Max–Planck–Institut für biologische Kybernetik
Max Planck Institute for Biological Cybernetics

MAX-PLANCK-GESELLSCHAFT

Technical Report No. 172

Consistent Nonparametric Tests of Independence

Arthur Gretton,[1] László Györfi[2]

July 2009

[1] Empirical Inference Department, MPI, and Machine Learning Department, CMU. Email: arthur.gretton@gmail.com [2] Budapest University of Technology and Economics, H-1521 Stoczek u. 2, Budapest, Hungary. Email: gyorfi@szit.bme.hu

Consistent Nonparametric Tests of Independence

Arthur Gretton, László Györfi

Abstract. Three simple and explicit procedures for testing the independence of two multi-dimensional random variables are described. Two of the associated test statistics (L_1, log-likelihood) are defined when the empirical distribution of the variables is restricted to finite partitions. A third test statistic is defined as a kernel-based independence measure. Two kinds of tests are provided. Distribution-free strong consistent tests are derived on the basis of large deviation bounds on the test statistcs: these tests make almost surely no Type I or Type II error after a random sample size. Asymptotically α-level tests are obtained from the limiting distribution of the test statistics. For the latter tests, the Type I error converges to a fixed non-zero value α, and the Type II error drops to zero, for increasing sample size. All tests reject the null hypothesis of independence if the test statistics become large. The performance of the tests is evaluated experimentally on benchmark data.

1. Introduction

Consider a sample of $\mathbb{R}^d \times \mathbb{R}^{d'}$-valued random vectors $(X_1, Y_1), \ldots, (X_n, Y_n)$ with independent and identically distributed (i.i.d.) pairs defined on the same probability space. The distribution of (X, Y) is denoted by ν, while μ_1 and μ_2 stand for the distributions of X and Y, respectively. We are interested in testing the null hypothesis that X and Y are independent,

$$\mathcal{H}_0 : \nu = \mu_1 \times \mu_2, \tag{1}$$

while making minimal assumptions regarding the distribution.

We consider two main approaches to independence testing. The first is to partition the underlying space, and to evaluate the test statistic on the resulting discrete empirical measures. Consistency of the test must then be verified as the partition is refined for increasing sample size. Previous multivariate hypothesis tests in this framework, using the L_1 divergence measure, include homogeneity tests (to determine whether two random variables have the same distribution), by Biau and Györfi (2005); and goodness-of-fit tests (for whether a random variable has a particular distribution), by Györfi and van der Meulen (1990); Beirlant et al. (1994). The log-likelihood has also been employed on discretised spaces as a statistic for goodness-of-fit testing, by Györfi and Vajda (2002). We provide generalizations of both the L_1 and log-likelihood based tests to the problem of testing independence, representing to our knowledge the first application of these techniques to independence testing.

We obtain two kinds of tests for each statistic: first, we derive *strong consistent* tests — meaning that both on $\mathcal{H}_0$ and on its complement the tests make a.s. no error after a random sample size[1] — based on large deviation bounds. While such tests are not common in the classical statistics literature, they are well suited to data analysis from streams, where we receive a sequence of observations rather than a sample of fixed size, and must return the best possible decision at each time using only current and past observations. Our strong consistent tests are *distribution-free*, meaning they require no conditions on the distribution being tested; and *universal*, meaning the test threshold holds independent of the distribution. Second, we obtain tests based on the asymptotic distribution of the L_1 and log-likelihood statistics, which assume only that ν is nonatomic. Subject to this assumption, the tests are *consistent*: for a given asymptotic error rate on $\mathcal{H}_0$, the probability of error on $\mathcal{H}_1$ drops to zero as the sample size increases. Moreover, the thresholds for the asymptotic tests are distribution-independent. We also present conjectures regarding the form taken by strong consistent and asymptotic tests based on the Pearson χ^2 statistic, using the goodness-of-fit results of Györfi and Vajda (2002) (further related test statistics include the power divergence family of Read and Cressie (1988), although we do not study

[1] In other words, denoting by $\mathbf{P}_0$ (*resp.* $\mathbf{P}_1$) the probability under the null hypothesis (*resp.* under the alternative), we have

$$\mathbf{P}_0\{\text{rejecting } \mathcal{H}_0 \text{ for only finitely many } n\} = 1 \tag{2}$$

and

$$\mathbf{P}_1\{\text{accepting } \mathcal{H}_0 \text{ for only finitely many } n\} = 1. \tag{3}$$

This concept relates to the definition of discernability introduced by Dembo and Peres (1994): two ensembles $\mathcal{H}_0$ and $\mathcal{H}_1$ of probability measures on $\mathbb{R}^k$ are said to be discernible if there exists a sequence $f_n : (\mathbb{R}^k)^n \to \{0, 1\}$ of Borel measurable functions achieving (2) and (3). Thus our test implies discernability of the set $\mathcal{H}_0$ in (1) and the set $\mathcal{H}_1$ of dependent random variables.

them here). We emphasize that our tests are explicit, easy to carry out, and require very few assumptions on the partition sequences.

Our second approach to independence testing is kernel-based. In this case, our test statistic has a number of different interpretations: as an L_2 distance between Parzen window estimates (Rosenblatt, 1975), as a smoothed difference between empirical characteristic functions (Feuerverger, 1993; Kankainen, 1995), or as the Hilbert-Schmidt norm of a cross-covariance operator mapping between functions of the random variables (Gretton et al., 2005a, 2008). Each test differs from the others regarding the conditions on the kernels: the Parzen window statistic requires the kernel bandwidth to decrease with increasing sample size, and has a different limiting distribution to the remaining two statistics; while the Hilbert-Schmidt approach uses a fixed bandwidth, and can be thought of as a generalization of the characteristic function-based test. We provide two new results: a strong consistent test of independence based on a tighter large deviation bound than that of Gretton et al. (2005a), and an empirical comparison of the limiting distributions of the kernel-based statistic for fixed and decreasing kernel bandwidth, as used in asymptotic tests.

Additional independence testing approaches also exist in the statistics literature. For $d = d' = 1$, an early nonparametric test for independence, due to Hoeffding (1948); Blum et al. (1961), is based on the notion of differences between the joint distribution function and the product of the marginals. The associated independence test is consistent under appropriate assumptions. Two difficulties arise when using this statistic in a test, however. First, quantiles of the null distribution are difficult to estimate. Second, and more importantly, the quality of the empirical distribution function estimates becomes poor as the dimensionality of the spaces $\mathbb{R}^d$ and $\mathbb{R}^{d'}$ increases, which limits the utility of the statistic in a multivariate setting. Further approaches to independence testing can be used when particular assumptions are made on the form of the distributions, for instance that they should exhibit symmetry. We do not address these approaches in the present study.

The current work is built on an earlier presentation by Gretton and Györfi (2008). Compared with this earlier work, the present study contains more detailed proofs of the main theorems, proofs of secondary theorems omitted by Gretton and Györfi (2008) due to space constraints, additional experiments on higher dimensional benchmark data, and an experimental comparison with the bootstrap approach for the L_1 and log-likelihood based tests (a similar comparison for the kernel-based test was made by Gretton et al., 2008).

The paper is organized as follows. Section 2 describes the large deviation and limit distribution properties of the L_1-test statistic. The large deviation result is used to formulate a distribution-free strong consistent test of independence, which rejects the null hypothesis if the test statistic becomes large. The limit distribution is used in an asymptotically α-level test, which is consistent when the distribution is nonatomic. Both a distribution-free strong consistent test and an asymptotically α-level test are presented for the log-likelihood statistic in Section 3. Section 4 contains a review of kernel-based independence statistics, and describes the associated hypothesis tests for both the fixed-bandwidth and variable-bandwidth cases. Finally, a numerical comparison between the tests is given in Section 5.

2. L_1-based statistic

Denote by ν_n, $\mu_{n,1}$ and $\mu_{n,2}$ the empirical measures associated with the samples $(X_1, Y_1), \ldots, (X_n, Y_n)$, $X_1, \ldots, X_n$, and $Y_1, \ldots, Y_n$, respectively, so that

$$\begin{aligned} \nu_n(A \times B) &= n^{-1}\#\{i : (X_i, Y_i) \in A \times B, i = 1, \ldots, n\}, \\ \mu_{n,1}(A) &= n^{-1}\#\{i : X_i \in A, i = 1, \ldots, n\}, \quad \text{and} \\ \mu_{n,2}(B) &= n^{-1}\#\{i : Y_i \in B, i = 1, \ldots, n\}, \end{aligned}$$

for any Borel subsets A and B. Given the finite partitions $\mathcal{P}_n = \{A_{n,1}, \ldots, A_{n,m_n}\}$ of $\mathbb{R}^d$ and $Q_n = \{B_{n,1}, \ldots, B_{n,m'_n}\}$ of $\mathbb{R}^{d'}$, we define the L_1 test statistic comparing ν_n and $\mu_{n,1} \times \mu_{n,2}$ as

$$L_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) = \sum_{A \in \mathcal{P}_n} \sum_{B \in Q_n} |\nu_n(A \times B) - \mu_{n,1}(A) \cdot \mu_{n,2}(B)|.$$

In the following two sections, we derive the large deviation and limit distribution properties of this L_1 statistic, and the associated independence tests.

2.1 Strongly consistent test

For testing a simple hypothesis versus a composite alternative, Györfi and van der Meulen (1990) introduced a related goodness of fit test statistic L_n defined as

$$L_n(\mu_{n,1}, \mu_1) = \sum_{A \in \mathcal{P}_n} |\mu_{n,1}(A) - \mu_1(A)|.$$

Beirlant, Devroye, Györfi, and Vajda (2001), and Biau and Györfi (2005) proved that, for all $0 < \varepsilon$,

$$\mathbf{P}\{L_n(\mu_{n,1}, \mu_1) > \varepsilon\} \le 2^{m_n} e^{-n\varepsilon^2/2}. \tag{4}$$

We now describe a similar result for our L_1 independence statistic.

Theorem 1 *Under $\mathcal{H}_0$, for all $0 < \varepsilon_1$, $0 < \varepsilon_2$ and $0 < \varepsilon_3$,*

$$\mathbf{P}\{L_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) > \varepsilon_1 + \varepsilon_2 + \varepsilon_3\} \le 2^{m_n \cdot m'_n} e^{-n\varepsilon_1^2/2} + 2^{m_n} e^{-n\varepsilon_2^2/2} + 2^{m'_n} e^{-n\varepsilon_3^2/2}.$$

Proof We bound $L_n(\nu_n, \mu_{n,1} \times \mu_{n,2})$ according to

$$\begin{aligned} L_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) &= \sum_{A \in \mathcal{P}_n} \sum_{B \in Q_n} |\nu_n(A \times B) - \mu_{n,1}(A) \cdot \mu_{n,2}(B)| \\ &\le \sum_{A \in \mathcal{P}_n} \sum_{B \in Q_n} |\nu_n(A \times B) - \nu(A \times B)| \\ &\quad + \sum_{A \in \mathcal{P}_n} \sum_{B \in Q_n} |\nu(A \times B) - \mu_1(A) \cdot \mu_2(B)| \\ &\quad + \sum_{A \in \mathcal{P}_n} \sum_{B \in Q_n} |\mu_1(A) \cdot \mu_2(B) - \mu_{n,1}(A) \cdot \mu_{n,2}(B)|. \end{aligned}$$

Under the null hypothesis $\mathcal{H}_0$, we have that

$$\sum_{A\in\mathcal{P}_n}\sum_{B\in Q_n}|\nu(A\times B)-\mu_1(A)\cdot\mu_2(B)|=0.$$

Moreover

$$\begin{aligned}&\sum_{A\in\mathcal{P}_n}\sum_{B\in Q_n}|\mu_1(A)\cdot\mu_2(B)-\mu_{n,1}(A)\cdot\mu_{n,2}(B)|\\ \le\ &\sum_{A\in\mathcal{P}_n}\sum_{B\in Q_n}|\mu_1(A)\cdot\mu_2(B)-\mu_1(A)\cdot\mu_{n,2}(B)|\\ &+\sum_{A\in\mathcal{P}_n}\sum_{B\in Q_n}|\mu_1(A)\cdot\mu_{n,2}(B)-\mu_{n,1}(A)\cdot\mu_{n,2}(B)|\\ =\ &\sum_{B\in Q_n}|\mu_2(B)-\mu_{n,2}(B)|+\sum_{A\in\mathcal{P}_n}|\mu_1(A)-\mu_{n,1}(A)|\\ =\ &L_n(\mu_{n,1},\mu_1)+L_n(\mu_{n,2},\mu_2).\end{aligned}$$

Thus, (4) implies

$$\begin{aligned}&\mathbf{P}\{L_n(\nu_n,\mu_{n,1}\times\mu_{n,2})>\varepsilon_1+\varepsilon_2+\varepsilon_3\}\\ \le\ &\mathbf{P}\left\{L_n(\nu_n,\nu)>\varepsilon_1\right\}+\mathbf{P}\left\{L_n(\mu_{n,1},\mu_1)>\varepsilon_2\right\}+\mathbf{P}\left\{L_n(\mu_{n,2},\mu_2)>\varepsilon_3\right\}\\ \le\ &2^{m_n\cdot m'_n}e^{-n\varepsilon_1^2/2}+2^{m_n}e^{-n\varepsilon_2^2/2}+2^{m'_n}e^{-n\varepsilon_3^2/2}.\end{aligned}$$

■

Theorem 1 yields a strong consistent test of independence, which rejects the null hypothesis if $L_n(\nu_n,\mu_{n,1}\times\mu_{n,2})$ becomes large. The test is distribution-free, i.e., the probability distributions ν, μ_1 and μ_2 are completely arbitrary; and the threshold is universal, i.e., it does not depend on the distribution.

