Asymptotic Distributions of Empirical Interaction Information

Mariusz Kubkowski & Jan Mielniczuk

Methodology and Computing in Applied Probability

ISSN 1387-5841
Volume 23
Number 1

Methodol Comput Appl Probab (2021)
23:291-315
DOI 10.1007/s11009-020-09783-0

Springer

Methodology and Computing in Applied Probability (2021) 23:291–315
https://doi.org/10.1007/s11009-020-09783-0

Asymptotic Distributions of Empirical Interaction Information

Mariusz Kubkowski[1,2] · **Jan Mielniczuk**[1,2]

Received: 20 February 2019 / Revised: 19 September 2019 / Accepted: 13 March 2020 /
Published online: 16 April 2020

Abstract
Interaction Information is one of the most promising interaction strength measures with many desirable properties. However, its use for interaction detection was hindered by the fact that apart from the simple case of overall independence, asymptotic distribution of its estimate has not been known. In the paper we provide asymptotic distributions of its empirical versions which are needed for formal testing of interactions. We prove that for three-dimensional nominal vector normalized empirical interaction information converges to the normal law unless the distribution coincides with its Kirkwood approximation. In the opposite case the convergence is to the distribution of weighted centred chi square random variables. This case is of special importance as it roughly corresponds to interaction information being zero and the asymptotic distribution can be used for construction of formal tests for interaction detection. The result generalizes result in Han (Inf Control 46(1):26–45 1980) for the case when all coordinate random variables are independent. The derivation relies on studying structure of covariance matrix of asymptotic distribution and its eigenvalues. For the case of $3 \times 3 \times 2$ contingency table corresponding to study of two interacting Single Nucleotide Polymorphisms (SNPs) for prediction of binary outcome, we provide complete description of the asymptotic law and construct approximate critical regions for testing of interactions when two SNPs are possibly dependent. We show in numerical experiments that the test based on the derived asymptotic distribution is easy to implement and yields actual significance levels consistently closer to the nominal ones than the test based on chi square reference distribution.

Keywords Interaction information · Asymptotic weighted chi square distribution · Linkage disequilibrium · Interaction detection

Mathematics Subject Classification (2010) Primary 62G20 · Secondary 60E99

✉ Jan Mielniczuk
j.mielniczuk@ipipan.waw.pl

[1] Institute of Computer Science Polish Academy of Sciences, Jana Kazimierza 5, 01-248 Warsaw, Poland

[2] Faculty of Mathematics and Information Science, Warsaw University of Technology, Koszykowa 75, 00-662 Warsaw, Poland

1 Introduction

Detection of interactions between explanatory variables occurring when predicting a response is an research issue of a fundamental importance which attracted widespread interest. This is due to recognition of the fact that frequently in natural sciences finding the main effects of predictors is not sufficient to explain their overall effect. When two predictors are considered, which is the case we focus on in the paper, one frequently observes their synergetic effect when their joint influence is larger than a sum of the individual effects. In the opposite case, one of the predictors may inhibit the effect of the other meaning that they together affect the response less strongly than the sum of their main effects would indicate. In particular, it may happen that neither of predictors influences individually the response (e.g. occurrence of a certain disease) but jointly they enhance or inhibit occurrence of its certain values. The premier example is detection of gene-gene or gene-environment interactions sought after in Genome Wide Association Studies (GWAS), see e.g. Cordell (2002, 2009.) In the case of a binary response the most popular approach to quantify strength of interaction is a method based on a general methodology due to R. Fisher which consists in fitting a logistic model and testing whether some of the coefficients corresponding to interaction terms are nonzero (see e.g. Agresti 2003). Many other measures have been proposed and there is no general agreement which one reflects adequately biochemical or physiological interaction between genes. However, one of indices which has gained popularity is an entropy-based Interaction Information (II) measure defined in Section 2. This is due to its intuitive definition and the fact that it is a nonparametric measure, thus its interpretation does not depend on any parametric model, which has to fit the data at hand. It is now routinely used in GWAS, see e.g Moore et al. (2006) or Chanda et al. (2008), Sucheston et al. (2010), where AMBIANCE package based on II is described. Moreover, it has been shown formally in Mielniczuk and Teisseyre (2018) by introduction of a partial order on a set of interaction measures, that in many situations Information Interaction is more discriminative than logistic interaction, in particular there may exist nontrivial interactions detected by II in additive logistic regression model which does not contain logistic interactions.

Besides a direct use for detecting of interactions, Interaction Information is widely used in related contexts such as variable selection based on mutual information, where many criterion functions used for selection take interactions into account. We mention e.g. CIFE criterion (cf Lin and Tang 2006) used in greedy selection which for a given candidate predictor is defined as a sum of its mutual information with the response and all information interactions between the candidate and already chosen variables. Many similar methods involving II are reviewed in Brown et al. (2012), see also Meyer et al. (2008).

One of the problems which hindered wider use of II in interaction detection is that statistical properties of its sample counterparts were not fully understood. In particular the asymptotic distribution of the plug-in estimate of II, called $\widehat{II}$ further on, was not known for the general case when II vanishes and thus it was not possible to construct correct rejection regions for such test. In practice, chi square distribution with appropriate number of degrees of freedom is used relying on the result due to Han (cf. Han 1980) who derived asymptotic distribution of II under the assumption of overall independence. This however may lead to a large number of false signals when the pertaining test is employed as the overall independence is only a very special situation when II vanishes. In particular, when one would like to test H_0 against the inhibition effect $H_1 : II < 0$, test based on chi

square reference distribution is intuitively inadequate, as such distribution is supported on positive half-line whereas the left tail of $\widehat{II}$ extends to negative values, resulting in too large critical thresholds. It was also noticed that when dependence between predictors is large then sampling distribution of $\widehat{II}$ deviates strongly from chi square distribution.

The present paper attempts to partially solve this problem when both predictors and the response are nominal variables by deriving the explicit form of asymptotic distribution of $\widehat{II}$ under more general scenario than in Han (1980) which allows for the dependence of predictors. The derivation is algebraic in nature and relies on studying structure of some matrices related to the asymptotic distribution and its eigenvalues. Situation of dependent predictors occurs frequently, e.g. in Genome Wide Associations Studies when dependence of SNPs (linkage disequilibrium) in close proximity is due to crossing-over mechanism and recent methods try to accommodate it (Duggal et al. 2008). It is shown in Theorem 2 that in the case when the underlying distribution is different from its Kirkwood approximation, for a sample consisting of n elements, the asymptotic distribution of $n^{1/2}(\widehat{II} - II)$ is approximately normal. In the opposite case, which includes the situation when both predictors are independent of the response, the distribution of $2n\widehat{II}$ is close to a certain weighted chi square distribution (cf. Section 3.2). For cases of special interest in GWAS studies, when predictors assume three values, the weights of the distribution are determined and the appropriate test is proposed. It is shown by simulations that for the proposed test of independence between predictors and the response actual significance levels (type I errors) are much closer to the nominal ones than significance levels for the test based on chi square distribution. In order to illustrate this point we give below actual significance levels when distribution of two predictors pertains to Clayton copula discussed in Section 4 with parameter θ reflecting dependence and predictors are independent of the binary response. Proposed tests denoted $W(\hat{\hat{\lambda}}_1, \hat{\hat{\lambda}}_2)$ and $Z(\hat{\Lambda})$ pertain to asymptotic distribution of $\widehat{II}$ derived in the paper and its approximation, respectively. They are discussed in Section 4. Chi square test is based on reference chi square distribution with 4 degrees of freedom and nominal significance level $\alpha = 0.05$. We see that the difference between actual and nominal levels is much smaller for the proposed tests than for chi square test (Table 1).

The paper is organized as follows. In Section 2 we introduce some information-theoretic concepts and discuss delta method which is the main technical tool to derive the distribution of $\widehat{II}$. In Sections 3.1 and 3.2 we discuss convergence to normal and weighted chi square laws and reprove Han's result using considered approach. In Section 3.3 we characterize asymptotic law for the case when both predictors assume three values. In Section 4 we introduce resulting test for interaction detection and check its actual levels for several trivariate distributions. Investigation of the power of the proposed test is left for a future research.

Table 1 Actual significance levels of considered test statistics for $H_0 : (X_1, X_2) \perp Y$, $H_1 : II < 0$ with Clayton copula for $n = 1000$, $P(Y = 1) = 0.95$, $\alpha = 0.05$

θ	$W(\hat{\hat{\lambda}}_1, \hat{\hat{\lambda}}_2)$	$Z(\hat{\Lambda})$	χ^2_4
–0.5	0.100	0.094	0.164
1.0	0.049	0.046	0.084
10.0	0.119	0.110	0.325
100.0	0.076	0.072	0.247

2 Preliminaries

2.1 Definitions and Notation

We will consider three-dimensional nominal variable (X_1, X_2, Y) with distribution $P = P_{X_1,X_2,Y}$ such that X_1, X_2 and Y have correspondingly n_{X_1}, n_{X_2} and n_Y possible values. Let $N = n_{X_1} \times n_{X_2} \times n_Y$. We purposefully denote the third coordinate as Y and not X_3 in order to underline the fact that in presented applications two first coordinates correspond to the values of explanatory variables (e.g. Single Nucleotide Polymorphisms (SNPs)) which we use to predict the value of Y. We will denote by p_{ijk} probability $P(X_1 = x_i, X_2 = x_j, Y = y_k)$. Accordingly, with a slight abuse of notion, p_{ij}, p_{ik}, p_{jk} will denote corresponding bivariate probabilities i.e. $p_{ij} = P(X_1 = x_i, X_2 = x_j)$ and p_i, p_j, p_k marginal probabilities i.e. $p_j = P(X_2 = x_j)$. We stress that the distributions of X_1 and X_2 may differ. We will use shorthand notation p for (p_{ijk}), (p_{ij}) or (p_i) depending on the context and throughout assume that all p_{ijk} are positive.

We assume that we observe n independent samples from P and denote by n_{ijk} number of samples with a particular value (x_i, x_j, y_k). Then vector (n_{ijk}) has multinomial distribution $Mult(n, \mathbf{p})$, where $\mathbf{p}$ is N dimensional vector of probabilities. Let $\hat{p}_{ijk} = n_{ijk}/n$.

2.2 Interaction Information

First we recall some concepts developed in Information Theory. Entropy of P is defined as

$$H(X_1, X_2, Y) = -\sum_{i,j,k} p_{ijk} \ln p_{ijk} = -E \ln P(X_1, X_2, Y),$$

with $H(X_1, X_2)$ and $H(X_1)$ defined accordingly. Also, we consider Mutual Information which is a measure quantifying the amount of information obtained about one random variable due to the knowledge of the other random variable. It is defined as

$$I(X_1, X_2) := \sum_{i,j} p_{ij} \log\left(\frac{p_{ij}}{p_i p_j}\right) \tag{1}$$

and thus can be regarded as measure of association for a pair of discrete variables. It determines how different the joint distribution is from the product of factored marginal distributions. More specifically, it is equal to Kullback-Leibler (KL) divergence $KL(P_{X_1,X_2}||P_{X_1} \times P_{X_2})$ between these two distributions

$$I(X_1, X_2) = KL(P_{X_1,X_2}||P_{X_1} \times P_{X_2}) = E_{X_2} KL(P_{X_1|X_2}||P_{X_1})$$

and can be also interpreted as averaged KL divergence between conditional distribution $P_{X_1|X_2}$ and P_{X_1}. The conditional Mutual Information is

$$I(X_1, X_2|Y) := \sum_k p_k \sum_{i,j} p(i, j|k) \log\left(\frac{p(i, j|k)}{p(i|k)p(j|k)}\right) = E_{X_2,Y} KL(P_{X_1|X_2,Y}||P_{X_1|Y}).$$

