Bayesian Analysis (2006) **1**, Number 4, pp. 687–698

Comment on article by Celeux et al.

Xiao-Li Meng*, and Florin Vaida†

Abstract. This discussion argues that any difficulty with DIC for missing data is due to DIC being intrinsically a large-sample measure and relying on point estimates. What is missing is not "missing data", but rather a set of coherent principles for DIC itself when the amount of data is not adequate to invoke quadratic approximation for a complex model. The non-uniqueness of data augmentation schemes for any observed-data model also argues for the importance of emphasizing inference "focus" in applying model complexity measures such as DIC. An attempt to bring in more Bayesian "flavor" into DIC also reveals that an insightful explanation is missing: neither pure Bayesian measure nor pure likelihood/sampling measure yield sensible results, but some hybrid ones do.

Keywords: Effective number of parameters, Information criteria, model selection, missing data, statistical principles

1 No, It's not the missing data...

A casual reader of this stimulating paper by Celeux, Forbes, Robert and Titterington (CFRT) might walk away with the impression that a principle is greatly needed to formulate DIC with *missing data* models. Surely eight variations of a single measure in one paper is an indication that a ripe fruit is still out of reach. A more astute reader, however, would sense that if there is a devil, it is not in the missing data, but rather in DIC itself. Indeed, in the world of data augmentation, as illustrated by CFRT with the introduction of the membership variable for the mixture model, the notion of *missing data models* covers every model under the sun — any probabilistic model we put down can be recast as a marginal model for infinitely many models on larger spaces.

So what is the point of making this almost tautological emphasis? The point is to help readers to see more easily the following key points of our discussion, all of which are inspired by CFRT.

(A) There is no need of a separate definition of DIC for missing data models, *if there are no qualms* with the original DIC definition(s) of Spiegehalter, Best, Carlin and van der Linde (SBCV; Spiegelhalter et al. (2002)).

(B) The missing-data formulation highlights the fundamental defect in DIC: its reliance on a point estimate to assess over-fitness of an entire model fitting process.

(C) When using a DIC-like quantity with missing data, a key question to ask, is *whether*

[*] Department of Statistics, Harvard University, Cambridge, MA, mailto:meng@stat.harvard.edu

[†] Division of Biostatistics, Dept. of Family and Preventive Medicine, UCSD School of Medicine, San Diego, CA, mailto:vaida@ucsd.edu

> *there is any change* of model or inferential focus, when we embed a marginal model into a joint one.

2 A Pandora's box in DIC?

To understand why there can be so many variations of DIC, in CFRT and elsewhere, it is useful to revisit SBCV's original derivation of DIC. Briefly, SBCV started with the "residual information" $-2\log[p(\mathbf{y}|\theta)]$ and an estimate $\tilde{\theta}(\mathbf{y})$ of the "pseudotrue" parameter θ^t; the use of a "pseudotrue" parameter is due to the fact that the "true" model, if it can ever be formulated, may lie outside the posited parametric family $\{p(\mathbf{y}|\theta) : \theta \in \Theta\}$. Under this setting, SBCV wrote

> "Then the excess of the true over the estimated residual information will be denoted
>
> $$d_\Theta\{\mathbf{y}, \theta^t, \tilde{\theta}(\mathbf{y})\} = -2\log[p(\mathbf{y}|\theta^t)] + 2\log[p(\mathbf{y}|\tilde{\theta}(\mathbf{y}))]. \qquad (3)$$
>
> This can be thought of as the reduction in surprise or uncertainty due to estimation, or alternatively the degree of 'overfitting' due to $\tilde{\theta}(\mathbf{y})$ adapting to the data $\mathbf{y}$. We now argue that d_Θ may form the basis for both classical and Bayesian measures of model dimensionality, with each approach differing in how it deals with the unknown true parameters in d_Θ."

