Kullback–Leibler divergence for Bayesian nonparametric model checking

Luai Al-Labadi, Vishakh Patel, Kasra Vakiloroayaei & Clement Wan

Journal of the Korean Statistical Society

ISSN 1226-3192

J. Korean Stat. Soc.
DOI 10.1007/s42952-020-00072-7

ONLINE FIRST

Springer

Journal of the Korean Statistical Society
https://doi.org/10.1007/s42952-020-00072-7

Online ISSN 2005-2863
Print ISSN 1226-3192

RESEARCH ARTICLE

Kullback–Leibler divergence for Bayesian nonparametric model checking

Luai Al-Labadi[1] · Vishakh Patel[1] · Kasra Vakiloroayaei[1] · Clement Wan[1]

Received: 8 July 2019 / Accepted: 1 May 2020

Abstract

Bayesian nonparametric statistics is an area of considerable research interest. While recently there has been an extensive concentration in developing Bayesian nonparametric procedures for model checking, the use of the Dirichlet process, in its simplest form, along with the Kullback–Leibler divergence is still an open problem. This is mainly attributed to the discreteness property of the Dirichlet process and that the Kullback–Leibler divergence between any discrete distribution and any continuous distribution is infinity. The approach proposed in this paper, which is based on incorporating the Dirichlet process, the Kullback–Leibler divergence and the relative belief ratio, is considered the first concrete solution to this issue. Applying the approach is simple and does not require obtaining a closed form of the relative belief ratio. A Monte Carlo study and real data examples show that the developed approach exhibits excellent performance.

Keywords Bayesian Non-parametric · Dirichlet process · Kullback–Leibler divergence · Model checking · Relative belief ratio

Mathematics Subject Classification 62F15 · 94A17 · 62F03

✉ Luai Al-Labadi
luai.allabadi@utoronto.ca

Vishakh Patel
vishakh.patel@mail.utoronto.ca

Kasra Vakiloroayaei
k.vakiloroayaei@mail.utoronto.ca

Clement Wan
cmclement.wan@mail.utoronto.ca

[1] Department of Mathematical and Computational Sciences, University of Toronto Mississauga, Mississauga, ON L5L 1C6, Canada

Published online: 04 June 2020

1 Introduction

Let $x = (x_1, \ldots, x_n)$ be a sample from a distribution F. The goal is to assess the hypothesis $\mathcal{H}_0 : F \in \{F_\theta : \theta \in \Theta\}$, where $\{F_\theta : \theta \in \Theta\}$ denotes the collection of continuous cumulative distribution functions (cdf's). This problem is known as *model checking* and it is quiet important in statistics. For instance, Jordan (2011) placed model checking and hypothesis testing as number one in a list of top-five open problems in Bayesian statistics.

Several Bayesian nonparametric procedures have been developed for model checking. A main approach considers embedding the proposed model as a null hypothesis in a larger family of distributions. Then priors are placed on the null and the alternative and a Bayes factor is computed. Using a Dirichlet process for the prior on the alternative can be found by Carota and Parmigiani (1996), and Florens et al. (1996). Verdinelli and Wasserman (1998), Berger and Guglielmi (2001) and McVinish et al. (2009) considered other types of priors on the alternative. Another important approach utilized for model testing is to place a prior on the true distribution that is generating the data and then measuring the distance between the posterior distribution and the hypothesized one. Swartz (1999) and Al-Labadi and Zarepour (2013a, b, 2014a, b) used the Dirichlet process as a prior and then considered the Kolmogorov distance in order to derive a goodness-of-fit test for continuous models. To test for discrete models, Viele (2007) used the Dirichlet process and the Kullback–Leibler (KL) divergence. For continuous model, Viele commented that his method "cannot be used for continuous data directly because the Dirichlet Process is discrete with probability 1. The KL information between any discrete distribution and any continuous distribution is infinity, and thus we must find a nonparametric method that produces continuous distributions. We employ a Dirichlet Process Mixture (DPM)." In fact working with the Dirichlet Process Mixture adds some complexity to the approach and makes it hard to implement by many users. Hsieh (2011) used the Pólya tree as a prior and the Kullback–Leibler divergence to test for continuous distributions. To judge whether a resulting divergence measure is large or small, he used normal approximations based on running a regression of the means and standard deviations. Al-Labadi and Evans (2018) established a new approach for model checking by utilizing the Dirichlet process and relative belief ratios. Then to measure the change from a priori to a posteriori they used Cramér-von Mises distance. See also Al-Labadi (2018), Al-Labadi et al. (2017), Al-Labadi et al. (2018) and Evans and Tomal (2018) for examples of using relative belief ratios in different hypothesis testing problems.

Although the KL divergence sits atop most distance/divergence measures (Viele 2007), it follows clearly from the previous discussion that its use alongside the Dirichlet process is very limited. In this paper, we propose address this issue. First, the Dirichlet process is considered as a prior on P (the true/sampling distribution). Then the concentration of the distribution of the KL divergence between the prior and the model of interest is compared to that between the posterior and the model. If the posterior is more concentrated about the model than the prior, then this is evidence in favor of the model and if the posterior is less concentrated, then this is evidence against the model. The comparison is made via a relative belief ratio about 0 (Evans 2015), which gives the evidence in the observed data for or against the model. Additionally, a measure of the strength of this evidence is also provided. So the methodology is based on a direct measure of statistical evidence. Implementing the approach is direct and does not require obtaining a closed form of the relative belief ratio. In addition, the methodology does not require the use of a prior on θ and so is truly a check on the model itself avoiding any issues with the prior on θ.

This paper is organized as follows. In Sects. 2 and 3, the relative belief ratio and the Dirichlet process are briefly reviewed, respectively. In Sect. 4, the Kullback–Leibler divergence between probability measures is discussed. In particular, a sample formula for computing the KL divergence between continuous distributions based on the Dirichlet process is developed. Section 5 discusses the proposed approach for model checking, where it is argued that a particular choice of the Kullback–Leibler divergence and the Dirichlet process should be employed. In Sect. 6, a computational algorithm for the implementation of the approach is outlined. Section 7 presents a number of examples where the behavior of the methodology is examined in some detail. Section 8 ends with a brief summary of the results.

2 Relative belief ratios

Consider $\{f_\theta : \theta \in \Theta\}$ to be a collection of densities on a sample space $\mathcal{X}$ and let π be a prior on Θ. Given the data x, the posterior distribution of θ is $\pi(\theta \,|\, x) = \pi(\theta)f_\theta(x)/\int_\Theta \pi(\theta)f_\theta(x)\,d\theta$. Let $\psi = \Psi(\theta)$ be the parameter of interest. Then the prior and posterior densities of ψ are denoted by π_Ψ and $\pi_\Psi(\cdot \,|\, x)$, respectively. The relative belief ratio (Evans 2015) for a value ψ is then defined as $RB_\Psi(\psi \,|\, x) = \lim_{\delta\to 0} \Pi_\Psi(N_\delta(\psi)|\, x)/\Pi_\Psi(N_\delta(\psi))$, where $N_\delta(\psi)$ is a sequence of neighbourhoods of ψ converging nicely (see, for example, Rudin 1974) to ψ as $\delta \to 0$. More commonly,

$$RB_\Psi(\psi \,|\, x) = \pi_\Psi(\psi \,|\, x)/\pi_\Psi(\psi), \tag{1}$$

is the ratio of the posterior density to the prior density at ψ. That is, $RB_\Psi(\psi \,|\, x)$ is measuring how beliefs have changed that ψ is the true value from *a priori* to *a posteriori*. Note that, a relative belief ratio is similar to a Bayes factor, as both are measures of evidence, but the latter measures this via the change in an odds ratio. A discussion about the relationship between relative belief ratios and Bayes factors is detailed in Baskurt and Evans (2013). In particular, when a Bayes factor is defined via a limit in the continuous case, the limiting value is the corresponding relative belief ratio.

