Chapter 2
Power Assessment of a New Test of Independence

P.N. Patil and D. Bagkavos

Abstract A new nonparametric test of independence between the components of bivariate random vectors (X, Y) is motivated and evaluated in practice. The test statistics is based on the fact that under independence, every quantile of Y given $X = x$ is constant. This is in contrast to the most commonly used basis that the joint probability density or distribution function of X and Y, equals to the product of their marginal probability densities or distributions, respectively. Emphasis is given on the small sample power properties of the test. Through numeric simulations with distributional data, the power of the test is benchmarked against standard independence tests, already existing in the literature.

Keywords Independence • Hypothesis test • Statistical power • Test size Quantile regression

2.1 Introduction

The present note is concerned with the general problem of testing the stochastic independence between the components of bivariate random vectors. Historically, the Cramer–Von Mises distance measure has provided the basis for several hypothesis tests on this topic, e.g., [2,4–6,11], and [12]. See also [14] for a broader view of the subject.

A test of independence which originates from a different basis is investigated here. The concept, first considered in [10] and further explored in [3] is that independence is implied if every regression quantile of Y versus $X = x$ is constant. Under the linear quantile regression framework, c.f. [1], this idea is put to action by applying density weighted conditional expectation on the first order condition for minimization of the least absolute deviation criterion and integrating across

P.N. Patil
School of Mathematics, University of Birmingham, Edgbaston, Birmingham, B152TT, UK

D. Bagkavos (✉)
Accenture, Rostoviou 39–41, 11526, Athens, Greece
e-mail: dimitrios.bagkavos@gmail.com

M.G. Akritas et al. (eds.), *Topics in Nonparametric Statistics*, Springer Proceedings in Mathematics & Statistics 74, DOI 10.1007/978-1-4939-0569-0_2

all possible quantiles. This results in a new condition which under independence between X and Y equals to zero, while otherwise takes large positive values. Naturally, this provides a proper candidate for developing a consistent independence test with its sample, kernel based, analogue as the proposed test statistic.

The purpose of this note is to provide insight on the practical performance of the suggested hypothesis test. Specifically, the small sample distribution of the test statistic under the null is approximated and then utilized in calculating the power of the test as a function of the level of correlation between X and Y. Furthermore, the power functions of the tests developed in [2] and [6] are used as a benchmark in assessing the practical performance of the test presented here. We note here that the test's theoretical properties, including its asymptotic distribution under both the null and alternative hypotheses, establishment of its consistency against all dependence alternatives as well as a bandwidth choice rule which controls the trade-off between the test's power and size functions will be provided in future work.

The rest of the paper is organized as follows. Section 2.2 discusses the development of the test and provides the test statistic. Numerical evidence on the power of the test and comparison with the powers of the [2] and [6] tests is given in Sect. 2.3.

2.2 Motivation and Test Statistic

Let $(X, Y) \in \mathbb{R} \times \mathbb{R}$ be a random variable with cumulative distribution function $F(x, y)$ and probability density function $f(x, y)$. Denote with $F_Y(y|x)$ the marginal distribution of Y conditional on $X = x$ and with $F_Y^{-1}(y|x)$ its inverse. The marginal (unconditional) distribution of Y is denoted by $F_Y(y)$ and by $F_Y^{-1}(y)$ its inverse. Obviously under independence between X and Y, $F_Y(y|x) = F_Y(y)$ and $F_Y^{-1}(y|x) = F_Y^{-1}(y)$.

The basis of the proposed test is that under independence, for every quantile p, where $0 < p < 1$, we have that $F_Y^{-1}(p|x) = c_p$ where c_p does not vary with x.

