Bayesian hypothesis tests with diffuse priors: Can we have our cake and eat it too?

J. T. Ormerod\textsuperscript{1}, M. Stewart\textsuperscript{1}, W. Yu\textsuperscript{2,*} and S. E. Romanes\textsuperscript{1}

\textit{The University of Sydney and The University of Melbourne}

Summary

We propose a new class of priors for Bayesian hypothesis testing, which we name ‘cake priors’. These priors circumvent the Jeffreys–Lindley paradox (also called Bartlett’s paradox) a problem associated with the use of diffuse priors leading to nonsensical statistical inferences. Cake priors allow the use of diffuse priors (having one’s cake) while achieving theoretically justified inferences (eating it too). We demonstrate this methodology for Bayesian hypotheses tests for various common scenarios. The resulting Bayesian test statistic takes the form of a penalised likelihood ratio test statistic. Under typical regularity conditions, we show that Bayesian hypothesis tests based on cake priors are Chernoff consistent, that is, achieve zero type I and II error probabilities asymptotically. We also discuss Lindley’s paradox and argue that the paradox occurs with small and vanishing probability as sample size increases.

\textbf{Key words:} asymptotic properties of hypothesis tests; Chernoff consistency; Jeffreys–Lindley–Bartlett paradoxes; improper priors; likelihood ratio tests; linear models.

1. Introduction

Determining appropriate parametric prior distributions is of paramount importance in Bayesian hypothesis testing. Bayesian hypothesis testing often centres around the concept of a Bayes factor, which was initially developed by Jeffreys (1935, 1961) and later popularised by Kass & Raftery (1995). The Bayes factor is simply the odds of the marginal likelihoods between two hypotheses and is analogous to the likelihood ratio statistic in classical statistics, where the model parameters are marginalised out instead of maximising the likelihoods with respect to the model parameters. In classical statistical theory, testing a simple point null hypothesis against a composite alternative is routine. However, such hypothesis tests can pose severe difficulties in the Bayesian inferential paradigm where the Bayes factors may exhibit undesirable properties unless parameter prior distributions are chosen with exquisite care. This paper proposes a solution to this difficulty.

*Author to whom correspondence should be addressed.

\textsuperscript{1}School of Mathematics and Statistics, The University of Sydney, Sydney, NSW 2006, Australia.

\textsuperscript{2}School of Mathematics and Statistics, The University of Melbourne, Parkville, VIC 3010, Australia. e-mail: weichang.yu@unimelb.edu.au

Acknowledgements. We would like to thank the reviewers for their comments which helped to improve the paper. The following source of funding is gratefully acknowledged: Australian Research Council Discovery Project grant (DP210100521) to J. T. Ormerod. Open access publishing facilitated by The University of Melbourne, as part of the Wiley - The University of Melbourne agreement via the Council of Australian University Librarians.

© 2024 The Authors. \textit{Australian & New Zealand Journal of Statistics} published by John Wiley & Sons Australia, Ltd on behalf of Statistical Society of Australia. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
Prior distributions can be chosen in an informative or uninformative fashion. Employing informative priors (based on either data from previously conducted experiments or eliciting priors from subject matter experts) can be impractical, particularly when the number of parameters in the model is large. Furthermore, informative priors can be criticised on grounds that such priors are inherently subjective or may not let the data from the current experiment speak for itself. However, using alternative priors can also lead to problems.

One such problem occurs when using overly diffuse or flat improper priors. In the former case, as priors become more diffuse the hypothesis corresponding to the smaller model becomes increasingly favoured regardless of the evidence provided by the data. This problem occurs due to the normalising constants of the priors dominating the expression for the Bayes factor, and is sometimes referred to as Bartlett’s paradox (e.g., Liang et al. 2008) or the Jeffreys–Lindley paradox (e.g., Robert 1993, 2014) named after the pioneering work of Jeffreys (1935, 1961), Lindley (1957) and Bartlett (1957), whose authors identified this and other related problems associated with Bayes factors. Discussions of this paradox and the related Lindley’s paradox can be found in Aitkin (1991), Bernardo (1999), Sprenger (2013), Spanos (2013) and Robert (2014).

An extension of Bartlett’s paradox occurs in the limiting case where diffuse priors become flat to the point of being improper. The use of flat improper priors gives rise to arbitrary constants in the numerator and the denominator of the Bayes factor (see DeGroot 1973). Such arbitrary constants are problematic since, without suitable modification, they could be chosen by the analyst to suit any preconceived conclusions preferred, and as such, are not suitable for scientific purposes. Techniques for selecting the arbitrary constants in Bayes factors in an acceptable way when employing flat improper priors have been developed in several papers. Bernardo (1980) proposes to derive a reference prior for the null hypothesis by maximising a measure of missing information. Spiegelhalter & Smith (1982) and Pettit (1992) use an imaginary data device leading to the arbitrary constants cancelling with other terms in the Bayes factor. A further approach to the problem of using diffuse priors was proposed by Robert (1993) who advocated for reweighing the prior odds to balance against parameter priors as prior hyperparameters become diffuse.

O’Hagan (1995) considers the problem of using flat improper priors in the calculation of the Bayes factor by splitting the data into a training and testing set. The training set is used to construct an informative prior. This informative prior is then used to calculate the Bayes factor based on the remaining portion of the data. These ideas have been refined in O’Hagan (1997), Berger & Pericchi (1996), and Berger & Pericchi (2001). A computational drawback of some of these approaches is that the same model is fitted multiple times. For models where Bayesian inferential procedures are considered too slow for fitting a single model, these approaches to Bayesian testing become infeasible from a practical viewpoint.

Other Bayesian hypothesis testing approaches abandon the Bayes factor altogether by constructing hypothesis testing criterion, which only enter the criterion through parameter posterior distributions themselves. These include information criteria type approaches such as the Bayesian information criterion (BIC) and the deviance information criterion (DIC). The BIC uses a Laplace approximation where the prior term is assumed to be asymptotically negligible as the sample size grows (Schwarz 1978). The DIC involves a linear combination of the log-likelihood evaluated at a suitably chosen Bayesian point.
estimator and the posterior expectation of the log-likelihood (Spiegelhalter et al. 2002).
Under such a construction, the DIC is not dominated by the prior as prior hyperparameters
diverge. Similarly, posterior Bayes factors proposed by Aitkin (1991) are based on the
posterior expectation of the likelihood function rather than the joint likelihood (comprising
the model likelihood and prior). Since this only involves the prior in the calculation of
the posterior distribution, the prior does not dominate posterior Bayes factors. Berger &
Pericchi (1996) criticised this approach because it employs a double use of the data that
are inconsistent with typical Bayesian logic.

An interesting alternative approach to Bayesian hypothesis testing is suggested in
section 6.3 of Gelman et al. (2013), which investigates the properties of the posterior
distributions of carefully chosen test statistics such that large values of a given test statistic
provide evidence against the null hypotheses. This idea is explored more formally in
Gelman, Meng & Stern (1996) and gives rise to the concept of posterior predictive
values—the probability that a test statistic of posterior predictive values is greater than the
observed value of the test statistic.

Bayes factors in the context of linear model selection (Zellner & Siow 1980; Mitchell
& Beauchamp 1988; George & McCulloch 1993; Fernández, Ley & Steel 2001; Liang
et al. 2008; Maruyama & George 2011; Bayarri et al. 2012) and generalised linear
model selection (Hansen & Yu 2001; Chen & Ibrahim 2003; Wang & George 2007;
Chen et al. 2008; Gupta & Ibrahim 2009; Bové & Held 2011; Hanson, Branscum &
Johnson 2014; Li & Clyde 2015) have received an enormous amount of attention. While
we defer discussion of the types of priors used in these contexts to Section 3.5, we will draw
special attention to Liang et al. (2008). Liang et al. (2008) consider several prior structures
in the context of linear models. They employ Zellner’s $g$-prior (Zellner & Siow 1980;
Zellner 1986) for the regression coefficients where $g$ is a prior hyperparameter. They
consider several choices of $g$ including setting $g$ to various constants, selecting $g$ using a
local and global empirical Bayes procedure, and assigning a hyperprior on $g$. Their results
suggest that in order for the resulting Bayes factors to be well-behaved (e.g. model selection
consistency), a hyperprior needs to be placed on $g$.

In this paper, we will construct a new class of priors, which was inspired by the
priors used in the context of linear and generalised linear models. This class of priors is
constructed in such a way that the resultant Bayes factor is equivalent (in the limiting case)
to a Bayes factor based on Jeffreys priors (Jeffreys 1946) and is therefore invariant to
any parameter transformation. In order to circumvent a Jeffreys–Lindley-like paradox from
occurring, the rate at which $g$ diverges is different in the null and alternative hypothesis in
such a way that results in the cancellation of problematic terms in both the numerator and
denominator of the Bayes factor.

Bayes factors using cake priors have several desirable properties. In the examples
we consider, the Bayes factor can be expressed as a difference in BIC values, that is, a
penalised version of the likelihood ratio test (LRT) statistic. Using properties of the LRT
statistic, we show that Bayesian hypothesis tests are Chernoff consistent (Shao 2003, section
2.13), that is, they achieve asymptotically zero type I and type II error probabilities as the
sample size diverges. In contrast, classical hypothesis testing procedures are usually chosen
to have a fixed type I error probability and are consequently not Chernoff consistent. In
this respect, our Bayesian hypothesis tests are superior to classical procedures where type
I error probability is held fixed. Due to the above properties, we call the priors we develop
‘cake priors’ since they allow the use of diffuse priors (having one’s cake) while being able to perform sensible statistical inferences (eating it too). We will also discuss Lindley’s paradox in the context of cake priors and argue that generally Lindley’s paradox will only occur with vanishingly small probability for large samples.

In Section 2, we review Bayes factors and discuss more specifically the problems associated with Bayes hypothesis testing using Bayes factors, including both Lindley’s and the Jeffreys–Lindley paradoxes. In Section 3, we describe cake priors and illustrate their use in several examples. We also provide suggestions on how we can overcome various challenges in implementing our proposed prior. In Section 4, we present asymptotic results for Bayesian hypothesis tests based on our proposed prior. Section 5 concludes.

