entropy

MDPI

Article

A Class of Association Measures for Categorical Variables Based on Weighted Minkowski Distance

Qingyang Zhang

Department of Mathematical Sciences, University of Arkansas, Fayetteville, AR 72701, USA; qz008@uark.edu

Received: 16 August 2019; Accepted: 10 October 2019; Published: 11 October 2019

Abstract: Measuring and testing association between categorical variables is one of the long-standing problems in multivariate statistics. In this paper, I define a broad class of association measures for categorical variables based on weighted Minkowski distance. The proposed framework subsumes some important measures including Cramér's V, distance covariance, total variation distance and a slightly modified mean variance index. In addition, I establish the strong consistency of the defined measures for testing independence in two-way contingency tables, and derive the scaled forms of unweighted measures.

Keywords: dependence measure; categorical variable; Minkowski distance; sparse contingency table; total variation distance; mean variance index

1. Introduction

Measuring and testing the association between categorical variables from observed data is one of the long-standing problems in multivariate statistics. The observed frequencies of two categorical variables are often displayed in a two-way contingency table, and a multinomial distribution can be used to model the cell counts. To be specific, let X and Y be two categorical random variables with finite sampling spaces $\mathcal{X}$ and $\mathcal{Y}$ ($|\mathcal{X}| < \infty, |\mathcal{Y}| < \infty$, where $|\cdot|$ stands for the cardinality of a set), and a simple random sample of size N can be summarized in a $|\mathcal{X}| \times |\mathcal{Y}|$ table with count N_{xy} in cell (x, y). Let $f(x, y)$, $f(x)$, and $f(y)$ be the joint and marginal probabilities of X and Y, i.e., $f(x, y) = P(X = x, Y = y), f(x) = P(X = x), f(y) = P(Y = y)$, then the statistical independence between X and Y can be defined as $f(x, y) = f(x)f(y)$ for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, i.e., all joint probabilities equal the product of their marginal probabilities. Pearson's chi-squared statistic,

$$X^2 = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \frac{(f_N(x, y) - f_N(x)f_N(y))^2}{f_N(x)f_N(y)/N},$$

where $f_N(x, y) = N_{xy}/N$, $f_N(x) = \sum_{y\in\mathcal{Y}} N_{xy}/N$, and $f_N(y) = \sum_{x\in\mathcal{X}} N_{xy}/N$, has been widely used to test independence in two-way contingency tables. Under independence and sufficient sample size, X^2 approximately follows a chi-squared distribution with $df = (|\mathcal{X}| - 1)(|\mathcal{Y}| - 1)$. However, for insufficient sample size (e.g., $\min_{x,y} N_{x+}N_{+y}/N < 5$, where $N_{x+} = \sum_{y\in\mathcal{Y}} N_{xy}, N_{+y} = \sum_{x\in\mathcal{X}} N_{xy}$), the chi-squared test tends to be conservative. Zhang (2019) suggested a random permutation test based on the test statistic

$$\mathcal{D}^2 = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} (f_N(x, y) - f_N(x)f_N(y))^2,$$

which is derived from the squared distance covariance, a measure of the dependence between two random vectors of any type (discrete or continuous) [1,2]. The $\mathcal{D}^2$ statistic is closely related to Pearson's chi-squared statistic, both measuring the squared distance between $f(x, y)$ and $f(x)f(y)$, $(x, y) \in \mathcal{X} \times \mathcal{Y}$. In the numerical study of Zhang (2019), the distance covariance test was evaluated in

terms of the statistical power and type I error rate under various settings (see Figures 1–3 in [1]). It is found that for relatively large sample sizes, the distance covariance test performs similarly well as Pearson's chi-squared test. However, for relatively small sample sizes, the distance covariance test is substantially more powerful and it controls the type I error rate at the nominal level. For small sample, Pearson's chi-squared test exhibits substantial conservativeness, in the sense that the type I error rate is much lower than the nominal level and it fails to reject many false hypotheses. For instance, in a simulation setting with 20 by 20 table and only 50 samples, the statistical power and type I error rate are both close to zero by Pearson's chi-squared test, indicating an extreme conservativeness.

Although the distance covariance test has better empirical performance than Pearson's chi-squared test, especially for small sample size, its theoretical properties have not been investigated. In addition, Zhang (2019) only studied two alternative measures, including distance covariance and projection correlation, but there are many other association measures in the literature remaining unexplored. To name a few, Goodman and Kruskal (1954) introduced two association measures for categorical variables, namely the concentration coefficient and the λ coefficient [3]. Cui et al. (2015) developed a generic association measure based on a mean-variance index [4]. Theil (1970) proposed measuring the association between two categorical variables by the uncertainty coefficient [5]. McCane and Albert (2008) introduced the symbolic covariance, which expresses the covariance between categorical variables in terms of symbolic expressions [6]. In addition, Reshef et al. (2011) proposed a pairwise dependence measure called maximal information coefficient (MIC) based on the grid that maximizes the mutual information gained from the data [7].

The purpose of this paper is to extend my previous work [1] to a broad class of association measures using a general weighted Minkowski distance, and numerically evaluate some selected measures from the proposed class. The proposed class unifies many existing measures including ϕ coefficient, Cramér's V, distance covariance, total variation distance and a slightly modified mean variance index. Furthermore, the strong consistency of the independence tests based on these measures was established, and the scaled forms of unweighted measures were derived. The proposed class provides a rich set of alternatives to the prevailing chi-squared statistic, and it has many potential applications. For instance, it can be applied to the correlation-based modeling, such as correlation-based deep learning [8]. As enlightened by a reviewer, the proposed method may also be applied to the pseudorandom number generator tests, and may improve some existing chi-squared based tests including the poker test and gap test [9].

The remainder of this paper is structured as follows: In Section 2, I introduce the defined class of association measures, and study some important special cases. The scaled forms of unweighted measures are also derived. Section 3 compares the performance of selected measures using simulated data. Section 4 discusses some extensions including the application to ordinal data and conditional independence test for three-way tables.

