Distributions of a general reduced-order dependence measure and conditional independence testing

Mariusz Kubkowski$^{[000-0002-1453-5589]}$, Małgorzata Łazęcka$^{1,2[0000-0003-0975-4274]}$, and Jan Mielniczuk$^{1,2[0000-0003-2621-2303]}$

[1] Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
[2] Warsaw University of Technology, Faculty of Mathematics and Information Sciences, Warsaw, Poland
m.kubkowski,miel,malgorzata.lazecka@ipipan.waw.pl

Abstract. We study distributions of a general reduced-order dependence measure and apply the results to conditional independence testing and feature selection. Experiments with Bayesian Networks indicate that using the introduced test in the Grow and Shrink algorithm instead of Conditional Mutual Information yields promising results for Markov Blanket discovery in terms of F measure.

Keywords: conditional mutual information · asymptotic distribution · feature selection · Markov blanket · reduced-order dependence measure.

1 INTRODUCTION

Consider a problem of selecting a subset of all potential predictors $\{X_1, \ldots, X_p\}$ to predict an outcome Y, which consists of all predictors significantly influencing it. Selection of active predictors leads to dimension reduction and is instrumental for many machine learning and statistical procedures, in particular in structure learning of dependence networks. Commonly for this task, such methods incorporate a sequence of conditional independence tests, among which the test based on Conditional Mutual Information (CMI) is the most frequent. In the paper we consider properties of a general information-based dependence measure $J^{\beta,\gamma}(X, Y|X_S)$ introduced in [2] in a context of constructing approximations to CMI. This is a reduced-order approximation which disregards approximations of order higher than 3. It can also be considered as a measure of predictive power of X for Y when variables $X_S = (X_s\,, s \in S)$ have been already chosen for this task. Special cases include Mutual Information Minimization (MIM), Minimum Redundancy Maximum Relevance (MrMR) [11], Mutual Information Feature Selection (MIFS) [1], Conditional Information Feature Extraction (CIFE) [7] and Joint Mutual Information (JMI) [14] criteria. They are routinely used in nonparametric approaches to feature selection, variable importance ranking and causal discovery (see e.g. [4], [13]). However, theoretical properties of such criteria remain largely unknown hindering study of associated selection methods.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50436-6_51

Here we show that $\hat{J}^{\beta,\gamma}(X, Y|X_S)$ exhibits dichotomous behaviour meaning that its distribution can be either normal or coincides with a distribution of a certain quadratic form in normal variables. The second case is studied in detail for binary Y. In particular for two popular criteria CIFE and JMI, conditions under which their distributions converge to distributions of quadratic form are made explicit. As two cases of dichotomy differ in behaviour of the variance of $\hat{J}^{\beta,\gamma}$, its order of convergence is used to detect which case is actually valid. Then a parametric permutation test (i.e. a test based on permutations to estimate parameters of the chosen distribution) is used to check whether candidate variable X is independent of Y given X_S.

2 PRELIMINARIES

2.1 Entropy and Mutual Information

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 $p(x, y) = P(X = x, Y = y)$. Entropy for discrete random variable X is defined as

$$H(X) = -\sum_{x\in\mathcal{X}} p(x) \log p(x). \tag{1}$$

Entropy quantifies the uncertainty of observing random values of X. In case of discrete X, $H(X)$ is non-negative and equals 0 when the probability mass is concentrated at one point. The above definition naturally extends to the case of random vectors (i.e. X can be multivariate random variable) by using multivariate probability instead of univariate probability. In the following we will frequently consider subvectors of $X = (X_1, \ldots, X_p)$ which is a vector of all potential predictors of class index Y. The conditional entropy of X given Y is written as

$$H(X|Y) = \sum_{y\in\mathcal{Y}} p(y)H(X|Y = y) \tag{2}$$

and the mutual information (MI) between X and Y is

$$I(X, Y) = H(X) - H(X|Y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}. \tag{3}$$

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 equals zero if and only if X and Y are independent and thus it is able to discover non-linear relationships. It is easily seen that $I(X, Y) = H(X) + H(Y) - H(X, Y)$. A natural extension of MI is conditional mutual information (CMI) defined as

$$I(X, Y|Z) = H(X|Z) - H(X|Y, Z), \tag{4}$$

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50436-6_51

which measures the conditional dependence between X and Y given Z. An important property is chain rule for MI which connects $I((X_1, X_2), Y)$ to $I(X_1, Y)$:

$$I((X_1, X_2), Y) = I(X_1, Y) + I(X_2, Y|X_1). \tag{5}$$

For more properties of the basic measures described above we refer to [3]. A quantity, used in next sections, is interaction information (II) [9]. The 3-way interaction information is defined as

$$II(X_1, X_2, Y) = I(Y, X_1|X_2) - I(Y, X_1), \tag{6}$$

which is consistent with an intuitive meaning of existence of interaction as a situation in which the effect of one variable on the class variable depends on the value of another variable.