2. Bayes factors

Suppose that we have observed the data vector \( \mathbf{x} = (x_1, \ldots, x_n)^\top \) which are samples from \( p^* (\cdot) = \prod_{i=1}^n p_i^* (\cdot) \) and we have two hypotheses \( H_0 \) and \( H_1 \) representing two models \( \mathcal{P}_j = \{ p_{ij} (\cdot | \theta_j, H_j) : i = 1, \ldots, n \}, \ j = 0, 1 \) that describes two potential models in which \( p^* \) belongs to, that is,

\[
H_0 : p^* \in \mathcal{P}_0 \quad \text{versus} \quad H_1 : p^* \in \mathcal{P}_1.
\]  

The models could potentially have distinct parameterisations and need not be nested. Let \( \pi (\theta_j | H_j) \) be the prior distribution under hypothesis \( H_j \) for \( j = 0, 1 \). The Bayes factor is then defined as

\[
BF_{01} = \frac{p (\mathbf{x} | H_0)}{p (\mathbf{x} | H_1)} = \frac{\int p (\mathbf{x} | \theta_0, H_0) \pi (\theta_0 | H_0) d \theta_0}{\int p (\mathbf{x} | \theta_1, H_1) \pi (\theta_1 | H_1) d \theta_1},
\]

where integrals are replaced with combinatorial sums for discrete random variables. A Bayesian hypothesis test function \( T (\mathbf{x}) \in \{ 0, 1 \} \) is based on

\[
T (\mathbf{x}) = \begin{cases} 
1 & \text{implies } H_1 \text{ is preferred; and} \\
0 & \text{implies } H_0 \text{ is preferred.}
\end{cases}
\]

When \( p (H_0) = p (H_1) = 1/2 \), the Bayes factors can be interpreted as such: when \( BF_{01} \) is above 1 the hypothesis \( H_0 \) is favoured and when \( BF_{01} \) is below 1 the hypothesis \( H_1 \) is favoured. The evidence strength may be determined through the magnitude of \( \lambda_{\text{Bayes}} \), and \( BF_{10} = 1/BF_{01} \) as suggested by Kass & Raftery (1995) (refer to Table 1 for details).

Table 1. Table of interpretation of Bayes factors offered by Kass & Raftery (1995).

| \( \lambda_{\text{Bayes}} \) | \( BF_{10} \) | Strength of evidence |
|-----------------|---------|---------------------|
| 0 to 2          | 1 to 3  | Not worth more than a bare mention |
| 2 to 6          | 3 to 20 | Positive            |
| 6 to 10         | 20 to 150| Strong              |
| > 10            | > 150   | Very strong         |
Paradoxes in Bayesian hypothesis testing

Paradoxes in Bayesian hypothesis testing due to poor choices of priors have been identified in Jeffreys (1935), Lindley (1957) and Bartlett (1957). More specifically, a poorly specified prior may lead to undesirable inconsistencies as pointed out by critics of Bayesian hypothesis tests. To provide context and illustrate potential problems, we consider the hypothesis testing problem introduced by Lindley (1957) in order to illustrate potential problems.

Consider the hypothesis test where the sample is modelled via $x_i | \mu \sim N(\mu, \sigma^2), 1 \leq i \leq n$ independently, where $\mu$ and $\sigma^2$ are the mean and variance parameters, respectively. Here $\mu$ is an unknown value to be estimated and $\sigma^2$ is a fixed known constant. Suppose that we wish to perform the hypothesis test

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0,$$  \hspace{1cm} (3)

where $\mu_0$ is a known constant. Under $H_0$ the values of all model parameters are fixed (so that under $H_0$ the model has zero unknown parameters), that is, $H_0$ is a simple point null hypothesis. Suppose that for $H_1$ we employ the prior $\mu | H_1 \sim N(\mu_0, \tau^2)$ where the prior variance $\tau^2$ is a known constant. The Bayes factor with the stated prior on $\mu$ is

$$BF_{01} = \sqrt{1 + \frac{n\tau^2}{\sigma^2}} \exp \left[ -\frac{n z(x)^2}{2(n + \sigma^2/\tau^2)} \right],$$ \hspace{1cm} (4)

where $z(x) = \sqrt{n(\bar{x} - \mu_0)/\sigma}$ is the standard $z$-test statistic (see Bernardo 1999).

Lindley’s paradox: Here, we define Lindley’s paradox as the discordance between the frequentist and Bayesian hypothesis test conclusions. In the context of the above example, Lindley’s paradox is equivalent to the following event:

$$\left\{ 1 < BF_{01} < \sqrt{1 + \frac{n\tau^2}{\sigma^2}} \exp \left[ -\frac{n \chi^2_{1,\alpha}}{2(n + \sigma^2/\tau^2)} \right] \right\} \cup \left\{ \sqrt{1 + \frac{n\tau^2}{\sigma^2}} \exp \left[ -\frac{n \chi^2_{1,\alpha}}{2(n + \sigma^2/\tau^2)} \right] < BF_{01} < 1 \right\},$$

where $\chi^2_{1,\alpha}$ denotes the upper $\alpha$-quantile of the $\chi^2_1$-distribution. Following Result 2 in Section 4, Lindley’s paradox turns out to be a minor problem to the Bayesian as it is only problematic when the null hypothesis is true and $n$ is small. In fact, Bayesian hypothesis tests based on $\lambda_{\text{Bayes}}$ have vanishing type I error probability as $n \to \infty$. Moreover, under the alternative hypothesis, a true Lindley’s paradox only occurs with vanishingly small probability as $n \to \infty$. Both points imply that the Bayesian test is consistent in the sense of Lehmann (2004, section 3.3).

From an epistemological perspective, Lindley’s paradox is consequential to an incoherent expectation that one should obtain the same conclusions for frequentist and Bayesian hypothesis tests. This expectation is incoherent because it disregards the fundamental and philosophical differences between the two paradigms. More specifically, the Bayesian paradigm posits preference towards a particular hypothesis, whereas frequentist
hypothesis testing asks whether the data could have plausibly been drawn from the null model.

**Remark:** While there is no need to modify the Bayesian hypothesis testing paradigm to address issues with Lindley’s paradox, it is of paramount importance that any proposed refinement to the Bayesian hypothesis testing framework must retain the desirable properties of the standard framework including its asymptotic consistency properties.

**Jeffreys–Lindley paradox:** Here, we define the Jeffreys–Lindley’s paradox as the conundrum where the Bayesian hypothesis test always prefers $H_0$ as the prior variances for the parameters under $H_1$ diverge. Formally, we state the following well-known result.

**Result 1:** Consider the setup in Lindley’s example. Then, for any finite $n$ and $\sigma^2 > 0$, we have $\lim_{\tau^2 \to \infty} BF_{01} = \infty$.

In the context of inferring an unknown Gaussian population mean, Jeffreys–Lindley paradox is paradoxical since as $\tau^2$ increases the prior on $\mu$ becomes increasingly vague regarding the location of $\mu$. However, in the attempt to be vague about the location of $\mu$, the prior becomes ‘informative’ in the sense of favouring $H_0$ as the preferred hypothesis regardless of the evidence provided by the data. Moreover, in the estimation context, the Bayesian point estimate based on a fully diffuse prior ($\tau^2 \to \infty$) is equivalent to the frequentist point estimate but this property does not hold in the hypothesis testing context (Robert 1993). Unlike Lindley’s paradox, we believe that Jeffreys–Lindley paradox is a real problem in practice since the use of diffuse priors can sometimes lead to testing procedures with extremely small power. In fact, to enable the widespread adoption of the Bayesian hypothesis testing framework, there is a need to refine the existing framework to address the Jeffreys–Lindley paradox.

3. Cake priors

We propose a prior that circumvents the Jeffreys–Lindley’s paradox as described in Section 2.1. Consider the general hypotheses setup in (1). Let $d_0$ and $d_1$ be the dimensions of $\theta_0$ and $\theta_1$, respectively. At present, we assume that $0 < d_0 \leq d_1$ (we consider the case $0 \leq d_0 \leq d_1$ in Section 3.6). Define the observed information and Fisher information matrices as $J(\theta) = -\sum_{i=1}^{n} \nabla^2_{\theta} \ln p(x_i|\theta)$ and $I(\theta) = E_{x_i|\theta}[-\nabla^2_{\theta} \ln p(x_i|\theta)]$, respectively. Define the mean observed information matrix as $\tilde{J}(\theta) = n^{-1}J(\theta)$ and the mean expected observed information matrix as $\tilde{I}(\theta) = n^{-1}E_{x_i|\theta}[J(\theta)]$. We denote the Fisher information matrix under the null and alternative hypotheses as $I_0(\theta_0)$ and $I_1(\theta_1)$ with similar use of subscripts to denote similar quantities such as $J$, $\tilde{I}$ and $\tilde{J}$. We define a Jeffreys prior as any density for $\theta$ such that $\pi(\theta) \propto |I(\theta)|^{1/2}$.

We design cake priors to mimic Jeffreys priors as the prior becomes increasingly diffuse using the following recipe:

1. Define the priors

$$\pi(\theta_j | H_j; g_j) = \exp \left[ -\frac{d_j}{2} \ln(2\pi g_j) + \frac{1}{2} \ln |P_j(\theta_j)| - \frac{1}{2g_j} \theta_j^\top P_j(\theta_j) \theta_j \right], \quad (5)$$

where $P_j(\theta_j)$ is a prior precision matrix (assumed to be full rank). For all of the examples considered in this paper we will use $P_j(\theta_j) = \tilde{I}_j(\theta_j)$. 

© 2024 The Authors. *Australian & New Zealand Journal of Statistics* published by John Wiley & Sons Australia, Ltd on behalf of Statistical Society of Australia.
2. Set $g_j = h^{1/d_j}$ where $h$ is a common hyperparameter (see below for discussion).
3. Calculate the Bayes Factor as

$$BF_{01}(h) = \frac{\int p(x|\theta_0, H_0)\pi(\theta_0|H_0; h^{1/d_0})d\theta_0}{\int p(x|\theta_1, H_1)\pi(\theta_1|H_1; h^{1/d_1})d\theta_1}.$$ 

When $P_j(\theta_j) \propto I_j(\theta_j)$, $j = 0, 1$, the limit of the Bayes factor in Step 3 ($h \to \infty$) is equivalent to a Bayes factor based on Jeffreys priors for both the null and alternative hypotheses. As $h \to \infty$, the priors on $\theta_j$ are made diffuse, but at a rate that depends on the $d_j$'s.

Firth (1993) showed that if the log-likelihood is penalised by Jeffreys’ invariant prior, then the resulting maximum penalised likelihood estimator has bias of smaller asymptotic order than that of the maximum likelihood estimator in general. Through this lens, the log-transform of the Bayes factor becomes

$$\exp\left\{\ln BF_{01}(h) - \ln BF_0(\theta_0)\right\} = \exp\left\{\ln BF_{01}(\theta_0) - \ln BF_0(\theta_0)\right\} = \exp\left\{\ln BF_{01}(\theta_1) - \ln BF_0(\theta_0)\right\}.$$

Unfortunately, it is unclear whether or not the proposed prior density in (5) is always bounded for all values of $\theta_j \in \Theta_j$. To ensure boundedness, we may replace $P_j(\theta_j)$ with $P_j(\hat{\theta}_j)$, where $\hat{\theta}_j$ is an estimator of $\theta_j$.

