THE INTRINSIC BAYES FACTOR DESCRIBED BY AN EXAMPLE

L. R. Pericchi, I. Fiteni and E. Presa

96-44

WORKING PAPERS

Universidad Carlos III de Madrid

Working Paper 96-44
Statistics and Econometrics Series 15
July 1996

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (341) 624-9849

THE INTRINSIC BAYES FACTOR DESCRIBED BY AN EXAMPLE

L.R. Pericchi, I. Fiteni and E. Presa*

Abstract

The Intrinsic Bayes Factor (IBF) has been recently introduced by Berger and Pericchi (1996) for automatic model selection and hypothesis testing in a Bayesian framework. A major result is the existence, in hypothesis testing problems, of an Intrinsic Proper Prior (IPP) that can be obtained from the IBF in an automatic way. In this article we describe the IBF and compute the IPP in a simple example. It is the hope that the present article will help in making Bayesian methods more widely used for Testing Hypothesis.

Key Words
Bayes Factor; Intrinsic Proper Prior.

*Pericchi, Departamento de Matemáticas, Universidad Simón Bolívar, Caracas; Fiteni, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid; Presa, Departamento de Economía, Universidad Carlos III de Madrid.

1 Introduction

For ease of exposition, asume that the data x is a random sample from one of the simplest (and still useful) model: the exponential model, with likelihood,

$$f(x_i|\theta) = \theta e^{-\theta x_i}, \quad x_i > 0, \quad \theta > 0, \quad \text{for} \quad \underline{x} = (x_1, \ldots, x_n)$$

1.1 Estimation.

Suppose first that the problem is one of the estimation.

If we do not have or do not wish to use subjective prior information, there are several automatic methods to assign a default prior for θ. The most widely used methods seems to be the Jeffreys' rule and the Berger and Bernardo (1992) reference prior algorithm. In this case both methods coincide and yield, calling $I(\theta)$ the Fisher's information,

$$\pi^N(\theta) \propto (detI(\theta))^{1/2} = \left(-\mathrm{E}\left[\frac{\partial^2 log f(x|\theta)}{\partial \theta^2}\right]\right)^{1/2} = \frac{1}{\theta} \tag{1}$$

or equivalently

$$\pi^N(\theta) = \frac{c}{\theta}$$

where c is an unspecified arbitrary positive constant.

Equivalently $\pi^N(log\theta) = c$ in the Real line. Note that $\pi^N(\theta)$ is improper, that is, it integrates infinity in $0 < \theta < \infty$.

This does not prevent us from calculating the posterior density for θ, $\pi(\theta|\underline{x})$, which is proper (if $n \geq 1$) and does not depend on the arbitrary constant c. To see this, note that application of Bayes Rule, yields

$$\pi(\theta|\underline{x}) = \frac{f(\underline{x}|\theta)\pi^N(\theta)}{\int f(\underline{x}|\theta)\pi^N(\theta)d\theta} = \frac{\theta^{n-1}e^{-\theta\sum x_i}}{\frac{\Gamma(n)}{(\sum x_i)^n}} \tag{2}$$

Thus, $\pi(\theta|\underline{x})$ is a Gamma distribution with parameter n and $\sum_{i=1}^n x_i$. This is clearly proper and does not involve the arbitrary constant c, which cancels out. As the sample size grows, the

posterior distribution above is closely approximated by a Normal distribution with the posterior mean and variance of θ , namely $1/\overline{x}$ and $1/(n\overline{x}^2)$ respectively.

On the other hand, the Maximum Likelihood inference based on the MLE (Maximum Likelihood Estimator) $\hat{\theta} = \frac{1}{\overline{x}}$ yields approximately that, $\hat{\theta} \sim N\left(\theta, \hat{I}(\underline{x})^{-1}\right)$, where $\hat{I}(\underline{x})$ is the observed Fisher Information evaluated at $\hat{\theta}$. In this case it gives,

$$\hat{\theta} \sim N\left(\theta, \frac{1}{n\overline{x}^2}\right) \tag{3}$$

For moderate to large samples, in view of the Central Limit Theorem, estimation inferences based on (2) and (3) are numerically quite similar. Still the interpretation of (2) is more satisfiying, in our opinion. However, it may be argued that the differences between automatic Bayes and Likelihood inference, is more academic than practical, for point estimation in regular low dimensional likelihoods and for not too small sample sizes, as often encountered in practice. This numerical similarity is typical in one dimensional estimation problem with regular likelihoods and reference priors.