2. Methods

2.1. A Class of Association Measures for Categorical Variables

As the strength of association between two categorical variables can be reflected by the distance between $f(x,y)$ and $f(x)f(y)$, here I define a class of measures based on the weighted Minkowski distance

$$\mathcal{L}_{r,\omega}(X,Y) = \left\{ \sum_{x\in\boldsymbol{\mathcal{X}}} \sum_{y\in\boldsymbol{\mathcal{Y}}} |f(x,y) - f(x)f(y)|^r \omega^r(x,y) \right\}^{\frac{1}{r}}, \tag{1}$$

where $r \geq 1$, $\omega(x,y) > 0$, and $\omega(x,y)$ only depends on the marginal distributions of X and Y. For $0 < r < 1$, the defined distance violates the triangle inequality therefore it is not a metric. However, $r = \infty$ is allowed, and I denote by $\mathcal{L}_{\infty,\omega}(X,Y)$ the maximum norm. It can be proved

that $\mathcal{L}_{1,\omega}(X,Y) \geq \mathcal{L}_{2,\omega}(X,Y) \geq ... \geq \mathcal{L}_{\infty,\omega}(X,Y)$ for a given weight $\omega(x,y)$. Throughout this paper, I denote by $\mathcal{L}_r(X,Y)$ the unweighted measures, i.e., $\omega(x,y) = 1$. The defined class is quite broad and I begin with some important special cases.

Firstly, most of the chi-squared-type measures belong to the defined class. For instance, the ϕ coefficient for 2×2 tables, i.e., $|\mathcal{X}| = |\mathcal{Y}| = 2$,

$$\phi(X,Y) = \left\{ \sum_{x \in \boldsymbol{\mathcal{X}}} \sum_{y \in \boldsymbol{\mathcal{Y}}} \frac{|f(x,y) - f(x)f(y)|^2}{f(x)f(y)} \right\}^{\frac{1}{2}},$$

is a special case of $\mathcal{L}_{2,\omega}(X,Y)$, where $\omega(x,y) = \{f(x)f(y)\}^{-1/2}$. Extensions of $\phi(X,Y)$ to $I \times J$ tables including Cramér's V and Tschuprow's T [10,11],

$$V(X,Y) = \left\{ \sum_{x \in \boldsymbol{\mathcal{X}}} \sum_{y \in \boldsymbol{\mathcal{Y}}} \frac{|f(x,y) - f(x)f(y)|^2}{f(x)f(y)} \right\}^{\frac{1}{2}} \left\{ \frac{1}{\min(|\mathcal{X}|-1, |\mathcal{Y}|-1)} \right\}^{\frac{1}{2}},$$

$$T(X,Y) = \left\{ \sum_{x \in \boldsymbol{\mathcal{X}}} \sum_{y \in \boldsymbol{\mathcal{Y}}} \frac{|f(x,y) - f(x)f(y)|^2}{f(x)f(y)} \right\}^{\frac{1}{2}} \left\{ \frac{1}{\sqrt{(|\mathcal{X}|-1)(|\mathcal{Y}|-1)}} \right\}^{\frac{1}{2}},$$

are also special cases of $\mathcal{L}_{2,\omega}(X,Y)$, where $\omega(x,y) = \{f(x)f(y)\min(|\mathcal{X}|-1, |\mathcal{Y}|-1)\}^{-1/2}$ for Cramér's V, and $\omega(x,y) = \{f(x)f(y)\sqrt{(|\mathcal{X}|-1)(|\mathcal{Y}|-1)}\}^{-1/2}$ for Tschuprow's T.

Distance covariance for categorical variables also belongs to the defined class. Distance covariance is a measure of statistical dependence between two random vectors X and Y. It is a special case of Hilbert-Schmidt independence criterion (HSIC) [12]. Let (X_1, Y_1), (X_2, Y_2) and (X_3, Y_3) be three independent copies of (X,Y), the distance covariance between X and Y is defined as the square root of

$$\mathrm{dCov}^2(X,Y) = \mathrm{cov}(\|X_1 - X_2\|, \|Y_1 - Y_2\|) - 2\mathrm{cov}(\|X_1 - X_2\|, \|Y_1 - Y_3\|), \tag{2}$$

where $\|\cdot\|$ represents distance between vectors, e.g., Euclidean distance. An alternative definition of distance covariance is given in Sejdinovic et al. (2013) [12], which only uses two independent copies of (X,Y). A proof of the equivalency between the two definitions is provided in Appendix A.1. One property of distance covariance is that $\mathrm{dCov}^2(X,Y) = 0$ if and only if X and Y are statistically independent, indicating its potential of measuring nonlinear dependence. Zhang (2019) studied the distance covariance for categorical variables under multinomial model. Define $\|X_1 - X_2\| = 0$ if $X_1 = X_2$ and 1 otherwise, one can show that

$$\mathrm{dCov}(X,Y) = \left\{ \sum_{x \in \boldsymbol{\mathcal{X}}} \sum_{y \in \boldsymbol{\mathcal{Y}}} |f(x,y) - f(x)f(y)|^2 \right\}^{\frac{1}{2}}, \tag{3}$$

and it is easy to see that $\mathrm{dCov}(X,Y) = \mathcal{L}_2(X,Y)$. A detailed proof of Equation (3) is provided in Appendix A.2.

Another special case is total variation distance, which is defined as the largest difference between two probability measures [13]. Let $\mu_0(\cdot)$ and $\mu_\alpha(\cdot)$ be the measures under independence and dependence respectively, the total variation distance between μ_0 and μ_α can be used to measure the dependence between variables X and Y

$$\delta(\mu_0, \mu_\alpha) = \max_{S \subset \mathcal{X} \times \mathcal{Y}} |\mu_0(S) - \mu_\alpha(S)|. \tag{4}$$

In the case of discrete sampling spaces, let $S^+ = \{(x,y), s.t., f(x,y) > f(x)f(y)\}$ and $S^- = \{(x,y), s.t., f(x,y) < f(x)f(y)\}$, then we have

$$\delta(\mu_0, \mu_\alpha) = |\mu_0(S^+) - \mu_\alpha(S^+)| = |\mu_0(S^-) - \mu_\alpha(S^-)| = \frac{1}{2}\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}|f(x,y) - f(x)f(y)|, \quad (5)$$

therefore $\delta(\mu_0, \mu_\alpha) = \mathcal{L}_{1,\omega}(X,Y)$, where $\omega(X,Y) = \frac{1}{2}$.

