entropy

Article

A Bayesian Motivated Two-Sample Test Based on Kernel Density Estimates

Naveed Merchant * and Jeffrey D. Hart

Department of Statistics, Texas A&M University, College Station, TX 77840, USA; hart@stat.tamu.edu
* Correspondence: naveedmerchant2@tamu.edu

Abstract: A new nonparametric test of equality of two densities is investigated. The test statistic is an average of log-Bayes factors, each of which is constructed from a kernel density estimate. Prior densities for the bandwidths of the kernel estimates are required, and it is shown how to choose priors so that the log-Bayes factors can be calculated exactly. Critical values of the test statistic are determined by a permutation distribution, conditional on the data. An attractive property of the methodology is that a critical value of 0 leads to a test for which both type I and II error probabilities tend to 0 as sample sizes tend to ∞. Existing results on Kullback–Leibler loss of kernel estimates are crucial to obtaining these asymptotic results, and also imply that the proposed test works best with heavy-tailed kernels. Finite sample characteristics of the test are studied via simulation, and extensions to multivariate data are straightforward, as illustrated by an application to bivariate connectionist data.

Keywords: Bayes factors; permutation tests; cross-validation; consistent tests; Kolmogorov–Smirnov test

Citation: Merchant, N.; Hart, J.D. A Bayesian Motivated Two-Sample Test Based on Kernel Density Estimates. *Entropy* **2022**, *24*, 1071. https://doi.org/10.3390/e24081071

Academic Editor: Udo Von Toussaint

Received: 1 July 2022
Accepted: 27 July 2022
Published: 3 August 2022

1. Introduction

Ref. [1] proposed the use of cross-validation Bayes factors in the classic two-sample problem of comparing two distributions. Their basic idea is to randomly divide the data into two distinct parts, call them A and B, and to define two models based on kernel density estimates from part A. One model assumes that the two distributions are the same and the other allows them to be different. A Bayes factor comparing the two part A models is then defined from the part B data. In order to stabilize the Bayes factor, Ref. [1] suggest that a number of different random data splits be used, and the resulting log-Bayes factors averaged.

In the current paper we consider a special case of this approach in which the part A data consists of all the available observations save one. If the sample sizes of the two data sets are m and n, this entails that a total of $m+n$ log-Bayes factors may be calculated. The average of these $m+n$ quantities becomes the test statistic here considered, and is termed ALB.

Although ALB is an average of log-Bayes factors, it does not lead to a consistent Bayes test because each of the log-Bayes factors is based on just a single observation. Ref. [1] suppose that the validation set size grows to ∞, while in our case it remains of size 1. This results in the ALB converging to the Kullback–Leibler divergence of the two densities, and not ∞ as in the case of [1]. We therefore use frequentist ideas to construct our test. The exact null distribution of ALB conditional on order statistics is obtained using permutations of the data. Doing so leads to a consistent frequentist test whose size is controlled exactly. The problem of bandwidth selection is dealt with by using leave-one-out likelihood cross-validation applied to the combination of the two data sets. This method is computationally efficient in that the resulting bandwidth is invariant to permutations of the combined data, and therefore has to be computed just once. Our methodology is easily extended to bivariate data, and we do so in a real data example.

Ref. [2] also use a permutation test based on kernel estimates for the two-sample problem, their statistic being based on an L_2 distance. Ref. [3] shows how other distances and divergences compare when applying them to the general k-sample problem, restricting their comparisons to the one-dimensional case. Our method mainly differs from these procedures by virtue of its Bayesian motivation. Existing methodology that most closely resembles ours is that of [4], who use a kernel-based marginal likelihood ratio to test goodness of fit of parametric models for a distribution. Their marginal likelihood employs a prior for a bandwidth, as does ours.

2. Methodology

We assume that $\mathbf{X} = (X_1, \ldots, X_m)$ are independent and identically distributed (i.i.d.) from density f, and independently $\mathbf{Y} = (Y_1, \ldots, Y_n)$ are i.i.d. from density g. We are interested in the problem of testing the null hypothesis that f and g are identical on the basis of the data $\mathbf{X}$ and $\mathbf{Y}$. Let $\mathbf{U} = (U_1, \ldots, U_k)$ be an arbitrary set of k scalar observations, and define a kernel density estimate by

$$\hat{f}_K(u|h, \mathbf{U}) = \frac{1}{kh} \sum_{i=1}^{k} K\left(\frac{u - U_i}{h}\right), \quad -\infty < u < \infty,$$

where K is the kernel and $h > 0$ the bandwidth.

2.1. The Test Statistic

Let $Z_i = X_i$, $i = 1, \ldots, m$, $Z_i = Y_{i-m}$, $i = m+1, \ldots, m+n$, $\mathbf{Z} = (Z_1, \ldots, Z_{m+n})$ and $\mathbf{Z}_i$ be the vector $\mathbf{Z}$ with all its components except Z_i, $i = 1, \ldots, m+n$. Furthermore, let $\mathbf{X}_i$ be all the components of $\mathbf{X}$ except X_i, $i = 1, \ldots, m$, and $\mathbf{Y}_j$ all the components of $\mathbf{Y}$ except Y_j, $j = 1, \ldots, n$. If we assume that f is identical to g, then potential models for f are $M_{0i} = \{\hat{f}_K(\,\cdot\,|h, \mathbf{Z}_i) : h > 0\}$, $i = 1, \ldots, m+n$. Suppose that $1 \le i \le m$. If we allow that f and g are different, then a model for the datum Z_i is $M_{1i} = \{\hat{f}_K(\,\cdot\,|a, \mathbf{X}_i) : a > 0\}$. In this case a legitimate Bayes factor for comparing M_{0i} and M_{1i} on the basis of the datum Z_i has the form

$$B_i = \frac{\int_0^\infty \pi(a) \hat{f}_K(Z_i|a, \mathbf{X}_i)\, da}{\int_0^\infty \pi(h) \hat{f}_K(Z_i|h, \mathbf{Z}_i) dh}, \quad i = 1, \ldots, m,$$

where, mainly for convenience, we have assumed that the bandwidth priors are the same in all cases. Likewise, if $i = m+1, \ldots, m+n$, then $M_{1i} = \{\hat{f}_K(\,\cdot\,|b, \mathbf{Y}_{i-m}) : b > 0\}$ is a model for the datum Z_i, and a Bayes factor for comparing M_{0i} and M_{1i} is

$$B_i = \frac{\int_0^\infty \pi(a) \hat{f}_K(Z_i|a, \mathbf{Y}_{i-m})\, da}{\int_0^\infty \pi(h) \hat{f}_K(Z_i|h, \mathbf{Z}_i) dh}, \quad i = m+1, \ldots, m+n.$$

When m and n are large, it is expected that M_{1i} will be a good model for f if $i = 1, \ldots, m$ and for g if $i = m+1, \ldots, m+n$. Likewise, each of M_{0i} will be a good model for the common density on the assumption that f and g are identical. However, none of $B_1, \ldots, B_{m+n}$ will be Bayes factors that can provide convincing evidence for either hypothesis simply because each one uses likelihoods based on a single datum. At first blush one might think that a solution to this problem is to take the average of the $m+n$ log-Bayes factors:

$$ALB = \frac{1}{(m+n)} \sum_{i=1}^{m+n} \log B_i. \tag{1}$$

However, this results in a statistic that will consistently estimate 0 or a positive constant in the respective cases $f \equiv g$ or $f \not\equiv g$. In neither case does the statistic have the property of Bayes consistency, i.e., the property that the Bayes factor tends to 0 and ∞ when $f \equiv g$ and $f \not\equiv g$, respectively.

The discussion immediately above points out a fundamental fact that seems not to have been widely discussed: combining a large number of inconsistent Bayes factors does not necessarily lead to a consistent Bayes factor. A guiding principle in [1] was that of averaging log-Bayes factors from different random splits of the data with the aim of producing a more stable log-Bayes factor. However, in order for this practice to yield a *consistent* Bayes factor, it is important that each of the log-Bayes factors being averaged is consistent. Furthermore, to ensure this consistency, it is necessary that the sizes of both the training and validation sets tend to ∞ with the samples sizes m and n. Obviously this is not the case when the size of each validation set is just 1, as in the current paper.

