

Metadata of the chapter that will be visualized in SpringerLink




Book Title	Modeling Decisions for Artificial Intelligence	
Series Title		
Chapter Title	Multiple Testing of Conditional Independence Hypotheses Using Information-Theoretic Approach	
Copyright Year	2021	
Copyright HolderName	Springer Nature Switzerland AG	
Author	Family Name	Lazęcka
	Particle	
	Given Name	Malgorzata
	Prefix	
	Suffix	
	Role	
	Division	Institute of Computer Science
	Organization	Polish Academy of Sciences
	Address	Jana Kazimierza 5, 01-248, Warsaw, Poland
	Division	Faculty of Mathematics and Information Science
	Organization	Warsaw University of Technology
	Address	Koszykowa 75, 00-662, Warsaw, Poland
	Email	malgorzata.lazeczka@ipipan.waw.pl
	ORCID	http://orcid.org/0000-0003-0975-4274
Corresponding Author	Family Name	Mielniczuk
	Particle	
	Given Name	Jan
	Prefix	
	Suffix	
	Role	
	Division	Institute of Computer Science
	Organization	Polish Academy of Sciences
	Address	Jana Kazimierza 5, 01-248, Warsaw, Poland
	Division	Faculty of Mathematics and Information Science
	Organization	Warsaw University of Technology
	Address	Koszykowa 75, 00-662, Warsaw, Poland
	Email	jan.mielniczuk@ipipan.waw.pl
	ORCID	http://orcid.org/0000-0003-2621-2303
Abstract	<p>In the paper we study the multiple testing problem for which individual hypotheses of interest correspond to conditional independence of the two variables X and Y given each of the several conditioning variables. Approaches to such problems avoiding inflation of probability of spurious rejections are widely studied and applied. Here we introduce a direct approach based on Joint Mutual Information (JMI) statistics which restates the problem as a problem of testing of a single hypothesis. The distribution of the test statistics JMI is established and shown to be well numerically approximated for a single data sample. The corresponding test is studied on artificial data sets and is shown to work promisingly when compared to general purpose multiple testing methods such as Bonferroni or Simes procedures.</p>	

Keywords
(separated by '-')

Conditional independence - Joint mutual information - Multiple testing - Weighted chi square distribution -
Dichotomous behaviour - Markov blanket - Dependence analysis



Multiple Testing of Conditional Independence Hypotheses Using Information-Theoretic Approach

Małgorzata Łazęcka^{1,2}  and Jan Mielniczuk^{1,2}  

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

{malgorzata.lazecka,jan.mielniczuk}@ipipan.waw.pl

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

Abstract. In the paper we study the multiple testing problem for which individual hypotheses of interest correspond to conditional independence of the two variables X and Y given each of the several conditioning variables. Approaches to such problems avoiding inflation of probability of spurious rejections are widely studied and applied. Here we introduce a direct approach based on Joint Mutual Information (JMI) statistics which restates the problem as a problem of testing of a single hypothesis. The distribution of the test statistics JMI is established and shown to be well numerically approximated for a single data sample. The corresponding test is studied on artificial data sets and is shown to work promisingly when compared to general purpose multiple testing methods such as Bonferroni or Simes procedures.

[AQ1](#)

Keywords: Conditional independence · Joint mutual information · Multiple testing · Weighted chi square distribution · Dichotomous behaviour · Markov blanket · Dependence analysis

1 Introduction

We focus here on multiple testing problem consisting in testing of conditional independence of two random variables given the third one, the later belonging to a group of variables of interest. The applications in this context are wide ranging. In studying human diseases one might be interested in checking whether occurrences of two diseases are independent given a third disease, where the later belongs to the group of diseases of interest possibly interacting with the first two. The same question may be asked when conditioning variables are characteristics of a patient such as age, gender or results of medical tests. Formally, the problem can be stated as testing p individual hypotheses

$$H_{0,i} : X \text{ and } Y \text{ are conditionally independent given } Z_i, \quad (1)$$

where $X \in \mathcal{X}$, $Y \in \mathcal{Y}$ and $Z_i \in \mathcal{Z}_i$ are some observed discrete random variables for $i = 1, \dots, p$. We want to construct a test which controls type I error under so called global null $H_0 = \cap_{i=1}^p H_{0,i}$ when all null hypotheses are true; i.e. we stipulate that $P(V \geq 1, H_{0,i} \text{ true}, i = 1, \dots, p)$ is smaller than the fixed level of significance α when V is the number of rejected hypotheses. Note also that simultaneous testing of (1) may be used as a proxy for testing the hypothesis $\tilde{H}_0 : X \perp Y | Z_1, Z_2, \dots, Z_p$ i.e. conditional independence of X and Y given all Z_1, \dots, Z_p . This is beneficial in the cases when the number of observations per one cell of $Z_1 = z_1, \dots, Z_p = z_p$ is small and the conditional independence tests loose power due to the curse of dimensionality. E.g. it is usually advised to have 5 observations per cell while using conditional chi-square test which results in number of observations at least 5×2^p on average when all variables are binary, whereas the use of the proposed test will require much less observations as the conditioning is done given individual variables. As a toy example of such situation consider binary random variables Z_0, Z_1, \dots, Z_{p+1} , where $Z_0 = Y$ and $Z_{p+1} = X$ which form a Markov chain $Z_0 \rightarrow Z_1 \dots \rightarrow Z_{p+1}$ such that $P(Z_{i+1} = 1 | Z_i = k)$ is q or $1 - q$ depending on whether $k = 1$ or $k = 0$. Then X and Y are conditionally independent given any individual Z_i , $i = 1, \dots, p$ but they are dependent.

