OPEN ACCESS Freely available online

PLOS ONE

A Novel Test for Independence Derived from an Exact Distribution of *i*th Nearest Neighbours

Sebastian Dümcke[1,2], Ulrich Mansmann[3], Achim Tresch[1,2]*

1 Institute for Genetics, Universität zu Köln, Cologne, Germany, **2** Max Planck Institute for Plant Breeding Research, Cologne, Germany, **3** IBE, Ludwig-Maximilians-Universität, Munich, Germany

Abstract

Dependence measures and tests for independence have recently attracted a lot of attention, because they are the cornerstone of algorithms for network inference in probabilistic graphical models. Pearson's product moment correlation coefficient is still by far the most widely used statistic yet it is largely constrained to detecting linear relationships. In this work we provide an exact formula for the *i*th nearest neighbor distance distribution of rank-transformed data. Based on that, we propose two novel tests for independence. An implementation of these tests, together with a general benchmark framework for independence testing, are freely available as a CRAN software package (http://cran.r-project.org/web/packages/knnIndep). In this paper we have benchmarked Pearson's correlation, Hoeffding's D, dcor, Kraskov's estimator for mutual information, maximal information criterion and our two tests. We conclude that no particular method is generally superior to all other methods. However, dcor and Hoeffding's D are the most powerful tests for many different types of dependence.

Citation: Dümcke S, Mansmann U, Tresch A (2014) A Novel Test for Independence Derived from an Exact Distribution of *i*th Nearest Neighbours. PLoS ONE 9(10): e107955. doi:10.1371/journal.pone.0107955

Editor: Holger Fröhlich, University of Bonn, Bonn-Aachen International Center for IT, Germany

Received April 22, 2014; **Accepted** August 18, 2014; **Published** October 2, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. The code required to run the tests and benchmarks is available as an R package 'knnIndep' on CRAN (http://cran.r-project.org/web/packages/knnIndep/). The data for the WHO dataset can be freely downloaded from http://www.exploredata.net/ftp/WHO.csv and the code to generate Fig. 4 is provided in the Supporting Information file.

Funding: AT was supported by the Bundesministerium für Bildung und Forschung (BMBF) e: Bio Syscore grant and by a Jeff Schell professorship from the Max Planck Institute for Plant Breeding Research and the University of Cologne. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: tresch@mpipz.mpg.de

Introduction

Dependence measures and tests for independence have recently attracted a lot of attention, because they are the cornerstone of algorithms for network inference in probabilistic graphical models. Pearson's product moment correlation coefficient is still by far the most widely used statistic in areas such as economy, biology and the social sciences. Yet Pearson's correlation is largely constrained to detecting linear relationships. Spearman [1] and Kendall [2] extended Pearson's work to monotonic dependencies. In 1948, Hoeffding [3] proposed a non parametric test for independence that is suited for many different functional relationships. Székely et al. [4] introduced the distance correlation (dcor) as a generalization of Pearson's correlation.

Other approaches build on mutual information (MI). MI characterizes independence in the sense that the MI of a joint distribution of two variables is zero if and only if these variables are independent. However, MI is difficult to estimate from finite samples. Kraskov et al. [5] proposed an accurate MI estimator derived from nearest neighbor distances. Reshef et al. [6] presented the maximal information coefficient (MIC), a measure of dependence for two-variable relationships which was heavily advertised [7] but lacks any statistical motivation.

dcor and Kraskov's estimator use the pair-wise distances of the points in a sample as a sufficient statistic. In this work we provide an exact formula for the *i*th nearest neighbor distance distribution of rank-transformed data ($i=1,2,...$). Based on that, we propose two novel tests for independence. An implementation of these tests, together with a general benchmark framework for independence testing, are freely available as a CRAN software package (http://cran.r-project.org/web/packages/knnIndep). In this paper we have benchmarked Pearson's correlation, Hoeffding's D, dcor, Kraskov's estimator for MI, MIC and our two tests. We conclude that no particular method is generally superior to all other methods. However, dcor and Hoeffding's D are the most powerful tests for many different types of dependence. Circular dependencies are best recognized by our tests. This type of dependence is fairly common, e.g., if two dependent periodic processes are monitored. An example from biology is the expression of a transcription factor and one of its target genes during the cell cycle [8].

Exact distribution of the *i*th nearest neighbour distances

Consider a set of $N \geq 4$ points that are distributed 'randomly' on a surface. In what follows, we derive the distribution (conditional distribution) of the $(i+1)$th nearest neighbor of a point (given the distance to its previous neighbors). We assume the points drawn from the following model: Let $X=(x_j)_{j=1,...,N}$ and $Y=(y_j)_{j=1,...,N}$ be permutations of the numbers $0,...,N-1$ that are drawn uniformly from the set of all permutations of $\{0,...,N-1\}$. The

points $z_j=(x_j,y_j)$, $j=1,...,N$, lie on a torus of size N which is endowed with the maximum distance as a metric. I.e., the distance between two points is given by

$$dist(z_1,z_2)=\max(\min(|x_1-x_2|,N-|x_1-x_2|),\ \min(|y_1-y_2|,N-|y_1-y_2|))$$

Fix a reference point z_1. Let d_i, $i=1,...,N-1$ denote the distance of the i-th nearest neighbor of z_1 to z_1 and D_i the random variable associated with it. Since this distance measure is translation invariant, let without loss $z_1=(x_1,y_1)=(0,0)$. Importantly, all points z_j have pairwise different x_j and y_j. A point at distance $d=dist(z,(0,0))$ to the origin must have at least one of its coordinates equal to d or $N-d$. This implies that there are at most 4 points exactly at distance d to the origin. Our target is the calculation of the joint probability of observing the whole sequence of nearest neighbor distances $P(D_0, D_1, \ldots, D_{N-1})$, of the conditional probability $P(D_{i+1}|\ D_i,\ldots,\ D_0)$ and the marginal $P(D_i)$. The main work will be the calculation of the probability $P(D_{i+1}\geq c,\ D_i=a,\ldots,\ D_{i-k+1}=a,\ D_{i-k}<a)$ for given values k, a and c. Once this is done, $P(D_0, D_1, \ldots, D_{N-1})$, $P(D_i)$ and $P(D_{i+1}|\ D_i,\ldots,\ D_0)$ can be derived by elementary calculations (section S1 in Methods S1).

