Econometrics and Statistics 26 (2023) 139–152

Contents lists available at ScienceDirect

Econometrics and Statistics

journal homepage: www.elsevier.com/locate/ecosta

A New Statistic for Bayesian Hypothesis Testing

Su Chen[a], Stephen G. Walker[b,*]

[a] Center of Transforming Data to Knowledge, United States
[b] Department of Mathematics, Department of Statistics and Data Sciences, University of Texas at Austin, United States

ARTICLE INFO

Article history:
Received 9 May 2021
Revised 9 October 2021
Accepted 21 October 2021
Available online 11 November 2021

Keywords:
Hypothesis Testing
Bayes factor
Kullback–Leibler divergence
Improper prior

ABSTRACT

A new Bayesian–inspired statistic for hypothesis testing is proposed which compares two posterior distributions; the observed posterior and the expected posterior under the null model. The Kullback–Leibler divergence between the two posterior distributions yields a test statistic which can be interpreted as a penalized log–Bayes factor with the penalty term converging to a constant as the sample size increases. Hence, asymptotically, the statistic behaves as a Bayes factor. Viewed as a penalized Bayes factor, this approach solves the long standing issue of using improper priors with the Bayes factor, since only posterior summaries are needed for the new statistic. Further motivation for the new statistic is a minimal move from the Bayes factor which requires no tuning nor splitting of data into training and inference, and can use improper priors. Critical regions for the test can be assessed using frequentist notions of Type I error.

1. Introduction

Since its early foundations, set out by Ronald Fisher in the 1920s, who developed the theory behind the p–value, and Jerzy Neyman and Egon Pearson in the 1930s, (see Biau et al. (2010)), hypothesis testing has been of fundamental importance to statistics. Nowadays, many of the problems which traditionally have been formulated in terms of hypothesis testing can also be cast as complex decision problems, such as variable selection in linear regression models.

The Bayesian approach to hypothesis testing was developed by Jeffreys (Jeffreys (1931, 1935, 1936, 1998)), which he called "significance tests": a methodology for quantifying the evidence in favor of a scientific theory. In his approach, statistical models are introduced to represent the probability of the data according to each of the two hypotheses, and Bayes's theorem is used to compute the posterior probability of each hypothesis and assess which one is more compatible with the data. The center piece of his theory is the "Bayes factor", which quantifies the evidence the data provide for one model over another. The Bayes factor has become the primary tool used in Bayesian hypothesis testing, variable selection, model selection, and model averaging (Hoeting et al. (1999); Berger (1997); Bayarri et al. (2012); Berger and Pericchi (2014)). In general, a Bayes factor is the ratio of posterior to prior odds in favor of one model over another, and offers a way of incorporating prior information into the evaluation of evidence about a hypothesis. We briefly review the Bayes factor here (for a detailed review, see Etz et al. (2017)).

Assume data $X = x$ have arisen under one of the two hypotheses; H_0 or H_1, according to probability distributions $\mathbb{P}(x \mid H_0)$ and $\mathbb{P}(x \mid H_1)$, respectively. Further assume some prior knowledge has suggested prior probabilities $\mathbb{P}(H_0)$ and $\mathbb{P}(H_1) = 1 - \mathbb{P}(H_0)$. Bayes Theorem combines the prior probabilities and the data to produce posterior probabilities $\mathbb{P}(H_0 \mid x)$ and

* Corresponding author.
E-mail addresses: su.chen@rice.edu (S. Chen), s.g.walker@math.utexas.edu (S.G. Walker).

https://doi.org/10.1016/j.ecosta.2021.10.009
2452-3062/

$\mathbb{P}(H_1 \mid x) = 1 - \mathbb{P}(H_0 \mid x)$. Therefore, the transformation of prior to posterior itself represents the evidence provided by the data, which takes a simple form,

$$\mathbb{P}(H_k \mid x) = \frac{\mathbb{P}(x \mid H_k)\mathbb{P}(H_k)}{\mathbb{P}(x \mid H_0)\mathbb{P}(H_0) + \mathbb{P}(x \mid H_1)\mathbb{P}(H_1)} \quad \text{for} \quad k = 0, 1,$$

so

$$\frac{\mathbb{P}(H_0 \mid x)}{\mathbb{P}(H_1 \mid x)} = \frac{\mathbb{P}(x \mid H_0)}{\mathbb{P}(x \mid H_1)} \frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)} = BF_{01} \frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)}, \tag{1}$$

where $BF_{01} = \mathbb{P}(x \mid H_0)/\mathbb{P}(x \mid H_1)$ denotes the Bayes factor, which can be described by words simply as: posterior odds = prior odds $\times$ Bayes factor.

From this definition, when a priori the two hypotheses H_0 and H_1 are assessed to be equally probable, the Bayes factor is the posterior odds in favor of H_0. When testing a "simple versus simple" hypothesis, the Bayes factor is the likelihood ratio. When there is an unknown parameter θ_k, corresponding to hypothesis H_k, the Bayes factor is the marginal likelihood ratio, where the marginal likelihood densities $m_k(x)$ are obtained by integrating over the parameter space with respect to the specified prior $\pi(\theta_k \mid H_k)$. For notation, we use subscript k and omit the conditioning on the hypothesis H_k; so

$$m_k(x) = \int f_k(x \mid \theta_k)\pi_k(\theta_k)d\theta_k, \quad \text{for} \quad k = 0, 1, \tag{2}$$

and $BF_{10} = m_1(x)/m_0(x)$.

However, there are difficulties with the Bayes factor when prior information about the unknown parameters of the models is weak, in particular, with the use of improper priors. This is due to an arbitrary constant involved with an improper prior. Assume improper priors $\pi_k^I(\theta_k)$ which can be written as $c_k h_k(\theta_k)$ with c_k being an arbitrary constant, and using the notation in (2), the Bayes factor can be written as

$$BF_{10} = \frac{c_1 \int f_1(x \mid \theta_1)h_1(\theta_1)d\theta_1}{c_0 \int f_0(x \mid \theta_0)h_0(\theta_0)d\theta_0}. \tag{3}$$

It is easy to see that whether improper priors are assigned to one or to both hypotheses, the corresponding Bayes factor depends on an unspecified constant $c = c_1/c_0$. Bartlett's paradox, Bartlett et al. (1957); Kass and Raftery (1995); Robert (2007), implies the Bayes factor is not well defined with improper priors.

Various solutions have been advocated for dealing with this problem. The most obvious one is to use proper priors. This means the hypotheses being tested under a Bayesian framework is not meaningful unless genuine prior information is available and such prior information can be represented by proper prior distributions. However, this restriction not only poses difficulty in practice, but further it is well known that Bayes factors are sensitive to the specification of prior distributions.

The approach of Aitkin (1991), also mentioned in Aitkin (1993), proposes the "Posterior Bayes factor", where the prior $\pi(\theta_k)$ in (2) is replaced by the posterior $\pi(\theta_k \mid x, H_k)$. This removes the issue of indeterminacy of the Bayes factor because posteriors are usually proper, even with improper priors. However, this double use of the data is at odds with standard statistical concepts.

A more concrete idea is to divide the observed data x into two parts $x = (y, z)$; the first part y can be combined with the improper prior distribution $\pi_k^I(\theta_i)$ to produce an intermediate posterior $\pi_k(\theta_k \mid y)$, which can serve as a proper prior for the remaining part of the data. The Bayes factor $BF(z \mid y)$ is then computed from the remaining part of the data z combined with this proper prior and is thus well–defined. The name "Partial Bayes factor" (PBF) is given to $BF(z \mid y)$ due to the use of partial data. It is straightforward to verify the following relationship between the "full" Bayes factor B_{01} and the partial Bayes factor, a consequence of the coherence of Bayes' Theorem under sequential updating: $BF(x) = BF(y)\,BF(z \mid y)$. PBF has been suggested by several authors including the earliest by Lempers (1971), then formally in Hagan (1991) and O'Hagan (1995).

In O'Hagan (1995) the author proposed using a fractional part of the entire likelihood instead of a training sample to calculate a "Fractional Bayes factor" (FBF). This can be viewed as another variant of the partial Bayes factor. Assume the entire dataset consists of n observations and the training dataset consists of m observations. Denote the proportion for training to be $b = m/n$. When both m and n are large enough, the likelihood $f_k(y \mid \theta_k)$ will approximate the full likelihood $f_k(x \mid \theta_k)$ raised to the power b. Formally, the FBF is defined as

$$BF_b(x) = \frac{\int f_1(x \mid \theta_1)^b \pi_1(\theta_1)d\theta_1}{\int f_0(x \mid \theta_0)^b \pi_0(\theta_0)d\theta_0} \frac{\int f_0(x \mid \theta_0)\pi_0(\theta_0)d\theta_0}{\int f_1(x \mid \theta_1)\pi_1(\theta_1)d\theta_1}. \tag{4}$$

This approach fixes the indeterminacy issue because, even if improper priors are used, the same arbitrary constants will cancel out. An advantage of this approach compared to partial Bayes factor is to avoid the arbitrariness of choosing a particular training sample.