By a basic principle of evidence, $RB_\Psi(\psi \,|\, x) > 1$ implies that the probability of ψ being correct increases after observing the data, and so there is evidence in favour of ψ. Else if $RB_\Psi(\psi \,|\, x) < 1$ then the data claims evidence of the ψ being incorrect and thus evidence of against ψ. Also if the $RB_{\Psi}(\psi \mid x) = 1$, then there is no evidence either way.

Therefore, the $RB_{\Psi}(\psi_0 \mid x)$ measures the evidence of the hypothesis $H_0 = \{\theta : \Psi(\theta) = \psi_0\}$. It is critical to rectify the degree of strength and weakness of this value. One of nicer calibration of the $RB_{\Psi}(\psi_0 \mid x)$ is suggested in Evans (2015), which considers the tail probability

$$\Pi_{\Psi}(RB_{\Psi}(\psi \mid x) \le RB_{\Psi}(\psi_0 \mid x) \mid x). \tag{2}$$

(2) can be interpreted as the posterior probability that the true value of ψ has a relative belief ratio no greater than that of the hypothesized value ψ_0. When $RB_{\Psi}(\psi_0 \mid x) < 1$, so there is evidence against ψ_0, then a small value for (2) indicates a large posterior probability that the true value has a relative belief ratio greater than $RB_{\Psi}(\psi_0 \mid x)$ and so there is strong evidence against ψ_0. When $RB_{\Psi}(\psi_0 \mid x) > 1$, so there is evidence in favor of ψ_0, then a large value for (2) indicates a small posterior probability that the true value has a relative belief ratio greater than $RB_{\Psi}(\psi_0 \mid x)$) and so there is strong evidence in favor of ψ_0, while a small value of (2) only indicates weak evidence in favor of ψ_0.

3 Dirichlet process

A relevant summary of the Dirichlet process is presented in this section. The Dirichlet process, formally introduced in Ferguson (1973), is considered the most well-known and widely used prior in Bayesian nonparametric inference. Specifically, consider $\mathfrak{X}$ a space with a $\sigma-$algebra $\mathcal{A}$ of subsets of $\mathfrak{X}$. Let G be a fixed probability measure on $(\mathfrak{X}, \mathcal{A})$, called the *base measure*, and a be a positive number, called the *concentration parameter*. Following Ferguson (1973), a random probability measure $P = \{P(A)\}_{A \in \mathcal{A}}$ is called a Dirichlet process on $(\mathfrak{X}, \mathcal{A})$ with parameters a and G, denoted by $DP(a, G)$, if for any finite measurable partition $\{A_1, \ldots, A_k\}$ of $\mathfrak{X}$ with $k \ge 2$, $\left(P(A_1), \ldots P(A_k)\right) \sim \text{Dirichlet}(aG(A_1), \ldots, aG(A_k))$. It is assumed that if $G(A_j) = 0$, then $P(A_j) = 0$ with a probability one. For any $A \in \mathcal{A}, P(A) \sim \text{Beta}(aG(A), (1 - G(A))$ and so $E(P(A)) = G(A)$ and $Var(P(A)) = G(A)(1 - G(A))/(1 + a)$. Thus, G plays the role of the center of the process, while a controls concentration, as, the larger value of a, the more likely that P will be close to G. Note that, for convenience, we do not distinguish between a probability measure and its cdf.

An important feature of the Dirichlet process is the conjugacy property. Specifically, if $x = (x_1, \ldots, x_n)$ is a sample from $P \sim DP(a, G)$, then the posterior distribution of P is $P \mid x = P_x \sim DP(a + n, G_x)$ where

$$G_x = a(a + n)^{-1}G + n(a + n)^{-1}F_n, \tag{3}$$

with $F_n = n^{-1}\sum_{i=1}^{n}\delta_{x_i}$ and δ_{x_i} the Dirac measure at x_i. Notice that G_x is a convex combination of the prior base distribution and the empirical distribution. Clearly, $G_x \to G$ as $a \to \infty$ while $G_x \to F_n$ as $a \to 0$. We refer the reader to Al-Labadi and

Zarepour (2013a, b, 2014a, 2017) for other interesting asymptotic properties of the Dirichlet process.

Following Ferguson (1973), $P \sim DP(a, G)$ has the following series representation

$$P = \sum_{i=1}^{\infty} J_i \delta_{Y_i}, \tag{4}$$

where $\Gamma_i = E_1 + \cdots + E_i, E_i \overset{i.i.d.}{\sim}$ exponential$(1), Y_i \overset{i.i.d.}{\sim} G$ independent of $\Gamma_i, L(x) = a\int_x^{\infty} t^{-1}e^{-t}dt, x > 0, L^{-1}(y) = \inf\{x > 0 : L(x) \geq y\}$ and $J_i = L^{-1}(\Gamma_i)/\sum_{i=1}^{\infty} L^{-1}(\Gamma_i)$. It follows clearly from (4) that a realization of the Dirichlet process is a discrete probability measure. This is correct even when G is absolutely continuous. Note that, one could resemble the discreteness of P with the discreteness of F_n. Since data is always measured to finite accuracy, the true distribution being sampled from is discrete. This makes the discreteness property of P with no practical significant limitation. Indeed, by imposing the weak topology, the support for the Dirichlet process is quite large. Precisely, the support for the Dirichlet process is the set of all probability measures whose support is contained in the support of the base measure. This means if the support of the base measure is $\mathfrak{X}$, then the space of all probability measures on $\mathfrak{X}$ is the support of the Dirichlet process. For instance, if G is the standard normal, then the Dirichlet process can choose any probability measure.

Recognizing that no closed form available for the inverse of Lévy measure $L(x)$, Sethuraman (1994) introduced the stick-breaking approach to define the Dirichlet Process. Specifically, let $(\beta_i)_{i\geq 1}$ be a sequence of i.i.d. random variables with a Beta$(1, \alpha)$ distribution. In (4), set

$$J_1 = \beta_1, \; J_i = \beta_i \prod_{k=1}^{i-1}(1-\beta_k), \quad i \geq 2. \tag{5}$$

and $(Y_i)_{i\geq 1}$ independent of $(\beta_i)_{i\geq 1}$. Unlike Ferguson's approach, the stick-breaking construction does not need normalization. By truncating the higher order terms in the sum to simulate Dirichlet process, we can approximate the Sethuraman stick breaking representation by

$$P_N = \sum_{i=1}^{N} J_{i,N} \delta_{Y_i}(\cdot). \tag{6}$$

In here, $(\beta_i)_{i\geq 1}, (J_{i,N})_{i\geq 1}$, and $(Y_i)_{i\geq 1}$ are as defined in (5) with $\beta_N = 1$. The assumption that $\beta_N = 1$ is necessary to make the weights add to 1, almost surely (Ishwaran and James 2001).