First we determine $P(D_{i+1}\geq c,\ D_i\leq a)$ by counting the number of admissible point configurations and dividing through $(N-1)!$, the number of all possible point configurations with $z_1=(0,0)$ fixed. When counting configurations, we repeatedly exploit the fact that each horizontal and each vertical grid line contains exactly one point from the sample. In case of $c>a$, we split the torus into 3 regions (Figure 1). Region I is a square of side length $2a+1$. It contains z_1 and i additional points at arbitrary positions. The number of possibilities to draw an i-tuple from $2a$ positions (recall that one position is already taken by z_1) without replacement is $\frac{(2a)!}{(2a-i)!}$. Thus, there are $\left(\frac{(2a)!}{(2a-i)!}\right)^2$ i-tuples describing an admissible configuration in region I. However, each configuration is counted $i!$ times, since the order of the points does not matter. Hence, the number of unique configurations in region I equals $\frac{1}{i!}\left(\frac{(2a)!}{(2a-i)!}\right)^2=\binom{2a}{i}^2 i!$. For the second region we have $N-2c+1$ possible y-coordinates and $2c-1-(i+1)=2c-i-2$ columns to be filled with sample points (note that the columns $-c$ and c belong to region III and that $i+1$ columns are already taken by points in region I). This yields $\frac{(N-2c+1)!}{(N-4c-i+3)!}$ unique configurations for region II. There are $N-2c+1$ points remaining which can be placed freely in the remaining $N-2c$ columns/rows, yielding $(N-2c+1)!$ possibilities. Together we obtain:

$$P(D_{i+1}\geq c,\ D_i\leq a)=\frac{1}{(N-1)!}\cdot\underbrace{\binom{2a}{i}^2 i!}_{\text{region I}}\cdot\underbrace{\frac{(N-2c+1)!}{(N-4c+i+3)!}}_{\text{region II}}\cdot\underbrace{(N-2c+1)!}_{\text{region III}}\quad N\geq 4 \tag{1}$$

In the case of $c=a$ there is one more complication, because we have a region R of points exactly at distance c, containing at least the i-th and $(i+1)$-th neighbor of z_1, where the region I overlaps with regions IIa and IIb (Figure 1). Let $r\in\{2,3,4\}$ be the number of points in region R and i_0 the number of points strictly inside the square of distance c. We derive a general formula for all admissible configurations in the case of $c=a$, $P(D_{i_0+r+1}>c,\ D_{i_0+r}=\cdots=D_{i_0+1}=c,\ D_{i_0}<c)$. Denote by $k(r,i_0,c)$ the number of admissible point configurations in region R (see section S2 in Methods S1 for a derivation of $k(r,i_0,c)$). Table 1 lists all possible admissible combinations of points in region R. Counting the admissible configurations strictly inside regions I, IIa, IIb and III is similar to the above cases (Equation 1). This leads to the following general formula for all admissible configurations:

$$P(D_{i_0+r+1}>c,\ D_{i_0+r}=\cdots=D_{i_0+1}=c,\ D_{i_0}<c)=\underbrace{\binom{2c-2}{i_0}^2 i_0!}_{\text{region I}}\cdot\underbrace{k(r,i_0,c)}_{\text{region R}}\cdot\left[\underbrace{\frac{(N-2c-1)!}{(N-4c+i_0+r-1)!}}_{\text{region IIa+IIb}}\right]^2\cdot\underbrace{(N+i_0+r-4c-1)!}_{\text{region III}} \tag{2}$$

The sum over all possible tuple (r,i_0) in Table 1 gives the probability $P(d_{i+1}=c,\ d_i=c)$ in the general case:

$$P(\mathrm{R})=\begin{cases}0 & \text{if } i_0>N-r\\ \sum_{(r,i_0)}\frac{1}{(N-1)!}P(D_{i_0+r+1}>c,\ D_{i_0+r}=\cdots=D_{i_0+1}=c,\ D_{i_0}<c) & \text{else}\end{cases} \tag{3}$$

The above calculations only hold if region R is a genuine square, for large values of c R degenerates to a pair of lines (one horizontal and one vertical line). These cases are covered in the extended formula

$$P(D_{i+1}=c,\ D_i=c)=\begin{cases} i=1 & \begin{cases} c=1: & P(d_3\geq 2, d_2\leq 1)\\ c>1: & P(\mathrm{R}),\text{Equation(3)}\end{cases}\\ 1<i<N-2 & \begin{cases} 1<c\leq\lfloor\frac{N}{2}\rfloor: & P(\mathrm{R}),\ \text{Equation(3)}\\ \text{else}: & 0\end{cases}\\ i=N-2 & \begin{cases} c=\frac{N}{2},\ N\ \text{even}: & \binom{N-2}{i-1}(i-1)!\\ \text{else}: & P(\mathrm{R}),\ \text{Equation(3)}\end{cases}\\ i=N-1 & \begin{cases} c=\lfloor\frac{N}{2}\rfloor: & 1\\ \text{else: } 0\end{cases}\end{cases}. \tag{4}$$

Analogously we can count the number of possible configurations where $D_{i+1}>c$, some k points $D_i,\ldots,D_{i-k+1}=a$ and all other points $D_{i-k}<a$ and deduce the following probability:

