Detection of conditional dependence between multiple variables using multiinformation

Jan Mielniczuk[1,2][0000−0003−2621−2303] and Paweł Teisseyre[1,2][0000−0002−4296−9819]

[1] Institute of Computer Science, Polish Academy of Sciences, Poland
[2] Faculty of Mathematics and Information Sciences, Warsaw University of Technology, Poland
{Jan.Mielniczuk, Pawel.Teisseyre}@ipipan.waw.pl

Abstract. We consider a problem of detecting the conditional dependence between multiple discrete variables. This is a generalization of well-known and widely studied problem of testing the conditional independence between two variables given a third one. The issue is important in various applications. For example, in the context of supervised learning, such test can be used to verify model adequacy of the popular Naive Bayes classifier. In epidemiology, there is a need to verify whether the occurrences of multiple diseases are dependent. However, focusing solely on occurrences of diseases may be misleading, as one has to take into account the confounding variables (such as gender or age) and preferably consider the conditional dependencies between diseases given the confounding variables. To address the aforementioned problem, we propose to use conditional multiinformation (CMI), which is a measure derived from information theory. We prove some new properties of CMI. To account for the uncertainty associated with a given data sample, we propose a formal statistical test of conditional independence based on the empirical version of CMI. The main contribution of the work is determination of the asymptotic distribution of empirical CMI, which leads to construction of the asymptotic test for conditional independence. The asymptotic test is compared with the permutation test and the scaled chi squared test. Simulation experiments indicate that the asymptotic test achieves larger power than the competitive methods thus leading to more frequent detection of conditional dependencies when they occur. We apply the method to detect dependencies in medical data set MIMIC-III.

Keywords: detection of conditional dependence · conditional multiinformation · information theory · weighted chi squared distribution · Kullback-Leibler divergence;

1 Introduction

Detecting conditional dependence is a fundamental problem in machine learning and statistics [7]. It has significant applications in several problems such as causal

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_51

inference [12], learning structure of Bayesian Networks [15], [18] and feature selection [3]. Most of the research focuses on testing the conditional independence between two variables given a third one (which can be multi-dimensional), (see [11]). The existing methods are based on different approaches, such as kernel methods [22], information theoretic measures [14], permutation methods [19],[2], generalized adversarial networks [1] and knockoffs [4].

In this work we investigate the more general problem of testing the conditional independence of multiple variables, i.e. we consider the null hypothesis H_0 of the form

$$P(X_1 = x_1, \ldots, X_d = x_d | Y = y) = \prod_{j=1}^{d} P(X_j = x_j | Y = y), \tag{1}$$

where $X_1, \ldots, X_d, Y$ are discrete random variables. The above hypothesis reduces to a classical one for $d = 2$. Surprisingly, such generalization attracted much less attention despite its wide potential applicability. For example, in epidemiology there is often a need to test the independence of multiple diseases. However the task is challenging, as one should take into account the possible confounding variables, such as age, gender or race. The occurrences of diseases can be independent, but dependent when conditioning on a confounding variable. In this case, an important information will be missed when we focus on unconditional dependence. On the other hand, the diseases may be dependent but become independent when conditioning on a confounding variable. In the latter case, there is a risk of finding spurious dependences due to ignoring the latent conditioning variables. Therefore, one should rather focus on testing the conditional independence between diseases given confounding variable/variables. The problem of testing (1) appears naturally in the context of supervised learning and the Naive Bayes (NB) method which is one of the simplest and most popular classifiers. In NB method, it is assumed that all features are conditionally independent given a class variable. Using this assumption, it is possible to avoid the challenging estimation of the joint conditional probabilities and instead estimate the univariate conditional probabilities which is much easier. Usually, in practice, the NB method is used without verifying the assumption, which may lead to poor predictive performance of the classifier. Finally, testing (1) can be useful in multi-label classification where the goal is to predict binary output variables (labels) $Y_1, \ldots, Y_K$ using feature X. If the labels are conditionally independent given X, then classification models can be independently fitted for each label. Otherwise, it is necessary to use more complex approaches that take into account conditional dependencies among labels.
In this work we consider the case of discrete random variables. Then the problem of detecting conditional independence is equivalent to the problem of independence testing on each strata $Y = y$. However, performing the test for each strata separately will lead to multiple testing problem and lack of control of false discovery rate. Thus the need for a specialised test for (1).

To test the hypothesis (1) we propose to use conditional multiinformation (CMI), which is a measure derived from Information Theory. The conditional

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_51

multiinformation reduces to conditional mutual information for $d = 2$. Although the latter attracts a great interest, the properties of conditional multiinformation and its usefulness to detect departures from (1) remain mainly unexplored. The present paper aims to fill the gap. We prove some new interesting theoretical properties of CMI. In particular we provide upper and lower bounds for it which are tight when $CMI = 0$. Most importantly, with the aim of reducing uncertainty concerning significance of positive value of sample CMI, we determine its asymptotic distribution (for both cases of H_0 and its alternative), which in particular allows to construct the asymptotic test for hypothesis (1). For $d = 2$, our result reduces to well known result for conditional mutual information, which states that the asymptotic distribution is chi squared with the number of degrees of freedom depending on the number of possible values of X_1, X_2, Y. For $d > 2$, the asymptotic distribution under null hypothesis is weighted sum of squared independent normally distributed variables. When $X_1, \ldots, X_d$ are conditionally dependent given Y, the asymptotic distribution is normal.