Note that the conditional Mutual Information is Mutual Information of X_1 and X_2 given Y averaged over values of Y. In the same vain as above $I(X_1; X_2|Y)$ can be interpreted as the expected decrease of amount of uncertainty of (X_1, X_2) when Y is known. The conditional Mutual Information is equal zero if and only if predictors are conditionally independent

given the outcome Y. Analogously to Eq. 1, we define the Mutual Information between a pair (X_1, X_2) and Y as

$$I[(X_1, X_2), Y] := \sum_{i,j,k} p_{ijk} \log\left(\frac{p_{ijk}}{p_{ij}\, p_k}\right) = E_Y KL(P_{X_1,X_2|Y} || P_{X_1,X_2}). \tag{2}$$

The main object of interest here, namely Interaction Information (II) (McGill 1954; Fano 1961) is defined as

$$II(X_1; X_2; Y) := I[(X_1, X_2), Y] - I(X_1, Y) - I(X_2, Y). \tag{3}$$

It follows from the above definition that II can be interpreted as a part of Mutual Information of (X_1, X_2) and Y which is due solely to interaction between X_1 and X_2 in predicting Y i.e. the part of $I[(X_1, X_2); Y]$ which remains after subtraction of individual informations between Y and X_1 and Y and X_2. In other words, II is obtained by removing the main effects from the term describing the overall dependence between Y and the pair (X_1, X_2). Below we discuss some properties of II, which will be used to prove our main results in the next section. Standard representation of $II(X_1; X_2; Y)$ is

$$\begin{aligned} II(X_1; X_2; Y) = &-H(X_1, X_2, Y) + H(X_1, X_2) + H(X_1, Y) + H(X_2, Y) \\ &- H(X_1) - H(X_2) - H(Y), \end{aligned} \tag{4}$$

which also shows that $II(X_1; X_2; Y)$ is symmetric. It is known that II is closely related to so-called Kirkwood superposition approximation (Matsuda 2000; Mielniczuk and Teisseyre 2018) defined as distribution P_K corresponding to mass function

$$p^K_{ijk} = \frac{p_{ij}\, p_{ik}\, p_{jk}}{p_i\, p_j\, p_k}, \tag{5}$$

where the upper index K stands for Kirkwood. Namely, it follows from Eq. 4 that the Interaction Information can be written using Kullback-Leibler divergence between the joint distribution of X_1, X_2, Y and its Kirkwood superposition approximation:

$$II(X_1; X_2; Y) \;=\; KL(P_{X_1,X_2,Y} || P_K) = \sum_{i,j,k} p_{ijk} \log\left(\frac{p_{ijk}}{p^K_{ijk}}\right). \tag{6}$$

Note that Kirkwood approximation is non-normalized i.e. it is not necessarily probability distribution. We define $\eta = \sum_{i,j,k} p^K_{ijk}$ to be the normalizing constant for P_K. The second important property of the Interaction Information is

$$II(X_1; X_2; Y) = I(X_1, X_2|Y) - I(X_1, X_2), \tag{7}$$

which indicates that II measures the influence of a variable Y on the amount of information shared between X_1 and X_2. In other words it quantifies how much Y influences the dependence between X_1 and X_2.

Observe that it follows from Eq. 7 that II in contrast to the Mutual Information can be either positive or negative. Positive value of II indicates that interactions between X_1 and X_2 enhance prediction of Y whereas negative values indicate that interactions diminish or inhibit such prediction. In other words, the conditional dependence is stronger than the unconditional one. The negative value of II indicates that Y weakens or inhibits the dependence between X_1 and X_2. For more detailed discussion and examples of positive and negative II we refer to Mielniczuk and Teisseyre (2018).

Empirical versions of the introduced quantities are defined as their plug-in variants. In particular, we have

$$\hat{H}(X_1) = -\sum_i \hat{p}_i \ln \hat{p}_i$$

$$\hat{I}(X_1, X_2) = \sum_{i,j} \hat{p}_{ij} \ln \frac{\hat{p}_{ij}}{\hat{p}_i \hat{p}_j}$$

$$\hat{I}[(X_1, X_2), Y] = \sum_{i,j,k} \hat{p}_{ijk} \ln \frac{\hat{p}_{ijk}}{\hat{p}_{ij} \hat{p}_k},$$

where $\hat{p}_{ijk} = n_{ijk}/n$ and $\hat{p}_{ij}$ and $\hat{p}_i$, $\hat{p}_j$ are defined analogously. Moreover,

$$\widehat{II}(X_1; X_2; Y) = \hat{I}[(X_1, X_2), Y] - \hat{I}(X_1, Y) - \hat{I}(X_2, Y). \tag{8}$$

Note that $n(\hat{p}_1, \ldots, \hat{p}_{n_{X_1}})$ has multinomial distribution $Mult(n, (p_1, \ldots, p_{n_{X_1}}))$ and analogous property holds for $n(\hat{p}_{11}, \ldots, \hat{p}_{n_{X_1} n_{X_2}})$. This allows us to to use a particular version of delta method stated in Theorem 1. Asymptotic distribution of empirical version of II has been derived by Han in his paper (Han 1980) for a special case when II vanishes, namely when all variables X_1, X_2 and Y are independent. The distribution is is chi-square with $(n_{X_1} - 1)(n_{X_2} - 1)(n_Y - 1)$ degrees of freedom. However, this case is too restrictive for testing purposes when we would like to allow for dependence between predictors X_1 and X_2. No significant progress has been made in this direction since Han's seminal paper. The main technical difficulty in deriving the distribution of $\widehat{II}$ is that its distribution does not follow directly from knowledge of asymptotic distributions of the terms on the rhs of Eq. 8 due to dependence between them. In the paper, we prove that if distribution of (X_1, X_2, Y) is different from its Kirkwood approximation, the limiting distribution of $\widehat{II}$ is normal. In the case when these distributions coincide, the asymptotic distribution, obtained for different normalization of $\widehat{II}$ coincides with distribution of sum of weighted chi square variables.

Finally, we note that modifications of II are used for interaction detection. For example, the measure defined as Kullback-Leibler divergence from normalized Kirkwood superposition approximation $\tilde{P}_K = P_K/\eta$ equals

$$KA(X_1; X_2; Y) := KL(P_{X_1,X_2,Y} || \tilde{P}_K) = II(X_1; X_2; Y) + \log(\eta) \tag{9}$$

can be used as a measure of interaction strength (Wan et al. 2010; Mielniczuk and Rdzanowski 2017).

2.3 Delta Method

Recall that $N = n_{X_1} \times n_{X_2} \times n_Y$. We present here an useful method for determining asymptotic distribution of a certain function of cell frequencies $f(\hat{\mathbf{p}})$ where $\hat{\mathbf{p}} = (\hat{p}_{ijk}) = (\hat{p}_1, \ldots, \hat{p}_N)$ is a vector of frequencies pertaining to multinomial distribution $Mult(n, \mathbf{p})$. It is based on the following Taylor expansion

$$n^{1/2}(f(\hat{\mathbf{p}}) - f(\mathbf{p})) = Df(\mathbf{p})' n^{1/2}(\hat{\mathbf{p}} - \mathbf{p}) + \frac{1}{2} n^{1/2}(\hat{\mathbf{p}} - \mathbf{p})' D^2 f(\mathbf{p})(\hat{\mathbf{p}} - \mathbf{p}) + r_n, \tag{10}$$

where $Df(\mathbf{p})$ and $D^2 f(\mathbf{p})$ denote the first and the second derivative at $\mathbf{p}$ and r_n is the remainder term. The asymptotic distribution of $n^{1/2}(\hat{\mathbf{p}} - \mathbf{p})$ is multivariate normal $N(0, \mathbf{\Sigma})$ with covariance matrix $\mathbf{\Sigma} = (\Sigma_{ijk}^{i'j'k'})$, where $\Sigma_{ijk}^{i'j'k'} = -p_{ijk}p_{i'j'k'} + p_{ijk}I((i, j, k) = (i', j', k'))$. Under appropriate conditions it is shown that the first term of the expansion determines the asymptotic law, which in this case is zero mean normal with a certain

variance σ^2. When σ^2 equals 0, the second term determines the law after increasing normalisation from $n^{1/2}$ to n. The formal result is as follows (see e.g. Agresti 2003 for part (i), with (ii) being an obvious extension of (i). Let $\stackrel{d}{\to}$ denote convergence in distribution and $I(A)$ is the indicator function of event A.

Theorem 1 (i) *Assume that $f : R^N \to R$ is continuously differentiable in the neighbourhood of $\boldsymbol{p} = (p_{ijk}) = (p_1, \ldots, p_N)$. Then*

$$n^{1/2}(f(\hat{\boldsymbol{p}}) - f(\boldsymbol{p})) \stackrel{d}{\to} N(0, \sigma^2), \tag{11}$$

when $n \to \infty$, where

$$\sigma^2 = \sum_{i,j,k} \Big(\frac{\partial f}{\partial p_{ijk}}\Big)^2 p_{ijk} - \Big(\sum_{i,j,k} \frac{\partial f}{\partial p_{ijk}} p_{ijk}\Big)^2 \tag{12}$$

(ii) *If σ^2 in* Eq. 12 *is 0 and f is is twice continuously differentiable in the neighbourhood of $\boldsymbol{p}$, then*

$$2n(f(\hat{\boldsymbol{p}}) - f(\boldsymbol{p})) \stackrel{d}{\to} \boldsymbol{W}' D^2 f(\boldsymbol{p}) \boldsymbol{W}. \tag{13}$$

where $\boldsymbol{W}$ has N-dimensional $N(0, \boldsymbol{\Sigma})$ distribution.

Note that the variance in Eq. 12 coincides with the variance of random variable which takes value $\partial f/\partial p_i$ with probability p_i. Let $\mathbf{H} = D^2 f(\mathbf{p})$. Observe that the distribution of the limit in Eq. 13 can be determined by writing $\mathbf{W} = \boldsymbol{\Sigma}^{1/2}\mathbf{Z}$, where $\mathbf{Z}$ is N-dimensional $N(0, \mathbf{I}_N)$ distribution. Then the asymptotic distribution of $2n(f(\hat{\mathbf{p}}) - f(\mathbf{p}))$ is

$$\mathbf{Z}'\boldsymbol{\Sigma}^{1/2}\mathbf{H}\boldsymbol{\Sigma}^{1/2}\mathbf{Z} = \sum_{i=1}^{N} \lambda_i (\mathbf{x}_i'\mathbf{Z})^2 \tag{14}$$

in view of spectral decomposition (see e.g. Schott 1997, p. 95) of $\boldsymbol{\Sigma}^{1/2}\mathbf{H}\boldsymbol{\Sigma}^{1/2} = \sum_{i=1}^{N} \lambda_i \mathbf{x}_i \mathbf{x}_i'$, where $\mathbf{x}_i$ and λ_i are eigenvectors and corresponding eigenvalues of $\boldsymbol{\Sigma}^{1/2}\mathbf{H}\boldsymbol{\Sigma}^{1/2}$ or equivalently $\mathbf{H}\boldsymbol{\Sigma}$. As $\mathbf{x}_i$ are orthonormal, random variables $\mathbf{x}_i'\mathbf{Z}$ are independent and $N(0, 1)$ distributed, thus the distribution in Eq. 14 equals distribution of a sum of weighted centred chi square variables, which we will call weighted centred chi square distribution.

3 Main Results

3.1 Convergence to Normal Law

We begin by studying $n^{1/2}$-convergence of $\widehat{II}$ and determine when for this normalisation the corresponding limiting normal law is non-degenerate. The equivalent condition is given in terms of probability vector $\mathbf{p} = (p_{ijk})$. We recall that mass function of P_K is defined in Eq. 5.

Theorem 2 *We have (i)*

$$n^{1/2}(\widehat{II} - II) \stackrel{d}{\to} N(0, \sigma_{II}^2), \tag{15}$$

where

$$\sigma_{II}^2 = \sum_{i,j,k} p_{ijk} \ln^2 \frac{p_{ijk}}{p_{ijk}^K} - II^2(X_1, X_2, Y) = Var\left(\ln \frac{p(X_1, X_2, Y)}{p_K(X_1, X_2, Y)}\right)$$

(ii) σ^2_{II} *equals 0 if and only if* $P = P_K$.

