On The Mathematics of The Jeffreys–Lindley Paradox

Cristiano Villa[*] and Stephen Walker[†]

Abstract

This paper is concerned with the well known Jeffreys–Lindley paradox. In a Bayesian set up, the so-called paradox arises when a point null hypothesis is tested and an objective prior is sought for the alternative hypothesis. In particular, the posterior for the null hypothesis tends to one when the uncertainty, i.e. the variance, for the parameter value goes to infinity. We argue that the appropriate way to deal with the paradox is to use simple mathematics, and that any philosophical argument is to be regarded as irrelevant.

Keywords Bayes factor, Bayesian hypothesis testing, Kullback–Leibler divergence, self-information loss.

1 Introduction

The literature on the *Jeffreys–Lindley paradox* has been prolific since it was brought to the attention of objective Bayesians by Lindley (1957) and by Bartlett (1957). Many authors have discussed this so-called paradox from varying perspectives; including not only statisticians, but philosophers too. Our aim is to consider the problem using simple mathematics.

Lindley (1957) shows that, for point null hypothesis testing, there may be a concern with the objective Bayesian approach. In the specific example used, if the prior for the location parameter, in the alternative model to the parameter being zero, has infinite variance, then the Bayesian

[*] University of Kent, School of Mathematics, Statistics and Actuarial Sciences, Cornwallis Building, University of Kent, Canterbury, Kent CT2 7NF UK. e-mail: cv88@kent.ac.uk

[†] Department of Mathematics, University of Texas at Austin, GDC 5.504, 1 University Station G2500 Austin, TX 78712. e-mail: s.g.walker@math.utexas.edu

will always select the null model, regardless of the observed data. This was first suggested as a warning against using improper priors, but the consequences have now become far reaching with a substantial amount of literature written about the observation.

Let us describe the mathematical setting of the problem. Suppose we wish to test the hypothesis

$$H_0 : \theta = 0 \quad \text{vs} \quad H_1 : \theta \neq 0$$

for the normal model $N(x|\theta, 1)$. Let $\rho_0 = P(M_0)$ be the prior probability assigned to the null hypothesis and let

$$\pi(\theta) = N(\theta|0, \sigma^2),$$

for some $\sigma > 0$, be the prior distribution for the unknown parameter θ under the alternative model. The problem is to set σ.

The Bayes factor for this problem is given by

$$B_{01} = \frac{N(x|0,1)}{\int N(x|\theta,1)\pi(\theta)\,d\theta},$$

which represents the odds in favour of the null hypothesis with respect to the alternative. The decision on whether one rejects H_0 in favour of H_1 is based on the posterior probability, given by

$$P(M_0|x) = \left[1 + \frac{1-\rho_0}{\rho_0}\frac{1}{B_{01}}\right]^{-1}.$$

This is the extent of the mathematical foundations to the problem. As Lindley noted, there are some combinations of (ρ_0, σ) yielding a $P(M_0|x)$ which one would not wish to countenance.

The natural objective choice for $\pi(\theta)$ involves taking $\sigma = \infty$. However, rather than a direct plug in of this value, a more general setting has been suggested and considered by Robert (1993) which is to let ρ_0 depend on σ, i.e. we have $\rho_0(\sigma)$, so then it is possible to study $P(M_0|x)$ as $\sigma \to \infty$. In this case we can identify three scenarios for $P(M_0|x)$ as $\sigma \to \infty$, all of which have associated problems. That is, there is no setting, i.e. choice of $\rho_0(\sigma)$, in which the choice $\sigma = \infty$

as an objective choice can work. What we mean by this is explained in Section 2. The conclusion is that the objective idea of $\sigma = \infty$ does not work and consequently the message is not to use it. On the other hand, we can set the pair $(\sigma < \infty, \rho_0(\sigma))$ objectively using ideas of Type I error calculations and a novel approach to the selection of priors for models.

It is beneficial for us to clarify what is meant by an objective approach; the most sensible definition, and at the same time the most pragmatic, can be found in Berger (2006). In short, objective Bayes is a collection of default methodologies which lead to the construct of a prior distribution through a set of automated processes. In this respect, our approach will be seen to be objective.