The Dirichlet process can also be obtained from the following finite mixture models developed by Ishwaran and Zarepour (2002). Let P_N has the from given in (4) with $(J_{1,N}, \dots, J_{N,N})$ ~Dirichlet$(a/N, \dots, a/N)$. Then $E_{P_N}(g) \to E_P(g)$ in distribution as $N \to \infty$, for any measurable function $g : \mathbb{R} \to \mathbb{R}$ with $\int_{\mathbb{R}} |g(x)|H(dx) < \infty$ and $P \sim DP(a, G)$. In particular, $(P_N)_{N\geq 1}$ converges in distribution to P, where P_N and P are random values in the space $M_1(\mathbb{R})$ of probability measures on $\mathbb{R}$ endowed with the topology of weak convergence. To generate $(J_{i,N})_{1\le i\le N}$ put $J_{i,N} = G_{i,N}/\sum_{i=1}^{N} G_{i,N}$, where $(G_{i,N})_{1\le i\le N}$ is a sequence of i.i.d. gamma$(a/N, 1)$ random variables independent of $(Y_i)_{1\le i\le N}$.

For other simulation methods for the Dirichlet process, see Bondesson (1982), Wolpert and Ickstadt (1998), Zarepour and Al-Labadi (2012), Al-Labadi and Zarepour (2014b).

4 Kullback–Leibler divergence

Let F and F_1 be two continuous cdf's with corresponding probability density functions (pdf's) f and f_1 (with respect to Lebesgue measure). Then Kullback–Leibler divergence or the Relative Entropy between F and F_1 is defined as

$$\begin{aligned} d_{KL}(F,F_1) &= \int_{-\infty}^{\infty} f(x)\log\big(f(x)/f_1(x)\big)dx \\ &= -H(F) - \int_{-\infty}^{\infty} f(x)\log f_1(x)dx, \end{aligned} \tag{7}$$

where

$$H(F) = -\int_{-\infty}^{\infty} f(x)\log f(x)dx = -E_F\left[\log f(x)\right] \tag{8}$$

is the *entropy* of F (Shannon 1948). It is well-know that $d_{KL}(F,F_1) \ge 0$ and the equality holds if and only if $f = f_1$. This property makes it suitable for model checking problems. Note that, the KL divergence is not a distance as it is not symmetric and does not satisfy the triangle inequality (Cover and Thomas 1991).

From practical viewpoint, one must estimate (7) from the data $x = (x_1, \ldots, x_n)$. In particular, estimating (8) is not a trivial task. Several frequentist estimators of (8) are offered in the literature. Vasicek (1976) noticed that (8) can be written as

$$H(F) = -\int_0^1 \log\left(\frac{d}{dt}F^{-1}(t)\right)dt.$$

If $x = (x_1, \ldots, x_n)$ is a sample from a distribution F, then, at each sample point x_i, the derivative of $F^{-1}(t)$ is estimated by the slope defined by

$$\frac{x_{(i+m)} - x_{(i-m)}}{F_n(x_{(i+m)}) - F_n(x_{(i-m)})} = \frac{x_{(i+m)} - x_{(i-m)}}{\frac{i+m}{n} - \frac{i-m}{n}} = \frac{x_{(i+m)} - x_{(i-m)}}{2m/n}, \tag{9}$$

where F_n is the empirical distribution function. Consequently, Vasicek (1976) estimator is given by

$$H^V_{m,n} = n^{-1}\sum_{i=1}^{n}\log\left(\frac{x_{(i+m)} - x_{(i-m)}}{2m/n}\right), \tag{10}$$

where m, called the window size, is a positive integer smaller than $n/2$ and $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$ are the order statistics of $x_1, x_2, \ldots, x_n$ with $x_{(i)} = x_{(1)}$ if $i < 1, x_{(i)} = x_{(n)}$ if $i > n$. Vasicek (1976) showed that $H^V_{m,n} \xrightarrow{p} H(F)$, where $\xrightarrow{p}$ denotes convergence in probability. Ebrahimi et al. (1994) noticed that (10) is not accurate when $i \le m$ or $i \ge n - m + 1$. They proposed the following revised version of (10):

$$H^{EPS}_{m,n} = n^{-1} \sum_{i=1}^{n} \log\left(\frac{x_{(i+m)} - x_{(i-m)}}{c_i m/n}\right), \tag{11}$$

where

$$c_i = \begin{cases} \frac{m+i-1}{m} & 1 \le i \le m \\ 2 & m+1 \le i \le N-m \\ \frac{N+m-i}{m} & N-m+1 \le i \le N \end{cases}. \tag{12}$$

They showed that $H^{EPS}_{m,N} \xrightarrow{p} H(F)$. For other nonparametric frequentist estimators of entropy consult, among others, the work of van Es (1992), Correa (1995), Wieczorkowski and Grzegorzewski (1999), Alizadeh Noughabi (2010), Alizadeh Noughabi and Arghami (2010), Bouzebda et al. (2013) and Al-Omari (2014), Al-Omari (2016). On the other hand, Bayesian estimation of entropy has received small consideration. Al-Labadi et al. (2018) derived the following Bayesian nonparametric estimator to entropy by using the Dirichlet process and adapting the estimators (10) and (11). Specifically, let $P_N = \sum_{i=1}^{N} J_{i,N}\delta_{Y_i}$ as defined in (6). Let m be a positive integer smaller than $N/2, Y_{(i)} = Y_{(1)}$ if $i < 1, Y_{(i)} = Y_{(N)}$ if $i > N, Y_{(1)} \le Y_{(2)} \le \cdots \le Y_{(N)}$ are the order statistics of $Y_1, Y_2, \ldots, Y_N$ and

$$H_{m,N,a} = \sum_{i=1}^{N} J_{i,N} \log\left(\frac{Y_{(i+m)} - Y_{(i-m)}}{c_{i,a}}\right), \tag{13}$$

where

$$c_{i,a} = \begin{cases} \sum_{k=2}^{i+m} J_{k,N} & 1 \le i \le m \\ \sum_{k=i-m+1}^{i+m} J_{k,N} & m+1 \le i \le N-m \\ \sum_{k=i-m+1}^{N} J_{k,N} & N-m+1 \le i \le N \end{cases}.$$

Then, as $N \to \infty, m \to \infty, m/N \to 0$ and $a \to \infty, H_{m,N,a} \xrightarrow{p} H(F)$.

