Nonparametric Independence Tests: Space Partitioning and Kernel Approaches

Arthur Gretton[1] and László Györfi[2]

[1] MPI for Biological Cybernetics, Spemannstr. 38, 72076 Tübingen, Germany, arthur@tuebingen.mpg.de

[2] Budapest University of Technology and Economics, H-1521 Stoczek u. 2, Budapest, Hungary. gyorfi@szit.bme.hu

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. All tests reject the null hypothesis of independence if the test statistics become large. The large deviation and limit distribution properties of all three test statistics are given. Following from these results, distribution-free strong consistent tests of independence are derived, as are asymptotically α-level tests. The performance of the tests is evaluated experimentally on benchmark data.

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,

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

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 [1]); and goodness-of-fit tests (for whether a random variable has a particular distribution, by Györfi and van der Meulen [2], and Beirlant et al. [3]). The log-likelihood has also been employed on discretised spaces as a statistic for goodness-of-fit testing [4]. 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: strong consistent tests[3] based on large deviation bounds, which make no assumptions about the distribution; and tests based on the asymptotic distribution of the test statistic, which assume only that the distribution is nonatomic. We also present a conjecture regarding the form taken by an asymptotic test based on the Pearson χ^2 statistic, using the goodness-of-fit results in [4] (further related test statistics include the power divergence family of Read and Cressie [6], although we do not study them here). The advantage of our test procedures is that, besides being explicit and easy to carry out, they require very few assumptions on the partition sequences, are consistent, and have distribution-independent thresholds.

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 [7], as a smoothed difference between empirical characteristic functions [8, 9], or as the Hilbert-Schmidt norm of a cross-covariance operator mapping between functions of the random variables [10, 11]. Each test differs from the others regarding the conditions required of 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 in [10], and an empirical comparison of the limiting distributions of the kernel-based statistic.

Additional independence testing approaches also exist in the statistics literature. For $d = d' = 1$, an early nonparametric test for independence, due to Hoeffding, Blum, Kiefer, and Rosenblatt [12, 13], 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 paper is organized as follows. Section 1 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

[3] A strong consistent test means that both on $\mathcal{H}_0$ and on its complement the test makes a.s. no error after a random sample size. This concept is close to the definition of discernability introduced by Dembo and Peres [5]. See [1] for further discussion.

and an asymptotically α-level test are presented for the log-likelihood statistic in Section 2, and a conjecture for an asymptotically α-level test based on the Pearson χ^2 statistic is described in Section 3. Section 4 contains a review of kernel-based independence statistics, and 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. More detailed proofs and further discussion may be found in an associated technical report [14].

1 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.

1.1 Large deviation properties

For testing a simple hypothesis versus a composite alternative, Györfi and van der Meulen [2] introduced a related goodness of fit test statistic L_n defined as

$$L_n(\mu_n, \mu) = \sum_{A \in \mathcal{P}_n} |\mu_n(A) - \mu(A)|.$$

Beirlant [15], and Biau and Györfi [1] proved that, for all $0 < \varepsilon$,

$$\mathbf{P}\{L_n(\mu_n, \mu) > \varepsilon\} \leq 2^{m_n} e^{-n\varepsilon^2/2}. \tag{1}$$

We now show that a similar result follows quite straightforwardly for the L_1 independence statistic.

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

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

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

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

The central term in the sum is zero under the null hypothesis. The proof is then completed by further applications of the triangle inequality, then using (1) on the resulting terms, and applying a union bound. ■

Theorem 1 yields a strong consistent test of homogeneity, 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. The proof of the following corollary is similar to that employed for the homogeneity test by Biau and Györfi [1].