Proof Note that in view of Eq. 8 the function $f(p)$ pertaining to II is

$$f(p) = \sum_{i,j,k} p_{ijk} \ln(p_{ijk}/p_{ij}p_k) - \sum_{i,k} p_{ik} \ln(p_{ik}/p_i p_k) - \sum_{j,k} p_{jk} \ln(p_{jk}/p_j p_k). \quad (16)$$

Denote the first term in the above decomposition by $f_1(\mathbf{p}) = \sum_{i,j,k} p_{ijk} \ln(p_{ijk}/p_{ij}p_k)$ and note that

$$\frac{\partial f_1(\mathbf{p})}{\partial p_{ijk}} = \ln \frac{p_{ijk}}{p_{ij}p_k} - 1$$

This follows from differentiating all summands involving p_{ijk} that is being functions of p_{ij} , $p_{ij'}$, $p_{i'j}$ or p_k. Similarly the derivative of the second term in Eq. 16 equals $-\ln(p_{ik}/p_i p_k) + 1$ and the derivative of the third is analogous with i replaced by j. Then

$$w_{ijk} = \frac{\partial f(p)}{\partial p_{ijk}} = \ln \frac{p_{ijk}}{p_{ij}p_k} - \ln \frac{p_{ik}}{p_i p_k} - \ln \frac{p_{jk}}{p_j p_k} + 1 \quad (17)$$

and thus (12) equals

$$\sum_{i,j,k} w^2_{ijk} p_{ijk} - (\sum_{i,j,k} w_{ijk} p_{ijk})^2$$

which coincides with σ^2_{II}.

In order to prove (ii) note that $\sigma^2_{II} = 0$ is equivalent to $p_{ijk}/p^K_{ijk} \equiv C$. We show that $C = 1$. Indeed, summing over k we have

$$p_{ij} = \frac{C p_{ij}}{p_i p_j} \sum_k \frac{p_{ik}p_{jk}}{p_k}$$

and thus ($p_{ij} > 0$)

$$p_i p_j = C \sum_k \frac{p_{ik}p_{jk}}{p_k}.$$

summing over j we thus have

$$p_i = C \sum_k \sum_j \frac{p_{ik}p_{jk}}{p_k} = C p_i$$

and thus $C = 1$. □

Remark 1 Note that that $P = P_K$ is a stronger condition than $II = 0$ and both are equivalent when normalizing constant η of Kirkwood approximation does not exceed 1 (cf Mielniczuk and Teisseyre 2018). It is also shown there that $P = P_K$ implies that P is so-called perfect distribution (cf Darroch 1974). In particular, we have that when (X_1, X_2) are independent of Y then $P = P_K$ and $\sigma^2_{II} = 0$. Note that the case when $II = 0$ and $\sigma^2_{II} > 0$ is possible. This makes behaviour of $\widehat{II}$ when $II = 0$ quite intricate as both $n^{-1/2}$ and n^{-1}-convergence rates are possible.

The situation when $\sigma^2_{II} = 0$ is studied in detail in the next section.

3.2 Convergence to Weighted Chi Square Distribution

In view of Theorem 1 and Eq. 14 when $P = P_K$ we have that

$$2n\widehat{II} \to \sum_{i=1} \lambda_i Z_i^2,$$

where Z_i are independent $N(0,1)$ random variables and λ_i are eigenvalues of $\mathbf{H\Sigma}$, where $\mathbf{\Sigma}$ and $\mathbf{H}$ are defined in Section 2.3. Thus asymptotically $2n\widehat{II}$ is distributed as weighted centred chi-square distribution. We now study the structure of $\mathbf{H\Sigma}$ in more detail. For any matrix $\mathbf{A}$ with row and column indices in $\{1,\ldots,n_{X_1}\}\times\{1,\ldots,n_{X_2}\}\times\{1,\ldots,n_Y\}$ we will denote by $A_{ijk}^{i'j'k'}$ its element with row index equal to ijk and column index $i'j'k'$. In order to keep the notation consistent B_{ij} will be denoted by B_i^j. We consider first $\mathbf{H} = D^2 f(\mathbf{p})$, where $f(\mathbf{p}) = f((p_{ijk}))$ is given in Eq. 16. Let

$$H_{ijk}^{i'j'k'} = \frac{\partial^2 f(\mathbf{p})}{\partial p_{ijk}\partial p_{i'j'k'}}.$$

Lemma 1

$$\begin{aligned} H_{ijk}^{i'j'k'} = 1 - \textstyle\sum_{s\subseteq z}\frac{(-1)^{|s|}}{p_s} = \frac{I(i=i')}{p_i} + \frac{I(j=j')}{p_j} + \frac{I(k=k')}{p_k} \\ - \frac{I(i=i',j=j')}{p_{ij}} - \frac{I(i=i',k=k')}{p_{ik}} - \frac{I(j=j',k=k')}{p_{jk}} \\ + \frac{I(i=i',j=j',k=k')}{p_{ijk}}, \end{aligned}$$

where $z = (\{i\}\cap\{i'\})\cup(\{j\}\cap\{j'\})\cup(\{k\}\cap\{k'\})$, $|s|$ denotes number of elements of s and inclusion $s\subseteq z$ is meant for each index i, j and k separately. This can be directly checked by differentiating (17).

Lemma 2 *Matrix $\boldsymbol{M} := \boldsymbol{H\Sigma}$ has the following form:*

$$M_{ijk}^{i'j'k'} = -p_{i'j'k'}\sum_{s\subseteq z}\frac{(-1)^{|s|}}{p_s}.$$

Proof Recall that $\Sigma_{ijk}^{i'j'k'} = p_{i'j'k'}I(i'=i',j'=j',k'=k') - p_{ijk}p_{i'j'k'}$. Using the previous lemma we have

$$\begin{aligned} M_{ijk}^{i'j'k'} &= \sum_{i'',j'',k''} H_{ijk}^{i''j''k''}\Sigma_{i''j''k''}^{i'j'k'} = \sum_{i'',j'',k''}\Big(\frac{I(i=i'')}{p_i} + \frac{I(j=j'')}{p_j} \\ &+ \frac{I(k=k'')}{p_k} - \frac{I(i=i'',j=j'')}{p_{ij}} - \frac{I(i=i'',k=k'')}{p_{ik}} - \frac{I(j=j'',k=k'')}{p_{jk}} \\ &+ \frac{I(i=i'',j=j'',k=k'')}{p_{ijk}}\Big)\cdot(p_{i'j'k'}I(i'=i'',j'=j'',k'=k'') - p_{i'j'k'}p_{i''j''k''}) = \\ &= p_{i'j'k'}\cdot\Big(\frac{I(i=i')}{p_i} + \frac{I(j=j')}{p_j} + \frac{I(k=k')}{p_k} - \frac{I(i=i',j=j')}{p_{ij}} - \frac{I(i=i',k=k')}{p_{ik}} \\ &- \frac{I(j=j',k=k')}{p_{jk}} + \frac{I(i=i',j=j',k=k')}{p_{ijk}}\Big) - p_{i'j'k'}\Big(\sum_{j'',k''}\frac{p_{ij''k''}}{p_i} + \sum_{i'',k''}\frac{p_{i''jk''}}{p_j} \\ &+ \sum_{i'',j''}\frac{p_{i''j''k}}{p_k} - \sum_{k''}\frac{p_{ijk''}}{p_{ij}} - \sum_{j''}\frac{p_{ij''k}}{p_{ik}} - \sum_{i''}\frac{p_{i''jk}}{p_{jk}} + \frac{p_{ijk}}{p_{ijk}}\Big) \\ &= p_{i'j'k'}\Big(-\sum_{s\subseteq z}\frac{(-1)^{|s|}}{p_s} + 1\Big) - p_{i'j'k'} = -p_{i'j'k'}\sum_{s\subseteq z}\frac{(-1)^{|s|}}{p_s}. \end{aligned}$$

□

Lemma 3 *We have*

$$\mathrm{tr}(\boldsymbol{M}) = (n_{X_1} - 1)(n_{X_2} - 1)(n_Y - 1)$$

Proof

$$\mathrm{tr}(\mathbf{M}) = \sum_{i,j,k} M_{ijk}^{ijk} = -\sum_{i,j,k} p_{ijk}\Big(1 - \frac{1}{p_i} - \frac{1}{p_j} - \frac{1}{p_k} + \frac{1}{p_{ik}} + \frac{1}{p_{jk}} + \frac{1}{p_{ij}} - \frac{1}{p_{ijk}}\Big)$$

$$= -\sum_{i,j,k} p_{ijk} + \sum_i \frac{1}{p_i}\sum_{j,k} p_{ijk} + \sum_j \frac{1}{p_j}\sum_{i,k} p_{ijk} + \sum_k \frac{1}{p_k}\sum_{i,j} p_{ijk} - \sum_{i,j} \frac{1}{p_{ij}}\sum_k p_{ijk}$$

$$- \sum_{i,k} \frac{1}{p_{ik}}\sum_j p_{ijk} - \sum_{j,k} \frac{1}{p_{jk}}\sum_i p_{ijk} + \sum_{i,j,k} \frac{p_{ijk}}{p_{ijk}}$$

$$= -1 + n_{X_1} + n_{X_2} + n_Y - n_{X_1}n_{X_2} - n_{X_1}n_Y - n_{X_2}n_Y + n_{X_1}n_{X_2}n_Y = (n_{X_1} - 1)(n_{X_2} - 1)(n_Y - 1)$$

□

From now on we will assume that (X_1, X_2) and Y are independent. The next lemma states the representation of $\mathbf{M}$ as a Kronecker product of two matrices (for the definition of Kronecker product and its properties we refer to Schott 1997).

Lemma 4 *If (X_1, X_2) and Y are independent, then*

$$\boldsymbol{M} = \boldsymbol{D} \otimes \boldsymbol{C},$$

where $\otimes$ is Kronecker product,

$$C_k^{k'} = -p_{k'} \sum_{s \subseteq \{k\} \cap \{k'\}} \frac{(-1)^{|s|}}{p_s},$$

$$D_{ij}^{i'j'} = p_{i'j'} \sum_{s \subseteq (\{i\} \cap \{i'\}) \cup (\{j\} \cap \{j'\})} \frac{(-1)^{|s|}}{p_s}.$$

Proof Independence of (X_1, X_2) and Y implies that:

$$\sum_{s \subseteq z} \frac{(-1)^{|s|}}{p_s} = 1 - \frac{I(i=i')}{p_i} - \frac{I(j=j')}{p_j} - \frac{I(k=k')}{p_k} + \frac{I(i=i')I(k=k')}{p_i p_k} + \frac{I(j=j')I(k=k')}{p_j p_k} + \frac{I(i=i')I(j=j')}{p_{ij}} - \frac{I(i=i')I(j=j')I(k=k')}{p_{ij} p_k}$$

$$= \Big(1 - \frac{I(k=k')}{p_k}\Big)\Big(1 - \frac{I(i=i')}{p_i} - \frac{I(j=j')}{p_j} + \frac{I(i=i', j=j')}{p_{ij}}\Big)$$

$$= \textstyle\sum_{s \subseteq \{k\} \cap \{k'\}} \frac{(-1)^{|s|}}{p_s} \sum_{s \subseteq (\{i\} \cap \{i'\}) \cup (\{j\} \cap \{j'\})} \frac{(-1)^{|s|}}{p_s}.$$

This ends the proof. □

Lemma 5 *The following properties of $\boldsymbol{C}$ hold:*

(i) $\boldsymbol{C}$ *is idempotent matrix, i.e.* $\boldsymbol{C}^2 = \boldsymbol{C}$.
(ii) $\lambda_1(\boldsymbol{C}) = \ldots = \lambda_{n_Y - 1}(\boldsymbol{C}) = 1, \lambda_{n_Y}(\boldsymbol{C}) = 0.$

Proof Denote locally in the proof $P(Y = y_i)$ by $p(y_i)$, $i = 0, \ldots, n_Y - 1$.