5 Model checking using relative belief

Let $\{F_\theta : \theta \in \Theta\}$ denote the collection of continuous cdf's for the model. Suppose that $x = (x_1, \ldots, x_n)$ is a sample from a distribution F. The goal is to test the hypothesis $\mathcal{H}_0 : F \in \{F_\theta : \theta \in \Theta\}$. Let the prior on F be $DP(a, G)$ for some choice of a and G. Then, by (3), the posterior distribution is $F \,|\, x \sim DP\big(a + n, G_x\big)$.

As pointed out in the introduction, if $\mathcal{H}_0$ is true, then the posterior distribution of the divergence between F and $\{F_\theta : \theta \in \Theta\}$ should be more concentrated about 0 than the prior distribution of this divergence. So this test will involve a comparison of the concentrations of the prior and posterior distributions of d_{KL} about 0 via a relative belief ratio with the interpretation as discussed in Sect. 2. A simple formula for computing the KL divergence between continuous distributions F and G is offered in Lemma 2. First we introduce the following result which plays a central rule in the proof Lemma 2.

Proposition 1 *Let* $P_N = \sum_{i=1}^N J_{i,N}\delta_{Y_i}$ *as defined in* (6), *where* $Y_1, Y_2, \ldots, Y_N \overset{i.i.d.}{\sim} G$. *As* $N \to \infty$,

$$\sum_{i=1}^{N} J_{i,N} \log\left\{g(Y_i)\right\} \overset{a.s.}{\to} E\left[\log\left\{g(Y_i)\right\}\right] = \sum_{i=1}^{\infty} J_i \log\left(g(Y_i)\right),$$

where $\overset{a.s.}{\to}$ *denotes convergence almost surely and* J_i *is the weight that correspond to* Y_i.

Proof Note that, since $(J_{i,N})_{1\le i\le N}$ are not independent, the standard strong law of large numbers cannot be applied. Instead, as described in Sect. 2, $J_{i,N} = G_{i,N}/\sum_{i=1}^N G_{i,N}$, where $(G_{i,N})_{1\le i\le N}$ is a sequence of i.i.d. gamma(a/N, 1) random variables independent of $(Y_i)_{1\le i\le N}$.

$$\begin{aligned}\sum_{i=1}^{N} J_{i,N} \log\left\{g(Y_i)\right\} &= \sum_{i=1}^{N} \frac{G_{i,N}}{\sum_{i=1}^N G_{i,N},} \log\left\{g(Y_i)\right\}\\ &= \frac{\sum_{i=1}^N G_{i,N} \log\left\{g(Y_i)\right\}}{\sum_{i=1}^N G_{i,N},}\\ &= \frac{\frac{1}{N}\sum_{i=1}^N NG_{i,N} \log\left\{g(Y_i)\right\}}{\frac{1}{N}\sum_{i=1}^N NG_{i,N}}.\end{aligned}$$

Now, by the strong law of large numbers and the independence between $G_{i,N}$ and Y_i, we have

$$\frac{1}{N}\sum_{i=1}^{N} NG_{i,N} \log\left\{g(Y_i)\right\} \overset{a.s.}{\to} E\left[NG_{1,N}\right]E\left[\log\left\{g(Y_1)\right\}\right] = aE\left[\log\left\{g(Y_1)\right\}\right]$$

and

$$\frac{1}{N}\sum_{i=1}^{N} NG_{i,N}, \overset{a.s.}{\to} E\left[NG_{1,N}\right] = a.$$

It follows by the continuous mapping theorem that

$$\sum_{i=1}^{N} J_{i,N} \log\left\{g(Y_i)\right\} \xrightarrow{a.s} \frac{aE\left[\log\left\{g(Y_1)\right\}\right]}{a} = E\left[\log\left\{g(Y_1)\right\}\right] = \sum_{i=1}^{\infty} J_i \log\left(g(Y_i)\right).$$

□

An estimator of the KL divergence between F and G is given in the following lemma.

Lemma 2 *Let $H_{m,N,a}$ be as defined in (13). Then, as $N \to \infty, m \to \infty, m/N \to 0$ and $a \to \infty$,*

$$\begin{aligned}\widehat{d}_{KL}(F,G) &= -H_{m,N,a} - \sum_{i=1}^{N} J_{i,N} \log\left(g(Y_i)\right) \\ &= -\sum_{i=1}^{N} J_{i,N} \log\left\{ \frac{\left(Y_{(i+m)} - Y_{(i-m)}\right) g\left(Y_i\right)}{2mc_{i,a}} \right\} \\ &\xrightarrow{p} d_{KL}(F,G).\end{aligned}$$

Proof By Lemma 2 of Al-Labadi et al. (2018), (13)$= -H_{m,N,a} \xrightarrow{p} -H(F)$, where $H(F)$ is the entropy of F. The result follows by Proposition 1 and the continuous mapping theorem. □

Still, to use the estimator given in Lemma 2, it is necessary to discuss appropriate values for m, a and G.

5.1 Measuring the divergence

Similar to Al-Labadi and Evans (2018), we compute $d_{KL}(F, F_{\theta(x)})$, where $F_{\theta(x)} \in \left\{F_\theta : \theta \in \Theta\right\}$ is the distribution that is best supported by the data. Since the evidence is being measured via relative belief ratios, $\theta(x)$ is the relative belief estimate of θ, which for the full model parameter is always the same as the maximum likelihood estimate (MLE). As such, the value $\theta(x)$ is completely independent of any prior placed on θ. Certainly this choice has some asymptotic justification as, under reasonable conditions, $\theta(x)$ will converge to the best choice (in terms of Kullback–Leibler divergence) of θ even when the model fails.

5.2 The choice of *m*

The value of m is required to compute (13). However, the ideal value m is still an open problem. As discussed in Vasicek (1976), with increasing N, the best value of m increases while the ratio m/N tends to zero. Grzegorzewski and Wieczorkowski (1999) proposed the following formula for optimal values of m

$$m = \lfloor \sqrt{N} + 0.5 \rfloor, \tag{14}$$

where $\lfloor y \rfloor$ is the largest integer less than or equal to y. Thus, for instance, by (14), if $N = 50$, the best choices of m is 7. In this paper, we will use the rule (14). Note that, the value of m in (14) is the value that will be used for the prior. For the posterior, N will be replaced by the number of distinct atoms in $P_N \mid x$, an approximation of $F|x$. It follows from (3) that if a/n is close to zero, then the number of distinct atoms in $P_N \mid x$ will typically be n, the sample size.

5.3 The choice of *G*

Following Al-Labadi and Evans (2018), we set $G = F_{\theta(x)}$. That is, $F \sim DP\big(a, F_{\theta(x)}\big)$). There are many benefits of this choice of G. First, it avoids prior-data conflict (Evans and Moshonov 2006; Al-Labadi and Evans 2017) as the existence of prior-data conflict may lead to the failure having an appreciable concentration of the posterior distribution of $d_{KL}\big(F, F_{\theta(x)}\big)$ about zero, even when $\mathcal{H}_0$ is true (Al-Labadi and Evans 2018). On the other hand, setting $G = F_{\theta(x)}$ would appear to induce a data dependent prior distribution for d_{KL}. The following lemma implies that this is not the case and so, with this choice, the approach is prior distribution-free.

