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.)

Integrated objective Bayesian estimation and hypothesis testing: a discussion

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

Summary

In this discussion, we congratulate Professor Bernardo for his all-encompassing perspective on intrinsic inference and focus on the case of nuisance parameters.

1. UNIFIED INFERENCE

The paper manages the *tour de force* of aggregating intrinsic loss functions with intrinsic (*aka* reference) priors. Thus, Professor Bernardo presents us with a unified picture of Bayesian analysis as he sees it and it is obviously fitting to see this cohesive perspective appearing in the Valencia 9 proceedings as a kind of third unification! We appreciated very much the paper and our comments will thus concentrate on minor issues rather than on the big picture, since we mostly agree with it. Although the tendency in Bayesian analysis, along the years, and in particular in the Valencia proceedings (see, e.g., Polson and Scott in this volume who discuss shrinkage without a loss function), has been to shy away from the decision-theoretic perspective (see, e.g., Gelman, 2008), it is worth reenacting this approach to the field, both because it sustains to a large extent the validation of a Bayesian analysis and because it avoids the deterioration of its scope into a mechanical data analysis tool.

2. DOWN WITH POINT MASSES!

The requirement that one uses a point mass as a prior when testing for point null hypotheses is always an embarassment and often a cause of misunderstanding in our classes. Rephrasing the decision to pick the simpler model as the result of a larger advantage is thus much more likely to convince our students. What matters in pointwise hypothesis testing is not whether or not $\theta = \theta_0$ holds but what the consequences of a wrong decision are. Of course, there is a caveat in the reformulation of Professor Bernardo, which is that, in the event the null hypothesis $\theta = \theta_0$ is accepted, one has to act with the model $\mathcal{M}_0$. One can of course assume that, given the model $\mathcal{M}_0$, the intrinsic Bayesian statistician would start from the reference

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

prior for $\mathcal{M}_0$, but this involves a dual definition of the prior for the *same* problem that remains a bit of an itch...

The case of compound hypotheses is only half-convincing in that the "natural" solution would seem to us to compare the posterior expected losses under both models, rather than singling out H_0 in a most unbalanced and unBayesian way. We actually take issue with the repeated use of infimums (infima?) in the definition of loss functions.

3. INTRINSIC LOSSES

Most obviously, we welcome the recentering of objective Bayes analyses around the intrinsic losses we developed in Robert (1996a). (Note that the severe lack of invariance of HPD regions was further studied in Druilhet and Marin, 2007, while integrating point estimation losses in the evaluation of credible regions was proposed in Robert and Casella, 1994.)

The handling of nuisance parameters always is a...nuisance, so Definition 5 is a possible solution to this nuisance. While it shies away from using the unsatisfactory argument of λ being "common" to both models, one of us (CPR) somehow dislikes the introduction of the infimum over all values of λ_0: a more agreable alternative would be to integrate over the λ_0's, using for instance an intrinsic prior $\pi(\lambda|\theta_0)$. We however acknowledge the relevance of projections in model comparisons, as illustrated by Robert and Rousseau (2002).

Another issue deals with cases when the nuisance parameter is ill-defined under the null hypothesis, as for instance in our favourite example of mixtures of distributions (Titterington et al., 1985, MacLachlan and Peel, 2000): When the null has several possible representations, the nuisance parameter varies from one representation to the next. A connected issue is the case when the parameter of interest is a function (functional) of the whole parameter vector that is such that there is no explicit way of breaking the whole parameter into a parameter of interest and a nuisance parameter, a setting that typically occurs in semi-parametric problems. Although a natural extension to Bernardo's approach is to define the intrinsic loss between the parameter $\theta = \theta(f)$ and θ_0 as

$$\delta(\theta_0, f) = \inf\{\min(k(f, f_0), k(f_0, f)); f_0 \in \mathcal{F} \text{ satisfies} \quad \theta(f_0) = \theta_0\}$$

such an approach seems impossible to implement in practice, even in simple semi-parametric problems.

When replacing regular testing with checking whether or not the new type of *regret* $\ell\{\theta_0, (\theta, \lambda)\} - \ell_0$ is positive, the so-called *context dependent positive constant* ℓ_0 is equal to

$$\int_\Theta \int_\Lambda \ell_h\{a_1, (\theta, \lambda)\} p(\theta, \lambda|z)\, \mathrm{d}\theta\, \mathrm{d}\lambda$$

in the original formulation. We therefore wonder why the special values $\ell_0 = \log 10^k$ for $k = 1, 2, 3, \ldots$, are of particular interest compared, say, with $\ell_0 = \log \sqrt{\pi}^k$ or $\ell_0 = \log e^k$... The calibration of ℓ_0 suffers from the same difficulty as the calibration of Bayes factors in that the choice of the decision boundary between acceptance and rejection is not based on a loss function. In particular, it is surprising that, in a *objective* context, ℓ_0 does not depend on the number of observations. Typically, the Kullback–Leibler divergence between the densities f_θ and $f_{\theta'}$ associated with n (not necessarily i.i.d) observations increases with n. Should ℓ_0 be rescaled as $n\ell_0$ and is

