arXiv:1303.5973v1 [stat.ME] 24 Mar 2013

On the Jeffreys–Lindley's paradox*

Christian P. Robert

Abstract: This note reassess the dual nature of the Jeffreys–Lindley's paradox and of its considerable impact on both classical and Bayesian statistics, as well as on existing resolutions of the paradox. It also examines a recent and critical viewpoint on the paradox by Spanos (2013).

Keywords and phrases: Bayesian inference, Testing statistical hypotheses, Type I error, significance level, p-value.

1. Understanding the paradox setting

Maybe paradoxically, my own understanding of the Jeffreys–Lindley's paradox has always been that it pointed at the poor and even unacceptable behaviour of vague prior distributions when testing point-null hypotheses. For instance, my own attempt at solving the paradox (Robert, 1993) was definitely written under this understanding and aimed at suppressing the impact of an arbitrary normalising constant in improper priors. It is only very recently that I became aware that most people (Dennis Lindley included) understand the paradox as an irreconcilable divergence between the Bayesian _and the frequentist (ƒ) resolutions of the point-null hypothesis testing problem, blaming one of those for the discrepancy. (It has been reasonably argued that there is no such thing as *one* Bayesian resolution or *one* frequentist resolution. While I agree on principle with this view, I will nonetheless restrict the discussion below to the opposition between the p-value and the posterior probability—or equivalently the Bayes factor, see e.g. Kass and Wasserman, 1996.)

I must acknowledge being rather surprised at this common focus as I see no reason why both approaches should agree: (a) one _is operating on the parameter space Θ, while the other (ƒ) is produced on the sample space $\mathcal{X}$, or, in other words, one (ƒ) is dealing with credibility while the other dabbles in confidence; (b) one (ƒ) relies solely on the point-null hypothesis H_0 and the corresponding distribution, while the other _opposes H_0 to a marginal version of H_1 (integrated over the parameter space Θ against a specific prior distribution); (c) following what may be the most famous quote from Jeffreys (1939, Section 7.2) one (ƒ) could rejects "a hypothesis that may be true (...) because it has not predicted observable results that have not occurred" ($\{X > x_{\text{obs}}\}$, say), while the other _conditions upon the observed value x_{obs}; (d) one (ƒ) resorts to an arbitrary fixed bound α on the p-value, while the other _refers to the boundary probability of $1/2$ (unless a genuine loss function is constructed) A consequent

*Christian P. Robert, CEREMADE, Université Paris-Dauphine, 75775 Paris cedex 16, France xian@ceremade.dauphine.fr. Research partly supported by the Agence Nationale de la Recherche (ANR, 212, rue de Bercy 75012 Paris) through the 2012–2015 grant ANR-11-BS01-0010 "Calibration" and by an Institut Universitaire de France chair.

literature (see, e.g. Berger and Sellke, 1987) has since then shown how divergent those two approaches could be (to the point of being asymptotically incompatible).

While the gap between frequentist and Bayesian degrees of evidence was certainly the reason for Lindley (1957) mentioning a statistical paradox, I thus remain convinced that the richest consequence of Jeffreys's (1939) and Lindley's (1957) exhibitions of this paradox is to highlight the genuine difficulty in using improper or very vague priors in testing settings: as stressed by Lindley (1957), "the only assumption that will be questioned is the assignment of a prior distribution of any type" (p.188). This were also the arguments made by both Shafer (1982) and DeGroot (1982) (see also DeGroot, 1973) in their discussion of the paradox. Note that Jeffreys does not address the general problem of using improper priors in testing, using ad-hoc solutions when available and developing a second (and under-appreciated) type of Jeffreys's priors otherwise (see Robert et al., 2009, Section 6.4, for a discussion).

The plan of this note is as follows: it reviews the paradox in Section 2, analyses the recent criticism on Spanos (2013) in Section 3, discusses the Bayesian aspects of the paradox in Section 4, and concludes in Section 5.