We compare the proposed asymptotic test with the permutation test as well as test based on the scaled chi-squared distribution. The advantage of asymptotic test over the permutation test is that we avoid generating permuted samples which is the main obstacle in applying permutation method.

2 Preliminaries

2.1 Conditional Multiinformation

We define first the main object of our interest here, namely conditional multiinformation. Let $(X_1, \ldots, X_d, Y)$ be $(d+1)$-dimensional random variable such that any of its coordinates admits finite number of values and the corresponding mass function $P(X_1 = x_1, \ldots, X_d = x_d, Y = y)$ is denoted by $p(x_1, \ldots, x_d, y)$. Moreover, define $p(x_i) = P(X_i = x_i), p(y) = P(Y = y)$ and $p(x_i|y) = P(X_i = x_i|Y = y)$. Let $X = (X_1, \ldots, X_d)$ and consider the discrete distribution $P_{X,Y}^{ind}$ having mass function $p(y)p(x_1|y) \cdots p(x_d|y)$ which fulfills (1). It means that for any random variable having this distribution, X_is are conditionally independent given Y. The common approach to measure the strength of dependence for (X, Y) is to study a distance of its distribution from some distribution which satisfies the required type of independence and is moreover similar in a specified way to the distribution of (X, Y). Note that $P_{X,Y}^{ind}$ satisfies this requirement as it has the first d components conditionally independent given the last one and also has the same bivariate marginal distributions as $P_{X,Y}$ i.e. $P_{X_i,Y} = P_{X_i,Y}^{ind}$ for any $i = 1, \ldots, d$. Recall that for any two discrete distributions having the same support, Kullback-Leibler (K-L) divergence (relative entropy) is defined as

$$D_{KL}(P||Q) = \sum_i p(x_i) \log \big(p(x_i)/q(x_i)\big),$$

where $p(x_i)$ and $q(x_i)$ denote the corresponding probability mass functions ([6]). K-L divergence plays a role of pseudo-distance between two distributions and is

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_51

frequently used to describe their discrepancy.
We define now conditional multiinformation (aka conditional total correlation) as

$$CMI(X|Y) = CMI(X_1, \ldots, X_d|Y) = D_{KL}(P_{X,Y}||P^{ind}_{X,Y}) \tag{2}$$

The term 'conditional' in the name CMI is explained by an equivalent definition of this quantity. Namely, note that it follows from definition (2) that

$$\begin{aligned} CMI(X|Y) &= \sum_{x_1,\ldots,x_d,y} p(x_1,\ldots,x_d,y) \log\left(\frac{p(x_1,\ldots,x_d,y)}{p(x_1|y)\cdots p(x_d|y)p(y)}\right) \\ &= \sum_y p(y) \sum_{x_1,\ldots,x_d} p(x_1,\ldots,x_d|y) \log\left(\frac{p(x_1,\ldots,x_d|y)}{p(x_1|y)\cdots p(x_d|y)}\right). \end{aligned} \tag{3}$$

Thus $CMI(X|Y) = E_{Y=y}(CMI(X|Y=y))$, where $CMI(X|Y=y)$ is Kullback-Leibler divergence between conditional distributions $P_{X|Y=y}$ and $P_{X_1|Y=y} \times \cdots \times P_{X_d|Y=y}$ and the expectation $E_{Y=y}$ is calculated with respect to the distribution of Y. Thus CMI is an *averaged* KL-divergence between these two conditional distributions. Note that, for $d=2$, CMI reduces to the conditional mutual information for two variables. It follows from non-negativity of K-L divergence ([6]) that CMI is non-negative. Moreover, the same argument implies that

$$CMI(X|Y) = 0 \quad \Leftrightarrow X_1 \perp X_2 \ldots \perp X_d|Y$$

as $CMI(X|Y) = 0$ entails that for any y in the support of Y distributions $P_{X|Y=y}$ and $P_{X_1|Y=y} \times \cdots \times P_{X_d|Y=y}$ coincide. We observe that $CMI(X|Y)$ can be de-constructed and represented as the combination of conditional entropies $H(X_i|Y)$ and $H(X|Y)$. Namely recalling that the conditional entropy of X given Y is defined as $H(X|Y) = -\sum_{x,y} p(x,y) \log p(x|y)$ we have from (3)

$$CMI(X|Y) = \sum_{i=1}^d H(X_i|Y) - H(X|Y) = \sum_{i=1}^d H(X_i,Y) - H(X,Y) - (d-1)H(Y). \tag{4}$$

