Three discussions on model choice

Christian P. Robert
CEREMADE, Université Paris Dauphine, and CREST, INSEE
xian@ceremade.dauphine.fr

January 23, 2006

Abstract

These three discussions, including one jointly written with Judith Rousseau and one written jointly with Gilles Celeux, Florence Forbes and Mike Titterington, are associated with three papers related to Bayesian model choice. This area of Bayesian statistics is still the object of intense and antagonistic debates and our vision of the field is reflected by these three texts.

Intrinsic credible regions: On objective Bayesian approach to interval estimation by José Miguel Bernardo A discussion *[published by TEST]*

Judith Rousseau and Christian P. Robert
CEREMADE, Université Paris Dauphine and CREST, INSEE

In this paper, José Miguel Bernardo presents a unified and structured objective approach to (decisional) statistical inference, based on information theoretic ideas he has used previously to define reference priors. He focusses here on the estimation of credible regions, keeping those values of the parameter that are the least costly rather than the most probable, as in HPD regions. This is an interesting and novel approach to an efficient construction of credible regions when lacking a decision-theoretic basis. As noted in Casella et al. (1993a,b) (see also Robert, 2001, Section 5.5.3, for a review), the classical decision-theoretic approaches to credible regions are quite behind their counterpart for point estimation and testing and incorporating loss perspectives in credible sets was then suggested in Robert and Casella (1994).

1 On invariance: link with HPD regions

A possible drawback of HPD regions, in particular in objective contexts, is their lack of invariance under re-parameterization as was pointed out by José Miguel Bernardo. Obviously, HPD regions are defined in terms of a volume-under-fixed-coverage loss and they do minimize the volume among q-credible regions. The lack of invariance hence stems from the lack of invariance in the definition of the volume, which is based on the Lebesgue measure for the considered parametrization θ. Therefore, simply considering a different type of volume based on an invariant measure would result in an *HPD* region that is invariant under reparameterization. A natural invariant measure in this setup is Jeffreys' measure, due to its geometric and information interpretations (among others). The resulting HPD region is thus constructed as the

region C that minimizes

$$\int_C \sqrt{i(\theta)}d\theta, \quad \text{u.c.} \quad P^\pi[C|X] \geq q. \tag{1}$$

This region also corresponds to the transform of the (usual) HPD region constructed using the reference parametrization as defined by José Miguel Bernardo.

Note that, in the above construction, there is absolutely no need in having the prior be Jeffreys prior and this construction could be used in (partially) informative setups. It is also interesting to note that, in regular cases, the above HPD region is asymptotically equivalent to the intrinsic credible region of José Miguel Bernardo. Which of both approaches is the most appealing is probably a question of taste or depends on how they will be used.

On a more philosophical basis, we think that invariance is less compelling an argument for (credible) regions than for point estimations. Indeed, while it is difficult to sell to a customer that the estimator of $h(\theta)$ is not necessarily the transform $h(\hat{\theta})$ of the estimator $\hat{\theta}$ of θ, the transform of a crebible region does remain a credible region, even though it is not always the optimal region. Moreover, invariance under reparameterization should be weighted against shape poor modifications. Indeed, if we impose that the credible region C_h on $h(\theta)$ is the transform by h of the credible region C_{id} on θ, we get exposed to strange shapes for less regular functions h! For instance, if the transform h is not monotonic (but still one-to-one), it is possible to obtain the transform of a credible interval as a collection of several disjoint intervals, always a puzzling feature! Connexity (and maybe to some extent convexity) should be part of the constraints on a credible region.

2 Asymptotic coverage : matching properties

Under regularity properties, the HPD region defined by (1) is a second order matching region for any smooth prior π, in the sense that its frequentist coverage is equal to its posterior coverage to the order $O(n^{-1})$. Third order coverage does not necessarily apply for Jeffreys' prior, though (see Datta and Mukerjee, 2004 or Rousseau, 1997). As José Miguel Bernardo's intrinsic credible region is asymptotically equivalent to the HPD region defined by (1)

there is a chance that second order matching is satisfied, which would explain the good small sample properties mentioned in the paper. In particular, the perturbation due to using the intrinsic loss, compared to using the posterior density, is of order $O(n^{-1})$, so second order asymptotics should be the same between (1) and the intrinsic credible region.