2. The paradox, paradoxes, or non-paradox

Let us first recall the setting set in Lindley (1957). If one considers a normal mean testing problem,

$$\bar{x}_n \sim \mathcal{N}(\theta, \sigma^2/n)\,, \quad H_0 : \theta = \theta_0\,,$$

using Jeffreys's (1939) choice of prior, $\theta \sim \mathcal{N}(\theta_0, \sigma^2)$, leads to the Bayes factor

$$\mathfrak{B}(t_n) = (1+n)^{1/2} \exp\left(-nt_n^2/2[1+n]\right)\,,$$

where $t_n = \sqrt{n}|\bar{x}_n - \theta_0|/\sigma$ is the classical t-test statistic.

The first level of the paradox is that, *when t_n is fixed and n to infinity, the Bayes factor goes to infinity while the p-value remains constant.* In Lindley's words, "we [can be] 95% confident that $\theta \neq \theta_0$ but have 95% belief that $\theta = \theta_0$" (p.187). As discussed previously in the literature, this is *not* a mathematical paradox as the quantities measure different objects (the probability measure of an event over the sample space versus the probability measure of an event over the parameter space, the former being conditional on the parameter value and the later on the observation of the sample) and this is *not* a statistical paradox in that observing a constant[1] t_n as n increases is not of interest: when H_0 is true, t_n has a limiting $\mathcal{N}(0,1)$ distribution, while, when H_0 does not hold, t_n converges almost surely to ∞, in which case the Bayes factor converges to 0. This behaviour is thus entirely compatible with the result of the consistency of the Bayes factor in this setting.[2]

[1] As pointed out by Lindley (1957): "5% in to-day's small sample does not mean the same as 5% in to-morrow's large one" (p.189).

[2] One could almost argue that the true paradox is that this consistency is overlooked in most commentaries on the Jeffreys–Lindley's paradox.

At a second level of interpretation for the above setting, if we shift the meaning of n from being a sample size to being a prior scale factor, namely if we set that the prior variance is n times larger than the observation variance (or that the prior is n times less precise),[3] the result derived from the above expression is that *when the scale n goes to infinity, the Bayes factor goes to infinity no matter what the value of the observation is*. (Note that both interpretations are mathematically equivalent.) Now, under this new light, n becomes what Lindley (1957) calls "a measure of lack of conviction about the null hypothesis" (p.189), a sentence that I re-interpret as the prior (under H_1) getting more and more diffuse as n grows. I must however stress that nowhere in the paper is the difficulty with improper (or very large variance) priors discussed.

In this perspective, I also consider that the phenomenon still is not a paradox *per se*: when the diffuseness of the (alternative) prior (i.e., under H_1) increases, the only relevant piece of information becomes that θ could be equal to θ_0, to the extent that it overwhelms any evidence to the contrary contained in the data. For one thing, and as put by Lindley (1957), "the value θ_0 is fundamentally different from any value of $\theta \neq \theta_0$, however near θ_0 it might be" (p.189).[4] For another thing, the mass of the prior distribution in the vicinity of any fixed neighbourhood of the null hypothesis and even in any set coherent with the data at hand vanishes to zero. There is therefore a deep coherence in the selection of the null hypothesis H_0 in this case: being completely indecisive about the alternative hypothesis means we simply should not chose it. It is impossible to pick the alternative hypothesis against the very special value θ_0 if we want to be "completely non-informative" about θ under H_1. Depending on one's perspective about Bayesian statistics, one might see this as a strength or as a weakness since Bayes factors and posterior probabilities do require a realistic model under the alternative when p-values and Bayesian predictives do not.

3. Don't be afraid...

Under the provocative[5] title of *"Who should be afraid of Lindley's paradox"*,[6] Spanos (2013) offers his frequentist reassessment of the paradox, arguing against both Bayesian and likelihood ratio approaches and in favour of the postdata severity evaluation he and Mayo have both been advocating since 2004.

First, let me stress that the notion of *evidence* is never defined throughout the paper, even though it is repeatedly mentioned therein. My experience is that the notion widely fluctuates according to its user, ranging from vague facts to

[3] Or yet that, in terms of de Finetti's imaginary observations, the prior corresponds to the information brought by *one single* imaginary observation, as opposed to n real observations.

[4] We will get back to this fundamental remark in the discussion of Spanos (2013) in the next section.

[5] Although the overall style of the paper is quite antagonistic, I will not produce here evidence towards the rethorical devices used therein, concentrating on the statistical aspects and on their bearings on a re-analysis of the foundations of our field.