In addition, I pointed out that the mean variance index (MV) recently developed by Cui et al. [4] also belongs to our defined class, subject to some slight modifications. The MV between two variables X and Y is defined as $MV(X|Y) = E_X(V_Y(F(X|Y)))$, where $F(x|y)$ stands for conditional distribution function. It can be proved that $MV(X|Y) = 0$ if and only if X and Y are independent. The MV measure is originally developed for continuous variables. To make it suitable for categorical variables while maintaining the main theoretical property, I slightly modified the definition of MV. First, I replaced the conditional c.d.f. $F(x|y)$ with conditional p.m.f. $f(x|y)$. Second, as the MV measure is generally asymmetric, i.e., $MV(X|Y) \neq MV(Y|X)$, I considered a symmetric version of the index, $MV(X,Y) = \frac{1}{2}(MV(X|Y) + MV(Y|X))$. With the two modifications, one can prove the following result (a detailed proof is provided in Appendix A.3)

$$\sqrt{\mathrm{MV}(X,Y)} = \left\{\sum_{x\in\boldsymbol{\mathcal{X}}}\sum_{y\in\boldsymbol{\mathcal{Y}}}\frac{1}{2}|f(x,y) - f(x)f(y)|^2\left(\frac{f(x)}{f(y)} + \frac{f(y)}{f(x)}\right)\right\}^{\frac{1}{2}},$$

therefore $\sqrt{\mathrm{MV}(X,Y)} = \mathcal{L}_{2,\omega}(X,Y)$, where $\omega(x,y) = \sqrt{\frac{1}{2}(\frac{f(x)}{f(y)} + \frac{f(y)}{f(x)})}$. As $\frac{1}{2}(\frac{f(x)}{f(y)} + \frac{f(y)}{f(x)}) \geq 1$, we also have $\sqrt{\mathrm{MV}(X,Y)} \geq \mathcal{L}_2(X,Y)$.

Similar as the MV index, the symmetric version of some other directional association measures (e.g., the concentration coefficient [3]), are also the special cases of $\mathcal{L}_{r,\omega}$.

2.2. Sample Estimate and Independence Test

Given a simple random sample of size N, one can estimate $\mathcal{L}_{r,\omega,N}(X,Y)$ using sample quantities

$$\mathcal{L}_{r,\omega,N}(X,Y) = \left\{\sum_{x\in\boldsymbol{\mathcal{X}}}\sum_{y\in\boldsymbol{\mathcal{Y}}}|f_N(x,y) - f_N(x)f_N(y)|^r\omega_N^r(x,y)\right\}^{\frac{1}{r}}, \quad (6)$$

where $f_N(x,y)$, $f_N(x)$ and $f_N(y)$ represent the maximum likelihood estimates of joint and marginal probabilities, respectively, i.e., $f_N(x,y) = N_{xy}/N$, $f_N(x) = \sum_{y\in\mathcal{Y}} N_{xy}/N$, and $f_N(y) = \sum_{x\in\mathcal{X}} N_{xy}/N$. The following theorem establishes the strong consistency of the independence test based on $\mathcal{L}_{r,\omega,N}(X,Y)$ (a detailed proof is provided in Appendix A.4):

Theorem 1. *Assume that the estimated weights are bounded above by a constant $C > 0$, i.e., $\sup_{x,y}\omega_N(x,y) = C$, then for any $r \geq 1$ and $\epsilon > 0$, we have $P\big(\mathcal{L}_{r,\omega,N}(X,Y) > \epsilon\big) < (2^{|\mathcal{X}||\mathcal{Y}|} + 2^{|\mathcal{Y}|} + 2^{|\mathcal{X}|})\exp(-N\epsilon^2/18C^2)$ under independence. The inequality also holds for maximum norm $\mathcal{L}_{\infty,\omega,N}(X,Y)$.*

It is noteworthy that the asymptotic null distribution of $\mathcal{L}_{r,\omega,N}(X,Y)$ is impratical to derive. The theorem above provides a simple way to compute the upper bound of p-value, however, the bound $(2^{|\mathcal{X}||\mathcal{Y}|} + 2^{|\mathcal{Y}|} + 2^{|\mathcal{X}|})\exp(-N\epsilon^2/18C^2)$ is generally not tight, thus the p-value could be largely overestimated. Here, I suggest a simple permutation procedure to evaluate the significance. One can randomly shuffle the observations of X (or equivalently, the observations of Y) for M times, and compute the test statistic $\mathcal{L}_{r,\omega,N}(X_{perm}, Y)$ for each permuted dataset. The permutation p-value can be computed as the proportion of $\mathcal{L}_{r,\omega,N}(X_{perm}, Y)$'s that exceed the actually observed one. I used the permutation p-value to evaluate statistical significant in our simulation studies.

2.3. Scaled Forms of Unweighted Measures

Motivated by the classic correlation coefficient, I define the following scaled form for unweighted measure $\mathcal{L}_r(X,Y)$:

$$\mathcal{L}_r^*(X,Y) = \frac{\mathcal{L}_r(X,Y)}{\sqrt{\mathcal{L}_r(X,X)}\sqrt{\mathcal{L}_r(Y,Y)}}, \tag{7}$$

where $\mathcal{L}_r(X,X) = \left\{ \sum_{x\in\boldsymbol{\mathcal{X}}} \sum_{x'\in\boldsymbol{\mathcal{X}}} |f(x,x') - f(x)f(x')|^r \right\}^{\frac{1}{r}}$, $f(x,x') = f(x)$ if $x = x'$ and $f(x,x') = 0$ otherwise.

The term $\mathcal{L}_r(X,X)$ can be written as

$$\mathcal{L}_r(X,X) = \left\{ \sum_{x\in\boldsymbol{\mathcal{X}}} |f(x) - f^2(x)|^r + \sum_{x\in\boldsymbol{\mathcal{X}}} \sum_{x'\in\boldsymbol{\mathcal{X}}\setminus x} |f(x)f(x')|^r \right\}^{\frac{1}{r}},$$

and as examples, the explicit expressions for $\mathcal{L}_1(X,X)$, $\mathcal{L}_2(X,X)$, and $\mathcal{L}_\infty(X,X)$ are given below