Corollary 2 *Consider the test which rejects $\mathcal{H}_0$ when*

$$L_n(\nu_n,\mu_{n,1}\times\mu_{n,2})>c_1\left(\sqrt{\frac{m_nm'_n}{n}}+\sqrt{\frac{m_n}{n}}+\sqrt{\frac{m'_n}{n}}\right)\approx c_1\sqrt{\frac{m_nm'_n}{n}},$$

where

$$c_1>\sqrt{2\ln 2}\approx 1.177. \tag{5}$$

Assume that conditions

$$\lim_{n\to\infty}\frac{m_nm'_n}{n}=0, \tag{6}$$

and

$$\lim_{n\to\infty}\frac{m_n}{\ln n}=\infty,\qquad \lim_{n\to\infty}\frac{m'_n}{\ln n}=\infty, \tag{7}$$

are satisfied. Then under $\mathcal{H}_0$, the test makes a.s. no error after a random sample size. Moreover, if

$$\nu\ne\mu_1\times\mu_2,$$

and for any sphere S centered at the origin,

$$\lim_{n\to\infty} \max_{A\in\mathcal{P}_n,\, A\cap S\neq 0} \operatorname{diam}(A) = 0 \tag{8}$$

and

$$\lim_{n\to\infty} \max_{B\in\mathcal{Q}_n,\, B\cap S\neq 0} \operatorname{diam}(B) = 0, \tag{9}$$

then after a random sample size the test makes a.s. no error.

Proof Under $\mathcal{H}_0$, we obtain from Theorem 1 a non-asymptotic bound for the tail of the distribution of $L_n(\nu_n, \mu_{n,1}\times\mu_{n,2})$, namely

$$\begin{aligned}
&\mathbf{P}\left\{ L_n(\nu_n, \mu_{n,1}\times\mu_{n,2}) > c_1 \left(\sqrt{\frac{m_n m_n'}{n}} + \sqrt{\frac{m_n}{n}} + \sqrt{\frac{m_n'}{n}} \right) \right\} \\
\le\ & 2^{m_n m_n'} e^{-c_1^2 m_n m_n'/2} + 2^{m_n} e^{-c_1^2 m_n/2} + 2^{m_n'} e^{-c_1^2 m_n'/2} \\
\le\ & e^{-(c_1^2/2-\ln 2) m_n m_n'} + e^{-(c_1^2/2-\ln 2) m_n} + e^{-(c_1^2/2-\ln 2) m_n'}
\end{aligned}$$

as $n\to\infty$. Therefore the conditions (7) imply

$$\sum_{n=1}^{\infty} \mathbf{P}\left\{ L_n(\nu_n, \mu_{n,1}\times\mu_{n,2}) > c_1 \left(\sqrt{\frac{m_n m_n'}{n}} + \sqrt{\frac{m_n}{n}} + \sqrt{\frac{m_n'}{n}} \right) \right\} < \infty,$$

and the proof under the null hypothesis is completed by the Borel-Cantelli lemma.

For the result under the alternative hypothesis, we first apply the triangle inequality

$$\begin{aligned}
L_n(\nu_n, \mu_{n,1}\times\mu_{n,2}) \ \ge\ & \sum_{A\in\mathcal{P}_n} \sum_{B\in\mathcal{Q}_n} |\nu(A\times B) - \mu_1(A)\cdot\mu_2(B)| \\
& - \sum_{A\in\mathcal{P}_n} \sum_{B\in\mathcal{Q}_n} |\nu_n(A\times B) - \nu(A\times B)| \\
& - \sum_{B\in\mathcal{Q}_n} |\mu_2(B) - \mu_{n,2}(B)| \\
& - \sum_{A\in\mathcal{P}_n} |\mu_1(A) - \mu_{n,1}(A)|.
\end{aligned}$$

The condition in (6) implies the three last terms of the right hand side tend to 0 a.s. Moreover, using the technique from Barron, Györfi, and van der Meulen (1992) we can prove that by conditions (8) and (9),

$$\sum_{A\in\mathcal{P}_n} \sum_{B\in\mathcal{Q}_n} |\nu(A\times B) - \mu_1(A)\cdot\mu_2(B)| \to 2\sup_C |\nu(C) - \mu_1\times\mu_2(C)| > 0$$

as $n\to\infty$, where the last supremum is taken over all Borel subsets C of $\mathbb{R}^d\times\mathbb{R}^{d'}$, and therefore

$$\liminf_{n\to\infty} L_n(\nu_n, \mu_{n,1}\times\mu_{n,2}) \ge 2\sup_C |\nu(C) - \mu_1\times\mu_2(C)| > 0 \tag{10}$$

a.s. ■

2.2 Asymptotic α-level test

Beirlant, Györfi, and Lugosi (1994) proved, under conditions

$$\lim_{n\to\infty} m_n = \infty, \qquad \lim_{n\to\infty} \frac{m_n}{n} = 0, \tag{11}$$

and

$$\lim_{n\to\infty} \max_{j=1,\ldots,m_n} \mu_1(A_{nj}) = 0, \tag{12}$$

that

$$\sqrt{n}\left(L_n(\mu_{n,1},\mu_1) - \mathbf{E}\{L_n(\mu_{n,1},\mu_1)\}\right)/\sigma \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

where $\xrightarrow{\mathcal{D}}$ stands for the convergence in distribution and $\sigma^2 = 1 - 2/\pi$. The technique of Beirlant, Györfi, and Lugosi (1994) involves a Poisson representation of the empirical process in conjunction with Bartlett's idea of partial inversion for obtaining characteristic functions of conditional distributions (see Bartlett, 1938). We apply these techniques in Appendix A to derive an asymptotic result for $L_n(\nu_n, \mu_{n,1} \times \mu_{n,2})$.

Theorem 3 *Assume that conditions (6) and*

$$\lim_{n\to\infty} \max_{A\in\mathcal{P}_n} \mu_1(A) = 0, \quad \lim_{n\to\infty} \max_{B\in Q_n} \mu_2(B) = 0, \tag{13}$$

are satisfied. Then, under $\mathcal{H}_0$, there exists a centering sequence $(C_n)_{n\geq 1}$ depending on ν such that

$$\sqrt{n}\left(L_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) - C_n\right)/\sigma \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

where $\sigma^2 = 1 - 2/\pi$.

Theorem 3 yields the asymptotic null distribution of a consistent independence test, which rejects the null hypothesis if $L_n(\nu_n, \mu_{n,1}\times\mu_{n,2})$ becomes large. In contrast to Corollary 2, and because of condition (12), this new test is *not* distribution-free. In particular, the measures μ_1 and μ_2 have to be nonatomic.

Corollary 4 *Let $\alpha \in (0,1)$. Consider the test which rejects $\mathcal{H}_0$ when*

$$\begin{aligned} L_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) \quad &> \quad c_2\sqrt{\frac{m_n m_n'}{n}} + \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(1-\alpha) \\ &\approx \quad c_2\sqrt{\frac{m_n m_n'}{n}}, \end{aligned}$$

where

$$\sigma^2 = 1 - 2/\pi \quad \textit{and} \quad c_2 = \sqrt{2/\pi} \approx 0.798,$$

and Φ denotes the standard normal distribution function. Then, under the conditions of Theorem 3, the test has asymptotic significance level α. Moreover, under the additional conditions (8) and (9), the test is consistent.

Before proceeding to the proof, we examine how the above test differs from that in Corollary 2. In particular, comparing c_2 above with c_1 in (5), both tests behave identically with respect to $\sqrt{m_n m'_n/n}$ for large enough n, but c_2 is smaller.

Proof According to Theorem 3, under $\mathcal{H}_0$,

$$\mathbf{P}\{\sqrt{n}(L_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) - C_n)/\sigma \leq x\} \approx \Phi(x),$$

therefore the error probability with threshold x is

$$\alpha = 1 - \Phi(x).$$

Thus the α-level test rejects the null hypothesis if

$$L_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) > C_n + \frac{\sigma}{\sqrt{n}} \Phi^{-1}(1 - \alpha).$$

As C_n depends on the unknown distribution, we apply an upper bound

$$C_n \leq \sqrt{2/\pi} \sqrt{\frac{m_n m'_n}{n}}$$

(see eq. (29) in Appendix A for the definition of C_n, and eq. (30) for the bound), so decreasing the error probability. ■

3. Log-likelihood statistic

In the literature on goodness-of-fit testing the *I-divergence statistic*, *Kullback-Leibler divergence*, or *log-likelihood statistic*,

$$I_n(\mu_{n,1}, \mu_1) = \sum_{j=1}^{m_n} \mu_{n,1}(A_{n,j}) \log \frac{\mu_{n,1}(A_{n,j})}{\mu_1(A_{n,j})},$$

plays an important role. For testing independence, the corresponding log-likelihood test statistic is defined as

$$I_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) = \sum_{A \in \mathcal{P}_n} \sum_{B \in \mathcal{Q}_n} \nu_n(A \times B) \log \frac{\nu_n(A \times B)}{\mu_{n,1}(A) \cdot \mu_{n,2}(B)}.$$

The large deviation and the limit distribution properties of $I_n(\nu_n, \mu_{n,1} \times \mu_{n,2})$ can be derived from the properties of

$$I_n(\nu_n, \nu) = \sum_{A \in \mathcal{P}_n} \sum_{B \in \mathcal{Q}_n} \nu_n(A \times B) \log \frac{\nu_n(A \times B)}{\nu(A \times B)}.$$

We have that under $\mathcal{H}_0$,

$$\begin{aligned} & I_n(\nu_n,\nu) - I_n(\nu_n,\mu_{n,1}\times\mu_{n,2}) \\ = & \sum_{A\in\mathcal{P}_n}\sum_{B\in Q_n} \nu_n(A\times B)\log\frac{\nu_n(A\times B)}{\nu(A\times B)} \\ & -\sum_{A\in\mathcal{P}_n}\sum_{B\in Q_n} \nu_n(A\times B)\log\frac{\nu_n(A\times B)}{\mu_{n,1}(A)\cdot\mu_{n,2}(B)} \\ = & \sum_{A\in\mathcal{P}_n}\sum_{B\in Q_n} \nu_n(A\times B)\log\frac{\mu_{n,1}(A)\cdot\mu_{n,2}(B)}{\nu(A\times B)} \\ = & \sum_{A\in\mathcal{P}_n}\sum_{B\in Q_n} \nu_n(A\times B)\log\frac{\mu_{n,1}(A)\cdot\mu_{n,2}(B)}{\mu_1(A)\cdot\mu_2(B)}, \end{aligned}$$

therefore

$$\begin{aligned} & I_n(\nu_n,\nu) - I_n(\nu_n,\mu_{n,1}\times\mu_{n,2}) \\ = & \sum_{A\in\mathcal{P}_n}\sum_{B\in Q_n} \nu_n(A\times B)\left(\log\frac{\mu_{n,1}(A)}{\mu_1(A)}+\log\frac{\mu_{n,2}(B)}{\mu_2(B)}\right) \\ = & \sum_{A\in\mathcal{P}_n}\mu_{n,1}(A)\log\frac{\mu_{n,1}(A)}{\mu_1(A)}+\sum_{B\in Q_n}\mu_{n,2}(B)\log\frac{\mu_{n,2}(B)}{\mu_2(B)} \\ = & I_n(\mu_{n,1},\mu_1)+I_n(\mu_{n,1},\mu_1) \\ \geq & 0. \end{aligned}$$

3.1 Strongly consistent test

We refer to Tusnády (1977) and Barron (1989) who first discussed the exponential character of the tails of I_n. Kallenberg (1985), and Quine and Robinson (1985) proved that, for all $\epsilon > 0$,

$$\mathbf{P}\{I_n(\mu_{n,1},\mu_1)>\epsilon\} \leq \binom{n+m_n-1}{m_n-1}e^{-n\epsilon} \leq e^{m_n\log(n+m_n)-n\epsilon}.$$

Note that using an alternative bound due to Barron (1989, eq. (3.5)), we obtain under (11) and (12) that

$$\mathbf{P}\{I_n(\mu_{n,1},\mu_1)>\epsilon\} = e^{-n(\epsilon+o(1))}, \tag{14}$$

such that

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbf{P}\{I_n(\mu_{n,1},\mu_1)>\epsilon\} = -\epsilon.$$

A large deviation based test can be introduced such that the test rejects the independence if

$$I_n(\nu_n,\mu_{n,1}\times\mu_{n,2}) \geq \frac{m_n m_n'(\log(n+m_n m_n')+1)}{n}.$$

Under $\mathcal{H}_0$, we obtain a non-asymptotic bound for the tail of the distribution of $I_n(\nu_n, \mu_{n,1} \times \mu_{n,2})$:

$$\begin{aligned} & \mathbf{P}\left\{ I_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) > \frac{m_n m_n'(\log(n + m_n m_n') + 1)}{n} \right\} \\ \leq\ & \mathbf{P}\left\{ I_n(\nu_n, \nu) > \frac{m_n m_n'(\log(n + m_n m_n') + 1)}{n} \right\} \\ \leq\ & e^{m_n m_n' \log(n+m_n m_n') - n\frac{m_n m_n'(\log(n+m_n m_n')+1)}{n}} \\ =\ & e^{-m_n m_n'}. \end{aligned}$$

Therefore condition (7) implies

$$\sum_{n=1}^{\infty} \mathbf{P}\left\{ I_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) > \frac{m_n m_n'(\log(n + m_n m_n') + 1)}{n} \right\} < \infty,$$

and by the Borel-Cantelli lemma we have strong consistency under the null hypothesis.