[6] Given the contents of the paper, the author presumably intends Bayesian statistics or Bayesians as the recipient of this question.

specific mathematical constructs (see, e.g., Skilling, 2006). Neither is the specific purpose of conducting a test (against, say, constructing a confidence interval) discussed at all. Reading the discourse of Spanos (2013) makes it sound as though there were an obvious truth (H_0 or H_1) and as though one and only one statistical approach could reach it, despite the evidence (!) to the contrary brought by the consistency of the three approaches in Lindley's (1957) setting.[7] Indeed, what differentiates tests from other aspects of inference is that (a) there is a question being asked about the statistical model under study and (b) the answer to this question will impact the subsequent actions of the individual who asked the question. Point (a) relates to Lindley's (1957) stress on the fact that θ_0 is very special indeed and quite different from any neighbouring value: it was select for a reason and with a motive, brought forward by a theoretical construct rather than inspired from the data. From a Bayesian perspective, this implies prior information is available as to why θ_0 is a special value of the parameter θ. Point (b) is about assessing the consequences of the answer to the questions, especially the wrong answer. Both from a frequentist and from a Bayesian perspective, this implies defining a loss or utility function that quantifies the impact of a wrong answer and eventually determines the boundary between acceptance and rejection.[8] Unfortunately, the remark "the problem does not lie with the p-value or the accept/reject rules as such, but with how such results are transformed into evidence for or against H_0 or a particular alternative" (p.76) does not proceed into a decisional step but instead into the introduction of a secondary p-value bound, the *severity evaluation*, coupled with a parameter value that requires a distance from the null and *in fine* an implicit loss function determining what is far and what is not. For instance, when Spanos (2013, p.75) states that "there is nothing fallacious or paradoxical about a small p-value or a rejection of the null, for a given significance level α; when n is large enough, since a highly sensitive test is likely to pick up on tiny (in a substantive sense) discrepancies from H_0", the "substantive sense" can only be gathered from a loss function. The conclusion that "what goes wrong is that the Bayesian factor and the likelihoodist procedures use Euclidean geometry to evaluate evidence for different hypotheses when in fact the statistical testing space is curved" (p.90) is mathematically meaningless when considering that the Bayes factor is invariant under one-to-one reparameterisation, hence impervious to the curvature of both the parameter and the sampling spaces.

Second, Spanos (2013) argues that the Jeffreys-Lindley paradox is demonstrating against the Bayesian (and likelihood) resolutions of the problem for failing to account for the large sample size.[9] I do not disagree with this perspective to the extent that I consider that the most important lesson learned

[7] Ironically, the numerical example used in the paper (borrowed from Stone, 1997, also father to the marginalisation paradoxes, see Dawid et al., 1973) is the very same as Bayes's billiard example (if with a larger value of n) and as Laplace's example on births (with a similar value of n).

[8] This is the simplest type of loss function: more advanced versions could include the case of a non-decision, calling for more observations, as in Berger (2003).

[9] The argument about the invariance of the Bayes factor to n (p.84) is found missing as the Bayes factor does depend on n as exhibited by $\mathfrak{B}(t_n)$ above.

from Lindley (1957) is that vague priors require special caution when conducting point-null hypothesis testing. There seems indeed to be little sense in arguing in favour of a procedure that would always conclude by picking the null, no matter what the value of the test statistics is. However, as already stressed in the introduction, considering a *fixed value* of the t statistic has little meaning in an asymptotic referential, i.e. when n increases to ∞. Either the t statistic converges in distribution to the t distribution under the null hypothesis H_0 or it diverges to infinity under the alternative H_1. This is the reason why both the Bayesian and the likelihood ratio approaches are consistent in this setting.[10]