- $\mathcal{L}_1(X,X) = \sum_{x\in\boldsymbol{\mathcal{X}}} \left[f(x) - f^2(x)\right] + \sum_{x\in\boldsymbol{\mathcal{X}}} \sum_{x'\in\boldsymbol{\mathcal{X}}\setminus x} \left[f(x)f(x')\right] = 2\left[1 - \sum_{x\in\boldsymbol{\mathcal{X}}} f^2(x)\right]$
- $\mathcal{L}_2(X,X) = \left\{ \sum_{x\in\boldsymbol{\mathcal{X}}} f^2(x)\left[\sum_{x\in\boldsymbol{\mathcal{X}}} f^2(x) + 1\right] - 2\sum_{x\in\boldsymbol{\mathcal{X}}} f^3(x) \right\}^{\frac{1}{2}}$
- $\mathcal{L}_\infty(X,X) = \max_{x\in\boldsymbol{\mathcal{X}}} \left[f(x) - f^2(x)\right] \vee \max_{x\in\boldsymbol{\mathcal{X}}, x'\in\boldsymbol{\mathcal{X}}\setminus x} \left[f(x)f(x')\right] = \max_{x\in\boldsymbol{\mathcal{X}}} \left[f(x) - f^2(x)\right]$

It can be seen that $\mathcal{L}_2^*(X,Y)$ is same as the distance correlation between X and Y [1], therefore $0 \le \mathcal{L}_2^*(X,Y) \le 1$, where $\mathcal{L}_r^*(X,Y) = 0$ if and only if X and Y are independent. In fact, for any $1 \le r < \infty$, if $f(x) > 0, f(y) > 0$ for $x \in \mathcal{X}, y \in \mathcal{Y}$, it can be proved that $0 \le \mathcal{L}_r^*(X,Y) \le 1$, where $\mathcal{L}_r^*(X,Y) = 0$ if and only if X and Y are independent, and $\mathcal{L}_r^*(X,Y) = 1$ if and only if X and Y have perfect association, i.e., $|\mathcal{X}| = |\mathcal{Y}|$ and for any $x \in \mathcal{X}$, there exists a unique $y \in \mathcal{Y}$, such that $f(x,y) = f(x) = f(y)$.

For $\mathcal{L}_\infty^*(X,Y)$, by Cauchy-Schwarz inequality,

$$\begin{aligned}
\mathcal{L}_\infty(X,Y) &= \max_{x\in\boldsymbol{\mathcal{X}}, y\in\boldsymbol{\mathcal{Y}}} |f(x,y) - f(x)f(y)| \\
&= \max_{x\in\boldsymbol{\mathcal{X}}, y\in\boldsymbol{\mathcal{Y}}} |\mathrm{cov}(I\{X = x\}, I\{Y = y\})| \\
&\le \max_{x\in\boldsymbol{\mathcal{X}}, y\in\boldsymbol{\mathcal{Y}}} \sqrt{V(I\{X = x\})V(I\{Y = y\})} \\
&= \max_{x\in\boldsymbol{\mathcal{X}}} \sqrt{V(I\{X = x\})} \max_{y\in\boldsymbol{\mathcal{Y}}} \sqrt{V(I\{Y = y\})} \\
&= \sqrt{\max_{x\in\boldsymbol{\mathcal{X}}} \left[f(x) - f^2(x)\right]} \sqrt{\max_{y\in\boldsymbol{\mathcal{Y}}} \left[f(y) - f^2(y)\right]} \\
&= \sqrt{\mathcal{L}_\infty(X,X)\mathcal{L}_\infty(Y,Y)},
\end{aligned}$$

therefore $0 \le \mathcal{L}_\infty^* \le 1$. However, in general, $\mathcal{L}_\infty^*(X,Y) = 1$ does not imply that X and Y are perfectly associated. I gave an example in Table 1, where $\mathcal{L}_\infty^*(X,Y) = 1$ but X and Y are not perfectly associated.

Table 1. An example that X and Y are not perfectly associated, but $\mathcal{L}_\infty^*(X,Y) = 1$.

	$Y=1$	$Y=2$	$Y=3$
$X=1$	1/2	0	0
$X=2$	0	1/8	1/8
$X=3$	0	1/8	1/8

3. Numerical Study

Two simulation studies were conducted to compare the performance of some selected measures from our defined class. In both simulations, I set $|\mathcal{X}| = |\mathcal{Y}| = 10$ and varied the sample size from 25 to 500, so that the simulated contingency tables were relatively large and sparse (average count $N/|\mathcal{X}||\mathcal{Y}|$ is between 0.25 and 5).

In the first simulation study, I considered the independence test based on different unweighted measures, including $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_4$ and $\mathcal{L}_\infty$, under the following multinomial settings:

- Setting 1: $f(x,y) = 0.05$ for 10 randomly selected cells and $f(x,y) = \frac{0.5}{90}$ for the remaining 90 cells
- Setting 2: $f(x,y) = 0.08$ for 10 randomly selected cells and $f(x,y) = \frac{0.2}{90}$ for the remaining 90 cells
- Setting 3: $f(x,y) = 0.1$ for one randomly selected cell and $f(x,y) = \frac{0.9}{99}$ for the remaining 99 cells
- Setting 4: $f(x,y) = 0.2$ for one randomly selected cell and $f(x,y) = \frac{0.8}{99}$ for the remaining 99 cells

For each test, the p-values were computed based on 2000 random permutations. Figure 1 summarizes the empirical statistical power of the four tests under significance level 0.05. It could be seen that, in settings 1 and 2, the $\mathcal{L}_2$ measure (Euclidean distance) performed consistently better than the other three (comparable to $\mathcal{L}_4$). The maximum norm $\mathcal{L}_\infty$ performs the worst in these two settings. In settings 3 and 4, where a single cell accounts for most deviation from independence, the maximum norm performs the best, while the $\mathcal{L}_1$ measure (Manhattan distance) gives the lowest power. Figure 2 summarizes the type I error rate, where it can be seen that all the four tests control the type I error rates at the nominal level of 0.05.