An advantage of the approach proposed herein is that the practitioner does not have to choose the size of the training sets. The cost is that the resulting statistic does not have the property of Bayes consistency. We thus propose that the statistic be used in frequentist fashion. An appealing way of doing so is to use a permutation test, which (save for certain practical issues to be discussed) leads to a test with exact type I error probability for all $m > 1$ and $n > 1$. Let $Z_{(1)} < Z_{(2)} < \cdots < Z_{(m+n)}$ be the order statistics for the combined sample. Let $\mathbf{j} = (j_1, \ldots, j_{m+n})$ be a random permutation of $1, \ldots, m+n$, and define $T(\mathbf{j})$ to be the statistic (1) when the X-sample is taken to be $Z_{j_1}, \ldots, Z_{j_m}$ and the Y-sample to be $Z_{j_{m+1}}, \ldots, Z_{j_{m+n}}$. It follows that, conditional on the order statistics $Z_{(1)}, \ldots, Z_{(m+n)}$, the $(m+n)!$ values taken on by $T(\cdot)$ are equally likely. Therefore, if $t_{m,n}$ is a $1-\alpha$ quantile of the empirical distribution of $T(\cdot)$, then the test that rejects $f \equiv g$ when $T \geq t_{m,n}$ will have an (unconditional) type I error probability of α. As will be shown in the Appendix A.3, ALB is negative with probability tending to 1 as $m, n \to \infty$, implying that for any $\alpha > 0$ $t_{m,n}$ will be negative for m and n large enough. From an evidentiary standpoint, it is nonsense to reject H_0 for a negative value of ALB. We therefore suggest using the critical value $\max(0, t_{m,n})$, which ensures that the test is sensible and has level α.

2.2. The Effect of Using Scale Family Priors

Let π_0 be an arbitrary density with support $(0, \infty)$. A possible family of priors is one that contains all rescaled versions of π_0. For $b > 0$, using the prior $\pi(h) = \pi_0(h/b)/b$ and making the change of variable $h/b = u$ in the denominator of B_i, we have

$$\int_0^\infty b^{-1}\pi_0(h/b)\hat{f}_K(Z_i|h, \mathbf{Z}_i)dh = \hat{f}_L(Z_i|b, \mathbf{Z}_i),$$

where the kernel L is

$$L(z) = \int_0^\infty u^{-1}\pi_0(u)K(z/u)\,du, \quad \text{for all } z. \tag{2}$$

So, by using this type of prior, each marginal likelihood comprising ALB becomes a kernel density estimate with bandwidth equal to the scale parameter of the prior. In one sense this is disappointing since it means that averaging kernel estimates with respect to a bandwidth prior does not actually sidestep the issue of choosing a smoothing parameter. One has simply traded bandwidth choice for choice of the prior's scale. However, it turns out that there is a quantifiable advantage to using a prior for the bandwidth of K. As detailed in the Appendix A.2, likelihood cross-validation is often more efficient when applied to $\hat{f}_L$ rather than to $\hat{f}_K$.

When using a scale family of priors, the result immediately above implies that

$$\begin{aligned}(m+n)ALB \quad = \quad & \sum_{i=1}^{m}\log(\hat{f}_L(X_i|b, \mathbf{X}^i)) + \sum_{j=1}^{n}\log(\hat{f}_L(Y_j|b, \mathbf{Y}^j)) \\ & - \sum_{i=1}^{m+n}\log(\hat{f}_L(Z_i|b, \mathbf{Z}^i)),\end{aligned} \tag{3}$$

and so the proposed statistic is proportional to the log of a likelihood ratio. The two likelihoods are cross-validation likelihoods, and the numerator and denominator of the ratio correspond to the hypotheses of different and equal densities, respectively.

In practice one must select both the kernel L and bandwidth b. For the moment we assume that L is given. The denominator of $\exp((m+n)ALB)$ as a function of b is the likelihood cross-validation criterion, as studied by [5], based on the combined sample. We propose using $b = \hat{b}$, the maximizer of this denominator. This bandwidth has the desirable property that it is invariant to the ordering of the data in the combined sample. Let ALB^* be the value of test statistic (1) for a permuted data set. One should use the principle that ALB^* is the same function of the permuted data as ALB is of the original data. So, in principle the bandwidth should be selected for every permuted data set, but because of the invariance of $\hat{b}$ to the ordering of the combined sample, this data-driven bandwidth equals $\hat{b}$ for every permuted data set. This results in a large computational savings relative to a procedure that selects the bandwidth differently for the X- and Y-samples. Using the same bandwidth under both null and alternative hypotheses also fits with the principle espoused by [6].

Concerning L, Ref. [5] showed that kernels must be relatively heavy-tailed in order for them to perform well with respect to likelihood cross-validation. In particular, he shows that likelihood cross-validation fails miserably as a method for choosing the bandwidth of a kde based on a Gaussian kernel. The tails of the kernel must be considerably heavier than those of a Gaussian density in order for likelihood cross-validation to be effective. Proposition A1 in the Appendix A.1 shows that under very general conditions L (as defined in (2)) has heavier tails than those of K. Therefore, the Bayesian notion of averaging commonly used kernel estimates with respect to a prior brings the resulting kernel estimate more in line with the conditions of [5]. This has a substantial benefit for our statistic inasmuch as we use a likelihood cross-validation bandwidth in its construction.

Consider the following kernel proposed by [5]:

$$L_0(u) = \frac{1}{\sqrt{8\pi e}\,\Phi(1)} \exp\left[-\frac{1}{2}(\log(1+|u|))^2\right].$$

Suppose that a kde is defined using kernel L_0 and its bandwidth is chosen by likelihood cross-validation. Ref. [5] shows that, in general, this cross-validation bandwidth will be asymptotically optimal in a Kullback–Leibler sense. We will therefore use L_0 in all subsequent simulations. Results in the Appendix A.2 provide a kernel K and corresponding prior that produce L_0.

2.3. Further Properties of ALB

In the Appendix A.3 we will show that the ALB test is consistent in the frequentist sense. In other words, for any alternative the power of an ALB test of fixed level tends to 1 as m and n tend to ∞.

Interestingly, ALB has the property of being sharply bounded above. It can be rewritten as follows:

$$\sum_{i=1}^{m} \log(\hat{f}_L(X_i|b,\mathbf{X}^i)/\hat{f}_L(X_i|b,\mathbf{Z}^i)) + \sum_{j=1}^{n} \log(\hat{f}_L(Y_j|b,\mathbf{Y}^j)/\hat{f}_L(Y_j|b,\mathbf{Z}^{m+j})).$$

Defining $p_{m,n} = (m-1)/(m+n-1)$,

$$\hat{f}_L(X_i|b,\mathbf{Z}^i) = p_{m,n}\hat{f}_L(X_i|b,\mathbf{X}^i) + (1-p_{m,n})\hat{f}_L(X_i|b,\mathbf{Y}), \quad i = 1,\ldots,m,$$

and therefore

$$\frac{\hat{f}_L(X_i|b,\mathbf{X}^i)}{\hat{f}_L(X_i|b,\mathbf{Z}^i)} = \frac{1}{p_{m,n}} \cdot \frac{p_{m,n}\hat{f}_L(X_i|b,\mathbf{X}^i)}{p_{m,n}\hat{f}_L(X_i|b,\mathbf{X}^i) + (1-p_{m,n})\hat{f}_L(X_i|b,\mathbf{Y})} \le \frac{1}{p_{m,n}}.$$

A similar bound applies for the other component of ALB, implying that

$$ALB \leq -\left[\left(\frac{m}{m+n}\right)\log(p_{m,n}) + \left(1-\frac{m}{m+n}\right)\log\left(\frac{n-1}{m+n-1}\right)\right]. \tag{4}$$

Using the fact that $-[x\log(x)+(1-x)\log(1-x)]$ has its maximum at $x = 1/2$ when $0 \leq x \leq 1$, bound (4) implies that

$$ALB \leq \log(2)\cdot\max\left(\frac{m}{(m-1)}, \frac{n}{(n-1)}\right).$$

Unless one of m and n is very small, the effective bound on ALB is $\log(2)$. This reinforces the fact that ALB does not have the property of Bayes consistency. While it is true that ALB is an average of Bayes factors, none of these Bayes factors can ever provide compelling evidence in favor of the alternative. To reiterate, this problem is overcome by employing ALB in frequentist fashion.