A more general case of FBF can be defined as:

$$BF_{a,b}(x) = \frac{\int f_1(x|\theta_1)^b \pi_1(\theta_1)d\theta_1}{\int f_0(x|\theta_0)^b \pi_0(\theta_0)d\theta_0} \frac{\int f_0(x|\theta_0)^a \pi_0(\theta_0)d\theta_0}{\int f_1(x|\theta_1)^a \pi_1(\theta_1)d\theta_1}. \tag{5}$$

where the regular Bayes factor is $BF_{1,0}(x)$, the FBF is $BF_{1,b}(x)$, and the posterior Bayes factor is $BF_{2,1}(x)$.

Yet another proposal, named the"Intrinsic Bayes factor" (IBF), is introduced in Berger and Pericchi (1996). The idea is to calculate many partial Bayes factors based on different training samples, then take the average, either arithmetically or geometrically. The authors view the correspondence of IBFs to actual Bayes factors with respect to intrinsic priors to be their strongest justification, but others argue that asymptotic dependence is not so important as the consistency of Bayes factor ensures it will identify the true model with probability one regardless of the choice of priors. Furthermore, intrinsic priors may not necessarily exist and are typically not unique when they do.

More recent developments deviate from using Bayes factor. For example, Bernardo and Rueda (2002) proposed the Bayesian reference criterion (BRC) where hypothesis testing is considered as a formal decision problem. A loss function is defined to be the symmetrical logarithmic divergence between the likelihood of two models under comparison, and the posterior mean of the loss function under reference prior, named the "intrinsic statistic", is the test function. Other alternatives consider scoring rules, since the Bayes factor is associated with the predictive performance under the logarithmic scoring rule; see Dawid et al. (2015, 2017); Shao et al. (2019).

In this paper, we propose a new framework for Bayesian hypothesis testing. We argue the difference between a classical and a Bayesian approach to hypothesis testing should only be that the Bayesian test statistic is allowed to depend on the data and prior information. The conclusion to the test, as is the norm, is to reject the null model if some distance between the observed and expected statistic is too large. To expand on this, consider a test with test statistic $T(X)$. The null hypothesis H_0 is rejected if $d\Big(T(X), \mathbb{E}\big[T(X)|H_0\big]\Big)$ is too large, for some distance d. For example, the appropriate statistic for testing a normal mean with known variance is $T(X) = \sqrt{n}\bar{X}/\sigma$. Under the hypothesis $H_0 : \theta = \theta_0$, $\mathbb{E}[T(X) \mid H_0] = \sqrt{n}\theta_0/\sigma$ and the hypothesis is rejected if $|T(X) - \mathbb{E}[T(X) \mid H_0]|$ is too large.

An appropriate statistic to use in a Bayesian context is the posterior distribution itself; i.e., $T(X) = \pi(\cdot|X)$. Along the lines of the testing procedure, we would consider a distance, or divergence, between the observed posterior and the expected posterior under the null hypothesis; i.e. $\pi_0(\cdot) = \mathbb{E}[\pi(\cdot \mid X) \mid H_0]$. Due to its connection with information, we use the Kullback–Leibler (KL) divergence between the posterior distributions; hence, we reject H_0 if $d_{KL}\big(\pi_0(\cdot), \pi(\cdot|X)\big)$ is too large.

Some comments are in order. First, there should by now be no issues concerning assessing a Bayesian test statistic using what might be regarded as a classical routine; i.e. reject the null if the difference between the observed statistic and its expectation under the null is too large. A number of Bayesian procedures are almost now exclusively studied from a frequentist perspective; most notably asymptotic studies. Second, since the posterior is assumed to exist whether the prior is improper or not, the test associated with our new procedure will still have a valid interpretation even with improper priors.

1.1. Merging of information

Before we look at specific cases, we look briefy at the general form of the statistic and how it behaves asymptotically. So consider

$$T_n = \frac{1}{n}\int\int \pi_0(\theta,\phi)\,\log\frac{\pi_0(\theta,\phi)}{\pi(\theta,\phi \mid X)}\,d\theta\,d\phi,$$

where θ can be multi–dimensional and ϕ is a nuisance parameter possibly also being multi–dimensional. Here we will be taking the $\pi_0(\theta,\phi) = \int \pi(\theta,\phi \mid y)\, f(y \mid \theta_0, \widehat{\phi})\, dy$, where y is of sample size n and $\widehat{\phi}$ is a consistent estimator of ϕ.

Following section 4 in Barron (1988), posterior distributions π_0 and $\pi(\cdot \mid X)$ are said to merge if

$$\lim_{n\to\infty}\frac{1}{n}\,d_{KL}(\pi_0, \pi(\cdot \mid X)) = 0.$$

Subject to some standard regularity conditions which almost certainly hold in parametric cases, the counter–examples exist only for nonparametric models, if π_0 and $\pi(\cdot \mid X)$ are derived from data sets which come from the same model, then $T_n \to 0$. Hence, if H_0 holds and $\widehat{\phi}$ converges to the true value, then $T_n \to 0$. On the other hand, if the datasets are not from the same source, i.e. the observed data X are not following H_0, then T_n does not converge to 0 and instead converges to a positive constant. This forms the basis of our test.

In some special cases we look at, we can be more specific with the asymptotics, particularly when we make comparisons with the connection to the Bayes factor. However, when such math is not so tractable we can rely on the above merging of posterior probabilities for the overarching asymptotic results.

1.2. Layout of paper

The remainder of this article proceeds as follows. In section 2 we illustrate this approach in detail with the general case of parametric test, and we discover the key test statistic is a log Bayes factor plus a penalty term. We explore the consistency and other theoretical properties of this penalized log Bayes factor and show its equivalence of classical tests using sufficient statistics for the exponential family.

We discuss the case of unknown nuisance parameters and the case of composite null hypothesis. Section 3 presents two numerical studies with tests involving the linear regression model and random effect model, and a real data application using the stochastic volatility model. We conclude with a discussion in section 4 and explore some possible future directions.

2. New Bayesian statistic

We first introduce the new statistic for the model $f(x \mid \theta)$, with $\theta \in \Theta$, and for the hypothesis

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta \neq \theta_0, \tag{6}$$

with data $X = (X_1, \ldots, X_n) \overset{\text{iid}}{\sim} f(X \mid \theta)$. From this, with prior $\pi(\theta)$, we can construct the observed posterior $\pi(\theta \mid X)$, assumed to be a proper density function even if the prior is improper. Although we consider the simple null hypothesis for now, for the sake of introduction, we extend to composite hypotheses and the inclusion of nuisance parameters later in this section.

Here we describe what we compare $\pi(\theta \mid X)$ with. Assume we can generate data $Y = (Y_1, \ldots, Y_n) \overset{\text{iid}}{\sim} f(Y \mid \theta_0)$ under the null hypothesis. With the same prior $\pi(\theta)$ we can construct $\pi(\theta \mid Y)$ and hence we can define the expected posterior under the null; i.e. $\pi_0(\theta) = \int \pi(\theta \mid y)\, f(y \mid \theta_0)\, dy$, where y is an n–vector.

The test statistic of interest is the KL divergence between the expected posterior under the null and the observed posterior; i.e.

$$T(X) = \int \pi_0(\theta)\, \log \frac{\pi_0(\theta)}{\pi(\theta \mid X)}\, d\theta. \tag{7}$$

We also write this as $T(X) = KL\big(\pi_0(\cdot), \pi(\cdot \mid X)\big)$. Clearly, we would doubt the veracity of the null hypothesis if $T(X)$ is too large. We now show how $T(X)$ is related to the Bayes factor.

Lemma 1. *It is that* $T(X) = K + \log B_{10}(X) + P(X)$, *where*

$$P(X) = \log f(X \mid \theta_0) - \int \log f(X \mid \theta)\, \pi_0(\theta)\, d\theta,$$

and K *is a constant not depending on* X, *but does depend on the sample size* n *and null hypothesis* H_0[1],

$$K = \int \pi_0(\theta)\, \log \frac{\pi_0(\theta)}{\pi(\theta)}\, d\theta,$$

the KL divergence between the expected posterior under the null and the prior.

The proof for Lemma 1 is trivial.