We also remark that (2) can be written as

$$\begin{aligned} CMI(X|Y) = &-\sum_y p(y) \sum_{x_1,\ldots,x_d} p(x_1|y)\cdots p(x_d|y) \log(p(x_1|y)\cdots p(x_d|y)) \\ &-(-\sum_{x,y} p(x,y) \log p(x,y)), \end{aligned}$$

which yields interpretation of CMI as the change of entropy for the conditionally independent and conditionally dependent (X,Y). We also note that when Y is independent from $(X_1,\ldots,X_d)$ and thus conditioning can be omitted in (3), then definition of $CMI(X|Y)$ coincides with definition of unconditional multiinformation $MI(X)$ ([21],[17]) which measure how much structure of dependence of X deviates from the unconditional dependence of its coordinates. $MI(X)$ is frequently applied to detect interactions in Genome Wide Association Studies (see [5]). Let $\widehat{CMI}(X|Y)$ be defined as plug-in estimator of $CMI(X|Y)$ i.e. probabilities $p(x_1,\ldots,x_d)$ are replaced by fraction based on iid sample from $P_{X,Y}$ consisting of n observations. Properties of $\widehat{MI}(X)$ were studied by Studený in [16]. In the case when X is multivariate normal $N(0,\Sigma)$, where $\Sigma=(\sigma_{ij})$ is the covariance matrix, we have $MI(X)=2^{-1}\sum_{i=1}^d \log\sigma_{ii}^2-\log(|\Sigma|)$ and the properties of its sample counterpart has been studied in [13].

2.2 Properties of Conditional Multiinformation

Below we list some properties of conditional muliinformation. The first two are well-known but we find it useful to state them together with (iii)-(v).

Theorem 1 *Let $X=(X_1,\ldots,X_d)$. We have*
(i) *For any $i<d$*

$$CMI(X_1,\ldots X_{i+1}|Y)\geq CMI(X_1,\ldots X_i|Y)$$

(ii)

$$CMI(X|Y)=\sum_{i=2}^d MI(X_i;X_1,\ldots,X_{i-1}|Y),$$

where $MI(X_i;X_1,\ldots,X_{i-1}|Y)$ denotes the mutual information between X_i and $(X_1,\ldots,X_{i-1})$ given Y ([6]). (iii) *We have*

$$CMI(X|Y)=\inf_{\tilde{X}_1,\ldots\tilde{X}_d} D_{KL}(P_{X|Y}||P_{\tilde{X}_1|Y}\times\cdots\times P_{\tilde{X}_p|Y}|Y),$$

where $(\tilde{X}_1,\ldots,\tilde{X}_d,Y)$ is any discrete random vector supported on $\mathcal{X}_1\times\cdots\mathcal{X}_d\times\mathcal{Y}$ with distribution of Y equal to P_Y.
(iv) *Let $P_{X,Y}^{ind}$ be a distribution with mass function $p(y)p(x_1|y)\cdots p(x_d|y)$. Then*

$$CMI(X|Y)=D_{KL}(P_{Y|X}||P_{Y|X}^{ind})+D_{KL}(P_X||P_X^{ind}) \tag{5}$$

(v) *We have*

$$\frac{1}{2}\Big(\sum_{x_1,\ldots,x_d,y}|p(x_1,\ldots,x_d,y)-p(x_1|y)\cdots p(x_d|y)p(y)|\Big)^2\leq CMI(X|Y)\leq\log\big(\chi^2+1\big),$$

where χ^2 index is defined as

$$\chi^2=\sum_{x_1,\ldots,x_d,y}\frac{(p(x_1,\ldots,x_d,y)-p(x_1|y)\cdots p(x_d|y)p(y))^2}{p(x_1|y)\cdots p(x_d|y)p(y)}.$$

LHS and RHS equal 0 for the conditional independence case (1).

We prove part (v) here, the remaining proofs are relegated to the on-line supplement [3].

[3] https://github.com/teisseyrep/cmi

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_51

Proof. Note that the RHS inequality in (v) follows from Jensen's inequality ([6])

$$\sum_{x_1,\ldots,x_d,y} p(x_1,\ldots,x_d,y)\log\left(\frac{p(x_1,\ldots,x_d,y)}{p(x_1|y)\cdots p(x_d|y)p(y)}\right) \le$$
$$\log\left(\sum_{x_1,\ldots,x_d,y}\frac{p^2(x_1,\ldots,x_d,y)}{p(x_1|y)\cdots p(x_d|y)}\right) = \log\bigl(\chi^2+1\bigr).$$

LHS is a direct consequence of Pinsker's inequality ([20]) and (2).

Note that (iii) may be interpreted as $P_{X_1|Y}\times\ldots\times P_{X_d|Y}$ is the closest distribution consisting of conditionally independent coordinates given Y to $P_{X_1,\ldots,X_d|Y}$.