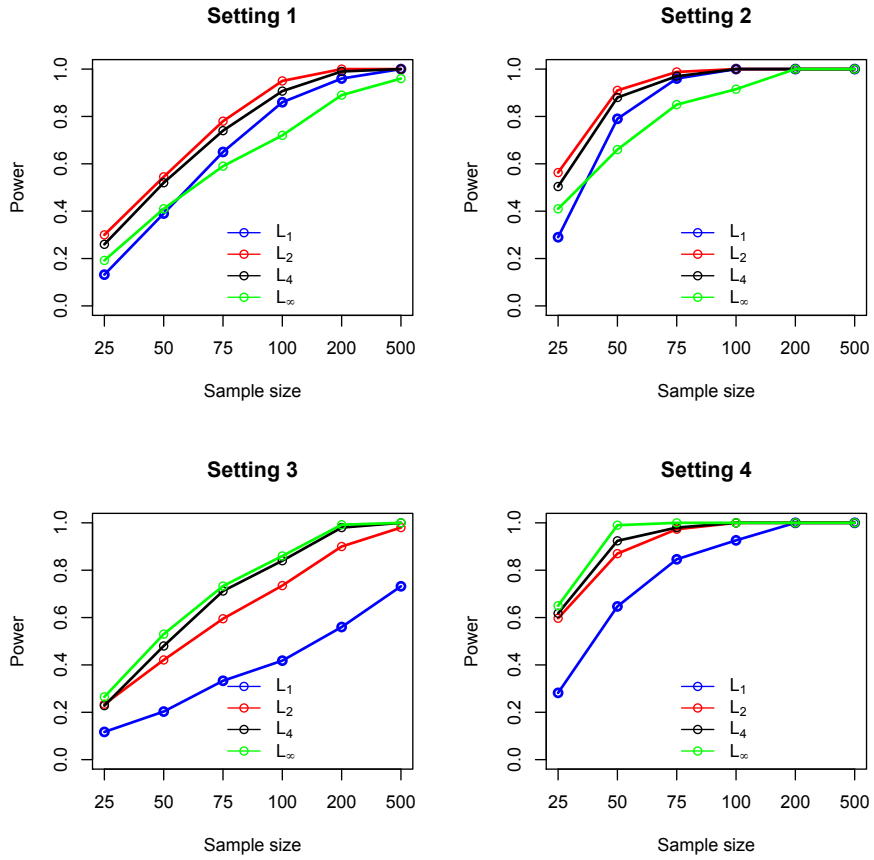

Figure 1. Empirical statistical power of four different measures including $\mathcal{L}_1$ (blue), $\mathcal{L}_2$ (red), $\mathcal{L}_4$ (black) and $\mathcal{L}_\infty$ (green), under settings 1–4. In each setting, sample sizes are $n = 25, 50, 75, 100, 200, 500$, and all results were based on 1000 replications.

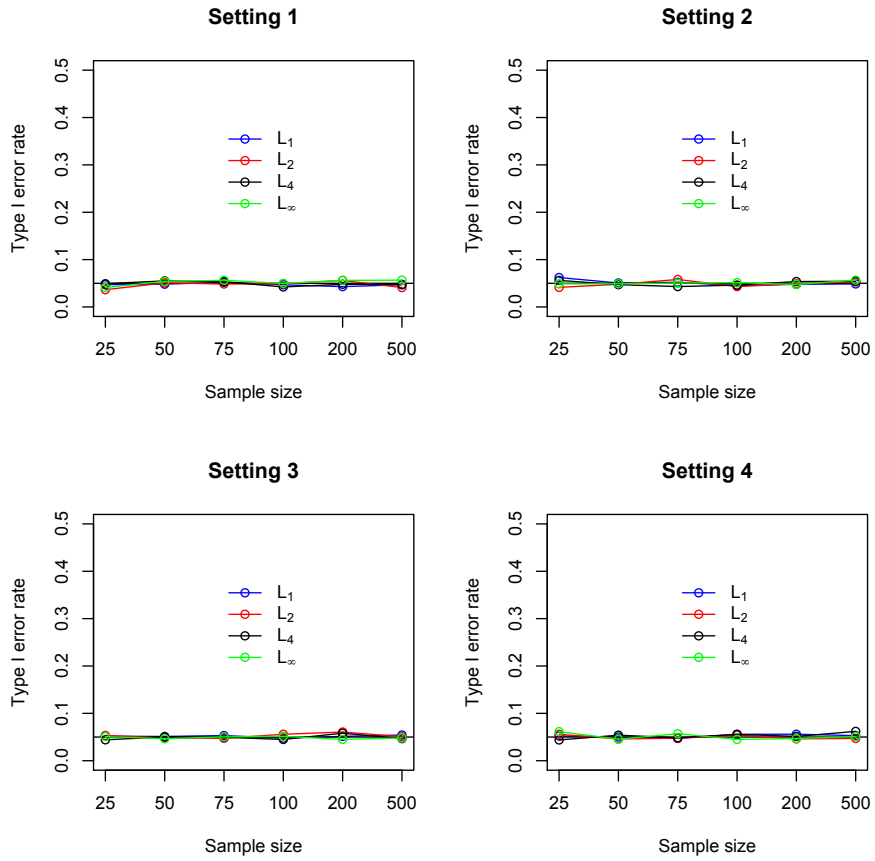

Figure 2. Empirical Type I error rate of four different measures including $\mathcal{L}_1$ (blue), $\mathcal{L}_2$ (red), $\mathcal{L}_4$ (black) and $\mathcal{L}_\infty$ (green), under settings 1–4. In each setting, sample sizes are $n = 25, 50, 75, 100, 200, 500$, and all results were based on 1000 replications.

In the second simulation study, I focused on $\mathcal{L}_{2,\omega}(X, Y)$ as it subsumes many popular measures. In particular, I compared three different weight functions, including $\omega(x, y) = 1$ (distance covariance), $\omega(x, y) = \{f(x)f(y)\}^{-1/2}$ (Pearson's chi-squared), and $\omega(x, y) = \sqrt{\frac{1}{2}(\frac{f(x)}{f(y)} + \frac{f(y)}{f(x)})}$ (modified mean variance index). Figure 3 shows the empirical statistical power of the three measures under settings 1 and 2, where it can be seen that the unweighted $\mathcal{L}_2$ compares favorably to the weighted ones.

Based on the simulation studies, I recommend to the unweighted $\mathcal{L}_r$ measures with a moderate choice of r, for instance, $r = 2, 3, 4$ for large sparse tables, because they could give satisfactory and stable statistical power in general scenarios. The maximum norm $\mathcal{L}_\infty$ is not recommended, unless one is very confident that there exist a very small number of cells that account for most deviation from independence.

