Journal of Machine Learning Research 22 (2021) 1-57 Submitted 7/19; Revised 2/21; Published 2/21

How to Gain on Power: Novel Conditional Independence Tests Based on Short Expansion of Conditional Mutual Information

Mariusz Kubkowski MARIUSZ.KUBKOWSKI@IPIPAN.WAW.PL
Institute of Computer Science
Polish Academy of Sciences
Warsaw, Jana Kazimierza 5, 01-248, Poland
and
Faculty of Mathematics and Information Science
Warsaw University of Technology
Warsaw, Koszykowa 75, 00-662 Poland

Jan Mielniczuk JAN.MIELNICZUK@IPIPAN.WAW.PL
Institute of Computer Science
Polish Academy of Sciences
Warsaw, Jana Kazimierza 5, 01-248, Poland
and
Faculty of Mathematics and Information Science
Warsaw University of Technology
Warsaw, Koszykowa 75, 00-662 Poland

Paweł Teisseyre PAWEL.TEISSEYRE@IPIPAN.WAW.PL
Institute of Computer Science
Polish Academy of Sciences
Warsaw, Jana Kazimierza 5, 01-248, Poland
and
Faculty of Mathematics and Information Science
Warsaw University of Technology
Warsaw, Koszykowa 75, 00-662 Poland

Editor: Peter Spirtes

Abstract

Conditional independence tests play a crucial role in many machine learning procedures such as feature selection, causal discovery, and structure learning of dependence networks. They are used in most of the existing algorithms for Markov Blanket discovery such as Grow-Shrink or Incremental Association Markov Blanket. One of the most frequently used tests for categorical variables is based on the conditional mutual information (CMI) and its asymptotic distribution. However, it is known that the power of such test dramatically decreases when the size of the conditioning set grows, i.e. the test fails to detect true significant variables, when the set of already selected variables is large. To overcome this drawback for discrete data, we propose to replace the conditional mutual information by Short Expansion of Conditional Mutual Information (called $SECMI$), obtained by truncating the Möbius representation of CMI. We prove that the distribution of $SECMI$ converges to either a normal distribution or to a distribution of some quadratic form in

©2021 Mariusz Kubkowski, Jan Mielniczuk and Paweł Teisseyre.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v22/19-600.html.

normal random variables. This property is crucial for the construction of a novel test of conditional independence which uses one of these distributions, chosen in a data dependent way, as a reference under the null hypothesis. The proposed methods have significantly larger power for discrete data than the standard asymptotic tests of conditional independence based on CMI while retaining control of the probability of type I error.

Keywords: conditional independence tests, mutual information, Möbius representation, feature selection, information-based selection criteria, interaction information

1. Introduction

In the paper we focus on testing conditional independence when a conditioning set is large. Conditional independence tests are ubiquitous in many problems of machine learning such as feature selection (see e.g. Guyon and Elyseeff 2006), causal discovery (see e.g. Spirtes et al. 2000), and structure learning of dependence networks (see e.g. Yu et al. 2018).

Such tests are used to decide whether to include a candidate variable into the set of active variables. As an example consider the problem of discovering a Markov Blanket (MB) of the target variable Y which is defined as the minimal set of variables among all considered, given which all other variables are independent of Y. MB is a key concept in structure learning of Bayesian (cf. Tsamardinos et al. 2003) and Markov networks (Bromberg et al., 2009; Schlüter, 2012) as well as in the feature selection (Brown et al., 2012). It was proposed by Pearl to identify MB of a class variable within the framework of Bayesian networks (Pearl, 1988) and was then transferred to non-causal case where it serves as the feature selection tool without establishing causal relationships (Koller and Sahami, 1995). A large class of such methods (the so-called constraint-based methods) use MB discovery as a step in learning the Bayesian network structure (see e.g. Pellet and Eliseeff 2008). Representative examples are: Grow-Shrink (GS) (Margaritis and Thrun, 1999), Incremental Association Markov Blanket (IAMB) and its variants (Tsamardinos et al., 2003) and HITON (Aliferis et al., 2003; Pena et al., 2007) among others.

Many authors focus on the optimality of proposed algorithms (i.e. algorithms which find a MB for any variable considered as a target) and in order to show that the logic of the algorithm is correct, assume that conditional tests are perfect, i.e. they establish with certainty whether a candidate variable is conditionally independent or not of the target given already chosen variables (see e.g. Pena et al. 2007 and Gao and Qiang 2017). Moreover, a putative reasoning is that even if an error is committed with a certain probability, given finite number of candidates, the overall error of MB discovery should not be too large. However, this argument may fail when either the number of variables is large or the error of deciding whether a candidate variable is conditionally independent from the target is large, and especially when both situations simultaneously occur. We argue that such situations may occur for large scale networks (i.e. those having many nodes) and small sample sizes. This concerns in particular a popular conditional mutual information (CMI) test based on an asymptotic distribution of sample CMI, which is approximately chi-square under the null hypothesis of conditional independence (see Theorem 1 below for specification of a number of degrees of freedom of the chi-square distribution).

We note in passing that, besides MB discovery, CMI-based inference for two features given the class attribute plays an important role in interaction detection when its value, if larger than the unconditional mutual information, indicates positive predictive interaction (see e.g. Mielniczuk and Teisseyre 2018).

However, it is widely known, although a systematic quantitative study is still missing, that the power of tests based on sample CMI dramatically decreases when the size of the conditioning set grows, i.e. the tests fail to detect the true significant variables, when the set of already selected variables is large. Consequently, the popular MB discovery algorithms may overlook the significant variables, which in turn leads to an inaccurate prediction of the network structure. To alleviate the drawback of the asymptotic test for CMI, some modifications have been proposed. For example, Tsamardinos and Borboudakis (2010) consider a test for CMI for which parameters of a chi-square reference distribution are calculated based on the permutation scheme. More specifically, when conditional independence of X and Y given Z is tested, the values of X are permuted on strata $Z = z$, thus keeping conditional distribution of Y given Z unchanged. A similar permutation scheme for CMI was used in Leppä-Aho et al. (2018) in the context of learning Markov network structures for continuous data and in Runge (2018) who tested conditional independence using a nearest-neighbour estimator of CMI.

We focus here on detecting conditional dependence using non-parametric information-based methods i.e. we do not make any assumptions on the distributions of (X, Z, Y) besides that X, Y and Z are discrete. In the case when the distribution of (X, Z) is unknown and one can sample from it, an approach based on the so-called knockoffs has been recently developed (see Barber and Candès 2015 and Candès et al. 2018). Here we assume that the sole information about the distribution of (X, Z, Y) is an iid sample (X_i, Z_i, Y_i) consisting of n observations. Let us mention in this context that the most powerful test for conditional independence against all alternatives does not exist when some coordinates of Z are continuous random variables (Shah and Peters, 2020). Even when Z is a discrete random variable, testing conditional independence is more difficult problem than its unconditional version as it involves testing independence of X and Y on each strata $Z = z$.

In this work, we discuss drawbacks of existing CMI-based procedures when the conditioning set consists of a large number of variables and in the view of them we propose a novel test procedure based on a sample analogue of $SECMI$ (Short Expansion of Conditional Mutual Information), obtained from Möbius representation of CMI by truncation. Möbius expansion allows one to write down CMI as a sum of low-dimensional terms pertaining to interactions between variables. The $SECMI$ is obtained by retaining the first two terms of the expansion, corresponding to the main effects and the second order interactions, and replacing them with their sample analogues, while omitting higher-order interactions. In this way we avoid the problem of calculating the sample mutual information for cells with small number of observations, which is the case for CMI.

Our aim here is to propose new test procedures based on the $SECMI$. To this end we first prove that the asymptotic distribution of $SECMI$ turns out to be either normal or coincides with distribution of a certain quadratic form of normal variable, according to dependence structure among variables. Additionally, for the case when Markov Blanket for binary Y is sought, we fully characterize the situation when a distribution of sample $SECMI$ is not normal. Moreover, it is empirically confirmed that, under null hypothesis, the distribution of a quadratic form is close to the chi square distribution. In contrast to CMI, we face a problem of choosing an appropriate reference distribution for $SECMI$-based conditional independence testing. In order to account for the aforementioned dichotomy of the asymptotic limit, we develop a data-driven choice of the reference distribution based on which the p-value is calculated. As the reference distributions depend on unknown parameters (the mean and the variance in the case of the normal distribution and the number of degrees of freedom in the case of the chi square distribution) they have to be estimated. We propose to estimate them using the permutation scheme. Since we only estimate parameters of these distributions and not the whole distribution we avoid prohibitive number of permutations to approximate well its quantiles. It follows from our experiments that the proposed testing procedures are well calibrated, i.e. the actual significance levels match the assumed ones. Moreover, our methods outperform CMI-based tests (the asymptotic one and the permutation based test) with respect to the power. We also combine our tests with MB discovery algorithms which results in significantly larger recall for various networks.

This paper is organized as follows. In Section 2 we discuss notation and recall basic definitions that are needed to define the proposed test statistic. Section 3 contains theoretical results on asymptotic distributions of CMI with self contained proofs as well as bounds on the sample sizes needed to ensure the desired power of the test. In Section 4 we introduce a novel statistic $SECMI$, discuss its properties and give theoretical results on the distribution of $SECMI$. In Section 5 we describe novel semi-parametric tests based on $SECMI$. Results of the experiments are described in Section 6, Section 7 concludes the paper.