There lies the source of the problem: d_Θ measures the "overfitting" due to a *point estimate* $\tilde{\theta}(\mathbf{y})$. When the model being entertained is simple enough (e.g., with enough built-in normality) or when the "data size" is large enough to validate normal asymptotic, a satisfactory choice, for the purposes of assessing the "over-fitness" of our *entire Bayesian model fitting process*, of the point estimator $\tilde{\theta}(\mathbf{y})$ often exists. This is largely, as far as we can see it, the theoretical underpinning of DIC as formulated by SBCV (e.g., their sections 3 and 4). In general, once we face more complex models, such as those investigated by CFRT, DIC appears to have a built-in Pandora's box: the choice of $\tilde{\theta}(\mathbf{y})$. The problem here is not much about how to choose $\tilde{\theta}$, but rather whether *a single point estimate can ever adequately represent an entire model fitting process.*

Indeed, SBCV acknowledged this problem. Their definition of the *effective dimension* p_D is the posterior mean of d_Θ, which can also be written as

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

where

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

$h(\mathbf{y})$ is a standardizing term, and the notation $\overline{A}$ denotes the posterior mean of A with respect to $p(\theta|\mathbf{y})$. As for the choice of $\tilde{\theta}$ in (1), SBCV wrote

> "In our examples we shall generally take $\tilde{\theta}(\mathbf{y}) = E(\theta|\mathbf{y}) = \bar{\theta}$, the posterior mean of the parameters. However, we note that it is not strictly necessary to use the posterior mean as an estimator of either d_Θ or θ, and the mode or median could be justified."

The trouble is that the issue appears to be much more complicated than merely "strictly necessary". Minimally, we should worry about the impossibility in assessing the overfitness of an entire Bayesian modelling process via a single point estimator, however it is chosen. Indeed, a key feature of Bayesian inference, in contrast to classic modes of inference (however defined or perceived), is its ability of focusing on estimands by directly accessing their entire posterior distributions, not on any particular point or even interval estimators. We are therefore in full agreement with the opening message Dawid (2002) conveyed in his discussion of SBCV:

> "This paper should have been titled 'Measures of Bayesian model complexity and fit', for it is the models, not the measures, that are Bayesian. Once the ingredients of a problem have been specified, any relevant question has a unique Bayesian answer. Bayesian methodology should focus on specification issues or on ways of calculating or approximating the answer. Nothing else is required."

3 History waiting to be repeated?

Indeed, one key reason that we have so many variations of DIC, and therefore violating the "uniqueness" of Bayesian answer as Dawid emphasized, is because DIC is fundamentally a "classic" measure for comparing models, despite its "Bayesian looks" from the use of posterior mean or mode or median. A good analogue for this is the *posterior predictive p-value* (ppp) (Meng 1994), which is fundamentally a frequentist measure, despite its B-looks. Indeed, the derivation of ppp followed almost the identical route as SBCV's DIC: start with a classic measure, which depends on some unknown parameter (in the ppp case, any parameter that is not specified by the null hypothesis), followed by posterior averaging. We mention this analogy to emphasize that we have nothing against blending classic and Bayesian methods. Indeed, like many others, we have advocated such blending especially in the context of model diagnostics and assessment (Gelman et al. (1996); Gelman and Meng (1995)). Nor do we believe that there is an all-encompassing *and practical* Bayesian solution to the problems that SBCV and CFRT (and many others) want to address. Rather, the purpose of our revisit to SBCV is to make it crystal clear that the difficulty here is not the missing data. All the problems CFRT revealed are problems with DIC for complex models even for "complete data", because of facing the same type of choices as any method that focuses on estimators, not on estimand. The latter is unique as soon as we specify our *focus of our inference or model*, a point we shall discuss in the next section.

Furthermore, this revisit suggests that we can call upon our knowledge and experience with non-Bayesian point estimators in making DIC to yield sensible results, as well as in our quest for a more unified *principle*. The CFRT's investigation reminded us that we are somewhat still in a "pre-EM" era regarding DIC or other similar measures. By "pre-EM", we refer to the era before Dempster et al. (1977) (DLR) EM formulation, when there was a great need to deal with each missing-data estimation problem separately by formulating estimators specifically for each case in order to accommodate the varying incomplete-data structures. The unification brought by DLR's EM algorithm prevented a tremendous amount of duplicated efforts. It also opened the door for dealing with many more, both in terms of number and complexity, missing-data problems. That is, many specific algorithms scattered in the literature turned out to be special cases or disguised versions of the EM algorithm, or rather "the EM principle" because it provides a general recipe for constructing algorithms. What is needed for DIC is a similar general recipe.