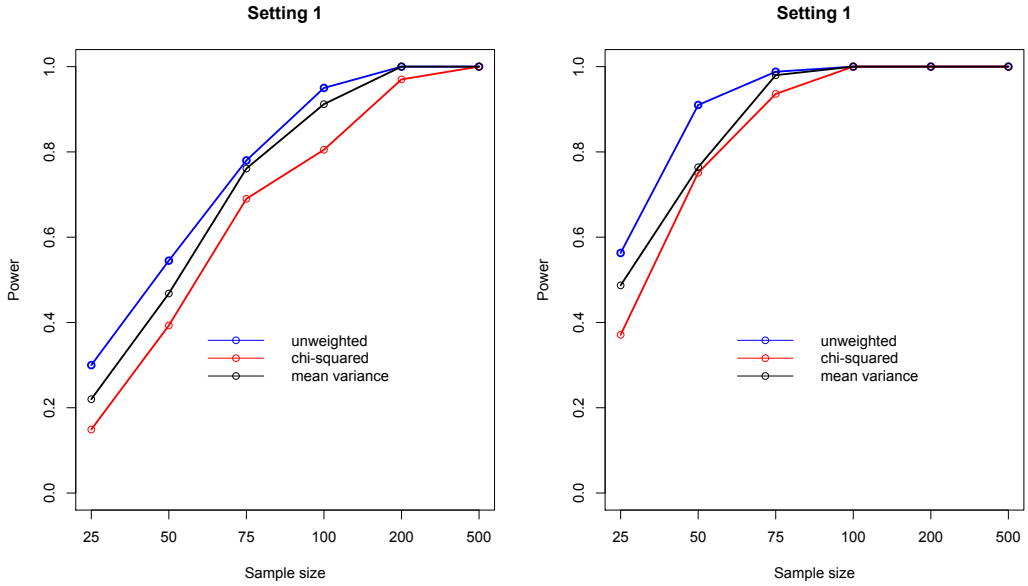

Figure 3. Empirical statistical power of three $\mathcal{L}_{2,\omega}$ measures including $\omega(x,y) = 1$ (distance covariance), $\omega(x,y) = \{f(x)f(y)\}^{-1/2}$ (chi-squared), and $\omega(x,y) = \sqrt{\frac{1}{2}(\frac{f(x)}{f(y)} + \frac{f(y)}{f(x)})}$ (symmetric mean variance index), under settings 1 and 2. In each setting, sample sizes are $n = 25, 50, 75, 100, 200, 500$, and all results were based on 1000 replications.

4. Discussion

In this work, I proposed a rich class of dependence measures for categorical variables based on weighted Minkowski distance. The defined class unifies a number of existing measures including Cramér's V, distance covariance, total variation distance and a slightly modified mean variance index. I provided the scaled forms of unweighted measures, which range from 0 (independence) to 1 (perfect association). Further, I established the strong consistency of the defined measures and suggested a simple permutation test for evaluating significance. Although I have used nominal and univariate categorical variables for illustrations, the proposed framework can be extended to other data types and problems:

First, the proposed measures can be used to detect ordinal association by assigning proper weights. Similar as Pearson's correlation coefficient, one may assign larger weights to more extreme categories of X and Y. To be specific, let $d(x,x')$ be the predefined distance between categories $X = x$ and $X = x'$, and $d(y,y')$ be the distance between y and y', and one could apply the following weight function

$$\omega(x,y) = E(d(x,X)d(y,Y)) = \sum_{x' \in \mathcal{X} \setminus x, y' \in \mathcal{Y} \setminus y} d(x,x')d(y,y')f(x')f(y'),$$

which assigns larger weights to cells in the corners but smaller weights to cells in the center of the table.

Second, my framework can be generalized to random vectors and multi-way tables. In the case of three-way table (X,Y,Z), one can define the following Minkowski distance between $f(x,y,z)$ and $f(x,y)f(z)$

$$\mathcal{L}_{r,\omega}((X,Y),Z) = \left\{ \sum_{x \in \boldsymbol{\mathcal{X}}} \sum_{y \in \boldsymbol{\mathcal{Y}}} \sum_{z \in \boldsymbol{\mathcal{Z}}} |f(x,y,z) - f(x,y)f(z)|^r \omega^r(x,y,z) \right\}^{\frac{1}{r}},$$

which can be used to test the joint independence between (X,Y) and Z, or equivalently, to test the homogeneity of the joint distribution of (X,Y) at different levels of Z. A similar permutation procedure

can be applied to evaluate the statistical significance. One can also define the distance between $f(x,y,z)$ and $f(x)f(y)f(z)$ to test the mutual independence of (X,Y,Z)

$$\mathcal{L}_{r,\omega}(X,Y,Z) = \left\{ \sum_{x\in\boldsymbol{\mathcal{X}}} \sum_{y\in\boldsymbol{\mathcal{Y}}} \sum_{z\in\boldsymbol{\mathcal{Z}}} |f(x,y,z) - f(x)f(y)f(z)|^r \omega^r(x,y,z) \right\}^{\frac{1}{r}},$$

Furthermore, the framework can be extended to conditional independence test in three-way tables [14], by defining distance between conditional joint probabilities $f(x,y|z)$ and the product of conditional marginal probabilities $f(x|z)f(y|z)$

$$\mathcal{L}_{r,\omega}(X,Y|Z) = \left\{ \sum_{x\in\boldsymbol{\mathcal{X}}} \sum_{y\in\boldsymbol{\mathcal{Y}}} \sum_{z\in\boldsymbol{\mathcal{Z}}} |f(x,y|z) - f(x|z)f(y|z)|^r \omega^r(x,y,z) \right\}^{\frac{1}{r}}.$$

Author Contributions: Q.Z. conceived of the presented idea, developed the theory, performed the computations and wrote the manuscript.

Funding: This research received no external funding.

Acknowledgments: The author would like to thank the editor and two reviewers for their thoughtful comments and efforts towards improving the manuscript.