Lemma 3 *If $F \sim DP(a, F_{\theta(x)})$, then the distribution of $d_{KL}(F, F_{\theta(x)})$ does not depend on $F_{\theta(x)}$.*

Proof By Lemma 2,

$$\widehat{d}_{KL}(F, F_{\theta(x)}) = -\sum_{i=1}^{N} J_{i,N} \log\Bigg(\frac{1}{c_{i,a}} \frac{Y_{(i+m)} - Y_{(i-m)}}{F_{\theta(x)}\big(Y_{(i+m)}\big) - F_{\theta(x)}\big(Y_{(i-m)}\big)} \times \Big[F_{\theta(x)}\big(Y_{(i+m)}\big) - F_{\theta(x)}\big(Y_{(i-m)}\big)\Big] f_{\theta(x)}(Y_i)\Bigg)$$

Note that, as $m \to \infty$ such that $m/N \to 0$, we have

$$\frac{Y_{(i+m)} - Y_{(i-m)}}{F_{\theta(x)}\big(Y_{(i+m)}\big) - F_{\theta(x)}\big(Y_{(i-m)}\big)} = \frac{1}{f_{\theta(x)}(Y_i)}, \tag{15}$$

where $f_{\theta(x)}$ is the pdf of $F_{\theta(x)}$. Also, since $(Y_i)_{i\geq 1}$ is a sequence of i.i.d. random variables with continuous distribution $F_{\theta(x)}$, for $i \geq 1$, we have $U_i \overset{d}{=} F_{\theta(x)}(Y_i)$, where $\big(U_i\big)_{i\geq 1}$ is a sequence of i.i.d. random variables with a uniform distribution on [0, 1]. Thus,

$$\widehat{d}_{KL}(F, F_{\theta(x)}) \overset{d}{=} -\sum_{i=1}^{N} J_{i,N} \log\left(\frac{U_{(i+m)} - U_{(i-m)}}{c_{i,a}}\right). \tag{16}$$

Now, as $N \to \infty, m \to \infty, m/N \to 0$, by Lemma 2, we conclude that the distribution of $d_{KL}(F, F_{\theta(x)})$ does not depend on $F_{\theta(x)}$. □

Note that, similar to Noughabi and Arghami (2013), if $G(y) = y + b$, $G(y) = ay$ or $G(y) = ay + b$, which involve the case of location, scale and location-scale families, then (15) holds without any condition.

The following result shows that the posterior distribution of $d_{KL}\left(F, F_{\theta(x)}\right)$ becomes concentrated around 0 as sample size increases if and only if $\mathcal{H}_0$ is true. The proof follows straightforwardly from the properties of the KL divergence and (3). Thus it is omitted.

Lemma 4 *Let* $x = (x_1, \ldots, x_n) \sim F$, *where* $F \sim DP(a, F_{\theta(x)})$. *Suppose that* $\theta(x) \overset{a.s.}{\to} \theta_0$, $\sup_y |F_{\theta(x)}(y) - F_{\theta_0}(y)| \overset{a.s.}{\to} 0$ *as* $n \to \infty$.

(i) *If* $\mathcal{H}_0$ *is true, then, as* $n \to \infty$, $d_{KL}(F|x, F_{\theta(x)}) \overset{a.s.}{\to} 0$.
(ii) *If* $\mathcal{H}_0$ *is false, then, as* $n \to \infty$, $\liminf d_{KL}(F|x, F_{\theta(x)}) \overset{a.s.}{>} 0$.

5.4 The choice of *a*

The selection of a is very important. In principle, larger values of a must be chosen to detect smaller deviations. Therefore, it is possible to consider several values of a. For example, one may start with $a = 1$. If, as the value of a is increased, the corresponding relative belief ratio drops rapidly below 1, then this is a clear indication against $\mathcal{H}_0$. As will be seen in the examples, when the null hypothesis is correct, the relative belief ratio always remains above 1 when larger values of a are considered. On the other hand, if the relative belief ratio is below than 1 and, as the value of a is increased (i.e., using a more concentrated prior), the corresponding relative belief ratio increases above 1, then this is a good indication in favour of $\mathcal{H}_0$. It is highly recommended to choose $a \le 0.5n$, however, otherwise the prior may become too influential. See Al-Labadi and Zarepour (2017) for the justification of this recommendation. It is noticed that, for most purposes, setting a between 1 and 10 is found satisfactory. This choice of a is also recommended by Holmes et al. (2015) when using the Pólya tree prior for the two-sample problem. This issue is further discussed in Table 1 of Sect. 7.

The following result is useful in the elicitation process of a.

Lemma 5 *If* $F \sim DP(a, F_{\theta(x)})$, *then*

$$\begin{aligned} E\left[\widehat{d}_{KL}(F, F_{\theta(x)})\right] =& \frac{2}{N}\sum_{i=1}^{m}\left[\psi\left(\frac{a(m+i-1)}{N}+1\right) - \psi(m+i-1)\right] \\ &+ \frac{N-2m}{N}\left(\psi\left(\frac{2am}{N}+1\right) - \psi(2m)\right) \\ &+ \psi(N+1) - \psi(a+1), \end{aligned} \tag{17}$$

where $\psi(x) = \Gamma'(x)/\Gamma(x)$ *is the digamma function.*

Table 1 Relative belief ratios and strengths for testing the location normal model with various alternatives and choices of a in Example 1

F_{true}	a	$d_{0.05}(pr)$	RB	Strength
$N(0, 1)$	1	0.5689	20	1
	5	0.1441	13.2041	0.3395
	10	0.0546	4.4124	0.7780
$N(10, 1)$	1	0.5690	20	1
	5	0.1440684	13.2041	0.3394
	10	0.0546	4.4124	0.7779
$N(0, 4)$	1	0.5512	0.8221	0.1355
	5	0.1283	0.03988	0.0015
	10	0.0494	0.0300	0.0000
$0.5N(-2, 1) + 0.5N(2, 1)$	1	0.5440	0.3878	0
	5	0.1286	0	0
	10	0.0548	0	0
$t_{0.5}$	1	0.570	0	0
	5	0.1331	0	0
	10	0.0569	0	0
t_3	1	0.5485	4.6007	0.9980
	5	0.1284	0.8618	0.3577
	10	0.0511	0.8819	0.5147

Proof By (16) and independence,

$$E\left[\widehat{d}_{KL}(F, F_{\theta(x)})\right] = -\sum_{i=1}^{N} E\left[J_{i,N}\right] E\left[\log\left(U_{(i+m)} - U_{(i-m)}\right)\right] + \sum_{i=1}^{N} E\left[J_{i,N} \log c_{i,a}\right]. \tag{18}$$