One key difference between EM and DIC formulations is that the former is just a computational unification, for the estimation principle underlies it, namely the maximum likelihood estimation, is the same regardless of whether one faces missing data or not. For DIC, the unification needed is more fundamental, as the question is how to assess the fitness and complexity of a model when relying on point estimators is simply inadequate. Perhaps the best analogy here is the replacement of maximum likelihood estimation by Bayesian estimation – both work for large-sample with essentially equivalent results, but the latter has a better chance of yielding scientifically more useful answers for smaller samples.

4 Stay focused or else...

Whereas we consider DIC's reliance on the choice of $\tilde{\theta}(\mathbf{y})$ an intrinsic problem, we view "(T)he diversity of the numerical answers associated with different focusses" a desirable feature of DIC, rather than a "real difficulty", as CFRT stated. As SBCV correctly emphasizes, any useful model fitness or complexity measure should be sensitive to what we use the model for, namely, to the "focus" of our models or more generally inference procedures. We surmise that CFRT were concerned with the same dilemma as those of us who provide statistical consultations to investigators with minimum training or interest in statistical methods. They demand the methods we advise only require a few simple lines in SAS or alike, with minimum "tuning parameters", but are most efficient/powerful for their specific problems.

There is nothing wrong with such a "consumer attitude" – we all carry it. And it is indeed very useful for those of us who work in methodological or computational statistics to always keep this attitude/demand in mind when conducting research. Nevertheless, we cannot push simplicity at the expense of functionality or validity. The examples in SBCV and CFRT clearly demonstrate that there cannot be any meaningful "focus-free" fitness or complexity measure, just as there cannot be a "model-free" EM implementation (as much as we have been repeatedly asked to develop one over the years!).

Although this is rather a trivial point, the data augmentation formulation should make this dependence on the purpose of inference even clearer. Since any "observed-data" model can be embedded into infinitely many "larger" models by introducing different kinds of latent variables (or other unobservable structures), it is evident that if our interests include learning about these unknowns, then our fitness or complexity measure is useless if it is invariant to the embedding. Surely useful answers cannot be robust to questions being asked. On the other hand, if the questions of interest concern only the observed data, then the measure should be invariant to which of the many possible data augmentation schemes is used, i.e. it should not depend on the missing data/latent variables which we don't care about. In this sense, $\text{DIC}_{4,5,6}$ of CFRT are not designed for dealing with this latter situation.

In a nutshell, our key emphasis here is the point (B) in Section 1 – when faced with missing data, ask if the focus has changed when moving from the observed data model to the augmented data model. If there is no such change of focus, and if we had no trouble to adopt SBCV's DIC for the "complete data", then logically we have no choice but to stick to the same definition, for the generic notation $-2\log[p(\mathbf{y}|\theta)]$ in SBCV, or $-2\log[f(\mathbf{y}|\theta)]$ in CFRT's notation, makes no reference to whether $\mathbf{y}$ is complete or not. The only requirement here is that the $f(\cdot|\theta)$ function captures all data selection mechanisms that are responsible for generating the $\mathbf{y}$ we *actually observe*. That is, $\mathbf{y}$ can even be a result of a *non-ignorable* missing-data mechanism operated on $\{\mathbf{y}, \mathbf{z}\}$, as long as the $f(\cdot|\theta)$ has included in its formulation this mechanism.

Evidently, this "logical coherence" requirement would immediately exclude CFRT's $D_{4,5,6,7,8}$ as possible "DIC" measures because each of them uses a different f function than the one that actually generates our observed $\mathbf{y}$. This leaves $D_{1,2,3}$ as the possible choices, with the differences among them precisely the differences of choosing $\tilde{\theta}$, at least for D_1 and D_2, namely, posterior mean versus posterior mode. As CFRT, we find D_3 quite sensible, as it is an attempt to move away from seeking a point estimate $\tilde{\theta}$ and then plug it in $f(\mathbf{y}|\theta)$. Rather, it replaces $f(\mathbf{y}|\theta)$ by an estimate of the entire f density function.