2.2 Approximations of Conditional Mutual Information

We consider a discrete class variable Y and p discrete features $X_1, \ldots, X_p$. Let X_S denote a subset of features indexed by a subset $S \subseteq \{1, \ldots, p\}$. We employ here greedy search for active features based on forward selection. Assume that S is a set of already chosen features, S^c its complement and $j \in S^c$ a candidate feature. In each step we add a feature whose inclusion gives the most significant improvement of the mutual information, i.e. we find

$$\arg\max_{j \in S^c} \left[I(X_{S \cup \{j\}}, Y) - I(X_S, Y) \right] = \arg\max_{j \in S^c} I(X_j, Y|X_S). \tag{7}$$

The equality in (7) follows from (5). Observe that (7) indicates that we select a feature that achieves the maximum association with the class given the already chosen features. For example, first-order approximation yields $I(X_j, Y)$, which is a simple univariate filter MIM, frequently used as a pre-processing step in high-dimensional data analysis. However, this method suffers from many drawbacks as it does not take into account possible interactions between features and redundancy of some features. When the second order approximation is used, the dependence score for candidate feature is

$$\begin{aligned} J(X_j) &= I(X_j, Y) + \sum_{i \in S} II(X_i, X_j, Y) \\ &= I(X_j, Y) + \sum_{i \in S} [I(X_i, X_j|Y) - I(X_i, X_j)]. \end{aligned} \tag{8}$$

The second equality uses (6). In literature (8) is known as CIFE (Conditional Infomax Feature Extraction) [7] criterion. Observe that in (8) we take into account not only relevance of the candidate feature, but also its possible interactions with the already selected features. However, frequently it is useful to scale down the corresponding term ([2]). Among such modifications the most popular is JMI

$$J(X_j) = I(X_j, Y) + \frac{1}{|S|} \sum_{i \in S} [I(X_i, X_j|Y) - I(X_i, X_j)] = \frac{1}{|S|} \sum_{i \in S} I(X_j, Y|X_i),$$

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50436-6_51

where the second equality follows from (5). JMI was also proved to be an approximation of CMI under certain dependence assumptions ([13]). Data-adaptive version of JMI will be considered in Section 4. In [2] it is proposed to consider a general information-theoretic dependence measure

$$J^{\beta,\gamma}(X_j, Y|X_S) = I(X_j, Y) - \beta \sum_{i\in S} I(X_i, X_j) + \gamma \sum_{i\in S} I(X_i, X_j|Y), \tag{9}$$

where β, γ are some positive constants usually depending in decreasing manner on the size $|S| = k$ of set S. Several frequently used selection criteria are special cases of (9). MrMR criterion ([11]) corresponds to $(\beta, \gamma) = (|S|^{-1}, 0)$ whereas more general MIFS (Mutual Information Feature Selection) criterion [1] corresponds to pair $(\beta, 0)$. Obviously, the simplest criterion MIM corresponds to $(0, 0)$ pair. CIFE defined above in (8) is obtained for $(1, 1)$ pair, whereas $(\beta, \gamma) = (1/|S|, 1/|S|)$ leads to JMI. In the following we consider asymptotic distributions of the sample version of $J^{\beta,\gamma}(X_j)$, namely

$$\hat{J}^{\beta,\gamma}(X_j) = \hat{I}(X_j, Y) - \beta \sum_{i\in S} \hat{I}(X_i, X_j) + \gamma \sum_{i\in S} \hat{I}(X_i, X_j|Y), \tag{10}$$

and show how the distribution depends on underlying parameters. In this way we gain a more clear idea what is an influence of β and γ on the behaviour of $\hat{J}^{\beta,\gamma}$. Sample version in (10) is obtained by plugging in fractions of observations instead of probabilities in (3) and (4).

3 DISTRIBUTIONS OF A GENERAL DEPENDENCE MEASURE

In the following we will state our theoretical results which study asymptotic distributions of $\hat{J}^{\beta,\gamma}(X, Y|Z)$ where $Z = (Z_1, \ldots, Z_{|S|})$ is possible multivariate discrete vector and then we apply it to previously introduced framework by putting $X := X_j$ and $Z := (X_1, \ldots, X_{|S|})$. We will show that its distribution is either approximately normal or, if the asymptotic variance vanishes, is approximately equal to distribution of quadratic form of normal variables. Let $p = (p(x, y, z))_{x,y,z}$ be a vector of probabilities for (X, Y, Z) and we assume whence forth that $p(x, y, z) > 0$ for any triple of (x, y, z) values in the range of (X, Y, Z). Moreover, $f(p)$ equals $J^{\beta,\gamma}(X, Y|Z)$ treated as a function of p, Df denotes a derivative of function f and $\xrightarrow{d}$ convergence in distribution. The special case of the result below for CIFE criterion has been proved in [6].

Theorem 1. *(i) We have*

$$n^{1/2}(\hat{J}^{\beta,\gamma}(X, Y|Z) - J^{\beta,\gamma}(X, Y|Z)) \xrightarrow{d} N(0, \sigma^2_{\hat{J}}), \tag{11}$$

where $\sigma^2_{\hat{J}} = Df(p)^T \Sigma Df(p) = \mathrm{Var}(Df(p)^T \hat{p})$ *and* $\Sigma = n\mathrm{Var}(\hat{p} - p)$.
(ii) If $\sigma^2_{\hat{J}} = 0$ *then*

$$2n(\hat{J}^{\beta,\gamma}(X, Y|Z) - J^{\beta,\gamma}(X, Y|Z)) \xrightarrow{d} V^T HV, \tag{12}$$

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50436-6_51

where V follows $N(0,\Sigma)$ distribution, $\Sigma_{xyz}^{x'y'z'} = p(x',y',z')(I(x=x',y=y',z=z') - p(x,y,z))/n$ and $H = D^2 f(p)$ is a Hessian of f.