Corollary 1. *Consider the test which rejects $\mathcal{H}_0$ when*

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

where $c_1 > \sqrt{2 \ln 2} \approx 1.177$. Assume conditions

$$\lim_{n \to \infty} m_n m_n'/n = 0, \qquad \lim_{n \to \infty} m_n / \ln n = \infty, \qquad \lim_{n \to \infty} m_n' / \ln n = \infty, \tag{2}$$

are satisfied. Then under $\mathcal{H}_0$, the test makes a.s. no error after a random sample size. Moreover, if $\nu \neq \mu_1 \times \mu_2$, and for any sphere S centered at the origin,

$$\lim_{n \to \infty} \max_{A \in \mathcal{P}_n,\, A \cap S \neq 0} \operatorname{diam}(A) = 0 \quad \text{and} \quad \lim_{n \to \infty} \max_{B \in Q_n,\, B \cap S \neq 0} \operatorname{diam}(B) = 0, \tag{3}$$

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

1.2 Asymptotic normality

Beirlant et al. [3] proved, under conditions

$$\lim_{n \to \infty} m_n = \infty, \qquad \lim_{n \to \infty} m_n / n = 0, \qquad \lim_{n \to \infty} \max_{j=1,\ldots,m_n} \mu(A_{nj}) = 0, \tag{4}$$

that

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

where $\xrightarrow{\mathcal{D}}$ stands for the convergence in distribution and $\sigma^2 = 1 - 2/\pi$. We adapt this proof to the case of independence testing (see Appendix for details).

Theorem 2. *Assume that conditions (2) and*

$$\lim_{n \to \infty} \max_{A \in \mathcal{P}_n} \mu_1(A) = 0, \quad \lim_{n \to \infty} \max_{B \in Q_n} \mu_2(B) = 0, \tag{5}$$

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

$$\sqrt{n} \left(L_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) - C_n \right) / \sigma \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \quad \text{where} \quad \sigma^2 = 1 - 2/\pi.$$

Theorem 2 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 1, and because of condition (4), this new test is *not* distribution-free. In particular, the measures μ_1 and μ_2 have to be nonatomic. The corollary below follows from Theorem 2, replacing C_n with the upper bound

$$C_n \;\le\; \sqrt{2m_n m_n'/(\pi n)}$$

(the original expression for C_n is provided in the Appendix, eq. (20)).

Corollary 2. *Let $\alpha \in (0,1)$. Consider the test which rejects $\mathcal{H}_0$ when*

$$L_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) > c_2\sqrt{m_n m_n'/n} + \sigma/\sqrt{n}\,\Phi^{-1}(1-\alpha) \approx c_2\sqrt{m_n m_n'/n},$$

where $\sigma^2 = 1 - 2/\pi$, $c_2 = \sqrt{2/\pi} \approx 0.798$, and Φ denotes the standard normal distribution function. Then, under the conditions of Theorem 2, the test has asymptotic significance level α. Moreover, under the additional conditions (3), the test is consistent.

2 Log-likelihood statistic

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

$$I_n(\mu_n, \mu) = 2\sum_{j=1}^{m_n} \mu_n(A_{n,j}) \log\left[\mu_n(A_{n,j})/\mu(A_{n,j})\right],$$

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

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

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

$$I_n(\nu_n, \nu) = 2\sum_{A\in\mathcal{P}_n}\sum_{B\in Q_n} \nu_n(A\times B)\log\left[\nu_n(A\times B)/\nu(A\times B)\right],$$

since under the null hypothesis it can easily be seen that

$$I_n(\nu_n, \nu) - I_n(\nu_n, \mu_{n,1}\times\mu_{n,2}) = I_n(\mu_{n,1}, \mu_1) + I_n(\mu_{n,2}, \mu_2) \ge 0.$$

For the large deviation bound, Kallenberg [16], and Quine and Robinson [17] proved that, for all $\epsilon > 0$,

$$\mathbf{P}\{I_n(\mu_n,\mu)/2 > \epsilon\} \le \binom{n+m_n-1}{m_n-1} e^{-n\epsilon} \le e^{m_n\log(n+m_n)-n\epsilon}.$$

