Bayesian Analysis (2015) **10**, Number 2, pp. 501–504

Comment on Article by Dawid and Musio*

Matthias Katzfuss† and Anirban Bhattacharya‡

The authors consider the interesting and important issue of Bayesian inference based on objective functions other than the likelihood. They focus on model selection in the low-dimensional setting using prequential local proper scoring rules.

1 General non-likelihood-based inference

There is a large and disparate literature on inference based on objective functions other than the likelihood. We will briefly mention some examples here, but we believe that a more thorough review and comparison would be a worthy endeavor.

Numerous objective functions have been proposed to replace the (log-)likelihood in pursuit of various inference goals. Proper scoring rules are a natural choice for serving as such objective functions, due to their property of being minimized (in expectation) under the true model. Depending on the goal of the analysis, certain well-known proper scoring rules can achieve robustness (e.g., continuous ranked probability score, or CRPS), have simple closed-form expressions (e.g., Dawid–Sebastiani score), or do not require densities (e.g., CRPS) or normalizing constants (e.g., Hyvärinen score, as in the present paper). See Gneiting and Katzfuss (2014) for a recent review of these and other scoring rules.

In a frequentist context, examples of approaches falling into this category of scoring-rule-based inference are minimum contrast estimation (e.g., Pfanzagl, 1969; Birgé and Massart, 1993), composite likelihood (e.g., Lindsay, 1988), and M-estimation (e.g., Huber and Ronchetti, 2009). Some further review is given in Dawid et al. (2014).

There have also been related approaches in the Bayesian framework. Shaby (2014) provides a nice review of Bayesian inference using general objective functions and, based on results of Chernozhukov and Hong (2003), he proposes an "open-faced sandwich adjustment" to obtain pseudo-posteriors with properly calibrated frequentist properties. Further, the "Gibbs posterior" (Jiang and Tanner, 2008; Li et al., 2013) has received considerable interest, where the negative log-likelihood is replaced by some "empirical risk" R_n (usually targeting the specific parameter to be estimated) to construct a pseudo-posterior of the form

$$Q(\theta) \propto \exp\{-\lambda R_n(\theta)\}\pi(\theta), \tag{1}$$

where λ is a positive scaling constant (often called "temperature"). Sampling from the pseudo-posterior Q can be performed via standard MCMC algorithms.

*Main article DOI: 10.1214/15-BA942.
†Department of Statistics, Texas A&M University, katzfuss@tamu.edu
‡Department of Statistics, Texas A&M University, anirbanb@stat.tamu.edu

 DOI: 10.1214/15-BA942A

2 Objective Bayesian model selection

In objective Bayesian model selection, a discrete prior is assumed on a (finite) class of models, and given a particular model, objective improper priors are placed on the model parameters. While improper priors are commonly used for analysis of a single model, one faces difficulties in comparing models via Bayes factors, since the marginal likelihoods of the competing models are only specified up to arbitrary constants. A number of remedies have been proposed in the literature to deal with this issue, such as fractional Bayes factors (O'Hagan, 1995) and intrinsic Bayes factors (Berger and Pericchi, 1996).

In the present paper, the authors take a different approach, which relies on replacing the (log-)marginal likelihood by a local proper scoring rule. The Hyvärinen score is recommended as a default. From the expression of the Hyvärinen score in the authors' equation (16), it can be seen that the arbitrary constant disappears. The authors look at examples where the Hyvärinen scores are analytically tractable and provide asymptotic orders for the difference in Hyvärinen scores assuming the respective models to be true.

Some clarification regarding practical implementation of the model selection procedure presented here would be helpful. When can we be sure that one model is truly better than another — or in other words, can anything be said about posterior model probabilities (also see Section 3 below)? Can the the necessary quantities be computed for models beyond the simple Gaussian examples considered in the paper?

3 Scaling issues

As indicated in (1) above, the literature on Gibbs posteriors typically includes a multiplicative scaling constant λ on the objective function. The choice of λ is considered a critical issue, as it has a direct effect on the (pseudo-)posterior uncertainty. Shaby (2014) does not consider a multiplicative scaling of the objective function, but his open-face-sandwich correction automatically adjusts for such scaling, and his approach is thus invariant to scaling. Without such a correction, the scaling issue also arises when the objective function is specified to be a proper scoring rule, including the Hyvärinen score. As implicitly acknowledged by the authors in their Footnote 2, the scaling of a proper scoring rule is arbitrary, in that any proper scoring rule is still proper when multiplied by a constant.

