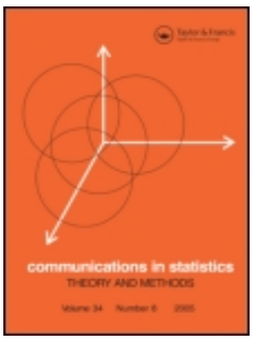

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: http://www.tandfonline.com/loi/lsta20

A Test of Independence in Two-Way Contingency Tables Based on Maximal Correlation

C. D. Yenigün , G. J. Székely & M. L. Rizzo

To cite this article: C. D. Yenigün , G. J. Székely & M. L. Rizzo (2011) A Test of Independence in Two-Way Contingency Tables Based on Maximal Correlation, Communications in Statistics - Theory and Methods, 40:12, 2225-2242, DOI: 10.1080/03610921003764274

To link to this article: http://dx.doi.org/10.1080/03610921003764274

Communications in Statistics—Theory and Methods, 40: 2225–2242, 2011

ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610921003764274

Taylor & Francis
Taylor & Francis Group

A Test of Independence in Two-Way Contingency Tables Based on Maximal Correlation

C. D. YENİGÜN[1], G. J. SZÉKELY[2], AND M. L. RIZZO[3]

[1]Department of Management, Bilkent University, Ankara, Turkey
[2]National Science Foundation, Arlington, Virginia, USA
[3]Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, Ohio, USA

Maximal correlation has several desirable properties as a measure of dependence, including the fact that it vanishes if and only if the variables are independent. Except for a few special cases, it is hard to evaluate maximal correlation explicitly. We focus on two-dimensional contingency tables and discuss a procedure for estimating maximal correlation, which we use for constructing a test of independence. We compare the maximal correlation test with other tests of independence by Monte Carlo simulations. When the underlying continuous variables are dependent but uncorrelated, we point out some cases for which the new test is more powerful.

Keywords Exact tests; Maximal correlation; Tests of independence.

Mathematics Subject Classification 62H20; 62H17.

1. Introduction

In virtually any field of statistics, there is a need for measuring the dependence between random variables. There are several measures of dependence in the statistical literature, which can be classified into three groups: dependence measures based on correlation, dependence measures based on distribution functions, and dependence measures for cross classifications. All these measures are often considered as an intermediate step for obtaining tests of independence. For a comprehensive survey on most important dependence measures, see Liebetrau (2005).

A basic tool for measuring dependence between two random variables X and Y is the product moment correlation coefficient $\rho(X, Y)$, which has some well-known drawbacks. For example, zero correlation does not imply independence. Gebelein

Received July 4, 2008; Accepted March 9, 2010
Address correspondence to C. D. Yenıgün, Department of Management, Bilkent University, 06800 Ankara, Turkey; E-mail: yenigun@bilkent.edu.tr

(1941) introduced the *maximal correlation* defined by

$$S(X, Y) = \sup_{f,g} \rho(f(X), g(Y)), \tag{1}$$

where the supremum is taken over all Borel-measurable functions of X and Y with finite and positive variance. Maximal correlation has several desirable properties as a dependence measure, including the fact that it vanishes if and only if the variables are independent.

Although maximal correlation is an attractive dependence measure, it is hard to evaluate it explicitly, except for a few special cases. Rényi (1959) gave the conditions such that the maximal correlation can be attained. Bell (1962) considered two normalizations of Shannon's mutual information as a measure of dependence and compared them with maximal correlation. Csáki and Fischer (1963) further studied the mathematical properties of maximal correlation and computed it for a number of examples. Abrahams and Thomas (1980) considered maximal correlation as a measure of dependence in stochastic processes. Breiman and Friedman (1985) provided an alternating conditional expectations algorithm for estimating the functions f_0 and g_0 which maximize the correlation between two random variables, thus their procedure provides a method for estimation maximal correlation from observations. A multivariate analog of maximal correlation was considered by Koyak (1987). For random variables that take only a finite number of values, Sethuraman (1990) gave a procedure to estimate the maximal correlation from the sample, and gave the asymptotic distribution of this estimate under the null hypothesis of independence. Gautam and Kimeldorf (1999) considered the calculation of maximal correlation in the case of $2 \times k$ contingency tables. Dembo et al. (2001) and Novak (2004) studied the maximal correlation between partial sums of independent and identically distributed random variables.

In this study, we discuss a new addition to the rare cases such that maximal correlation can be computed. We first discuss how maximal correlation is computed for two categorical variables that are cross classified. We then use this to construct an independence test for two-way contingency tables. Analysis of contingency tables has been a very active research field in statistics for a long time, due to its enormous applicability. A principal interest in many studies regarding contingency tables is to test if the variables are independent. Although many good tests are available, no single test is known to be optimal for all independence problems. Our purpose is to point out cases such that maximal correlation independence test is preferable.

This article will proceed as follows. In Sec. 2, we present some preliminary notation and basic tools on the analysis of contingency tables which will be used throughout this study. In Sec. 3, we present the maximal correlation. In Sec. 4, we discuss the computation of maximal correlation for the case of contingency tables, and in Sec. 5 we introduce the maximal correlation independence test. Section 6 contains empirical power comparisons, followed by the conclusions in Sec. 7.

2. Preliminaries

In this section, we present some basic notation and some of the well-known tools in contingency table analysis that we use in the remaining of the article, namely, loglinear models and exact tests. Consider two categorical response variables X and Y having I and J levels, respectively. When we classify subjects on both variables,

the responses (X, Y) of a randomly selected subject have a probability distribution which can be displayed in a rectangular table having I rows for categories of X and J columns for categories of Y. Let π_{ij} denote the probability that (X, Y) falls in the cell in row i and column j. The probability distribution $\{\pi_{ij}\}$ is the joint distribution of X and Y. When the cells contain frequency counts of outcomes, denoted by $\{n_{ij}\}$, the table is called a *contingency table*. In this study, we consider the multinomial sampling model, where the sample size is fixed and the probability mass function (pmf) of the contingency table is a multinomial distribution characterized by the sample size n and the cell probabilities $\{\pi_{ij}\}$. The null hypothesis of statistical independence is

$$H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}, \tag{2}$$

for $i = 1, \ldots, I$, $j = 1, \ldots, J$ and the subscript "$\cdot$" denotes the sum over the index it replaces.