Therefore under the condition $m_n \log n = o(n)$, which is stronger than (4),

$$\mathbf{P}\{I_n(\mu_n,\mu)/2 > \epsilon\} = e^{-n(\epsilon+o(1))}. \tag{6}$$

A test based on this result can be introduced which rejects independence if

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

Under $\mathcal{H}_0$, we obtain the non-asymptotic bound

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

Therefore condition (2) implies

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

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 Pinsker's inequality,

$$L_n^2(\nu_n, \mu_{n,1} \times \mu_{n,2}) \leq I_n(\nu_n, \mu_{n,1} \times \mu_{n,2}). \tag{7}$$

Concerning the limit distribution, Györfi and Vajda [4] proved under (4),

$$(nI_n(\mu_n, \mu) - m_n)\,(2m_n)^{-1/2} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1).$$

This implies that for any real valued x, under conditions (2) and (5),

$$\mathbf{P}\left\{\frac{nI_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) - m_n m'_n}{\sqrt{2m_n m'_n}} \leq x\right\} \leq \mathbf{P}\left\{\frac{nI_n(\nu_n, \nu) - m_n m'_n}{\sqrt{2m_n m'_n}} \leq x\right\} \to \Phi(x),$$

from which an asymptotically α-level test follows straightforwardly.

3 Pearson χ^2 statistic

Another statistic for testing independence is the Pearson χ^2 test statistic,

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

For the associated goodness of fit test, Quine and Robinson [17] provide a large deviation bound for the statistic

$$\chi_n^2(\mu_n, \mu) = \sum_{j=1}^{m_n} \frac{(\mu_n(A_{n,j}) - \mu(A_{n,j}))^2}{\mu(A_{n,j})}, \tag{8}$$

from which it may be possible to obtain a strong consistent distribution-free test of independence. The asymptotic distribution of (8) under conditions (4) was addressed by Györfi and Vajda [4], who proved

$$\left(n\chi_n^2(\mu_n, \mu) - m_n\right)(2m_n)^{-1/2} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1).$$

We conjecture that under the conditions (2) and (5),

$$\left(n\chi_n^2(\nu_n, \mu_{n,1} \times \mu_{n,2}) - m_n m'_n\right)(2m_n m'_n)^{-1/2} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1),$$

from which an asymptotically α-level test follows.

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 [7], 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

$$f_n(x, y) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) K'_h(y - Y_i) \text{ and } f_{n,1}(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i), \quad (9)$$

with $f_{n,2}(y)$ defined by analogy. Rosenblatt [7] introduced the kernel-based independence statistic

$$T_n = \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} (f_n(x, y) - f_{n,1}(x) f_{n,2}(y))^2 dx\, dy. \quad (10)$$

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

$$\begin{aligned} T_n = & \tfrac{1}{n^2} \textstyle\sum_{i=1}^{n} \sum_{j=1}^{n} L_h(X_i - X_j) L'_h(Y_i - Y_j) \\ & - \tfrac{2}{n^3} \textstyle\sum_{i=1}^{n} \left(\sum_{j=1}^{n} L_h(X_i - X_j) \right) \left(\sum_{j=1}^{n} L'_h(Y_i - Y_j) \right) \\ & + \left(\tfrac{1}{n^2} \textstyle\sum_{i=1}^{n} \sum_{j=1}^{n} L_h(X_i - X_j) \right) \left(\tfrac{1}{n^2} \textstyle\sum_{i=1}^{n} \sum_{j=1}^{n} L'_h(Y_i - Y_j) \right). \end{aligned} \quad (11)$$

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

$$\begin{aligned} \mathbf{E}\{T_n\} &= \frac{n-1}{n^2} \left(L_h(0) - \mathbf{E}\{L_h(X_1 - X_2)\} \right) \left(L'_h(0) - \mathbf{E}\{L'_h(Y_1 - Y_2)\} \right) & (12) \\ &\leq n^{-1} L_h(0) L'_h(0) \; = \; (n h^d h^{d'})^{-1} \|K\|^2 \|K'\|^2. & (13) \end{aligned}$$

A second interpretation of the above statistic is as a smoothed difference between the joint characteristic function and the product of the marginals [8]. 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 [9]. 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 [7, conditions a.1-a.4].