Investing further the higher order matching properties of this credible region would be worthwhile though. Regarding the discrete case, however, things are more complicated than what was mentioned by José Miguel Bernardo since there is usually no matching to orders higher than $O(n^{-1/2})$ or sometimes $o(n^{-1/2})$ for higher dimensional cases. Whether reference posterior q-credible regions provide the best available solution for this particular problem is somehow doubtful as there are many criteria which could reasonably be considered for comparing credible regions or their approximations in the discrete case, see Brown et al. (2002).

3 Computations

Adopting this approach to credible set construction obviously makes life harder than computing HPD regions: while HPD regions do simply require the derivation of a posterior level ϱ for the set $\{\theta : \pi(\theta|x) \geq \varrho\}$ to have coverage q, an intrinsic credible set involves the intrinsic loss—not easily computed outside exponential families—, the posterior intrinsic loss—possibly integrated over a large dimensional space—, the posterior coverage of the corresponding region and at last the bound on $d(\theta|x)$ that garantees q coverage. In large dimensional settings or outside expential frameworks, the tasks simply seems too formidable to be contemplated, especially given that standard numerical features like convexification cannot be taken for granted since the credible region is not necessarily convex or even connected.

Deviance Information Criteria
for Missing Data Models:
A rejoinder *[published by Bayesian Analysis]*

G. Celeux[1], F. Forbes[2], C.P. Robert[3] and D.M. Titterington[4]
[1]*INRIA FUTURS, Orsay,* [2]*INRIA Rhône-Alpes, Grenoble*
[3]*CREST and CEREMADE, Université Paris Dauphine,*
and [4]*Department of Mathematics and Statistics, University of Glasgow*

We are grateful to all discussants for their comments and to the editor Rob Kass for initiating this discussion. Rather than addressing each discussion separately, we identify several themes of interest and contention among the discussants that we now develop separately.

1 Foundations of DIC

A theme common to all discussions is that DIC is so far more of a plausible (?) measure of complexity than a well-grounded criterion. We completely agree with this perspective and even share the more radical prognosis of Meng and Vaida that DIC may simply lack a theoretical foundation. Indeed, there are deeper concerns with DIC than just that of a definition in the missing data case. In this regard, we do agree with Carlin that our "casework" analysis cannot solve the problem of defining a proper DIC for missing data and even less in general. Therefore, Carlin's point that *"authors do not refer at all to any derivation, nor to any subsequent interpretation of model complexity"* is both true and meaningless: if DIC as originally defined is a universal way of evaluating model fit or model complexity, it should also apply in the missing data setting and we showed here that it clearly does not. The main conclusion of our paper is thus that DIC lacks a natural generalisation outside exponential families or, alternatively, that it happened to work within exponential families while lacking a true theoretical foundation. Similarly, regarding Meng and Vaida's criticisms about our proposal of an almost tautological emphasis, we (obviously!) cannot agree: in the paper, we are considering models that can be *fruitfully* regarded as missing data models, that is models for which there is a many to one mapping linking the complete data and the observed data.

Some discussants attempt to provide alternatives that could establish theoretical foundations for DIC. For instance, van der Linde focusses on DIC as an approximate estimated loss, in the same way that BIC is an approximate log Bayes factor, even though she is obviously less critical of DIC in exponential families. She seems to envisage our developments as the result of various approximations. (The sentence *"the purpose of a model [is] independent of the sampling scheme"* remains a mystery to us.) In that perspective, we could wonder what is the whole point of producing such criteria. If the approximation (of a posterior loss?) cannot be evaluated, we should then consider other models in which no approximation is required and then check the appropriateness of each approximation... Further, while using true loss functions is usually sensible (Celeux et al., 2000), it remains to be seen which loss functions correspond to each of the DIC_i's, if any. (In this regard, DIC_2 could be described in a sense as being a more robust version of the basic DIC_1.) This obviously does not relate to the hair(y) loss mentioned by Meng and Vaida!