$$\mathbf{C} = \begin{pmatrix} 1 - p(y_0) & -p(y_1) & \ldots & -p(y_{n_Y-1}) \\ -p(y_0) & 1 - p(y_1) & \ldots & -p(y_{n_Y-1}) \\ \vdots & \vdots & \ddots & \vdots \\ -p(y_0) & -p(y_1) & \ldots & 1 - p(y_{n_Y-1}) \end{pmatrix} = \mathbf{I}_{n_Y} - \mathbb{1}_{n_Y} \mathbf{p}_{n_Y}^T,$$

where $\mathbf{I}_{n_Y} \in \mathbb{R}^{n_Y \times n_Y}$ is identity matrix, $\mathbb{1}_{n_Y} \in \mathbb{R}^{n_Y \times 1}$ is the vector of 1's, $\mathbf{p}_{n_Y} = (p(y_0), \ldots, p(y_{n_Y-1}))^T$. We observe that $\mathbf{p}_{n_Y}^T \mathbb{1}_{n_Y} = 1$ and thus we have: $(\mathbf{I}_{n_Y} - \mathbf{C})^2 = \mathbb{1}_{n_Y}(\mathbf{p}_{n_Y}^T \mathbb{1}_{n_Y})\mathbf{p}_{n_Y}^T = \mathbb{1}_{n_Y}\mathbf{p}_{n_Y}^T = \mathbf{I}_{n_Y} - \mathbf{C}$. This means that $\mathbf{I}_{n_Y} - \mathbf{C}$ is idempotent and consequently, $\mathbf{C}$ is idempotent. To prove (ii), note that $\mathrm{tr}(\mathbf{C}) = n_Y - 1$ and $\mathbb{1}_{n_Y}$ is an eigenvector with eigenvalue $\lambda_{n_Y}(\mathbf{C}) = 0$. As $\mathbf{C}$ is idempotent this means that $\lambda_1(\mathbf{C}) = \ldots = \lambda_{n_Y-1}(\mathbf{C}) = 1, \lambda_{n_Y}(\mathbf{C}) = 0$. □

We now show that the proved properties imply Han's theorem (cf Han 1980).

Theorem 3 *If X_1, X_2, Y are all independent, then:*

$$2n\widehat{II} \xrightarrow{d} \chi^2_{(n_{X_1}-1)(n_{X_2}-1)(n_Y-1)}.$$

Proof It is enough to show that $\mathbf{M}$ is idempotent since then from Lemma 3 it follows that it has $(n_{X_1} - 1)(n_{X_2} - 1)(n_Y - 1)$ eigenvalues equal 1 and remaining ones are 0. Thus the asymptotic distribution of $2n\widehat{II}$ is $\chi^2_{(n_{X_1}-1)(n_{X_2}-1)(n_Y-1)}$, as $P_K = P$ in the case of independent variables and thus $\sigma_{II}^2 = 0$. Analogously, as in the proof of Lemma 4, we observe that:

$$M_{ijk}^{i'j'k'} = -p_{i'}p_{j'}p_{k'}\Big(1 - \frac{I(i = i')}{p_i}\Big)\Big(1 - \frac{I(j = j')}{p_j}\Big)\Big(1 - \frac{I(k = k')}{p_k}\Big).$$

This means, that $\mathbf{M} = \mathbf{A} \otimes \mathbf{B} \otimes \mathbf{C}$, where:

$$A_i^{i'} = -p_{i'}\Big(1 - \frac{I(i = i')}{p_i}\Big),$$

$$B_j^{j'} = -p_{j'}\Big(1 - \frac{I(j = j')}{p_j}\Big),$$

and $\mathbf{C}$ is defined in Lemma 5. From the proof of Lemma 5 we know that $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are idempotent. Hence from the mixed-product property of Kronecker product it follows that:

$$\mathbf{M}^2 = (\mathbf{A} \otimes \mathbf{B} \otimes \mathbf{C})(\mathbf{A} \otimes \mathbf{B} \otimes \mathbf{C}) = (\mathbf{A}^2) \otimes (\mathbf{B}^2) \otimes (\mathbf{C}^2) = \mathbf{A} \otimes \mathbf{B} \otimes \mathbf{C} = \mathbf{M}.$$

Thus $\mathbf{M}$ is idempotent. □

Lemma 6 *Matrix $\boldsymbol{D}$ defined in Lemma 4 has at least 1 eigenvalue equal 0 and at least $(n_{X_1} - 1)(n_{X_2} - 1)$ eigenvalues equal 1.*

Proof Eigenvector for the eigenvalue 0 is $\mathbb{1}_{n_{X_1}n_{X_2}}$, as for each row of $\mathbf{D}$ indexed by i, j we have:

$$\begin{aligned}\sum_{i',j'} D_{ij}^{i'j'} &= \sum_{i',j'} p_{i'j'} \sum_{s\subseteq(\{i\}\cap\{i'\})\cup(\{j\}\cap\{j'\})} \frac{(-1)^{|s|}}{p_s} \\ &= \sum_{i',j'} p_{i'j'}\Big(1 - \tfrac{I(i=i')}{p_i} - \tfrac{I(j=j')}{p_j} + \tfrac{I(i=i',j=j')}{p_{ij}}\Big) \\ &= \sum_{i',j'} p_{i'j'} - \sum_{j'} p_{ij'}\tfrac{1}{p_i} - \sum_{i'} p_{i'j}\tfrac{1}{p_j} + 1 = 1 - 1 - 1 + 1 = 0.\end{aligned}$$

Moreover, for any i_0, j_0, such that $i_0 \neq n_{X_1}$, $j_0 \neq n_{X_2}$ we define the vector $\boldsymbol{\alpha} = (\alpha_{ij})$:

$$\alpha_{ij} = \frac{I(i=i_0, j=j_0)}{p_{i_0 j_0}} - \frac{I(i=i_0, j=n_{X_2})}{p_{i_0 n_{X_2}}} - \frac{I(i=n_{X_1}, j=j_0)}{p_{n_{X_1} j_0}} + \frac{I(i=n_{X_1}, j=n_{X_2})}{p_{n_{X_1}n_{X_2}}}.$$

We show that $\boldsymbol{\alpha}$ is an eigenvector of $\mathbf{D}$ corresponding to eigenvalue 1. Indeed, observe that

$$\begin{aligned}\sum_{i',j'} \alpha_{i'j'} D_{ij}^{i'j'} &= 1 - \frac{I(i=i_0)}{p_{i_0}} - \frac{I(j=j_0)}{p_{j_0}} + \frac{I(i=i_0, j=j_0)}{p_{i_0 j_0}} - 1 + \frac{I(i=i_0)}{p_{i_0}} \\ &+ \frac{I(j=n_{X_2})}{p_{n_{X_2}}} - \frac{I(i=i_0, j=n_{X_2})}{p_{i_0 n_{X_2}}} - 1 + \frac{I(i=n_{X_1})}{p_{n_{X_1}}} + \frac{I(j=j_0)}{p_{j_0}} - \frac{I(i=n_{X_1}, j=j_0)}{p_{n_{X_1} j_0}} + 1 \\ &- \frac{I(i=n_{X_1})}{p_{n_{X_1}}} - \frac{I(j=n_{X_2})}{p_{n_{X_2}}} + \frac{I(i=n_{X_1}, j=n_{X_2})}{p_{n_{X_1}n_{X_2}}} = \frac{I(i=i_0, j=j_0)}{p_{i_0 j_0}} \\ &- \frac{I(i=i_0, j=n_{X_2})}{p_{i_0 n_{X_2}}} - \frac{I(i=n_{X_1}, j=j_0)}{p_{n_{X_1} j_0}} + \frac{I(i=n_{X_1}, j=n_{X_2})}{p_{n_{X_1}n_{X_2}}} = \alpha_{ij}.\end{aligned}$$

Thus $\boldsymbol{\alpha}$ is an eigenvector of $\mathbf{D}$. Since there are $(n_{X_1}-1)(n_{X_2}-1)$ such vectors for different i_0, j_0 and they are linearly independent, the lemma is proved. □

3.3 Special Cases

We state now the result which is the main conclusion of the paper.

Theorem 4 *Let $n_{X_1} = n_{X_2} = 3$, $n_Y \geq 2$ and (X_1, X_2) be independent of Y. Then:*

$$2n\widehat{II} \xrightarrow{d} W,$$

where:

$$W = T_1 + \lambda_5(\boldsymbol{M})(T_2 - T_3) + \lambda_7(\boldsymbol{M})(T_4 - T_5),$$

$$T_1 \sim \chi^2_{4(n_Y-1)}, T_2, T_3, T_4, T_5 \sim \chi^2_{n_Y-1},$$

T_1, T_2, T_3, T_4, T_5 are all independent, $\lambda_5(\boldsymbol{M}) = \lambda_5(\boldsymbol{D})$, $\lambda_7(\boldsymbol{M}) = \lambda_7(\boldsymbol{D})$ and $\boldsymbol{D}$ is the matrix defined in Lemma 4.

Proof From the Lemma 4 we know that $\mathbf{M} = \mathbf{D} \otimes \mathbf{C}$, thus eigenvalues of $\mathbf{M}$ have the form $\lambda(\mathbf{M}) = \lambda(\mathbf{D})\lambda(\mathbf{C})$. Thus from the Lemmas 5 and 7 it follows that:

$$\lambda_{l\cdot n_{X_1}n_{X_2}+m}(\mathbf{M}) = \begin{cases} \lambda_m(\mathbf{D}) & l \in \{0,\ldots,n_Y-2\},\ m \in \{1,\ldots,n_{X_1}n_{X_2}\} \\ 0 & l = n_Y - 1,\ m \in \{1,\ldots,n_{X_1}n_{X_2}\}. \end{cases}$$

Note that in this case $P = P_K$, this means that using the delta method we obtain: $2n\widehat{II} \xrightarrow{d} W$, where

$$W = \sum_{i=1}^{n_{X_1} n_{X_2} n_Y} \lambda_i(\mathbf{M}) Z_i^2 = \sum_{i=1}^{n_{X_1} n_{X_2} (n_Y - 1)} \lambda_i(\mathbf{M}) Z_i^2$$

and $(Z_i)_i$ are independent standard normal random variables. Note that 1 is an eigenvalue of multiplicity $(n_{X_1} - 1)(n_{X_2} - 1)(n_Y - 1) = 4(n_Y - 1)$ and 0, $\lambda_5(\mathbf{D})$, $-\lambda_5(\mathbf{D})$, $\lambda_7(\mathbf{D})$ and $-\lambda_7(\mathbf{D})$ have all multiplicity $n_Y - 1$. Now we define:

$F_j = \sum_{k=1}^{n_Y - 1} Z_{(k-1)n_{X_1} n_{X_2} + j} \sim \chi^2_{n_Y - 1}$ for $j \in \{1, \ldots, n_{X_1} n_{X_2}\}$,

$T_1 = F_1 + F_2 + F_3 + F_4 \sim \chi^2_{4(n_Y - 1)}$, $T_k = F_{k+3}$ for $k = 2, 3, 4, 5$. We do not need to define T_6, as $\lambda_9(\mathbf{D}) = 0$. This ends the proof. □

In the special case when Y is binary i.e $n_Y = 2$ we have $W = T_1 + \lambda_5(\mathbf{D})(Z_1^2 - Z_2^2) + \lambda_7(\mathbf{D})(Z_3^2 - Z_4^2)$ where $T_1 \sim \chi^2_4$ and $Z_1, \ldots, Z_4$ are independent $N(0, 1)$ random variables independent of T_1.

Corollary 1 *Let $n_{X_1}, n_{X_2} \in \{2, 3\}$ and $n_{X_1} = 2$ or $n_{X_2} = 2$. Let $n_Y \geq 2$ and (X_1, X_2) be independent of Y. Then:*

$$2n\widehat{II} \xrightarrow{d} W,$$

where:

$$W = T_1 + \sqrt{H_1}(T_2 - T_3),$$
$$T_1 \sim \chi^2_{(n_{X_1} - 1)(n_{X_2} - 1)(n_Y - 1)},\ T_2, T_3 \sim \chi^2_{n_Y - 1}$$

and T_1, T_2 and T_3 are independent.

