

Asymptotic hypotheses testing for the colour blind problem

Laura Dumitrescu and Estate V. Khmaladze

*School of Mathematics and Statistics
Victoria University of Wellington
PO Box 600, Wellington 6140, New Zealand
e-mail: laura.dumitrescu@vuw.ac.nz
e-mail: estate.khmaladze@vuw.ac.nz*

Abstract: Within a nonparametric framework, we consider the problem of testing the equality of marginal distributions for a sequence of independent and identically distributed bivariate data, with unobservable order in each pair. In this case, it is not possible to construct the corresponding empirical distributions functions and yet this article shows that a systematic approach to hypothesis testing is possible and provides an empirical process on which inference can be based. Furthermore, we identify the linear statistics that are asymptotically optimal for testing the hypothesis of equal marginal distributions against contiguous alternatives. Finally, we exhibit an interesting property of the proposed stochastic process: local alternatives of dependence can also be detected.

MSC 2010 subject classifications: Primary 62G10; secondary 62G20.

Keywords and phrases: Asymptotically optimal test, contiguous alternatives, dependence alternatives, empirical process, goodness of fit, Kolmogorov–Smirnov statistics, unordered pairs.

Received March 2019.

Contents

1	Introduction	4574
1.1	Related work	4577
2	The two-sample problem. The colour blind process	4578
2.1	The asymptotic behaviour under the null hypothesis	4579
2.2	Goodness of fit test statistics	4580
3	Linear statistics	4582
3.1	Description of local alternatives	4584
3.2	Optimal linear statistics	4586
3.3	Tests based only on maxima	4588
4	Local alternatives of dependence	4591
	Acknowledgements	4594
	References	4594

1. Introduction

Motivated by data that appears in genetics, biostatistics and biology, we consider the problem of testing hypotheses for *unordered* pairs of observations. For example, in each nucleus of a somatic human cell, there are 23 pairs of chromosomes. Within each pair, one chromosome is derived from the mother DNA and the other is derived from the father DNA. In karyotype analysis, measurements of different characteristics (such as the spiralisation coefficient – see the pioneering paper of [6]) are collected on homologous chromosomes, and the question of interest is to determine if there exist significant differences between the chromosomes derived from the mother and those derived from the father. However, visually the chromosomes in the pair are not distinguishable; an example of a sequence of (unordered) pairs of normalized measurements in the C-band area of the number 9 chromosome appears in [11].

Another example arises in controlled trials, where “blinding” is often employed. This refers to the situation where some or more individuals (the analyst and/or the participant) involved in the study are unaware of the assigned treatment (see e.g. [5] for a detailed description and motivation). Blind assessment of the treatment outcome during the course of a clinical trial comparing, for example, the effects of two treatments on individuals, generates a sequence of unordered pairs. The (blind) analysis of interim data is often required as, in some cases, the availability of preliminary results, may help to avoid the risk of further experimentation, see also [1] and [13] for additional details.

Formally, we assume that the observed data consists of a sequence of unordered pairs, denoted as $\{X_i, Y_i\}_{1 \leq i \leq n}$, and consider the problem of testing if measurements are independent and have the same distribution

$$H_0 : P_1 = P_2 \text{ (and equal to some } \textit{unspecified} \text{ distribution function } Q).$$

Here, P_1 and P_2 denote the distribution functions of the random variables $\{X_i\}_{1 \leq i \leq n}$ and $\{Y_i\}_{1 \leq i \leq n}$, respectively, and we assume that they are continuous.

In the classical two-sample problem, the class of test statistics with distributions that are independent of the common distribution Q (provided it is continuous), is well known. Namely, if P_{1n} and P_{2n} are the empirical distribution functions of $\{X_i\}_{1 \leq i \leq n}$ and $\{Y_i\}_{1 \leq i \leq n}$, respectively, any statistic based on the empirical processes

$$\sqrt{n}(P_{1n} - P_{2n}), \text{ or } \sqrt{n} \left[P_{1n} - \frac{1}{2}(P_{1n} + P_{2n}) \right],$$

which are invariant under Kolmogorov time transformation $Q(x) = t$, has the same distribution, regardless of the form of Q . If the research interest lies only in the change between the expected values, then the Student’s statistic $\sqrt{n}(\bar{X}_n - \bar{Y}_n)$ (with proper normalization) will provide a good test. If only testing the independence of variables within each pair is of interest, then the inference can be based on

$$\sqrt{n}(\mathbb{P}_n - P_{1n}P_{2n}),$$

where \mathbb{P}_n denotes the bivariate empirical distribution function of the pairs $\{(X_i, Y_i)\}_{1 \leq i \leq n}$.

However, in the case of unobservable order in each pair, it is not possible to construct the empirical distribution functions P_{1n} and P_{2n} , or, even the averages \bar{X}_n and \bar{Y}_n . We refer to this situation as the color blind problem and in this article we aim to answer the following question. Being color blind, when observing a sequence of pairs of balls with random diameters, one red and one green ball in each pair, is it possible to distinguish if the diameters of the red balls were generated from the same distribution as the diameters of the green balls?

In this article we show that a systematic approach to testing is possible and provide an empirical process, having a distribution under the null that is independent of the unknown common Q . This property allows the convenient use of omnibus goodness of fit statistics. In construction of this process, no statistical information is lost and, as in the classical theory, the statistics of the asymptotically most powerful tests are just linear functionals from this process. Furthermore, we explain what is the price to be paid for “colour blindness” in terms of the power of our tests and find the statistic of the asymptotically most powerful test for testing H_0 against a sequence of local alternatives.

A heuristic justification for the form of the proposed approach is based on a couple of simple remarks. The only data that a colour blind observer can collect is a sequence of pairs $\{(U_i, V_i)\}_{1 \leq i \leq n}$, where $U_i = \max\{X_i, Y_i\}$, $V_i = \min\{X_i, Y_i\}$, and so any statistic for testing H_0 will have to be constructed based on their empirical distributions. Regardless of whether the two marginal distributions are equal or not, the random variables $\{U_i\}_{1 \leq i \leq n}$ and $\{V_i\}_{1 \leq i \leq n}$ form two sequences of i.i.d. random variables and, assuming that the variables in each pair are independent, their cumulative distribution functions, under H_0 are, respectively,

$$P^{(2)}(x) = P_1(x)P_2(x), \quad P^{(1)}(x) = P_1(x) + P_2(x) - P_1(x)P_2(x), \quad x \in \mathbb{R}.$$

Note that if $P_1 = P_2$, the distributions $P^{(2)}$ and $P^{(1)}$ cannot be arbitrarily different: they are tied by the relation $P^{(2)}(x) = \left[1 - \sqrt{1 - P^{(1)}(x)}\right]^2$. For arbitrary distributions P_1 and P_2 , the following inequalities hold, for any $x \in \mathbb{R}$,

$$P^{(2)}(x) \leq \left[\frac{P^{(1)}(x) + P^{(2)}(x)}{2}\right]^2 \leq \left[1 - \sqrt{1 - P^{(1)}(x)}\right]^2, \quad (1.1)$$

or, equivalently,

$$P_1(x)P_2(x) \leq \left[\frac{P_1(x) + P_2(x)}{2}\right]^2 \leq \left[1 - \sqrt{(1 - P_1(x))(1 - P_2(x))}\right]^2,$$

with equality if and only if H_0 is true. However, the inequalities in (1.1) are surprisingly tight even when the difference between P_1 and P_2 is considerable; Figure 1 illustrates this property for $P_1(x) = x$ and $P_2(x) = x^2$. Therefore,

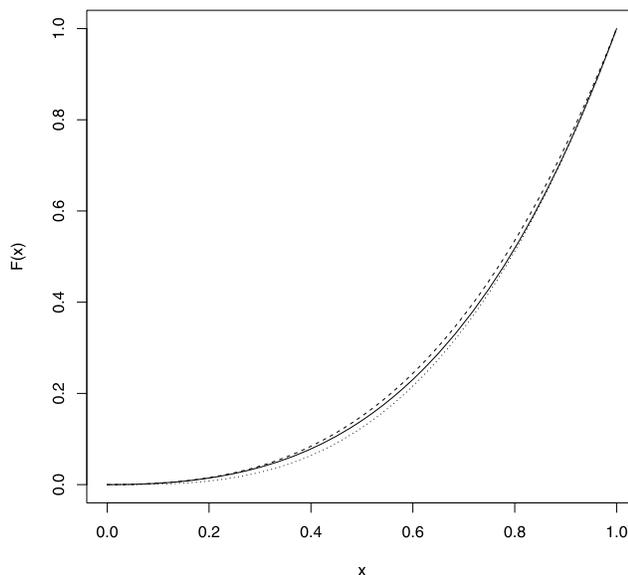


FIG 1. The case when $P_1(x) = x$ and $P_2(x) = x^2$, with $0 \leq x \leq 1$. The top dotted curve shows the graph of $[1 - \sqrt{1 - P^{(1)}(x)}]^2$, the solid curve is the graph of $[(P_1(x) + P_2(x))/2]^2$, whereas the bottom dotted curve gives the plot $P^{(2)}(x)$.

testing H_0 on the basis of the empirical distributions of $\{U_i\}_{1 \leq i \leq n}$ and $\{V_i\}_{1 \leq i \leq n}$ will not be easy. We will see below that, although we use a different and broader basis for testing H_0 , it remains a rather difficult statistical problem.