5 But it does get fuzzy once out of focus...

The issues do become more complicated once the focus changes, i.e. the questions of interest concern not only the original θ, but also part of the missing data $\mathbf{z}$. The complication comes from the fact that in SBCV's original definition there are only two ingredients: the observed data $\mathbf{y}$ and parameter θ. When the focus is on θ, the augmented data $\mathbf{z}$, however useful, say, for computation or even for the modelling exercise itself, gets integrated out in our final model $f(\mathbf{y}|\theta)$.

But once $\mathbf{z}$ itself is of interest ("in focus"), in what way can SBCV's DIC accommodate $\mathbf{z}$? By virtue of $\mathbf{z}$ being in focus, it becomes part of the estimand *by definition* – we prefer to use *estimand* than *parameter* to avoid the unnecessary but thorny distinction between "parameter" and "latent variables" in Bayesian formulation. In other words, the estimand changes from θ to $\theta_N = (\theta, \mathbf{z})$. What shall we do then?

In principle all we need to do is to just replace θ by θ_N, which results in

$$p_D = -2E[\log[f(\mathbf{y}|\theta, \mathbf{z})|\mathbf{y}] + 2\log[f(\mathbf{y}|\hat{\theta}, \hat{\mathbf{z}})]. \tag{3}$$

The trouble is, as before, what should be used as $\hat{\mathbf{z}}$? Even if the size of $\mathbf{y}$ is large enough to render an approximate validity of using a single $\hat{\theta}$ for DIC purposes, the size of $\mathbf{y}$ cannot possibly be large enough to rend the same for $\mathbf{z}$ in general, for $\mathbf{z}$ can be of arbitrary dimension and size as we please, because it is an artificial model construction, at least in theory. And it is known in general that maximizing over missing data or latent variable is a rather dangerous practice, even though occasionally it might provide an acceptable answer (Little and Rubin (1983)).

So the "solution", if it can ever be precisely defined, must depend on the specific construction of $\mathbf{z}$. A simple example may illustrate this point. Consider the following hierarchical model, similar to (2) in SBCV,

$$y_{ij} = x_{ij}\beta + b_i + \epsilon_{ij}, \quad b_i \sim N(0, \tau^2), \quad \epsilon_{ij} \sim N(0, \sigma^2), \tag{4}$$

where $j = 1, \ldots, J_i, i = 1, \ldots, I$. In a "marginal" interpretation the focus is on the population parameters, $\theta = (\beta, \tau^2, \sigma^2)$, and $b = \{b_i; i = 1 \ldots I\}$ is treated as missing data. In contrast, in a "conditional" interpretation, b is part of the estimand, that is, $\theta_N = (\theta, b)$. In the latter case DIC has a frequentist equivalent in the conditional AIC of Vaida and Blanchard (2005).

Since DIC has asymptotic justification, it will work for θ_N as long as there is a corresponding asymptotic scenario under which the dimensionality in θ_N does not grow indefinitely with the sample size; in the current case this means that the correct asymptotics need to assume growing J_i, but fixed I. Otherwise we run the risk of overfitting, as in the well-known Neyman-Scott problem, and the DIC breaks down. The situation needs to be assessed on a case-by-case basis.

In the mixture model of CFRT, in a search for an "overall" criterion for the model we are not interested in the number of components of the mixture, nor the parameters of each component, but rather in the predictive distribution. This specific focus is addressed by CFRT's DIC$_3$. If, on the other hand, the parameters of the mixture are relevant (e.g., they have a substantive interpretation which is under study), then this "missing data" is "in focus" and is part of the parameter θ_N. In this situation, CFRT's DIC$_7$ is a more appropriate choice.