We discuss now the case when $n_{X_1} = n_{X_2} = 3$ and n_Y equals 2 which corresponds to an important case of two SNPs interacting with an binary outcome. We will show that in this case, or more generally, when Y admits n_Y values, the distribution of Y can be explicitly described.

Lemma 7 *Let $n_{X_1} = n_{X_2} = 3$,*

$$H_1 = \sum_{i,j} \frac{p_{ij}^2}{p_i p_j} - 1 = \sum_{i,j} \frac{(p_{ij} - p_i p_j)^2}{p_i p_j},$$

$$H_2 = \sum_{i,j,i',j'} \frac{p_{ij} p_{ij'} p_{i'j'} p_{i'j}}{p_i p_j p_{i'} p_{j'}} - 1.$$

Then the eigenvalues of matrix $\boldsymbol{D} \in \mathbb{R}^{9 \times 9}$ are as follows: $\lambda_1(\boldsymbol{D}) = \ldots = \lambda_4(\boldsymbol{D}) = 1$, $\lambda_9(\boldsymbol{D}) = 0$, $\lambda_5(\boldsymbol{D}) = -\lambda_6(\boldsymbol{D})$, $\lambda_7(\boldsymbol{D}) = -\lambda_8(\boldsymbol{D})$, and

$$\lambda_5^2(\boldsymbol{D}) = \frac{H_1 + \sqrt{\Delta}}{2},\ \lambda_7^2(\boldsymbol{D}) = \frac{H_1 - \sqrt{\Delta}}{2}, \tag{18}$$

where $\Delta = 2H_2 - H_1^2$. If X_1, X_2 and Y are independent then $\lambda_5(\boldsymbol{D}) = \ldots = \lambda_9(\boldsymbol{D}) = 0$.

Technical proof of the lemma is moved to the Appendix.

Remark 2 (i) Note that all eigenvalues $\lambda_i(\mathbf{D})$ are real as they are eigenvalues of the symmetric matrix $\boldsymbol{\Sigma}^{1/2} \mathbf{H} \boldsymbol{\Sigma}^{1/2}$, thus it follows that $\Delta \geq 0$. We also remark that Δ may be 0

for dependent X_1, X_2 as it happens for $X_1 = X_2$ having any nondegenerate distribution on $\{1, 2, 3\}$ since then $H_1 = H_2 = 2$. Furthermore, for arbitrary n_{X_1} and n_{X_2} from the inequality

$$H_1 \leq \sum_{i,j} \frac{p_{ij} p_i}{p_i p_j} - 1 = \sum_j 1 - 1 = n_{X_2} - 1$$

it follows that $H_1 \leq \min(n_{X_1}, n_{X_2}) - 1$ and the upper bound is attained. The same inequality holds for H_2. Besides, from the Jensen inequality we have

$$I(X_1, X_2) = \sum_{i,j} p_{ij} \ln\left(\frac{p_{ij}}{p_i p_j}\right) \leq \ln\left(\sum_{i,j} \frac{p_{ij}^2}{p_i p_j}\right) = \ln(H_1 + 1).$$

(ii) Moreover, observe that H_2 defined in the last Lemma can be represented as

$$H_2 = \sum_{i,j,i',j'} \frac{(p_{ij} - p_i p_j)(p_{ij'} - p_i p_{j'})(p_{i'j} - p_{i'} p_j)(p_{i'j'} - p_{i'} p_{j'})}{p_i p_j p_{i'} p_{j'}}$$

Note also that H_1 equals chi square index of the distribution of (X_1, X_2).

By using the same method as for Lemma 7, we obtain

Corollary 2 *The following statements hold:*

(i) *For* $(n_{X_1}, n_{X_2}) = (3, 2)$ *or* $(n_{X_1}, n_{X_2}) = (2, 3)$: $\lambda_1(\boldsymbol{D}) = \lambda_2(\boldsymbol{D}) = 1,\ \lambda_5(\boldsymbol{D}) = \lambda_6(\boldsymbol{D}) = 0,\ \lambda_3(\boldsymbol{D}) = \sqrt{H_1},\ \lambda_4(\boldsymbol{D}) = -\sqrt{H_1}$.

(ii) *For* $n_{X_1} = n_{X_2} = 2 : \lambda_1(\boldsymbol{D}) = 1,\ \lambda_4(\boldsymbol{D}) = 0,\ \lambda_2(\boldsymbol{D}) = \sqrt{H_1},\ \lambda_3(\boldsymbol{D}) = -\sqrt{H_1}$.

Remark 3 We stress that in Theorem 3 the response Y can take one of an arbitrary number of discrete values, thus the result is applicable e.g. to multiclass classification problems. For the case when X_1 or X_2 admit more than 3 values we note that matrix D defined in Lemma 4 has at least $(n_{X_1} - 1)(n_{X_2} - 1)$ eigenvalues equal to 1 and one equal to 0. Thus in order to determine the distribution of W we need to compute $n_{X_1} + n_{X_2} - 2$ remaining eigenvalues of D. In view of the proof of the lemma this would involve determining $n_{X_1} + n_{X_2} - 2$ powers of D which is computationally challenging for larger values of n_{X_i}. However, as a polynomial Q (cf Eq. 21) has in general case every second coefficient equal to 0, we conjecture that explicit formulae for λ_i should be calculable for $n_{X_1} + n_{X_2} \leq 10$. The alternative method of determining distribution of W using permutations is described in Remark 4.

4 Numerical Experiments

In the following we apply Theorem 4 to construct test for the hypothesis that (X_1, Y_1) are independent of Y:

$$H_0 : (X_1, X_2) \perp Y$$

against two-sided alternative

$$H_1 : II \neq 0$$

Note that it follows from the discussion in Section 2.2 that H_0 implies $II = 0$ and thus the null and the alternative hypotheses describe disjoint events. Such set-up is useful in cases when a researcher focuses on detection of interaction between explanatory variables in predicting the response and not on their main effects. We remark that it would be also desirable to develop tests for a broader null hypotheses $H_0^{(1)} : II = 0$ (no interaction),

or $H_0^{(2)} : X_1 \perp Y \ \& X_2 \perp Y \ \& \ II = 0$ (no interaction and no main effects), but the asymptotic distributions of $\widehat{II}$ under such null hypotheses are intractable. Note that H_0 allows for dependence of explanatory variables which is a common case e.g. in Genome Wide Association Studies (GWAS).

To fix ideas, we consider here the case of a binary response Y i.e. the case when $n_Y = 2$. We consider the test with critical region

$$\mathcal{C} = \{\widehat{II} : 2n\widehat{II} < W_{\alpha/2} \text{ or } 2n\widehat{II} > W_{1-\alpha/2}\}, \tag{19}$$

where $W_{\alpha/2}$ is an quantile of order α for distribution W defined in Theorem 4, that is we reject H_0 when the observed value $\widehat{II}$ belongs to $\mathcal{C}$. It follows from Theorem 4 that such test has asymptotic significance level α. We investigate how the quantiles needed in test construction can be approximated and in consequence how well actual significance levels pertaining to asymptotic quantiles correspond to nominal ones. We also consider an analogous test based on χ^2_4 approximation frequently used in GWAS (cf. e.g. Chanda et al. 2008 and Wan et al. 2010). We note that the asymptotic distribution of $2n\widehat{II}$ is χ^2_4 only under assumption that all three variables X_1, X_2 and Y are independent. This is very restrictive and the null hypothesis above is is much more general. We will show in the following that using χ^2_4 distribution instead of distribution of W leads to significant increase of false rejections under H_0, especially in the case when two-sided alternative is considered.

Table 2 Quantiles of different test statistics for normal copula

Reference distribution	θ	Quantile order			
		0.025	0.05	0.95	0.975
χ^2_4	–	0.4844	0.7107	9.4877	11.1433
$W(\bar{\lambda}_1, \bar{\lambda}_2)$	0.0	0.4844	0.7107	9.4877	11.1433
$W(\hat{\bar{\lambda}}_1, \hat{\bar{\lambda}}_2)$		0.4713	0.7031	9.4901	11.1449
$Z(\Lambda)$		0.4844	0.7107	9.4877	11.1433
$Z(\hat{\Lambda})$		0.4747	0.7027	9.4918	11.1467
$W(\bar{\lambda}_1, \bar{\lambda}_2)$	0.3	0.3107	0.6020	9.5351	11.1918
$W(\hat{\bar{\lambda}}_1, \hat{\bar{\lambda}}_2)$		0.3002	0.5956	9.5382	11.1951
$Z(\Lambda)$		0.3428	0.5948	9.5466	11.1936
$Z(\hat{\Lambda})$		0.3340	0.5877	9.5502	11.1968
$W(\bar{\lambda}_1, \bar{\lambda}_2)$	0.5	0.0004	0.4154	9.6318	11.2938
$W(\hat{\bar{\lambda}}_1, \hat{\bar{\lambda}}_2)$		–0.0089	0.4101	9.6351	11.2972
$Z(\Lambda)$		0.0767	0.3786	9.6577	11.2909
$Z(\hat{\Lambda})$		0.0687	0.3721	9.6612	11.2939
$W(\bar{\lambda}_1, \bar{\lambda}_2)$	0.7	–0.5460	0.0819	9.8142	11.4900
$W(\hat{\bar{\lambda}}_1, \hat{\bar{\lambda}}_2)$		–0.5506	0.0788	9.8164	11.4926
$Z(\Lambda)$		–0.3812	0.0100	9.8535	11.4688
$Z(\hat{\Lambda})$		–0.3858	0.0063	9.8556	11.4709
$W(\bar{\lambda}_1, \bar{\lambda}_2)$	0.9	–1.5503	–0.6177	10.2204	11.9410
$W(\hat{\bar{\lambda}}_1, \hat{\bar{\lambda}}_2)$		–1.5499	–0.6172	10.2208	11.9416
$Z(\Lambda)$		–1.2806	–0.7077	10.2649	11.8670
$Z(\hat{\Lambda})$		–1.2806	–0.7077	10.2651	11.8673

4.1 Approximation of Distribution of *W*

In the following we denote by W the limiting distribution in Theorem 4. Note that for the $3 \times 3 \times 2$ case W is parametrized by λ_5 and λ_7 i.e. $W = W(\lambda_5, \lambda_7)$, where λ_5, λ_7 are given by Eq. 18. In order to simplify the notation we let $\bar{\lambda}_1 := \lambda_5$ and $\bar{\lambda}_2 := \lambda_7$. Thus for testing purposes $W(\bar{\lambda}_1, \bar{\lambda}_2)$ is approximated by $W(\hat{\bar{\lambda}}_1, \hat{\bar{\lambda}}_2)$, where $\hat{\bar{\lambda}}_i$ are the plug-in estimators of $\bar{\lambda}_i$ based on Eq. 18. The values of quantiles of $W(\hat{\bar{\lambda}}_1, \hat{\bar{\lambda}}_2)$ can be numerically calculated using R package `distr` or Python package `Pacal`. The differences between the results are minor and the results presented below are obtained with the aid of `distr`. We checked that for sample sizes larger than 100 and when the fixed type of dependence is considered (and thus $\bar{\lambda}_1$ and $\bar{\lambda}_2$ are fixed), quantiles of order 0.95 and 0.975 as well as 0.025 and 0.05 of $W(\bar{\lambda}_1, \bar{\lambda}_2)$ are very well approximated by quantiles of $W(\hat{\bar{\lambda}}_1, \hat{\bar{\lambda}}_2)$. In Table 2 below we show how expected values of quantiles $W(\hat{\bar{\lambda}}_1, \hat{\bar{\lambda}}_2)$ compare with quantiles of $W(\bar{\lambda}_1, \bar{\lambda}_2)$ when distribution of (X_1, X_2) is given by normal copula discussed below for $n = 1000$, $P(Y = 1) = 0.9$ and number of repetitions $L = 1000$. Parameter θ corresponds to correlation coefficient. The analogous results for Clayton copula with parameter θ discussed in Section 4.2 are given in Table 3. Thus, despite more complicated form of distribution of $W(\bar{\lambda}_1, \bar{\lambda}_2)$ than chi square distribution approximation of its quantiles does not pose any computational difficulties.

