Bayesian Analysis (2015) **10**, Number 2, pp. 511–515

Comment on Article by Dawid and Musio*

C. Grazian[†], I. Masiani[‡], and C. P. Robert[§]

Abstract. This note is a discussion of the article "Bayesian model selection based on proper scoring rules" by A. P. Dawid and M. Musio, to appear in *Bayesian Analysis*. While appreciating the concepts behind the use of proper scoring rules, we point out here some possible practical difficulties with the advocated approach.

Keywords: Bayesian model choice, proper scoring rules, Bayes factor.

The[1] frustrating issue of Bayesian model selection preventing improper priors (DeGroot, 1982) and hence most objective Bayes approaches has been a major impediment to the development of Bayesian statistics in practice (Marin and Robert, 2007), as the failure to provide a "reference" answer is an easy entry for critics who point out the strong dependence of posterior probabilities on prior assumptions. This was presumably not forecasted by the originator of the Bayes factor, Harold Jeffreys, who customarily and informally used improper priors on nuisance parameters in his construction of Bayes factors (Robert et al., 2009). (The expansion (4) in the paper, while worth recalling, is unlikely to convince such critics.) It is therefore a very welcome item of news that a truly Bayesian approach can allow for improper priors.

As also pointed out in the paper, there exist a wide range of "objective Bayes" solutions in the literature (Robert, 2001), all provided with validating arguments of sorts, but this range by itself implies that such solutions are doomed in that they cannot agree for a given dataset and a given prior.

Finding a criterion that does not depend on the normalising constant of the predictive possibly is the unravelling key to handle improper priors, and we congratulate the authors for this finding of the Hyvärinen score and related proper scoring rules. Some difficulties deriving from the use of improper prior distributions in model choice may be solved by applying the approach proposed in the paper. There are nonetheless some issues with this solution:

- (Calibration difficulties) Once the score value is computed, the calibration of its strength very loosely relates to a loss function, hence makes decision in favour of a model difficult;
- (A clear dependence on parameterisation) Changing x into the transform $\mathfrak{h}(x)$ produces a different score;

*Main article DOI: 10.1214/15-BA942.

[†] Università La Sapienza–Roma, Université Paris-Dauphine

[‡] Università La Sapienza–Roma, Université Paris-Dauphine

[§] Université Paris-Dauphine, University of Warwick, xian@ceremade.dauphine.fr

[1] This work is based on a Master research thesis written by the second author under the supervision of the first and third authors, at Université Paris-Dauphine. The first author is a PhD candidate at Università La Sapienza–Roma and Université Paris-Dauphine.