We also argue that the Bayesian component of the Bayes factor test is restricted to the construction of the Bayesian test statistic; i.e. the Bayes factor B_{01} itself. Following this, the Bayes factor test must be scrutinized according to standard testing procedures such as Type I errors. The procedure is fully automated once the Type I error has been set, alongside the value for which the posterior probability of the null hypothesis is regarded as just too large for it to be rejected.

The layout of the paper is as follows. In Section 2 we formalise the Jeffreys–Lindley paradox and discuss Robert (1993) solution to it. Section 3 is dedicated to the our approach, and Section 4 is reserved to conclusions and final comments.

2 Formalisation of the paradox

In order to discuss approaches to the Jeffreys–Lindley paradox, let us first formalise it and, at the same time, define the notation. The aim is to compare the two normal models,

$$M_0 = \left\{ N(x|0,1) = (2\pi)^{-1/2} \exp(-\tfrac{1}{2}x^2) \right\},$$

$$M_1 = \left\{ N(x|\theta,1) = (2\pi)^{-1/2} \exp\{-\tfrac{1}{2}(x-\theta)^2\} \right\}.$$

To apply the Bayesian approach, as described in Section 1, we need to define both the priors; i.e. the value of σ, for the unknown parameter θ, and the prior for the null hypothesis; i.e. the value of ρ_0. To be most general we will assume that ρ_0 can depend on σ and hence we write it as $\rho_0(\sigma)$.

With this information we can compute the Bayes factor representing the odds for the null hypothesis H_0. That is

$$B_{01} = \frac{N(x|0,1)}{\int N(x|\theta,1) \cdot N(\theta|0,\sigma^2)\, d\theta} = \frac{e^{-\frac{1}{2}x^2}}{e^{-\frac{1}{2}x^2/(1+\sigma^2)}} \sqrt{1+\sigma^2},$$

so the posterior probability for the null hypothesis is given by

$$P(M_0|x) = \left[1 + \frac{1-\rho_0}{\rho_0}\frac{1}{B_{01}}\right]^{-1} = \left[1 + \frac{1-\rho_0}{\rho_0}\frac{e^{-\frac{1}{2}x^2/(1+\sigma^2)}}{e^{-\frac{1}{2}x^2}}\frac{1}{\sqrt{1+\sigma^2}}\right]^{-1}. \tag{1}$$

We note in (1) that the quantity

$$m(\sigma) = \frac{1-\rho_0(\sigma)}{\rho_0(\sigma)}\frac{1}{\sqrt{1+\sigma^2}} \tag{2}$$

is the key term and opens the way to understanding the paradox, which we recall being associated with $\sigma = \infty$. As we want a prior for the null hypothesis that goes to zero for $\sigma \to \infty$, reflecting an increasing uncertainty about $\theta = 0$, we can write

$$\rho_0(\sigma) = O\left(\sigma^{-\beta}\right), \qquad \beta \geq 0,$$

and identify the following three cases where the decision maker specifies β to control the speed at which $\rho_0(\sigma)$ goes to zero as $\sigma \to \infty$:

(i) $0 \leq \beta < 1$. Under this scenario we have the undesirable result that $P(M_0|x)$ converges to 1 regardless of the x value. This is the so-called paradoxical result. In fact, from (2) we see that $m(\sigma) \to \infty$ for $\sigma \to \infty$; so, if the prior on the null hypothesis is too large as $\sigma \to \infty$, the posterior probability in the null hypothesis will converge to one.

(ii) $\beta = 1$. From (2) we have that $m(\sigma) \to c$ for some constant $0 < c < \infty$. Under this scenario it is that, for large σ,

$$\rho_0(\sigma) = \frac{1}{1 + c\sqrt{1+\sigma^2}} \approx \frac{1}{1+c\sigma}.$$

In particular, Robert (1993) presents an objective argument for

$$\rho_0(\sigma) = \frac{1}{1+\sqrt{2\pi}\sigma}.$$

However, this idea leads to an undesirable inconsistency in that $\rho_0(\sigma) \to 0$ yet $P(M_0|x)$ is converging to a constant bounded away from 0. Thus, with $\sigma = \infty$, we have $P(M_0) = 0$ but $P(M_0|x) \neq 0$, which are incoherent choices.