For example, $F_Y^{-1}(p|x)$ can be modeled (see also [1]) by

$$F_Y^{-1}(p|x) = \beta x + F_Y^{-1}(p). \tag{2.1}$$

Under independence between X and Y, for any fixed $p \in (0, 1)$ the conditional quantile function of Y given $X = x$ does not depend on x and therefore

$$F_Y^{-1}(p|x) = F_Y^{-1}(p).$$

Now, let $\psi_p(u) = \text{sign}(u) + 2p - 1$ and let (X_1, Y_1) and (X_2, Y_2) be two independent random vectors with common probability density function $f(x, y)$. Now set

$$J(p) = \mathbb{E}\left\{K_h(X_1, X_2)\psi_p\left(Y_2 - F_Y^{-1}(p)\right)\psi_p\left(Y_1 - F_Y^{-1}(p)\right)\right\}$$

where $K_h(X_1, X_2) = h^{-1}K((X_1 - X_2)h^{-1})$, K is a second order kernel and h denotes bandwidth, i.e., the spread of the kernel. Observe that for every fixed $p \in (0, 1)$, under the hypothesis of independence of X and Y,

$$g_p(x) = E\{\psi_p(Y_1 - F_Y^{-1}(p))|X_1 = x\} = 0.$$

Therefore, under independence of X and Y, for every $p \in (0,1)$ we have,

$$J(p) = \mathbb{E}\left\{K_h(X_1, X_2)\psi_p(Y_2 - F_Y^{-1}(p))\mathbb{E}\{\psi_p(Y_1 - F_Y^{-1}(p))|X_1\}\right\} = 0$$

and consequently

$$J = \int_0^1 J(p)\, dp = 0.$$

Under dependence we have

$$g_p(X_1) = E\{\psi_p(Y_1 - F_Y^{-1}(p))|X_1\} \neq 0 \text{ a.s.}$$

for at least one $p \in (0,1)$. Assuming that g_p is twice differentiable,

$$\begin{aligned} J(p) &= \mathbb{E}\left\{K_h(X_1, X_2)\psi_p\left(Y_2 - F_Y^{-1}(p)\right)\psi_p\left(Y_1 - F_Y^{-1}(p)\right)\right\} \\ &= \mathbb{E}\left\{K_h(X_1, X_2)\mathbb{E}\left\{\psi_p\left(Y_2 - F_Y^{-1}(p)\right)\psi_p\left(Y_1 - F_Y^{-1}(p)\right)|(X_1, X_2)\right\}\right\} \\ &= \int_0^1 g_p^2(x) f_X^2(x) dx + O(h^2). \end{aligned}$$

Thus as $h \to 0$, $J(p) > 0$ and since $J(p)$ is a continuous function of p,

$$J = \int_0^1 J(p) dp > 0.$$

Therefore, J can be used for the following hypothesis test

$$H_0 : \bigcap_{0<p<1} H_{0p}, \quad H_{0p} : F_Y^{-1}(p|x) = c_p$$

with the alternative specified by

$$H_1 : \bigcup_{0<p<1} H_{1p}, \quad H_{1p} : F_Y^{-1}(p|x) = c(x)$$

where now $c(x)$ varies with x for at least one $p \in (0,1)$.

For deriving a test statistic, assume a sample $(X_i, Y_i), i = 1, \ldots, n$ from $F(x, y)$ and denote by $\hat{F}_Y^{-1}(y)$ the inverse of the empirical marginal distribution of Y, $\hat{F}_Y(y)$. The density weighted conditional expectation

$$\mathbb{E}\left\{\psi_p\left(Y - F_Y^{-1}(p)\right)|x\right\} f_X(x),$$

can be reasonably estimated (fixing $x = X_i$) by

$$\frac{1}{(n-1)h} \sum_{j=1,\, j\neq i}^{n} K\left(\frac{X_i - X_j}{h}\right)\psi_p(Y_j - \hat{F}^{-1}(p)) \tag{2.2}$$

where the real valued function K is called kernel and integrates to 1, while h is called bandwidth and controls the spread of the kernel and therefore the amount of smoothing applied. Based on (2.2), the sample version of J is