Moreover, it is known that distributions of weighted chi square random variables can be adequately approximated by $\alpha\chi^2_d + \beta$, where χ^2_d is generalized chi square distribution with $d \in R^+$ and parameters α, d and β are chosen to match three first moments of $W(\bar{\lambda}_1, \bar{\lambda}_2)$

Table 3 Quantiles of different test statistics for Clayton copula

Reference distribution	θ	Quantile order			
		0.025	0.05	0.95	0.975
χ^2_4	–	0.4844	0.7107	9.4877	11.1433
$W(\bar{\lambda}_1, \bar{\lambda}_2)$	-0.5	0.0006	0.4162	9.6329	11.2949
$W(\hat{\bar{\lambda}}_1, \hat{\bar{\lambda}}_2)$		–0.0077	0.4114	9.6356	11.2978
$Z(\Lambda)$		0.0744	0.3767	9.6587	11.2917
$Z(\hat{\Lambda})$		0.0675	0.3712	9.6616	11.2944
$W(\bar{\lambda}_1, \bar{\lambda}_2)$	1	–0.0788	0.3731	9.6603	11.3242
$W(\hat{\bar{\lambda}}_1, \hat{\bar{\lambda}}_2)$		–0.0859	0.3690	9.6626	11.3267
$Z(\Lambda)$		0.0036	0.3195	9.6885	11.3183
$Z(\hat{\Lambda})$		–0.0018	0.3151	9.6909	11.3205
$W(\bar{\lambda}_1, \bar{\lambda}_2)$	10	–2.2559	–1.1236	10.5610	12.3365
$W(\hat{\bar{\lambda}}_1, \hat{\bar{\lambda}}_2)$		–2.2606	–1.1270	10.5636	12.3397
$Z(\Lambda)$		–1.9447	–1.2367	10.5979	12.2079
$Z(\hat{\Lambda})$		–1.9493	–1.2404	10.6004	12.2106
$W(\bar{\lambda}_1, \bar{\lambda}_2)$	100	–3.1107	–1.7543	11.0131	12.8741
$W(\hat{\bar{\lambda}}_1, \hat{\bar{\lambda}}_2)$		–3.1108	–1.7544	11.0131	12.8742
$Z(\Lambda)$		–2.7525	–1.8822	11.0387	12.6769
$Z(\hat{\Lambda})$		–2.7526	–1.8823	11.0387	12.6769

(see Zhang 2005). It follows from formulas (5) in Zhang's paper that in the case considered here when the first four non-zero eigenvalues are 1 and remaining four form two symmetric pairs $\pm\bar{\lambda}_i$, the approximating distribution depends only on $\Lambda = \bar{\lambda}_1^2 + \bar{\lambda}_2^2$ and thus it is called $Z(\Lambda)$. Analogously, plug-in version of $Z(\Lambda)$ is denoted by $Z(\hat{\Lambda})$. It will follow that for purposes of approximating critical region $\mathcal{C}$ quantiles of $Z(\hat{\Lambda})$ can be used instead that of $W(\hat{\bar{\lambda}}_1, \hat{\bar{\lambda}}_2)$ (see Tables 2 and 3) which also contains values of quantiles of corresponding $Z(\Lambda)$ and averaged values of quantiles of $Z(\hat{\Lambda})$). The only exception is when left tail is considered and the dependence is strong (see e.g. the results in Table 2 for quantiles of order 0.025 and 0.05 and $\theta = 0.9$).

When $W(\bar{\lambda}_1, \bar{\lambda}_2)$ is compared with χ_4^2 distribution we see that in the case of normal copula mainly the left tails differ when the dependence becomes strong. In the case of Clayton copula both tails of these two distribution are significantly different when θ is large.

Remark 4 When X_1 or X_2 admit more than three values we established that distribution of W coincides with distribution of weighted sum of independent chi squares with weights λ_i. As λ_is are unknown the distribution of W can be alternatively approximated using permutation method. This would involve calculation of L permutations of the original sample with (X_1, X_2) being permuted. Then as obtained samples satisfy null hypothesis H_0 the empirical distribution of values $\widehat{II}_1, \ldots, \widehat{II}_L$ will approximate that of W. A main problem with this approach is that as we are actually interested in quantiles $W_{\alpha/2}$ and $W_{1-\alpha/2}$ we would need large number of permuted sampled to approximate them precisely which is computationally demanding for larger n (for discussion of this in the context of testing for conditional independence see e.g. Tsamardinos and Borboudakis 2010). The way to circumvent this problem would be to approximate permutation distribution using weighted chi square distribution which would involve only estimation of moments and would require significantly smaller number of permutations. This problem will be the subject of of future research.

4.2 Dependence Models of (X_1, X_2)

In order to check how dependence between X_1 and X_2 affects the asymptotic distribution W and the way it influences actual significance levels of the tests based on $\widehat{II}$ we investigated discretized versions of four distributions pertaining to popular copulas. Copula C is defined as a bivariate function defined on the unit square which satisfies

$$F(x_1, x_2) = C(F_1(x_1), F_2(x_2)), \tag{20}$$

where F is a distribution function corresponding to P_{X_1,X_2} and F_1 and F_2 are the corresponding marginal distribution functions. Here we consider normal copula together with Clayton, Gumbel and Frank copulas. They are described in standard reference texts (see e.g. Nelsen 2006), here it suffices to state that they are all parametrized by a parameter θ, which in case of discretized normal copula corresponds to correlation coefficient of original normal variables. We assume that X_1 and X_2 take one of the values 1,2 or 3, $P(X_1 = 1) = 0.25$, $P(X_1 = 2) = 0.5$, $P(X_1 = 3) = 0.25$ and $X_2 \stackrel{d}{=} X_1$. Thus the marginals satisfy Hardy-Weinberg hypothesis with $p = q = 0.5$. The distribution function given in Eq. 20 is then discretized to atoms (i, j) with $i, j = 1, 2, 3$ e.g. $P(X_1 = 1, X_2 = 1) = C(0.75, 0.25)$, $P(X_1 = 2, X_2 = 1) = C(0.75, 0.25) - C(0.25, 0.25)$ and so on. Binary response Y is generated independently from (X_1, X_2) and such that $P(Y = 0) = 0.05$. In genetic applications $P(Y = 0)$ may correspond to prevalence of a disease. Thus $H_0 : II = 0$ is satisfied.

We also consider the following joint distributions of (X_1, X_2) (which we call later circular distribution and quasi-diagonal distribution, respectively), given by the matrices:

$$\begin{bmatrix} P(X_1=1,X_2=1) & P(X_1=1,X_2=2) & P(X_1=1,X_2=3) \\ P(X_1=2,X_2=1) & P(X_1=2,X_2=2) & P(X_1=2,X_2=3) \\ P(X_1=3,X_2=1) & P(X_1=3,X_2=2) & P(X_1=3,X_2=3) \end{bmatrix} = \begin{bmatrix} \frac{1}{8}-\frac{\theta}{2} & \theta & \frac{1}{8}-\frac{\theta}{2} \\ \theta & \frac{1}{2}-2\theta & \theta \\ \frac{1}{8}-\frac{\theta}{2} & \theta & \frac{1}{8}-\frac{\theta}{2} \end{bmatrix},$$

for $\theta \in \left(0, \frac{1}{4}\right)$, and

$$\begin{bmatrix} P(X_1=1,X_2=1) & P(X_1=1,X_2=2) & P(X_1=1,X_2=3) \\ P(X_1=2,X_2=1) & P(X_1=2,X_2=2) & P(X_1=2,X_2=3) \\ P(X_1=3,X_2=1) & P(X_1=3,X_2=2) & P(X_1=3,X_2=3) \end{bmatrix} = \begin{bmatrix} r\theta & rq & rq \\ rq & r^2\theta & rq \\ rq & rq & r\theta \end{bmatrix},$$

for $\theta \in \left(0, \frac{1}{8}\right)$, where $r = (2(\sqrt{\theta(\theta+1)} - \theta))^{-1}$, $q = \frac{-3\theta+\sqrt{\theta(\theta+1)}}{4}$.

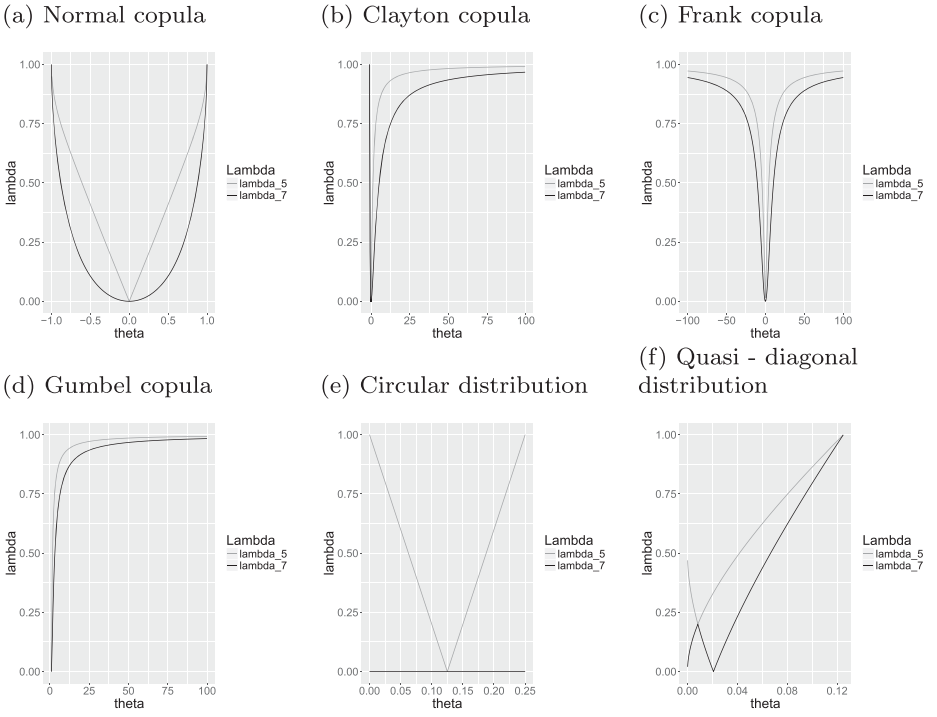

Fig. 1 Behaviour of $\lambda_5 = \bar{\lambda}_1$ and $\lambda_7 = \bar{\lambda}_2$ for different copulas

It turns out that $\bar{\lambda}_1$ and $\bar{\lambda}_2$ behave very differently as a function of θ for considered copulas and two introduced distributions (see Fig. 1). Recall that the difference between their squared values is described by $\sqrt{\Delta} = \bar{\lambda}_1^2 - \bar{\lambda}_2^2$.

We also show in Fig. 2 how parameter θ influences dependence between X_1 and X_2 measured by their mutual information. Note that dependence of X_1 and X_2 corresponds to linkage disequilibrium which is of interest in genetics.

Moreover Fig. 3 shows the the discrepancies between the true distribution and its two approximations for quasi-diagonal copula. It is worthwhile to compare this figure and the corresponding panel in Fig. 4 for this distribution to see the influence of the lack of fit on actual type I error. It can be also seen that that the approximation of the empirical distribution by the asymptotic distribution is the least accurate for the strongest dependence between predictors ($\theta = 0.12$ Fig. 2).

