Acta Math. Hungar.
https://doi.org/10.1007/s10474-021-01164-4

RÉNYI 100, QUANTITATIVE AND QUALITATIVE (IN)DEPENDENCE

M. ARATÓ[1,2], GY. O. H. KATONA[3], GY. MICHALETZKY[1], T. F. MÓRI[3,*], J. PINTZ[3], T. RUDAS[4], G. J. SZÉKELY[5,3] and G. TUSNÁDY[3]

[1]Department of Probability Theory and Statistics, Loránd Eötvös University, Pázmány Péter s. 1/C, H-1117 Budapest, Hungary
e-mails: miklos.arato@ttk.elte.hu, gyorgy.michaletzky@ttk.elte.hu

[2] Department of Mathematics and Computational Sciences, Széchenyi István University, Egyetem tér 1, H-9026 Győr, Hungary

[3]Alfréd Rényi Institute of Mathematics, Reáltanoda u. 13–15, H-1053 Budapest, Hungary
e-mails:
katona.gyula.oh@renyi.hu, mori.tamas@renyi.hu, pintz.janos@renyi.hu, tusnady.gabor@renyi.hu

[4]Department of Statistics, Faculty of Social Sciences, Loránd Eötvös University, Pázmány Péter s. 1/A, H-1117 Budapest, Hungary
e-mail: trudas@elte.hu

[5]National Science Foundation, 2415 Eisenhower Avenue, Alexandria, VA 22314, USA
e-mail: gszekely@nsf.gov

(Received February 19, 2021; revised April 22, 2021; accepted April 29, 2021)

We dedicate this paper to the memory of Alfréd Rényi who was born on March 20, 1921 and died on February 1, 1970 at the age of 49. He is the founding father of modern probability theory, information theory, and mathematical statistics in Hungary.

Abstract. We discuss recent developments in the following important areas of Alfréd Rényi's research interest: axiomatization of quantitative dependence measures, qualitative independence in combinatorics, conditional qualitative independence in statistics/data science and in measure theory/probability theory, and finally, prime gaps that are responsible for Rényi's early career reputation. Most authors of this paper are main contributors to the new developments.

* Corresponding author.

Supported by the Hungarian National Research, Development and Innovation Office (NKFIH) grants No. K119528 and KKP133819 to JP, K125569 to TFM.

Key words and phrases: axioms of dependence measures, qualitative independence, conditional qualitative independence, prime gaps.

Mathematics Subject Classification: primary 00-02, secondary 05D05, 11N05, 60E05, 62H15, 62H17, 62H20.

1. Introduction: Rényi and dependence measures

It is hard to overestimate the importance of dependence measures in statistics and in science. When we try to find the cause X that is (partly) responsible for an effect Y then it is a natural first step to find out if X and Y are statistically dependent. Thus, it is not surprising that Pearson's linear correlation r is responsible for many important causal discoveries like smoking and lung cancer. Unfortunately if $r = 0$ then we might suspect that there is no causal relationship between X and Y even when there is. Pearson's correlation $r = 0$ does not imply independence. This is a typical problem when the relationship between the variables is highly nonlinear, not even monotonic. The importance of dependence measures led Rényi to introduce seven axioms of dependence measures in [113]. We will discuss these quantitative dependence measures in Sections 2–7. The next block, Sections 8–10 is on qualitative independence. Rényi discussed this notion and its relevance to combinatorics in his seminars and also in his book [115]. The last section is on prime gaps, a classical topic in number theory that brought international reputation for Rényi in 1947–48. The "randomness" of prime gaps connects primes and independence.

2. Rényi's axioms of dependence

For real valued random variables X, Y, $\Delta(X, Y)$ is a dependence measure if it satisfies Rényi's axioms:

(A) $\Delta(X, Y)$ is defined for all random variables X and Y, neither of them being constant with probability 1.

(B) $\Delta(X, Y) = \Delta(Y, X)$ (symmetry).

(C) $0 \leq \Delta(Y, X) \leq 1$.

(D) $\Delta(X, Y) = 0$ if and only if X and Y are independent.

(E) $\Delta(X, Y) = 1$ if there is a strict dependence between X and Y; that is, either $X = g(Y)$ or $Y = f(X)$, where g and f are Borel measurable functions.

(F) If the Borel measurable functions $f(x)$ and $g(x)$ map the real axis in a 1–1 way onto itself, $\Delta(f(X), g(Y)) = \Delta(X, Y)$.

(G) If the joint distribution of X and Y is normal, then $\Delta(X, Y) = |r(X, Y)|$ where $r(X, Y)$ is the correlation coefficient of X and Y.

All these axioms are satisfied by maximal correlation which is the supremum of all correlations $r(f(X), g(Y))$ for which the correlation of $f(X)$ and $g(Y)$ exists. Maximal correlation seemed to be the final word in this topic and it seemed that maximal correlation was the dependence measure everybody was looking for. In [113] Rényi proved that under explicit technical conditions the maximal correlation is actually attained and characterized. Namely, let $\mathcal{L}_X^2$ denote the Hilbert space of all random variables of the

form $f(X)$ for which $\mathbb{E}f(X) = 0$ and $\mathrm{Var}(f(X))$ is finite, and similarly, $\mathcal{L}_Y^2$ the Hilbert space of all those random variables $g(Y)$ for which $\mathbb{E}g(Y) = 0$ and $\mathrm{Var}(g(Y))$ is finite. Let us put for any $f = f(X) \in \mathcal{L}_X^2$

$$Af = \mathbb{E}\big(\mathbb{E}(f(X) \mid Y) \mid X\big). \tag{2.1}$$

Then A is a bounded linear transformation of the Hilbert space $\mathcal{L}_X^2$; moreover, it is self-adjoint and positive definite.

THEOREM 2.1 [113]. *If the transformation A defined by* (2.1) *is completely continuous, then the maximal correlation of X and Y is attained for $f_0(X)$ and $g_0(Y)$ where f_0 is an eigenfunction belonging to the greatest eigenvalue λ of A and $g_0(Y) = \lambda^{-1/2}\,\mathbb{E}(f_0(X) \mid Y)$.*

The condition that A should be completely continuous is not easy to verify in concrete cases. Therefore the following theorem is useful.

THEOREM 2.2 [113]. *If the joint distribution $Q_{(X,Y)}$ of X and Y is absolutely continuous with respect to the direct product $Q_X \times Q_Y$ of their distributions, and*

$$\int_{\mathbb{R}^2} \Big(\frac{dQ_{X,Y}}{d(Q_X \times Q_Y)} - 1\Big)^2 d(Q_X \times Q_Y) < \infty,$$

then the transformation A is completely continuous and thus the maximal correlation of X and Y can be attained.

Rényi's paper was cited more than 800 times. Rényi himself applied the notion of maximal correlation to a probabilistic generalization of Linnik's "large sieve", a very nice extension of a classical tool. Many years later it turned out that the empirical maximal correlation is (almost) always 1 no matter what the statistical sample is; thus, if the maximal correlation were (weakly) continuous then it would be identically 1 which contradicts axiom (D). A remedy for the non-continuity of maximal correlation is distance correlation, see Section 3.

3. New axioms

Let S be a nonempty set of pairs of nondegenerate random variables X, Y taking values in Euclidean spaces or in real, separable Hilbert spaces H. (Nondegenerate means that the random variable is not constant with probability 1.) Then $\Delta\colon S \to [0,1]$ is called a dependence measure on S if the following four axioms hold. In the axioms below we need similarity transformations of H. Similarity of H is defined as a bijection (1–1 correspondence) from H onto itself that multiplies all distances by the same positive real number (scale). Similarities are known to be compositions of a translation, an

orthogonal linear mapping, and a uniform scaling. We assume that if $(X, Y) \in S$ then $(LX, MY) \in S$ for all similarity transformations L, M of H.

(i) $\Delta(X, Y) = 0$ if and only if X and Y are independent.

(ii) $\Delta(X, Y)$ is invariant with respect to all similarity transformations of H; that is, $\Delta(LX, MY) = \Delta(X, Y)$ where L, M are similarity transformations of H.

(iii) $\Delta(X, Y) = 1$ if and only if $Y = LX$ with probability 1, where L is a similarity transformation of H.

(iv) $\Delta(X, Y)$ is continuous; that is, if for some positive constant K we have $E\big(|X_n|^2 + |Y_n|^2\big) \le K$, $n = 1, 2, \dots$ and (X_n, Y_n) converges weakly (converges in distribution) to (X, Y) then $\Delta(X_n, Y_n) \to \Delta(X, Y)$. (The condition on the boundedness of second moments can be replaced by any other condition that guarantees the convergence of expectations: $\mathbb{E}(X_n) \to \mathbb{E}(X)$ and $\mathbb{E}(Y_n) \to \mathbb{E}(Y)$; such a condition is uniform integrability of X_n, Y_n which follows from the boundedness of second moments.)

The goal of the new system of axioms is not to characterize a single dependence measure. The new system of axioms is "minimalist" in the sense that all good dependence measures can be expected to satisfy them. In [95] it is shown that even this "minimalist" system of axioms can disqualify several classical measures and also some recently introduced measures of dependence; for example, neither the maximal correlation coefficient nor the recently introduced maximal information coefficient satisfy axiom (iv). The same axiom fails to hold for the correlation ratio.

Remark 3.1. (a) Functions of independent random variables are independent, thus property $\Delta(X, Y) = 0$ is invariant with respect to all 1–1 Borel measurable transformations of H. On the other hand we do not suppose this 1–1 invariance for other values of Δ. As we shall see, such a strong condition would contradict axiom (iv). In axiom (ii) and (iii) one can try to replace the invariance with respect to similarities by other groups of invariances, particularly, when the statistical problem in question exhibits symmetries/invariances in the sense of [83, Ch. 6], see also [39]. It is up to the statistician to choose the right level of invariance. Too much invariance is not necessarily good. Even if a very strong invariance of Δ does not contradict other important axioms it might decrease the power of Δ in testing independence. If $H = \mathbb{R}$, the real line, affine transformations coincide with similarities. In higher dimensions, however, affine invariance for all bounded nonconstant random variables contradicts axiom (iv) as it is proved in Theorem 6.1. This makes the choice of similarity invariance in our axioms even more natural.

(b) Rényi did not assume axiom (iv). Theorem 3.1 below explains that if he did then no dependence measure would have satisfied all his axioms.

(c) Why did we suppose that S does not contain random variables that are constant with probability 1? Because if Y is such a random variable

then it is independent of all random variables X and thus by axiom (i) we have $\Delta(X,Y)=0$. On the other hand, for all $X\in S$ axiom (iii) implies $\Delta(X,X/n)=1$ for $n=1,2,\dots$. But for bounded random variables X the limit of X/n is 0 and $\Delta(X,0)=0$ which contradicts axiom (iv). In axiom (A) Rényi also assumes that the random variables X and Y are not constant with probability 1, i.e., their distributions are nondegenerate. This assumption guarantees that Δ cannot be discontinuous at degenerate distributions because Δ is simply not defined there. Thus Rényi did not overlook the importance of weak continuity of Δ, he just could not assume it because it would have been inconsistent with his other axioms.

In what follows we will see that 1–1 invariance is not compatible with our new axiom of continuity (iv). But why is continuity so natural that one should suppose it as an axiom? If there is a tiny little change/perturbation in the distribution of (X,Y) and this tiny little perturbation changes $\Delta(X,Y)$ dramatically, e.g., changes it from 1 to 0 then Δ has no stability. We cannot rely our statistical inference on such an unstable Δ because a minor perturbation, no matter how small it is, can result in a completely different statistical inference. This can be viewed as a violation of distributional robustness. If we replace weak convergence by stronger forms of convergence then of course this would allow more measures of dependence to be continuous but these measures might violate distributional robustness. We do not need to disregard all nonrobust measures but we need to be aware of this deficiency.

THEOREM 3.1. *Suppose S is a set of pairs of non-constant random variables and if $(X,Y)\in S$ then $(LX,MY)\in S$ for all affine transformations L, M of H. If the dependence measure $\Delta(X,Y)$ on S is invariant with respect to all affine transformations L, M of H where $\dim H>1$ then axiom* (iv) *cannot hold. If $\dim H=1$ then affinity is the same as similarity and in this case distance correlation is affine invariant. On the other hand, if $\Delta(X,Y)$ is invariant with respect to all* 1–1 *Borel measurable functions of H then even if $\dim H=1$, axiom* (iv) *cannot hold.*

Recall that Euclidean geometry is characterized by invariances with respect to the Euclidean group of transformations (translations, rotations, and reflections). Similarity geometry deals with geometrical objects with the same shape. We can obtain one object from another by scaling (enlarging or shrinking). Similarity transformations consist of all Euclidean transformations and all (nonzero) scaling; that is, changing the measurement units. Instead of 1–1 invariance, in our axioms we suppose similarity invariance only. Similarity invariance is something we do not want to weaken because changing the scale, (that is, changing the measurement unit), should not affect the degree of dependence. Luckily, similarity invariance does not contradict continuity. Let us see that our system of new axioms is not contradictory

when S is the set of all nondegenerate random variables with finite expectation. For this it is sufficient to define a dependence measure that satisfies the four axioms. Such a measure is *distance correlation*, which was introduced in [124].

First of all recall the definition of the sample distance correlation. Take all pairwise distances between sample values of one variable, and do the same for the second variable. Rigid motion invariance is automatically guaranteed if instead of sample elements we work with their distances. Another advantage of working with distances is that they are always real numbers even when the data are vectors of possibly different dimensions. Once we have computed the distance matrices of both samples, double-center them (so each has column and row means equal to zero). Then average the entries of the matrix which holds componentwise products of the two centered distance matrices. This is the square of the sample distance covariance. If we denote the centered distances by A_{ij}, $i,j=1,\dots,n$ and B_{ij}, $i,j=1,\dots,n$ where n is the sample size, then the squared sample distance covariance is

$$\frac{1}{n^2}\sum_{i,j=1}^{n} A_{i,j}B_{i,j}.$$

This definition is very similar to, and almost equally simple as, the definition of Pearson's covariance, except that here we have double indices.

The population squared distance covariance can be reduced to the following form if $\mathbb{E}|X|^2$ and $\mathbb{E}|Y|^2$ are finite [124]. Let (X,Y), (X',Y'), (X'',Y'') be independent and identically distributed then the distance covariance is the square root of

$$\begin{aligned}(3.1)\qquad \mathrm{dCov}^2(X,Y) :=\ & \mathbb{E}(|X-X'|\,|Y-Y'|) + \mathbb{E}(|X-X'|)\mathbb{E}(|Y-Y'|)\\ & - \mathbb{E}(|X-X'|\,|Y-Y''|) - \mathbb{E}(|X-X''|\,|Y-Y'|).\end{aligned}$$

In the above referred paper it is proved that $\mathrm{dCov}(X,Y)$ is a metric, and the distance variance, $\mathrm{dCov}(X,X)$ is zero if and only if X is constant with probability 1. Once we defined distance covariance and distance variance we can define distance correlation the same way as we defined correlation with the help of covariance and variance. If the random variables X,Y have finite expected values and they are not constant with probability 1 then the definition of *population distance correlation* is the following:

$$\mathrm{dCor}(X,Y) := \frac{\mathrm{dCov}(X,Y)}{\sqrt{\mathrm{dCov}(X,X)\,\mathrm{dCov}(Y,Y)}}\,.$$

If $\mathrm{dCov}(X,X)\,\mathrm{dCov}(Y,Y)=0$ then define $\mathrm{dCor}(X,Y)=0$. Distance correlation equals zero if and only if the variables are independent, whatever

be the underlying distributions and whatever be the dimension of the two variables (for a transparent explanation see below). This fact and the simplicity of the statistic make distance correlation an attractive candidate for measuring dependence. For generalizations to metric spaces see [63,84,85].

In [124] an alternative formula for $\mathrm{dCov}^2(X,Y)$ was given in terms of characteristic functions $f_{X,Y}$, f_X and f_Y of (X,Y), X, and Y respectively. If the random variable X takes values in a p-dimensional Euclidean space $\mathbb{R}^p$ and Y takes values in $\mathbb{R}^q$ and both variables have finite expectations we have

$$\mathrm{dCov}^2(X,Y) := \frac{1}{c_p c_q} \int_{\mathbb{R}^{p+q}} \frac{|f_{X,Y}(t,s) - f_X(t) f_Y(s)|^2}{|t|_p^{1+p}\ |s|_q^{1+q}}\, dt\, ds.$$

where c_p and c_q are constants. This formula clearly shows that independence of X and Y is equivalent to $\mathrm{dCov}(X,Y) = 0$. It is interesting to note that in Hoeffding's dissertation [58] the following expression is proved for the Pearson's covariance of real valued X and Y with finite variances:

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [F_{X,Y}(x,y) - F_X(x)F_Y(y)]\, dx\, dy,\end{aligned}$$

where F denotes the cumulative distribution functions. Thus we might want to define a sign or rather a direction of distance covariance and distance correlation as the argument of the complex number

$$z := \int_{\mathbb{R}^{p+q}} [f_{X,Y}(t,s) - f_X(t) f_Y(s)]\, w(t,s)\, dt\, ds,$$

where $w(t,s)$ is a suitable weight function. In the most natural case of $w(-t,-s) = w(t,s)$, this z is always real, so its direction is not more than a sign. Unfortunately in the most natural choice for w when $w(s,t) = \left(|t|_p^{1+p}|s|_q^{1+q}\right)^{-1}$, it is not trivial that z exists at all. We also note that in [59] a test of independence was introduced, based on

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [F_{X,Y}(x,y) - F_X(x)F_Y(y)]^2\, dF_{X,Y}(x,y).$$

If the expectations of X, Y do not exist, we can generalize distance correlation for random variables with finite $\alpha > 0$ moments, see Section 4. It is easy to see that the population distance correlation, $\mathrm{dCor}(X,Y)$, satisfies axioms (ii) and (iv). For the proof that $\mathrm{dCor}(X,Y)$ satisfies (i) and (iii), see [124].

In the special case when (X, Y) are jointly distributed as bivariate normal, distance correlation is a deterministic function of Pearson correlation $r = r(X, Y)$ [124, Theorem 7], namely,

$$\mathrm{dCor}^2(X, Y) = \frac{r \arcsin r + \sqrt{1 - r^2} - r \arcsin(r/2) - \sqrt{4 - r^2} + 1}{1 + \pi/3 - \sqrt{3}}.$$

Note that this is a strictly increasing, convex function of $|r|$, and $\mathrm{dCor}(X, Y) \leq |r(X, Y)|$ with equality when $r = 0$ or $r = \pm 1$. Thus $\mathrm{dCor}(X, Y)$ does not satisfy Rényi's axiom (G). It is also clear that if Δ satisfies our four axioms then $h(\Delta)$ also satisfies them whenever h is a strictly increasing, continuous function, $h(0) = 0$, $h(1) = 1$, and $0 < h(x) < 1$ for $0 < x < 1$. In the definition of partial distance correlation $h(x) = x^2$ is applied [127]. In this case the distance standard deviations of the random variables X, Y are measured in the same units as the X distances and Y distances. If we insisted on axiom (G) we would disqualify distance correlation and also its square and instead would have accepted a complicated function of distance correlation as "legitimate".

