Stat Comput (2013) 23:149–162
DOI 10.1007/s11222-011-9298-0

A nonparametric assessment of model adequacy based on Kullback-Leibler divergence

Ping-Hung Hsieh

Received: 15 October 2010 / Accepted: 24 October 2011 / Published online: 19 November 2011

Abstract A discrepancy measure to assess model fitness against a nonparametric alternative is proposed. First, a Polya tree prior is constructed so that the centering distribution is the null. Second, the prior is updated in the light of data to obtain the posterior centering distribution as the alternative. Third, a Kullback-Leibler divergence type of test statistic is derived to assess the discrepancy between the two centering distributions. The properties of the test statistic are derived, and a power comparison with several well-known test statistics is conducted. The use of the test statistic is illustrated using network traffic data.

Keywords Goodness of fit · Nonparametric alternative · Packet train · Polya tree · Teletraffic data

1 Introduction

Many empirical studies in network traffic modeling (e.g., Leland et al. 1994; Willinger et al. 1997, 1998; Hsieh 2002) have demonstrated that the classical queuing assumption of normal tails is inappropriate for network data and a long-tailed distribution is, in general, more suitable for such data. However, due to the design of a network, the underlying distribution of traffic between a source (e.g., a server) and a particular destination (e.g., a classroom computer) may or may not significantly deviate from normality, and thus, the well developed classical queuing theory and widely available tools may still be applicable. It is the main goal of this paper to develop a discrepancy measure that quantifies the adequacy of an assumed model.

Assessing the adequacy of a posited model or of assumptions in general is fundamental to statistical analyses. The posited model must be evaluated routinely in the light of empirical evidence. In many such evaluations, the experimenter may be concerned less with whether the posited model is correct and more with whether the model remains adequate (defined as closeness to the null posited model) and, if not, how significant the departure from the posited model is. Thus, the objective of this article is to develop a discrepancy measure, $\rho(F_0, F_a)$, to assess model adequacy when an alternative distribution is not specified, where the distribution $F_0(x|\theta)$ is the posited model for a random variable X with a vector of real-valued parameters θ, and F_a is a nonparametric alternative updated from F_0 via data. In other words, the discrepancy measure ρ suggests the "distance" between F_0 and F_a; thus, a small value of ρ quantifies the notion of model adequacy whereas a large value of ρ indicates the inadequacy of the current posited model F_0 and the nonparametric alternative F_a can serve as the new F_0. Similar ideas can be found in, for example, Mengerson and Robert (1996), Verdinelli and Wasserman (1998), and Goutis and Robert (1998).

Two issues in this approach must be discussed: (1) the construction of an alternative distribution F_a and (2) the choice of a discrepancy measure. To address the first issue, we consider the approach provided by Berger and Guglielmi (2001) in which the nonparametric alternative F_a is embedded in a mixture of Polya tree distributions. The choice of a nonparametric alternative over a parametric one is certainly subjective. But, if possible, it is always desirable in an analysis to make fewer assumptions about the underlying distribution. Between the two popular Bayesian nonparametric priors, the Polya trees and the Dirichlet processes, our choice of

P.-H. Hsieh (✉)
College of Business, Oregon State University, Corvallis, OR 97331-2603, USA
e-mail: hsiehph@bus.orst.edu

the former is based mainly on the assumption that the possible alternative distribution is continuous. In other words, we would like the alternative F_a to be a continuous distribution even if the normal distribution (F_0) assumed by the classical queuing theory is rejected. Lavine (1992) and Mauldin et al. (1992) showed that the Polya trees, when properly constructed, assign a probability of one to the set of continuous distributions. In contrast (e.g., Viele 2000), the Dirichlet processes expect duplicate observations and therefore often lead to unappealing results when the testing of an absolutely continuous model is of primarily concern (see Carota and Parmigiani 1996). Furthermore, in the course of this study, we confirmed various attractive properties, such as computational and robustness advantages, inherent in the use of Polya tree processes, as have been documented by Lavine (1992, 1994) and Berger and Guglielmi (2001).

To address the second issue, we propose a test statistic derived from the Kullback-Leibler divergence to discriminate between F_0 and F_a. Although various discrepancy measures are available, for example, the Kolmogorov-Smirnov D, Cramér-von Mises W^2, Watson U^2, and Anderson-Darling A^2 statistics, we employ a Kullback-Leibler based discrepancy measure because its sampling distribution can be reasonably approximated by a normal distribution, even for sample size as small as 20. The property of normal approximation is advantageous because we eventually will need to decide whether a resulting divergence measure is large or small. Elaborative simulations to create *ad hoc* tables of critical values at different levels of significance or sample sizes can thus be avoided.

The main idea of this article parallels Viele's (2000), in which a mixture of Dirichlet processes is used. In addition to Berger and Guglielmi (2001), the model embedding approach to test a parametric model against a nonparametric alternative has been considered by, for example, Carota and Parmigiani (1996), who applied a mixture of Dirichlet processes, and Verdinelli and Wasserman (1998), who utilized a mixture of Gaussian processes. The use of a divergence measure, particularly entropy-based measures such as the Kullback-Leibler divergence, in the hypothesis testing framework has also been studied. For example, motivated by Vasicek's (1976) paper on sample entropy estimation, Arizono and Ohta (1989), Dudewicz and van der Meulen (1981), and Ebrahimi et al. (1992) derived entropy-based test statistics specifically for tests of normality, uniformity, and exponentiality, respectively. However, critical values for these test statistics often depend on, say, the sample size, the alternative distribution, and some auxiliary parameters and can be tabulated only through Monte Carlo simulations.

The organization of this paper is as follows: In Sect. 2, we define Polya trees and mixtures of Polya trees and summarize their relevant properties. In Sect. 3, the proposed test statistic is introduced, and its theoretical properties are derived. A simulation study to explore the strengths and weaknesses of the proposed test statistic appears in Sect. 4. The application of the proposed statistic is illustrated using network data collected from a student laboratory in Sect. 5, followed by concluding remarks in Sect. 6.

2 Polya tree distributions

2.1 Definitions

We follow the notation used by Lavine (1992) to define a mixture of Polya tree distributions. Let $E = \{0, 1\}$, $E^0 = \emptyset$, E^m be the m-fold product $E \times \cdots \times E$, and $E^* = \bigcup_0^\infty E^m$. Let $\Pi = \{\Pi_m, m = 0, 1, \ldots\}$ be a separating binary tree of partitions of real line $\Pi_0 = \Re$. That is, for the sequence of partitions $\Pi_0, \Pi_1, \ldots$ of $\Re$, $\bigcup_0^\infty \Pi_m$ generates the Borel sets, and every $B \in \Pi_{m+1}$ is obtained by splitting some $B' \in \Pi_m$ into two pieces. For all $\varepsilon = \varepsilon_1 \cdots \varepsilon_m \in E^*$, let $B_{\varepsilon 0}$ and $B_{\varepsilon 1}$ be the two pieces into which B_ε is split. Hence, a random probability measure $\mathcal{P}$ on $\Re$ is said to have a Polya tree distribution, or a Polya tree prior, with parameters Π and $\mathcal{A}$, denoted by $\mathcal{P} \sim PT(\Pi, \mathcal{A})$, if there exist nonnegative numbers $\mathcal{A} = \{\alpha_\varepsilon : \varepsilon \in E^*\}$ and random variables $\mathcal{Y} = \{Y_\varepsilon : \varepsilon \in E^*\}$ such that the following conditions hold:

- All random variables in $\mathcal{Y}$ are independent;
- For every $\varepsilon \in E^*$, Y_ε has a beta distribution with parameters $\alpha_{\varepsilon 0}$ and $\alpha_{\varepsilon 1}$; and
- For every $m = 1, 2, \ldots$ and every $\varepsilon \in E^*$,

$$\mathcal{P}(B_{\varepsilon_1 \cdots \varepsilon_m}) = \prod_{j=1, \varepsilon_j = 0}^{m} Y_{\varepsilon_1 \cdots \varepsilon_{j-1}} \times \prod_{j=1, \varepsilon_j = 1}^{m} (1 - Y_{\varepsilon_1 \cdots \varepsilon_{j-1}}),$$

where the first term in the products is interpreted as $Y_\emptyset$ or $1 - Y_\emptyset$.

Note that, on the basis of the above conditions, we may derive

$$E[\mathcal{P}(B_0)] = 1 - E[\mathcal{P}(B_1)] = \frac{\alpha_0}{\alpha_0 + \alpha_1}, \tag{1}$$

and

$$E[\mathcal{P}(B_{\varepsilon 0} | B_\varepsilon)] = 1 - E[\mathcal{P}(B_{\varepsilon 1} | B_\varepsilon)] = \frac{\alpha_{\varepsilon 0}}{\alpha_{\varepsilon 0} + \alpha_{\varepsilon 1}}, \tag{2}$$

where $\varepsilon \in E^m$ and $m = 1, 2, \ldots$.