Proof. Note that $f(p) = J^{\beta,\gamma}(X,Y|Z)$ equals

$$I(X,Y) - \sum_{s\in S}(\beta I(X,Z_s) - \gamma I(X,Z_s|Y)) = \sum_{x,y,z} p(x,y,z)\Bigg(\ln\left(\frac{p(x,y)}{p(x)p(y)}\right) - \sum_{s\in S}\left(\beta\ln\left(\frac{p(x,z_s)}{p(x)p(z_s)}\right) - \gamma\ln\left(\frac{p(x,y,z_s)p(y)}{p(x,y)p(y,z_s)}\right)\right)\Bigg).$$

After some calculations one obtains that $\frac{\partial f(p)}{\partial p(x,y,z)}$ equals for $z=(z_1,\ldots,z_{|s|})$

$$\ln\left(\frac{p(x,y)}{p(x)p(y)}\right) - \beta\sum_{s\in S}\left(\ln\left(\frac{p(x,z_s)}{p(x)p(z_s)}\right) - 1\right) + \gamma\sum_{s\in S}\ln\left(\frac{p(x,y,z_s)p(y)}{p(x,y)p(y,z_s)}\right) - 1. \tag{13}$$

Let $\hat{p}(x,y,z) = n(x,y,z)/n$, $\hat{p} = (\hat{p}(x,y,z))_{x,y,z}$. Then $\hat{J}^{\beta,\gamma}(X,Y|Z) = f(\hat{p})$. The remaining part of the proof relies on Taylor's formula for $f(\hat{p}) - f(p)$. Details are given in supplemental material [5].

We characterize the case when $\sigma_{\hat{J}}^2 = 0$ in more detail for binary Y and $\beta=\gamma\neq 0$ which encompasses CIFE and JMI criteria. Note that binary Y case covers an important situation of distinguishing between cases ($Y=1$) and control ($Y=0$). We define two scenarios:

- Scenario 1 (S1): $X \perp Y|Z_s$ for any $s\in S$ and $X\perp Y$ ($X\perp Y|Z$ denotes conditional independence of X and Y given Z).
- Scenario 2 (S2): $\exists W\subset S$ such that $W\neq\emptyset$ and for $s\in W$ $Z_s\perp Y|X$, $X\not\perp Y|Z_s$ and for $s\in W^c$ we have $X\perp Y|Z_s$.

Define W as

$$W = \left\{s\in S : \exists_{x,y,z_s}\frac{p(x,y,z_s)p(z_s)}{p(x,z_s)p(y,z_s)} \neq 1\right\}. \tag{14}$$

We will study in detail the case when $\sigma_{\hat{J}}^2 = 0$ and either $\beta=\gamma\neq 0$ or at least one of the parameters β,γ equal 0. We note that all cases of used information-based criteria fall in one of these categories ([2]). We have

Theorem 2. *Assume that $\sigma_{\hat{J}}^2 = 0$ and $\beta=\gamma\neq 0$. Then we have:*
(i) If $|S|>1$ and $\beta^{-1}\in\{1,2,\ldots,|S|-1\}$ then one of the above scenarios holds with W defined in (14).
(ii) If $\beta^{-1} = |S|$ or $\beta^{-1}\notin\{1,2,\ldots,|S|-1\}$ then Scenario 1 is valid.

The analogous result can be stated for the case when at least one of the parameters β or γ equals 0 (details are given in supplement [5]).

DOI: 10.1007/978-3-030-50436-6_51

3.1 Special Case: JMI

We state below corollary for criterion JMI. Note that in view of Theorem 2 Scenario 2 holds for JMI. Let

$$\sigma^2_{\widehat{JMI}} = \sum_{x,y,z} p(x,y,z) \left(\frac{1}{|S|} \sum_{s \in S} \ln \frac{p(x,y,z_s)p(z_s)}{p(x,z_s)p(y,z_s)} \right)^2 - (JMI)^2. \tag{15}$$

Corollary 1. *Let Y be binary. (i) If $\sigma^2_{\widehat{JMI}} \neq 0$ then*

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

(ii) If $\sigma^2_{\widehat{JMI}} = 0$ then $JMI = 0$ and

$$2n\widehat{JMI} \xrightarrow{d} V^T HV,$$

where V and H are defined in Theorem 1. Moreover in this case Scenario 1 holds.

Note that $\sigma^2_{\widehat{JMI}} = 0$ implies $JMI = 0$ as in this case Scenario 2 holds. The result for CIFE is analogous (see supplemental material [5]).
In both cases we can infer the type of limiting distribution if the corresponding theoretical value of the statistic is nonzero. Namely, if $JMI \neq 0$ ($CIFE \neq 0$) then $\sigma^2_{\widehat{JMI}} \neq 0$ (respectively, $\sigma^2_{\widehat{CIFE}} \neq 0$) and the limiting distribution is normal. Checking that $JMI \neq 0$ is simpler than $CIFE \neq 0$ as it is implied by $X \not\perp Y|Z_s$ for at least one $s \in S$. Actually, $JMI = 0$ is equivalent to conditional independence of X and Y given Z_s for any $s \in S$ which in its turn is equivalent to $\sigma^2_{\widehat{JMI}} = 0$. In the next section we will use a behaviour of the variance to decide which distribution to use as a benchmark for testing conditional independence. In a nutshell, the corresponding switch which is constructed in data-adaptive way and is based on different order of convergence of the variance to 0 in both cases. This is exemplified in the Figure 1 which shows boxplots of the empirical variance of JMI multiplied by sample size in two cases, when the theoretical variance is 0 (model M2 discussed below) or not (model M1). The Figure clearly indicates that the switch can be based on the behaviour of the variance.