An important generalization of distance correlation is introduced in [123]. This is related to a generalized distance correlation where the distance is a more general metric than the Euclidean one. These generalizations under some natural conditions like scale invariance also satisfy axioms (i)–(iv).

On a completely different axiomatic approach of dependence measures see [65].

4. On the relationship between Pearson's correlation and distance correlation

Since both the absolute value of Pearson's correlation coefficient and the distance correlation coefficient are used in applications to quantify strength of dependence, it is important to understand how large the differences between these two measures can possibly be. It is immediately clear that $\mathrm{dCor}(X, Y) = 0$ implies $r(X, Y) = 0$ and that $|r(X, Y)| = 1$ if and only if $\mathrm{dCor}(X, Y) = 1$. It is also straightforward to show that the distance correlation coefficient of random variables only having two possible values coincides with their absolute Pearson correlation coefficient. Moreover, several results for bivariate parametric distributions have been derived, see [38,124].

The fact that dCor is defined for X and Y with finite first moments, while $r(X, Y)$ requires finite second moments leads us to the conjecture that the Pearson correlation coefficient is more sensitive to dependence in the tails than the distance correlation. This conjecture motivated the construction of a specific mixture distribution in [40] showing that, up to trivial exceptions,

all possible values for the Pearson correlation coefficient r and dCor can be simultaneously achieved.

This result can be proved in a more general form. If the expectations of X, Y do not exist, we can generalize distance correlation for random variables with finite $\alpha > 0$ moments by taking the α-th powers of the distances in (3.1), with $0 < \alpha < 2$, see [124,125]. That is, let

$$\begin{aligned} \mathrm{dCov}_\alpha^2(X,Y) &= \mathbb{E}\big(|X-X'|^\alpha |Y-Y'|^\alpha\big) \\ &+ \mathbb{E}\big(|X-X'|^\alpha\big)\,\mathbb{E}\big(|Y-Y'|^\alpha\big) - 2\mathbb{E}\big(|X-X'|^\alpha |Y-Y''|^\alpha\big). \end{aligned} \tag{4.1}$$

By this definition, the α-distance correlation coefficient can be expressed in the usual way, as

$$\mathrm{dCor}_\alpha(X,Y) = \frac{\mathrm{dCov}_\alpha(X,Y)}{\sqrt{\mathrm{dCov}_\alpha(X,X)\,\mathrm{dCov}_\alpha(Y,Y)}}$$

provided the denominator is positive. dCor_α shares all advantageous properties of dCor; in particular $\mathrm{dCor}_\alpha(X,Y) = 0$ if and only if X and Y are independent [125]. Moreover, while definition (4.1) only holds for random variables X and Y with finite moments of order 2α, the definition of $\mathrm{dCor}_\alpha(X,Y)$ can be straightforwardly extended to X, Y with moments of order α. For $\alpha = 1$ we get back the distance correlation coefficient of (3.1).

THEOREM 4.1 [40]. *Let $0 < \alpha < 2$. For every pair (r_1, r_2), $-1 < r_1 < 1$, $0 < r_2 < 1$, there exist random variables X, Y with finite moments of order 2α, such that $r(X,Y) = r_1$, $\mathrm{dCor}_\alpha(X,Y) = r_2$.*

In addition to the set above, the only possible values of the pair $\big(r(X,Y), \mathrm{dCor}_\alpha(X,Y)\big)$ are $(-1,1)$, $(0,0)$ and $(1,1)$.

Let $0 < \alpha < \beta \le 2$. We conjecture that for every pair $0 < r_1, r_2 < 1$ there exist random variables X, Y with finite moments of order α and β, resp., such that $\mathrm{dCor}_\alpha(X,Y) = r_1$ and $\mathrm{dCor}_\beta(X,Y) = r_2$.

5. The earth mover's correlation and why we need it

In Section 3 we explained that for real valued or even for separable Hilbert space valued random variables distance correlation is a very good measure of dependence. In many new areas of statistical applications like brain research or network analysis the underlying metric spaces are not Hilbert spaces. For some of these spaces distance correlation works but for many others it does not. When it does not, we need a new measure of dependence that hopefully works in all metric spaces. The Earth mover's correlation we are about to introduce in this section is a good candidate for

such a dependence measure. It works in all metric spaces but we need to pay the price for that which is computational complexity. Let us see the details.

Distance correlation can be generalized to metric spaces $(\mathcal{M}, \delta)$ that are of negative type [84]. A metric space $(\mathcal{M}, \delta)$ is called of negative type if the metric possesses the "conditional negative definite" property, namely that for all integers $n \geq 1$ and for all sets of n points $x_i \in \mathcal{M}$ and $x_i' \in \mathcal{M}$ $(i = 1, 2, \ldots, n)$ and for all real numbers $a_1, a_2, \ldots, a_n$ such that their sum is 0 we have

$$\sum_{i,j} a_i a_j \delta(x_i, x_i') \leq 0.$$

Strong negative type metric spaces satisfy this with equality iff $a_1 = \cdots = a_n = 0$. However, for the strong negative type property we need somewhat more, namely for all probability measures μ and ν defined on the Borel sets of $\mathcal{M}$

$$\int \delta(x, y)\, d(\mu - \nu)^2(x, y) \leq 0$$

with equality iff $\mu = \nu$.

According to a classical theorem of Schoenberg [120,121] a necessary and sufficient condition for negative type of $(\mathcal{M}, \delta)$ is that $(\mathcal{M}, \sqrt{\delta})$ is isometrically embeddable into a Hilbert space. Obviously this property does not hold for every metric space. When it does then in these "nice" metric spaces we can apply distance correlation, for all others we need to make new efforts.

We can try to work with functions of δ, say $\delta^*(\delta)$, that satisfies our axioms. If the only problem is that the metric is not of strong negative type, only of negative type then it is easy to find a remedy: take the square root (or any other power $0 < \alpha < 1$) of the metric and this new metric becomes of strong negative type, see [84].

For arbitrary finite $\mathcal{M}$ one can show, see [127], that for a suitably large number K the new distance $\delta^*(x, y) = \delta(x, y) + K$ whenever $x \neq y$ and 0 otherwise, is always conditionally negative definite. On top of that, this simple transformation of the metric does not change the unbiased estimator of dCov which is simply invariant with respect to this additive constant K.

For infinite $\mathcal{M}$ there does not always exist a strictly monotone increasing function $\delta^*(\delta)$ such that $(\mathcal{M}, \delta^*)$ is of negative type. Take e.g. two disjoint infinite sets, A and B, and let $\mathcal{M}$ be their union. Define the distance of two distinct elements to be 1 if they are in different sets, and 2 if they are in the same set. The function δ^* must have the following form: $\delta^*(1) = u$, $\delta^*(2) = v$, $0 < u < v$. Define $a_i := 1$ for n elements of A and $a_i := -1$ for n elements of B. Then the sum we need to check is $n(n-1)v - n^2 u$, which is positive for large enough n.

Another approach is this. If all we want from our dependence measure is to test independence then it is acceptable to change the distances

in $(\mathcal{M}, \delta)$ and thus change the distance correlation so long as we do not change $\mathrm{dCor}(X, Y) = 0$. If f is an arbitrary 1–1 Borel function on $(\mathcal{M}, \delta)$ and X, Y are $(\mathcal{M}, \delta)$ valued random variables then they are independent iff $f(X), f(Y)$ are independent. But every metric space is Borel isomorphic to a "nice" metric space that is embeddable isomorphically into a Hilbert space. According to Kuratowski's theorem two complete separable Borel spaces are Borel isomorphic iff they have the same cardinality. They are Borel isomorphic either to $\mathbb{R}$, or to $\mathbb{Z}$ or to a finite metric space. Denote this Borel isomorphism by f. If we can construct it then we can check the independence of the real valued random variables $f(X), f(Y)$ via distance correlation and this is equivalent to testing the independence of X, Y that take values in general metric spaces. We might want to make f continuous to avoid the negative effect of minor noise. In this case we can choose f to be a homeomorphism between our metric space and a subspace of a Hilbert cube. This f exists if and only if our metric space is separable. Here is how to construct such an f.

Assume $\delta \leq 1$ (otherwise, use $\delta/(\delta + 1)$). Choose a dense countable sequence (x_n) from $\mathcal{M}$ which exists because the metric space is separable, and define $f(x) := (\delta(x, x_n)/n)_{n \geq 1}$, a point in the Hilbert cube and here we can apply distance correlation for testing independence.

These tricks can help to solve some of the problems in testing independence but they do not solve the problem of finding a general measure of dependence applicable to general metric space valued random variables. The following quantity, called *the earth mover's correlation* was introduced in [96], with the goal to define a dependence measure that applies to the "rest of the universe" (with one of John von Neumann's favorite expressions).

First, let us define the population value of the earth mover's correlation.

Recall the definition of the earth mover's distance for probability measures μ, ν on general metric spaces $(\mathcal{M}, \delta)$. We suppose that the topology of this metric space and the probability measures on the Borel sets are "compatible", that is, we suppose that the probability measures are Radon measures (finite on compact sets, outer regular and inner regular).

Heuristically, if we have two (Radon) probability distributions, μ and ν on $(\mathcal{M}, \delta)$ then the earth mover's distance is the minimum cost of turning one pile of dust or dirt with distribution μ into the other with distribution ν. The cost is proportional to the transport distance and also to the amount of dirt we transport.

This distance was considered by [66,67,92,107,130,133], and many others, and in mathematical circles it is typically called Wasserstein distance. Most statisticians and computer scientists call it earth mover's distance. On a recent survey see [97].

Denote by $\mathcal{P}(\mathcal{M})$ the set of all (Radon) probability measures μ on $\mathcal{M}$. Suppose that for some $x_0 \in \mathcal{M}$ we have

$$\int_{\mathcal{M}} \delta(x, x_0)\, d\mu(x) < +\infty.$$

Then the earth mover's distance or Wasserstein distance of the probability measures μ and ν can be equivalently defined as

$$e(\mu, \nu) := \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{M}\times\mathcal{M}} \delta(x, y) d\gamma(x, y),$$

where $\Gamma(\mu, \nu)$ is the set of all possible couplings of probability measures μ and ν, that is, the set of all joint distributions γ of (X, Y) with marginal distributions μ and ν, respectively. Equivalently,

$$e(\mu, \nu) = e(X, Y) := \inf_{\gamma \in \Gamma(\mu,\nu)} \mathbb{E}[\delta(X, Y)],$$

where again the infimum is taken for all joint distributions of (X, Y) with marginal distributions μ and ν, respectively.

Mathematically this is not an easy minimization problem to solve. Even if $(\mathcal{M}, \delta)$ is an Euclidean space where the transportation cost is the Euclidean distance, the solution is related to the so-called Monge–Ampère difference equation, see [17,29,30]. For real valued random variables X, Y, however, there is a simple formula for the earth mover distance. Denote $F(x) = \mathbb{P}(X \le x)$ and $G(y) = \mathbb{P}(Y \le y)$ the cdf's of X and Y and consider their generalized inverses $F^{-1}(u)$, $G^{-1}(u)$, defined as $F^{-1}(u) = \sup\{t : F(t) \le u\}$. Then

$$e(X, Y) = \mathbb{E}\big|\, F^{-1}(U) - G^{-1}(U)\big|$$

$$= \int_0^1 \big|\, F^{-1}(u) - G^{-1}(u)\big|\, du = \int_{-\infty}^{\infty} \big|\, F(t) - G(t)\big|\, dt.$$

Define a metric d on the space $\mathcal{M} \times \mathcal{M}$, e.g. d can be the Manhattan distance: $d\big[(x, y), (u, v)\big] = \delta(x, u) + \delta(y, v)$. Then the *earth mover's covariance* of random variables X, Y taking values in $(\mathcal{M}, \delta)$ is defined as the earth mover's distance between the joint distribution and the product of its marginals:

$$\text{(5.1)} \qquad \mathrm{eCov}(X, Y) = \inf_{\gamma \in \Gamma} \mathbb{E} d\big[(X, Y), (X', Y')\big] = e\big[(X, Y), (X', Y')\big],$$

where Γ is the set of all possible joint distributions of the random variables X, Y, X', Y' such that X' and X are identically distributed, Y' and Y

are also identically distributed, and X', Y' are independent (and the joint distribution of X and Y is given).

In the following we do not really need that d is a Manhattan distance; what we need is more general, namely that $(\mathcal{M} \times \mathcal{M}, d)$ with a metric d is a metric space such that

$$d\big[(x,u),(x,v)\big] = \delta(u,v), \quad d\big[(x,u),(y,u)\big] = \delta(x,y),$$
$$d\big[(x,x),(u,v)\big] \geq \delta(u,v).$$

The following inequality is of Cauchy–Bunyakovsky–Schwarz type:

$$e^2\big[(X,Y),(X',Y')\big] \leq e\big[(X,X),(X,X')\big]\, e\big[(Y,Y),(Y,Y')\big],$$

where X and X' are iid, as well as Y and Y', and X', Y' are independent.

In fact, one can show more, namely that

THEOREM 5.1 [96].

$$e\big[(X,Y),(X',Y')\big] \leq \min\big\{e\big[(X,X),(X,X')\big],\, e\big[(Y,Y),(Y,Y')\big]\big\}.$$

On the right-hand side, $e\big[(X,X),(X,X')\big] = \mathrm{eCov}(X,X)$ will be called *the earth mover's variance* and denoted by $\mathrm{eVar}(X)$. It can be shown that the earth mover variance is the same as Gini's mean difference:

$$\mathrm{eVar}(Y) = \mathbb{E}\delta(Y,Y'), \tag{5.2}$$

where Y and Y' are iid.

Based on Theorem 5.1 we can now introduce the definition of a new type of correlation. The *earth mover's correlation* of the distributions of X and Y is defined as

$$\mathrm{eCor}(X,Y) = \frac{\mathrm{eCov}(X,Y)}{\min\big\{\mathrm{eVar}(X), \mathrm{eVar}(Y)\big\}}. \tag{5.3}$$

We do not define $\mathrm{eCor}(X,Y)$ when $\min\big\{\mathrm{eVar}(X), \mathrm{eVar}(Y)\big\} = 0$.

REMARK 5.1. (a) By the previous theorem in the formula for eCor the denominator $\min\big\{\mathrm{eVar}(X), \mathrm{eVar}(Y)\big\} = \min\big\{\mathbb{E}\delta(X,X'), \mathbb{E}\delta(Y,Y')\big\} = 0$ iff at least one of X, Y is constant with probability 1. In this case we do not define eCor. It is interesting to note that for real valued random variables eVar is easy to compute. It is known, see e.g. [135], that

$$\mathrm{eVar}(X) = 2\int_{-\infty}^{\infty} F(x)(1-F(x))\,dx,$$

where $F(x) = \mathbb{P}(X \leq x)$ is the cdf of the random variable X.

(b) Let us apply the Manhattan distance for pairs. Then by the triangle inequality for δ we have $\delta(X,X')+\delta(Y,Y')\geq|\delta(X,Y)-\delta(X',Y')|$, thus

$$\mathrm{eCov}(X,Y)\geq\inf_{(X',Y')}\mathbb{E}\big|\,\delta(X,Y)-\delta(X',Y')\big|\;\geq\big|\mathbb{E}\delta(X,Y)-\mathbb{E}\delta(X',Y')\big|.$$

For an example let us consider dependent indicators. Let X and Y be indicators, $\mathbb{P}(X=1)=1-\mathbb{P}(X=0)=p_X$, $\mathbb{P}(Y=1)=1-\mathbb{P}(Y=0)=p_Y$, $\mathbb{P}(X=Y=1)=p_{XY}$. Let us apply the Euclidean metric in $\mathbb{R}$ and the Manhattan distance for pairs. Then

$$\mathrm{eCor}(X,Y)=\frac{|p_{XY}-p_Xp_Y|}{\min\big\{\,p_X(1-p_X),\,p_Y(1-p_Y)\big\}}$$

see [96, Example 3.4].

The absolute value of Pearson's correlation for indicators is

$$|r(X,Y)|=\frac{|p_{XY}-p_Xp_Y|}{\sqrt{p_X(1-p_X)\,p_Y(1-p_Y)}}$$

thus for indicators X and Y we have $|r(X,Y)|\leq\mathrm{eCor}(X,Y)$ (and we have equality iff $p_X=p_Y$). Based on this observation one can suspect that $|r(X,Y)|\leq\mathrm{eCor}(X,Y)$ for all real valued random variables with finite variance. This conjecture is also supported by the fact that the independence of X,Y implies their uncorrelatednes. In the other extreme case when $r(X,Y)=\pm1$ we know that $Y=f(X)$ where f is a similarity (here a linear function) and it can be proved that in this case we have $\mathrm{eCor}(X,Y)=1$.

However, this conjecture can easily be disproved. If the joint distribution of X,Y is bivariate normal, the opposite inequality holds. Namely, in [96] it is shown that for bivariate normal (X,Y) with correlation $r(X,Y)=r$

$$\mathrm{eCor}(X,Y)\leq\Big[1-\sqrt{1-r^2}\,\Big]^{1/2},$$

which is strictly less than $|r|$, apart from the trivial cases of $r=0$ or $r=\pm1$, where $\mathrm{eCor}(X,Y)=|r|$. We conjecture that the inequality above holds with equality.

Concerning the lower bound of $\mathrm{eCor}(X,Y)$, if $\sigma_X=\sigma_y$ then Remark 5.1(b) provides the following inequality:

$$\big|\,1-\sqrt{1-\varrho}\,\big|\leq\mathrm{eCor}(X,Y). \tag{5.4}$$

It seems to be true (though we cannot prove) that in computing the infimum $\mathrm{eCov}(X,Y)=\inf_{(X',Y')}\mathbb{E}\big[\,\delta(X,X')+\delta(Y,Y')\big]$, under "general conditions" we can suppose $X=X'$ or $Y=Y'$. This conjecture is not true

without some restrictions, because if X and Y are 1–1 functions of each other then the conjecture would imply that $\mathrm{eCor}(X,Y)=1$. Indeed, Y is a function of $X=X'$ thus Y is independent of Y'. Hence $\mathrm{eCov}(X,Y)=\min\{\mathrm{eVar}(X),\mathrm{eVar}(Y)\}$. Thus in case of continuous marginals the empirical eCor would always be 1 because for continuous marginals no vertical or horizontal lines can contain more than one sample points with probability one. This is, however, not true as is shown by the following sample of four elements: $(1,4)$, $(2,2)$, $(3,3)$, $(4,1)$.

It can be proved that this conjecture implies that (5.4) holds with equality in the bivariate normal case [96].