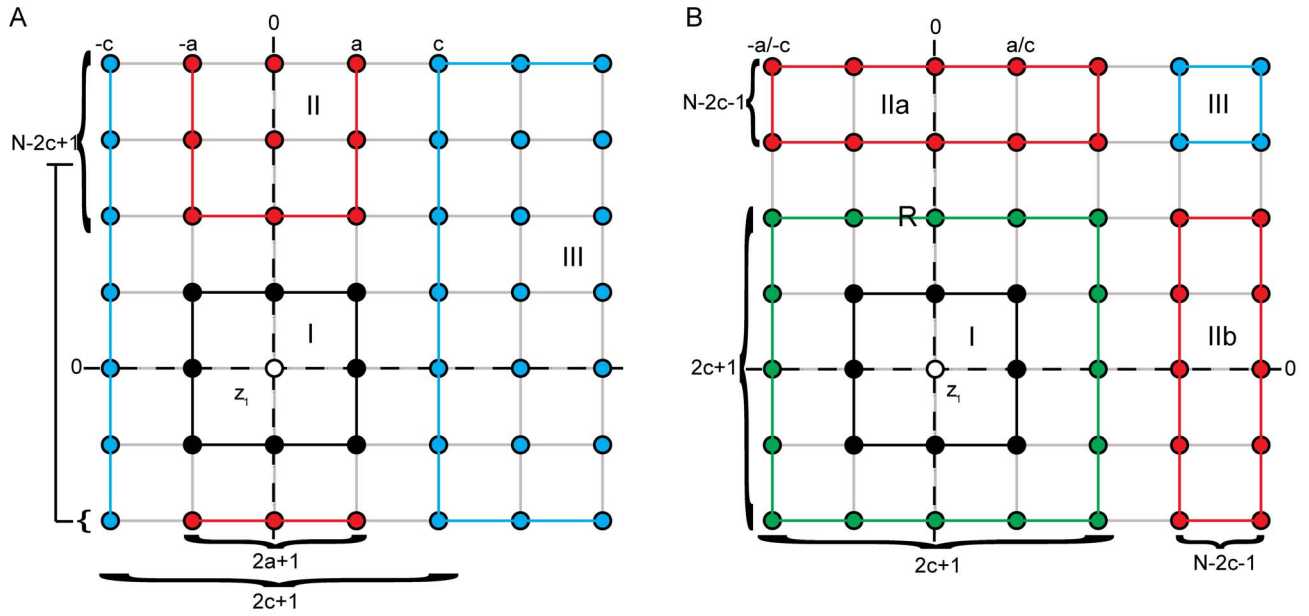

Figure 1. Diagrams explaining Equations 1 and 2 for $N=7$, $a=1$ and $c=2$ (panel A) and $a=c=2$ (panel B) with the reference point z_1 at coordinates $(0,0)$**. A**: We define 3 regions I, II and III (black, red and blue points respectively). Region I has the least number of constraints and the number of admissible configurations is the number of possibilities to draw i points from $2a$ positions without replacement nor ordering: $\binom{2a}{i}^2 i!$. The number of admissible configurations for region II is given by the number of rows $n_r=N-2c+1$ available and the number of columns which remain to be filled $n_c=2c-i-2$ according to $\frac{n_r!}{(n_r-n_c)!}$. Region III has the remaining $N-2c+1$ points freely distributed, yielding $(N-2c+1)!$ admissible configurations. **B**: In the case $a=c$ we add an additional region R of r points exactly at distance c (green points). There can be $r=2$, 3 or 4 such points. Region I has size $(2(c-1))^2$ and $\binom{2c-2}{i_0}^2 i_0!$ admissible configurations with i_0 the number of points strictly inside the square of distance c. Region IIa and IIb are symmetric and handled analogous to region II in panel A with $n_r=N-2c-1$ and $n_c=2c-i_0-r$. Region III has $(N+i_0+r-4c-1)!$ admissible configurations analogous to panel A.
doi:10.1371/journal.pone.0107955.g001

$$P(D_{i+1}\geq c,\ D_i=a,\ldots,D_{i-k+1}=a,\ D_{i-k}<a) = \frac{1}{(N-1)!}\underbrace{\binom{2a-2}{i-k}^2 (i-k)!}_{\text{region I}}\cdot\underbrace{k(k,i-k,a)}_{\text{region R}}\cdot\underbrace{\frac{(N-2c+1)!}{(N-4c+i+3)!}}_{\text{region II}}\cdot\underbrace{(N-2c+1)!}_{\text{region III}} \quad (5)$$

Since the above formulas involve tedious calculations, we validated the formulas for $N=7$ and $N=8$ by counting the occurrence of each possible configuration among all $N!$ configurations. Additionally, we checked the validity of our formula for larger N $(N=20)$ by taking 10^6 random configurations and comparing the empirical frequency $h(d_i)$ with $P(d_i)$ (section S3 in Methods S1).

Figure 2 shows the distribution of $P(d_{i+1}|d_i)$ and $P(d_i)$. The conditional distribution is shown for $i=50$. The marginal distribution is highly peaked with a low variance that decreases with increasing i (and reaches 0 for $i=N$).

The formulas have been implemented in the statistical language R [9] with emphasis on a numerically stable implementation as we deal with small numbers. The implementation is vectorized for speed. Still there is a computational penalty through the many factorials and logarithms that have to be calculated. For a sample of size 320, calculating all $P(d_{i+1}|d_i)$ takes 4.1 seconds on a single workstation (single thread, Intel Core i5-2500 CPU @ 3.30GHz). Runtime for larger samples is shown in Figure S1 in File S1 and

Table 1. Counts for points lying exactly on the border region R.

r	i_0	$k(r,i_0,c)$; **let** $\epsilon=2c-2-i_0$	**condition**
1	$i-1$	$4\epsilon+4$	if $i_0<N-r$
2	$i-1$	$2\epsilon(\epsilon-1)+4\epsilon^2+8\epsilon+2=6\epsilon^2+6\epsilon+2$	if $i_0<N-r$
3	$i-1,i-2$	$4\epsilon^2(\epsilon-1)+4\epsilon^2=4\epsilon^3$	if $i_0<N-r$
4	$i-1,i-2,i-3$	$\epsilon^2(\epsilon-1)^2$	if $i_0<N-r$

For each possible number of points $r=2,3,4$ on the border region R and each possible number of points i_0 strictly inside of region I, we give the the number of admissible combinations of points in region R. The derivations of the number of admissible combinations is shown in the section S2 in Methods S1.
doi:10.1371/journal.pone.0107955.t001

indicates a practical limit on the sample size of $N<3000$ (which takes up to 3 minutes) and a complexity of $O(N^2)$.