6 One curiosity leads to another...

An astute reader may have already noted that in Section 4 we have argued that if we accept SBCV's definition of d_Θ of their (3), as quoted in Section 2, then its first term $-2\log[p(\mathbf{y}|\theta)]$ is not subject to any modification. Only the second term $2\log[p(\mathbf{y}|\tilde{\theta})]$ is, because of the unsettling nature of $\tilde{\theta}$. However, since our goal is to assess *Bayesian* model fitness or complexity, we at least wonder if the use of log-posterior instead of log-likelihood could lead to more sensible measures. After all, prior is an integrated part of a Bayesian model, and theoretically we cannot exclude point-mass prior, which clearly will reduce the number of "free parameters" in the model. That is, in a Bayesian model, the number of parameters, however measured, cannot be invariant to the prior. An added benefit is that

$$p_D^B \equiv -2E[\log\{p(\theta|\mathbf{y})\}|\mathbf{y}] + 2\log[p(\hat{\theta}|\mathbf{y})], \tag{5}$$

where $\hat{\theta}$ is the posterior mode, will be non-negative regardless of the choice of prior, as long as $\hat{\theta}$ is the (global) posterior mode. In this sense, p_D^B, where the superscript B signifies its Bayesian nature (at least when compared to SBCV's p_D), is the more general version of CFRT's D_2, which was guaranteed to be non-negative only when the prior is constant (in which case of course our p_D^B is the same as SBCV's p_D with $\tilde{\theta}$ being the MLE).

Since the normalizing constant for the posterior density gets cancelled in the right-hand side of (5), p_D^B can also be motivated by Moody's (1992) "effective number of parameters," which was constructed via "penalized likelihood". For a Bayesian, a penalized likelihood is essentially a disguised version of posterior, and, furthermore, it would be natural to replace Moody's (1992) sampling expectation by the posterior expectation. The resulting modified measure then would be exactly the p_D^B in (5), barring that Moody (1992) invoked a quadratic approximation, which is unnecessary in our Bayesian formulation (see next section for more discussion).

Given the discussion of Moody's (1992) method in SBCV in their Section 2.4, and given the more or less obvious (and seemingly advantageous) modification as discussed above, we were a bit curious why this clearly more Bayesian route was not followed up by SBCV, nor by any of their discussants (excluding those who questioned the entire formulation of DIC, such as Dawid (2002)). Granted, SBCV's p_D does depend on the prior, but it is only through the posterior averaging for dealing with the unknown θ^t. So surely the direct use of log-posterior in measuring "residual information" should provide a more sensible complexity measure for Bayesian model fitting?

Indeed we were so sure that p_D^B would yield more sensible results than p_D, until we actually checked with the following simple example. Consider a simple random sample $\mathbf{y} = \{X_1, \ldots, X_n\} \sim N(\theta, 1)$, with a normal prior $\theta \sim N(0, \tau^2)$, where τ^2 is known. Then it is well-known that the posterior distribution for θ is $N(\hat{\theta}_n, \sigma_n^2)$, where

$$\hat{\theta}_n = \frac{n\bar{X}}{n + \tau^{-2}} \quad \text{and} \quad \sigma_n^2 = \frac{1}{n + \tau^{-2}}. \tag{6}$$

Then, using the notation in (2), we see that the difference in using the log-likelihood and log-posterior amounts to choosing two different D functions:

$$D^{(1)}(\theta) = -2\log f(\mathbf{y}|\theta) + 2\log f(\mathbf{y}|\bar{X}) = n(\theta - \bar{X})^2, \tag{7}$$

and, *as long as* $\tau^2 > 0$,

$$D^{(2)}(\theta) = -2\log p(\theta|\mathbf{y}) + 2\log p(\hat{\theta}_n|\mathbf{y}) = \frac{(\theta - \hat{\theta}_n)^2}{\sigma_n^2}. \tag{8}$$

It follows then,

$$p_D = E\left[D^{(1)}(\theta) - D^{(1)}(\hat{\theta}_n)\Big|\ \mathbf{y}\right] = nV(\theta|\mathbf{y}) = n\sigma_n^2 = \frac{n\tau^2}{n\tau^2+1}. \tag{9}$$

In comparison, when $\tau^2 > 0$,

$$p_D^B = E\left[D^{(2)}(\theta) - D^{(2)}(\hat{\theta}_n)\Big|\ \mathbf{y}\right] = \frac{V(\theta|\mathbf{y})}{\sigma_n^2} = 1, \tag{10}$$

and $p_D^B = 0$ when $\tau^2 = 0$ because then $D^{(2)}$ is only defined on $\Theta = \{\theta = 0\}$ with value zero.