### 3.1. How does cake prior overcome the Jeffreys–Lindley paradox?

We provide some intuition for how cake priors avoid the Jeffreys–Lindley paradox via the following heuristic argument. Let $P_j(\theta_j) = P_j$, that is, the prior precision matrices are constant, then letting $BF_{01} = \left[\int p(x|\theta_0, H_0)\pi(\theta_0|H_0; g_0)d\theta_0\right]/\left[\int p(x|\theta_1, H_1)\pi(\theta_1|H_1; g_1)d\theta_1\right]$ (which depends on $g_0$ and $g_1$ rather than $h$). Then the Bayesian test statistic is

$\lambda_{\text{Bayes}} = 2 \ln \left[\frac{\int p(x|\theta_1, H_1)d\theta_1}{\int p(x|\theta_0, H_0)d\theta_0}\right] + d_0 \ln(2\pi g_0) - d_1 \ln(2\pi g_1) + \ln\left(\frac{|P_1|}{|P_0|}\right) + O_p(g_0^{-1} + g_1^{-1}),$

where the second line is obtained using a Taylor series argument in $g_0$ and $g_1$. Ignoring the dependency of $O_p(g_j^{-1})$ terms on the $\theta_j$'s, using Laplace’s method on the numerator and denominator of the first term in the second line above, and setting $P_j = \tilde{J}_j(\hat{\theta}_j)$ (where $\hat{\theta}_j$ are the MLEs for the $\theta_j$'s) leads to

$$\lambda_{\text{Bayes}} = \lambda_{\text{LRT}} - \nu \ln(n) + d_0 \ln(g_0) - d_1 \ln(g_1) + O_p\left(g_0^{-1} + g_1^{-1} + n^{-1}\right). \quad (6)$$

The $O_p(n^{-1})$ error follows from the relative error of the Laplace’s method applied to the numerator and denominator (Tierney, Kass & Kadane 1989; Kass, Tierney & Kadane 1990). Suppose that $g_0 = g_1 = g$. By noting that the asymptotic distribution of $\lambda_{\text{LRT}}$ is $\chi^2_\nu$, the
level of the test based on (6) is \( \alpha = \Pr [X^2_\nu \geq v \ln(n g)] \). Observe that the power of the test increases from 0 as \( g \to \infty \). Moreover, observe that setting \( g_0 \) to be a large constant (making the prior for \( \theta_0 \) diffuse) leads to a preference for \( H_0 \) while making \( g_1 \) large leads to a preference for \( H_1 \). Hence, the relative rates that \( g_0 \) and \( g_1 \) diverge must be considered.

Setting \( g_j = h^{1/d_j} \) means that \( d_0 \ln(g_0) = d_1 \ln(g_1) \) and leads to \( \lambda_{\text{Bayes}} = \lambda_{\text{LRT}} - v \ln(n) + O\left(h^{-1/d_0} + h^{-1/d_1} + n^{-1}\right) \). For sufficiently large \( h \) and \( n \) we have \( \lambda_{\text{Bayes}} \approx \lambda_{\text{LRT}} - v \ln(n) \). The level of the test becomes approximately \( \alpha = \Pr [X^2_\nu \geq v \ln(n)] \). The choice \( g_j = h^{1/d_j} \) is not driven by any theoretical desiderata other than leading to algebraic simplification.

Finally, we briefly discuss the choice of \( P_j \). Setting \( P_j = I \) leads to \( \lambda_{\text{Bayes}} = \lambda_{\text{LRT}} - v \ln(n) + \ln|\mathbf{J}_0(\hat{\theta}_0)| - \ln|\mathbf{J}_1(\hat{\theta}_1)| + O\left(h^{-1/d_0} + h^{-1/d_1} + n^{-1}\right) \). This would be undesirable because of the additional computational burden of the log-determinant terms (which can be considerable in some contexts), and because \( \lambda_{\text{LRT}} \approx v \ln(n) \) we would prefer the model with larger dispersion. For this reason we would like \( P_j \approx \mathbf{J}_j(\hat{\theta}_j) \) so that approximate cancellation occurs.

Remark: The cake prior in (5) may not always be a proper prior. If a proper prior is required, one may replace \( P_j(\theta_j) \) with \( P_j(\hat{\theta}_j) \) which is equivalent to the choice of prior: \( \theta_j | H_j, g_j \sim N(\theta_j, g_j P_j(\hat{\theta}_j)^{-1}) \), where \( \hat{\theta}_j \) is an estimator of \( \theta_j \). Here, a possible choice for \( \hat{\theta}_j \) is the MLE. Other frequentist estimators for \( \theta_j \) may also be possible. While reliance on data through an estimator of \( \theta_j \) is not desirable, such other priors, for example, the prior suggested by Hansen & Yu (2003), Gupta & Ibrahim (2009), and the generalised g-prior of Li & Clyde (2015), have also done this. However, the impact of such a choice becomes small as \( h \to \infty \).

### 3.2. Complications with the cake prior construction procedure

Complications may arise from the cake prior construction procedure for certain examples. These include:

(A) Arbitrary constants. In the above we have chosen \( g_j = h^{1/d_j} \). Choosing say, \( g_j = \kappa_j h^{1/d_j} \) for some constant \( \kappa_j \) leads to a different limit as \( h \to \infty \). The choice of \( \kappa_j \) is arbitrary.

(B) Model parameters may not be defined on the whole real line, for example, variances. Priors of the form (5) are not appropriate for such model parameters.

(C) The priors \( \pi(\theta_j | g_j, H_j) \) may not be proper densities.

(D) If \( d_0 = 0 \) using \( g = h^{1/d_j} \) is problematic.

Section 3.2 discusses (A) and ways to overcome it. Sections 3.3 and 3.4 will illustrate how complications (B) & (C) can be handled. Complication (D) is addressed in Section 3.6.

### 3.3. Arbitrary constants

Since cake priors may be improper, we may encounter the arbitrary constants issue. More specifically, consider the conditional density \( p(x|\theta_j, H_j) \), where \( \pi(\theta_j | H_j) = \kappa_j f_j(\theta_j) \) for some \( \kappa_j > 0 \) and \( \int f_j(\theta_j)d\theta_j = \infty \) for \( j = 0, 1 \). Then

\[
B_{01} = \frac{\kappa_0}{\kappa_1} \times \frac{\int p(x|\theta_0, H_0) f_0(\theta_0)d\theta_0}{\int p(x|\theta_1, H_1) f_1(\theta_1)d\theta_1}.
\] (7)
Observe that this Bayes factor depends on two arbitrary constants $\kappa_0$ and $\kappa_1$. Hence, a criticism of the use of improper priors is that either the null or the alternative model can be made to be preferred by artificially changing $\kappa_0$ or $\kappa_1$ to suit the a priori preferred conclusion. While the focus of our paper is to address the Jeffreys–Lindley paradox, we point out that our proposed cake prior framework may be modified to address the arbitrary constant issue. In particular, our proposed cake prior can be subsumed under an appropriate existing solution that addresses this arbitrary constant issue. Some suitable existing solutions include the frameworks for intrinsic Bayes factor (Berger & Pericchi 1996), the fractional Bayes factor (O’Hagan 1995) and the objective Bayes factor (Villa & Walker 2022). A brief review of these existing Bayes factor frameworks is provided in Appendix A.

3.4. One sample test for equal means (with unknown variance)

Consider the hypothesis test $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$, that is, (3) where $x_i | \mu, \sigma^2 \sim N(\mu, \sigma^2)$, $1 \leq i \leq n$, where $\mu$ and $\sigma^2$ are the mean and variance parameters, respectively. Suppose now that both $\mu$ and $\sigma^2$ are unknown parameters to be estimated (unlike the example in Section 2 where $\sigma^2$ was assumed to be known).

The mean expected information matrices for the null and alternative hypothesis respectively are

$$\tilde{I}_0(\sigma^2) = 1/(2\sigma^4)$$

and

$$\tilde{I}_1(\mu, \sigma^2) = \begin{bmatrix} \sigma^{-2} & 0 \\ 0 & 1/(2\sigma^4) \end{bmatrix},$$

which coincides with the Fisher information matrices in this case. We cannot directly use the methodology in Section 3 to construct our cake priors as $\sigma^2 > 0$. To handle this complication, we use the transformation $\sigma^2 = \exp(s)$. Under this transformation, the mean expected information matrices become $\tilde{I}_0(s) = 1/2$ and $\tilde{I}_1(\mu, s) = \text{diag}(\exp(-s), 1/2)$. Using the steps for Section 3 under this transformation, we have $s|H_0 \sim N(0, 2g_0)$, that is, $\pi(s|H_0) = [4\pi g_0]^{-1/2} \exp[-s^2/(4g_0)]$. Transforming back to the $\sigma^2$ parameterisation gives us $\sigma^2|H_0 \sim \text{LN}(0, 2g_0)$, where LN($m, s$) denotes the log-normal law with mean parameter $m$ and variance parameter $s$.

Note that for $H_1$, the upper left entry of $\tilde{I}_1(\mu, s)$ depends on $s$. This implies a conditional dependence of $\mu$ on $s$ which leads to $\mu|s, H_1 \sim N(0, g_1 \exp(s))$ and $s|H_1 \sim N(0, 2g_1)$. Transforming back from the parameterisation in $s$ to the parameterisation using $\sigma^2$ gives $\mu|\sigma^2, H_1 \sim N(0, g_1 \sigma^2)$ and $\sigma^2|H_1 \sim \text{LN}(0, 2g_1)$.

The marginal distributions of $x$ given $H_0$ and $H_1$ are

$$p(x|H_0) = \int_0^\infty c \exp \left[ -\frac{n + 1}{2} \ln(\sigma^2) - \frac{n \hat{\sigma}_0^2}{2\sigma^2} - \frac{\ln(4\pi g_0)}{2} \right] d\sigma^2$$

and

$$p(x|H_1) = \int_0^\infty c \exp \left[ -\frac{n + 2}{2} \ln(\sigma^2) - \frac{n \hat{\sigma}_0^2}{2\sigma^2} - \frac{\ln(4\pi g_1^2)}{2} \right] d\sigma^2,$$

where $c = 1/(\sqrt{2\pi})^n$, $\hat{\sigma}_0^2 = n^{-1}||x - \mu_0 1||^2$, and $\hat{\sigma}_1^2 = n^{-1}||x||^2 - (n\bar{x})^2/(n + g^{-1})$.
\( BF_{01} \to \infty \) as \( g \to \infty \). If we instead use \( g_0 = h \) and \( g_1 = h^{1/2} \), then the Bayes factor simplifies to

\[
BF_{01}(h) = \frac{\int_0^\infty \exp \left[ -\left( \frac{n}{2} + 1 \right) \ln(\sigma^2) - \frac{n\hat{\sigma}_0^2}{2\sigma^2} \right] d\sigma^2}{\int_0^\infty \exp \left[ -\left( \frac{n}{2} + 1 \right) \ln(\sigma^2) - \frac{n\hat{\sigma}_0^2}{2\sigma^2} - \left( \frac{\ln \sigma^2}{4h} \right)^2 \right] d\sigma^2}
\]

which can be evaluated using univariate quadrature or other methods for any fixed \( h > 0 \).