Under the alternative hypothesis the proof of strong consistency follows from the inequality, also called Pinsker's inequality, which upper bounds the L_1 error in terms of I-divergence (c.f. Csiszár, 1967; Kemperman, 1969; Kullback, 1967),

$$L_n(\nu_n, \mu_{n,1} \times \mu_{n,2})^2 \leq 2 I_n(\nu_n, \mu_{n,1} \times \mu_{n,2}). \tag{15}$$

Therefore,

$$\begin{aligned} \liminf_{n\to\infty} 2 I_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) &\geq (\liminf_{n\to\infty} L_n(\nu_n, \mu_{n,1} \times \mu_{n,2}))^2 \\ &\geq 4 \sup_C |\nu(C) - \mu_1 \times \mu_2(C)|^2 > 0 \end{aligned}$$

a.s., where the supremum is taken over all Borel subsets C of $\mathbb{R}^d \times \mathbb{R}^{d'}$. In fact, under conditions (8), (9), and

$$I(\nu, \mu_1 \times \mu_2) < \infty,$$

one may get

$$\lim_{n\to\infty} I_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) = I(\nu, \mu_1 \times \mu_2) > 0$$

a.s. (see Barron et al., 1992). Note that due to the form of the universal test threshold, strong consistency under $\mathcal{H}_1$ requires the condition

$$\lim_{n\to\infty} \frac{m_n m_n'}{n} \log(n + m_n m_m') = 0,$$

as compared to (6).

3.2 Asymptotic α-level test

Concerning the limit distribution, Inglot et al. (1990), and Györfi and Vajda (2002) proved that under (11) and (12),

$$\frac{2n I_n(\mu_{n,1}, \mu_1) - m_n}{\sqrt{2m_n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1). \tag{16}$$

This implies that for any real valued x, under the conditions (6) and (13),

$$\begin{aligned}\mathbf{P}\left\{\frac{2nI_n(\nu_n,\mu_{n,1}\times\mu_{n,2})-m_nm'_n}{\sqrt{2m_nm'_n}}\le x\right\} &\le \mathbf{P}\left\{\frac{2nI_n(\nu_n,\nu)-m_nm'_n}{\sqrt{2m_nm'_n}}\le x\right\}\\ &\to \Phi(x),\end{aligned}$$

which results in a test rejecting the independence if

$$\frac{2nI_n(\nu_n,\mu_{n,1}\times\mu_{n,2})-m_nm'_n}{\sqrt{2m_nm'_n}}\ge\Phi^{-1}(1-\alpha),$$

or equivalently

$$I_n(\nu_n,\mu_{n,1}\times\mu_{n,2})\ge\frac{\Phi^{-1}(1-\alpha)\sqrt{2m_nm'_n}+m_nm'_n}{2n}.$$

Note that unlike the L_1 case, the ratio of the strong consistent threshold to the asymptotic threshold increases for increasing n.

4. Kernel-based statistic

We now present a second class of approaches to independence testing, based on a kernel statistic. We can derive this statistic in a number of ways. The most immediate interpretation, introduced by Rosenblatt (1975), defines the statistic as the L_2 distance between the joint density estimate and the product of marginal density estimates. Let K and K' be density functions (called kernels) defined on $\mathbb{R}^d$ and on $\mathbb{R}^{d'}$, respectively. For the bandwidth $h>0$, define

$$K_h(x)=\frac{1}{h^d}K\left(\frac{x}{h}\right)\quad\text{and}\quad K'_h(x)=\frac{1}{h^{d'}}K'\left(\frac{x}{h}\right).$$

The Rosenblatt-Parzen kernel density estimates of the density of (X,Y) and X are respectively

$$f_n(x,y)=\frac{1}{n}\sum_{i=1}^{n}K_h(x-X_i)K'_h(y-Y_i)\ \text{and}\ f_{n,1}(x)=\frac{1}{n}\sum_{i=1}^{n}K_h(x-X_i),\tag{17}$$

with $f_{n,2}(y)$ defined by analogy. Rosenblatt (1975) introduced the kernel-based independence statistic

$$T_n=\int_{\mathbb{R}^d\times\mathbb{R}^{d'}}(f_n(x,y)-f_{n,1}(x)f_{n,2}(y))^2dx\,dy.\tag{18}$$

Alternatively, defining

$$L_h(x)=\int_{\mathbb{R}^d}K_h(u)K_h(x-u)du=\frac{1}{h^d}\int_{\mathbb{R}^d}K(u)K(x-u)du$$

and $L'_h(x)$ by analogy, we may write the kernel test statistic

$$\begin{aligned}T_n=&\tfrac{1}{n^2}\textstyle\sum_{i=1}^n\sum_{j=1}^n L_h(X_i-X_j)L'_h(Y_i-Y_j)\\ &-\tfrac{2}{n^3}\textstyle\sum_{i=1}^n\left(\sum_{j=1}^n L_h(X_i-X_j)\right)\left(\sum_{j=1}^n L'_h(Y_i-Y_j)\right)\\ &+\left(\tfrac{1}{n^2}\textstyle\sum_{i=1}^n\sum_{j=1}^n L_h(X_i-X_j)\right)\left(\tfrac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n L'_h(Y_i-Y_j)\right).\end{aligned}\tag{19}$$

Note that at independence, the expected value of the statistic is not zero, but

$$\mathbf{E}\{T_n\} \;=\; \frac{n-1}{n^2}\left(L_h(0) - \mathbf{E}\{L_h(X_1 - X_2)\}\right)\left(L'_h(0) - \mathbf{E}\{L'_h(Y_1 - Y_2)\}\right) \tag{20}$$

$$\leq\; n^{-1}L_h(0)L'_h(0) \;=\; (nh^d h^{d'})^{-1}\|K\|^2\|K'\|^2. \tag{21}$$

A second interpretation of the above statistic is as a smoothed difference between the joint characteristic function and the product of the marginals (Feuerverger, 1993). The characteristic function and Rosenblatt-Parzen window statistics can be quite similar: in fact, for appropriate smoothing and kernel choices and fixed n, they may be identical (Kankainen, 1995). For increasing n, the main differences between the approaches are that the kernel bandwidth h must decrease in the Rosenblatt test for consistency of the kernel density estimates, and the more restrictive conditions on the Rosenblatt-Parzen test statistic (Rosenblatt, 1975, conditions a.1-a.4).

A further generalization of the statistic is presented by Gretton et al. (2005a, 2008), in terms of covariances between feature mappings of the random variables to reproducing kernel Hilbert spaces (RKHSs). We now briefly review this interpretation, beginning with some necessary terminology and definitions. Let $\mathcal{F}$ be an RKHS, with the continuous feature mapping $\phi(x) \in \mathcal{F}$ for each $x \in \mathbb{R}^d$, such that the inner product between the features is given by the positive definite kernel function $L_h(x, x') := \langle \phi(x), \phi(x') \rangle_{\mathcal{F}}$. Likewise, let $\mathcal{G}$ be a second RKHS on $\mathbb{R}^{d'}$ with kernel $L'_h(\cdot, \cdot)$ and feature map $\psi(y)$. Following Baker (1973); Fukumizu et al. (2004), the cross-covariance operator $C_\nu : \mathcal{G} \to \mathcal{F}$ for the measure ν is defined such that for all $f \in \mathcal{F}$ and $g \in \mathcal{G}$,

$$\langle f, C_\nu g \rangle_{\mathcal{F}} \;=\; \mathbf{E}\left([f(X) - \mathbf{E}(f(X))]\,[g(Y) - \mathbf{E}(g(Y))]\right).$$

The cross-covariance operator can be thought of as a generalisation of a cross-covariance matrix between the (potentially infinite dimensional) feature mappings $\phi(x)$ and $\psi(y)$.

To see how this operator may be used to test independence, we recall the following characterization of independence (see e.g. Jacod and Protter, 2000, Theorem 10.1e):

Theorem 5 *The random variables X and Y are independent if and only if* $\operatorname{cov}(f(X), g(Y)) = 0$ *for any pair (f, g) of bounded, continuous functions.*

While the bounded continuous functions are too rich a class to permit the construction of a covariance-based test statistic on a sample, Fukumizu et al. (2008); Sriperumbudur et al. (2008) show that when $\widetilde{\mathcal{F}}$ is the unit ball in a *characteristic*[2] RKHS $\mathcal{F}$, and $\widetilde{\mathcal{G}}$ the unit ball in a characteristic RKHS $\mathcal{G}$, then

$$\sup_{f \in \widetilde{\mathcal{F}}, g \in \widetilde{\mathcal{G}}} \mathbf{E}\left([f(X) - \mathbf{E}(f(X))]\,[g(Y) - \mathbf{E}(g(Y))]\right) = 0 \qquad \Longleftrightarrow \qquad \nu = \mu_1 \times \mu_2.$$

In other words, the spectral norm of the covariance operator C_ν between characteristic RKHSs is zero only at independence. Rather than the maximum singular value, we may

[2] The reader is referred to (Fukumizu et al., 2008; Sriperumbudur et al., 2008) for conditions under which an RKHS is characteristic. We note here that the Gaussian kernel on $\mathbb{R}^d$ has this property, and provide further discussion below.

use the squared Hilbert-Schmidt norm (the sum of the squared singular values), which has a population expression

$$\begin{aligned}\mathrm{H}(\nu;\mathcal{F},\mathcal{G}) &= \mathbf{E}\{L_h(X_1-X_2)L'_h(Y_1-Y_2)\} - 2\mathbf{E}\left\{\mathbf{E}\{L_h(X_1-X_2)|X_1\}\mathbf{E}\{L_h(Y_1-Y_2)|Y_1\}\right\} \\ &\quad +\mathbf{E}\{L_h(X_1-X_2)\}\mathbf{E}\{L'_h(Y_1-Y_2)\}\end{aligned}$$

(if the expectations exist; see Gretton et al., 2005a, Lemma 1): we call this the Hilbert-Schmidt independence criterion (HSIC).

The test statistic in (19) is then interpreted as a biased empirical estimate of $\mathrm{H}(\nu;\mathcal{F},\mathcal{G})$. Clearly, when K_h and K'_h are continuous and square integrable densities, the induced kernels L_h and L'_h are continuous positive definite RKHS kernels. However, as long as L_h and L'_h are characteristic kernels, then $\mathrm{H}(\nu;\mathcal{F},\mathcal{G}) = 0$ iff X and Y independent. The Gaussian and Laplace kernels are characteristic on $\mathbb{R}^d$ (Fukumizu et al., 2008), and universal kernels (in the sense of Steinwart, 2001) are characteristic on compact domains (Gretton et al., 2005a, Theorem 6). Sriperumbudur et al. (2008) provide a simple necessary and sufficient condition for a translation invariant kernel to be characteristic on $\mathbb{R}^d$: the Fourier spectrum of the kernel must be supported on the entire domain. Note that characteristic kernels need not be inner products of square integrable probability density functions: an example is the kernel

$$L_h(x_1,x_2) = \exp(x_1^T x_2/h)$$

from Steinwart (2001, Section 3, Example 1), which is universal, hence characteristic on compact subsets of $\mathbb{R}^d$. Finally, an appropriate choice of kernels allows testing of dependence in non-Euclidean settings, such as distributions on strings and graphs (Gretton et al., 2008).

4.1 Strongly consistent test

The empirical statistic T_n was previously shown by Gretton et al. (2005a) to converge in probability to its expectation with rate $1/\sqrt{n}$. Given $0 \le L_h(0)L'_h(0) \le 1$, the corresponding result is

$$\mathbf{P}(T_n - \mathbf{E}(T_n) \ge \epsilon^2) \le 3e^{-0.24n\epsilon^4},$$

which follows from the straightforward application of a bound by Hoeffding (1963, p. 25). We now provide a more refined bound which scales better with ϵ, and is thus tighter when the bandwidth h decreases.