Note that the problem of testing H_0 is a special case of the multiple testing problem, which due to its importance is analysed intensively in machine learning and statistics [1, 2, 7, 13]. There are several off-the-shelf generic methods of testing multiple hypotheses $H_{0,i}$ such as Bonferroni correction or Simes method described below which are known to perform well when test statistics for individual tests are mutually independent. This in case of testing (1) is hardly realistic and would require having independent samples for testing the individual hypothesis (see [11]). In general such methods may perform rather poorly at detecting violations of H_0 when no strong signal is available for any i resulting in low rejection rate in such situation. Thus true weak associations may be overlooked. In the special case of testing (1) for all i we show that it is possible to design a special purpose test statistic which would control type I error rate and have high true rejection rate when moderate and weak signals occur.

The paper is structured as follows: we introduce some information-theoretic concepts and define Joint Mutual Information (*JMI*) statistic designed for testing H_0 . In Sect. 3 we establish asymptotic distribution of sample *JMI* which leads to a novel test of H_0 (Sect. 4). In Sect. 5 the behaviour of the test procedure is investigated using synthetic and real data sets. The main contribution is to show that introduced *JMI*-based test of simultaneous conditional independence usually works better than the generic tests.

2 Preliminaries

2.1 Conditional Mutual Information

We introduce some information theoretic concepts leading to the conditional mutual information definition for discrete random variables.

We denote by $p(x) := P(X = x)$, $x \in \mathcal{X}$ a probability mass function corresponding to X , where \mathcal{X} is a domain of X and $|\mathcal{X}|$ is its cardinality. Joint probability will be denoted by $p(x, y) = P(X = x, Y = y)$ and $p(x, y|z)$ is $P(X = x, Y = y|Z = z)$. The sample estimate of $p(x)$ is denoted by $\hat{p}(x)$.

The mutual information (MI) between X and Y is

$$I(X, Y) = H(X) - H(X|Y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}, \quad (2)$$

where $H(X)$ and $H(X|Y)$ are the entropy and the conditional entropy, respectively [5]. This can be interpreted as the amount of uncertainty in X which is removed when Y is known, which is consistent with an intuitive meaning of mutual information as the amount of information that one variable provides about another. MI is non-negative and equals 0 if and only if X and Y are independent. We can extend the definition of $I(X, Y)$ to the conditional mutual information $I(X, Y|Z = z)$ of X and Y given $Z = z$ by replacing unconditional probabilities appearing in (2) by their conditional counterparts given $Z = z$. Then averaging $I(X, Y|Z = z)$ wrt distribution of Z yields the conditional mutual information (CMI)

$$I(X, Y|Z) = E_{Z=z}(I(X, Y|Z = z)) = \sum_{x,y,z} p(x, y, z) \log \left(\frac{p(x, y|z)}{p(x|z)p(y|z)} \right). \quad (3)$$

It follows from $MI = 0$ being equivalent to independence that $I(X, Y|Z) = 0$ if and only if X and Y are conditionally independent given Z which will be denoted by $X \perp Y|Z$ in the following. The construction of the test statistic JMI below relies on this fundamental fact. Moreover, the following chain rule holds:

$$I((X, Z), Y) = I(X, Y) + I(X, Y|Z). \quad (4)$$

For more properties of the basic measures described above we refer to [5].

2.2 Multiple Conditional Independence Testing and JMI Statistic

Intuitively, specially designed statistic should measure the cumulative effect of violating several null hypotheses $H_{0,i}$. In accordance with this heuristics we define

$$JMI = \frac{1}{p} \sum_{i=1}^p I(X, Y|Z_i). \quad (5)$$

Note that as the summands in (5) are non-negative, JMI averages violation effects of $H_{0,i}$. Note that for $p = 1$ JMI reduces to CMI . JMI has been introduced in [15] in the context of feature selection when Y is a target variable to be explained by a subset of potentially useful predictors, $(Z_i)_{i=1}^p$ are predictors already chosen and X is a potential candidate. It is also shown to be an approximation of $I(X, Y|Z_1, \dots, Z_p)$ under certain dependence conditions imposed on (X, Y, Z_1, \dots, Z_p) [14]. We stress however, that testing H_0

is not equivalent to testing \widehat{H}_0 of conditional independence of X and Y given (Z_1, \dots, Z_p) although for many dependence structures the former implies the latter (see e.g. [4], Sect. 13.6). We also note that testing H_0 requires less data than testing \widehat{H}_0 as the number of elements satisfying $Z_i = z_i$ for fixed z_i is usually larger than satisfying $Z_1 = z_1, Z_2 = z_2, \dots, Z_p = z_p$. The following lemma states some properties of JMI (with the proof confined to the online supplement¹). Define χ^2 measure of conditional dependence between X and Y given Z_i as

$$\chi_i^2 = \sum_{x,y,z_i} \frac{(p(x,y,z_i) - p(x|z_i)p(y|z_i)p(z_i))^2}{p(x|z_i)p(y|z_i)p(z_i)}.$$

Lemma 1.