1.2 Hypothesis Testing

Suppose now that we are told that there is reason to test the hypothesis,

$$M_0 : \theta = \theta_0 \qquad vs \qquad M_1 : \theta \neq \theta_0$$

How the Bayesian analysis is going to proceed? Formally, there is no doubt about it. Define the Bayes Factor of M_1 vs M_0 as,

$$B^{\pi}_{10}(\underline{x}) = \frac{m^{\pi}_1(\underline{x})}{f(\underline{x}|\theta_0)} = \frac{\int f(\underline{x}|\theta)\pi(\theta)d\theta}{f(\underline{x}|\theta_0)} \tag{4}$$

where,

$$m^{\pi}_1(\underline{x}) = \int f(\underline{x}|\theta)\pi(\theta)d\theta,$$

is the marginal or predictive of the observations under M_1.

Then after the assessment of $p_0 = P(M_0)$ and $p_1 = P(M_1)$, $p_0 + p_1 = 1$, for example $p_0 = p_1 = 1/2$ for a default analysis, it follows that

$$P(M_0|\underline{x}) = \frac{1}{\left(1 + \frac{p_1}{p_0}B_{10}(\underline{x})\right)} \tag{5}$$

and $P(M_1|\underline{x}) = 1 - P(M_0|\underline{x})$.

Equation (5) appears to solve the testing problem in a most satisfying manner. It gives the posterior probability of the alternative models, based on the relative adequacy of predicting the data actually observed, that is the Bayes Factor (4). Furthermore to perform predictions of future observations we are not forced to decide for either one model or the other, but we may keep both under consideration weighting individual model predicitions by their posterior probabilities. However, in the quest for automatic default methods, that we can actually be compared with frequentist testing on equal footing, (4) is a *formal* solution to the problem not a definite one. To see this assume the automatic choice (1),$\pi^N(\theta) = \frac{c}{\theta}$. Then,

$$B_{10}^{\pi^N}(\underline{x}) = c\frac{\int f(\underline{x}|\theta)\frac{d\theta}{\theta}}{f(\underline{x}|\theta_0)}$$

which depends upon the arbitrary constant c. Thus, automatic improper choice of $\pi^N(\theta)$, leaves the Bayes Factor undetermined.

On the other hand, there is a sense in which an automatic or default choice in Testing is drastically different from that in Estimation. It can be forcefully argued in a Testing scenario, that the fact that the null model M_0 is seriously considered gives a definite piece of information, and now θ_0 is a distinguished point. Therefore it is needed a general automatic method to assign a proper default prior $\pi(\theta)$ under the alternative model, in view that the null model $M_0 : \theta = \theta_0$ has been definetely suggested.

The surprising fact is that up to now, there was no general method for assessing automatic proper priors for testing, equivalent to the Jeffrey's Rule or the Berger-Bernardo algorithm that assess improper priors for estimation. In words, although there is a formally flawless Bayesian methods for testing, it is not attainable, unless substantial prior information exist. This fact may explain a paradox encountered in the statistical practice. Bayesian methods are much more developed for estimation problems than for testing. But as explained previously , for estimation in regular univariate likelihoods, Bayesian and Likelihood methods tipically rapidly converge numerically, although not in interpretation. On the other hand in testing problems, Bayesian and frequientist methods differ dramatically, and this difference tipically grows with the sample size, see for example Berger and Sellke (1987). And it is frequentist methods for testing which seems to be at fault. Among other things frequentist measure of evidence for models have a difficulty in incorporating the well accepted scientific principle of "Ockham's Razon", i.e. the notion that if two models predicts the data at hand approximately equally well, the simpler model is to be preferred. In fact, for large sample sizes and fixed levels of type one error, the simpler hypothesis is typically rejected in practice, see for example Allenby (1990). To

solve this paradox in practice, and to make cohere the statistical measures of evidence with well established scientific principles, an strategic aim is to develop and disseminate automatic general Bayesian methods for testing hypothesis.