From a somewhat different side, observe the following fact: although it is obvious that the maximum value within each pair will be greater than the minimum value in the same pair, comparing cross values can be informative. More precisely, if $i \neq j$, under H_0 ,

$$P(U_j > V_i) = P(Q(U_j) > Q(V_i)) = \frac{5}{6}, \quad (1.2)$$

while this probability would be larger if, for example, P_1 is stochastically dominated by P_2 . In particular, if $P_1(x) = x$, $P_2(x) = x^2$, then

$$P(U_j > V_i) = \frac{51}{60},$$

which is only larger by $1/60$, in spite of big difference between P_1 and P_2 .

These arguments suggest that a nonparametric approach to hypothesis testing in the colour blind problem is possible, but deviations from H_0 will be difficult to detect. They also seem to suggest that a natural approach is to base our inference on a statistical version of the contrast

$$P_1(x)P_2(x) - \left[\frac{P_1(x) + P_2(x)}{2} \right]^2, \quad (1.3)$$

leading to some sort of empirical process in one-dimensional time. However, we prefer to first study the joint behaviour of U_i and V_i which requires the study of the empirical process in two-dimensional time. This, in turn, opens a possibility to test the independence of X_i and Y_i in each pair. In Section 4 we consider local alternatives of dependence, described in terms of copula functions, and show that, in the colour blind set up, detecting these is essentially an easier problem. Yet, difficulties due to the colour blind situation are still possible, as it appears in Section 4, and may occur when the alternative copulas are not symmetric.

This article is organised as follows. The empirical process, on which our inference is based, is introduced in Section 2. In Section 2.2 we show the behaviour of the Kolmogorov–Smirnov goodness of fit statistic, whereas in Section 3.2, we derive linear statistics from this process, which are optimal for particular sequences of local alternatives of different marginals. Furthermore, in Section 3.3, we consider an empirical process based on the largest observation within each pair: this could be thought of as an empirical analogue of (1.3) and the first and the most direct object to consider. However, as we said, the roundabout way through the empirical process with two-dimensional time defined in Section 2 is actually simpler and more natural. It also has the advantage of allowing us to consider testing the null against alternatives of dependence, which we describe in Section 4.

1.1. Related work

When the observations are assumed to be collected from two independent normal populations, a likelihood ratio statistic for testing the equality of means of unordered pairs of data was proposed in [8], and for several populations, in [2].

The more general case, of two marginal distributions that belong to the same parametric family, was considered in [15] and a statistic for a locally most powerful rank test, within this parametric family, was derived. It was also shown there that the degree of separation, which can make the alternative hypothesis distinguishable from the null, is of order $n^{-1/4}$. As it will be seen in Section 3.1, an analogous finding is encountered under the nonparametric approach.

Under the assumption that the observations within pairs are independent and their distributions belong to the family of Lehmann alternatives, a test for verifying their equality was discussed in [4]. The authors proposed a modified Mann-Whitney test statistic, by counting the number of times the minimum in a pair exceeded the maximum from another pair (a closely related note is observed in (1.2) of the present manuscript).

More recently, a semiparametric approach was taken in [12]. The authors assume an exponential tilting model for the density ratio and, based on a remark that, when the variables in each pair are independent, testing the equality of marginal distributions is related to an independence testing problem, a test based on the empirical Shannon’s mutual information was proposed.

2. The two-sample problem. The colour blind process

In the sequel we assume that $\{(X_i, Y_i)\}_{1 \leq i \leq n}$ are i.i.d observations from an unspecified, continuous distribution, and are located in $[0, 1]^2$. Moreover, for distributions functions P_1 and P_2 , we employ the notation $P_1 \times P_2$ for their product, i.e. $(P_1 \times P_2)(x, y) = P_1(x)P_2(y)$, with $x, y \in [0, 1]^2$.

For each $n \geq 1$ and $0 \leq x \leq 1$, $0 \leq y \leq 1$, let

$$\mathbb{P}_n(x, y) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{0 \leq X_i \leq x, 0 \leq Y_i \leq y\}}$$

be the empirical distribution function of the pairs, $\{(X_i, Y_i)\}_{1 \leq i \leq n}$, where we use the notation $\mathbb{1}_A$ for the indicator function of a set A . Define $\mathbb{Q}_n = Q_n \times Q_n$, with marginals given by

$$Q_n(x) = \frac{P_{1n}(x) + P_{2n}(x)}{2},$$

where, $P_{1n}(x) = 1/n \sum_{i=1}^n \mathbb{1}_{\{0 \leq X_i \leq x\}}$ and $P_{2n}(x) = 1/n \sum_{i=1}^n \mathbb{1}_{\{0 \leq Y_i \leq x\}}$ are the empirical marginal distributions of $\{X_i\}_{1 \leq i \leq n}$ and $\{Y_i\}_{1 \leq i \leq n}$, respectively. Note that the distribution function Q_n can be computed from the data, since $P_{1n} + P_{2n} = P_n^{(2)} + P_n^{(1)}$, where $P_n^{(2)}$ and $P_n^{(1)}$ denote the empirical distribution functions based on the maxima $\{U_i\}_{1 \leq i \leq n}$ and the minima $\{V_i\}_{1 \leq i \leq n}$, respectively.

Prompted by the discussion in Section 1, we consider the empirical process defined as

$$\mathbb{R}_n(x, y) = \sqrt{n} [\mathbb{P}_n(x, y) - \mathbb{Q}_n(x, y)], \quad 0 \leq x \leq 1, \quad 0 \leq y \leq 1. \quad (2.1)$$

Although the empirical distribution function \mathbb{P}_n , can not be obtained in the colour blind problem, it can be obtained within the class of symmetric Borel sets, i.e., sets B for which $(x, y) \in B$ implies $(y, x) \in B$. Indeed, if B is symmetric, then

$$P((U_i, V_i) \in B) = P((X_i, Y_i) \in B),$$

and a similar equality holds for the empirical distribution functions. Therefore, we consider the process \mathbb{R}_n on symmetric sets, as in the next definition. Denote by $S_{x,y} = [0, x] \times [0, y] \cup [0, y] \times [0, x]$ the symmetrised version of the rectangle $[0, x] \times [0, y]$; clearly, if $u = \max\{x, y\}$ and $v = \min\{x, y\}$ then $S_{u,v} = S_{x,y}$.

Definition 2.1. Let \mathcal{B} denote the class of symmetric Borel subsets of $[0, 1]^2$. Then, the restriction of \mathbb{R}_n to \mathcal{B} is called a colour blind empirical process

$$\mathbb{R}_n^s(B) = \mathbb{R}_n(B), \quad \text{where } B \in \mathcal{B}.$$

For $S_{u,v}$ as defined above, we write

$$\mathbb{R}_n^s(u, v) = \mathbb{R}_n(S_{u,v}) = \mathbb{R}_n(u, v) + \mathbb{R}_n(v, u) - \mathbb{R}_n(v, v),$$

and so the process $\mathbb{R}_n^s(u, v)$ is defined on the simplex $0 \leq v \leq u \leq 1$.

2.1. The asymptotic behaviour under the null hypothesis

In this section we will show that the limit in distribution of the colour blind process is the restriction of a Brownian pillow to the class of symmetric sets. We first consider the asymptotic behaviour of \mathbb{R}_n and show how the functional central limit theorem can be used to obtain its convergence in distribution to a suitable Gaussian process, as $n \rightarrow \infty$.

Let v_n denote the classical empirical process, based on i.i.d. pairs (X_i, Y_i) , from a specified distribution \mathbb{Q}

$$v_n(x, y) = \sqrt{n}(\mathbb{P}_n - \mathbb{Q})(x, y),$$

and $v_{\mathbb{Q}}$ denote the Brownian bridge in time \mathbb{Q} . Then, the functional central limit theorem gives $v_n \xrightarrow{d} v_{\mathbb{Q}}$, as $n \rightarrow \infty$, on the space of càdlàg functions on $[0, 1]^2$, equipped with the Skorohod topology.