2. Preliminaries

First we recall definitions of basic quantities considered in Information Theory, which will be used in the next sections. Throughout we consider univariate or multivariate nominal random variables having a finite number of discrete values. The mutual information (MI) between X and Y is defined as (log stands for logarithm to the base 2)

$$I(Y,X) = \sum_{x,y} P(X=x,Y=y)\log\frac{P(X=x,Y=y)}{P(X=x)P(Y=y)} = H(Y) - H(Y|X), \tag{1}$$

where $H(Y) = -\sum_y P(Y=y)\log P(Y=y)$ and $H(Y|X) = \sum_x P(X=x)H(Y|X=x)$ are the entropy and the conditional entropy, respectively. MI evaluates how similar the joint distribution is to the product of marginal distributions and thus can be considered a measure of strength of dependence between X and Y. It is symmetric, non-negative and is equal zero if and only if X and Y are independent. It is also easily seen that

$$I(Y,X) = H(Y) + H(X) - H(Y,X). \tag{2}$$

The conditional mutual information (CMI)

$$I(Y,X|Z) = \sum_{x} P(Z=z)\sum_{x,y} P(X=x,Y=y|Z=z)\log\frac{P(X=x,Y=y|Z=z)}{P(X=x|Z=z)P(Y=y|Z=z)} \tag{3}$$

measures the strength of the conditional dependence between X and Y given Z. Note that the conditional mutual information is the mutual information of X and Y given $Z=z$

averaged over the values of Z. It is equal zero if and only if X and Y are conditionally independent given Z. The conditional independence of X and Y given Z is denoted by $X \perp Y|Z$. For more properties of the basic measures described above we refer to Cover and Thomas (2006) and Yeung (2002). The above definitions can be naturally extended to the case of random vectors (i.e. when X, Y and Z are multivariate random variables) by using a multivariate mass function instead of a univariate one. It is easily seen that

$$I((X,Z),Y) = I(Z,Y) + I(X,Y|Z). \tag{4}$$

Another important quantity, used in the next sections, is the interaction information II (McGill, 1954). The 3-way interaction information is defined as

$$II(X,Y,Z) = -H(X) - H(Y) - H(Z) + H(X,Y) + H(X,Z) + H(Y,Z) - H(X,Y,Z). \tag{5}$$

It can be easily proved that II can be also written as

$$II(X,Y,Z) = I((X,Z),Y) - I(X,Y) - I(Z,Y), \tag{6}$$

and thus II can be interpreted as the part of the mutual information of (X,Z) and Y which is due solely to the interaction between X and Z in predicting Y i.e. the part of $I((X,Z),Y)$ which remains after subtraction of the amount of individual informations between Y and X and that of Y and Z. In other words, II is obtained by removing the main effects from the term $I((X,Z),Y)$ which describes the overall dependence between Y and the pair (X,Z). Interaction information can be also written as

$$II(X,Y,Z) = I(Y,X|Z) - I(Y,X), \tag{7}$$

which is consistent with an intuitive meaning of existence of an interaction as the situation in which the effect of a one variable on the class variable Y depends on a value of another variable. The positive value of II indicates the existence of complementarity, e.g. for $Y = XOR(X,Z)$, being indicator function of the event $\{X \neq Z\}$ we have $II(X,Y,Z) = \log(2) > 0$, when all three variables are binary and take their values with probability 1/2. On the other hand, the negative value of II indicates redundancy, e.g. for $Y = X = Z$, we have $II(X,Y,Z) = -\log(2) < 0$. The 3-way II can be extended to the general case of m variables. The m-way interaction information (Ting, 1960; Han, 1980) is

$$II(Z_1,\ldots,Z_m) = -\sum_{T \subseteq \{1,\ldots,m\}} (-1)^{m-|T|} H(Z_T), \tag{8}$$

where Z_T is the subvector of Z with indices in T and $|T|$ is a number of elements of T. For $m = 1$, equality (8) reduces to $II(Z_1) = -H(Z_1)$, in the case of $m = 2$ it yields mutual information, whereas for $m = 3$ it reduces to (5). It turns out that the conditional mutual information can be represented as a sum of interaction informations, see Section 4.1. It seems intuitive and helpful that the strength of dependence between multivariate Z and the class variable Y is decomposed into parts corresponding to interactions between subvectors of Z of dimensionality r and Y for $r = 1, 2, \ldots, m$. Using this decomposition we define a novel test statistic, called $SECMI$, by truncating the decomposition and retaining interaction informations of Y with individual Z_is and with pairs (Z_i, Z_j).

3. Testing Conditional Independence Using CMI

3.1 Asymptotic Distribution of CMI

A frequently applied test of the conditional independence $X \perp Y|Z$, where $Z = (Z_1, \ldots, Z_m)$, uses sample conditional mutual information $\hat{I}(X, Y|Z)$ as a test statistic obtained by plugging in frequencies for probabilities in (3). It is based on the useful fact that under conditional independence $X \perp Y|Z$ its asymptotic distribution does not depend on that of (X, Y, Z) and is approximately chi square with the known number of degrees of freedom (see Theorem 1 below). On the other hand, when the conditional independence does not hold, the limiting distribution of $\hat{I}(X, Y|Z) - I(X, Y|Z)$ is normal with the variance depending on the underlying probability distribution. We stress that speeds of convergence of $\hat{I}(X, Y|Z)$ to $I(X, Y|Z)$ are different in both cases: they equal n^{-1} in the first case and $n^{-1/2}$ in the second. The test based on CMI is a popular tool in dependence analysis, in particular for Markov Blanket discovery. It comes under different guises and names among which G^2 test is the most popular (see e.g. Agresti 2002). X^2 denotes the second order approximation of CMI which turns out to be the conditional chi square test and has the same asymptotic distribution as $\hat{I}(X, Y|Z)$ discussed below (this can be shown similarly to the unconditional case, see ibidem). It is frequently noted that CMI test lacks power when the average number of observations per cell is too small (e.g. smaller than 5) but these statements are rarely supported by quantitative analysis. Actually, the lack of power of CMI can be quite dramatic. It has important consequences for performance of Markov Blanket discovery procedures e.g. GS or IAMB - such as a poor detection of influential variables and consequently low Recall. We support this claim, which is exemplified by analysis of behaviour of CMI-based test in synthetic models of Section 6, by straightforward derivation of a behaviour of the test statistic on the alternative using normal approximation (see Theorem 1 (i)) which turns out to yield surprisingly good approximation of the power of the test.

We now introduce some notations. Let $p(x, y, z) := P(X = x, Y = y, Z = z)$ and analogously $p(x, y) := P(X = x, Y = y)$, $p(x) := P(X = x)$, etc. We assume throughout that $p(x, y, z) > 0$ for any x, y, z. Moreover, we let X, Y, Z take I, J, K possible values, respectively. Note that, since Z is a m-dimensional vector, K is the number of all possible combinations of values of its coordinates. For example if all Z_is take b possible values, then $K = b^m$. We assume throughout that the estimation of $I(X, Y|Z)$ is based on n iid samples from the distribution of (X, Z, Y). The following known result is frequently used in dependence analysis (compare Kullback 1978 and, e.g.,Tsamardinos and Borboudakis 2010 p. 325; see also Shao 2003 for the statement of the result).

Theorem 1 *(i) Assume that $I(X, Y|Z) \neq 0$. Then we have*

$$n^{1/2}(\widehat{I}(X, Y|Z) - I(X, Y|Z)) \xrightarrow{d} N(0, \sigma^2_{CMI}), \tag{9}$$

where

$$\sigma^2_{CMI} = \sum_{x,y,z} p(x, y, z) \log^2 \frac{p(x, y, z)p(z)}{p(x, z)p(y, z)} - I^2(X, Y|Z) = \mathrm{Var}\left(\log \frac{p(X, Y, Z)p(Z)}{p(X, Z)p(Y, Z)}\right)$$

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

$$2n\widehat{I}(X,Y|Z) \xrightarrow{d} \chi^2_d, \tag{10}$$

where $d = (I-1)(J-1)K$.

A closely related result to Theorem 1 (ii), goes back to Fisher (see Fisher 1922) who considered chi square statistic in place of mutual information. We include a proof of Theorem 1 in the Appendix as, surprisingly, it is not easy to find its complete and self-contained version. The reason is that the usual way to justify it is rather circuitous. It relies on representing conditional independence hypothesis

$$p(x,y|z) = \frac{p(x,y,z)}{p(z)} = p(x|z)p(y|z) = \frac{p(x,z)p(y,z)}{p^2(z)}$$

as a parametric hypothesis whose parameters are given by the conditional marginal probabilities $p(x|z)$ and $p(y|z)$. Then CMI is derived as the Likelihood Ratio Statistic (LRT) and asymptotic results on LRT are used to find its limit (see Shao 2003, p. 438). In the appendix we give a self-contained proof of this result based on the delta method (see Agresti 2002, Section 14.2.1) in the course of which we develop a groundwork for proving our main result, Theorem 2, which can thus be viewed as a generalization of Theorem 1. We note that scaling factor $n^{1/2}$ in (i) needs to be replaced by a larger scaling factor n in (ii) in order to obtain non-degenerate asymptotic law of the centered $\widehat{I}(X,Y|Z)$. We also note that this result is an analogue of the result for the estimator of unconditional information $I(X,Y)$ which states that its law is asymptotically normal provided $I(X,Y) \neq 0$ and chi square distributed in the opposite case (compare Agresti 2002). In the first case, the variance of $\hat{I}(X,Y)$ equals $\mathrm{Var}(\log p(X,Y)/p(X)p(Y))$ and in the second case the number of degrees of freedom is $(I-1)(J-1)$.

Let us note that LRT tests under appropriate assumptions are uniformly most powerful tests for testing the conditional independence against the lack of it (see e.g. Agresti 2002). However, we show in the following sections that when the true dependence structure does not exhibit higher order interactions or if they are negligible, the proposed $SECMI$ test is frequently more powerful that the CMI based test.

3.2 Lack of Power of CMI

In this section we address the problem of the lack of power of CMI, that is its poor performance at distinguishing departures from the conditional independence of X and Y given Z unless the sample size is very large. We show here that this is consistent with Theorem 3 (i). First, we report on a very good correspondence between the power calculated in Monte Carlo experiments for synthetic data and the theoretical power derived below. To show this fact, we performed an experiment on synthetic data, generated as follows.

Synthetic data D0. We draw $Y \in \{1,2\}$ from the Bernoulli distribution with the success probability $P(Y=1) = P(Y=2) = 0.5$. Then we generate $Z_1,\ldots,Z_m \in \{1,2,3\}$ with the marginal distribution $P(Z_i = 1) = 0.25$, $P(Z_i = 2) = 0.5$, $P(Z_i = 3) = 0.25$ and such that the joint distribution (Y, Z_i) is described by a discretized normal copula, with covariance matrix $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, where γ is a parameter. Finally, we generate variable $X \in \{1, 2\}$ from the Bernoulli distribution with success probability $P(X = 1|Z_1, \ldots, Z_m, Y) = \sigma(-\sum_{i=1}^{m}(Z_i + Y))$, where $\sigma(s) = (1 + \exp(-s))^{-1}$ is a logistic function. We consider the null hypothesis $Y \perp X|Z_1, \ldots, Z_m$.

Figure 1 shows how the theoretical and the empirical power depend on n, for $m = 10$, $\gamma = 1$ and $\alpha = 0.05$ for synthetic data D0. We observe a very good agreement between curves corresponding to the theoretical and the empirical power.