In the context of model selection between models M_1 and M_2 with scores S_{M_1} and S_{M_2}, respectively, the scaling constant can arbitrarily inflate or deflate the pseudo Bayes factor,

$$\text{PBF} = \frac{\exp(\lambda S_{M_1})}{\exp(\lambda S_{M_2})}$$

and thus the amount of evidence in favor of M_1 over M_2 (cf. Kass and Raftery, 1995). This also makes it challenging to compute pseudo posterior model probabilities, such as

$$\widetilde{P}(M_1|\mathbf{x}) = \frac{\exp(\lambda S_{M_1})}{\exp(\lambda S_{M_1}) + \exp(\lambda S_{M_2})}. \tag{2}$$

If S_{M_1} is larger than S_{M_2}, (2) can be arbitrarily close to 0.5 or 1 by choosing λ to be very small or very large, respectively.

In light of these scaling issues, how should model selection be calibrated and interpreted? Moreover, is it possible to handle more than two competing models or even high-dimensional settings, where the number of competing models may grow exponentially with the sample size? In the high-dimensional linear regression context, Johnson and Rossell (2012) showed that a number of commonly used procedures (including fractional and intrinsic Bayes factors) assign vanishingly small posterior probabilities to the true model with increasing sample size. The scaling issue may assume an even more important role in such cases.

References

Berger, J. O. and Pericchi, L. R. (1996). "The intrinsic Bayes factor for model selection and prediction." *Journal of the American Statistical Association*, 91: 109–122. MR1394065. doi: http://dx.doi.org/10.2307/2291387. 502

Birgé, L. and Massart, P. (1993). "Rates of convergence for minimum contrast estimators." *Probability Theory and Related Fields*, 97: 113–150. MR1240719. doi: http://dx.doi.org/10.1007/BF01199316. 501

Chernozhukov, V. and Hong, H. (2003). "An MCMC approach to classical estimation". *Journal of Econometrics*, 115: 293–346. MR1984779. doi: http://dx.doi.org/10.1016/S0304-4076(03)00100-3. 501

Dawid, P., Musio, M., and Ventura, L. (2014). "Minimum scoring rule inference." arXiv:1403.3920. 501

Gneiting, T. and Katzfuss, M. (2014). "Probabilistic forecasting." *Annual Review of Statistics and Its Application*, 1(1): 125–151. doi: http://dx.doi.org/10.1146/annurev-statistics-062713-085831. 501

Huber, P. and Ronchetti, E. (2009). *Robust Statistics*. Hoboken, NJ: Wiley, 2nd edition. MR2488795. doi: http://dx.doi.org/10.1002/9780470434697. 501

Jiang, W. and Tanner, M. A. (2008). "Gibbs posterior for variable selection in high-dimensional classification and data mining." *The Annals of Statistics*, 36(5): 2207–2231. MR2458185. doi: http://dx.doi.org/10.1214/07-AOS547. 501

Johnson, V. E. and Rossell, D. (2012). "Bayesian model selection in high-dimensional settings." *Journal of the American Statistical Association*, 107(498): 649–660. MR2980074. doi: http://dx.doi.org/10.1080/01621459.2012.682536. 503

Kass, R. and Raftery, A. E. (1995). "Bayes factors." *Journal of the American Statistical Association*, 90(430): 773–795. doi: http://dx.doi.org/10.1080/01621459.1995.10476572. 502

Li, C., Jiang, W., and Tanner, M. A. (2013). "General oracle inequalities for Gibbs posterior with application to ranking." *Journal of Machine Learning Research: Workshop and Conference Proceedings 30*, 512–521. 501

Lindsay, B. (1988). "Composite likelihood methods." *Contemporary Mathematics*, 80: 221–239. MR0999014. doi: http://dx.doi.org/10.1090/conm/080/999014. 501

O'Hagan, A. (1995). "Fractional Bayes factors for model comparison." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57: 99–138. MR1325379. 502

Pfanzagl, J. (1969). "On the measurability and consistency of minimum contrast estimates." *Metrika*, 14(1): 249–272. doi: http://dx.doi.org/10.1007/BF02613654. 501

Shaby, B. A. (2014). "The open-faced sandwich adjustment for MCMC using estimating functions." *Journal of Computational and Graphical Statistics*, 23(3): 853–876. MR3224659. doi: http://dx.doi.org/10.1080/10618600.2013.842174. 501, 502