It is easy to see that eCor as a new measure of dependence satisfies at least two of our axioms for dependence measures. In metric spaces axiom (iv) should read as

(iv*) $\Delta(X,Y)$ is continuous; that is, if for some positive constant K and $x_0\in\mathcal{M}, y_0\in\mathcal{M}$ we have $\mathbb{E}\big(\delta^2(X_n,x_0)+\delta^2(Y_n,y_0)\big)\le K$, $n=1,2,\dots$ and (X_n,Y_n) converges weakly (i.e., converges in distribution) to (X,Y) then $\Delta(X_n,Y_n)\to\Delta(X,Y)$.

In [96] axioms (i) and (iv*) are shown to hold. Concerning (ii) and (iii), only the following weaker versions are proved.

(ii*) $\mathrm{eCor}(X,Y)=\mathrm{eCor}(f(X),f(Y))$ for every similarity transformation f of $(\mathcal{M},\delta)$.

(iii*) $\mathrm{eCor}(X,Y)=1$ if $Y=f(X)$ with probability 1, where f is a similarity transformation of $(\mathcal{M},\delta)$.

In [96] a counterexample is also presented to show that axiom (iii) cannot be true for eCor in an arbitrary metric space $(\mathcal{M},\delta)$. However, it is conjectured that axiom (iii) is satisfied for Banach spaces valued random variables.

One can easily define the earth mover's correlation for more than two variables. The population version of eCov for three variables is as follows:

$$\mathrm{eCov}(X,Y,Z)=\inf_{(X',Y',Z')}\mathbb{E}d\big[(X,Y,Z),(X',Y',Z')\big].$$

Here in distribution $X=X'$, $Y=Y'$, $Z=Z'$, and X',Y',Z' are independent, and we take the inf over all joint distributions of (X,Y,Z) and (X',Y',Z').

The population version of the three-variate earth mover's correlation is

$$\mathrm{eCor}(X,Y,Z)=\frac{\mathrm{eCov}(X,Y,Z)}{\min\big\{\mathrm{eVar}(X),\mathrm{eVar}(Y),\mathrm{eVar}(Z)\big\}}.$$

Thus we have a natural measure for mutual dependence of more than two random variables.

Let us turn to the empirical version of eCov.

The earth mover's metric suggests the following earth mover's distance definition between two sequences $x := (x_1, x_2, \ldots, x_n)$ and $y := (y_1, y_2, \ldots, y_n)$:

$$\mathcal{E}(x, y) := \inf_{\pi} \sum_{i=1}^{n} \delta(x_i, y_{\pi(i)}),$$

where the infimum is taken for all permutation π on the integers $1, \ldots, n$. One can easily see that for real valued data, if the ordered sample is denoted by subscripts in brackets, then

$$\mathcal{E}(x, y) := \sum_{i=1}^{n} |x_{(i)} - y_{(i)}|.$$

The empirical version of eCov is the minimum transportation cost between the following two mass distributions or probability distributions:

(Q_1) $1/n$ mass at each point (x_i, y_i), $i = 1, 2, \ldots, n$

and

(Q_2) $1/n^2$ mass at each point (x_i, y_j), $i, j = 1, 2, \ldots, n$.

It is easy to see that the empirical eVar is the arithmetic average of the distances $\delta(x_i, x_j)$ because the cost to transport $1/n^2$ mass from the point (x_i, x_j) to the main diagonal (x, x) is at least $\delta(x_i, x_j)/n^2$ and we can achieve this via "horizontal" transportation only. This is not the case if we want to transport to n general points, not necessarily on the main diagonal. The "naive" computational complexity of the empirical eVar which is essentially Gini's mean difference is $O(n^2)$ but for real valued random variables we can decrease it to $O(n \log n)$.

The complexity of the computation of the empirical eCov is less obvious.

Our transportation problem can be reduced to an assignment problem between two sets of n^2 points, thus according to the "Hungarian algorithm" [81] this optimization can be solved in polynomial time. It was shown by [41] and [128] that the algorithmic complexity of assignment problem for two sets of n points is $O(n^3)$ thus in our case the complexity can be reduced to $O(n^6)$.

This is not very encouraging. A better complexity, $O(n^3 \log^2 n)$, is in [70]. Here the authors show that for the (linear) transportation problem with m supply nodes, n demand nodes and k feasible arcs there is an algorithm which runs in time proportional to $m \log m(k + n \log n)$ assuming without loss of generality that $m \geq n$, still at least one order of magnitude worse than the algorithmic complexity, $O(n^2)$, of computing the distance covariance or the distance correlation. This is the price we need to pay for the generality of eCov and eCor. The AMPL (A Mathematical Programming Language) code is easy to apply for computing empirical eCov and then eCor. In [2] it was shown that given n random blue and n random red

points on the unit square, the transportation cost between them is typically $\sqrt{n \log n}$. Our problem is to find the optimal transportation costs when the distance is the Manhattan distance and the number of red points is different from the number of blue points (the total mass is the same). A recent paper of Agarwal et al. [1] suggests that our task of computing the earth mover's distance between two sets of size n^2 can be done with the first algorithm in the cited paper with $O(\log^2(1/\varepsilon))$ approximation error bound in $O(n^{2+\varepsilon})$ steps, for any $\varepsilon > 0$. On related algorithmic optimizations see [4,5].

6. Representations by uncorrelated random variables

Let $X = (X_1, X_2, \dots, X_n)$ be an arbitrary random vector where the coordinates X_i, $i = 1, 2, \dots, n$ have finite variances. Then we can diagonalize the covariance matrix of X and thus we can find a linear transformation A such that $Y = AX$ becomes a random vector with uncorrelated coordinates. If the inverse A^{-1} exists then $X = A^{-1}Y$ is a representation of X with the help of uncorrelated random variables. But here Y is a mixture of the X coordinates and in many cases we cannot interpret these mixtures, e.g., if X_1 is the squared velocity and X_2 is the mass. Instead, let us consider representations that are univariate functions of the coordinates, not their mixtures. The idea that estimators of unrelated parameters should be unrelated (in some sense) is an old problem. The most natural notion of "unrelatedness" is independence. A classical theorem is that the maximum likelihood estimators of the mean and variance of a Gaussian distribution are independent. See [64] for many related results. Another classical approach is the parameter orthogonality, see [33]. In what follows unrelatedness means uncorrelatedness.

THEOREM 6.1 [94]. *Every random vector $X = (X_1, \dots, X_n)$ can be represented as functions of uncorrelated random variables $Y_1, \dots, Y_n$, i.e., we can always find $\mathbb{R} \to \mathbb{R}$ functions $f_1, \dots, f_n$ such that $(X_1, \dots, X_n)$ has the same distribution as $(f_1(Y_1), , \dots, f_n(Y_n))$.*

The functions f_i, $i = 1, 2, \dots, n$ cannot always be one-to-one because [93] can be reformulated as follows.

THEOREM 6.2 [93]. *A necessary and sufficient condition for random variables X_1, X_2 not to have the same distribution as $f_1(Y_1)$, $f_2(Y_2)$ where Y_1 and Y_2 are uncorrelated random variables and f_1, f_2 are one-to-one functions is that X_i, $i = 1, 2$ have the representation (equality in distribution)*

$$X_i = Z_i + c_i V_i \, \mathbb{1}(Z_i = b_i), \quad i = 1, 2,$$

where $\mathbb{1}(\cdot)$ denotes the indicator of the event in brackets, V_1 and V_2 are dependent (correlated) indicator functions, Z_1, Z_2, (V_1, V_2) are independent, b_i and c_i are real numbers, $c_i \neq 0$, and $\mathbb{P}(Z_i = b_i) > 0$, $i = 1, 2$.

On the other hand, the following proposition shows that with very few exceptions for *all random variables* X one can find a 1–1 real function f such that X and $f(X)$ are uncorrelated.

THEOREM 6.3 [95]. *Let X be a square integrable random variable defined on an arbitrary probability space. Suppose the distribution of X is not concentrated on three or less points. Then there exists a measurable injective function $f : \mathbb{R} \to \mathbb{R}$ such that X and $f(X)$ are uncorrelated. This f can be chosen piecewise linear.*

Such an f cannot exist if X takes on exactly two values, because in this case uncorrelatedness is equivalent to independence. When the distribution of X is supported on exactly 3 points then a necessary and sufficient condition for f to exist is $\mathbb{P}(X = \mathbb{E}X) = 0$.

This is another reason for not assuming 1–1 invariance of Δ in axiom (ii). The 1–1 invariance would imply the existence of many *uncorrelated* random variables X, Y for which $\Delta(X, Y) = 1$, which is counterintuitive.

THEOREM 6.4 [94]. *Let $X_1, \dots, X_n$ be arbitrary random variables with zero means, finite variances, and absolutely continuous distributions. Then there exist Borel sets $B_i \subset \mathbb{R}^+$, $i = 1, \dots, n$, such that if we define $f_i(t) = -t$ for $|t| \in B_i$, and $f_i(t) = t$ otherwise, then the random variables $Y_i := f_i(X_i)$ are uncorrelated, $\mathbb{E}(Y_i) = 0$ and $\mathrm{Var}(Y_i) = \mathrm{Var}(X_i)$. Since the functions f_i are idempotent ($f_i(f_i(x)) = x$), we have that $X_i = f_i(Y_i)$ is a one-to-one piecewise linear function of uncorrelated random variables.*

The next result on bivariate Gaussian vectors is already folklore.

PROPOSITION 6.1. *Let (X_1, X_2) be and arbitrary bivariate normal random variable with standard marginals. Then the one-to-one and piecewise linear function $f(x) = x$ for $|x| \geq c$ and $f(x) = -x$ for $|x| < c$ with a suitable constant $c = 1.539\dots$ makes X_1 and $f(X_2)$ uncorrelated, and $(X_1, X_2) \equiv (X_1, f(f(X_2)))$. Here the function does not depend on the correlation of X_1 and X_2.*

Can this result be generalized to n-variate Gaussian random variables? For this generalization one would need a partition of the set of positive integers into n disjoint subset N_i such that if H_k denotes the k-th Hermite polynomial, that is, $H_0(x) \equiv 1$, and $H_k(x)n(x) = (-1)^k(\frac{d}{dx})^k n(x)$, $k \geq 1$, where $n(x) = (2\pi)^{-1/2}e^{-x^2/2}$ denotes the standard normal p.d.f., then

$$f_i(x) := \sum_{k \in N_i} a_{ki} H_k(x)$$

is a one-to-one function for $i = 1, 2, \dots, n$. The explanation is given in the following theorem.

THEOREM 6.5 [94]. *Let $(X_1, \dots, X_n)$ be an arbitrary n-variate Gaussian random variable with standard marginals. Then $f_i(X_i)$, $i = 1, \dots, n$, are uncorrelated regardless of the correlation of the X variables if and only if the set of positive integers can be partitioned into n disjoint subsets N_i, $i = 1, \dots, n$, such that the Hermite expansion of $f_i(x)$ only contains terms $H_k(x)$ with indices from N_i, that is,*

$$f_i(x) = \sum_{k \in N_i} a_{ki} H_k(x).$$

In case $n = 2$, the partition $N_1 = \{1\}$, $N_2 = \{2, 3, \dots\}$ will do (as we have seen above), but at the moment we do not know if there exist n one-to-one functions f_i with the property above for $n > 2$. This seems to be an interesting open problem.

7. Tests for multivariate independence

Testing for independence between components (coordinates) of random vectors is one of the most classical problems in multivariate statistics. It has almost innumerable applications in economics, finance, life sciences, geology and other fields. The importance of the topic has grown especially since copula methods have gained popularity. This is because, in fact, any study based on copulas should be preceded by a test of independence with respect to all components of the observed vectors because completely different methods have to be used to investigate vectors with independent coordinates. Another potentially important statistical application is the time series analysis. For example, in the dynamic factor analysis (see [21]) it is essential that the estimated factors show independent behavior.

In this section our random vectors are p-dimensional with a continuous distribution function. Assume that we have N iid observations:

$$\underline{\xi}_i = (\xi_{i,1}, \dots, \xi_{i,p}), \quad i = 1, 2, \dots, N,$$

and let

$$G(\underline{x}) = \mathbb{P}(\xi_{i,1} \le x_1, \dots, \xi_{i,p} \le x_p), \quad F_j(x) = \mathbb{P}(\xi_{i,j} \le x), j = 1, \dots, p,$$

for all i. Our hypothesis of independence

$$H_0 : G(\underline{x}) = F_1(x_1) \cdot \ldots \cdot F_p(x_p)$$

is to be tested against the alternative

$$H_1 : G(\underline{x}) \neq F_1(x_1) \cdot \ldots \cdot F_p(x_p)$$

for at least one $\underline{x}$.

A natural approach is to find a measure for mutual dependence with "good" properties for random vectors and use this measure of dependence to construct our test statistic. The so called eCor introduced in the previous section is a suitable candidate for this measure.

The basic question of this approach lies in the properties that are required for the measure. We suggest the application of the new axioms introduced in Section 3 that provide an appropriate basis for tackling these questions. The underlying axioms of course should be reformulated for more than two variables. Another way to study the properties of the measure of dependence is based on copulas. Schmid et al. refer to these measures as association measures and list several possible good properties [119].

Here we only selected two mutual dependence measures that are frequently used to construct independence tests. The first one is the so-called dCor mentioned earlier, for details see [8]. The other measure is the so-called total distance multivariance described by [25] which is also included in the Multivariance R package for an independence test. An interesting attempt could be to generalize the kernel-based independence measure developed in [54] to several variables.

The above methods were well preceded in time by the approach that measures the difference between the empirical distribution function G_N of G and the product of the marginal empirical distribution functions $F_{N,j}$:

$$GG_N(\underline{x}) := \sqrt{N}\Big(G_N(\underline{x}) - \prod_{j=1}^{p} F_{N,j}(x_j) \Big).$$

Blum et al. [20] used the Cramér–von Mises statistic derived from the process GG_N

$$\int GG_N^2 \, dG_N,$$

and computed the limiting distribution of this statistic for $p = 2$. They pointed out the complexity of the GG process and determined the following useful decomposition:

$$GG_N(\underline{x}) = \sum_{A \subset \{1,\dots,p\}, |A|>1} G_{A,N}(\underline{x}) \prod_{j \in \{1,\dots,p\} \setminus A} F_{N,j}(x_j),$$

where

$$G_{A,N}(\underline{x}) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \prod_{i \in A} \big(\mathbb{1}(\xi_{i,j} \le x_j) - F_{N,j}(x_j) \big).$$

The G_A statistics or the combinations thereof could be used for testing mutual independence. Note, however, that their limiting distributions depend

on the marginals. Tests based on empirical characteristic functions, maximum variance statistics, and the exact boundary distribution of the test statistic can be found in [35].

The basic idea of tests most commonly used recently is that the independence hypothesis should be replaced by its copula counterpart

$$G(\underline{x}) = F_1(x_1) \cdots F_p(x_p) \Longleftrightarrow C(u_1, \ldots, u_p) = \prod_{i=1}^{p} u_i,$$

where C is the copula of G:

$$G(\underline{x}) = C(F_1(x_1), \ldots, F_p(x_p).$$

Deheuvels [36] suggested to replace the original observations $\underline{\xi}_1, \ldots, \underline{\xi}_N$ by their associated rank vectors

$$R_{i,j} = \sum_{l=1}^{N} \mathbb{1}(\xi_{l,j} \leq \xi_{i,j})$$

and the empirical distribution function G_N by the empirical copula

$$C_N(\underline{u}) = \frac{1}{N} \sum_{i=1}^{N} \prod_{j=1}^{p} \mathbb{1}(R_{i,j} \leq N u_j).$$

Deheuvels computed the limiting distribution of the copula process

$$CC_N(\underline{u}) = \sqrt{N}\left(C_N(\underline{u}) - \prod_{j=1}^{p} u_j \right).$$

Genest et al. [49] investigated the behavior of the CC_N-based Cramér–von Mises statistic and the rank analogues of $G_{A,N}$. This paper computed power comparisons for different tests of mutual independence.

In the following, we propose a new test whose power turned out to be very good in our simulation studies. Introduce the rank vectors

$$\underline{R}_i = (R_{i,1}, \ldots, R_{i,p}), \quad i = 1, \ldots, N.$$

In our test we will use the vectors

$$\underline{R}_1/(N+1), \quad \ldots, \quad \underline{R}_N/(N+1).$$

These vectors, we will simply refer to them as observations, are all located in a p-dimensional unit cube. We will test how the location of observations obtained from vectors with independent coordinates differs from the

location of observations obtained from vectors with non-independent coordinates. Our basic idea is very simple: a p-dimensional vector (with values in $(0,1)^p$) divides the p-dimensional unit cube into 2^p parts and we can detect how many observations fall into each part.

Our procedure is as follows.

(1) Simulate p-dimensional random vectors with iid $U(0,1)$ marginals K times. $\Rightarrow$ We have K different partitions of the unit cube.

(2) For each part determine the probability that the random vector with independent uniform marginals on $\frac{1}{N+1}, \dots, \frac{N}{N+1}$ is included in that particular part.

(3) Save the partitions and probabilities.

(4) Calculate the Kullback–Leibler divergence

$$\mathrm{D}(s\|q) = \sum_{\ell=1}^{2^p} s_\ell \log\Big(\frac{s_\ell}{q_\ell}\Big)$$

of the observed (s) and expected (q) frequencies for each of the K partitions.

(5) Multiply the sample via permuting the coordinates and then compute how many of these new vectors fall into the partitions of the unit cube.

(6) Calculate the mean of the divergences. This is our test statistic.

(7) The critical values of the test are computed by generating M samples (of size N) with independent marginals.

In simulation studies the power of the tests is usually computed by using well-known copulas. In part, we also followed this procedure, using Gauss and Clayton

$$C_\theta(\underline{u}) = \Big(\sum_{j=1}^{p} u_j^{-\theta} - p + 1\Big)^{-1/\theta}$$

as well as Gumbel

$$C_\theta(\underline{u}) = \exp\Big(-\big[(-\log u_1)^\theta + \dots + (-\log u_p)^\theta\big]^{1/\theta}\Big)$$

copulas.

However, we consider it equally important to use different conditional distributions for which the dependence is eventually even more apparent. One of our suggestions is the conditional exponential model. We generate $\xi_1, \dots, \xi_p$ independent exponentially distributed variables with parameter 1, but under the condition that $\xi_1 + \dots + \xi_p < v$. In this case, it is clear that the smaller the value of v, the stronger the dependence. Not all features of this model are easy to determine; however, the pairwise correlation can be

accurately described. Let $S_k = \xi_1 + \cdots + \xi_k$, then

$$\operatorname{corr}(\xi_1, \xi_2 \mid S_p < v) = 1 - \frac{\mathbb{P}(S_{p+2} < v)\,\mathbb{P}(S_p < v)}{2\mathbb{P}(S_{p+2} < v)\,\mathbb{P}(S_p < v) - \mathbb{P}(S_{p+1} < v)^2}\,.$$

The simulation study was performed by the R 3.6.2 statistical software package [106]. Our test was compared with the tests in the Copula R package [77]. These can be done using the so-called indepTest function. We utilized the fisher.pvalue, a p-value resulting from a combination à la Fisher of the subset statistic p-values, the tippett.pvalue, a p-value resulting from a combination à la Tippett of the subset statistic p-values, and the global.statistic, value of the global Cramér–von Mises statistic derived directly from the independence empirical copula process (see [49]).