Furthermore, assuming the independence of components within each pair, under the null hypothesis of equal marginals, we have $\mathbb{Q}(x, y) = Q(x)Q(y)$ and the empirical process introduced in (2.1) can be written as

$$\begin{aligned} \mathbb{R}_n(x, y) &= v_n(x, y) - \sqrt{n}[Q_n(x)Q_n(y) - Q(x)Q(y)] \\ &= v_n(x, y) - \sqrt{n}[Q_n(x) - Q(x)]Q(y) - \sqrt{n}[Q_n(y) - Q(y)]Q(x) + r_n(x, y) \quad (2.2) \\ &= v_n(x, y) - \frac{Q(y)}{2}[v_n(x, 1) + v_n(1, x)] - \frac{Q(x)}{2}[v_n(1, y) + v_n(y, 1)] + r_n(x, y), \end{aligned}$$

with

$$r_n(x, y) = \sqrt{n}[Q_n(x) - Q(x)][Q_n(y) - Q(y)],$$

and $\sup_{(x,y) \in [0,1]^2} |r_n(x, y)| = o_{\mathbb{Q}}(1)$.

Since the leading term in the right hand side of (2.2) is a linear transformation of v_n , then, as $n \rightarrow \infty$, we have $\mathbb{R}_n \xrightarrow{d} \mathbb{R}$, where

$$\mathbb{R}(x, y) = v_{\mathbb{Q}}(x, y) - \frac{1}{2}[v_{\mathbb{Q}}(x, 1) + v_{\mathbb{Q}}(1, x)]Q(y) - \frac{1}{2}[v_{\mathbb{Q}}(1, y) + v_{\mathbb{Q}}(y, 1)]Q(x). \quad (2.3)$$

Note that the transformation of v_n in the right hand side of (2.2) (as well as the transformation of $v_{\mathbb{Q}}$ in (2.3)) is, actually, a projection.

By Definition 2.1 and the limiting behaviour of \mathbb{R}_n we obtain

$$\mathbb{R}_n^s \xrightarrow{d} \mathbb{R}^s,$$

where \mathbb{R}^s is defined as the restriction of the process \mathbb{R} (given in (2.3)) to \mathcal{B} . Moreover, for a better insight in the nature of the process \mathbb{R}_n^s , let us consider a different projection of $v_{\mathbb{Q}}$, given by the operator

$$\mathcal{L}\alpha(x, y) = \alpha(x, y) - Q(x)\alpha(1, y) - Q(y)\alpha(x, 1) + Q(x)Q(y)\alpha(1, 1). \quad (2.4)$$

It projects a function α onto the class of functions equal to zero everywhere on the boundary of $[0, 1]^2$ and so, the projection of $v_{\mathbb{Q}}$ is given by

$$\begin{aligned} z_{\mathbb{Q}}(x, y) &= \mathcal{L}v_{\mathbb{Q}}(x, y) \\ &= v_{\mathbb{Q}}(x, y) - Q(x)v_{\mathbb{Q}}(1, y) - Q(y)v_{\mathbb{Q}}(x, 1) + Q(x)Q(y)v_{\mathbb{Q}}(1, 1). \end{aligned} \quad (2.5)$$

The process $z_{\mathbb{Q}}$ is called a *Brownian pillow* on $[0, 1]^2$ (or a *bivariate tied-down Brownian bridge* or *completely tucked Brownian sheet*) and it usually appears as the limit process for testing the independence of components of continuous bivariate random vectors (see e.g. [3] and Section 3.8 in [16]). With the time transformation $t = Q(x)$, $s = Q(y)$, it can be mapped into a *standard* Brownian pillow in t and s , i.e. a Gaussian process $z(s, t)$, with covariance function given by $E(z(s', t')z(s'', t'')) = (\min\{s', s''\} - s's'')(\min\{t', t''\} - t't'')$, $0 \leq s', s'', t', t'' \leq 1$. Its finite n version is, obviously, given by $z_n = \mathcal{L}v_n$.

Our interest in $z_{\mathbb{Q}}$ and z_n stems from the following fact.

Proposition 2.2. *For symmetric sets $B \in \mathcal{B}$ we have*

$$\mathbb{R}_n^s(B) = z_n(B) + o_{\mathbb{Q}}(1), \text{ as } n \rightarrow \infty, \text{ and } \mathbb{R}^s(B) = z_{\mathbb{Q}}(B).$$

In particular, for any $n \geq 1$ and symmetrised rectangles $S_{u,v}$, the following relationship holds

$$\mathbb{R}_n^s(u, v) = z_n(S_{u,v}) + r_n(S_{u,v}),$$

where

$$\begin{aligned} z_n(S_{u,v}) &= z_n(u, v) + z_n(v, u) - z_n(v, v), \\ r_n(S_{u,v}) &= 2\sqrt{n}[Q_n(u) - Q(u)][Q_n(v) - Q(v)] - \sqrt{n}[Q_n(v) - Q(v)]^2 = o_{\mathbb{Q}}(1). \end{aligned}$$

2.2. Goodness of fit test statistics

Proposition 2.2 shows that, asymptotically, the colour blind process \mathbb{R}_n^s is equivalent to a Brownian pillow on symmetric sets. We also noted that the Brownian pillow $z_{\mathbb{Q}}$ can be transformed into the standard Brownian pillow, i.e., a Brownian pillow when \mathbb{Q} is the Lebesgue measure on $[0, 1]^2$. Therefore, the classical goodness of fit statistics, such as the Kolmogorov–Smirnov statistic, and other statistics based on \mathbb{R}_n^s and invariant with respect to time transformation $t = Q(x)$, $s = Q(y)$, will have limiting distribution, which does not depend on the unknown Q . Intuitively, we would expect that among the two statistics

$$D_n = \sup_{(x,y) \in [0,1]^2} |\mathbb{R}_n(x, y)|, \quad D_n^s = \sup_{(u,v) \in [0,1]^2, v < u} |\mathbb{R}_n^s(u, v)|,$$

the second one will be stochastically smaller. Figure 2 shows the graphs of their distributions and confirms that this intuition is correct. To illustrate the situation in terms of the power of the goodness of fit tests, we consider in Figure 3 the shift of D_n^s under the alternatives described in Figure 1. It was illustrated in Figure 1 that, in the colour blind situation, the alternative distributions look surprisingly difficult to distinguish from the null. Yet, according to Figure 3, with $n = 500$, the Kolmogorov–Smirnov test will have some power. However, to

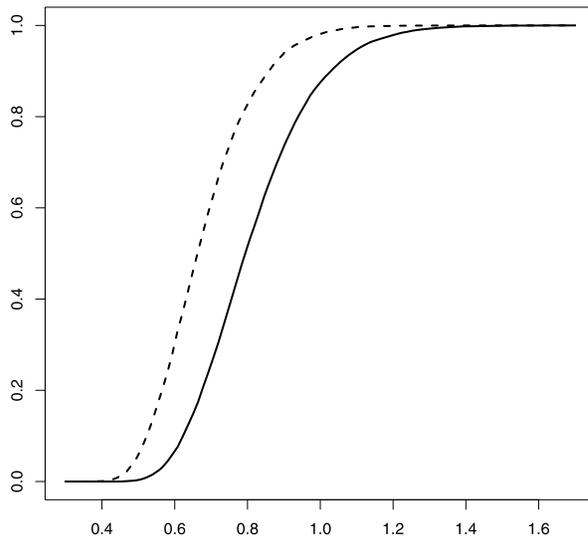


FIG 2. The dotted line represents the graph of the (simulated) distribution function of D_n^s , while the solid line represents that of D_n , with $n = 1,000$.