© 2015 International Society for Bayesian Analysis DOI: 10.1214/15-BA942C

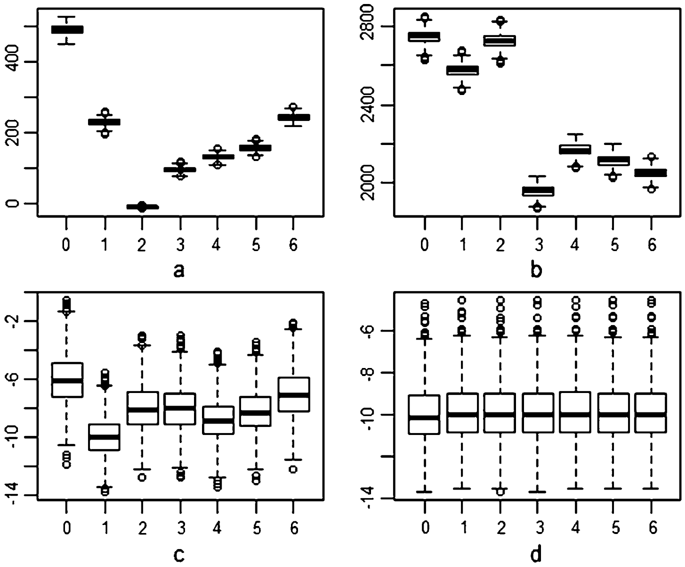

Figure 1: Boxplots over 1,000 simulations of the sample distributions of the scores of seven models under analysis, depending on the true model. Model selection is performed in the case of nested linear normal models. The data was simulated from one of seven nested linear models with up to six covariates. The design matrix is denoted by $\mathbf{X}$. While M_0 is the model that uses zero covariate, M_1 to M_6 use the first, the first two, up to all of the covariates. The values of the covariates were simulated from normal proposals, except for the first column, made of 1's. The data $\mathbf{y} = (y_1, \dots, y_n)$ have distribution $\mathbf{y}|\boldsymbol{\theta} \sim N(\mathbf{X}\boldsymbol{\theta}, \sigma^2)$, with $n = 100$ and $\sigma^2 = 10$. In (a) the true model used for the generations is M_2 (which considers a single regressor), in (b) it is M_3 which considers the first two regressors, in (c) it is M_1 which considers only the constant, while in (d) the correct model is M_0 which has no parameter.

- (A dependence on the dominating measure) As exhibited in the case of exponential families and (30), changing the dominating measure modifies the score function;
- The arbitrariness of the Hyvärinen score, which is indeed independent of the normalising constant, but offers limited arguments in favour of this particular combination of derivatives. Since there exists an immense range of possible score functions, a stronger connection with inferential properties is a clear requirement;
- As noted above, consistency is not a highly compelling argument for the layperson, as it does not help in the calibration and selection of the score. Having an inconsistent multivariate score, while the prequential score remains consistent, is highlighting this difficulty.

Furthermore, the only application of the method presented in the paper is within the setting of the Normal linear model, and we worry that the approach may not be easily

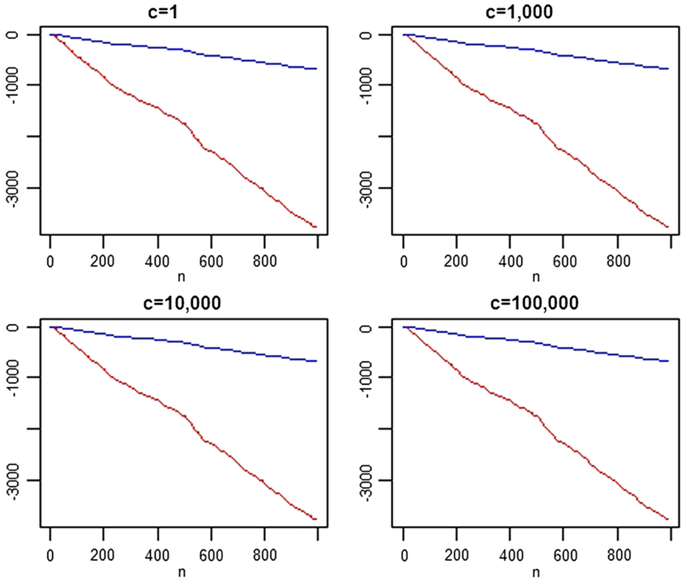

Figure 2: (Linear model) Log-Bayes factor (red) and difference of the scores (blue) as a function of an increasing sample size $n = 1, \dots, 1000$, and of the prior variance on θ, $V = c\sigma^2$, where $\sigma^2 = 10$ is known. Given simulated data $\mathbf{y} = (y_1, \dots, y_n)$ with conditional distribution $\mathbf{y}|\theta \sim N(\mathbf{X}\theta, \sigma^2)$, we consider one regressor and two possible models for generating the data: $M_0 : \theta = 0$ and $M_1 : \theta = 1$ when the true model is M_1.

extended to other types of models. In particular, the representation of the precision matrix of the marginal distribution in (33), based on the Woodbury matrix inversion lemma, is essential to easily apply the proper scoring rule approach to model choice with an improper prior, given that an improper prior may then be seen as a limiting version of a conjugate prior and its influence disappears in the following computations. However, the approach overcomes the singularity of the precision matrix of the marginal distribution.

We first performed some simulation studies when applying the proposed method to models that differ from the Normal linear model. When choosing between two different models with no covariates, we observed that the proposed approach can perform well as, for instance, when a Gamma model is opposed to a Normal model (well in the sense of comparing with a standard Bayes factor). However, when a Pareto distribution and a Normal distribution are compared, the approach does not often select the right model when data are generated from the Pareto distribution, while the Bayes factor always yields the right model. In addition, we came to the realisation that the method based on the Hyvärinen scoring rule may not be applied to some models, for example, when data come from a Laplace distribution, which is not differentiable at 0, or for discrete models.

