dependent prior that integrates to less than one. Also, if one criterion is always smaller than the other, then it is clear that none of them can be better than the other in all circumstances. It is easy to provide numerical evidence of this sort in Example 6.

Finally, in Example 6, the advantage that the AIBF has over the FBF is bought at a price. For, being exactly equal to a BF with respect to a prior may be preferable to being so only asymptotically.

We too have our own preferences, namely, the AIBF, a trimmed AIBF and the median IBF but believe it's too early to come to any definite conclusion.

Example 1 (Section 4.1) illustrates the difficulties with the class of models with "improper likelihoods" such as the mixture models for which the IBF and FBF cannot be directly employed. It refers to an unpublished work of Shui (1996) that considers modifications of the IBF and the FBF approaches to deal with the mixture models. Hopefully, the interesting work of Shui would be published soon.

5. *Teaching Nonsubjective Bayes Testing.* How should one teach nonsubjective Bayes testing in an undergraduate course? What would be the best way to communicate these ideas to the students who cannot be expected to understand all the subtleties of an IBF? At least in the classical examples ($N(\mu, 1)$ or $N(\mu, \sigma^2)$) it may be easier to motivate and use a BF based on a default or intrinsic prior but one would still have to motivate the prior. Do Berger and Pericchi have any suggestions?

Fulvio De Santis
Università di Roma, "La Sapienza"

## 1. Introduction.

Model selection and hypothesis testing are difficult topics. In these problems, as we depart from the usual assumption of explicitly stated underlying model, fundamental statistical principles (e.g., likelihood, sufficiency, etc.) begin to fade and we are left with no clear direction. Substantial debate over the appropriate model selection and hypothesis testing problems has taken place inside and outside the Bayesian community. Within the Bayesian approach, controversies arise on how model selection should be performed, even in the *ideal* situation where prior information is available. For example, the recent renewed interest in the development of *default* model selection methods has

[0]Fulvio De Santis is Assistant Professor, Dipartimento di Statistica, Probabilità e Statistiche Applicate , Università di Roma, "La Sapienza", P.le A. Moro, 5, 00185 - Roma, Italy; email: fulvio.desantis@uniroma1.it

witnessed many conflicts: not only is the statistical problem hard, but some statisticians make it even harder by the controversial use of improper priors! The objective Bayesian approach is of great theoretical and applied importance, not only for its connections to non-Bayesian analysis, but because it has been able to produce original, useful and sensible statistical tools. Furthermore, until a few years ago, little had been done for default model selection, compared to standard estimation. Nevertheless, in the last few years, important advances have been made towards developing useful strategies for objective model selection, as the present paper by Berger and Pericchi (B&P, from now on) demonstrates. This paper is an important piece of work for at least three reasons. First, it summarises the extensive experience and important contributions of the two authors and their collaborators. Furthermore, the paper critiques most of the methods currently available for default model selection: formal properties, behaviour in important classes of problems, compatibility with subjective Bayesian methods, difficulties in implementation and computation. Third, and most importantly, the authors provide a fair comparison of different model selection methods. This allows them to recognize limits of the IBF approach and to acknowledge merits of competitive methods.

The goal of this discussion is to outline possible strategies for selecting among default Bayes factors (DBFs) or for comparing such methods. Next sections propose a possible complement to the authors' presentation, not an alternative.

This note consists of two sections. Section 2 is devoted to two possible ways for comparing DBFs. Namely, the discussion focuses on finite-sample properties of DBFs (Section 2.1) and on the use of a frequentist pre-experimental analysis to perform a *neutral* comparison between competing methods (Section 2.2). Section 3 considers the more specific issue of the selection of the fraction(s) in the FBF approach.

## 2. Strategies for comparing and selecting default Bayes factors.

The rich literature on DBFs has given a lot of attention to comparative issues, and the paper of B&P reviews the main approaches to this goal. Comparative analyses of DBFs have focused, in general, on two aspects: *a*) coherence of the methods, mainly thought of as the ability of DBFs to satisfy some typical properties of *true* BFs; and *b*) asymptotic correspondence to real BFs (intrinsic prior theory). The next two subsections propose to widen the comparison of DBFs from two different points of view.