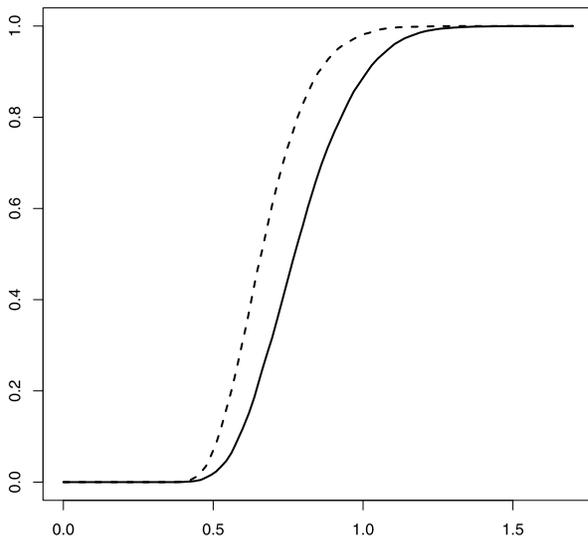


FIG 3. The dotted line shows the simulated distribution function of D_n^s under the null hypothesis, while the solid line shows its distribution under the alternative $A_1 \times A_2$, with $A_1(x) = x$, $A_2(x) = x^2$ and $n = 500$.

better illustrate the consequence of colour blindness, in Figure 4, we present the simulated distribution functions of D_n , under the null and under the alternatives

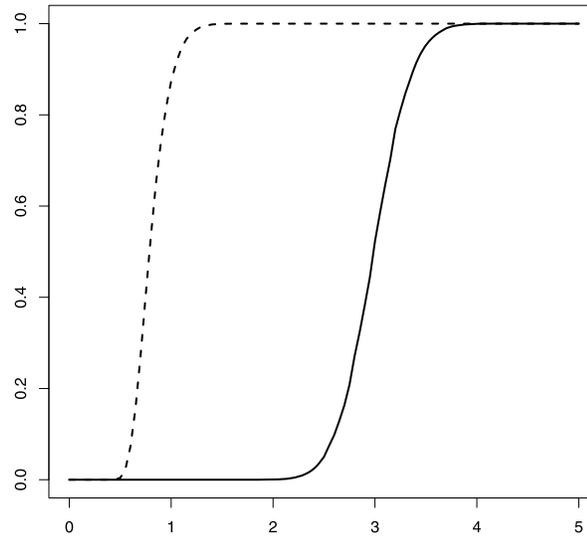


FIG 4. The graph shows the simulated distribution functions of D_n under the null hypothesis (dotted line) and under the alternative $A_1(x) = x, A_2(x) = x^2$ (solid line), with $n = 500$.

in Figure 1. We observe that the discrimination between the two, for $n = 500$, could have been absolutely obvious.

3. Linear statistics

Unlike the goodness of fit tests, which are of omnibus nature and typically have some power against a very wide class of alternatives, the tests based on linear statistics may have asymptotically no power against the “majority” of alternatives, but are asymptotically most powerful against a certain form of alternatives. In Section 3.2 we derive the form of such statistic, which is optimal for testing H_0 against a sequence of local alternatives of different marginals, described in Section 3.1.

Let $\varphi \in L^2(Q \times Q)$ and consider the function-parametric version of the colour blind process

$$\mathbb{R}_n^s(\varphi) = \int_0^1 \int_{v < u} \varphi(u, v) d\mathbb{R}_n^s(u, v).$$

In search for the optimal linear functional, the next result shows that we can restrict our attention to symmetric functionals from z_n .

Proposition 3.1. *For every $\varphi \in L^2(Q \times Q)$, as $n \rightarrow \infty$,*

$$\int_0^1 \int_{v < u} \varphi(u, v) d\mathbb{R}_n^s(u, v) = \int_0^1 \int_0^1 \tilde{\varphi}(x, y) dz_n(x, y) + o_{\mathbb{Q}}(1), \quad (3.1)$$

where

$$\tilde{\varphi}(x, y) = \begin{cases} \varphi(x, y), & x \geq y \\ \varphi(y, x), & x < y \end{cases}.$$

Proof. We apply Proposition 2.2 and first note that the planar integral from the term $z_n(v, v)$ is null. Moreover, the integral from the residual term is indeed small, i.e. for any $\varphi \in L^2(Q \times Q)$

$$\int_0^1 \int_0^1 \varphi(x, y) \sqrt{n} [dQ_n(x) - dQ(x)] [dQ_n(y) - dQ(y)] = o_{\mathbb{Q}}(1)$$

and hence, the left hand side of (3.1) becomes

$$\int_0^1 \int_{v < u} \varphi(u, v) d\mathbb{R}_n^s(u, v) = \int_0^1 \int_{v < u} \varphi(u, v) [dz_n(u, v) + dz_n(v, u)] + o_{\mathbb{Q}}(1). \tag{3.2}$$

The main term in the right hand side of (3.1) is

$$\begin{aligned} \int_0^1 \int_0^1 \tilde{\varphi}(x, y) dz_n(x, y) &= \int_0^1 \int_{x > y} \varphi(x, y) dz_n(x, y) + \int_0^1 \int_{y > x} \varphi(y, x) dz_n(x, y) \\ &= \int_0^1 \int_{u > v} \varphi(u, v) dz_n(u, v) + \int_0^1 \int_{u > v} \varphi(u, v) dz_n(v, u), \end{aligned}$$

where we used the fact that $\tilde{\varphi}$ is symmetric. The proof is concluded by noting that the right hand sides of the last two displays are equal. \square

Denote by $(Q\varphi)(x) = \int_0^1 \varphi(x, y) dQ(y)$ and consider the following projection of φ

$$(\mathcal{L}^* \varphi)(x, y) = \varphi(x, y) - (Q\varphi)(x) - (Q\varphi)(y) + E_{\mathbb{Q}}[\varphi(x, y)].$$

The process z_n was introduced as a projection of v_n and so $z_n = \mathcal{L}v_n$, with $(\mathcal{L}v_n)(x, y) = v_n(x, y) - Q(x)v_n(1, y) - Q(y)v_n(x, 1) + Q(x)Q(y)v_n(1, 1)$. The following proposition shows that operator \mathcal{L}^* can be viewed as the adjoint operator (see [10]) of \mathcal{L} , defined in (2.4).

Proposition 3.2. *We have the following*

$$\begin{aligned} z_n(\varphi) = \mathcal{L}v_n(\varphi) &= \int_0^1 \int_0^1 \varphi(x, y) dz_n(x, y) \\ &= \int_0^1 \int_0^1 (\mathcal{L}^* \varphi)(x, y) dv_n(x, y) = v_n(\mathcal{L}^* \varphi). \end{aligned}$$

Proof. By direct computation,

$$\begin{aligned} z_n(\varphi) &= \int_0^1 \int_0^1 \varphi(x, y) dv_n(x, y) - \int_0^1 \int_0^1 \varphi(x, y) dQ(x) dv_n(1, y) \\ &\quad - \int_0^1 \int_0^1 \varphi(x, y) dQ(y) dv_n(x, 1) \\ &= \int_0^1 \int_0^1 \left[\varphi(x, y) - \int_0^1 \varphi(x, y) dQ(x) - \int_0^1 \varphi(x, y) dQ(y) \right] dv_n(x, y). \quad \square \end{aligned}$$

As a result,

$$\mathbb{R}_n^s(\varphi) = v_n(\mathcal{L}^* \tilde{\varphi}) + o_{\mathbb{Q}}(1), \quad (3.3)$$

and the well-known central limit theorem for the function-parametric empirical process v_n can be used to describe the asymptotic behaviour of $\mathbb{R}_n^s(\varphi)$.

3.1. Description of local alternatives

We are interested in detecting small departures from the null hypothesis and, assuming independence between the coordinates, such deviations will be specified by a sequence of probability distribution functions of the form $\{A_{1n} \times A_{2n}\}_{n \geq 1}$.

For an arbitrary fixed probability distribution Q , let A_{1n} and A_{2n} be probability distributions which are defined as asymptotically “small” departures from Q

$$\frac{dA_{1n}}{dQ}(x) = 1 + \varepsilon_n h_{1n}(x), \quad \frac{dA_{2n}}{dQ}(x) = 1 + \varepsilon_n h_{2n}(x), \quad (3.4)$$

where $\varepsilon_n \rightarrow 0$ as $n \rightarrow \infty$, and the functions $h_{kn}, k = 1, 2$ converge to square integrable functions $h_k(x)$

$$\int_0^1 [h_{kn}(x) - h_k(x)]^2 dQ(x) \rightarrow 0, \quad \int_0^1 h_k^2(x) dQ(x) < \infty, \quad k = 1, 2.$$

From the definition of h_{kn} , it follows that, for all $n \geq 1$, $\int_0^1 h_{kn}(x) dQ(x) = 0$, $k = 1, 2$, and their limits inherit this property.

We shall see that ε_n will not be of order $n^{-1/2}$ as it is typically the case within the theory of contiguity (see [14] and [7]), but will need to decrease slower. Therefore, we will eventually be outside the contiguity theory, and, therefore, we can neglect using square roots from the Radon–Nikodym derivatives in (3.4), and consider them as they are, which is somewhat simpler. It is more interesting to recall that, although both A_{1n} and A_{2n} tend to Q , in testing H_0 , they will look differently.