For practical reasons, we assumed that the points lie on a torus (distances on the torus are translation-invariant and therefore our formulas for $P(d_{i+1}|d_i)$ and $P(d_i)$ hold for all points in the sample). This will bias results when applied to points on a plane, as points on the border will have different nearest neighbors when projected on the torus. The bias is less pronounced for close neighbors (i small), thus we limit our statistics to $i_{max}=N/2$. We do not expect to lose statistical power, since the information content of $P(d_i)$ for large i approaches zero (see Figure 2).

The derivation of $P(D_0, D_1, \ldots, D_{N-1})$, $P(D_i)$ and $P(D_{i+1}|D_i, \ldots, D_0)$ is based on Equations (1–5), see section S1 in Methods S1.

Tests based on the *i*th nearest neighbour distribution

It has been shown that the distance of the ith nearest neighbour of some point z can be used to estimate the local (log) density at z [5]. Our idea is to use the full sequence of nearest neighbour distances for assessing local density. For a sample point z, let $(D_0=d_0^z=0, D_1=d_1^z, D_2=d_2^z, \ldots, D_{N-1}=d_{N-1}^z)$ the sequence of neighbour distances. If z lies in a dense region, we expect this sequence to increase slower than in a region with lower density.

Distributional tests

The sequence of nearest neighbor distances of a point z, $(D_0=d_0^z=0, D_1=d_1^z, D_2=d_2^z, \ldots, D_{N-1}=d_{N-1}^z)$ is a 4th order Markov chain, i.e.,

$$P(d_0^z, d_1^z, d_2^z, \ldots, d_{N-1}^z)=\prod_{i=0}^{N-2} P(d_{i+1}^z|d_i^z, d_{i-1}^z, d_{i-2}^z, d_{i-3}^z)$$

That way, taking z as the center point, the distances d_{i+1}^z, given the four previously observed distances $(d_i^z, d_{i-1}^z, d_{i-2}^z, d_{i-3}^z)$, are pairwise independent for all i. On the other hand this is not true for the distances $d_{i+1}^{z_1}$ and $d_{i+1}^{z_2}$(not even if we condition the four previously observed distances). This follows from the triangle inequality in metric spaces, $dist(z_1,x)\le dist(z_2,x)+dist(z_1,z_2)$, which implies that $d_{i+1}^{z_1}\le d_{i+1}^{z_2}+dist(z_1,z_2)$.

Let the random variable C_i be defined by the process of drawing a point Z uniformly from 1,...,N and then drawing C_i according to the distribution $P(D_i|D_{i-1}=d_{i-1}^z, \ldots, D_0=d_0^z)$. Let f_i denote the probability function of C_i, it is given by

$$\begin{aligned} f_i(c) &= P(C_i=c) \\ &= \sum_{z=1}^{N} P(C_i=c|Z=z)\cdot P(Z=z) \\ &= \sum_{z=1}^{N} P(D_i|D_{i-1}=d_{i-1}^z, \ldots, D_0=d_0^z)\cdot P(Z=z) \\ &= \frac{1}{N}\sum_{z=1}^{N} P(D_i|D_{i-1}=d_{i-1}^z, \ldots, D_0=d_0^z) \\ &\approx \frac{1}{N}\sum_{z=1}^{N} P(D_i|D_{i-1}=d_{i-1}^z) \end{aligned}$$

We consider the observed values d_i^z, $z=1,...,N$, as (not necessarily independent) realizations of D_i. Their empirical frequency e_i is

$$e_i(c)=\frac{1}{N}\sum_{z=1}^{N}\mathbf{I}[d_i^z=c]$$

where $\mathbf{I}[.]$ denotes the indicator function with values in $\{0,1\}$. Pearson's χ^2 test [10] can be used to test for the fit of f_i to e_i:

$$X_i=\sum_{c=1}^{\lfloor\frac{N}{2}\rfloor}\frac{(e_i(c)-f_i(c))^2}{f_i(c)}\sim\chi^2_{\phi_i-1}$$

X_i is a χ^2-distributed test statistic with ϕ_i-1 degrees of freedom where ϕ_i is the number distances c with $f_i(c)$ strictly positive. Our final test statistic is:

$$\sum_{i}^{N-1} X_i\sim\chi^2_{\sum_i^{N-1}(\phi_i-1)}$$

Alternatively the empirical and theoretical cumulative distributions corresponding to e_i and f_i can be compared by an Anderson-Darling [11] or a Cramér-von Mises test, which proved inferior to Pearson's χ^2 test (section S4 in Methods S1).

Test for location

We have the idea to compare the distribution of the ith neighbour distances observed in a sample with a suitable null distribution by means of their location. The most robust measures of location are mean or median, however in our studies of samples taken from joint distributions with low mutual information, we realized that many points do not show exceptional nearest neighbour distances. The difference to a sample drawn from independent X and Y distributions was made up by few points that had extreme nearest neighbour distances. This lead us to use extreme values as a test for location. The pvalue of a two-sided test based on $P(D_i^z|d_{i-1}^z,\ldots)$ is $p_i^z=2\min(v_i^z, 1-v_i^z)$, with $v_i^z=P(D_i^z\le d_i^z|d_{i-1}^z,\ldots)$. We summarize, for all ith neighbours, the 2-sided pvalues by their minimum

$$V_i=\min(p_i^z;\ z=1,...,N)$$

Our test statistic V is obtained by aggregating the V_i values: $V=-2\sum_{i=1}^{N-1}\ln V_i$.