4 JMI-BASED CONDITIONAL INDEPENDENCE TEST AND ITS BEHAVIOUR

4.1 JMI-Based Conditional Independence Test

In the following we use $\hat{J} = \widehat{JMI}$ as a test statistic for testing conditional independence hypothesis

$$H_0 : X \perp Y|X_S. \tag{16}$$

where X_S denotes set of X_i with $i \in S$. A standard way of testing it is to use

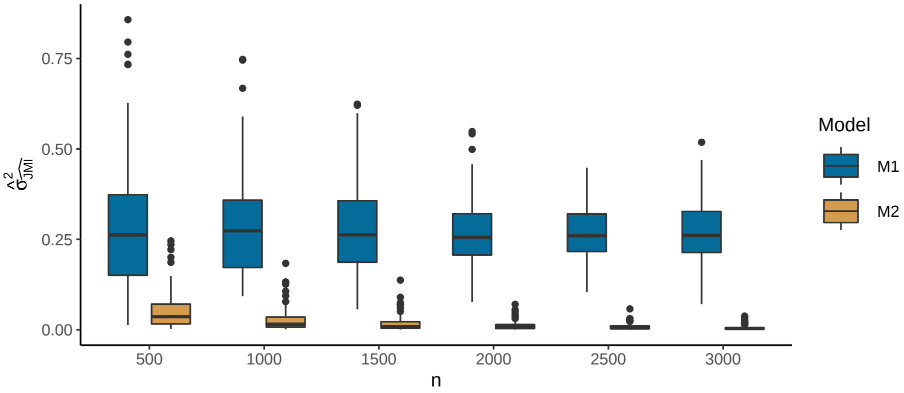

Fig. 1. Behaviour of the empirical variance multiplied by n in the case when corresponding value of $\sigma^2_{\widehat{JMI}}$ is zero (yellow) or not (blue). Models: M1, M2 (see text), JMI $= JMI(X_1^{(1)}, Y|X_1, \ldots, X_5)$, $n = 1000$, $\rho = 0$, $\gamma = 1$.

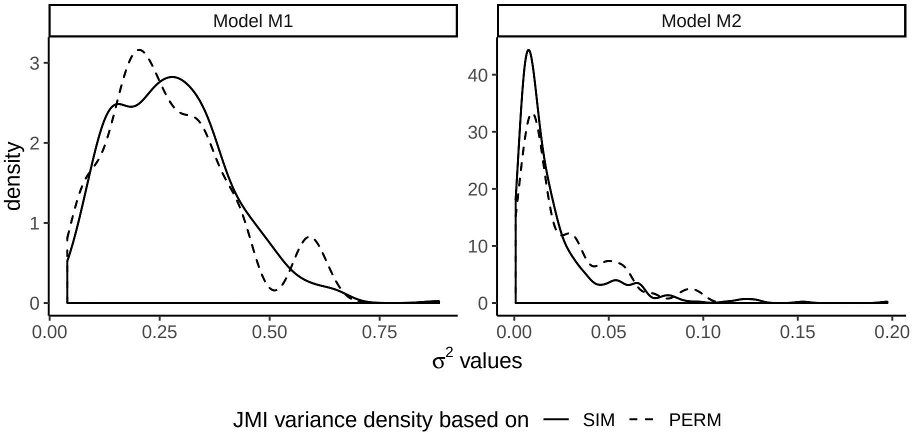

Fig. 2. Comparison of variances' distributions under conditional independence hypothesis. SIM corresponds to distribution of $\hat{\sigma}^2_{\widehat{JMI}}$ based on $N = 500$ simulated samples. PERM is based on $N = 50$ simulated samples. For each of them X was permuted on strata ($B = 1$) and $\hat{\sigma}^2_{\widehat{JMI}}$ was calculated. Models: M1, M2 (see text), JMI $= JMI(X_1^{(1)}, Y|X_1, \ldots, X_5)$, $n = 1000$, $\rho = 0$, $\gamma = 1$

Conditional Mutual Information (CMI) as a test statistic and its asymptotic distribution to construct the rejection region. However, it is widely known that such test loses power when the size of conditioning set grows due to inadequate estimation of $p(x, y|X_S = x_S)$ for all strata $\{X_S = x_S\}$. Here we use as a test statistic $\hat{J}$ which does not suffer from this drawback as it involves conditional probabilities given univariate strata $\{X_s = x_s\}$ for $s \in S$. As behaviour of $\hat{J}$ is