We will obtain our results for the semi-statistic

$$\tilde{T}_n = \|f_n(\cdot,\cdot) - \mathbf{E}f_n(\cdot,\cdot)\|^2,$$

since under the null hypothesis,

$$\begin{aligned}\sqrt{T_n} &= \|f_n(\cdot,\cdot) - f_{n,1}(\cdot)f_{n,2}(\cdot)\| \\ &\le \|f_n(\cdot,\cdot) - \mathbf{E}f_n(\cdot,\cdot)\| + \|f_{n,1}(\cdot)f_{n,2}(\cdot) - \mathbf{E}f_{n,1}(\cdot)\mathbf{E}f_{n,2}(\cdot)\| \\ &\le \sqrt{\tilde{T}_n} + \|f_{n,1}(\cdot)(f_{n,2}(\cdot) - \mathbf{E}f_{n,2}(\cdot))\| + \|(f_{n,1}(\cdot) - \mathbf{E}f_{n,1}(\cdot))\mathbf{E}f_{n,2}(\cdot)\| \\ &= \sqrt{\tilde{T}_n} + \|f_{n,1}(\cdot)\|\,\|f_{n,2}(\cdot) - \mathbf{E}f_{n,2}(\cdot)\| + \|f_{n,1}(\cdot) - \mathbf{E}f_{n,1}(\cdot)\|\,\|\mathbf{E}f_{n,2}(\cdot)\| \\ &\approx \sqrt{\tilde{T}_n}.\end{aligned}$$

Theorem 6 *For any $\epsilon > 0$,*

$$\mathbf{P}\left\{\tilde{T}_n \geq \left(\epsilon + \mathbf{E}\left\{\sqrt{\tilde{T}_n}\right\}\right)^2\right\} \leq e^{-n\epsilon^2 \big/ (2L_h(0)L_h'(0))}.$$

Proof We apply the McDiarmid inequality (c.f. McDiarmid, 1989): Let $Z_1, \ldots, Z_n$ be independent random variables taking values in a set A and assume that $f : A^n \to \mathbb{R}$ satisfies

$$\sup_{\substack{z_1,\ldots,z_n,\\ z_i' \in A}} |f(z_1, \ldots, z_n) - f(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n)| \leq c_i,\ 1 \leq i \leq n.$$

Then, for all $\epsilon > 0$,

$$\mathbf{P}\left\{f(Z_1, \ldots, Z_n) - \mathbf{E}f(Z_1, \ldots, Z_n) \geq \epsilon\right\} \leq e^{-2\epsilon^2 \big/ \sum_{i=1}^n c_i^2}.$$

Because of

$$\begin{aligned}
\sqrt{\tilde{T}_n} &= \|f_n(\cdot, \cdot) - \mathbf{E}f_n(\cdot, \cdot)\| \\
&= \|\frac{1}{n}\sum_{i=1}^n K_h(\cdot - X_i)K_h'(\cdot - Y_i) - \mathbf{E}f_n(\cdot, \cdot)\| \\
&\leq \|\frac{1}{n}K_h(\cdot - X_1)K_h'(\cdot - Y_1)\| + \|\frac{1}{n}\sum_{i=2}^n K_h(\cdot - X_i)K_h'(\cdot - Y_i) - \mathbf{E}f_n(\cdot, \cdot)\|
\end{aligned}$$

we can apply McDiarmid inequality with

$$\frac{2}{n}\|K_h(\cdot - X_1)K_h'(\cdot - Y_1)\| = \frac{2}{n}\sqrt{L_h(0)L_h'(0)} =: c_i = c_1,$$

where we note that the c_i are independent of i, and can be replaced by a single c_1. Thus,

$$\begin{aligned}
\mathbf{P}\left\{\sqrt{\tilde{T}_n} - \mathbf{E}\left\{\sqrt{\tilde{T}_n}\right\} \geq \epsilon\right\} &\leq e^{-2\epsilon^2 \big/ \sum_{i=1}^n c_i^2} \\
&= e^{-2\epsilon^2 \big/ (nc_1^2)} \\
&\leq e^{-n\epsilon^2 \big/ (2L_h(0)L_h'(0))}.
\end{aligned}$$

This implies

$$\mathbf{P}\left\{\tilde{T}_n \geq \left(\epsilon + \mathbf{E}\left\{\sqrt{\tilde{T}_n}\right\}\right)^2\right\} \leq e^{-n\epsilon^2 \big/ (2L_h(0)L_h'(0))}.$$

■

From these inequalities we can derive a test of independence. Choose ϵ such that

$$n\epsilon^2 \big/ (2L_h(0)L_h'(0)) = 2\ln n.$$

Because of

$$\mathbf{E}\{\tilde{T}_n\} \approx \mathbf{E}\{T_n\} \leq \frac{L_h(0)L'_h(0)}{n},$$

we choose the threshold

$$\left(\sqrt{\frac{L_h(0)L'_h(0)4\ln n}{n}} + \sqrt{\frac{L_h(0)L'_h(0)}{n}}\right)^2 = \frac{L_h(0)L'_h(0)}{n}(\sqrt{4\ln n} + 1)^2,$$

i.e., we reject the hypothesis of independence if

$$T_n > \frac{\|K\|^2\|K'\|^2}{nh^d h^{d'}}(\sqrt{4\ln n} + 1)^2.$$

It follows from

$$\begin{aligned}
& \mathbf{P}\left\{T_n \geq \frac{L_h(0)L'_h(0)}{n}(\sqrt{4\ln n} + 1)^2\right\} \\
\approx\ & \mathbf{P}\left\{\tilde{T}_n \geq \left(\sqrt{\frac{L_h(0)L'_h(0)4\ln n}{n}} + \sqrt{\frac{L_h(0)L'_h(0)}{n}}\right)^2\right\} \\
\leq\ & \mathbf{P}\left\{\tilde{T}_n \geq \left(\sqrt{\frac{L_h(0)L'_h(0)4\ln n}{n}} + \sqrt{\mathbf{E}\{\tilde{T}_n\}}\right)^2\right\} \\
\leq\ & \mathbf{P}\left\{\tilde{T}_n \geq \left(\sqrt{\frac{L_h(0)L'_h(0)4\ln n}{n}} + \mathbf{E}\left\{\sqrt{\tilde{T}_n}\right\}\right)^2\right\} \\
\leq\ & e^{-2\ln n}
\end{aligned}$$

that this test of independence is strongly consistent.

Under the alternative hypothesis, there are two cases:

- If $h \to 0$ and the density f exists and is square integrable, then

  $$T_n \to \|f - f_1 f_2\|^2 > 0$$

  a.s. The strong consistency is not distribution-free, since ν must have a square integrable density.

- If h is fixed, the strong law of large numbers implies

  $$\begin{aligned}
  T_n \quad \to \quad & \mathbf{E}\{L_h(X_1 - X_2)L'_h(Y_1 - Y_2)\} - 2\mathbf{E}\left\{\mathbf{E}\{L_h(X_1 - X_2)|X_1\}\mathbf{E}\{L_h(Y_1 - Y_2)|Y_1\}\right\} \\
  & \quad + \mathbf{E}\{L_h(X_1 - X_2)\}\mathbf{E}\{L'_h(Y_1 - Y_2)\} \qquad (22) \\
  =: \quad & \mathrm{H}(\nu; \mathcal{F}, \mathcal{G})
  \end{aligned}$$

  If K_h and K'_h are continuous and square integrable densities, the induced kernels L_h and L'_h are continuous positive definite kernels: $\mathrm{H}(\nu; \mathcal{F}, \mathcal{G})$ is then the squared Hilbert-Schmidt norm of the covariance operator for ν. We may replace L_h and L'_h with any *characteristic kernels* (in the sense of Fukumizu et al., 2008; Sriperumbudur et al., 2008), however, and retain the property $\mathrm{H}(\nu; \mathcal{F}, \mathcal{G}) = 0$ iff X and Y independent. In this case, the strong consistency is distribution-free.

4.2 Approximately α-level tests

We now describe the asymptotic limit distribution of the test statistic T_n in (19). We address two cases: first, when the kernel bandwidth decreases, and second, when it remains fixed.

Let us consider the case where $K_h(x)$ and $K'_h(y)$ are intended to be used in a Rosenblatt-Parzen density estimator, as in (17). The corresponding density estimates in T_n are mean square consistent if $h = h_n$ such that

$$h_n \to 0 \quad \text{and} \quad n h_n^d h_n^{d'} \to \infty. \tag{23}$$

Based on the results of Hall (1984); Cotterill and Csörgő (1985); Beirlant and Mason (1995), we expect that, under these consistency conditions,

$$\frac{T_n - \mathbf{E}\{T_n\}}{\sqrt{\operatorname{var}(T_n)}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

We next calculate $\operatorname{var}(T_n) \approx \operatorname{var}(\tilde{T}_n)$. Under the null hypothesis,

$$\begin{aligned}
\tilde{T}_n &= \|f_n(\cdot,\cdot) - \mathbf{E} f_n(\cdot,\cdot)\|^2 \\
&= \left\| \frac{1}{n} \sum_{i=1}^{n} (K_h(\cdot - X_i) K'_h(\cdot - Y_i) - \mathbf{E}\{K_h(\cdot - X) K'_h(\cdot - Y)\}) \right\|^2 \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \Big((K_h(\cdot - X_i) K'_h(\cdot - Y_i) - \mathbf{E}\{K_h(\cdot - X) K'_h(\cdot - Y)\}) \times \\
&\qquad (K_h(\cdot - X_j) K'_h(\cdot - Y_j) - \mathbf{E}\{K_h(\cdot - X) K'_h(\cdot - Y)\}) \Big) \\
&=: \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} M_h(X_i, Y_i, X_j, Y_j),
\end{aligned}$$

and therefore

$$\operatorname{var}(\tilde{T}_n) = \frac{1}{n^4} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{i'=1}^{n} \sum_{j'=1}^{n} \operatorname{cov}(M_h(X_i, Y_i, X_j, Y_j), M_h(X_{i'}, Y_{i'}, X_{j'}, Y_{j'})).$$

One can check that

$$\operatorname{cov}(M_h(X_i, Y_i, X_j, Y_j), M_h(X_{i'}, Y_{i'}, X_{j'}, Y_{j'})) = 0$$

unless $(i,j) = (i',j')$ or $(i,j) = (j',i')$. Thus,

$$\begin{aligned}
\operatorname{var}(\tilde{T}_n) &= \frac{1}{n^4} \left(n \operatorname{var}(M_h(X_1, Y_1, X_1, Y_1)) + 2n(n-1) \operatorname{var}(M_h(X_1, Y_1, X_2, Y_2)) \right) \\
&\approx \frac{2}{n^2} \operatorname{var}(M_h(X_1, Y_1, X_2, Y_2)).
\end{aligned}$$

If $h \to 0$ then

$$\frac{2}{n^2} \operatorname{var}(M_h(X_1, Y_1, X_2, Y_2)) \approx \frac{2\|f\|^2}{n^2 h^d h^{d'}}, \tag{24}$$

therefore a possible form for the asymptotic normal distribution is

$$nh^{d/2}h^{d'/2}(T_n - \mathbf{E}\{T_n\})/\sigma \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

where

$$\sigma^2 = 2\|f\|^2.$$

Thus the asymptotic α-level test rejects the null hypothesis if

$$T_n > \mathbf{E}\{T_n\} + \frac{\sigma}{nh^{d/2}h^{d'/2}}\,\Phi^{-1}(1-\alpha),$$

where $\mathbf{E}\{T_n\}$ may be replaced by its upper bound,

$$L_h(0)L'_h(0)/n = \|K\|^2\|K'\|^2/(nh^d h^{d'}).$$

The only problem left is that the threshold is not distribution-free: σ depends on the unknown f. The simplest distribution-free bound for the variance,

$$\sigma^2 \leq \frac{\|K\|^4\|K'\|^4}{n^2h^{2d}h^{2d'}}$$

is unsatisfactory since its performance as a function of h is worse than the result (24). An improved distribution-free bound on the variance (for both fixed and decreasing h) is a topic for future research: we give an empirical estimate below (eq. 26) for use in asymptotic hypothesis tests.

We now consider the case of fixed h. Following Feuerverger (1993); Serfling (1980), the distribution of T_n under $\mathcal{H}_0$ is

$$nT_n \xrightarrow{D} \sum_{l=1}^{\infty} \lambda_l z_l^2, \tag{25}$$

where $z_l \sim \mathcal{N}(0,1)$ i.i.d., and λ_l are the solutions to an eigenvalue problem depending on the unknown distribution of X and Y (see Gretton et al., 2008, Theorem 2 for details).

A difficulty in using the statistic (19) in a hypothesis test therefore arises due to the form of the null distribution of the statistic, which is a function of the unknown distribution over X and Y, whether or not h is fixed. In the case of h decreasing according to (23), we may use an empirical estimate of the variance of T_n under $\mathcal{H}_0$ due to Gretton et al. (2008, Theorem 4). Denoting by $\odot$ the entrywise matrix product and $A^{\cdot 2}$ the entrywise matrix power,

$$\mathrm{var}(T_n) = \mathbf{1}^\top\left(\mathbf{B} - \mathrm{diag}(\mathbf{B})\right)\mathbf{1}, \tag{26}$$

where

$$\mathbf{B} = \left((\mathbf{HLH}) \odot \left(\mathbf{HL'H}\right)\right)^{\cdot 2},$$

$\mathbf{L}$ is a matrix with entries $L_h(X_i - X_j)$, $\mathbf{L}'$ is a matrix with entries $L'_h(Y_i - Y_j)$, $\mathbf{H} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^\top$ is a centering matrix, and $\mathbf{1}$ an $n \times 1$ vector of ones.