In a global pondering about hypothesis testing, I would actually argue that the Jeffreys-Lindley paradox expresses difficulties for all of the three methodological threads: when following Fisher's approach, there is a theoretical and practical difficulty as to one should decrease the acceptace bound $\alpha = \alpha(n)$ on the p-value when n increases. It fails to provide a principle on which this bound (or sequence of bounds) $\alpha(n)$ should be chosen. For instance, the paper mentions (p.78) that because "of the large sample size, it is often judicious to choose a small type I error, say $\alpha = .003$" but this sentence simply points at the arbitrariness of this numerical value. Or, worse, that it was dictated by the data since the observed p-value takes the nearby .0027 value. In addition, I find the argument of consistence inconvincing in that case since both the Bayes factor and the likelihood ratio tests are then consistent testing procedures. In the Neyman-Pearson referential, I have a difficulty in finding a proper balance or imbalance between Type I and Type II errors, since such balance is not provided by the theory, which settles for the sub-optimal selection of a *fixed* Type I error. In addition, I have troubles with the whole notion of power, due to the fact that it is a function that depends on the unknown parameter. In particular, the power decreases to the Type I error at the boundary of the parameter set between the null and the alternative hypotheses. Without a prior distribution, giving a meaning to something like (eon. (25), p.87)

$$\mathbb{P}(\mathbf{x}; d(\mathbf{X}) < d(\mathbf{x}_0); \theta > \theta_1 \text{ is false})$$

seems impossible.[11] As discussed further in other sections of this note, apart from the genuine difficulty in setting a prior distribution, following a standard Bayesian approach with a flat prior on the binomial probability infered about in Spanos (2013) leads to a Bayes factor of 8.115 (p.80). Since this is neither a huge nor a tiny quantity *per se*, the very difficulty is in calibrating it, Jeffreys's (1939, Appendix) scale being highly formal.

Third, Spanos (2013) uses the failures (or fallacies?) of all three main approaches to address the difficulties with the Jeffreys–Lindley paradox to advocate his own criterion the "postdata severity evaluation" introduced in an

[10] In connection with this point, I fail to understand why a Bayes factor would "ignore the sampling distribution (...) by invoking the likelihood principle" (p.90): the Bayes factor incorporates the sampling distribution by integrating out against the associated prior under the alternative hypothesis.

[11] After an exchange with D. Mayo (2013, personal communication), it appears that this probability is computed under the distribution of $\mathbf{X}$ associated with the parameter θ_1.

earlier paper with Deborah Mayo (Mayo and Spanos, 2004).[12] The notion of severe tests has been advocated by Mayo and Spanos over the past years, but it has not yet had any impact on the practice of statistics: in my opinion, the solution seems to require even further calibration than the regular p-value and it is thus bound to confuse practitioners. Indeed, the severity evaluation as explained[13] in Spanos (2013) implies defining for each departure from the null, $\theta_1 = \theta_0 + \gamma$ the probability that a dataset associated with this parameter values "accords less with $\theta > \theta_1$ than x_0 does" (p.87). (Note that the two-sided alternative has been turned postdatum into a one-sided version.) This notion is therefore a mix of p-value and of type II error that is supposed to "provide the 'magnitude' of the warranted discrepancy from the null" (p.88), i.e. to decide about how close (in distance) to the null we can get and still be able to discriminate the null from the alternative hypotheses. As discussed in the paper, the value of this closest discrepancy γ—which is thus a bound on when we can discriminate between H_0 and H_1 at a given sample size—does depend on another arbitrary tail probability, the "severity threshold",

$$\mathbb{P}_{\theta_1}\{d(\mathbf{X}) \leq d(x_0)\}\,,$$

since this probability has to be chosen by the experimenter without being more intuitive than the initial acceptance bound on the p-value.[14] Further, once the resulting discrepancy γ is found, whether it is far enough from the null is a matter of informed opinion as, as duly noted by Spanos (2013), whether it is "substantially significant (...) pertains to the substantive subject matter" (p.88), implying once more some sort of loss function that is ignored (or implicit) throughout the paper.[15]

In connection with the special meaning of the value θ_0, several parts of Spanos' discussion of the Bayesian approach argue (see, e.g., p.81) about other values of θ that are supported and even better supported by the data than the null value θ_0. This is a surprising argument as it pertains to the construction of Bayesian credible intervals but not to testing. While it is correct that the observed data $\mathbf{x}_0$ does "favor certain values more strongly" (p.81) than θ_0, those

[12] Section 6 starts with the mathematically puzzling argument that, since we have observed x_0, the sign of $x_0 - \theta_0$ "indicates the relevant direction of departure from H_0". First, random variables may take values both sides of θ_0 for most values of θ. Second, the fact that one is testing H_0 against a two-sided or a one-sided alternative hypothesis pertains to the motivation of the test, not to the direction suggested by the data. The contentious modification of the testing setting *once* the data is observed is an issue with Spanos' (2013) perspective that we will discuss further.