Neither result is wrong, technically. Indeed, one could even argue that p_D^B gives a better answer. Mathematically speaking, as long as $\tau^2 > 0$, there is one unknown parameter, θ, to be estimated. When $\tau^2 = 0$, both measures give zero, which is also correct for there is no unknown parameter in the model. However, from a Bayesian model complexity point of view, the result given by p_D of (9) is more appealing, because it monotonically and *continuously* decreases from 1 to 0 as τ^2 decreases from ∞ to 0. As τ approaches zero, the model complexity, in terms of its potential changes as a functional of the unknowns, should reduce, because we have stronger and stronger prior information. Consequently, the "effective number of parameters" for our Bayesian model should decrease as well. In that sense, the result given by p_D^B is puzzling, for it delivers either 1 or 0. We surely expected it to be more sensitive to the value of τ^2, given it is "purely Bayesian". All three ingredients in forming the p_D^B are Bayesian: the choice of "$\log p$" is log-posterior, as in (8), the choice of the point estimator is posterior mode, as in (10), and the expectation is with respect to the posterior distribution of θ, also as in (10). So why does p_D^B give a less sensible result than p_D, which only uses two out of the three Bayesian ingredients, as (7) is based on likelihood only? Is this a case of "too much of a good thing"?

Puzzled, we decided to examine all eight possibilities (not to be confused with the CFRT's eight variations!). That is, all three ingredients can take the Bayesian version, as detailed above, or non-Bayesian version: (1) sampling expectation for the "over bar" operation in $\bar{D}$ instead of posterior expectation, (2) MLE for $\tilde{\theta}$ instead of posterior mode, and (3) log-likelihood for "log p" instead of log-posterior, resulting a 2 × 2 × 2 design as displayed in Table 1 and Table 2. In these two tables,

$$p_{ijk} = E_i[d_\Theta^{(k)}(\mathbf{y}, \theta, \tilde{\theta}_j(\mathbf{y}))], \quad i, j, k = 1, 2. \tag{11}$$

In particular, $p_{222} = p_D^B$ and $p_{221} = p_D$. Here, using the notation of SBVC's (3), as quoted in Section 2,

$$d_\Theta^{(k)}(\mathbf{y}, \theta, \tilde{\theta}(\mathbf{y})) = D^{(k)}(\theta) - D^{(k)}(\tilde{\theta}), \quad k = 1, 2,$$

where $D^{(k)}$'s are as given by (7) and (8), $\tilde{\theta}_1$ is the MLE, $\tilde{\theta}_2$ is the posterior mode, E_1 is with respect to the sampling distribution of $\mathbf{y}$, as in Moody (1992), and E_2 is with respect to the posterior distribution of θ, as in SBVC. Note that because the

Table 1: "Effective number" under Sampling Expectation.

Choice of p in "$\log p$" =	Likelihood	Posterior
Point Est. = MLE	$p_{111} = 1$	$p_{112} = 1 - \frac{1}{n\tau^2}$
Point Est. = Post. Mode	$p_{121} = 1 - \frac{1+n\theta^2}{(1+n\tau^2)^2}$	$p_{122} = \frac{n\tau^2}{1+n\tau^2} + \frac{\theta^2}{\tau^2(1+n\tau^2)}$

Table 2: "Effective number" under Posterior Expectation.

Choice of p in "$\log p$" =	Likelihood	Posterior
Point Est. = MLE	$p_{211} = \frac{n\tau^2}{1+n\tau^2} + \frac{n\bar{X}^2}{(1+n\tau^2)^2}$	$p_{212} = 1 - \frac{\bar{X}^2}{\tau^2(1+n\tau^2)}$
Point Est. = Post. Mode	$p_{221} = \frac{n\tau^2}{1+n\tau^2}$	$p_{222} = 1$

two expectations are on entirely different spaces, the $d_\Theta(\mathbf{y}, \theta, \tilde{\theta}(\mathbf{y}))$ notation is more appropriate as it shows the dependence on both $\mathbf{y}$ and θ, whereas the $D(\theta)$ notation is adequate only for the posterior expectation.

All the results in the two tables are exact, and are based on the assumption that $\tau^2 > 0$. Under this assumption, we have the following observations.