Since $U(i) = U(1)$ for $i < 1$ and $U(i) = U(N)$ for $i > N$ and using the well-known fact that $U_{(s)} - U_{(r)} \sim Beta(s - r, N - s + r + 1)$ we have:

$$\begin{aligned}
&\sum_{i=1}^{N} E\left[\log(y_{(i+m)} - y_{(i-m)})\right] \\
&\quad = \sum_{i=1}^{m} E\left[\log(U_{(i+m)} - U_{(i-m)})\right] \\
&\qquad + \sum_{i=m+1}^{N-m} E\left[\log(U_{(i+m)} - U_{(1)})\right] + \sum_{N-m+1}^{N} E\left[\log(U_{(N)} - U_{(i-m)})\right] \\
&\quad = (N - 2m)(\psi(2m) - \psi(N+1)) + 2\sum_{i=1}^{m} (\psi(i + m - 1) - \psi(N + 1)) \\
&\quad = (N - 2m)\psi(2m) - N\psi(N + 1) + 2\sum_{i=1}^{m} \psi(i + m - 1).
\end{aligned} \tag{19}$$

On the other hand,

$$\sum_{i=1}^{N} E\left[J_{i,N}\log c_{i,a}\right] = \sum_{i=1}^{m} E\left[J_{i,N}\log\left(\sum_{k=2}^{i+m} J_{k,N}\right)\right] \tag{20}$$

$$+\sum_{i=m+1}^{N-m} E\left[J_{i,N}\log\left(\sum_{k=i-m+1}^{i+m} J_{k,N}\right)\right] \tag{21}$$

$$+\sum_{i=N-m+1}^{N} E\left[J_{i,N}\log\left(\sum_{k=i-m+1}^{N} J_{k,N}\right)\right]. \tag{22}$$

From the proof of Lemma 1 of Al-Labadi et al. (2018) and using the facts that $\psi(x+1) = \psi(x) + 1/x$ (Abramowitz and Stegun 1972), we have

$$\begin{aligned}(20) &= \frac{1}{N}\sum_{i=1}^{m}\psi\left(\frac{a(m+i-1)}{N}+1\right) - \frac{m}{N}\psi(a+1),\\ (21) &= \frac{1}{N}\sum_{i=m+1}^{N-m}\psi\left(\frac{2am}{N}+1\right) - \frac{N-2m}{N}\psi(a+1)\\ &= \frac{N-2m}{N}\psi\left(\frac{2am}{N}+1\right) - \frac{N-2m}{N}\psi(a+1)\end{aligned}$$

and

$$\begin{aligned}(22) &= \frac{1}{N}\sum_{i=N-m+1}^{N}\psi\left(\frac{a(N+m-i)}{N}+1\right) - \frac{m}{N}\psi(a+1)\\ &= \frac{1}{N}\sum_{i=1}^{m}\psi\left(\frac{a(m+i-1)}{N}+1\right) - \frac{m}{N}\psi(a+1).\end{aligned}$$

Substitute (19), (20), (21) and (22) in (18), we get the result. □

6 Computational algorithm

To use (1), closed forms of the prior and posterior densities of $D = d_{KL}(F, F_{\theta(x)})$ are required, which is typically not available. Consequently, the relative belief ratio needs to be approximated via simulation. A particular attention here should be given to the case when both $\pi_D(0 \mid x)$ and $\pi_D(0)$ are close to 0. In such a case, determining

$RB_D(0\,|\,x)$ is challenging. However, as discussed in Sect. 2, the formal definition of $RB_D(0\,|\,x)$ is given as a limit and this limit can be approximated by $RB_D([0, d_*)\,|\,x)$, the ratio of the posterior to prior probability that $0 \le D \le d_*$, for a suitably small value of d_*.

We adapt the procedure outlined in Al-Labadi and Evans (2018). This approach is based on M quantiles of the prior distribution of D, namely, the i-th interval is $[d_{i/M}(pr), d_{(i+1)/M}(pr))$ where $d_{i/M}(pr)$ is the (i/M)-th quantile for $i = 0, \dots, M$. Note that values in the left tail of this distribution correspond to those F, in the population of distributions, that, according to the prior at least, do not differ materially from 0. As such we will consider the left-tail quantile of this prior distribution, such as the 0.05-quantile or the 0.01-quantile, so $d_* = d_{i_0/M}(pr)$ where $i_0/M \approx 0.05$ or $i_0/M \approx 0.01$.

The following gives a computational algorithm for assessing $\mathcal{H}_0$.

Algorithm A *(Relative belief algorithm for model checking):*

(i) Generate a sample from P_N, where P_N is an approximation of $F \sim DP(a, F_{\theta(x)})$. See Sect. 3.

(ii) Compute $d(pr) = \widehat{d}_{KL}(P_N, F_{\theta(x)})$ as described in Lemma 2.

(iii) Repeat steps (i) and (ii) to obtain a sample of r_1 values from the prior of D.

(iv) Generate a sample from $P_N|x$, where $P_N|x$ an approximation of $F|x \sim DP(a+n, G_x)$.

(v) Compute $d(po) = \widehat{d}_{KL}(P_N|x, F_{\theta(x)})$ as described in Lemma 2.

(vi) Repeat steps (iv)-(v) to obtain a sample of r_2 values from the posterior of D.

(vii) For a fixed positive number M, let $\widehat{F}_D$ denote the empirical cdf of D based on the prior sample in (iii) and for $i = 0, \dots, M$, let $\widehat{d}_{i/M}(pr)$ be the estimate of $d_{i/M}(pr)$, the (i/M)-th prior quantile of D. Here $\widehat{d}_0(pr) = 0$, and $\widehat{d}_1(pr)$ is the largest value of $d(pr)$. Let $\widehat{F}_D(\cdot\,|\,x)$ denote the empirical cdf of D based on the posterior sample of $d(po)$ in (vi). For $d \in [\widehat{d}_{i/M}(pr), \widehat{d}_{(i+1)/M}(pr))$, estimate $RB_D(d\,|\,x)$ by the ratio of the estimates of the posterior and prior contents of $[\widehat{d}_{i/M}(pr), \widehat{d}_{(i+1)/M}(pr))$. Specifically,

$$\widehat{RB}_D(d\,|\,x) = M\{\widehat{F}_D(\widehat{d}_{(i+1)/M}(pr)\,|\,x) - \widehat{F}_D(\widehat{d}_{i/M}(pr)\,|\,x)\}, \tag{23}$$

Moreover, estimate $RB_D(0\,|\,x)$ by $\widehat{RB}_D(0\,|\,x) = M\widehat{F}_D(\widehat{d}_{i_0/M}(pr)\,|\,x)$ where i_0 is chosen so that i_0/M is not too small (typically $i_0/M \approx 0.05$).

(viii) Estimate the strength $DP_D(RB_D(d\,|\,x) \le RB_D(0\,|\,x)\,|\,x)$ by the finite sum

$$\sum_{\{i \ge i_0 : \widehat{RB}_D(\widehat{d}_{i/M}(pr)\,|\,x) \le \widehat{RB}_D(0\,|\,x)\}} (\widehat{F}_D(\widehat{d}_{(i+1)/M}(pr)\,|\,x) - \widehat{F}_D(\widehat{d}_{i/M}(pr)\,|\,x)). \tag{24}$$

For fixed M, as $r_1 \to \infty, r_2 \to \infty$, then $\widehat{d}_{i/M}(pr)$ converges almost surely to $d_{i/M}(pr)$, (23) converge almost surely to $RB_D(d\,|\,x)$ and (24) converge almost surely to $DP_D(RB_D(d\,|\,x) \le RB_D(0\,|\,x)\,|\,x)$ (Al-Labadi and Evans 2018).