- (i) $JMI = 0 \iff H_0 = \cap_i H_{0i}$ holds,
- (ii) $JMI = I(X, Y) + \frac{1}{p} \sum_{i=1}^p (I(X, Z_i|Y) - I(X, Z_i)),$
- (iii) $\frac{1}{2} \sum_{i=1}^p \left(\sum_{x,y,z_i} |p(x,y,z_i) - p(x|z_i)p(y|z_i)p(z_i)| \right)^2 \leq p \times JMI \leq \sum_{i=1}^p \log(\chi_i^2 + 1)$

and both inequalities are tight when H_0 holds.

Observe that statistics defined as $\sum_{i=1}^p \chi_i^2$ also enjoys analogous property to (i).

Let us mention that JMI statistic is frequently used in feature selection and Markov blanket discovery (see e.g. [3]) in order to test conditional independence of the response and the candidate predictor given the already chosen predictors. Here our aim is different as we want to test multiple individual conditional independence hypotheses. Given a sample $(X_i, Y_i, Z_i), i = 1, \dots, n$ of independent observations sampled from distribution $P_{X,Y,Z}$ we denote by \widehat{JMI} plug-in counterpart of JMI defined above obtained by replacing $I(X, Y|Z_i)$ by their empirical versions $\hat{I}(X, Y|Z_i)$. For $p = 1$ \widehat{JMI} reduces to the empirical CMI . In this case, provided conditional independence of X and Y given Z holds, it is asymptotically chi square distributed with $(|\mathcal{X}| - 1)(|\mathcal{Y}| - 1)(|\mathcal{Z}|)$ degrees of freedom (see e.g. [10]). We will derive the distribution of \widehat{JMI} in the next section: note that it does not follow in straightforward manner from the latter result as the summands $\hat{I}(X, Y|Z_i)$ of \widehat{JMI} are dependent.

3 Main Result: Dichotomous Behaviour of Test Statistic Statistic \widehat{JMI}

In the following we explicitly state the asymptotic distribution of \widehat{JMI} when H_0 holds. The general formula for distribution of \widehat{JMI} has been already stated in [9]. We derive below its explicit form which is amenable to computations for

¹ github.com/lazeckam/JMI_GlobalNull.

moderate p and derive some of its properties. Moreover, we indicate that when H_0 fails the behaviour of \widehat{JMI} and its distribution is fundamentally different from that under H_0 suggesting that the resulting test should have a reasonable power.

Let $K = |\mathcal{X}| \times |\mathcal{Y}| \times \prod_{i=1}^p |\mathcal{Z}_i|$ be the number of levels of random variable (X, Y, Z_1, \dots, Z_p) and $z = (z_1, \dots, z_p)$. Let $A_{x,y,z}^{x',y',z'}$ denote the element of $K \times K$ matrix A with the row index x, y, z and the column index x', y', z' . Finally, $H_{0,i}^c$ is the opposite hypothesis to $H_{0,i}$. $\mathbb{I}(A)$ denotes the indicator function of set A :

Theorem 1. (i) Assume that the global null H_0 holds. Then

$$2n\widehat{JMI} \xrightarrow{d} \sum_{i=1}^K \lambda_i(M) Z_i^2, \quad (6)$$

where Z_i are independent $N(0, 1)$ random variables and $\lambda_i(M), i = 1, \dots, K$ are eigenvalues of matrix M with the elements

$$M_{x,y,z}^{x',y',z'} = \frac{1}{p} p(x', y', z') \sum_{i=1}^p \left[\frac{\mathbb{I}(z_i = z'_i)}{p(z_i)} - \frac{\mathbb{I}(x = x', z_i = z'_i)}{p(x, z_i)} - \frac{\mathbb{I}(y = y', z_i = z'_i)}{p(y, z_i)} + \frac{\mathbb{I}(x = x', y = y', z_i = z'_i)}{p(x, y, z_i)} \right], \quad (7)$$

where $z = (z_1, \dots, z_p)$ and $z' = (z'_1, \dots, z'_p)$. Moreover, the trace of M equals $p^{-1}(|\mathcal{X}| - 1)(|\mathcal{Y}| - 1) \sum_i |\mathcal{Z}_i|$.