While ALB can take on positive values when the null hypothesis is true, our proof of frequentist consistency shows that, under H_0, $P(ALB < 0) \rightarrow 1$ as $m, n \rightarrow \infty$. This implies that if 0 is used as a critical value, then the resulting test level tends to 0 as $m, n \rightarrow \infty$. So, even though $|ALB|$ does not tend to ∞, the *sign* of ALB provides compelling evidence for the hypotheses of interest when the sample sizes are large.

The exact conditional distribution of ALB is known under the null hypothesis, as we use a permutation test. Nonetheless, it is of some interest to have an impression of the *unconditional* distribution of ALB. To this end, we randomly select two normal mixture densities that differ. The number of components M in the first mixture is between 2 and 20 and chosen from a distribution such that the probability of m is proportional to m^{-1}, $m = 2, \ldots, 20$. Given $M = m$, mixture weights are drawn from a Dirichlet distribution with all m parameters equal to $1/2$. Given $M = m$ and mixture weights, variances $\sigma_1^2, \ldots, \sigma_m^2$ of the normal components are a random sample from an inverse gamma distribution with both parameters equal to $1/2$. Finally, means $\mu_1, \ldots, \mu_m$ of the normal components are such that $\mu_1, \ldots, \mu_m$ given $\sigma_1, \ldots, \sigma_m$ are independent with $\mu_j|\sigma_j \sim N(0, \sigma_j^2)$, $j = 1, \ldots, m$. The second normal mixture is independently selected using exactly the same mechanism. Random selection of densities in this manner for simulation studies has been proposed and explored in [7].

We draw a sample of size 100 from each of the two randomly generated densities (so that $m = n = 100$), and then compute ALB. This procedure is replicated on the same two densities 100 times. After this, we repeat the whole procedure for nine more pairs of randomly selected densities. The results are seen in Figure 1. Save for case 3, the proportion of positive ALBs is nearly 1 in all cases.

We repeated a similar procedure for the null hypothesis setting. The simulation was exactly the same except that in each of the ten cases, only one density was generated, and a pair of independent samples (of size 100 each) was selected from this same density. The resulting ALB distributions can be seen in Figure 2. The proportion of the cases where $ALB < 0$ for the 10 densities were, respectively, 0.89, 0.83, 0.83, 0.84, 0.85, 0.87, 0.91, 0.84, 0.84, and 0.76. These results are consistent with the fact that $P(ALB < 0)$ tends to 1 with sample size.

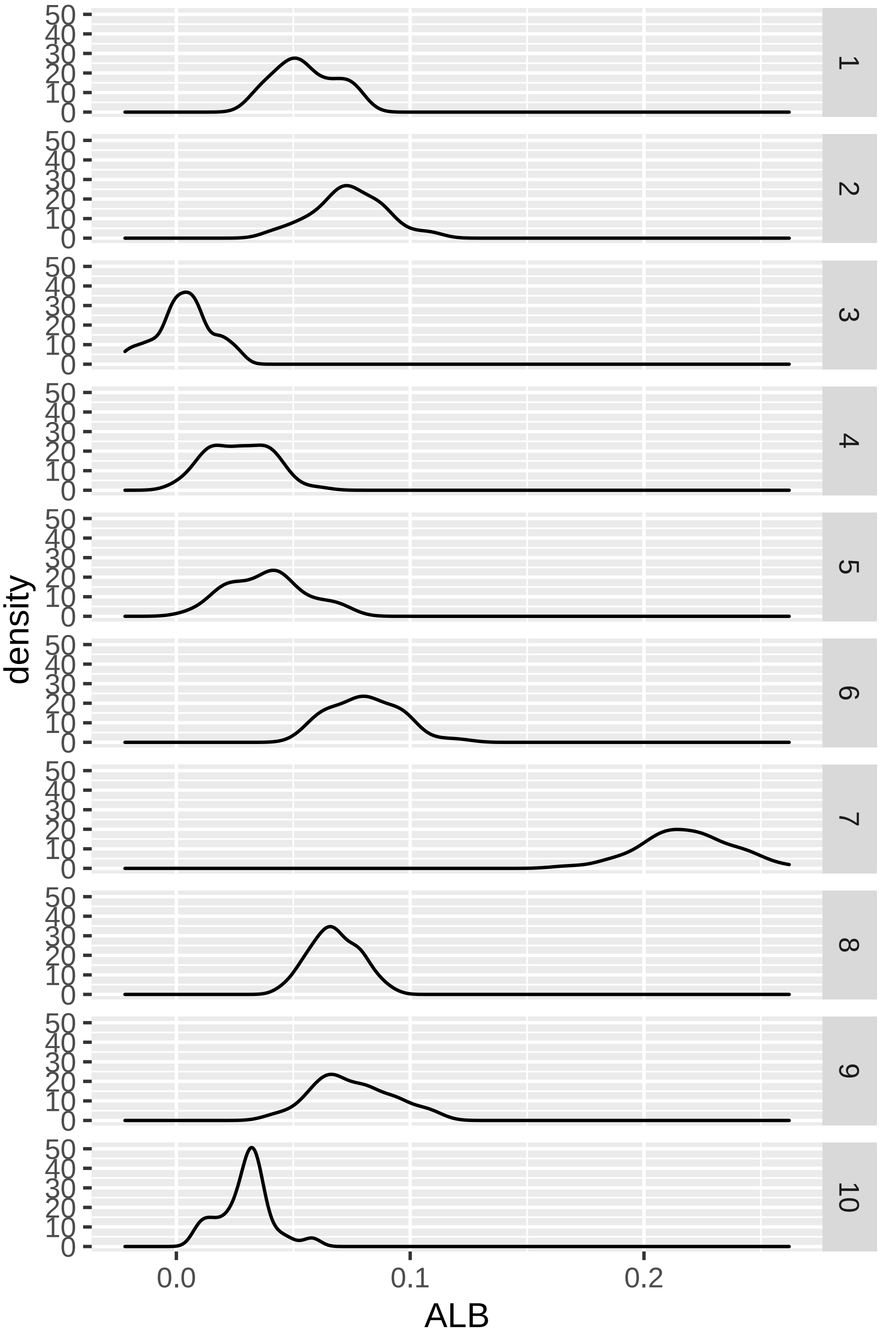

Figure 1. Distribution of ALB under various alternative hypotheses.

We feel that ALB has potential for screening variables in a binary classification problem. Since ALB is negative with high probability under H_0, we feel that 0 is a nicely interpretable cutoff for variable inclusion. However, we leave this topic for future research.