Let $\boldsymbol{X} = \{x_1, \ldots, x_n\}$ be a random sample drawn from a random probability measure $\mathcal{P} \sim PT(\Pi, \mathcal{A})$. The posterior distribution $PT(\Pi, \mathcal{B} | \boldsymbol{X})$ can be easily obtained by updating

the parameter $\mathcal{A}$ such that $\beta_\varepsilon = \alpha_\varepsilon + n_\varepsilon$, where $\varepsilon \in E^*$, $\beta_\varepsilon \in \mathcal{B}$, $\alpha_\varepsilon \in \mathcal{A}$, and n_ε is the number of observations in $\boldsymbol{X}$ that belong to B_ε.

Moreover, the distribution of a random probability measure $\mathcal{P}$ is said to be a mixture of Polya trees if there is a random variable Θ (called the index variable) with mixing distribution H and Polya tree parameters $(\Pi_\theta, \mathcal{A}_\theta)$ such that $[\mathcal{P}|\Theta = \theta] \sim PT(\Pi_\theta, \mathcal{A}_\theta)$. Finally, for a given $\theta \in \Theta$, the posterior distribution of $[\mathcal{P}|\theta, \boldsymbol{X}]$ is obtained by updating the parameter $\mathcal{A}_\theta$ such that $\beta_\varepsilon = \alpha_\varepsilon + n_\varepsilon$ and is denoted by $PT(\Pi_\theta, \mathcal{B}_\theta|\boldsymbol{X})$.

2.2 Constructing a Polya tree

Given a continuous distribution $F_0(x|\theta)$ with an inverse function $F_0^{-1}(x|\theta)$, we are interested in constructing a Polya tree $PT(\Pi_\theta, \mathcal{A}_\theta)$ such that, for a fixed θ,

$$E[\mathcal{P}|\theta] = F_0(x|\theta), \tag{3}$$

where $[\mathcal{P}|\theta] \sim PT(\Pi_\theta, \mathcal{A}_\theta)$. That is, we wish the "center" of the probability measure $\mathcal{P}$ to be F_0.

Lavine (1992) described a canonical construction of Polya trees such that (3) is true. Conditional on $\Theta = \theta$, we define

$$\Pi_{\theta,m} = \left\{ \left[F_0^{-1}\left(\frac{k}{2^m}\middle|\theta\right), F_0^{-1}\left(\frac{k+1}{2^m}\middle|\theta\right)\right], \right.$$
$$\left. k = 0, 1, \ldots, 2^m - 1 \right\}, \tag{4}$$

and

$$\mathcal{A}_{\theta,m} = \left\{ \alpha_\varepsilon = m^2, \varepsilon \in E^m \right\}. \tag{5}$$

Thus, we obtain a Polya tree with parameters $\Pi_\theta = \{\Pi_{\theta,m}, m = 1, 2, \ldots\}$ and $\mathcal{A}_\theta = \{\mathcal{A}_{\theta,m}, m = 1, 2, \ldots\}$ that satisfies (3). Note that the canonical construction yields a $\mathcal{P}$ that is absolutely continuous with a probability of one and enjoys the best theoretical properties, such as invariance, as discussed by Lavine (1992). For different ways of constructing Polya trees that are tailored to the specific data at hand, we refer readers to Ferguson (1974), Mauldin et al. (1992), Lavine (1992, 1994), and Walker and Muliere (1997).

3 The proposed discrepancy measure $\log \rho$

3.1 Derivation: case when θ is known

Recall that $F_0(x|\theta)$ is the posited model and that we wish to assess whether F_0 is still adequate in light of recent observations by measuring the "distance" between F_0 and an unspecified alternative F_a. To derive F_a, we first construct a Polya tree $PT(\Pi_\theta, \mathcal{A}_\theta)$ using the canonical construction such that the expected value of the Polya tree is $F_0(x|\theta)$. We then update the Polya tree to obtain the posterior $PT(\Pi_\theta, \mathcal{B}_\theta|\boldsymbol{X})$ and define the nonparametric alternative as the expected value of the updated Polya tree (i.e., $F_a = E[\mathcal{P}|\theta, \boldsymbol{X}]$), where $[\mathcal{P}|\theta, \boldsymbol{X}] \sim PT(\Pi_\theta, \mathcal{B}_\theta|\boldsymbol{X})$.

Let dF_0 and dF_a be the Radon-Nikodym derivatives of F_0 and F_a, respectively, with regard to the Lebesgue measure. The Kullback-Leibler divergence is defined as

$$\rho(F_0, F_a|\theta, \boldsymbol{X}) = \int \log \frac{dF_0}{dF_a} dF_0 \tag{6}$$

and is considered here to measure the discrepancy between F_0 and F_a. According to Property 1, the divergence measure ρ can be estimated by

$$\rho_m(F_0, F_a|\theta, \boldsymbol{X}) = \sum_{\varepsilon \in E^m} F_0(B_\varepsilon|\theta) \log \frac{F_0(B_\varepsilon|\theta)}{F_a(B_\varepsilon)}. \tag{7}$$

Property 1 (Chaganty and Karandikar 1996) *Let $PT(\Pi_\theta, \mathcal{A}_\theta)$ and $PT(\Pi_\theta, \mathcal{B}_\theta|\boldsymbol{X})$ be the Polya tree prior and posterior, respectively, of a random probability measure $\mathcal{P}$. Let $m' < m$. Since $\Pi_{\theta,m}$, as defined in (4), is a finer partition than $\Pi_{\theta,m'}$, in that each set of $\Pi_{\theta,m'}$ can be written as the union of disjoint sets in $\Pi_{\theta,m}$, we find that*

$$\rho(F_0, F_a|\theta, \boldsymbol{X}) = \lim_{m\to\infty} \rho_m(F_0, F_a|\theta, \boldsymbol{X}), \tag{8}$$

where F_0 is the null posited model defined in (3) and F_a is a nonparametric alternative defined as the expected value of the updated Polya tree.

Instead of specifying an entire Polya tree, Property 1 enables us to approximate ρ with a partially specified Polya tree, namely, a Polya tree that is only updated to a predetermined level m. Thus, calculations and computer programs may be simplified, and the error of approximation can be either estimated or bounded.

Inevitably, we are faced with a decision of whether a particular divergence measure is too large to justify the use of the posited model F_0. We conducted a simulation study to explore the sampling distribution of $\log \rho_m$. Random samples of different sizes were generated from a uniform distribution F_0 with the domain [0, 1), as denoted by $U(0, 1)$. For each random sample, the Polya tree was updated to a level m to obtain the alternative F_a and the test statistic $\log \rho_m$. In Table 1, we report the means, standard deviations, and percentiles of $\log \rho_m$ for $m = 5$, 10, and 12, as well as the sample size between 10 and 1000. By comparing the simulated percentiles to those of a standard normal distribution, we can see that the sampling distribution of $\log \rho_m$ can be reasonably approximated by a normal distribution for sample size as small as 20. The simulation also indicates that

Table 1 Sampling distribution of $\log \rho_m(F_0, F_a|\theta, \boldsymbol{X})$, where $\boldsymbol{X} \sim \text{Uniform}(0, 1)$. 5000 random samples were drawn for each sample size n and the updating level m

n	Mean	Std.	Percentiles (Standardized $\log \rho_m$)									
		Dev.	0.5%	1%	2.5%	5%	10%	90%	95%	97.5%	99%	99.5%
$m = 12$												
10	−3.26	0.91	−2.43	−2.42	−2.16	−1.95	−0.99	1.48	1.59	1.72	2.29	2.34
20	−3.39	0.75	−2.55	−2.29	−1.92	−1.70	−1.31	1.27	1.63	1.85	2.29	2.35
30	−3.52	0.68	−2.54	−2.34	−1.96	−1.66	−1.28	1.30	1.64	1.95	2.23	2.46
50	−3.62	0.58	−2.62	−2.30	−2.00	−1.67	−1.29	1.28	1.63	1.91	2.26	2.51
100	−3.86	0.46	−2.56	−2.37	−2.03	−1.71	−1.31	1.26	1.59	1.90	2.24	2.45
200	−4.07	0.38	−2.60	−2.36	−1.97	−1.67	−1.30	1.28	1.59	1.92	2.26	2.45
500	−4.35	0.28	−2.50	−2.32	−1.95	−1.66	−1.29	1.27	1.66	1.96	2.29	2.40
1000	−4.57	0.22	−2.55	−2.25	−1.98	−1.67	−1.30	1.28	1.65	1.94	2.28	2.59
$m = 10$												
10	−3.26	0.91	−2.43	−2.42	−2.16	−1.95	−0.99	1.48	1.59	1.72	2.29	2.34
20	−3.39	0.75	−2.55	−2.29	−1.92	−1.70	−1.31	1.27	1.63	1.85	2.29	2.35
30	−3.52	0.68	−2.54	−2.34	−1.96	−1.66	−1.28	1.30	1.64	1.95	2.23	2.46
50	−3.62	0.58	−2.62	−2.30	−2.00	−1.67	−1.29	1.28	1.63	1.91	2.26	2.51
100	−3.86	0.46	−2.56	−2.37	−2.03	−1.71	−1.31	1.26	1.59	1.90	2.24	2.45
200	−4.07	0.38	−2.60	−2.36	−1.97	−1.67	−1.30	1.28	1.59	1.92	2.26	2.45
500	−4.35	0.28	−2.50	−2.33	−1.95	−1.66	−1.29	1.27	1.66	1.96	2.29	2.40
1000	−4.57	0.22	−2.56	−2.25	−1.99	−1.67	−1.30	1.28	1.65	1.94	2.28	2.59
$m = 5$												
10	−3.26	0.91	−2.43	−2.42	−2.16	−1.95	−0.99	1.48	1.59	1.72	2.29	2.34
20	−3.40	0.76	−2.56	−2.30	−1.92	−1.70	−1.31	1.27	1.63	1.85	2.28	2.35
30	−3.52	0.68	−2.55	−2.35	−1.96	−1.66	−1.28	1.30	1.64	1.95	2.23	2.46
50	−3.63	0.58	−2.63	−2.31	−2.01	−1.68	−1.29	1.27	1.63	1.91	2.25	2.50
100	−3.88	0.47	−2.60	−2.39	−2.05	−1.71	−1.32	1.25	1.58	1.89	2.22	2.44
200	−4.12	0.40	−2.68	−2.43	−1.99	−1.70	−1.29	1.27	1.59	1.90	2.22	2.41
500	−4.48	0.32	−2.57	−2.36	−2.04	−1.68	−1.29	1.24	1.62	1.90	2.20	2.33
1000	−4.83	0.29	−2.76	−2.50	−2.05	−1.75	−1.30	1.26	1.60	1.91	2.20	2.40
Percentiles under standard normal			−2.58	−2.33	−1.96	−1.65	−1.28	1.28	1.65	1.96	2.33	2.58