To see this fact, specific to the two-sample problem (and not to colour blindness as such), consider the expected value of the classical two-sample process $\sqrt{n}(P_{1n} - P_{2n})$ under the alternative $A_{1n} \times A_{2n}$. Introduce the functions H_{1n} and H_{2n} as

$$H_{kn}(x) = \int_0^x h_{kn}(y) dQ(y), \quad \text{so that } H_{kn}(0) = H_{kn}(1) = 0, \quad k = 1, 2,$$

and so

$$E_a \sqrt{n}[P_{1n}(x) - P_{2n}(x)] = \sqrt{n}[A_{1n}(x) - A_{2n}(x)] = \sqrt{n} \varepsilon_n [H_{1n}(x) - H_{2n}(x)].$$

It can be shown that choosing $\varepsilon_n = n^{-1/2}$ and the linear statistic

$$\int_0^1 \phi(x) \sqrt{n}[dP_{1n}(x) - dP_{2n}(x)], \quad \text{where } \phi(x) = h_{1n}(x) - h_{2n}(x) \quad (3.5)$$

leads to the asymptotically most powerful test, among those based on $\sqrt{n}(P_{1n} - P_{2n})$. However, the power of this test is only less than or equal to the power of the optimal (Neyman–Pearson) test in the problem of discriminating between the alternative $A_{1n} \times A_{2n}$ and the hypothesis $Q \times Q$ (see [7]) and to obtain an equality, we have to change Q to $Q_{an} = (A_{1n} + A_{2n})/2$. Then, the test based on the linear statistic in (3.5) becomes asymptotically equivalent to the Neyman–Pearson test for discriminating $A_{1n} \times A_{2n}$ from the hypothesis $Q_{an} \times Q_{an}$. The Radon–Nikodym derivatives of A_{1n} and A_{2n} with respect to Q_{an} possess a symmetric form

$$\frac{dA_{1n}}{dQ_{an}}(x) = 1 + \varepsilon_n h_n(x), \quad \frac{dA_{2n}}{dQ_{an}}(x) = 1 - \varepsilon_n h_n(x),$$

with

$$h_n = \frac{(h_{1n} - h_{2n})/2}{1 + \varepsilon_n(h_{1n} + h_{2n})/2}.$$

The dependence of Q_{an} on n , which itself converges to Q , is immaterial and we can assume from now on that in (3.4) we have $h_{1n} = -h_{2n}$, i.e., we will restrict our attention to the class of local alternatives of the form

$$H_a : \frac{dA_{1n}}{dQ}(x) = 1 + \varepsilon_n h_n(x), \quad \frac{dA_{2n}}{dQ}(x) = 1 - \varepsilon_n h_n(x), \quad (3.6)$$

for some continuous distribution Q on $[0, 1]$, assuming that there exists a function $h \in L^2(Q)$ such that

$$\int_0^1 h(x)dQ(x) = 0 \text{ and } \int_0^1 [h_n(x) - h(x)]^2 dQ(x) \rightarrow 0.$$

The function h determines the direction in which the alternative distribution $A_{1n} \times A_{2n}$ approaches the diagonal $\{Q \times Q, Q\text{-continuous}\}$; Figure 5 illustrates the situation.

We now consider the rate of convergence of ε_n in the colour blind problem. Under H_a ,

$$\begin{aligned} E_a[v_n(x, y)] &= E_a\{\sqrt{n}[\mathbb{Q}_n(x, y) - \mathbb{Q}(x, y)]\} \\ &= \sqrt{n}[-\varepsilon_n Q(x)H_n(y) + \varepsilon_n H_n(x)Q(y) - \varepsilon_n^2 H_n(x)H_n(y)], \end{aligned}$$

where, as above, $H_n(x) = \int_0^x h_n(y)dQ(y)$. Since $H_n(1) = 0$, it follows that

$$E_a[\mathbb{R}_n(x, y)] = \sqrt{n}[-\varepsilon_n Q(x)H_n(y) + \varepsilon_n H_n(x)Q(y) - \varepsilon_n^2 H_n(x)H_n(y)],$$

which shows that the statistics based on the process $\mathbb{R}_n(x, y)$ can distinguish alternatives with $\varepsilon_n = O(n^{-1/2})$. However, in the case of the symmetrised process $\mathbb{R}_n^s(u, v)$, under H_a , the linear term in ε_n becomes zero and we have

$$E_a[\mathbb{R}_n^s(u, v)] = \sqrt{n}\varepsilon_n^2[-2H_n(u)H_n(v) + H_n(v)^2].$$

This shows the loss of power when using the colour blind statistic: only alternatives with $\varepsilon_n \sim n^{-1/4}$ can be detected in the colour blind problem.

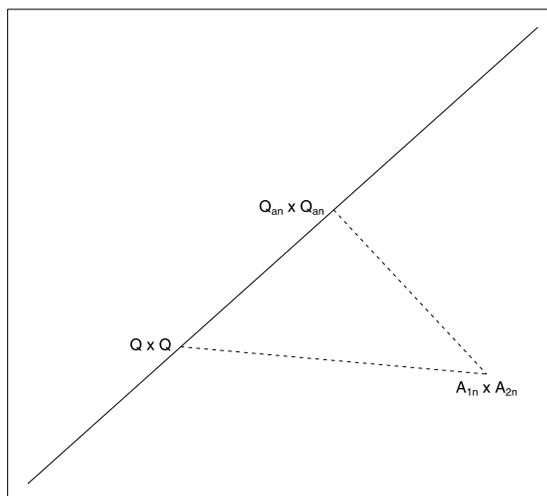


FIG 5. The sequence of alternatives $A_{1n} \times A_{2n}$ may be some distance away from the hypothetical pair $Q \times Q$, to which it converges, but statistics from the empirical process will “react” on this alternative as much as it deviates from its projection $Q_{an} \times Q_{an}$.

3.2. Optimal linear statistics

In this section we describe the statistic of the asymptotically optimal test for testing H_0 against the sequence of alternatives in (3.6). Here,

$$\langle \varphi, \psi \rangle_{Q \times Q} = \int_0^1 \int_0^1 \varphi(x, y) \psi(x, y) dQ(x) dQ(y),$$

denotes the inner product in $L^2(Q \times Q)$, and we recall that the variance of the Gaussian random variable $v_Q(\tilde{\varphi})$ is

$$\text{Var}(v_Q(\tilde{\varphi})) = \langle \tilde{\varphi}, \tilde{\varphi} \rangle_{Q \times Q} - \langle \tilde{\varphi}, 1 \rangle_{Q \times Q}^2.$$

Under H_a , the expected value of $v_n(\tilde{\varphi})$ is not zero and we have

$$\begin{aligned} v_n(\tilde{\varphi}) &= \sqrt{n} \int_0^1 \int_0^1 \tilde{\varphi}(x, y) [d\mathbb{P}_n(x, y) - dA_{1n}(x) dA_{2n}(y)] \\ &+ \sqrt{n} \varepsilon_n \int_0^1 \int_0^1 \tilde{\varphi}(x, y) [h_n(x) - h_n(y)] dQ(x) dQ(y) \\ &- \sqrt{n} \varepsilon_n^2 \int_0^1 \int_0^1 \tilde{\varphi}(x, y) h_n(x) h_n(y) dQ(x) dQ(y). \end{aligned}$$

The first integral on the right side, which contains the centered part of $v_n(\tilde{\varphi})$, converges in distribution to $v_Q(\tilde{\varphi})$, and because $\tilde{\varphi}(x, y)$ is symmetric, the middle

integral is null. Therefore, if $\varepsilon_n = n^{-1/4}$, we have, as $n \rightarrow \infty$, under H_0 and under H_a , respectively,

$$v_n(\tilde{\varphi}) \xrightarrow{d} v_{\mathbb{Q}}(\tilde{\varphi}), \quad \text{and} \quad v_n(\tilde{\varphi}) \xrightarrow{d} v_{\mathbb{Q}}(\tilde{\varphi}) - \langle \tilde{\varphi}, h \times h \rangle_{Q \times Q}. \quad (3.7)$$

To make a judgment about the asymptotic power of the linear statistics, consider the distance in total variation between two Gaussian distributions with different means and equal variances. This is given by

$$\sup_C |N_{(\mu_1, \sigma^2)}(C) - N_{(\mu_2, \sigma^2)}(C)|,$$

where the supremum is taken over all measurable sets (or critical regions of tests) on the real line. By its definition, this measure gives the largest possible difference between the power and the level of tests, among all those that can discriminate between the two distributions. Its advantage is that there is no need to specify a particular level of a test. We have

$$\sup_C |N_{(\mu_1, \sigma^2)}(C) - N_{(\mu_2, \sigma^2)}(C)| = 2N_{(0,1)}\left(\frac{|\mu_1 - \mu_2|}{2\sigma}\right) - 1,$$

where $N_{(0,1)}(x)$ denotes the standard normal distribution function. The ratio

$$T = \frac{|\mu_1 - \mu_2|}{\sigma},$$

(especially when $\mu_1 = 0$), is often called the signal to noise ratio; the larger this ratio is, the greater the difference between the power and the level.

Recalling that $U_i = \max\{X_i, Y_i\}$ and $V_i = \min\{X_i, Y_i\}$, $1 \leq i \leq n$, the following result gives the form of the optimal test statistic.