In Section 2, we describe the automatic and general Bayesian method for testing hypothesis and selecting models put forward by Berger and Pericchi (1996), the Intrinsic Bayes Factor (IBF). A key result of the IBF theory, is the existence of a Intrinsic Proper Prior (IPP), for which constructive equations exist. For the simple testing problem described in the introduction, we compute the IBF and find the IPP, which is argued to be appropiate to the problem. In the last Section we advance some conclusions. It is the hope of the present article, to help to disseminate in practice, default Bayesian methods for hypothesis testing.

2 The Intrinsic Bayes Factor for the Exponential Model.

Assume that $x_1, x_2, \ldots, x_n$ is a sample from the Exponential distribution, as in the Introduction and suppose the framework and notation as in Section (1.2).

Hence the Bayes Factor based on $\pi^N(\theta)$ is,

$$B_{10}^N = \frac{m_1^N(\underline{x})}{f(\underline{x}|\theta_0)} = \frac{c\int f(\underline{x}|\theta)\frac{d\theta}{\theta}}{f(\underline{x}|\theta_0)} \tag{6}$$

From (6) it is concluded, as in the Introduction that B_{10}^N is undetermined since it depends on the arbitrary constant c.

A solution to this problem is to use a subset at the observations, say $\underline{x}(l)$ to make the prior for θ proper, and perform the discrimination with the remaining observations $\underline{x}(-l)$. In this example it follows that it is enough to take training samples of size 1 to make the prior proper. Thus, $\underline{x}(l) = x_l > 0$ and $\underline{x}(-l)$ is the original data set taking away the data x_l.

For such a training sample,

$$\begin{aligned} \pi(\theta|x_l) &= \frac{f(x_l|\theta)\pi^N(\theta)}{\int_0^\infty f(x_l|\theta)\pi^N(\theta)d\theta} = \\ &= \frac{\theta e^{-\theta x_l}\frac{1}{\theta}}{\int_0^\infty \theta e^{-\theta x_l}\frac{1}{\theta}d\theta} = x_l e^{-\theta x_l} \end{aligned} \tag{7}$$

Thus, $\pi(\theta|x_l)$ is exponential with parameter x_l, wich is obviously proper.

Using $\pi(\theta|x_l)$ to define a proper Bayes Factor leads to (using Bayes Rule),

$$\begin{aligned} B_{10}(l) &= \frac{\int_0^\infty f(x(-l)|\theta)\,\pi(\theta|x_l)d\theta}{f(x(-l)|\theta_0)} = \\ &= \frac{m_1^N(\underline{x})}{f(\underline{x}|\theta_0)}\frac{f(x_l|\theta_0)}{m_1^N(x_l)} = B_{10}^N B_{01}^N(x_l). \end{aligned} \tag{8}$$

Note that in (8), the arbitrary constant c, has cancel out.

Notice that the more complex model has been placed in the numerator in the Bayes Factor, which is not the usual practice. For explanation of this see Berger and Pericchi (1996) and below.

The Bayes Factor, $B_{10}(l)$, is well defined, but it depends on the arbitrary choice of training sample x_l. To eliminate such a dependence and to increase stability, Berger and Pericchi (1996) propose to average all the posible $B_{10}(l)$. Two averages are put forward:

a) The Aritmetic IBF (AIBF), is defined as

$$\begin{aligned} B_{10}^{AI} &= \frac{1}{n}\sum_{l=1}^{n} B_{10}(l) = \\ &= B_{10}^N \frac{1}{n}\sum_{l=1}^{n} B_{01}^N(x_l). \end{aligned} \tag{9}$$

Computation gives,

$$B_{10}^{AI} = \left[\frac{\Gamma(n)}{e^{-\theta_0\sum_{i=1}^n x_i}\left(\sum_{i=1}^n x_i\right)^n}\right]\left[\frac{\theta_0}{n}\sum_{l=1}^{n} x_l e^{-\theta_0 x_l}\right]. \tag{10}$$

b) The Geometric IBF (GIBF), is defined as,

$$\begin{aligned} B_{10}^{GI} &= \left(\prod_{l=1}^{n} B_{10}(l)\right)^{1/n} = B_{10}^N\left[\prod_{l=1}^{n} B_{01}^N(x_l)\right]^{1/n} = \\ &= B_{10}^N exp\left[\frac{1}{n}\sum_{l=1}^{n} log B_{01}^N(x_l)\right] = \\ &= \left[\frac{\Gamma(n)}{e^{-\theta_0\sum_{i=1}^n x_i}\left(\sum_{i=1}^n x_i\right)^n}\right]\left[\theta_0^n e^{-\theta_0\sum_{l=1}^n x_l}\prod_{l=1}^{n} x_l\right]^{1/n} \end{aligned} \tag{11}$$

