BAYESIAN STATISTICS 9,
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)

On moment priors for Bayesian model choice: a discussion

Judith Rousseau & Christian P. Robert
CREST, Paris and Université Paris-Dauphine, CEREMADE
xian@ceremade.dauphine.fr

Summary

In this discussion, we address some difficulties we have with Consonni and La Rocca's proposal and we propose a new interpretation of their approach.

We cannot but agree with the authors that "Bayesian model choice is an important and fascinating area" and we applaud this new attempt at providing an objective answer to the variable selection problem, although we disagree with some aspects of the solution they adopt. We sympathise with the idea of separating both hypothesis, as an approach to the difficult problem of using pointwise hypotheses as approximations of interval null hypothesis, i.e. replacing

$$H_0 : d(\theta, \theta_0) < \epsilon, \quad \text{by} \quad H_0' : \theta = \theta_0.$$

It is indeed a difficult issue and the Bayes factor associated to the problem H_0' is not a satisfactory approximation of the Bayes factor associated to H_0, see Rousseau (2007) for a discussion on the subject. The assymmetry between the asymptotic behaviours of the Bayes factor under H_0' and H_1 comes from this problem.

In his approach to the same goal of defining a general framework to the model choice problem, José Bernardo adopts a somehow opposed perspective with which we much more readily agree, namely that an objective Bayes principle should start from an encompassing model rather than seeking priors on every possible submodel. We refer the reader to McCulloch and Rossi (1993), Mengersen and Robert (1996), Goutis and Robert (1998), Dupuis and Robert (2003), Marin and Robert (2007) for some arguments of ours on this perspective.

The definition of *local priors* and hence of *non-local priors* does not appeal very much to us as the notion of the prior density π_1 being non-zero in a neighbourhood of the null hypothesis does not qualify how much the alternative prior weights this neighbourhood of the null. We also find it quite disturbing to use such a prior for estimation. Furthermore, while getting a closed-form expression in Theorem 1 is a neat feat, the extreme dependence of the Bayes factor on the power h makes it

J. Rousseau and C.P. Robert are supported by the 2007–2010 grant ANR-07-BLAN-0237-01 "SP Bayes".

difficult to advocate the use of this prior without further guidance upon the choice of h.

Using the well-known dichotomy between prior and loss selection (Rubin, 1987, Robert, 2001), we think that the definition of non-local priors should be replaced by the use of loss functions that take into account the distance to the null, once again reverting to the principles exposed in José Bernardo's paper in this volume. This perspective was actually pursued in Robert and Casella (1994) (see also Goutis and Robert, 1997), where Bayes procedures were constructed, with the additional incentive of allowing the use of improper priors without resorting to pseudo-Bayes factors (O'Hagan, 1995, Berger and Pericchi, 1996). In particular the Bayes factor associated to the non local prior proposed by the authors and prior to them by Johnson and Rossell (2010)

$$\tilde{\pi}_h(\theta) \propto g_h(\theta,\theta_0)\pi(\theta), \quad g_h(\theta,\theta_0) = |\theta-\theta_0|^h$$

is the Bayesian solution associated to the prior $\pi(\theta)$ and the loss function

$$L(\delta,\theta) = \left\{ \begin{array}{cccc} 1 & \text{if} & (\delta=1 & \& \quad \theta=\theta_0) \\ |\theta-\theta_0|^h & \text{if} & (\delta=0 & \& \quad \theta\neq\theta_0) \end{array} \right.$$

Presented as such, the solution makes much more sense and also leads to wider generalisations and more interesting perspectives. One such is the use of other distances than $|\theta-\theta_0|$ in problems where the question can be formalised on other parameterisations and for which invariant distances such as $d(f_\theta, f_{\theta_0})$—where d is either the Kullback-Leibler divergence (or a symmetric version of it), or the L_1 or Hellinger distances—would be more appropriate. We recall that Robert and Casella (1994) contains a detailed study of losses jointly addressing testing and simulation: denoting by φ the estimate of $\mathbb{I}_{\theta_0}(\theta)$, i.e. the indicator of the null hypothesis,