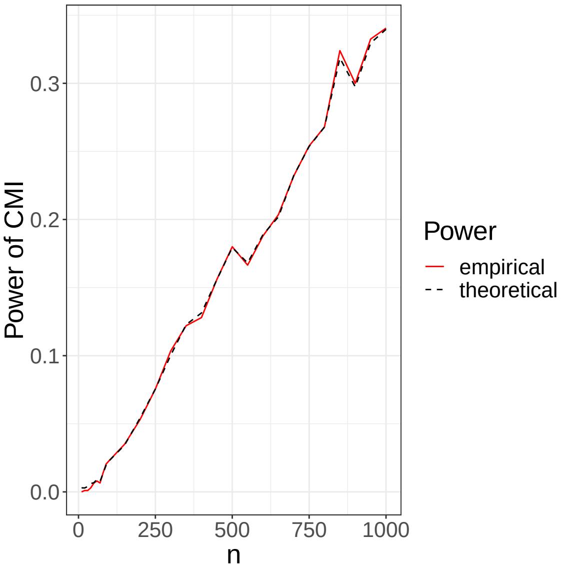


Figure 1: Theoretical and empirical power of CMI, for $m = 10$, $\gamma = 1$ and $\alpha = 0.05$ for synthetic data D0 described in Section 3.2.

Now we calculate theoretically the sample size necessary for the power to be at least β when the test is performed at significance level α for hypothesis $H_0 : I(X, Y|Z) = 0$. The rejection region is $\mathcal{C} = \{\hat{I}(X, Y|Z) \geq t_\alpha)$, where t_α is the critical value corresponding to level α calculated from (10). First, in order to control a significance level at α, we note that, in view of Theorem 1 (ii), we have

$$\alpha = P_{H_0}(\hat{I}(X, Y|Z) \geq t_\alpha) = P_{H_0}(2n\hat{I}(X, Y|Z) \geq 2nt_\alpha) \approx 1 - F_{\chi^2_d}(2nt_\alpha), \tag{11}$$

where $d = (I-1)(J-1)K$ and $F_{\chi^2_d}$ is a distribution function of χ^2_d with d degrees of freedom. Hence $t_\alpha \approx F^{-1}_{\chi^2_d}(1 - \alpha)/2n$. Standard calculations (see Appendix A) show that, in order to have the power of the test to be at least β, or equivalently $P(\mathcal{C}) \geq \beta$, sample size n should satisfy

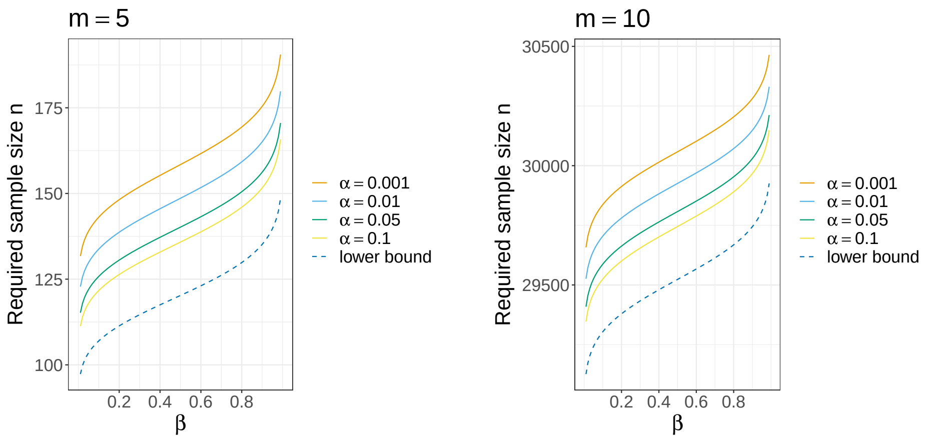
$$n \geq \left(\frac{\sqrt{4\sigma^2_{CMI} \left(\Phi^{-1}(1-\beta)\right)^2 + 8I(X, Y|Z)F^{-1}_{\chi^2_d}(1-\alpha)} - 2\sigma_{CMI}\Phi^{-1}(1-\beta)}{4I(X, Y|Z)} \right)^2. \tag{12}$$

If $\beta = 0.5$ which corresponds to the probability at least 0.5 of detecting conditional dependence, as $\Phi^{-1}(1/2) = 0$, this reduces to:

$$n \geq \frac{F^{-1}_{\chi^2_d}(1-\alpha)}{2I(X,Y|Z)}. \tag{13}$$

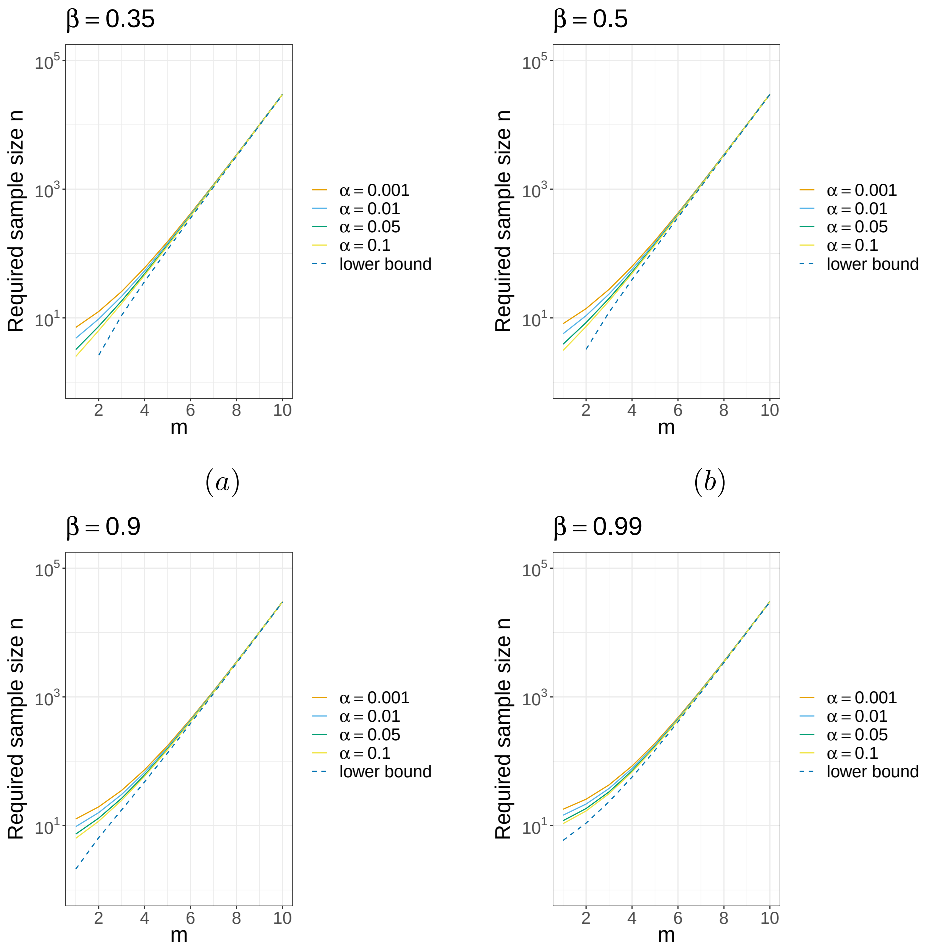
In order to visualize how the required sample size for CMI test depends on parameters m, β, α we consider a simple situation in which X and Y are binary and each $Z_1, \ldots, Z_m$ takes three possible values, and thus $d = 3^m$ (compare Theorem 1). In addition, we assume that $I(X,Y|Z) = 1$ and $\sigma^2_{CMI} = 1$. The lower bound for the required sample size is derived by bounding from below the right hand side of (13) by $d - 5/2$ for $\alpha \leq 0.17$ (easily obtainable from Inglot 2010, Proposition 5.1). Figure 2 shows how the required sample size increases with β for fixed m ($m = 5$ and $m = 10$). Note that for $m = 5$ and $\beta = 0.5$ (Figure 2 (a)) the required sample sizes oscillate around 150, whereas for $m = 10$ and $\beta = 0.5$ ((Figure 2 (b)) they oscillate around 30000. Figure 3 shows the required sample sizes to achieve the power $\beta = 0.35, 0.5, 0.9$ and 0.99 (logarithmic scale is used on the Y axis). It is clearly seen that the required sample size increases exponentially with m . The above analysis shows that we need a large amount of data to detect the conditional dependence among variables, even for a moderate size of the conditioning set, e.g. for $m = 10$ we need around 30000 observations in order to obtain the power of CMI-based test to be at least 0.5. Note also that for smaller α, when we try to control significance level of the procedure consisting of several tests using e.g. the Bonferroni correction, the required sample size will be considerably larger.

In the view of the discussed lack of power of the CMI-based test for moderate sample sizes and large conditioning sets it becomes necessary to look for its alternatives which do not suffer from this drawback. One possible solution based on the truncated Möbius expansion is discussed in the following section.



(a) (b)

Figure 2: Required sample size for asymptotic CMI test wrt to β for $m = 5$ (a) and $m = 10$ (b).



(a) (b)

(c) (d)

Figure 3: Required sample size for the power of asymptotic CMI test to be at least $\beta = 0.35, 0.5, 0.9, 0.99$ wrt to the size of conditioning set m.

4. Expansions of Conditional Mutual Information

In the view of the previous discussion we consider a different statistic than CMI in order to build a more powerful test of conditional dependence and construct the corresponding rejection region based on its approximate distribution. Its derivation is based on the so-called Möbius expansion of $I(X, Y|Z)$, where $Z = (Z_1, \ldots, Z_m)$. By truncating the Möbius expansion we reduce the sizes of conditioning sets for the summands which are the source of the lack of power of the CMI. First we give some preliminaries on the Möbius representation of the conditional mutual information.