The very idea of loss function is nonetheless very central to the debate, since DIC appears as a portmanteau substitute for well-defined loss functions. While debating about DIC, we are so far forgetting a central issue, namely what we plan to do with the output of a model comparison exercise. In fact, there is a "dark history" of Bayesian model assessment waiting to be told, in that almost all attempts have stepped outside Bayesian boundaries in order to evaluate the fit of a model. These attempts include that of Robert and Rousseau (2002) and involve p-values that are not strictly Bayesian, or that are not evaluated via a Bayesian perspective. We can therefore truly wonder whether or not it is possible to compare or even to define model complexity within the Bayesian paradigm. At a naïve level, an obvious answer is that we cannot, since we cannot look at a model without standing outside this model. At another level, however, we could answer positively, since tools like Bayes factors and even BIC are already available. But this is not really a less naïve answer! Plummer's alternative is thus interesting in this respect as (a) it does not depend on parameterisation and (b) it is a quantity that can be evaluated a posteriori. Its main drawbacks are that it does not necessarily relate to the original problem, and also that it uses the replica distribution rather than the predictive distribution, which has been advocated in Bayesdom as paramount; see for example van der Linde's discussion or Robert and Rousseau (2002). Also, this only defines a particular type of complexity (or of true dimension) but it does not allow for the comparison of models.

2 Complexity and focus

As noted in Plummer's discussion, an interesting point in Spiegelhalter et al. (2002) is the concept of *focus*. Missing data models clearly give rise to different types of focus, as stressed by both van der Linde and Meng and Vaida (Sections 4 and 5). This feature makes a big difference with ordinary models since possible focusses for missing data models are multifaceted and (much) more numerous than those of standard models, assuming that we do not introduce an articial level of completion!

We thus appreciate the different focusses proposed by Plummer, although they only apply in simple problems: as the hierarchy becomes more and more complex, the number of possible focusses simply explodes. They highlight the complexity of the notion of ... complexity rather than truly solving the problem. Indeed, Plummer's empirical results are rather unhelpful, seeming;y not behaving satisfactorily as K increases. For instance, in Plummer's Figure 1, we could introduce a fourth focus where (μ, τ) would come down at the level of Z, even if this may be a completely artificial represention.

In the case of mixtures, this has the interesting effect of reminding us of the very different nature of p compared with both other parameters. As already stated in Celeux et al. (2000), some natural loss functions for mixture estimation simply omit the parameter p if for instance allocation is taken into account. There is therefore is something delicate and indefinite about p. Note that in Table 2 of Plummer the expected p_D is strikingly close to $2K$ (excluding p then), except for $K = 3, 4$. The last column of Table 3 in Plummer's discussion is also intriguing: p_D and DIC move in such a non-monotonic way that the argument about a simple-and-good-enough model vs. a complex-but-better-fitting model is far from convincing.

To answer van der Linde's question, the complexity of a predictive density is for us the complexity of the underlying model, since the degree(s) of complexity (in the posterior distribution) has been integrated out in the calculation of the predictive. (Think for instance of model averaging which is a *proper* Bayes solution: the weighted sum of predictive densities of different complexities has no well-defined complexity.) We also fail to see how DIC has brought a *"quantification of the reduction of model complexity due to the information in a prior",* although this would suggest using instead Meng and Vaida's posterior version.

A question raised when reading the discussion is whether or not the nuisance parameters in a model are appropriately treated by DIC. In a sense,

this is another type of problem where the definitions of p_D and DIC are unclear, the missing data taking the place of the nuisance parameters. Section 5 of Meng and Vaida's discussion as well as Plummer take alternative positions on that problem, and there are possibly many more others.

3 Plug-in estimates

Without going so far as to agree fully with Dawid's complete dismissal of DIC in his discussion of Spiegelhalter et al. (2002), we concede that using a *plug-in* estimate disqualifies the technique from being properly Bayesian. In the case of mixture models, the problem runs deeper since there is not even a clear-cut estimate without an associated loss function. (This difficulty with DIC is stressed both by Meng and Vaida and by Plummer.) If we want to keep using DIC, it seems that the Bayes estimate of the density is more appropriate for reasons stated in the original paper. If instead we use the predictive then another term should replace the plug-in.

Carlin's suggestion of replacing a plug-in degree of freedom by its posterior distribution is obviously most appealing from a Bayesian point of view, even though the implementation of this principle in a unified methodology may also be *"a few years away"*.