3 Main theoretical result

Let $CMI(X|Y)$ and $\widehat{CMI}(X|Y)$, where $X=(X_1,\ldots,X_d)$ be defined as before and assume that $p(x_1,\ldots,x_d,y)>0$ for all $(x_1,\ldots,x_d,y)\in\mathcal{X}_1\times\cdots\mathcal{X}_d\times\mathcal{Y}$. Obviously, even in the case when X_is are conditionally independent given Y $\widehat{CMI}(X|Y)$ will be strictly larger than 0 and we need to assess whether the deviations from 0 are due to random error caused by estimation of probabilities $p(x_1,\ldots,x_d)$ or to the fact that $CMI(X|Y)>0$. In order to account for this uncertainty, the distribution of $\widehat{CMI}(X|Y)$ under conditional independence is needed. We state now the result below supplementing it by the behaviour of the estimator when conditional independence is violated. It basically says that the distribution of $\widehat{CMI}(X|Y)$ when $X_1,\ldots X_d$ are not conditionally independent given Y, the asymptotic distribution is normal whereas in the opposite case of the null hypothesis $(X_1\perp X_2\ldots\perp X_d|Y)$ the distribution is weighted sum of squared independent normally distributed variables. Moreover, in the latter case $\widehat{CMI}$ converges to its theoretical value CMI more quickly: the rate of convergence is n^{-1} instead of $n^{-1/2}$. The result reduces to the known result for Conditional Mutual Information when $d=2$ for which the weights are equal to ones and the distribution coincides with chi squared distributed random variable with $k=(|\mathcal{X}_1|-1)(|\mathcal{X}_2|-1)|\mathcal{Y}|$ degrees of freedom. However, for $d>2$ this simplification does not hold. It also generalises the result by M. Studený ([16]) for $|\mathcal{Y}|=1$ i.e. for the case of unconditional multiinformation.

Theorem 2 *(i) Assume that $CMI(X|Y)\neq 0$. Then we have*

$$n^{1/2}(\widehat{CMI}(X|Y)-CMI(X|Y))\xrightarrow{d}N(0,\sigma^2_{\widehat{CMI}}), \tag{6}$$

where $\xrightarrow{d}$ denotes convergence in distribution, $\sigma^2_{\widehat{CMI}}$ equals

$$\sum_{x_1,\ldots,x_d,y} p(x_1,\ldots,x_d,y)\log^2\frac{p(x_1,\ldots,x_d|y)}{p(x_1|y)\cdots p(x_d|y)} - CMI^2(X|Y)$$
$$=\mathrm{Var}\left(\log\frac{p(X_1,\ldots X_d|Y)}{p(X_1|Y)\cdots p(X_d|Y)}\right)$$

DOI: 10.1007/978-3-030-77980-1_51

and $\sigma^2_{\widehat{CMI}} > 0$.
(ii) Assume that $CMI(X|Y) = 0$. *Then*

$$2n\widehat{CMI}(X|Y) \xrightarrow{d} \sum_{i=1}^{l} \lambda_i(M) Z_i^2, \tag{7}$$

where $l = |\mathcal{X}_1| \cdots |\mathcal{X}_d||\mathcal{Y}|$ *and* Z_i *are independent* $N(0,1)$ *and* $\lambda_i(M)$ *are eigenvalues of the matrix* M *defined as*

$$M_{x_1 \ldots x_d y}^{x'_1 \ldots x'_d y'} = I(x_1 = x'_1, \ldots, x_d = x'_d, y = y') - \sum_{i=1}^{d} I(x_i = x'_i, y = y') \frac{p(x'_1, \ldots, x'_d, y')}{p(x_i, y)}$$
$$+ I(y = y') \frac{p(x'_1, \ldots, x'_d, y')}{p(y)}.$$

with $M_{x_1 \ldots x_d y}^{x'_1 \ldots x'_d y'}$ *denoting element of* M *with row index* $x_1 \ldots x_d y$ *and column index* $x'_1 \ldots x'_d y'$; $I()$ *is an indicator function.*

The proof of the above Theorem can be found in on-line supplement. Note that M is a sparse matrix as its elements are non-zero only if one or more row and column indices coincide.

For $d = 2$ as $X_1, \ldots, X_d$ are independent given Y, the above formula reduces to:

$$M_{x_1 x_2 y}^{x'_1 x'_2 y'} = I(y = y') \left(I(x_1 = x'_1) - \frac{p(x'_1, y)}{p(y)} \right) \left(I(x_2 = x'_2) - \frac{p(x'_2, y)}{p(y)} \right).$$

and M can be shown to be indempotent ($M^2 = M$). Thus all its eigenvalues are 0 and 1 and as trace of M equals $(|\mathcal{X}_1| - 1)(|\mathcal{X}_2| - 1)|\mathcal{Y}|$ this yields the known result about asymptotic distribution of conditional mutual information under conditional independence ([10]). For general d, matrix M is not necessarily idempotent and asymptotic distribution of $\widehat{CMI}$ deviates from chi-squared distribution. We note that M can be estimated from the sample by its plug-in estimator $\widehat{M}$ and its eigenvalues $\lambda_i(\widehat{M})$ numerically determined. In this way we approximate limiting distribution in (7) and the approximation will serve as a limiting distribution for the proposed test of the conditional independence.

4 Detection of conditional dependence: permutation versus asymptotic method

4.1 Permutation method

A popular method of checking whether H_0 is violated is the permutation method adapted to the present problem. For a given sample generated from $P_{X,Y}$ and each strata $Y = y$ and $i = 1, \ldots, d$ we randomly permute values of ith coordinate of observations such that the corresponding value of Y equals y (see [19]).