Two approaches have been proposed in the case of fixed h to obtain appropriate quantiles of the null distribution for hypothesis testing: repeated shuffling of the sample (Feuerverger,

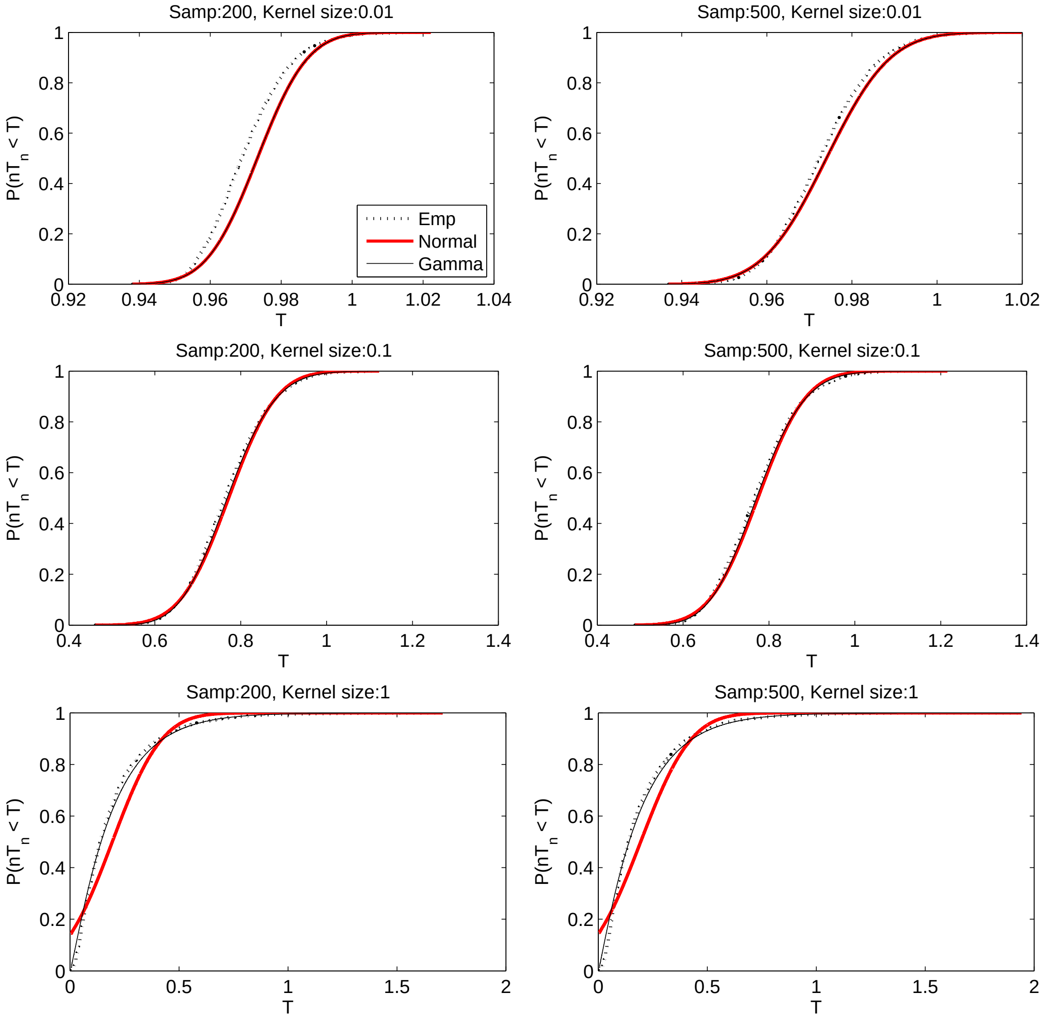

Figure 1: Simulated cumulative distribution function of T_n (*Emp*) under $\mathcal{H}_0$ for $n = 200$ (left column) and $n = 500$ (right column), compared with the two-parameter Gamma distribution (*Gamma*) and the Normal distribution (*Normal*). The empirical CDF was obtained empirically using 5000 independent draws of T_n. Both the parametric approximations are fit using the mean and variance in equations (20) and (26). “Samp” is the number n of samples, and the bandwidth is h.

1993), and approximation by a two-parameter Gamma density (Kankainen, 1995),

$$\begin{aligned} nT_n &\sim \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha\Gamma(\alpha)} \\ \text{where} \quad \alpha &= \frac{(\mathbf{E}\{T_n\})^2}{\mathrm{var}(T_n)}, \quad \beta = \frac{n\mathrm{var}(T_n)}{\mathbf{E}\{T_n\}}, \end{aligned}$$

and $\mathbf{E}\{T_n\}$ is given in (20). This Gamma approximation was found by Gretton et al. (2008) to perform identically on the Section 5 benchmark data to the more computationally expensive approach of Feuerverger (1993). We emphasize, however, that this approximation is a heuristic: no guarantees are provided regarding the asymptotic performance of this approximation in terms of Type II error, nor is it established under what conditions the approximation fails.

We end this section with an empirical comparison between the Normal and two-parameter Gamma null distribution approximations, and the null CDF generated by repeated independent samples of T_n. We chose X and Y to be independent and univariate, with X having a uniform distribution and Y being a symmetric bimodal mixture of Gaussians. Both variables had zero mean and unit standard deviation. Results are plotted in Figure 1.

We observe that as the bandwidth increases, the Gamma approximation of T_n becomes more accurate (although it is always good for large quantiles, which is the region most important to a hypothesis test). The Normal approximation is very close to the Gamma approximation for small bandwidths, but is less accurate (with respect to both the Gamma distribution and the simulated CDF) for larger bandwidths. Finally, for the smallest bandwidth ($h = 0.01$), both approximate null distributions become more accurate for increasing n (for larger kernel sizes, the effect is too small to see on the plots). We will return to these points in the next section when analysing our experimental results.

5. Numerical results

In comparing the independence tests, we made use of the multidimensional benchmark data proposed by Gretton et al. (2008). We tested the independence in two, four, and six dimensions (i.e. $d \in 1, 2, 3$ and $d = d'$). The data were constructed as follows. First, we generated n samples of two independent univariate random variables, each drawn at random from the ICA benchmark densities of Bach and Jordan (2002, Figure 5): these included super-Gaussian, sub-Gaussian, multimodal, and unimodal distributions, with the common property of zero mean and unit variance. The densities are described in Table 5, as reproduced from Gretton et al. (2005b, Table 3). Second, we mixed these random variables using a rotation matrix parametrised by an angle θ, varying from 0 to $\pi/4$ (a zero angle meant the data were independent, while dependence became easier to detect as the angle increased to $\pi/4$: see the two plots in Figure 2). Third, in the cases $d = 2$ and $d = 3$, independent Gaussian noise of zero mean and unit variance was used to fill the remaining dimensions, and the resulting vectors were multiplied by independent random two- or three-dimensional orthogonal matrices, to obtain random vectors X and Y dependent across all observed dimensions. We emphasise that classical approaches (such as Spearman's ρ or Kendall's τ) are unable to find this dependence, since the variables are uncorrelated; nor can we recover the subspace in which the variables are dependent using PCA, since this subspace has the same second order properties as the noise. We investigated sample sizes $n = 128, 512, 1024$, and 2048.

We compared three different asymptotic independence testing approaches based on space partitioning: the L_1 test, denoted *L1*; the log likelihood test *Like*; and a third test, *Pears*,

Label	Definition	Kurtosis
a	Student's t distribution, 3 DOF	∞
b	Double exponential	3.00
c	Uniform	-1.20
d	Students's t distribution, 5 DOF	6.00
e	Exponential	6.00
f	Mixture, 2 double exponentials	-1.70
g	Symmetric mixture 2 Gauss., multimodal	-1.85
h	Symmetric mixture 2 Gauss., transitional	-0.75
i	Symmetric mixture 2 Gauss., unimodal	-0.50
j	Asymm. mixture 2 Gauss., multimodal	-0.57
k	Asymm. mixture 2 Gauss., transitional	-0.29
l	Asymm. mixture 2 Gauss., unimodal	-0.20
m	Symmetric mixture 4 Gauss., multimodal	-0.91
n	Symmetric mixture 4 Gauss., transitional	-0.34
o	Symmetric mixture 4 Gauss., unimodal	-0.40
p	Asymm. mixture 4 Gauss., multimodal	-0.67
q	Asymm. mixture 4 Gauss., transitional	-0.59
r	Asymm. mixture 4 Gauss., unimodal	-0.82

Table 1: Labels of distributions used in the independence test benchmarks, and their respective kurtoses. All distributions have zero mean and unit variance.

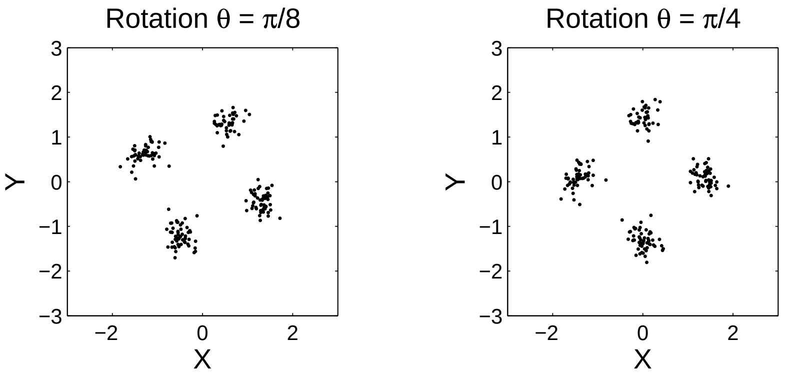

Figure 2: Example dataset for $d = d' = 1$, $n = 200$, and rotation angles $\theta = \pi/8$ (left) and $\theta = \pi/4$ (right). In this case, both sources are mixtures of two Gaussians (source *(g)* in Gretton et al., 2005b, Table 3).

based on a conjecture regarding the asymptotic distribution of the Pearson χ^2 statistic

$$\chi_n^2(\nu_n, \mu_{n,1} \times \mu_{n,2}) \quad = \quad \sum_{A \in \mathcal{P}_n} \sum_{B \in Q_n} \frac{(\nu_n(A \times B) - \mu_{n,1}(A) \cdot \mu_{n,2}(B))^2}{\mu_{n,1}(A) \cdot \mu_{n,2}(B)}$$

(see Appendix B for details, and for a further conjecture regarding a strongly consistent test for the χ_n^2 statistic). The number of discretisations per dimension was set at $m_n = m_n' = 4$, besides in the $n = 128, d = 2$ case and the $d = 3$ cases, where it was set at $m_n = m_n' = 3$: for the latter values of n and d, there were too few samples per bin when a greater number of partitions were used, causing poor performance. We divided our spaces $\mathbb{R}^d$ and $\mathbb{R}^{d'}$ into roughly equiprobable bins. Further increases in the number of partitions per dimension, where sufficient samples were present to justify this (i.e., the $n = 512, d = 1$ case), resulted only in very minor shifts in performance.

We compared the partitioning approaches with the kernel approach from Section 4, using both the Gamma *Ker(g)* and Normal *Ker(n)* approximations to the null distribution. Our kernels were Gaussian for both X and Y, with bandwidths set to the median distance between samples of the respective variables. Note that a more sophisticated but computationally costly approach to bandwidth selection is described by Fukumizu et al. (2008), which involves matching the closed-form expression for the variance of T_n in (26) with an estimate obtained by data shuffling.

Results are plotted in Figure 3 (average over 500 independent generations of the data). The y-intercept on these plots corresponds to the acceptance rate of $\mathcal{H}_0$ at independence, or $1-$ (Type I error), and should be close to the design parameter of $1 - \alpha = 0.95$. Elsewhere, the plots indicate acceptance of $\mathcal{H}_0$ where the underlying variables are dependent, i.e. the Type II error.