A generalization of the statistic to include non-Euclidean domains is presented by Gretton et al. [10, 11]. The test statistic in (11) is then interpreted as a biased empirical estimate of the Hilbert-Schmidt norm of a cross-covariance operator between reproducing kernel Hilbert spaces (RKHS),[4] $\|C_{xy}\|^2_{\mathrm{HS}}$. 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* (in the sense of Fukumizu et al. [18]; see also Sriperumbudur et al. [19]), then $\|C_{xy}\|^2_{\mathrm{HS}} = 0$ iff X and Y independent: these kernels need not be inner products of square integrable probability density functions. The Gaussian kernel is characteristic on $\mathbb{R}^d$ [18], and universal kernels (in the sense of Steinwart [20]) are characteristic on compact domains [10]. Note that universal kernels exist that may not be written as inner products of kernel density functions: see examples in [20, Section 3]. An appropriate choice of kernels allows testing of dependence in general settings, such as distributions on strings and graphs [11].

4.1 Large deviation property

The empirical statistic T_n was previously shown by Gretton et al. [10] to converge in probability to its expectation with rate $1/\sqrt{n}$. We now provide a more refined bound, which is tighter when the bandwidth h decreases. We will obtain our results for the statistic

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

since under the null hypothesis,

$$\begin{aligned}\sqrt{T_n} &= \|f_n(\cdot,\cdot) - f_{n,1}(\cdot)f_{n,2}(\cdot)\| \\ &\leq \sqrt{\tilde{T}_n} + \|f_{n,1}(\cdot)\|\,\|f_{n,2}(\cdot) - \mathbf{E}f_{n,2}(\cdot)\| + \|f_{n,1}(\cdot) - \mathbf{E}f_{n,1}(\cdot)\|\,\|\mathbf{E}f_{n,2}(\cdot)\| \approx \sqrt{\tilde{T}_n}.\end{aligned} \tag{14}$$

Theorem 3. *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 McDiarmid's inequality [21]: 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

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

[4] Given RKHSs $\mathcal{F}$ and $\mathcal{G}$, the cross-covariance operator $C_{xy} : \mathcal{G} \to \mathcal{F}$ for the measure ν is defined such that for all $f \in \mathcal{F}$ and $g \in \mathcal{G}$, $\langle f, C_{xy}g\rangle_{\mathcal{F}} = \mathbf{E}\left\{[f(x) - \mathbf{E}\{f(x)\}]\,[g(y) - \mathbf{E}\{g(y)\}]\right\}$.

Then, for all $\epsilon > 0$,

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

Given the function $f = \sqrt{\tilde{T}_n}$, the result follows from

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

From these inequalities we can derive a test of independence. Given ϵ such that $n\epsilon^2/(2L_h(0)L_h'(0)) = 2\ln n$, and recalling (13) and (14), a strongly consistent test rejects independence if

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

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. If h is fixed, the strong law of large numbers implies $T_n \to \|C_{xy}\|^2_{\mathrm{HS}} > 0$ for characteristic kernels, and the test is strongly consistent. In both cases the strong consistency is not distribution-free.