the resulting summary statistics are insensitive to the choice of m, given $m \geq 10$. Note that the closed form formulas for the mean and standard deviation of $\log \rho_m$ are difficult to derive. We approximate the formulas by running a regression of the means and standard deviations from the simulation with $m = 10$ for the sample sizes $10 \leq n \leq 1000$. We obtain

$$\hat{\mu}_0 = -\exp(1.01 + 0.07 \log n), \quad \text{and} \tag{9}$$

$$\log \hat{\sigma}_0 = 0.65 - 0.31 \log n, \tag{10}$$

where the R^2 statistic, defined as the sum of squares due to regression divided by the total sum of squares, for both regression equations exceeds 0.996.

3.2 Derivation: case when θ is unknown

An intuitive approach to dealing with the case of unknown parameter θ is to construct the alternative F_a with its maximum likelihood estimate (M.L.E.) $\hat{\theta}$. As n increases, $\hat{\theta} \to \theta$ and the Kullback-Leibler divergence ρ can be approximated by $\rho_m(F_0, F_a|\hat{\theta}, \boldsymbol{X})$. To demonstrate that the distribution of $\log \rho$ can be fairly approximated by a normal distribution, we generated 5000 random samples of size $n \in \{10, 20, 30, 50, 100, 200, 500, 1000\}$ from a standard normal distribution, i.e., $\boldsymbol{X} \sim N(\mu = 0, \sigma = 1)$, and calculate $\log \rho$ with $m = 10$ under four possible scenarios: (1) both μ and σ are known (Case 0), (2) μ is unknown but σ is known (Case 1), (3) μ is known but σ is unknown (Case 2), and (4) neither μ nor σ is known (Case 3). The unknown

Table 2 Sampling distribution of $\log \rho_m(F_0, F_a|\boldsymbol{X})$ with $m = 10$ levels. 5000 samples were generated from a standard normal distribution

n	Mean	Std. Dev.	Percentiles (Standardized $\log \rho_m$)									
			0.5%	1%	2.5%	5%	10%	90%	95%	97.5%	99%	99.5%
Case 0. Both μ and σ are known												
10	−3.29	0.92	−2.36	−2.35	−2.09	−1.89	−1.44	1.48	1.59	1.73	2.29	2.34
20	−3.40	0.75	−2.67	−2.30	−1.95	−1.73	−1.33	1.26	1.54	1.85	2.29	2.40
30	−3.50	0.66	−2.53	−2.30	−2.00	−1.70	−1.30	1.29	1.61	1.89	2.23	2.45
50	−3.64	0.58	−2.56	−2.32	−2.02	−1.68	−1.29	1.27	1.63	1.93	2.27	2.44
100	−3.85	0.47	−2.49	−2.32	−2.02	−1.69	−1.31	1.29	1.62	1.93	2.25	2.51
200	−4.06	0.38	−2.66	−2.38	−1.98	−1.65	−1.30	1.29	1.64	1.95	2.31	2.48
500	−4.35	0.28	−2.63	−2.37	−2.02	−1.70	−1.31	1.26	1.63	1.91	2.29	2.51
1000	−4.57	0.22	−2.61	−2.35	−1.98	−1.67	−1.27	1.28	1.60	1.94	2.31	2.61
Case 1. μ is unknown but σ is known												
10	−3.83	0.85	−1.99	−1.94	−1.93	−1.70	−1.60	1.37	1.64	1.72	1.94	2.10
20	−3.82	0.72	−2.91	−2.41	−1.93	−1.62	−1.26	1.30	1.58	1.88	2.18	2.32
30	−3.83	0.64	−2.53	−2.35	−1.93	−1.64	−1.29	1.31	1.67	1.89	2.18	2.42
50	−3.91	0.57	−2.44	−2.24	−1.93	−1.68	−1.31	1.29	1.64	1.94	2.28	2.50
100	−4.04	0.46	−2.51	−2.30	−1.94	−1.68	−1.28	1.28	1.67	1.97	2.32	2.58
200	−4.18	0.37	−2.53	−2.34	−1.98	−1.64	−1.29	1.31	1.67	1.96	2.28	2.48
500	−4.43	0.28	−2.56	−2.35	−1.99	−1.65	−1.29	1.31	1.63	1.89	2.21	2.43
1000	−4.61	0.22	−2.56	−2.33	−1.99	−1.62	−1.31	1.26	1.63	1.95	2.38	2.60
Case 2. μ is known but σ is unknown												
10	−3.37	0.96	−2.21	−2.17	−1.96	−1.87	−1.55	1.50	1.61	1.74	2.27	2.33
20	−3.49	0.78	−2.60	−2.32	−1.93	−1.66	−1.37	1.29	1.57	1.91	2.34	2.46
30	−3.60	0.68	−2.55	−2.29	−1.95	−1.65	−1.30	1.30	1.65	1.94	2.22	2.51
50	−3.73	0.59	−2.55	−2.36	−2.01	−1.66	−1.29	1.29	1.61	1.92	2.26	2.49
100	−3.94	0.48	−2.53	−2.32	−1.99	−1.67	−1.29	1.28	1.64	1.91	2.26	2.50
200	−4.13	0.38	−2.58	−2.32	−1.99	−1.63	−1.29	1.28	1.60	1.92	2.34	2.54
500	−4.40	0.28	−2.55	−2.32	−2.00	−1.67	−1.29	1.28	1.62	1.95	2.29	2.47
1000	−4.60	0.22	−2.75	−2.41	−1.98	−1.61	−1.28	1.25	1.65	1.95	2.27	2.54
Case 3. Neither μ nor σ is known												
10	−3.98	0.86	−1.78	−1.78	−1.72	−1.67	−1.44	1.25	1.73	1.85	1.89	2.07
20	−3.96	0.71	−3.09	−2.66	−1.85	−1.55	−1.20	1.29	1.63	1.91	2.29	2.41
30	−3.96	0.64	−2.63	−2.40	−1.96	−1.59	−1.28	1.30	1.63	1.98	2.26	2.52
50	−4.03	0.57	−2.51	−2.27	−1.97	−1.67	−1.28	1.29	1.64	2.00	2.32	2.57
100	−4.15	0.46	−2.59	−2.33	−1.99	−1.67	−1.28	1.27	1.62	1.96	2.35	2.51
200	−4.26	0.37	−2.61	−2.33	−1.96	−1.65	−1.25	1.26	1.64	1.97	2.25	2.49
500	−4.48	0.28	−2.47	−2.29	−1.93	−1.66	−1.30	1.30	1.65	1.96	2.28	2.51
1000	−4.65	0.22	−2.58	−2.26	−1.94	−1.64	−1.27	1.28	1.66	1.97	2.31	2.56
Percentiles under standard normal			−2.58	−2.33	−1.96	−1.65	−1.28	1.28	1.65	1.96	2.33	2.58

parameters are replaced with the corresponding maximum likelihood estimates and are used to construct a Polya tree prior. The simulation results are given in Table 2. Similar to the findings in Table 1, the asymptotic normality can be reasonably obtained for sample size as small as 20 across all four cases. Again, the closed form formulas for the asymptotic mean and variance of $\log \rho_m$, similar to (9) and (10), are not available. We approximate the mean and standard deviation of $\log \rho_m$, with $m = 10$, by running a regression of the simulated means and standard deviations on the sample size

Table 3 The mean and variance estimators of $\log \rho_m(F_0, F_a|\theta, \boldsymbol{X})$ as a function of the sample size n, where $m = 10$ and $\theta = \{\mu, \sigma\}$ is replaced by its maximum likelihood estimate for Cases 1, 2, and 3. All estimates of regression coefficients are significant with the p-value < 0.0001

	Case	Estimator	R^2
The mean estimators			
0.	μ is known σ is known	$\hat{\mu}_0 = -\exp(1.01 + .07 \log n)$	.997
1.	μ is unknown σ is known	$\hat{\mu}_1 = -\sqrt{13.64 + .25\sqrt{n}}$	.988
2.	μ is known σ is unknown	$\hat{\mu}_2 = -\exp(1.05 + .07 \log n)$	.998
3.	μ is unknown σ is unknown	$\hat{\mu}_3 = -\sqrt{14.85 + .22\sqrt{n}}$	.987
The standard deviation estimators			
0.	μ is known σ is known	$\log \hat{\sigma}_0 = .65 - .31 \log n$	.999
1.	μ is unknown σ is known	$\log \hat{\sigma}_1 = .54 - .29 \log n$	.996
2.	μ is known σ is unknown	$\log \hat{\sigma}_2 = .71 - .32 \log n$	.999
3.	μ is unknown σ is unknown	$\log \hat{\sigma}_3 = .56 - .30 \log n$	.997

$10 \le n \le 1000$. Table 3 summarizes the approximate means and standard deviations for all four cases. Equations (9) and (10) are certainly identical to the mean and standard deviation approximation formulas under Case 0 because both parameters μ and σ are known. All but two regression fits result in a R^2 greater than 0.99.