As expected, we observe dependence becomes easier to detect as θ increases from 0 to $\pi/4$, when n increases, and when d decreases. Although no tests are reliable for small θ, several tests do well as θ approaches $\pi/4$ (besides the case of $n = 128$, $d = 2$). The L_1 test has a lower Type II error than the χ^2 test when the number of samples per partition

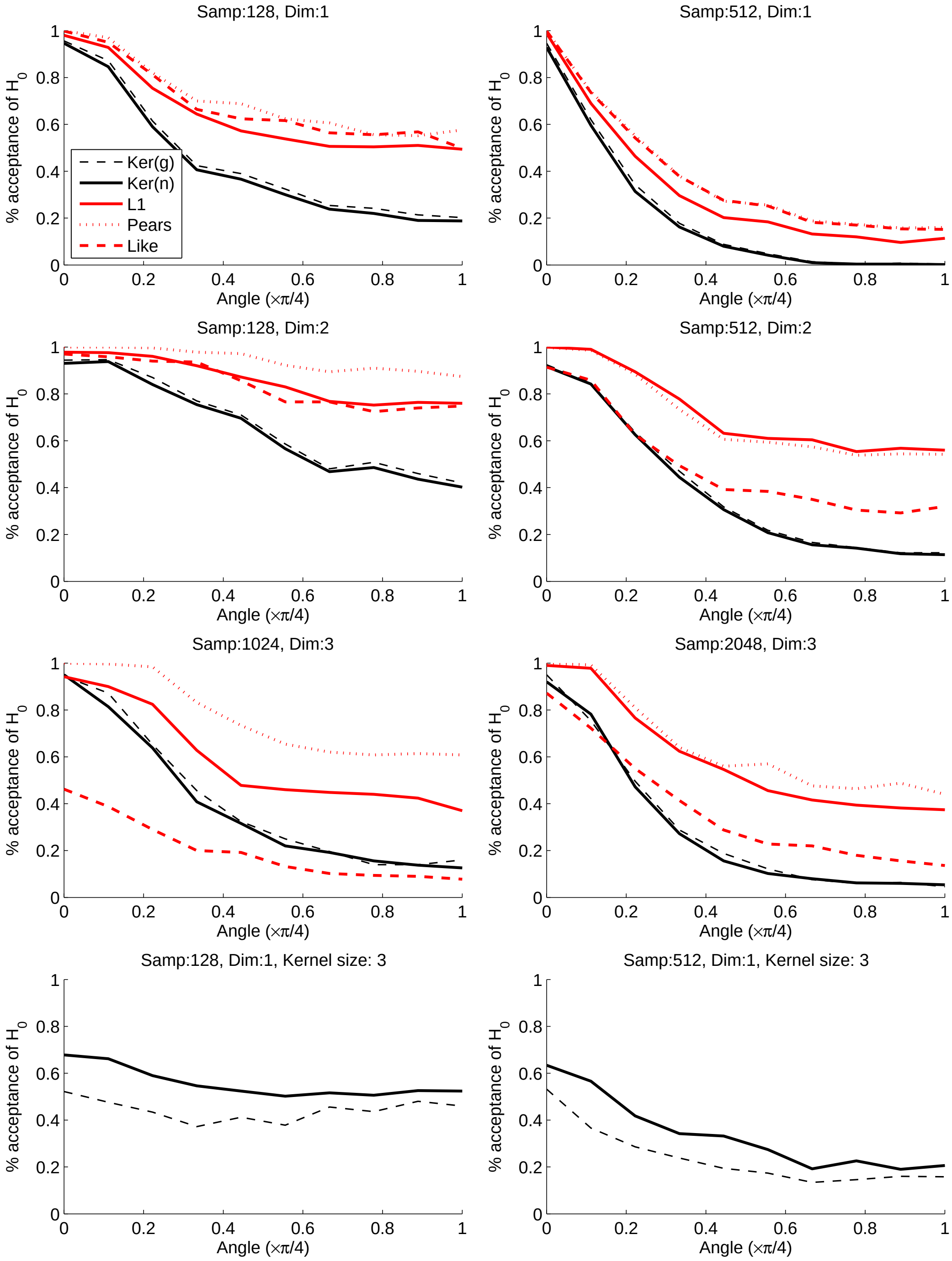

Figure 3: Rate of acceptance of $\mathcal{H}_0$ for the *Ker(g)*, *Ker(n)*, *L1*, *Pears*, and *Like* tests. "Samp" is the number n of samples, and "dim" is the dimension $d = d'$ of x and y. In the final row, the performance of the *Ker(g)* and *Ker(n)* tests is plotted for a large bandwidth $h = 3$, and $\tilde{\alpha} = 0.5$, to illustrate the difference between the Normal and two-parameter Gamma approximations to the null distribution.

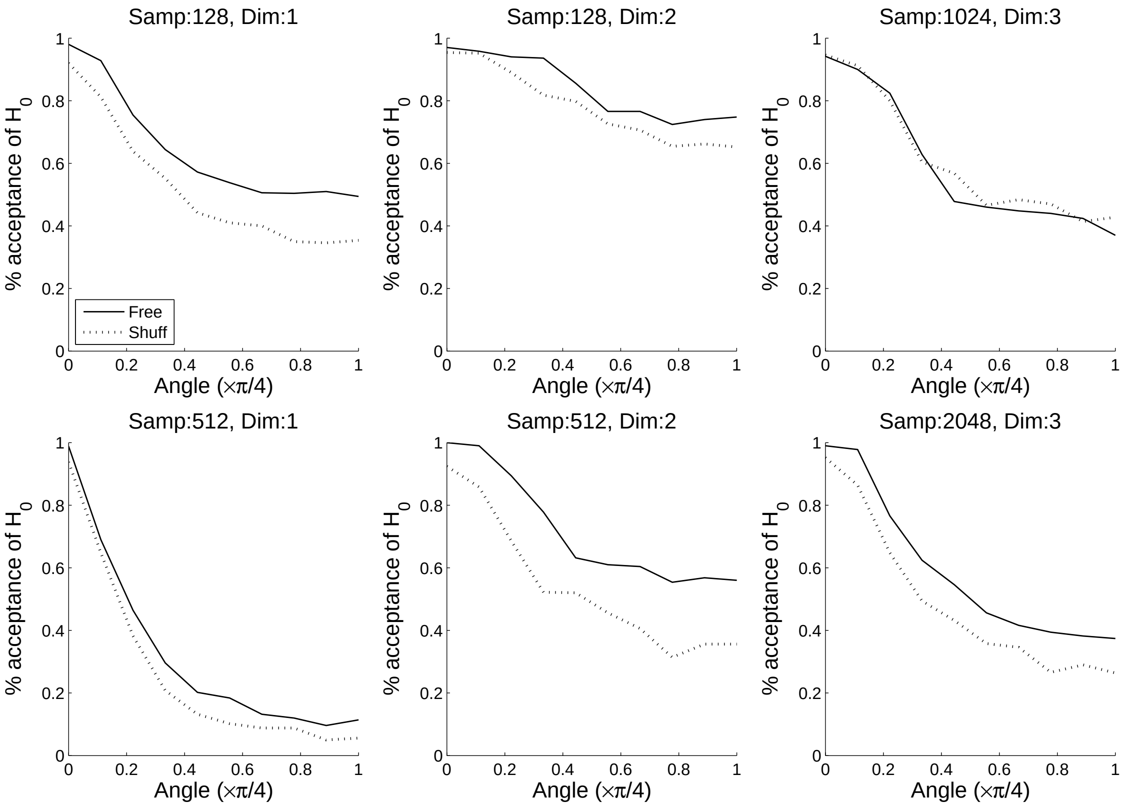

Figure 4: Rate of acceptance of $\mathcal{H}_0$ for the distribution-free (*Free*) and shuffling-based (*Shuff*) null distribution quantiles, using the L1 test statistic. "Samp" is the number n of samples, and "dim" is the dimension $d = d'$ of x and y.

is small ($n = 128, d = 1$, $n = 128, d = 2$, and $n = 1024, d = 3$), but this advantage is lessened for larger numbers of samples per partition. The log-likelihood test generally has the lowest Type II error of the three partition-based tests, however it gives a Type I error larger than the design parameter of 0.05 when the number of samples per bin is insufficient: this problem is severe in the case $n = 1024$ and $d = 3$, but can also be observed at $n = 2048, d = 3$ (for larger sample sizes $n = 3072, d = 3$ and $n = 4096, d = 3$, the Type I error of the log-likelihood test was at or below the design value). This suggests the log-likelihood test is more susceptible to bias for small numbers of samples per bin than the L_1 and χ^2 tests. In the remaining cases, performance of the log-likelihood test and the L_1 test is comparable, besides in the case $n = 512, d = 2$, where the log-likelihood test has an advantage.

The superior performance of the log-likelihood test compared with the χ^2 test (in the cases $d = 1$ and $d = 2$) might arise due to the different convergence properties of the two test statistics. In particular, we note the superior convergence behaviour of the goodness-of-fit statistic for the log likelihood (eq. 14), as compared with the χ^2 statistic (eq. 31 in Appendix B), in terms of the dependence of the latter on the number m_n of partitions used. By analogy, we anticipate the log-likelihood independence statistic $I_n(\nu_n, \mu_{n,1} \times \mu_{n,2})$ will also converge faster than the Pearson χ^2 independence statistic $\chi^2_n(\nu_n, \mu_{n,1} \times \mu_{n,2})$, and

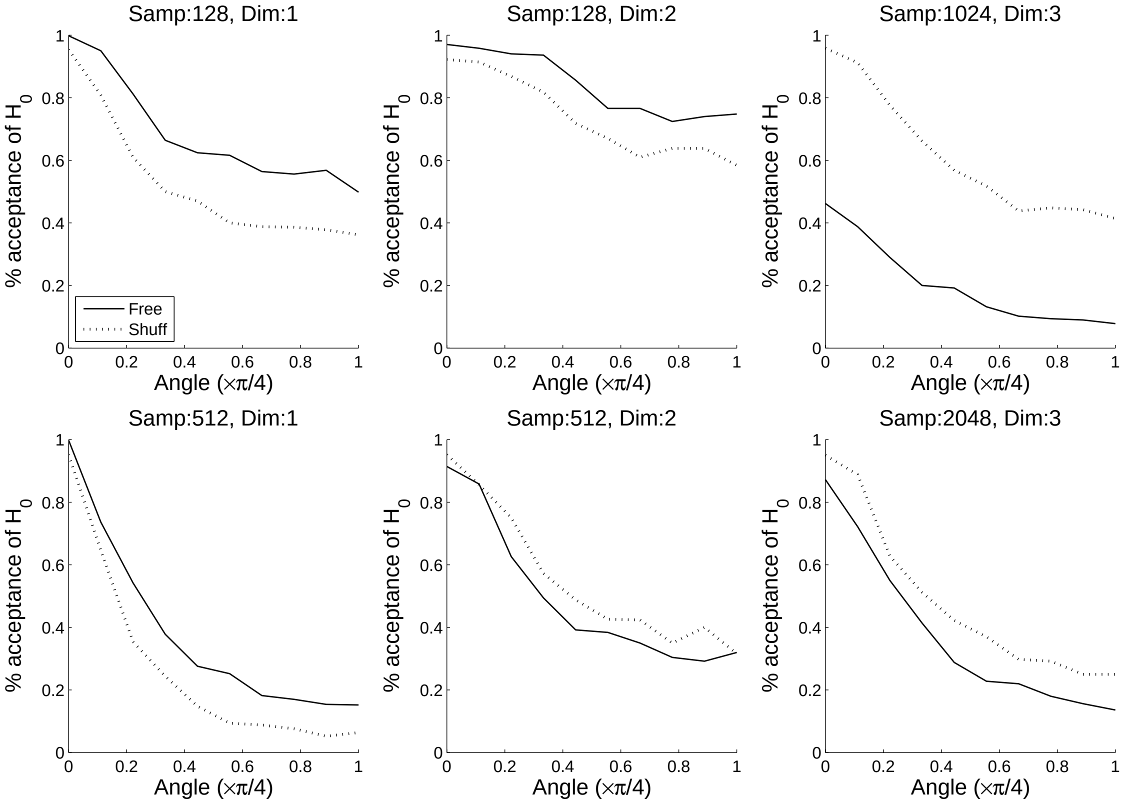

Figure 5: Rate of acceptance of $\mathcal{H}_0$ for the distribution-free (*Free*) and shuffling-based (*Shuff*) null distribution quantiles, using the log-likelihood test statistic. "Samp" is the number n of samples, and "dim" is the dimension $d = d'$ of x and y.

thus provide better test performance. A more formal discussion of this behaviour is a topic for future research.

In all cases, the kernel-based test has the lowest Type II error.[3] That said, one should bear in mind the kernel test thresholds require $\mathbf{E}\{T_n\}$ and $\text{var}(T_n)$, which are unknown and must be estimated from the data using equations (20) and (26), respectively. In other words, unlike the L_1 and log likelihood tests, the kernel test thresholds in our experiments are themselves finite sample estimates (which we have not attempted to account for, and which could impact on test performance). Moreover, the Gamma approximation to the null distribution is simply a heuristic, with no asymptotic guarantees.

It is of interest to further investigate the null distribution approximation strategies for the kernel tests, and in particular to determine the effect on test performance of the observations made in Figure 1. Since the median distance between sample points was small enough in our previous experiments for the Normal and Gamma estimates to be very similar, we used an artificially high kernel bandwidth $h = 3$. In addition, we employed a much lower $\tilde{\alpha} = 0.5$, since this provided a more visible performance difference. The final row of Figure 3 shows the resulting test performance. We recall from Figure 1 that for large

[3] Aside from $n = 1024$ and $d = 3$, where the log-likelihood has a lower Type II error: we disregard this result since it is due to the log-likelihood test being affected by bias, as discussed above.

kernel sizes and $\tilde{\alpha} = 0.5$, the Gaussian approximation returns a larger threshold than the true CDF would require, and thus the Normal distribution has a lower Type I error (the error for very small values of α is in the opposite direction, but had a less pronounced effect in our experiments). The large bandwidth required to observe this behaviour results in a substantial performance penalty on the Type II error, however, and would not be used in practice.

An alternative approach to obtaining null distribution quantiles for test thresholds is via a shuffling procedure: the ordering of the $Y_1, \ldots, Y_n$ sample is permuted repeatedly while that of $X_1, \ldots, X_n$ sample is kept fixed, and the $1 - \alpha$ quantile is obtained from the resulting estimated cumulative distribution function of the test statistic. Again, we emphasize that unlike the asymptotic L1 and log-likelihood tests we have proposed, the resulting test threshold is an empirical estimate, and the convergence behaviour of this estimate is not accounted for. In our final experiments, we compared the performance of our asymptotic tests for *L1* and *Like* with this shuffling approach, for the same data as in our Figure 3 experiments.[4] We used $p = 200$ permutations in obtaining the approximation to the null distribution. Results for the *L1* case are plotted in Figure 4, and those for the *Like* case in Figure 5.

