{y}, \mathbf{Z}|\hat{\theta}(\mathbf{y}))|\mathbf{y}, \hat{\theta}(\mathbf{y})]$$

Rewriting this for our models, we get:

$$\text{DIC} = -4E_{\alpha, \beta, \theta_1, \theta_2}[\log f(\delta, \theta_1, \theta_2|\alpha, \beta)|\delta] + 2E_{\theta_1, \theta_2}[\log f(\delta, \theta_1, \theta_2|\hat{\alpha}, \hat{\beta})|\delta, \hat{\alpha}, \hat{\beta}]$$

From this point on, since our focus is application rather than theory, we drop several redundant conditionals from the equations to improve readability. This simplifies this above to

$$\text{DIC} = -4E_{\alpha, \beta, \theta_1, \theta_2}[\log f(\delta, \theta_1, \theta_2|\alpha, \beta)] + 2E_{\theta_1, \theta_2}[\log f(\delta, \theta_1, \theta_2|\hat{\alpha}, \hat{\beta})]$$

The first component of this expression can be rewritten as follows:

$$\begin{aligned} & -4E_{\alpha, \beta, \theta_1, \theta_2}[\log f(\delta, \theta_1, \theta_2|\alpha, \beta)] \\ &= -\frac{4}{M} \sum_{i=1}^M \log f(\delta, \theta_1^{(i)}, \theta_2^{(i)}|\alpha^{(i)}, \beta^{(i)}) \\ &= -\frac{4}{M} \sum_{i=1}^M \sum_{k=1}^N \log f(\delta_k, \theta_{1(k)}^{(i)}, \theta_{2(k)}^{(i)}|\alpha^{(i)}, \beta^{(i)}) \\ &= -\frac{4}{M} \sum_{i=1}^M \sum_{k=1}^N \log f(\delta_k|\theta_{1(k)}^{(i)}, \theta_{2(k)}^{(i)}, \alpha^{(i)}, \beta^{(i)}) + \log f_{\theta}(\theta_{1(k)}^{(i)}) + \log f_{\theta}(\theta_{2(k)}^{(i)}) \end{aligned}$$

$$= -\frac{4}{M} \sum_{i=1}^M \sum_{k=1}^N \log f_{\tau}(\delta_k + \theta_{1(k)}^{(i)} - \theta_{2(k)}^{(i)} | \alpha^{(i)}, \beta^{(i)}) + \log f_{\theta}(\theta_{1(k)}^{(i)}) + \log f_{\theta}(\theta_{2(k)}^{(i)})$$

where  $M$  is the number of iterations in the thinned converged Markov chain, with a superscript  $(i)$  denoting the  $i$ th iteration;  $N$  is the number of infector-infectee pairs represented in the serial interval data set, with a subscript  $(k)$  denoting the  $k$ th pair;  $f_{\tau}$  is the generation interval distribution (with parameters  $\alpha$  and  $\beta$ ); and  $f_{\theta}$  is the prior for the incubation period.

The second component of the expression can similarly be rewritten:

$$\begin{aligned} & 2E_{\theta_1, \theta_2}[\log f(\boldsymbol{\delta}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \hat{\alpha}, \hat{\beta})] \\ &= \frac{2}{M'} \sum_{i=1}^{M'} \log f(\boldsymbol{\delta}, \boldsymbol{\theta}_1'^{(i)}, \boldsymbol{\theta}_2'^{(i)} | \hat{\alpha}, \hat{\beta}) \\ &= \frac{2}{M'} \sum_{i=1}^{M'} \sum_{k=1}^N \log f(\delta_k, \theta_{1(k)}'^{(i)}, \theta_{2(k)}'^{(i)} | \hat{\alpha}, \hat{\beta}) \\ &= \frac{2}{M'} \sum_{i=1}^{M'} \sum_{k=1}^N \log f(\delta_k | \theta_{1(k)}'^{(i)}, \theta_{2(k)}'^{(i)}, \hat{\alpha}, \hat{\beta}) + \log f_{\theta}(\theta_{1(k)}'^{(i)}) + \log f_{\theta}(\theta_{2(k)}'^{(i)}) \\ &= \frac{2}{M'} \sum_{i=1}^{M'} \sum_{k=1}^N \log f_{\tau}(\delta_k + \theta_{1(k)}'^{(i)} - \theta_{2(k)}'^{(i)} | \hat{\alpha}, \hat{\beta}) + \log f_{\theta}(\theta_{1(k)}'^{(i)}) + \log f_{\theta}(\theta_{2(k)}'^{(i)}) \end{aligned}$$

where  $\hat{\alpha}$  and  $\hat{\beta}$  are the parameter estimates (in this case, posterior means) from the original MCMC, and apostrophes indicate quantities from a secondary MCMC in which  $\alpha$  and  $\beta$  are fixed at  $\hat{\alpha}$  and  $\hat{\beta}$ , respectively.