(iii) $\beta > 1$. Under this scenario we have $m(\sigma) \to \infty$, resulting in $P(M_0|x) \to 0$. This at least now becomes consistent with the prior probability since $\rho_0(\sigma) \to 0$ in this case. Yet undesirable in that with $\sigma = \infty$, $P(M_0|x) = 0$.

These considerations clearly exclude the choice $\sigma = \infty$. It simply does not work. Thus a finite choice of σ is required. In the next section we will demonstrate how we can set $(\sigma < \infty, \rho_0(\sigma))$ objectively.

3 An objective choice for $(\sigma, \rho_0(\sigma))$

Given a value of σ we first, in Section 3.1, show how to obtain an objective choice for $\rho_0(\sigma)$. Then, in Section 3.2, we show how $\sigma < \infty$ can be selected objectively.

3.1 The prior $\rho_0(\sigma)$

Our approach consists in measuring the *worth* of the alternative hypothesis M_1 with respect to the null M_0, as outlined in Villa and Walker (2015). The quantification of the *worth* comes from a result in Berk (1966) which states that, if a model is misspecified, the posterior distribution would asymptotically accumulate on the model which is the nearest, in terms of the Kullback–Leibler divergence (Kullback and Leibler , 1951), to the true model. Therefore, if we were to choose M_0

(and we knew θ), when M_1 is the true model, the loss in information we would incur is given by the divergence $D_{KL}(N(x|\theta,1)\|N(x|0,1))$, which can also be interpreted as the utility in keeping M_1, that is its *worth*. In other words, the larger the value of $D_{KL}(N(x|\theta,1)\|N(x|0,1))$ the greater the utility (or equivalently, the smaller the loss) of choosing the alternative hypothesis M_1. However, since we do not know θ, but we can assign a prior $\pi(\theta)$ to it, we compute the expected loss as

$$l(M_1) = -\int_\Theta D_{KL}\Big(N(x|\theta,1)\|N(x|0,1)\Big)\pi(\theta)\,\mathrm{d}\theta = -\int \tfrac{1}{2}\theta^2\,\pi(\theta)\,\mathrm{d}\theta = -\tfrac{1}{2}\sigma^2. \tag{3}$$

Now, if we consider the mass the be put on the alternative hypothesis, this can be linked to the *worth* of M_1 via the *self-information* loss function. The *self-information* loss function (also known as the *log-loss* function) measures the performance of a probability statement with respect to an outcome. For example, if we quantify the probability of outcome A of being true as $P(A)$, the associated *self-information* loss would be given by $-\log P(A)$. More details and properties of this particular loss function can be found, for example, in Merhav and Feder (1998). We then have a measure of the information loss of M_1 related to its *worth*, given by (3), and a measure of the same loss related to the *self-information*, given by $-\log P(M_1)$. By equating the two measures we have

$$-\log P(M_1) = -\int_\Theta D_{KL}\Big(N(x|\theta,1)\|N(x|0,1)\Big)\pi(\theta)\,\mathrm{d}\theta,$$

which gives

$$1-\rho_0(\sigma) \propto \exp\{\sigma^2/2\}.$$

Given that the hypothesis M_0 is nested into the alternative hypothesis M_1, the loss in information in choosing M_1 when M_0 is the true model is zero; therefore, by applying the above framework, we have $P(M_0)\propto 1$, and so we have

$$\rho_0(\sigma) = \frac{1}{1+\exp\{\frac{1}{2}\sigma^2\}}.$$

This then fits into category (iii) for large σ, which implies that $P(M_0|x,\sigma)$ goes to zero as $P(M_0) \to 0$. Thus there is coherence in this approach; however, we are not advocating the choice of large σ. It has to be noted that the idea of linking the *self-information* with the expected loss is not new, as in fact is originated in Bernardo (1979a) and Bernardo (1979b) where it forms the basis for reference analysis.

3.2 Determining σ

In any classical test the Type I error is of key importance. We can use this quantity to objectively set the value for σ. The Type I error needs to be set, and a valid objective Bayesian criterion is to match classical benchmarks and quantities.