Before looking more closely at $P(X)$ in general, here we consider a simple toy example whereby $f(x \mid \theta)$ is normal with unknown mean θ and known variance σ^2. We take a flat improper prior for θ; i.e. $\pi(\theta) = 1$. Then $\pi(\theta \mid X) = \mathrm{N}(\theta \mid \bar{X}, \sigma^2/n)$, and easy computations give $\pi_0(\theta) = \mathrm{N}(\theta \mid \theta_0, 2\sigma^2/n)$. Then it is easy to see that $P(X) = 1$, and so is a constant, hence, in this case, $T(X) = \frac{1}{2} + \log\frac{1}{2} + \frac{1}{2} n\bar{X}^2/\sigma^2$, while the log Bayes factor is $\log B_{10}(X) = \log(\sigma/\sqrt{n}) + \frac{1}{2} n\bar{X}^2/\sigma^2$. To get the $T(X)$ from the Bayes factor, we can take the "arbitrary" c appearing in the Bayes factor as $\log c = \frac{1}{2} + \log\frac{1}{2} - \log(\sigma/\sqrt{n})$. So writing $\widetilde{B}_{10} = c\, B_{10}$ with this c we recover $T(X) = \log \widetilde{B}_{10}$. The important difference here is that $T(X)$, even with improper priors, is calibrated, as it is the Kullback–Leibler divergence between two density functions. On the other hand, the Bayes factor with improper priors has an arbitrary constant and hence it is difficult to specify a value for deciding when to accept the null or alternative. Thus, if H_0 is rejected for $T(X) > \lambda$, the λ can be motivated as an information loss from the expected posterior, under the null model, to the observed posterior. Such a λ for the Bayes factor is problematic. We also note that while the log Bayes factor tends to 0 under the null as $n \to \infty$, it is rather the difference between the statistic under the null and alternative which is important, and this is of order n for both $\log B_{10}(X)$ and $T(X)$.

Another toy example to examine the $P(X)$ is provided by the exponential family; $f(x \mid \theta) = c(x) \exp\{\theta a(x) - b(\theta)\}$ and we test the hypotheses (6). We can derive the KL divergence focusing on the key terms that depend on the observed data; i.e.

$$T(X) = K + \log\left[\int \pi(\theta) \exp\{\theta A(x) - n\, b(\theta)\} d\theta\right] - \theta_0 A(X), \tag{8}$$

where $A(X) = \sum_{i=1:n} a(X_i)$, and where it is assumed the possibly improper prior is constructed such that $\mathrm{E}[\theta \mid \theta_0] = \theta_0$. So $T(X)$ is convex in the sufficient statistic $A(X)$, which makes it formally equivalent to the classical test. To consider $P(X)$ in this example, assume we can write $\theta = \theta_0 + \xi/\sqrt{n}$, from $\pi_0(\theta)$, where $\mathrm{E}[\xi] = 0$ and $\mathrm{Var}(\xi) = \tau^2$. So $P(X) = n\,[\int b(\theta)\pi(\theta \mid \theta_0)\, d\theta - b(\theta_0)]$, and using a Taylor expansion we get $P(X) \to \frac{1}{2} b''(\theta_0)\tau^2$. Once again, we can use this result to calibrate the Bayes factor with the improper prior; since $K + \frac{1}{2} b''(\theta_0)\tau^2 + \log B_{10}(X)$ represents a KL loss in information.

This latter toy example provides a picture as to how we can obtain a general result for $P(X)$. Assume $\pi(\theta \mid \theta_0)$ can still be written as $\theta = \theta_0 + \xi/\sqrt{n}$ with $\mathrm{E}[\xi] = 0$ and $\mathrm{Var}(\xi) = \tau^2$. This will follow from, for example, the asymptotic normality of $\pi_0(\theta)$ and would be a standard result.

[1] we suppress any subscript of n or H_0 and denote this term as K for simplicity of notation

Lemma 2. *Under regularity conditions on* $f(x \mid \theta)$[2], *and the previous condition for* $\pi_0(\theta)$, *it is that*

$$\int \log f(X \mid \theta)\, \pi_0(\theta)\, d\theta = \log f(X \mid \theta_0) + \tfrac{1}{2}\frac{\tau^2}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log f(X_i \mid \theta_0) + o(1/n).$$

Hence, under the null model, $P(X) \to \frac{1}{2}\tau^2 I(\theta_0)$, *where* $I(\theta)$ *denotes the Fisher information. Under an alternative with true parameter value* θ^*, $P(X) \to \frac{1}{2}\tau^2 I(\theta^*, \theta_0)$, *where*

$$I(\theta^*, \theta_0) = -\int \{\partial^2/\partial\theta^2 \log f(x \mid \theta)\}\, f(x \mid \theta^*)\, dx. \tag{9}$$

PROOF: See Appendix.

So once again we see that an adjustment to the Bayes factor with possibly improper priors, i.e.

$$T(X) = K + \log B_{10}(X) + \tfrac{1}{2}\tau^2 I(X \mid \theta_0) + o(1/n),$$

provides calibration for the Bayes factor, Here $I(X \mid \theta_0)$ is the sample Fisher information, i.e. $I(X \mid \theta_0) = n^{-1}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log f(X_i \mid \theta_0)$. That is, we are able to determine λ for which we reject H_0 when $T(X) > \lambda$, due to the interpretation of $T(X)$. There is a strong connection between the test statistic (7) and its asymptotic properties with the notion of merging of information with respect to posterior distributions, see Blackwell and Dubins (1962); Barron (1988). These articles are concerned with the Kullback–Leibler divergence between different posterior distributions based on the same sample; whereas we are looking at the divergence between the observed posterior and the expected posterior under the null model.

2.1. Multivariate parameter

In the d–dimensional case; i.e. $\Theta \subset \mathbb{R}^d$, a straightforward extension of Lemma 2 yields $P(X) \to \frac{1}{2}\text{trace}(I(\theta^*, \theta_0)\,\Sigma)$, where now $\theta = \theta_0 + \xi/\sqrt{n}$ and $\mathrm{E}[\xi] = 0$ and $\mathrm{E}[\xi\xi'] = \mathrm{Cov}(\xi) = \Sigma$, and $I(\theta^*, \theta_0)$ now corresponds to the $d \times d$ information matrix analogue of (9), so the (j,k)th element of I now is

$$I_{j,k}(\theta^*, \theta_0) = -\int \{\partial^2/\partial\theta_j\partial\theta_k \log f(x \mid \theta_0)\}\, f(x \mid \theta^*)\, dx.$$

In this case we have

$$T(X) = K + \log B_{10}(X) + \tfrac{1}{2}\text{trace}(I(X \mid \theta_0)\,\Sigma) + o(1/n).$$

2.2. Nuisance parameter

Here we look at the case when we consider the same hypothesis but now the model parameters are (θ, ϕ); i.e. we have the data model as $f(\cdot \mid \theta, \phi)$ where ϕ is deemed a nuisance parameter. We now take

$$\pi_0(\theta, \phi) = \int \pi(\theta, \phi \mid y)\, f\big(y \mid \theta_0, \widehat{\phi}\big)\, dy, \quad y = (y_{1:n}),$$

where $\widehat{\phi}$ is for example the maximum likelihood estimator from the data X. Using a plug–in point estimator to deal with nuisance parameters is widely adopted in both classical and Bayesian approaches. We only use this point estimator $\widehat{\phi}$ to simulate data under the null and we also assume that the distribution of $\widehat{\phi}$ does not depend on the true value of θ.

The statistic of interest here is

$$T(X) = \int\int \pi_0(\theta, \phi)\, \log \frac{\pi_0(\theta, \phi)}{\pi(\theta, \phi \mid X)}\, d\phi\, d\theta.$$

As previously, we have $T(X) = K + \log B_{10}(X) + P(X)$ where

$$B_{10}(X) = \frac{\int\int \pi(\theta, \phi)\, f(X \mid \theta, \phi)\, d\theta d\phi}{\int \pi(\phi)\, f(X \mid \theta_0, \phi)\, d\phi},$$

$$P(X) = \log \int \pi(\phi)\, f(X \mid \theta_0, \phi)\, d\phi - \int\int \pi_0(\theta, \phi)\, \log f(X \mid \theta, \phi)\, d\theta\, d\phi,$$

and

$$K = \int\int \pi_0(\theta, \phi)\, \log \frac{\pi_0(\theta, \phi)}{\pi(\theta, \phi)}\, d\theta\, d\phi.$$

Note now that K does depend on the data X. However, its distribution is the same under the null and the alternative, as it does not depend on the true value of θ. Note also we can accommodate improper priors for both θ and ϕ. The asymptotic results here essentially mimic those from Lemma 2.

[2] Essentially those which permit a Taylor expansion of $\log f(x \mid \theta)$

2.2.1. Illustration

Here we consider a normal mean test when the variance is unknown. So the $(X_i)_{i=1:n}$ are independent and identically distributed as $\mathrm{N}(\theta, \sigma^2)$ and we use the parameterization with $\lambda = 1/\sigma^2$ and prior $\pi(\theta, \lambda) = 1/\lambda$. Then

$$\pi(\theta, \lambda \mid X) = \mathrm{N}\Big(\theta \mid \bar{X}, 1/(n\lambda)\Big)\, G(\lambda \mid (n-1)/2, (n-1)S^2/2),$$

where $\bar{X}$ is the sample mean and S^2 the sample variance.

Under the null hypothesis $H_0 : \theta = \theta_0$ we have $\bar{X} = \mathrm{N}(\theta_0, \sigma^2/n)$ and so to get $\pi_0(\theta, \lambda)$ we use the estimator for the variance; so $f(\bar{X} \mid \theta_0, \lambda)$ is normal with mean θ_0 and variance S^2/n. Hence,

$$\pi_0(\theta, \lambda) = \mathrm{N}\Big(\theta \mid \theta_0, S^2/n + 1/(n\lambda)\Big)\, G(\lambda \mid (n-1)/2, (n-1)S^2/2).$$

Here we focus on the part of $T(X)$ which changes according to the null or alternative and show this remaining term is the classical t–statistic.