1 As $n \to \infty$, all $p_{ijk} \to 1$, reinforcing the well-known fact that when the sample size is getting large, the impact of the prior, as long as it is not mathematically "dominating" (e.g., a singleton mass), becomes more and more negligible. Same is true when $\tau \to \infty$, as then the prior becomes constant and hence the likelihood and Bayesian methods coincide. The two cases, that is, either large n or large τ, essentially refer to the same condition: the prior information is negligible compared to that from the likelihood.

2 Among all p_{ijk}'s, besides the trivial cases of $p_{111} = 1$ and $p_{222} = 1$, only p_{221}, which is the same as SBCV's p_D or CFRT's D_2, always stays inside the desired interval $[0, 1]$. In particular, as we discussed previously, p_{221} has the nice property of approaching zero as $\tau^2 \to 0$, with no need of any further condition (see below).

3 Both "pure" methods give sensible answers in their own right. The pure Bayesian measure p_{222}, which is just p_D^B, is discussed previously. The pure sampling/likelihood measure, p_{111}, provides the correct measure for the number of parameters in the sampling/likelihood model, even when $\tau^2 = 0$, for the value of τ does not enter the picture.

4 Compared to p_{111} and p_{221} discussed above, when $\tau \to 0$, the other measures corresponding to the "Likelihood" column, that is, with $k = 1$, converge to zero *only* *when the prior information is correct.* That is, when $\tau \to 0$, if this strong prior information is correct, then both the true θ and $\bar{X}$ will also converge (almost surely in the case of $\bar{X}$) to the prior mean zero. Otherwise, either of them can take nonsensical values because, when $\tau^2 \to 0$, $p_{121} \to -n\theta^2$ and $p_{211} \to n\bar{X}^2$.

5 Besides $p_{222} = 1$, the behavior of the measures corresponding to "Posterior" column, that is, with $k = 2$, is even more troublesome when $\tau^2 \to 0$. The assumption of "correct prior information" is no longer enough. For p_{112}, it is simply beyond "rescue" – it will converge to $-\infty$ as $\tau^2 \to 0$. For p_{122} and p_{212}, we need to impose $\theta/\tau \to 0$ or $\bar{X}/\tau \to 0$ when $\tau \to 0$ in order to ensure their converging to zero. Clearly, these assumptions would have little statistical meaning or practical relevance.

Some readers might be puzzled by the above discussion. Besides the two "pure measures", p_{111} and p_{222}, which do seem to have good theoretical basis and indeed deliver quite sensible results, why should anyone expect anything sensible from the remaining six "hybrid" ones? Granted, p_{221} is the one that has been given a good amount of theoretical underpinning by SBCV, and hence its good performance should not come as a surprise. But SBCV's theoretical justifications are essentially large-sample ones, which could be applied to any of the rest of five hybrids. In any event, these large-sample results provide little insight regarding how the hybrid ones work or do not work when the prior information cannot be ignored, which are the cases, arguably, where Bayesian modelling is most needed.

7 An even bigger puzzle...

In case a reader is not convinced that there is a puzzle here, let us bring in another curiosity by examining the "dual" measure to p_{221}, that is, p_{122}. The reason we labelled it "dual" should be clear shortly, but for now let's just notice that the subscript of one is the mirror reflection of the other, signifying that we have made a "trade-off" in our Bayesian/non-Bayesian choices in the three ingredients. The reason we investigate p_{122} is that it is actually the very measure Moody (1992) attempted to approximate. This is discussed in Section 2 of SBCV, who pointed out that Moody's (1992) approximation p^* is identical to SBCV's p_D for a class of ANOVA models, for which our simple model is a special case.