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50436-6_51

dichotomous on (16) we consider a data-dependent way of determining which of the two distributions: normal or distribution of quadratic form (abbreviated to d.q.f. further on) is closer to distribution of $\hat{J}$. Here we propose a switch based on the connection between distribution of the statistics and its variance (see Theorem 1). We consider the test based on JMI as in this case $\sigma^2_{\widehat{JMI}} = 0$ is equivalent to $JMI = 0$. Namely, it is seen from Theorem 1 that normality of asymptotic distribution corresponds to the case when the asymptotic variance calculated for samples of size n and $n/2$ should be approximately the same and should be strictly smaller for a larger sample otherwise. For each strata $X_S = x_S$ we permute corresponding n_{X_S} values of X B times and for each permutation we obtain value of $\widehat{JMI}$ as well as an estimator of its asymptotic variance v_n. The permutation scheme is repeated for randomly chosen subsamples of original sample of size $n/2$ and B values of $v_{n/2}$ are calculated. We than compare the mean of v_n with the mean of $v_{n/2}$ using t-test. If the equality of the means is not rejected we bet on normality of asymptotic distribution, in the opposite case d.q.f. is chosen. Note that permuting samples for a given value $X_S = x_S$ we generate data (X_{perm}, Y, X_S) which follows null hypothesis (16) while keeping the distribution $P_{X|X_S} = P_{X_{perm}|X_S}$ unchanged. In Figure 2 we show that when conditional independence hypothesis is satisfied then distribution of estimated variance $\hat{\sigma}^2_{\widehat{JMI}}$ based on permuted samples follows closely distribution of $\hat{\sigma}^2_{\widehat{JMI}}$ based on independent samples. Thus indeed using permutation scheme described above we can approximate the distribution of the variance of JMI under H_0 for a fixed conditional distribution $\hat{\sigma}^2_{\widehat{JMI}}$.

Now we approximate sample distribution of $\widehat{JMI}$ by $N(\hat{\mu}, \hat{\sigma}^2)$ when normal distribution has been picked or when d.q.f. has been picked approximation is $\chi^2_{\hat{\mu}}$ (with $\hat{\mu}$ being the empirical mean of $\widehat{JMI}$) or scaled chi square $\hat{\alpha}\chi^2_{\hat{d}} + \hat{\beta}$ where parameters are based on three first empirical moments of the permuted samples ([15]). Then the observed value $\widehat{JMI}$ is compared to quantile of the above benchmark distribution and conditional independence is rejected when this quantile is exceeded. Note that as parametric permutation test is employed we need much smaller B than in the case of non-parametric permutation test and we use $B = 50$. Algorithm will be denoted by JMI(norm/chi) or JMI(norm/chi_scale) depending on whether chi square or scaled chi square is considered in the switch. The pseudo-code of the algoritm is given below in Algorithm 1 and the code itself is available in [5]. For comparison we consider two tests: asymptotic test for CMI (called CMI) and semi-parametric permutation test (called CMI(sp)) proposed in [12]. In CMI(sp) the permutation test is used to estimate the number of degrees of freedom of reference chi square distribution.

4.2 Numerical Experiments

We investigate the behaviour of the proposed test in two generative tree models shown in the left and the right panel of Figure 3 which will be called M1 and M2. Note that in model M1 $X_1^{(1)} \perp Y|X_1, \ldots, X_k$ whereas for model M2 the

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50436-6_51

Algorithm 1: $JMI(chi/norm)$

```
Input      : Training data $D_0 = (X, Y, Z)$ of size $n$ ($Z$ with $p$ columns),
             number of permutations $B$.
Let:
$CRIT_i(X, Y, Z) := (JMI(X, Y|Z))_{i=1}^n = \sum_{j=1}^{p} \log \frac{\hat{p}(x_i, y_i, z_{i,j})\hat{p}(z_i)}{(\hat{p}(y_i, z_{i,j})\hat{p}(x_i, z_{i,j})}$
Compute:
$JMI^{(0)} = \frac{1}{n}\sum_{i=1}^{n} CRIT_i(X, Y, Z)$
for $b = 1, \ldots, B$ do
    Randomly permute $X$ (on each strata on $Z$) to obtain permuted sample
      $D^{(b)} = (X^{(b)}, Y, Z)$
    Compute:
    $JMI^{(b)} = \frac{1}{n}\sum_{i=1}^{n} CRIT_i(X^{(b)}, Y, Z)$,
    $VAR^{(b)} = \frac{1}{n-1}\sum_{i=1}^{n}(CRIT_i(X^{(b)}, Y, Z) - JMI^{(b)})^2$
for $b = 1, \ldots, B$ do
    Randomly permute $X$ (on each strata on $Z$) and randomly choose [n/2]
      observations to obtain permuted sample $D_{1/2}^{(b)} = (X_{1/2}^{(b)}, Y_{1/2}, Z_{1/2})$
    Compute:
    $JMI_{1/2}^{(b)} = \frac{2}{n}\sum_{i=1}^{n/2} CRIT_i(X_{1/2}^{(b)}, Y_{1/2}, Z_{1/2})$,
    $VAR_{1/2}^{(b)} = \frac{1}{n/2-1}\sum_{i=1}^{n/2}(CRIT_i(X_{1/2}^{(b)}, Y_{1/2}, Z_{1/2}) - JMI_{1/2}^{(b)})^2$
Let:
$T(\cdot, \cdot)$ two sample t-test statistic
$p_T(\cdot, \cdot)$ p-value of the two sample t-test statistic
$F_{N(\hat{\mu}, \hat{\sigma})}(s)$ theoretical distribution function of $N(\hat{\mu}, \hat{\sigma})$
$F_{\chi^2_{\hat{\mu}}}(s)$ theoretical distribution function of $\chi^2_{\hat{\mu}}$
Compute:
$p_T := p_T(VAR^{(1:B)}, VAR_{1/2}^{(1:B)})$
$\hat{\mu} := \frac{1}{B}\sum_{b=1}^{B} JMI^{(b)}$
$\hat{\sigma}^2 := \frac{1}{B}\sum_{b=1}^{B} VAR^{(b)}$
if $p_T > 0.05$ or $\hat{\mu} \leq 0$ then
    $p = 1 - F_{N(\sqrt{n}\hat{\mu}, \hat{\sigma})}(\sqrt{n} JMI^{(0)})$
else
    $p = 1 - F_{\chi^2_{2n\hat{\mu}}}(2n JMI^{(0)})$
Output     : p-value $p$
```

stronger condition $X_1^{(1)} \perp (Y, X_1, \ldots, X_k)$ holds. We consider the performance of JMI based test for testing hypothesis $H_{01} : X_1^{(1)} \perp Y|X_1, \ldots, X_k$ when the sample size and parameters of the model vary. As H_{01} is satisfied in both models this contributes to the analysis of the size of the test.