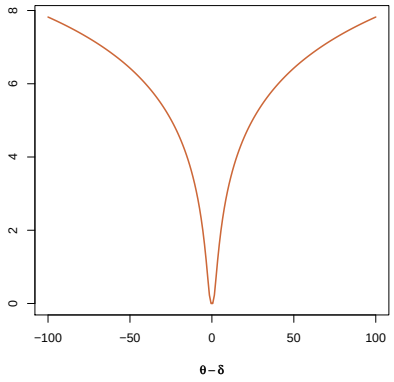

Figure 1: Kullback–Leibler loss function $\ell(\theta, \delta)$ associated with a Cauchy distribution with location parameter θ.

such a scaling appropriate in general? We argue that rescaling by n as such is as *arbitrary* as considering Jeffreys prior as default prior.

A last point of interest to us is whether or not an integrated reference analysis is always possible. Bypassing the issue of finding a reference prior, We wonder if there exist settings where the posterior Kullback–Leibler loss is *uniformly* infinite, thus preventing the choice of a Bayes estimator. For instance, when observing a Cauchy variate x, the intrinsic loss is of the form represented in Figure 1. Since the posterior under the flat prior is a Cauchy distribution with location parameter x, the loss may be increasing too fast for the Bayes estimator to exist A family of models where the Kullback–Leibler loss cannot be applied corresponds to cases where the densities have supports that depend on the parameters in a non-trivial way, i.e.

$$f_\theta(x) = \mathbb{I}_{L(\theta)} g_\theta(x), \quad \text{where} \quad L(\theta) \cap L(\theta')^c \neq \emptyset \quad \text{and} \quad L(\theta') \cap L(\theta)^c \neq \emptyset$$

and $g_\theta(x) > 0$ everywhere.

In conclusion, our point here is to emphasize that, although the Kullback–Leibler loss has compelling features such as additivity, it also suffers from drawbacks, related to the requirement of comparing absolutely continuous distributions (one way or the other) and to its unboundedness. Some other *natural* intrinsic losses could be considered, in particular the Hellinger distance (Robert, 1996b). How would both losses compare and what would their relative merits be? It seems to us that the *natural calibrations* found in Professor Bernardo's proposal could not be used with an Hellinger loss. Now, ould that be such a bad thing...?!

4. REFERENCE PRIORS

Although we essentially agree with most of the construction of reference priors, we are doubtful about the systematic use of repeated (indentically and independently)

data sets. Indeed, in cases where the observations are modelled as a dependent process, say a time series, a part of the parameter vector addresses the dependence structure. Then, first, repeated iid sampling from the model will not provide useful knowledge about these parameters, since they can only be infered correctly by letting the sample size increase to infinity. Second, for a fixed sample size, the Fisher information matrix depends in a non-trivial way on n and it usually has a non-explicit representation. Therefore, the reference prior under repeated sampling does not have an interesting formulation. For instance, when sampling from a stationary Gaussian process with spectral density f_θ, the Fisher information matrix associated with the covariance matrix includes terms of the form

$$\mathrm{tr}\left[\left(T_n(f_\theta)^{-1}T_n(\nabla f_\theta)\right)^2\right] ,$$

where $T_n(f)$ is the n dimensional Toeplitz matrix associated with the function f and ∇f_θ is the first derivative of the spectral density, see Philippe and Rousseau (2003). This expression is not *user-friendly*, to say the least!, whereas the reference prior—obtained by letting the sample size go to infinity—actually corresponds to the limit of the above terms:

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}(\nabla \log f_\theta)^2(x)\,\mathrm{d}x$$

which are much more satisfactory for the construction of a prior distribution. The latter can also be obtained by considering the limit of the reference priors as n goes to infinity, however it is not clear whether it should be interpreted as the reference directly obtained from increasing n in the sampling or as the limit of Professor Bernardo's reference prior when n goes to infinity. These two approaches might indeed lead to quite different results, for instance for non-stationary models.

REFERENCES

Druilhet, P. and J.-M. Marin. 2007. Invariant HPD credible sets and MAP estimators. *Bayesian Analysis* 2(4): 681–692.

Gelman, A. 2008. Objections to Bayesian statistics. *Bayesian Analysis* 3(3): 445–450.

MacLachlan, G. and D. Peel. 2000. *Finite Mixture Models*. New York: John Wiley.

Philippe, A. and J. Rousseau. 2003. Non-informative priors for Gaussian long-memory processes. *Bernoulli* 8: 451–473.

Robert, C. 1996a. Intrinsic loss functions. *Theory and Decision* 40(2): 191–214.

—. 1996b. *Méthodes de Monte Carlo par Chaînes de Markov*. Paris: Economica.

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

Robert, C. and J. Rousseau. 2002. A Mixture Approach to Bayesian Goodness of Fit. Tech. rep., Cahiers du CEREMADE, Université Paris Dauphine.

Titterington, D., A. Smith, and U. Makov. 1985. *Statistical Analysis of Finite Mixture Distributions*. New York: John Wiley.