From a Bayesian perspective, we may estimate the (log) Kullback-Leibler divergence $\log \rho(F_0, F_a|\boldsymbol{X})$ using

$$\log \rho_m^*(F_0, F_a|\boldsymbol{X}) = \int_\theta \log \rho_m(F_0, F_a|\theta, \boldsymbol{X}) \times H(\theta|\boldsymbol{X}) d\theta, \tag{11}$$

where $H(\theta|\boldsymbol{X})$ is the posterior distribution of Θ. An analytical solution to (11) is difficult to derive except in a few special cases (e.g., Goutis and Robert 1998), and thus, the use of the Monte Carlo technique may be considered. A random sample of $\theta_1, \ldots, \theta_L$ is generated from $H(\theta|\boldsymbol{X})$. For each θ_j, a Polya tree prior is constructed with $F_0(x|\theta_j)$, and $\log \rho_m(F_0, F_a|\theta_j, \boldsymbol{X})$ is calculated using (7). Finally, $\log \rho_m^*$ can be approximated by $[\sum_{j=1}^L \log \rho_m(F_0, F_a|\theta_j, \boldsymbol{X})]/L$.

Alternatively, we may consider the posterior predictive assessment technique described by Gelman et al. (1996). Again, a random sample of $\theta_1, \ldots, \theta_L$ is generated from the posterior distribution $H(\theta|\boldsymbol{X})$. Let $\boldsymbol{X}^{rep}$ be replicated data that could have been observed under the posited model F_0. For each θ_j, we construct a Polya tree prior and draw a replicated data $\boldsymbol{X}_j^{rep}$ from $F_0(x|\theta_j)$. We then obtain two updated Polya tree posterior distributions from the observed and simulated data, $\boldsymbol{X}$ and $\boldsymbol{X}_j^{rep}$, respectively, and, as a result, two discrepancy measures, $\log \rho_m(F_0, F_a|\theta_j, \boldsymbol{X})$ and $\log \rho_m(F_0, F_a|\theta_j, \boldsymbol{X}_j^{rep})$. With L pairs of discrepancy measures, we can estimate the *posterior predictive* p-value (Rubin 1984; Meng 1994) by the proportion of the L pairs for which $\log \rho_m(F_0, F_a|\theta_j, \boldsymbol{X}) < \log \rho_m(F_0, F_a|\theta_j, \boldsymbol{X}_j^{rep})$.

A similar simulation study was conducted to investigate the distribution of $\log \rho$ with the use of non-informative priors under all four cases. For each case, random samples of various sizes are generated from a standard normal distribution to update the posterior distribution $H(\theta|\boldsymbol{X})$ of the parameter Θ. $L = 250$ θs are generated from $H(\theta|\boldsymbol{X})$ to construct Polya tree priors. The non-informative priors considered for each cases are $H(\mu) \propto$ constant, $H(\sigma^2) \propto 1/\sigma^2$, and $H(\mu, \sigma^2) \propto 1/\sigma^2$, for Cases 1, 2, and 3, respectively.

For a large sample, $H(\theta|\boldsymbol{X})$ degenerates toward θ and one may consider substituting θ with its M.L.E. However, the simulation shows that the sample size required for the normal approximation to be as good as the M.L.E. case is rather large and, in general, the distribution of $\log \rho_m$ is skewed to the right even for sample size as large as 200. This may be attributed to the choice of $L = 250$ in our simulation. A larger sample of $\theta_1, \ldots, \theta_L$ from the posterior distribution $H(\theta|\boldsymbol{X})$ may provide a more accurate description of the distribution; however, the computational time required for sampling additional θs, on top of the already computationally intensive calculation of $\log \rho_m$, makes the Bayesian consideration of unknown θ less practical. As a results, we will focus on the M.L.E. approach in the following sections. The simulation results are available upon request.

3.3 An illustration

Consider the logrithms of 100 stress-rupture lifetimes of Kevlar pressure vessels on p. 183 of Andrews and Herzberg (1985). Evans and Swartz (1994) fitted the data to the family of polynomial-normal densities and concluded that "it is clear that this is a highly non-normal dataset." Verdinelli and Wasserman (1998) calculated a Bayes factor of 0.10 meaning that the odds are 10 to 1 against the underlying distribution being normal. Using the proposed test procedure, we first update the Polya tree prior to $m = 2$, 5, 10, 15, and 20 levels and calculate the corresponding $\log \rho_m$. The discrepancy measure $\log \rho_m$ are then standardized using the $\hat{\mu}_i$ and $\hat{\sigma}_i$ formulas, $i = 0, 1, 2, 3$, in Table 3. For Cases 0, 1 and 2, the assumed known parameter values, either μ, σ, or both, are set to be equal to the sample mean and/or sample standard deviation. The summary in Table 4 strongly supports the conclusion of prior studies. That is, the data are not from

Table 4 An illustration using the logrithms of 100 stress-rupture lifetimes of Kevlar pressure vessels

Case i	$\hat{\mu}_i$	$\hat{\sigma}_i$	m	$\log \rho_m$	z statistic	p-value
0	−3.790	0.460	2	−3.365	0.924	0.1776
			5	−2.913	1.909	0.0281
			10	−2.905	1.926	0.0271
			15	−2.905	1.926	0.0271
			20	−2.905	1.926	0.0271
1	−4.017	0.451	2	−3.365	1.445	0.0742
			5	−2.913	2.448	0.0072
			10	−2.905	2.465	0.0069
			15	−2.905	2.465	0.0069
			20	−2.905	2.465	0.0069
2	−3.945	0.466	2	−3.365	1.244	0.1068
			5	−2.913	2.215	0.0134
			10	−2.905	2.231	0.0128
			15	−2.905	2.231	0.0128
			20	−2.905	2.231	0.0128
3	−4.129	0.440	2	−3.365	1.737	0.0412
			5	−2.913	2.766	0.0028
			10	−2.905	2.784	0.0027
			15	−2.905	2.784	0.0027
			20	−2.905	2.784	0.0027

a normal distribution. Other than $m = 2$, all cases show large z statistics and convincing p-values. Since the discretization is not fine enough for $m = 2$, we do not expect $\log \rho_2$ to be able to detect the discrepancy. In addition, this example again supports our assertion that $\log \rho_m$ is relatively robust to the choice of m, given $m \geq 10$. The resulting $\log \rho_m$ values are identical to 4 decimal places for $m \geq 10$.

4 Power comparison

4.1 Test for the uniform distribution

The Monte Carlo simulation of the power of the proposed test procedure was carried out for several alternative distributions. For each sample size $5 \leq n \leq 500$, 5,000 samples of size n were generated from $U(0, 1)$ (i.e., the posited model F_0 is assumed to be uniform). The Polya tree prior was constructed using the canonical construction and then updated to the $m = 10$th level to obtain the Polya tree posterior. The alternative model F_a was derived by taking the expected value of the Polya tree posterior, and the test statistic $\log \rho_m(F_0, F_a|\boldsymbol{X})$ was calculated. At the 5% significance level, the empirical power of the proposed test procedure was estimated by the proportion of the 5,000 samples that falls into the critical region $\hat{\sigma}_0^{-1}(\log \rho_m - \hat{\mu}_0) > 1.645$, where $\hat{\mu}_0$ and $\hat{\sigma}_0$ are defined in (9) and (10).

Independent samples were generated from the following continuous distributions

$$A_k(x) = 1 - (1 - x)^k \quad \text{if } 0 \leq x \leq 1; \tag{12}$$

$$B_k(x) = \begin{cases} 2^{k-1}x^k & \text{if } 0 \leq x \leq 0.5, \\ 1 - 2^{k-1}(1 - x)^k & \text{if } 0.5 \leq x \leq 1; \end{cases} \tag{13}$$

$$C_k(x) = \begin{cases} 0.5 - 2^{k-1}(0.5 - x)^k & \text{if } 0 \leq x \leq 0.5, \\ 0.5 + 2^{k-1}(x - 0.5)^k & \text{if } 0.5 \leq x \leq 1; \text{ and} \end{cases} \tag{14}$$

$$D_k(x) = x + \frac{1}{k\pi}\sin(kx\pi)$$

$$\text{if } 0 \leq x \leq 1 \text{ and } k \text{ is an integer.} \tag{15}$$

As described by Stephens (1974), A_k provides points closer to zero than would be expected by the hypothesis of uniformity and, thus, can be interpreted as a shift in the mean. Both B_k and C_k, in contrast, can be considered changes toward a smaller and a larger variance, respectively, because B_k gives more points near 0.5, whereas C_k provides two clusters close to 0 and 1. D_k, which has been studied by Swartz (1992) and Ledwina (1994), is designed to represent "spiky" data that can arise in mixtures of distributions. Figure 1 illustrates these densities with selected values of k.