To reduce computational complexity, we let \( h \to \infty \). The marginal distributions of \( x \) may be expressed as

\[
p(x|H_0) \xrightarrow{h \to \infty} \int_0^\infty \exp \left[ -\left( \frac{n}{2} + 1 \right) \ln(\sigma^2) - \frac{n\hat{\sigma}_0^2}{2\sigma^2} \right] d\sigma^2, \quad \text{and}
\]

\[
p(x|H_1) \xrightarrow{h \to \infty} \int_0^\infty \exp \left[ -\left( \frac{n}{2} + 1 \right) \ln(\sigma^2) - \frac{n\hat{\sigma}_1^2}{2\sigma^2} - \frac{1}{2} \ln(n) \right] d\sigma^2,
\]

where the notation \( \xrightarrow{h \to \infty} \) is used to denote ‘equality in the limit as \( h \to \infty \) after terms related to \( h \) cancel in the numerator and denominator of the corresponding Bayes factor, or terms related to \( h \) vanish as \( h \) diverges in the Bayes factor.’ The Bayesian test statistic using the cake prior becomes \( \lambda_{\text{Bayes}} = \lambda_{\text{LRT}} - \ln(n) \).

We compare the performances of Bayesian tests using our proposed cake Bayes factor, fractional Bayes factor (O’Hagan 1995) based on Jeffreys prior, objective Bayes factor (Villa & Walker 2022) based on Jeffreys prior, and the intrinsic Bayes factor (Berger & Pericchi 1996) based on Jeffreys prior. Following Appendix A, the fractional Bayes factor for this example is given by

\[
BF_f = \frac{1}{\sqrt{\Delta}} \left( \frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2} \right)^{-n(1-\Delta)/2}
\]

where \( \Delta = n_0/n \) denotes the minimal sample size proportion and \( n_0 \) denotes the minimal sample size, that is, the smallest sample size such that

\[
0 < \int p(x | \sigma^2, H_0) \pi_{0, \text{Jeffreys}}(\sigma^2 | H_0) \, d\sigma^2 < \infty
\]

and

\[
0 < \int p(x | \mu, \sigma^2, H_1) \pi_{1, \text{Jeffreys}}(\mu, \sigma^2 | H_1) \, d\mu \, d\sigma^2 < \infty.
\]

Here, the Jeffreys priors are given by \( \pi_{0, \text{Jeffreys}}(\sigma^2 | H_0) \propto \sigma^{-2} \) and \( \pi_{1, \text{Jeffreys}}(\mu, \sigma^2 | H_0) \propto \sigma^{-3} \). Consequently, we have \( n_0 = 2 \). The objective Bayes factor (Villa & Walker 2022) for this example is given by

\[
BF_O = 2^n \sum_{i=1}^n \xi_{i/22}(x) - \sum_{i=1}^n \{\xi_{i/21}(x)\}^2 - 2^n \sum_{i=1}^n \xi_{i/12}(x) + \sum_{i=1}^n \{\xi_{i/11}(x)\}^2,
\]

© 2024 The Authors. *Australian & New Zealand Journal of Statistics* published by John Wiley & Sons Australia, Ltd on behalf of Statistical Society of Australia.
where

\[ \zeta_{i11}(x) = -\frac{n(x_i - \mu_0)}{\sum_i (x_i - \mu_0)^2}, \quad \zeta_{i12}(x) = -\frac{n}{\sum_i (x_i - \mu_0)^2} \left\{ 1 - \frac{(n + 2)(x_i - \mu_0)^2}{\sum_i (x_i - \mu_0)^2} \right\}, \]

\[ \zeta_{i21}(x) = -\frac{x_i - \bar{x}}{\hat{\sigma}_1^2}, \quad \zeta_{i22}(x) = -\frac{1}{\hat{\sigma}_1^2} \left\{ 1 - \frac{1}{n} \frac{(n + 2)(x_i - \bar{x})^2}{n\hat{\sigma}_1^2} \right\}. \]

The arithmetic intrinsic Bayes factor for this example is given by

\[ AIBF = \frac{1}{\sqrt{\Delta}} \left( \frac{\hat{\sigma}_1^2}{\hat{\sigma}_0^2} \right)^{n/2} \frac{1}{L} \sum_{\ell \in \mathcal{L}} \left( \frac{\hat{\sigma}_{0\ell}^2}{\hat{\sigma}_{1\ell}^2} \right)^{n_0/2}, \]

where \( \mathcal{L} \) denotes the set of indices of all partition subsets of \( \{1, \ldots, n\} \), \( L = |\mathcal{L}| = n!/\{n_0!(n - n_0)!\} \), \( \hat{\sigma}_{0\ell}^2 \) and \( \hat{\sigma}_{1\ell}^2 \) are the variance estimates based on the partition \( x_i(\ell) \subset \{1, \ldots, n\} \) under \( H_0 \) and \( H_1 \), respectively.

We conduct the following simulation study to illustrate the differences between the LRT and Bayesian hypothesis tests for this problem. Letting \( \mu_0 = 0 \), we simulate a single set of data from \( x_i \sim N(\mu_{\text{true}}, 1) \), \( 1 \leq i \leq n \). After simulating \( 10^6 \) such datasets for all values of \( \mu_{\text{true}} \) in the set \( \{0, 0.1, 0.25, 0.5\} \) and a grid of \( n \) from \( n = 10 \) to \( n = 1000 \), we plot in Figure 1 the empirical probabilities of rejecting the null hypothesis (for the LRT test) using \( \alpha = 0.05 \) or preferring the alternative hypothesis (for the Bayesian tests).

Figure 1. Empirical probabilities of rejecting the null hypothesis or preferring the alternative hypothesis for the simulation described in Section 3.3 comparing the LRT and Bayesian tests when \( \mu_{\text{true}} \in \{0, 0.1, 0.25, 0.5\} \).
From Figure 1, we see empirically that the limiting type I error probabilities using cake-based Bayes factors, FBF and AIBF are equal to 0 as \( n \) diverges. It can be shown that the OBF type I error probability converges to 0. Nonetheless, this convergence behaviour is not evident from Figure 1 as its rate of convergence is very slow. The LRT test has, by design, a limiting type I error probability of 0.05 and hence does not exhibit the same convergence property as the Bayesian tests. When \( H_1 \) is true, that is, \( \mu_{\text{true}} \in \{0.1, 0.25, 0.5\} \) and \( \ln(n) < \chi^2_{1,\alpha} \) the Bayesian test is more powerful than the LRT test, and when \( H_1 \) is true and \( \ln(n) > \chi^2_{1,\alpha} \) the LRT test is more powerful than the cake-based Bayesian test. When \( \mu_{\text{true}} \in \{0.25, 0.5\} \), all tests appear to have a limiting statistical power of 1 as \( n \) diverges. For the case \( \mu_{\text{true}} = 0.5 \) when \( \ln(n) > \chi^2_{1,\alpha} \) LRT, cake-based Bayes tests, FBF and OBF tests have very similar power. However, AIBF-based tests appear to be severely underpowered here.

### 3.5. Two sample test for equal means

Suppose we have data \( \mathbf{x} = (x_1, \ldots, x_n)^\top \). We want to test whether the first \( n_0 \) samples \( \mathbf{x}_0 = (x_1, \ldots, x_{n_0})^\top \) from class 0 come from the same normal population as the second \( n_1 \) samples \( \mathbf{x}_1 = (x_{n_0+1}, \ldots, x_n)^\top \) from class 1 with \( n_0 + n_1 = n \). We wish to test

\[
H_0 : x_i | \mu, \sigma^2 \sim N(\mu, \sigma^2), \quad 1 \leq i \leq n, \quad \text{versus} \\
H_1 : \begin{cases} 
  x_i | \mu_0, \sigma^2_0 \sim N(\mu_0, \sigma^2_0), & 1 \leq i \leq n_0, \\
  x_i | \mu_1, \sigma^2_1 \sim N(\mu_1, \sigma^2_1), & n_0 + 1 \leq i \leq n,
\end{cases}
\]

(8)

where \( \mu, \sigma^2, \mu_0, \sigma^2_0, \mu_1, \sigma^2_1 \) are the means and variances under the one and two group hypotheses, respectively. Here \( \mathbf{\theta}_0 = (\mu, \sigma^2)^\top \) with \( d_0 = 2 \) and \( \mathbf{\theta}_1 = (\mu_0, \mu_1, \sigma^2_0, \sigma^2_1)^\top \) with \( d_1 = 4 \).

Using similar arguments as in Section 3.3 with \( \mathbf{P}_0(\mathbf{\theta}_0) = \tilde{\mathbf{I}}_0(\mathbf{\theta}_0) = \text{diag}[\sigma^{-2}, 1/(2\sigma^4)] \) and \( \mathbf{P}_1(\mathbf{\theta}_1) = \tilde{\mathbf{I}}_1(\mathbf{\theta}_1) = \text{diag}[n_0/(n\sigma^2_0), n_1/(n\sigma^2_1), n_0/(2n\sigma^4_0), n_1/(2n\sigma^4_1)] \) leads to the priors

\[
\begin{align*}
\mu | \sigma^2, H_0 & \sim N(0, g_0\sigma^2), \\
\mu_0 | \sigma^2_0, H_1 & \sim N(0, g_1(n/n_0)\sigma^2_0), \\
\mu_1 | \sigma^2_1, H_1 & \sim N(0, g_1(n/n_1)\sigma^2_1), \\
\sigma^2 | H_0 & \sim \text{LN}(0, 2g_0), \\
\sigma^2_0 | H_1 & \sim \text{LN}(0, 2(n/n_0)g_1), \quad \text{and} \quad \sigma^2_1 | H_1 \sim \text{LN}(0, 2(n/n_1)g_1).
\end{align*}
\]

Setting \( g_j = h^{1/d_j} \), \( j = 0, 1 \) results in

\[
p(\mathbf{x}|H_0) = \int_0^\infty c \exp \left[ -\frac{n + 2}{2} \ln(\sigma^2) - \frac{n\hat{\sigma}_h^{2\alpha}}{2\sigma^2} - \frac{\ln(4\pi h)}{4} - \frac{1}{4} \ln(n + 1) \right] d\sigma^2
\]

\[
\Rightarrow \int_0^\infty c \exp \left[ -\frac{n + 2}{2} \ln(\sigma^2) - \frac{n\hat{\sigma}_h^{2\alpha}}{2\sigma^2} - \frac{\ln(4\pi)}{4} - \frac{1}{2} \ln(n) \right] d\sigma^2
\]

\[
= \exp \left[ \ln p(\mathbf{x}|\widehat{\mathbf{\theta}}_0) + \xi \left( \frac{n}{2} \right) - \frac{1}{2} \ln(2) - \frac{1}{2} \ln(n) \right],
\]

where \( c = (2\pi)^{-n/2}, \) \( \xi(x) = \ln \Gamma(x) + x - x \ln(x) - (1/2) \ln(2\pi) \) and \( \ln p(\mathbf{x}|\widehat{\mathbf{\theta}}_0) = -(n/2) \ln(2\pi \hat{\sigma}^2) - n/2 \) is the log-likelihood of the null model evaluated at its MLE \( \widehat{\mathbf{\theta}}_0 \).
Similarly, \( p(x_0|H_1, h) \) is

\[
p(x_0|H_1) = \int_0^\infty \frac{1}{\sqrt{2\pi n_0}} \exp \left[ - \left( \frac{n_0}{2} + 1 \right) \ln(\sigma_0^2) - \frac{1}{2\sigma_0^2} \left\{ \|x_0\|^2 - \frac{(n_0\bar{x}_0)^2}{n_0 + (n_0/n)h^{-1/4}} \right\} \right] \\
- \frac{1}{2} \ln(n + h^{-1/4}) - \frac{1}{2} \ln(2\pi(2n/n_0)h^{1/2}) - \frac{(\ln\sigma_0^2)^2}{4(n/n_0)h^{1/4}} d\sigma_0^2.
\]