(ii) Assume that the alternative $H_1 = \cup_{i=1}^p H_{0,i}^c$ to the global null is valid and Y is binary. Then

$$\sigma_{\widehat{JMI}}^2 = \text{Var} \left(\frac{1}{p} \log \prod_{i=1}^p \frac{p(X, Y, Z_i) p(Z_i)}{p(X, Z_i) p(Y, Z_i)} \right) > 0$$

and

$$n^{1/2}(\widehat{JMI} - JMI) \xrightarrow{d} N(0, \sigma_{\widehat{JMI}}^2). \quad (8)$$

The result above states an exact dichotomy of asymptotic behaviour which makes the construction of the test possible: the asymptotic distribution of \widehat{JMI} is either that of quadratic form in normal variables as in (6) or normal (cf. (8)) depending on whether H_0 is satisfied or not.

Proof. (i) Let $f(p) = p^{-1} \sum_{i=1}^p p(x, y, z_i) \log(p(x, y, z_i) p(z_i) / p(x, z_i) p(y, z_i))$, where $p = p(x, y, z_1, \dots, z_p)$. Note that when H_0 holds then $\sigma_{\widehat{JMI}}^2 = 0$ and it follows from the delta method (cf. Corollary 1 in [9]) that the asymptotic distribution of $2n\widehat{JMI}$ is the distribution of $Z^T M Z$ where $Z \in R^p$ has $N(0, I)$ distribution, $M = H \Sigma$, $\Sigma_{x,y,z}^{x',y',z'} = p(x', y', z') (I(x = x', y = y', z = z') - p(x, y, z)) / n$ and $H = D^2 f(p)$ is the Hessian of $f(p)$. By direct calculation we have

$$Df(p)_{xyz} = \frac{1}{p} \sum_{i=1}^p \log \left(\frac{p(x, y, z_i) p(z_i)}{p(x, z_i) p(y, z_i)} \right),$$

$$H_{xy'z'}^{x'y'z'} = D^2 f(p)_{xy'z'}^{x'y'z'} = \frac{1}{p} \sum_{i=1}^p \left[\frac{\mathbb{I}(z_i = z'_i)}{p(z_i)} - \frac{\mathbb{I}(x = x', z_i = z'_i)}{p(x, z_i)} - \frac{\mathbb{I}(y = y', z_i = z'_i)}{p(y, z_i)} + \frac{\mathbb{I}(x = x', y = y', z_i = z'_i)}{p(x, y, z_i)} \right]$$

where $z = (z_1, \dots, z_p)$ and M is obtained by the direct multiplication of H and Σ resulting in (7). The trace of M equals $p^{-1} \sum_{x,y,z} \sum_{i=1}^p (p(x, y|z_i) - p(x|y, z_i) - p(y|x, z_i) + 1)$ which yields the result. (ii) is proved in Corollary 1 in [9].

4 Asymptotic Versus Generic Methods

4.1 Asymptotic Method

For a given sample chosen from P_{XYZ} we calculate \widehat{JMI} and plug-in estimator \widehat{M} of matrix M defined in Theorem 1. We use now the fact that the asymptotic distribution W of \widehat{JMI} under H_0 given in (6) is determined by the eigenvalues $\lambda_i(M)$ and we approximate it by \widehat{W} plugging in $\lambda_i(\widehat{M})$ for $\lambda_i(M)$, where $\lambda_i(\widehat{M})$ are numerically calculated. Then the rejection region for a given significance level α is given by $\{\widehat{JMI} \geq q_{\widehat{W}, 1-\alpha}\}$, where $q_{\widehat{W}, 1-\alpha}$ is quantile of the order $1 - \alpha$ of \widehat{W} . A function `eigen` from R package `base` has been used to calculate the eigenvalues and package `CompQuadForm` [6] for quantiles of \widehat{W} .

4.2 Generic Methods

We use two generic methods to cope with controlling type I error while performing multiple tests, namely Bonferroni correction and Simes method (see e.g. [12] and [7]).

- Bonferroni correction: individual tests are performed with level of significance α/p , where p is the number of tests performed thus bounding probability $P(V \geq 1, \forall_i H_{0i} \text{ true})$ by α . It is known to work well when the test statistics used to test individual hypotheses are independent, but in a general case is conservative leading to the low power when H_0 fails. Individual tests are \widehat{MI} -based tests based on chi square benchmark distribution described at the end of Sect. 2.2.
- Simes method: p-values of individual test p_1, \dots, p_p are calculated and ordered: $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(p)}$. H_0 is rejected when for certain $i \leq p$ we have $p_{(i)} \leq i\alpha/p$, or equivalently if $\min_i p_{(i)}/i \leq \alpha/p$. Individual tests considered are the same as for Bonferroni correction method.

5 Simulation Study

5.1 Artificial Data Sets

We discuss first the dependence structures which we use to generate data (see Fig. 1). Below $Z \sim \text{Bern}(p)$ stands for Z being distributed as the Bernoulli

distribution with probability of success p and Φ is the cumulative distribution function (CDF) of the standard normal distribution.