7 Examples

In this section, the approach is illustrated through three examples. In all the examples, the prior was taken to be $DP\big(a, F_{\theta(x)}\big)$ and, in Algorithm A, we set $r_1 = r_2 = 2000, N = 200, M = 20$ and $i_0 = 1$. A critical factor here for success are the choices of a as the prior has to be sufficiently concentrated about the family. The sensitivity to the choice of a is investigated and we record only a few values in the tables.

Example 1 *Location normal model.*

In this example, samples of $n = 20$ was generated from the distribution F_{true} in Table 1. Then the methodology was applied to assess whether or not the correct model is $\{F_\theta : \theta \in \Theta\} = \{N(\theta, 1) : \theta \in \mathbb{R}\}$ and so $\theta(x) = \bar{x}$. Thus, by Lemma 2,

$$\widehat{d}_{KL}(F, F_{\theta(x)}) = -H_{m,N,a} - \sum_{i=1}^{N} J_{i,N} \log\left(f_{\theta(x)}(Y_i)\right),$$

where

$$f_{\theta(x)}(Y_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(Y_i - \bar{x})^2}.$$

It follow that

$$\widehat{d}_{KL}(F, F_{\theta(x)}) = -H_{m,N,a} + \frac{1}{2}\log(2\pi) + \frac{1}{2}\sum_{i=1}^{N} J_{i,N}\left(Y_i - \bar{x}\right)^2.$$

In Table 1 the relative belief ratios and the strengths are recorded for testing the location normal model against a variety of alternatives with two choices of the hyperparameter a and m. Recalling that we want $RB > 1$ and the strength close to 1 when $\mathcal{H}_0$ is true and $RB < 1$ and the strength close to 0 when $\mathcal{H}_0$ is false, it is seen that the methodology using $d_{KL}(F, F_{\theta(x)})$ performs well in every instance.

Example 2 *The Gumbel Model.*

In this example, we consider the Gumbel model. This model is commonly used in environmental sciences, hydrology in the modeling of heavy rain, floods and industrial applications. A random variable Y is said to have the Gumbel distribution if its probability density function has the form

$$f(y; \xi, \beta) = \frac{1}{\beta} \exp\left\{ -\frac{y - \xi}{\beta} - \exp\left(-\frac{y - \xi}{\beta} \right) \right\}, \quad y, \xi \in \mathbb{R}, \quad \beta > 0.$$

Here ξ represents the location parameter and β represents the scale parameter. The following dataset gives the annual maxima of daily rainfall (in mm) during the period 1967–2001 recorded at the Álamo, Veracruz, meteorological station, México

Table 2 Relative belief ratios and strengths for testing the Gumbel model in Example 2

a	1	5	10	15	20
$d_{0.05}(pr)$	0.5573	0.1209	0.0499	0.0306	0.0215
RB	20	13.2132	5.7211	3.7904	3.0154
Strength	1	1	1	1	1

Table 3 Relative belief ratios and strengths for testing the normality of the Kevlar data and various choices of a in Example 3

a	1	5	10	15	20	25	30
$d_{0.05}(pr)$	0.5451	0.1194	0.0501	0.0316	0.0240	0.0146	0.0130
RB	17.65	1.8107	0.6847	0.4782	0.3915	0.1784	0.1697
Strength	1	0.7009	0.0342	0.7009	0.1037	0.0355	0.0089

maximum flood levels of the Susquehanna River at Harrisburg, Pennsylvania, over four-year periods (1890–1969) in millions of cubic feet per second.

$$86.8, 78.5, 93.1, 95.5, 78.1, 89.9, 109.5, 161.6, 187.6, 89.9, 73.4, 78.1, 73.3, 130.1, 188.3, 113.9, 42.5, 80.0, 142.6, 42.9, 60.2, 100.0, 129.0, 98.0, 116.4, 37.9, 60.7, 48.7, 39.7, 80.3, 30.7, 120.0, 160.0, 64.3, 80.0$$

According to Pérez-Rodríguez et al. (2009), the maximum likelihood estimators for ξ and β are 74.5432 and 32.4328, respectively. The goal is to test whether the underlying distribution is a Gumbel distribution. The results in Table 2 indicate that indeed the data can be considered as coming from a Gumbel distribution as there is evidence in favor of this model. See also Sect. 5 of Pérez-Rodríguez et al. (2009, Sect. 5) for similar results.

Example 3 *Lifetimes of Kevlar pressure vessels.*

Consider the data of 100 stress-rupture lifetimes of Kevlar pressure vessels presented in Andrews and Herzberg (1985). The goal is to test whether the underlying distribution is normal. That is, $\{F_\theta : \theta \in \Theta\} = \{N(\mu, \sigma^2) : \theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)\}$ and so $\theta(x) = (\bar{x}, \sum_{i=1}^{n}(x - \bar{x})^2/n)$. For this data set, $\theta(x) = (209.171, 37606.56)$. Previous studies such as Evans and Swartz (1994), Verdinelli and Wasserman (1998) and Al-Labadi and Evans (2018, Table 4) suggested that model is not correct. The results in Table 3 support the non-normality of this data set only when using a more concentrated prior.

8 Conclusions

A general procedure for model checking based on integrating the Dirichlet process, the Kullback–Leibler divergence and the relative belief ratio has been considered. Applying the approach is simple and does not require obtaining a closed form of the relative belief ratio. Numerous examples are presented in which the proposed approach shows excellent performance.

References

Abramowitz, M., & Stegun, I. A. (1972). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. New York: Dover.

Alizadeh Noughabi, H. (2010). A new estimator of entropy and its application in testing normality. *Journal of Statistical Computation and Simulation*, *80*, 1151–1162.

Alizadeh Noughabi, H., & Arghami, N. R. (2010). A new estimator of entropy. *Journal of the Iranian Statistical Society*, *9*, 53–64.

Al-Labadi, L. (2018). The two-sample problem via relative belief ratio. https://arxiv.org/abs/1805.07238.

Al-Labadi, L., & Abdelrazeq, I. (2017). On functional central limit theorems of Bayesian nonparametric priors. *Statistical Methods and Applications*, *26*, 215–229.

Al-Labadi, L., & Evans, M. (2017). Optimal robustness results for relative belief inferences and the relationship to prior-data conflict. *Bayesian Analysis*, *12*, 705–728.

Al-Labadi, L., & Evans, M. (2018). Prior-based model checking. *Canadian Journal of Statistics*, *46*, 380–398.

Al-Labadi, L., Patel, V., Vakiloroayaei, K., & Wan, C.(2018). A Bayesian nonparametric estimation to entropy. https://arxiv.org/abs/1903.00655.