Our simulation studies have also covered linear models, both nested and non-nested. The details of the simulation models are given in the captions of Figures 1–3. The

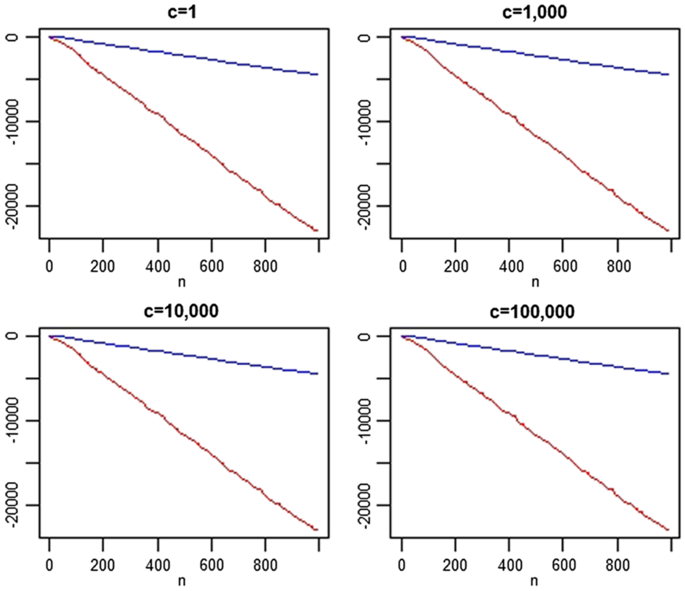

Figure 3: (Nested models) Log-Bayes factor (red) and difference of the scores (blue) as a function of an increasing sample size $n = 1, \ldots, 1000$, and of the prior variance on θ, $V = c\sigma^2$, where $\sigma^2 = 10$. The setting is the same as Figure 1, where we consider six possible regressors and we compare model M_3 which considers the first three regressors against model M_6 which considers all the regressors (M_3 is the true model in our simulations).

performance of the multivariate Hyvärinen score when comparing Normal linear models is excellent, as shown in Figure 1, even when using an improper prior, provided the sample size is larger than the number of parameters in the model. Following repeated simulations, we observed that the method is always able to choose the right model. We, however, noticed that, when the true model does not involved covariates, the ability of the method to discriminate between models is reduced. Although this approach shows a consistent behaviour and chooses the right model with higher and higher certainty when the sample size increases, our simulations have also shown that the log-proper scoring rule tends to infinity more slowly than the Bayes factor or than the likelihood ratio. It is approximately four times slower, all priors being equal, as shown in Figures 2 and 3, which represent the comparison between the approach based on the log-Bayes factor and the one based on the difference between the score functions for the case of linear models, both nested (Figure 3) and non-nested (Figure 2).

As a final remark, we would like to point out the alternative and recent proposal of Kamary et al. (2014) for correctly handling partly improper priors in testing settings through the tool of mixture modelling, each model under comparison corresponding to a component of the mixture distribution. Testing is then handled as an estimation problem in an encompassing model. Therein, the authors show consistency in a wide range of situations. We currently appreciate the approach through mixture estimation as the most compelling for the many reasons advanced in Kamary et al. (2014), in particular because the posterior distribution of the weight of a model is easily interpretable and scalable towards selecting this very model or an alternative one. Furthermore, it returns posterior probabilities for the models under comparison without the need to resort to specific prior probability weights.

References

DeGroot, M. (1982). "Discussion of Shafer's 'Lindley's paradox'." *Journal of the American Statistical Association*, 378: 337–339. 511

Kamary, K., Mengersen, K., Robert, C., and Rousseau, J. (2014). "Testing hypotheses as a mixture estimation model." arXiv:1214.2044. 514, 515

Marin, J. and Robert, C. (2007). *Bayesian Core*. Springer-Verlag, New York. MR2723361. 511

Robert, C. (2001). *The Bayesian Choice*. Springer-Verlag, New York, second edition. MR1835885. 511

Robert, C., Chopin, N., and Rousseau, J. (2009). "Theory of Probability revisited (with discussion)." *Statistical Science*, 24(2): 141–172 and 191–194. 511