- **Model A.** Parameters: $\alpha_x \geq 0, \alpha_y \geq 0$
 $Z_i \sim \text{Bern}(0.5)$ for $i \in \{1, 2, \dots, p\}$, $\bar{Z} := \frac{1}{p} \sum_{i=1}^p Z_i$
 $X|\bar{Z} = z \sim \text{Bern}(1 - \Phi(\alpha_x(\frac{1}{2} - z)))$, $Y|\bar{Z} = z \sim \text{Bern}(1 - \Phi(\alpha_y(\frac{1}{2} - z)))$
Model A' is a modification of model **Model A** for which $\bar{Z} := \frac{1}{s} \sum_{i=1}^s Z_i$ and $s < p$. $Z_i \perp (X, Y)$ for $i \in \{s+1, s+2, \dots, p\}$.
- **Model B.** Parameters: $\alpha_x \in [0, 1], \alpha_z \geq 0$
 $X \sim \text{Bern}(0.5)$ and $Y \sim \text{Bern}(0.5)$,
 $Z_i | (\alpha_x X + (1 - \alpha_x)Y) = w \sim \text{Bern}(1 - \Phi(\alpha_z(\frac{1}{2} - w)))$ for $i \in \{1, 2, \dots, p\}$
Model B' is a modification of model **Model B** for which the dependence of Z_i on X and Y defined above holds only for $i \in \{1, 2, \dots, s\}$, for $i \in \{s+1, s+2, \dots, p\}$ $Z_i \perp (X, Y)$ and $Z_i \sim \text{Bern}(0.5)$
- **Models C.** Parameters: $q \in [0, 1], q_{XY} \in [0, 1]$
C(q): $Y \sim \text{Bern}(0.5)$, $Z_1|Y = y \sim \text{Bern}(q^y(1 - q)^{1-y})$, $Z_{i+1}|Z_i = z \sim \text{Bern}(q^z(1 - q)^{1-z})$ for $i \in \{1, 2, \dots, p\}$ and $Z_{p+1} = X$.
C(q, q_{XY}): while retaining the conditional distribution $P_{X|Z_p}$ as above, the distribution of (X, Y) is modified so that H_1 is satisfied:
 $P(X = z, Y = z|Z_p = z) = q - P(X = z, Y = 1 - z|Z_p = z) = q_{XY}q$,
 $P(X = 1 - z, Y = 1 - z|Z_p = z) = 1 - q - P(X = 1 - z, Y = z|Z_p = z) = q_{XY}(1 - q)$.

Model A corresponds to the situation when variables Z_1, \dots, Z_p influence X and Y simultaneously. Parameters α_x and α_y control how strong the dependence between the variables Z_i and X or Y is. If at least one of the parameters equals zero then X and Y are independent and conditionally independent given any Z_i , otherwise X and Y are (conditionally) dependent. In Model A' the role of parameter p is taken over by s and the additional variables $Z_i, i = s+1, \dots, p$ are independent of X and Y . In Model B the dependence structure is reversed and both variables X and Y influence variables Z_i . The parameter α_x measures the strength of influence of X compared to that of Y , whereas the parameter α_z controls the strength of the joint dependence of Y and X on Z_i . In the model X and Y are independent but they are conditionally dependent given Z_i unless $\alpha_x \in \{0, 1\}$ or $\alpha_z = 0$. Model B' is constructed analogously to A'. Model C(q) is a Markov chain for which due to Markov property X and Y are conditionally independent given any in-between variable Z_i . Here, q denotes the probability that the previous variable equals the next one. If $q = 0.5$, then any two adjacent variables are independent and if it increases (decreases) the variables become positively (negatively) dependent. By introduction of an additional parameter q_{XY} , we obtain model C(q, q_{XY}) for which H_0 is violated.

Our main aim is to study the actual type I error of the considered procedures (i.e. probability rejection when H_0 is true) and the power (probability of rejection when H_0 is false) for the assumed significance level α using the fractions of rejections for artificial data sampled from the above models. We also studied ROC-type curves for all three considered procedures. ROC-type curves are based

on two models: the one for which H_0 holds and the second for which H_1 is true, and the report *the actual* type I error and the power approximated by means of simulations for varying α . In this way y values of three ROC curves for the fixed x value correspond to the power for *the same* actual type I error (see Fig. 5).

We present results in Figs. 2, 3, 4, 5 and Table 1 (the chosen parameters represent various strengths of dependence for the structures considered, see discussion in the online supplement). Figure 2 shows the behaviour of the true asymptotic distribution of \widehat{JMI} (see (6)) and its estimate. The left panel depicts boxplots of sorted eigenvalues $\lambda_i(\widehat{M})$, the right compares averaged CDFs corresponding to $\lambda_i(\widehat{M})$ and 90% confidence bands for the true CDF based on them with the true asymptotic CDF and the empirical CDF based on \widehat{JMI} values. In Figs. 3 and 4 the behaviour of the power of the considered procedures is compared against one of the model's varying parameters when the remaining ones are held fixed and the significance level is set to $\alpha = 0.05$. Table 1 indicates how the power and the type I error for the considered procedures depend on the sample size n .