The way Plummer defines p_D is also sensible and the numerical illustrations for the galaxy benchmark dataset are of interest. However, for focus F3, the decrease in DIC for $K \geq 5$ is hard to explain: it could be related to numerical imprecision when deriving its p_D proposal. (We take the opportunity to address here Carlin's last comment about MCMC convergence. While we completely agree that non-identifiable settings are usually welcomed in terms of MCMC convergence, we are rather confident that our sampler has converged within the number of simulations we ran and thus that the exotic behaviour of some DIC_i's is not the result of lack of convergence.)

A puzzling part of Meng and Vaida's discussion is their Section 6, where they happily start mixing even further Bayesian and frequentist tools and objects! The fact that the (more convincing) posterior equivalent of p_D is not working as well is indeed quite intriguing although Plummer somehow gives the hint of an answer in his first paragraph, namely that there are many ways of decomposing a joint distribution into $f(y|\theta)f(\theta)$, just as the number of missing data representations are infinite. First note that using the posterior instead of the likelihood in DIC is nominally Bayesian but

not truly Bayesian as the concept is still frequentist. (The fact that p_D^B is constant in the example is not a difficulty per se: after all this really is a one-parameter problem and it is difficult to look at it otherwise.) There is also the issue that incorporating the prior into the complexity measure confounds the complexity due to the model with the complexity due to the prior and this is very confusing when different models are being compared because we need to use one prior for each model. The final part of Meng and Vaida's Section 6 also makes limited sense (to us at least) because of its systematic alternation between [OR intermingling of?] Bayes and non-Bayes rules and concepts. The only conclusion we could derive from this part is that *ad hoc* criteria can breed even more criteria with seemingly the same validity, which is not necessarily the conclusion expected by the authors...

4 Missing data specifics

For missing data models and in particular for the mixture model, several discussants (Carlin, Meng and Vaida, Plummer) seem to prefer DIC_7 when the focus emphasizes the ability of the model to classify the observed data accurately into groups because, as noted by Carlin, this criterion treats $\mathbf{Z}$ and θ symmetrically. However, a potential default of DIC_7 is that it treats the missing data as parameters. Thus, the number of parameters to be estimated grows to infinity with the sample size for many models including the mixture model. Moreover, it can be remarked that in full Bayesian approaches of the mixture model (see Marin et al., 2004, for a recent survey) the $\mathbf{Z}$ are not treated as parameters (with a prior distribution) but as missing data. In this context, our favorite criterion remains DIC_4 even though this criterion is not invariant to the choice of $\mathbf{Z}$, as noted in the paper and as stressed by Plummer. In our opinion, this problem is essentially formal: when the focus is on imputing values for the missing data, the choice of $\mathbf{Z}$ does not suffer from any ambiguity from a practical point of view.

The point of the last section of Plummer's discussion about the missing or arbitrary function of the data Y was altogether missed by us, although it also replicates a statement in Spiegelhalter et al. (2002). We indeed have trouble in understanding why $f(y, z|\theta)$ is not defined exactly. Is this problem deeper than a mere measure-theoretic subtlety? We would also take issue with the last paragraph of this discussion in that we are not completely convinced that we should use *any* of the DIC's we examined!

5 Conclusion

It seems to us that, if DIC is to 'work' in general then the basic approach, in other words DIC_1 (or arguably DIC_2), should produce satisfactory results, since this is Spiegelhalter *et al.*'s (2002) criterion. In this paper, we have highlighted in some detail the problems in applying DIC beyond the exponential family case. Our goal was not to find a 'cure-all', so that the existence of a generally-applicable measure remains an open question. In other words, the definition of a deviance information criterion, albeit immensely desirable, remains *ad hoc* at this stage and is not even close to being a well-defined ideal criterion or the solution of a well-defined optimisation problem. There is thus a need to reappraise its properties or to start afresh with a new deviance information criterion based on decision theoretic grounds.

On the frequentist and Bayesian approaches to hypothesis testing
by Elías Moreno and F. Javier Girón
A discussion *[published by SORT]*

CHRISTIAN P. ROBERT[1]
CEREMADE, Université Paris Dauphine and CREST, INSEE, Paris

1 Warning

While the authors have made a great job of exposing the advantages of using Bayes factors for hypothesis testing (compared with classical solutions like UMP tests or p-values) and should be congratulated for a new review paper on the objective Bayes approach to testing, let me first take the oratory precaution to state that I will play here devil's advocate by arguing that objective Bayesian theory has not yet reached a satisfactory position with regards to hypothesis testing. (Obviously, I do not consider the p-values as valid answers either, but I will not discuss their shortcomings here since they are already discussed in Robert, 2001, Chapter 6.)