4.2 Limit distribution

We now describe the asymptotic limit distribution of the test statistic T_n in (11). 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 (9). The corresponding density estimates in T_n are mean square consistent if $h = h_n$ such that

$$h_n \to 0 \quad \text{and} \quad nh_n^d h_n^{d'} \to \infty. \tag{15}$$

Based on the results in [22, 23, 24], we expect T_n to be asymptotically normally distributed. For an independence test, we require $\mathrm{var}(T_n) \approx \mathrm{var}(\tilde{T}_n)$. If $h \to 0$,

$$\mathrm{var}(\tilde{T}_n) \approx 2\|f\|^2 n^{-2}h^{-d}h^{-d'}. \tag{16}$$

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 $\sigma^2 = 2\|f\|^2$. While an α-level test may be obtained by replacing $\mathbf{E}\{T_n\}$ with its upper bound in (13), the resulting threshold is still not distribution-free, since σ depends on the unknown f. The simplest distribution-free bound for the variance, $\sigma^2 \leq \|K\|^4\|K'\|^4 n^{-2}h^{-2d}h^{-2d'}$, is unsatisfactory since its performance as a function of h is worse than the result in (16). An improved distribution-free bound on the variance is a topic for future research: we give an empirical estimate below (eq. 18) for use in asymptotic hypothesis tests.

We now consider the case of fixed h. Following [8], the distribution of T_n under $\mathcal{H}_0$ is

$$nT_n \xrightarrow{D} \sum_{l=1}^{\infty} \lambda_l z_l^2, \tag{17}$$

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 [11, Theorem 2]). A difficulty in using the statistic (11) in a hypothesis test therefore arises due to the form of the null distribution, 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 (15), we may use an empirical estimate of the variance of T_n under $\mathcal{H}_0$ due to Gretton et al. [11, Theorem 4]. Denoting by $\odot$ the entrywise matrix product and $A^{\cdot 2}$ the entrywise matrix power,

$$\mathrm{var}(T_n) = \mathbf{1}^\top \left(\mathbf{B} - \mathrm{diag}(\mathbf{B})\right) \mathbf{1}, \quad \text{where} \quad \mathbf{B} = \left((\mathbf{HLH}) \odot (\mathbf{HL'H})\right)^{\cdot 2}, \tag{18}$$

$\mathbf{L}$ 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 quantiles of the null distribution (17) for hypothesis testing: repeated shuffling of the sample [8], and approximation by a two-parameter Gamma density [9],

$$nT_n \sim x^{\alpha-1} e^{-x/\beta} / \left(\beta^\alpha \Gamma(\alpha)\right), \quad \alpha = (\mathbf{E}\{T_n\})^2 / \mathrm{var}(T_n), \quad \beta = n\mathrm{var}(T_n) / \mathbf{E}\{T_n\},$$

using $\mathbf{E}\{T_n\}$ from (12). This Gamma approximation was found by [11] to perform identically on the Section 5 benchmark data to the more computationally expensive approach of Feuerverger [8]. We emphasize, however, that this approximation is a heuristic, with no guarantees on asymptotic performance.

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 kernel size 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 close to the Gamma approximation for small kernel sizes, but is less accurate for larger kernel sizes (where "small" and "large" will depend on the measure ν).

5 Experiments

In comparing the independence tests, we made use of the multidimensional benchmark data proposed by Gretton et al. [11]. We tested the independence of both one-dimensional and two-dimensional random variables (i.e. $d = d' = 1$ and $d = d' = 2$). The data were constructed as follows. First, we generated n samples of two univariate random variables, each drawn at random from the

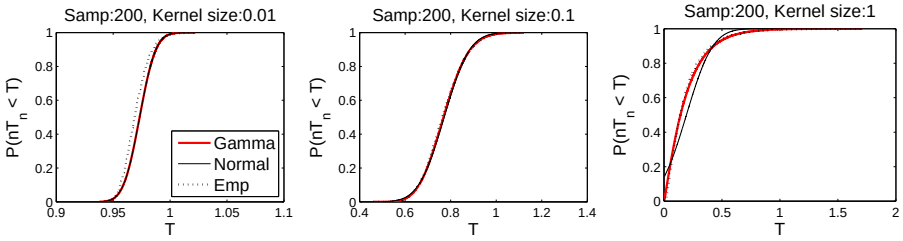

Fig. 1. Simulated cumulative distribution function of T_n (*Emp*) under $\mathcal{H}_0$ for $n = 200$, compared with the two-parameter Gamma distribution (*Gamma*) and the Normal distribution (*Normal*). The empirical CDF was obtained using 5000 independent draws of T_n.