The proposed test procedure was compared with four well-known test statistics for testing uniformity: Kolmogorov-Smirnov D, Cramér-von Mises W^2, Watson U^2, and Anderson-Darling A^2 statistics. We refer readers to d'Agostino and Stephens (1986, Sect. 4.2) for the definitions, computational formulas, and critical values of these test statistics. The simulation results are summarized in Table 5.

The results shown in the table indicate that the power of the proposed test statistic $\log \rho_m$ increases rather quickly as n increases with respect to the powers of D, W^2, and A^2. For example, for $C_{1.5}$, the power of $\log \rho_m$ is slightly higher than that of W^2 but lower than those of D and A^2 for $n = 30$. However, as n increases to 50, the power of $\log \rho_m$ increases more than twofold and exceeds those of D, W^2, and A^2. Similar examples can be found in the cases of A_k, B_k, and D_k distributions.

For A_k, which represents a shift in the mean, $\log \rho_m$ clearly dominates U^2 in all sample sizes. Although it does not have a higher power than do D, W^2, and A^2, the proposed test procedure is still compatible. Even for $n = 30$, the difference in power between $\log \rho_m$ and W^2 under $A_{1.5}$ is only 0.07. The difference becomes trivial as n increases. For B_k and C_k, which represent changes in variance, the Watson statistic U^2 dominates all test statistics. This is to be expected, because U^2 has been shown in several power comparison articles (e.g., Quesenberry and Miller 1977) to perform well in detecting changes in variance. The proposed $\log \rho_m$, however, provides the same stability as U^2 regardless of whether the change is towards a smaller (B_k) or larger (C_k) variance. This stability is not observed for the

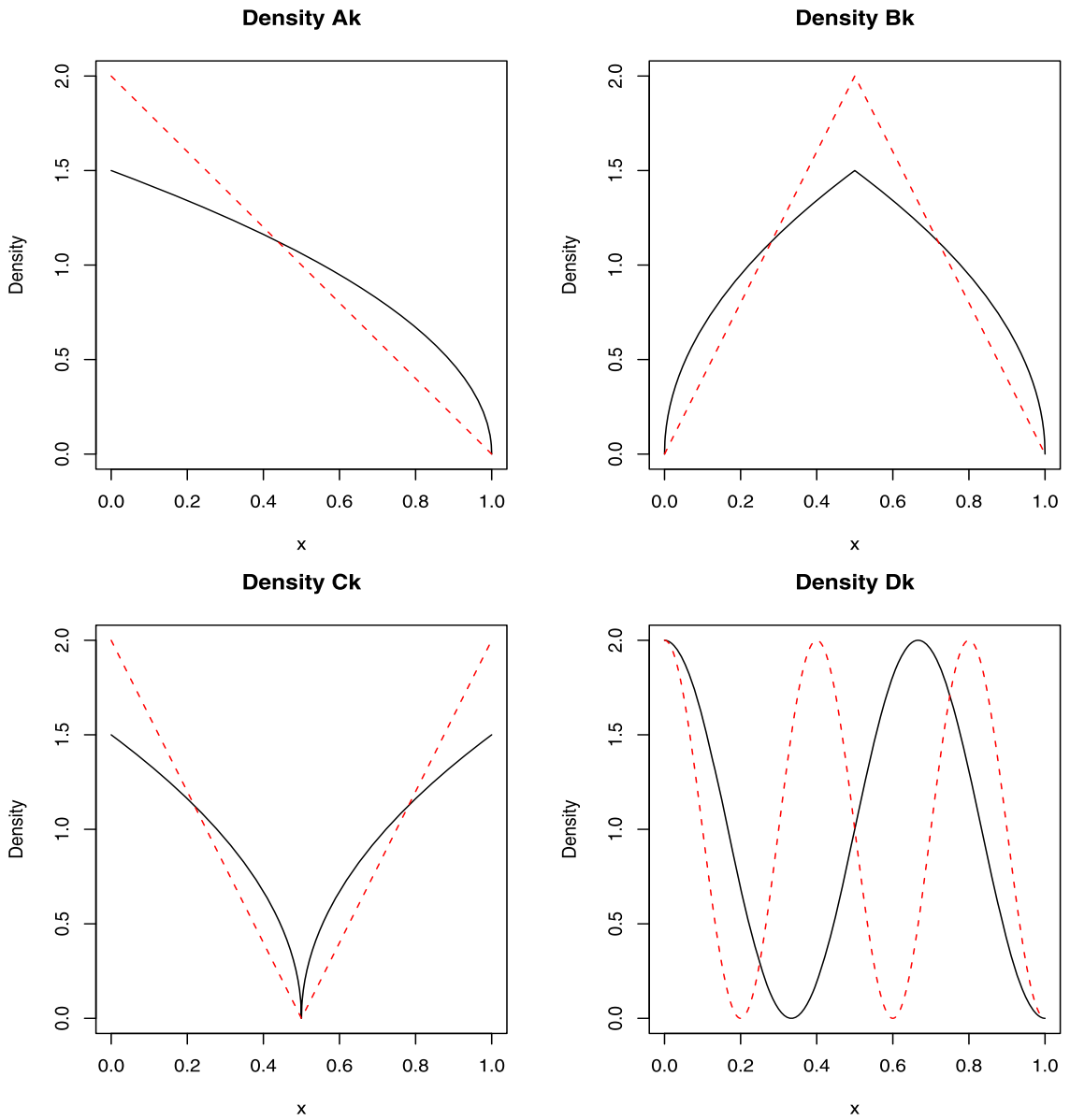

Fig. 1 Graphs of A_k, B_k, C_k, and D_k. Except for D_k, densities with the parameters $k = 1.5$ and $k = 2.0$ are plotted as *solid* and *dotted lines*, respectively. For D_k, $D_{3.0}$ is represented as the *solid line* and $D_{5.0}$ as the *dotted line*

test statistics D, W^2, and A^2, all of which have lower powers than $\log \rho_m$ in almost all cases. Finally, the simulation shows that the proposed test procedure has distinctly higher power than do the other test statistics in the case of "spiky" D_k. In many cases, the power of $\log \rho_m$ can be several times higher than that of its competitors.

4.2 Tests for the normal distribution

If the parameter value θ in the posited model $F_0(\theta)$ is specified, then the measure of discrepancy between $F_0(\theta)$ and the non-parametric alternative $F_a(\theta)$, $\rho(F_0, F_a)$, can be restated as $\rho(U_0, U_a)$, where $U_0 = F_0(\theta)$ and $U_a = F_a(\theta)$ through the probability integral transformation. As a result, the conclusions of the power comparison in the previous section will still hold. To illustrate, 1000 samples of size $n = 30, 50$, and 100 were generated from a normal distribution with a given mean μ and a given standard deviation σ, i.e. $\boldsymbol{X} \sim N(\mu, \sigma)$, $\theta = \{\mu, \sigma\}$. Setting $H_0 : F_0(\theta) = \text{Normal}(0, 1)$, we varied $-0.4 \le \mu \le 0.4$ and $0.65 \le \sigma \le 1.45$ to examine the power of ρ under the change of location and variation, similar to A_k, B_k, and C_k in the previous section. For each sample, the observation x_i is first converted to $u_i = F_0(x_i | \mu = 0, \sigma = 1)$ and the nonparametric alternative U_a is then updated to level m based on $u_1, \ldots, u_n$. The discrepancy measure $\log \rho_m$ is then calculated based on (7),