$$T_n(h) = \int_0^1 \frac{1}{n}\sum_{i=1}^{n} \psi_p(Y_i - \hat{F}_Y^{-1}(p)) \frac{1}{(n-1)h} \sum_{j\neq i} K\left(\frac{X_i - X_j}{h}\right) \psi_p(Y_j - \hat{F}^{-1}(p))\, dp$$

$$= \frac{1}{n(n-1)h} \sum_{1\leq i<j\leq n}\sum K\left(\frac{X_i - X_j}{h}\right) \int_0^1 \psi_p\left(Y_i - \hat{F}_Y^{-1}(p)\right) \psi_p(Y_j - \hat{F}^{-1}(p))\, dp$$

which is also the proposed test statistic. By denoting with R_{ni} the rank of Y_i after ordering the random sample $(X_i, Y_i), i = 1, 2, \ldots, n$ with respect to Y_i and after slightly modifying the definition of the empirical distribution function from $\hat{F}_Y$ to $n(n+1)^{-1}\hat{F}_Y$, a suitable for computational purposes form of the statistic is

$$T_n(h) = \frac{2}{n(n-1)h} \sum_{1\leq i<j\leq n}\sum K\left(\frac{X_i - X_j}{h}\right)$$
$$\times \left\{ \frac{\min(R_{ni}, R_{nj})^2}{(n+1)^2} + \frac{n - \max(R_{ni}, R_{nj})^2}{(n+1)^2} - \frac{1}{3} \right\}. \tag{2.3}$$

The next section discusses the test's operational characteristics and provides numerical evidence on its power.

2.3 Numerical Evaluation of the Test's Power

In this section, the implementation details of the suggested test are discussed and then distributional data is used to exhibit the performance of the proposed test's power properties and asses its practical performance.

Throughout this section $T_n(h)$ is calculated on bivariate samples of size 50 by (2.3) with K being the uniform kernel. The bandwidth, h, is chosen so as to maximize the test's power under the null, subject to keeping the significance level constant. Specifically, in each $T_n(h)$ implementation, the bandwidth is given by

$$h = ch^*. \tag{2.4}$$

In (2.4), for each given sample, h^* is the optimal MISE regression bandwidth of [13], implemented in package `lokern, R`. The factor c is determined by a grid search in a probe analysis and applies to all bandwidth calculations with samples from the same distribution.

The probe analysis is designed to find c so that the test's size under the null matches the user supplied level confidence level a. Specifically, for each of the 30

equidistant c points in $(0, 3)$, 10,000 test statistic values, $T_n(ch^*)$, under the null are calculated. The probability $a(ch^*) = P(T_n(ch^*) > l_a)$ is then calculated where l_a is the $(1-a)100\,\%$ quantile of the 10,000 T_n values, calculated as described in the next paragraph. The c that corresponds to the highest among the 30 $a(ch^*)$'s is selected and used in (2.4).

The bandwidth choice rule employed here implicitly works as a selection rule based on bootstrap/edgeworth expansion ideas considered in the past by [8]. The benefit of such an approach is that it offers the means to control both power and size. An additional advantage of this bandwidth procedure is that it offers the best bandwidth both under the null and the alternative hypothesis.

As a cut-off point in the test's power function approximation, we use the $(1-a)100\,\%$ quantile of the numerical distribution of $T_n(h)$. For this purpose, definition 7 of [7], which is readily implemented in `R` by the `quantile()` function, is applied on 100,000 $T_n(h)$ values. The $T_n(h)$ values result by applying (2.3), implemented as described in the previous paragraph, on 100,000 uncorrelated bivariate samples.The desired number of uncorrelated samples is obtained by repeatedly generating bivariate samples $(X_i, Z_i), i = 1, \ldots, 50$ and keeping only those for which the correlation between X and Z is less than 0.001.