Notice that $B_{10}^{GI} \leq B_{10}^{AI}$, since the Geometric mean is less or equal than the Aritmetic mean, so that B_{10}^{GI} will favor more the null hypothesis than B_{10}^{AI}. For a justification of the GIBF see Smith (1995) and Berger and Pericchi (1994). This justification is based on a decision theoretic argument, without assuming that one of the models, M_0 or M_1, is necessarely the sampling model. This is what Bernardo and Smith (1993) call the "Open Model Perspective". To see this, consider in a decision framework the Kullback-Leibler loss funtion predicting the observed data, under the true sampling model $m_T(x)$.

$$\int log\left[\frac{m_0(\underline{x})}{m_1(\underline{x})}\right] m_T(\underline{x})dx$$

and choose M_1 iff the expectation is grater than 0. For improper priors this can not be computed since the marginals are not proper. Take a training sample of size s in order to make all the marginals proper. The problem is of course that we do not know the true sampling model m_T. But as the sampling size grows this can be approximated by

$$\frac{1}{L}\sum_{l=i}^{L} log\left[\frac{m_0(x_{n-s}(-l)|x_s(l)}{m_1(x_{n-s}(-l)|x_s(l)}\right],$$

where

$$m_j(x_{n-s}(-l)|x_s(l)) = \int f_j(x_{n-s}(x(-l)|\theta, x_s(l))\pi(\theta|x_s(l))d\theta.$$

For Minimal Training Samples this turns out to be the Geometric IBF.

In the present paper we concentrate mainly in explorating and justifying the AIBF, because they generate proper Intrinsic priors, see below. However the Geometric IBF has also an important role to play. Appart from the justification in the Open Model Perspective, they are automatically coherent across models in the sense that considering a further model M_2,

$$B_{10}^{G} = B_{20}^{G}/B_{21}^{G}$$

which is not the the case for the Arithmetic IBF which needs an adjustment, Berger and Pericchi (1996). Unfortunately, the Geometric IBF generates Intrinsic Measures that do not integrate up to one, although they do not tend to be far off, see below.

c) The Expected AIBF. (B_{10}^{EAI}).

For very small samples the arithmetic mean of the AIBF's may have a large variability. On the other hand for very large samples the computation of the AIBF may be computationally expensive. An attractive alternative is then to substitute, the average of correction factor by its expectation under the larger model:

$$B_{10}^{EAI} = B_{10}^{N} \mathrm{E}_{M_1}^{\theta} \left[B_{01}^{N}(x_l) \right]$$

Note that quite generally, the law of large numbers gives that,

$$\frac{1}{L} \sum_{l=1}^{L} B_{01}^{N}(x_l) \underset{\mathrm{L} \to \infty}{\longrightarrow} \mathrm{E}_{M_1}^{\theta} \left[B_{01}^{N}(x_l) \right] \quad .$$

both when M_1 or M_0 are the sampling model, provided the limit exists.

Coming back to our example we find that

$$\mathrm{E}_{M_1}^{\theta} \left[B_{01}^{N}(x_l) \right] = \mathrm{E}_{M_1}^{\theta} \left[\theta_0 x_l e^{-\theta_0 x_l} \right] = \tag{12}$$
$$= \int_0^{\infty} \theta_0 x_l e^{-\theta_0 x_l} f(x_l|\theta) dx_l = \frac{\theta \theta_0}{(\theta_0 + \theta)^2}$$

Now since θ is unknown, we estimate it by the M.L.E. estimator $\hat{\theta} = \frac{1}{\bar{x}}$, and we find,

$$\hat{B}_{10}^{EAI} = \frac{\Gamma(n)}{n^n} \frac{\theta_0 e^{\theta_0 n \bar{x}}}{\bar{x}^{n-1} (\bar{x}\theta_0 + 1)^2} \tag{13}$$

Notice that for $B_{10}^{N}(x_l)$, the expectation does not exist for all θ. This is why the more complex model is placed above.

d) The Expected Geometric IBF. Similarly the Expected IBF is defined as

$$B_{10}^{EGI} = B_{10}^{N} exp \left[\mathrm{E}_{M_1}^{\theta} \left(log B_{01}^{N}(x_l) \right) \right]$$

In the present example computation yields

$$B_{10}^{EGI} = B_{10}^{N} \frac{\theta_0}{\theta} exp \left(\psi(1) - \frac{\theta_0}{\theta} \right),$$

where ψ is the digamma function. Anagously, $\hat{B}^{EGI}$ is obtained replacing above θ by $\hat{\theta}$.