Earlier applications in the analysis of contingency tables are on testing independence, by the well-known Pearson chi-square test of independence and its modifications, as well as the likelihood ratio test. By the 1960s and 1970s, the attention of researchers shifted from testing to modeling and *loglinear models* gained significant attention. With the loglinear approach, the cell counts in a contingency table are modeled in terms of the associations between the variables. The *saturated loglinear model* for $I \times J$ contingency tables is given by

$$\log m_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}, \tag{3}$$

where $m_{ij} = n\pi_{ij}$ is the expected frequency in cell (i, j), and the parameters satisfy $\sum_i \lambda_i^X = \sum_j \lambda_j^Y = \sum_i \lambda_{ij}^{XY} = \sum_j \lambda_{ij}^{XY} = 0$, for $i = 1, \ldots, I$, $j = 1, \ldots, J$. Note that the logarithm of the expected frequency in cell (i, j) is an additive function of an ith row effect λ_i^X, a jth column effect λ_j^Y, and an interaction effect λ_{ij}^{XY}. A special case of the saturated model is the *loglinear model of independence*, in which all the interaction effects λ_{ij}^{XY} equal zero. For multinomial sampling, the cell probabilities of the multinomial distribution corresponding to the saturated loglinear model (3) is given by

$$\pi_{ij} = \frac{\exp(\mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY})}{\sum_i \sum_j \exp(\mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY})}. \tag{4}$$

Given a contingency table, the loglinear model parameters can be estimated by using maximum likelihood methods. In the loglinear approach, testing independence corresponds to testing whether the interaction term is needed in the model or not; see, e.g., Agresti (2002) for more on loglinear models.

When working with contingency tables with a small number of observations or sparse data, *exact inferential methods* provide an alternative to large sample methods. Under the null hypothesis of independence, the pmf of the contingency table $\{n_{ij}\}$ includes nuisance parameters $\{\pi_{i\cdot}\}$ and $\{\pi_{\cdot j}\}$, thus it has a limited use. These parameters can be eliminated by conditioning on sufficient statistics for them, $\{n_{i\cdot}\}$ and $\{n_{\cdot j}\}$. The pmf of $\{n_{ij}\}$ conditional on the sufficient statistics is given by

$$\frac{(\prod_i n_{i\cdot}!)(\prod_j n_{\cdot j}!)}{n! \prod_i \prod_j n_{ij}!}. \tag{5}$$

Exact tests require finding all possible contingency tables that have the same row and column sums as the observed contingency table. These tables are then ordered according to some measure of dependence. The p-value of the observed contingency table is simply the sum of probabilities of observing the tables which are more extreme than the observed table, i.e., the ones that are farther from independence in terms of the dependence measure used. The related table probabilities are calculated from (5).

A well-known example of exact tests is the Fisher exact test for 2×2 contingency tables (Fisher, 1934), where exact enumeration by hand is possible. The availability of computational power makes exact tests possible for higher dimensional tables, however, in most cases complete enumeration is still impossible with current computational power. In such cases, one can simulate contingency tables with given marginals and approximate the p-value of the independence test. For two-dimensional tables, the algorithm given by Patefield (1981) is widely used. For higher dimensional tables, one can use the algorithm given by Diaconis and Strumfels (1998). Exact tests are available in most statistical software packages such as SPSS, SAS, and R. StatXact is a statistical package that specializes in exact tests. For a survey of exact tests, see Agresti (1992).

3. Maximal Correlation

Consider two random variables, X and Y, defined on a given probability space. According to Rényi (1959), a measure of dependence $\delta(X, Y)$ of these variables should satisfy the following postulates.

(A) $\delta(X, Y)$ is defined for any X, Y neither of which is constant with probability 1.
(B) $\delta(X, Y) = \delta(Y, X)$.
(C) $0 \leq \delta(X, Y) \leq 1$.
(D) $\delta(X, Y) = 0$ if and only if X and Y are independent.
(E) $\delta(X, Y) = 1$ if either $X = g(Y)$ or $Y = f(X)$, where $g(\cdot)$ and $f(\cdot)$ are Borel-measurable functions.
(F) If the Borel-measurable functions $g(\cdot)$ and $f(\cdot)$ map the real axis in a one-to-one way to itself, then $\delta(f(X), g(Y)) = \delta(X, Y)$.
(G) If the joint distribution of X and Y is normal, then $\delta(X, Y) = |\rho(X, Y)|$, where $\rho(X, Y)$ is the product moment correlation coefficient of X and Y.

After listing these seven postulates, Rényi (1959) considered five measures of dependence, and notes that only the maximal correlation satisfies all seven postulates. Note that product moment correlation satisfies B, C, and G only.

The maximal correlation (1) between X and Y cannot be evaluated explicitly except for special cases, since there does not always exist functions $f_0(x)$ and $g_0(x)$ such that $S(X, Y) = \rho(f_0(X), g_0(Y))$. If this equality holds for some f_0 and g_0, we say that the *maximal correlation of X and Y can be attained.* Let $\mathscr{L}_X^2$ denote the Hilbert space of all random variables of the form $f(X)$ for which $E(f(X)) = 0$ and $\mathrm{Var}(f(X))$ is finite. Similarly, let $\mathscr{L}_Y^2$ denote the Hilbert space of all random variables of the form $g(Y)$ for which $E(g(Y)) = 0$ and $\mathrm{Var}(g(Y))$ is finite. For any $f = f(X) \in \mathscr{L}_X^2$, consider the transformation

$$Af = E[E(f(X) \mid Y) \mid X]. \tag{6}$$

Rényi (1959) showed that if the transformation A defined in (6) is completely continuous, then the maximal correlation between X and Y is attained for $f_0(X)$ and $g_0(Y)$, where f_0 is an eigenfunction belonging to the greatest eigenvalue $S^2 = S^2(X, Y)$ of A and $g_0(Y) = S^{-1}E(f_0(X) \mid Y)$. Rényi (1959) also noted that if the dependence between X and Y is *regular* and the mean square contingency is finite, then the transformation A is completely continuous. Here, *regular* dependence of the variables means that the joint distribution of the variables is absolutely continuous with respect to the direct product of their distributions.

3.1. *An Example: Lissajous Curve Case*

Here, we introduce a new example such that maximal correlation can be computed. In this example we consider two dependent but uncorrelated random variables, and show that the maximal correlation between them equals one. We will revisit this example in Sec. 6.

Let the random variable W have uniform distribution over the interval $[0, 2\pi]$. Let $X = \sin aW$ and $Y = \sin bW$ where a and b are integers and $a \neq b$. The variables X and Y are clearly dependent, and one can show that they are uncorrelated. The plot of the relationship between the variables is a special case of the well-known Lissajous curve, as illustrated in Fig. 1, therefore we will refer to this example as the *Lissajous curve case*.

In order to compute the maximal correlation between X and Y, we use the *Chebyshev polynomials of the first kind*, which are defined by $T_n(x) = \cos(n \arccos x)$, where $x \in [-1, 1]$, and n is the degree of the polynomials. Replacing x by $\cos\theta$, where $\theta \in \mathbb{R}$, we have $T_n(\cos\theta) = \cos(n\theta)$. Here, $\cos(n\theta)$ is a polynomial of degree n in $\cos(\theta)$.

Proposition 3.1. *The maximal correlation between X and Y is* 1.