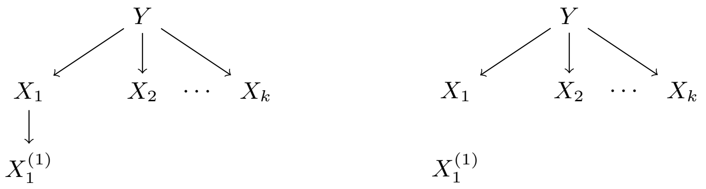

Fig. 3. Models under consideration in an experiment I. The models in the left and right panel will be called M1 and M2.

Observations in M1 are generated as follows: first, Y is chosen from Bernoulli distribution with success probability $P(Y = 1) = 0.5$. Then $(Z_1, \ldots, Z_k)$ are generated from $N(0, \Sigma)$ given $Y = 0$ and $N(\gamma, \mathbf{\Sigma})$ given $Y = 1$, where elements of $\mathbf{\Sigma}$ are equal $\sigma_{ij} = \rho^{|i-j|}$ and $\gamma = (1, \gamma, \ldots, \gamma^{k-1})^T$ with $0 \le \rho < 1$ and $0 < \gamma \le 1$ some chosen values. Then Z values are discretised to two values (0 and 1) to obtain $X_1, \ldots X_k$. In the next step $Z_1^{(1)}$ is generated from conditional distribution $N(X_1, 1)$ given X_1 and then $Z_1^{(1)}$ is discretised to $X_1^{(1)}$.We note that such method of generation yields that $Z_1^{(1)}$ and Y are conditionally independent given X_1 and the same is true for $X_1^{(1)}$. Observations in M2 are generated similarly, the only difference being that $Z_1^{(1)}$ is now generated independently of $(Y, X_1, \ldots, X_k)$.

We will also check the power of the tests in M1 for testing hypotheses $H_{02} : X_1^{(1)} \perp Y|X_2, \ldots, X_k$ and $H_{03} : X_1 \perp Y|X_2, \ldots, X_k$ as neither of them is satisfied in M1. Note however, that since information $I(X_1^{(1)}, Y|X_2, \ldots, X_k)$ and $I(X_1, Y|X_2, \ldots, X_k)$ decreases when k (or γ, ρ) increases the task becomes more challenging for larger k (or γ, ρ, respectively) which will result in a loss of power for large k when sample size is fixed.
Estimated tests sizes and powers are based on $N = 200$ times repeated simulations.

We first check how the switch behaves for JMI test while testing H_{01} (see Figure 4). In M1 for $k = 1$ as $X_1^{(1)} \perp Y$ given X_1 and $JMI = I(X_1^{(1)}, Y|X_1) = 0$ asymptotic distribution is d.q.f. and we expect switching to d.q.f. which indeed happens in almost 100%. For $k \ge 2$, $JMI \ne 0$ asymptotic distribution is normal which is reflected by the fact that the normal distribution is chosen with large probability. Note that this probability increases with n as summands $\hat{I}(X_1^{(1)}, Y|X_i)$ of $\widehat{JMI}$ for $i \ge 2$ converge to normal distributions due to Central Limit Theorem. The situation is even more clear-cut for M2 where $JMI = 0$ for all k and the switch should choose d.q.f.
Figure 5 shows the empirical sizes of the test when theoretical size has been fixed at $\alpha = 0.05$ and $\rho = 0$ and $\gamma = 1$. We see that empirical size is controlled fairly well for CMI(sp) and for the proposed methods, with the switch (norm/chi_scale) working better than the switch (norm/chi). A superiority of the former is even more pronounced for $0 < \gamma < 1$ and when $X_1, \ldots, X_k$ are dependent (not shown). Note erratic behaviour of size for CMI, which significantly exceeds 0.1 for certain k and then drops to 0. Figures 6 and 7 show the power of the considered meth-

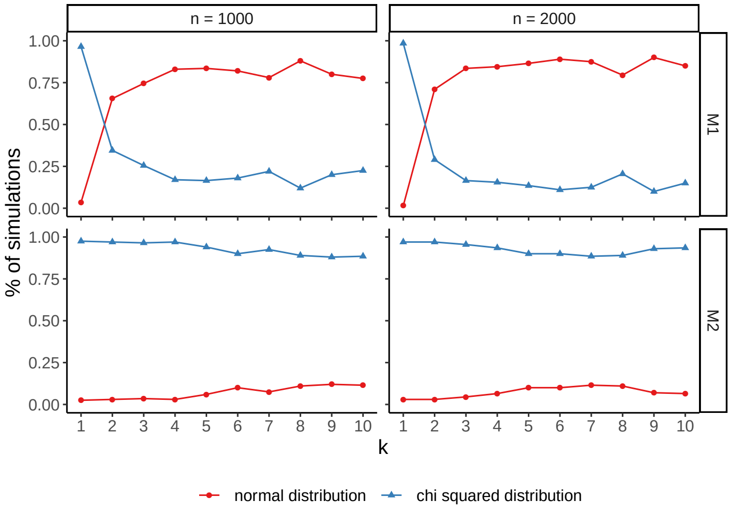

Fig. 4. The behaviour of the switch for testing H_{01} in M1 and M2 models ($\rho = 0$, $\gamma = 1$, $n = 1000$).