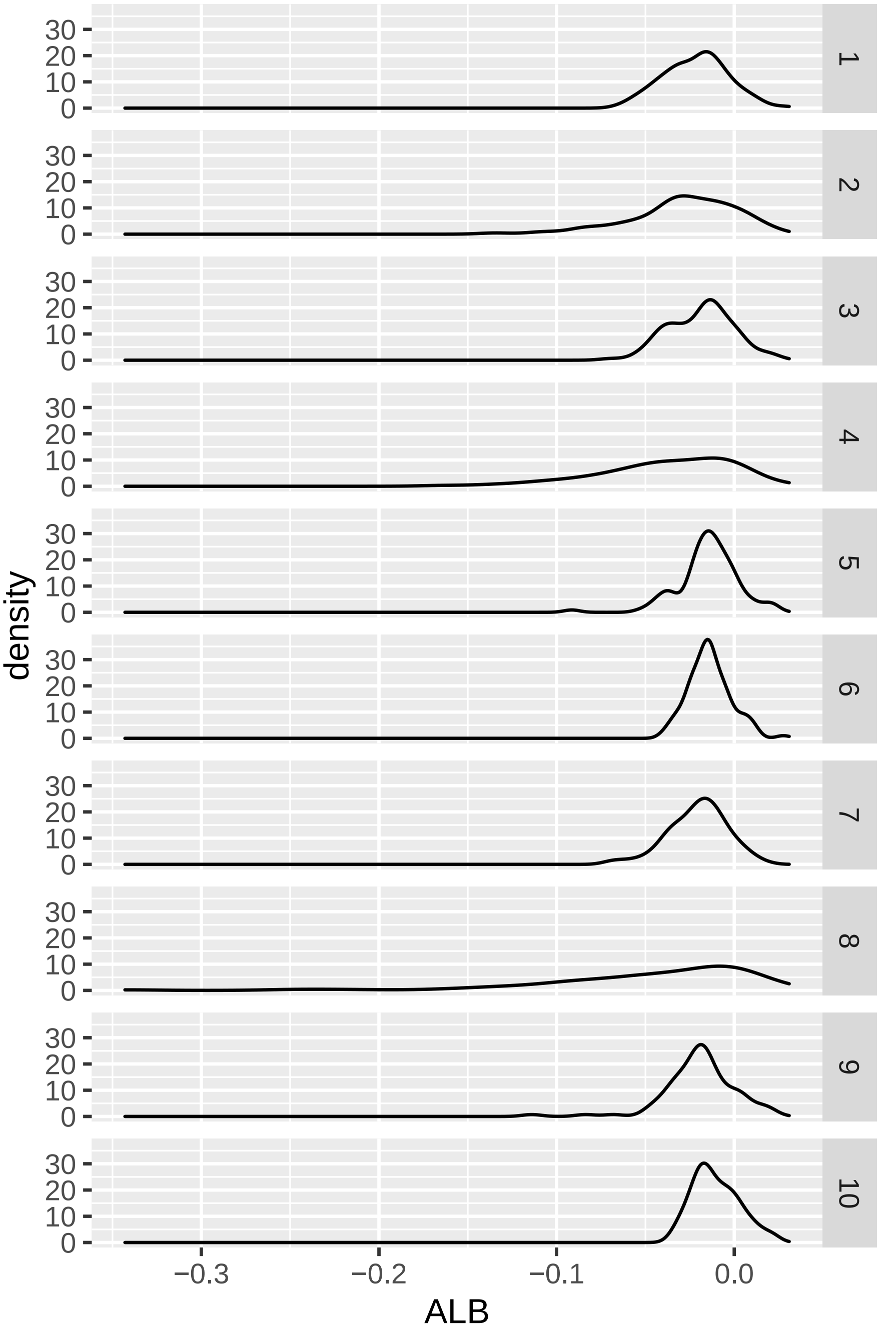

Figure 2. Distribution of ALB under various null hypotheses.

3. Simulations

We perform a small simulation study to investigate the size and power of our test. To explore the effect of the number of permutations, we generate 500 pairs of data sets, with one data set being a random sample of size $m = 50$ from a standard normal distribution, and the other a random sample of size $n = 50$ from a normal distribution with mean 0 and standard deviation 2. For each of the 500 pairs of data sets, the 95th percentile of ALBs is approximated using a range of different numbers (N) of permutations starting at 100 and increasing by a factor of 1.5 up to 3845. Results are indicated by the boxplots in Figure 3. The percentiles are centered at approximately the same value for all N. Not surprisingly, the variability of the percentiles becomes smaller as N increases. This implies a certain amount of mismatch between percentiles at $N = 3845$ and those at smaller N.

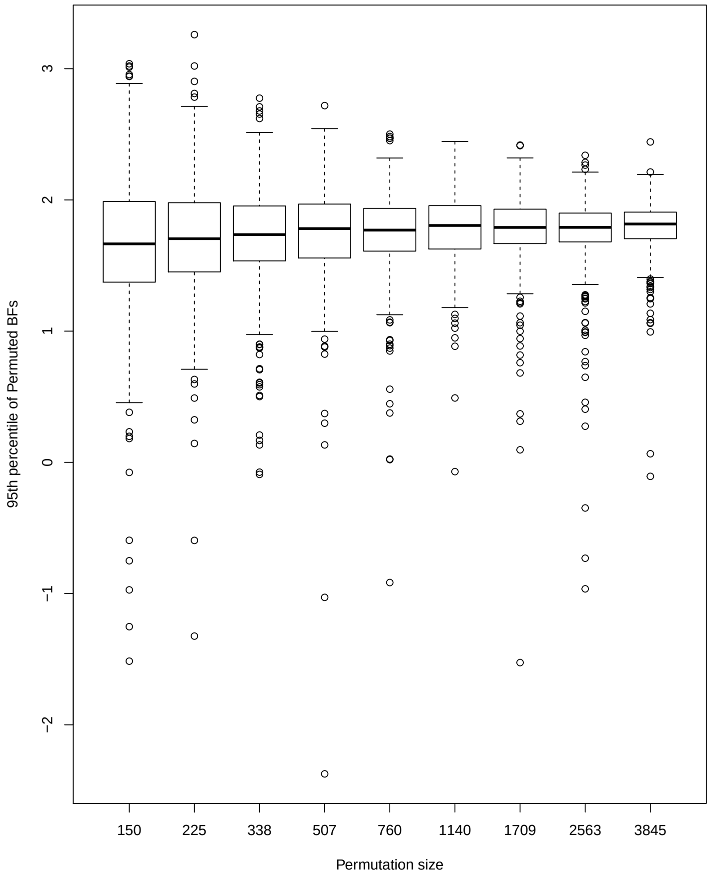

Figure 3. Effect of number of permutations on the 95th percentile of permutation distributions.

The consequence of the mismatch just alluded to can be investigated by determining the true conditional and unconditional levels of tests based on small N. For the null case, two data sets, each of size 50, are generated from a common normal distribution. Since the distribution of ALB is invariant to location and scale in the null case, we use a standard normal without loss of generality. For each pair of data sets, the data are randomly permuted 338 times, which leads to 338 values of ALB. A second set of 3845 permutations is then performed, leading to 3845 more values of ALB. The proportion of ALBs from the second set that exceed the 95th percentile of the ALBs formed from the first set is then determined. This proportion is approximately equal to the conditional level of the test based on 338 permutations. This same procedure is used for each of 500 data sets, and the resulting distribution of approximate levels is shown in Figure 4.

The histogram is centered near 0.05, and 87% of the conditional levels are between 0.03 and 0.07. Furthermore, an approximation to the unconditional level is $\sum_{i=1}^{500} \hat{\alpha}_i/500 = 0.053$, where $\hat{\alpha}_i$ is the approximate conditional level for the ith data set, $i = 1, \ldots, 500$. Based on these results, use of only 338 permutations is arguably adequate.

The same experiment is repeated except now the two data sets are drawn from different distributions, a standard normal and a normal with mean 0 and standard deviation 2. Results from this experiment are given in Figure 5.

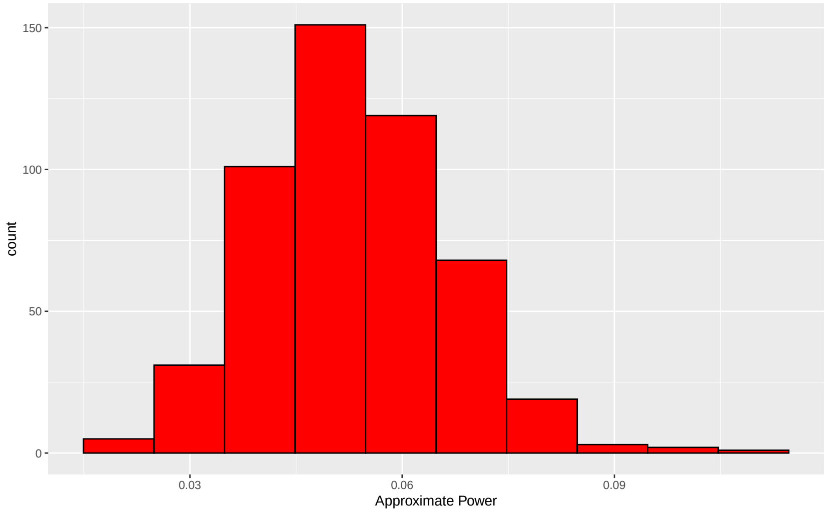

Figure 4. Distribution of approximate conditional levels of permutation tests under the null hypothesis. Each conditional level is the proportion of 3845 ALBs from permuted data sets that exceed the 95th percentile of ALBs formed from 338 permuted data sets. Results are based on 500 replications in each of which both distributions are standard normal.