Proof. Let $Z = \cos(2abW)$. Then there exists a Chebyshev polynomial T_a such that

$$Z = \cos(2abW) = T_a(\cos(2bW)) = T_a(1 - 2\sin^2(bW)) = T_a(1 - 2Y^2).$$

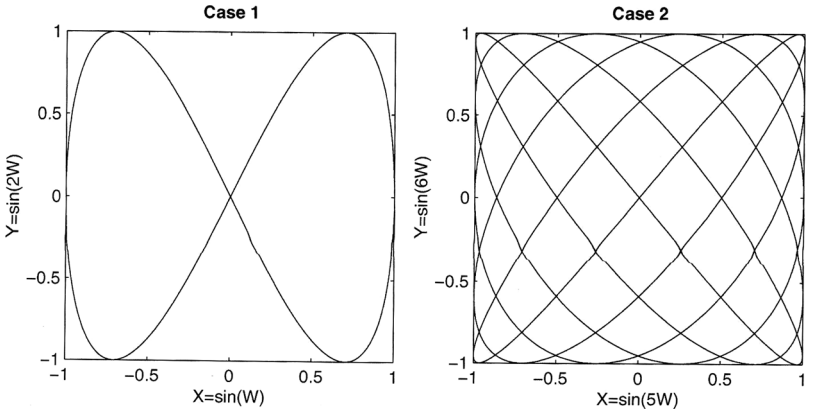

Figure 1. Two illustrations for the Lissajous curve. In Case 1, $a = 1$ and $b = 2$; in Case 2, $a = 5$, $b = 6$.

Similarly, there exists a Chebyshev polynomial T_b such that

$$Z = \cos(2abW) = T_b(\cos(2aW)) = T_b(1 - 2\sin^2(aW)) = T_b(1 - 2X^2).$$

Thus, Z is a function of both X and Y. The correlation of Z with itself is 1, thus, the maximal correlation between X and Y is 1.

4. Maximal Correlation in Case of Contingency Tables

In this section, we introduce the computation of maximal correlation for two-dimensional contingency tables, which is another addition to the rare cases such that maximal correlation can be computed. Consider two categorical response variables, X and Y, taking values $\alpha_1, \ldots, \alpha_I$ and $\beta_1, \ldots, \beta_J$ respectively. Without loss of generality, suppose $I \leq J$. Consider the cross-classification as described in Sec. 2. Assume that the matrix $\{\pi_{ij}\}$ is positive, in other words there are no structural zeroes in the contingency table. Let $\mathbf{I}_X = (\mathbf{1}_{\alpha_1}(X), \ldots, \mathbf{1}_{\alpha_I}(X))'$ and $\mathbf{I}_Y = (\mathbf{1}_{\beta_1}(Y), \ldots, \mathbf{1}_{\beta_J}(Y))'$, where $\mathbf{1}$ denotes the indicator function. Since X and Y can take only a finite number of outcomes, the functions f and g in their most general forms can be written as

$$f(X) = \mathbf{a}'\mathbf{I}_X, \tag{7}$$

$$g(Y) = \mathbf{b}'\mathbf{I}_Y, \tag{8}$$

where $\mathbf{a} = (a_1, \ldots, a_I)'$, $\mathbf{b} = (b_1, \ldots, b_J)'$, and a_i and b_j are arbitrary real numbers for $i = 1, \ldots, I$ and $j = 1, \ldots, J$. Our task is to find f_0 and g_0 such that the maximal correlation between X and Y is attained, which is equivalent to finding proper vectors $\mathbf{a}$ and $\mathbf{b}$. We begin with writing the transformation (6) explicitly.

Let $\mathbf{r}_i = (\pi_{i1}, \ldots, \pi_{iJ})'$ for $i = 1, \ldots, I$, which is the transpose of the ith row of $\{\pi_{ij}\}$. Similarly, let $\mathbf{c}_j = (\pi_{1j}, \ldots, \pi_{Ij})'$ for $j = 1, \ldots, J$, which is the jth column of $\{\pi_{ij}\}$. Then we have

$$U_j := E(f(X) \mid Y = \beta_j) = \frac{1}{\pi_{\cdot j}}\mathbf{c}_j'\mathbf{a},$$

for $j = 1, \ldots, J$. Let $\mathbf{U} = (U_1, \ldots, U_J)'$. Then

$$V_i := E\left[E(f(X) \mid Y) \mid X = \alpha_i\right] = \frac{1}{\pi_{i\cdot}}\mathbf{r}_i'\mathbf{U},$$

for $i = 1, \ldots, I$. Let $\mathbf{V} = (V_1, \ldots, V_I)'$. Note that $\mathbf{V}$ is the right-hand side of Eq. (6).

Proposition 4.1. *The vector* $\mathbf{V}$ *can be factored such that* $\mathbf{V} = \mathscr{A}\mathbf{a}$, *where* $\mathscr{A}$ *is an* $I \times I$ *matrix with the general term*

$$\mathscr{A}_{kl} = \sum_{r=1}^{J} \frac{\pi_{kr}\pi_{lr}}{\pi_{k\cdot}\pi_{\cdot r}}, \tag{9}$$

where $k, l = 1, \ldots, I$.

Proof.

$$\mathbf{V} = \begin{bmatrix} \frac{1}{\pi_{1\cdot}}\mathbf{r}_1'\mathbf{U} \\ \vdots \\ \frac{1}{\pi_{I\cdot}}\mathbf{r}_I'\mathbf{U} \end{bmatrix} = \begin{bmatrix} \frac{1}{\pi_{1\cdot}}(\pi_{11}U_1 + \pi_{12}U_2 + \cdots + \pi_{1J}U_J) \\ \vdots \\ \frac{1}{\pi_{I\cdot}}(\pi_{I1}U_1 + \pi_{I2}U_2 + \cdots + \pi_{IJ}U_J) \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{\pi_{1\cdot}}\left(\frac{\pi_{11}}{\pi_{\cdot 1}}\mathbf{c}_1'\mathbf{a} + \frac{\pi_{12}}{\pi_{\cdot 2}}\mathbf{c}_2'\mathbf{a} + \cdots + \frac{\pi_{1J}}{\pi_{\cdot J}}\mathbf{c}_J'\mathbf{a}\right) \\ \vdots \\ \frac{1}{\pi_{I\cdot}}\left(\frac{\pi_{I1}}{\pi_{\cdot 1}}\mathbf{c}_1'\mathbf{a} + \frac{\pi_{I2}}{\pi_{\cdot 2}}\mathbf{c}_2'\mathbf{a} + \cdots + \frac{\pi_{IJ}}{\pi_{\cdot J}}\mathbf{c}_J'\mathbf{a}\right) \end{bmatrix}$$

$$= \begin{bmatrix} \left(\frac{\pi_{11}}{\pi_{1\cdot}\pi_{\cdot 1}}(\pi_{11}, \pi_{21}, \ldots, \pi_{I1}) + \frac{\pi_{12}}{\pi_{1\cdot}\pi_{\cdot 2}}(\pi_{12}, \pi_{22}, \ldots, \pi_{I2}) \right. \\ \left. + \cdots + \frac{\pi_{1J}}{\pi_{1\cdot}\pi_{\cdot J}}(\pi_{1J}, \pi_{2J}, \ldots, \pi_{IJ})\right)\mathbf{a} \\ \vdots \\ \left(\frac{\pi_{I1}}{\pi_{I\cdot}\pi_{\cdot 1}}(\pi_{11}, \pi_{21}, \ldots, \pi_{I1}) + \frac{\pi_{I2}}{\pi_{I\cdot}\pi_{\cdot 2}}(\pi_{12}, \pi_{22}, \ldots, \pi_{I2}) \right. \\ \left. + \cdots + \frac{\pi_{IJ}}{\pi_{I\cdot}\pi_{\cdot J}}(\pi_{1J}, \pi_{2J}, \ldots, \pi_{IJ})\right)\mathbf{a} \end{bmatrix}$$

