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.

Reply to Professor De Santis:

2.1. De Santis rightly observes that it would be nice to have formal ways of evaluating objective Bayes factors in small sample settings. To date, our efforts in this direction have been limited to determining the intrinsic prior (based on asymptotics) and then investigating, for small samples, the extent to which the objective Bayes factor is close to the Bayes factor from the intrinsic prior. If the two Bayes factors are close, one can rest easy. But if they differ, one is not sure what to conclude.

De Santis investigates this issue within the context of partial prior information, as specified by a class Γ of prior distributions. He considers several possible ways of measuring compatibility of objective Bayes factors with the information in Γ.

A variant of this idea that is in tune with our strategy for development of objective Bayes factors is to directly utilize the partial prior information in the construction of the Bayes factor. One natural approach is to first calculate the reference prior, subject to the restriction of being in the class Γ (cf. Sun and Berger 1998). If the reference prior is proper, it can be immediately used to calculate the BF. Otherwise, one could use the constrained reference prior to compute a default Bayes factor (e.g., an IBF). Of course, if the prior is unconstrained, this reduces to the ordinary definition of an objective Bayes factor.

2.2. As argued in the chapter, our recommendation for evaluating an objective Bayes factor method is simply: discover which prior is effectively being used when applying the method, and informally judge whether or not this prior is reasonable. Our experience is that, by looking at this intrinsic prior, one can gain a great deal of insight into possible biases or inadequacies of the corresponding objective Bayes factor method.

In contrast, comparison of operating characteristics of objective Bayes factors rarely seems to yield clear insights. The problem is that the results are, of necessity, highly dependent on the particular operating characteristic that one considers. And even then, one objective Bayes factor will rarely uniformly dominate another. After all, virtually any Bayesian procedure in testing is formally admissible from a frequentist perspective, meaning that uniform domination cannot be attained.

Consider, for instance, the use of 'discriminatory power' as discussed by De Santis. For the particular situation he considered, the IBF happened to have higher discriminatory power than the FBF. But both the IBF and the FBF in this example have proper intrinsic priors, so that presumably a different choice of the 'design' prior for θ (the prior under which discriminatory power is computed) or a different choice of k (or allowing different k for choice of different models) could easily reverse this finding. Another issue here is that we feel the 'design' prior must be fixed in carrying out pre-

experimental comparisons and, in particular, should not equal the intrinsic prior for a procedure (as is apparently done in section 3 of the discussion). In other words, one must fix the pre-experimental 'truth' and then judge various procedures against this truth, rather than allowing the truth to shift with the procedure.

Of course, part of the message of De Santis is that one should formally consider experimental design, with the goal of ensuring that the discriminatory power of the procedure to be used is adequate, and we completely agree. It is just that we do not feel that generic comparison of objective Bayes factors can easily be carried out in this way.

3. De Santis reaffirms the need for modifying the FBF to allow for multiple fractions. Indeed, De Santis and Spezzaferri (1999) propose a quite compelling method for determining the multiple fractions. We have two comments. First, no method can overcome the difficulties of FBFs in irregular models, such as our Example 2. Second, when observations can be dependent, it is something of a misnomer to call the method an FBF method, since it produces a prior that cannot then be written in terms of fractions of multiplicative parts of the full likelihood. Indeed, their method is more closely related to what is known as use of the *empirical expected posterior prior* (use of (5.2) with m^* chosen to be the empirical distribution of minimal training samples); (5.2) can then be viewed as the arithmetic average of minimal training sample posteriors, while the approach of De Santis and Spezzaferri leads to a geometric average of these training sample posteriors. These connections are all very interesting and affirm the basic point made by De Santis that all these objective Bayes factors are based on much the same principle.

Reply to Professors Ghosh and Samanta:

1. We agree with Ghosh and Samanta that the direct intuitive appeal of certain of the objective Bayes factors can actually lend support to use of the corresponding intrinsic prior. It is indeed useful to think of the justification as a two-way street. We also agree that, in situations in which the number of parameters is allowed to grow with the sample size, the existing theory of Intrinsic Priors need not apply (although it can sometimes be directly modified in an appropriate fashion, as was done in our Example 4).

Ghosh and Samanta raise the interesting issue of propriety of the conditional Intrinsic Distribution $\pi_2^I(\eta|\psi)$. We actually seek much more than just propriety of this distribution; we also want the (typically improper) marginal intrinsic priors for the nuisance parameters under the two models to be properly 'calibrated.' Dass (2000) shows that this can be done in problems with a suitable group structure, as long as the initial noninformative priors that are used to derive the IBF are the right Haar priors.