2 Bayes solutions

Even though the authors mention the standard Bayes solution against a classical $0-1$ loss at the beginning of their paper, they evacuate quite forcibly this solution in favour of alternative albeit less well-defined procedures. This is quite unfortunate in that this is the only Bayesian solution to the testing problem if one wants to come up with a "yes/no" answer. (Whether or not this is a good decisional setup is another story, as discussed below, but this is often the type of answers required from statisticians!) Indeed, a Bayes factor, a posterior probability or even a p-value may be used in a decision process but they are intrinsically *not* decision-theoretic procedures. This is also why I regret the early dismissal of the likelihood ratio: when both hypotheses are simple, the likelihood ratio (or rather the comparison of the likelihood ratio

[1]Discussion written during a visit to the Department of Statistics, University of Florida, Gainesville. The author is grateful to the organisers of the 8th Winter Statistics Conference and in particular to George Casella and to Jim Hobert for their hospitality.

with a certain bound c) is a Bayes rule, no matter the value of c. When at least one hypothesis is composite, the likelihood ratio is no longer Bayes, but if we follow the asymptotic arguments of the authors, it can be taken as a first order approximation in some situations.

Why, thus, is there a problem with the Bayes solution, especially when considering it is *consistent*? The main point is that consistency is a very weak criterion, because a Bayes rule can choose to accept the null hypothesis when the likelihood ratio (or the ratio of marginals) is above about any bound, depending on the prior model. Is this bad or wrong? Formally speaking, this is completely acceptable as the range of Bayes rules usually covers the range of (admissible) possible answers. This only starts looking bad when one seeks a universal or common answer or, in other words, an "objective" procedure. Always from a Bayesian point of view, this universal perspective can be disputed as non-Bayesian, because of excluding all prior inputs.

I am afraid that it may be the case that the inconsistencies discussed below are simply beacons that signal the impossibility of Bayesian non-informative or objective procedures. (The fact that these procedures exist for estimation problems simply is a reflection on the smoother topological structure of estimation setups.) Or, more radicaly, it may be argued that the whole approach to testing as a $\{0, 1\}$ problem is flawed in that the true consequences of a decision are not properly modeled. (Hence the common confusion bewtween scientific hypothesis testing—as in *is the age of the Universe larger than* 13 *billion years?*—and model choice—as in *how close to the family of probit models is the true model?*)

It indeed seems to me that model choice calls for a completely different decisional approach that integrates both the complexity of the models under comparison and the consequences of choosing one model rather than another. Since the authors are not adopting this perspective in variable selection, they have to ressort to cascades of Bayesian tests without properly evaluating the consequences of this repetition of tests. There is for instance no uncertainty evaluation on the ranking and the selection of the most likely model. Nor is the sequence of decisions evaluated sequentialy or conditionaly. Model choice is in fact much more an estimation problem for the difference between the true model and an hypothetic collection of models than a testing model. Methods based on function divergences (Robert, 2001, Chapter 7) should thus be prefered as they ascertain the different consequences of picking the "wrong" model, even though complexity summaries like AIC, BIC or DIC are standing far away from the Bayesian paradigm. I tend to agree with the authors that a

Bayesian approach is more likely to account automatically for the complexity of a model than frequentist and likelihood perspectives but I also think that a more thorough assessment of the consequences of model choice should be undertaken, rather than trusting blindly a dimension-free Bayes factor. In this regard, the recent paper by Yong (2005) is quite illuminating in that it exposes the conflict between model selection and parameter estimation, establishing in particular that model averaging (Raftery, 1996) cannot reach optimal convergence rates.

Note at last that under a completely different decision-theoretic perspective, namely when losses of the type $(\delta - \mathbb{I}_{H_0}(\theta))^2$ are used, the posterior probabilities are themselves Bayes rules (Robert, 2001) and that there exist cases where p-values are admissible as truncations of Bayes rules (Hwang et al., 1992). Although these are rather formal results, I think they are still worth mentioning.