DOI: 10.1007/978-3-030-77980-1_51

Permutations for each i are performed independently. Consequently, performing this operation separately for any value of y occurring in the original sample, we obtain a sample from distribution $P_{X,Y}^{ind}$ which satisfies H_0. We repeat this operation N times drawing N permuted samples in total and calculate corresponding values $\widehat{CMI}_k(X|Y)$ for $k = 1, \ldots, N$. Then empirical p-value

$$\hat{p} = \frac{\#\{k : \widehat{CMI}_k(X|Y) \geq \widehat{CMI}(X|Y)\}}{N},$$

where $\widehat{CMI}(X|Y)$ is empirical CMI, is calculated. Small value of $\hat{p}$ indicates conditional dependence.

4.2 Asymptotic method

For a given sample pertaining to P_{XY} we calculate $\widehat{CMI}(X|Y)$ and plug-in estimator $\hat{M}$ of matrix M defined in the previous section. We use now the fact that the asymptotic distribution W of $\widehat{CMI}(X|Y)$ is given in (7) and we approximate it by $\widehat{W}$ by plugging in $\lambda_i(\widehat{M})$ for $\lambda_i(M)$, where $\lambda_i(\widehat{M})$ are numerically calculated. Then rejection region for a given significance level α is given by $\{\widehat{CMI}(X|Y) \geq q_{\widehat{W},1-\alpha}\}$, where $q_{\widehat{W},1-\alpha}$ is quantile of order $1-\alpha$ of distribution $\widehat{W}$. We note that by using asymptotic distribution we avoid the main drawback of permutation method, namely generation of many samples for every value of $Y = y$ which may be very time consuming. R package `eigen` has been used to calculate the eigenvalues and package `COMpQuadForm` for quantiles of $\widehat{W}$.

5 Simulation study

5.1 Artificial data sets

The aim of the simulation experiments was to compare the performance of the tests described in previous section: asymptotic test and permutation test. In addition, we consider semi-parametric test based on the scaled chi squared distribution. It is defined as distribution of $\alpha\chi_d^2 + \beta$, where χ_d^2 is a chi square distribution with $d > 0$; parameters α, d, β are calculated based on permutation samples (see [9])

To assess the performance of the tests we use ROC-type curves which are generated in the following way. In each simulation, we generate two samples: D_0 and D_1 conforming to the null hypothesis and alternative hypothesis, respectively. So, D_0 is generated from distribution for which $X_1, \ldots, X_d$ are conditionally independent given Y, whereas D_1 is generated from distribution for which $X_1, \ldots, X_d$ are conditionally dependent given Y. Then, we run a considered test for both D_0 and D_1 using significance level $\alpha \in (0, 1)$ and report whether the null hypothesis has been rejected. Importantly, the reference distributions of empirical CMI under null hypothesis are different for D_0 and D_1. The above steps are repeated 1000 times which allows to approximate probabilities of rejection in the cases when samples conform and do not conform to H_0, respectively. Each point on the curve corresponds to a different value of α. Observe that the first coordinate of each point is an approximation of type I error for some α obtained using D_0, wheres the second coordinate is an approximation of the power of the test obtained from D_1 for the same value of α. We also report Area Under the Curve (AUC), the larger the value of AUC the better is the performance of the test (we observe larger power for a fixed value of type I error). The above method has two important advantages. First, we control simultaneously the type I error and the corresponding power of the test. Secondly, it is possible to analyze power and type I error for different values of significance levels α. As we found that the actual levels of significance of the asymptotic test may exceed assumed levels of significance in some cases for medium sample sizes we view presented ROC curve analysis to be a more objective way of comparing tests, as they enable comparison of methods at the same level of significance.

We consider the following simulation models.

1. **Simulation model 1**.
   - Sample D_0: Generate $Y \sim B(1, 0.5)$ and $X_1, \ldots, X_d | Y = y \sim N(y, 1)$.
   - Sample D_1: Generate $Y \sim B(1, 0.5)$, $X_1, \ldots, X_{d-1} | Y = y \sim N(y, 1)$ and $X_d | X_{d-1} = x_{d-1} \sim N(\gamma x_{d-1}, 1)$, where γ is a parameter.

   Simulation model 1p. Modification of model 1. The only difference is that in D_1, $X_d | X_{d-1} = x_{d-1}, Y = y \sim N(\gamma x_{d-1} + y, 1)$
2. **Simulation model 2**.
   - Sample D_0: Generate $Y \sim B(1, 0.5)$ and $X_1, \ldots, X_d | Y = y \sim N(y, 1)$.
   - Sample D_1: Generate $Y \sim B(1, 0.5)$ and $Z \sim B(1, 0.5)$, where $Y \perp Z$. Next, generate $X_1, \ldots, X_d | Y = y, Z = z \sim N(\gamma(z + y), 1)$, where γ is a parameter.
3. **Simulation model 3**.
   - Sample D_0: Generate $X_1, \ldots, X_d \sim N(0, 1)$ and $Y \sim B(1, 0.5)$.
   - Sample D_1: Generate $X_1, \ldots, X_d \sim N(0, 1)$ and then $Y | X = x \sim B(1, \sigma(\gamma \cdot x^T \mathbf{1}))$ where γ is a parameter and $\mathbf{1} = (1, \ldots, 1)^T$.