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MIC	maximum information coefficient
HSIC	Hilbert-Schmidt independence criterion
MV	mean variance index
c.d.f.	cumulative distribution function
p.m.f.	probability mass function

Appendix A. Technical Details

Appendix A.1. Equivalency between Two Definitions of Distance Covariance

- Definition by Szekely et al. (2007): $\mathrm{dCov}^2(X,Y) = E_{X_1X_2Y_1Y_2}(\|X_1-X_2\|\|Y_1-Y_2\|) + E_{X_1X_2}(\|X_1-X_2\|)E_{Y_1Y_2}(\|Y_1-Y_2\|) - 2E_{X_1X_2Y_1Y_3}(\|X_1-X_2\|\|Y_1-Y_3\|)$
- Definition by Sejdinovic et al. (2013): $\mathrm{dCov}^2(X,Y) = E_{X_1X_2Y_1Y_2}(\|X_1-X_2\|\|Y_1-Y_2\|) + E_{X_1X_2}(\|X_1-X_2\|)E_{Y_1Y_2}(\|Y_1-Y_2\|) - 2E_{X_1Y_1}(E_{X_2}(\|X_1-X_2\|)E_{Y_2}(\|Y_1-Y_2\|))$

The first two terms are the same, and the equivalency between the two definitions can be showed as follows:

$$\begin{aligned}
&E_{X_1Y_1}(E_{X_2}(\|X_1-X_2\|)E_{Y_2}(\|Y_1-Y_2\|))\\
&= \int_{x_1}\int_{y_1}\left[\int_{x_2}\|x_1-x_2\|f(x_2)dx_2\int_{y_2}\|y_1-y_2\|f(y_2)dy_2\right]f(x_1,y_1)dx_1dy_1\\
&= \int_{x_1}\int_{y_1}\left[\int_{x_2}\|x_1-x_2\|f(x_2)dx_2\int_{y_3}\|y_1-y_3\|f(y_3)dy_3\right]f(x_1,y_1)dx_1dy_1\\
&= \int_{x_1}\int_{x_2}\int_{y_1}\int_{y_3}\|x_1-x_2\|\|y_1-y_3\|f(x_1,y_1)f(x_2)f(y_3)dx_1dx_2dy_1dy_3\\
&= E_{X_1X_2Y_1Y_3}(\|X_1-X_2\|\|Y_1-Y_3\|)
\end{aligned}$$

Appendix A.2. Derivation of Equation (3)

Following Zhang (2019), I rewrite categorical variables X and Y as two random vectors of dimensions $|\boldsymbol{\mathcal{X}}|$ and $|\boldsymbol{\mathcal{Y}}|$, $\boldsymbol{X} = \{I(X = x)\}_{x\in\boldsymbol{\mathcal{X}}}$ and $\boldsymbol{Y} = \{I(Y = y)\}_{y\in\boldsymbol{\mathcal{Y}}}$, where $I(\cdot)$ stands for the indicator function. Let $\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|$ equal 0 if $\boldsymbol{X}_1 = \boldsymbol{X}_2$ and 1 otherwise. Let $(\boldsymbol{X}_1, \boldsymbol{Y}_2)$, $(\boldsymbol{X}_2, \boldsymbol{Y}_2)$, $(\boldsymbol{X}_3, \boldsymbol{Y}_3)$ be three independent copies of $(\boldsymbol{X}, \boldsymbol{Y})$. By Equation (2), the squared distance covariance can be also expressed as

$$\mathrm{dCov}^2(\boldsymbol{X}, \boldsymbol{Y}) = E(\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|\|\boldsymbol{Y}_1 - \boldsymbol{Y}_2\|) + E(\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|)E(\|\boldsymbol{Y}_1 - \boldsymbol{Y}_2\|) - 2E(\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|\|\boldsymbol{Y}_1 - \boldsymbol{Y}_3\|).$$

Under multinomial sampling scheme, it is straightforward to show that

$$\begin{aligned}
E(\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|\|\boldsymbol{Y}_1 - \boldsymbol{Y}_2\|) &= P(\boldsymbol{X}_1 \neq \boldsymbol{X}_2, \boldsymbol{Y}_1 \neq \boldsymbol{Y}_2) = \sum_{x\in\boldsymbol{\mathcal{X}}}\sum_{y\in\boldsymbol{\mathcal{Y}}} f(x,y)[1 - f(x) - f(y) + f(x,y)], \\
E(\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|)E(\|\boldsymbol{Y}_1 - \boldsymbol{Y}_2\|) &= P(\boldsymbol{X}_1 \neq \boldsymbol{X}_2)P(\boldsymbol{Y}_1 \neq \boldsymbol{Y}_2) = [1 - \sum_{x\in\boldsymbol{\mathcal{X}}} f^2(x)][1 - \sum_{y\in\boldsymbol{\mathcal{Y}}} f^2(y)], \\
E(\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|\|\boldsymbol{Y}_1 - \boldsymbol{Y}_3\|) &= P(\boldsymbol{X}_1 \neq \boldsymbol{X}_2, \boldsymbol{Y}_1 \neq \boldsymbol{Y}_3) = \sum_{x\in\boldsymbol{\mathcal{X}}}\sum_{y\in\boldsymbol{\mathcal{Y}}} f(x,y)[1 - f(x)][1 - f(y)].
\end{aligned}$$

Summarizing the results above, I have

$$\mathrm{dCov}(X, Y) = \Big\{ \sum_{x\in\boldsymbol{\mathcal{X}}}\sum_{y\in\boldsymbol{\mathcal{Y}}} |f(x,y) - f(x)f(y)|^2 \Big\}^{\frac{1}{2}}.$$

Appendix A.3. Derivation of the Modified Mean Variance Index

The symmetric mean variance index is defined as

$$MV(X, Y) = \frac{1}{2}(MV(X|Y) + MV(Y|X)) = \frac{1}{2}(E_X(V_Y(f(X|Y))) + E_Y(V_X(f(Y|X)))).$$

I first derived the explicit formula for $E_X(V_Y(f(X|Y)))$:

$$\begin{aligned}
E_X(V_Y(f(X|Y))) &= E_X E_Y(f(X|Y))^2 - E_X E_Y^2(f(X|Y)) \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} f^2(x,y)\frac{f(x)}{f(y)} - \sum_{x\in\mathcal{X}} f^3(x) \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \frac{f(x)}{f(y)}(f^2(x,y) - f^2(x)f^2(y)) \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \frac{f(x)}{f(y)}\{(f(x,y) - f(x)f(y))^2 - 2f^2(x)f^2(y) + 2f(x,y)f(x)f(y)\} \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \frac{f(x)}{f(y)}(f(x,y) - f(x)f(y))^2
\end{aligned}$$

Similarly, it can seen that $E_Y(V_X(f(Y|X))) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \frac{f(y)}{f(x)}(f(x,y) - f(x)f(y))^2$, therefore

$$MV(X, Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \frac{1}{2}\left(\frac{f(y)}{f(x)} + \frac{f(x)}{f(y)}\right)(f(x,y) - f(x)f(y))^2.$$

Appendix A.4. Proof of Theorem 1

Because $\mathcal{L}_{1,\omega,N} \geq \mathcal{L}_{2,\omega,N} \geq ... \geq \mathcal{L}_{\infty,\omega,N}$, I only need prove the strong consistency for $\mathcal{L}_{1,\omega,N}$. For categorical variable X, let $f(x)_{x\in\mathcal{X}}$ be the probability mass function, N be the sample size, and $f_N(x)$ be the sample estimate, Biau and Gyorfi (2005) [15] proved the following result

Lemma A1. *For any* $\epsilon > 0$, $P(\sum_{x\in\mathcal{X}} |f_N(x) - f(x)| > \epsilon) < 2^{|\mathcal{X}|} e^{-\frac{N\epsilon^2}{2}}$.