By construction, the \( \ln(h^{1/2}) \) term cancels out from the numerator and the denominator of the Bayes factor. For computational convenience, we let \( h \to \infty \) in the above computations. Following Stirling’s approximation (Abramowitz & Stegun 1972) and some algebraic manipulation, the Bayesian hypothesis test statistic simplifies to

\[
\lambda_{\text{Bayes}} = \lambda_{\text{LRT}} - 2 \ln(n) + O(n_0^{-1} + n_1^{-1}).
\]

It is important to note that all of the constant terms have cancelled from the asymptotic approximation for \( \lambda_{\text{Bayes}} \). This has been achieved by incorporating the \( (n/n_0) \) and \( (n/n_1) \) factor in the priors for \( \mu_0 \) and \( \mu_1 \), and the \( (2n/n_0) \) and \( (2n/n_1) \) factors in the priors for \( \sigma_0^2 \) and \( \sigma_1^2 \). Without these factors, cancellation of \( O(1) \) and larger terms in the expression \( \lambda_{\text{Bayes}} \) would not occur.

We conducted a numerical experiment to compare our Bayesian test with the LRT. Note that a comparison with existing well-known Bayes factor methods (O’Hagan 1995; Berger & Pericchi 1996; Villa & Walker 2022) has been omitted. We simulated \( 10^6 \) datasets where \( n_0 = n_1 = 50 \) with the true parameter values

(i) \( \mu_0 = 0, \sigma_0 = \sigma_1 = 1 \) and \( \mu_1 \in \{0, 0.25, 0.5, 1\} \); or
(ii) \( \mu_0 = 0, \mu_1 = 0, \sigma_0 = 1 \) and \( \sigma_1 \in \{0, 1.25, 1.5, 2.5\} \).

The empirical probabilities of rejecting the null (in the LRT case) or preferring the alternative (in the Bayesian test) are illustrated in Figure 2. Note that under \( H_0 \), the type I error probability approaches zero as \( n \to \infty \) for the Bayesian test, and under \( H_1 \) the type II error probabilities approach zero as \( n \to \infty \) for both the Bayesian and LRT tests. When \( H_0 \) is true, the LRT has a fixed 5% type I error probability.

3.6. Linear models

We will now consider hypothesis testing for linear models. Consider the base model

\[
y|\alpha, \beta, \sigma^2 \sim N(\alpha 1 + X\beta, \sigma^2 I),
\]

where \( y \) is a response vector of length \( n \), \( \beta \) is a coefficient vector of length \( p \), \( \sigma^2 \) is a positive scalar, \( X \) is a full-rank \( n \) by \( p \) matrix of covariates, and \( I \) is the identity matrix of appropriate dimension. In order to simplify some calculations, we will transform \( y \) and \( X \) so that \( y \) and the columns of \( X \) are standardised, that is, \( \bar{y} = 0, \|y\|^2 = y^\top y = n, X_j^\top 1 = 0, \) and \( \|X_j\|^2 = n \) where \( X_j \) is the \( j \)th column of \( X \). Let \( y \) be a binary vector of length \( p \), and let \( X_j \) be the submatrix \( X \) comprising the columns of \( X \) whose corresponding elements of
Figure 2. The empirical probabilities of rejecting the null (in the LRT case) or preferring the alternative (in the Bayesian test) when simulating two normal populations with $n_0 = n_1 = 50$, and (i) $\mu_0 = 0, \sigma_0 = \sigma_1 = 1$ and $\mu_1 \in \{0, 0.25, 0.5, 1\}$ (left four panels); or (ii) $\mu_0 = 0, \mu_1 = 0, \sigma_0 = 1$ and $\sigma_1 \in \{0, 1.25, 1.5, 2.5\}$ (right four panels).
\( \gamma \) are non-zero. Let \( \beta_{\gamma} \) denote the subvector of \( \beta \) whose corresponding elements of \( \gamma \) are non-zero. Consider the hypothesis test

\[
H_0 : \gamma = \gamma_0 \quad \text{versus} \quad H_1 : \gamma = \gamma_1, \tag{9}
\]

where \( \gamma_0 \) and \( \gamma_1 \) denote the models under the null and alternative hypotheses, respectively with \( 0 < |\gamma_0| < |\gamma_1| \). Note that we do not require the model under \( H_0 \) to be nested within the model under \( H_1 \).

To simplify the exposition for this example, we will only use cake priors for \( \alpha \) and \( \beta_{\gamma} \). Since \( \sigma^2 \) is a common parameter across all parameters, we can use the typical improper (Jeffreys) priors for \( \sigma^2 \) given by \( \pi(\sigma^2) \propto (\sigma^2)^{-1}I(\sigma^2 > 0) \). This choice has been formally justified in Berger, Pericchi & Varshavsky (1998). Cake priors can be used for all parameters for this example, but the working out is lengthy and unnecessarily obfuscates the exposition. Using \( P(\alpha, \beta_{\gamma}) = \tilde{I}(\alpha, \beta_{\gamma}) = \text{diag}(\sigma^{-2}, \sigma^{-2}X_{\gamma}^\top X_{\gamma}/n) \) for a particular model \( \gamma \) leads to

\[
\alpha | \sigma^2, g \sim N(0, \sigma^2), \quad \text{and} \quad \beta_{\gamma} | \sigma^2, g \sim N \left( 0, \sigma^2 \left( \frac{1}{n} X_{\gamma}^\top X_{\gamma} \right)^{-1} \right). \tag{10}
\]

Further, we use \( \pi(\beta_{-\gamma}) = \prod_{j: \gamma_j = 0} \delta(\beta_j; 0) \) where \( \delta(x; a) \) is the Dirac delta function with location \( a \), where \( \beta_{-\gamma} \) denotes the subvector of \( \beta \) with corresponding elements of \( \gamma \) equal to zero. The prior on \( \beta_{\gamma} \) is simply the Zellner \( g \)-prior (Zellner 1986) where the prior covariance is scaled by a factor of \( n \). The prior on \( \beta_{\gamma} \) combined with the prior on \( \beta_{-\gamma} \) is a spike and slab prior for \( \beta \).

Marginalising over \( \alpha, \beta \) and \( \sigma^2 \) for a particular model \( \gamma \), we obtain after simplification

\[
p(\gamma | \gamma, g) = \frac{\Gamma(n/2)}{(n\pi)^{n/2}} \exp \left[ -\frac{1}{2} \frac{|\gamma|^2}{\ln(g) - 1} - \frac{1}{2} \ln(n + g^{-1}) - \frac{n}{2} \ln \left( 1 - \frac{g}{1 + g - R_{\gamma}^2} \right) \right],
\]

where \( R_{\gamma}^2 \) is the usual \( R^2 \) statistic for model \( \gamma \). This is equivalent to the \( g = h^{1/d_h} \) of Section 3 using the hyperpriors \( \pi(g | \gamma_j) = \delta(g; h^{1/(1+|\gamma_j|)}) \), \( j = 0, 1 \). After marginalising over \( g \), the Bayes factor (as a function of \( h \)) simplifies to

\[
BF_{01}(h) = \exp \left[ -\frac{n}{2} \ln \left( 1 - \frac{h^{1/(1+|\gamma_0|)}}{1 + h^{1/(1+|\gamma_0|)}R_{\gamma_0}^2} \right) + \frac{n}{2} \ln \left( 1 - \frac{h^{1/(1+|\gamma_1|)}}{1 + h^{1/(1+|\gamma_1|)}R_{\gamma_1}^2} \right) \right. \\
\left. \frac{1}{2} \ln(n + h^{-1/(1+|\gamma_0|)}) + \frac{1}{2} \ln(n + h^{-1/(1+|\gamma_1|)}) \right].
\]

Taking \( h \to \infty \) we use the fact that \( 1 - R_{\gamma}^2 = \hat{\sigma}_{\gamma}^2 \) (where \( \hat{\sigma}_{\gamma}^2 \) is the MLE for \( \sigma^2 \) under the model \( \gamma \)) to obtain

\[
\lambda_{\text{Bayes}} = \left[ -n \ln(\hat{\sigma}_{\gamma_1}^2) - |\gamma_1| \ln(n) \right] - \left[ -n \ln(\hat{\sigma}_{\gamma_0}^2) - |\gamma_0| \ln(n) \right] = \text{BIC}_{\gamma_0} - \text{BIC}_{\gamma_1} = \lambda_{\text{LRT}} - \nu \ln(n),
\]

where \( \text{BIC}_\gamma = n \ln (2\pi \hat{\sigma}_\gamma^2) - n + |\gamma| \ln(n) = -2 \ln p(\gamma | \tilde{\alpha}_\gamma, \tilde{\beta}_\gamma, \hat{\sigma}_{\gamma}^2) + |\gamma| \ln(n) \), and \( \tilde{\alpha}_\gamma \) and \( \tilde{\beta}_\gamma \) are the MLEs for \( \alpha \) and \( \beta \) under model \( \gamma \), \( \lambda_{\text{LRT}} = n \ln(\hat{\sigma}_{\gamma}^2) - n \ln(\hat{\sigma}_{\gamma_1}^2) \) is the LRT statistic corresponding to the hypotheses (9) and \( \nu = |\gamma_1| - |\gamma_0| \). Hence, for these models and prior structures, the Bayesian test statistic is simply the difference between two BIC values.
Note that as $h \to \infty$, the posteriors are

$$\alpha|\gamma, y \sim t_n(0, \hat{\sigma}_\gamma^2/n), \beta_\gamma|\gamma, y \sim \hat{\beta}_\gamma \hat{\sigma}_\gamma^2 \left( X_\gamma^\top X_\gamma \right)^{-1}, \text{ and } \sigma^2|\gamma, y \sim IG\left( \frac{n}{2}, \frac{n}{2} \hat{\sigma}_\gamma^2 \right),$$

where $\hat{\beta}_\gamma$ and $\hat{\sigma}_\gamma^2$ are the MLEs corresponding to model $\gamma$. Our proposed prior results in a Bayes factor that is a simple function of the BIC and is model-selection consistent for independent and identically distributed (iid) data (under some additional mild assumptions, Yang 2005).