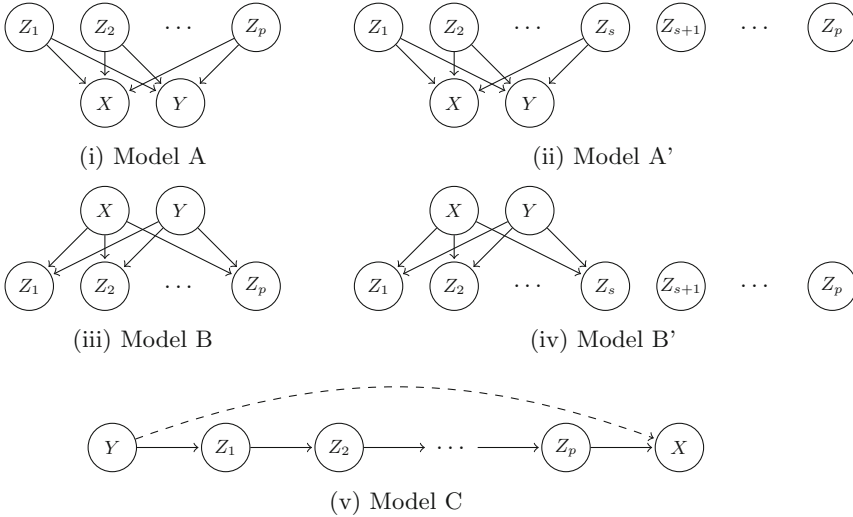


Fig. 1. Graphical representation of the dependence structures

Results. For the sample sizes $n = 500$ and larger the eigenvalues of the estimated matrix \widehat{M} approximate very closely the eigenvalues of the theoretical matrix M and therefore the plots of the averaged CDFs based on eigenvalues $\lambda_i(\widehat{M})$ and CDF using eigenvalues $\lambda_i(M)$ almost overlap (Fig. 2). Such sample sizes are sufficient to ensure the adequate approximation of the distribution of \widehat{JMI} by its asymptotic counterpart. It follows from Table 1 that starting from

Table 1. Estimated powers and type I errors based on $N = 5000$ simulations for varying n and the tests considered. Parameters in Models A, B and C are the same as in Fig. 5.

Mod.	Proc.	Estimated power						Estimated type I error					
		$n = 50$	100	250	500	1000	2000	$n = 50$	100	250	500	1000	2000
A	Bonf.	0.115	0.157	0.368	0.676	0.945	0.999	0.056	0.035	0.037	0.037	0.038	0.033
	Simes	0.129	0.185	0.419	0.732	0.964	1.000	0.063	0.043	0.043	0.043	0.042	0.038
	JMI	0.159	0.249	0.541	0.828	0.983	1.000	0.064	0.052	0.048	0.048	0.052	0.047
B	Bonf.	0.147	0.241	0.487	0.807	0.983	1.000	0.058	0.074	0.053	0.050	0.047	0.046
	Simes	0.162	0.259	0.524	0.837	0.987	1.000	0.062	0.078	0.056	0.052	0.048	0.048
	JMI	0.276	0.333	0.649	0.907	0.997	1.000	0.212	0.089	0.056	0.053	0.049	0.053
C	Bonf.	0.083	0.106	0.180	0.335	0.667	0.952	0.051	0.054	0.040	0.039	0.044	0.039
	Simes	0.096	0.116	0.200	0.363	0.696	0.957	0.060	0.058	0.045	0.042	0.048	0.042
	JMI	0.178	0.139	0.238	0.408	0.747	0.971	0.135	0.072	0.058	0.051	0.050	0.044

the moderate sample sizes ($n \geq 250$) *JMI* controls well type I error whereas Bonferroni and Simes methods are conservative in some cases (such as Model A for $n = 1000, 2000$). Moreover, it consistently yields the largest power among these three methods. For Fig. 3 H_1 holds and in models A, B (on-line supplement) and C *JMI*-based test on the whole works better than mutual information-based individual tests with correction applied. As expected, when there is only one strong signal i.e. null hypothesis $X \perp Y|Z_i$ is strongly violated for just one i (model B' with $s = 1$, middle panel of Fig. 4), Bonferroni correction and Simes procedure work well. The novel test does not detect the dependence as frequently as the other two. The situation changes, however, when number of hypotheses that should be rejected increases (see Fig. 4, panels 1 and 3). Comparison of the ROC curves in Fig. 5 indicates that even when the *actual* significance levels of the three tests are matched, *JMI*-based test remains the most powerful (H_1 hypotheses for the panels correspond to the first column of Fig. 4 for $p = 5$). This is also reflected in the largest values of Area Under Curve (AUC) for *JMI*.

5.2 Medical Data Set Example

We show an example of the application of the novel test and Bonferroni and Simes procedures to a real medical dataset MIMIC-III [8]. The dataset contains information about patients requiring intensive care and it includes, among others, 10 binary variables representing the presence or absence of the following diseases: **hypertension**, kidney failure (**kidney**), disorders of fluid electrolyte balance (**fluid**), **hypotension**, disorders of lipid metabolism (**lipoid**), liver disease (**liver**), **diabetes**, thyroid disease (**thyroid**), chronic obstructive pulmonary disease (**copd**) and **thrombosis**. We select two diseases, liver disease and thrombosis for which conditional mutual informations given any of the other eight diseases are approximately the same (see the first panel of Fig. 6) to analyse the situation for which all null hypotheses are rejected with approximately the same strength for