Al-Labadi, L., & Zarepour, M. (2013a). A Bayesian nonparametric goodness of fit test for right censored data based on approximate samples from the beta-Stacy process. *Canadian Journal of Statistics*, *41*, 466–487.

Al-Labadi, L., & Zarepour, M. (2013b). On asymptotic properties and almost sure approximation of the normalized inverse-Gaussian process. *Bayesian Analysis*, *8*, 553–568.

Al-Labadi, L., & Zarepour, M. (2014a). Goodness of fit tests based on the distance between the Dirichlet process and its base measure. *Journal of Nonparametric Statistics*, *26*, 341–357.

Al-Labadi, L., & Zarepour, M. (2014b). On simulations from the two-parameter Poisson–Dirichlet process and the normalized inverse-Gaussian process. *Sankhyā A*, *76*, 158–176.

Al-Labadi, L., & Zarepour, M. (2017). Two-sample Kolmogorov–Smirnov test using a Bayesian nonparametric approach. *Mathematical Methods of Statistics*, *26*, 212–225.

Al-Labadi, L., Zeynep, B., & Evans, M. (2017). Goodness of fit for the logistic regression model using relative belief. *Journal of Statistical Distributions and Applications*,. https://doi.org/10.1186/s40488-017-0070-7.

Al-Labadi, L., Zeynep, B., & Evans, M. (2018). Statistical reasoning: Choosing and checking the ingredients, inferences based on a measure of statistical evidence with some applications. *Entropy*, *20*, 289. https://doi.org/10.3390/e20040289.

Al-Omari, A. I. (2014). Estimation of entropy using random sampling. *Journal of Computation and Applied Mathematics*, *261*, 95–102.

Al-Omari, A. I. (2016). A new measure of entropy of continuous random variable. *Journal of Statistical Theory and Practice*, *10*, 721–735.

Andrews, D. F., & Herzberg, A. M. (1985). *Data–A collection of problems from many fields for the student and research worker*. Berlin: Springer.

Baskurt, Z., & Evans, M. (2013). Hypothesis assessment and inequalities for Bayes factors and relative belief ratios. *Bayesian Analysis*, *8*, 569–590.

Berger, J. O., & Guglielmi, A. (2001). Bayesian testing of a parametric model versus nonparametric alternatives. *Journal of the American Statistical Association*, *96*, 174–184.

Bondesson, L. (1982). On simulation from infinitely divisible distributions. *Advances in Applied Probability*, *14*, 885–869.

Bouzebda, S., Elhattab, I., Keziou, A., & Lounis, T. (2013). New entropy estimator with an application to test of normality. *Communications in Statistics - Theory and Methods*, *42*, 2245–2270.

Carota, C., & Parmigiani, G. (1996). On Bayes factors for nonparametric alternatives. In J. M. Bernardo, J. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian Statistics 5*. London: Oxford University Press.

Correa, J. C. (1995). A new estimator of entropy. *Communications in Statistics—Theory and Methods*, *24*, 2439–2449.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory* (2nd ed.). New York: Wiley.

Ebrahimi, N., Pflughoeft, K., & Soofi, E. (1994). Two measures of sample entropy. *Statistics and Probability Letters*, *20*, 225–234.

Evans, M. (2015). *Measuring Statistical Evidence Using Relative Belief, Monographs on Statistics and Applied Probability* (Vol. 144). Boca Raton: CRC Press, Taylor and Francis Group.

Evans, M., & Moshonov, H. (2006). Checking for prior-data conflict. *Bayesian Analysis*, *1*, 893–914.

Evans, M., & Swartz, T. (1994). Distribution theory and inference for polynomial-normal densities. *Communications in Statistics—Theory and Methods*, *23*, 1123–1148.

Evans, M., & Tomal, J. (2018). Measuring statistical evidence and multiple testing. *FACET*, *3*, 563–583.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, *1*, 209–230.

Florens, J. P., Richard, J. F., & Rolin, J. M. (1996). Bayesian encompassing specification tests of a parametric model against a nonparametric alternative. Technical Report 9608, Universitsé Catholique de Louvain, Institut de statistique.

Grzegorzewski, P., & Wieczorkowski, R. (1999). Entropy-based goodness-of-fittest for exponentiality. *Communications in Statistics—Theory and Methods*, *28*, 1183–1202.

Holmes, C. C., Caron, F., Griffin, J. E., & Stephens, D. A. (2015). Two-sample Bayesian nonparametric hypothesis testing. *Bayesian Analysis*, *2*, 297–320.

Hsieh, P. (2011). A nonparametric assessment of model adequacy based on Kullback–Leibler divergence. *Statistics and Computing*, *23*, 149–162.

Ishwaran, H., & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, *96*, 161–173.

Ishwaran, H., & Zarepour, M. (2002). Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, *30*, 269–283.

Jordan, M. I. (2011). What are the open problems in Bayesian statistics? *ISBA Bulletin*, *18*, 1–4.

McVinish, R., Rousseau, J., & Mengersen, K. (2009). Bayesian goodness of fit testing with mixtures of triangular distributions. *Scandivavian Journal of Statistics*, *36*, 337–354.

Noughabi, H. A., & Arghami, N. R. (2013). General treatment of goodness-of-fit tests based on Kullback–Leibler information. *Journal of Statistical Computation and Simulation*, *83*, 1556–1569.

Pérez-Rodríguez, P., Vaquera-Huerta, H., & Villaseñor-Alva, J. A. (2009). A goodness-of-fit test for the Gumbel distribution based on Kullback–Leibler information. *Communications in Statistics: Theory and Methods*, *38*, 842–855.

Rudin, W. (1974). *Real and Complex Analysis* (2nd ed.). New York: McGrawHill.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, *4*, 639–650.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(379–423), 623–656.

Swartz, T. B. (1999). Nonparametric goodness-of-fit. *Communications in Statistics: Theory and Methods*, *28*, 2821–2841.

van Es, B. (1992). Estimating functionals related to a density by a class of statistics based on spacings. *Scandinavian Journal of Statistics*, *19*, 61–72.

Vasicek, O. (1976). A test for normality based on sample entropy. *Journal of Royal Statistical Society B*, *38*, 54–59.

Verdinelli, I., & Wasserman, L. (1998). Bayesian goodness-of-fit testing using finite-dimensional exponential families. *Annals of Statistics*, *26*, 1215–1241.

Viele, K. (2007). Nonparametric estimation of Kullback–Leibler information illustrated by evaluating goodness of fit. *Bayesian Analysis*, *2*, 239–280.

Wieczorkowski, R., & Grzegorzewski, P. (1999). Entropy estimators-improvements and comparisons. *Communications in Statistics—Simulation and Computation*, *28*, 541–567.

Wolpert, R. L., & Ickstadt, K. (1998). Simulation of Lévy random fields. In D. Day, P. Muller, & D. Sinha (Eds.), *Practical nonparametric and semiparametric Bayesian statistics* (pp. 227–242). Berlin: Springer.

Zarepour, M., & Al-Labadi, L. (2012). On a rapid simulation of the Dirichlet process. *Statistics and Probability Letters*, *82*, 916–924.