The above simulation models are chosen to represent various dependency structures. Figure 1 and 2 show the graphs corresponding to distributions conforming to H_0 and H_1, respectively, for simulation models 1-3. Models 1-2 are generative models, as we first generate Y and then $X_1, \ldots, X_d$, whereas model 3 is a discriminative model, corresponding to scenario of supervised classification. For models 1-2 and sample D_0, variables $X_1, \ldots, X_d$ are conditionally independent given Y, but at the same time they are not unconditionally independent. For model 3 and D_0, all considered variables are independent and also conditionally independent. In all three models, parameter γ controls the difficulty of the problem. For larger γ it is easier to reject the null hypothesis for sample D_1.

Figure 3 shows the ROC-type curves for simulation models 1-3, for $n = 500$ and $d = 7$ (results for other values of d are placed in supplement). As expected, AUC increases with parameter γ, which is due to the fact that for larger γ the conditional dependence is stronger. The proposed asymptotic test works better

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_51

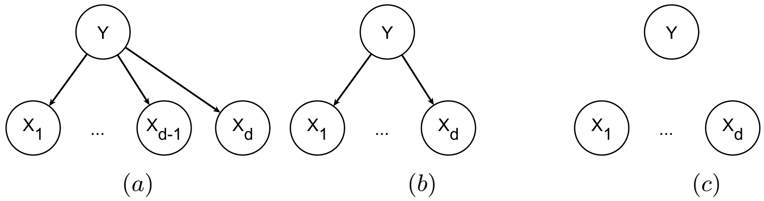

Fig. 1. Dependency structures corresponding to distributions conforming to H_0, for simulation model 1 (a), model 2 (b) and model 3 (c).

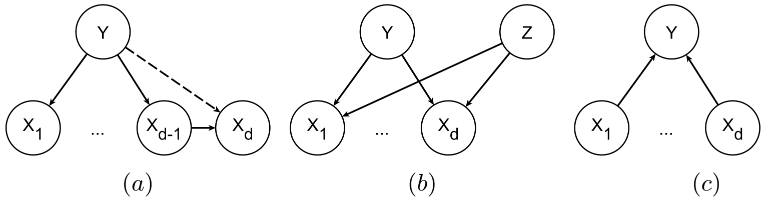

Fig. 2. Dependency structures corresponding to distributions conforming to H_1, for simulation model 1 (a), model 2 (b) and model 3 (c). Dashed line in (a) corresponds to model 1p.

(in terms of AUC) than the remaining tests for all simulation models except model 1p for which it works on par with permutation test. The advantage of asymptotic test is most pronounced for model 2. The test based on scaled chi squared distribution performs worse than the two competitors, which confirms our theoretical results (see Theorem 2) indicating that the reference distribution of CMI under null hypothesis significantly deviates from chi squared distribution for $d > 2$ and is poorly approximated by the scaled chi squared distribution. We also experimented with smaller $d = 3, 5$, for which all considered methods perform similar (the results are placed in supplement).

5.2 Analysis of medical data set MIMIC-III

We also illustrate the problem of conditional independence testing using real medical dataset MIMIC-III [8] containing information about patients from the intensive care units. We are interested in finding out whether the occurrences of some diseases are conditionally independent given gender. We consider 10 diseases (shortened names and the prevalences estimated from data are given in brackets): hypertension (66%), kidney failure (kidney; 35%), disorders of fluid electrolyte balance (fluid; 37%), hypotension (14%), disorders of lipoid metabolism (lipoid; 37%), liver disease (liver; 7%), diabetes (32%), thyroid disease (thyroid; 14%), chronic obstructive pulmonary disease (copd; 23%) and thrombosis (6%). We analysed all triples, i.e. we tested the null hypothesis of conditional independence between X_1, X_2, X_3 given Y, where X_1, X_2, X_3 denote occurrences of three out of ten of the above diseases and Y is gender. So, in total, we performed $\binom{10}{3} = 120$ tests. We present the results for triples with the largest (kidney, fluid, diabetes) and the smallest (hypotension, liver, diabetes) value of CMI (Figures 4 and 5, respectively). Each figure shows the values of joint conditional probabilities $P(X_1, X_2, X_3|Y)$ (red bars) and products of marginal conditional probabilities $P(X_1|Y)P(X_2|Y)P(X_3|Y)$ (blue bars). According to the asymptotic test, we reject the null hypothesis in the case of triple: kidney, fluid, diabetes (p-value equal to 0.0004 is smaller than 0.05/120) and we do not reject the null hypothesis in the case of diseases (hypotension, liver, diabetes, p-value equal to 0.4719). We assumed standard significance level $\alpha = 0.05$ and used Bonferroni correction in order to account for multiple tests. As expected, in the latter case, the joint conditional probabilities are very close to the corresponding products of marginal conditional probabilities (see Figure 5). When the null hypothesis is rejected, the differences between values of the probabilities are significantly larger (Figure 4). In particular, for kidney, fluid, diabetes, the probability of the occurrence of all diseases at the same time for females is $P(X_1 = 1, X_2 = 1, X_3 = 1|Y = \text{female}) = 8\%$, whereas the corresponding product is $P(X_1 = 1|Y = \text{female})P(X_2 = 1|Y = \text{female})P(X_3 = 1|Y = \text{female}) = 4\%$. Further analysis using Conditional Mutual Information detects pairwise conditional dependencies between kidney and fluid and kidney and diabetes (both p-values of order 10^{-9}) but no dependence is detected between fluid and diabetes (p-value 0.36).