4.1 The Möbius Representation of CMI

Let Z_T denote the subvector of $Z = (Z_1, \ldots, Z_m)$ for indices belonging to $T \subseteq \{1, \ldots, m\}$. In order to justify the Möbius expansion, we note that $II(Z_1, \ldots, Z_m)$ defined in (8) is a special case of the so-called difference operator $\Delta f(Z_S)$ for $f(Z_S) = -H(Z_S)$, where $\Delta f(Z_S)$ is defined by (Rota, 1964; Han, 1980)

$$\Delta f(Z_S) = \sum_{T \subseteq S} (-1)^{|S|-|T|} f(Z_T).$$

Then the following Möbius inversion formula holds

$$f(Z_S) = \sum_{T \subseteq S} \Delta f(Z_T), \tag{14}$$

(see Rota 1964, Corollary 1 and Principle of Inclusion-Exclusion). For $f(Z_S) = -H(Z_S)$ equality (14) yields

$$-H(Z_S) = \sum_{T \subseteq S} II(Z_T). \tag{15}$$

Observe that in the view of (2), we have $I(Z_S, Y) = H(Z_S) + H(Y) - H(Z_S, Y)$. Applying (15) for the first and the third term of the last equality with $S = \{1, \ldots, m\}$ and recalling that $II(Y) = -H(Y)$ we obtain the Möbius representation of $I(Z, Y)$

$$I(Z, Y) = \sum_{k=1}^{m} \sum_{\{t_1, \ldots, t_k\} \subseteq \{1, \ldots, m\}} II(Z_{t_1}, \ldots, Z_{t_k}, Y). \tag{16}$$

Thus, in the view of (4), we have

$$I(X, Y|Z) = I((Z, X), Y) - I(Z, Y) = I(X, Y) + \sum_{k=1}^{m} \sum_{\{t_1, \ldots, t_k\} \subseteq \{1, \ldots, m\}} II(X, Z_{t_1}, \ldots, Z_{t_k}, Y). \tag{17}$$

4.2 Short Expansion of Conditional Mutual Information ($SECMI$)

We define Short Expansion of Conditional Mutual Information ($SECMI$) as the truncated Möbius expansion (17) which incorporates the leading term $I(X, Y)$ and interactions of order 2 i.e. the terms $II(X, Z_k, Y)$. Let $\widehat{SECMI}$ be the sample version of $SECMI$. Thus we have in view of (7)

$$SECMI(X, Y|Z) = I(X, Y) + \sum_{k=1}^{m} II(X, Z_k, Y) = (1 - m)I(X, Y) + \sum_{k=1}^{m} I(X, Y|Z_k) \tag{18}$$

and

$$\widehat{SECMI}(X, Y|Z) = \hat{I}(X, Y) + \sum_{k=1}^{m} \widehat{II}(X, Z_k, Y) = (1 - m)\hat{I}(X, Y) + \sum_{k=1}^{m} \hat{I}(X, Y|Z_k), \tag{19}$$

where $\hat{I}(X,Y)$ and $\hat{I}(X,Y|Z)$ are the plug-in estimators of $I(X,Y)$ and $I(X,Y|Z)$, respectively and $\widehat{II}$ is defined using (7) and the plug-in estimators of MI and CMI. The asymptotic distributions of $\widehat{II}(X,Y,Z)$ when $X \perp (Y,Z)$ have been derived in Kubkowski and Mielniczuk (2020). The main advantage of $\widehat{SECMI}$ over CMI when used as a test statistic is that it involves conditioning on single variables only thus making approximation of its target value more precise than in the case of CMI. Note that, in view of (18), for $m=1$, $SECMI$ reduces to $I(X,Y|Z)$ and $\widehat{SECMI}$ reduces to $\hat{I}(X,Y|Z)$. It also follows from (17) that CMI reduces to $SECMI$ when high-order interactions are not present, i.e. $II(X,Z_{t_1},\ldots,Z_{t_k},Y)=0$ for all $k\geq 2$. The $SECMI$ can be naturally extended to take into account higher-order interactions by including more than the two first terms in (17), see Sections 4.4 and 6.4 for examples and discussion. We note that when $X \perp (Y,Z)$ then it follows that the distribution of $\hat{I}(X,Y)$ is asymptotically chi square, and the same is true for $\hat{I}(X,Y|Z_k)$ (see Theorem 1). However, even in this special case determination of the distribution of $\widehat{SECMI}(X,Y|Z)$ in (19) does not follow from these two facts as the summands in (19) are dependent.

The $SECMI$ criterion introduced here is identical with $CIFE$ (Conditional Infomax Feature Selection) criterion used in variable selection, see Lin and Tang (2006); Brown et al. (2012). Here, our aim is to consider $\widehat{SECMI}$ in a broader context as a test statistic for testing the conditional independence and to construct a corresponding test and study its actual level of significance and power. Then established properties of $\widehat{SECMI}$ can be used in particular in variable selection (see Section 5.2).

4.3 Properties of $SECMI$

We discuss now some properties of $\widehat{SECMI}$ and show that it exhibits the dichotomous behaviour under the conditional independence hypothesis, depending on which of the two possible scenarios is valid. The asymptotic distributions appearing in Theorem 2 below will serve as the reference distributions in the introduced $SECMI$-based test. Define first the following random variables for $i=1,\ldots,m$

$$W_i = \log\left(\frac{p(X,Y,Z_i)p(Z_i)p(X)p(Y)}{p(Z_i,X)p(Z_i,Y)p(X,Y)}\right), \qquad U_i = \log\left(\frac{p(Z_i,X,Y)p(Z_i)}{p(Z_i,X)p(Z_i,Y)}\right) \tag{20}$$

and

$$U_0 = (1-m)\log\left(\frac{p(X,Y)}{p(X)p(Y)}\right). \tag{21}$$

Let $\mathbb{E}U$ denote the expected value of random variable U. Observe that $SECMI$ can be written as

$$SECMI = \sum_{k=0}^{m}\mathbb{E}U_k = (1-m)^{-1}\mathbb{E}U_0 + \sum_{k=1}^{m}\mathbb{E}W_k. \tag{22}$$

This follows from

$$\mathbb{E}U_0 = (1-m)I(X,Y),\ \mathbb{E}U_i = I(X,Y|Z_i),\ \mathbb{E}W_i = II(X,Z_i,Y),$$

for $i = 1, \ldots, m$, since for example

$$I(X, Y|Z_i) = \sum_{x,y,z_i} p(x, y, z_i) \log \Big(\frac{p(x, y, z_i)p(z_i)}{p(z_i, x)p(z_i, y)} \Big)$$

and

$$II(X, Y, Z_i) = \sum_{x,y,z_i} p(x, y, z_i) \log \Big(\frac{p(x, y, z_i)p(z_i)p(x)p(y)}{p(z_i, x)p(z_i, y)p(x, y)} \Big)$$

(see equation (7)).
Let

$$\sigma^2_{\widehat{SECMI}} = \mathrm{Var}\Bigg(\log \Bigg[\Big(\frac{p(X, Y)}{p(X)p(Y)} \Big)^{1-m} \prod_{i=1}^{m} \Bigg(\frac{p(Z_i, X, Y)p(Z_i)}{p(Z_i, X)p(Z_i, Y)} \Bigg) \Bigg] \Bigg) = \mathrm{Var}\Big(\sum_{i=0}^{m} U_i \Big). \quad (23)$$

We state the two main theoretical results. The first exhibits the dichotomous behaviour of $\widehat{SECMI}$ showing that its limit can be either normal or may have a distribution of quadratic form in normal random variables. The second result fully characterizes the latter case for binary Y. The proofs of the results are given in the Appendix. Let $\hat{\mathbf{p}}$ be a vector of sample fractions corresponding to the vector of probabilities $\mathbf{p} = (p(x, y, z))$. Note that the case $m = 1$ is covered by the previous result.

Theorem 2 *(i) Assume that $\sigma^2_{\widehat{SECMI}} > 0$ and $m > 1$. Then we have*

$$n^{1/2}(\widehat{SECMI} - SECMI) \xrightarrow{d} N(0, \sigma^2_{\widehat{SECMI}}). \quad (24)$$

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

$$2n(\widehat{SECMI} - SECMI) \xrightarrow{d} W^T HW, \quad (25)$$

where W has $N(0, \Sigma)$ distribution, $\Sigma = n\mathrm{Var}(\hat{\mathbf{p}} - \mathbf{p})$ and H is a Hessian matrix defined in (44).

It follows from the proof of Theorem 1 that the distribution of $W^T HW$ in (25) coincides with the distribution of a weighted sum of squared independent normally distributed random variables, where the weights are equal to the eigenvalues of ΣH. Theorem 2 asserts that $\widehat{SECMI}$ exhibits the dichotomous behaviour. Depending on whether $\sigma^2_{\widehat{SECMI}} \neq 0$ or $\sigma^2_{\widehat{SECMI}} = 0$ the asymptotic distribution is either normal or coincides with a distribution of a certain quadratic form in normal variables. Such type of behaviour is typical for the Likelihood Ratio Statistics (see e.g. Vuong 1989).

Let $S = \{1, \ldots, m\}$. As the limiting distribution depends on whether $\sigma^2_{\widehat{SECMI}}$ is zero or not, it is of importance to characterize this condition. In this way we learn dependence structures which are possible when the asymptotic distribution of $\widehat{SECMI}$ is not normal. This can be investigated in detail for the binary Y. Namely, the following result holds for all $m \in N$.

Theorem 3 *Assume that Y is binary random variable, X, Z are discrete and $\sigma^2_{\widehat{SECMI}} = 0$.*
(i) Then we have the following exclusive possibilities:
1) for all $s \in S$, variables X, Y are conditionally independent given Z_s, moreover X, Y are independent. This case always holds when $m = 1$;
2) there exists exactly one $s_0 \in S$ such that X, Y are conditionally independent given Z_{s_0}, for all $s \in S \setminus \{s_0\}$ Y, Z_s are independent, (Y, Z_s) are conditionally independent given X, X, Y are conditionally dependent given Z_s. Additionally, X and Y are dependent.
(ii) Conversely, if 1) or 2) holds then $\sigma^2_{\widehat{SECMI}} = 0$.
(iii) We have $SECMI(X, Y|Z) = 0$.

Observe that in the view of (i)-(ii) above, possibilities 1) and 2) give a complete description of the dependence structures for which $\widehat{SECMI}$ does not have an asymptotically normal law. Moreover, for binary Y in the view of (iii) the centering in (25) can be omitted.
The following corollary follows directly from Theorem 4.3.

Corollary 1 *Assume that Y is binary and $m \geq 2$. If there exist at least two Z_i such that $X \perp Y|Z_i$ and $\sigma^2_{\widehat{SECMI}} = 0$ then $X \perp Y|Z_i$ for all $i = 1, \ldots, m$ and X, Y are independent.*

Indeed, it follows from the assumptions and Theorem that 1) of (ii) holds in this case.
We discuss now the consequences of Theorem 4.3 for the feature selection. Let Z be a vector of already chosen variables and X a candidate. Then case 2) of (i) states in particular that (Y, Z_s) are conditionally independent given X for $s \neq s_0$ and this means that X should be chosen before all such Z_s in Markov Blanket discovery for Y. Thus remaining Z_s are redundant if X and Z_{s_0} are already chosen. Whence this situation is excluded in the greedy selection context and thus only the case 1) is possible. In other words, for $m > 1$ limiting distribution of $\widehat{SECMI}$ is not normal under the restrictive set-up of the case 1) of (i). Thus for $m > 1$, Y binary with Z_S being the set of chosen variables $\sigma^2_{\widehat{SECMI}} = 0$ is equivalent to 1) and can be re-expressed as the following Scenario 1.
Scenario 1.

$$\mathbb{E}U_0 = \mathbb{E}U_1 = \cdots = \mathbb{E}U_m = 0.$$

The alternative scenario is
Scenario 2. There exists $0 \leq i \leq m$ such that $\mathbb{E}U_i \neq 0$.
For binary Y in the view of Theorem 4.3 the number of $0 \leq i \leq m$ such that $EU_i \neq 0$ in Scenario 2 equals m as $EU_i \neq 0$ for $i = 0$ and $i \in S \setminus \{s_0\}$. Moreover, we note that in view of (iii) of Theorem 4.3 for binary Y, $SECMI \neq 0$ is sufficient condition for the asymptotic normality of $\widehat{SECMI}$.