Table 5 Powers of the Proposed $\log \rho_m$, Kolmogorov-Smirnov D, Cramér-von Mises W^2, Watson U^2, and Anderson-Darling A^2 Test Statistics. The posited model is assumed to be uniform, $H_0 : F_0 \sim$ Uniform(0, 1) and A_k, B_k, C_k and D_k are the true distribution F that generate the sample $\boldsymbol{X}$

n	True F	$\log \rho_m$	D	W^2	U^2	A^2
30	$A_{1.5}$	0.39	0.40	0.46	0.23	0.45
	$A_{2.0}$	0.87	0.87	0.92	0.65	0.92
	$B_{1.5}$	0.17	0.09	0.07	0.38	0.07
	$B_{2.0}$	0.51	0.24	0.25	0.85	0.30
	$C_{1.5}$	0.16	0.19	0.14	0.37	0.20
	$C_{2.0}$	0.51	0.44	0.39	0.86	0.55
	$D_{3.0}$	0.74	0.25	0.15	0.62	0.24
	$D_{5.0}$	0.32	0.13	0.07	0.18	0.11
50	$A_{1.5}$	0.61	0.62	0.70	0.39	0.70
	$A_{2.0}$	0.99	0.99	0.99	0.89	1.00
	$B_{1.5}$	0.35	0.15	0.13	0.58	0.17
	$B_{2.0}$	0.88	0.51	0.63	0.98	0.76
	$C_{1.5}$	0.35	0.27	0.21	0.59	0.29
	$C_{2.0}$	0.87	0.68	0.70	0.99	0.82
	$D_{3.0}$	0.98	0.43	0.34	0.92	0.59
	$D_{5.0}$	0.71	0.19	0.10	0.38	0.17
100	$A_{1.5}$	0.90	0.91	0.95	0.70	0.95
	$A_{2.0}$	1.00	1.00	1.00	1.00	1.00
	$B_{1.5}$	0.75	0.35	0.41	0.90	0.58
	$B_{2.0}$	1.00	0.94	0.99	1.00	1.00
	$C_{1.5}$	0.76	0.49	0.47	0.90	0.58
	$C_{2.0}$	1.00	0.96	0.99	1.00	0.99
	$D_{3.0}$	1.00	0.89	0.96	1.00	1.00
	$D_{5.0}$	1.00	0.36	0.19	0.90	0.45

and is standardized using the mean and standard deviation formulas for Case 0 in Table 3. The discrepancy between F_0 and F_a is considered to be large if the standardized $\log \rho_m$ is greater than 1.645 for the 5% significance level. Tables 6, 7, and 8 summarize the power of $\log \rho_m$, D, W^2, U^2, and A^2 for $n = 30$, 50, and 100, respectively, under the column heading $H_0 : F_0(\theta) = \text{Normal}(\mu = 0, \sigma = 1)$. The conclusions are similar to those discussed in the previous section where the parameter θ is assumed known. Specifically, the test statistic U^2 clearly outperforms the others in the case of changing variation σ. The power of the proposed $\log \rho_m$ in the case of varying variation is higher than those of D and W^2. Although the power of A^2 is higher than that of $\log \rho_m$ as σ increases, it falls short of the power of $\log \rho_m$ as σ decreases. $\log \rho_m$ clearly outperforms U^2 in the case of varying μ, but its performance in detecting the mean shift is not better than that of D, W^2, and A^2.

If the parameter values μ and σ in the above simulation are unspecified, we propose to estimate the parameters with their corresponding maximum likelihood estimates, $\hat{\mu}$ and $\hat{\sigma}$, to ease the computational burden of the proposed $\log \rho_m$ as discussed in Sect. 3 before applying the probability integral transformation. The resulting $\log \rho_m$ should be standardized, however, using the mean and standard deviation formulas for Case 3 in Table 3, and the standardized discrepancy between $F_0(\hat{\theta})$ and $F_a(\hat{\theta})$ is considered to be large if it exceeds 1.645 for the 5% significance level. Hence, the number of false rejection is expected to be around 50 out of 1000 samples in the simulation. The results reported in Tables 6, 7, and 8 under the column heading $H_0 : F_0(\hat{\theta}) = \text{Normal}(\hat{\mu}, \hat{\sigma})$ are in line with the expectation for various sample sizes and parameter values. The simulation also highlights two advantages of the proposed $\log \rho_m$ over the other test statistics. First, modifications to the test statistics D, W^2, U^2 and A^2 are required for the case of unspecified θ, see d'Agostino and Stephens (1986, Table 4.7); whereas modification is not required of the proposed $\log \rho_m$. Second, as demonstrated in Table 2, normal approximation to the sampling distribution still holds for Case 3, as a result, the p-value of the discrepancy measure $\log \rho_m$ can be easily approximated.

4.3 Summary of power comparison

The power of several test statistics in detecting shifts in mean or changes in variance was examined in this section. Although none of the test statistics clearly dominate the others in all cases considered, the proposed $\log \rho_m$ provides a stable performance. When compared to $\log \rho_m$, it is not surprising that U^2 performs well in detecting changes in variance but it falls short in detecting mean shift. Similarly, while A^2 enjoys higher power in detecting *increasing* variance, it suffers in detecting *decreasing* variance. Both D and W^2 were designed to detect mean shift and did have higher power in those cases; however, they did not compete well in detecting changes in variance. Furthermore, in the case of "spiky" D_k when both the mean and variance vary together, the proposed $\log \rho_m$ clearly outperforms the other test statistics. In real life situations where a change in mean or in variance is typically not known *a priori*, the stability of the proposed $\log \rho_m$ provides a reasonable alternative for assessing model adequacy.

5 Adequacy of the normal model for network data

A popular packet-train model (e.g., Leland et al. 1994; Willinger et al. 1997, 1998) assumes that the (log-transformed) lengths of packet transmission between a source and destination pair during an "active" state (or ON-period) and "idle" state (or OFF-period) are governed by two distributions, F_{ON} and F_{OFF}, respectively. Several studies have demonstrated that the tails of F_{ON} and F_{OFF} deviate from that

Table 6 Powers of the Proposed $\log \rho_m$, Kolmogorov-Smirnov D, Cramér-von Mises W^2, Watson U^2, and Anderson-Darling A^2 Test Statistics. 1000 independent samples of size $n = 30$ were generated and the Polya tree was updated to level $m = 10$. Significance level = 5%

True distribution $X \sim \text{Normal}(\mu, \sigma)$		Case when θ is specified $H_0 : F_0(\theta) = \text{Normal}(\mu = 0, \sigma = 1)$					Case when θ is estimated by M.L.E. $H_0 : F_0(\hat{\theta}) = \text{Normal}(\hat{\mu}, \hat{\sigma})$				
μ	σ	$\log \rho_m$	D	W^2	U^2	A^2	$\log \rho_m$	D	W^2	U^2	A^2
0.00	0.65	0.294	0.099	0.094	0.630	0.116	0.055	0.048	0.044	0.047	0.044
0.00	0.75	0.126	0.075	0.051	0.300	0.048	0.058	0.048	0.049	0.048	0.053
0.00	0.85	0.071	0.047	0.042	0.119	0.034	0.041	0.048	0.056	0.050	0.060
0.00	0.95	0.040	0.047	0.048	0.056	0.042	0.042	0.048	0.049	0.051	0.046
0.00	**1.00**	**0.052**	**0.052**	**0.055**	**0.050**	**0.052**	**0.041**	**0.053**	**0.048**	**0.040**	**0.047**
0.00	1.05	0.044	0.049	0.049	0.061	0.054	0.042	0.053	0.058	0.053	0.056
0.00	1.15	0.066	0.080	0.072	0.097	0.124	0.044	0.066	0.062	0.062	0.062
0.00	1.25	0.089	0.106	0.106	0.187	0.222	0.042	0.045	0.054	0.057	0.052
0.00	1.35	0.119	0.148	0.155	0.306	0.346	0.047	0.051	0.051	0.051	0.052
0.00	1.45	0.152	0.202	0.214	0.447	0.526	0.051	0.039	0.039	0.041	0.041
0.40	1.00	0.397	0.472	0.528	0.242	0.546	0.058	0.043	0.057	0.054	0.056
0.30	1.00	0.224	0.270	0.312	0.139	0.326	0.044	0.053	0.055	0.056	0.052
0.20	1.00	0.151	0.181	0.208	0.107	0.215	0.057	0.042	0.043	0.046	0.040
0.10	1.00	0.060	0.070	0.074	0.063	0.082	0.039	0.046	0.058	0.054	0.057
0.05	1.00	0.044	0.050	0.054	0.053	0.047	0.040	0.041	0.048	0.047	0.042
0.00	**1.00**	**0.052**	**0.052**	**0.055**	**0.050**	**0.052**	**0.041**	**0.053**	**0.048**	**0.040**	**0.047**
−0.05	1.00	0.069	0.074	0.071	0.058	0.072	0.047	0.050	0.047	0.047	0.042
−0.10	1.00	0.074	0.081	0.080	0.059	0.079	0.056	0.050	0.055	0.049	0.061
−0.20	1.00	0.144	0.174	0.200	0.100	0.200	0.052	0.059	0.051	0.051	0.064
−0.30	1.00	0.240	0.287	0.348	0.150	0.361	0.036	0.039	0.040	0.041	0.043
−0.40	1.00	0.393	0.451	0.525	0.237	0.548	0.046	0.068	0.053	0.050	0.057

of a normal distribution, and the classical queuing assumption is not appropriate. In this section, we apply the proposed $\log \rho_m$ to measure the discrepancy between F's and the normal distribution and identify the source-destination pairs whose underlying distributions most deviate from the normal distributions.

We collected a traffic trace (data set) from the student computer laboratory at the author's host college. The trace, which consists of a total of 1,420,758 packets and 196 megabytes of data, contains the time stamp and source-destination addresses of each packet in and out of a server (source) that connects to 174 computers (destinations) in the student laboratory. The aggregate trace is separated into 174 individual traces, each of which represents the traffic flow between the server and a host computer.

In the following analysis, we define an OFF-period of a trace as any intertrain gap longer than a threshold (t seconds) that does not contain any packet transmission. In turn, this definition defines the ON-period (packet-train length). Two thresholds, $t = 2$ seconds and $t = 0.075$ seconds, are considered. Due to lack of packet transmission in some of the source-destination pairs, we only analyze and report the findings on the 100 most active source-destination pairs, which account for more than 95% of total packets recorded. Table 9 provides summary statistics of the lengths (in seconds) of intertrain gaps during the OFF period and the lengths (in seconds) of packet trains during the ON period from the 100 most active source-destination pairs. There are at least 2475 records in the ON period but the sample sizes vary between 20 and 953 for $t = 0.075$ and between 3 and 400 for $t = 2$.