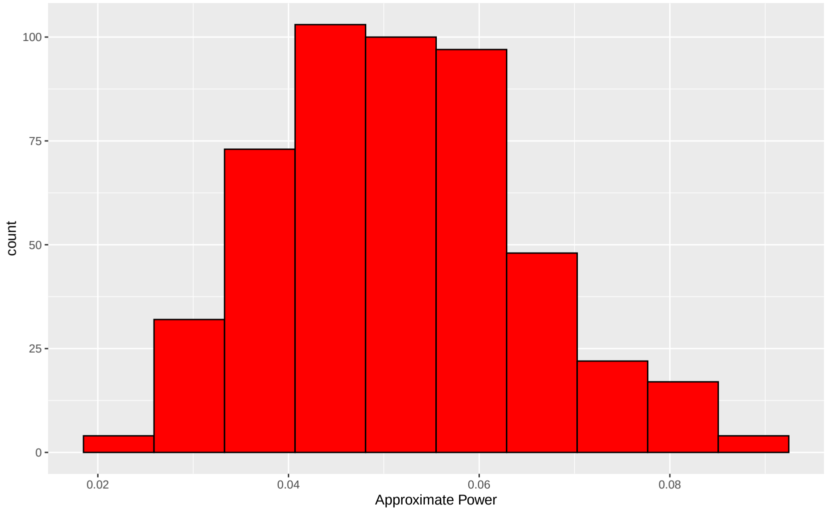

Figure 5. Distribution of approximate conditional levels of permutation tests under an alternative hypothesis. Each conditional level is the proportion of 3845 ALBs from permuted data sets that exceed the 95th percentile of ALBs formed from 338 permuted data sets. Results are based on 500 replications in each of which one distribution is standard normal and the other is normal with mean 0 and standard deviation 2.

As in the null case, the conditional levels based on the use of 338 permutations are quite good. Eighty-eight percent of the levels are between 0.03 and 0.07, and the approximate unconditional level is 0.051.

The proportion of ALBs from permuted data sets that are larger than the ALB computed from the original data provides a p-value. The p-values obtained with our method (based on 3845 permutations) are compared to the p-values obtained with the Kolmogorov–Smirnov test and Bowman's two-sample test. Results are summarized in Figures 6 and 7. In 98% of the replications the K-S p-value was larger than the ALB p-value, and in 57% of the cases the Bowman p-value was equal to or larger than the ALB p-value. These results suggest that in this case our test has much better power than that of the Kolmogorov–Smirnov test and power at least comparable to that of Bowman's test.

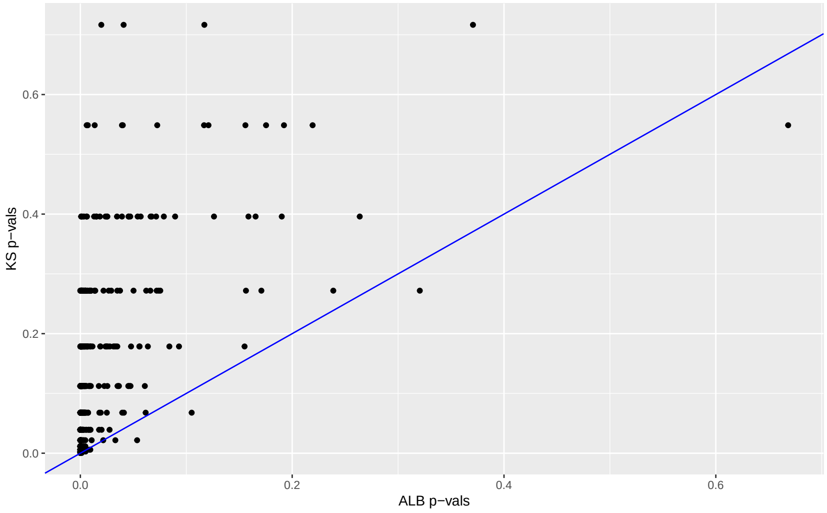

Figure 6. Kolmogorov–Smirnov p-values versus ALB p-values. Results are based on 500 data sets in each of which one distribution is standard normal and the other is normal with mean 0 and standard deviation 2. The ALB p-value is less than the KS-test p-value in 98% of cases. There are only 183 p-values from the KS-test that are less than 0.05.

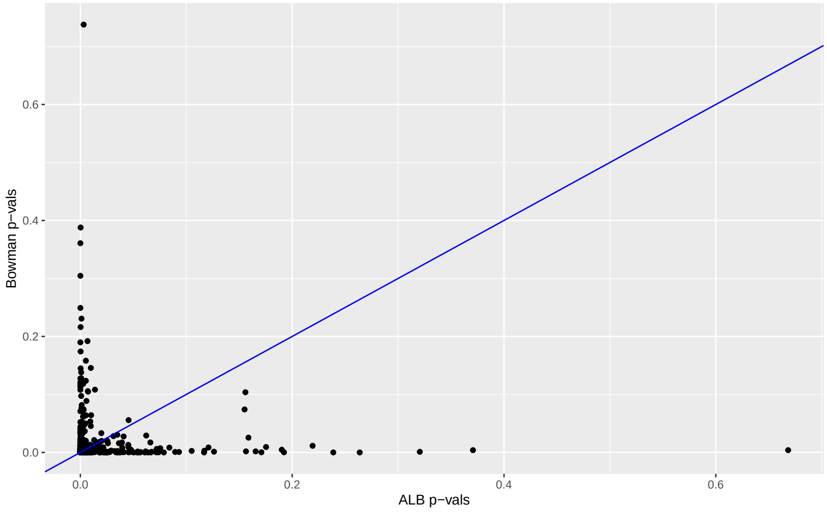

Figure 7. Bowman p-values versus ALB p-values. Results are based on 500 data sets in each of which one distribution is standard normal and the other is normal with mean 0 and standard deviation 2. The number of p-values less than 0.05 for Bowman's test and the ALB test are 454 and 458, respectively. The ALB p-value is less than, more than and equal to the Bowman p-value in 49%, 43% and 8% of cases, respectively.

4. A Bivariate Extension of the Two-Sample Test and Application to Connectionist Bench Data

Our method can be extended to the bivariate case by using a bivariate kernel density estimate. Assume now that $\mathbf{X} = (X_1, ..., X_m)$ are independent and identically distributed from density f and $\mathbf{Y} = (Y_1, ..., Y_m)$ are independent and identically distributed from g, where X_i and Y_j are each bivariate observations, $i = 1, \ldots, m$, $j = 1, \ldots, n$.

A product kernel K will be used, i.e., the bivariate kernel K is the product of two univariate kernels. For k arbitrary bivariate observations $\mathbf{U} = (U_1, \ldots, U_k)$, $U_i = (U_{i1}, U_{i2})$, $i = 1, \ldots, k$, and $u = (u_1, u_2)$, the kernel estimate is defined by

$$\hat{f}_K(u|h, \mathbf{U}) = \frac{1}{kh_1h_2}\sum_{i=1}^{k} K\left(\frac{u_1 - U_{i1}}{h_1}\right)K\left(\frac{u_2 - U_{i2}}{h_2}\right),$$

where $-\infty < u_1 < \infty$, $-\infty < u_2 < \infty$ and $h = (h_1, h_2)$ is a two-vector of (positive) bandwidths.

We will use the same sort of notation as before, i.e., $Z_i = X_i$, $i = 1, \dots, m$, $Z_i = Y_{i-m}$, $i = m+1, \dots, m+n$, $\mathbf{Z} = (Z_1, \dots, Z_{m+n})$ and $\mathbf{Z}_i$ is the object $\mathbf{Z}$ with all its components except Z_i, $i = 1, \dots, m+n$. In this case the ith Bayes factor is defined as

$$B_i = \frac{\int_0^\infty \int_0^\infty \pi(h_1, h_2) \hat{f}_K(Z_i | h, \mathbf{X}_i)\, dh_1 dh_2}{\int_0^\infty \int_0^\infty \pi(h_1, h_2) \hat{f}_K(Z_i | h, \mathbf{Z}_i) dh_1 dh_2}, \quad i = 1, \dots, m,$$

and similarly for $i = m+1, \dots, m+n$. As before the test statistic is $ALB = \sum_{i=1}^{m+n} \log B_i / (m+n)$.