4.3 Actual Significance Levels

Below we present analysis how significance levels of the test based on $\widehat{II}$ differ from the nominal levels when the reference distribution is either $W(\hat{\bar{\lambda}}_1, \hat{\bar{\lambda}}_2)$ or χ^2_4. The figures below are based on $L = 1000$ repetitions, for each repetition critical values for critical region $\mathcal{C}$ in Eq. 19 were calculated based on $W(\hat{\bar{\lambda}}_1, \hat{\bar{\lambda}}_2)$ distribution. Nominal level was set at $\alpha = 0.05$. It follows from the figures that the proposed test based on $W(\hat{\bar{\lambda}}_1, \hat{\bar{\lambda}}_2)$ distribution yields actual levels of significance much closer to the nominal one than the test based on χ^2_4 distribution. This is due to mainly to much better approximation of the left tail of the distribution of $\widehat{II}$, which extends to the negative values, by the left tail of $W(\hat{\bar{\lambda}}_1, \hat{\bar{\lambda}}_2)$. In contrast the

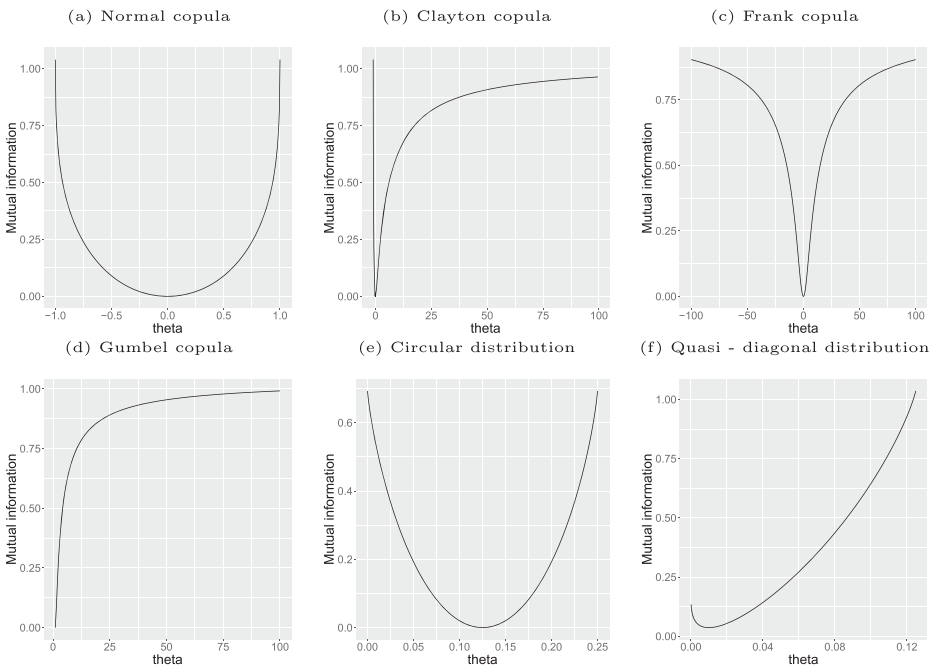

Fig. 2 Behaviour of $I(X_1, X_2)$ for different copulas

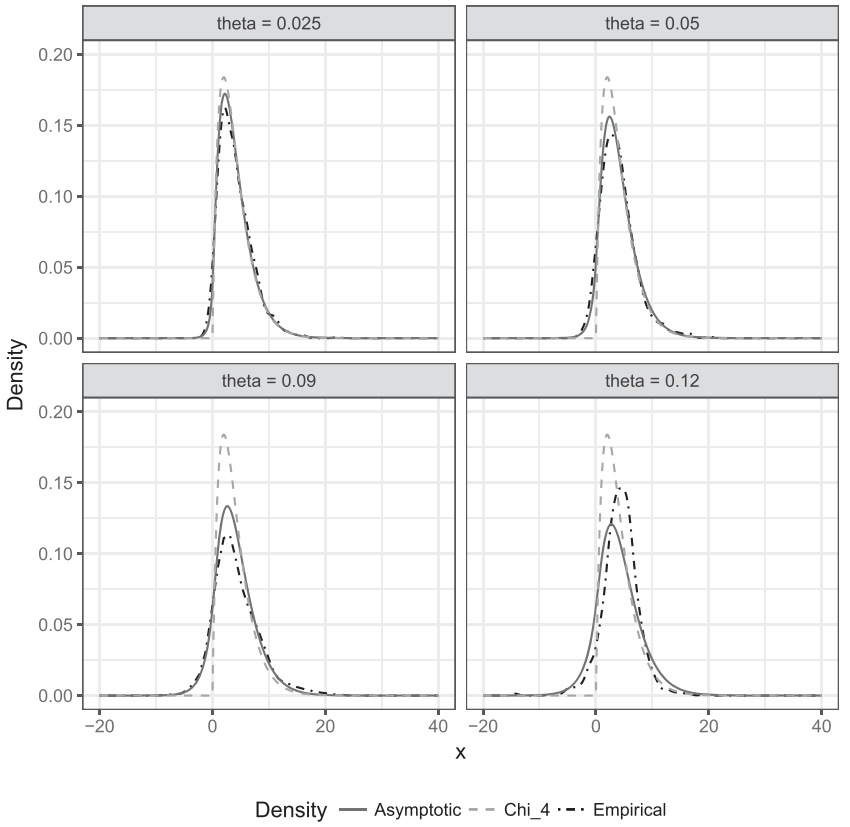

Fig. 3 Quasi-diagonal copula - comparison of empirical density of $2n\widehat{II}$ with the densities of χ^2_4 and $W(\bar{\lambda}_1, \bar{\lambda}_2)$ for $n = 4000$ and $P(Y = 1) = 0.95$

distribution χ^2 is supported on positive part of a real line and it yields poor approximation of lower quantiles of distribution of $\widehat{II}$. Upper quantiles of $\widehat{II}$ are moderately well approximated by that of χ^2 and this explains why for one sided alternative $H_1 : II > 0$ the differences between nominal level based on χ^2_4 approximation and actual leveles are often smaller (cf Mielniczuk and Rdzanowski 2017). In contrast, for two-sided tests approximation of both lower and upper quantiles plays significant role and then test based on chi square approximation performs poorly. We note that in all the cases considered actual significance levels for chi square tests were larger than for the proposed test and larger than the nominal level 0.05. Thus when the power of these tests is compared this would lead to an erroneous conclusion that the chi square test is more powerful, which is solely due to the lack of control of the significance level. We also note there are cases when the proposed test, although it performs better than chi square test is still much too liberal even for large sample sizes: see e.g. the case of normal copula for $\theta = 0.7$ and quasi-diagonal distribution for $\theta = 0.09$. These cases correspond to situations when dependence between X_1 and X_2 becomes strong and this affects the speed with which distribution of $2n\widehat{II}$ converges to the asymptotic distribution. In general, the control of significance level is more accurate when the asymptotic distribution is close to chi square distribution. This include the cases of weak linkage disequilibrium and the cases when both λ_1 and λ_2 are close to 1.

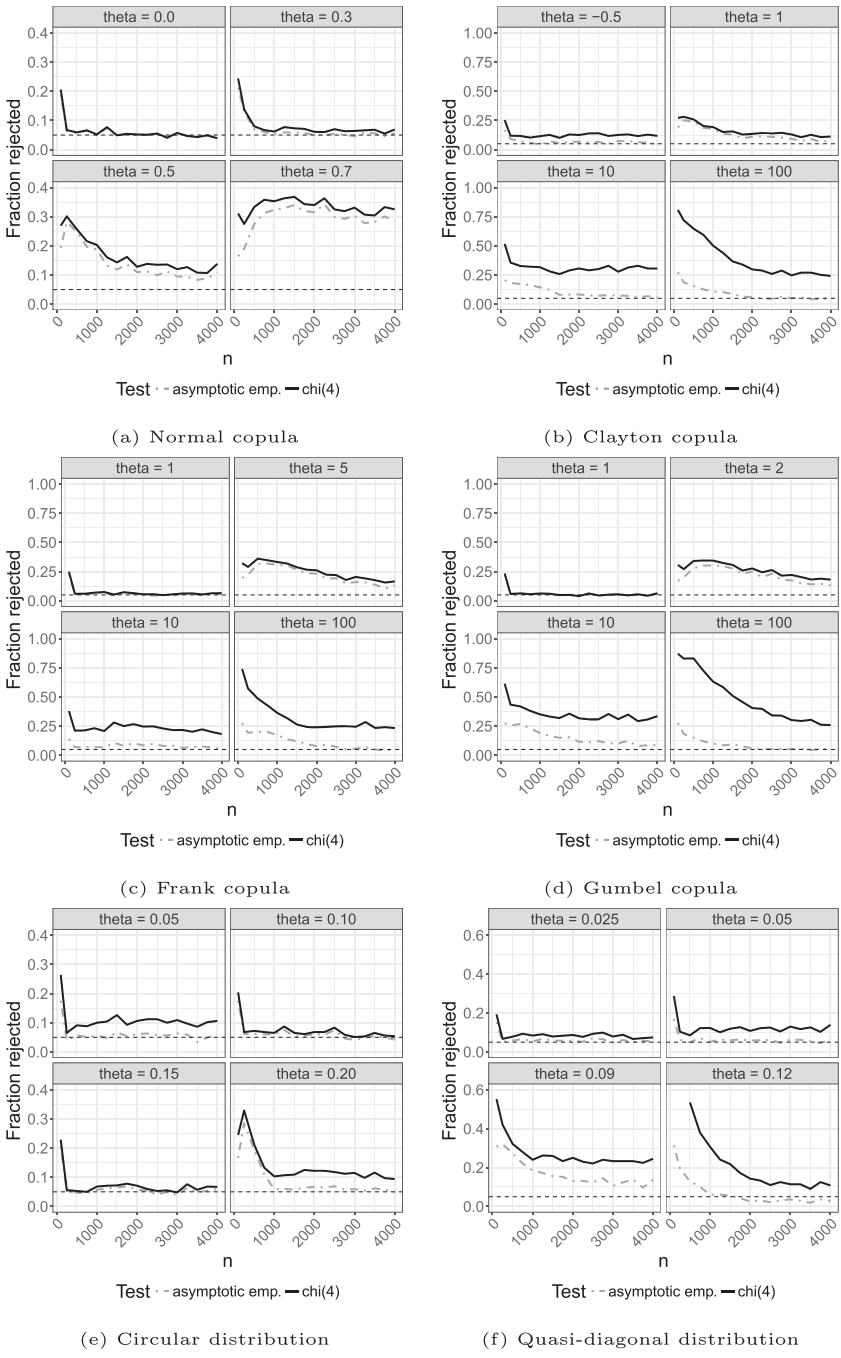

Fig. 4 Percent of rejections of $H_0 : II = 0$, when (X_1, X_2) and Y are independent for different copulas

Table 4 Computation times (in milliseconds) of interaction tests for (X_1, X_2) corresponding to normal copula with $\rho = 0.5$

Method\n	100	500	1000	5000	10000	100000
$W(\hat{\tilde{\lambda}}_1, \hat{\tilde{\lambda}}_2)$	90.76	90.81	91.16	93.46	96.2	147.62
χ^2_4	2.73	2.97	3.25	5.48	8.24	59.83
$Z(\hat{\Lambda})$	3.15	3.39	3.68	5.9	8.63	59.03

Entries are medians of times based on 100 repetitions

We note that computational cost of calculating $\widehat{II}$ is $O(n_y + n) = O(n)$, where $O(n)$ stems from calculating $\hat{p}_{ijk}$. Calculation of quantiles of $W(\hat{\tilde{\lambda}}_1, \hat{\tilde{\lambda}}_2)$ takes time $O(1)$ which does not depend on N. Table 4 shows computation times for testing H_0 for normal copula with $\rho = 0.5$. Note that although calculation of $W(\hat{\tilde{\lambda}}_1, \hat{\tilde{\lambda}}_2)$ takes from 35 to 3 times longer then that of χ^2_4 test depending on a sample size, the times for χ^2_4 and $Z(\hat{\Lambda})$ which is approximation $W(\hat{\tilde{\lambda}}_1, \hat{\tilde{\lambda}}_2$ of are comparable. We also note that the ratio of computing times for the proposed test and χ^2_4 diminishes with the sample size and is less than 3 for $n = 10^5$. Our experiments also indicate that the times depend only insignificantly on dependence of two predictors.

