Technical Report 1701

testforDEP: An **R** Package for distribution-free tests and visualization tools for independence

Jeffrey C. Miecznikowski, En-shuo Hsu, Yanhua Chen, Albert Vexler

Department of Biostatistics, the State University of New York,

Buffalo,NY 14214,USA

Abstract

This article introduces **testforDEP**, a portmanteau **R** package containing several tests and visualization tools to examine independence between two variables. This new package combines classical tests including Pearson's product moment correlation coefficient method, Kendall's τ rank correlation coefficient method and Spearman's ρ rank correlation coefficient method with modern tests consisting of density-based empirical likelihood ratio test, Kallenberg data-driven test, Maximal information coefficient test, Hoeffding's independence test, empirical likelihood based test, and continuous analysis of variance test. For two variables the function `testforDEP` provides an interface to those tests and returns test statistics, corresponding p values, and bootstrap confidence intervals. The function `AUK` provides an interface for Kendall plots and computes the area under the Kendall curve. In this paper, we present the **testforDEP** package and perform a power analysis via Monte-Carlo simulations ultimately concluding that classical tests are superior for simple linear dependence structures while the more modern tests are more powerful for non-linear and random-types of dependence.

*Keywords:*testing for independence, empirical likelihood ratio test, Kallenberg data-driven rank test, maximal information coefficient, Kendall plot

1 Introduction

In this article, we present the **testforDEP** package, a package for testing dependence between two random variables in **R**. This package addresses a need

for implementing *both* classical and modern tests of independence, as well as visualization in easy to implement functions. The function `cor.test` offered in the base **R** package R Core Team (2013) gives classical tests for association/-correlation between two samples using the Pearson product moment correlation coefficient (Pearson, 1920), Kendall τ rank correlation coefficient (Kendall, 1938) and Spearman ρ rank correlation coefficient (Spearman, 1904). The function `cor.test` is helpful to test for independence between two variables when the variables are linearly dependent or monotonically associated. However the function `cor.test` is less powerful to detect general structures of dependence between two random variables, including non-linear and/or random-effect dependence structures. Many modern statistical methodologies have been proposed to detect general structure of dependence. These methods include the density-based empirical likelihood ratio test for independence (Vexler et al., 2014), data-driven rank test for independence (Kallenberg and Ledwina, 1999), maximal information coefficient (Reshef et al., 2011), empirical likelihood based test (Einmahl and McKeague, 2003), and continuous analysis of variance test (Wang et al., 2015). These methods are useful to detect complex structures of dependence and until now there were no **R** packages available to implement these modern tests. Hence, we propose the new package **testforDEP** combing both classical and modern tests. The package **testforDEP** also provides visualization tools such as the Kendall plot and area under the Kendall curve (Vexler et al., 2015). Moreover, we develop an exact test based on the maximal information coefficient to detect dependence between two random variable and we perform a power analysis for these different tests.

The remainder of the article is as follows. Section 2 details the various approaches used to test for independence including the classical tests, modern tests and area under Kendall plot. Section 3 outlines the components of the package **testforDEP** and how to use the package **testforDEP** to detect dependence. Section 4 provides power analysis for the different tests and Section 5 gives the package availability. Section 6 offers an example analysis based on real data. Finally, Section 7 provides a brief summary and future directions.

2 Tests for independence

Independence and dependence are key concepts related to many statistical procedures. We will focus on tests of bivariate independence of two random variables X and Y for n subjects with observations $(X_1, X_2, \ldots, X_n)$ and $(Y_1, Y_2, \ldots, Y_n)$, respectively. We are interested in testing whether X and Y are independent. If X and Y have cumulative distribution functions $F_X(x)$ and $F_Y(y)$ and probability density functions $f_X(x)$ and $f_Y(y)$ we say X and Y are independent if and only if the joint random variable (X, Y) has a joint cumulative distribution function $F(x, y) = F_X(x)F_Y(y)$ or equivalently, if the joint density exists, $f(x, y) = f_X(x)f_Y(y)$ for all $x, y \in \mathbb{R}$. We are interested in testing the null hypothesis:

$$H_0 : X \text{ and } Y \text{are independent}, \tag{1}$$

which, in most cases, is equivalent to

$$H_0 : F(x, y) = F_X(x)F_Y(y) \text{ for all } x, y \in \mathbb{R}. \tag{2}$$

In the following subsections we outline the classical and modern tests for independence.

2.1 Pearson product moment correlation coefficient

The Pearson product-moment correlation coefficient γ is a measure of linear correlation (or dependence) between two random variables (Pearson, 1920; Hauke and Kossowski, 2011) defined as:

$$\gamma = \frac{cov(X, Y)}{\sigma_X \sigma_Y}, \tag{3}$$

where, $cov(X, Y)$ is the covariance of X and Y, σ_X is the standard deviation of X, σ_Y is the standard deviation of Y. Note, by definition, $\gamma \in [-1, 1]$.

An estimator for γ, $\hat{\gamma}$ is,

$$\hat{\gamma} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}, \tag{4}$$

where $\bar{X}$ and $\bar{Y}$ are the sample means of the observed values $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$, respectively.

Since $\gamma = 0$ implies linear independence, testing the null hypothesis in (1) is equivalent to testing:

$$H_0 : \gamma = 0. \tag{5}$$

We use the statistic T_γ defined as:

$$T_\gamma = \hat{\gamma}\sqrt{\frac{n-2}{1-\hat{\gamma}^2}}, \tag{6}$$

where T_γ asymptotically follows a t distribution with $n-2$ degrees of freedom under null hypothesis in (5). Accordingly, a size α rejection rule is:

$$|T_{\hat{\gamma}}| > t_{1-\alpha/2}, \tag{7}$$

where $t_{\alpha/2}$, is the $\alpha/2$ quantile for t distribution with $n-2$ degree of freedom. Note that package **testforDEP** returns $T_{\hat{\gamma}}$ as test statistic instead of $\hat{\gamma}$.

2.2 Kendall rank correlation coefficient

The Kendall rank correlation coefficient τ was proposed in Kendall (1938) as a nonparametric measure of monotonic association between two variables. In cases of no ties in the variables X and Y, statistic τ can be defined as:

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\frac{1}{2}n(n-1)}, \tag{8}$$

where two pairs of data points (X_i, Y_i) and (X_j, Y_j) are concordant if and only if $\{X_i > X_j \ , \ Y_i > Y_j\}$ or $\{X_i < X_j \ , \ Y_i < Y_j\}$. Note (X_i, Y_i) and (X_j, Y_j) are discordant if and only if $\{X_i > X_j \ , \ Y_i < Y_j\}$ or $\{X_i < X_j \ , \ Y_i > Y_j\}$. The range for τ is $[-1, 1]$. If X and Y are independent, τ is expected to be approximately 0. When ties are present, the formula for τ is more complicated, see (**?**).

The test statistic Z_τ is:

$$Z_\tau = \frac{3\tau\sqrt{n(n-1)}}{\sqrt{2(2n+5)}} \tag{9}$$

where Z_τ approximately follows a standard normal distribution under null hypothesis. A level α rejection rule for the null hypothesis is as follows:

$$|Z_\tau| > z_{1-\alpha/2} \tag{10}$$

where $z_{\alpha/2}$ is $\alpha/2$ quantile for a standard normal distribution.
Note that package **testforDEP** returns Z_τ as test statistic instead of τ.

2.3 Spearman rank correlation coefficient

Spearman's rank correlation coefficient ρ proposed by Spearman (1904) is a nonparametric measure of statistical dependence between two variables. Spearman's rank correlation measures the monotonic association between two variables (Spearman, 1904; Hauke and Kossowski, 2011). The statistic ρ is defined as:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} {d_i}^2}{n(n^2-1)}, \quad d_i = R_i - S_i, \tag{11}$$

where R_i, S_i are the ranks of X_i and Y_i, respectively. When $\rho > 0$ it suggests a monotonically increasing association in data while ρ ¡ 0 represents a monotonically decreasing association. To conduct a test for statistical independence we use the test statistic T_ρ defined as:

$$T_\rho = \rho\sqrt{\frac{n-2}{1-\rho^2}}, \tag{12}$$

which is distributed approximately as a Student's t distribution with $n-2$ degrees of freedom under the null hypothesis. Accordingly, a level α rejection rule is:

$$|T_\rho| > t_{1-\alpha/2}. \tag{13}$$

Note that package **testforDEP** returns T_ρ as test statistic instead of ρ.