The dimension p of our random vectors was 3 and 5, respectively, and the sample size was $N = 100$. Critical values for both the new test and the indepTest function were computed from 1000 experiments. The power of the tests at the 5% level was also estimated based on 1000 trials. In our new test, we divided the unit cube $K = 2000$ times into $2^3 = 8$ and $2^5 = 32$ parts, respectively.

For dependent exponentials, the cuts were $v = 3, 4, \ldots, 10$. For the Gumbel, Clayton, and Gauss copulas, we considered cases where the Kendall's τ coefficient was 0.05, 0.1, 0.15, and 0.2, respectively.

Table 1 summarizes the power of the tests multiplied by 1000. In the table we can see that the proposed new test produces better results in essentially all cases. This is particularly evident in moderately dependent cases, such as the conditional exponential model for $v = 3$ and 4 and the copula for $\tau = 0.15$.

We can experience a very similar situation in dimension $p = 5$. Table 2 shows that the new test always performs better than the previously developed tests especially in the case $\tau = 0.1$.

In the future, we plan to compare our test with further tests and determine the limiting distribution of our statistics.

8. Qualitative independence in finite sets

Let Ω be a finite set of n elements. The following notion was introduced by Marczewski [88]. The subsets $A, B \subset \Omega$ are called *qualitatively independent* (QI) if they divide Ω into four non-empty parts that is $A \cap B$, $\bar{A} \cap B$, $A \cap \bar{B}$, $\bar{A} \cap \bar{B}$ are all non-empty. Of course, in the language of probability theory A and B are events in the probability space Ω. The significance of this notion in *search theory* lies in the consequence that after knowing if an unknown x is in A or not, in both cases, x may be in B or in $\bar{B}$. In other words, all the 4 pairs of answers may happen. In terms of probability theory, neither the occurrence nor the non-occurrence of the event A implies

the occurrence or the non-occurrence of the event B. The family $\mathcal{F} \subset 2^\Omega$ is called *qualitatively independent* if their members are pairwise qualitatively independent. Rényi asked in his seminar (later in his book [115]) the question what is the maximum size of a qualitatively independent family in an n-element set. The answer was found by one of the present authors in [68]. Later it turned out that several authors found the same result, around the same time, with different motivations and formulations: Brace and Daykin [26], Bollobás [22] and Kleitman and Spencer [71] (see also the closely related result of [122]).

THEOREM 8.1 [22,26,68,71]. *The maximum number of pairwise qualitatively independent sets in an n-element set is*

$$\binom{n-1}{\lceil \frac{n}{2} \rceil}.$$

A simple proof, using the cycle method was found in [9].

In fact, Kleitman and Spencer asked and asymptotically solved a more general problem. They introduced the notion of k-qualitative independence. A family $\mathcal{F} \subset 2^\Omega$ is called *k-qualitatively independent* iff

$$\bigcap_{i=1}^{k} A_i^{\varepsilon_i} \neq \emptyset$$

holds for any choice of distinct $A_1, \ldots, A_k \in \mathcal{F}$ and $\varepsilon_i = 0, 1$, where $A^0 = A$, $A^1 = \bar{A}$, that is, when any k members divide Ω into 2^k non-empty parts. The probabilistic interpretation of this condition is that knowing for any $k-1$ of the events (members of the family) which one occurred and which one did not, in all 2^{k-1} cases the kth event may or may not occur.

Let $f(n,k)$ denote the maximum size of a k-qualitatively independent family. The following theorem was proved.

THEOREM 8.2 (Kleitman and Spencer [71]).

$$2^{c_1 2^{-k} k^{-1} n} \leq f(n,k) \leq 2^{c_2 2^{-k} n}.$$

Noga Alon [3] gave an explicit construction.

However, in his book [115], Rényi actually asked a more general question. A set $A \subset \Omega$ defines a partition $(A, \bar{A})$. This is generalized in the following way. Consider partitions $(U_1, U_2, \ldots, U_r)$ of Ω into r parts, in other words, *r-partitions*. Two r-partitions $(U_1, U_2, \ldots, U_r)$ and $(V_1, V_2, \ldots, V_r)$ are called *qualitative independent*, if all the r^2 intersections $U_i \cap V_j$ of the classes are non-empty. Of course, such a partition can be identified with a random variable. Two such partitions are qualitatively independent iff the corresponding

random variables ξ and η possess the following property: whatever is the value of ξ, all the r values are possible for η. The maximum number of pairwise qualitatively independent r-partitions is denoted by $g(n,r)$.

It should be mentioned that Poljak and Rödl [102] independently introduced the same concept under the name orthogonal partition and rediscovered Theorem 8.1. The paper [101] shows the connection of this problem to several other problems in combinatorics and graph theory.

An important development was the following theorem of Poljak and Tuza [103].

THEOREM 8.3. *We have*

$$g(n,r) \le \frac{1}{2}\binom{\lfloor \frac{2n}{r} \rfloor}{\lfloor \frac{n}{r} \rfloor}.$$

Observe that this upper bound coincides with the exact value in Theorem 8.1 if n is even. Improvements of the lower bound were obtained in the same paper and by Körner and Simonyi [80], and Gargano, Körner and Vaccaro [47]. Since the value of $g(n,r)$ is exponential in n and an exact formula for it in the case $r \ge 3$ is hopeless, it is sufficient to consider the exponent:

$$q_r = \limsup_{n\to\infty} \frac{1}{n} \log g(n,r).$$

Theorem 8.3 gives the upper bound $q_r \le \frac{2}{r}$. The lower estimate of [103] was weaker. With strong techniques borrowed from information theory Gargano, Körner and Vaccaro succeeded to prove that the upper estimate is sharp.

THEOREM 8.4 (Gargano, Körner and Vaccaro [48]). $q_r = \frac{2}{r}$ $(2 \le r)$.

The difficulty of the problem is illustrated by the fact that the analogous problem for the case when every three r-partitions are qualitatively independent is still unsolved. But there are other possible generalizations.

Körner and Monti [79] suggested a weakening of the problem. Three qualitatively independent sets (equivalently, 2-partitions) divide Ω into 8 non-empty sets. If it is only supposed that at least 6 out of 8 parts are non-empty, we say that the family is $\frac{6}{8}$ or $\frac{3}{4}$-qualitatively independent. [79] gives good estimates on the size of the largest such family.

There is another weakening. Suppose that there is a qualitatively independent pair A, B among any m members of the family $\mathcal{F} \subset 2^\Omega$. Then the family is called an *m-weak qualitatively independent* family. As an example look at the family of all $\frac{n}{2}$-element subsets. In this family there are complementing pairs of members, that is, it is not qualitative independent. Choosing 3 members, two of them must be qualitatively independent, that is, this family is 3-weak qualitatively independent. Balázs in [9] proved that the size of an m-weak qualitatively independent family cannot exceed the

sum of the $\frac{m-1}{2}$ largest binomial coefficients of order n if m is odd and this estimate is sharp when $\frac{3m-1}{2} \leq n$. The case m is even is not completely settled, only good estimates are given in [9].

The following one is a strengthening of the condition. A family $\mathcal{S} \subset 2^\Omega$ is *s-strongly separating* iff all four intersections $A \cap B$, $A \cap \bar{B}$, $\bar{A} \cap B$, $\bar{A} \cap \bar{B}$ are of size at least s for any two distinct members $A, B \in \mathcal{S}$. In terms of probability theory we suppose not only that all four combinations of the two events occur, but they occur with probability at least $\frac{s}{n}$. The maximum size of an s-strongly separating family is denoted by $h(n, s)$. Its determination was proposed in [69]. It has been asymptotically answered by Frankl for fixed s.

THEOREM 8.5 (Frankl [45]).

$$d_1(s)\frac{2^n}{n^{s-\frac{1}{2}}} \leq h(n,s) \leq d_2(s)\frac{2^n}{n^{s-\frac{1}{2}}}$$

where

$$d_1(s) = \sqrt{\frac{2}{\pi}}\,\frac{1}{2^s} - \varepsilon \quad \textit{and} \quad d_2(s) = \sqrt{\frac{2}{\pi}}\,2^{s-2}(s-1)! + \varepsilon.$$

The situation is very different when s is about pn where p is a fixed positive ($\leq \frac{1}{4}$) number. Let us start with a very special case. Suppose that n is divisible by 4 and $s = \frac{n}{4}$. Let $\mathcal{S}$ be an $\frac{n}{4}$-strongly separating family. Then $A, B \in \mathcal{S}$ $(A \neq B)$ divide Ω into four equal parts of size $\frac{n}{4}$ each. Associate a vector with coordinates $1, -1$ with a member A of $\mathcal{S}$ writing 1 in the ith position iff the ith element of Ω is in A. Denote the vectors obtained in this way from the members of $\mathcal{S}$ by $v_1, v_2, \ldots, v_m$. It is easy to see that the inner product $v_i v_j$ is 0 for $1 \leq i < j \leq m$. Let v_0 have 1's in each coordinate. Then $v_0 v_i = 0$ also holds $(1 \leq i \leq m)$. That is, $v_0, v_1, \ldots v_m$ are pairwise orthogonal vectors in an n-dimensional space. We obtained the following statement.

PROPOSITION 8.1. *Let n be divisible by 4. Then*

$$h\Big(n, \frac{n}{4}\Big) \leq n - 1$$

with equality iff there is a Hadamard matrix of order n.

For general p we have the following upper bound.

THEOREM 8.6 [69]. *Suppose that $0 < p < 0.099$, then*

$$h(n, pn) \leq 2^{n\left(-\frac{1}{2}\log p - 1.099\right) + o(n)}.$$

Unfortunately, the method applied in [69] does not work for the cases $0.099 < p < \frac{1}{4}$. No reasonable lower bound is known.

Combining all the problems above we arrive to the following general question. Let $M(n, r, k, s)$ be the maximum number of r-partitions of an n-element set under the condition that any k of these partitions divide the underlying set into r^k parts, each of them having size at least s. Give good estimates on $M(n, r, k, s)$.

A closely related problem was formulated in [69]. The "distance" between two subsets (2-partitions) is given now by the entropy. For a pair A, B of subsets define the probabilities

$$p_1 = \frac{|A \cap B|}{n}, \quad p_2 = \frac{|\bar{A} \cap B|}{n}, \quad p_3 = \frac{|A \cap \bar{B}|}{n}, \quad p_4 = \frac{|\bar{A} \cap \bar{B}|}{n}$$

and then the entropy of the pair:

$$H(A, B) = H(p_1, p_2, p_3, p_4) = \sum_{i=1}^{4} (-p_i) \log p_i$$

where $p \log p$ is defined to be 0 for $p = 0$. Find the maximum size of a family $\mathcal{F}$ of subsets satisfying $q \leq H(A, B)$ for every pair of distinct members of $\mathcal{F}$.

9. Qualitative conditional independence of finite partitions

Under Qualitative Independence (QI) of two algebras (families of subsets of a set, containing the empty set and closed under complements and finite unions and intersections), an extension of measures given on them to the generated algebra is always possible. The generated measure is of a product structure. In this section, existence of product extensions of measures given on algebras or only on some sets are investigated. The underlying space is assumed to be finite and nonempty sets are assumed to have positive measures. The developments reported here are partly motivated by statistical applications, some of which are also presented.

Let Ω be a set and let $\mathcal{A}$ and $\mathcal{B}$ be algebras of subsests of Ω, both generated by finite partitions. Then $\mathcal{C} = \mathcal{A} \cap \mathcal{B}$ is also an algebra of sets, and it is also generated by a finite partition of Ω.

In [16] the following definition was given: $\mathcal{A}$ and $\mathcal{B}$ are qualitatively conditionally independent (QCI), if for every $\emptyset \neq A \in \mathcal{A}$ and $\emptyset \neq B \in \mathcal{B}$, such that $A \cap B = \emptyset$, there exists a $C \in \mathcal{C}$ such that

$$A \subseteq C \quad \text{and} \quad B \subseteq C^c. \tag{9.1}$$

If $\mathcal{A}$ and $\mathcal{B}$ are such that $\mathcal{C} = \{\emptyset, \Omega\}$, then QCI of $\mathcal{A}$ and $\mathcal{B}$ implies QI.

Let P be a measure on $\mathcal{A}$, and Q be a measure on $\mathcal{B}$. P and Q are called weakly compatible, if $P(C) = Q(C)$ for all $C \in \mathcal{C}$. Because of the finite algebras considered, measures are assumed to be finitely additive only. Let $(\mathcal{A} \cup \mathcal{B})$ denote the smallest algebra containig $\mathcal{A}$ and $\mathcal{B}$. As the intersection of algebras is an algbra, $(\mathcal{A} \cup \mathcal{B})$ exists. A measure R on $(\mathcal{A} \cup \mathcal{B})$ is called an extension of P and Q, if $R(A) = P(A)$ for all $A \in \mathcal{A}$ and $R(B) = Q(B)$ for all $B \in \mathcal{B}$. If such an extension exists, P and Q are called strongly compatible.

Then one has the the following result.

THEOREM 9.1 [16]. *The following statements are equivalent*:

(i) *The algebras $\mathcal{A}$ and $\mathcal{B}$ are QCI.*

(ii) *If two measures defined on $\mathcal{A}$ and $\mathcal{B}$ are weakly compatible, then they are also strongly compatible.*

Two algebras need to be QCI, indeed, for the existence of an extension for all pairs of (weakly compatible) measures. For instance, if $\Omega = \{1, 2, 3\}$, and $\mathcal{A}$ is generated by the partition $\{\{1\}, \{2, 3\}\}$, and $\mathcal{B}$ is generated by the partition $\{\{1, 2\}, \{3\}\}$, then a P with $P(1) = 0.6$ and a Q with $Q(3) = 0.6$, with $P(1, 2, 3) = Q(1, 2, 3) = 1$, cannot have a common extension. Such an example may always be constructed, whenever QCI does not hold.

The generated algebra $(\mathcal{A} \cup \mathcal{B})$ has atoms of the form $A \cap B$, where A and B are atoms of the respective algebras. For every atom A, there is an atom C_A of $\mathcal{C}$, which contains A. Indeed, otherwise $\emptyset \neq A \cap C \subsetneq A$ would hold for some atom C of $\mathcal{C}$, but then A could not be an atom of $\mathcal{A}$. Similarly, an atom C_B of $\mathcal{C}$ contains B. If $C_A \neq C_B$, then $A \cap B = \emptyset$ and set $R(A \cap B) = 0$. Otherwise, $C_A = C_B$ and

$$R(A \cap B) = P(A|C_A)Q(B|C_B)P(C_A) = P(A|C_A)Q(B|C_B)Q(C_B) \tag{9.2}$$

It was shown in [16] that R defines a finitely additive measure on $(\mathcal{A} \cup \mathcal{B})$ which extends both P and Q. When QCI is QI, under the extension obtained, $\mathcal{A}$ and $\mathcal{B}$ are independent.

The concept of QCI includes the traditional product case. To see this, let $\mathcal{A}'$, $\mathcal{B}'$, $\mathcal{C}'$ be finitely generated algebras of subsets of Ω. Define

$$\mathcal{A} = \mathcal{A}' \times \mathcal{C}' \times \Omega, \quad \mathcal{B} = \Omega \times \mathcal{C}' \times \mathcal{B}'.$$

The algebras $\mathcal{A}$ and $\mathcal{B}$ defined above are QCI, with

$$\mathcal{A} \cap \mathcal{B} = \Omega \times \mathcal{C}' \times \Omega.$$

The results above extend to several algebras on the same space. Let Ω be an arbitrary set and for $i = 1, \ldots, k$, let $\mathcal{A}_i$ be an algebra of subsets of Ω generated by a finite partition. Call the algebras decomposable, if there is an order $\mathcal{A}_{i_1}, \mathcal{A}_{i_2}, \ldots, \mathcal{A}_{i_k}$, such that for every $j \geq 2$,

$$(\mathcal{A}_{i_1} \cup \cdots \cup \mathcal{A}_{i_{j-1}}) \text{ and } \mathcal{A}_{i_j} \text{ are QCI}$$

and there is an $i_1 \le i_l \le i_{j-1}$, such that

$$(\mathcal{A}_{i_1} \cup \cdots \cup \mathcal{A}_{i_{j-1}}) \cap \mathcal{A}_{i_j} = \mathcal{A}_{i_l} \cap \mathcal{A}_{i_j}.$$

Then, one has the following result.

PROPOSITION 9.1. *Let the algebras $\mathcal{A}_i$, $i = 1, \ldots, k$, be such that none of them is a subalgebra of another one. Let the algebras be decomposable and let the measures P_i be given on $\mathcal{A}_i$ such that on the intersection of any two algebras the given measures coincide. Then there exists a common extension of all these measures to $(\mathcal{A}_1 \cup \cdots \cup \mathcal{A}_k)$.*

PROOF. The proof is straightforward induction on k. For $k = 2$, Theorem 9.1 implies the result. Theorem 9.1 also applies in the induction step. □

That QCI is needed for the existence of extension of all weakly compatible measures, was shown above. But pairwise QCI is not sufficient in the product case, which is implied by the known counterexample for a three-dimensional space, see, e.g., [18].

The measure R in (9.2) may also be written in a product form as

$$\begin{aligned} R(A \cap B) &= P(A|C_A)\sqrt{P(C_A)}Q(B|C_B)\sqrt{P(C_B)} \\ &= h_{\mathcal{A}}(A, C_A)h_{\mathcal{B}}(B, C_B), \end{aligned} \tag{9.3}$$

with functions $h_{\mathcal{A}}$, depending on A and $h_{\mathcal{B}}$, depending on B.

In the sequel, the existence of an extension will be assumed, and existence and properties of multiplicative extensions generalizing (9.3) are investigated.

Families of distributions which generalize (9.3) on product spaces play an important role in statistical analysis. The related statistical models are usually defined on a product space generated by ranges of discrete (finitely generated) variables and there is a graph given, where the nodes are the variables. Each of the functions that enter the representation, like $h_{\mathcal{A}}$ and $h_{\mathcal{B}}$ in (9.3), depends on variables which are cliques, that is, maximal complete subgraphs of the graph. A distribution with this structure is called Gibbsian with respect to the graph. A distribution is called Markovian with respect to a graph, if it has conditional independence (Markov) properties which can be read off from the graph. These Markov properties generalize the conditional independence in (9.2). A celebrated result, called the Hammersley–Clifford theorem, is that a distribution is Markovian, if and only if it is Gibbsian, see, e.g., [82] or [118].

Similar structures may also be defined when the underlying space is not a Cartesian product, and instead of variables, one has finitely generated algebras of subsets. A further generalization to be considered is when, instead

of measures on algebras, the probabilities of selected subsets are specified, to be extended to a measure on the generated algebra. First, two examples are presented.