To determine an appropriate value for σ based on the classical concept of Type I error, we would select σ so that

$$P_0(\text{reject } M_0) = \alpha,$$

where $\alpha \in (0,1)$ and P_0 is the probability under the null hypothesis. Regardless of the surroundings, all Bayesian experimenters in this problem would need to assign an α_B value for which one would reject H_0 if $P(M_0|x) < \alpha_B$. As is usual with the paradox, it is studied with a single observation x and so α_B needs to be set bearing this in mind. Hence, a small α_B, we consider a value of 0.05, is required. As the sample size grows we would indeed increase α_B to $\frac{1}{2}$ as a desirable large sample value.

To have

$$P_0\biggl(P(M_0|x) < \alpha_B\biggr) = \alpha, \tag{4}$$

we require

$$\begin{aligned}
\frac{1}{1+m(\sigma)\exp\left\{\frac{1}{2}x^2\frac{\sigma^2}{1+\sigma^2}\right\}} &< \alpha_B, \\
\text{i.e. } \exp\left\{\frac{1}{2}x^2\frac{\sigma^2}{1+\sigma^2}\right\} &> \frac{1/\alpha_B-1}{m(\sigma)} \\
\frac{1}{2}x^2\frac{\sigma^2}{1+\sigma^2} &> \log\left(\frac{\alpha_B^{-1}-1}{m(\sigma)}\right) \\
x^2 &> \frac{2(1+\sigma^2)}{\sigma^2}\log\left(\frac{\alpha_B^{-1}-1}{m(\sigma)}\right). \qquad (5)
\end{aligned}$$

Therefore, if we write

$$\psi(\sigma) = \frac{2(1+\sigma^2)}{\sigma^2}\log\left(\frac{\alpha_B^{-1}-1}{m(\sigma)}\right),$$

we have (4) as

$$P_0\Big(x^2 > \psi(\sigma)\Big) = 2\left[1-\Phi\left(\sqrt{\psi(\sigma)}\right)\right] = \alpha. \qquad (6)$$

The key here is that, for $\alpha_B < 1/2$, $\psi(\sigma)$ is decreasing as σ increases, so there is a one-to-one correspondence between α and σ satisfying (6). The following theorem, the proof of which is in Appendix A, gives the above result.

Theorem 1. *The quantity $\psi(\sigma) = 2(1+\sigma^2)/\sigma^2 \cdot \log\left(\{\alpha_B^{-1}-1\}/m(\sigma)\right)$, where $m(\sigma)$ is defined in* (2)*, is decreasing for $\sigma \to \infty$, for all $\alpha_B < 1/2$.*

Figure 1 shows the behaviour of $\log\psi(\sigma)$, given $\alpha_B = 0.05, 0.40, 0.49$. As it must be that $\psi(\sigma) > 0$, we compute the functions $\log\psi(\sigma)$ up to the appropriate value of σ which ensures $m(\sigma) < \alpha_B^{-1} - 1$.

For each value of α_B, expression (6) has to be solved numerically. So, for $\alpha_B = \alpha = 0.05, 0.40, 0.49$, we would have $\sigma = 2.11, 0.51, 0.15$, respectively. In other words, we can be objective about σ with a finite value. The notion therefore that an objective σ and $\sigma = \infty$ is the only choice is wrong. An objective classical test requires an α value and it is this which can be linked to the (finite) objective choice for σ.