2.1 Intrinsic Priors

The Intrinsic Priors are defined as those measures that gives approximately (as n grows) the same answer as the IBF's. Assuming as in the leiv motiff example of this article, a simple null model, we are led via Laplace expansions of the integrals involved to the following equation. First, assume a proper prior $\pi(\theta)$. Then,

$$B_{10} = \frac{\int f(\underline{x}|\theta)\frac{\pi(\theta)}{\pi^N(\theta)}\pi^N(\theta)d\theta}{f(x|\theta_0} \approx \frac{B_{01}^N}{f(x|\theta_0)}\frac{\pi(\hat{\theta})}{\pi^N(\hat{\theta})}.$$

Thus in order that $\pi(\theta)$ would yield approximately the same value as the IBF's we have, from (9) and (11),

$$\frac{\pi(\hat{\theta})}{\pi^N(\hat{\theta})} \approx \frac{1}{n}\sum_{l=1}^{n} B_{01}(x_l)$$

and,

$$\frac{\pi(\hat{\theta})}{\pi^N(\hat{\theta})} \approx exp\left[\frac{1}{n}\sum_{l=1}^{n} log\left(B_{01}(x_l)\right)\right],$$

for the Arithmetic or the Geometric IBF, repectively.

The asymptotic solution of this equation is

$$\pi^I(\theta) = \pi^N(\theta)\mathrm{E}_{M_1}^{\theta}\left[B_{01}^N(x_l)\right]$$

or

$$\pi^I(\theta) = \pi^N(\theta)exp\left[\mathrm{E}_{M_1}^{\theta}\left(logB_{01}^N(x_l)\right)\right]$$

for the Arithmetic and Geometric averages, respectively. These are called Intrinsic Priors.

a) The Arithmetic Intrinsic Prior.

$$\pi^I(\theta) = \pi^N(\theta)\mathrm{E}_{M_1}^{\theta}\left[B_{01}^N(x_l)\right] = \frac{\theta_0}{(\theta_0+\theta)^2}. \tag{14}$$

This prior is quite appealing in more than one sense. First of all it integrates up to one as it is easy to check. Secondly, its median is precisely θ_0, the distinguished point specified by the null model. Thirdly it is quite flat over the whole range with heavy tails. In Figure 1, it is depicted this prior for $\theta_0 = 5$.. It is seen how different it is with respect to the original "non informative" prior. In fact it may be argued that the Arithmetic Intrinsic Prior is the appropriate "default or authomatic" prior given the information that the null model is seriously considered.

The fact that this Intrinsic Prior is proper, is not a fortunate fact particular to this example as the following argument shows. Consider any simple Hypothesis M_0. Then integrating the Intrinsic Prior we find,

$$\begin{aligned}\int \pi^I(\theta)d\theta &= \int \pi^N(\theta) \int \frac{f(X(l)|\theta_0)}{m_1^N(X(l))} f(X(l)|\theta)dX(l)d\theta = \\ &= \int \frac{f(X(l)|\theta_0)}{m_1^N(X(l))} \int \pi^N(\theta) f(X(l)|\theta)d\theta dX(l) = \\ &= \int f(X(l)|\theta_0)dX(l) = 1\end{aligned}$$

This can be seen as a procedure for obtaining priors appropriate for Testing Hypothesis, and might be thought as the equivalent procedure for obtaining reference or default priors for testing hypothesis than Jeffreys or Berger-Bernardo algorithm for obtaining reference priors for estimation problems.

Once obtained they can be used as a prior and the perform tha corresponding integration (often numerically) to obtain the proper Intrinsic Bayes Factors. Alternatively, approximations can be performed, obtaining in this example, the following simple approximation

$$B_{10}^N = \frac{\int f(\underline{x}|\theta)\pi^I(\theta)d\theta}{f(\underline{x}|\theta_0)} \approx B_{10}^N \frac{\theta_0\hat{\theta}}{(\hat{\theta}+\theta_0)^2}.$$

b) The Geometric Intrinsic Prior. The same method as above, but now using the average of the Geometric IBF, leads in this example to,

$$\pi^I(\theta) = \frac{\theta_0}{\theta^2} exp\left[\psi(1) - \frac{\theta_0}{\theta}\right].$$

Integration of this prior gives $\exp(\psi(1)) = 0.561$, that is it is integrable but not proper as anticipated above.