EXAMPLE 9.1. Let the elements of Ω be denoted by numbers and let $\Omega = \{1, 2, 3, 4\}$. Set

$$\mathcal{A} = \{\emptyset, \{1, 2\}, \{3, 4\}, \Omega\}, \quad \mathcal{B} = \{\emptyset, \{1, 3\}, \{2, 4\}, \Omega\}.$$

Then, the two algebras are QI and the atoms of the generated algebra are the elements of Ω. A probability distribution on Ω may be defined as

$$\log P(i) = f_{\mathcal{A}}(i) + g_{\mathcal{B}}(i), \tag{9.4}$$

where $f_{\mathcal{A}}(i)$ depends on the atom A of $\mathcal{A}$, which contains i, in the sense that if $i, j \in A$, then $f_A(i) = f_A(j)$ and similarly for $g_B(i)$. The values of the functions f_A and g_B for the respective atoms may be seen as parameters in the representation (9.4). The distribution P may be written as

$$\begin{pmatrix} \log P(1) \\ \log P(2) \\ \log P(3) \\ \log P(4) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} f_{\{1,2\}} \\ f_{\{3,4\}} \\ g_{\{1,3\}} \\ g_{\{2,4\}} \end{pmatrix} \tag{9.5}$$

An overparameterized version is obtained by including a parameter o present everywhere, called the overall effect:

$$\begin{pmatrix} \log P(1) \\ \log P(2) \\ \log P(3) \\ \log P(4) \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} o \\ f_{\{1,2\}} \\ f_{\{3,4\}} \\ g_{\{1,3\}} \\ g_{\{2,4\}} \end{pmatrix}$$

This overparameterized version makes it possible to balance the parameters in the sense of assuming that

$$f_{\{1,2\}} + f_{\{3,4\}} = 0 \quad \text{and} \quad g_{\{1,3\}} + g_{\{2,4\}} = 0.$$

The so called corner parameterization of the same family of distributions is

$$\begin{pmatrix} \log P(1) \\ \log P(2) \\ \log P(3) \\ \log P(4) \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} o \\ e_{\{1,2\}} \\ e_{\{1,3\}} \end{pmatrix} \tag{9.6}$$

Let now a probability measure S be given on Ω. Then, in all of the parametrizations, the parameters may be selected in such a way that

$$\begin{cases} P(1,2) = S(1,2), \ P(3,4) = S(3,4), \\ P(1,3) = S(1,3), \ P(2,4) = S(2,4). \end{cases} \tag{9.7}$$

Thus, P is the independent extension of the restrictions implied by S to $\mathcal{A}$ and to $\mathcal{B}$, which always exists because the two algebras are QI. Such an independent distribution is characterized by the following odds ratio (see, e.g., [118]) being equal to 1:

$$\frac{P(1)P(4)}{P(2)P(3)} = 1.$$

The next example illustrates the situation when QI does not hold.

EXAMPLE 9.2. In this example, $\Omega = \{1,2,3\}$. Set

$$\mathcal{A} = \{\emptyset, \{1,2\}, \{3\}, \Omega\}, \quad \mathcal{B} = \{\emptyset, \{1,3\}, \{2\}, \Omega\}.$$

In this case, the corner parametrization like the one in (9.6)

$$\begin{pmatrix} \log P(1) \\ \log P(2) \\ \log P(3) \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} o \\ e_{\{1,2\}} \\ e_{\{1,3\}} \end{pmatrix}$$

is not restrictive as it has 3 parameters for the 3 probabilities. If a measure S is given on Ω, similarly to (9.7),

$$P(1,2) = S(1,2), \quad P(1,3) = S(1,3), \tag{9.8}$$

may be achieved.

If one makes the model restrictive by omitting the overall effect

$$\begin{pmatrix} \log P(1) \\ \log P(2) \\ \log P(3) \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} e_{\{1,2\}} \\ e_{\{1,3\}} \end{pmatrix} \tag{9.9}$$

then (9.8) cannot be achieved. Instead, one has

$$P(1,2) = \lambda S(1,2), \quad P(1,3) = \lambda S(1,3),$$

for some positive λ.

In the case of distributions of the form (9.9), instead of the odds ratio being equal to 1, one has

$$\frac{P(1)}{P(2)P(3)} = 1.$$

This is a non-homogeneous generalization of the odds ratio, see [72].

While the space in Example 9.1 may be seen as generalizing the product structure because of QI, the space in Example 9.2 is not of a product nature and it can be seen as an incomplete Cartesian product, see Table 3. Such data structures occur often in statistical problems, for instance in the analysis of register data, see [76]. In a register, data about relevant events, like a baby being born with one or more congenital abnormalities or a driver being fined for one or several violations or an online purchase of one or several goods being made by a customer, are collected. Each record in a register describes an event and contains which features (abnormalities or violations or goods) are present. A fundamental characteristic of such registers is that every record has at least 1 feature present, otherwise, there would be no event to be entered into the registry, see [76]. Therefore, the summary structure where the frequencies of every feature combination may be recorded is an incomplete Cartesian product, and the pairs of algebras generated by the lack or presence of each feature are not QI. Example 9.2 is a minimal illustration of such a structure with 2 features. In Table 3, cell 1 counts the cases where both features are present, cell 2 counts the cases when only feature $\mathcal{A}$ is present and cell 3 counts the cases when only feature $\mathcal{B}$ is present.

Relevant statistical models, see [76], may associate effects with a feature present, but no effect is implied if a feature is not present. A straightforward justification for considering such structures is parsimony. Every record in a register is characterized by many possible features. For instance, an online store may have 10000 different items for sale, but perhaps 99.99% of the purchases contain not more than 10 goods. If effects are only associated with the features present, that is goods purchased, the probability of most purchases would be modeled with 10 parameters (and possible interactions). If effects are also associated with the features not present, one would have 10000 parameters in every cell (plus possible interactions), which would not lead to useful simplifications. In Example 9.2, the 2 parameters in (9.9) are the effects associated with each of the features present.

In a general formulation of this problem, let Ω be a finite set and let I be its elements arranged in a vector. Probability distributions on Ω of the following form are considered:

$$\log \mathbf{P}(I) = \mathbf{A}\boldsymbol{\beta}, \tag{9.10}$$

where $\mathbf{P}(I)$ is the vector of probabilities, $\mathbf{A}$ is a design matrix, $\boldsymbol{\beta}$ is a vector of parameters. The design matrix is a $0-1$ matrix, with at least one 1 in every row. The family of distributions in (9.10) generalizes the usual log-linear model, see [19] or [118], in three aspects. The space Ω is not assumed to be a Cartesian product, the parameters $\boldsymbol{\beta}$ are not assumed to be associated with groups of variables, and an overall effect is not assumed to be present. In particular, the model in (9.10) generalizes (9.3).

The distributions in (9.10) are said to have an overall effect, if the column space of $\mathbf{A}$ contains the vector of 1's. In this case, the model may be reparameterized to have a parameter present in every cell, like the parameter o in (9.6). Such a reparameterization of (9.9) is not possible.

Let now $\mathbf{a}_j$, $j = 1, \ldots, k$ denote the columns of the design matrix $\mathbf{A}$. Each of these vectors may be interpreted as the indicator vector of a subset of Ω. If $\mathcal{A}$ contains an overall effect, some of these subsets may be seen as indicators of the atoms of an algebra (that is indicators of a partition of Ω). This is seen in Example 9.1, where the first 2 columns of the design matrix in (9.5) are the indicators of the atoms of $\mathcal{A}$ and the last 2 columns are indicators of $\mathcal{B}$. Conversely, when some of the columns of $\mathbf{A}$ are indicators of a partition of Ω, then the $\mathbf{A}$ may be changed to contain an overall effect, without changing the family of distributions defined in (9.10).

Suppose the probabilities of the subsets specified by the columns of $\mathbf{A}$ are given as

$$Q(\mathbf{a}_j), \quad j = 1, \ldots, k. \tag{9.11}$$

The most general extension considered here, is extending the probabilities of subsets given in (9.11) to a probability measure of the product (linear on the logarithmic scale) form in (9.10), even if the subsets do not form a partition, thus one does not have measures given on algebras.

Obviously, the probabilities in (9.11) need to be strongly compatible for the existence of an extension with a product structure. A simple way to achieve this compatibility, is to assume that the probabilities in (9.11) are taken from a distribution on Ω. Also, the compatibility condition is equivalent to the existence of a positive solution of the following system of linear equations in $\mathbf{x}$:

$$\mathbf{A}^*\mathbf{x} = \mathbf{Q}^*,$$

where $\mathbf{A}^*$ is the transpose of $\mathbf{A}$ with a row of 1's added as the last row and $\mathbf{Q}^*$ is the column vector with the components given in (9.11) with a 1 added as the last component. Then, one has the following result.

THEOREM 9.2 [72]. *Let the probabilities of the subsets with indicators* $\mathbf{a}_j$, $j = 1, \ldots, k$ *given in* (9.11) *be strongly compatible.*

(i) *Then, a unique distribution* $\mathbf{P}$ *of the form* (9.10) *where the design matrix* $\mathbf{A}$ *consists of the columns* $\mathbf{a}_j$, $j = 1, \ldots, k$ *with*

$$P(\mathbf{a}_j) = Q(\mathbf{a}_j), \quad j = 1, \ldots, k \tag{9.12}$$

exists if and only if $\mathbf{A}$ *has the vector of* 1*'s in its column space.*

(ii) *If* $\mathbf{A}$ *does not have the vector of* 1*'s in its column space, then there exists a unique probability distribution of the form* (9.10) *with*

$$P(\mathbf{a}_j) = \lambda Q(\mathbf{a}_j), \quad j = 1, \ldots, k \tag{9.13}$$

for some positive λ.

The families of distributions P with the properties in (9.12) and (9.13) are linear, and the family in (9.10) is exponential. In particular, the family defined in (9.10) is a regular exponential family when the model contains an overall effect, and is a curved exponential family when this is not the case, see [75].

To determine the distribution $\mathbf{P}$ in Theorem 9.2, one, in general, needs to apply an iterative algorithm which was described in [73]. In statistical applications, Q in (9.11) is the observed probability of the event of having an observation in the subsets $\mathbf{a}_j$, and $\mathbf{P}$ is the maximum likelihood estimate in the model (9.10), see [72].

For results related to the closure of the family (9.10) to allow distributions with zero probabilities, see [74].

10. Qualitative conditional independence in non-atomic measure spaces

There are various notions in the literature generalizing the notion of qualitative indepence, or — as originally introduced by Marczewski [88] — of *independent fields*.

In the notion of *almost independence of σ-fields* the assumption on the two given σ-fields $\mathcal{A}$ and $\mathcal{B}$ equipped with measures P and Q, respectively, is that the intersections of two sets $A \in \mathcal{A}$, $B \in \mathcal{B}$ with positive measures should have non-empty intersection. The beautiful example given by [57] shows that this assumption is not enough for assuring the existence of a common extension of the two measures under which the two σ-fields become independent. In his example he considers a subset $T \subset [0,1]^2$ with planar measure less than 1. For example,

$$T = \big\{ (x,y) \mid 0 \le x, y \le 1,\ y - x \text{ is rational} \big\} .$$

Set

$$\mathcal{A} = \big\{ T \cap (A \times [0,1]) \mid A \subset [0,1] \text{ is a Borel-set} \big\} ,$$
$$\mathcal{B} = \big\{ T \cap ([0,1] \times B) \mid B \subset [0,1] \text{ is a Borel-set} \big\} ,$$

and

$$P\big(T \cap (A \times [0,1])\big) = \lambda(A) , \quad Q\big(T \cap ([0,1] \times B)\big) = \lambda(B) ,$$

where λ is the linear Lebesgue measure. Then, since if $\lambda(A) > 0$ and $\lambda(B) > 0$ then the set of points $y - x$, $x \in A$, $y \in B$ covers some interval, and so contains a rational point. Hence

$$\emptyset \ne T \cap (A \times B) = (T \cap (A \times [0,1])) \cap (T \cap ([0,1] \times B))$$

that is, the σ-fields $\mathcal{A}$ and $\mathcal{B}$ are almost independent. But, on the other hand, since the (Lebesgue) measure of T is less than 1, it can be covered by a disjoint union $\bigcup_i (A_i \times B_i)$ of measure less than 1. Now,

$$T \subset \bigcup_i (A_i \times B_i) = \bigcup_i (T \cap (A_i \times [0,1]) \cap ([0,1] \times B_i)) \,.$$

Assuming that there is a common extension R of the measures P and Q we have that

$$R(T) \le \sum_i \lambda(A_i)\lambda(B_i) < 1 \,.$$

However, $T = T \cap [0,1]^2$ thus $R(T) = (\lambda([0,1]))^2 = 1$. This contradiction shows that R cannot be a (σ-additive) measure.

In order to generalize Theorem 9.1 to non-atomic probability spaces we are going to use the notion of lifting introduced by von Neumann, analyzed in a series of papers by A. Ionescu Tulcea and C. Ionescu Tulcea [61,62] and also the notion of regular conditional probability considered by Doob. The connection between lifting and regular version of conditional probability is discussed e.g. in [60].

Let us consider a probability space $(\Omega, \mathcal{F}, Q)$. We shall use the notation $A \equiv B$, when $Q(A \circ B) = 0$, where $A \circ B$ denotes the symmetric difference of the sets A and B. This is an equivalence relation.

DEFINITION 10.1. For a sub-σ-field $\mathcal{G} \subset \mathcal{F}$ the map $\rho : \mathcal{G} \to \mathcal{G}$ is a *lifting*, if the following properties hold.

(i) $\rho(A) \equiv A$;
(ii) $A \equiv B$ implies that $\rho(A) = \rho(B)$;
(iii) $\rho(\emptyset) = \emptyset$, $\rho(\Omega) = \Omega$;
(iv) $\rho(A \cap B) = \rho(A) \cap \rho(B)$;
(v) $\rho(A \cup B) = \rho(A) \cup \rho(B)$.

It was proved by Maharam [86] that if $(\Omega, \mathcal{F}, Q)$ is a complete probability space, then the lifting exists for any sub-σ-field. A proof was given in [61] based on the martingale convergence theorem. The lifting can be extended to the Banach-algebra $L_\infty(\Omega, \mathcal{G}, Q)$ resulting in a mapping $T : L_\infty(\Omega, \mathcal{G}, Q) \to L_\infty(\Omega, \mathcal{G}, Q)$ with the properties:

(i) $T(f) \equiv f$;
(ii) $f \equiv g$ implies that $T(f) = T(g)$;
(iii) $T(1) = 1$;
(iv) $f \ge 0$ implies that $T(f) \ge 0$;
(v) $T(\alpha f + \beta g) = \alpha T(f) + \beta T(g)$;
(vi) $T(fg) = T(f)T(g)$,

where $f \equiv g$ means that $Q(f = g) = 1$. Note, that the connection between ρ and T is given by the formula $\mathbb{1}(\rho(A)) = T(\mathbb{1}(A))$ using the notation $\mathbb{1}$ for

the indicator function of a set. Property (iv) implies that the supremum norm of the function $T(f)$ equals the ess-sup norm of $f \in L_\infty(\Omega, \mathcal{G}, Q)$.

Next, let us recall the notion of the regular conditional probability. Let $\mathcal{C} \subset \mathcal{B}$ be two sub-σ-fields of $\mathcal{F}$. The function $Q(B, \omega)$ for $B \in \mathcal{B}$, $\omega \in \Omega$ is a *regular conditional probability* with respect to $\mathcal{C}$, if

(i) $Q(\,\cdot\,, \omega)$ is a probability measure on $\mathcal{B}$ for any $\omega \in \Omega$;

(ii) $Q(B, \,\cdot\,)$ is a $\mathcal{C}$-measurable random variable for any $B \in \mathcal{B}$;

(iii) $Q(B \cap C) = \int_C Q(B, \omega)\, dQ(\omega)$ for any $C \in \mathcal{C}$.

If there exists a lifting T in $L_\infty(\Omega, \mathcal{C}, Q)$ then considering the values at any fixed $\omega \in \Omega$ of the random variables $T(E(X \mid \mathcal{C}))$, for $X \in L_\infty(\Omega, \mathcal{B}, Q)$ we get a positive, linear functional. Hoffman-Jørgensen [60] discusses the situation, when these functionals can be induced by probability measures defining in this way a regular conditional probability. According to Theorem 1 in that paper, if Ω is a Hausdorff space, $\mathcal{F}$ is the σ-field of Borel sets, Q is a regular measure, then the regular conditional probability can be constructed from the lifting defined on $L_\infty(\Omega, \mathcal{F}, Q)$.

DEFINITION 10.2. In the present paper we shall use the terminology of *regular conditional probability induced by lifting.*

Notice that if the σ-field is generated by a partition then a lifting obviously exists and generates a regular conditional probability.

Let us consider now a set Ω with two σ-fields denoted $\mathcal{A}$ and $\mathcal{B}$. Set $\mathcal{C} = \mathcal{A} \cap \mathcal{B}$. Let P and Q be weakly compatible probability measures on $\mathcal{A}$ and $\mathcal{B}$, respectively. Assume the existence of a lifting T defined on $L_\infty(\Omega, \mathcal{A}, P)$ and $L_\infty(\Omega, \mathcal{B}, Q)$ giving the same value on $L_\infty(\Omega, \mathcal{C}, P) = L_\infty(\Omega, \mathcal{C}, Q)$ and inducing a regular conditional probability on $\mathcal{B}$ with respect to $\mathcal{C}$. Let us denote by ρ the corresponding lifting defined on the sets from $\mathcal{A}$ and $\mathcal{B}$.

In the proof of the following proposition we follow the ideas presented in [115, Section 3.3.].

PROPOSITION 10.1. *Under the previous assumptions if the following stronger version of qualitative conditional independence holds for the σ-fields $\mathcal{A}$ and $\mathcal{B}$:*

if $A \in \mathcal{A}$, $B \in \mathcal{B}$, $A \cap B = \emptyset$, then there exists a set $C \in \mathcal{C}$ such that $B \subset C$, $A \cap \rho(C) = \emptyset$,

then the measures P and Q are strongly compatible.

PROOF. Let us introduce the notation R for the measure obtained from P or Q restricted to the σ-field $\mathcal{C}$. Since P and Q are — according to the assumption — weakly compatible, the measure R is well-defined. The conditional probability of a set $A \in \mathcal{A}$ with respect to $\mathcal{C}$ will be denoted by $P(A \mid \mathcal{C})$, and similarly $Q(B \mid \mathcal{C})$ denotes the conditional probability of $B \in \mathcal{B}$ with respect to $\mathcal{C}$. Here we do not assume that they are regular conditional probabilities.

We are going to prove that there is a common extension μ of the measures P, Q, R which is defined on the σ-field $\sigma(\mathcal{A}, \mathcal{B})$.