3 Inconsistencies

If we now turn to the alternative solutions provided in the paper, there is a lot to be said against Bayes factors, pseudo-Bayes factors and intrinsic priors.

First, as stated above, the Bayes factor is not even on the same scale as the Bayes solution given that it is dimension free. The authors often switch back and forth between Bayes factors and posterior probabilities. But doing this implies the choice of a particular prior ratio $\pi(H_0)/\pi(H_1) = 1$, a choice that is never discussed in the paper, while being paramount for the comparison with the p-values. In particular, there is no clear reason why $\pi(H_0) = 1/2$ should be considered as a "non-informative" solution (Robert, 1993). In some instances, intrinsic priors are associated with unbalanced prior weights $\pi(H_0)$ and $\pi(H_1)$ for example. If the complexity of the model under hypothesis H_0 is much higher than under hypothesis H_1, the prior weight of H_0 could be lowered as an consequence of Occam's razor rule. (A personnal aside: I never really understood the need to call for this rule. In fact, while being interesting from an epistemiological point of view, Occam's key sentence *Pluralitas non est ponenda sine neccesitate* does not constitute an operational principle and about anything can be justified on this vague sentence.)

If we now turn to pseudo-Bayes factors, they seem to cumulate the shortcomings of Bayes factors and of pseudos! While providing a workable solution to the impossibility of using improper σ-finite measures under both hypothe-

ses, they are suffering from a high level of adhocquery that is reflected by the myriad of versions found in the literature (as discussed in Robert, 2001, Chapter 6). Pseudo-Bayes are clever mathematical constructs but they do not enjoy the same justifications as true Bayes factors. While I do not think that the (minor) criticism by the authors that "the arithmetic intrinsic Bayes factor (...) reuses the sample observations" is particularly true [many genuine Bayes procedures appear as weighted sums of averages on part of the data, think for instance of mixtures of distributions], both the lack of symmetry between H_0 and H_1 and the possible difficulty in defining acceptable subsamples and minimal sample sizes maintain the pseudo-Bayes factors at a considerable distance from genuine Bayesian inference.

The construction of the intrinsic prior proposed in the review is clever and reminiscent of Pérez and Berger (2002). Note that this prior can also be written as

$$\pi_1^I(\theta_1) = \int \pi^I(\theta_1|\theta_0)\pi_0^N(\mathrm{d}\theta_0) = \int \mathbb{E}_{\theta_0}[\pi_1^N(\theta_1|X)]\pi_0^N(\mathrm{d}\theta_0)\,,$$

a representation which somehow is more intuitive, apparently works for nonnested models, but also exposes the limitation of the device. Indeed, once θ_0 is integrated out in the above equation, we are faced with a new improper prior and the mathematical result that

$$B_{10}^I(\mathbf{x}) = \frac{\int f(\mathbf{x}|\theta_1)\pi^I(\mathrm{d}\theta_1|\theta_0)\pi_0^N(\mathrm{d}\theta_0)}{\int f(\mathbf{x}|\theta_0)\pi_0^N(\mathrm{d}\theta_0)}$$

does not depend on the normalising constant of π_0^N does not translate so easily to the ratio

$$B_{10}^I(\mathbf{x}) = \frac{\int f(\mathbf{x}|\theta_1)\pi_1^I(\mathrm{d}\theta_1)}{\int f(\mathbf{x}|\theta_0)\pi_0^N(\mathrm{d}\theta_0)}\,,$$

although they are mathematicaly the same. In other words, were we given π_1^I and π_0^N separately as in a regular improper prior hypothesis testing we would not know how to normalise both priors. (This is a dilemna common to missing data problems: while the density of the observables can be written as a marginal of another distribution, the dummy variables used in the marginalisation have no specific meaning.)

We also note that the proposed solution is not symmetric in (H_0, H_1), which is a drawback shared by most pseudo-Bayes procedures. In addition, when more than two hypotheses are in competition, this approach requires

either a change of priors for each pair of hypotheses or the subjective choice of a *reference* hypothesis H_0 under which all other intrinsic priors are constructed.