$$\begin{aligned} L_1(\theta;\varphi,\hat{\theta}) &= d(\theta,\hat{\theta})\,\mathbb{I}(\varphi=0) + d(\theta_0,\theta)\,\mathbb{I}(\varphi=1) \\ L_2(\theta;\varphi,\hat{\theta}) &= d(\theta,\hat{\theta})\,(1-\varphi)^2 + d(\theta_0,\theta)\,\varphi^2 \\ L_3(\theta;\varphi,\hat{\theta}) &= d(\theta,\hat{\theta})\,(\mathbb{I}_{\theta_0}(\theta)-\varphi)^2 + d(\theta_0,\theta)\,\varphi^2 \\ L_4(\theta;\varphi,\hat{\theta}) &= 2d(\theta,\hat{\theta})\,\mathbb{I}(\varphi=0) + \left\{ d(\theta_0,\theta) + d(\theta_0,\hat{\theta}) \right\} \mathbb{I}(\varphi=1) \end{aligned}$$

out of which only L_4 provides a sensible answer for $d(t)=t^2$:

$$(\varphi^\pi(x),\hat{\theta}^\pi(x)) = \begin{cases} (0,\delta^\pi(x)) & \text{if } \mathrm{var}^\pi(x) < (\delta^\pi(x)-\theta_0)^2 \\ (1,\theta_0)^2 & \text{otherwise} \end{cases}$$

where $\delta^\pi(x)$ is the regular Bayes estimator, namely the posterior mean.

As a last minor remak, we do wonder whether or not the use of those non-local priors (or should we say loss functions?) could help in solving the Lindley–Jeffreys paradox because they exclude some neighbourhoods of the null hypothesis. Note that, although Robert (1993) often gets quoted in relation with this Lindley–Jeffreys paradox, we no longer find it to be a satisfactory solution as it suffers from the same measure-theoretic difficulty as the Savage–Dickey paradox, as exposed in Marin and Robert (2010).

REFERENCES

Berger, J. and L. Pericchi. 1996. The Intrinsic Bayes Factor for Model Selection and Prediction. *J. American Statist. Assoc.* 91: 109–122.

Dupuis, J. and C. Robert. 2003. Model choice in qualitative regression models. *J. Statistical Planning and Inference* 111: 77–94.

Goutis, C. and C. Robert. 1997. Choice among hypotheses using estimation criteria. *Ann. Econom. Statist.* 46: 1–22.

—. 1998. Model choice in generalized linear models: a Bayesian approach via Kullback–Leibler projections. *Biometrika* 85: 29–37.

Johnson, V. and D. Rossell. 2010. On the use of non-local prior densities in Bayesian hypothesis tests. *J. Royal Statist. Society Series B* 72: 143–170.

Marin, J. and C. Robert. 2010. Resolution of the Savage–Dickey paradox. *Electronic Journal of Statistics* 4: 1–9. (To appear.).

Marin, J.-M. and C. Robert. 2007. *Bayesian Core*. Springer-Verlag, New York.

McCulloch, R. and P. Rossi. 1993. Bayes factors for nonlinear hypotheses and likelihood distributions. *Biometrika* 79: 663–673.

Mengersen, K. and C. Robert. 1996. Testing for mixtures: A Bayesian entropic approach (with discussion). In *Bayesian Statistics 5*, eds. J. Berger, J. Bernardo, A. Dawid, D. Lindley, and A. Smith, 255–276. Oxford University Press, Oxford.

O'Hagan, A. 1995. Fractional Bayes factors for model comparisons. *J. Royal Statist. Society Series B* 57: 99–138.

Robert, C. 1993. A Note on Jeffreys-Lindley paradox. *Statistica Sinica* 3(2): 601–608.

—. 2001. *The Bayesian Choice*. 2nd ed. Springer-Verlag, New York.

Robert, C. and G. Casella. 1994. Distance penalized losses for testing and confidence set evaluation. *Test* 3(1): 163–182.

Rousseau, J. 2007. Approximating interval hypotheses: p-values and Bayes factors. In *Bayesian Statistics 8*, eds. J. M. Bernardo, M. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West. Oxford: Oxford University Press.

Rubin, H. 1987. A weak system of axioms for rational behavior and the nonseparability of utility from prior. *Statist. Decision* 5: 47–58.