So consider

$$\int\int \pi_0(\theta, \lambda)\, \log \pi(\theta, \phi \mid X) =$$

$$C + \int\int \pi_0(\theta, \lambda) \left[\tfrac{1}{2} \log \lambda - \tfrac{1}{2}\, n\lambda(\theta - \bar{X})^2 + a \log S^2 - a\lambda S^2 + (a-1) \log \lambda\right] d\theta\, d\lambda,$$

where $a = (n-1)/2$ and C does not depend on the data. The only part of this which depends on the true value of θ, and hence changes according to the null or alternative hypothesis, is

$$n \int\int \pi_0(\theta, \lambda)\, \lambda\, (\theta - \bar{X})^2\, d\theta\, d\lambda.$$

This is given by

$$n \int \pi_0(\lambda)\, \lambda \left(\theta_0^2 + S^2/n + 1/(n\lambda) - 2\theta_0 \bar{X} + \bar{X}^2\right) d\lambda = \psi(S^2) + n\, (\theta_0 - \bar{X})^2/S^2,$$

where ψ is a function of S^2 only, and n.

Given that the distribution of S^2 does not depend on whether the null or alternative hypothesis is true, the decision is only determined by the usual t–statistic; i.e. $n(\theta_0 - \bar{X})^2/S^2$.

2.3. Composite hypotheses

Here we consider the case when testing hypothesis $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \notin \Theta_0$. As with the point null hypothesis, we take the statistic of interest to be

$$T(X) = \int \pi(\theta \mid \Theta_0)\, \log \frac{\pi(\theta \mid \Theta_0)}{\pi(\theta \mid X)} d\theta,$$

where now we take

$$\pi(\theta \mid \Theta_0) = \int \pi(\theta \mid y)\, f(y \mid \widehat{\theta}_0)\, dy,$$

$y = (y_{1:n})$ and $\widehat{\theta}_0$ is the data driven maximum likelihood estimator of θ restricted to Θ_0. Clearly, for the point null hypothesis, i.e. $\Theta_0 = \{\theta_0\}$, we recover $\pi(\theta \mid \Theta_0) = \pi(\theta \mid \theta_0)$. Further, arbitrary constants associated with an improper prior still cancel out.

The new test statistic is

$$T(X) = K + \log m(X) - \int \log f(X \mid \theta)\, \pi(\theta \mid \Theta_0)\, d\theta,$$

where $m(X)$ is the marginal likelihood defined in (2), i.e. $m(X) = \int f(X \mid \theta)\, d\theta$. We will focus on, as we did in the nuisance parameter case, the part of $T(X)$ which has a distribution depending on the hypothesis, so we consider

$$S(X) = n^{-1}\left[\log m(X) - \int \log f(X \mid \theta)\, \pi(\theta \mid \Theta_0)\, d\theta\right],$$

and the aim now is to show that if $\theta^* \in \Theta_0$, where θ^* is the true value of θ, then $S(X) \to 0$, whereas if $\theta^* \notin \Theta_0$ then $S(X)$ converges to a positive number.

According to Theorem 1 in Barron (1985), under mild regularity conditions, the term $n^{-1} \log m(X)$ converges to

$$\int \log f(x \mid \theta^*)\, f(x \mid \theta^*)\, dx.$$

We present this result here;

Theorem 1. *If the prior $\pi(\theta)$ puts positive mass on all KL neighborhoods of θ^*; i.e.*

$$\Pi\Big(\theta : KL\big(f(\cdot \mid \theta^*), f(\cdot \mid \theta\big) < \epsilon\Big) > 0,$$

for all $\epsilon > 0$ then

$$n^{-1} \log m(X) \to \int \log f(x \mid \theta^*)\, f(x \mid \theta^*)\, dx \quad \text{a.s.}$$

Proof. Define

$$I_n = \frac{m(X)}{\prod_{i=1}^n f(X_i \mid \theta^*)}.$$

Following Schwartz (1965), the fact that $\mathbb{E}_{\theta^*}[I_n] = 1$, and an application of the Markov inequality, it is that $\sum_{n=1}^{\infty} \mathbb{P}(I_n < e^{-nc} \cup I_n > e^{nc}) < \infty$. By Borel–Cantelli theorem, we have $e^{-nc} \le I_n \le e^{nc}$ a.s. for all large n, for any $c > 0$. Hence, $-c \le n^{-1} \log I_n \le c$ a.s. for all large n for any $c > 0$ indicating that $n^{-1} \log I_n \to 0$ a.s. Hence, provided $\int \log f(x \mid \theta^*)\, f(x \mid \theta^*)\, dx$ is finite, the proof is complete. □

We now turn attention to $\int \log f(X \mid \theta)\, \pi(\theta \mid \Theta_0)\, d\theta$ and first consider $\pi(\theta \mid \Theta_0)$. Let $\widehat{\theta}$ be the maximum likelihood estimator of $\prod_{i=1}^n f(Y_i \mid \theta)$, and assume

$$n^{-1} \sum_{i=1}^{n} \log \frac{f(Y_i \mid \widehat{\theta})}{f(Y_i \mid \widehat{\theta}_0)} \to 0 \quad \text{a.s.} \tag{10}$$

This is a reasonable condition since the (Y_i) are from $f(\cdot \mid \widehat{\theta}_0)$. Further, we assume, with the standard conditions on a maximum likelihood estimator, that $\widehat{\theta}_0 \to \theta_\infty$, where

$$\theta_\infty = \begin{cases} \theta^* & \theta^* \in \Theta_0 \\ \widetilde{\theta} & \theta^* \notin \Theta_0, \end{cases}$$

and $\widetilde{\theta} = \arg\inf_{\theta \in \Theta_0} KL(f(\cdot \mid \theta^*), f(\cdot \mid \theta))$, (see White (1982)), and that

$$n^{-1} \sum_{i=1}^{n} \log \frac{f(Y_i \mid \theta_\infty)}{f(Y_i \mid \widehat{\theta}_0)} \to 0 \quad \text{a.s.} \tag{11}$$

Lemma 3. *Under conditions (10) and (11), and if the prior $\pi(\theta)$ puts positive mass on all KL neighborhoods of θ^*; i.e.*

$$\pi\Big(\theta : KL(f(\cdot \mid \theta^*), f(\cdot \mid \theta) < \epsilon\Big) > 0, \tag{12}$$

for all $\epsilon > 0$, we have $\pi(A_\epsilon \mid \Theta_0) \to 0$ a.s. for all $\epsilon > 0$, where $A_\epsilon = \{\theta : d_H(f(\cdot \mid \theta), f(\cdot \mid \theta_\infty) \ge \epsilon\}$ and

$$d_H(f, g) = \left(\int (\sqrt{f} - \sqrt{g})^2 \right)^{1/2},$$

is the Hellinger distance between density functions f and g.

PROOF: See Appendix.

The implication is that $\pi(\theta \mid \Theta_0)$ converges to a point mass at θ_∞. Hence, assuming the variance of $\log f(X \mid \theta)$ exists for all θ, we have

$$n^{-1} \int \log f(X \mid \theta)\, \pi(\theta \mid \Theta_0)\, d\theta \to \int \log f(x \mid \theta_\infty)\, f(x \mid \theta^*)\, d\theta.$$

Hence, as claimed, if $\theta^* \in \Theta_0$ then $S(X) \to 0$, whereas if $\theta^* \notin \Theta_0$ then $S(X)$ converges to a positive number.

Finally we consider $K = \int \pi(\theta \mid \Theta_0) \log \pi(\theta \mid \Theta_0)\, d\theta$ and assume that for large n we have an asymptotic normal approximation to $\pi(\theta \mid \Theta_0)$ given by $N(\widehat{\theta}_0, v^2)$. It is then easy to show that K does not depend on $\widehat{\theta}_0$ and so asymptotically K does not depend on which hypothesis is true.

2.3.1. Illustration

Here we consider the model for which (X_i) are independent and identically distribution from a normal with mean θ and known variance σ^2. The hypothesis is $H_0 : \theta > 0$ vs $H_1 : \theta < 0$. Then with a flat prior for θ, we have $\pi(\theta \mid X) = \mathrm{N}(\bar{X}, \sigma^2/n)$. Also, $\pi(\theta \mid \Theta_0) = \mathrm{N}(\widehat{\theta}_0, 2\sigma^2/n)$, where $\widehat{\theta}_0 = \bar{X}\, \mathbf{1}(\bar{X} > 0)$. Then the key part of $T(X)$ which changes depending on which hypothesis is true is

$$-\tfrac{1}{2} \frac{n}{\sigma^2} \int \pi(\theta \mid \Theta_0)\, (\theta - \bar{X})^2\, d\theta = C - \tfrac{1}{2} \frac{n}{\sigma^2} (\widehat{\theta}_0 - \bar{X})^2,$$

where C is a constant. This then would be equivalent to use of the classical test statistic based on $\bar{X}$ since $(\widehat{\theta}_0 - \bar{X})^2 = 0$ if $\bar{X} > 0$ and is $\bar{X}^2$ otherwise.