6 Conclusions

In this paper we investigated the properties of conditional multiinformation (CMI), which is a natural measure of a strength of conditional dependence between multiple variables. Our main theoretical contribution is deriving asymptotic distribution of sample CMI (Theorem 2). It is a generalization of well known result for the case of two variables ($d = 2$). Moreover, we constructed a statistical test based on the distribution. Importantly, the asymptotic distribution of sample CMI significantly deviates from chi squared distribution for $d > 2$. This explains why the simple test based on scaled chi squared distribution works poorly when more than two variables are taken into account. The proposed asymptotic test usually outperforms permutation test in terms of power, when the number of variables is moderate. Its advantage over permutation test is that we avoid generating many permutation samples. On the other hand, asymptotic test requires numerical calculation of eigenvalues using matrix whose size significantly increases with the number of variables. Thus, the method may fail when the number of variables d increases. Therefore, the proposed asymptotic test is strongly recommended for moderate number of variables, whereas for larger d we recommend permutation test.

References

1. A. Bellot and M. van der Schaar. Conditional independence testing using generative adversarial networks. In *Advances in Neural Information Processing Systems 32*,

pages 2199–2208, 2019.

2. T. B. Berrett, Y. Wang, R. F. Barber, and R. J. Samworth. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):175–197, 2020.
3. P. Bühlmann and S. van de Geer. *Statistics for High-dimensional Data*. Springer, 1st edition, 2015.
4. E. Candès, Y. Fan, L. Janson, and J. Lv. Panning for gold: model-x knockoffs for high-dimensional controlled variable selection. *Journal of the Royal Statistical Society B*, 80:551–577, 2018.
5. P. Chanda and et al. Ambience: A novel approach and efficient algorithm for identifying informative genetic and environmental associations with complex phenotypes. *Genetics*, 180:1191–2010, 2008.
6. T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.
7. A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(1):1–15, 1979.
8. A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, Anthony C. L., and R. G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:1–9, 2016.
9. M. Kubkowski and J. Mielniczuk. Asymptotic distributions of interaction information. *Methodology and Computing in Applied Probability*, 2020.
10. S. Kullback. *Information Theory and Statistics*. Peter Smith, 1978.
11. C. Li and X. Fan. On nonparametric conditional independence tests for continuous variables. *WIREs Computational Statistics*, 12:1–11, 2020.
12. J. Pearl. *Causality*. Cambridge University Press, 2009.
13. T. Rowe and D. Troy. The sampling distribution of the total correation for multivariate gaussian random variables. *Entropy*, 21:–, 2019.
14. J. Runge. Conditional independence testing based on a nearest neighbour estimator of conditional mutual information. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, PMLR-84, pages 938–947, 2018.
15. P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.
16. M Studený. Asymptotic behaviour of empirical multiinformation. *Kybernetika*, 23:124–135, 1987.
17. M. Studený and J. Vejnarová. The multiinformation as a tool for measuring stochastic dependence. In *Learning in Graphical Models*, MIT Press, pages 66–82, 1999.
18. I. Tsamardinos, C. Aliferis, and A. Statnikov. Algorithms for large scale Markov Blanket discovery. In *FLAIRS Conference*, pages 376–381, 2003.
19. I. Tsamardinos and G. Borboudakis. Permutation testing improves on Bayesian network learning. In *Proceedings of ECML PKDD 2010*, pages 322–337, 2010.
20. A. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 1st edition, 2009.
21. S. Watanabe. Information theoretical analysis of multivariate correlation. *IBM J. Res. Dev.*, 4:66–82, 1960.
22. K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, page 804–813, 2011.