### 2.1. Ordinary and default Bayes factors: a finite-sample analysis of compatibility:

According to B&P (see Section 3), DBFs must be judged in the light of their correspondence to actual Bayes factors. Since, in most of the cases, correspondence cannot be established for finite sample sizes, the authors argue that such a correspondence might

be established asymptotically. This consideration motivates the intrinsic prior methodology. However, in a finite sample setting, it can be of interest to evaluate how "far" a DBF is from a real BF. For a given set of data, we can evaluate the compatibility of a DBF with an ordinary BF computed with a proper prior. This is of particular interest in the presence of weak prior information. Specifically, suppose that we want to compare a fully specified model $f_1(\cdot)$ with a second model $f_2(\cdot|\theta_2)$, with unknown $\theta_2$. Let us assume that, in the presence of partial prior information, we succeed in eliciting a class $\Gamma$ of priors for $\theta_2$, but we are not able to select any specific prior in this set. (Note that, if prior information is totally lacking, $\Gamma$ is the class of all the distributions for $\theta_2$). In this context the most natural approach is to look at the range of the standard BF over $\Gamma$ but, as it is often the case, such a range might be so large not to lead to decisive evidence in favour of either one of the two models. This is a context in which DBF methodology can be of some help, even in the presence of partial prior information.

A first way to resort to a DBF, $B_{21}^D$, is to look at its range over $\Gamma$, rather than considering the range of the standard BF. Note that we are suggesting to compute $B_{21}^D$, originally proposed to be used with improper priors, using the proper priors in $\Gamma$. Often the range of $B_{21}^D$ is more informative than the range of $B_{21}$, as it would be the case in the example under consideration, if $\Gamma$ contained flat priors for an unbounded parameter space for $\theta_2$. This fact was extensively discussed in De Santis and Spezzaferri (1997), among others. See Liseo (2000) for a recent discussion on robustness isssues in Bayesian model selection.

A second approach is the following: Given the class $\Gamma$ of priors for $\theta_2$, we can decide to use a DBF with a non-informative prior, $\pi^N$. One may study the compatibility of the DBF with the class $\Gamma$ in order to determine if the method is *sensible*. Given a set of data $\mathbf{x}_n$ of size $n$, we say that $B_{21}^D$ is $\Gamma-$compatible if there is at least one prior $\pi^*$ in $\Gamma$ such that $B_{21}^D$ equals the BF computed with $\pi^*$. Hence, using either the standard BF with $\pi^*$ or the DBF with $\pi^N$, we would anyhow be using a true Bayesian method. Of course, in most of the cases such a prior $\pi^*$ does not exist. It is, however, interesting to establish how far we are from the class $\Gamma$, when we use $B_{21}^D$. As noted above, the difficulty in determining a $\pi^*$ in $\Gamma$ for finite $n$ is the motivation for looking at intrinsic priors. In this case $\Gamma$ is the class of all the distributions and the compatibility between $B_{21}^D$ and $\Gamma$ is established only approximatively.

To illustrate how the approach outlined above can be used for comparing alternative DBFs, let us get back to our finite sample set-up and let $d[B_{21}^D(\pi^N), B_{21}(\pi)]$ be some sort of *distance* between the DBF and the true BF for $\pi \in \Gamma$ (here we stress the dependence of BF and DBF on the respective priors). Also, let $\bar{B}_{21}^D$ be an alternative DBFs for the

same problem. Then, for a given data set $\mathbf{x}_n$, we prefer $B_{21}^D$ to $\bar{B}_{21}^D$ if

$$\inf_{\pi\in\Gamma} d[B_{21}^D(\pi^N), B_{21}(\pi)] < \inf_{\pi\in\Gamma} d[\bar{B}_{21}^D(\pi^N), B_{21}(\pi)].$$

If $\inf_{\pi\in\Gamma} d[B_{21}^D(\pi^N), B_{21}(\pi)] = 0$, $B_{21}^D$ is $\Gamma$-compatible; otherwise it is to be preferred to the alternative DBF since it is closer to the true BF.

For example, consider the simple testing problem of Illustration 3 in the paper. In this problem, the point null hypothesis, $\theta = 0$, for a normal mean with variance equal to one (model $M_1$) is tested against a two-sided alternative (model $M_2$). Consider for the prior under the alternative the standard class of conjugate priors, $\Gamma_{Con}$, with mean zero and variance $\tau^2 \in I\!R^+$. Suppose we are interested in comparing 4 typical DBFs: fractional BF (with $b = 1/n$), expected arithmetic IBF, BIC and posterior BF (POBF, Aitkin, 1991). For example, suppose that $n = 10$ and $\sqrt{n}\bar{x} = 1.96$ (corresponding to the classical .05 $p$-value). It is easy to check that in this case the values of the 4 DBFs are 1.78, 1.53, 2.15 and 4.83 respectively, but the range of the standard BF in the conjugate class is $(0, 2.11)$ Therefore, with respect to the observed data, FBF and *expected* IBF are both $\Gamma_{Con}$−compatible, while BIC and POBF are not. However, being BIC "closer" to the possible values of BF in $\Gamma_{Con}$, it is preferable to POBF.