ods for hypotheses H_{02} and H_{03}. It is seen that for $\gamma = 1$, $\rho = 0$ the expected decrease of power with respect to k is much more moderate for the proposed methods than for CMI and CMI(sp). JMI(norm/chi_scale) works in most cases slightly better than JMI(norm/chi). For H_{03} power of CMI(sp) is similar to that of CMI but it exceeds it for large k, however, it is significantly smaller than the power of both proposed methods. For H_{03} superiority of JMI-based tests is visible only for large k when n is moderate ($n = 500, 1000$), whereas for H_{02} it is also evident for small k. With changing ρ and γ superiority of the proposed methods is still evident (see Figure 7). Note that for fixed γ the power of all methods decreases when ρ increases.

5 APPLICATION TO FEATURE SELECTION

Finally, we illustrate how the proposed test can be applied for Markov Blanket (MB, see e.g. [10]) discovery of Bayesian Networks (BN). MB for a target Y is defined as the minimal set of predictors given which Y and remaining predictors are conditionally independent ([2]). We have used the JMI test (with normal/scaled chi square switch) in the Grow and Shrink (GS, see e.g. [8]) algorithm for MB discovery and compared it with GS using CMI and CMi(sp). GS algorithm finds a large set of potentially active features in the Grow phase and then whittles it down in the Shrink phase. In the real data experiments we used another estimator of σ^2 equal to the empirical variance of JMIs calculated for permuted samples which behaved more robustly. The results were evaluated by F measure (the harmonic mean of precision and recall). We have considered several

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50436-6_51

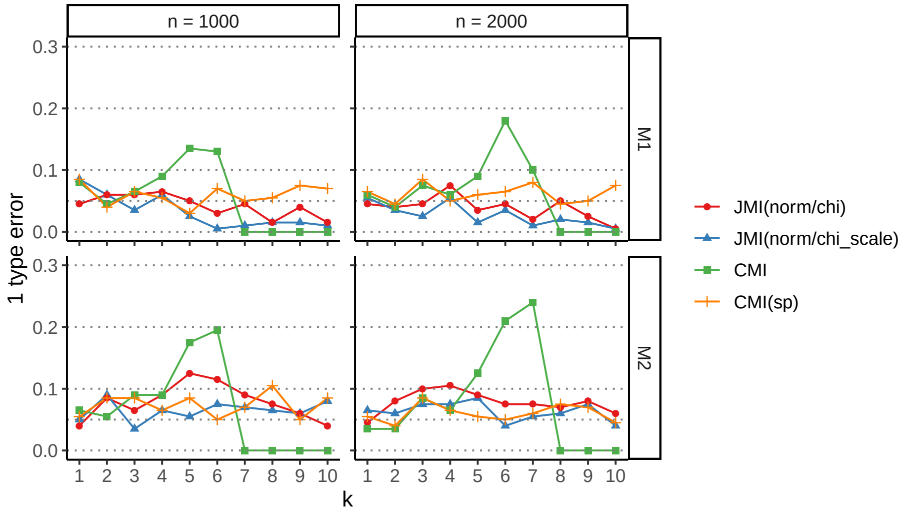

Fig. 5. Test sizes for testing H_{01} in M1 and M2 models ($\rho = 0$, $\gamma = 1$) for fixed $\alpha = 0.05$.

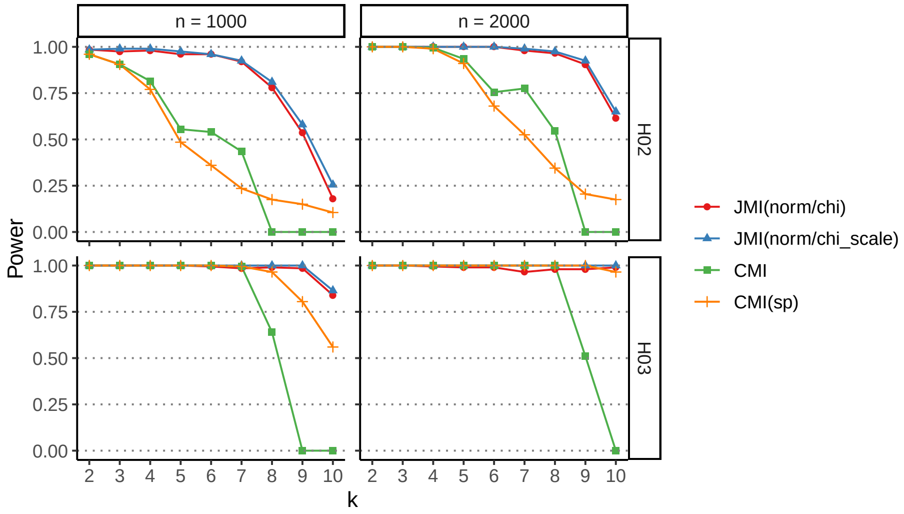

Fig. 6. Power for testing H_{02} and H_{03} in M1 model ($\rho = 0$, $\gamma = 1$).