$$= \begin{bmatrix} \left(\sum_{r=1}^{J} \frac{\pi_{1r}\pi_{1r}}{\pi_{1\cdot}\pi_{\cdot r}}, \sum_{r=1}^{J} \frac{\pi_{1r}\pi_{2r}}{\pi_{1\cdot}\pi_{\cdot r}}, \ldots, \sum_{r=1}^{J} \frac{\pi_{1r}\pi_{Ir}}{\pi_{1\cdot}\pi_{\cdot r}}\right)\mathbf{a} \\ \vdots \\ \left(\sum_{r=1}^{J} \frac{\pi_{Ir}\pi_{1r}}{\pi_{I\cdot}\pi_{\cdot r}}, \sum_{r=1}^{J} \frac{\pi_{Ir}\pi_{2r}}{\pi_{I\cdot}\pi_{\cdot r}}, \ldots, \sum_{r=1}^{J} \frac{\pi_{Ir}\pi_{Ir}}{\pi_{I\cdot}\pi_{\cdot r}}\right)\mathbf{a} \end{bmatrix} = \mathscr{A}\mathbf{a}.$$

Therefore, in the case of contingency tables, the transformation A in (6) is represented by the matrix $\mathscr{A}$. This transformation is completely continuous since the dependence between the response variables is always regular and the mean square contingency is finite for the case of contingency tables. Therefore, the maximal correlation between X and Y can be computed.

According to Rényi (1959), the largest eigenvalue of transformation (6) is the square of the maximal correlation between X and Y. However, his approach makes the assumption that $E(f(X)) = 0$ and $\mathrm{Var}(f(X))$ is finite. We impose these assumptions on our calculation as follows.

It is easy to check that $\mathscr{A}$ is a positive stochastic matrix. Then by the Perron–Frobenius theorem (see, e.g., Aldrovandi, 2001, p. 47), $\mathscr{A}$ has a single unit eigenvalue which is larger than the absolute value of any other eigenvalue. Let $\lambda_1 = 1 > |\lambda_2| \geq \cdots \geq |\lambda_I| \geq 0$ denote the eigenvalues of $\mathscr{A}$ sorted in a decreasing fashion, and let $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_I$ denote the corresponding column eigenvectors. It is easy to see that $\mathbf{e}_1 = \mathbf{1}_I$, where $\mathbf{1}_I$ denotes the I-dimensional column vector, all of whose components are one. Then if we set $\mathbf{a} = \mathbf{e}_1$ we have

$$E[f(X)] = E(\mathbf{a}'\mathbf{I}_X) = E(\mathbf{e}_1'\mathbf{I}_X) = (\pi_{1\cdot}, \ldots, \pi_{I\cdot})\mathbf{e}_1 = 1,$$

therefore the assumption $E[f(X)] = 0$ is violated. Also note that, in this case, $\mathrm{Var}(f(X)) = 0$. When we set $\mathbf{a} = \mathbf{e}_i$ for $i = 2, \ldots, I$, we have $E(f(X)) = 0$ and $0 < \mathrm{Var}(f(X)) < \infty$. So we discard the largest eigenvalue of $\mathscr{A}$ and conclude that the maximal correlation S between X and Y is the square root of the second largest

eigenvector of $\mathscr{A}$. Formally, we have

$$S(X, Y) = \sqrt{\lambda_2}. \tag{10}$$

Now let us compute f_0 and g_0 such that the maximal correlation is attained, i.e., let us find the vectors **a** and **b**. We have

$$f_0(X) = \mathbf{e}_2' \mathbf{I}_X, \tag{11}$$

which means that we must set $\mathbf{a} = \mathbf{e}_2$ in (7). Moreover, we have $g_0(Y) = S^{-1}E(f_0(X) \mid Y)$. Let $\mathbf{d} = (d_1, \ldots, d_J)'$ where

$$d_j = E(f_0(X) \mid Y = \beta_j) = (\mathbf{c}_j' \mathbf{e}_2)/\pi_{\cdot j},$$

for $j = 1, \ldots, J$. Then we have

$$g_0(Y) = S^{-1} \mathbf{d}' \mathbf{I}_Y, \tag{12}$$

where $S^{-1} = 1/\sqrt{\lambda_2}$. So we must set $\mathbf{b} = S^{-1}\mathbf{d}$ in (8). We have shown the following.

Proposition 4.2. *Let the categorical variables X and Y take values* $\alpha_1, \ldots, \alpha_I$ *and* $\beta_1, \ldots, \beta_J$, *respectively, with* $I \leq J$. *Consider the cross-classification defined above, where the cell probabilities of the joint distribution is given in the positive matrix* $\{\pi_{ij}\}$. *Then the* population maximal correlation *between X and Y is the square root of the second largest eigenvalue of the matrix* $\mathscr{A}$ *defined in* (9). *The maximal correlation is attained when* f_0 *and* g_0 *are as defined in* (11) *and* (12), *respectively.*

In practice, we want to estimate the maximal correlation from an observed contingency table. Let $\{n_{ij}\}$ denote an observed contingency table, as described above. Let $\widehat{\mathscr{A}}$ denote the estimator of $\mathscr{A}$ obtained by replacing π_{ij}'s in (9) by their maximum likelihood estimators $\hat{\pi}_{ij} = n_{ij}/n$. Then $\widehat{\mathscr{A}}$ is an $I \times I$ matrix with general term

$$\widehat{\mathscr{A}}_{kl} = \sum_{r=1}^{J} \frac{n_{kr} n_{lr}}{n_{k\cdot} n_{\cdot r}}, \tag{13}$$

where $k, l = 1, \ldots, I$. The matrix $\widehat{\mathscr{A}}$ is well defined when the observed contingency table does not have zero row or column sums. If a row or column sum equals zero for an observed contingency table, we use the convention of replacing the zeros in the denominator of (13) by $\varepsilon = 10^{-8}$, without changing the table dimensions for the analysis. The *sample maximal correlation* S_n between X and Y is the square root of the second largest eigenvalue of $\widehat{\mathscr{A}}$. Formally, we have

$$S_n(X, Y) = \sqrt{\hat{\lambda}_2}, \tag{14}$$

where $\hat{\lambda}_2$ is the second largest eigenvalue of $\widehat{\mathscr{A}}$. The functions f_0 and g_0 can be computed analogous to the above approach. We give the following remarks before we proceed to the maximal correlation test of independence.