Remark 2 *We comment now on the general case of arbitrary triple (X, Y, Z) when Y is not necessarily binary. Let $M = \{i = 0, \ldots, m : \mathbb{E}U_i \neq 0\}$. Then M is non-empty and letting $\hat{I}(X, Y|Z_0) := (1 - m)\hat{I}(X, Y)$ we can write*

$$n^{1/2}(\widehat{SECMI} - SECMI) = n^{1/2} \sum_{i \in M} (\hat{I}(X, Y|Z_i) - \mathbb{E}U_i) + n^{1/2} \sum_{i \in M^c} \hat{I}(X, Y|Z_i).$$

In view of Theorem 1 (i) and the remark below it, each term in the first sum has asymptotic normal distribution and the second sum is asymptotically negligible due to the difference *in normings in (9) and (10). Thus in Scenario 2 contribution of the first sum eventually prevails and conclusion of Theorem 2 (i) holds unless $Z_i, i \in M$ are strongly negatively dependent in the sense that the variance of the corresponding sum tends more quickly to 0 than n^{-1}. In contrast, in the case of Scenario 1 for all Z_i we have that X and Y are conditionally independent given Z_i and moreover, X is independent of Y. In this case in view of Theorem 1 (ii) and the decomposition of $\widehat{SECMI}$ in (19) all terms U_i will have approximately the chi square distribution for $i \geq 1$ and scaled chi square distribution for $i = 0$. As all these terms are dependent the asymptotic distribution of the sum is in general case more complicated than chi square distribution, namely the distribution of the quadratic form given in (25).*

Note that if X is such that $X \perp (Y, Z)$ then Scenario 1 holds as this implies $X \perp Y | Z_i\,, i = 1, \ldots, m$ and $X \perp Y$. This particular case has been considered in synthetic models investigated in Mielniczuk and Teisseyre (2019), where distribution of $\widehat{SECMI}$ has been approximated by the chi square distribution.

We now consider behaviour of $\widehat{SECMI}$ under the null hypothesis

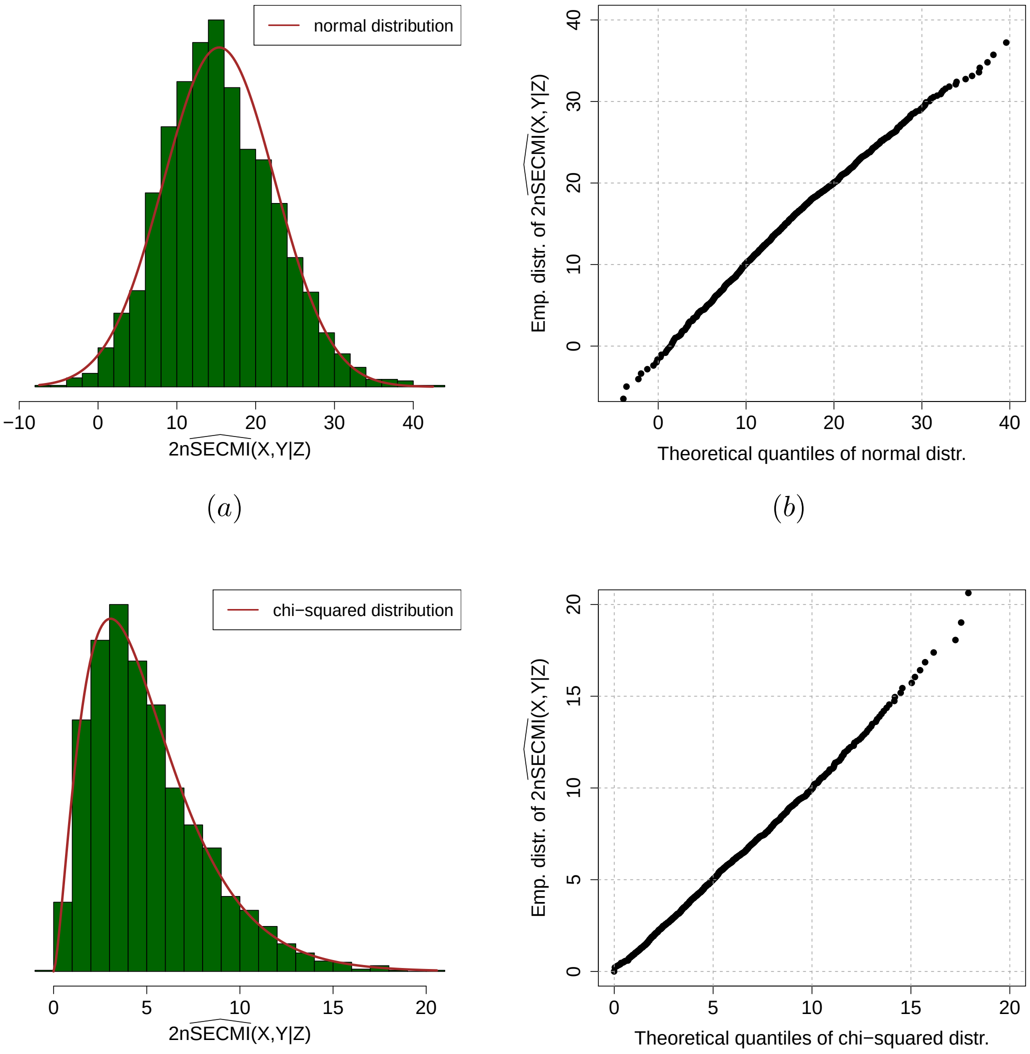
$$H_0 : \quad X \perp Y | Z$$

and under the alternative

$$H_1 : \quad X \not\perp Y | Z.$$

We note that under the null hypothesis both scenarios, Scenario 1 and Scenario 2 are possible. Distribution of $SECMI$ when H_0 holds may differ depending on distribution of the vector (X, Y, Z). However it follows from Theorem 4.3, that if there are at least two, but not all Z_is such that $X \perp Y | Z_i$ then, for the binary Y, the distribution of $\widehat{SECMI}$ is asymptotically normal as then the case 2) is excluded. In the view of discussion above, in the case of Scenario 2, we approximate distribution of $\widehat{SECMI}$ by the normal distribution (i.e. its asymptotic limit.) In the case of Scenario 1, we use the chi square distribution or, alternatively, a scaled and shifted chi square distribution. Thus in the second case, the distribution of the quadratic form is approximated by one of these distributions. As the parameters of these distributions are unknown, we propose to estimate them using permutation scheme which generates data conforming to the null hypothesis and preserves sample distributions of X and Y given Z. Figure 4 shows distributions and quantile plots of $\widehat{SECMI}$ for $m = 2$, $n = 5000$ and simulation models E1 and E2, described in Section 6. In both cases null hypothesis H_0 holds. In the first case, however, $SECMI \neq 0$ and thus $\sigma^2_{\widehat{SECMI}} \neq 0$ (see Theorem 4.3). Whence distribution of $\widehat{SECMI}$ is asymptotically normal whereas in the second case $X \perp (Y, Z)$ holds, $\sigma^2_{\widehat{SECMI}} = 0$ and the asymptotic distribution coincides with distribution of quadratic form in normal variables.

Remark 3 *Note that the null hypothesis H_0 for $m > 1$ is not implied in general by neither $X \perp Y | Z_i$ for $i = 1, \ldots, m$ or $X \perp Y$. However, this is true for a vast class of distributions which are faithful to some undirected graph (see Section 13.6 in Bühlmann and van de Geer 2015).*



(a) (b)

(c) (d)

Figure 4: Histogram and quantile plot of $SECMI$, for $X \perp Y|Z$ (a,b) and $X \perp (Y,Z)$ (c,d), for $m = 2$ and $n = 5000$. Data is generated from simulation models E1 and E2, described in Section 6.

4.4 Higher-Order Expansions of Conditional Mutual Information

The main advantage of $SECMI$ defined in (18) lies in conditioning on single variables only, which leads to more powerful tests in the situations when the dependence structure among the variables is not very complex, i.e. high-order interactions (of order > 2) among variables are not present or are negligible. On the other hand, it should be stressed that $SECMI$ ignores the higher-order terms (it is obtained by truncation of the Möbius representation of CMI, see formula (17)) and thus it will fail to detect dependence when low-dimensional terms are zero and only high-order interactions contribute to CMI. In such case $CMI > 0$ and at the same time $SECMI = 0$. To deal with the above situation ($CMI > 0$, $SECMI = 0$) it is possible to generalize $SECMI$ in order to take into account higher-order interactions. Such generalization can be obtained by taking the first k terms ($k < m$) in the Möbius representation (17). For example a test statistic containing the first three components

$$SECMI3(Y,X|Z) := I(X,Y) + \sum_{k=1}^{m} II(X,Z_k,Y) + \sum_{k_1<k_2} II(X,Z_{k_1},Z_{k_2},Y), \tag{26}$$

where $II(X,Z_{k_1},Z_{k_2},Y)$ is defined in (8), will detect the interactions of order 3. Such generalization was already considered in the context of feature selection, see Vinh et al. (2016) and Pawluk et al. (2019). We note that application of (8) yields that (see Appendix)

$$SECMI3(Y,X|Z) = A \times I(X,Y) + B\sum_{s\in S} I(X,Y|Z_s) + \sum_{s<s',s,s'\in S} I(X,Y|Z_s,Z_{s'}),$$

where A and B are defined as where

$$A = 1 - \binom{|S|}{1} + \binom{|S|}{2}, \qquad B = 1 - \binom{|S|-1}{1}. \tag{27}$$

Define $\widehat{SECMI}3$ to be empirical counterpart of (26) defined analogously to (19) and let

$$\sigma^2_{\widehat{SECMI}3} = \mathrm{Var}\Big(\log\Big(\Big[\frac{p(X,Y)}{p(X)p(Y)}\Big]^A\Big[\prod_{s\in S}\frac{p(X,Y,Z_s)}{p(X,Z_s)p(Y,Z_s)}\Big]^B\prod_{s<s',s,s'\in S}\frac{p(X,Y,Z_s,Z_{s'})p(Z_s,Z_{s'})}{p(X,Z_s,Z_{s'})p(Y,Z_s,Z_{s'})}\Big). \tag{28}$$

Then the following analogue of Theorem 2 holds

Theorem 4 *(i) Assume that* $m > 2$ *and* $\sigma^2_{\widehat{SECMI}3} > 0$. *Then we have*

$$n^{1/2}(\widehat{SECMI}3 - SECMI3) \xrightarrow{d} N(0, \sigma^2_{\widehat{SECMI}3}). \tag{29}$$

(ii) If $\sigma^2_{\widehat{SECMI}3} = 0$ *then*

$$2n(\widehat{SECMI}3 - SECMI3) \xrightarrow{d} W^T H W, \tag{30}$$

where W *has* $N(0,\Sigma)$ *distribution,* Σ *is defined in Theorem 2 and* H *is a Hessian matrix of the function defined in (70).*

The proof of Theorem 4 is given in the Appendix. The result is a theoretical justification of choosing one of the two reference distributions under the hypothesis of the conditional independence when the test statistic is $\widehat{SECMI}3$. We note that when the existence of an interaction between specified variables is suspected a variable corresponding to this interaction may be introduced as a new predictor. Also, construction of a test of existence of interactions of higher order is possible by extending results of Kubkowski and Mielniczuk (2020), however the question how to use it efficiently avoiding multiple testing problem remains unresolved.