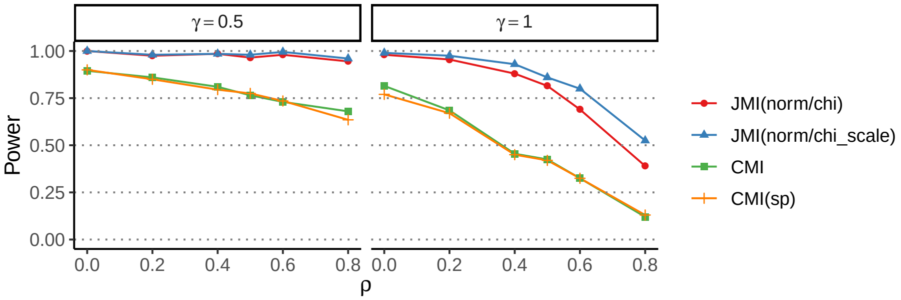

Fig. 7. Power for testing H_{02} in M1 model ($n = 1000$, $k = 4$).

benchmark BNs from BN repository `htttp://www.bnlearn.com/bnrepository` (asia, cancer, child, earthquake, sachs, survey). For each of them Y has been chosen as the variable having the largest MB. The results are given in Table 5. It is seen that with respect to F in the majority of cases GS-JMI method is the winner and ties with one of the other methods and the more detailed analysis indicates that this is due to its largest recall in comparison with GS-CMI and GS-CMI(sp) (see supplement [5]). This agrees with our initial motivation of considering such method which was the lack of power (i.e. missing important variables) by CMI-based tests.

Table 1. Values of F measure for GS algorithm using JMI, CMI and CMIsp tests. The winner is in bold.

Dataset	Y	MB size	JMI	CMI(sp)	CMI
asia	either	5	**0.58**	0.57	**0.58**
cancer	Cancer	4	**0.78**	0.65	0.56
child	Disease	8	0.55	**0.74**	0.55
earthquake	Alarm	4	**0.87**	**0.87**	0.76
sachs	PKA	7	0.83	**0.88**	0.59
survey	E	4	**0.81**	0.52	0.54

6 CONCLUSIONS

We have proposed a new test of conditional independence based on approximation JMI of the conditional mutual information CMI and its asymptotic distributions. We have shown using synthetic data that the introduced test is more powerful than tests based on asymptotic or permutation distributions of CMI when a conditioning set is large. In our analysis of real data sets we have indicated that the proposed test used in GS algorithm yields promising results in MB discovery problem. Drawback of such a test is that it disregards interactions between predictors and target variables of order higher than 3. Further research topics include systematic study of $\hat{J}^{\beta,\gamma}$ and especially how its parameters influence the power of the associated tests and feature selection procedures. Moreover, studying tests based on extended JMI including higher order terms is worthwhile.

References

1. Battiti, R.: Using Mutual Information for Selecting Features in Supervised Neural-Net Learning. IEEE Transactions on Neural Networks **5**(4), 537–550 (1994)
2. 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)
3. Cover, T.M., Thomas, J.A.: Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing). Wiley-Interscience (2006)
4. Guyon, I., Elyseeff, A.: An introduction to feature selection. Feature Extraction, Foundations and Applications **207**(Studies in Fuzziness and Soft Computing), 1–25 (2006)
5. Kubkowski, M., Łazęcka, M., Mielniczuk, J.: Distributions of a general reduced-order dependence measure and conditional independence testing: supplemental material (2020), github.com/lazeckam/JMI_CondIndTest
6. Kubkowski, M., Mielniczuk, J., Teisseyre, P.: How to gain on power: novel conditional independence tests based on short expansion of conditional mutual information (2019), submitted
7. Lin, D., Tang, X.: Conditional infomax learning: An integrated framework for feature extraction and fusion. In: Proceedings of the 9th European Conference on Computer Vision - Volume Part I. pp. 68–82. ECCV'06 (2006)
8. Margaritis, D., Thrun, S.: Bayesian network induction via local neighborhoods. In: Proceedings of the 12th International Conference on Neural Information Processing Systems. pp. 505–511. NIPS'99 (1999)
9. McGill, W.J.: Multivariate information transmission. Psychometrika **19**(2), 97–116 (1954)
10. Pena, J.M., Nilsson, R., Bjoerkegren, J., Tegner, J.: Towards scalable and data efficient learning of markov boundaries. International Journal of Approximate Reasoning **45**(2), 211 – 232 (2007)
11. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(1), 1226–1238 (2005)
12. Tsamardinos, I., Borboudakis, G.: Permutation testing improves on Bayesian network learning. In: Proceedings of ECML PKDD 2010. pp. 322–337 (2010)
13. Vergara, J., Estevez, P.: A review of feature selection methods based on mutual information. Neural Computing and and Applications **24**(1), 175–16 (2014)
14. Yang, H.H., Moody, J.: Data visualization and feature selection: new algorithms for nongaussian data. Advances in Neural Information Processing Systems **12**, 687–693 (1999)
15. Zhang, J.T.: Approximate and asymptotic distributions of chi-squared type mixtures with applications. Journal of American Statistical Association **100**, 273–285 (2005)