Of course, compatibility of a DBF with a class of priors is not necessarily guaranteed over the sample space and for any given sample size. In the example under consideration, it is easy to check that FBF is uniformly $\Gamma_{Con}$−compatible, regardless of $n$, but the remaining DBFs (*expected* IBF, BIC and POBF) are $\Gamma$−compatible only if the sample mean is less than $\sqrt{2}e^{-1/2}$, $e^{-1/2}$ and $\sqrt{2/n}e^{-1/2}$, respectively.

Let us extend this idea. In the pre-experimental set-up, it might be of some interest to look at the probability of obtaining a DBF that is $\Gamma$−compatible. In the previous simple example, let us focus on *expected* IBF and BIC. It is easy to check that, under the null hypothesis, the probabilities that such DBFs are $\Gamma_{Con}$− compatible are $\Phi(\sqrt{2n}e^{-1/2})$ and $\Phi(\sqrt{n}e^{-1/2})$, respectively, where $\Phi(\cdot)$ is the c.d.f. of a standard normal. Therefore, in this case, *expected* IBF is uniformly more likely to be $\Gamma_{Con}$−compatible than BIC, under the null. This fact implies that, had we to choose the sample size in order to be guaranteed to have a $\Gamma_{Con}$−compatible DBF at a given probability level, less data are needed for *expected* IBF than for BIC.

This analysis is admittedly crude since the frequentist pre-experimental behaviour of DBFs should also be studied under the alternative. However, it might give an idea of how to bridge the asymptotic analysis of compatibility between DBFs and real BFs, represented by the intrinsic prior theory, to the finite-sample necessity of evaluating DBFs, in the presence of partial prior information

**2.2. A neutral comparison of default Bayes factors in the presence of proper priors: a frequentist analysis:**

As mentioned by B&P (Section 5.5) and also discussed in the previous section, DBFs can be used with proper priors as robust methods. It is intuitively easy to understand that gain in robustness of the prior results in loss of discriminatory power (d.p., in the following) of the model choice criteria. Of course, between two fairly robust methods, the one whose loss in d.p. is less should be preferred. Loss in d.p. of a BF can be quantified by extending ideas given in Verdinelli (1996) and Royall (1997). Following Verdinelli (1996), we say that a BF (ordinary or default) is *decisive* if it provides clear evidence in favor of $M_1$ or $M_2$. Given some data and chosen a suitable threshold $k > 1$, a BF is decisive if it is greater than $k$ or smaller than $1/k$. However, discriminating ability of BFs must be established before the experiment is performed. Hence, the idea is to evaluate the frequentist probabilities that such criteria are decisive. Two alternative DBFs can then be compared as follows: if, in order to achieve a certain probability of being decisive, a DBF requires less data than another, the former has a greater d.p. than the latter. In the standard test of a normal mean, already considered in the previous section, assuming equal probabilities for the two hypotheses and a prior variance $\tau^2 = 1.5$, the mimimal sample sizes required to have decisive ordinary, fractional and *expected* intrinsic BFs are 11, 27 and 18 respectively when $k = 3$. Therefore, even though both FBF and *expected* IBF's d.p. are less than ordinary BF's d.p. as expected, FBF seems to be more conservative than *expected* IBF as a choice criterion. Of course, that the choice of $k$ is crucial and calibration of thresholds for DBFs deverves investigation.

In principle, the above analysis might be extended to the comparison of DBFs defined with improper priors. However, problems arise in the computation of the frequentist probabilities since these require the use of the marginals of the data that are not defined when improper priors are used. This problem is considered in De Santis (2000).

**3. On the choice of the fraction(s) for FBFs.**

B&P clearly point out that in the FBF approach the choice of the fraction $b$ is crucial. They show that in the Neyman and Scott testing problem, illustrated in Section 4.4, the basic definition of FBF must be extended in order to achieve consistency. A similar problem has been pointed out by Iwaki (1997). In the specific context of linear models, even though the same can be proved in more general set-ups, whenever the data are not exchangeable, De Santis and Spezzaferri (1999) show how the use of a unique fraction for the likelihoods in the FBF's correction factor might lead to inconsistencies. They also propose a constructive method to derive a *multiple-fractions*, consistent FBF. In my opinion, FBF as well as IBFs can simply be seen as the result of a suitable combination of

partial Bayes factors: in the FBF, a geometric mean of the likelihoods for all the different training samples is performed, while, in the IBF approach, suitable averages of the entire correction terms are computed. In this way, at least in relevant problems, the selection of $b$, or of multiple fractions $b_i$'s, is automatically made once partial BFs are defined. In fact, the choice of the fraction(s) is the automatic result of the likelihood-combination process. From this perspective, the computation fo the fraction(s) is, at most, as hard as it is the determination of the terms to average in the IBF approach.