Specifically, the equation (4) of SBCV, in our notation and with log posterior in place of log-likelihood, as in Moody (1992), is

$$p_{122} = E_1[d_\Theta^{(2)}(\mathbf{y}, \theta, \tilde{\theta}_2(\mathbf{y}))] \approx p^* \equiv \mathrm{trace}(KJ^{-1}), \tag{12}$$

where

$$J = -E_1\left[\frac{\partial^2 \log p(\theta|\mathbf{y})}{\partial\theta^2}\right], \quad \text{and} \quad K = V_1\left[\frac{\partial \log p(\theta|\mathbf{y})}{\partial\theta}\right], \tag{13}$$

and both E_1 and V_1 operations are with respect to the sampling distribution. Because of the quadratic nature of the log-posterior, as given in (8), it is trivial to verify that, by noting (6),

$$J = \frac{1}{\sigma^2} \quad \text{and} \quad K = V_1\left[\frac{(\hat{\theta}_n - \theta)}{\sigma_n^2}\right] = V_1(n\bar{X}) = n. \tag{14}$$

Consequently,

$$p^* = KJ^{-1} = n\sigma_n^2 = p_D, \tag{15}$$

a fact that was emphasized by SBCV (Section 2).

At this point we hope that at least one reader would be curious enough to notice that although $p^* = p_D$, which seems quite assuring, the exact value p_{122} that p^* tries to approximate, is always larger than p_D:

$$p_{122} - p^* = \frac{\theta^2}{\tau^2(1 + n\tau^2)}. \tag{16}$$

Although one might argue that this difference is of order n^{-1}, and hence negligible, one must keep in mind that this argument itself does not explain why p^* is the "right approximation"—indeed, the difference between any two of the p_{ijk}'s in the tables is of order n^{-1} (when $\tau^2 > 0$).

We surmise that the real reason that p^* works better than the measure it actually tries to approximate is because by invoking the large-sample quadratic approximations, the sampling calculation underlying p^* becomes closer to the posterior calculation because of the symmetry in the sampling and posterior distributions under normality. However, this argument would suggest that we should try to be as Bayesian as possible, which would lead to $p_D^B = p_{222}$, not p_{221}, yet we have seen that p_{222} is not as sensible as p_{221}, from the Bayesian model complexity point of view.

Even more puzzling is the fact that both p^* and p_{221} achieve sensible results by giving up being Bayesian in one of the three choices, but they give up different ones. For p^* (and for the p_{122} it approximates), the log posterior is used in formulating the d (or D) measure, but the expectation is with respect to the sampling distribution. For p_{221} the opposite is the case, the log part uses the sampling density/likelihood function, but the expectation is with respect to the posterior density. This is what we meant by "duality". Our simple example suggests that giving up the "purity" is necessary as otherwise one ends up with either p_{111} or p_{222}, neither of them as sensible as p_{122} or p^*. But what is the fundamental principle behind such a "hybrid" method? Without a sound principle, where is the assurance that we are not lost in the sea of ad hoc methods, such as our "Eights" or CFRT's "Eights", when we do not have the luxury of scrutinizing various measures as given in Table 1 and Table 2?

We are puzzled, very much. We hope that CFRT's rejoinder could help to slow down the rate at which our hairs are leaving us......

References

Dawid, A. P. (2002). "Discussion of "Bayesian measures of model complexity and fit", by Spiegelhalter *et al.*" *Journal of the Royal Statistical Society, Series B*, 64: 624. 689, 693

Dempster, A., Laird, N., and Rubin, D. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion)." *Journal of the Royal Statistical Society, Series B*, 39(1): 1–38. 690

Gelman, A. and Meng, X.-L. (1995). "Model checking and model improvement." In *Practical Markov Chain Monte Carlo*, 189–201. Chapman and Hall, London. 689

Gelman, A., Meng, X.-L., and Stern, H. (1996). "Posterior predictive assessment of model fitness via realized discrepancies (with discussion)." *Statistica Sinica*, 6: 733–806. 689

Little, R. and Rubin, D. B. (1983). "On jointly estimating parameters and missing data by maximizing the complete-data likelihood." *The American Statistician*, 37: 218–220. 692

Meng, X.-L. (1994). "Posterior predictive p-values." *Annals of Statistics*, 22: 1142–1160. 689

Moody, J. (1992). "The effective number of parameters: an analysis of generalization and regularization in nonlinear learning systems." In *Advances in Neural Information Processing Systems 4*, volume 4. San Mateo, CA. 693, 694, 696

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*, 64: 1–34. 687

Vaida, F. and Blanchard, S. (2005). "Conditional Akaike Information for Mixed Effects Models." *Biometrika*, 92(2): 351–370. 692

Acknowledgments

Meng thanks NSF for partial support.