[13] Let me remark that typos in both the last line in p.87, which is mixing the standardised and the non-standardised versions of the test statistic, and Table 1, which introduces a superfluous minus sign, do not help in clarifying the issue.

[14] When considering the severity as a function of θ_1, complement to a probability cdf in θ_1, the most natural interpretation would be Bayesian, the bound being a quantile. However, this solution is quite improbable to meet with the authors' approval.

[15] While this is very much unlikely to be advocated either by the author or by Bayesian statisticians, we note that, as a statistics, i.e. a transform of the data, both the Bayes factor and the likelihood ratio could be processed in exactly the same way to produce severity thresholds of their own.

values are (a) driven by the data, i.e. will vary from one repetition of the experiment to the next, and (b) of no particular relevance for conducting a test, meaning that the experimenter or the scientist behind the experiment had not expressed a particular interest in those values before they were exposed by the data. The tested value, $\theta_0 = 0.2$ say, is chosen prior to the experiment because it has some special meaning for the problem at hand. The fact that the likelihood and the posterior are larger in other values of θ does not "constitute conflicting evidence" against the fact that the null hypothesis holds. Or does not hold. It simply reflects on the fact that the likelihood function is a random function of the parameter θ, whose mode also varies with the data and is almost surely not located at the true value of the parameter. Even under the null.

4. On some resolutions of the Bayesian version

While the divergence between the frequentist and Bayesian answers is reflecting upon the difference between the paradigms in terms of purpose and evaluation, the (Bayesian) debate about constructing limiting Bayes factors or posterior probabilities that include improper prior modelling stands both open and relevant. DeGroot's (1982) warning that "diffuse prior distributions (...) must be used with care" has now been impressed upon generations of students and it is indeed a fair warning. There remains nonetheless a crucial need to produce assessments of null hypotheses from a Bayesian perspective and under limited prior information, once again without any incentive whatsoever to mimic, reproduce or even come close to frequentist solutions like p-values. (I will therefore abstain from covering here the notion of *matching priors*, whose sole purpose is to bring frequentist and Bayesian coverages as close as possible, see e.g. Datta and Mukerjee, 2004.)

In Robert (1993), I suggested selecting the prior weights of the two hypotheses, $(\varrho_0, 1-\varrho_0)$, in order to compensate for the increased mass brought by the alternative hypothesis prior.[16] While the solution therein produced numerical results that brought a proximity with the p-value, its construction is flawed from a measure-theoretic point of view since the determination of the weights involves the value of the prior density π_1 at the point-null value θ_0,

$$\varrho_0 = (1-\varrho_0)\pi_1(\theta_0)\,,$$

a difficulty also shared by the (related) Savage–Dickey paradox (Robert and Marin, 2009).[17] I nonetheless remain of the opinion that the degree of freedom represented by the prior weight ϱ_0 in the Bayesian formalism should not be neglected to overcome the difficulty in using improper priors.[18]

[16] The compensation cannot be probabilistic in that the overall mass of an improper prior will remain improper.

[17] A solution to the measure-theoretic difficulty is to impose a version of π_1 that is continuous at θ_0 so that $\pi_1(\theta_0)$ is uniquely defined. It however equates the values of two density functions under two orthogonal measures.

[18] Some will object at this choice on Bayesian grounds as it implies that the prior does depend on the sample size n.

Another direction worth pursuing is Berger et al.'s (1998) partial validation of the use of *identical* improper priors on the nuisance parameters, a notion already entertained by Jeffreys (1939, see the discussion in Robert et al., 2009, Section 6.3). While arguing about the "same" constant in both models towards using the "same" improper prior for both models has no mathematical nor statistical validation, using the same prior eliminates quite conveniently the major thorn in the side of Bayesian testing of hypotheses. As demonstrated in Marin and Robert (2007) and Celeux et al. (2012), it allows in particular for the use of a partly improper g-prior in linear and generalised linear models (Zellner, 1986).[19]