2. We agree with the observation that increasing the size of the training sample will imply more peaked Intrinsic Priors. This is of particular relevance because of the next comments of the discussants.

3. The discussion of the 'Scale of the Priors' is fascinating. It may, indeed, frequently be the case that, in well-designed experiments, the pre-experimentally chosen sample size, n_0, is such that 'local alternatives' like $\theta = \delta/\sqrt{n_0}$ are those that are apriori viewed to be likely, and objective Bayes factors would need to adjust to this scale. The various technical mechanisms discussed by Ghosh and Samanta for achieving this adjustment (such as increasing the training sample size in IBFs) are quite clever.

There remains, however, the outstanding practical issue of determining when a scale adjustment is necessary. One possibility - seemingly that envisaged by Ghosh and Samanta - is to subjectively elicit the appropriate scale, and then embed this scale in a suitable default procedure. This is entirely reasonable, but does require some subjective thinking.

One might, of course, begin by computing the answers arising from both a 'local alternative' scale and the usual scale; it is only if these yield contrasting conclusions that one would need to make a subjective decision as to which scale is most appropriate. By the 'local alternative' scale here we effectively mean that which would arise as a lower bound from a robust Bayesian analysis with respect to a reasonable class of priors. Unfortunately, it is not clear how one can automatically find this scale through use of modifications of IBF type procedures.

This discussion is also related to the idea of 'local' vs 'global' alternatives in Smith and Spiegelhalter (1980). Also of interest, from that paper, is the observation that use of local alternatives can lead to AIC type approximations to Bayes factors; this is related to the observation of Ghosh and Samanta that local scales can bring Bayesian and frequentist answers closer together. While this is true asymptotically, it should be pointed out that, for moderate sample sizes, a significant discrepancy remains between, say, p-values and Bayes factors (at any scale).

4. *Examples:* As pointed out by the discussants, one always has to consider comparison-by-example with caution. The main point of the examples in the chapter was to indicate the types of things that could go wrong with default procedures (so that one could be properly cautious in their use) rather than to try to 'prove' which procedures were better. Our surprise that the IBF seemed to automatically overcome all obstacles probably led us to emphasize the comparison aspect a bit too much.

Concerning the FBF, we should mention that we have always thought of the FBF as a multitude of procedures, especially when multiple fractions are allowed. Thus we

introduced multiple factions in examples (such as Example 4) where it seemed to be necessary.

Our purpose in Example 6 *When Neither Model is True* was apparently not stated very clearly. We certainly did not mean to suggest that the GIBF is better than the FBF because it is smaller. In the example we were, instead, reacting to the comment in O'Hagan (1997) to the effect that the appearance of the sample variance in the GIBF leads to *"...not intuitively reasonable behaviour"*. We were simply suggesting that the situation is far from clear, and that the appearance of the sample variance in the GIBF can be motivated through consideration of robustness to the assumption $\sigma^2 = 1$. The effect, on the FBF, of violation of this assumption is quite serious, while the GIBF seems to compensate rather well to its violation. We also wanted to present the example to point out that IBFs may well have advantages over, say, the use of the corresponding intrinsic prior, when it is suspected that none of the models under consideration may be true.

5. *Teaching Non-Subjective Bayes Testing.* We have certainly been thinking about possible ways to teach this material. Part of the problem is that one might choose to emphasize quite different methods for different audiences. In a Bayesian course emphasizing MCMC, it would be natural to focus on objective testing and model comparison based on use of expected posterior priors (often equivalent to intrinsic priors from AIBFs), since they can usually be directly incorporated into MCMC schemes. For a low level undergraduate service course, one might settle for simply teaching students to calibrate p-values via the $BF = -ep\log(p)$ formula of Sellke, Bayarri and Berger (2001). At a higher level undergraduate course, one might emphasize the idea of training samples and present the median IBF as a general purpose testing and model selection tool. There is surely also a role for approximations, such as BIC and its possible generalizations or reformulations.

ADDITIONAL REFERENCES

Dass, S.C. (2000). Propriety of intrinsic priors in invariant testing situations. Technical Report, Michigan State Univ.

Sun, D. and Berger, J. (1998). Reference priors with partial information. *Biometrika* **85**, 55-71.