2.4 Density-based empirical likelihood ratio test for independence

A density-based empirical likelihood ratio test was proposed in Vexler et al. (2014) as a nonparametric test of dependence of two variables. The likelihood

ratio $\lambda(\underline{x})$ is the ratio of supremums of two likelihood functions defined as:

$$\lambda(\underline{x}) = \frac{\sup\{L(\theta|\underline{x}) : \theta \in \Theta\}}{\sup\{L(\theta|\underline{x}) : \theta \in \Theta_0\}}, \quad \underline{x} = (x_1, \ldots, x_n), \tag{14}$$

where Θ is the unrestricted parameter space, Θ_0 is a subset of the Θ under the null hypothesis, θ denotes a member of the parameter space and $\underline{x}$ is the vector of observed values $(X_1, X_2, \ldots, X_n)$ with $L(\theta|\underline{x})$ being the value of the likelihood function at a given $\underline{x}$.

Empirical likelihood (EL) was introduced as nonparametric alternatives to parametric likelihood methods. A thorough introduction to empirical likelihood methods can be found in Owen (2001). The derivation of the empirical likelihood ratio test for independence can be found in Vexler, Tsai, and Hutson (2014). The general form likelihood ratio test statistic defined in (14) can be adapted for testing independence as

$$\begin{aligned}
\lambda(\underline{x}, \underline{y}) &= \frac{\sup\{L(\theta|\underline{x}, \underline{y}) : \theta \in \Theta\}}{\sup\{L(\theta|\underline{x}, \underline{y}) : \theta \in \Theta_0\}} \\
&= \prod_{i=1}^{n} \frac{f_{XY}(X_i, Y_i)}{f_X(X_i) f_Y(Y_i)} \\
&= \prod_{i=1}^{n} \frac{f_{XY}(X_{t(i)}, Y_{(i)})}{f_X(X_{t(i)}) f_Y(Y_{(i)})} \\
&= \prod_{i=1}^{n} \frac{f_{Y|X}(Y_{(i)} \mid X_{t(i)})}{f_Y(Y_{(i)})},
\end{aligned} \tag{15}$$

where $Y_{(1)} \leq Y_{(2)}, \ldots, \leq Y_{(n)}$ are the order statistics of $Y_1, Y_2, \ldots Y_n$ and $X_{t(i)}$ is the concomitant of the ith order statistic defined in David and Nagaraja (1970). Since in general the joint and marginal density functions in (15) are unknown, our focus is on the nonparametric approximation to the likelihood function by applying the density-based empirical likelihood methodology. Following Vexler et al. (2014), to test the null hypothesis in (1) we use the test statistic VT_n defined as:

$$VT_n = \max_{0.5n^{\beta_2} \leq m \leq \gamma_n} \max_{0.5n^{\beta_2} \leq r \leq \gamma_n} \prod_{i=1}^{n} \frac{n\widetilde{\Delta}_i(m, r)}{2m}, \tag{16}$$

where $\gamma_n = \min(n^{0.9}, n/2), 0.75 < \beta_2 < 0.9$ and $\widetilde{\Delta}_i(m, r)$ is defined as,

$$\widetilde{\Delta}_i(m, r) \equiv \frac{\hat{F}(X_{(s_i+r)}, Y_{(i+m)}) - \hat{F}(X_{(s_i-r)}, Y_{(i+m)}) - \hat{F}(X_{(s_i+r)}, Y_{(i-m)}) + \hat{F}(X_{(s_i-r)}, Y_{(i-m)}) + n^{-\beta_1}}{\hat{F}_X(X_{(s_i+r)}) - \hat{F}_X(X_{(s_i-r)})}, \tag{17}$$

where $\hat{F}$ is the empirical joint distribution of (X, Y), $\hat{F}_X$ is the empirical marginal distribution of X and s_i is the integer number such that $X_{(s)} = X_{t(i)}$, $\beta_1 \in (0, 0.5)$.
The statistic VT_n reaches its maximum with respect to $m \geq 0.5n^{\beta_2}$ and $r \geq$

$0.5n^{\beta_2}$ at $m = 0.5n^{\beta_2}$ and $r = 0.5n^{\beta_2}$ (Vexler et al., 2014). Thus, we simplify (16) to obtain

$$VT_n = \prod_{i=1}^{n} n^{1-\beta_2} \widetilde{\Delta}_i \left([0.5n^{\beta_2}], [0.5n^{\beta_2}] \right), \tag{18}$$

where the function $[x]$ denotes the nearest integer to x.
Accordingly, a size α rejection rule of the test is:

$$\log(VT_n) > C_\alpha, \tag{19}$$

where C_α is an α-level test threshold. It is established in Vexler et al. (2014) that VT_n is distribution free under (1) and the critical values C_α can be estimated by Monte Carlo simulations from $X_1, \ldots, X_n \sim Uniform[0,1]$ and $Y_1, \ldots, Y_n \sim Uniform[0,1]$. Note that package **testforDEP** returns $\log(VT_n)$ as test statistic.

2.5 Kallenberg data-driven test for independence

Kallenberg and Ledwina (1999) propose two data-driven rank tests for independence based on statistics $TS2$ and V. The $TS2$ statistic is derived from the intermediate statistic T_k where:

$$T_k = \sum_{j=1}^{k} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} b_j \left(\frac{R_i - \frac{1}{2}}{n} \right) b_j \left(\frac{S_i - \frac{1}{2}}{n} \right) \right\}^2, \tag{20}$$

where b_j denotes the jth orthonormal Legendre polynomial. The selection of the order k in T_k is done by a modified Schwarz's rule given by

$$S2 = \min\{1 \le k \le d(n), T_k - k \log n \ge T_j - j \log n,\ j = 1, \ldots, d(n)\}, \tag{21}$$

where $d(n)$ is a sequence of numbers tending to infinity as $n \to \infty$. The data-driven smooth test statistic for testing the null hypothesis in (1) is,

$$TS2 = \sum_{j=1}^{S2} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} b_j \left(\frac{R_i - \frac{1}{2}}{n} \right) b_j \left(\frac{S_i - \frac{1}{2}}{n} \right) \right\}^2. \tag{22}$$

It was found in Kallenberg and Ledwina (1999) that there is almost no change in the critical value of $TS2$ for $d(n) > 2$. By default, we choose $d(n) = 4$. The decision rule to reject the null hypothesis in (2) is

$$TS2 > C_\alpha, \tag{23}$$

where C_α is an α-level test threshold.

In Kallenberg and Ledwina (1999) the $TS2$ test statistic in (22) was called the "diagonal" test statistic. The other test statistic, V was called the "mixed" statistic due to the fact that it involves "mixed" products. To derive the V

statistic, we only consider the case $d(n) = 2$ and have sets of indexes $\{(1,1)\}$, $\{(1,1),(i,j)\}$, where $(i,j) \neq (1,1)$. Let Λ be one of these sets and define

$$T_\Lambda = \sum_{(r,s)\in\Lambda} V(r,s), \tag{24}$$

where

$$V(r,s) = \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} b_r \left(\frac{R_i - \frac{1}{2}}{n} \right) b_s \left(\frac{S_i - \frac{1}{2}}{n} \right) \right\}^2. \tag{25}$$

Letting $|\Lambda|$ denote the cardinality of Λ, we now search for a model, say $\Lambda^{(max)}$, for which $T_\Lambda - |\Lambda| \log n$ is maximized. Having obtained $\Lambda^{(max)}$, the mixed data-driven test statistic is

$$V = T_{\Lambda^{(max)}}. \tag{26}$$

It can be easily seen that the test statistic V can be rewritten (for $d(n) = 2$) in the simple form,

$$V = \begin{cases} V(1,1) & \text{, if} \max\Big\{V(1,2), V(2,1), V(2,2)\Big\} < \log n \\ V(1,1) + \max\Big\{V(1,2), V(2,1), V(2,2)\Big\} & \text{, otherwise.} \end{cases}$$

A size α rejection rule for the mixed test is

$$V > C_\alpha, \tag{27}$$

where C_α is a size α critical value. We note that Kallenberg and Ledwina (1999) develop their test under the assumption that the observed samples are distributed following a joint distribution function that belongs to the exponential family. Thus while still technically distribution-free, this assumption may limit the power of their tests for certain alternatives.