Yet another resolution to the paradox is apparently found in DeGroot's (1982) recommendation to keep "in mind that the assignment of a prior distribution to the parameter θ induces a predictive distribution for the observation" (p.337), as comparing predictives allows for an assessment of Bayesian models (meaning that either the sampling or the prior distribution may be inadequate). However, I think Morrie DeGroot meant in this text using the prior predictive,

$$m(y) = \int_\Theta \pi(\theta) f(y|\theta)\,\mathrm{d}\theta\,.$$

in which case this approach is essentially equivalent to the Bayes factor, hence does not solve the improperness issue, and suffers from the same calibration difficulty. If, instead, one considers the posterior predictive, this is the solution advocated in, among others, Gelman et al. (2003), under the name of *posterior predictive checking*, but it implies using the data twice (once for building the posterior and one for deriving the assessment), and has been reinterpreted by Aitkin (1991, 2010) in his integrated likelihood theory, drawing strong criticism from many, including Dennis Lindley's now famous "One hardly advances the respect with which statisticians are held in society by making such declarations" (1991, p.131). (See also Gelman et al., 2013.)[20]

A last direction worth investigating is the recent development of the use of score functions $S(x, m)$ that extend the log score function associated with the Bayes factor:

$$\log B_{12}(x) = \log m_1(x) - \log m_2(x) = S_0(x, m_1) - S_0(x, m_2)\,,$$

where m_i is the prior predictive associated with model $\mathfrak{M}_i$. Indeed, there exists a whole family of proper scoring rules that are independent from the normalising

[19] Once again, choosing $g = n$ should attract criticism from some Bayesian corners for being dependent on the sample size, even though it boils down to picking an imaginary sample (Smith and Spiegelhalter, 1982) size of 1. See Liang et al. (2008) for an alternative approach setting an hyperprior on g.

[20] Although a huge literature has been dedicated to partial Bayes factors like fractional and intrinsic Bayes factors where a part of the dataset is used to make the posterior distribution well-defined and the remainder addresses the testing question, as started in Smith and Spiegelhalter (1982), I will not pursue this direction as (a) it is very rarely a truly Bayesian procedure, i.e. cannot be expressed as a genuine Bayes factor against a pair of proper prior distributions, and (b) it suffers from facing too many competing variants to be advocated. See e.g. Berger and Pericchi (2001) or Robert (2001) for a review.

constant of the prior predictive (Parry et al., 2012) and can thus be used on improper priors as well. For instance, Hyvärinen's (2005) score is one of these scores. While the scores are delicate to calibrate, i.e. the magnitude of $S(x, m_1) - S(x, m_2)$ is not absolute, they provide a consistent method for selecting models (REF) and avoid the delicate issue of selecting priors that differ for model selection and for regular inference (conditional on the model).

5. Reflections

The appeal of great paradoxes[21] is to exhibit foundational issues in a field, either to reinforce the arguments in favour of a given theory or, on the opposite, to cast serious doubts on its validity. The fact that the Jeffreys–Lindley's paradox is still discussed in papers (as exemplified by the recent Spanos, 2013) and blogs, by statisticians and non-statisticians alike, is a testimony to its impact on the debate about the very nature of (statistical) testing. The irrevocable opposition between frequentist and Bayesian approaches to testing, but also the persistent impact of the prior modelling in this case, are fundamental questions that have not yet met with definitive answers. And they presumably never will for, as put by Lad (2003), "the weight of Lindley's paradoxical result (...) burdens proponents of the Bayesian practice". However, this is a burden with highly positive features in that it paradoxically (!) drives the field to higher grounds.[22]

References

Aitkin, M. (1991). Posterior Bayes factors (with discussion). *J. Royal Statist. Society Series B*, **53** 111–142.

Aitkin, M. (2010). *Statistical Inference: A Bayesian/Likelihood approach*. CRC Press, Chapman & Hall, New York.

Berger, J. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, **18** 1–32.

Berger, J. and Pericchi, L. (2001). Objective Bayesian methods for model selection: introduction and comparison. In *Model Selection* (P. Lahiri, ed.), vol. 38 of *Lecture Notes – Monograph Series*. Institute of Mathematical Statistics, Beachwood Ohio, 135–207.

Berger, J., Pericchi, L. and Varshavsky, J. (1998). Bayes factors and marginal distributions in invariant situations. *Sankhya A*, **60** 307–321.

Berger, J. and Sellke, T. (1987). Testing a point-null hypothesis: the irreconcilability of significance levels and evidence (with discussion). *J. American Statist. Assoc.*, **82** 112–122.