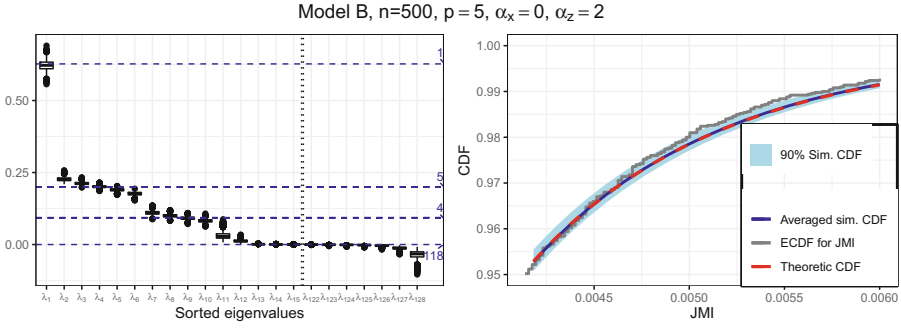


Fig. 2. Left: Box-plots of the empirical values $\lambda_i(\widehat{M}), i = 1, \dots, 128$ for Model B ($n = 500, p = 5, \alpha_x = 0.5, \alpha_z = 2$). Eigenvalues $\lambda_i(M)$ approximately equal to 0 (multiplicity 118), 0.093 (multiplicity 4), 0.2 (multiplicity 5) and 0.627 (multiplicity 1) are marked by the horizontal lines. Right: values of theoretical CDF, the empirical CDF of \widehat{JMI} and the average of CDFs corresponding to $\lambda_i(\widehat{M})$ for the values of JMI greater than 0.95th quantile of \widehat{JMI} .

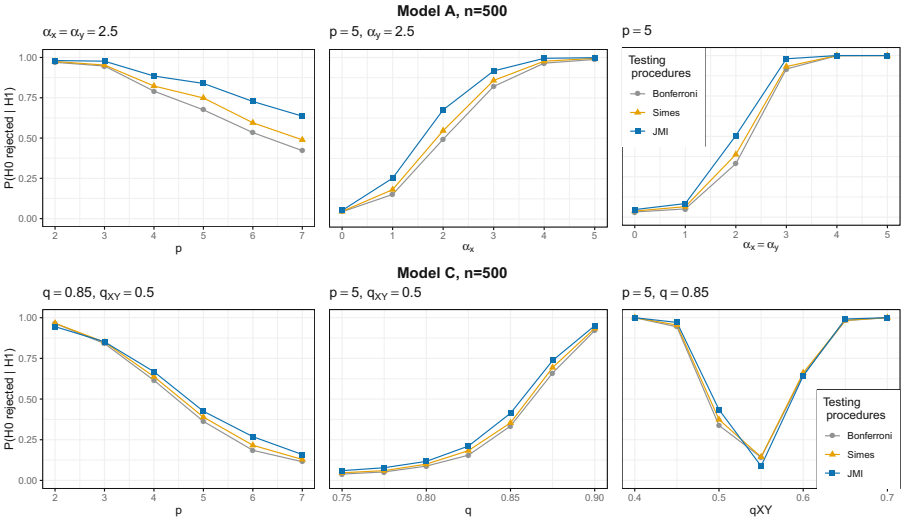


Fig. 3. Power against the changing number of variables and the parameter values for models A and C based on $N = 1000$ simulations.

the whole data set consisting of 10000 observations. In the second panel of Fig. 6 we present how often the null hypothesis that liver disease and thrombosis are conditionally independent is rejected for smaller sample size scenarios for which conditional dependencies are much harder to reject. The estimation is based on $N = 200$ samples randomly sub-sampled from the original data set for each n ranging from 250 to 5000. The asymptotic test works uniformly better than Bonferroni

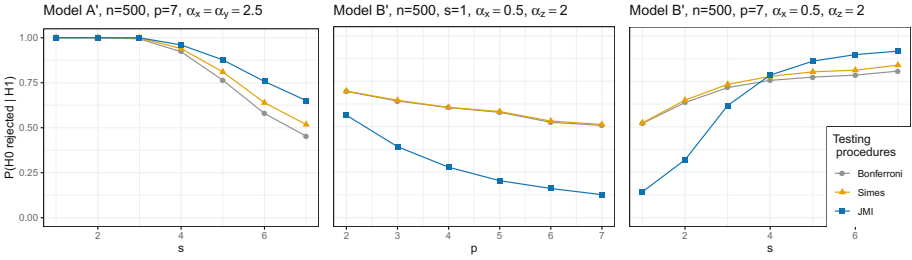


Fig. 4. Power against the number of variables and the number of significant variables in models A' and B' (see text) based on $N = 1000$ simulations.

