n° 2013-43

Two discussions of the paper "Bayesian Measures of Model Complexity and Fit" by D. Spiegelhalter et al., Read before the Royal Statistical Society at a meeting organized by the Research Section on Wednesday March 13th 2002

E. MORENO[1] − F. J. VAZQUEZ-POLO[2]
C. P. ROBERT[3]

December 2013

[1] Universidad de Granada.

[2] Universidad de Las Palmas de Gran Canaria, TiDES Institute, Spain.

[3] Université Paris-Dauphine, CEREMADE ; CREST and University of Warwick, UK. Email : xian@ceremade.dauphine.fr

Two discussions of the paper "Bayesian measures of model complexity and fit" by D. Spiegelhalter et al., Read before The Royal Statistical Society at a meeting organized by the Research Section on Wednesday, March 13th, 2002

Elías Moreno[1], Francisco–José Vázquez–Polo[2,3], and Christian P. Robert[4,5]

[1] *Universidad de Granada,* [2] *Universidad de Las Palmas de Gran Canaria,* [3] *TiDES Institute, Spain,* [4] *University of Warwick, UK, and* [5] *Université Paris-Dauphine, France*

Summary. These are the written discussions of the paper "Bayesian measures of model complexity and fit" by D. Spiegelhalter et al. (2002), following the discussions given at the Annual Meeting of the Royal Statistical Society in Newcastle-upon-Tyne on September 3rd, 2013.

1. Discussion by E. Moreno and F.-J. Vázquez–Polo

This is an interesting paper, in which a new dimension correction to penalise over-fit models is presented. It has given rise to considerable discussion; here, we focus on the DIC model selection procedure defined in the paper.

Eleven years later, model selection for complex models remains an open problem. The weak link of the Bayesian model selection approach is the elicitation of the prior over models and over the model parameters to be used in the procedure. Several priors have been proposed for interesting model selection problems, such as variable selection in high dimensional regression, clustering, change points and classification, but none of them satisfy all reasonable requirements. Thus, we fully agree with the authors' claim in justifying DIC that "*full elicitation of informative priors and utilities is simply not feasible in most situations*". However, this does not imply that in model selection we can avoid the use of priors in a coherent way (Berger and Pericchi, 2001).

1.1. Does the DIC have a justification from a decision theory viewpoint?

In model selection we have a sample $\mathbf{y}_n$ of size n, a discrete class of k competing sampling models $\mathfrak{M}$, the sampling density of model M_i is $f(\mathbf{y}_n|\theta_i, M_i)$, and a prior for models and model parameters $\pi(\theta_i, M_i) = \pi(\theta_i|M_i)\,\pi(M_i)$, where $\theta_i \in \Theta_i$. The parameter spaces are typically continuous.

In model selection the quantity of interest is the model, and therefore the decision space is $\mathfrak{D} = \{d_j, j = 1, ..., k\}$, where d_j is the decision to choose model M_j, and the states of nature is the class of models $\mathfrak{M}$. Given a loss function $\mathfrak{L}(d_i, M_j)$, $\mathfrak{L} : \mathfrak{D} \times \mathfrak{M} \longrightarrow \mathbb{R}^+$, the optimal Bayesian decision is to choose the model M^π such that

$$M^\pi = \arg \min_{i=1,...,k} \sum_{j=1}^k \mathfrak{L}(d_i, M_j)\pi(M_j|\mathbf{y}_n),$$

where

$$\pi(M_j|\mathbf{y}_n) = \frac{m_j(\mathbf{y}_n)\pi(M_j)}{\sum_{j=1}^k m_j(\mathbf{y}_n)\pi(M_j)},$$

and the marginal $m_i(\mathbf{y}_n) = \int_{\Theta_i} f(\mathbf{y}_n|\theta_i, M_i)\pi(\theta_i|M_i)d\theta_i$, is the likelihood of model $M_i, i = 1, ..., k$. This means that whatever loss function $\mathfrak{L}(d_i, M_j)$ we use, the optimal decision depends on the posterior model probabilities; that is, the decision formulation takes into account the uncertainty of the model. However, the DIC does not depend on $\pi(M_j|\mathbf{y}_n), j = 1, ..., k$.