Camus, A. (1942). *Le Mythe de Sisyphe*. Gallimard, Paris.

[21] I use this term despite my reluctance to call such phenomena "paradoxes", since they do not correspond to logical impossibilities, but rather to contradictions in reasoning or, in the current case, to attempts to bring two different paradigms together.

[22] To conclude with a literary quote, "il faut imaginer Sisyphe heureux" (Camus, 1942).

Celeux, G., Anbari, M. E., Marin, J. and Robert, C. (2012). Regularization in regression: Comparing Bayesian and frequentist methods in a poorly informative situation. *Bayesian Analysis*, **7** 477–502.

Datta, G. and Mukerjee, R. (2004). *Probability matching priors: higher order asymptotics*. Springer-Verlag, New York.

Dawid, A., Stone, N. and Zidek, J. (1973). Marginalization paradoxes in Bayesian and structural inference (with discussion). *J. Royal Statist. Society Series B*, **35** 189–233.

DeGroot, M. (1973). Doing what comes naturally: Interpreting a tail area as a posterior probability or as a likelihood ratio. *J. American Statist. Assoc.*, **68** 966–969.

DeGroot, M. (1982). Discussion of Shafer's 'Lindley's paradox'. *J. American Statist. Assoc.*, **378** 337–339.

Gelman, A., Carlin, J., Stern, H. and Rubin, D. (2003). *Bayesian Data Analysis*. 2nd ed. Chapman and Hall, New York, New York.

Gelman, A., Robert, C. and Rousseau, J. (2013). Inherent difficulties of non-Bayesian likelihood-based inference, as revealed by an examination of a recent book by Aitkin. *Statistics and Risk Modelling*. To appear.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, **6** 695–709. (electronic).

Jeffreys, H. (1939). *Theory of Probability*. 1st ed. The Clarendon Press, Oxford.

Kass, R. and Wasserman, L. (1996). Formal rules of selecting prior distributions: a review and annotated bibliography. *J. American Statist. Assoc.*, **91** 343–1370.

Lad, F. (2003). Appendix: the Jeffreys–Lindley paradox and its relevance to statistical testing. Tech. rep., Conference on Science and Democracy, Palazzo Serra di Cassano, Napoli.

Liang, F., Paulo, R., Molina, G., Clyde, M. and Berger, J. (2008). Mixtures of g priors for Bayesian variable selection. *J. American Statist. Assoc.*, **103** 410–423.

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

Lindley, D. (1991). Discussion of the paper by Aitkin. *J. Royal Statist. Society Series B*, **53** 130–131.

Marin, J. and Robert, C. (2007). *Bayesian Core*. Springer-Verlag, New York.

Mayo, D. and Spanos, A. (2004). Methodology in practice: Statistical misspecification testing. *Philosophy of Science*, **71** 1007–1725.

Parry, M., Dawid, A. and Lauritzen, S. (2012). Proper local scoring rules. *Ann. Statist.* 561–592.

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

Robert, C. (2001). *The Bayesian Choice*. 2nd ed. Springer-Verlag, New York.

Robert, C., Chopin, N. and Rousseau, J. (2009). Theory of Probability revisited (with discussion). *Statist. Science*, **24(2)** 141–172 and 191–194.

Robert, C. and Marin, J.-M. (2009). On resolving the Savage–Dickey paradox. *Elect. J. Statistics*.

Shafer, G. (1982). On Lindley's paradox (with discussion). *J. American Statist. Assoc.*, **378** 325–351.

Skilling, J. (2006). Nested sampling for general Bayesian computation. *Bayesian Analysis*, **1(4)** 833–860.

Smith, A. and Spiegelhalter, D. (1982). Bayes factors for linear and log-linear models with vague prior information. *J. Royal Statist. Society Series B*, **44** 377–387.

Spanos, A. (2013). Who should be afraid of the Jeffreys–Lindley paradox? *Philosophy of Science*, **80** 73–93.

Stone, M. (1997). Discussion of aitkin (1997). *Statistics and Computing*, **7** 263–264.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distribution regression using Bayesian variable selection. In *Bayesian inference and decision techniques: Essays in Honor of Bruno de Finetti*. North-Holland / Elsevier, 233–243.