Proposition 3.3. *The statistic of the asymptotically most powerful test for testing H_0 against the sequence of alternatives $A_{1n} \times A_{2n}$ is of the form*

$$\mathbb{R}_n^s(h \times h) = \frac{1}{\sqrt{n}} \sum_{i=1}^n h(U_i)h(V_i) + o_{\mathbb{Q}}(1). \quad (3.8)$$

The distance in total variation between its asymptotic distributions, under the null and under the alternatives $A_{1n} \times A_{2n}$, is equal to $N_{(0,1)}(T/2)$, where T represents the limit of $\sqrt{n}\varepsilon_n^2 \|h \times h\|_{Q \times Q} = \sqrt{n}\varepsilon_n^2 \|h\|_Q^2$.

Proof. Using (3.7) and the fact that $\langle \mathcal{L}^* \tilde{\varphi}, 1 \rangle_{Q \times Q} = 0$, the asymptotic power of the test based on the linear statistic $v_n(\mathcal{L}^* \tilde{\varphi})$ is equal to $N_{(0,1)}(T_{\varphi})$, where, the “signal to noise ratio” is

$$T_{\varphi} = \frac{\langle \mathcal{L}^* \tilde{\varphi}, h \times h \rangle_{Q \times Q}}{\langle \mathcal{L}^* \tilde{\varphi}, \mathcal{L}^* \tilde{\varphi} \rangle_{Q \times Q}^{1/2}}.$$

Now note that $h \times h$ is both symmetric and passes through \mathcal{L}^* . Therefore, we have

$$\langle \mathcal{L}^* \tilde{\varphi}, h \times h \rangle_{Q \times Q} = \langle \mathcal{L}^* \tilde{\varphi}, \mathcal{L}^*(h \times h) \rangle_{Q \times Q},$$

which implies that T_φ is maximised at $\mathcal{L}^*\tilde{\varphi} = \mathcal{L}^*(h \times h) = h \times h$. The statistic $v_n(h \times h)$ is equal to the sum in the display formula in (3.8). On the other hand, (3.3) gives $\mathbb{R}_n^s(h \times h) = v_n(h \times h) + o_{\mathbb{Q}}(1)$. \square

Example 3.4. In the case illustrated in Figure 1, with $A_1(x) = x$, $A_2(x) = x^2$ and $Q(x) = (x + x^2)/2$, we have $h(x) = \frac{1-2x}{1+2x}$ and $\varepsilon = 1$. We can also choose the sequence of $\{A_{1\varepsilon} \times A_{2\varepsilon}\}$ as $\varepsilon \rightarrow 0$, keeping the same $h(x)$; in particular,

$$A_{1\varepsilon}(x) = \frac{1+\varepsilon}{2}x + \frac{1-\varepsilon}{2}x^2 \text{ and } A_{2\varepsilon}(x) = \frac{1-\varepsilon}{2}x + \frac{1+\varepsilon}{2}x^2.$$

The proposed test statistic for testing H_0 against the alternative $A_{1\varepsilon} \times A_{2\varepsilon}$ is of the form

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{(2U_i - 1)(2V_i - 1)}{(1 + 2U_i)(1 + 2V_i)}, \quad (3.9)$$

and its signal to noise ratio, for $n = 400$ and $\varepsilon = 1$, is given by $\sqrt{n\varepsilon^2} \|h\|_Q^2 = 1.98$ and therefore, the power of this linear test, directed to the chosen alternatives is essentially higher than the general (not directed) Kolmogorov-Smirnov test, as it appears in Figure 3. However, the test may have very low or no power for testing H_0 against other alternatives.

3.3. Tests based only on maxima

In this section we consider an alternative approach to testing H_0 , which is based on the empirical distribution function of the maxima $U_i = \max\{X_i, Y_i\}$,

$$P_n^{(2)}(u) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{U_i \leq u\}}.$$

The above is naturally centered by $Q_n^2(u)$, thus leading to an empirical process (see also (1.3))

$$R_n^{(2)}(u) = \sqrt{n}[P_n^{(2)}(u) - Q_n^2(u)], \quad (3.10)$$

whose construction, together with the form of the linear functionals from $P_n^{(2)}$,

$$\int_0^1 \alpha(u) dP_n^{(2)}(u) = \frac{1}{n} \sum_{i=1}^n \alpha(U_i),$$

can prompt us to speak about “statistics”, or tests, “based only on maxima”, even though this is not an accurate expression. Indeed, the term $Q_n^2(u) = [P_n^{(1)}(u) + P_n^{(2)}(u)]^2/4$ certainly incorporates information about minima as well. Yet, the process in (3.10) may look as the first choice to base test statistics upon and has some nice properties. For example, its covariance function retains a symmetric structure,

$$E_0[R_n^{(2)}(u)R_n^{(2)}(u')] = Q^2(u)[1 - Q(u')]^2, \quad 0 \leq u \leq u' \leq 1,$$

which is the covariance function of the product of two independent Q -Brownian bridges.

The natural way to study the process $R_n^{(2)}$ and its function parametric version is to embed it into the colour blind process, introduced in Definition 2.1. First, we see that $R_n^{(2)}$ is a restriction of \mathbb{R}_n^s to rectangles $[0, u] \times [0, u]$

$$R_n^{(2)}(u) = \mathbb{R}_n(u, u) = \sqrt{n} [\mathbb{P}_n(u, u) - Q_n^2(u)],$$

which, in particular, implies that the expected value of $R_n^{(2)}(u)$ under H_a is

$$E_a[R_n^{(2)}(u)] = -\sqrt{n}\varepsilon_n^2 H_n^2(u). \tag{3.11}$$

We now choose the functional argument of $\mathbb{R}_n(\cdot)$ as $\varphi(x, y) = \alpha(\max(x, y))$ and note that the function $\varphi = \varphi_\alpha$ is symmetric and belongs to $L^2(Q \times Q)$ if and only if $\alpha \in L^2(Q^2)$. Indeed, we have

$$\int_0^1 \int_0^1 \alpha^2(\max\{x, y\}) dQ(x) dQ(y) = \int_0^1 \alpha^2(u) dQ^2(u),$$

or $\langle \varphi_\alpha, \varphi_\alpha \rangle_{Q \times Q} = \langle \alpha, \alpha \rangle_{Q^2}$, leading to $R_n^{(2)}(\alpha) = \mathbb{R}_n(\varphi_\alpha)$. Denoting

$$\mathcal{C} = \{\varphi \in L^2(Q \times Q) : \varphi(x, y) = \alpha(\max\{x, y\}), \alpha \in L^2(Q^2)\},$$

then studying $R_n^{(2)}(\alpha)$, with $\alpha \in L^2(Q^2)$, becomes equivalent to studying $\mathbb{R}_n(\varphi)$, where $\varphi \in \mathcal{C}$. Then, (2.2), (2.5) and Proposition 3.2 imply that

$$R_n^{(2)}(\alpha) = v_n(\mathcal{L}^* \varphi_\alpha) + o_{\mathbb{Q}}(1),$$

and we can focus on the linear statistics appearing on the right side. The problem of finding the optimal linear statistic requires the maximisation of a different signal to noise ratio and opens up an interesting structure.

Introduce the Radon–Nikodym derivative

$$q(x) = \frac{dH^2(x)}{dQ^2(x)} = h(x) \frac{H(x)}{Q(x)},$$

and denote by $\varphi_q(x, y) = q(\max(x, y))$. Using (3.7), we derive that

$$\begin{aligned} E_a[R_n^{(2)}(\alpha)] &= E_a[v_n(\mathcal{L}^* \varphi_\alpha)] + o(1) = \langle \mathcal{L}^* \varphi_\alpha, h \times h \rangle_{Q \times Q} + o(1) \\ &= \langle \varphi_\alpha, h \times h \rangle_{Q \times Q} + o(1). \end{aligned}$$

Unless the trivial case (when $h = 0$), the functions of the form $(h \times h)(x, y) = h(x)h(y)$ do not belong to the class \mathcal{C} , but the expression of the expected values above suggests that the projection of $h \times h$ on \mathcal{C} would be useful to consider. This projection is given by the function φ_q . Indeed, we have

$$\langle \varphi_\alpha, h \times h - \varphi_q \rangle_{Q \times Q} = 0,$$

so that

$$\langle \varphi_\alpha, h \times h \rangle_{Q \times Q} = \langle \varphi_\alpha, \varphi_q \rangle_{Q \times Q} = \langle \alpha, q \rangle_{Q^2},$$

and would now seem straightforward to choose $\alpha = q$ as an optimal test statistics. However, more clarifications are needed.