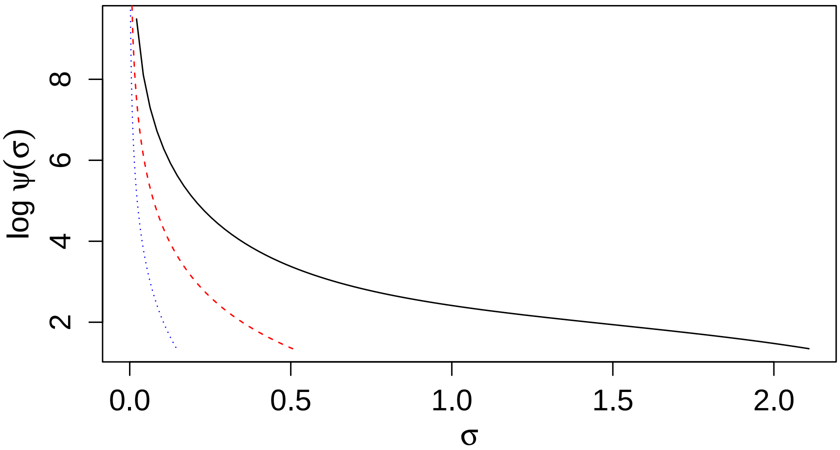

Figure 1: Plot of $\log \psi(\sigma)$, with $\alpha_B = 0.05$, continuous black line, $\alpha_B = 0.40$, dashed red line, and $\alpha_B = 0.49$, dotted blue line.

4 Discussion

The findings of this paper can be summarised as follows. The posterior for the point null hypothesis is driven by the quantity $m(\sigma)$; in particular, if $\sigma = \infty$ is desired as an objective criterion then the behaviour of $m(\sigma)$ as $\sigma \to \infty$ is the key. If the prior $\rho_0(\sigma)$ is fixed, e.g. is equal to $\frac{1}{2}$, then the Jeffreys–Lindley paradox arises, since the posterior probability $P(M_0|x)$ goes to one. Robert (1993) proposed to solve the issue by having $m(\sigma)$ to converge to a positive constant. Although the direct paradox is avoided, the approach gives an incoherent result as the posterior mass on H_0 is positive whereas the prior mass is zero. Our approach gives a quantity $m(\sigma)$ which goes to infinity, for σ going to infinity, which both solves the paradox and yields zero posterior mass for H_0 when $\rho_0(\sigma) = 0$, implying the prior mass for H_0 is zero.

It is clear that the three types of behaviour of $m(\sigma)$ for large σ rule out the possibility of having $\sigma = \infty$. As such, σ has to be determined to have a finite value. For $\rho_0(\sigma)$, the choice can be either objective or subjective. Our approach allows $\rho_0(\sigma)$ to be determined in an objective fashion by considering the loss in information if the true model is removed. Dellaportas et al. (2012), on the other hand, propose a prior for the null hypothesis that is subjective.

Dellaportas et al. (2012) focus on models for which the use of a multivariate normal prior is appropriate, such as linear regression models, generalised linear models and standard time series models. The idea is to set the multiplicative constant for the prior dispersion matrix, c_M, which will indicate the level of prior uncertainty. The authors aim to reduce the sensitivity of the posterior model probabilities to the scale of the prior by suitably specifying the prior model probability. This is done by setting

$$P(M) \propto P'(M) c_M^{d_M},$$

where d_M is the dimension of the model M and $P'(M)$ is a suitably determined base line prior model probability. Dellaportas et al. (2012) recommend $P'(M) \propto 1$, although other choices are possible. We see that the core of the whole approach is to make a prior model probability dependent on the variance of the prior in the parameters, avoiding the Jeffreys–Lindley paradox. In this aspect, hour approach and the one in Dellaportas et al. (2012) are analogous: in both cases the paradoxical effect is avoided by incorporating the effect of the prior variance in the model probability. However, while they use a penalty based on the dimension, we specify the model comparison (i.e. H_0 vs. H_1) by choosing the Type I error.

The conclusion is that it is not possible to be objective for $\pi(\theta)$ by setting $\sigma = \infty$. This is not the sole case where objective Bayes fails to deliver adoptable solutions. For example, Jeffreys' rule prior for multidimensional parameter spaces gives prior distribution with poor performance properties (Bernardo and Smith , 1994). It is common practice not to use Jeffreys prior in these type of problems and opt for a different solution, such as reference priors.

However, an objective and finite value of σ can be assigned by exploiting thinking behind classical tests and setting the Type I error. That is, there is a one-to-one correspondence between σ and the Type I error α and it is this correspondence which permits the interpretation and assignment of σ.

Surprisingly, or not, there have been philosophical papers attempting to find some hidden profound explanation behind the paradox; see, for example, the recent papers of Spanos (2013) and Sprenger (2013). We argue that it is not necessary to philosophize, as the mathematics of the problem are quite straightforward and a clear picture of what is happening can be understood solely by mathematical considerations.