By assuming the posited model is a normal distribution, i.e., the log-transformed lengths during the ON and OFF periods are normally distributed, we proceed to calculate the proposed $\log \rho_m$ statistic to quantify the discrepancy between the posited model and its nonparametric alternative. Table 10 reports the testing results of the F_{ON} and F_{OFF} for all 100 source-destination Paris under $t = 0.075$, $t = 2$, $m = 10$ and the significance level $\alpha = 5\%$. Case 0 assumes the parameters are known and equal to the sample mean and sample standard deviation whereas Case 3 considers both parameters are unknown. As a result, different formulas are used (see Table 3) to standardized $\log \rho_m$. The table clearly shows that the F_{ON} distribution of each source-destination pair and almost every F_{OFF} under $t = 0.075$ significantly deviate from normal. However, for $t = 2$, we do not have

Table 7 Powers of the Proposed $\log \rho_m$, Kolmogorov-Smirnov D, Cramér-von Mises W^2, Watson U^2, and Anderson-Darling A^2 Test Statistics. 1000 independent samples of size $n = 50$ were generated and the Polya tree was updated to level $m = 10$. Significance level = 5%

True distribution $X \sim \text{Normal}(\mu, \sigma)$		Case when θ is specified $H_0 : F_0(\theta) = \text{Normal}(\mu = 0, \sigma = 1)$					Case when θ is estimated by M.L.E. $H_0 : F_0(\hat{\theta}) = \text{Normal}(\hat{\mu}, \hat{\sigma})$				
μ	σ	$\log \rho_m$	D	W^2	U^2	A^2	$\log \rho_m$	D	W^2	U^2	A^2
0.00	0.65	0.604	0.260	0.304	0.886	0.449	0.041	0.049	0.049	0.050	0.051
0.00	0.75	0.251	0.109	0.096	0.494	0.122	0.049	0.040	0.046	0.045	0.045
0.00	0.85	0.094	0.059	0.042	0.174	0.040	0.052	0.051	0.053	0.049	0.053
0.00	0.95	0.059	0.045	0.036	0.066	0.029	0.062	0.051	0.054	0.054	0.057
0.00	**1.00**	**0.043**	**0.053**	**0.040**	**0.050**	**0.048**	**0.035**	**0.055**	**0.061**	**0.055**	**0.057**
0.00	1.05	0.052	0.062	0.063	0.064	0.076	0.052	0.051	0.050	0.054	0.044
0.00	1.15	0.084	0.092	0.087	0.146	0.136	0.048	0.061	0.050	0.052	0.053
0.00	1.25	0.138	0.112	0.111	0.297	0.286	0.039	0.051	0.047	0.053	0.046
0.00	1.35	0.236	0.213	0.207	0.520	0.506	0.048	0.048	0.042	0.039	0.042
0.00	1.45	0.335	0.303	0.306	0.698	0.740	0.036	0.047	0.049	0.047	0.054
0.40	1.00	0.611	0.691	0.755	0.399	0.781	0.059	0.046	0.056	0.049	0.056
0.30	1.00	0.381	0.451	0.503	0.230	0.523	0.049	0.057	0.056	0.048	0.060
0.20	1.00	0.181	0.226	0.245	0.130	0.262	0.043	0.057	0.053	0.053	0.048
0.10	1.00	0.075	0.100	0.101	0.070	0.100	0.054	0.046	0.057	0.055	0.042
0.05	1.00	0.052	0.048	0.054	0.051	0.056	0.045	0.044	0.051	0.051	0.046
0.00	**1.00**	**0.043**	**0.053**	**0.040**	**0.050**	**0.048**	**0.035**	**0.055**	**0.061**	**0.055**	**0.057**
−0.05	1.00	0.058	0.055	0.053	0.055	0.055	0.055	0.057	0.069	0.060	0.068
−0.10	1.00	0.070	0.086	0.096	0.068	0.094	0.042	0.050	0.063	0.060	0.062
−0.20	1.00	0.171	0.209	0.259	0.107	0.266	0.043	0.046	0.051	0.050	0.051
−0.30	1.00	0.365	0.418	0.490	0.227	0.513	0.046	0.052	0.039	0.038	0.040
−0.40	1.00	0.589	0.670	0.743	0.396	0.758	0.051	0.036	0.040	0.038	0.038

strong enough evidence to conclude that the posited model during the OFF period significantly deviates from the normal assumption for at least 25% of the source-destination pairs (see, for example, Case 3). Note that the smaller sample size may be the contributing factor of the insignificance. We also identify the top 5 source-destination pairs with the largest standardized $\log \rho_m$ values. The standardized values are greater than 10 during the OFF period and are greater than 70 during the ON period, exhibiting large deviation from the normal assumption. In summary, it is clear that the normal distribution is inappropriate to model network data.

6 Concluding remarks

A discrepancy statistic to assess model adequacy was proposed. The discrepancy statistic $\log \rho_m$ was developed to measure the "distance" between a posited model F_0 and a nonparametric alternative F_a. To measure the discrepancy, we first constructed a Polya tree prior such that the centering distribution is the posited model (i.e., $F_0 = E[\mathcal{P}|\theta]$). We then updated the prior to obtain a new centering distribution $F_a = E[\mathcal{P}|\theta, \boldsymbol{X}]$ and applied (7) to calculate $\log \rho_m$. Finally, normal approximation to the sampling distribution of $\log \rho_m$ provides us a convenient decision rule by which to judge whether a resulting $\log \rho_m$ is too large to justify the use of the current model F_0.

The advantages of using the proposed statistic are several. First, it can be applied to a wide class of distributions F_0 whose inverse function F_0^{-1} exists. Second, unlike many popular test statistics, *ad hoc* tables of critical values for different levels of significance are not required of the proposed $\log \rho_m$. Third, as shown in the power study, its power increases quickly as the sample size increases and maintains stable powers across a wide range of alternatives. It has distinctly higher power than several popular test statistics in the case of "spiky" alternatives, dominates the well-known Kolmogorov-Smirnov D statistic in all cases examined in the article, and competes well with the highly recommended Watson statistic U^2 (Quesenberry and Miller 1977) in the case of a variance change. The proposed $\log \rho_m$ statistic is applied to network data collected from a computer laboratory and the most active source-destination pairs whose ON- and OFF-distributions significantly deviate from the normal distribution are identified.

Table 8 Powers of the Proposed $\log \rho_m$, Kolmogorov-Smirnov D, Cramér-von Mises W^2, Watson U^2, and Anderson-Darling A^2 Test Statistics. 1000 independent samples of size $n = 100$ were generated and the Polya tree was updated to level $m = 10$. Significance level = 5%

True distribution $X \sim \text{Normal}(\mu, \sigma)$		Case when θ is specified $H_0 : F_0(\theta) = \text{Normal}(\mu = 0, \sigma = 1)$					Case when θ is estimated by M.L.E. $H_0 : F_0(\hat{\theta}) = \text{Normal}(\hat{\mu}, \hat{\sigma})$				
μ	σ	$\log \rho_m$	D	W^2	U^2	A^2	$\log \rho_m$	D	W^2	U^2	A^2
0.00	0.65	0.983	0.706	0.831	1.000	0.968	0.047	0.051	0.053	0.057	0.055
0.00	0.75	0.646	0.272	0.292	0.860	0.493	0.053	0.053	0.051	0.047	0.046
0.00	0.85	0.213	0.094	0.071	0.356	0.096	0.058	0.037	0.051	0.047	0.049
0.00	0.95	0.059	0.039	0.036	0.068	0.034	0.057	0.040	0.038	0.037	0.043
0.00	**1.00**	**0.060**	**0.050**	**0.058**	**0.052**	**0.061**	**0.055**	**0.047**	**0.043**	**0.044**	**0.044**
0.00	1.05	0.038	0.047	0.061	0.062	0.075	0.053	0.052	0.052	0.052	0.049
0.00	1.15	0.139	0.091	0.084	0.248	0.198	0.066	0.059	0.054	0.053	0.050
0.00	1.25	0.334	0.215	0.217	0.577	0.529	0.049	0.045	0.044	0.040	0.042
0.00	1.35	0.632	0.387	0.438	0.829	0.808	0.050	0.044	0.049	0.045	0.042
0.00	1.45	0.780	0.569	0.628	0.940	0.942	0.048	0.042	0.034	0.036	0.035
0.40	1.00	0.898	0.932	0.965	0.725	0.973	0.040	0.066	0.057	0.058	0.052
0.30	1.00	0.651	0.747	0.833	0.453	0.855	0.039	0.064	0.055	0.055	0.053
0.20	1.00	0.300	0.390	0.470	0.206	0.487	0.045	0.042	0.044	0.043	0.050
0.10	1.00	0.098	0.127	0.151	0.063	0.153	0.053	0.060	0.056	0.060	0.055
0.05	1.00	0.053	0.069	0.078	0.059	0.081	0.051	0.049	0.053	0.050	0.043
0.00	**1.00**	**0.060**	**0.050**	**0.058**	**0.052**	**0.061**	**0.055**	**0.047**	**0.043**	**0.044**	**0.044**
−0.05	1.00	0.058	0.067	0.070	0.053	0.070	0.033	0.052	0.054	0.056	0.049
−0.10	1.00	0.106	0.130	0.152	0.082	0.153	0.061	0.051	0.045	0.047	0.048
−0.20	1.00	0.298	0.401	0.470	0.205	0.482	0.057	0.041	0.048	0.048	0.046
−0.30	1.00	0.649	0.748	0.806	0.421	0.834	0.045	0.052	0.057	0.054	0.052
−0.40	1.00	0.889	0.935	0.964	0.702	0.973	0.051	0.052	0.052	0.051	0.053