1.2. Does the DIC correspond to a Bayesian procedure?

The Bayesian procedures automatically penalise model complexity without any adjustment (Dawid, 2002), and this is a good reason to require a model selection procedure to be Bayesian. Another reason is that the competing models can be averaged, with the weights being the model posterior probabilities. On the other hand, for Schwarz's Bayesian information criterion (BIC), to compare model M_i with M_j,

$$-2 \log BIC_{ij}(\mathbf{y}_n) = -2 \log \frac{f(\mathbf{y}_n|\hat{\theta}_i(\mathbf{y}_n), M_i)}{f(\mathbf{y}_n|\hat{\theta}_j(\mathbf{y}_n), M_j)} + (d_i - d_j) \log n,$$

where d_i, d_j are the dimensions of Θ_i and Θ_j, there is a Bayes factor B_{ij} such that $|-2\log B_{ij} - 2\log BIC_{ij}| = O_P(n^{-1/2})$ (Kass and Wasserman, 1995), and thus the BIC asymptotically corresponds to a Bayes factor, we do not see that a similar correspondence can be established with the

$$DIC_{ij}(\mathbf{y}_n) = -2\log\frac{f(\mathbf{y}_n|\bar{\theta}_i(\mathbf{y}_n), M_i)}{f(\mathbf{y}_n|\bar{\theta}_j(\mathbf{y}_n), M_j)} + \text{Correction}_{ij}$$

where $\bar{\theta}_i(\mathbf{y}_n) = E_{\theta_i|\mathbf{y}_n}\theta_i$, and

$$\text{Correction}_{ij} = 4\left\{E_{\theta_i|\mathbf{y}_n}\log f(\mathbf{y}_n|\theta_i, M_i) - E_{\theta_j|\mathbf{y}_n}\log f(\mathbf{y}_n|\theta_j, M_i)\right\} + 4\log\frac{f(\mathbf{y}_n|\tilde{\theta}_i(\mathbf{y}_n), M_i)}{f(\mathbf{y}_n|\tilde{\theta}_j(\mathbf{y}_n), M_j)}.$$

We note that under mild conditions $|\hat{\theta}(\mathbf{y}_n) - \bar{\theta}(\mathbf{y}_n)| = O_P(n^{-1})$, and hence the main difference between BIC and DIC comes from the correction term. As a result of this term, the DIC does not correspond to a Bayesian procedure.

1.3. Asymptotic.

The DIC is not a consistent model selection procedure and although it is a negative property of the procedure, this does not seem to worry the authors, who argue that "*we neither believe in a true model nor would expect the list of models being considered*". This implies that the probability of a model has no meaning, as no model space is considered. However, the point is that if we applied the DIC to a case in which the class of models were known, we would have consistency.

On the other hand, some statisticians, for instance Fraser (2011), have suggested that the sampling properties of the Bayesian methods should be studied. In this respect, Wasserman (2011) asserts that "we must be vigilant and pay careful attention to the sampling properties of procedures". We agree with both these views. Moreover, consistency is a very useful sampling property that allows us to compare the behaviour of alternative Bayesian model selection procedures for complex models.

Consistency in a model selection procedure for a given class of models $\mathfrak{M}$ means that when sampling from a model in $\mathfrak{M}$, the posterior probability of this model tends to one as the sample size tends to infinity. Bayesian procedures for model selection are typically consistent when the dimension of the models is small compared with the sample size (David, 1992; Casella et al., 2009). Furthermore, when the model from which we are sampling is not in the class $\mathfrak{M}$, the Bayesian procedure asymptotically chooses a model in $\mathfrak{M}$ that is as close as possible to the true one, in the Kullback–Leibler distance.