2.6 Maximal Information Coefficient

The maximal information coefficient (MIC) proposed in Reshef et al. (2011) is a measure of the strength of the linear and non-linear association between two variables X and Y. The maximal information coefficient uses binning as a means to apply mutual information on continuous random variables. Defining $\mathcal{D}$ as a finite set of ordered pairs, we can partition the x-values of $\mathcal{D}$ into x bins and the y-values of $\mathcal{D}$ into y bins, allowing empty bins. We call such partition an x-by-y grid, denoted G. For a fixed $\mathcal{D}$, different grids G results in different distributions $\mathcal{D}|G$. The MIC of a set $\mathcal{D}$ of two-variable data with sample size n and grid size less than $B(n)$ is defined as:

$$MIC(\mathcal{D}) = \max_{xy<B(n)} \{M(\mathcal{D})_{x,y}\}, \tag{28}$$

where x and y are observed values of variables X and Y, $\omega_{(1)} < B(n) \leq o(n^{1-\varepsilon})$ for some $0 < \varepsilon < 1$ (see Reshef et al. (2011) for more details). Note

$M(\mathcal{D})_{x,y}$ is called the characteristic matrix of a set $\mathcal{D}$ of two-variable data x, y and defined as:

$$M(\mathcal{D})_{x,y} = \frac{I^*(\mathcal{D}, x, y)}{\log \min\{x, y\}}, \tag{29}$$

where $I^*(\mathcal{D}, x, y)$ is defined as:

$$I^*(\mathcal{D}, x, y) = \max I(\mathcal{D}|G), \tag{30}$$

for a finite set $\mathcal{D} \subset \mathbb{R}^2$ and positive integers x, y. We denote $I(\mathcal{D}|G)$ as the mutual information of $D|G$, the expected values of the point-wise mutual information (PMI) and define the mutual information of two discrete random variables X and Y as:

$$I(\mathcal{D}|G) = \sum_{y \in Y} \sum_{x \in X} \hat{F}(x, y) \log \left(\frac{\hat{F}(x, y)}{\hat{F}_X(x) \hat{F}_Y(y)} \right). \tag{31}$$

The point-wise mutual information (PMI) is $\log \left(\frac{\hat{F}(x,y)}{\hat{F}_X(x)\hat{F}_Y(y)} \right)$. The statistic PMI is 0 when X and Y are independent. The statistic MIC is computed using the **R** package **minerva** (see Filosi et al. (2014)). To our knowledge there is no hypothesis test based on MIC for detecting the general structure of dependence. We use an approach similar to the one in Simon and Tibshirani (2014) to develop an exact test based on the MIC statistic. A size α MIC rejection rule is,

$$MIC(\mathcal{D}) > C_\alpha \tag{32}$$

where C_α is a size α critical value.

To evaluate our approach, we simulate 5000 Monte Carlo sets of independent random variables X and Y of size n from a standard normal distribution, exponential distribution and reverse of the standard normal distribution. The C_α cutoffs are in Table 1. Regardless of the data distribution the cutoff values for a given sample size are very similar indicating that an Monte Carlo approach to determine the cutoff is reasonable for several different types of dependence.

Sig level \ n	10	35	75	100
	N — E — RN	N — E — RN	N — E — RN	N — E — RN
0.01	0.61—0.61—0.61	0.50—0.52—0.50	0.38—0.37—0.38	0.33—0.33—0.33
0.05	0.61—0.61—0.61	0.43—0.43—0.43	0.33—0.33—0.33	0.30—0.30—0.30
0.1	0.61—0.61—0.61	0.41—0.40—0.41	0.31—0.31—0.31	0.28—0.28—0.28
0.5	0.24—0.24—0.24	0.31—0.30—0.31	0.25—0.25—0.25	0.23—0.24—0.23
0.75	0.24—0.24—0.24	0.27—0.26—0.27	0.23—0.23—0.23	0.21—0.21—0.21
0.9	0.12—0.12—0.12	0.24—0.23—0.24	0.21—0.21—0.21	0.20—0.20—0.20
0.95	0.12—0.12—0.12	0.23—0.22—0.23	0.20—0.20—0.20	0.19—0.19—0.19
0.99	0.11—0.11—0.11	0.20—0.20—0.20	0.18—0.18—0.18	0.17—0.17—0.17

Table 1: Cutoff values from 5000 Monte Carlo simulations for a normal distribution (N), exponential distribution (E), and reverse normal distribution (RN).

2.7 Hoeffding's test for independence

Hoeffding's test for dependence was proposed in Hoeffding (1948) as a test for two random variables with continuous distribution functions (see also Harrell Jr and Dupont (2006)). Hoeffding's D is a nonparametric measure of the distance between the joint distribution $F(x, y)$ and the product of marginal distributions $F_X(x)F_Y(y)$. The coefficient D is:

$$D(x, y) = F(x, y) - F_X(x)F_Y(y). \tag{33}$$

We then define Δ as:

$$\Delta = \Delta(F) = \int D^2(x, y)dF(x, y). \tag{34}$$

We let Ω' be the class of (X, Y)'s where the joint density function $F(x, y)$ is continuous and we let Ω'' be the class of (X, Y)'s where the joint and marginal probability density functions are continuous. It has been shown that if $F(x, y)$ belongs to Ω'', $\Delta(F) = 0 \iff D(x, y) = 0$. The random variables X and Y are independent if and only if $D(x, y) = 0$.

It can be shown (see Harrell Jr and Dupont (2006)) that D is such that

$$-\frac{1}{60} \leq D \leq \frac{1}{30}, \tag{35}$$

where larger value of D suggest dependence. An estimator of D, $\hat{D}$ is defined as:

$$\hat{D}(x, y) = \hat{F}(x, y) - \hat{F}_X(x)\hat{F}_Y(y). \tag{36}$$

To implement the Hoeffding test we use $\hat{D}$ as test statistic and the **R** package **Hmisc** developed by Harrell Jr and Dupont (2006).

Note the test statistic $\hat{D}'$ returned by **Hmisc** is 30 times the original $\hat{D}$ in Hoeffding (1948). That makes $\hat{D}'$ range from -0.5 to 1 with a size α test given by:

$$\hat{D}' > C_\alpha, \tag{37}$$

where C_α is a size α critical value.

The manuscript for Hoeffding's test was published in 1948. Due to the limiting computing tools, the author only provided cutoff tables for small sample sizes. With advanced computing power and algorithms, we compute the C_α cutoffs for $n = 10, 20, 50, 100, 200, 500$. The results are shown in Table 2.

2.8 Empirical Likelihood based test for independence

Another empirical likelihood based test for independence was proposed by Einmahl and McKeague (2003). It is a nonparametric test for two variables. The main approach is localizing the empirical likelihood with one or more 'time' variables implicit in the given null hypothesis and construct an omnibus test

Sig level \ n	10	20	50	100	200	500
0.01	0.2976	0.1145	0.0377	0.0191	0.0086	0.0037
0.05	0.1548	0.0589	0.0209	0.0098	0.0045	0.0020
0.1	0.0952	0.0378	0.0131	0.0061	0.0028	0.0013
0.2	0.0437	0.0184	0.0060	0.0028	0.0011	0.0006
0.8	-0.0635	-0.0232	-0.0078	-0.0037	-0.0018	-0.0007
0.9	-0.0794	-0.0294	-0.0097	-0.0045	-0.0021	-0.0008

Table 2: Hoeffding cutoff values for $\hat{D}'$ based on 5000 Monte Carlo simulations.

statistic by integrating the log-likelihood ratio over those variables. We outline the localization approach in Einmahl and McKeague (2003) where we first consider a null hypothesis,

$$H_0 : F_X = F_0, \tag{38}$$

where F_0 is a fully specified distribution function then we define the localized empirical likelihood ratio $R(x)$ as:

$$R(x) = \frac{sup\{L(\tilde{F_X}) : \tilde{F_X}(x) = F_0(x)\}}{sup\{L(\tilde{F_X})\}}, \tag{39}$$

where $\tilde{F}$ is an arbitrary distribution function, $L(\tilde{F_X}) = \prod_{i=1}^{n}(\tilde{F_X}(X_i) - \tilde{F_X}(X_i-))$. The supremum in the denominator is achieved when $\tilde{F} = \hat{F}$, the empirical distribution function. The supremum in the numerator is attained by putting mass $F_0/(n\hat{F}(x))$ on each observation up to and including x and mass $(1 - F_0(x))/(n(1 - \hat{F}(x)))$ on each observation beyond x (Einmahl and McKeague, 2003). This gives the log localized empirical likelihood ratio:

$$\log R(x) = nF(x) \log \frac{F_0(x)}{\hat{F}(x)} + n(1 - \hat{F}(x)) \log \frac{1 - F_0(x)}{1 - \hat{F}(x)}. \tag{40}$$

Note that $0 \log(a/0)$ is defined to be 0.