Then, the power function of $T_n(h)$ is approximated by

$$P(T_n(h) > \text{cut} - \text{off}) = \frac{\#T_n(h) > \text{cut} - \text{off}}{m}, \tag{2.5}$$

for $m = 100$ replications. Specifically, 40 equidistant correlation levels, ρ, between 0 and 1 are determined. For each ρ, 100 independent (i.e., $\text{corr}(X, Z) < 0.001$) bivariate samples $(X_i, Z_i), i = 1, \ldots, 50$ are drawn. The actual samples used by $T_n(h)$ are $(X_i, Y_i), i = 1, \ldots, 50$ where the Y_i's are obtained by the transformation

$$Y = \rho X + Z\sqrt{1-\rho^2}.$$

The provision of drawing samples with $\text{corr}(X, Z) < 0.001$ is sought because this ensures that the above transformation will return samples (X_i, Y_i) with the desired level of correlation. Then for each ρ, $T_n(h)$ is calculated 100 times using the (X_i, Y_i) samples and the empirical power is calculated by (2.5).

The tests of [2] (noted as B_n) and [6] (noted as D_n) are used for benchmarking $T_n(h)'$s behavior. The B_n test is calculated as described in [9, p. 43]. Its power function is approximated by (2.5) with $T_n(h)$ replaced by B_n and with cut-off points found in Table 2 of [9]. The D_n test is implemented by the `hoeffd` function of `R` (package `Hmisc`) which also returns its p-value for the given sample. It's power function is approximated by $\{\#\text{of}D_n\text{p} - \text{values} < a\}/m$.

Now, three examples are presented next to exhibit the test's power behavior and asses, its practical performance. For each distribution utilized in each example, the power functions presented result by an average of 40 power functions calculated as described above. Further, all three tests are always calculated on the same samples.

The first example (Fig. 2.1) utilizes the bivariate distribution with p.d.f.

$$f_1(x, y) = \exp(-(x + y)), \quad x > 0, \ y > 0.$$

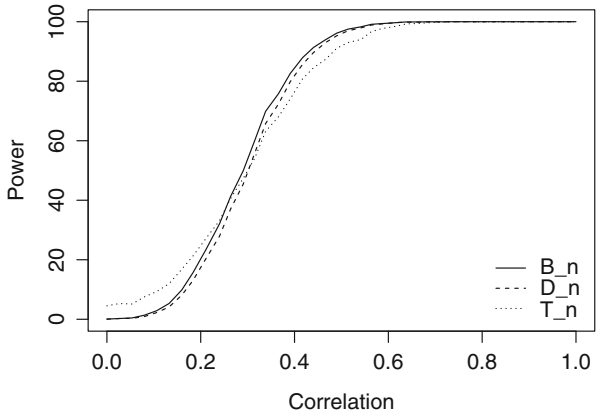

Fig. 2.1 Empirical test powers for T_n (*dotted line*), B_n (*solid line*), and D_n (*dashed line*), $n = 50$ with data from $f_1(x, y)$

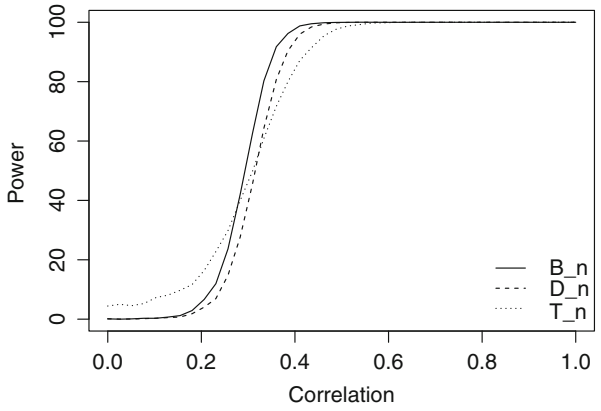

Fig. 2.2 Empirical test powers for T_n (*dotted line*), B_n (*solid line*), and D_n (*dashed line*), $n = 50$ with data from $f_2(x, y)$

In implementing the power function of $T_n(h)$, the bandwidth factor $c = 0.32$ has been found optimal. The significance level is $a = 1\,\%$.