5. Testing Conditional Independence Using $SECMI$

5.1 Permutation Test for $SECMI$

In the following we discuss construction of $SECMI$-based test using permutations of the given data. In view of the dichotomous behaviour of $\widehat{SECMI}$ (see Theorem 2) we propose data-dependent determination which of the two distributions is closer to the distribution of $\widehat{SECMI}$ under the null hypothesis and the chosen distribution is used to decide the outcome of the test. Namely, for each strata $Z = z$ consisting of n_z observations we permute B times corresponding n_z values of X and for each permutation we calculate value $\widehat{SECMI}_k$ for $k = 1, \ldots, B$.

Using the sample of $\widehat{SECMI}_k$, we calculate the two first central moments which yield estimators of μ and σ^2 for a normal distribution $N(\mu, \sigma^2)$ and a number of degrees of freedom d for χ^2_d which is estimated as a sample mean. Then the sample distribution of $\widehat{SECMI}_k$ is compared to $N(\hat{\mu}, \hat{\sigma}^2)$ and $\chi^2_{\hat{d}}$ and the closer of the two in supremum metrics is chosen. When the sample mean is non positive then the normal distribution is chosen as the reference distribution. Then p-value of the observed value of $\widehat{SECMI}$ is calculated with respect to the chosen distribution. This version of the algorithm is described in detail in Algorithm 1. We also consider $SECMI(chi\_s)$ in which the chi square distribution is replaced with a scaled and shifted chi square distribution. It is defined as the distribution of $\alpha\chi^2_d + \beta$, where χ^2_d is the chi square distribution with $d \in R^+$; parameters α, d, β are calculated based on the three first moments of $\{\widehat{SECMI}_k\}$ (see Buckley and Eagleson 1988 and Zhang 2005).
Note that the conditional permutation method is based on a simple principle that under H_0 the distribution $P_{X|Y,Z}$ equals $P_{X|Z}$. Thus if r.v. $\tilde{X}$ is such that $P_{\tilde{X}|Z} = P_{X|Z}$ then $P_{X,Y,Z} = P_{\tilde{X},Y,Z}$. Observations following $P_{\tilde{X}|Z}$ are e.g. obtained by randomly permuting the sample generated from distribution $P_{X|Z}$.

Our operational premise is that distribution of $\widehat{SECMI}$ has essentially two forms: is either approximately normal or chi square. We only use permutations to estimate parameters of these distributions from the distribution of $\widehat{SECMI}$ calculated for the permuted samples and not to estimate the whole distribution. Thus, we avoid prohibitive number of permutations to approximate accurately quantiles of this distribution. The proposed method is a variant of the semi-parametric permutation test used also in the Markov Blanket discovery (see Tsamardinos and Borboudakis 2010) but differs in two important aspects: the switch between two parametric distributions is proposed based on theoretical considerations to ensure better fit to permutation distribution, and, secondly, the conditional information $\hat{I}$ test statistics is changed to $\widehat{SECMI}$ to obtain more powerful tests.

We experimentally tested the switch mode in $SECMI$ procedure, described in Algorithm 1. We consider the following simple simulation model. We first generate $Z_1, \ldots, Z_m$ independently from discrete uniform distribution on $\{-1, 0, 1\}$. We generate $Y \in \{0, 1\}$ from Bernoulli distribution with success probability $P(Y = 1|Z_1, \ldots, Z_m) = \sigma(Z_1 + \ldots + Z_m)$ and X which follows the Bernoulli distribution with success probability $P(X = 1|Z_1) = \sigma(\delta Z_1)$, where δ is a parameter. Parameter δ controls the dependence strength between X and Z_1. Value $\delta = 0$ corresponds to $X \perp (Y, Z_1, \ldots, Z_m)$ and $\delta > 0$ corresponds to $X \perp Y|Z_1, \ldots, Z_m$ and Scenario 2. Thus in the first case the asymptotic distribution of

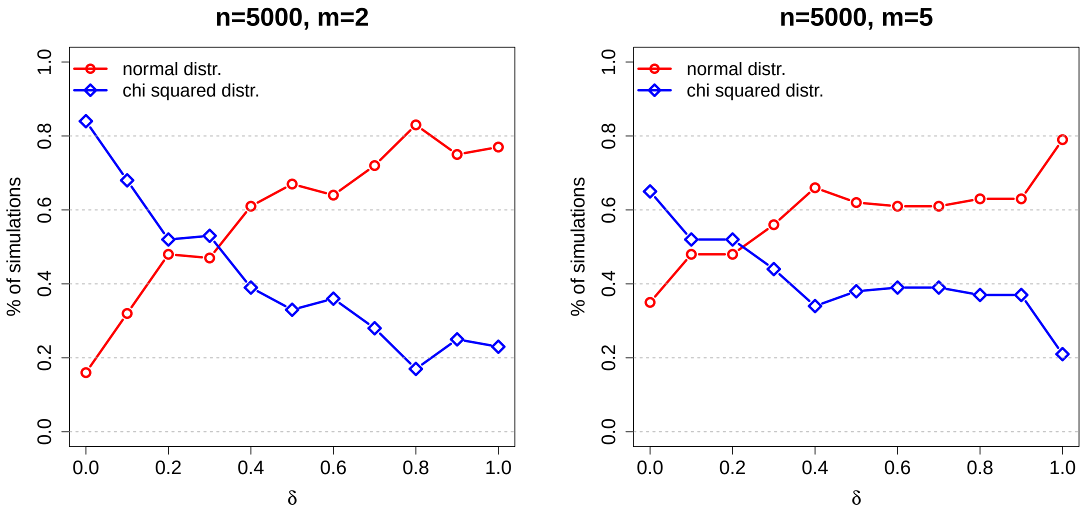
```
Algorithm 1: $SECMI$
  Input      : Training sample $D$ of size $n$, drawn from distribution of $(X, Y, Z)$;
               number of permutations $B$.
  Compute: $\widehat{SECMI}_0 = \widehat{SECMI}(X, Y|Z)$
  for $k = 1, \ldots, B$ do
      Randomly permute $X$ (on each strata on $Z$); denote by $X^{(k)}$ the variable $X$
        with permuted values.
      Compute: $\widehat{SECMI}_k := \widehat{SECMI}(X^{(k)}, Y|Z)$
  Compute:
  $\hat{\mu} := \frac{1}{B}\sum_{k=1}^{B} \widehat{SECMI}_k$
  $\hat{\sigma}^2 := \frac{1}{B-1}\sum_{k=1}^{B}(\widehat{SECMI}_k - \hat{\mu})^2$.
  Let:
  $F_B(s)$ empirical distribution function of $\widehat{SECMI}_k$, $k = 1, \ldots, B$.
  $F_{N(\hat{\mu},\hat{\sigma})}(s)$ theoretical distribution function of $N(\hat{\mu}, \hat{\sigma}^2)$
  $F_{\chi^2_{\hat{\mu}}}(s)$ theoretical distribution function of $\chi^2_{\hat{\mu}}$
  Compute:
  $D_{N(\hat{\mu},\hat{\sigma})} := \sup_s |F_B(s) - F_{N(\hat{\mu},\hat{\sigma}^2)}(s)|$
  $D_{\chi^2_{\hat{\mu}}} := \sup_s |F_B(s) - F_{\chi^2_{\hat{\mu}}}(s)|$
  if $D_{N(\hat{\mu},\hat{\sigma})} < D_{\chi^2_{\hat{\mu}}}$ or $\hat{\mu} \leq 0$ then
      $p = 1 - F_{N(\hat{\mu},\hat{\sigma}^2)}(\widehat{SECMI}_0)$
  else
      $p = 1 - F_{\chi^2_{\hat{\mu}}}(\widehat{SECMI}_0)$
  Output     : p-value $p$
```

$\widehat{SECMI}$ is non-normal in contrast to the case of large $\delta > 0$. Figure 5 shows the fraction of simulations in which the chi square or the normal distribution were chosen in Algorithm 1 for the above simulation model. When $X \perp (Y, Z_1, \ldots, Z_m)$, the chi square distribution is chosen in 84% of simulations (for $m = 2$) and 65% (for $m = 5$). For larger δ, the normal distribution is chosen more often. When the size of the conditioning set m increases, the chi square distribution becomes similar to the normal distribution, which explains why the curves do not separate so clearly for $\delta = 0$.

It is well known that the asymptotic results are hard to apply in the case of the small sample sizes. We stress that in the following only qualitative result about the dichotomous behaviour of $\widehat{SECMI}$ is used and parameters of the benchmark distributions are estimated using permutation scheme. In Section 6 we check in numerical experiments that such approach leads to satisfactory control of type I error and investigate the power of the underlying test.

5.2 Using $SECMI$-Based Test for Variable Selection

We discuss now possible uses of $SECMI$-based test in variable selection. We have stated already that the $SECMI$ criterion is identical with the $CIFE$ criterion introduced in Lin and



(a) (b)

Figure 5: Fraction of simulations in which chi squared or normal distribution were chosen in Algorithm 1 $SECMI$ for simulation model described in Section 5.1. Value $\delta = 0$ corresponds to $X \perp (Y, Z_1, \ldots, Z_m)$ and $\delta > 0$ corresponds to $X \perp Y | Z_1, \ldots, Z_m$.