For a test of independence the local likelihood ratio test statistic is:

$$R(x, y) = \frac{sup\{L(\tilde{F}) : \tilde{F}(x, y) = \tilde{F}_X(x)\tilde{F}_Y(y)\}}{sup\{L(\tilde{F})\}}, \tag{41}$$

for $(x, y) \in \mathbb{R}^2$, with $L(\tilde{F}) = \prod_{i=1}^{n} \tilde{P}(\{X_i\})$, where $\tilde{P}$ is the probability measure

corresponding to $\tilde{F}$. Log local likelihood ratio test statistic is then:

$$\begin{aligned}\log R(x,y) =& n\hat{P}(A_{11}) \log \frac{\hat{F}_X(x)\hat{F}_Y(y)}{\hat{P}(A_{11})} + \\ & n\hat{P}(A_{12}) \log \frac{\hat{F}_X(x)(1-\hat{F}_Y(y))}{\hat{P}(A_{12})} + \\ & n\hat{P}(A_{21}) \log \frac{(1-\hat{F}_X(x))\hat{F}_Y(y)}{\hat{P}(A_{21})} + \\ & n\hat{P}(A_{22}) \log \frac{(1-\hat{F}_X(x))(1-\hat{F}_Y(y))}{\hat{P}(A_{22})},\end{aligned} \tag{42}$$

where $\hat{P}$ is the empirical measure of joint probability, and

$$\begin{aligned} A_{11} &= (-\infty, x] \times (-\infty, y], \\ A_{12} &= (-\infty, x] \times (y, \infty), \\ A_{21} &= (x, \infty) \times (-\infty, y], \\ A_{22} &= (x, \infty) \times (y, \infty). \end{aligned} \tag{43}$$

The test statistic T_{el} is defined as:

$$T_{el} = -2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \log R(x,y) d\hat{F}_X(x)\hat{F}_Y(y). \tag{44}$$

T_{el} is clearly distribution-free. We reject H_0 in (1) when

$$T_{el} > C_\alpha, \tag{45}$$

where C_α is a size α critical value.

2.9 Kendall Plot and Area Under Kendall Plot

The Kendall plot, also called K-plot, is a visualization of dependence in a bivariate random sample. It was proposed in Genest and Boies (2012). Similar to a Chi-plot which detects association in random samples from continuous bivariate distributions, the K-plot adapts the concept of a probability plot to detect dependence. K-plot deals with rank orders of samples thus, like a Chi-plot, it is invariant to monotone transformations of the samples. The horizontal axis in a K-plot $W_{i:n}$ is the expectation of the ith order statistic in a random sample of size n from the distribution K_0 of the H_i under the null hypothesis in (1). $W_{i:n}$ can be computed as:

$$W_{i:n} = n \binom{n-1}{i-1} \int_0^1 w\{K_0(w)\}^{i-1} \times \{1-K_0(w)\}^{n-i} dK_0(w), \tag{46}$$

for all $1 \leq i \leq n$. Note K_0 is given as:

$$K_0(w) = w - w\log(w), 0 \leq w \leq 1. \tag{47}$$

The vertical axis is the sorted H_i where H_i is:

$$H_i = \frac{1}{n-1} \sum_{i=1}^{n} Ind(X \leq X_i, Y \leq Y_i), \tag{48}$$

where Ind is the indicator function. Note that H_i is similar to $\hat{F}(x, y)$. Let $0 \leq p \leq 1$, $p \in \mathbb{R}$ and note some properties of K-plot are:

1. When $Y = X$, all points fall on curve $K^{-1}(p)$,
2. When $Y = -X$, all points fall on a line with slope$= 0$,
3. When X and Y are independent, the graph is linear.

The area under the Kendall plot (AUK) is proposed in Vexler et al. (2015) as an index to evaluate Kendall plots. It applies area-under-curve analysis and computes the area under the Kendall plot as a measure of independence. Some properties of AUK are listed below:

1. When $Y = X$, $AUK = 0.75$,
2. When $Y = -X$, $AUK = 0$,
3. When X and Y are independent, $AUK = 0$.

where (2) and (3) taken together shows that $AUK = 0$ does not imply independence.

3 What is the package testforDEP

The **R** Package **testforDEP** includes the functions `testforDEP()` and `AUK()`, one data frame `LSAT`, and implements the procedures described in Section 2.

3.1 testforDEP

The function `testforDEP()` is the interface that implements each of the following tests: Pearson test, Kendall test, Spearman test, density-based empirical likelihood ratio test ($log(VT_n)$), Kallenberg data-driven test ($TS2$, and V), maximal information coefficient test (MIC), Hoeffding's test, and empirical likelihood based test (T_{el}). The function `testforDEP()` takes two vectors X, Y as input and returns an S4 object of class `testforDEP_result` which contains:

1. test statistic,
2. p-value,
3. bootstrap confidence intervals.

Interface of `testforDEP()` is:

```
testforDEP(x, y, data, test = c( "PEARSON", "KENDALL", "SPEARMAN,
"VEXLER", "TS2", "V", "MIC", "HOEFFD", "EL", "CANOVA"),
p.opt = c("dist", "MC", "table"), num.MC = 10000,
BS.CI = 0, rm.na = FALSE, set.seed = FALSE)
```

where `x` and `y` are two equal-length numeric vectors of input data. The input `data` is an alternative that takes a data frame with two columns representing X and Y. When `x` or `y` are not provided the parameter `data` is taken as input. The parameter `test` specifies the hypothesis test to implement. Note that `"VEXLER"` refers to $log(VT_n)$ test, `"HOEFFD"` refers to Hoeffding's test, `"EL"` refers to T_{el} test. The parameter `p.opt` is the option for computing p-values in which p-values can be computed from the (asymptotic) null distribution of the test statistic (applicable for Pearson, Kendall, and Spearman only) or by an exact Monte Carlo (MC) method (applicable for all tests), or by pre stored MC simulated tables derived by the exact method. By default, `p.opt = "MC"`. Parameter `num.MC` gives the Monte Carlo simulation number and will only be taken when `p.opt = "MC"`. When `p.opt = "dist"` or `p.opt = "table"`, `num.MC` will be ignored. To balance accuracy and computation time `num.MC` must $\in [100, 10000]$ with `num.MC = 10000` as default. Parameter `BS.CI` specifies α for an α bootstrap confidence intervals. The normal, percentile, and pivotal bootstrap intervals are produced except when `BS.CI = 0` then confidence intervals will not be computed. Parameter `rm.na` is a flag for removing rows with missing data. Parameter `set.seed` is a flag for setting seed.

3.2 AUK

The function `AUK()` is the interface for Kendall plots and AUK. It takes two vectors X, Y and returns a list containing:

1. AUK,
2. $W_{i:n}$,
3. sorted H_i,
4. bootstrap confidence intervals.

The interface of `AUK()` is:

```
AUK(x, y, plot = FALSE, main = "Kendall plot", Auxiliary.line = TRUE,
BS.CI = 0, set.seed = FALSE)
```

where `x` and `y` are two equal-length numeric vectors of input data. The parameter `plot` is a flag for drawing Kendall plot. Parameter `main` determines the title of the plot. If `plot = FALSE`, `main` will be ignored. Parameter `Auxiliary.lie` is a flag for auxiliary line. Parameter `BS.CI` specifies α for α bootstrap confidence intervals, e.g. $\alpha = 0.95$ will produce a 95 percent confidence interval. When `BS.CI = 0`, confidence intervals will not be computed. Parameter `set.seed` is a flag for setting seed.