Remark 4.1. The eigenvalues of the matrix $\mathscr{A}$ are invariant with respect to the row and column permutations of the matrix $\{\pi_{ij}\}$, so is the maximal correlation. Same holds for an observed matrix $\widehat{\mathscr{A}}$ and the sample maximal correlation. This makes maximal correlation suitable for analyzing data on nominal scale.

Remark 4.2. Since one can observe empty cells in a contingency table, the stochastic matrix $\widehat{\mathscr{A}}$ is not always positive. Therefore, in some special cases, $\widehat{\mathscr{A}}$ may have more than one unit eigenvalues. One such case is the case for which the cell n_{ij} is the only nonzero cell in ith row and jth column. In such cases, the sample maximal correlation is defined as the square root of the largest non unity eigenvalue of $\widehat{\mathscr{A}}$.

Remark 4.3. Maximal correlation is the first canonical correlation between $\mathbf{I}_X$ and $\mathbf{I}_Y$.

Remark 4.4. For 2×2 contingency tables, we have the following relation between the classical *Pearson chi-square statistic*, X^2, and the sample maximal correlation:

$$X^2 = nS_n^2.$$

5. Maximal Correlation Test of Independence

We now develop an independence test for two-way contingency tables, based on maximal correlation. Sethuraman (1990) considered the maximal correlation of variables that take only a finite number of values. This work does not provide an explicit evaluation of maximal correlation from an observed sample, however, it gives the limiting distribution of sample maximal correlation under the null hypothesis of independence. We adapt these distributional results to contingency tables as follows.

Consider categorical variables X and Y having I and J levels, respectively. Consider the cross-classification of X and Y which leads to a contingency table $\{n_{ij}\}$. We would like to test the independence hypothesis (2). Without loss of generality, assume that $I \leq J$. Let n denote the sample size and let S_n denote the sample maximal correlation between X and Y. Assume X and Y are independent. Then by Sethuraman (1990), the limiting distribution of nS_n^2 as $n \to \infty$ is the distribution of τ_1, where τ_1 is the maximum eigenvalue of W which has a Wishart distribution $W(\mathbf{I}_{I-1}, J-1)$. Here, $\mathbf{I}_a$ denotes the identity matrix of size a.

Thus for large samples, a size α maximal correlation test of independence can be constructed as follows.

1. Compute the sample maximal correlation S_n using (14), obtain the test statistic nS_n^2.
2. Reject H_0 if $nS_n^2 > C(\alpha)$, where $C(\alpha)$ is the $100(1-\alpha)\%$ point of the limiting distribution of the test statistic nS_n^2.

The critical points $C(\alpha)$ can be obtained from Table 51 of Pearson and Hartley (1972, p. 352) (set $\nu = I-1$ and $p = J-1$ or vice versa), which gives the percentage points of the extreme eigenvalues of a Wishart matrix. When the dimensions or the significance level of interest cannot be found on this table, one can simulate the null distribution of nS_n^2 and obtain the critical values. For several choices of I and J,

we generated Wishart matrices with scale parameter $\mathbf{I}_{I-1}$ and degrees of freedom $J-1$ by using the `rwishart` function available in the statistical package R. The empirical 90%, 95%, and 99% percentile points of the maximum eigenvalues of the 100,000 generated Wishart matrices are given in Table 1. Here, we may note that for 2×2 contingency tables maximal correlation test of independence is identical to the classical Pearson chi-square test of independence.

When working with contingency tables with a small number of observations or sparse data, it may not be appropriate to use the above independence test. This is also the case for other commonly used independence tests that are based on large sample results, such as chi-square test or likelihood ratio test. In such situations, exact inferential methods provide an alternative to the large sample methods. As mentioned in Sec. 2, exact tests require an ordering criterion. Employing maximal correlation as the ordering criterion, a size α exact maximal correlation test of independence can be constructed as follows:

1. Observe a contingency table with frequencies $\{n_{ij}\}$ and calculate the row and column sums $n_{i\cdot}$ and $n_{\cdot j}$.
2. Find all possible contingency tables $\{a_{ij}\}$ such that $a_{i\cdot} = n_{i\cdot}$ and $a_{\cdot j} = n_{\cdot j}$. Compute the probabilities of obtaining such tables by plugging a_{ij} for n_{ij} in (5).
3. Order all tables with respect to the maximal correlation between the row and the column variables.
4. The p-value of the exact test is the sum of probabilities of obtaining contingency tables which yield equal or greater maximal correlation compared to the observed table. Reject the null hypothesis of independence if p-value is less than α.

Exact test algorithms require enumerating all possible tables that have the same marginal sums as the observed contingency table. We have written codes in the statistical package R to enumerate the required tables for small contingency tables. When there is a large number of row or column categories, enumerating all these tables may be a numerically challenging task. To illustrate, in order to carry out an exact test for a 4×4 table with 100 observations, one has to deal with roughly 7 billion contingency tables. To overcome this difficulty, several algorithms have been proposed to simulate contingency tables that have the same row and column sums as a given contingency table. One example is the Patefield's (1981) algorithm, which can be implemented by using the `r2dtable` function in the statistical package R.

Table 1
Critical values of nS_n^2 for $I \times J$ contingency tables

I	J	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
3	3	6.998	8.599	12.057
4	4	11.229	13.137	17.179
5	5	15.441	17.584	21.976
6	6	19.634	21.953	26.706
8	8	27.872	30.359	35.636
10	10	36.122	38.875	44.710
12	12	44.233	47.215	53.330
14	14	52.506	55.686	62.133
16	16	60.583	63.892	70.359

When we would like to construct an exact maximal correlation test of independence, but the complete enumeration is infeasible, we approximate the p-value of the exact test by using the Patefield's (1981) algorithm. Then a size α independence test can be constructed as follows:

1. Observe a contingency table with frequencies $\{n_{ij}\}$ and calculate the row and column sums $n_{i\cdot}$ and $n_{\cdot j}$.
2. Using Patefield's (1981) algorithm, generate a reasonable amount of contingency tables $\{a_{ij}\}$ such that $a_{i\cdot} = n_{i\cdot}$ and $a_{\cdot j} = n_{\cdot j}$. Compute the probabilities of obtaining such tables by plugging a_{ij} for n_{ij} in (5).
3. Order all tables with respect to the maximal correlation between the row and the column variables.
4. An approximation to the p-value of the exact test is the ratio of the sum of probabilities of obtaining contingency tables which yield equal or greater maximal correlation compared to the observed table, to the sum of probabilities of obtaining all the contingency tables that have been generated. Reject the null hypothesis of independence if the approximate p-value is less than α.