Tang (2006). However, $CIFE$ is commonly understood as a greedy feature selection procedure, which works as follows. Assume the supervised learning scenario in which $X_1, \ldots, X_p$ are features and Y is the target variable. The method starts from an empty set of features. Next, in each step it selects feature X_j, $j \in \{1, \ldots, p\} \setminus S$ which maximizes the following criterion

$$J(X_j, S) := I(Y, X_j) + \sum_{i \in S} II(X_i, X_j, Y), \tag{31}$$

where S is the set of features selected as relevant in the previous steps. The first term in (31) corresponds to marginal dependence between a candidate variable X_j and the target variable Y. The second term is related to interactions between the already selected variables and the candidate variable. A conditional independence test based on $SECMI$ can be naturally used in the feature selection task. Firstly it can be easily incorporated into $CIFE$ algorithm as a stopping rule, namely, we stop adding new variables in $CIFE$ when $X_j \perp Y | S$, for all $j \in \{1, \ldots, p\} \setminus S$, where the null hypothesis of the conditional independence is verified using $SECMI$. Importantly, note that using $SECMI$ test in the context of the feature selection is not limited to $CIFE$ algorithm. The proposed test can be combined with all popular Markov Blanket (MB) discovery algorithms such as GS or IAMB (Tsamardinos et al., 2003), which are based on conditional independence testing. The GS and IAMB consist of two steps. In the first step, we sequentially add features using a series of the conditional independence tests. In the second (backward) step, the variables are sequentially removed from the current set of the selected variables. The second step is used to limit the number of variables falsely included in the MB (for the review of various forward-backward techniques to achieve this we refer to Borbudakis and Tsamardinos 2019, see also Mielniczuk and Teisseyre 2019). Finally, note that there is also a close relationship between $CIFE$ algorithm and IAMB algorithm when IAMB is combined with $SECMI$. Specifically, denote by IAMB+$SECMI$ an IAMB algorithm in which $SECMI$ is used as a test of conditional independence. Observe that when we ignore correction for a multiple testing, $CIFE$ with a stopping rule determined by $SECMI$ test is identical to the first (forward) step of IAMB+$SECMI$.

6. Experiments

The aim is now to verify how $SECMI$ works in practice for three problems: (i) as a test statistic for the test of conditional independence, (ii) for the feature selection in supervised classification and (iii) for Markov blanket discovery.

6.1 Power and Type I Error Comparison

In this Section we apply the conditional independence tests considered above to verify the null hypothesis

$$H_0 : X \perp Y | Z_1, \ldots, Z_m. \tag{32}$$

We investigate how a power (probability of rejection of H_0 when it is false) and a type I error (probability of rejection of H_0 when it is true) of the proposed tests depend on the sample size n, the size of conditioning set m and the dependence structure among variables. We artificially generate variables $X, Y, Z_1, \ldots, Z_m$ using various dependence schemes described below. We run $L = 500$ simulations and report the empirical power and the type I error as a fraction of simulations in which the null hypothesis stated in (32) has been rejected. We compare the proposed tests $SECMI$, $SECMI(chi\_s)$ and $SECMI3$ (see Section 4.4). As a reference we use two tests: (1) asymptotic test for CMI (called simply CMI) and (2) semi-parametric test for CMI (called $CMI(sp)$) proposed in Tsamardinos and Borboudakis (2010). In $CMI(sp)$, permutation scheme is used to estimate a number of degrees of freedom of the reference chi square distribution. In CMI test, the number of degrees of freedom is determined by asymptotic distribution, see Theorem 1. Note that in both CMI and $CMI(sp)$ the same statistic is used. The methods only differ in the choice of the reference distribution. To avoid a significant computational burden, we use only $B = 50$ permutations in $SECMI$, $SECMI(chi\_s)$, $SECMI3$ and $CMI(sp)$. As in $CMI(sp)$ we only estimate parameters of reference distributions (the chi square or the normal), it is not necessary to take very large B. For CMI and $CMI(sp)$ we used implementations available in R package `bnlearn` (Scutari, 2010). The proposed methods were also implemented in R language.

6.1.1 Simulation Models

We consider the following simulation models for the power comparison (models P1-P4) and the type I error comparison (models E1-E4). Figures 6 and 7 illustrate graphically the considered probability structures. First we define models for power comparison which were chosen to represent a wide spectrum of dependence structures. We note that when Y is treated as the target variable, models P1 and P4 below are discriminative models, in the sense that we first generate variables Z_i and X and then the target variable Y is generated using Z_i and X (see e.g. Barber 2014, Chapter 13). In contrast, models P2 and P3 belong to the class of so-called generative models in which the order of generation is opposite to that for discriminative models, i.e. we first generate the target variable Y and then generate Z_i using the conditional distributions of Z_i given Y. Moreover, model P3 is often referred to as a 'collider' as the arrows indicating direction of dependence 'collide' at vertex X (see Figure 6 (c) and e.g. Barber 2014, Definition 3.2).

Simulation model P1 We first generate $Z_1, \ldots, Z_m, X$ independently from the discrete the uniform distribution on $\{-1, 0, 1\}$. Then we generate $Y \in \{0, 1\}$ from the Bernoulli distribution with success probability $P(Y = 1|Z_1, \ldots, Z_m, X) = \sigma(Z_1 + \ldots + Z_m + \gamma \cdot X)$, where $\sigma(s) = 1/(1 + \exp(-s))$ is the logistic function and γ is a parameter which controls the strength of conditional dependence between Y and X given $Z_1, \ldots, Z_m$. See Figure 6 (a).

Simulation model P2 Variable $Y \in \{0, 1\}$ is generated from the Bernoulli distribution with success probability 0.5. Next we generate independent auxiliary variables $U_1, \ldots, U_m, U_{m+1}$ independently from Y according to the normal distribution $N(0, 1)$. We let $Z_i = h(U_i + Y)$ for $i = 1, \ldots, m$ and $X = h(U_{m+1} + Y)$, where $h(x)$ discretizes x to 3 values $0, 1, 2$ (Φ denotes below $N(0, 1)$ cdf):

$$h(x) = \begin{cases} 0 & x < \Phi^{-1}(0.25), \\ 1 & x \in [\Phi^{-1}(0.25), \Phi^{-1}(0.75)], \\ 2 & x > \Phi^{-1}(0.75). \end{cases} \tag{33}$$

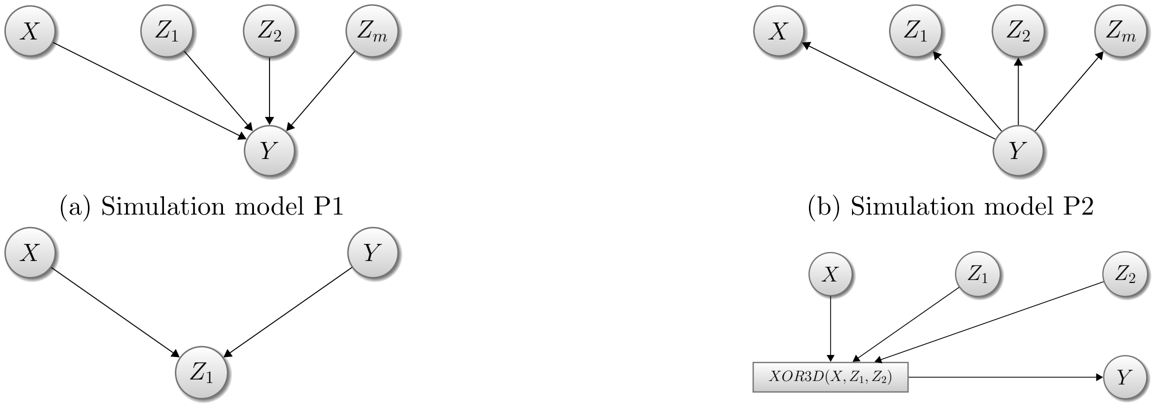
Simulation model P3 (collider) First we generate $Y \in \{0, 1\}$ from the Bernoulli distribution with success probability 0.5. Variable X is generated from the discrete uniform distribution on $\{-1, 0, 1\}$. Then we generate $Z_1 \in \{0, 1\}$ according to $P(Z_1 = 1|X, Y) = 0.7$ if $X + Y \geq 0$ and $P(Z_1 = 1|X, Y) = 0.3$ if $X + Y < 0$. Finally, $Z_2, \ldots, Z_m$ are generated independently from the discrete uniform distribution on $\{-1, 0, 1\}$.

Simulation model P4 Binary variables X, Z_1, Z_2 are generated from the uniform distribution on $\{0, 1\}$ and $Y = XOR3D(X, Z_1, Z_2)$ is defined as 1 when the sum $X_1 + Z_1 + Z_2$ is odd and 0 otherwise. Thus XOR3D is three-dimensional version of the usual XOR. To make the testing task harder, we next change $Y = 1$ to $Y = 0$ with probability 0.3 if $X + Z_1 + Z_2$ is odd and we change $Y = 0$ to $Y = 1$ with probability 0.3 if $X + Z_1 + Z_2$ is even. We additionally generate $Z_3, \ldots, Z_m$ from the uniform distribution on $\{0, 1, 2\}$, independently from X, Y, Z_1, Z_2.

Observe that models P1, P2 and P3 are constructed in such a way that the interactions between variables of order 3 and higher are absent. Therefore, we expect $SECMI$ to work on par or even better than CMI as it avoids estimation of non-existing effects. $SECMI3$ takes into account the 3rd order interaction terms, whose theoretical values are zero for models P1-P3. So $SECMI3$ is also expected to work correctly for these models. For model P4, however, assumptions of $SECMI$ are not met, i.e. the theoretical value of $SECMI$ is zero, whereas theoretical values of CMI and $SECMI3$ are positive and equal. So in this case, $SECMI$ will fail to detect the true conditional dependence between variables, whereas $SECMI3$ should work correctly.

Now we define models for Type I error comparison.

Simulation model E1 We first generate $Z_1, \ldots, Z_m$ independently from the discrete uniform distribution on $\{-1, 0, 1\}$. We generate $Y \in \{0, 1\}$ from the Bernoulli distribution



(a) Simulation model P1

(b) Simulation model P2

(c) Simulation model P3

(d) Simulation model P4

Figure 6: Graphs corresponding to simulation models P1-P4 described in Section 6.1.1.

with success probability $P(Y = 1|Z_1, \ldots, Z_m) = \sigma(Z_1 + \ldots + Z_m)$ and X which follows the Bernoulli distribution with success probability $P(X = 1|Z_1) = \sigma(Z_1)$. In this case Scenario 2 holds.

Simulation model E1bis We define $Y, Z_1, \ldots, Z_m$ as in model E1, but X is now generated from the Bernoulli distribution with success probability $P(X = 1|Z_1, \ldots, Z_m) = \sigma(Z_1 + \ldots + Z_m)$.