3.3 LSAT

The data frame `LSAT` contains the data analysis example. See Section 6 for details.

4 Power Analysis

We conduct a simulation study to evaluate the power for the tests in **testforDEP** package. We simulate data under various alternative hypotheses of dependence. Simulations are divided into 3 groups: group 1: non-linear correlation; group 2: linear correlation; and group 3: other bivariate distributions including Pearson Type VII, Morgenstern, Plackett and Cauchy distribution. Both group 1 and group 2 include random-effect models. Table 3 shows details of the designs. For each design, test, and sample size, a simulation of 5000 MCs is performed to estimate the power. Table 4 shows our results.

Alternative Designs		Model Description	
		$X_i, i = 1...n$	$Y_i, i = 1...n$
Group 1	Design 1.1	$N(0,1)$	$1+0.2*X_i + 0.8 * {X_i}^2 + \epsilon_i$
	Design 1.2	$N(0,1)$	$0.5+0.1*X_i + {X_i}^2 + \gamma_i * X_i + \epsilon_i$
	Design 1.3	$N(0,1)$	$log(1 + \lvert X_i \rvert)$
(non-linear)	Design 1.4	$N(0,1)$	$log(1 + \lvert X_i \rvert) * \gamma_i$
	Design 1.5	$N(0,1)$	$2+0.1*\epsilon_i/X_i$
	Design 1.6	$N(0,1)$	$1/X_i$
	Design 1.7	$N(0,1)$	$1/X_i^2$
Group 2	Design 2.1	$Lognormal(0,1)$	$1+\gamma_i * X_i$, $\gamma_i \sim N(0,1)$
	Design 2.2	$Lognormal(0,1)$	$1+4 * \gamma_i * X_i + 0.1 * X_i + \epsilon_i$, γ_i, $\epsilon_i \sim N(0,1)$
	Design 2.3	$N(0,1)$	$2+0.1*X_i + \epsilon_i$, $\epsilon_i \sim N(0,1)$
	Design 2.4	$U(0,1)$	$2+0.5*X_i + \epsilon_i$, $\epsilon_i \sim N(0,1)$
	Design 2.5	$U(0,1)$	$2+0.5*X_i + \epsilon_i + \gamma_i * X_i$, $\epsilon_i \sim N(0,1)$, $\gamma_i \sim N(0,2^2)$
(linear)	Design 2.6	$U(0,1)$	$2+X_i + \epsilon_i$, $\epsilon_i \sim N(0,1)$
	Design 2.7	$U(0,1)$	$2+X_i + \epsilon_i + \gamma_i * X_i$, $\epsilon_i \sim N(0,1)$, $\gamma_i \sim N(0,2^2)$
Group 3	Design 3.1	$Morgenstern(\alpha = 1)$	Johnson (1987),pp.180-190
	Design 3.2	$Plackett(\psi = 3.5)$	Johnson (1987) ,pp.191-197
(bivariate)	Design 3.3	Pearson Type VII	Johnson (1987),pp.117-121
	Design 3.4	The multivariate Cauchy distribution	Johnson (1987),pp.44

Table 3: Distributions of X and Y in each design.

Tests	Design 1.1						Design 1.2					
	Sample size (n)						Sample size (n)					
	20	25	30	35	50	70	20	25	30	35	50	70
Pearson	0.25	0.29	0.29	0.3	0.33	0.36	0.24	0.28	0.28	0.29	0.28	0.3
Kendall	0.13	0.15	0.15	0.17	0.18	0.23	0.1	0.13	0.12	0.14	0.12	0.14
Spearman	0.13	0.13	0.14	0.15	0.17	0.20	0.10	0.10	0.10	0.12	0.10	0.12
$log(VT_n)$	0.35	0.47	0.57	0.68	0.84	0.96	0.35	0.46	0.56	0.67	0.84	0.96
TS_2	0.13	0.16	0.19	0.23	0.35	0.56	0.16	0.21	0.25	0.32	0.51	0.74
V	0.6	0.74	0.84	0.91	0.98	1	0.54	0.66	0.77	0.85	0.97	0.99
MIC	0.25	0.34	0.39	0.42	0.64	0.78	0.23	0.29	0.34	0.35	0.53	0.65
Hoeffding	0.21	0.29	0.36	0.45	0.67	0.89	0.18	0.25	0.28	0.36	0.54	0.78
T_{el}	0.26	0.41	0.52	0.67	0.87	0.99	0.21	0.31	0.41	0.55	0.76	0.93
Tests	Design 1.3						Design 1.4					

	Sample size (n)						Sample size (n)					
	20	25	30	35	50	70	20	25	30	35	50	70
Pearson	0.19	0.19	0.19	0.19	0.18	0.19	0.19	0.19	0.2	0.19	0.19	0.21
Kendall	0.21	0.24	0.22	0.25	0.23	0.25	0.13	0.13	0.12	0.13	0.13	0.14
Spearman	0.14	0.14	0.13	0.15	0.13	0.14	0.10	0.10	0.10	0.10	0.10	0.10
$log(VT_n)$	1	1	1	1	1	1	0.32	0.41	0.54	0.66	0.86	0.97
TS_2	0.18	0.19	0.21	0.21	0.26	0.41	0.41	0.56	0.69	0.79	0.94	0.99
V	1	1	1	1	1	1	0.42	0.55	0.65	0.75	0.92	0.99
MIC	1	1	1	1	1	1	0.09	0.09	0.11	0.11	0.12	0.15
Hoeffding	1	1	1	1	1	1	0.12	0.13	0.14	0.14	0.18	0.22
T_{el}	1	1	1	1	1	1	0.11	0.12	0.14	0.15	0.18	0.26
Tests	Design 1.5						Design 1.6					
	Sample size (n)						Sample size (n)					
	20	25	30	35	50	70	20	25	30	35	50	70
Pearson	0	0	0	0	0	0	0.05	0.05	0.05	0.05	0.05	0.0
Kendall	0	0	0	0	0	0	0.01	0	0.01	0	0	0
Spearman	0.01	0.01	0.01	0.01	0.01	0.01	0.74	0.77	0.90	0.96	1	1
$log(VT_n)$	0.14	0.12	0.24	0.38	0.76	0.93	1	1	1	1	1	1
TS_2	0.42	0.66	0.82	0.91	0.99	1	0.89	0.96	0.96	0.99	1	1
V	0.4	0.61	0.78	0.89	0.99	1	1	1	1	1	1	1
MIC	0.11	0.13	0.13	0.14	0.18	0.29	1	1	1	1	1	1
Hoeffding	0.03	0.04	0.05	0.06	0.11	0.31	0.26	0.69	1	1	1	1
T_{el}	0.03	0.03	0.05	0.08	0.17	0.56	1	1	1	1	1	1
Tests	Design 1.7						Design 2.1					
	Sample size (n)						Sample size (n)					
	20	25	30	35	50	70	20	25	30	35	50	70
Pearson	0	0	0	0	0	0	0.45	0.48	0.5	0.51	0.54	0.56
Kendall	0.21	0.24	0.22	0.25	0.23	0.25	0.09	0.1	0.1	0.11	0.1	0.1
Spearman	0.14	0.14	0.13	0.15	0.13	0.14	0.08	0.09	0.10	0.10	0.09	0.09
$log(VT_n)$	1	1	1	1	1	1	0.51	0.64	0.77	0.86	0.97	1
TS_2	0.18	0.19	0.21	0.21	0.26	0.41	0.06	0.05	0.07	0.06	0.07	0.07
V	1	1	1	1	1	1	0.78	0.89	0.96	0.99	1	1
MIC	1	1	1	1	1	1	0.45	0.59	0.68	0.72	0.9	0.98
Hoeffding	1	1	1	1	1	1	0.3	0.46	0.6	0.73	0.94	1
T_{el}	1	1	1	1	1	1	0.33	0.55	0.73	0.88	0.99	1
Tests	Design 2.2						Design 2.3					
	Sample size (n)						Sample size (n)					
	20	25	30	35	50	70	20	25	30	35	50	70
Pearson	0.46	0.47	0.5	0.51	0.54	0.57	0.07	0.07	0.09	0.09	0.1	0.13
Kendall	0.09	0.1	0.1	0.11	0.1	0.1	0.07	0.07	0.07	0.08	0.09	0.12
Spearman	0.09	0.09	0.09	0.10	0.09	0.09	0.07	0.07	0.08	0.08	0.10	0.13
$log(VT_n)$	0.44	0.57	0.69	0.79	0.93	0.99	0.05	0.06	0.06	0.06	0.06	0.07
TS_2	0.06	0.06	0.07	0.08	0.09	0.11	0.06	0.06	0.06	0.06	0.08	0.09
V	0.71	0.85	0.93	0.98	1	1	0.06	0.06	0.06	0.06	0.07	0.08