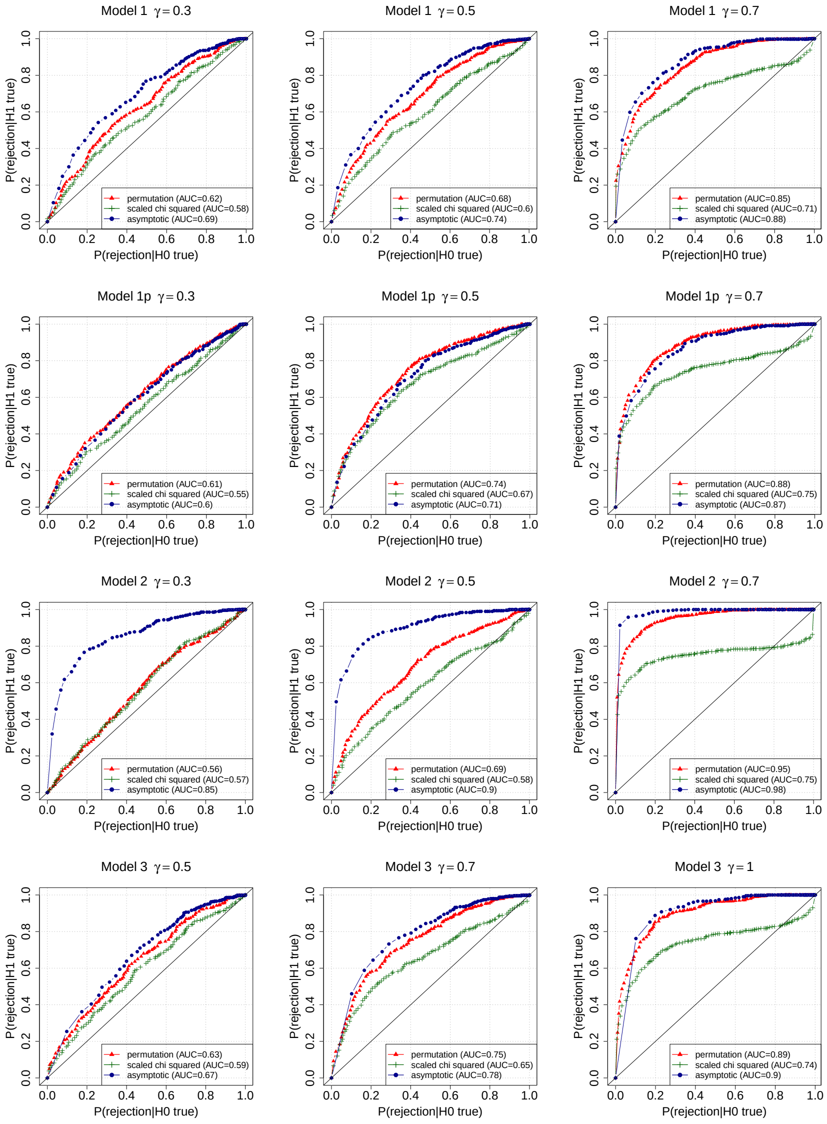

Fig. 3. ROC-type curves for simulation models 1, 1p, 2, 3 and permutation test (red), scaled chi-squared test (green) and asymptotic test (blue). Number of variables $d = 7$ and sample size $n = 500$.

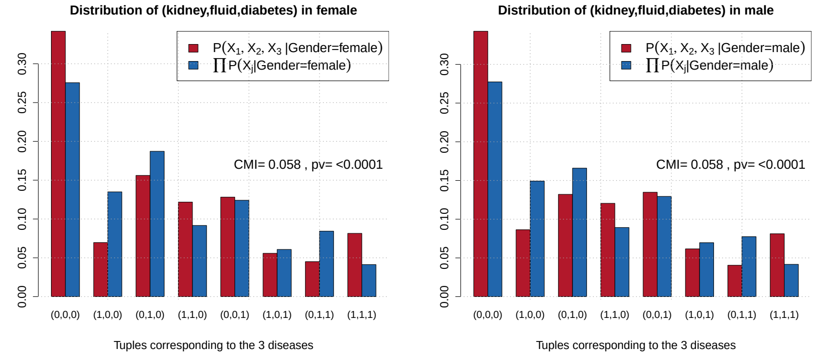

Fig. 4. Example based on MIMIC-III database. The bars correspond to: (1) conditional distribution $P(X_1 = x_1, X_2 = x_2, X_3 = x_3|Y = y)$ of three diseases (kidney, fluid, diabetes) given gender and (2) product of the conditional for all values (x_1, x_2, x_3). The null hypothesis of conditional independence is rejected for $\alpha = 0.05$ (p-value< 0.0001).

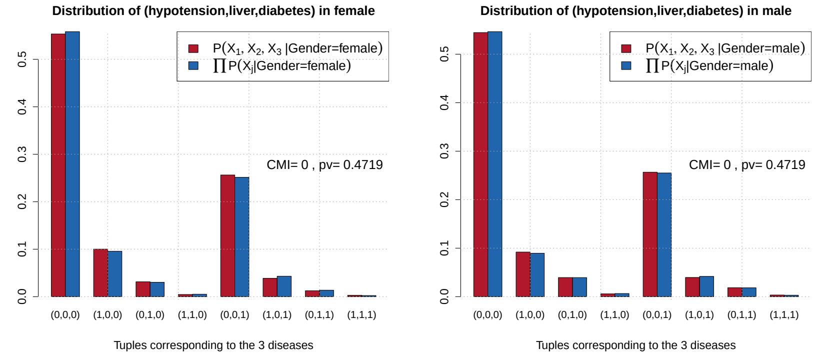

Fig. 5. Example based on MIMIC-III database. The bars correspond to: (1) conditional distribution $P(X_1 = x_1, X_2 = x_2, X_3 = x_3|Y = y)$ of three diseases (hypotension, liver, diabetes) given gender and (2) product of the conditional distributions for all values (x_1, x_2, x_3). The null hypothesis of conditional independence is not rejected for $\alpha = 0.05$ (p-value$= 0.4719$).