On the other hand, consistent Bayesian procedures for low dimensional models are not necessarily consistent for high dimensional models. For example: (a) Schwarz's approximation to the Bayes factor BIC is not necessarily consistent in high dimensional settings (Berger, 2003; Moreno et al., 2010). (b) When the number of models increases with the sample size, as occurs in clustering, change point or classification problems, consistency of the Bayesian model selection procedure depends not only on the prior over the model parameters but also on the prior over the models. In fact, default priors commonly used for discrete spaces may give an inconsistent Bayesian model selection procedure, as occurs in clustering when using the uniform prior over the models (Casella et al., 2012). (c) In variable selection in regression when the number of regressors p increases with the sample size, i.e., $p = O(n^b), 0 \le b \le 1$, some priors that are commonly used over the model parameters and over the model space make the Bayesian procedures inconsistent. For instance, the $g-$priors (Zellner, 1986) with $g = n$ produce an inconsistent Bayesian procedure. The mixture of $g-$priors with respect to the InverseGamma$(g|1/2, n/2)$, or the intrinsic priors (Moreno et al., 1998) over the model parameters when combined with the independent Bernoulli prior on the model space (George and McCulloch, 1997; Raftery et al., 1997) may also provide an inconsistent Bayesian procedure.

These results show that consistency can be a very useful property for the difficult task of selecting priors for model selection in complex models.

2. Discussion by C.P. Robert

The main issue with DIC undoubtedly is the question of its worth for (or within) Bayesian decision analysis (since I doubt there exist many proponents of DIC outside the Bayesian community). The appeal of DIC is, I presume, to deliver a *single* summary per model for all models under comparison and to allow therefore for a complete ranking of those models. I however object at the worth of simplicity for simplicity's sake: models are complex (albeit less than reality) and their usages are complex as well. To consider that model A is to be preferred upon model B just because $DIC(A) = 1228 < DIC(B) = 1237$ is a mimicry of the complex mechanisms at play behind model choice, especially given the wealth of information provided by a Bayesian

framework. (Non-Bayesian paradigms may be more familiar with procedures based on a single estimator value.) And to abstain from accounting for the significance of the difference between $DIC(A)$ and $DIC(B)$ clearly makes matters worse.

This is not even discussing the stylised setting where one model is considered as "true" and where procedures are compared by their ability to recover the "truth". David Spiegelhalter repeatedly mentioned during his talk that he was not interested in this. This stance inevitably brings another objection, though, namely that models–as tools instead of approximations to reality–can only be compared against their predictive abilities, which DIC seems unable to capture. Once again, what is needed in this approach to model comparison is a multi-factor and all-encompassing criterion that evaluates the predictive models in terms of their recovery of some features of the phenomenon under study. Or of the process being conducted. (Even stooping down to a one-dimensional loss function that is supposed to summarise the purpose of the model comparison does not produce anything close to the DIC function, unless one agrees to massive approximations.)

Obviously, considering that asymptotic consistency is of no importance whatsoever (as repeated by David Spiegelhalter in his presentation) easily avoids some embarrassing questions, except the (still embarrassing) one about the true purpose of statistical models and procedures. How can those be compared if no model is true and if accumulating data from a given model is not meaningful? How can simulation be conducted in such a barren landscape? I find this minimalist attitude the more difficult to accept that models are truly used as if they were or could be true, at several stages in the process. It also prevents the study of the criterion under model misspecification, which would clearly be of interest.

Another point worth discussing, already exposed in Celeux et al. (2006)), is that there is no unique driving principle for constructing DICs. In that paper inspired from the discussion by De Iorio and Robert (2002), we examined eight different and natural versions of DIC for mixture models, resulting in highly diverging values for DIC and the effective dimension of the parameter, I believe that such a lack of focus is bound to reappear in any multi-modal setting and fear that the answer about (eight) different focus on what matters in the model is too cursory and lacks direction for the hapless practitioner.