MIC	0.39	0.52	0.59	0.64	0.85	0.95	0.05	0.06	0.06	0.07	0.06	0.06
Hoeffding	0.25	0.37	0.48	0.62	0.86	0.98	0.06	0.06	0.07	0.07	0.09	0.11
T_{el}	0.29	0.47	0.65	0.81	0.96	1	0.06	0.06	0.07	0.08	0.09	0.11
Tests	Design 2.4						Design 2.5					
	Sample size (n)						Sample size (n)					
	20	25	30	35	50	70	20	25	30	35	50	70
Pearson	0.09	0.1	0.12	0.13	0.17	0.22	0.08	0.09	0.08	0.09	0.09	0.1
Kendall	0.08	0.1	0.11	0.13	0.14	0.21	0.06	0.08	0.07	0.08	0.08	0.08
Spearman	0.09	0.11	0.12	0.13	0.16	0.21	0.07	0.08	0.07	0.08	0.08	0.09
$log(VT_n)$	0.06	0.07	0.08	0.09	0.09	0.11	0.17	0.21	0.28	0.35	0.49	0.69
TS_2	0.07	0.07	0.08	0.09	0.1	0.14	0.05	0.05	0.05	0.06	0.06	0.06
V	0.07	0.06	0.07	0.08	0.1	0.12	0.25	0.35	0.47	0.57	0.8	0.94
MIC	0.06	0.06	0.07	0.07	0.08	0.07	0.17	0.22	0.23	0.26	0.4	0.51
Hoeffding	0.08	0.09	0.1	0.12	0.14	0.19	0.11	0.14	0.16	0.21	0.32	0.55
T_{el}	0.08	0.09	0.11	0.13	0.14	0.19	0.11	0.16	0.22	0.3	0.47	0.76
Tests	Design 2.6						Design 2.7					
	Sample size (n)						Sample size (n)					
	20	25	30	35	50	70	20	25	30	35	50	70
Pearson	0.22	0.27	0.34	0.38	0.51	0.66	0.1	0.12	0.12	0.13	0.14	0.19
Kendall	0.19	0.26	0.29	0.35	0.46	0.62	0.08	0.1	0.1	0.12	0.13	0.17
Spearman	0.21	0.26	0.31	0.36	0.47	0.63	0.09	0.11	0.11	0.12	0.13	0.17
$log(VT_n)$	0.11	0.15	0.17	0.22	0.24	0.35	0.18	0.23	0.29	0.38	0.52	0.73
TS_2	0.11	0.15	0.18	0.22	0.33	0.49	0.05	0.06	0.07	0.07	0.09	0.1
V	0.12	0.15	0.17	0.2	0.3	0.42	0.26	0.36	0.48	0.58	0.8	0.94
MIC	0.09	0.12	0.13	0.15	0.16	0.21	0.17	0.23	0.25	0.27	0.42	0.53
Hoeffding	0.17	0.22	0.27	0.31	0.42	0.57	0.13	0.17	0.2	0.26	0.39	0.62
T_{el}	0.19	0.24	0.29	0.35	0.45	0.6	0.13	0.19	0.25	0.34	0.53	0.8
Tests	Design 3.1						Design 3.2					
	Sample size (n)						Sample size (n)					
	20	25	30	35	50	70	20	25	30	35	50	70
Pearson	0.3	0.38	0.46	0.52	0.69	0.83	0.42	0.52	0.61	0.68	0.82	0.93
Kendall	0.25	0.34	0.39	0.49	0.64	0.81	0.39	0.5	0.57	0.67	0.81	0.93
Spearman	0.28	0.37	0.43	0.50	0.66	0.82	0.40	0.50	0.58	0.66	0.81	0.93
$log(VT_n)$	0.15	0.21	0.24	0.3	0.37	0.5	0.24	0.35	0.4	0.47	0.59	0.75
TS_2	0.13	0.19	0.24	0.32	0.48	0.7	0.24	0.32	0.41	0.5	0.7	0.86
V	0.14	0.19	0.23	0.28	0.44	0.6	0.25	0.32	0.39	0.47	0.65	0.81
MIC	0.13	0.17	0.19	0.22	0.27	0.33	0.17	0.24	0.28	0.33	0.38	0.48
Hoeffding	0.23	0.31	0.36	0.44	0.6	0.77	0.35	0.45	0.53	0.62	0.79	0.91
T_{el}	0.26	0.34	0.41	0.49	0.63	0.8	0.38	0.47	0.56	0.65	0.78	0.91
Tests	Design 3.3						Design 3.4					
	Sample size (n)						Sample size (n)					
	20	25	30	35	50	70	20	25	30	35	50	70
Pearson	0.74	0.76	0.78	0.81	0.84	0.87	0.59	0.62	0.64	0.68	0.7	0.76
Kendall	0.77	0.88	0.93	0.97	0.99	1	0.12	0.13	0.12	0.13	0.14	0.14
Spearman	0.69	0.81	0.87	0.92	0.98	1.00	0.10	0.10	0.10	0.10	0.11	0.11

$log(VT_n)$	0.8	0.92	0.97	0.99	1	1	0.23	0.3	0.37	0.47	0.62	0.8
TS_2	0.83	0.93	0.98	0.99	1	1	0.43	0.57	0.69	0.79	0.93	0.99
V	0.84	0.93	0.98	0.99	1	1	0.44	0.55	0.67	0.77	0.91	0.98
MIC	0.4	0.56	0.67	0.7	0.83	0.94	0.06	0.06	0.06	0.07	0.06	0.06
Hoeffding	0.78	0.89	0.95	0.98	1	1	0.1	0.11	0.11	0.11	0.14	0.16
T_{el}	0.72	0.83	0.91	0.95	0.99	1	0.12	0.13	0.14	0.16	0.18	0.25

Table 4: The estimated power for the tests in the package **testforDEP** based on 5000 Monte-Carlo simulations.

In group 1, in general, the Kallenberg tests (V and $TS2$ test) have the largest power. As the designs in group 1 are nonlinear, the poor power for the classical tests (Pearson, Spearman, and Kendall) is not surprising. As is well-known, the Pearson product moment correlation coefficient, γ is a measure of the strength of linear relationship between two random variables (Pearson, 1920). The Spearman rank correlation coefficient, ρ is a measure of monotonic association between two random variables (Spearman, 1904; Hauke and Kossowski, 2011). The Kendall rank correlation coefficient, τ measures dependence based on monotonic functions (Wang et al., 2015). These classical tests are not suitable for detecting non-monotonic dependence between two variables. The T_{el} test, Hoeffding test, and MIC test have good performance in designs 1.1-1.3 but poor performance in designs 1.4-1.5. This indicates that those tests perform well for fixed coefficient of x but often fail to detect dependence for random coefficients of x.

In group 2, we can divide the designs into two sub-groups one containing random coefficients on x: designs 2.1, 2.2, 2.5, and 2.7 and those containing fixed x coefficients: designs 2.3, 2.4 and 2.6. For designs with fixed x coefficients the classical methods are more powerful then others. Note in designs 2.3, 2.4 that when the x coefficient increased from 0.1 to 0.5, the classical tests have a larger increase in power than modern tests. For designs with random X coefficients, the results are similar to that in group 1, non-linear designs, where the Kallenberg V test outperforms the others.

Group 3 consists of different bivariate distributions as alternatives to independence. In designs 3.1 and 3.2, the classical tests tend to have the highest power. In designs 3.3 and 3.4, the Kallenberg tests V and $TS2$ dominate all others.

Based on our simulations, we conclude the classical tests are the most powerful in detecting linear or monotonic relationships while $log(VT_n)$ test and V test are more powerful when non-linearity or random effects are involved. The Hoeffding's test and T_{el} test have intermediate power under most of the situations.