In the case of the *L1* statistic, we observe the distribution-free approach is conservative in terms of the Type I error, generally setting it slightly lower than the target value. The shuffling approach returns a lower Type II error, however it is notable that the performance difference is not particularly large with respect to our distribution-free threshold, and that apart from an offset, the error as a function of angle takes the same form. We should further bear in mind that the shuffling approach has a substantially greater computational cost (p times the cost of the distribution-free test). In the case of the *Like* statistic, we observe similar behaviour to *L1* in the cases $d = 1$ and $d = 2$. In the $d = 3$ case, however, the *Like* test gives too large a Type I error, and thus the Type II performance of the two approaches cannot be compared (although for $n = 2048$, the *Like* test is observed to approach the asymptotic regime, and the Type I performance is closer to the target value).

6. Conclusion

We have described distribution-free strong consistent tests of independence, and asymptotically α-level tests, based on three statistics: the L_1 distance, the log-likelihood, and a kernel-based distance. The asymptotic L_1 and log-likelihood tests require that the distributions be non-atomic, but make no assumptions apart from this: in particular, the test thresholds are *not* functions of the distribution. The kernel statistic is interpretable as either an L_2 distance between kernel density estimates (if the kernel bandwidth shrinks for increasing sample size), or as the Hilbert-Schmidt norm of a covariance operator between reproducing kernel Hilbert spaces (if the kernel bandwidth is fixed). We have provided a novel strong consistent test for the kernel statistic, as well as reviewing two asymptotically α-level tests (for both fixed and shrinking kernel bandwidth). Unlike the L_1 and log-likelihood tests, the thresholds for the kernel asymptotic tests are distribution depen-

[4] This comparison was made for the kernel statistic on these data by Gretton et al. (2008), and no performance difference was found.

dent. We also gave conjectures regarding the strong consistent test and asymptotically α-level test for the Pearson χ^2 distance.

Our experiments showed the asymptotic tests to be capable of detecting dependence for both univariate and multi-dimensional variables (of up to three dimensions each), for variables having no linear correlation. The kernel tests had lower Type II error than the L_1 and log-likelihood tests for a given Type I error, however we should bear in mind that the kernel test thresholds were finite sample estimates, and the resulting convergence issues have not been addressed. The log-likelihood test appeared to suffer more from bias than the L_1 test, in cases where there were few samples per partition (this effect was most visible in high dimensions).

This study raises a number of questions for future research. First, the χ^2 tests remain conjectures, and proofs should be established. Second, there is as yet no distribution-free asymptotic threshold for the kernel test, which could be based on a tighter bound on the variance of the test statistic under the null distribution. Third, the asymptotic distribution of the kernel statistic with fixed bandwidth is presently a heuristic: it would therefore be of interest to replace this with a null distribution estimate having appropriate convergence guarantees.

Acknowledgments

The research of László Györfi is supported by the Hungarian Academy of Sciences (MTA SZTAKI). This work is also supported by the IST Program of the EC, under the FP7 Network of Excellence, ICT-216886-NOE. Part of this work was completed while Arthur Gretton was a project scientist at CMU, under grants DARPA IPTO FA8750-09-1-0141, ONR MURI N000140710747, and ARO MURI W911NF0810242.

Appendix A. Proof of Theorem 3

The main difficulty in proving Theorem 3 is that it states the asymptotic normality of $L_n(\nu_n, \mu_{n,1} \times \mu_{n,2})$, which is a sum of *dependent* random variables. To overcome this problem, we use a “Poissonization” argument originating from the fact that an empirical process is equal in distribution to the conditional distribution of a Poisson process given the sample size (for more on Poissonization techniques, we refer the reader to Beirlant, Györfi, and Lugosi, 1994).

We begin by introducing the necessary terminology. For each $n \geq 1$, denote by N_n a Poisson(n) random variable, defined on the same probability space as the sequences $(X_i)_{i\geq 1}$ and $(Y_i)_{i\geq 1}$, and independent of these sequences. Denote by ν_{N_n}, $\mu_{N_n,1}$ and $\mu_{N_n,2}$ the Poissonized version of the empirical measures associated with the samples $\{(X_i, Y_i)\}$, $\{X_i\}$ and $\{Y_i\}$, respectively, so that

$$\nu_{N_n}(A \times B) = \frac{\#\{i : (X_i, Y_i) \in A \times B, i = 1, \ldots, N_n\}}{n},$$

$$\mu_{N_n,1}(A) = \frac{\#\{i : X_i \in A, i = 1, \ldots, N_n\}}{n},$$

and

$$\mu_{N_n,2}(B) = \frac{\#\{i : Y_i \in B, i = 1, \ldots, N_n\}}{n}$$

for any Borel subsets A and B. The Poissonized version $\tilde{L}_n(\nu_n, \mu_{n,1} \times \mu_{n,2})$ of $L_n(\nu_n, \mu_{n,1} \times \mu_{n,2})$ is then

$$\tilde{L}_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) = \sum_{A \in \mathcal{P}_n} \sum_{B \in Q_n} |\nu_{N_n}(A \times B) - \mu_{N_n,1}(A) \cdot \mu_{N_n,2}(B)|.$$

Clearly,

$$n\nu_{N_n}(A \times B) = \#\{i : (X_i, Y_i) \in A \times B, i = 1, \ldots, N_n\},$$
$$n\mu_{N_n,1}(A) = \#\{i : X_i \in A, i = 1, \ldots, N_n\},$$

and

$$n\mu_{N_n,2}(B) = \#\{i : Y_i \in B, i = 1, \ldots, N_n\}$$

are Poisson random variables.

Key to the proof of Theorem 3 is the following property, which is a slight extension of the proposition of Beirlant, Györfi, and Lugosi (1994, p. 311).

Proposition 7 *Let g_{njk} ($n \geq 1$, $j = 1, \ldots, m_n$, $k = 1, \ldots, m'_n$) be real measurable functions, and let*

$$M_n := \sum_{j=1}^{m_n} \sum_{k=1}^{m'_n} g_{njk} \left(\nu_{N_n}(A_{nj} \times B_{nk}) - \mu_{N_n,1}(A_{nj})\mu_{N_n,2}(B_{nk})\right).$$

Assume that, under the null hypothesis,

$$\mathbf{E}\{g_{njk} \left(\nu_{N_n}(A_{nj} \times B_{nk}) - \mu_{N_n,1}(A_{nj})\mu_{N_n,2}(B_{nk})\right)\} = 0,$$

and that

$$\left(M_n, \frac{N_n - n}{\sqrt{n}}\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\left[\begin{array}{c} 0 \\ 0 \end{array}\right], \left[\begin{array}{cc} \sigma^2 & 0 \\ 0 & 1 \end{array}\right]\right) \tag{27}$$

as $n \to \infty$, where σ is a positive constant and $\mathcal{N}(\mathbf{m}, \mathbf{C})$ is a normally distributed random variable with mean $\mathbf{m}$ and covariance matrix $\mathbf{C}$. Then

$$\frac{1}{\sigma} \sum_{j=1}^{m_n} \sum_{k=1}^{m'_n} g_{njk} \left(\nu_n(A_{nj} \times B_{nk}) - \mu_{n,1}(A_{nj})\mu_{n,2}(B_{nk})\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1).$$

Proof The proof is in sketch form, along the lines of Biau and Györfi (2005). Define the two characteristic functions

$$\Phi_n(t, v) := \mathbf{E}\left\{\exp\left(\imath t M_n + \imath v \frac{N_n - n}{\sqrt{n}}\right)\right\}$$

and

$$\Psi_n(t) := \mathbf{E}\left\{\exp\left(\imath t \sum_{j=1}^{m_n} \sum_{k=1}^{m'_n} g_{njk} \left(\nu_n(A_{nj} \times B_{nk}) - \mu_{n,1}(A_{nj})\mu_{n,2}(B_{nk})\right)\right)\right\}.$$

We begin with the result

$$\mathbf{E}\left\{\exp\left(\imath tM_n + \imath uN_n\right)\right\} = \sum_{l=0}^{\infty} \mathbf{E}\{\exp(\imath tM_n)|N_n = l\}e^{\imath ul}p_n(l),$$

where $p_n(l)$ is the probability distribution of the Poisson(n) random variable N_n,

$$p_n(l) = \mathbf{P}\{N_n = l\} = e^{-n}n^l/l!,$$

and

$$\Psi_n(t) = \mathbf{E}\{\exp(\imath tM_n)|N_n = n\}.$$

Taking the inverse Fourier transform,

$$\mathbf{E}\{\exp(\imath tM_n)|N_n = n\} = \frac{1}{2\pi p_n(n)}\int_{-\pi}^{\pi} e^{-\imath un}\mathbf{E}\left\{\exp\left(\imath tM_n + \imath uN_n\right)\right\}du.$$

We now replace $n!$ with the Stirling approximation to obtain

$$2\pi p_n(n) = \frac{2\pi e^{-n}n^n}{n!} \approx \sqrt{\frac{2\pi}{n}} \quad \text{as} \quad n \to \infty.$$

Then, substituting $v = u\sqrt{n}$, we get

$$\Psi_n(t) = \frac{1}{\sqrt{2\pi}}(1 + o(1))\int_{-\pi\sqrt{n}}^{\pi\sqrt{n}} \Phi_n(t, v)dv.$$

By assumption,

$$\Phi_n(t, v) \to e^{-t^2\sigma^2/2}e^{-v^2/2}$$

as $n \to \infty$. The result follows from Rao (1973, p. 136). ■

We now use Proposition 7 to prove

$$\frac{\sqrt{n}}{\sigma}(L_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) - \mathbf{E}\{\tilde{L}_n(\nu_n, \mu_{n,1} \times \mu_{n,2})\}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \tag{28}$$

where we recall $\sigma^2 = 1 - 2/\pi$. This provides the result in Theorem 3 with the centering constant

$$C_n = \mathbf{E}\{\tilde{L}_n(\nu_n, \mu_{n,1} \times \mu_{n,2})\} = \sum_{A \in \mathcal{P}_n}\sum_{B \in Q_n} \mathbf{E}\{|\nu_{N_n}(A \times B) - \mu_{N_n,1}(A) \cdot \mu_{N_n,2}(B)|\}. \tag{29}$$

To apply Proposition 7, we must prove assumption (27) holds. Define

$$g_{njk}(x) = \sqrt{n}\left(|x| - \mathbf{E}\left|\nu_{N_n}(A_{nj} \times B_{nk}) - \mu_{N_n,1}(A_{nj})\mu_{N_n,2}(B_{nk})\right|\right).$$

Let

$$\begin{aligned} S_n \quad := \quad & t\sqrt{n}\sum_{j=1}^{m_n}\sum_{k=1}^{m'_n}\Big(\left|\nu_{N_n}(A_{nj} \times B_{nk}) - \mu_{N_n,1}(A_{nj})\mu_{N_n,2}(B_{nk})\right| \\ & -\mathbf{E}\left|\nu_{N_n}(A_{nj} \times B_{nk}) - \mu_{N_n,1}(A_{nj})\mu_{N_n,2}(B_{nk})\right|\Big) \\ & +v\sqrt{n}\left(\frac{N_n}{n} - 1\right). \end{aligned}$$

Our goal is to prove the assumption in (27) holds. In particular, we require the variance of the Poissonized statistic S_n. After this variance is calculated, the asymptotic normality in (27) can be proved by verifying the Lyapunov conditions as in (Beirlant, Györfi, and Lugosi, 1994). From the definitions of ν_{N_n}, μ_1, and μ_2, we have

$$\frac{N_n}{n} - 1 = \sum_{A\in\mathcal{P}_n}\sum_{B\in Q_n} \nu_{N_n}(A\times B) - \sum_{A\in\mathcal{P}_n}\sum_{B\in Q_n} \mu_1(A)\mu_2(B),$$

and thus the variance of S_n is

$$\begin{aligned}
\mathrm{var}(S_n) &= t^2 n \sum_{A\in\mathcal{P}_n}\sum_{B\in Q_n} \mathrm{var}\,|\nu_{N_n}(A\times B) - \mu_{N_n,1}(A)\mu_{N_n,2}(B)| \\
&+ 2tvn \sum_{A\in\mathcal{P}_n}\sum_{B\in Q_n} \mathbf{E}\big\{\, |\nu_{N_n}(A\times B) - \mu_{N_n,1}(A)\mu_{N_n,2}(B)| \\
&\quad \cdot (\nu_{N_n}(A\times B) - \mu_1(A)\mu_2(B))\,\big\} \\
&+ v^2.
\end{aligned}$$