Since $\text{Var}(v_n(\mathcal{L}^* \varphi_\alpha)) = \langle \mathcal{L}^* \varphi_\alpha, \mathcal{L}^* \varphi_\alpha \rangle_{Q \times Q}$, to find the asymptotically most powerful test against the sequence of alternatives $A_{1n} \times A_{2n}$, we need to maximise the absolute value of signal to noise ratio

$$T_\alpha = \frac{\langle \varphi_\alpha, \varphi_q \rangle_{Q \times Q}}{\langle \mathcal{L}^* \varphi_\alpha, \mathcal{L}^* \varphi_\alpha \rangle_{Q \times Q}^{1/2}} = \frac{\langle \alpha, q \rangle_{Q^2}}{\langle \mathcal{L}^* \varphi_\alpha, \mathcal{L}^* \varphi_\alpha \rangle_{Q \times Q}^{1/2}}.$$

A useful step in this direction will be to express the denominator in terms of the inner product in $L^2(Q^2)$. It can be verified that

$$\langle \mathcal{L}^* \varphi_\alpha, \mathcal{L}^* \varphi_\alpha \rangle_{Q \times Q} = \langle \alpha, \alpha \rangle_{Q^2} - \langle \alpha, S\alpha \rangle_{Q^2},$$

where the operator S is given by

$$S\alpha(x) = \alpha(x)Q(x) + 4 \int_x^1 \alpha(t)dQ(t),$$

and we need to maximise the absolute value of

$$T_\alpha = \frac{\langle \alpha, q \rangle_{Q^2}}{[\langle \alpha, \alpha \rangle_{Q^2} - \langle \alpha, S\alpha \rangle_{Q^2}]^{1/2}}.$$

One can go into this problem, for example – as a problem of calculation of support function for the convex set

$$\{\alpha : \langle \alpha, \alpha \rangle_{Q^2} - \langle \alpha, S\alpha \rangle_{Q^2} \leq 1\}.$$

However, it is much simpler to *reverse* the point of view: for a given α , find an alternative for which $\mathbb{R}_n(\varphi_\alpha)$ will be an optimal statistic. In the previous section, this reversal will not produce a different result, whereas in the present case the maximisation in q becomes simple.

Proposition 3.5. *Consider a cone*

$$\mathcal{M} = \{\alpha : \int_0^u \alpha(z)dQ^2(z) \geq 0 \text{ for all } u > 0\}.$$

If $\alpha \in \mathcal{M}$, then the power of the statistic $\mathbb{R}_n(\varphi_\alpha)$ is largest for the local alternatives in (3.6), with the corresponding q equal α .

In this case, the functions H and h which describe the alternatives are given by

$$H(x) = \pm \sqrt{\int_0^u \alpha(z)dQ^2(z)} \quad \text{and} \quad h(x) = \frac{Q(x)\alpha(x)}{2H(x)}.$$

Proof. The choice of $\alpha = q$ is valid if and only if $\alpha \in \mathcal{M}$. Indeed,

$$\alpha(x) = q(x) = \frac{dH^2(x)}{dQ^2(x)} \quad \text{or} \quad \int_0^u \alpha(x)dQ^2(x) = H^2(u)$$

and, therefore, the integral has to be non-negative for all $u > 0$, which implies that α has to belong to \mathcal{M} . The forms of H and h follow directly. \square

We may see \mathcal{M} as the class of *admissible* or *effective* α -s. The need to specify such a class, i.e. to choose φ_α more narrowly than from the linear space \mathcal{C} , is visible when we try to connect α and q . Actually, this fact provided the proof of the proposition.

Example 3.6. For the sake of numerical comparison, we consider the case in Example 3.4, where $A_1(x) = x$, $A_2(x) = x^2$ and $Q(x) = (x + x^2)/2$. This leads to

$$h(x) = \frac{2x - 1}{1 + 2x}, \quad H(x) = -\frac{1}{2}x(1 - x) \quad \text{and, therefore,} \quad q(x) = \frac{(x - 1)(2x - 1)}{(1 + x)(1 + 2x)}.$$

The form of the statistic resembles the form in (3.9), but requires centering,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{(U_i - 1)(2U_i - 1)}{(1 + U_i)(1 + 2U_i)} - \sqrt{n} \int_0^1 q(z)dQ_n^2(z).$$

The variance of this statistic is equal to 0.0030, while the shift under the alternatives becomes $\sqrt{n} \int_0^1 q(z)dz^3 = \sqrt{n} 0.0048$. As a result, the signal to noise ratio, for $n = 400$, is 1.74 and hence, the power of the linear test here is less than what it was in Example 3.4, although not by much.

4. Local alternatives of dependence

Interestingly, the assumption of independence within the pairs can also be tested in the colour blind situation. In this section, we assume $\{(X_i, Y_i)\}_{1 \leq i \leq n}$ to be a sample from a distribution with equal marginals that may not be independent and for which, again, the order in each pair is unobservable. The tests can be based on the same process introduced in Definition 2.1 and the problem is actually easier than the problem consider in the previous sections: it is possible to detect local alternatives converging to the null hypothesis, with a rate of $n^{-1/2}$.

We introduce a new class of alternative hypotheses

$$H'_a : \mathbb{A}'_n = \mathbb{Q} + \varepsilon_n \mathbb{G}_n, \tag{4.1}$$

where

$$\frac{d\mathbb{A}'_n(x, y)}{d\mathbb{Q}(x, y)} = 1 + \varepsilon_n g_n(x, y), \quad g_n \in L^2(Q \times Q), \tag{4.2}$$

and there exists $g \in L^2(Q \times Q)$ such that

$$\int_0^1 \int_0^1 [g_n(x, y) - g(x, y)]^2 dQ(x) dQ(y) \rightarrow 0, \text{ as } n \rightarrow \infty.$$

The function $g(x, y)$ describes the functional direction, from which the sequence of alternatives $\{\mathbb{A}'_n\}_{n \geq 1}$ approaches \mathbb{Q} .

However, as it is defined so far, deviations of \mathbb{A}'_n from \mathbb{Q} may be different for different Q . It will bring some economy in the description and more clarity, if we group alternatives according to how they deviate from the product of their marginals. Moreover, it is natural to assume that the two marginals are equal (see below).

Therefore, we restrict the choice of g_n by two conditions: there exists $g_0 \in L^2([0, 1]^2)$, such that

$$\int_0^1 \int_0^1 [g_n(x, y) - g_0(Q(x), Q(y))]^2 dQ(x) dQ(y) \rightarrow 0, \text{ as } n \rightarrow \infty, \quad (4.3)$$

and

$$\int_0^1 g_n(x, y) dQ(x) = \int_0^1 g_n(x, y) dQ(y) = 0. \quad (4.4)$$

The last condition guarantees, that the marginal distributions of X_i and Y_i under \mathbb{A}'_n are the same as under \mathbb{Q} , while, in (4.3), the function g_0 describes a functional direction from which the copula function $\mathbb{A}'_n(Q^{-1}(t), Q^{-1}(s))$, corresponding to \mathbb{A}'_n , approaches the product ts . This function inherits the property (4.4):

$$\int_0^1 g_0(t, s) dt = \int_0^1 g_0(t, s) ds = 0. \quad (4.5)$$

The geometric interpretation of this condition shows that it can be always made without loss of generality. Indeed, the local alternatives introduced in (3.6) for testing the identical marginal distributions assumption, approach \mathbb{Q} from the functional direction $h(x) - h(y)$. This difference, as a function in $L^2(Q \times Q)$, is orthogonal to the function $g_0(Q(x), Q(y))$ if g_0 satisfies condition (4.5). Any alternative defined by (4.1) can be decomposed into a part, which satisfies (4.4), and a part, which describes the changes in marginal distributions, already discussed in the previous section. Therefore, the new part in testing independence is represented by the functions g_0 satisfying (4.5). In a somewhat different situation, this was nicely expressed in Section 5 of [9]. In describing hypothesis testing for product measures the authors suggested to regard any deviations from the marginals as nuisance within the null hypothesis of independence.

In dependence alternatives, as we said above, ε_n can again have the classical rate of $n^{-1/2}$. Indeed, with $\mathbb{G}_n(x, y) = \int_0^x \int_0^y g_n(x', y') dQ(x') dQ(y')$, we have

$$E_{a'}[v_n(x, y)] = E_{a'}\{\sqrt{n}[\mathbb{P}_n(x, y) - \mathbb{Q}(x, y)]\} = \sqrt{n}\varepsilon_n \mathbb{G}_n(x, y),$$

and noting that $\mathbb{G}_n(1, y) = \mathbb{G}_n(x, 1) = 0$, we obtain

$$\begin{aligned} E_{a'}[\mathbb{R}_n(x, y)] &= \sqrt{n}\varepsilon_n\mathbb{G}_n(x, y), \\ E_{a'}[\mathbb{R}_n^s(u, v)] &= \sqrt{n}\varepsilon_n[\mathbb{G}_n(u, v) + \mathbb{G}_n(v, u) - \mathbb{G}_n(v, v)]. \end{aligned}$$

Therefore the rate $n^{-1/2}$ can render the shift $E_{a'}(\mathbb{R}_n^s)$ non-zero.