Table 1
Examples of some classical tests.

Hypothesis test	$T^c(X)$	$T(X,Y)$	prior
Binomial testTest[a]	$\sum_{i=1:n} X_i$	$\log\left[\frac{(\sum_{i=1:n} X_i)!(n-\sum_{i=1:n} X_i)!}{(\sum_{i=1:n} Y_i)!(n-\sum_{i=1:n} Y_i)!}\right]$	$\pi(\theta) \propto 1$
χ^2–test	$\frac{\sum_{i=1:n} X_i^2}{\sigma_0^2}$	$n\left(\frac{SS_X}{SS_Y} - \log SS_X\right)$[b]	$\pi(\sigma^2) \propto 1/\sigma^2$

[a] $H_0 : \theta = 1/2$ vs. $H_1 : \theta \neq 1/2$
[b] $SS_X = \sum_{i=1}^n (X_i - \mu)^2$

2.4. Classical tests

We have already seen in a number of illustrations how we obtain classical test statistics using the KL statistic based on improper priors. In this subsection we illustrate how to derive some further classical tests. To this end, we introduce the following notation:

$$T(X,Y) = \int \pi(\theta \mid Y) \log \frac{\pi(\theta \mid Y)}{\pi(\theta \mid X)}\, d\theta.$$

where data $Y = (Y_1, \ldots, Y_n)$ are simulated under H_0. Therefore $T(X,Y)$ is defined the same way as $T(X)$ but replacing $\pi_0(\theta)$ by $\pi(\theta \mid y)$, i.e.,

$$T(X) = \mathbb{E}_{f(y|\theta_0)}[T(X,Y)].$$

In practice, we would compare $T(X,Y)$ with $T(Y^{(m)},Y)$ by repeatedly sample $\{Y^{(m)}, m = 1, \ldots, M\}$ in the same way we sample Y. Here $T(Y^{(m)},Y)$ is the value of the test statistic as if the observed data were actually from $f(y \mid \theta_0)$, and represents the empirical sampling distribution of the KL statistic under the null. When there is nuisance parameter involved, this is the same idea as parametric bootstrap. The decision rule would be using quantile of the sampling distribution as cut-off. This whole process can be repeated by simulating more Y and the bootstrap for each Y. In this way, the cut-off quantile could be motivated by controlling the empirical Type I error rate. We demonstrate this implementation in Section 3 using simulation study and a real data application.

In the examples shown in Table 1, we calculate $T(X,Y)$ and compare to the classical test statistic $T^c(X)$.

For these tests of a simple null hypothesis and no nuisance parameter case, it is easy to verify $T(X,Y)$ is equivalent to the classical test statistic once we take the expectation of Y.

2.4.1. Exponential family

Here we consider the general form for exponential family

$$f(x \mid \theta) = c(x)\, \exp\{\theta T(x) - b(\theta)\},$$

and the test of the hypothesis $H_0 : \theta = \theta_0$. With conjugate prior $\pi(\theta) \propto \exp\{\alpha\theta - \beta b(\theta)\}$ the posterior is

$$\pi(\theta \mid X) = \frac{\exp\{(\alpha + T_n(X))\theta - (\beta + n)b(\theta)\}}{\gamma(\alpha + T_n(X), \beta + n)},$$

where $\gamma(\alpha,\beta) = \int \exp\{\alpha\theta - \beta b(\theta)\}\, d\theta$ and $T_n(X) = \sum_{i=1:n} X_i$. Note that $d\gamma/d\alpha = \int \theta \exp\{\alpha\theta - \beta b(\theta)\}\, d\theta$.

In finding $\int \pi(\theta \mid Y) \log(\pi(\theta \mid X)\, d\theta$, we obtain the part of $T(X)$ which depends on X as $\log \gamma(\alpha + T, \beta + n) - T\, E(\theta \mid Y)$, where we now write $T = T_n(X)$ as the sufficient test statistic. Hence, our test statistic is

$$\phi(T) = \log \gamma(\alpha + T, \beta + n) - T\, E_0\left[\frac{d}{dt}\gamma(\alpha + t, \beta + n)\Big|_{t=T_n(Y)}\right],$$

where E_0 is expectation with respect to the null value, so $E_0\, d\phi/dT = 0$.

3. Numerical Studies

3.1. Linear regression model

In this numerical study with simulated data, we compare results with the Fractional Bayes Factor (FBF) and the Intrinsic Bayes Factor (IBF) in linear regression model when testing a single regression coefficient. This example is to show empirically the behavior of KL statistic under the null and the alternative using small sample size, compared to FBF and IBF.

Suppose we have observed data X from a multivariate normal distribution $p(\cdot \mid \beta, \tau) \sim \mathrm{N}(\cdot \mid Z\beta, (\tau V)^{-1})$, and assume Z and V are known, and we are interested in testing the following hypothesis,

$$H_0 : (\beta, \tau) = (\beta_0, \tau_0) \quad \text{vs} \quad H_1 : (\beta, \tau) \neq (\beta_0, \tau_0). \tag{13}$$

Fig. 1. Comparison of KL statistic with Fractional and Intrinsic Bayes Factor

Using the conjugate normal–gamma prior, we can calculate the KL divergence between the expected posterior and the observed posterior. So $\pi(\beta,\tau)=\mathrm{N}(\beta\mid\mu_0,(\tau\Lambda_0)^{-1})\,\mathrm{Ga}(\tau\mid a_0,b_0)$ and hence

$$\pi(\beta,\tau\mid X)=\mathrm{N}(\beta\mid\mu_n,(\tau\Lambda_n)^{-1})\,\mathrm{Ga}(\tau\mid a_n,b_n),$$

where $\mu_n=\Lambda_n^{-1}(Z'VX+\Lambda_0\mu_0)$, $\Lambda_n=Z'VZ+\Lambda_0$, $a_n=a_0+n/2$, and

$$b_n=b_0+\tfrac{1}{2}\left(X'VX+\mu_0'\Lambda_0\mu_0-\mu_n'\Lambda_n\mu_n\right).$$

For simplicity, we assume $\mu_0=0$, $a_0=b_0=0$, $V=I$, then we have $\tilde{H}=Z(Z'Z+\Lambda_0)^{-1}Z'$, and $Y\sim\mathrm{N}(\cdot\mid X\beta_0,(\tau_0V)^{-1})$. Denote $\mathrm{RSS}_X=X^T(I-\tilde{H})X$, then we have

$$T(X,Y)=\frac{n}{2}\left[\frac{\mathrm{RSS}_X}{\mathrm{RSS}_Y}-\log\mathrm{RSS}_X+\frac{(Y-X)^T\tilde{H}(Y-X)}{\mathrm{RSS}_Y}\right]. \tag{14}$$

Data is simulated from the following linear regression model;

$$y_i=\beta_1x_{i1}+\beta_2x_{i2}+\beta_3x_{i3}+\beta_4x_{i4}+\beta_5x_{i5}\quad\text{for } i=1,\ldots,n \tag{15}$$

where the number of dimensions is $p=5$ and the sample size n increases from 20 to 200. The true coefficients $\beta_2=\beta_4=0$, representing null hypothesis being true case; while β_1,β_3 and β_5 are non-zero values, representing the alternative hypothesis being true. For each sample size, we calculate the KL statistic, and compare with FBF and IBF, both under the log scale, see Fig. 1.

We further illustrate some settings in the simulation for Fig. 1 and make the following observations:

- All three are calculated under the default improper prior for normal linear model.
- For the IBF, we pick the training sample sized to be fixed at 10, when the sample size of the entire dataset increases from 20 to 200. This is not necessarily the minimal training sample, but a fixed training sample size is required to achieve consistency for both IBF and FBF. Here IBF is calculated by randomly split the dataset into training and remaining 100 times and taking the arithmetic average of Bayes Factor.
- Similarly, the fraction for FBF is fixed at $10/n$ to make it comparable to IBF. According the asymptotic theory of FBF, such choice is adequate to achieve good asymptotic behavior as long as outliers is not a concern for the data.
- The KL statistic is calculated by simulating data under the null model 100 times and using the average value. Therefore it is not exactly the definition in (7) but rather a Monte Carlo approximation.
- Under the null, both IBF and FBF converges to 0 as sample size increases, while the KL statistic converges to some constant. This is the case according to Lemma 1, as KL statistic can be written as a sum of a constant, a log Bayes factor and a penalty term. We argue this is not a concern because the focus should be the difference of the test statistic under the null and the alternative, rather than the absolute values.
- The penalty term converges according to Lemma 2, and importantly the difference between the KL statistic under the null and alternative is of the order n: log Bayes factor under the alternative increases as order of n.
- Compared to IBF and FBF, the KL statistic does not require data splitting or a choice of fraction, where this fraction is not really interpretable.