5.1. *A Numerical Illustration*

Table 2 is taken from Snee (1974), which presents the hair color and eye color of 264 males. We would like to test the independence hypothesis (2) using the maximal correltaion test. Using (13) we have

$$\widehat{\mathscr{M}} = \begin{bmatrix} 0.394022 & 0.313580 & 0.187000 & 0.105397 \\ 0.257698 & 0.437607 & 0.168808 & 0.135889 \\ 0.330235 & 0.362757 & 0.184119 & 0.122867 \\ 0.260590 & 0.415905 & 0.175035 & 0.143970 \end{bmatrix}.$$

The eigenvalues and the corresponding eigenvectors of $\widehat{\mathscr{M}}$ are

$$e_1 = 1, \quad v_1 = (0.5, 0.5, 0.5, 0.5)',$$
$$e_2 = 0.14399, \quad v_2 = (0.7168, -0.5261, 0.1644, -0.4268)',$$
$$e_3 = 0.01295, \quad v_3 = (-0.1434, -0.3335, 0.3697, 0.8552)',$$
$$e_4 = 0.00276, \quad v_4 = (0.2297, -0.0088, -0.7877, 0.5714)'.$$

Table 2
Hair color and eye color for 264 males

	Eye color			
Hair color	Brown	Blue	Hazel	Green
Black	32	11	10	3
Brown	38	50	25	15
Red	10	10	7	7
Blond	3	30	5	8

Then by (14), the sample maximal correlation is $S_n = \sqrt{0.14399} = 0.3794$, and the test statistic is $nS_n^2 = 38.015$. From Table 1 the critical values for this 4×4 case are 11.229, 13.137, and 17.179 for significance levels 0.1, 0.05, and 0.01, respectively. The null hypothesis of independence is rejected at all three significance levels.

6. Empirical Results

In this section, we present some empirical results to illustrate the performance of maximal correlation test of independence and compare it with two most commonly used independence tests for contingency tables, namely Pearson chi-square test and likelihood ratio test. We first report a numerical study which compares the empirical significance level of maximal correlation test of independence with the nominal significance level. In this study we simulated contingency tables using a loglinear independence model with $\lambda_1^X = 0.2$, $\lambda_2^X = -0.4$, $\lambda_3^X = 0.2$, $\lambda_1^Y = 0.1$, $\lambda_2^Y = -0.3$, $\lambda_3^Y = 0.1$, and $I = J = 3$, and carried out three independence tests: maximal correlation test of independence (labeled M), Pearson chi-squared (P), and likelihood ratio (L) tests of independence. Table 3 presents the rejection proportions for all three tests at nominal significance levels 10%, 5%, and 1%, based on 10,000 simulations. For all significance levels considered, the empirical significance level of maximal correlation independence test and the two other tests are consistent with the nominal significance levels, as the nominal levels are within the corresponding 95% confidence intervals.

Next, we report a numerical study which compares the power performance of maximal correlation test of independence with Pearson chi-square and likelihood ratio tests of independence. Under a given dependence structure, we determine the empirical power of each independence test by simulating a large number of contingency tables and computing the proportion of times the independence hypothesis is rejected at a given significance level α. We report the results by empirical power curves, which are obtained by smoothing the scatter plots of sample sizes versus empirical powers. The smoothing is obtained by the `loess` function in the statistical package R. For each graph we report the sample sizes considered, for example, $n = 20{:}35{:}5$ means sample sizes 20, 25, 30, and 35 are considered. Unless otherwise indicated, large sample results are used for the independence tests. We consider five examples with different dependence structures. In the first example, we

Table 3
Empirical significance of Pearson chi-square (P), likelihood ratio (L), and maximal correlation (M) tests of independence, under loglinear independence model

	$\alpha = 0.1$			$\alpha = 0.05$			$\alpha = 0.01$		
n	P	L	M	P	L	M	P	L	M
50	0.105	0.133	0.098	0.047	0.071	0.047	0.006	0.013	0.007
60	0.095	0.119	0.093	0.045	0.063	0.044	0.008	0.014	0.008
70	0.099	0.119	0.097	0.048	0.062	0.046	0.008	0.013	0.008
80	0.097	0.116	0.099	0.047	0.060	0.046	0.009	0.014	0.008
90	0.101	0.114	0.100	0.048	0.057	0.049	0.009	0.013	0.008
100	0.095	0.107	0.093	0.048	0.056	0.046	0.008	0.011	0.009

generate contingency tables by using a loglinear model. In the remaining examples we assume that the categorical variables have an underlying continuous distribution.

Example 6.1. The first example is based on a saturated loglinear model for 3×3 contingency tables, where we control the dependence by the interaction term λ_{ij}^{XY}. For the case with loglinear parameters $\lambda_i^X = \lambda_i^Y = 0$ for $i = 1, 2, 3$, $\lambda_{11}^{XY} = 0.4$, $\lambda_{12}^{XY} = -0.2$, $\lambda_{13}^{XY} = -0.2$, $\lambda_{21}^{XY} = 0.8$, $\lambda_{22}^{XY} = -0.4$, $\lambda_{23}^{XY} = -0.4$, $\lambda_{31}^{XY} = -1.2$, $\lambda_{32}^{XY} = 0.6$, and $\lambda_{33}^{XY} = 0.6$, contingency tables are generated using (4). Empirical power comparisons at significance level $\alpha = 0.05$ are summarized in Fig. 2. The results are based on 1,000 exact tests for sample sizes $n = 20{:}35{:}5$, and 10,000 tests based on large sample results for sample sizes $n = 40{:}90{:}5$. The simulation results show that all three tests are comparable in terms of power. Likelihood ratio test is slightly more powerful. We carried out similar simulation studies for several other loglinear parameter settings and observed similar results.

Example 6.2. In this example, we consider the continuous variables X and Y which are centered at points uniformly distributed along the unit circle, with $N(0, \sigma^2)$ noise in both coordinates. Let $W \sim U[0, 2\pi]$, $e_1 \sim N(0, \sigma^2)$ and $e_2 \sim N(0, \sigma^2)$. Let $X = \cos(W) + e_1$ and $Y = \sin(W) + e_2$. Here, the random variables of interest are X and Y. A scatter plot of observed X and Y forms a circular cloud of points. For several cases, we generated X and Y, collapsed them into contingency tables, and carried out tests of independence based on the contingency tables. We report the empirical power comparison for the following case.

Let $\sigma = 0.4$, consider 6×6 tables. Independence tests are carried out based on these contingency tables. The empirical power plots of three tests at significance level 0.05 are presented in Fig. 3. The results are based on 10,000 tests for sample sizes $n = 70{:}220{:}10$. The empirical study shows that the likelihood ratio test is more

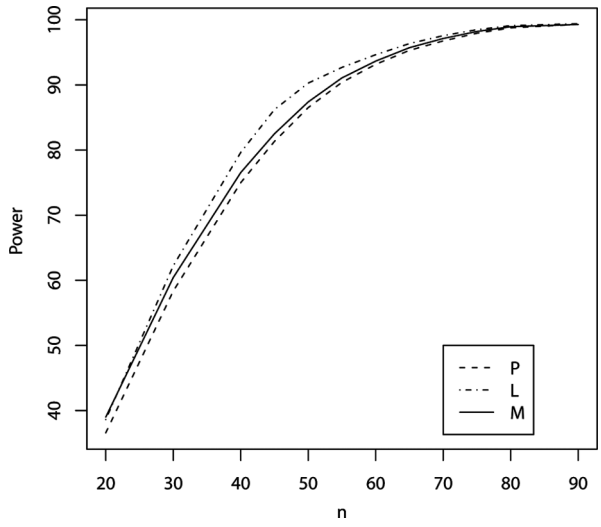

Figure 2. Empirical power of Pearson chi-square (P), likelihood ratio (L), and maximal correlation (M) tests of independence for Example 6.1. Significance level $\alpha = 0.05$.