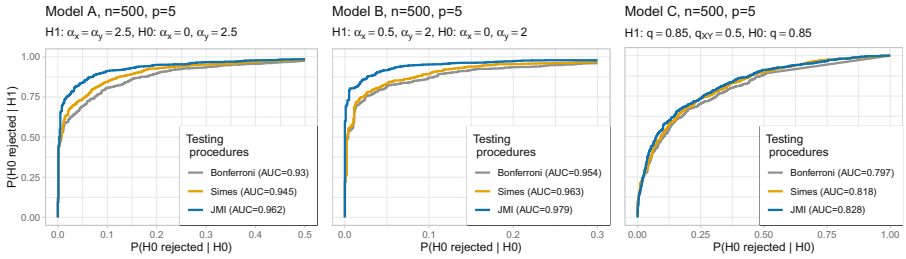


Fig. 5. ROC-type curves and the corresponding values of AUC for models A, B and C and for H_0 and H_1 indicated in the header and $n = 500$.

and Simes procedures. This holds even for small sample sizes for which approximation of the distribution of \widehat{JMI} by its limit is likely to be worse than for larger sample sizes.

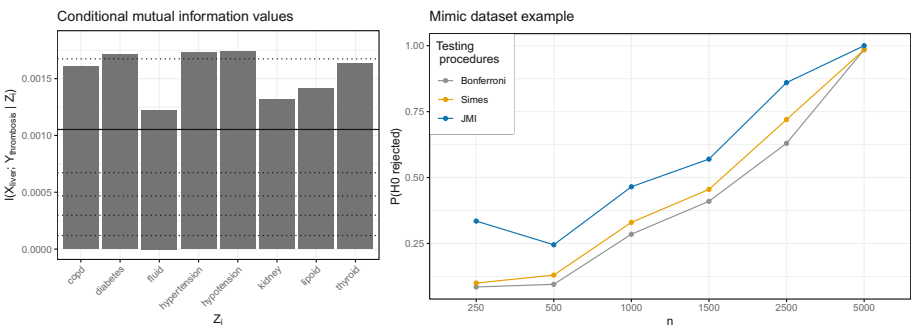


Fig. 6. Left: $\hat{I}(X, Y | Z_i)$, where X and Y denote liver and thrombosis and Z_i is one of the eight remaining variables. Right: the estimated probability of rejection.

5.3 Conclusion

In the paper we have constructed a test for multiple conditional independence which relies on approximating the asymptotic distribution of \widehat{JMI} . It follows from numerical experiments that \widehat{JMI} -based test is a promising alternative to procedures based on individual test which are modified for multiple testing, especially when one expects several weak violations of individual conditional independence hypotheses. The proposed test has consistently the largest power in such cases, while controlling for type I error. Its superiority is retained even when Bonferroni and Simes methods are calibrated to have exactly the same value of type I error as JMI -based test. The method is numerically stable and reasonably quick for $p \leq 8$, for larger p eigen function has to be modified.

References

1. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stat. Soc. B* **57**, 289–300 (1995)
2. Benjamini, Y., Yakutieli, D.: The control of false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**(4), 1165–1188 (2001)
3. Brown, G., Pocock, A., Zhao, M., Luján, M.: Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *J. Mach. Learn. Res.* **13**(1), 27–66 (2012)
4. Buhlmann, P., de Geer, S.: *Statistics for High-Dimensional Data*. Springer, New York (2006). <https://doi.org/10.1007/978-3-642-20192-9>
5. Cover, T.M., Thomas, J.A.: *Elements of Information Theory* (Wiley Series in Telecommunications and Signal Processing). Wiley-Interscience, New York (2006)
6. Duchesne, P., Lafaye de Micheaux, P.: Computing the distribution of quadratic forms: further comparisons between the Liu-Tang-Zhang approximation and exact methods. *Comput. Stat. Data Anal.* **54**, 858–862 (2010)
7. Dudoit, S., van der Laan, M.J.: *Multiple Testing Procedures with Applications to Genomics*. Springer, New York (2009). <https://doi.org/10.1007/978-0-387-49317-6>
8. Johnson, A.: MIMIC-III, a freely accessible critical care database. *Scientific Data* **3**, 1–9 (2016)
9. Kubkowski, M., Łazęcka, M., Mielniczuk, J.: Distributions of a general reduced-order dependence measure and conditional independence testing. In: Krzhizhanovskaya, V.V., et al. (eds.) *ICCS 2020*. LNCS, vol. 12143, pp. 692–706. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-50436-6_51
10. Kullback, S.: *Information Theory and Statistics*. Smith, P. (1978)
11. Moskvina, V., Schmidt, K.: On multiple-testing correction in genome-wide association studies. *Genet. Epidemiol.* **32**, 1567–573 (2008)
12. Simes, R.: An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751–754 (1986)
13. Storey, J.: A direct approach to false discovery rates. *J. Royal Stat. Soc. B* **64**(3), 479–498 (2002)
14. Vergara, J., Estevez, P.: A review of feature selection methods based on mutual information. *Neural Comput. Appl.* **24**(1), 175–186 (2014)
15. Yang, H., Moody, J.: Data visualization and feature selection: new algorithms for nongaussian data. *Adv. Neural. Inf. Process. Syst.* **12**, 687–693 (1999)

Author Queries

Chapter 7

Query Refs.	Details Required	Author's response
AQ1	This is to inform you that corresponding author has been identified as per the information available in the Copyright form.	