Another way to calibrate the KL statistic is to divide by $\sqrt{n}$, so that under null it goes to 0 at a rate of $1/\sqrt{n}$ and under alternative it goes to infinity at speed of $\sqrt{n}$. This scaling corrects the uneven convergence rates of Bayes factor under the null and the alternative, and ensures the convergence to 0 under the null. The following Fig. 2 shows this scaled KL statistic under the same setting.

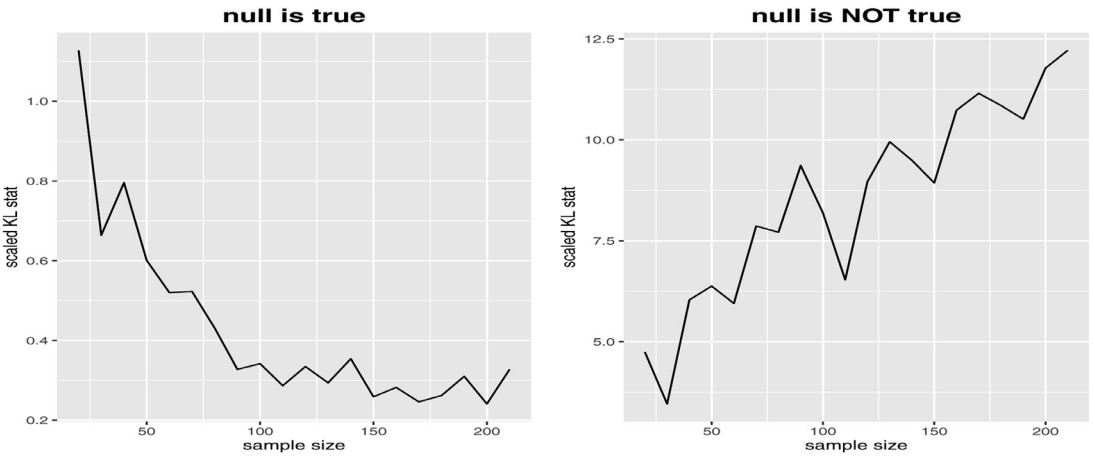

Fig. 2. Scaled KL statistic divided by $\sqrt{n}$

3.2. Random effects model

In this numerical study with simulated data, we present a random effect model where Markov chain Monte Carlo is carried out to compute the KL statistic, and compare with its empirical distribution under the null hypothesis to show its easy calibration. We consider the following model:

$$\mu_j \sim \mathrm{N}(\mu, \tau^2) \quad \text{for } j = 1, \ldots, J, \quad x_{ij} \sim \mathrm{N}(\mu_j, \sigma^2) \quad \text{for } i = 1, \ldots, N_j, \tag{16}$$

and we test the following hypothesis, equivalent to a one–way ANOVA test,

$$H_0 : \tau^2 = 0 \quad \text{vs} \quad H_1 : \tau^2 \neq 0.$$

So $(\mu, \mu_1 \ldots \mu_J, \sigma^2)$ are nuisance parameters, and our test statistic is given by: $T(X, Y) =$

$$\int\int \pi(\mu, \tau^2, \mu_1, \ldots, \mu_J, \sigma^2 \mid Y) \log \frac{\pi(\tau^2 \mid \mu, \mu_1, \ldots, \mu_J, \sigma^2, Y)}{\pi(\tau^2 \mid \mu, \mu_1, \ldots, \mu_J, \sigma^2, X)} \, d\mu \, d\mu_1 \ldots d\mu_J \, d\sigma^2,$$

where the simulated data Y under the null is sampled as

$$y_{ij} \sim \mathrm{N}\big(\bar{X}_j, \mathrm{Var}(X)\big) \quad \text{for } i = 1, \ldots, N_j, \quad j = 1, \ldots, J,$$

where $\bar{X}_j = N_j^{-1} \sum_{i=1}^{N_j} x_{ij}$ and $\mathrm{Var}(X)$ is the sample variance of the entire dataset. We put standard improper priors on parameters:

$$\pi(\mu, \tau^2) \propto 1/\tau^2 \quad \text{and} \quad 1/\sigma^2 \propto 1,$$

and estimate $T(X, Y)$ with a Monte Carlo approximation using 1000 MCMC samples carried out via a Gibbs sampler. Notice the posterior in this case is proper.

We first simulate data with $J = 3$ and sample size of each group $N_j = 100$. In Fig. 3 and 4, the histogram represents the sampling distribution of the KL test statistic under the null: distribution of $T(Y^{(m)}, Y)$ with simulated $(Y^{(m)})_{m=1:100}$ setting $\tau^2 = 0$ (sample mean and standard deviation of $T(Y^{(m)}, Y)$ marked with vertical lines). The red dot is the same test statistic $T(X, Y)$ on observed data X, simulated with different value of τ/σ to represent different signal to noise ratio. As we see in Fig. 3, even when signal to noise ratio is only 0.5, $T(X, Y)$ is more than 3 standard deviation away from the mean of $T(Y^{(m)}, Y)$, indicating great power to reject the null hypothesis. When signal to noise ratio is as small as 0.1, which translate to within group variance being one hundred times as large as between group variance, we fail to reject the null.

We carry out the same simulation with $J = 10$ and identical setting otherwise, see Fig. 4. Now the total sample size increased from 300 to 1000, which increased the power, especially for the case where signal to noise ratio is 0.1: now our test statistic is more than 2 standard deviation away compared to Fig. 3.

3.3. Real Data

In this real data application, we consider the log–normal stochastic volatility model as described in Taylor (2008):

$$[y_t \mid x_t] \sim \mathrm{N}(0, \exp(x_t)), \quad [x_t \mid \eta_t] = \theta x_{t-1} + \eta_t, \quad \eta_t \sim \mathrm{N}(0, \sigma^2),$$

where the (y_t) are stock index returns. The (x_t) are the log–volatilities which are unobserved, and represent the random and uneven flow of new information. The (η_t) are independent Gaussian white noise with variance σ^2.

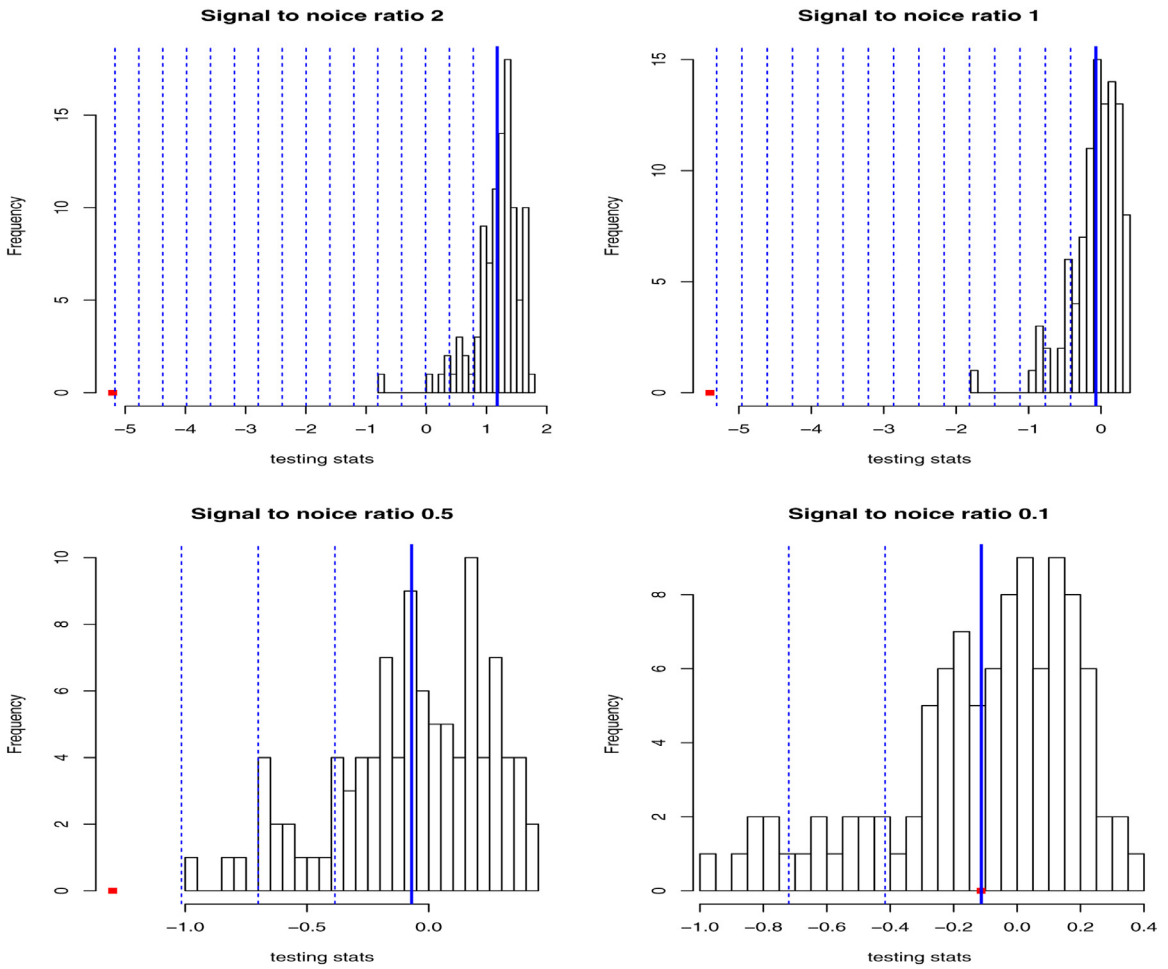

Fig. 3. Histogram of $T(Y_k)$: blue solid line is mean of $T(Y_k)$, blue dash line is one standard deviation away from from mean, red mark is $T(X)$

The data we use are the daily records of the NYSE Composite Index from the R package "fBasics"; see Wuertz et al. (2017). The stock index return y_t is calculated on a continuously compounded basis and expressed as a percentage, i.e., $y_t = 100\log(P_t/P_{t-1}))$, where P_t denotes the stock index on day t. The real data include the daily index value from 01/04/1966 to 12/31/2002. For computational issues, we only use the most recent 500 observations.