This form may seem daunting, but reduces to a more familiar form if we take $\pi(h_1, h_2) = \pi_0(h_1/b_1)\pi_0(h_2/b_2)/(b_1 b_2)$. In this case, proceeding exactly as in Section 2, B_i has the form

$$B_i = \frac{\hat{f}_L(Z_i | b, \mathbf{X}_i)}{\hat{f}_L(Z_i | b, \mathbf{Z}_i)}, \quad i = 1, \dots, m,$$

and similarly for $i = m+1, \dots, m+n$, where $b = (b_1, b_2)$ and L is defined by (2).

We will analyze a subset of the connectionist bench data, which consist of measurements obtained after bouncing sonar waves off of either rocks or metal cylinders. The data may be found at the UCI Machine Learning repository, Ref. [8]. There are 60 variables in the data set, with $m = 111$ and $n = 97$ measurements of each variable for the metal cylinders and rocks, respectively. Variable numbers (1 to 60) correspond to increasing aspect angles at which signals are bounced off of either metal or rock, and each of the 60 numbers is an amount of energy within a particular frequency band, integrated over a certain period of time. We will apply our test to see if the first two variables (corresponding to the smallest aspect angles) have a different distribution for rocks than they do for metal cylinders. In our analysis K is taken to be ϕ, the standard normal density, and π_0 to be of the form (A1). In this event L is a t-density with ν degrees of freedom. We will use $\nu = 3$, leading to a fairly heavy-tailed kernel, which is desirable for reasons discussed previously.

The data for each variable are inherently between 0 and 1, and bivariate kernel estimates display boundary effects along the lines $x = 0$ and $y = 0$, with the largest bias near the origin. We therefore use a reflection technique to reduce bias along these two lines. Suppose one has k observations $(x_1, y_1), \dots, (x_k, y_k)$ on the unit square. Each observation (x_i, y_i) is reflected to create three new observations: $(x_i, -y_i)$, $(-x_i, -y_i)$ and $(-x_i, y_i)$, $i = 1, \dots, k$. One then simply computes, at points in the unit square, a standard kernel density estimate from the data set of size $4k$, and multiplies it by 4 to ensure integration to 1. The value of ALB is computed as described previously except that each leave-out estimate leaves out four values: the observation at which the estimate is evaluated plus its three reflected versions. In this way the kde is constructed from data that are independent of the value at which the kde is evaluated.

Kernel density estimates for variables 1 and 2 in the form of heat maps are shown in Figures 8 and 9, and contours of the estimates are given in Figure 10. The latter figure suggests that the distributions for metal cylinders and rock are different. The value of ALB turned out to be 0.013, and an approximate p-value based on 10,000 permuted data sets was 0.0076. So, there is strong evidence of a difference between the rock and metal bivariate distributions. Interestingly, the percentage of negative ALBs among the 10,000 permutations was 0.9785. A kernel density estimate based on the 10,000 values of ALB^* is shown in Figure 11.

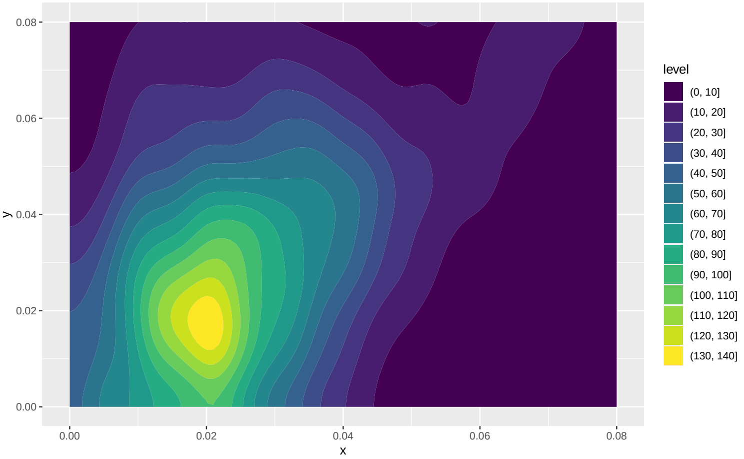

Figure 8. A heat map of the first two variables for the signals bounced off the metal cylinder. Variables x and y correspond to the smallest and next to smallest aspect angles, respectively.

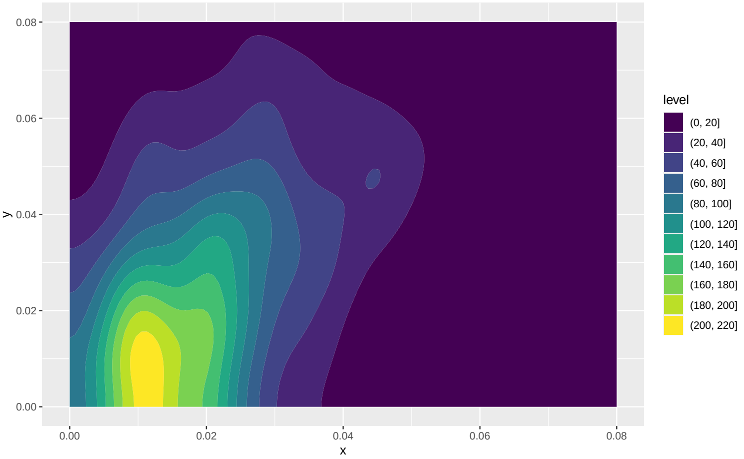

Figure 9. A heat map of the first two variables for the signals bounced off the rock object. Variables x and y are as defined in Figure 8.

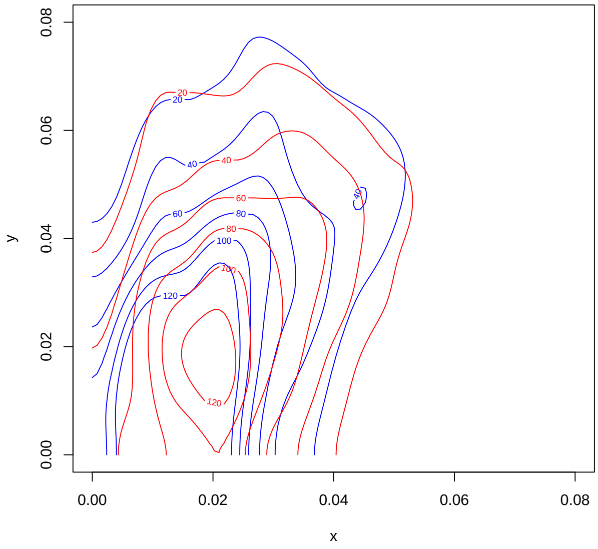

Figure 10. Contour plots of the first two variables of both rock and cylinder objects. The blue contours correspond to the rock measurements and red to the cylinder measurements. Variables x and y are as defined in Figure 8.

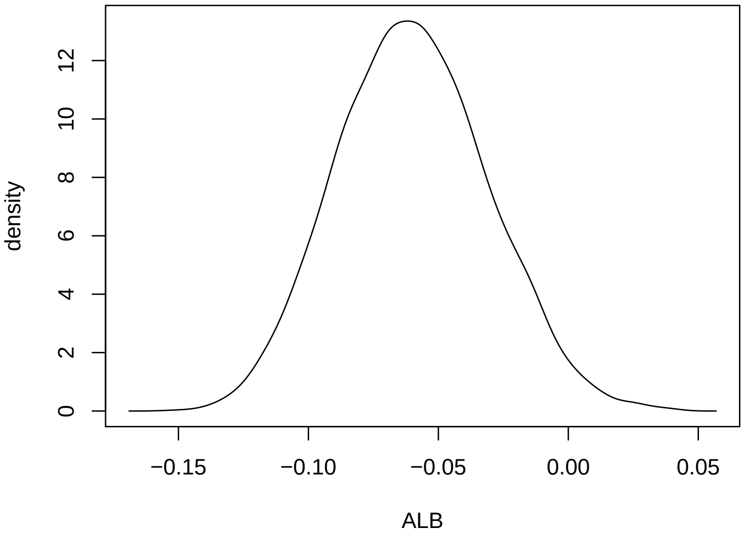

Figure 11. A kernel density estimate computed using 10,000 values of ALB from permuted data sets. The value of ALB for the original data set was 0.013.