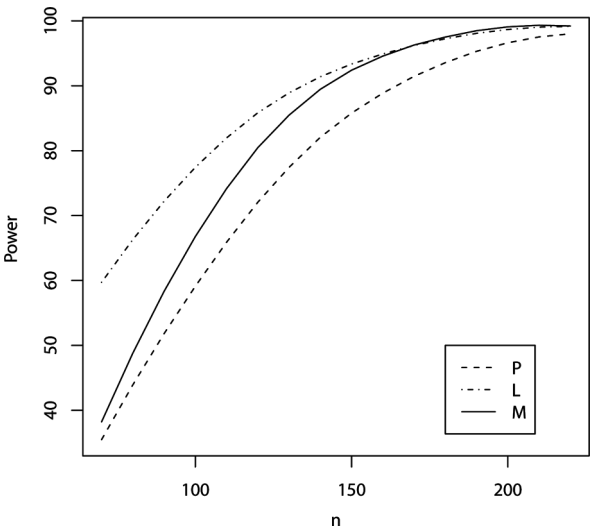

Figure 3. Empirical power of Pearson chi-square (P), likelihood ratio (L), and maximal correlation (M) tests of independence for Example 6.2. Significance level $\alpha = 0.05$.

powerful for sample sizes 70–150. For larger samples, the maximal correlation test is as powerful as the likelihood ratio test.

Example 6.3. In this example we revisit the *Lissajous curve case* discussed in Sec. 3.1. Let the random variable W have uniform distribution over the interval $[0, 2\pi]$. Let $X = \sin aW$ and $Y = \sin bW$ where a and b are integers and $a \neq b$. The random variables of interest are X and Y, which are clearly dependent. Recall from Sec. 3.1 that X and Y are uncorrelated and their maximal correlation is one. For several cases we generated X and Y by transforming from the generated W and adding noise $N(0, \sigma^2)$) on both coordinates. We then collapsed the observations into contingency tables and carried out tests of independence based on the contingency tables.

We will present two Lissajous curve cases here. In the first case, we let $a = 1$, $b = 2$, and $\sigma = 0.03$, which yields a relationship between X and Y on a Lissajous curve (see Fig. 1, Case 1) with some noise added. We collapse the generated X and Y on 5×5 contingency tables and carry out independence tests based on the contingency tables. In this case, exact tests are used where the p-value of the tests are approximated based on 500 simulated tables. For significance level $\alpha = 0.05$, the empirical power plots of three tests are presented in Fig. 4. The results are based on 10,000 tests for sample sizes $n = 50{:}110{:}5$. We observe that maximal correlation test is more powerful in this example.

In the second Lissajous curve case, we set $a = 5$, $b = 6$ (see Fig. 1, Case 2), $\sigma = 0.03$, and we consider 16×16 contingency tables. The empirical power plots of three tests at significance level $\alpha = 0.05$ are presented in Fig. 5. The results are based on 10,000 tests for sample sizes $n = 200{:}300{:}10$. The maximal correlation test is more powerful in this example, and the difference is larger compared to the first case.

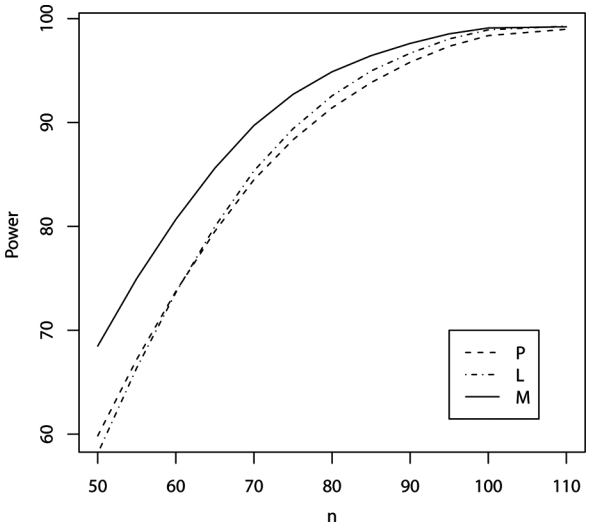

Figure 4. Empirical power of Pearson chi-square (P), likelihood ratio (L), and maximal correlation (M) tests of independence for Example 6.3, Case 1. Significance level $\alpha = 0.05$.

Motivated by Example 6.3, we investigated other cases for which the underlying continuous distributions are uncorrelated but dependent, and we present two of them here.

Example 6.4. Let $U \sim N(0, 1)$ and $V = |U|$. One can show that the dependent variables U and V are uncorrelated. In this example we consider the variables U and V with some noise added. As in the previous examples, we collapse the observations

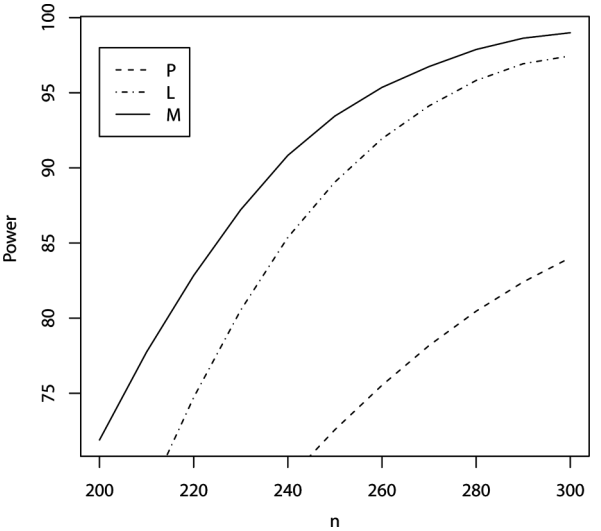

Figure 5. Empirical power of Pearson chi-square (P), likelihood ratio (L), and maximal correlation (M) tests of independence for Example 6.3, Case 2. Significance level $\alpha = 0.05$.

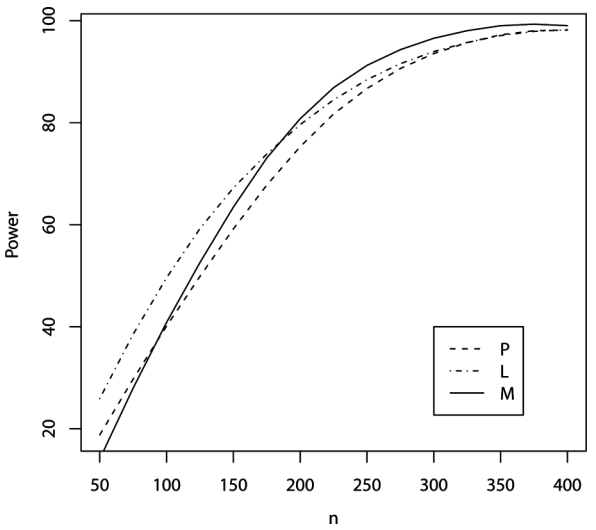

Figure 6. Empirical power of Pearson chi-square (P), likelihood ratio (L), and maximal correlation (M) tests of independence for Example 6.4. Significance level $\alpha = 0.05$.

into contingency tables and perform independence tests based on the contingency tables. We performed empirical power comparisons for several cases and we report the following case.