Let us first claim that if $B \in \mathcal{B}$ then

$$\rho(\{Q(B \mid \mathcal{C}) = 1\}) \subset \rho(B) \subset \rho(\{Q(B \mid \mathcal{C}) > 0\})\,; \tag{10.1}$$
$$\{T(Q(B \mid \mathcal{C})) > 0\} \subset \rho(\{Q(B \mid \mathcal{C}) > 0\})\,.$$

In fact, using the identity

$$\int_{\{Q(B|\mathcal{C})=1\}} \mathbb{1}(B)\, dQ = \int_{\{Q(B|\mathcal{C})=1\}} Q(B \mid \mathcal{C})\, dQ$$
$$= \int_{\{Q(B|\mathcal{C})=1\}} 1\, dQ = Q(\{Q(B \mid \mathcal{C}) = 1\})$$

we obtain that $Q\left(\{Q(B \mid \mathcal{C}) = 1)\} \setminus B\right) = 0$ implying the first inclusion. On the other hand

$$\int_{\{Q(B|\mathcal{C})=0\}} \mathbb{1}(B)\, dQ = 0\,,$$

thus $Q(\{Q(B \mid \mathcal{C}) = 0\} \cap B) = 0$, giving the second inclusion. Finally, for any random variable ξ with $0 \le \xi \le 1$ we have that

$$T(\xi) = T(\xi\, \mathbb{1}(\xi > 0)) = T(\xi) T(\mathbb{1}(\xi > 0)) = T(\xi)\, \mathbb{1}\big(\rho(\xi > 0)\big)\,.$$

Consequently,

$$\{T(\xi) > 0\} \subset \rho(\xi > 0)\,,$$

proving the last inclusion for $\xi = Q(B \mid \mathcal{C})$.

Now, if $A \in \mathcal{A}$, $B \in \mathcal{B}$ then set

$$\mu(A \cap B) = \int_\Omega P(A \mid \mathcal{C}) Q(B \mid \mathcal{C})\, dR. \tag{10.2}$$

If $A \cap B = \emptyset$, then the stronger version of qualitative conditional independence implies that there exists a set $C \in \mathcal{C}$ for which $B \subset C$, $A \cap \rho(C) = \emptyset$. Consequently,

$$\int_{\Omega \setminus C} Q(B \mid \mathcal{C})\, dR = \int_{\Omega \setminus C} \mathbb{1}(B)\, dR = 0$$

thus the random variable $Q(B \mid \mathcal{C})$ is almost surely zero on the set $\Omega \setminus C$. Similarly,

$$\int_C P(A \mid \mathcal{C})\, dR = \int_C \mathbb{1}(A)\, dR = \int_{\rho(C)} \mathbb{1}(A)\, dR = 0$$

using that $R(\rho(C) \circ C) = 0$. This implies that the random variable $P(A \mid \mathcal{C})$ is almost surely zero on the set C. Thus the product $P(A \mid \mathcal{C})Q(B \mid \mathcal{C}) = 0$ with R-probability 1, giving that $\mu(A \cap B) = 0$, if the sets A, B are disjoint.

Next, we show that if for the sets $A, A' \in \mathcal{A}$ and $B, B' \in \mathcal{B}$ we have that $A \cap B = A' \cap B'$, then $\mu(A \cap B) = \mu(A' \cap B')$. To this aim let us introduce the notation $A_0 = A \cap A'$, $B_0 = B \cap B'$. Then $A_0 \cap B_0 = A \cap B = A' \cap B'$ and

$$(A \setminus A_0) \cap B_0 = \emptyset\,, \quad A_0 \cap (B \setminus B_0) = \emptyset\,, \quad \text{and} \quad (A \setminus A_0) \cap (B \setminus B_0) = \emptyset\,.$$

Using that

$$P(A \mid \mathcal{C}) = P(A \setminus A_0 \mid \mathcal{C}) + P(A_0 \mid \mathcal{C})$$

and

$$Q(B \mid \mathcal{C}) = Q(B \setminus B_0 \mid \mathcal{C}) + Q(B_0 \mid \mathcal{C})$$

with R probability 1, the previous observation for disjoint sets implies that

$$\mu(A \cap B) = \mu(A_0 \cap B_0)\,.$$

Applying the same argument for the sets A', B' we obtain that

$$\mu(A' \cap B') = \mu(A_0 \cap B_0)\,,$$

thus expression (10.2) gives the same value for $\mu(A \cap B)$ if $A' \cap B' = A \cap B$.

Let us extend the set function μ to the algebra generated by the intersections $A \cap B$, where $A \in \mathcal{A}$, $B \in \mathcal{B}$. Using Theorem 1.3.1 in [115] we obtain that any element is this algebra has the form

$$D = \bigcup_{i=1}^{k} (A_i \cap B_i)\,,$$

where $A_1, \dots, A_k \in \mathcal{A}$ form a partition of Ω and $B_1, \dots, B_k \in \mathcal{B}$. Set

$$\mu(D) = \sum_{i=1}^{k} \mu(A_i \cap B_i)\,. \tag{10.3}$$

We are going to show that this is well-defined, that is, if the set D has another representation in the form $D = \bigcup_{j=1}^{\ell} \left(A'_j \cap B'_j\right)$, then the expression (10.3) gives the same value.

Let us consider a common refinement of the two partitions $A_1, \dots, A_k$ and $A'_1, \dots, A'_\ell$ denoted by $A_{i,j} = A_i \cap A'_j$ for $i = 1, \dots, k$, $j = 1, \dots, \ell$. Sim-

ilarly, set $B_{i,j} = B_i \cap B'_j$. Then we have a third representation for the set D as

$$D = \bigcup_{i=1}^{k} \bigcup_{j=1}^{\ell} (A_{i,j} \cap B_{i,j}) \ .$$

Furthermore, since the sets $A_1, \dots, A_k$ are disjoint, we have that

$$A_i \cap B_i = \bigcup_{j=1}^{\ell} (A_{i,j} \cap B_{i,j}) \, . \tag{10.4}$$

Consequently, it is enough to prove that

$$\mu(A_i \cap B_i) = \sum_{j=1}^{\ell} \mu(A_{i,j} \cap B_{i,j}) \, ,$$

because applying a similar argument for $A'_j \cap B'_j$ and taking the sum of these identities with respect to i and j, respectively, we obtain that both representations of the set D give the same value for $\mu(D)$.

Considering the intersection of both sides of (10.4) with $A_{i,r}$ and using that the sets $A_{i,j}$ for $j = 1, \dots, \ell$ are disjoint, we obtain that

$$A_{i,r} \cap B_i = A_{i,r} \cap B_{i,r} \, .$$

As we have proved, this implies that $\mu(A_{i,r} \cap B_i) = \mu(A_{i,r} \cap B_{i,r})$. Consequently,

$$\sum_{r=1}^{\ell} \mu(A_{i,r} \cap B_{i,r}) = \sum_{r=1}^{\ell} \mu(A_{i,r} \cap B_i) = \sum_{r=1}^{\ell} \int P(A_{i,r} \mid \mathcal{C}) Q(B_i \mid \mathcal{C}) \, dR$$

$$= \int P(A_i \mid \mathcal{C}) Q(B_i \mid \mathcal{C}) \, dR = \mu(A_i \cap B_i) \, ,$$

using that $A_i = \bigcup_{r=1}^{\ell} A_{i,r}$ and the sets $A_{i,r}$ for $r = 1, \dots, \ell$ are disjoint.

Summarizing, the identity (10.3) defines a well-defined, additive set function on the algebra generated by the σ-algebras $\mathcal{A}$ and $\mathcal{B}$.

We show that μ can be extended to the generated σ-algebra as a σ-additive measure. To this aim — using Lemmata 2.2.2 and 2.2.1 in [115] — it is enough to prove that for any decreasing sequence of sets $D_1 \supset D_2 \supset \cdots$ from the algebra generated by $\mathcal{A}$ and $\mathcal{B}$ such that $\bigcap_{\ell=1}^{\infty} D_\ell = \emptyset$ we have that $\lim_{\ell \to \infty} \mu(D_\ell) = 0$.

So, let us consider a decreasing sequence of events $D_1 \supset D_1 \supset \cdots$, where $D_\ell = \bigcup_{i=1}^{k_\ell} \big(A_i^{(\ell)} \cap B_i^{(\ell)} \big)$. We might assume that the partition determined by

the sets $A_j^{(\ell+1)}$, $j = 1, \dots, k_{\ell+1}$ is finer than that of $A_i^{(\ell)}$, $i = 1, \dots, k_\ell$ and furthermore, if $A_j^{(\ell+1)} \subset A_i^{(\ell)}$ then $B_j^{(\ell+1)} \subset B_i^{(\ell)}$.

Assume that there exists a value $a > 0$ such that $\mu(D_\ell) > a$ for all $\ell \geq 1$. We show that in this case $\bigcap_\ell D_\ell \neq \emptyset$.

Consider the $\mathcal{A}$-measurable random variable

$$X_\ell = \sum_{i=1}^{k_\ell} \mathbb{1}(A_i^{(\ell)}) T\big(Q(B_i^{(\ell)} \,|\, \mathcal{C})\big)\,, \quad \text{for } \ell \geq 1\,.$$

The value of the random variable $X_{\ell+1}$ on the set $A_j^{(\ell+1)}$ is given by $T\big(Q(B_j^{(\ell+1)} \,|\, \mathcal{C})\big)$. There is a uniquely defined i for which $A_j^{(\ell+1)} \subset A_i^{(\ell)}$. On $A_i^{(\ell)}$ the value of X_ℓ is determined by $T\big(Q(B_i^{(\ell)} \,|\, \mathcal{C})\big)$. Since $B_j^{(\ell+1)} \subset B_i^{(\ell)}$ we have that

$$T\big(Q(B_j^{(\ell+1)} \,|\, \mathcal{C})\big) \leq T\big(Q(B_i^{(\ell)} \,|\, \mathcal{C})\big)\,,$$

consequently $X_{\ell+1} \leq X_\ell$.

On the other hand

$$\int X_\ell \, dP = \mu(D_\ell) > a\,.$$

Consider the sets $F_\ell = \big\{X_\ell \geq \frac{a}{2}\big\}$. Then $F_{\ell+1} \subset F_\ell$, and

$$a < \mu(D_\ell) \leq \frac{a}{2}\,(1 - P(F_\ell)) + P(F_\ell)\,,$$

implying that

$$P(F_\ell) \geq \frac{a}{2-a}\,.$$

Since P is a probability measure on $\mathcal{A}$, Lemma 2.2.1 in [115] implies that $\bigcap_\ell F_\ell \neq \emptyset$. Let us consider an elementary event $\omega^* \in \bigcap_\ell F_\ell$. Then for each ℓ there exists a unique i_ℓ for which $\omega^* \in A_{i_\ell}^{(\ell)}$. Since the partition determined by the sets $A_j^{(\ell+1)}$, $j = 1, \dots, k_{\ell+1}$ is finer than that of $A_i^{(\ell)}$, $i = 1, \dots, k_\ell$, we have that $A_{i_{\ell+1}}^{(\ell+1)} \subset A_{i_\ell}^{(\ell)}$. Set

$$A^* = \bigcap_\ell A_{i_\ell}^{(\ell)} \quad \text{and} \quad B^* = \bigcap_\ell B_{i_\ell}^{(\ell)}\,.$$

We have that $A^* \in \mathcal{A}$, $A^* \neq \emptyset$ and $B^* \in \mathcal{B}$. According to our assumption concerning the representation of the sequence D_1, D_2, …, the inclusion $A_{i_{\ell+1}}^{(\ell+1)} \subset A_{i_\ell}^{(\ell)}$ implies that $B_{i_{\ell+1}}^{(\ell+1)} \subset B_{i_\ell}^{(\ell)}$.

Since $X_\ell(\omega^*) \geq a/2$, we obtain that $T\big(Q(B_{i_\ell}^{(\ell)} \mid \mathcal{C})\big)(\omega^*) \geq a/2$. We have assumed that the lifting T induces a regular conditional probability, in other words the set function $B \mapsto T\big(Q(B \mid \mathcal{C})\big)(\omega^*)$ for $B \in \mathcal{B}$ is a probability measure on $\mathcal{B}$. The sequence $B_{i(1)}^1 \supset B_{i(2)}^2 \supset \cdots$ is a decreasing sequence of events, thus we have that

$$T\big(Q(B^* \mid \mathcal{C})\big)(\omega^*) = \lim_{\ell\to\infty} T\big(Q(B_{i_\ell}^{(\ell)} \mid \mathcal{C})\big)(\omega^*) \geq \frac{a}{2}\,.$$

We get that

$$B^* = \bigcap_\ell B_{i_\ell}^{(\ell)} \neq \emptyset\,.$$

Since $A^* \cap B^* \subset \bigcap_\ell D_\ell$, in order to finish the proof of the proposition it is enough to show that $A^* \cap B^* \neq \emptyset$.

Assume, on the contrary, that $A^* \cap B^* = \emptyset$. Our assumptions imply that in this case there exists a set $C \in \mathcal{C}$ for which $B^* \subset C$ and $A^* \cap \rho(C) = \emptyset$. The inclusion $B^* \subset C$ implies that $\rho(B^*) \subset \rho(C)$ and thus $\rho(\{Q(B^* \mid \mathcal{C}) > 0\}) \subset \rho(C)$ using that $C \in \mathcal{C}$. Consequently, the intersection $A^* \cap \rho(\{Q(B^* \mid \mathcal{C}) > 0\})$ should be empty. Using identity (10.1) we obtain that $A^* \cap \{T(Q(B^* \mid \mathcal{C})) > 0\} = \emptyset$, contradicting that ω^* belongs to both events. Thus $A^* \cap B^* \neq \emptyset$, so $\bigcap_\ell D_l$ is non-empty. Consequently, the measure μ can be extended to the σ-field generated by $\mathcal{A}$ and $\mathcal{B}$. This concludes the proof of the proposition. □

11. Primes, prime gaps and independence

Several of Rényi's first papers were devoted to the theory of primes. In fact, these works brought a high international reputation for him as early as in 1947–48 [108,109]. He wrote his PhD thesis (in fact, his Candidate of Science thesis) under the guidance of Linnik who was one of the greatest figures of 20$^{\text{th}}$ century mathematics, in probability theory as well as in number theory.

The mentioned works containing the results of his thesis dealt with approximations to two of the oldest and most celebrated (still open) problems of mathematics: the (binary) Goldbach problem and the twin prime conjecture. Goldbach's conjecture arose in the correspondence of Euler and Goldbach in 1742 and it asserts that every even integer larger than two can be written as the sum of two primes. The twin prime conjecture may have its origin in the ancient Greek mathematics and it states the existence of infinitely many twin primes, i.e. primes p for which $p+2$ is also a prime.

One can easily see the similarity of the two problems, and in fact the methods which brought advances in one of the mentioned conjectures led

to analogous results in the other problem too, at least until the work about small gaps between primes [50], whose method worked for the approximation of the twin prime problem but gave no results for the Goldbach problem.

The first successful attack for these problems was made by Brun [28] exactly hundred years ago. He showed that every sufficiently large even number can be written as the sum of two numbers having at most nine prime factors. Analogously he proved the existence of infinitely many pairs of integers $(n, n+2)$ such that both have at most nine prime factors. His tool was a sieve method, invented by him, called today Brun's sieve. In the following two decades the above result was improved in several steps to almost primes having at most four prime factors. The tool remained Brun's sieve.

Rényi was the first who succeeded to show in his above mentioned works these theorems in the form that one of the terms in the sum (or difference, respectively) can be a prime and the other a number having at most K prime factors with an unspecified large absolute constant K. The value of K was diminished in several steps, until in 1966 Chen Jing-Run [31,32] could show this with $K = 2$.

The tools Rényi used were Brun's sieve and a new version of the large sieve of Linnik, also often called the large sieve of Linnik and Rényi (see e.g. [24, §1]). Besides a direct self-contained approach [112], Rényi gave also a proof for his sieve [110] where he derives his result from a general probabilistic theorem. Rényi's large sieve is formulated in the book of Bombieri [24] in the following form (in what follows p will always denote primes):

Suppose that $1 \le n_1 < n_2 < \cdots < n_Z \le N$, $Z < N$, are arbitrary integers,

$$Z(p,a) = \#\big\{\, i : n_i \equiv a (\mathrm{mod}\ p)\big\}\,.$$

Then we have for any $X \le (N/12)^{1/3}$

$$V := \sum_{p \le X} p \sum_{a=1}^{p} \Big(Z(p,a) - \frac{Z}{p} \Big)^2 \le 2NZ.$$

This was improved to $V \le (N + X^2)Z$ for arbitrary values of X and extended for composite moduli. Its most important application is the celebrated Bombieri–Vinogradov theorem (Bombieri [23] and Vinogradov [131, 132]) which showed that prime numbers up to X are on average equidistributed in residue classes modulo $q \le Q$ if $Q \le X^{1/2}(\log X)^{-B}$ with an average error of size $X^{1/2}(\log X)^{C}$ if the worst residue class a (mod q) is considered for every $q \le Q$, i.e.

$$\sum_{q \le X^{1/2} \log^{-B} X} \max_{y \le X} \max_{(a,q)=1} \left| \pi(y,q,a) - \frac{\operatorname{li} y}{\varphi(q)} \right| = O\big(X \log^{-A} X \big)$$

if $B = B(A) > 0$ is sufficiently large, $\pi(y, q, a)$ denotes the number of primes $p \equiv a (\mathrm{mod}\ q)$ with $p \le y$, $\mathrm{li}\, x = \int_0^x dt/\log t$, and $\varphi(n)$ is Euler's totient function: $\varphi(n) = \#\{m \le n : \gcd(m, n) = 1\}$. This result without averaging over q is of the same strength as the Generalized Riemann Hypothesis. Barban [13] used the large sieve of Linnik and Rényi to prove the above relation for moduli $q \le X^{1/6-\varepsilon}$. Two years later he extended this for moduli up to $X^{1/3-\varepsilon}$ [14]. For a survey and an extension to $3/8 - \varepsilon$ see [15].

The above theorem of Bombieri–Vinogradov has countless applications. For example, Bombieri [24, §9, Théorème 19] uses it to prove Rényi's theorem in the stronger form that every sufficiently large even integer is the sum of a prime and a number with at most four prime factors.

Set $p_0 = 1$ and for $n > 0$ denote the nth prime by p_n. The differences between two successive prime numbers are called prime gaps and denoted by $d_n = p_n - p_{n-1}$. Properties of prime gaps have been in limelight for a long time. Rényi also devoted a couple of papers to this problem [44,111,112], and this is still a vital research direction in our days too, see e.g. [78].

Introduce the notation

$$x_n := (p_n - p_{n-1})/\log(p_n), \quad n = 1, 2, \dots. \tag{11.1}$$

Note that the merit of the prime gap $p_{n+1} - p_n$ is usually defined as $(p_{n+1} - p_n)/\log(p_n)$, that is, the gap is divided by the natural logarithm of the smaller prime. It does not make a difference asymptotically.