For the second example (Fig. 2.2) the bivariate distribution with p.d.f.

$$f_2(x, y) = 2, \quad 0 \le x \le y \le 1,$$

is employed, $a = 5\,\%$ and the bandwidth factor for $T_n(h)$ is found to be $c = 1.34$.

The last example (Fig. 2.3) uses the bivariate normal distribution and $a = 5\%$. $T_n(h)$ is implemented with bandwidth factor $c = 0.7$.

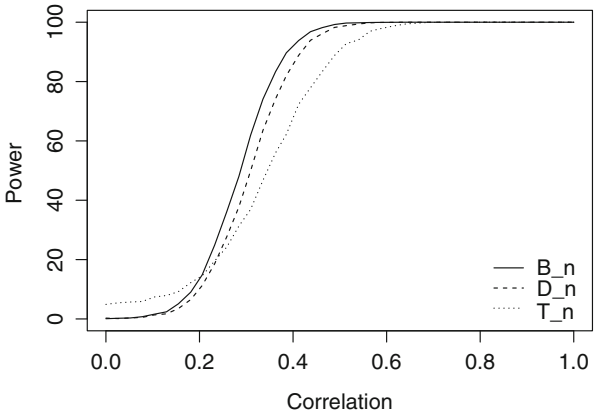

Fig. 2.3 Empirical test powers for T_n (*dotted line*), B_n (*solid line*), and D_n (*dashed line*), $n = 50$ with data from the bivariate normal distribution

References

1. Basset, G., Koenker, R.: An empirical quantile function for linear models with i.i.d. errors. J. Am. Stat. Assoc. **77**, 407–415 (1982)
2. Blum, J.R., Kiefer, J., Rosenblatt, M.: Distribution free tests for independence based on the sample distribution function. Ann. Math. Stat. **32**, 485–498 (1961)
3. Chan, E.: Testing constancy of regression quantiles using generalised sign test. Ph.D. thesis, University of Birmingham (2007)
4. Einmahl, J., McKeauge, I.: Empirical likelihood based hypothesis testing. Bernoulli **9**, 267–290 (2003)
5. Feurverger, A.: A consistent test for bivariate dependence. Int. Stat. Rev. **61**, 419–433 (1993)
6. Hoeffding, W.: A nonparametric test for independence. Ann. Math. Stat. **19**, 546–557 (1948)
7. Hyndman, R., Fan, Y.: Sample quantiles in statistical packages. Am. Statist. **50**, 361–365 (1996)
8. Li, Q., Wang, S.: A simple consistent bootstrap test for parametric regression function. J. Econ. **87**, 145–165 (1998)
9. Mudholkar, G.S., Wilding, G.E.: On the conventional wisdom regarding two consistent tests of bivariate independence. Statistician **52**, 41–57 (2003)
10. Patil, P., Sengupta, D.: On testing constancy of regression quantiles in non parametric regression models via kernel smoothing. Research Report CMA-SRR24-94, Centre for Mathematics and its Applications. The Australian National University (2003)
11. Rosenblatt, M.: A quadratic measure of deviation of two-dimensional density estimates and a test of independence. Ann. Stat. **3**, 1–14 (1975)
12. Rosenblatt, M., Wahlen, B.: A nonparametric measure of independence under a hypothesis of independent components. Stat. Probab. Lett. **15**, 245–252 (1992)
13. Ruppert, D., Sheather, S.J., Wand, M.P.: An effective bandwidth selector for local least squares regression. J. Am. Stat. Assoc. **90**, 1257–1270 (1995)
14. Tjostheim, D.: Measures of dependence and tests of independence. Statistics **28**, 249–284 (1996)

http://www.springer.com/978-1-4939-0568-3

Topics in Nonparametric Statistics
Proceedings of the First Conference of the
International Society for Nonparametric Statistics
Akritas, M.G.; Lahiri, S.N.; Politis, D.N. (Eds.)
2014, XVI, 367 p. 64 illus., 24 illus. in color., Hardcover
ISBN: 978-1-4939-0568-3