ICA benchmark densities in Figure 5 of Bach and Jordan [25]: these included super-Gaussian, sub-Gaussian, multimodal, and unimodal distributions. 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 case of $d = 2$, a second dimension was appended to each of the mixed variables, consisting of independent Gaussian noise of zero mean and unit standard deviation; and each resulting vector was multiplied by an independent random two-dimensional orthogonal matrix, to obtain vectors 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 $n = 128, 512$.

We compared three different asymptotic independence testing approaches based on space partitioning: the L_1 test, denoted *L1*; the Pearson χ^2 test *Pears*; and the log likelihood test *Like*. The number of discretisations per dimension was set at $m_n = m'_n = 4$, besides in the $n = 128, d = 2$ case, where it was set at $m_n = m'_n = 3$: in the latter case, there were too few samples per bin when a greater number of partitions were used. 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 also compared 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 h and h' set to the median distances between samples of the respective variables, following Gretton et al. [11].

Results are plotted in Figure 2 (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, dependence becomes easier to detect as θ increases from 0 to $\pi/4$, when

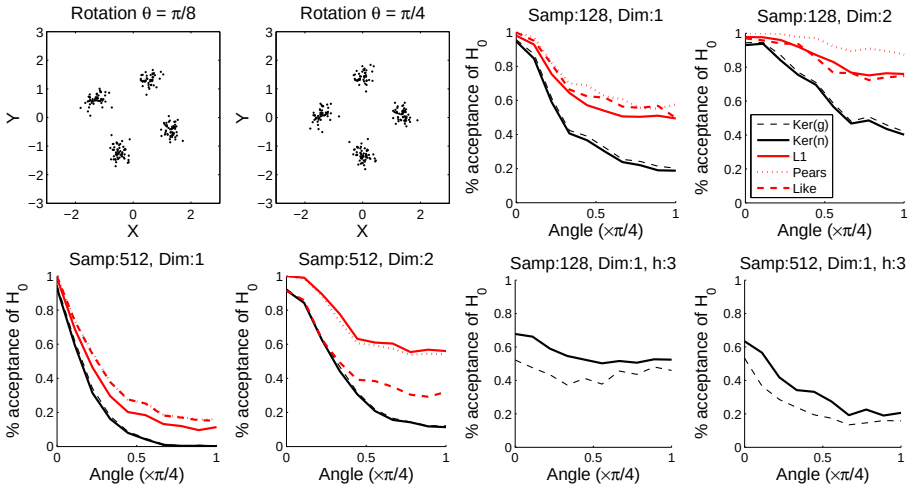

Fig. 2. Top left plots: Example dataset for $d = d' = 1$, $n = 200$, and rotation angles $\theta = \pi/8$ (left) and $\theta = \pi/4$ (right). In this case, both distributions prior to rotation are mixtures of two Gaussians. **Next four plots:** Rate of acceptance of $\mathcal{H}_0$ for the *PD*, *fCorr*, *HSICp*, and *HSICg* tests. "Samp" is the number n of samples, and "dim" is the dimension $d = d'$ of x and y. **Bottom right plots** Performance of the *Ker(g)* and *Ker(n)* tests for a large kernel size $h = 3$, and $\alpha = 0.5$, to show the difference between the Normal and two-parameter Gamma approximations to the null distribution.