We test the hypothesis: $H_0 : \theta = 0$ vs $H_1 : \theta \neq 0$. The following steps are carried out in order to simulate data under the null and calculate the KL statistic:

- Posterior inference with standard non-informative prior $\pi(\theta, \sigma^2) \propto 1/\sigma^2$ via MCMC:
  - Use Gibbs sampler to sample θ and σ^2 under conditional conjugacy.
  - Use Metropolis-Hastings step to sample x_t with a random walk proposal.
- Data under the null is simulated using the posterior mean as the point estimator of σ^2, and the null hypothesis assumption that $\theta = 0$.
- Posterior densities estimated using kernel density estimate with function"kde" using the default setting in the R package "ks".
- Posterior inference and posterior density estimates need to be done for the observed data as well as every set of simulated data.
- The KL divergence is calculated using Monte Carlo approximation and the posterior samples of θ under the null hypothesis.

The Fig. 5 shows the histogram of $T(Y^{(m)}, Y)$, i.e. the KL statistic under the null hypothesis, with Y and $Y^{(m)}, m = 1, \ldots, 100$, all simulated under the null hypothesis; calculated using kernel density estimators and Monte Carlo approximation. The observed KL statistic using the "nyse" data is $T(X, Y) = 12.54$. This observed value is in excess of any simulated value under the null hypothesis and hence the outcome is the clear choice to reject the hypothesis $H_0 : \theta = \theta_0$. While we

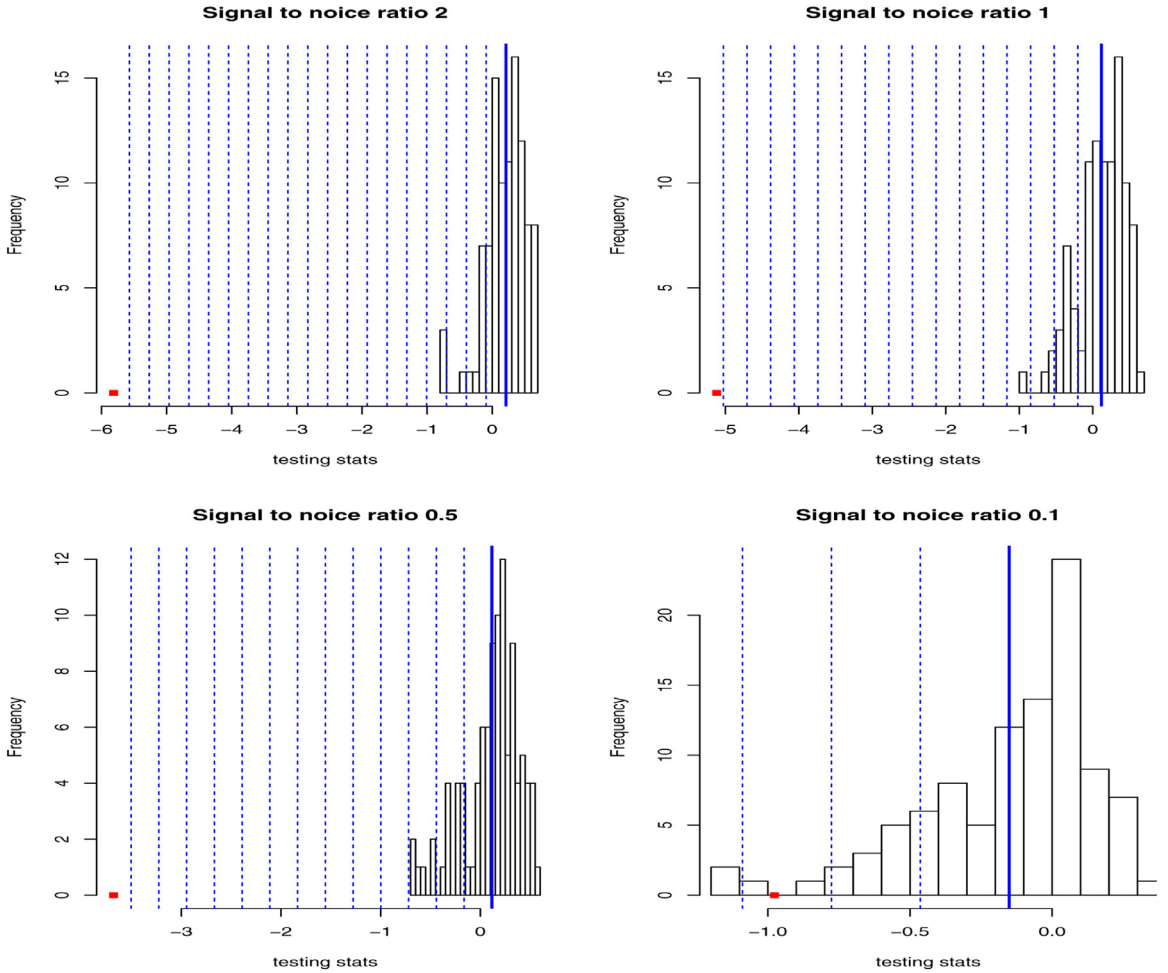

Fig. 4. Same setting as in Fig. 3, but with $J = 10$.

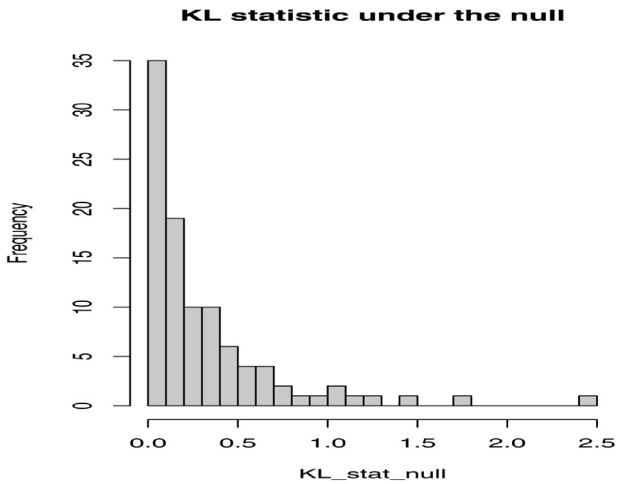

Fig. 5. Histogram of $T(Y^{(m)}, Y)$ with $m = 1, \ldots, 100$

did this using a single simulated data set Y in order to get the approximate distribution of the KL statistic under the null assumption, repetition of the experiment with different Y led to the same conclusion.

To be more precise concerning a "correct" procedure, we want to compute

$$I = \int \log \frac{\pi_0(\theta)}{\pi(\theta \mid X)} \pi_0(\theta)\, d\theta,$$

when the posterior can only be sampled; i.e. it can not be written down. To sample from $\pi_0(\theta)$ we would take Y from the null model and run a small chain to get a single θ from the posterior given Y. We repeat this to get many samples from π_0; say $\theta^{(j)}$ for $j = 1, \ldots, N$. We also have samples from the observed posterior and with the samples from the π_0 we construct the corresponding kernel density estimators, write as $\widehat{\pi}(\cdot \mid X)$ and $\widehat{\pi}_0(\cdot)$, respectively. Hence, we approximate I by

$$\widehat{I} = N^{-1} \sum_{i=1}^{N} \log \left(\frac{\widehat{\pi}_0(\theta^{(j)})}{\widehat{\pi}(\theta^{(j)} \mid X)} \right).$$

We should find the distribution, or get a sample from, the distribution of I if X is from the null model. This is easy to do, sample many $(X_b)_{b=1}^{B}$ from the null model and for each b get $\widehat{I}_b$ in the same way we got $\widehat{I}$. We then compare the observed statistic with the statistics distributed under the null.

To conclude, the KL statistic in this example of stochastic volatility model is easy to implement with MCMC and kernel density estimation. By generating a sampling distribution of the test statistic under the null hypothesis, it provides meaningful calibration and interpretation which facilitates decision making.