We will not provide any numerical examples due to the close relationship between our Bayes factors and the BIC, and the fact that most papers on model selection for linear models use the BIC in its comparisons. We direct the interested reader to any of the papers in the discussion below, all of which make comparisons with the BIC as a model selection criteria.

Our proposed prior differs from the existing priors that have been used in the literature for linear models. First, we have assigned cake priors for the entire coefficient vector ($\alpha, \beta_\gamma$), whereas the typical choice of prior on $\alpha$ in the literature is the Jeffreys prior $p(\alpha) \propto 1$ as advocated by Berger, Pericchi & Varshavsky (1998). If we were to assign the Jeffreys prior for the intercept and were only to use cake priors for $\beta_\gamma$, then $g_0 = h^{1/|\gamma_0|}$ instead of $g_0 = h^{1/(1+|\gamma_0|)}$. The consequence of this would be that $g_0$ would become problematic to calculate for the null model ($\gamma = 0$).

Second, our proposed prior for $\beta_\gamma$ does not degenerate to a point mass. This asymptotic behaviour is in line with a recommendation by Bayarri et al. (2012). More specifically, under mild conditions, our proposed prior variance for $\beta_\gamma$ converges almost surely to $g\sigma^2 \left( E(X_\gamma^\top X_\gamma) \right)^{-1}$. On the other hand, a popular Bayesian approach to model selection for linear models is to use the Zellner $g$-prior where

$$\beta_\gamma|\sigma^2, g \sim N \left( 0, g\sigma^2 \left( X_\gamma^\top X_\gamma \right)^{-1} \right).$$

Observe that the prior variance degenerates to a point mass which is against the recommendation by Bayarri et al. (2012).

The third difference is the choice of prior on $g$. As stated in the introduction, Liang et al. (2008) argue for a non-point-mass hyperprior to be assigned to $g$. Liang et al. (2008) considers the hyper $g$-prior, the hyper $g/n$-prior, and the Zellner-Siow prior (equivalent to a particular inverse-gamma on $g$) (Zellner & Siow 1980). Maruyama & George (2011) use a different prior to (10) or (11) and a beta-prime prior with specially chosen prior hyperparameter values. These non-point-mass hyperpriors lead to intractable expressions for the marginal likelihood. On the other hand, our proposed prior involves a careful specification of a point mass prior for $g$ to avoid the Jeffreys–Lindley paradox.

3.7. Handling zero parameter in the null model

We now return to Lindley's example posed by Lindley (1957) described in Section 2.1. In order to apply the methodology of Section 3, the null model needs to have a non-zero number of parameters. We provide the following novel artificial construct to handle this case in order to augment the problems so that both hypotheses have a non-zero number of parameters.
1. Introduce a second sample of hypothetical data, say \( z \).

2. Modify the null and alternative hypotheses by adding a clause that the hypothetical data have the same distribution under the null and alternative hypotheses.

3. Apply the methodology of Section 3 to the augmented problem.

In order to illustrate this approach, suppose we have a second sample of hypothetical data \( z = (z_1, \ldots, z_n)^T \) and consider the augmented hypotheses

\[
H_0 : x_1, \ldots, x_n \sim N(\mu_0, \sigma^2) \quad \text{and} \quad z_1, \ldots, z_n | \widehat{\mu} \sim N(\widehat{\mu}, \sigma^2) \quad \text{versus} \\
H_1 : x_1, \ldots, x_n | \mu \sim N(\mu, \sigma^2) \quad \text{and} \quad z_1, \ldots, z_n | \widehat{\mu} \sim N(\widehat{\mu}, \sigma^2),
\]

where \( \mu_0 \) and \( \sigma^2 \) have known fixed values, and \( \widehat{\mu} \) is an artificial mean parameter corresponding to the sample \( z \). This modification of the original hypotheses in (3) has the same logical implication as the hypotheses (3) for the observed sample \( x \) since the hypothetical data has the same hypothetical models under the null and alternative hypotheses. For the augmented problem we have \( \theta_0 = \widehat{\mu} \) with \( d_0 = 1 \), and \( \theta_1 = (\mu, \widehat{\mu})^T \) and \( d_1 = 2 \) which allows us to avoid the problem of dividing by zero. The cake priors become \( \widehat{\mu} | H_0 \sim N(0, g_0 \sigma^2) \), \( \mu | H_1 \sim N(0, g_1 \sigma^2) \) and \( \widehat{\mu} | H_1 \sim N(0, g_1 \sigma^2) \). For \( g_0 \) and \( g_1 \) we use \( g_0 = h \) and \( g_1 = h^{1/2} \). Then

\[
\ln p(x | H_0) = c - \frac{\|x - \mu_0 1\|^2}{2\sigma^2},
\]

\[
\ln p(z | H_0) = c - \frac{1}{2\sigma^2} \left[ \|z\|^2 - \frac{(n \widetilde{z})^2}{n + h^{-1}} \right] - \frac{1}{2} \ln (h) - \frac{1}{2} \ln (n + h^{-1}),
\]

\[
\ln p(x | H_1) = c - \frac{1}{2\sigma^2} \left[ \|x\|^2 - \frac{(n \widetilde{\mu})^2}{n + h^{-1/2}} \right] - \frac{1}{2} \ln (h^{1/2}) - \frac{1}{2} \ln \left( n + \frac{1}{\sqrt{h}} \right)
\]

\[
\ln p(z | H_1) = c - \frac{1}{2\sigma^2} \left[ \|z\|^2 - \frac{(n \widehat{\mu})^2}{n + h^{-1/2}} \right] - \frac{1}{2} \ln (h^{1/2}) - \frac{1}{2} \ln \left( n + \frac{1}{\sqrt{h}} \right),
\]

where \( c = -\frac{n}{2} \ln (2\pi \sigma^2) \). The Bayes factor in the limit as \( h \to \infty \) is

\[
\lambda_{\text{Bayes}} = \lim_{h \to \infty} -2 \ln \left[ \frac{p(x | H_0)p(z | H_0)}{p(x | H_1)p(z | H_1)} \right] = \lambda_{\text{LRT}} - \ln (n),
\]

where \( \lambda_{\text{LRT}} = \sigma^{-2}(\|x - \mu_0 1\|^2 - \|x - \widehat{\mu} 1\|^2) \) is the likelihood ratio test statistic corresponding to the hypothesis (3). We conducted a simulation study identical to the simulation study in Section 3.3 with the exception that \( \sigma^2 \) was treated as known. The resulting figure and interpretation was nearly identical to that in Section 3.3 (not shown).

4. Theory

We consider the asymptotic properties of Bayesian hypothesis tests based on cake priors. Let \( X = (X_1, \ldots, X_n)^T \) be independent random samples from \( T^i = \{ p^i(\cdot) : i = 1, \ldots, n \} \). The type I and type II error probabilities are defined by \( \alpha_T(P) = \Pr ( T(X) = 1 ) \) when \( P \in \mathcal{P}_0 \) and \( 1 - \alpha_T(P) = \Pr ( T(X) = 0 ) \) when \( P \in \mathcal{P}_1 \) respectively. Fix the level of
significance \( \alpha \) such that \( \sup_{P \in P_0}\{\alpha_T(P)\} \leq \alpha \). Following (6), we may express the cake Bayes factor as:

\[
\lambda_{\text{Bayes}} = \lambda_{\text{LRT}} - \nu \ln(n) + O_p(n^{-1}).
\] (12)

Note that the above equality holds in most contexts when cake priors are used (recall that the above equality holds in all examples in Section 3). Following the above equality, the asymptotic behaviour of Bayes factors based on the cake prior depends on the behaviour of the LRT test statistic. A detailed exposition on the characterisation of the asymptotic distribution of the LRT statistic under quite general conditions, including when \( H_0 \) and/or \( H_1 \) is mis-specified, and whether the hypotheses are nested or non-nested can be found in Vuong (1989).

In the rest of this section, we adopt notations and definitions from Shao (2003). We denote \( T_n(X) \equiv T(X) \) and consider scenarios where \( n \) diverges. In our ensuing discussion we use the following definitions.

**Definitions from 2.13 of Shao (2003):**

(i) If \( \lim_{n \to \infty} \sup_{P \in P_0}\{\alpha_{T_n}(P)\} \leq \alpha \) then \( \alpha \) is an asymptotic significance level of \( T_n \).

(ii) If \( \lim_{n \to \infty} \sup_{P \in P_0}\{\alpha_{T_n}(P)\} \) exists, then it is called the limiting size of \( T_n \).

(iii) The sequence of tests \( \{T_n\}_{n \geq 1} \) is called consistent if and only if the type II error probability converges to 0, i.e., \( \lim_{n \to \infty} \{1 - \alpha_{T_n}(P)\} = 0 \), for any \( P \in P_1 \).

(iv) The sequence of tests \( \{T_n\}_{n \geq 1} \) is called Chernoff consistent if and only if \( T_n \) is consistent and the type I error probability converges to 0, i.e., \( \lim_{n \to \infty} \{\alpha_{T_n}(P)\} = 0 \), for any \( P \in P_0 \). Furthermore, \( T_n \) is called strongly Chernoff consistent if and only if \( T_n \) is consistent and the limiting size of \( T_n \) is 0.

We note that any reasonable test which is consistent where the level \( \alpha \) is controllable can be made Chernoff consistent by letting \( \alpha_n = \alpha \to 0 \) as \( n \to \infty \).

### 4.1. Main results

Consider the log-likelihood \( \ell(\theta) = \sum_{i=1}^{n} \ln p_i(X_i|\theta) \), the MLE and the ‘pseudo-true’ values of \( \theta \)

\[
\hat{\theta} = \arg \max_\theta \{ \ell(\theta) \} \quad \text{and} \quad \theta^* = \arg \max_\theta \left[ E\{n^{-1}\ell(\theta)\} \right],
\]

respectively (assuming both are well-defined), where the above expectation is taken with respect to \( \prod_{i=1}^{n} p_i^*(.) \). Assume also that \( E\{n^{-1}\ell(\theta^*)\} \to C \), for some finite \( C > 0 \). It can be shown that \( \prod_{i=1}^{n} p_i^*(\cdot|\theta^*) \) minimises the KL divergence \( E[\sum_{i=1}^{n} \log p_i^*(X_i) - \ell(\theta)] \). Hence, \( \prod_{i=1}^{n} p_i^*(\cdot|\theta^*) \) is the ‘best’ distribution in the sense that it results in the smallest Kullback–Leibler (KL) divergence. We assume that these best distributions are unique. We summarize the key results of the Bayesian hypothesis test based on cake priors as follows:

- **Result 2:** If \( H_0 \) is true, then \( \lambda_{\text{LRT}} \) is \( O_p(1) \) and \( \Pr(T(X) = 1) \to 0 \) as \( n \to \infty \). If \( H_1 \) is true, then \( \lambda_{\text{LRT}} \) is \( \Omega_p(n) \) and \( \Pr(T(X) = 0) \to 0 \) as \( n \to \infty \). Hence, a Bayesian test of the form (12) is Chernoff consistent.