5. Conclusions and Future Work

We have proposed a new nonparametric test of the null hypothesis that two densities are equal. An attractive property of the test is that its critical values are defined by a permutation distribution, allaying essentially any concern about test validity. The fact that the statistic is an average of log-Bayes factors leads to another attractive property: a critical value of 0 leads to a test with type I error probability tending to 0 with sample size. A simulation study showed the new test to have much better power than the Kolmogorov–Smirnov test in a case where the two densities differed with respect to scale. An application to connectionist data illustrated the usefulness of our methodology for bivariate data.

Future work includes efforts to increase the speed of computing the test statistic and its permutation distribution, especially for large data sets. We are also interested in applying the new test to the problem of screening variables prior to performing binary classification. A common method of doing so is to compute a two-sample test statistic for each variable, and to then select variables whose statistics exceed some threshold. An inherent problem in this approach is objectively choosing a threshold. Results of the current paper suggest that 0 would be a natural and effective threshold for variable screening.

Author Contributions: Conceptualization, N.M. and J.D.H.; Methodology, N.M. and J.D.H.; Software, N.M. and J.D.H.; Investigation, N.M. and J.D.H.; Writing, N.M. and J.D.H. All authors have read and agreed to the published version of the manuscript.

Funding: The APC was funded by the Department of Statistics, Texas A&M University.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Appendix A.1. Relationship of K and L

By far the most popular choice of kernel in practice is the Gaussian kernel, $K(x) = \phi(x)$, $-\infty < x < \infty$, where ϕ is the standard normal density. For $\nu > 0$, define

$$\pi_0(u) = \frac{2(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} u^{-(\nu+1)} \exp\left(-\frac{\nu}{2u^2}\right), \quad u > 0. \tag{A1}$$

If one takes K to be the the standard normal kernel and uses prior (A1), then the corresponding kernel L is a t-density with ν degrees of freedom. An interesting aspect of these kernels is that they have heavier tails than those of the Gaussian kernel. This is especially true for the more diffuse, or noninformative priors, i.e., those for which ν is small. (The mean and variance of (A1) exist for $\nu > 2$. At $\nu = 3$, the two are 1.382 and 1.090, respectively, and as $\nu \to \infty$ they converge to 1 and 0).

The fact that the kernel L is more heavy-tailed than K in the previous example is not an isolated phenomenon, as indicated by the following proposition (which is straightforward to prove):

Proposition A1. *If π_0 has support $(0, C)$ with $1 < C \leq \infty$ and the tails of K decay exponentially, then the tails of L are heavier than those of K in that $K(u)/L(u) \to 0$ as $u \to \infty$.*

In principle, many different choices of π_0 and K could produce the same kernel L. Or, one might ask "given kernel K, what prior π_0 would produce a specified L?" When K is Gaussian, the latter question is answered by solving an integral equation. Unfortunately, doing so, at least in a general sense, exceeds our mathematical abilities. In the case where K is uniform, though, an elegant solution exists, as seen in the next section.

Appendix A.2. When K Is Uniform

In the special case where K is uniform on the interval $(-1/2, 1/2)$, it is easy to check that, for all u,

$$L(u) = \int_{2|u|}^{\infty} \alpha^{-1} \pi_0(\alpha)\, d\alpha. \tag{A2}$$

If π_0 has support $(0, \infty)$, then L has support $(-\infty, \infty)$, and hence we see again that averaging kernels with respect to a prior leads to a more heavy-tailed kernel.

Since our statistic ends up being a log-likelihood ratio based on kernel L, an interesting question is "what prior π_0 gives rise to a specified kernel L?" Taking $u \geq 0$, (A2) implies that

$$\pi_0(2u) = -uL'(u). \tag{A3}$$

When L is decreasing on $[0, \infty)$ it follows that π_0 is a density. (Under mild tail conditions on L and assuming that $L'(0+)$ exists finite, it is easy to show using integration by parts that (A3) integrates to 1 on $(0, \infty)$.)

Suppose that a kde is defined using the Hall kernel L_0 and its bandwidth is chosen by likelihood cross-validation. Ref. [5] shows that, in general, this cross-validation bandwidth will be asymptotically optimal in a Kullback–Leibler sense. *In contrast, using cross-validation to choose the bandwidth of a uniform kernel kde will produce a bandwidth that diverges to ∞ as the sample size tends to ∞.*

Using (A3) the prior, shown in Figure A1, that produces L_0 is

$$\pi_0(2u) = L_0(u) \frac{u \log(1 + u)}{1 + u}.$$

This shape for the bandwidth prior could be considered canonical inasmuch as L' will be similarly shaped for kernels that are decreasing on $(0, \infty)$.

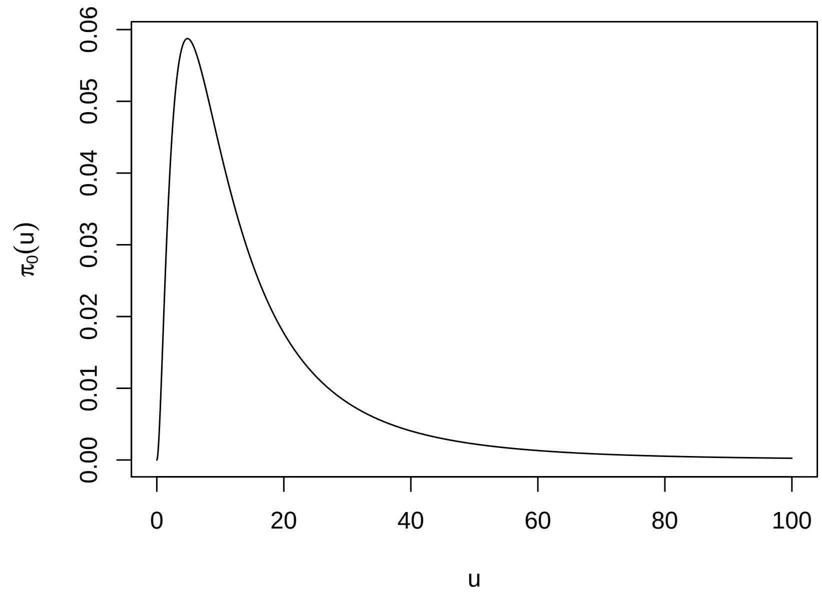

Figure A1. The prior that produces the Hall kernel when K is uniform.

Appendix A.3. Consistency

Here we prove

R1. frequentist consistency of our test, and
R2. $P(ALB < 0) \to 1$ as $m, n \to \infty$.

Our proof uses the following assumptions.

A1. Under the null and alternative hypotheses the following integrals exist finite:

$$I_X = \int_{-\infty}^{\infty} f(x) \log f(x)\, dx \quad \text{and} \quad I_Y = \int_{-\infty}^{\infty} g(y) \log g(y)\, dy.$$

When the alternative hypothesis is true, f and g are assumed to be different in the sense that the total variation distance, $\delta(f, g)$, is positive.

A2. The kernel L in ALB (expression (3)) is the Hall kernel, L_0.
A3. The combined data likelihood cross-validation is maximized over an interval of the form $[(m+n)^{-1+\epsilon}, (m+n)^{-\epsilon}]$, where ϵ is an arbitrarily small positive constant. The maximizer of this cross-validation is denoted $\hat{b}_{m+n}$.
A4. The ratio $m/(m+n)$ tends to ρ, $0 < \rho < 1$, as m, n tend to ∞.
A5. The densities f, g and $\rho f(x) + (1-\rho) g(x)$ satisfy the conditions of [5] that are needed for the asymptotic optimality of a likelihood cross-validation bandwidth.
A6. Under the null hypothesis, let $\ell_k(b)$ be the Kullback–Leibler risk of a kernel density estimate based on sample size k, kernel L_0 and bandwidth b. Then ℓ_k satisfies

$$\ell_k(b) = C_V (nb)^{-1+a} + C_B b^4 + o\Big((nb)^{-1+a} + b^4\Big)$$

for positive constants a, C_V and C_B with $0 < a < 1$.