At the same time, the symmetrisation, again, can make an alternative undetectable. To see that, let

$$g_n^a(x, y) = [g_n(x, y) - g_n(y, x)]/2$$

be the anti-symmetric part of g_n . This part will make zero contribution to the shift of $E_{a'}(\mathbb{R}_n^s)$, so that if g_n is itself anti-symmetric, then the shift of $E_{a'}(\mathbb{R}_n^s)$ becomes 0, and the alternatives \mathbb{A}'_n become undetectable. For example, alternatives of the form

$$\mathbb{A}'_n(x, y) = xy + \varepsilon_n xy(1-x)(1-y)(y-x), \quad 0 \leq x \leq 1, \quad 0 \leq y \leq 1$$

cannot be detected by \mathbb{R}_n^s . Hence, in what follows, we restrict our attention to alternatives of dependence, converging from a functional direction, g that has a non-zero symmetric part.

As in Section 3.2, under the alternative of dependence H'_a , for any symmetric $\tilde{\varphi} \in L^2(Q \times Q)$ we have

$$\begin{aligned} v_n(\tilde{\varphi}) &= \sqrt{n} \int_0^1 \int_0^1 \tilde{\varphi}(x, y) [d\mathbb{P}_n(x, y) - d\mathbb{A}'_n(x, y)] \\ &+ \sqrt{n}\varepsilon_n \int_0^1 \int_0^1 \tilde{\varphi}(x, y)g_n(x, y)dQ(x)dQ(y). \end{aligned}$$

The first term of the sum contains the centered part of $v_n(\tilde{\varphi})$ and converges in distribution to $v_Q(\tilde{\varphi})$, and if $\varepsilon_n = n^{-1/2}$, then, as $n \rightarrow \infty$, under H_0 and under H'_a , respectively, we obtain

$$v_n(\tilde{\varphi}) \xrightarrow{d} v_Q(\tilde{\varphi}), \quad \text{and} \quad v_n(\tilde{\varphi}) \xrightarrow{d} v_Q(\tilde{\varphi}) + \langle \tilde{\varphi}, g \rangle_{Q \times Q}.$$

We write $g(x, y) = g^s(x, y) + g^a(x, y)$, where $g^s(x, y) = (1/2)[g(x, y) + g(y, x)]$ and $g^a(x, y) = (1/2)[g(x, y) - g(y, x)]$ are, respectively, its symmetric and anti-symmetric parts and note that, due to the symmetry of $\tilde{\varphi}$, we have

$$\langle \tilde{\varphi}, g \rangle_{Q \times Q} = \langle \tilde{\varphi}, g^s \rangle_{Q \times Q}.$$

The next result gives the form of the optimal linear test statistic that can be used for testing the null hypothesis against the sequence of dependence alternatives, H'_a .

Proposition 4.1. *The statistic of the asymptotically most powerful test for testing H_0 against the sequence of alternatives \mathbb{A}'_n is of the form*

$$\mathbb{R}_n^s(g^s) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{g(U_i, V_i) + g(V_i, U_i)}{2} + o_{\mathbb{Q}}(1).$$

The distance, in total variation, between its asymptotic distributions, under the null and under the symmetric alternatives \mathbb{A}'_n , is equal to $N_{(0,1)}(T/2)$, where T represents the limit of $\sqrt{n}\varepsilon_n\|g_0\|$.

Proof. Obviously, for any function g , its symmetric part satisfies $\tilde{g}^s = g^s$ and due to (4.4), we have $\mathcal{L}^*g^s = g^s$. Then, the conclusion follows along the lines of the proof of Proposition 3.3. \square

Example 4.2. The statistic of the asymptotically most powerful test for testing H_0 against the sequence of dependence alternatives

$$\mathbb{A}'_n(x, y) = xy + \varepsilon_n xy(1-x)(1-y), \quad |\varepsilon_n| \leq 1$$

is given by

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (1-2U_i)(1-2V_i).$$

Here the function $g(x, y) = g_0(x, y) = (1-2x)(1-2y)$ satisfies the condition (4.4) and is symmetric. Hence, the detection of \mathbb{A}'_n from its copula is possible and the test has the same power, regardless of whether the colours can be distinguished or not. For the sequence of dependence alternatives

$$\mathbb{A}'_n(x, y) = xy + \varepsilon_n xy(1-x)(1-y^2), \quad |\varepsilon_n| \leq 1$$

the function $g(x, y) = (1-2x)(1-3y^2)$ also satisfies (4.4) but is not symmetric. Hence, the detection of this sequence of alternatives in the colour blind situation, i.e. using the statistic

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n [(1-2U_i)(1-3V_i^2) + (1-3U_i^2)(1-2V_i)]$$

is connected with the loss of power. An example of a sequence of alternatives, undetectable with $\varepsilon_n \sim n^{-1/2}$ in the colour blind situation was given above.

Acknowledgements

The authors would like to thank two anonymous referees for valuable suggestions, which lead to an improved version of the article. Furthermore, the authors thank Adrian Kennedy and Joel Bancolita for their help with figures and numerical calculations. This help was wider than what eventually is included in this text.

References

- [1] BANERJEE, T., CHATTOPADHYAY, G. and BANERJEE, K. (2017). Two stages test of means of unordered pairs. *Statistics in Medicine*, **36**, 2466–2480. [MR3660144](#)

- [2] BERNSTEIN, A. V. and SIDOROV, A. A. (1972). Estimates of the set of expectations for a normal population. *Theory of Probability and Its Applications*, **17**, 723–726. [MR0312628](#)
- [3] BLUM, J. R., KIEFER, J. and ROSENBLATT, M. (1961). Distribution free tests of independence based on the sample distribution function. *Annals of Mathematical Statistics*, **32**, 485–496. [MR0125690](#)
- [4] DAVIES, P. and PHILLIPS, A. J. (1988). Nonparametric tests of population differences and estimation of the probability of misidentification with unidentified paired data. *Biometrika*, **75**, 753–760. [MR0995117](#)
- [5] DAY, S. J. and ALTMAN, D. G. (2000). Blinding in clinical trials and other studies. *BMJ*, **321**, 504.
- [6] GINDILIS, V. M. (1966). Mitotic chromosome spiralization and karyogrammic analysis in man. *Citologia*, **8**, 144–157.
- [7] HÁJEK, J., ŠIDÁK, Z. and SEN, P. K. (1999). *Theory of rank tests*, 2nd edition. Academic Press. [MR0334369](#)
- [8] HINKLEY, D. V. (1973). Two-sample tests with unordered pairs. *Journal of the Royal Statistical Society, Series B*, **35**, 337–346. [MR0334369](#)
- [9] JANSSEN, A. and RAHNENFÜHRER, J. (2002). A hazard-based approach to dependence tests for bivariate censored models. *Mathematical Methods of Statistics*, **11**, 297–322. [MR1964449](#)
- [10] KUO, H. -H. (1975). *Gaussian measures in Banach spaces*, Lecture Notes in Mathematics. Springer-Verlag, Berlin-New York. [MR0461643](#)
- [11] LAUDER, I. J. (1977). Tracing quantitative measurements on human chromosomes in family studies. *Annals of Human Genetics*, **41**, 77–82.
- [12] LI, P. and QIN, J. (2011). A new nuisance-parameter elimination method with application to the unordered homologous chromosome pairs problem. *Journal of the American Statistical Association*, **106**, 1476–1484. [MR2896850](#)
- [13] MILLER, F., FRIEDE, T. and KIESER, M. (2009). Blinded assessment of treatment effects utilizing information about the randomization block length. *Statistics in Medicine*, **28**, 1690–1706. [MR2675245](#)
- [14] OOSTERHOFF J. and VAN ZWET W. R. (2012). A note on contiguity and hellinger distance. In: van de Geer S., Wegkamp M. (eds) *Selected Works of Willem van Zwet. Selected Works in Probability and Statistics*. Springer, New York, NY. [MR0561267](#)
- [15] PARSADANISHVILI, E. G. and KHMALADZE, E. V. (1982). The testing of statistical hypotheses on unidentifiable objects. *Theory of Probability and Its Applications*, **27**, 175–182. [MR0645140](#)
- [16] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak convergence and empirical processes: With applications to statistics*. Springer Series in Statistics. Springer-Verlag, New York. [MR1385671](#)