Table 9 Summary statistics of length (in seconds) of intertrain gap for the OFF period and packet trains for the ON period from the 100 most active source-destination pairs

Lengths in seconds	Top 100 Pairs	OFF-Period		ON-Period	
		$t = 0.075$	$t = 0.2$	$t = 0.075$	$t = 0.2$
Mininum	0.000002	0.075	2	0.000002	0.000002
1st Quartile	0.000145	0.1373	4.807	0.000142	0.000144
Median	0.000379	0.1941	5.006	0.000375	0.000377
Mean	0.1716	10.27	31.66	0.001287	0.004863
3rd Quartile	0.001041	4.128	20.52	0.000948	0.001011
Maximum	1309	1309	1309	0.07499	2
Standard deviation	5.900	44.651	74.899	0.004	0.056
Sample size					
Total # of records	1351365	22428	7120	1328937	1344245
Minimum #	2648	20	3	2475	2582
1st Quartile	4294	109	27.75	4190.75	4246.5
Median #	8379	148	43.5	8298	8355
3rd Quartile	19552.5	263	85	19323.5	19414.25
Maximum #	75682	953	400	74890	75600
Standard deviation	13225.95	204.06	79.82	13100.97	13187.57

Table 10 Percentage of the source-destination pairs whose underlying distributions significantly deviate from the normal distribution based on the proposed $\log \rho_m$. Significance level = 5%, and $m = 10$

	OFF-Period		ON-Period	
	$t = 0.075$	$t = 2$	$t = 0.075$	$t = 2$
Case 0	97%	55%	100%	100%
Case 1	99%	71%	100%	100%
Case 2	97%	58%	100%	100%
Case 3	99%	74%	100%	100%
Top 5 pairs with the largest standardized $\log \rho_m$ values				
Cases 0 and 2	8, 39, 41	8, 22, 39	8, 41, 43, 77, 159	
	43, 159	43, 159		
Cases 1 and 3	8, 39, 41	8, 22, 39	8, 11, 43, 139, 159	
	43, 159	43, 159		

The current study may lead to various research directions. As suggested by a reviewer, it would be interesting to examine the impact of the selected priors on the derived discrepancy measure. The current study chose the Polya trees prior based on a continuous alternative F_a assumption, but what will be the pros and cons of using, say, a Dirichlet process in constructing a discrepancy measure? At the miminum, a comprehensive simulation study must be carried out to address this issue. Furthermore, the current construction of the Polya trees alternative can be modified, according to the outline by Lavine (1992) and Neath (2003), to accommodate censored or grouped data often observed in survival analysis. The sampling distribution of the test statistic $\log \rho_m$ might not be tractable in this case; nevertheless, its critical values can be obtained through Monte Carlo simulations. Finally, if certain subsets of the underlying distribution are of major concern, for example, routine estimation of the tail area probabilities in risk analysis, the proposed $\log \rho_m$ can be modified to sum over the tail areas of the distribution and thereby obtain the discrepancy in the tails between F_0 and F_a.

Acknowledgements The author acknowledges financial support by the Vivian Bales Research Felloship and the assistance of Information Services group at the College of Business, Oregon State University. In addtion, the author wishes to thank the Editor, AE, and two anonymous referees for constructive suggestions that have improved the article significantly.

References

Andrews, D.F., Herzberg, A.M.: Data: A Collection of Problems from Many Fields for the Student and Research Worker. Springer, Berlin (1985)

Arizono, I., Ohta, H.: A test for normality based on Kullback-Leibler information. Am. Stat. **43**, 20–22 (1989)

Berger, J.O., Guglielmi, A.: Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives. J. Am. Stat. Assoc. **96**, 174–184 (2001)

Carota, C., Parmigiani, G.: On Bayes factors for nonparametric alternatives. In: Bernardo, J.M., Berger, J.O., David, A.P., Smith, A.F.M. (eds.) Bayesian Statistics 5, pp. 507–511. Clarendon Press, Oxford (1996)

Chaganty, N.R., Karandikar, R.L.: Some properties of the Kullback-Leibler number. Sankhyā, Ser. A **58**, 69–80 (1996)

d'Agostino, R.B., Stephens, M.A.: Goodnesso-of-fit techniques. Statistics: Textbooks and Monographs, vol. 68. Marcel Dekker, New York (1986)

Dudewicz, E.J., van der Meulen, E.C.: Entropy-based tests of uniformity. J. Am. Stat. Assoc. **76**, 967–974 (1981)

Ebrahimi, N., Habibullah, M., Soofi, E.S.: Testing exponentiality based on Kullback-Leibler information. J. R. Stat. Soc., Ser. B **54**, 739–748 (1992)

Evans, M., Swartz, T.: Distribution theory and inference for polynomial-normal densities. Commun. Stat., Theory Methods **23**(4), 1123–1148 (1994)

Ferguson, T.S.: Prior distributions on spaces of probability measures. Ann. Stat. **2**, 615–629 (1974)

Gelman, A., Meng, X.-L., Stern, H.: Posterior predictive assessment of model fitness via realized discrepancies. Stat. Sin. **6**, 733–807 (1996)

Goutis, C., Robert, C.: Model choice in generalized linear models: a Bayesian approach via Kullback-Leibler projections. Biometrika **85**, 29–37 (1998)

Hsieh, P.-H.: An exploratory first step in teletraffic data modeling: evaluation of long-run performance of parameter estimators. Comput. Stat. Data Anal. **40**, 263–283 (2002)

Lavine, M.: Some aspects of polya tree distributions for statistical modelling. Ann. Stat. **20**, 1222–1235 (1992)

Lavine, M.: More aspects of polya tree distributions for statistical modelling. Ann. Stat. **22**, 1161–1176 (1994)

Ledwina, T.: Data-driven version of Neyman's smooth test of fit. J. Am. Stat. Assoc. **89**, 1000–1005 (1994)

Leland, W.E., Taqqu, M.S., Willinger, W., Wilson, D.V.: On the self-similar nature of ethernet traffic (Extended Version). IEEE/ACM Trans. Netw. **2**, 1–15 (1994)

Mauldin, R.D., Sudderth, W.D., Williams, S.C.: Polya trees and random distributions. Ann. Stat. **20**, 1203–1221 (1992)

Meng, X.L.: Posterior predictive p-values. Ann. Stat. **22**, 1142–1160 (1994)

Mengerson, K., Robert, C.: Testing for mixtures: a Bayesian entropic approach. In: Bernardo, J.M., Berger, J.O., David, A.P., Smith, A.F.M. (eds.) Bayesian Statistics 5, pp. 255–276. Clarendon Press, Oxford (1996)

Neath, A.A.: Polya tree distributions for statistical modeling of censored data. J. Appl. Math. Decis. Sci. **7**(3), 175–186 (2003)

Quesenberry, C.P., Miller Jr., F.L.: Power studies of some tests for uniformity. J. Stat. Comput. Simul. **5**, 169–191 (1977)

Rubin, D.B.: Bayesianly justifiable and relevant frequency calculations for the applied statistician. Ann. Stat. **12**, 1151–1172 (1984)

Stephens, M.A.: EDF statistics for goodness of fit and some comparisons. J. Am. Stat. Assoc. **69**, 730–737 (1974)

Swartz, T.: Goodness-of-fit tests using Kullback-Leibler information. Commun. Stat. Part. B, Simul. Comput. **21**, 711–729 (1992)

Vasicek, O.: A test for normality based on sample entropy. J. R. Stat. Soc., Ser. B **38**, 54–59 (1976)

Verdinelli, I., Wasserman, L.: Bayesian goodness-of-fit testing using infinite-dimensional exponential families. Ann. Stat. **26**(4), 1215–1241 (1998)

Viele, K.: Evaluating fit using Dirichlet processes. Technical Report 384, Department of Statistics, University of Kentucky (http://web.as.uky.edu/statistics/techreports/techreports.html) (2000)

Walker, S., Muliere, P.: A characterisation of polya tree distributions. Stat. Probab. Lett. **31**, 163–168 (1997)

Willinger, W., Taqqu, M.S., Sherman, R., Wilson, D.V.: Self-similarity through high-variability: statistical analysis of ethernet LAN traffic at the source level (Extended Version). IEEE/ACM Trans. Netw. **5**, 71–86 (1997)

Willinger, W., Paxson, V., Taqqu, M.S.: Self-similarity and heavy tails: structural modeling of network traffic. In: Adler, R., Feldman, R., Taqqu, M.S. (eds.) A Practical Guide to Heavy Tails: Statistical Techniques and Applications, pp. 27–53. Birkhäuser, Boston (1998)