Let $U \sim N(0, 1)$, $e_1 \sim N(0, \sigma^2)$ and $e_2 \sim N(0, \sigma^2)$, where $\sigma = 0.8$. Let $X = U + e_1$ and $Y = |U| + e_2$. The variables X and Y are generated, and then they are collapsed into 6×6 contingency tables. The empirical power plots of all three tests at significance level 0.05 is presented in Fig. 6. The results are based on 10,000

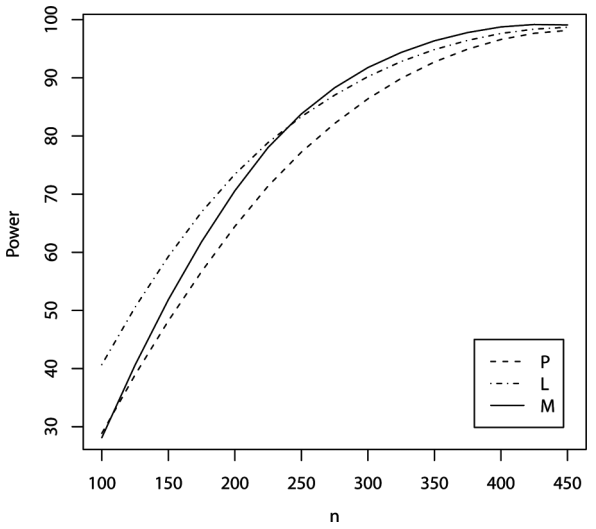

Figure 7. Empirical power of Pearson chi-square (P), likelihood ratio (L), and maximal correlation (M) tests of independence for Example 6.5. Significance level $\alpha = 0.05$.

tests for sample sizes $n = 50{:}400{:}25$. Empirical study shows that, likelihood ratio test is more powerful for sample sizes up to 180. For larger sample sizes, maximal correlation test is more powerful.

Example 6.5. In this example, we consider another case for which the underlying continuous variables are dependent but not correlated. Let $U \sim U(-1, 1)$ and $V = U^2$. One can show that the dependent variables U and V are uncorrelated. Consider the variables U and V with some noise added. We collapse the observations into contingency tables and perform independence tests based on the contingency tables. We performed empirical power comparisons for several cases and we report the following case.

Let $U \sim U(-1, 1)$, $e_1 \sim N(0, \sigma^2)$ and $e_2 \sim N(0, \sigma^2)$, where $\sigma^2 = 0.3$. Let $X = U + e_1$ and $Y = U^2 + e_2$. The variables X and Y are generated, and then they are collapsed into 4×4 contingency tables. Figure 7 presents the empirical power plots of all three tests at significance level 0.05. The results are based on 10,000 tests for sample sizes $n = 100{:}40{:}25$. Similar to Example 6.4, empirical study shows that the likelihood ratio test is more powerful for sample sizes up to 250. For larger sample sizes, maximal correlation test is slightly more powerful.

7. Conclusions

Being a dependence measure with several desirable properties, maximal correlation has been studied by many authors in the statistical literature. In this article, we discussed how maximal correlation can be computed for two-way contingency tables, and we constructed an independence test based on maximal correlation. Given a two-way contingency table, the maximal correlation between the row and column variables has a compact form. Moreover, under independence, the asymptotic distribution of the maximal correlation test statistic has been tabulated since it is related with the distribution of extreme eigenvalues of a Wishart matrix.

We carried out a simulation study to see the empirical power performance of the maximal correlation test and compare it with Pearson chi-squared and likelihood ratio tests of independence. The simulation study consists of several cases with different dependence structures between the row and column variables. When we considered contingency tables generated from loglinear models, we observed that maximal correlation test is comparable to the other two tests. A major advantage of maximal correlation is that, unlike correlation, it vanishes if and only if the variables are independent. When we generated contingency tables such that the underlying continuous variables are uncorrelated but dependent, our simulation results pointed out some cases for which the maximal correlation test appears to be more powerful. A natural extension of this work is the application to higher dimensional contingency tables, and we will consider this as a future project.

References

Abrahams, J., Thomas, J. B. (1980). Properties of the maximal correlation function. *J. Franklin Inst.* 310:317–323.

Agresti, A. (1992). A survey of exact inference for contingency tables. *Statist. Sci.* 7:131–153.

Agresti, A. (2002). *Categorical Data Analysis*. New York: Wiley.

Aldrovandi, R. (2001). *Special Matrices of Mathematical Physics*. Singapore: World Scientific.

Bell, C. B. (1962). Mutual information and maximal correlation as measures of dependence. *Ann. Mathemat. Statist.* 33:587–595.

Breiman, L., Friedman, J. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *J. Amer. Statist. Assoc.* 80:580–619.

Csáki, P., Fischer, J. (1963). On the general notion of maximum correlation. Magyar Tudományos Akad. Mat. Kutató Intézetenk Közleményei (publ. *Math. Inst. Hungar. Acad. Sci.*) 8:27–51.

Dembo, A., Kagan, A., Shepp, L. (2001). Remarks on the maximum correlation coefficient. *Bernoulli* 7:343–350.

Diaconis, P., Strumfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.* 26:363–397.

Fisher, R. A. (1934). *Statistical Methods for Research Workers*. Edinburg: Oliver and Boyd.

Gautam, S., Kimeldorf, G. (1999). Some results on the maximal correlation in 2×2 contingency tables. *Amer. Statist.* 53:336–341.

Gebelein, H. (1941). Das statistische problem der korrelation als variations – und eigenwerthproblem und sein zusammenhang mit der ausgleichsrechnung. *Z. Angew. Math. Mech.* 21:364–379.

Koyak, R. (1987). On measuring internal dependence in a set of random variables. *Ann. Statist.* 15:1215–1228.

Liebetrau, A. M. (2005). *Measures of Association*. London: SAGE Publications.

Novak, S. (2004). On Gebelein's correlation coefficient. *Statist. Probab. Lett.* 69:299–303.

Patefield, W. (1981). Algorithm AS159. An efficient method of generating $r \times c$ tables with given row and column totals. *Appl. Statist.* 30:91–97.

Pearson, E., Hartley, H. (1972). *Biometrika Tables for Statisticians*. Cambridge: Cambridge University Press.

Rényi, A. (1959). On measures of dependence. *Acta Math. Acad. Sci. Hungar.* 10:441–451.

Sethuraman, J. (1990). The asymptotic distribution of Rényi maximal correlation. *Commun. Statist. Theor. Meth.* 19:4291–4298.

Snee, R. D. (1974). Graphical display of two-way contingency tables. *Amer. Statist.* 28:9–12.