As $\sup_{x,y} \omega(x,y) = C > 0$, I have

$$\mathcal{L}_{1,\omega,N}(X,Y) \leq C\Big(\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} |f_N(x,y) - f(x,y)| + \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} |f(x,y) - f(x)f(y)| + \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} |f_N(x)f_N(y) - f(x)f(y)|\Big).$$

Under independence, I have $\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} |f(x,y) - f(x)f(y)| = 0$. By Lemma A1, the first term $\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} |f_N(x,y) - f(x,y)|$ satisfies that

$$P\left(C\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} |f_N(x,y) - f(x,y)| > \frac{\epsilon}{3}\right) < 2^{|\mathcal{X}||\mathcal{Y}|} e^{-\frac{N\epsilon^2}{18C^2}},$$

The third term can be bounded as follows:

$$\begin{aligned}\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} |f_N(x)f_N(y) - f(x)f(y)| &\leq \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} |f(x)f_N(y) - f(x)f(y)| + \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} |f_N(x)f_N(y) - f(x)f_N(y)| \\ &= \sum_{y\in\mathcal{Y}} |f_N(y) - f(y)| + \sum_{x\in\mathcal{X}} |f_N(x) - f(x)|\end{aligned}$$

By Lemma A1, I have

$$P\left(C\sum_{y\in\mathcal{Y}} |f_N(y) - f(y)| > \frac{\epsilon}{3}\right) < 2^{|\mathcal{Y}|} e^{-\frac{N\epsilon^2}{18C^2}},$$

$$P\left(C\sum_{x\in\mathcal{X}} |f_N(x) - f(x)| > \frac{\epsilon}{3}\right) < 2^{|\mathcal{X}|} e^{-\frac{N\epsilon^2}{18C^2}},$$

and summarizing the results above, I have

$$\begin{aligned}P\Big(\mathcal{L}_{1,\omega,N}(X,Y) > \epsilon\Big) &\leq P\left(C\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} |f_N(x,y) - f(x,y)| > \frac{\epsilon}{3}\right) \\ &+ P\left(C\sum_{y\in\mathcal{Y}} |f_N(y) - f(y)| > \frac{\epsilon}{3}\right) + P\left(C\sum_{x\in\mathcal{X}} |f_N(x) - f(x)| > \frac{\epsilon}{3}\right) \\ &< 2^{|\mathcal{X}||\mathcal{Y}|} e^{-\frac{N\epsilon^2}{18C^2}} + 2^{|\mathcal{Y}|} e^{-\frac{N\epsilon^2}{18C^2}} + 2^{|\mathcal{X}|} e^{-\frac{N\epsilon^2}{18C^2}} \\ &= (2^{|\mathcal{X}||\mathcal{Y}|} + 2^{|\mathcal{X}|} + 2^{|\mathcal{Y}|}) e^{-\frac{N\epsilon^2}{18C^2}}.\end{aligned}$$

References

1. Zhang, Q. Independence test for large sparse contingency tables based on distance correlation. *Stat. Probab. Lett.* **2019**, *148*, 17–22. [CrossRef]
2. Szekely, G.; Rizzo, M.; Bakirov, N. Measuring and testing dependence by correlation of distances. *Ann. Stat.* **2007**, *35*, 2769–2794. [CrossRef]
3. Goodman, L.; Kruskal, W. Measures of association for cross classifications, part I. *J. Am. Stat. Assoc.* **1954**, *49*, 732–764.

4. Cui, H.; Li, R.; Zhong, W. Model-Free Feature Screening for Ultrahigh Dimensional Discriminant Analysis. *J. Am. Stat. Assoc.* **2015**, *110*, 630–641. [CrossRef] [PubMed]
5. Theil, H. On the estimation of relationships involving qualitative variables. *Am. J. Sociol.* **1970**, *76*, 103–154. [CrossRef]
6. McCane, B.; Albert, M. Distance functions for categorical and mixed variables. *Pattern Recognit. Lett.* **2008**, *29*, 986–993. [CrossRef]
7. Reshef, D.; Reshef, Y.; Finucane, H.; Grossman, S.; McVean, P.; Turnbaugh, E.; Lander, M.; Mitzenmacher, M.; Sabeti, P. Detecting novel associations in large data sets. *Science* **2011**, *334*, 1518–1524. [CrossRef]
8. Moews, B.; Herrmann, M.; Ibikunle, G. Lagged correlation-based deep learning for directional trend change prediction in financial time series. *arXiv* **2018**, arXiv:1811.11287.
9. Knuth, D. *The Art of Computer Programming*, 3rd ed.; Addison-Wesley: Boston, MA, USA, 1997.
10. Cramér, H. *Mathematical Methods of Statistics*; Princeton Press: Princeton, NJ, USA, 1946.
11. Tschuprow, A. Principles of the Mathematical Theory of Correlation. *Bull. Am. Math. Soc.* **1939**, *46*, 389.
12. Sejdinovic, D.; Sriperumbudur, B.; Gretton, A.; Fukumizu, K. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Stat.* **2013**, *41*, 2263–2291. [CrossRef]
13. Sriperumbudur, B.; Fukumizu, K.; Gretton, A.; Scholkopf, B.; Lanckriet, G. On the empirical estimation of integral probability metric. *Electron. J. Stat.* **2012**, *6*, 1550–1599. [CrossRef]
14. Zhang, Q.; Tinker, J. Testing conditional independence and homogeneity in large sparse three-way tables using conditional distance covariance. *Stat* **2019**, *8*, 1–9. [CrossRef]
15. Biau, G.; Gyorfi, L. On the asymptotic properties of a nonparametric l1-test statistic of homogeneity. *IEEE Trans. Inf. Theory* **2005**, *51*, 3965–3973. [CrossRef]