This prior, after being normalized for comparison, is also shown in Figure 1. Except close to $\theta = 0$, the Geometric and Arithmetic priors are quite close.

3 Discussion

For ease of exposition we have restricted ourselves to describe some of the ideas behind the Intrinsic Strategy for Bayesian model comparison in a simple example. Far more complex situations are addressed in Berger and Pericchi (1994, 1996).

When the expectations involved in the Intrinsic Prior Equations are feasable this strategy gives a procedure for obtaining prior measures suitable for automatic analyses in the comparison of models in a proper Bayesian way.

More generally, Arithmetic and Geometric IBF's, are often easely computable, and this paves the way for practical Bayesian Hypothesis Testing. For large data sets, and a large number of models computations might be expensive, but recently several approximating inexpensive procedures are being put forward (Varshavsky, 1995).

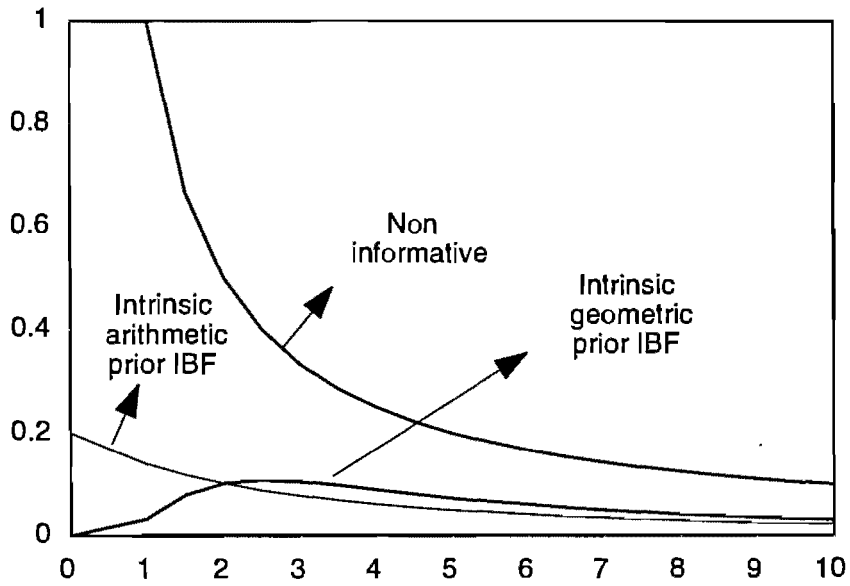

Figure 1.

References

[1] Allenby G.M. (1990), "Hypothesis testing with scanner data: the advantage at Bayesian Methods", *Journal of Marketing Research*, 27, 379-389.

[2] Berger J.O. and Sellke T. (1987), "Testing a point null hypothesis: the irreconciability of p-values and evidence", *Journal of American Statistical Association*, 82, 112-122.

[3] Berger J.O. and Bernardo J.M. (1992), "On the development of the reference prior method", *Bayesian Statistics IV*, (eds. J.M. Bernardo et al.), London: Oxford University Press.

[4] Berger J.O. and Pericchi L.R. (1994), "The intrinsic Bayes Factor for Linear Models", *To appear in Bayesian Statistics V*, (eds. J.M. Bernardo et al.), London: Oxford University Press.

[5] Berger J.O. and Pericchi L.R. (1996), "The intrinsic Bayes Factor for Model Selection and Prediction", *Journal of American Statistical Association*, in press.

[6] Bernardo J.M. and Smith (1993), *"Bayesian Theory"*, J. Wiley and Sons, New York.

[7] Smith, A.F.M. (1995), Contribution on the discussion of: O'Hagan A. "Fractional Bayes factors for model comparisons", *J. Royal Statistical Society B*, 57, 1, 120-122.

[8] Varshavsky, J.A. (1995), "On the development of Intrinsic Bayes Factors", *Technical Report, Statistics Department*, Purdue University, Indiana, U.S.A..