4.4 Analysis of a Real-World Data Set

We perform an analysis of a real data set on pancreatic cancer considered in Tan et al. (2008) downloaded from the addrees (SNPsyn 2011). The data consist of 208 observations (121 cases ($Y = 1$) and 87 controls ($Y = 0$)) with values of 901 SNPs. We chose predictors with at least two values (there are 499 such predictors in the data set) and consider all $K = 499 \cdot 498/2 = 124251$ pairs. We have applied all three discussed two-sided tests for all pairs and $\alpha = 0.05$ with Bonferroni correction resulting in an individual level of significance $0.05/K = 4.03 \times 10^{-7}$. It turns out that $W(\hat{\tilde{\lambda}}_1, \hat{\tilde{\lambda}}_2)$ test detects 77 significant interactions, whereas chi square test detected 24283 and $Z(\hat{\Lambda})$ test 21289 interactions.

Table 5 Top ten pairs discovered by $W(\hat{\tilde{\lambda}}_1, \hat{\tilde{\lambda}}_2)$ test for pancreatic cancer data set

X_1	X_2	$\widehat{II}$	Adj. p-value $\times 10^4$
rs3217922	rs3771527	–0.0140	0.146
rs1131854	rs7374	0.1215	0.694
rs1061282	rs3771527	–0.0100	0.844
rs1045485	rs3128	–0.0148	1.653
rs3128	rs3217922	–0.0086	1.917
rs1045485	rs2429467	–0.0127	2.660
rs14804	rs7201	0.1135	3.502
rs1045485	rs3773606	–0.0116	5.253
rs2429467	rs3217922	–0.0062	6.765
rs1058213	rs6115	0.1089	8.569

The last column gives adjusted p-value times 10^4

Much larger number of the detected interactions in the last two cases stems from the fact that majority of the interactions discovered is negative and their values due to positive support of chi square were considered significant in the case of χ^2 square test. The similar phenomenon occurs for $Z(\hat{\Lambda})$. Thus majority of negative interactions discovered by those two tests is likely to be spurious. This suggests that using $W(\hat{\bar{\lambda}}_1, \hat{\bar{\lambda}}_2)$ for negative interactions will have much smaller false discovery rate. Table 5 shows 10 of the most significant pairs with 7 of them being negative. Note that the most significant pair has negative interaction. Three the highest ranked pairs with positive interactions occupy three first places in ranking with respect to p-values when one sided alternative $II > 0$ is considered.

5 Conclusions

We have derived asymptotic distributions of interaction information for general trivariate nominal distribution (X_1, X_2, Y) as shown that it is weighted chi square distribution and have determined it weights in the case when (X_1, X_2) are independent of binary Y and both X_1 and X_2 have at most three values. We have shown numerically that using quantiles of asymptotic distribution W with estimated parameters yields actual significance level consistently much closer to nominal ones than in the case when quantiles of χ^2_4 distribution are used. This is especially pronounced in the case of two-sided alternative due to significant difference between left tails of χ^2_4 and $W(\bar{\lambda}_1, \bar{\lambda}_2)$. This can lead to much larger fraction of false rejections than expected when χ^2_4 based test is used. Putting it differently, one may detect many spurious interactions which do not actually exist. On the negative side, convergence of $2n\widehat{II}$ to $W(\bar{\lambda}_1, \bar{\lambda}_2)$ can be slow even for large n. We observed such behaviour mainly in situations when dependence between X_1 and X_2 becomes strong. Thus we recommend using the proposed test with critical value base on quantiles of W with estimated parameters when the dependence of both predictors, measured e.g. by their mutual information is not too strong. On the computational side, despite more complicated form of asymptotic distribution then in the case of overall independence construction of a critical region does not pose any significant hurdles.

Acknowledgements We are grateful to Łukasz Smaga for pointing out (Zhang 2005) to us.

Appendix

Below we prove Lemma 7.

Proof From the previous lemma we know that $\lambda_1(\mathbf{D}) = \ldots = \lambda_4(\mathbf{D}) = 1$, $\lambda_9(\mathbf{D}) = 0$.

By lengthy calculations we obtain a sequence of the following equalities:

$$\mathbf{D}^2 = \mathbf{D} + \mathbf{Z},$$

$$\mathbf{D}^3 = \mathbf{D} + \mathbf{Z} + \mathbf{ZD},$$

$$\mathbf{D}^4 = \mathbf{D} + \mathbf{Z} + 2\mathbf{ZD} + \mathbf{Z}^2,$$

where $\mathbf{Z} = (Z_{ij}^{i'j'})$ and:

$$Z_{ij}^{i'j'} = p_{i'j'}\Big(\frac{p_{i'j}}{p_{i'}p_j} + \frac{p_{ij'}}{p_i p_{j'}} - 2\Big),$$

$$\mathrm{tr}(\mathbf{Z}) = 2\sum_{i,j}\frac{p_{ij}^2}{p_i p_j} - 2 = 2H_1,$$

$$\mathrm{tr}(\mathbf{ZD}) = -\mathrm{tr}(\mathbf{Z}),$$

$$\mathrm{tr}(\mathbf{Z}^2) = -4 + 2\sum_{i,j,i',j'}\frac{p_{ij}p_{i'j}p_{i'j'}p_{ij'}}{p_i p_j p_{i'} p_{j'}} + 2\sum_{i,j}\frac{p_{ij}^2}{p_i p_j} = 2H_1 + 2H_2,$$

$$\mathrm{tr}(\mathbf{D}) = (n_{X_1} - 1)(n_{X_2} - 1) = 4,$$

$$\mathrm{tr}(\mathbf{D}^2) = \mathrm{tr}(\mathbf{D}) + \mathrm{tr}(\mathbf{Z}) = (n_{X_1} - 1)(n_{X_2} - 1) + 2\sum_{i,j}\frac{p_{ij}^2}{p_i p_j} - 2 = 2 + 2\sum_{i,j}\frac{p_{ij}^2}{p_i p_j} = 4 + 2H_1,$$

$$\mathrm{tr}(\mathbf{D}^3) = \mathrm{tr}(\mathbf{D}) = (n_{X_1} - 1)(n_{X_2} - 1) = 4,$$

$$\mathrm{tr}(\mathbf{D}^4) = (n_{X_1} - 1)(n_{X_2} - 1) - 2 + 2\sum_{i,j,i',j'}\frac{p_{ij}p_{i'j}p_{i'j'}p_{ij'}}{p_i p_j p_{i'} p_{j'}} = 2 + 2\sum_{i,j,i',j'}\frac{p_{ij}p_{i'j}p_{i'j'}p_{ij'}}{p_i p_j p_{i'} p_{j'}} = 4 + 2H_2.$$

Note that e.g. $\mathrm{tr}(\mathbf{D}) = (n_{X_1} - 1)(n_{X_2} - 1)$ follows from Lemmas 3-5 as it follows from Lemma 4 that $tr(\mathbf{M}) = -tr(\mathbf{C}) \cdot tr(\mathbf{D})$ and $tr(\mathbf{M}) = (n_{X_1} - 1)(n_{X_2} - 1)(n_Y - 1)$ and $tr(\mathbf{D}) = 1 - n_Y$. As trace of square matrix is sum of its eigenvalues from the above equalities it follows that:

$$\begin{cases} \lambda_5(\mathbf{D}) + \lambda_6(\mathbf{D}) + \lambda_7(\mathbf{D}) + \lambda_8(\mathbf{D}) = 0, \\ \lambda_5^2(\mathbf{D}) + \lambda_6^2(\mathbf{D}) + \lambda_7^2(\mathbf{D}) + \lambda_8^2(\mathbf{D}) = 2H_1, \\ \lambda_5^3(\mathbf{D}) + \lambda_6^3(\mathbf{D}) + \lambda_7^3(\mathbf{D}) + \lambda_8^3(\mathbf{D}) = 0, \\ \lambda_5^4(\mathbf{D}) + \lambda_6^4(\mathbf{D}) + \lambda_7^4(\mathbf{D}) + \lambda_8^4(\mathbf{D}) = 2H_2. \end{cases}$$

From the Newton-Girard identities we obtain:

$$\sum_{5 \le i_1 < i_2 \le 8} \lambda_{i_1}(\mathbf{D})\lambda_{i_2}(\mathbf{D}) = -H_1,$$

$$\sum_{5 \le i_1 < i_2 < i_3 \le 8} \lambda_{i_1}(\mathbf{D})\lambda_{i_2}(\mathbf{D})\lambda_{i_3}(\mathbf{D}) = 0,$$

$$\lambda_5(\mathbf{D})\lambda_6(\mathbf{D})\lambda_7(\mathbf{D})\lambda_8(\mathbf{D}) = \frac{1}{2}H_1^2 - \frac{1}{2}H_2.$$

This means that $\lambda_5(\mathbf{D})$, $\lambda_6(\mathbf{D})$, $\lambda_7(\mathbf{D})$, $\lambda_8(\mathbf{D})$ are the roots of the polynomial:

$$Q(x) = x^4 - H_1 x^2 + \frac{1}{2}H_1^2 - \frac{1}{2}H_2. \tag{21}$$

Note that in view of its definition $H_1 \ge 0$ and if X_1 and X_2 are independent then $H_1 = H_2 = \Delta = 0$. From this the lemma follows. □

References

Agresti A (2003) Categorical data analysis. Wiley, New York
Brown G, Pocock A, Zhao MJ, Luján M (2012) Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. J Mach Learn Res 13:27–66
Chanda P et al (2008) Ambience: a novel approach and efficient algorithm for identifying informative genetic and environmental associations with complex phenotypes. Genetics 180:1191–1210
Cordell HJ (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Mol Genet 11(20):2463–2468
Cordell HJ (2009) Detecting gene-gene interactions that underlie human diseases. Nat Rev Gen 10(20):392–404
Darroch J (1974) Multiplicative and additive interaction in contingency tables. Biometrika 9:207–214
Duggal P, Gillanders E, Holmes T, Bailey-Wilson J (2008) Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. BMC Genomics 9:516–613
Fano F (1961) Transmission of information: statistical theory of communication. MIT Press, Cambridge
Han TS (1980) Multiple mutual informations and multiple interactions in frequency data. Inf Control 46(1):26–45
Lin D, Tang X (2006) Conditional infomax learning: an integrated framework for feature extraction and fusion. European Conference on Computer Vision
Matsuda H (2000) Physical nature of higher-order mutual information: intrinsic correlations and frustration. Phys Rev E - Stat Phys Plasmas Fluids Related Interdiscip Topics 62(3 A):3096–3102
McGill WJ (1954) Multivariate information transmission. Psychometrika 19(2):97–116
Meyer P, Schretter C, Bontempi G (2008) Information-theoretic feature selection in microarray data using variable complementarity. IEEE J Selected Topics in Signal Process 2:261–274
Mielniczuk J, Rdzanowski M (2017) Use of information measures and their approximations to detect predictive gene-gene interaction. Entropy 19:1–23
Mielniczuk J, Teisseyre P (2018) A deeper look at two concepts of measuring gene-gene interactions: logistic regression and interaction information revisited. Genet Epidemiol 42(2):187–200
Moore JH, Gilbert JC, Tsai CT, Chiang FT, Holden T, Barney N, White BC (2006) A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. J Theor Biol 241(2):256–261
Nelsen R (2006) An introduction to copulas, 2nd edn. Springer, London
Schott J (1997) Matrix analysis for statistics wiley series in probabiliy and statistics. Wiley, New York
SNPsyn (2011) Data set GSE8054 http://snpsyn.biolab.si/examples/gse8054.tab.gz, (date of access: August 29, 2019)
Sucheston L, Chanda P, Zhang A, Tritchler D, Ramanathan M (2010) Comparison of information-theoretic to statistical methods for gene-gene interactions in the presence of genetic heterogeneity. BMC Genom 11:1–12
Tan A, Fan J, Karikari C et al (2008) Allele-specific expression in the germline of patients with familial pancreatic cancer: an unbiased approach to cancer gene discovery. Cancer Biol Ther 7:135–144
Tsamardinos I, Borboudakis G (2010) Permutation testing improves on Bayesian network learning. In: Proceedings of ECML PKDD 2010, pp 322–337
Wan X, Yang C, Yang Q, Xue T, Fan X, Tang N, Yu W (2010) Boost: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. Amer J Human Genetics 87(3):325–340
Zhang JT (2005) Approximate and asymptotic distributions of chi-squared type mixtures with applications. J Am Stat Assoc 100(469):273–285