Construction of a benchmark set

Benchmarking was done on distributions (X,Y) given by $X\sim U[0,1]$, and $Y\sim f(X)+\mathcal{N}(0,\sigma^2)$. Here, $U[0,1]$ denotes a uniform distribution on the interval [0,1], and $\mathcal{N}(0,\sigma^2)$ denotes a Gaussian distribution with mean 0 and variance σ^2. The function f was chosen as one of the following: linear, quadratic, cubic, sine with period 0.5, circular, $f(x)=x^{1/4}$ and a step function (see Figure S2 in File S1). This choice was inspired by a comment by Simon & Tibshirani (http://statweb.stanford.edu/tibs/reshef/script.R, [12]) to the publication of the method MIC by Reshef et al. [6]. The noise parameter σ^2 determines the degree of dependence between X and Y, i.e., the mutual information

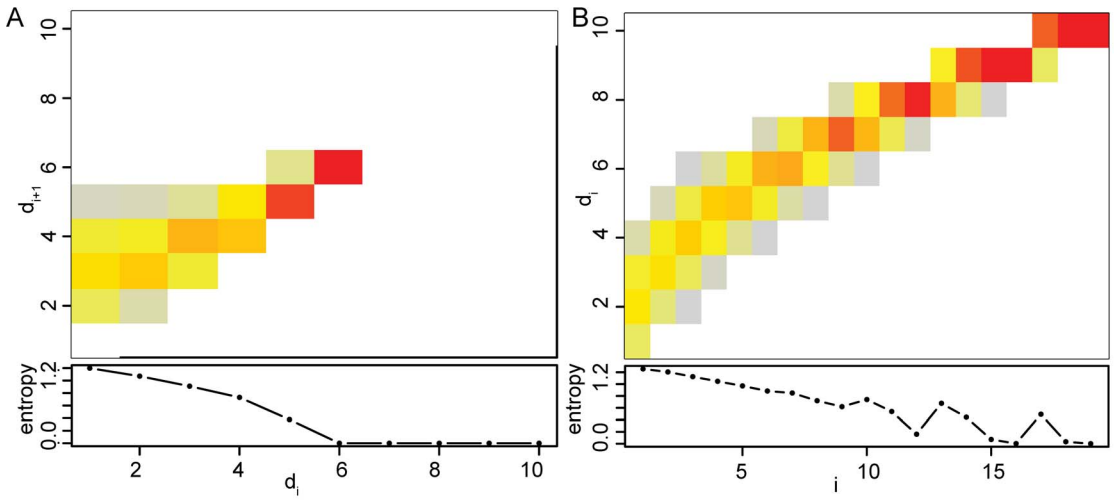

Figure 2. A: Conditional distribution $p_c = P(D_{i+1} = d_{i+1} | D_i = d_i)$ for $i = 2$, $N = 21$ (*top*) and the entropy $-\sum_{d_{i+1}=1}^{\lfloor \frac{N}{2} \rfloor} p_c \log p_c$ (*bottom*). The probability p_c of observing large (d_{i+1}, d_i) is zero for distances larger than (6,6) when $i = 2$. The lower triangle is empty because $d_{i+1} \geq d_i$ and the entropy is constantly decreasing for increasing values of d_i because the possible (d_{i+1}, d_i) decrease towards (6,6). **B**: Marginal distribution $P(d_i)$ for $N = 21$ (*top*) and entropy $-\sum_{d_i=1}^{\lfloor \frac{N}{2} \rfloor} P(d_i) \log P(d_i)$ (*bottom*). With increasing i, the distribution becomes narrower and the entropy tends towards 0, as the number of possible distances to the ith nearest neighbor decrease. The non-monotonic behavior of the entropy for large values of i is due to downstream constraints imposed by the maximal distance $\frac{N}{2}$. For testing independence, we advise using all $P(D_{i+1} | D_i)$ until the value of i where the entropy starts increasing again ($i = 9$ in this example).
doi:10.1371/journal.pone.0107955.g002

$MI(X, Y; f, \sigma^2)$. The latter was estimated using an approximation $q_{XY}(X, Y)$ to the density $p(X, Y)$ for which the mutual information can easily be calculated. We make q_{XY} a piecewise-constant density on a sufficiently fine quadratic grid $\{(\epsilon x, \epsilon y) | x, y \in \mathbb{Z}\}$ with $q_{XY}(x, y) = p(\epsilon \lfloor \frac{x}{\epsilon} + 0.5 \rfloor, \epsilon \lfloor \frac{y}{\epsilon} + 0.5 \rfloor)$. In our case, $\epsilon = 0.01$ yielded sufficient precision. It is elementary to calculate the mutual information of q by

$$MI = \epsilon^2 \cdot \sum_{x,y \in \mathbb{Z}} q_{XY}(\epsilon x, \epsilon y) \cdot \log \frac{q_{XY}(\epsilon x, \epsilon y)}{q_X(\epsilon x) q_Y(\epsilon y)}$$

Here, q_X and q_Y denote the marginal densities with respect to x and y.

To make the results comparable for different f, we fixed an MI value M and chose $\sigma^2_{f,M}$ such that $MI(X, Y; f, \sigma^2_{f,M}) = M$. This was done for 20 MI values, M ranging from 0.01 to 0.5. The noise levels $\sigma^2_{f,M}$ are listed in section S5 in Methods S1. Samples from all dependencies f with $M = 0.5$ is shown in Figure S2 in File S1.