Let us consider the closure J of the set of all x_n's. Erdős [43] formulated the conjecture that $J = [0, \infty]$. Before this only the point infinity was known to belong to J [134], and the result of Erdős [42] that J contains numbers less than 1. In the above mentioned work [43] and in a paper of Ricci [117] it was shown that J has positive Lebesgue measure $\mu(J)$. However, no finite limit point of J was explicitly known until 2005 when $0 \in J$ was shown by Goldston, Pintz and Yıldırım [50]. This is equivalent to

$$\liminf_{n\to\infty} x_n = \liminf_{n\to\infty} \frac{d_n}{\log p_n} = 0.$$

This was improved in the celebrated works of Yitang Zhang [136] and Maynard [90] to

$$\liminf_{n\to\infty} d_n \le C$$

with $C = 7 \cdot 10^7$ [136] and $C = 600$ [90]. The latter method was refined in a Polymath project [104,105] to $C = 246$. These methods opened the way to obtain further information about the set J of limit points. Pintz [99] has shown that J contains an interval $[0, c]$ for some ineffective constant $c > 0$.

(Which means that we still do not know any concrete element of J besides 0 and 1). The method of Maynard led Banks et al. [12] to

$$\mu(J_T) := \mu(J \cap [0,T]) \geq \frac{T(1+o(1))}{8}$$

which was improved in further works of Pintz to $\mu(J_T) \geq (1+o(1))T/4$ [100] and Merikoski to $\mu(J_T) \geq T/3$ [91].

The Prime Number Theorem [55,129]

$$\pi(x) = \sum_{p_i \leq x} 1 \sim \operatorname{li} x = \int_0^x \frac{dt}{\log t} \sim \frac{x}{\log x}$$

suggests that we may expect in a short interval $[x, x+h]$ about λ primes if $h \sim \lambda \log x$ where $h, x \to \infty$.

If we consider bounded differences then a far-reaching generalisation of the twin prime conjecture was formulated by Dickson [37] and later in a quantitative form by Hardy and Littlewood [56]. The prime k-tuple conjecture of Hardy and Littlewood asserts that if $\mathcal{H} = \{h_i\}_{i=1}^k$ is a k-tuple of non-negative integers ($h_i < h_{i+1}$), then (denoting the set of primes by $\mathcal{P}$)

$$\pi_{\mathcal{H}}(x) = \sum_{\substack{n \leq x \\ n+h_i \in \mathcal{P}}} 1 = \big(\mathfrak{S}(\mathcal{H}) + o(1)\big) \frac{x}{(\log x)^k}.$$

We call $\mathcal{H}$ *admissible* if the so-called singular series is positive, that is,

$$\mathfrak{S}(\mathcal{H}) = \prod_p \Big(1 - \frac{\nu_{\mathcal{H}}(p)}{p}\Big)\Big(1 - \frac{1}{p}\Big)^{-k} > 0,$$

where $\nu_{\mathcal{H}}(p)$ denotes the number of distinct residue classes occupied by $h_1, \ldots, h_k \bmod p$. The above condition is equivalent to the property that the polynomial $P(n) = (n+h_1)\cdots(n+h_k)$ has no fixed prime divisor. Thus if $\mathcal{H}$ is admissible the Conjecture above predicts an infinity of numbers n with $\{n+h_i\}_{i=1}^k \in \mathcal{P}^k$.

Gallager [46] showed that if the above conjecture holds in a stronger uniform form, then the number $P_r(h,N)$ of $n \leq N$ such that the interval $[n, n+h]$ contains exactly r primes is under the condition $h \sim \lambda \log N$ asymptotically

$$P_r(h,N) \sim N e^{-\lambda} \frac{\lambda^k}{k!},$$

i.e. it satisfies a Poisson distribution with parameter λ.

The phenomenon that primes seem to follow globally a given random distribution, was formulated as early as in the 1930's by Cramér [34]. Gauss

observed already at an age of 15–16 years (based on prime number tables up to three million) that primes around x occur with a frequency $1/\log x$. He never published anything about this but his observation was confirmed by the mentioned proof of the Prime Number Theorem, the asymptotic relation $\pi(x) \sim \operatorname{li} x$, proved by Hadamard [55] and Vallée-Poussin [129] about hundred years later. Cramér [34] suggested a probabilistic model for primes according to which every natural number $n > 2$ is chosen as prime independently with probability $1/\log n$. This model can clearly not reflect all properties of the deterministic sequence of primes. For example, the model would suggest asymptotically the same number $x/(2\log x)$ of even and odd primes below x.

Cramér used his model to conjecture

$$\max_{p_n \le x}(p_n - p_{n-1}) \sim c_0 \log^2 x \quad \text{with } c_0 = 1.$$

Numerical calculations suggest that the order of growth, $\log^2 x$ is correct but we are uncertain whether $c_0 = 1$. Some theoretic considerations in [51,52] suggest that $c_0 > 1$, may be $c_0 = 2e^{-\gamma}$, whereas in the range of calculations (i.e. up to 10^{18}) we obtain a constant ≈ 0.9206, i.e. less than 1 (as calculated by T. R. Nicely).

The general belief was that Cramér's model gives a good prediction in those cases where we do not have a simple reason for its falsity (like the mentioned case of even and odd primes, for example). It was therefore a great surprise when Helmut Maier [87] showed that the Prime Number Theorem is not true in short intervals of type

$$\left[x, x + (\log x)^A\right], \quad A > 0 \text{ arbitrary},$$

whereas Cramér's model would suggest its truth with probability 1 for every $A > 2$. Later a modification of the model eliminated this contradiction [51, 52]. Nevertheless it was shown [98] that essentially any reasonable model based on independent random variables makes wrong predictions in some global problems as well as, for example, the average order of the error term of the Prime Number Theorem, i.e. the average order of $\Delta(x) = \pi(x) - \operatorname{li} x$. An alternative for Cramér's model was recently suggested by Banks et al. [11].

However, in some other cases, like the value of the mentioned singular series $\mathfrak{S}(\mathcal{H})$ in Hardy–Littlewood's prime k-tuple conjecture the model does make good predictions if we modify it slightly (see e.g., [98]). We may mention that the original conjecture of Hardy and Littlewood was based on more complicated arguments, namely on a heuristic application of the circle method and a summation of the corresponding (multiple) singular series.

Let us now turn to the following question: what can we say about the ranks of the sequence x_n defined in (11.1)? In other words, consider the permutation $r(1), \ldots, r(n)$ of $1, \ldots, n$ such that

$$x_{r(1)} < x_{r(2)} < \cdots < x_{r(n)}.$$

Denote this permutation by $\mathcal{R}(n)$. What can we say about these permutations as $n \to \infty$?

Consider the permutation $\mathcal{R}(n)$ for fixed n. Then the changing speed of the differences d_k is proportional to the numbers themselves. The d_k's are of course even for $k > 2$ and it is conjectured that they take each value infinitely many times. Thus the smallest possible values persist but initially the ratio of changes is much larger than the changes of the value of the logarithm in their neighborhood; because of that, threads are formed in the permutations.

The limit distribution of the x_n's is conjectured to be standard exponential. What we do not know is if we have a sort of independence. If the x_n's were iid then the permutations above would have uniform distribution and we would know a lot about their dynamics. Rényi [114] discussed a similar problem: instead of the prime gaps he considered the (suitably normalized) gaps between consecutive numbers coprime with and less than n. He found that they behave similarly to a homogeneous Poisson process. He proved his statements via the Poisson model. Rényi liked to use the fact that in the limit the ranks of iid uniform variables follow the Poisson process. (The conditional distribution of the points of a Poisson process in an interval given that there are n of them is the same as the distribution of an ordered sample of size n from the uniform distribution on that interval.) The Hungarian stochastic school prefers referring to this method as Rényi's method but it was also discovered by others before Rényi. Rényi gave a talk about the applications of stochastic methods in different areas of mathematics at a conference in Canada. Unfortunately we were unable to find this paper.

Let us fix n and denote the kth elements of the ordered sample of the x's by z_k, i.e. $z_k = x_{r(k)}$. Let us label the element z_k of the ordered sample by the difference $d_{r(k)}$. For each possible label $d = 1, 2, \ldots$ the corresponding x's are monotone decreasing starting with x_F where F is the serial number of the first appearance of d and ending by x_L while d_L is the last difference equal to the chosen d. Because $x_L = d/\log(p_L)$ with a denominator barely changing with d these numbers are practically proportional to d.

Since the size of the maximal gap g_n is small with respect to p_n (more precisely $g_n \le p_n^\theta$, where θ can be taken to be 0.525 according to [7]) for the different values of d the corresponding decreasing runs in the permutation $\mathcal{R}(n)$ are in general exceptionally long (with the obvious exception for

$d = 1$). As we have pointed out, the smallest values in these runs are approximately proportional to d. Now, these runs are merged together to form the complete permutation $\mathcal{R}(n)$.

Now, for any fixed difference d let us consider the appearance of this value in the whole sequence $d_{r(k)}$. Denote by h the number of different values between two consecutive occurrences of the same value d. In particular, if $h = 0$ then $d_{r(k)} = d_{r(k+1)}$. The second columns in Table 4 contain the cumulative tail distribution function of the h values. (The total number of primes considered is 50 847 533, and in $39\,504\,660 = 50\,847\,533 - 11\,342\,873$ cases a value d was followed by the same value in the sequence $d_{r(k)}$, $k = 1, \dots, 50\,847\,533$. In 7 114 470 cases there was only one other value between two consecutive occurrences of the same d value, and so on.) Thus in the sequence of prime gaps, when they are ordered according to the modified merits $x_{r(k)}$ the same d values are sticked together.

In order to check how particular behavior is this one, the same set of prime gaps was arranged into another order using a random modification of the original sequence $d_{r(k)}$. For some fixed value n_0 the position of quantities $d_{r(k+n_0)}$ and $d_{r(n+1-k)}$ changed places with probabilities $1/2$, independently for the different k values ($k = 1, \dots, (n - n_0)/2$). The third columns of Table 4 show that after these random changes in the sequence the cumulative tail distribution function of the statistics h became fatter, thus the position of the same d values became more scattered. This shows the particularly interesting behavior of the prime gaps. On this topic see also [27]. Related interesting reads are [6] and [53].

Final words

One of Rényi's last papers is joint with his wife, Kató. On this topic Rényi gave a breathtaking conference talk in 1969, a few days after Kató's death on August 23 (see [116]).

Acknowledgements. On behalf of Project "Multivariate hypothesis testing" we thank for the usage of ELKH Cloud (https://science-cloud.hu/) that significantly helped us achieve the results published in Section 7. Authors are also indebted to Imre Z. Ruzsa and Kati Bognár for their valuable remarks.

Model	Rtest. global.statistic	Rtest. fisher	Rtest. tippett	New test
Exponential, v = 3	779	519	332	930
Exponential, v = 4	464	243	155	629
Exponential, v = 5	214	109	76	294
Exponential, v = 6	83	62	59	119
Exponential, v = 7	60	51	50	79
Exponential, v = 8	62	54	55	75
Exponential, v = 9	49	50	45	60
Exponential, v = 10	52	43	43	62
Clayton, tau = 0.2	988	965	885	997
Clayton, tau = 0.15	903	819	627	969
Clayton, tau = 0.15	610	448	312	761
Clayton, tau = 0.05	255	145	94	338
Gauss, tau = 0.2	991	967	874	998
Gauss, tau = 0.15	919	791	638	961
Gauss, tau = 0.1	625	419	309	681
Gauss, tau = 0.05	256	141	108	297
Gumbel, tau = 0.2	975	974	905	998
Gumbel, tau = 0.15	842	844	682	943
Gumbel, tau = 0.1	562	518	355	761
Gumbel, tau = 0.05	176	160	120	285

Table 1: Power of the tests at the 5% level ($p = 3$)

Model	Rtest. global.statistic	Rtest. fisher	Rtest. tippett	New test
Exponential, v = 3	987	528	227	1000
Exponential, v = 4	925	358	159	996
Exponential, v = 5	703	209	123	957
Exponential, v = 6	415	134	78	706
Exponential, v = 7	195	93	56	412
Exponential, v = 8	85	66	44	176
Exponential, v = 9	38	59	44	77
Exponential, v = 10	36	56	42	50
Clayton, tau = 0.2	1000	999	978	1000
Clayton, tau = 0.15	996	978	764	1000
Clayton, tau = 0.15	922	729	344	989
Clayton, tau = 0.05	516	239	99	653
Gauss, tau = 0.2	1000	1000	966	1000
Gauss, tau = 0.15	999	954	718	1000
Gauss, tau = 0.1	956	675	369	982
Gauss, tau = 0.05	558	202	105	609
Gumbel, tau = 0.2	999	998	977	1000
Gumbel, tau = 0.15	972	986	855	999
Gumbel, tau = 0.1	755	857	529	965
Gumbel, tau = 0.05	360	431	179	694

Table 2: Power of the tests at the 5% level ($p = 5$)

Product structure			No product structure		
	$\mathcal{B}$ yes	$\mathcal{B}$ no		$\mathcal{B}$ yes	$\mathcal{B}$ no
$\mathcal{A}$ yes	1	2	$\mathcal{A}$ yes	1	2
$\mathcal{A}$ no	3	4	$\mathcal{A}$ no	3	-

Table 3: The structures of the spaces in Examples 9.1 and 9.2

h	prime	control	h	prime	control	h	prime	control
0	50847533	50847533	42	-	476618	84	-	2486
1	11342873	45719214	43	-	421002	85	-	2162
2	4228403	41345257	44	-	372074	86	-	1893
3	1847795	37594880	45	-	328219	87	-	1652
4	889304	34325944	46	-	289738	88	-	1460
5	457644	31444604	47	-	256106	89	-	1281
6	247113	28862927	48	-	226285	90	-	1121
7	139129	26520179	49	-	199833	91	-	1000
8	80547	24368685	50	-	176625	92	-	892
9	47983	22377756	51	-	156333	93	-	781
10	29255	20517902	52	-	138336	94	-	688
11	18266	18771664	53	-	122574	95	-	610
12	11606	17130188	54	-	108533	96	-	545
13	7429	15587438	55	-	96250	97	-	481
14	4833	14135235	56	-	85202	98	-	419
15	3179	12774156	57	-	75447	99	-	361
16	2147	11504408	58	-	66738	100	-	320
17	1398	10325180	59	-	58824	101	-	281
18	956	9233525	60	-	52029	102	-	239
19	638	8231895	61	-	45999	103	-	211
20	457	7320402	62	-	40421	104	-	188
21	329	6494564	63	-	35470	105	-	166
22	238	5751255	64	-	31260	106	-	153
23	169	5083791	65	-	27402	107	-	132
24	123	4492357	66	-	24049	108	-	117
25	81	3967685	67	-	21137	109	-	99
26	54	3503412	68	-	18663	110	-	88
27	32	3094779	69	-	16468	111	-	73
28	24	2734341	70	-	14515	112	-	59
29	18	2415488	71	-	12853	113	-	49
30	12	2133496	72	-	11317	114	-	41
31	7	1884466	73	-	9994	115	-	38
32	5	1664544	74	-	8838	116	-	35
33	4	1469978	75	-	7774	117	-	31
34	4	1297565	76	-	6900	118	-	28
35	3	1144877	77	-	6085	119	-	25
36	3	1009986	78	-	5336	120	-	20
37	2	890668	79	-	4712	121	-	16
38	1	785864	80	-	4196	122	-	7
39	1	693008	81	-	3691	123	-	5
40	-	611292	82	-	3250	124	-	4
41	-	539700	83	-	2826			

Table 4: Ordered Prime Gaps

References

[1] P. K. Agarwal, K. Fox, D. Panigrahi, K. R. Varadarajan and A. Xiao, Faster algorithms for the geometric transportation problem, in: B. Aronov and M. J. Katz (eds.), *33rd International Symposium on Computational Geometry (SoCG 2017)*, Leibniz International Proceedings in Informatics (LIPIcs) 77, Schloss Dagstuhl–Leibniz-Zentrum für Informatik (Dagstuhl, Germany, 2017), pp. 7:1–7:16.

[2] M. Ajtai, J. Komlós and G. Tusnády, On optimal matchings, *Combinatorica*, **4** (1984), 259–264.

[3] N. Alon, Explicit construction of exponential sized families of k-independent sets, *Discrete Math.*, **58** (1986), 191–193.

[4] J. Altschuler, J. Niles-Weed and P. Rigollet, Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration, in: I. Guyon et al. (eds), *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, Curran Associates Inc. (Red Hook, NY, 2017), pp. 1964–1974.

[5] J. Altschuler, F. Bach, A. Rudi and J. Weed, Massively scalable Sinkhorn distances via the Nyström method, in: H. Wallach et al. (eds), *Advances in Neural Information Processing Systems 32 (NIPS 2019)*, Curran Associates Inc. (Red Hook, NY, 2019), pp. 4427–4437.

[6] R. Arratia, A. D. Barbour and S. Tavaré, *Logarithmic Combinatorial Structures: a Probabilistic Approach*, EMS Monographs in Mathematics, European Mathematical Society (2003).

[7] R. C. Baker, G. Hartman and J. Pintz, The difference between consecutive primes II, *Proc. London Math. Soc.*, **83** (2001), 532–562.

[8] N. K. Bakirov, M. L. Rizzo and G. J. Székely, A multivariate nonparametric test of independence, *J. Multivariate Anal.*, **97** (2006), 1742–1756.

[9] B. Balázs, m-Qualitatively independent families, *J. Stat. Theory Pract.*, **9** (2015), 733–740.

[10] S. Banach, On measures in independent fields (edited by S. Hartman), *Studia Math.*, **10** (1948), 159–177.

[11] W. Banks, K. Ford and T. Tao, Large prime gaps and probabilistic models, arXiv:1908.08613 [math.NT] preprint (2019).

[12] W. D. Banks, T. Freiberg and J. Maynard, On limit points of the sequence of normalized prime gaps, *Proc. Lond. Math. Soc. (3)*, **113** (2016), 515–539.

[13] M. B. Barban, New applications of the "great sieve" of Yu. V. Linnik, *Akad. Nauk Uzbek. SSR Trudy Inst. Mat.*, **22** (1961), 1–20 (in Russian).

[14] M. B. Barban, The "density" of the zeros of Dirichlet L-series and the problem of the sum of primes and "near primes", *Mat. Sb. (N.S.)*, **61** (1963), 418–425 (in Russian).

[15] M. B. Barban, The "large sieve" method and its application to number theory, *Uspehi Mat. Nauk*, **21** (1966), 51–102 (in Russian).

[16] P. Bártfai and T. Rudas, Conditionally independent extension of measures, *Preprint 57/1998*, Mathematical Institute, Hungarian Academy of Sciences (Budapest, 1998)

[17] J. D. Benamou and Y. Breiner, A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem, *Numer. Math.*, **84** (2000), 375–393.

[18] W. P. Bergsma and T. Rudas, Marginal models for categorical data, *Ann. Stat.*, **30** (2002), 140–152.

[19] Y. N. Bishop, S. E. Fienberg and P. W. Holland, *Discrete Multivariate Analysis: Theory and Practice*, Springer (New York, 2007).