The authors mention several times that having no moment is a "nice property to be expected from an objective prior". This is rather peculiar a remark since the lack of moments must depend on the parameterisation of the model: if we use a bounded parameterisation, the corresponding parameter will have finite moments. To continue about puzzling remarks, I do not understand the point about the one-to-one relationship between p-values and posterior probabilities: both quantities depend on the same (multinomial) sufficient statistic but since there is no one-to-one relationship between the p-value and the sufficient statistic, this does not seem possible.

4 Conclusions and perspectives

This paper provides a (partial) summary of the large literature on the comparison between frequentist (restricted to p-values) and Bayesian (restricted to Bayes factors), but it may constitutes too restricted a vision of the challenges met by both approaches in both hypothesis testing and model comparison. To find that p-values do not provide the same numerical answer than posterior probabilities or Bayes factors is not a fundamental difficulty in that they are not to be treated as similar objects.

On the one hand, I definitely acknowledge the urgent need for objective Bayes procedures in testing problems and do not want to disparage the past work led by the authors and others on this topic. On the other hand, there currently exist much more pressing challenges in hypothesis testing, for instance with the emergence of massively multiple tests in Genomics (Genovese and Wasserman, 2002) and in other fields. Even in the case of model selection, the complexity (or the combinatorics) of the decision space often prevents a perfect exploration and new tools are necessary to ascertain whether or not important models and situations are not forgotten. About ten years ago (Madigan and Raftery, 1994), *model averaging* appeared as a potentially rich tool for handling multiple models simultaneously. While this is not a panacea, in that it does not directly allow for pruning in a large collection of models and may lead to subefficiencies (Yong, 2005), this direction should not either be abandoned altogether.

References

L.D. Brown, T.T. Cai, and A. DasGupta. Confidence intervals for a binomial proportion and Edgeworth expansion. *Annals Statist.*, 30(1):160–201, 2002.

G. Casella, J.T. Hwang, and C.P. Robert. A paradox in decision-theoretic set estimation. *Statist. Sinica*, 3:141–155, 1993a.

G. Casella, J.T. Hwang, and C.P. Robert. Loss function for set estimation. In J.O. Berger and S.S. Gupta, editors, *Stat. Decision Theo. Rel. Topics V*, pages 237–252. Springer Verlag, New York, 1993b.

G. Celeux, M.A. Hurn, and C.P. Robert. Computational and inferential difficulties with mixtures posterior distribution. *J. American Statist. Assoc.*, 95(3):957–979, 2000.

G.S. Datta and R. Mukerjee. *Probability Matching Priors: Higher Order Asymptotics*, volume 178 of *Lecture Notes in Statistics*. Springer Verlag, New York, 2004.

C. Genovese and L. Wasserman. Bayesian and frequentist multiple testing. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 7*, page to appear. Oxford University Press, 2002.

C.R. Hwang, G. Casella, C.P. Robert, M.T. Wells, and R. Farrel. Estimation of accuracy in testing. *Ann. Statist.*, 20:490–509, 1992.

D. Madigan and A.E. Raftery. Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. American Statist. Assoc.*, 89:1535–1546, 1994.

J.M. Marin, K.L. Mengersen, and C.P. Robert. Bayesian modelling and inference on mixtures of distributions. In C.R. Rao and D. Dey, editors, *Handbook of Statistics*, volume 25 (to appear). Springer-Verlag, New York, 2004.

J. M. Pérez and J. Berger. Expected posterior prior distributions for model selection. *Biometrika*, 89:491–512, 2002.

A.E. Raftery. Hypothesis testing and model selection. In W.R. Gilks, D.J. Spiegelhalter, and S. Richardson, editors, *Markov Chain Monte Carlo in Practice*, pages 163–188. Chapman and Hall, New York, London, 1996.

C. P. Robert and J. Rousseau. A mixture approach to Bayesian goodness of fit. Technical Report 2002-9, Université Paris Dauphine, 2002.

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

C.P. Robert. *The Bayesian Choice*. Springer-Verlag, New York, second edition, 2001.

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

J. Rousseau. Expansions of penalized likelihood ratio statistics and consequences on matching priors for hpd regions. Technical report, CREST, INSEE, Paris, 1997.

D. J. Spiegelhalter, N.G. Best, Carlin B.P., and A. van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, 64:583–640, 2002.

Y. Yong. Can the strengths of AIC and BIC be shared? *Biometrika*, 92(4): 937–950, 2005.