So far performance evaluation of measure of dependence was only done on functional dependencies. Here we introduce "patchwork copulas" as a new non-functional dependence of x and y. Fix a grid size B, say $B = 10$. Our density q will be a piecewise constant function defined on a rectangular 2D grid on the unit square (with uneven grid line spacing) such that its marginal distributions are uniform (i.e., we will define a copula). The parameters of our distribution are the values p_{ij}, $i, j = 1, \ldots, B$, with $\sum_{i,j=1}^{B} p_{ij} = 1$. Let $p_{i*} = \sum_{j=1}^{B} p_{ij}$ and $p_{*j} = \sum_{i=1}^{B} p_{ij}$. Let (I, J) be a random variable which selects the grid rectangle (i, j) with probability p_{ij}, i.e., $P((I, J) = (i, j)) = p_{ij}$, $i, j = 1, \ldots, B$. Our distribution (X, Y) is then defined by $X \sim \sum_{i=1}^{I-1} p_{i*} + U_I$, $U_I \sim U[0, p_I]$, and $Y \sim \sum_{j=1}^{J-1} p_{*j} + V_J$, $V_J \sim U[0, p_J]$. The density in the grid rectangle (i, j) can be computed as $q_{ij} = \frac{p_{ij}}{p_{i*} p_{*j}}$. It is elementary to verify that the marginals of q are uniform and that the mutual information of (X, Y) is

$$MI(X, Y; (p_{ij})) = \sum_{i,j=1}^{B} p_{ij} \log \left(\frac{p_{ij}}{p_{i*} p_{*j}} \right)$$

To generate samples with a desired MI value, we choose suitable values for α and β. We draw i.i.d. samples $p_{ij} \sim \mathrm{Beta}(\alpha, \beta)$, $i, j = 1, \ldots, B$, and then rescale the p_{ij} by dividing them by their sum. This process is repeated with different α, β until $MI(X, Y; (p_{ij}))$ is close enough to the desired MI value. The resulting dependence resembles a patchwork quilt of dense and spread out point clouds (Figure S3 in File S1).

Typically the points are considered embedded in Euclidean spaces [5], however the distance function can easily be adapted to model the geometry of a torus. We benchmarked some methods on both geometries (Euclidean plane and torus) and found that all methods were sensitive to changes of geometry.

We made the benchmark framework publicly available under a GPL3.0+ license. It is implemented in R [9] and contains code for generating the dependence structures as well as plotting the results. An example is given in section S6 in Methods S1.

Comparison of methods

We compared both our tests (based on χ^2 and extreme paths) to Pearson's product moment correlation coefficient, distance correlation (dcor, [4]), Hoeffding's D [3], Kraskov's estimator for mutual information [5] and MIC [6]. For each type of dependence and each given value of MI, we generated a test set of 500 samples each consisting of 320 points from the respective dependence type. Test statistics were calculated for each sample. Additionally we generated a reference set of 500 samples with x and y values drawn independently which is used to calculate the cutoff value corresponding to a significance level of 5%. The power of each

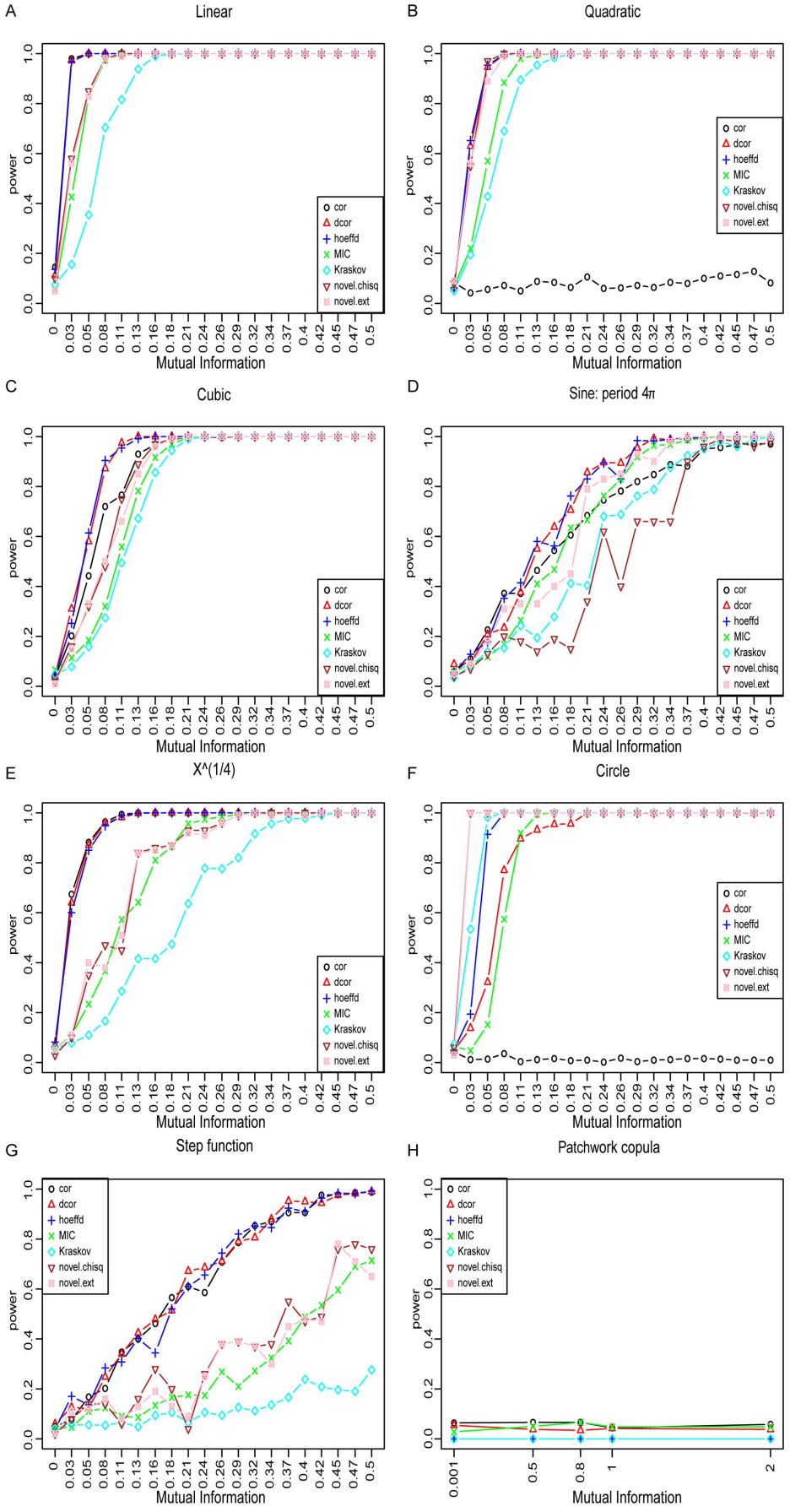
A
Linear
B
Quadratic
C
Cubic
D
Sine: period 4π
E
X^(1/4)
F
Circle
G
Step function
H
Patchwork copula
power
Mutual Information
cor
dcor
hoeffd
MIC
Kraskov
novel.chisq
novel.ext