[20] J. R. Blum, J. Kiefer and M. Rosenblatt, Distribution free tests of independence based on the sample distribution function, *Ann. Math. Stat.*, **32** (1961), 485–498.
[21] M. Bolla and A. Kurdyukova, Dynamic factors of macroeconomic data, *Ann. Univ. Craiova, Math. Comp. Sci. Ser.*, **37** (2010), 18–28.
[22] B. Bollobás, Sperner systems consisting of pairs of complementary subsets, *J. Comb. Theory A*, **15** (1973), 363–366.
[23] E. Bombieri, On the large sieve, *Mathematika*, **12** (1965), 201–225.
[24] E. Bombieri, *Le grand crible dans la théorie analytique des nombres*, Astérisque, No. 18, Société Mathématique de France (Paris, 1974).
[25] B. Böttcher, M. Keller-Ressel and R. L. Schilling, Distance multivariance: new dependence measures for random vectors, *Ann. Stat.*, **47** (2019), 2757—2789.
[26] A. Brace and D. E. Daykin, Sperner-type theorems for finite sets, in: D. J. A. Welsh and D. R. Woodall (eds), *Combinatorics: Being the Proceedings of the Conference on Combinatorial Mathematics Held at the Mathematical Institute, Oxford*, Inst. Math. Appl. (Southend-on-Sea, UK, 1972), pp. 18–37.
[27] A. Breitzman Sr, A new look at Pólya's prime gap heuristics, *Math. Scientist*, **42** (2017), 38–42.
[28] V. Brun, Le crible d'Eratosthène et le théorème de Goldbach, *Videnskaps. Skr., I. Mat.-Naturv. Kl. Kristiania* 1920, No. 3, 36 pp.
[29] L. A. Caffarelli, The Monge-Ampère equation and optimal transportation, an elementary review, in: L. Ambrosio et al. (eds), *Optimal Transportation and Applications*, Lecture Notes in Mathematics, 1813, Springer (Berlin, Heidelberg, 2003), pp. 1–10.
[30] L. A. Caffarelli and R. J. McCann, Free boundaries in optimal transport and Monge-Ampère obstacle problems, *Ann. of Math (2)*, **171** (2010), 673–730.
[31] J. Chen, On the representation of a large even integer as the sum of a prime and the product of at most two primes, *Kexue Tongbao*, **17** (1966), 385–386.
[32] J. Chen, On the representation of a larger even integer as the sum of a prime and the product of at most two primes, *Sci. Sinica*, **16** (1973), 157–176.
[33] D. R. Cox and N. Reid, Parameter orthogonality and approximate conditional inference, *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **49** (1987), 1–39.
[34] H. Cramér, Prime numbers and probability, *Skand. Math. Kongr.*, **8** (1935), 107–115.
[35] S. Csörgő, Testing for independence by the empirical characteristic function, *J. Multivariate Anal.* **16** (1985), 290–299.
[36] P. Deheuvels, An asymptotic decomposition for multivariate distribution-free tests of independence, *J. Multivariate Anal.*, **11** (1981), 102–113.
[37] L. E. Dickson, A new extension of Dirichlet's theorem on prime numbers, *Messenger of Math.*, **33** (1904), 155–161.
[38] J. Dueck, D. Edelmann and D. Richards, Distance correlation coefficients for Lancaster distributions. *J. Multivariate Anal.*, **154** (2017), 19–39.
[39] M. L. Eaton, *Group Invariance. Applications in Statistics*, NSF-CBMS Regional Conference Series in Probability and Statistics 1, IMS (Hayward, CA, 1989)
[40] D. Edelmann, T. F. Móri and G. J. Székely, On relations between Pearson's correlation and distance correlation, *Stat. Probab. Lett.*, **169** (2021), 108960.
[41] J. Edmonds and R. M. Karp, Theoretical improvements in algorithmic efficiency for network flow problems, *J. ACM*, **19** (1972), 248–264.
[42] P. Erdős, The difference of consecutive primes, *Duke Math. J.*, **6** (1940), 438–441.
[43] P. Erdős, Some problems on the distribution of prime numbers, In: *C.I.M.E. Teoria dei Numeri, Math. Congress Varenna*, Istituto Matematico dell'Università (Roma, 1955), pp. 79–88.
[44] P. Erdős and A. Rényi, Some problems and results on consecutive primes, *Simon Stevin*, **27** (1950), 115–125.

[45] P. Frankl, An extremal problem of coding type, *Ars Combinatoria*, **1** (1976), 53–55.
[46] P. X. Gallagher, On the distribution of primes in short intervals, *Mathematika*, **23** (1976), 4–9.
[47] L. Gargano, J. Körner and U. Vaccaro, Sperner theorems on directed graphs and qualitative independence, *J. Comb. Theory A*, **61** (1992), 173–192.
[48] L. Gargano, J. Körner and U. Vaccaro, Sperner capacities, *Graph. Combin.*, **9** (1993), 31–56.
[49] C. Genest, J-F. Quessy and B. Rémillard, Asymptotic local efficiency of Cramér–von Mises tests for multivariate independence, *Ann. Stat.*, **35** (2007), 166–191.
[50] D. A. Goldston, J. Pintz and C. Y. Yıldırım, Primes in tuples I, *Ann. of Math.*, **170** (2009), 819–862.
[51] A. Granville, Harald Cramér and the distribution of prime numbers. Harald Cramér Symposium (Stockholm, 1993), *Scand. Actuar. J.*, **1** (1995), 12–28.
[52] A. Granville, Unexpected irregularities in the distribution of prime numbers, in: S. D. Chatterji (ed) *Proceedings of the International Congress of Mathematicians (Zürich, 1994)*, Vols. 1, 2, Birkhäuser (Basel, 1995), pp. 388–399.
[53] A. Granville, The anatomy of integers and permutations, preprint, Université de Montréal, Canada (2008).
[54] A. Gretton and L. Győrfi, Consistent nonparametric tests of independence, *J. Mach. Learn. Res.*, **11** (2010), 1391–1423.
[55] J. Hadamard, Sur la distribution des zéros de la fonction $\zeta(s)$ et ses conséquences arithmétiques, *Bull. Soc. Math. France*, **24** (1896), 199–220.
[56] G. H. Hardy and J. E. Littlewood, Some problems of 'Partitio Numerorum'; III: On the expression of a number as sum of primes, *Acta Math.*, **44** (1923), 1–70.
[57] H. Helson, Remark on measures in almost independent fields, *Studia Math.*, **10** (1948), 182–183.
[58] W. Hoeffding, Masstabinvariante Korrelationstheorie, *Schr. Math. Inst. und Inst. Angew. Math. Univ. Berlin*, **5** (1940), 181–233.
[59] W. Hoeffding, A non-parametric test of independence, *Ann. Math. Stat.*, **19** (1948), 546–557.
[60] J. Hoffmann-Jørgensen, Existence of conditional probabilities, *Math. Scand.*, **28** (1971), 257–264.
[61] A. Ionescu Tulcea and C. Ionescu Tulcea, On the lifting property (I), *J. Math. Anal. Appl.*, **3** (1961), 537–546.
[62] A. Ionescu Tulcea and C. Ionescu Tulcea, On the lifting property (IV). Disintegration of measures, *Ann. Inst. Fourier*, **14(2)** (1964), 445–472.
[63] M. E. Jakobsen, Distance Covariance in Metric Spaces: Non-Parametric Independence Testing in Metric Spaces (Master's thesis), arXiv:1706.03490 [math.ST] preprint (2017).
[64] A. M. Kagan, Yu. V. Linnik and C. R. Rao, *Characterization Problems in Mathematical Statistics*, Wiley (New York, 1973).
[65] A. M. Kagan and G. J. Székely, Calibrating dependence between random elements, *J. Theor. Probab.*, **34** (2021), 784–790.
[66] L. V. Kantorovich, On the translocation of masses, *Dokl. Akad. Nauk SSSR*, **37** (1942), 227–229 (in Russian).
[67] L. V. Kantorovich and G. S. Rubinstein, On a space of completely additive functions, *Vestnik Leningrad Univ. Ser. Mat. Mekh. Astron.*, **13** (1958), 52–59 (in Russian).
[68] G. O. H. Katona, Two applications (for search theory and truth functions) of Sperner type theorems, *Period. Math. Hungar.*, **3** (1973), 19–26.
[69] G. O. H. Katona, Strong qualitative independence, *Discrete Appl. Math.*, **137** (2004), 87–95.

[70] P. Kleinschmidt and H. Schannath, A strongly polynomial algorithm for the transportation problem, *Math. Program.*, **68** (1955), 1–13.
[71] D. Kleitman and J. Spencer, Families of k-independent sets, *Discrete Math.* **6** (1973), 255–262.
[72] A. Klimova, T. Rudas and A. Dobra, Relational models for contingency tables, *J. Multivariate Anal.*, **104** (2012), 159–173.
[73] A. Klimova and T. Rudas, Iterative scaling in curved exponential families, *Scand. J. Stat.*, **42** (2015), 832–847.
[74] A. Klimova and T. Rudas, On the closure of relational models, *J. Multivariate Anal.*, **143** (2016), 440–452.
[75] A. Klimova and T. Rudas, On the role of the overall effect in exponential families, *Electron. J. Stat.*, **12** (2018), 2430–2453.
[76] A. Klimova and T. Rudas, Hierarchical Aitchison–Silvey models for incomplete binary sample spaces, *J. Multivariate Anal.* (2021), to appear. arXiv:2002.00357 [stat.ME] preprint (2020).
[77] I. Kojadinovic and J. Yan, Modeling Multivariate Distributions with Continuous Margins Using the copula R Package, *J. Stat. Softw.*, **34** (2010), 1–20.
[78] A. Kourbatov and M. Wolf, Predicting maximal gaps in sets of primes, *Mathematics*, **7** (2019), 400.
[79] J. Körner and A. Monti, Delta-systems and qualitative (in)dependence, *J. Comb. Theory A*, **99** (2002), 75—84.
[80] J. Körner and G. Simonyi, A Sperner-type theorem and qualitative independence, *J. Comb. Theory A*, **59** (1992), 90–103.
[81] H. W. Kuhn, The Hungarian method for the assignment problem, *Naval Res. Logist.*, **2** (1955), 83–97.
[82] S. L. Lauritzen, *Graphical Modeling*, Oxford University Press (Oxford, 1996).
[83] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses* (3rd ed.), Springer (New York, 2005).
[84] R. Lyons, Distance covariance in metric spaces, *Ann. Probab.*, **41** (2013), 3284–3305.
[85] R. Lyons, Errata to "Distance covariance in metric spaces", *Ann. Probab.*, **46** (2018), 2400–2405.
[86] D. Maharam, On a theorem of von Neumann, *Proc. Amer. Math. Soc.*, **9** (1958), 987–994.
[87] H. Maier, Primes in short intervals, *Michigan Math. J.*, **32** (1985), 221–225.
[88] E. Marczewski, Indépendance d'ensembles et prolongement de mesures (Résultats et problèmes), *Colloq. Math.*, **1** (1948), 122–132.
[89] E. Marczewski, Measures in almost independent fields, *Fundamenta Math.*, **38** (1951), 217–229.
[90] J. Maynard, Small gaps between primes, *Ann. Math.*, **181** (2015), 383–413.
[91] J. Merikoski, Limit points of normalized prime gaps, *J. Lond. Math. Soc. (2)*, **102** (2020), 99–124.
[92] G. Monge, *Mémoire sur la théorie des déblais et des remblais*, De l'Imprimerie Royale (Paris, 1781).
[93] T. F. Móri, Essential correlatedness and almost independence, *Stat. Probab. Lett.*, **15** (1992), 169–172.
[94] T. F. Móri and G. J. Székely, Representations by uncorrelated random variables, *Math. Method. Stat.*, **26** (2017), 149–153.
[95] T. F. Móri and G. J. Székely, Four simple axioms of dependence measures, *Metrika*, **82** (2019), 1–16.
[96] T. F. Móri and G. J. Székely, The earth mover's correlation, *Ann. Univ. Sci. Budapest, Sect. Comput.*, **50** (2020), 349–268.
[97] V. M. Panaretos and Y. Zemel, Statistical aspects of Wasserstein distances, *Annu. Rev. Stat. Appl.*, **6** (2019), 405–431.

[98] J. Pintz, Cramér vs. Cramér. On Cramér's probabilistic model for primes, *Funct. Approx. Comment. Math.*, **37** (2007), 361–376.
[99] J. Pintz, Polignac numbers, conjectures of Erdős on gaps between primes, arithmetic progressions in primes, and the bounded gap conjecture, in: J. Sander et al. (eds), *From Arithmetic to Zeta-Functions: Number Theory in Memory of Wolfgang Schwarz*, Springer (Cham, 2016) pp. 367–384.
[100] J. Pintz, A note on the distribution of normalized prime gaps, *Acta Arith.*, **184** (2018), 413–418.
[101] S. Poljak, A. Pultr and V. Rödl, On qualitatively independent partitions and related problems, *Discrete Appl. Math.*, **6** (1983), 193–205.
[102] S. Poljak and V. Rödl, Orthogonal partitions and covering of graphs, *Czech Math. J.*, **30** (1980), 475–485.
[103] S. Poljak and Z. Tuza, On the maximum number of qualitatively independent partitions, *J. Comb. Theory A*, **51** (1989), 111–116.
[104] D. H. J. Polymath, Variants of the Selberg sieve, and bounded intervals containing many primes, *Res. Math. Sci.*, **1** (2014), Art. 12, 83 pp.
[105] D. H. J. Polymath, Erratum to: Variants of the Selberg sieve, and bounded intervals containing many primes, *Res. Math. Sci.*, **2** (2015), Art. 15, 2 pp.
[106] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing (Vienna, Austria, 2020), https://www.R-project.org/
[107] S. T. Rachev and L. Rüschendorf, *Mass Transportation Problems: Volume I: Theory*, Springer-Verlag (New York, 1998).
[108] A. Rényi, On the representation of an even number as the sum of a single prime and a single almost-prime number, *Doklady Akad. Nauk SSSR (N.S.)*, **56** (1947), 455–458 (in Russian).
[109] A. Rényi, On the representation of an even number as the sum of a single prime and single almost-prime number, *Izvestiya Akad. Nauk SSSR. Ser. Mat.*, **12** (1948), 57–78 (in Russian).
[110] A. Rényi, Un nouveau théorème concernant les fonctions indépendantes et ses applications à la théorie des nombres, *J. Math. Pures Appl.*, **28** (1949), 137–149.
[111] A. Rényi, On a theorem of Erdős and Turán, *Proc. Amer. Math. Soc.*, **1** (1950), 7–10.
[112] A. Rényi, On the large sieve of Yu. V. Linnik, *Compositio Math.*, **8** (1950), 68–75.
[113] A. Rényi, On measures of dependence, *Acta Math. Acad. Sci. Hungar.*, **10** (1959), 441–451.
[114] A. Rényi, On the distribution of numbers prime to n, in: P. Turán (ed) *Number Theory and Analysis*, Springer (Boston, MA, 1969), pp. 269–278.
[115] A. Rényi, *Foundations of Probability*, Holden-Day, Inc. (San Francisco, CA, 1970).
[116] C. Rényi and A. Rényi, The Prüfer code for k-trees, in: P. Erdős et al. (eds), *Combinatorial Theory and its Applications* (Proc. Colloq., Balatonfüred, Hungary, August 24–29, 1969), Vol. III, North-Holland (Amsterdam, 1970) pp. 945–971.
[117] G. Ricci, Recherches sur l'allure de la suite $(p_{n+1} - p_n)/\log p_n$, in: *Coll. Th. Nombres Bruxelles 1955*, G. Thone (Liège, 1956) pp. 93–106.
[118] T. Rudas, *Lectures on Categorical Data Analysis*, Springer (New York, 2018).
[119] F. Schmid, R. Schmidt, T. Blumentritt, S. Gaißer and M. Ruppert, Copula-Based Measures of Multivariate Association, in: P. Jaworski et al. (eds), *Copula Theory and Its Applications*, Lecture Notes in Statistics, vol. 198, Springer (Berlin, Heidelberg, 2010) pp. 209–236.
[120] I. J. Schoenberg, On certain metric spaces arising from Euclidean spaces by a change of metric and their imbedding in Hilbert space, *Ann. of Math. (2)*, **38** (1937), 787–793.

[121] I. J. Schoenberg, Metric spaces and positive definite functions, *Trans. Amer. Math. Soc.*, **44** (1938), 522–536.
[122] J. Schönheim, On a problem of Purdy related to Sperner systems, *Canad. Math. Bull.*, **17** (1974), 135–136.
[123] D. Sejdinovic, B. Sriperumbudur, A. Gretton and K. Fukumiyu, Equivalence of Distance-Based and RKHS-based Statistics in Hypothesis Testing, *Ann. Stat.*, **41** (2013), 2263–2291.
[124] G. J. Székely, M. L. Rizzo and N. K. Bakirov, Measuring and testing independence by correlation of distances, *Ann. Stat.*, **35** (2007), 2769–2794.
[125] G. J. Székely and M. L. Rizzo, Brownian Distance Covariance, *Ann. Appl. Stat.*, **3** (2009), 1236–1265.
[126] G. J. Székely and M. L. Rizzo, The distance correlation t-test of independence in high dimension, *J. Multivariate Anal.*, **117** (2013), 193-–213.
[127] G. J. Székely and M. L. Rizzo, Partial distance correlation with methods for dissimilarities, *Ann. Stat.*, **42** (2014), 2382–2412.
[128] N. Tomizawa, On some techniques useful for solution of transportation network problems, *Networks*, **1** (1971), 173–194.
[129] C. J. de la Vallée-Poussin, Recherches analytique sur la théorie des nombres premiers, I–III, *Ann. Soc. Sci. Bruxelles*, **20** (1896), 183–256, 281–362, 363–397.
[130] C. Villani, *Optimal Transport: Old and New*, Springer-Verlag (Berlin, Heidelberg, 2009).
[131] A. I. Vinogradov, The density hypothesis for Dirichet L-series, *Izv. Akad. Nauk SSSR Ser. Mat.*, **29** (1965), 903–934 (in Russian).
[132] A. I. Vinogradov, Correction to the paper of A. I. Vinogradov "On the density hypothesis for the Dirichlet L-series", *Izv. Akad. Nauk SSSR Ser. Mat.*, **30** (1966), 719–720 (in Russian).
[133] L. N. Wasserstein, Markov processes over denumerable products of spaces describing large systems of automata, *Probl. Inf. Transm.*, **5** (1969), 47–52.
[134] E. Westzynthius, Über die Verteilung der Zahlen die zu den n ersten Primzahlen teilerfremd sind, *Comment. Phys.-Math. Helsingsfors*, **5** (1931), 1–37.
[135] S. Yitzhaki, Gini's mean difference: a superior measure of variability for non-normal distributions, *Metron*, **61** (2003), 285–316.
[136] Y. Zhang, Bounded gaps between primes, *Ann. Math.*, **179** (2014), 1121–1174.