One can check that there exist standard normal random variables $Z_{A\times B}$, Z_A, and Z_B such that

$$\nu_{N_n}(A\times B) \overset{\mathcal{D}}{\approx} Z_{A\times B}\sqrt{\frac{\mu_1(A)\mu_2(B)}{n}} + \mu_1(A)\mu_2(B),$$

$$\mu_{N_n,1}(A) \overset{\mathcal{D}}{\approx} Z_A\sqrt{\frac{\mu_1(A)}{n}} + \mu_1(A),$$

and

$$\mu_{N_n,2}(B) \overset{\mathcal{D}}{\approx} Z_B\sqrt{\frac{\mu_2(B)}{n}} + \mu_2(B),$$

which implies

$$\begin{aligned}
& \nu_{N_n}(A\times B) - \mu_{N_n,1}(A)\mu_{N_n,2}(B) \\
\overset{\mathcal{D}}{\approx}\ & Z_{A\times B}\sqrt{\frac{\mu_1(A)\mu_2(B)}{n}} + \mu_1(A)\mu_2(B) \\
& - \left(Z_A\sqrt{\frac{\mu_1(A)}{n}} + \mu_1(A)\right)\left(Z_B\sqrt{\frac{\mu_2(B)}{n}} + \mu_2(B)\right) \\
=\ & \sqrt{\frac{\mu_1(A)\mu_2(B)}{n}}\left(Z_{A\times B} - Z_A Z_B\frac{1}{\sqrt{n}} - Z_A\sqrt{\mu_2(B)} - Z_B\sqrt{\mu_1(A)}\right) \\
\approx\ & Z_{A\times B}\sqrt{\frac{\mu_1(A)\mu_2(B)}{n}}.
\end{aligned}$$

Thus,

$$\begin{aligned}
& \operatorname{var}(S_n) \\
\approx\ & t^2 n \sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n} \operatorname{var}\left|Z_{A\times B}\sqrt{\frac{\mu_1(A)\mu_2(B)}{n}}\right| \\
+\ & 2tvn \sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n} \mathbf{E}\left\{\left|Z_{A\times B}\sqrt{\frac{\mu_1(A)\mu_2(B)}{n}}\right|\cdot\left(Z_{A\times B}\sqrt{\frac{\mu_1(A)\mu_2(B)}{n}}\right)\right\} \\
+\ & v^2 \\
=\ & t^2 \sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n} \operatorname{var}\left|Z_{A\times B}\right|\mu_1(A)\mu_2(B) \\
+\ & 2tv \sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n} \mathbf{E}\left\{\left|Z_{A\times B}\right| Z_{A\times B}\right\}\mu_1(A)\mu_2(B) \\
+\ & v^2 \\
=\ & t^2(1-2/\pi)+v^2.
\end{aligned}$$

Finally, we use the variable $Z_{A\times B}$ in defining a distribution-free upper bound on C_n, which we use in our asymptotically α-level independence test,

$$\begin{aligned}
C_n &= \sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n} \mathbf{E}\{|\nu_{N_n}(A\times B)-\mu_{N_n,1}(A)\cdot\mu_{N_n,2}(B)|\} \\
&\approx \sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n} \mathbf{E}\{|Z_{A\times B}|\}\sqrt{\mu_1(A)\mu_2(B)/n} \\
&\le \sqrt{2/\pi}\sqrt{\frac{m_n m_n'}{n}} \qquad (30)
\end{aligned}$$

Appendix B. Conjectured large sample properties of the Pearson χ^2 statistic

For a real parameter λ, the *power divergence statistic* is defined as

$$D_{n,\lambda}(\mu_{n,1},\mu_1)=\frac{2}{\lambda(\lambda+1)}\sum_{j=1}^{m_n}\mu_{n,1}(A_{n,j})\left[\left(\frac{\mu_{n,1}(A_{n,j})}{\mu_1(A_{n,j})}\right)^{\lambda}-1\right]$$

provided $\lambda\neq 0$ and $\lambda\neq 1$ (cf. Read and Cressie, 1988). One can check that

$$\lim_{\lambda\to 0} D_{n,\lambda}(\mu_{n,1},\mu_1)=I_n(\mu_{n,1},\mu_1).$$

For $\lambda=1$, we have the Pearson χ^2 statistic:

$$\chi_n^2(\mu_{n,1},\mu_1)=D_{n,1}(\mu_{n,1},\mu_1)=\sum_{j=1}^{m_n}\frac{(\mu_{n,1}(A_{n,j}))-\mu_1(A_{n,j}))^2}{\mu_1(A_{n,j})}.$$

For testing independence, we employ the Pearson χ^2 test statistic

$$\chi_n^2(\nu_n,\mu_{n,1}\times\mu_{n,2}) \;=\; \sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n}\frac{(\nu_n(A\times B)-\mu_{n,1}(A)\cdot\mu_{n,2}(B))^2}{\mu_{n,1}(A)\cdot\mu_{n,2}(B)}.$$

B.1 Strongly consistent test

Quine and Robinson (1985) proved that, for all $\epsilon > 0$,

$$\mathbf{P}\{\chi_n^2(\mu_{n,1}, \mu_1) > \epsilon\} \leq \binom{n + m_n - 1}{m_n - 1} e^{-\frac{n \log m_n}{2\sqrt{m_n}} \sqrt{\epsilon}} \leq e^{m_n \log(n+m_n) - \frac{n \log m_n}{2\sqrt{m_n}} \sqrt{\epsilon}}. \tag{31}$$

A large deviation-based test can be introduced that rejects independence if

$$\chi_n^2(\nu_n, \mu_{n,1} \times \mu_{n,2}) \geq \left(\frac{2(m_n m_n')^{3/2}(\log(n + m_n m_n') + 1)}{n \log(m_n m_n')} \right)^2.$$

Under $\mathcal{H}_0$, we conjecture a non-asymptotic bound for the tail of the distribution of $\chi_n^2(\nu_n, \mu_{n,1} \times \mu_{n,2})$,

$$\begin{aligned} & \mathbf{P}\left\{ \chi_n^2(\nu_n, \mu_{n,1} \times \mu_{n,2}) > \left(\frac{2(m_n m_n')^{3/2}(\log(n + m_n m_n') + 1)}{n \log(m_n m_n')} \right)^2 \right\} \\ \leq \;& e^{m_n m_n' \log(n+m_n m_n') - \frac{n \log(m_n m_n')}{2\sqrt{m_n m_n'}} \frac{2(m_n m_n')^{3/2}(\log(n+m_n m_n')+1)}{n \log(m_n m_n')}} \\ = \;& e^{-m_n m_n'}. \end{aligned}$$

Therefore the conditions (7) imply

$$\sum_{n=1}^{\infty} \mathbf{P}\left\{ \chi_n^2(\nu_n, \mu_{n,1} \times \mu_{n,2}) > \left(\frac{2(m_n m_n')^{3/2}(\log(n + m_n m_n') + 1)}{n \log(m_n m_n')} \right)^2 \right\} < \infty,$$

and by the Borel-Cantelli lemma we have strong consistency under the null hypothesis.

Under the alternative hypothesis the proof strong consistency follows from the proof for the information divergence since

$$I_n(\nu_n, \mu_{n,1} \times \mu_{n,2})/2 \leq \chi_n^2(\nu_n, \mu_{n,1} \times \mu_{n,2})$$

(c.f. Györfi et al., 1998).

B.2 Asymptotic α-level test

Morris (1975), Inglot et al. (1990), and Györfi and Vajda (2002) proved that under (11) and (12),

$$\frac{n\chi_n^2(\mu_{n,1}, \mu_1) - m_n}{\sqrt{2m_n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1),$$

which is the same asymptotic normality result as for $2I_n(\mu_{n,1}, \mu_1)$ (see eq. (16) in Section 3.2). We conjecture that under the conditions (6) and (13),

$$\frac{n\chi_n^2(\nu_n, \mu_{n,1} \times \mu_{n,2}) - m_n m_n'}{\sqrt{2m_n m_n'}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1).$$

Thus, as for the log-likelihood statistic, the hypothesis of independence is rejected if

$$\chi_n^2(\nu_n, \mu_{n,1} \times \mu_{n,2}) \geq \frac{\Phi^{-1}(1 - \alpha)\sqrt{2m_n m_n'} + m_n m_n'}{n}.$$

References

F. R. Bach and M. I. Jordan. Kernel independent component analysis. *J. Mach. Learn. Res.*, 3:1–48, 2002.

C. Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.

A. R. Barron. Uniformly powerful goodness of fit tests. *Ann. Statist.*, 17:107–124, 1989.

A. R. Barron, L. Györfi, and E. C. van der Meulen. Distribution estimation consistent in total variation and in two types of information divergence. *IEEE Trans. Inform. Theory*, 38:1437–1454, 1992.

M. S. Bartlett. The characteristic function of a conditional statistic. *J. London Math. Soc.*, 13, 1938.

J. Beirlant and D. M. Mason. On the asymptotic normality of l_p-norms of empirical functionals. *Math. Methods Statist.*, 4:1–19, 1995.

J. Beirlant, L. Györfi, and G. Lugosi. On the asymptotic normality of the l_1- and l_2-errors in histogram density estimation. *Canad. J. Statist.*, 22:309–318, 1994.

J. Beirlant, L. Devroye, L. Györfi, and I. Vajda. Large deviations of divergence measures on partitions. *J. Statist. Plann. Inference*, 93:1–16, 2001.

G. Biau and L. Györfi. On the asymptotic properties of a nonparametric l_1-test statistic of homogeneity. *IEEE Trans. Inform. Theory*, 51:3965–3973, 2005.

J. R. Blum, J. Kiefer, and M. Rosenblatt. Distribution free tests of independence based on the sample distribution function. *Ann. Math. Statist.*, 32:485–498, 1961.

D. S. Cotterill and M. Csörgő. On the limiting distribution of and critical values for the Hoeffding, Blum, Kiefer, Rosenblatt independence criterion. *Statistics and Decisions*, 3: 1–48, 1985.

I. Csiszár. Information-type measures of divergence of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

A. Dembo and Y. Peres. A topological criterion for hypothesis testing. *Ann. Statist.*, 22: 106–117, 1994.

A. Feuerverger. A consistent test for bivariate dependence. *International Statistical Review*, 61(3):419–433, 1993.

K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems 20*, pages 489–496, Cambridge, MA, 2008. MIT Press.

A. Gretton and L. Györfi. Nonparametric independence tests: Space partitioning and kernel approaches. In *Algorithmic Learning Theory: 19th International Conference*, pages 183–198. Springer, 2008.

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Proc. Intl. Conf. on Algorithmic Learning Theory*, pages 63–78, 2005a.

A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, 2005b.

A. Gretton, K. Fukumizu, C.-H. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20*, pages 585–592, Cambridge, MA, 2008. MIT Press.

L. Györfi and I. Vajda. Asymptotic distributions for goodness of fit statistics in a sequence of multinomial models. *Statistics and Probability Letters*, 56:57–67, 2002.

L. Györfi and E. C. van der Meulen. A consistent goodness of fit test based on the total variation distance. In G. Roussas, editor, *Nonparametric Functional Estimation and Related Topics*, pages 631–645. Kluwer, Dordrecht, 1990.

L. Györfi, F. Liese, I. Vajda, and E. C. van der Meulen. Distribution estimates consistent in χ^2-divergence. *Statistics*, 32:31–57, 1998.

P. Hall. Central limit theorem for integrated square error of multivariate nonparametric density estimators. *Journal of Multivariate Analysis*, 14:1–16, 1984.

W. Hoeffding. A nonparametric test for independence. *The Annals of Mathematical Statistics*, 19(4):546–557, 1948.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

T. Inglot, T. Jurlewitz, and T. Ledwina. Asymptotics for multinomial goodness of fit tests for simple hypothesis. *Theory Probab. Appl.*, 35:797–803, 1990.

J. Jacod and P. Protter. *Probability Essentials.* Springer, New York, 2000.

W. C. M. Kallenberg. On moderate and large deviations in multinomial distributions. *Annals of Statistics*, 13:1554–1580, 1985.

A. Kankainen. *Consistent Testing of Total Independence Based on the Empirical Characteristic Function.* PhD thesis, University of Jyväskylä, 1995.

J. H. B. Kemperman. An optimum rate of transmitting information. *Ann. Math. Statist.*, 40:2156–2177, 1969.

S. Kullback. A lower bound for discrimination in terms of variation. *IEEE Trans. Information Theory*, 13:126–127, 1967.

C. McDiarmid. On the method of bounded differences. In *Survey in Combinatorics*, pages 148–188. Cambridge University Press, 1989.

C. Morris. Central limit theorems for multinomial sums. *Annals of Statistics*, 3:165–188, 1975.

M.P. Quine and J. Robinson. Efficiencies of chi-square and likelihood ratio goodness-of-fit tests. *Ann. Statist.*, 13:727–742, 1985.

C. R. Rao. *Statistical Inference and its Applications*. Wiley, New York, second edition, 1973.

T. Read and N. Cressie. *Goodness-Of-Fit Statistics for Discrete Multivariate Analysis*. Springer-Verlag, New York, 1988.

M. Rosenblatt. A quadratic measure of deviation of two-dimensional density estimates and a test of independence. *The Annals of Statistics*, 3(1):1–14, 1975.

R. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, New York, 1980.

B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Injective hilbert space embeddings of probability measures. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 111–122, 2008.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

G. Tusnády. On asymptotically optimal tests. *Annals of Statistics*, 5:385–393, 1977.