Figure 3. Benchmark of all methods. cor denotes Pearson's product moment correlation coefficient, dcor distance covariance, hoeffd Hoeffding's D, MIC denotes MIC, novelTest.chisq is our test based on Pearson's χ^2 test and novelTest.ext is our test based on extreme paths. Each plot shows the power (on the y-axis) against the MI (x-axis). We examine 8 different types of dependence: linear, quadratic, cubic, sine with period 4π, $x^{1/4}$, circle, step function and the dependence called "patchwork copula" (**A–H**)
doi:10.1371/journal.pone.0107955.g003

method was estimated as the fraction of samples that were called significant according to the cutoff. Results are shown in Figure 3. Additionally we generated receiver operating curves (ROC) for each type of dependence and MI value (Figures S4–S9 in File S1).

The method of Hoeffding and dcor perform well throughout all types of dependence considered except for the circular dependence. Our methods have a performance that places them after dcor and Hoeffding's method and before MIC. In the case of the circular dependence, our methods perform best, achieving maximum power at mutual information of 0.03. We suspect that is due to the fact that a circle geometrically resembles two crossing lines when projected onto a torus (Figure S11 in File S1). To test this hypothesis we projected all types of dependence onto the torus and reran the whole benchmark (Figure S10 in File S1). We observe that the cubic, sine and step functions are not detected by any method, even at the same MI.