Before proceeding to the proof, remarks about assumption A6 are in order. This condition is needed only in proving R2, and represents a subset of the cases studied by [5]. It has been assumed merely to allow a more concise proof of R2, which remains true under more general conditions on ℓ_k.

The critical values of a test with fixed size $\alpha > 0$ will tend to 0 as m, n tend to ∞ so long as ALB tends to 0 in probability under the null hypothesis. Therefore, the power of the test will tend to 1 if we can show that ALB tends to a positive constant under the alternative. Our proof of consistency thus boils down to showing that, as m, n tend to ∞,

ALB converges in probability to 0 and a positive number under the null and alternative hypotheses, respectively.

For data $\mathbf{U} = (U_1, \dots, U_k)$, define

$$CV(b|\mathbf{U}) = \frac{1}{k}\sum_{i=1}^{k} \log(\hat{f}_L(U_i|b, \mathbf{U}^i)), \quad b > 0.$$

The statistic ALB may then be written

$$ALB = \left(\frac{m}{m+n}\right)CV(\hat{b}|\mathbf{X}) + \left(\frac{n}{m+n}\right)CV(\hat{b}|\mathbf{Y}) - CV(\hat{b}|\mathbf{Z}),$$

where $\hat{b}$ maximizes $CV(b|\mathbf{Z})$ for $b \in [(m+n)^{-1+\epsilon}, (m+n)^{-\epsilon}]$.

Now suppose that $\mathbf{U}$ is a random sample from density d, $\ell_k(b)$ is the expectation of the Kullback–Leibler loss of $\hat{f}_L(\cdot|b, \mathbf{U})$ and define

$$Q(k) = \frac{1}{k}\sum_{i=1}^{k} \log d(X_i) - \int d(x) \log d(x)\, dx,$$

where $\int d(x) \log d(x)\, dx$ exists finite. Then if d satisfies the conditions of [5] and $k \to \infty$,

$$CV(b|\mathbf{U}) = \int d(x) \log d(x)\, dx - \ell_k(b) + Q(k) + o_p(\ell_k(b)) \tag{A4}$$

uniformly in $b \in [k^{-1+\epsilon}, k^{-\epsilon}]$, where ϵ is arbitrarily small. By the strong law of large numbers $Q(k)$ converges to 0 in probability. Furthermore, $\max_{b\in[k^{-1+\epsilon},k^{-\epsilon}]} \ell_k(b)$ tends to 0 as $k \to \infty$. If the maximizer $\tilde{b}$ of $CV(b|\mathbf{U})$ is in $[k^{-1+\epsilon}, k^{-\epsilon}]$ it therefore follows that $CV(\tilde{b}|\mathbf{U})$ converges in probability to $\int d(x) \log d(x)\, dx$ as $k \to \infty$.

In the null case, (A4) implies that

$$\left(\frac{m}{m+n}\right)CV(b|\mathbf{X}) + \left(\frac{n}{m+n}\right)CV(b|\mathbf{Y}) - CV(b|\mathbf{Z}) =$$

$$-\left[\left(\frac{m}{m+n}\right)\ell_m(b) + \left(\frac{n}{m+n}\right)\ell_n(b)\right] + \ell_{m+n}(b) + o_p(\ell_m(b)), \tag{A5}$$

uniformly in $b \in [(m+n)^{-1+\epsilon}, (m+n)^{-\epsilon}]$, where we have used all of A1–A5. Since $\hat{b}_{m+n} \in [(m+n)^{-1+\epsilon}, (m+n)^{-\epsilon}]$, (A5) implies that ALB converges to 0 in probability as $m, n \to \infty$, which proves one part of R1.

To prove R2, we first observe that the bias component of $\ell_k(b)$ is free of sample size, and hence the first order term of (A5) is free of bias components. Along with A3 and A6, this implies that

$$\left(\frac{m}{m+n}\right)CV(b|\mathbf{X}) + \left(\frac{n}{m+n}\right)CV(b|\mathbf{Y}) - CV(b|\mathbf{Z}) =$$

$$-C_V((m+n)b)^{-1+a}(\rho^a + (1-\rho)^a - 1) + o_p\Big(((m+n)b)^{-1+a} + b^4\Big), \tag{A6}$$

uniformly in $b \in [(m+n)^{-1+\epsilon}, (m+n)^{-\epsilon}]$. By A5, $\hat{b}_{m+n}$ is asymptotic in probability to b_{m+n}, the minimizer of the Kullback–Leibler risk ℓ_{m+n}. Along with (A6), this implies that

$$\begin{aligned} ALB &= -C_V((m+n)b_{m+n})^{-1+a}(\rho^a + (1-\rho)^a - 1) \\ &\quad + o_p\Big(((m+n)b_{m+n})^{-1+a} + b_{m+n}^4\Big). \end{aligned}$$

By A6, we have

$$b_{m+n} \sim C_0(m+n)^{-(1-a)/(5-a)},$$

where

$$C_0 = \left[\frac{C_V(1-a)}{4C_B}\right]^{1/(5-a)}.$$

Combining the previous results yields

$$\begin{aligned} ALB &= -\left(\frac{C_V}{C_0^{1-a}}\right)(\rho^a + (1-\rho)^a - 1)(m+n)^{-4(1-a)/(5-a)} \\ &\quad + o_p((m+n)^{-4(1-a)/(5-a)}). \end{aligned}$$

Using the fact that $(\rho^a + (1-\rho)^a - 1) > 0$ it now follows that $P(ALB < 0) \to 1$ as $m, n \to \infty$.

Turning to the alternative case, we apply (A4) to conclude that $CV(\hat{b}|\mathbf{X})$, $CV(\hat{b}|\mathbf{Y})$ and $CV(\hat{b}|\mathbf{Z})$ are consistent for $\int f(x)\log f(x)\,dx$, $\int g(x)\log g(x)\,dx$, and $\int f_\rho(x)\log f_\rho(x)\,dx$, respectively, where

$$f_\rho(x) = \rho f(x) + (1-\rho)g(x).$$

It follows that ALB is consistent for $\Delta = \rho KL(f, f_\rho) + (1-\rho)KL(g, f_\rho)$, where $KL(f_1, f_2)$ denotes the Kullback–Leibler divergence between f_1 and f_2. By the Csiszár-Kemperman-Kullback-Pinsker inequality,

$$\begin{aligned} \Delta &\geq \frac{\log e}{2} \cdot \left[\rho\delta(f, f_\rho)^2 + (1-\rho)\delta(g, f_\rho)^2\right] \\ &= \frac{\log e}{2} \cdot \left[\rho(1-\rho)^2\delta(f,g)^2 + (1-\rho)\rho^2\delta(f,g)^2\right] \\ &= \left(\frac{\log e}{2}\right)\rho(1-\rho)\delta(f,g)^2 > 0, \end{aligned}$$

with the last inequality following by assumption. This completes the proof of R1.

References

1. Merchant, N.; Hart, J.; Choi, T. Use of cross-validation Bayes factors to test equality of two densities. *arXiv* **2020**, arXiv:2003.06368.
2. Bowman, A.W.; Azzalini, A. *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*; OUP Oxford: New York, NY, USA, 1997; Volume 18.
3. Baranzano, R. Non-Parametric Kernel Density Estimation-Based Permutation Test: Implementation and Comparisons. Ph.D. Thesis, Uppsala University, Uppsala, Sweden, 2011.
4. Hart, J.D.; Choi, T.; Yi, S. Frequentist nonparametric goodness-of-fit tests via marginal likelihood ratios. *Comput. Stat. Data Anal.* **2016**, *96*, 120–132. [CrossRef]
5. Hall, P. On Kullback-Leibler loss and density estimation. *Ann. Stat.* **1987**, *15*, 1491–1519. [CrossRef]
6. Young, S.G.; Bowman, A.W. Non-parametric analysis of covariance. *Biometrics* **1995**, *51*, 920–931. [CrossRef]
7. Hart, J.D. Use of BayesSim and smoothing to enhance simulation studies. *Open J. Stat.* **2017**, *7*, 153–172. [CrossRef]
8. Dua, D.; Graff, C. UCI Machine Learning Repository. School of Information and Computer Sciences, University of California, Irvine. 2017. Available online: http://archive.ics.uci.edu/ml (accessed on 15 March 2022).