4. Discussion

In this paper we have proposed a new approach for Bayesian hypothesis testing using the KL divergence between the expected posterior distribution under the null model and the observed posterior. As a statistic it is calibrated and interpretable and hence cut–off values for determining whether to accept or reject the null can be well set. Further, since it is based solely on posterior distributions, improper priors can be used. As we have shown, a component of the statistic is the log Bayes factor and so the asymptotic properties of the new statistic follows a well trodden path.

We believe this approach, given its interpretation, proximity to the Bayes factor, and ability to accommodate improper priors, as well as it being easy to compute using Monte Carlo methods, makes it a significant contribution to Bayesian hypothesis testing.

Finally, we comment on the Bayesianity of the procedure. Just as with the Bayes factor, the new statistic is a combination of prior information and data. Establishing a critical region for a Bayesian statistic is not so clear nor well documented in the literature. There is for example only a rule for thumb for the Bayes factor. It is not necessarily a violation of Bayesian thinking to use a frequentist style critical region based on Type I errros since it could equally be argued the Bayesianity of the test is the form of the statistic.

Acknowledgments

The authors are grateful for the detailed comments and suggestions of two anonymous reviewers.

Appendix A

A1. Proof of Lemma 2

Using the order 2 Taylor expansion of $\log f(x \mid \theta)$ around θ_0:
$\log f(x \mid \theta_0) +$

$$\left[\frac{\partial}{\partial \theta} \log f(x \mid \theta) \mid_{\theta=\theta_0} \right] (\theta - \theta_0) + \frac{1}{2} \left[\frac{\partial^2}{\partial \theta^2} \log f(x \mid \theta) \mid_{\theta=\theta_0} \right] (\theta - \theta_0)^2 + o((\theta - \theta_0)^2).$$

We assume the posterior $\pi(\theta \mid \theta_0)$ admits Gaussian approximation locally around θ_0, i.e., can be written as $\theta = \theta_0 + \xi/\sqrt{n}$ with $\mathrm{E}[\xi] = 0$ and $\mathrm{Var}(\xi) = \tau^2$, the rest of the proof is trivial.

A2. Proof of Lemma 3

The proof relies on ideas from Walker and Hjort (2001) who discuss posterior asymptotic in a parametric setting when maximum likelihood estimators exist and behave suitably. The posterior is given by:

$$\pi(A_\epsilon \mid Y_{1:n}) = \frac{\int_{A_\epsilon} R_n(\theta)\, \pi(\theta)\, d\theta}{\int R_n(\theta)\, \pi(\theta)\, d\theta},$$

where $R_n(\theta) = \prod_{i=1:n} f(Y_i \mid \theta)/f(Y_i \mid \widehat{\theta}_0)$. We will consider the lower bound on the denominator first. It is well known that (12) implies

$$\int \prod_{i=1}^{n} \frac{f(Y_i \mid \theta)}{f(Y_i \mid \theta_\infty)} \pi(\theta)\, d\theta > e^{-nc},$$

a.s. for all large n for any $c > 0$. Hence, using (11), we can also deduce that $\int R_n(\theta)\pi(\theta)\, d\theta > e^{-nc}$ for all large n, for any $c > 0$.

For the numerator, we find the initial upper bound as

$$R_n(\widehat{\theta})^{\frac{1}{2}} \int_{A_\epsilon} R_n(\theta)^{\frac{1}{2}} \pi(\theta)\, d\theta.$$

Then (10) implies $R_n(\widehat{\theta})^{\frac{1}{2}} < e^{nc'}$ a.s. for all large n for any $c' > 0$. Further

$$E \int_{A_\epsilon} R_n(\theta)^{\frac{1}{2}} \pi(\theta)\, d\theta = \int_{A_\epsilon} \left(1 - \tfrac{1}{2} d_H^2(f(\cdot \mid \theta), f(\cdot \mid \widehat{\theta}_0)\right)^n \pi(\theta)\, d\theta, \tag{A.1}$$

where the expectation is with respect to the $Y_{1:n}$ while conditioning on $\widehat{\theta}_0$. For suitably large n, we have $d_H(f(\cdot \mid \widehat{\theta}_0, f(\cdot \mid \theta_\infty)) < \frac{1}{2}\epsilon$, and using the triangular inequality we can show that (A.1) is upper bounded by $\exp(-n\epsilon^2/8)$. Hence, by Borel–Cantelli,

$$\int_{A_\epsilon} R_n(\theta)^{\frac{1}{2}} \pi(\theta)\, d\theta < \exp(-c'' n\epsilon^2), \qquad \text{a.s.}$$

for all large n for some $c'' > 0$. Putting all these results together yields the conclusion that $\pi(A_\epsilon \mid \Theta_0) \to 0$ a.s.

References

Aitkin, M., 1991. Posterior Bayes factors. Journal of the Royal Statistical Society, Series B (Methodological) 53 (1), 111–128.
Aitkin, M., 1993. Posterior Bayes factor analysis for an exponential regression model. Statistics and Computing 3 (1), 17–22.
Barron, A.R., 1985. The strong ergodic theorem for densities: Generalized Shannon–McMillan–Breiman theorem. The Annals of Probability 13, 1292–1303.
Barron, A.R., 1988. The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Technical Report 7, University of Illinois, Urbana.
Bartlett, P.L., Jordan, M.I., Mcauliffe, J.D., 1957. A comment on D. In: V. Lindleys Statistical Paradox, Biometrika. Citeseer.
Bayarri, M.J., Berger, J.O., Forte, A., García-Donato, G., et al., 2012. Criteria for Bayesian model choice with application to variable selection. The Annals of Statistics 40 (3), 1550–1577.
Berger, J., 1997. Bayes factors. in the Encyclopedia of Statistical Sciences. Update 3, 20–29.
Berger, J., Pericchi, L., 2014. Bayes factors. Wiley StatsRef: Statistics Reference Online 1–14.
Berger, J.O., Pericchi, L.R., 1996. The intrinsic Bayes factor for model selection and prediction. Journal of the American Statistical Association 91 (433), 109–122.
Bernardo, J.M., Rueda, R., 2002. Bayesian hypothesis testing: A reference approach. International Statistical Review 70 (3), 351–372.
Biau, D.J., Jolles, B.M., Porcher, R., 2010. P value and the theory of hypothesis testing: an explanation for new researchers. Clinical Orthopaedics and Related Research® 468 (3), 885–892.
Blackwell, D., Dubins, L., 1962. Merging of opinions with increasing information. Annals of Mathematical Statistics 33, 882–886.
Dawid, A.P., Musio, M., Columbu, S., 2017. A note on Bayesian model selection for discrete data using proper scoring rules. Statistics & Probability Letters 129, 101–106.
Dawid, A.P., Musio, M., et al., 2015. Bayesian model selection based on proper scoring rules. Bayesian Analysis 10 (2), 479–499.
Etz, A., Wagenmakers, E.-J., et al., 2017. Jbs haldaneas contribution to the Bayes factor hypothesis test. Statistical Science 32 (2), 313–329.
Hagan, A., 1991. Discussion on posterior Bayes factors (by m. aitkin). Journal of the Royal Statistical Society, Series B 53, 136.
Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Statistical Science 382–401.
Jeffreys, H., 1931. Scientific Inference. Cambridge University Press.
Jeffreys, H., 1935. Some Tests of Significance, Treated by the Theory of Probability. In: Mathematical Proceedings of the Cambridge Philosophical Society, Vol. 31. Cambridge University Press, pp. 203–222.
Jeffreys, H., 1936. Further Significance Tests. In: Mathematical Proceedings of the Cambridge Philosophical Society, Vol. 32. Cambridge University Press, pp. 416–445.
Jeffreys, H., 1998. The Theory of Probability. OUP Oxford.
Kass, R.E., Raftery, A.E., 1995. Bayes factors. Journal of the American Statistical Association 90 (430), 773–795.
Lempers, F. B., 1971. Posterior Probabilities of Alternative Linear Models.
O'Hagan, A., 1995. Fractional Bayes factors for model comparison. Journal of the Royal Statistical Society: Series B (Methodological) 57 (1), 99–118.
Robert, C., 2007. The Bayesian choice: from decision–theoretic foundations to computational implementation. Springer Science & Business Media.
Schwartz, L., 1965. On Bayes procedures. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 4 (1), 10–26.
Shao, S., Jacob, P.E., Ding, J., Tarokh, V., 2019. Bayesian model comparison with the hyvärinen score: Computation and consistency. Journal of the American Statistical Association.
Taylor, S.J., 2008. Modelling Financial Time Series. world scientific.
Walker, S., Hjort, N.L., 2001. On Bayesian consistency. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63 (4), 811–821.
White, H., 1982. Maximum likelihood estimation of misspecified models. Econometrica 50 (1), 1–25.
Wuertz, D., Setz, T., Chalabi, Y., Maechler, M., Setz, M.T., 2017. Package fBasics. Rmetrics-Markets and Basic Statistics. R Foundation for Statistical Computing.