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 $n = 128,\ d = 2$). For smaller numbers of samples ($n = 128$), the L_1 test performs the same as or slightly better than the log likelihood test; the Pearson χ^2 test always performs worst. For larger numbers of samples ($n = 512$), the L_1 test has a slight advantage at $d = 1$, but the log-likelihood test displays far better performance for $d = 2$. The superior performance of the log-likelihood test compared with the χ^2 test 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, as compared with the χ^2 statistic, in terms of the dependence of the latter on the number m_n of partitions used [15]. In all cases, the kernel-based test outperforms the remaining methods, and behaviour under the Normal and Gamma null distribution models is virtually identical. That said, we should bear in mind the kernel test thresholds require $\mathbf{E}\{T_n\}$ and $\mathrm{var}(T_n)$, which are unknown and must be estimated from the data: thus, unlike the L_1, χ^2, and log likelihood tests, the kernel test thresholds are not distribution-independent.

It is of interest to further investigate the null distribution approximation strategies for the kernel tests. We used an artificially high kernel bandwidth $h = 3$, and a lower $\alpha = 0.5$, to make visible the performance difference. Results are shown in the final row of Figure 2. In accordance with Figure 1, the Gaussian approximation yields a larger threshold than the true CDF would require, and consequently has a Type I error below the design level α.

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.

References

[1] Biau, G., Györfi, L.: On the asymptotic properties of a nonparametric l_1-test statistic of homogeneity. IEEE Trans. Inform. Theory **51** (2005) 3965–3973
[2] Györfi, L., van der Meulen, E.C.: A consistent goodness of fit test based on the total variation distance. In Roussas, G., ed.: Nonparametric Functional Estimation and Related Topics, Kluwer, Dordrecht (1990) 631–645
[3] Beirlant, J., Györfi, L., Lugosi, G.: On the asymptotic normality of the l_1- and l_2-errors in histogram density estimation. Canad. J. Statist. **22** (1994) 309–318
[4] Györfi, L., Vajda, I.: Asymptotic distributions for goodness of fit statistics in a sequence of multinomial models. Stat. Prob. Lett. **56** (2002) 57–67
[5] Dembo, A., Peres, Y.: A topological criterion for hypothesis testing. Ann. Statist. **22** (1994) 106–117
[6] Read, T., Cressie, N.: Goodness-Of-Fit Statistics for Discrete Multivariate Analysis. Springer-Verlag, New York (1988)
[7] Rosenblatt, M.: A quadratic measure of deviation of two-dimensional density estimates and a test of independence. The Annals of Statistics **3**(1) (1975) 1–14
[8] Feuerverger, A.: A consistent test for bivariate dependence. International Statistical Review **61**(3) (1993) 419–433
[9] Kankainen, A.: Consistent Testing of Total Independence Based on the Empirical Characteristic Function. PhD thesis, University of Jyväskylä (1995)
[10] Gretton, A., Bousquet, O., Smola, A., Schölkopf, B.: Measuring statistical dependence with Hilbert-Schmidt norms. In: ALT. (2005) 63–78
[11] Gretton, A., Fukumizu, K., Teo, C.H., Song, L., Schölkopf, B., Smola, A.: A kernel statistical test of independence. In: NIPS 20. (2008)
[12] Hoeffding, W.: A nonparametric test for independence. The Annals of Mathematical Statistics **19**(4) (1948) 546–557
[13] Blum, J.R., Kiefer, J., Rosenblatt, M.: Distribution free tests of independence based on the sample distribution function. Ann. Math. Stat. **32** (1961) 485–498
[14] Gretton, A., Györfi, L.: Consistent nonparametric tests of independence. Technical Report 172, MPI for Biological Cybernetics (2008)
[15] Beirlant, J., Devroye, L., Györfi, L., Vajda, I.: Large deviations of divergence measures on partitions. J. Statist. Plan. Inference **93** (2001) 1–16
[16] Kallenberg, W.C.M.: On moderate and large deviations in multinomial distributions. Annals of Statistics **13** (1985) 1554–1580
[17] Quine, M., Robinson, J.: Efficiencies of chi-square and likelihood ratio goodness-of-fit tests. Ann. Statist. **13** (1985) 727–742
[18] Fukumizu, K., Gretton, A., Sun, X., Schölkopf, B.: Kernel measures of conditional dependence. In: NIPS 20. (2008)
[19] Sriperumbudur, B.K., Gretton, A., Fukumizu, K., Lanckriet, G.R.G., Schölkopf, B.: Injective hilbert space embeddings of probability measures. In: COLT. (2008) 111–122
[20] Steinwart, I.: On the influence of the kernel on the consistency of support vector machines. Journal of Machine Learning Research **2** (2001) 67–93
[21] McDiarmid, C.: On the method of bounded differences. In: Survey in Combinatorics. Cambridge University Press (1989) 148–188