5 Availability

The **testforDEP** package is available from the Comprehensive **R** Archive Network at

https://cran.r-project.org/web/packages/testforDEP/index.html

and is also available for download at the author's department webpage:

https://sphhp.buffalo.edu/content/dam/sphhp/biostatistics/Documents/techreports/testforDEP_0.2.0.tar.gz

with a technical report available at

https://sphhp.buffalo.edu/content/dam/sphhp/biostatistics/Documents/techreports/UB-Biostatistics-TR1701.pdf

6 Data Analysis Example

In this section we present a data analysis example to demonstrate the practical use for **testforDEP**. The data we use is average law school admission test (LSAT) and grade point average (GPA) from 82 law schools (details described in Efron and Tibshirani (1994)). The aim is to assess the dependence between LSAT and GPA using our package. Table 5 shows the data and a scatter plot of GPA and LSAT is shown in Figure 1.

Figure 1 suggests a linear relationship between LSAT and GPA. To confirm this, we further draw Kendall plot and compute AUK. Kendall plot is shown in Figure 2. It shows a curve above the diagonal and AUK is 0.665, which is close to 0.75. This is consistent to a potential positive correlation between LSAT and GPA.

Now consider the dependence tests provided in package **testforDEP**. Table 6 shows test results: test statistics and p-values. Obviously, all tests, classical and modern, suggest dependence between LSAT and GPA. We conclude that

School	LSAT	GPA	School	LSAT	GPA	School	LSAT	GPA
1	622	3.23	28	632	3.29	56	641	3.28
2	542	2.83	29	587	3.16	57	512	3.01
3	579	3.24	30	581	3.17	58	631	3.21
4	653	3.12	31	605	3.13	59	597	3.32
5	606	3.09	32	704	3.36	60	621	3.24
6	576	3.39	33	477	2.57	61	617	3.03
7	620	3.10	34	591	3.02	62	637	3.33
8	615	3.40	35	578	3.03	63	572	3.08
9	553	2.97	36	572	2.88	64	610	3.13
10	607	2.91	37	615	3.37	65	562	3.01
11	558	3.11	38	606	3.20	66	635	3.30
12	596	3.24	39	603	3.23	67	614	3.15
13	635	3.30	40	535	2.98	68	546	2.82
14	581	3.22	41	595	3.11	69	598	3.20
15	661	3.43	42	575	2.92	70	666	3.44
16	547	2.91	43	573	2.85	71	570	3.01
17	599	3.23	44	644	3.38	72	570	2.92
18	646	3.47	45	545	2.76	73	605	3.45
19	622	3.15	46	645	3.27	74	565	3.15
20	611	3.33	47	651	3.36	75	686	3.50
21	546	2.99	48	562	3.19	76	608	3.16
22	614	3.19	49	609	3.17	77	595	3.19
23	628	3.03	50	555	3.00	78	590	3.15
24	575c	3.01	51	586	3.11	79	558	2.81
25	662	3.39	52	580	3.07	80	611	3.16
26	627	3.41	53	594	2.96	81	564	3.02
27	608	3.04	54	594	3.05	82	575	2.74
			55	560	2.93			

Table 5: LSAT data from Efron and Tibshirani (1994).

```
>   library(testforDEP)
>   lsat = testforDEP::LSAT$LSAT
>   gpa = testforDEP::LSAT$GPA
>   plot(x = gpa, y = lsat, xlab = "GPA", ylab = "LSAT")
>   abline(lm(lsat~gpa))
```

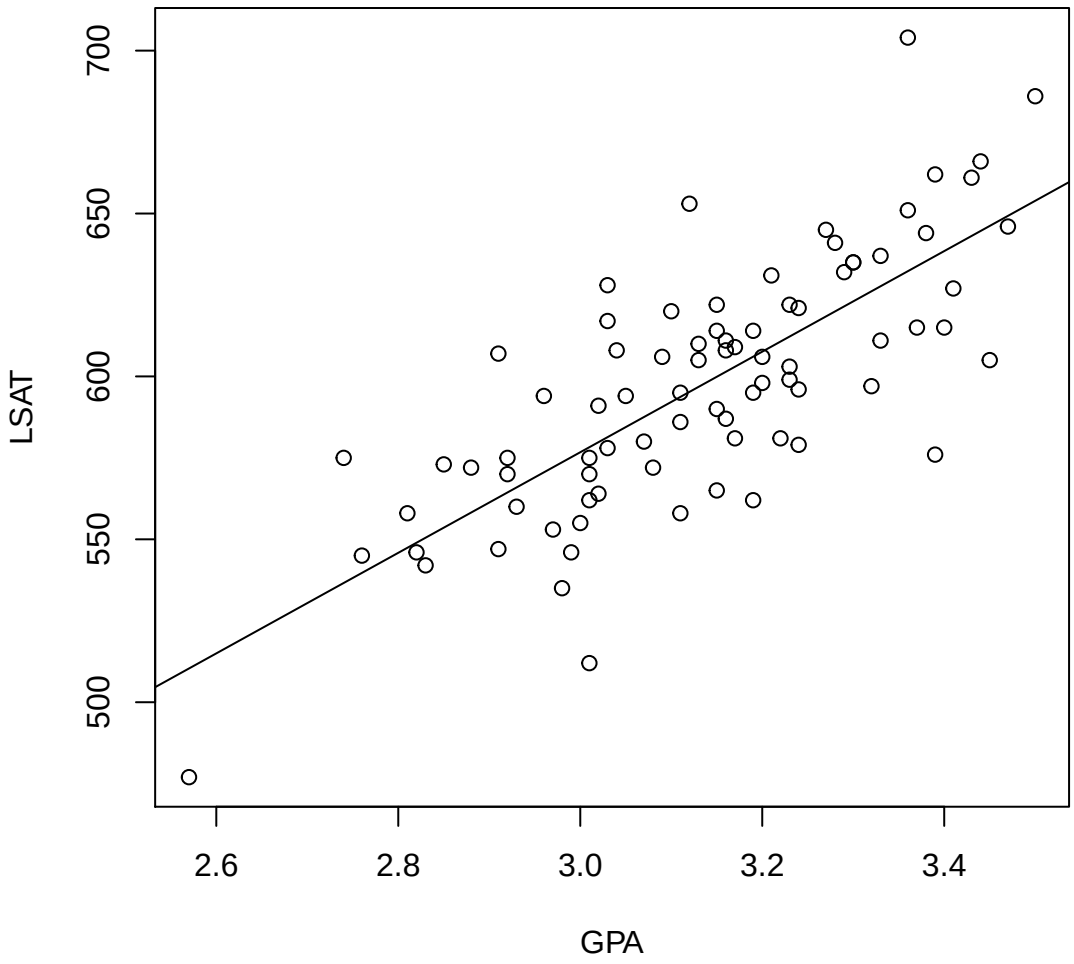

Figure 1: Scatter plot based on data in Table 5.

```
>   library(testforDEP)
>   lsat = testforDEP::LSAT$LSAT
>   gpa = testforDEP::LSAT$GPA
>   result = testforDEP::AUK(lsat, gpa, plot = TRUE, set.seed = TRUE)
>   result$AUK

[1] 0.6646519
```