To discuss some of the philosophy, Spanos (2013) says: “The question that generally arises is why the Bayesian and the likelihoodist approaches give rise to the above conflicting and confusing results”. However, we have $P(M_0|x) < \alpha_B \Leftrightarrow x^2 > \psi(\sigma)$, which is precisely the form of the classical test!

The classical test is: reject H_0 if $x^2 > c_\alpha$, where $P_0(x^2 > c_\alpha) = \alpha$. We can then set $\psi(\sigma) = c_\alpha$ to ensure a standard value for the Type I error. Consequently, Bayes makes no contribution to this problem, since even a subjective Bayesian approach will yield a classical test, but with perhaps a non-standard Type I error. Such an observation between Bayesian and classical tests has been made by Shively and Walker (2013).

Although the aim of this paper is to discuss on the mathematics of the Jeffreys–Lindley paradox, in particular on the original problem as in Lindley (1957), the above result can be seen to hold in more general cases. Let us assume that we want to test the null hypothesis $H_0 : M_0 = f(x|\theta_0)$ versus the alternative hypothesis $H_1 : M_1 = f(x|\theta)$, where $\theta, \theta_0 \in \Theta$, and $f(\cdot|\theta)$ is a given probability distribution. The Bayes factor test statistic, given prior π, will then be

$$B = \frac{\int_\Theta f(x|\theta)\pi(d\theta)}{f(x|\theta_0)}.$$

The method to obtain the prior value to be assigned to each hypothesis, given the prior on the parameter θ, is discussed in Villa and Walker (2015). As discussed in Section 3, the prior on the hypothesis (i.e. the models) depends on the choice of π, and it considers the expected loss in information deriving from the wrong choice of hypothesis, which is represented by the expected minimum Kullback–Leibler divergence from one model to the other. Therefore, for fixed α and α_B, it is possible to choose a prior π such that

$$P_0\bigg(P(M_0|x,\pi) < \alpha_B\bigg) = \alpha. \tag{7}$$

In summary our general approach is as follows:

- Set the pair of values (α, α_B), noting that all Bayesians need to set a value for α_B.
- Obtain the prior value $P(M_0|\pi)$ as discussed in Villa and Walker (2015)
- Select the π such that (7) holds.

It is straightforward to see that, should we remove the assumption of known variance, the proposed approach still holds. Let us then consider the more general normal model $N(x|\theta, \lambda)$, where λ is the precision. The Bayes factor representing the odds in favour of the null hypothesis, assuming $\lambda \sim \text{Ga}(\varepsilon, \varepsilon)$, becomes

$$B_{01} = \left\{ \frac{1 + T^2/(1 + n\sigma^2)}{1 + T^2} \right\}^{n/2} \sqrt{1 + n\sigma^2},$$

where

$$T^2 = \frac{n\overline{X}^2}{(n-1)S^2},$$

with $\overline{X} = \sum_{i=1}^n X_i/n$ being the sample mean and S^2 the sample variance. Given that the normal model now has two unknown parameters, and to have S^2 properly defined, we consider $n = 2$. Furthermore, the choice of a gamma prior for λ allows us to consider the Jeffreys prior $\pi(\lambda) \propto \lambda^{-1}$ by taking $\varepsilon \to 0$. Similarly to the case with known variance, we have that the key quantity in understanding the paradox is

$$m'(\sigma) = \frac{1 - \rho_0(\sigma)}{\rho_0(\sigma)} \frac{1}{\sqrt{1 + 2\sigma^2}},$$

and we can replicate the considerations about the behaviour of the posterior $P(M_0|x)$ discussed in Section 1. The form of $\rho_0(\sigma)$ does not change; now (3) becomes

$$\begin{aligned} l(M_1) &= -\int_\Lambda \int_\Theta D_{KL}\Big(N(x|\theta, \lambda) \| N(x|0, \lambda)\Big) \pi(\theta)\, \pi(\lambda)\, d\theta\, d\lambda \\ &= -\int_\Lambda \int_\Theta \frac{\theta^2 \lambda}{2} \pi(\theta) \pi(\lambda)\, d\theta\, d\lambda \\ &= -\frac{1}{2}\sigma^2\, \mathbb{E}\lambda = -\frac{1}{2}\sigma^2, \end{aligned}$$