My final and critical remark about DIC is that the criterion shares very much the same perspective as Murray Aitkin's integrated likelihood, as already stressed in Robert and Titterington (2002). Both Aitkin (1991, 2010) and Spiegelhalter et al. (2002) consider a posterior distribution on the likelihood function, taken as a function of the parameter but omitting the delicate fact that it also depends on the observable and hence does not exist a priori. See Gelman et al. (2013)) for a detailed review of Aitkin's (2010) book, since most of the criticisms therein equally apply to DIC, and I will not reproduce them here, except for pointing out that DIC escapes the Bayesian framework (and thus requires even more its own justifications).

References

M. Aitkin. Posterior Bayes factors (with discussion). *J. Royal Statist. Society Series B*, 53:111–142, 1991.

M. Aitkin. *Statistical Inference: A Bayesian/Likelihood approach.* CRC Press, Chapman & Hall, New York, 2010.

Berger, J.O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? (with discussion). *Statist. Sci.*, **18**, 1–32.

Berger, J.O and Pericchi, L. (2001). Objective Bayesian methods for model selection: Introduction and comparison (with discussion). In *IMS Lecture Notes-Monograph Series*, Vol 38, pp.135–203. Amsterdam: North–Holland.

Casella, G., Girón, F.J., Martinez, M.L. and Moreno, E. (2009). Consistency of Bayesian procedures for variable selection. *Ann. Statist.*, **37**, 1207–1228.

Casella, G., Moreno, E., Girón, F.J. (2012). Cluster Analysis, Model Selection, and Prior Distributions on Models. *Technical Report*. University of Florida.

G. Celeux, F. Forbes, C.P. Robert, and D.M. Titterington. Deviance information criteria for missing data models (with discussion). *Bayesian Analysis*, 1(4):651–674, 2006.

Dawid, A.P. (2002). Discussion on the Paper by Spiegelhalter, Best, Carlin and van der Linde, *J. R. Statist. Soc. Ser. B*, **64**, 583–639.

Fraser, D.A.S. (2011). Is Bayes posterior just quick and dirty confidence? *Statist. Science*, **26**, 299–316.

M. De Iorio and C.P. Robert. Discussion of "Bayesian measures of model complexity and fit", by spiegelhalter et al. *J. Royal Statist. Society Series B*, 64(4):629–630, 2002.

A. Gelman, C.P. Robert, and J. Rousseau. Inherent difficulties of non-Bayesian likelihood-based inference, as revealed by an examination of a recent book by Aitkin (with a reply from the author). *Statistics & Risk Modeling*, 30:1001–1016, 2013.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinic.*, **7**, 339–374.

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Amer. Statist. Assoc.*, **90**, 928–934.

Moreno, E., Bertolino, F. and Racugno, W. (1998). An intrinsic limiting procedure for model selection and hypothesis testing. *J. Amer. Statist. Assoc.*, **93**, 1451–1460.

Moreno, E., Girón, F. J. and Casella, G. (2010). Consistency of objective Bayes factors as the model dimension grows. *Ann. Statist.*, **38**, 1937–1952.

Raftery, A., Madigan, D. and Hoeting, J. (1997). Bayesian model averaging for linear regression models. *J. Amer. Statist. Assoc.*, **92**, 179–191.

C.P. Robert and D.M. Titterington. Discussion of "Bayesian measures of model complexity and fit", by spiegelhalter et al. *J. Royal Statist. Society Series B*, 64(4):621–622, 2002.

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit (with discussion). *J. Royal Statist. Society Series B*, 64(2):583–639, 2002.

Wasserman, L. (2011). Frasian inference (discussion on the paper by Fraser, 2011). *Statistical Science*, **26**, 322–325.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g–prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti, (eds. P. K. Goel and A. Zellner)*, pp. 233–243. Amsterdam: North–Holland.