- **Result 3:** The asymptotic probability of the occurrence of Lindley’s paradox is \( \Pr(\chi^2_{\nu,\alpha} < \lambda_{\text{LRT}} < \nu \ln n) \). If \( H_0 \) is true, the probability converges to \( \alpha \) as \( n \to \infty \). If \( H_1 \) is true, the probability converges to 0.
• **Result 4:** Suppose we have $M$ competing hypotheses $H_j$, $j = 1, \ldots, M$. For each model, we have $E\left[ n^{-1} \ell_j(\theta_j^*) \right] \to \ell_j^*$ for some constants $\ell_j^*$. We will call a hypothesis $H_j$ correct if $j \in C$ where

$$C = \left\{ j : \ell_j^* = \max_{k=1,\ldots,M} \ell_k^* \right\}.$$  

The hypotheses $H_j$ such that $j \in C$ correspond to correct models in the sense that all such models are closest in terms of their KL-divergence to the data generating distribution. Using (6) to compare models not in $C$ with models in $C$ leads to

$$\lambda_{LRT}(X) = \Omega_p(n)$$

and the test preferring the model in $C$. Comparing any two models in $C$ will asymptotically prefer the model with the smallest size.

• **Result 5:** For our Bayesian tests, the cut-off value $\nu \ln(n)$ grows with $n$ implying that the level of the test decays with $n$. A consequence is that our Bayesian tests offers protection against a sequential analysis (where samples are collected and hypothesis testing performed sequentially).

Results 2, 3 and 5 may be deduced by using standard results from Vuong (1989) and Shao (2003) regarding the asymptotic behaviour of $\lambda_{LRT}$ and taking note of the asymptotic equality in (12). A proof of Result 4 is provided in the appendix.

5. Conclusion

We introduced ‘cake priors’—a new class of priors which possess a number of desirable properties. Cake priors can be made arbitrarily diffuse without the Bayes factor favouring the null or alternative hypotheses as the prior becomes increasingly diffuse. In the examples we considered, the limiting Bayes factors take the form of penalised likelihood ratio statistics, one of the most thoroughly understood quantities in Statistics. Due to their close link with Jeffreys priors, cake priors are parametrisation invariant. The resulting Bayesian test avoids the need to specify a $p$-value cut-off and are asymptotically Chernoff consistent. Unlike approaches that split the dataset up into parts or use imaginary data, cake priors are transparent, uncomplicated and easily implementable. As briefly discussed, the cake prior construction recipe can be subsumed under an existing solution framework (e.g. fractional Bayes factor, objective Bayes factor and intrinsic Bayes factor) to resolve the arbitrary constants issue. Finally, Bayesian tests based on cake priors provide some protection against over-conservativeness in sequential testing as the sample size grows.

A limitation of our proposed approach is that cake priors may be improper. While a natural solution to ensure properness is to set the cake prior pseudo-precision matrix $P_j(\theta_j)$ as $P_j(\widetilde{\theta}_j)$ for some estimator $\widetilde{\theta}_j$, this solution may lead to a loss of desirable theoretical properties such as the prior invariance property and the smaller asymptotic order of bias. Moreover, it is unclear whether the desirable theoretical properties of the cake prior is retained if it is subsumed under an existing Bayes factor framework to address the arbitrary constants issue. A further limitation occurs under the assumption that the parameter dimensions under each hypothesis remain fixed as $n$ grows. This assumption may not hold even under i.i.d. settings.

Taken together, we believe all the advantageous properties should make cake priors a reasonable choice when performing Bayesian hypothesis tests for hypothesis consisting
of a simple point null against a composite alternative for parametric models with iid data, under the regularity conditions stated.

Appendix A

Appendix: Brief review of other Bayes factor frameworks

Intrinsic prior modification: An approach to circumventing the arbitrary constants issue is the intrinsic priors framework of Berger & Pericchi (1996). Given an observed dataset \( \{x_1, \ldots, x_n\} \), we define an equal-sized partition of the dataset \( \{x(\ell)\}_{\ell \in \mathcal{L}} \), where \( \mathcal{L} \) is the set of indices of all partition subsets. Consider the prior

\[
\pi_\ell^I(\theta_j | H_j) = p(\theta_j | H_j, x(\ell)), \quad j = 0, 1,
\]

where

\[
p(\theta_j | H_j, x_\ell) \propto p(x(\ell) | \theta_j, H_j)p(\theta_j | H_j).
\]

Here, \( x_\ell \) is chosen to be the minimum required sample size to ensure that \( p_\ell^I(\theta_j | H_j) \) is proper for all \( j = 0, 1 \). The resultant intrinsic Bayes factor may be expressed as

\[
B_{01}^I(x(\ell)) = \frac{\int_{\Theta_\ell} p(x(\ell) | \theta_0, H_0) \pi_\ell^I(\theta_0 | H_0) \, d\theta_0}{\int_{\Theta_\ell} p(x(\ell) | \theta_1, H_1) \pi_\ell^I(\theta_1 | H_1) \, d\theta_1},
\]

where

\[
\pi_\ell^I(\theta_0 | H_0) = \frac{\int_{\Theta_\ell} p(x(\ell) | \theta_0, H_0)f_0(\theta_0) \, d\theta_0}{\int_{\Theta_\ell} p(x(\ell) | \theta_1, H_1)f_1(\theta_1) \, d\theta_1} \times \frac{\int_{\Theta_\ell} p(x(\ell) | \theta_0, H_0)f_0(\theta_0) \, d\theta_0}{\int_{\Theta_\ell} p(x(\ell) | \theta_0, H_0)f_0(\theta_0) \, d\theta_0},
\]

where \( x(\ell) = \{x_1, \ldots, x_n\} \setminus x(\ell) \). Observe that the above Bayes factor does not depend on the arbitrary constants and that the intrinsic Bayes factor is sensitive to the choice of \( x(\ell) \) through the second term on the RHS. To resolve this sensitivity issue, we can average over all possible subsets which yields the arithmetic intrinsic Bayes factor

\[
\text{AIBF}_{01} = \frac{\int_{\Theta_\ell} p(x | \theta_0, H_0)f_0(\theta_0) \, d\theta_0}{\int_{\Theta_\ell} p(x | \theta_1, H_1)f_1(\theta_1) \, d\theta_1} \times \frac{1}{\mathcal{L}} \sum_{\ell \in \mathcal{L}} BF_{10}(x(\ell)),
\]

where

\[
BF_{10}(x(\ell)) = \frac{\int_{\Theta_\ell} p(x(\ell) | \theta_1, H_1)f_1(\theta_1) \, d\theta_1}{\int_{\Theta_\ell} p(x(\ell) | \theta_0, H_0)f_0(\theta_0) \, d\theta_0}.
\]

Fractional Bayes factors: Define

\[
p(x | H_j, \Delta) = \int p(x | \theta_j, H_j) \pi(\theta_j | H_j) \, d\theta_j.
\]

where \( \Delta \) is a small positive constant interpreted as the marginal likelihood using a fraction \( (\Delta) \) of the likelihood and \( \pi(\theta_j | H_j) \) is an improper prior. The fractional Bayes factor (FBF) for testing a pair of hypotheses \( H_j \) versus \( H_k \) is given by

\[
\text{FBF}_{jk}(x) = \left[ \frac{p(x | H_j, 1 - \Delta)}{p(x | H_j, \Delta)} \right] / \left[ \frac{p(x | H_k, 1 - \Delta)}{p(x | H_k, \Delta)} \right] = \frac{p(x | H_j, 1 - \Delta)}{p(x | H_k, 1 - \Delta)} \frac{p(x | H_k, \Delta)}{p(x | H_j, \Delta)}.
\]

© 2024 The Authors. Australian & New Zealand Journal of Statistics published by John Wiley & Sons Australia, Ltd on behalf of Statistical Society of Australia.
The ratios \( p(x|H_j, 1 - \Delta)/p(x|H_j, \Delta) \) can be interpreted as the fractional marginal likelihood using \((1 - \Delta) \times 100\% \) of the data. The FBF is then analogous to (2). Note that a consequence of considering the normalised marginal likelihood is that any arbitrary constants are cancelled.

**Objective Bayes factors:** The objective Bayes factors (Villa & Walker 2022) arises from matching the Kullback–Leibler distance with the Fisher distance in order to solve the arbitrary constant problem. Their reasoning leads to a Bayes factor of the form

\[
BF_{01} = \exp \left[ 2 \sum_{i=1}^{n} \frac{\partial^2 p(x|H_1)/\partial x_i^2}{p(x|H_1)} - \left( \sum_{i=1}^{n} \frac{\partial p(x|H_1)/\partial x_i}{p(x|H_1)} \right)^2 \right] - \frac{2}{n} \sum_{i=1}^{n} \frac{\partial^2 p(x|H_0)/\partial x_i^2}{p(x|H_0)} + \left( \sum_{i=1}^{n} \frac{\partial p(x|H_0)/\partial x_i}{p(x|H_0)} \right)^2].
\]

**Appendix B**

Appendix: Proof of Result 4

If \( \mathcal{P} \) is suitably regular, with \( \Theta \) a nice subset of \( d \)-dimensional Euclidean space, then certain derivatives exist and various statements can be made: the Euclidean norm of \( \hat{\theta} - \theta^* \) is \( O_p(n^{-1/2}) \), and writing \( \nabla \ell(\theta) \) and \( \nabla^2 \ell(\theta) \) for the first and second order partial derivatives we may expand \( \ell(\theta^*) \) about \( \hat{\theta} \) to get

\[
\ell(\theta^*) = \ell(\hat{\theta}) + (\theta^* - \hat{\theta})^\top \nabla \ell(\hat{\theta}) + \frac{1}{2} (\theta^* - \hat{\theta})^\top \nabla^2 \ell(\hat{\theta}) (\theta^* - \hat{\theta})
\]

\[
= \ell(\hat{\theta}) - \frac{1}{2} [n^{1/2} (\hat{\theta} - \theta^*)]^\top E \left[ - \frac{1}{n} \nabla^2 \ell(\theta^*) \right] [n^{1/2} (\hat{\theta} - \theta^*)] + o_p(1),
\]

since \( \nabla \ell(\hat{\theta}) \equiv 0 \), for some \( \tilde{\theta} = w\theta^* + (1 - w)\hat{\theta} \), where \( 0 < w < 1 \) and the quadratic form is \( \|\nabla^2 \ell(\theta^*)\| = O_p(n) \). Thus, we may decompose the maximised log-likelihood into the following terms:

\[
\ell(\hat{\theta}) = E\{\ell(\theta^*)\} + \left\{ \ell(\theta^*) - E\{\ell(\theta^*)\} \right\} + O_p(1).
\]

The first term on the right hand side is asymptotic to \( nC \), the second term is a random sum of \( n \) terms with expectation zero and so under further regularity conditions is \( O_p(n^{1/2}) \). We refer to these three terms as the ‘\( O_p(n) \)’, ‘\( O_p(n^{1/2}) \)’ and ‘\( O_p(1) \)’ terms, respectively (from left to right). Precise regularity conditions guaranteeing all of this nice behaviour may be found in Vuong (1989); see also Chapter 5 of van der Vaart (1998). They are all easily satisfied in all of our examples.