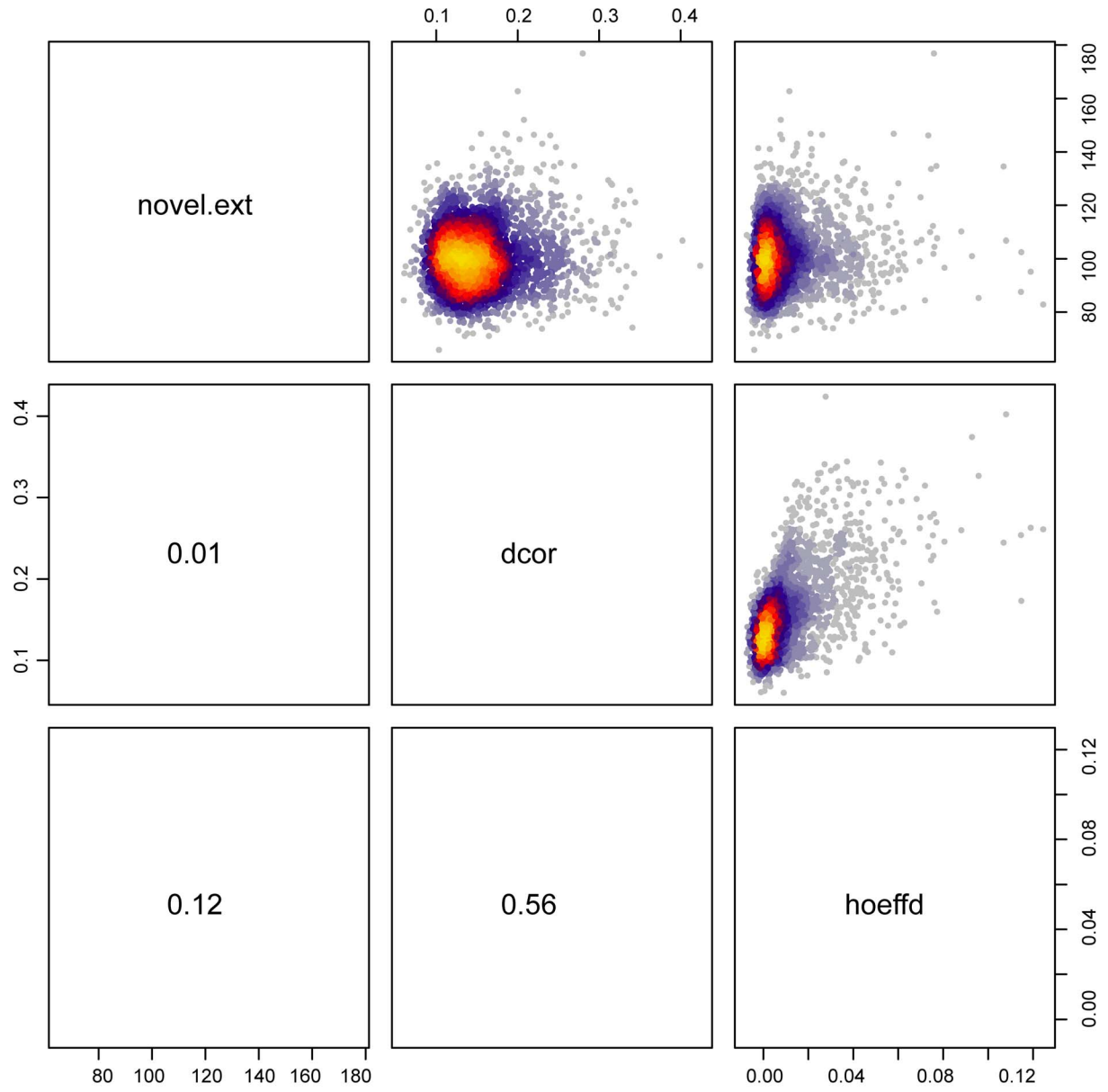

Figure 4. Performance on WHO data. novelTest.ext denotes our test based on extreme paths, dcor distance covariance and hoeffd Hoeffding's D. All methods were applied to all comparison between pairwise variables which had Pearson's product moment correlation coefficient near zero to exclude linear relationships. Only pairwise complete observations were used as most methods cannot handle missing vallues. All comparisons include ate least 81 datapoints. In total we compare all 3 methods on 2971 variable pairs.
doi:10.1371/journal.pone.0107955.g004

The scaling of the plots in Figure 3 to the MI of the underlying joint distribution, enables the direct visual comparison of different dependence types. On the one hand this reveals that some types of dependence seem to be more difficult to detect for all methods (step function, sine curve and the "patchwork copula"). On the other hand each method performs best on different types of dependence.

We compared method of Hoeffding, dcor and our test based on extreme paths on a dataset from the World Health Organization and partner organizations. This dataset is available at http://www.exploredata.net/ftp/WHO.csv. We ran the methods on all pairwise comparisons that have a squared Pearson's product moment correlation coefficient lower than 0.001 to exclude any linear relationships. As most method cannot handle missing values, we further restricted the comparisons to have at least 81 pairwise complete observations. This leads to 2971 pairwise comparison shown in Figure 4. All test statistics are uncorrelated for the pairs in which no linear dependency was detected leading again to the conclusion that no method is uniformly more powerful.

Discussion

We have derived an exact formula for the distribution of the distances of the i th nearest neighbour of a given point. This distribution assumes rank transformed bivariate data from two independent variables. While this result is of independent interest, we used it to construct two non-parametric tests of independence for bivariate data. Similar to Kraskov's estimator, our test statistic is purely based on nearest neighbour distances. In contrast to Kraskov's estimator which requires an arbitrarily fixed i, we simultaneously take into account the whole sequence of ith nearest neighbours ($i=1,2,...$). This improves on Kraskov's estimator, if used as a score for independence testing. Our tests use rank transformed data, because this is a prerequisite for applying the exact nearest neighbour distributions derived in this paper. The rank transformation is often used as a primary step to estimating mutual information, therefore we consider it an uncritical step in our procedure. Our tests perform almost as well as the best competitors dcor and Hoeffding's D and they perform better than the recently proposed MIC statistic. We believe that the power of our method could be further improved in the Euclidean plane if our ith neighbour statistic would be adapted to account for boundary effects in the Euclidean plane. Although our methods try to account for the dependence of the variables D_i^z, $z=1,...,N$, we necessarily lose power because their exact dependence structure is not known. Alternatively we propose to take all distances d_i^z for a point z and apply a sequential testing approach for calling points that are located in dense regions. The number of these points could serve as a test statistic. The rationale is that under the null hypothesis of independence there should be fewer points z considered significant in the sequential test than for dependent samples.

Next we reviewed competing methods and presented a benchmark framework for performance testing on different types of dependence structures and topologies (Euclidean and toroidal). The benchmark framework and our novel tests for independence are publicly available as an R [9] package on CRAN (http://cran.r-project.org/web/packages/knnIndep). By scaling each type of dependence to a common set of mutual information values we allow comparison between all dependence types. Remarkably, when benchmarked on patchwork copulas, all methods fail. This is particularly intriguing for MIC as by design it should detect the grid structure of the data. In the case of the circular dependence, our methods perform best, while the method of Hoeffding and dcor perform well throughout all types of dependence considered. This in turn shows, that all tests we investigated are biased towards the detection of certain types of dependence structures.

Supporting Information

Code S1 R code of the analysis of the WHO dataset.
(PDF)

File S1 Figures supporting results from the main text.
(PDF)

Methods S1 Supporting Methods for the main results.
(PDF)

Acknowledgments

We thank Arijit Das for helpful discussions and Philipp Eser for providing the biological example and data.

Author Contributions

Conceived and designed the experiments: UM AT. Performed the experiments: SD. Analyzed the data: SD. Wrote the paper: SD AT UM.

References

1. Spearman C (1904) The proof and measurement of association between two things. The American Journal of Psychology 15: 72–101.
2. Kendall MG (1938) A new measure of rank correlation. Biometrika 30: 81–93.
3. Hoeffding W (1948) A non-parametric test of independence. The Annals of Mathematical Statistics 19: 546–557.
4. Székely GJ, Rizzo ML, Bakirov NK (2007) Measuring and testing dependence by correlation of distances. The Annals of Statistics 35: 2769–2794.
5. Kraskov A, Stögbauer H, Grassberger P (2004) Estimating mutual information. Phys Rev E 69: 066138.
6. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, et al. (2011) Detecting novel associations in large data sets. Science 334: 1518–1524.
7. Speed T (2011) A correlation for the 21st century. Science 334: 1502–1503.
8. Eser P, Demel C, Maier KC, Schwalb B, Pirkl N, et al. (2014) Periodic mrna synthesis and degradation co-operate during cell cycle gene expression. Molecular Systems Biology 10: n/a–n/a.
9. R Core Team (2013) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
10. Pearson K (1900) X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine Series 5 50: 157–175.
11. Anderson TW, Darling DA (1952) Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. The Annals of Mathematical Statistics 23: 193–212.
12. Simon N, Tibshirani R (2014) Comment on" detecting novel associations in large data sets" by reshef et al., science dec 16, 2011. arXiv preprint arXiv:14017645.