given that we have set $\pi(\lambda) = \text{Ga}(\varepsilon, \varepsilon)$, meaning $\mathbb{E}(\lambda) = \varepsilon/\varepsilon = 1$. To satisfy (4), we need

$$\begin{aligned} P(M_0|x) &= \left[1 + m'(\sigma)\left\{\frac{1+T^2}{1+T^2/(1+2\sigma^2)}\right\}\right]^{-1} < \alpha_B \\ \text{i.e.} \quad T^2 &> \frac{(\alpha_B^{-1}-1)/m'(\sigma)-1}{1-(\alpha_B^{-1}-1)/[m'(\sigma)*(1+2\sigma^2)]}. \end{aligned}$$

By setting the right-hand-side to $\psi'(\sigma)$, we recover the equivalent of (6) as

$$P_0\Big(T^2 > \psi'(\sigma)\Big) = 2\left[1 - F^{-1}\left(\sqrt{\psi'(\sigma)}\right)\right] = \alpha,$$

where F^{-1} is the inverse of a Student-t with 1 degree of freedom, and which numerical solution will allow to set σ at the appropriate value.

A Proof of Theorem 1

We start by noting that we can write

$$m(\sigma) = \frac{e^{\sigma^2/2}}{\sqrt{1+\sigma^2}},$$

which is increasing for σ^2 increasing and $m(\sigma) > 1$. We then write

$$\psi(\sigma) = 2\left(\frac{1}{\sigma^2}+1\right)\log\left(\frac{c}{m(\sigma)}\right),$$

which is decreasing on $(0, \widehat{\sigma})$, where $c = \alpha_B^{-1} - 1$, and $\widehat{\sigma}$ satisfies $m(\widehat{\sigma}) = c$. Also, note that $\psi(0) = \infty$. Hence, we can find a suitable σ provided $c > 1$, i.e. provided $\alpha_B < \frac{1}{2}$.

References

Bartlett, M. S. (1957). Comment on D. V. Lindleys statistical paradox. *Biometrika* **44**,533–534.

Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis* **1**,385–402.

Bernardo, J. M. (1979a). Expected information as expected utility. *The Annals of Statistics* **7**, 686–690.

Bernardo, J. M. (1979b). Reference posterior distributions for Bayesian inference. *J. Royal Statistical Society B* **41**, 113–147.

Bernardo, J. M. (1997). Noninformative priors do not exist: a discussion. *J. of Statistic. Plan. and Inf.* **65**, 159–189.

Bernardo, J. M., and Smith, A. F. M. (1994). *Bayesian Theory.* John Wiley & Sons.

Berk, R. H. (1966). Limiting behaviour of posterior distributions when the model is incorrect. *Ann. of Math. Statist.* **37**, 51–58.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2012). Joint Specification of model space and parameter space prior distributions. *Statistical Science* **27**,232–246.

Kullback, S., and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics* **2**,79–86.

Lindley, D. V. (1957). A statistical paradox. *Biometrika* **44**,187–192.

Merhav, N., and Feder, M. (1998). Universal prediction. *IEEE Trans. Inf. Theory* **44**,2124–2147.

Robert, C. P. (1993). A Note on Jeffreys-Lindley Paradox. *Statistica Sinica* **3**,601–608.

Shively, T., and Walker, S. G. (2013). On the Equivalence between Bayesian and Classical Hypothesis Testing. arXiv:1312.0302v1 [stat.ME].

Spanos, A. (2013). Who Should Be Afraid of the Jeffreys–Lindley Paradox? *Philosophy of Science* **80**,73–93.

Sprenger, J. (2013). Testing a Precise Null Hypothesis: The Case of Lindley's Paradox. *Philosophy of Science* **80**,733–744.

Villa C., and Walker, S. G. (2015). An objective Bayesian criterion to determine model prior probabilities. *Scandinavian Journal of Statistics.* **42**, 947–966