A further approach is to simply regard $b$ (let us consider now, for simplicity, the simple case of a unique fraction) as a constant in $(0,1)$, not necessarily related to the size of the training sample. It has been often noted that the choice of $b$ has an effect on both the sensitivity to the priors and on the d.p. of the criterion (Gilks 1995, O'Hagan 1995). Conigliani and O'Hagan (2000) study the effect of the choice of $b$ on the sensitivity of the standard FBF to both proper and improper priors. They conclude that, on the grounds of sensitivity to the prior, the choice $b = \ell_0/n$, where $\ell_0$ is the minimal training sample size, is often appropriate, but not necessarily the unique. The authors point out correctly that, in addition to the effect on the sensitivity to the priors, the effect on the d.p. must be taken into account in the choice of $b$. This last aspect is however hard to quantify. A possible, natural way to pursue this is again based on a pre-experimental analysis of FBF. We can look at the choice of $b$ as a design problem, and dicriminatory power of the FBF can then be quantified by the pre-experimental probability of observing decisive evidence in favour of $M_1$ or $M_2$. The idea is to determine the probability of having decisive evidence (i.e. strong discriminatory power) as a function of $b$ to be used in order to select, before performing the experiment, the optimal fraction. As a simple example, suppose again that, under $M_1$, $X \sim N(0,1)$, and, under $M_2$, $X \sim N(\theta,1)$. In this case, if $\pi_2^N(\theta) \propto 1$, the resulting FBF corresponds to a Bayes factor obtained using, as a prior under $M_2$, a $N(0,(b-1)/bn)$ density, with $b \in (0,1)$. Therefore, noting that, marginally, $\bar{x} \sim N(0,1/n)$ under $M_1$ while, under $M_2$, $\bar{x} \sim N(0,(1-b)/bn)$, computation of the probability of observing a decisive FBF, as a function of $b$, is straightforward. Table 1 shows the probabilities of obtaining decisive evidence with FBF, when $k = 3$ and equal prior probabilities for the two hypotheses are assumed, for some values of $n$ and $b$.

Table 1. *Probabilities of Decisive FBF*

| $n$ | 10 | 20 | 30 | 50 | 100 | 200 | 500 |
|---|---|---|---|---|---|---|---|
| $b = 1/n$ | 0.429 | 0.712 | 0.787 | 0.849 | 0.904 | 0.937 | 0.964 |
| $b = 2/n$ | 0.179 | 0.429 | 0.634 | 0.757 | 0.849 | 0.904 | 0.945 |

It is clear that reduction in discriminatory power of FBF depends strongly on the sample size: it is substantial for small sample sizes ($n = 10, 20$) but less and less influent as the

sample size increases.

Beyond the above stardard oversimplistic example, such an analysis might be the starting point to develop an objective quantitative measure of discriminatory power of the FBF, as a function of $b$. This measure could be combined with measures of sensitivity of the FBF to the prior, such the ones proposed in Conigliani and O'Hagan (2000), in a unifying tool to be used to choose $b$.

Two final comments are in order. First, in principle the above analysis can be also performed in the presence of *multiple fractions* FBF. Secondly, and more importantly, as noted above computation of the probabilities to be used to set the fraction(s) requires the knowledge of the marginal distributions of the data under the two models, and this is, in general, much more complicated than it is in this problem. The use of fractional priors might be, at least in some cases, of help (De Santis, 2000).

## ADDITIONAL REFERENCES

Conigliani, C. and O'Hagan, A. (2000). Sensitivity measures of the fractional Bayes factor to prior distributions. *Canad. J. Statist.* **28**.

De Santis, F. (2000). Statistical evidence and sample size for robust and default Bayes testing. Technical Report, Univ. of Rome, "La Sapienza".

Gilks, W.R. (1995). Discussion of O'Hagan. *J. Roy. Statist. Soc. Ser. B* 57 118-120.

Liseo, B. (2000). Robustness issues in Bayesian model selection. In: *Robust Bayesian Analysis.* Lectures Notes in Statistics, 152, 197-222. Springer-Verlag.

Royall, M.R. (1997). *Statistical Evidence: A Likelihood Paradigm.* Chapman and Hall, London.

Verdinelli, I. (1996). Bayesian designs of experiments for the linear model. PhD dissertation, Dept. of Statistics, Carnegie Mellon Univ.

# REJOINDER

J. O. Berger and L. R. Pericchi

We thank the discussants for their very interesting comments and viewpoints. We respond to each in turn, using the numbering scheme of the discussants. If we do not mention a section of a discussion, it is because we appreciate and agree with the points mentioned therein.