[22] Hall, P.: Central limit theorem for integrated square error of multivariate nonparametric density estimators. Journal of Multivariate Analysis **14** (1984) 1–16
[23] Cotterill, D.S., Csörgő, M.: On the limiting distribution of and critical values for the Hoeffding, Blum, Kiefer, Rosenblatt independence criterion. Statistics and Decisions **3** (1985) 1–48
[24] Beirlant, J., Mason, D.M.: On the asymptotic normality of l_p-norms of empirical functionals. Math. Methods Statist. **4** (1995) 1–19
[25] Bach, F.R., Jordan, M.I.: Kernel independent component analysis. J. Mach. Learn. Res. **3** (2002) 1–48

A Proof of Theorem 2

The main difficulty in proving Theorem 2 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 (see [3] for details). 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. Further define ν_{N_n}, $\mu_{N_n,1}$ and $\mu_{N_n,2}$ as the Poissonized version of the empirical measures associated with the samples $\{(X_i, Y_i)\}$, $\{X_i\}$ and $\{Y_i\}$, respectively,

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

for any Borel subsets A and B. Clearly, $n\nu_{N_n}(A\times B)$, $n\mu_{N_n,1}(A)$, and $n\mu_{N_n,2}(B)$ are Poisson random variables. 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

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

Key to the proof of Theorem 2 is the following result, which extends the proposition of [3, p. 311].

Proposition 1. *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

$$\left(M_n, \frac{N_n - n}{\sqrt{n}}\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\sigma^2 & 0\\ 0 & 1\end{bmatrix}\right) \tag{19}$$

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

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

The proof of the proposition is a simple extension of that by Beirlant et al. for the goodness-of-fit case [3, pp. 311–313]. We now turn to the proof of Theorem 2.

Proof (*Theorem 2, sketch only*) We will show the theorem with the centering constant

$$C_n = \mathbf{E}\{\tilde{L}_n(\nu_n, \mu_{n,1}\times\mu_{n,2})\} = \sum_{A\in\mathcal{P}_n}\sum_{B\in Q_n}\mathbf{E}\{|\nu_{N_n}(A\times B) - \mu_{N_n,1}(A)\cdot\mu_{N_n,2}(B)|\}. \tag{20}$$

Define

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

Our goal is to prove that the assumption in (19) holds. In particular (see [3, 1] for details), we require a central limit result to hold for the Poissonized statistic

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

Once we obtain $\mathrm{var}(S_n)$, the asymptotic normality in (19) can be proved by verifying the Lyapunov conditions as in Beirlant et al. [3]. We have that

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

and therefore the variance of S_n is

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

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

$$\begin{aligned} \nu_{N_n}(A\times B) &\overset{\mathcal{D}}{\approx} Z_{A\times B}\sqrt{\mu_1(A)\mu_2(B)/n} + \mu_1(A)\mu_2(B), \\ \mu_{N_n,1}(A) &\overset{\mathcal{D}}{\approx} Z_A\sqrt{\mu_1(A)/n} + \mu_1(A), \end{aligned}$$

with $\mu_{N_n,2}(B)$ and Z_B defined by analogy. Making these substitutions and simplifying,

$$\mathrm{var}(S_n) \approx t^2(1 - 2/\pi) + v^2.$$

■