Finally, we note that the first term can be written as:

\[
E\{\ell(\theta^*)\} = E\left[ \sum_{i=1}^{n} \ln \frac{p_i(X_i|\theta^*)}{p_i^*(X_i)} \right] + E\left[ \sum_{i=1}^{n} \ln p_i^*(X_i) \right]
\]

\[
= - \text{KL}(p_i^* \| p_i) + E\left[ \sum_{i=1}^{n} \ln p_i^*(X_i) \right],
\]

© 2024 The Authors. *Australian & New Zealand Journal of Statistics* published by John Wiley & Sons Australia, Ltd on behalf of Statistical Society of Australia.
where the first term corresponds to the negative KL-divergence between \( p_1^* \) and \( p_i \), and the second term is related to the entropy of \( p_i^* \). We see from the above equation that maximising \( E[\ell(\theta)] \) with respect to \( \theta \) is equivalent to minimising the KL-divergence between \( \prod_{i=1}^n p_i^*(\cdot) \) and \( \prod_{i=1}^n p_i(\cdot|\theta) \).

Suppose now that we have two competing models \( \mathcal{P}_j = \{ p_{ij}(\cdot|\theta_j, H_j) : i = 1, \ldots, n, \theta_j \in \Theta \}, j = 0, 1 \), with corresponding log-likelihoods \( \ell_0(\cdot) \) and \( \ell_1(\cdot) \), MLEs \( \hat{\theta}_0 \) and \( \hat{\theta}_1 \), and pseudo-true values \( \theta_0^* \) and \( \theta_1^* \). For convenience we will write \( \tilde{\ell}_j = \ell_j(\hat{\theta}_j) \).

As assumed in this result for \( j = 0, 1 \), we have \( E[n^{-1}\ell_j(\theta_j^*)] \rightarrow \ell_j^* \).

There are two main cases:

1. If \( \ell_1^* > \ell_0^* \), then the model under \( H_1 \) has a smaller KL-divergence from the true distribution than the model under \( H_0 \), and we have immediately

\[
 n^{-1}\lambda_{LRT}(X) = n^{-1}(\hat{\ell}_1 - \hat{\ell}_0) \xrightarrow{P} \ell_1^* - \ell_0^* > 0 \, ,
\]

that is the LRT statistic \( \lambda_{LRT}(X) = 2(\hat{\ell}_1 - \hat{\ell}_0) \) is of order \( n \) in probability.

2. If \( \ell_1^* = \ell_0^* \), then immediately we see that the two ‘\( O_p(n) \)’ terms in the difference between the log-likelihoods would, at least asymptotically, cancel out and that the ‘\( O_p(n^{1/2}) \)’ terms would ‘dominate’. However, in many practical examples the ‘\( O_p(n^{1/2}) \)’ terms also cancel out, in which case

\[
 \lambda_{LRT}(X) = 2(\hat{\ell}_1 - \hat{\ell}_0) = O_p(1) \, .
\]

(see in particular theorem 3.3 of Vuong 1989). This occurs when the ‘best’ member of each model yields the same distribution, that is when \( \{ p_{i0}(\cdot|\theta_0^*, H_0) \} = \{ p_{i1}(\cdot|\theta_1^*, H_1) \} \). The parameterisations may be completely different, but nonetheless the distributions corresponding to the pseudo-true parameter values are identical. This occurs when the two models have some overlap, that is, \( \mathcal{P}_0 \cap \mathcal{P}_1 \) is non-empty as a subset of all possible sets of distributions \( \{ p_i(\cdot) \} \). In such a case, the models may or may not be nested, and may or may not be correctly specified; however, the ‘best’ distribution in both is the same (and is part of the overlap).

References

Abramowitz, M. & Stegun, I.A. (1972). Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. New York: Dover Publications.

Aitkin, M. (1991). Posterior Bayes factors (with discussion). Journal of the Royal Statistical Society, Series B, 53, 111–142.

Bartlett, M.S. (1957). A comment on D.V. Lindley’s statistical paradox. Biometrika, 44, 533–534.

Bayarri, M.J., Berger, J.O., Forte, A. & García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. The Annals of Statistics, 40, 1550–1577.

Berger, J.O. & Pericchi, L.R. (1996). The intrinsic Bayes factor for model selection and prediction. Journal of the American Statistical Association, 91, 109–122.

Berger, J.O. & Pericchi, L.R. (2001). Objective Bayesian methods for model selection: Introduction and comparison. IMS Lecture Notes - Monograph Series, 38, 135–207.

Berger, J.O., Pericchi, L.R. & Varshavsky, J.A. (1998). Bayes factors and marginal distributions in invariant situations. Sankhyā: The Indian Journal of Statistics, Series A, 60, 307–321.

Bernardo, J.M. (1980). A Bayesian analysis of classical hypothesis testing. Trabajos de Estadistica Y de Investigacion Operativa, 31, 605–647.
BERNARDO, J.M. (1999). Nested hypothesis testing: The Bayesian reference criterion. *Bayesian Statistics, 6*, 101–130.

BOVÉ, D.S. & HELD, L. (2011). Hyper-g priors for generalized linear models. *Bayesian Analysis, 6*, 387–410.

CHEN, M.H., HUANG, L., IBRAHIM, J.G. & KIM, S. (2008). Bayesian variable selection and computation for generalized linear models with conjugate priors. *Bayesian Analysis, 3*, 585–614.

CHEN, M.H. & IBRAHIM, J.G. (2003). Conjugate priors for generalized linear models. *Statistica Sinica, 13*, 461–476.

DEGROOT, M.H. (1973). Doing what comes naturally: Interpreting a tail area as a posterior probability or as a likelihood ratio. *Journal of the American Statistical Association, 68*, 966–969.

FERNÁNDEZ, C., LEY, E. & STEEL, M.F.J. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics, 100*, 381–427.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. & Rubin, D.B. (2013). Bayesian Data Analysis. In *Chapman & Hall/CRC Texts in Statistical Science*, 3rd edn. Boca Raton, FL: Taylor & Francis.

Gelman, A., Meng, X.L. & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6*, 733–760.

George, E.I. & McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association, 88*, 881–889.

Gupta, M. & Ibrahim, J.G. (2009). An information matrix prior for Bayesian analysis in generalized linear models with high dimensional data. *Statistica Sinica, 19*, 1641–1663.

Hansen, M.H. & Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association, 96*, 746–774.

Hansen, M.H. & Yu, B. (2003). Minimum description length model selection criteria for generalized linear models. *Lecture Notes-Monograph Series, 40*, 145–163.

Hanson, T.E., Branscum, A.J. & Johnson, W.O. (2014). Informative g-priors for logistic regression. *Bayesian Analysis, 9*, 597–612.

Hauck, W.W., Jr. & Donner, A. (1977). Wald’s test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association, 72*, 851–853.

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophical Society, 31*, 203–222.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 186*, 453–461.

Jeffreys, H. (1961). *Theory of Probability*, Vol. 2. Oxford, UK: Oxford University Press.

Kass, R.E. & Raftery, A. (1995). Bayes Factors. *Journal of the American Statistical Association, 91*, 773–795.

Kass, R.E., Tierney, L. & Kadane, J.B. (1990). The validity of posterior expansions based on Laplace’s method. In *Essays in Honor of George Bernard*, eds., Geisser, S., Hodges, J.S., Press, S.J., & Zellner, A., pp. 473–488. Amsterdam, Netherlands: North-Holland.

Lehmann, E. (2004). *Elements of Large-Sample Theory*. Springer Texts in Statistics. New York: Springer.

Li, Y. & Clyde, M.A. (2015). Mixture of g-priors in generalized linear models, arXiv preprint, arXiv:1503.06913.

Liang, F., Paulo, R., Molina, G., Clyde, M.A. & Berger, J.O. (2008). Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association, 103*, 410–423.

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

Maruyama, Y. & George, E.I. (2011). Fully Bayes factors with a generalized g-prior. *The Annals of Statistics, 39*, 2740–2765.

Mitchell, T.J. & Beauchamp, J.J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association, 83*, 1023–1032.

O’Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society, Series B, 57*, 99–138.

O’Hagan, A. (1997). Properties of intrinsic and fractional Bayes factors. *Test, 6*, 101–118.

Petitt, L.I. (1992). Bayes factors for outlier models using the device of imaginary observations. *Journal of the American Statistical Association, 87*, 541–545.

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

Robert, C.P. (2014). On the Jeffreys-Lindley paradox. *Philosophy of Science, 81*, 216–232.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464.
SHAO, J. (2003). Mathematical Statistics. In Springer Texts in Statistics. New York: Springer.
SPANOS, A. (2013). Who should be afraid of the Jeffreys-Lindley paradox? Philosophy of Science, 80, 73–93.
SPIEGELHALTER, D.J., BEST, N.G., CARLIN, B.P. & VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society, Series B, 64, 583–639.
SPIEGELHALTER, D.J. & SMITH, A.F.M. (1982). Bayes factors for linear and log-linear models with vague prior information. Journal of the Royal Statistical Society, Series B, 44, 377–387.
SPRENGER, J. (2013). Testing a precise null hypothesis: The case of Lindley’s paradox. Philosophy of Science, 80, 733–744.
TIERNEY, L., KASS, R.E. & KADANE, J.B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. Journal of the American Statistical Association, 84, 710–716.
VAN DER VAART, A.W. (1998). Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics, Vol. 3. Cambridge, UK: Cambridge University Press.
VILLA, C. & WALKER, S.G. (2022). An objective Bayes factor with improper priors. Computational Statistics and Data Analysis, 168, 107404.
VUONG, Q.H. (1989). Likelihood ratio tests for model selection and nonnested hypotheses. Econometrica, 57, 307–333.
WANG, X. & GEORGE, E.I. (2007). Adaptive Bayesian criteria in variable selection for generalized linear models. Statistics Sinica, 17, 667–690.
YANG, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. Biometrika, 92, 937.
ZELLNER, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti, pp. 233–243. New York: Elsevier Science Publishers, Inc.
ZELLNER, A. & SIOW, A. (1980). Posterior odds ratios for selected regression hypotheses. Trabajos de Estadistica Y de Investigacion Operativa, 31, 585–603.