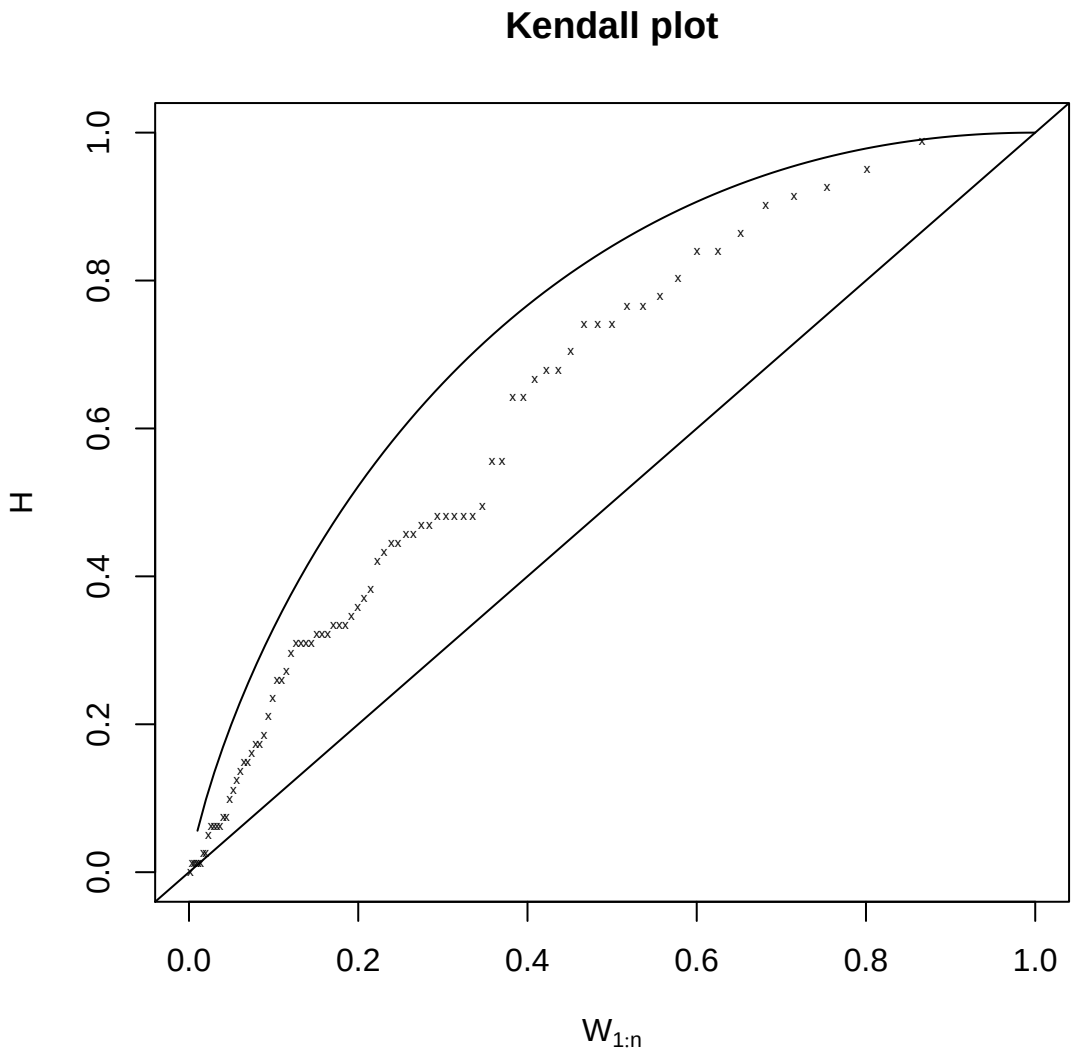

Figure 2: Kendall plot of LSAT and GPA.

LSAT and GPA are dependent.

```
>   library(testforDEP)
>   library(xtable)
>   lsat = testforDEP::LSAT$LSAT
>   gpa = testforDEP::LSAT$GPA

>   #test if there's tie
>   if(length(lsat) != length(unique(lsat)) ||    length(gpa) != length(unique(gpa)))
>   print("tie detected in data!")

>   #compute test statistics and p-values using methods in "testforDEP" package
>   #Since tie is detected in data, Spearman test will not be taken.
>   tests = c("PEARSON", "KENDALL", "VEXLER",
+             "TS2", "V", "MIC", "HOEFFD", "EL")
>   testNames = list("Pearson", "Kendall", "$\\log(VT_n)$",
+                    "$TS_2$", "$V$", "MIC", "Hoeffding",
+                    "$T_{el}$")
>   result = list()
>   for(i in 1:length(tests))
+     result[[i]] = testforDEP(lsat, gpa, test = tests[i], p.opt = "MC", set.seed = T)
>   #write results into table
>   table = matrix(0, nrow = length(tests), ncol = 3)
>   for(i in 1:length(tests)){
+     table[i,1] = testNames[[i]]
+     table[i,2] = round(result[[i]]@TS, digits = 3)
+     p.val = result[[i]]@p_value
+     table[i,3] = ifelse(p.val == 0, "< .0001", p.val)
+   }
>   colnames(table) = c("test", "statistic", "p-value")
>   xtab = xtable(table, caption = "Test results based on the LSAT data.",
+                 label = "Table:Example")
>   print(xtab, include.rownames = FALSE, sanitize.text.function=identity)
```

test	statistic	p-value
Pearson	10.459	$<.0001$
Kendall	7.464	$<.0001$
$\log(VT_n)$	65.698	$<.0001$
TS_2	80.762	$<.0001$
V	69.038	$<.0001$
MIC	0.534	$<.0001$
Hoeffding	0.222	$<.0001$
T_{el}	13.641	$<.0001$

Table 6: Test results based on the LSAT data.

7 Conclusions

The package **testforDEP** provides a new test and a convenient way to detect general structure of dependence. This new package is not only useful to analyze monotonically associated data and complex structures of non-linear or random-type independence, but also to visualize dependence. Moreover, a novel exact method based on the MIC measurement have been proposed in the package **testforDEP**. Future work is necessary to further develop a formal testing structure based on the MIC statistic. The Monte-Carlo simulation study shows the modern tests are more powerful to detect the non-linear structure while the classical tests are more powerful to test the structure of linear dependence. We believe that the package **testforDEP** will help investigators identify dependence using a cadre of tests designed to detect dependency.

References

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL http://www.R-project.org/.

Karl Pearson. Notes on the history of correlation. *Biometrika*, pages 25–45, 1920.

Maurice G Kendall. A new measure of rank correlation. *Biometrika*, pages 81–93, 1938.

Charles Spearman. The proof and measurement of association between two things. *The American journal of Psychology*, 15(1):72–101, 1904.

Albert Vexler, Wan-Min Tsai, and Alan D Hutson. A simple density-based empirical likelihood ratio test for independence. *The American Statistician*, 68(3):158–169, 2014.

Wilbert CM Kallenberg and Teresa Ledwina. Data-driven rank tests for independence. *Journal of the American Statistical Association*, 94(445):285–301, 1999.

David N Reshef, Yakir A Reshef, Hilary K Finucane, Sharon R Grossman, Gilean McVean, Peter J Turnbaugh, Eric S Lander, Michael Mitzenmacher, and Pardis C Sabeti. Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524, 2011.

John HJ Einmahl and Ian W McKeague. Empirical likelihood based hypothesis testing. *Bernoulli*, pages 267–290, 2003.

Yi Wang, Yi Li, Hongbao Cao, Momiao Xiong, Yin Yao Shugart, and Li Jin. Efficient test for nonlinear dependence of two continuous variables. *BMC Bioinformatics*, 16(1):1, 2015.

Albert Vexler, Xiwei Chen, and Alan D Hutson. Dependence and independence: Structure and inference. *Statistical Methods in Medical Research*, page 0962280215594198, 2015.

Jan Hauke and Tomasz Kossowski. Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones Geographicae*, 30(2):87–93, 2011.

Art B Owen. *Empirical Likelihood*. CRC Press, 2001.

Herbert Aron David and Haikady Navada Nagaraja. *Order Statistics*. Wiley Online Library, 1970.

Michele Filosi, Roberto Visintainer, and Davide Albanese. *Minerva: Minerva: Maximal Information-Based Nonparametric Exploration R Package for Variable Analysis*, 2014. URL http://CRAN.R-project.org/package=minerva. R package version 1.4.1.

Noah Simon and Robert Tibshirani. Comment on" detecting novel associations in large data sets" by reshef et al, science dec 16, 2011. *arXiv Preprint arXiv:1401.7645*, 2014.

Wassily Hoeffding. A non-parametric test of independence. *The annals of Mathematical Statistics*, pages 546–557, 1948.

Frank E Harrell Jr and Maintainer Charles Dupont. The hmisc package. *R package version*, 3:0–12, 2006.

Christian Genest and Jean-Claude Boies. Detecting dependence with kendall plots. *The American Statistician*, 2012.

M. E. Johnson. *Multivariate Statistical Simulation*. Wiley, 1987.

Nail K Bakirov, Maria L Rizzo, and Gábor J Székely. A multivariate nonparametric test of independence. *Journal of Multivariate Analysis*, 97(8): 1742–1756, 2006.

Bradley Efron and Robert J Tibshirani. *An Introduction to the Bootstrap*. CRC Press, 1994.