Simulation model E2 We define $Y, Z_1, \ldots, Z_m$ as in model E1, but X is now generated independently from $(Y, Z_1, \ldots, Z_m)$ and has the Bernoulli distribution with probability 1/2.

Simulation model E3 First we generate $Y \in \{0, 1\}$ from the Bernoulli distribution with success probability 0.5. Next we generate auxiliary variables $U_1, \ldots, U_m, U_{m+1}$ independently from the normal distribution $N(0, 1)$. We let $Z_i = h(U_i + Y)$ for $i = 1, \ldots, m$ and $X = h(Z_1 - 2 + U_{m+1})$.

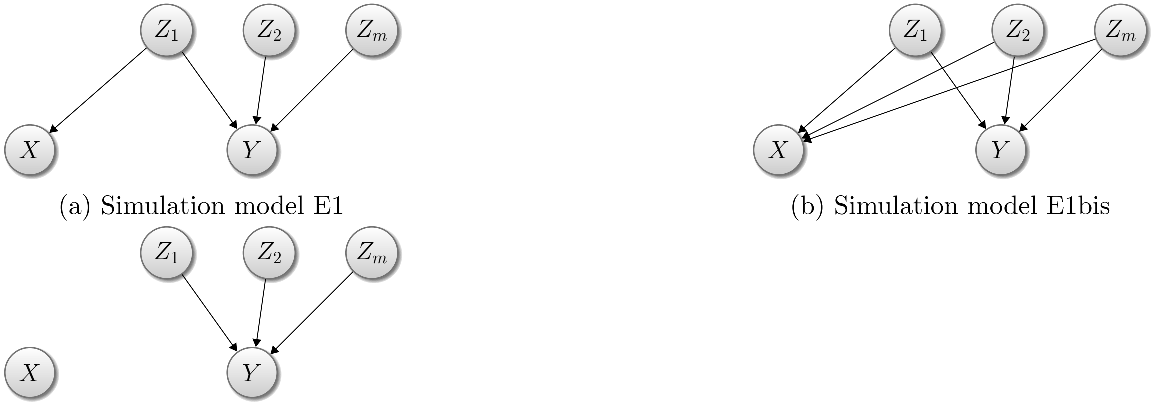
Simulation model E3bis We generate Y, $U_1, \ldots, U_m, U_{m+1}$ and Z_i for $i = 1, \ldots, m$ as in model E3, but X is defined as $X = h(Z_1 + \ldots + Z_m - 2m + U_{m+1})$.

Simulation model E4 We define $Y, Z_1, \ldots, Z_m$ as in model E3, however $X = h(U_{m+1})$ now. In this case Scenario 1 holds.

Note that in E1 and E3 we have $X \perp Y|Z_1, \ldots, Z_m$, whereas in E2 and E4 a stronger condition holds, namely $X \perp (Y, Z_1 \ldots, Z_m)$. Models E1bis and E3bis are modifications of E1 and E3, respectively. Note that in E1 $X \perp Y|Z_1, \ldots, Z_m$ and also $X \perp Y|Z_1$ but $X \not\perp Y|Z_i$ for $i \geq 2$. In E1bis, $X \perp Y|Z_1, \ldots, Z_m$ but $X \not\perp Y|Z_i \, i = 1, \ldots, m$. So in E1bis, X and Y are conditionally independent given all variables $Z_1, \ldots, Z_m$ and conditionally dependent given individual Z_is. An analogous relationship exists between models E3 and E3bis.

6.1.2 Results

Figures 9 and 10 show the power wrt size of the conditioning set m, for simulation models P1-P4 defined in Section 6.1.1, and two sample sizes: $n = 1000$ and $n = 5000$. In model P1, parameter $\gamma = 1$. Figure 11 shows the power wrt to γ for model P1. In addition we present how the power varies with the sample size for fixed size of the conditioning set $m = 5$, see

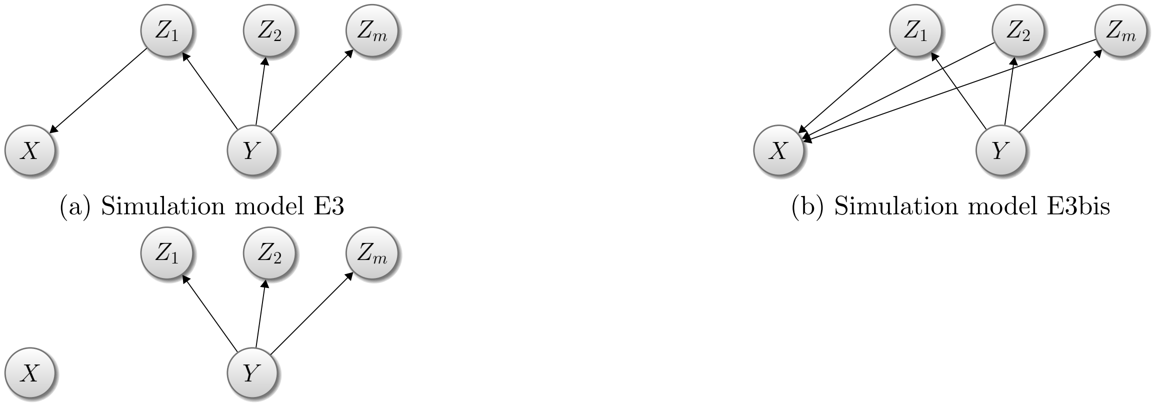


(a) Simulation model E1

(b) Simulation model E1bis

(c) Simulation model E2

Figure 7: Graphs corresponding to simulation models E1, E1bis and E2 described in Section 6.1.1.



(a) Simulation model E3

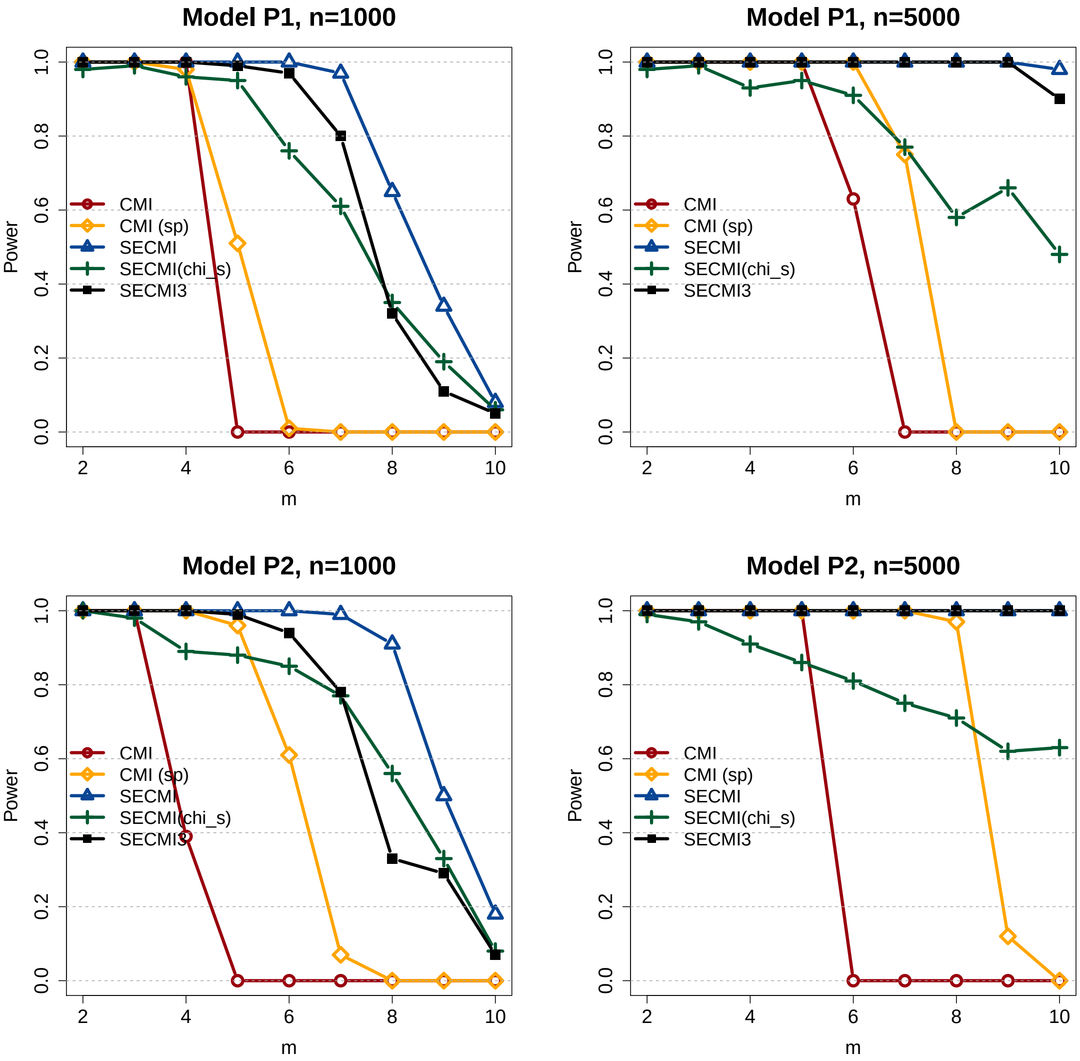
(b) Simulation model E3bis

(c) Simulation model E4

Figure 8: Graphs corresponding to simulation models E3, E3bis and E4 described in Section 6.1.1.

Figure 12. First it is clearly seen that the power of CMI decreases to 0 for relatively small m. For example, in the case of models P1-P2, the power of CMI is equal 0 already for $m = 5$ and $n = 1000$. The power of $CMI(sp)$ is usually slightly larger, but it is also close to 0 for relatively small m. For example, in the case of model P1 the power of $CMI(sp)$ approaches 0 already for $m = 6$ when $n = 1000$. The power for the proposed methods is much larger than for the CMI-based tests. Among the proposed methods, $SECMI$ achieves the largest power for models P1-P3. The test based on the sample version of $SECMI3$ defined in (26) is the second best for models P1, P2 and P3 (in the last case only for $n = 5000$). This indicates that an unnecessary inclusion of higher-order terms in the Möbius representation diminishes the power of the test. Despite this, $SECMI3$ is still better than CMI and $CMI(sp)$ for all considered models. As expected, $SECMI$ fails in the case of model P4 (its power oscillates around the significance level) as this model contains an interaction of order 3, which is not captured by $SECMI$. The CMI-based methods work for small m, for model P4. Importantly, when m increases, CMI also fails to detect the dependence for P4. However, $SECMI3$ works very well in this case, notably even in situations when CMI fails. The other proposed method $SECMI(chi\_s)$ usually works worse than $SECMI$ and $SECMI3$ wrt to the power, although it controls the type I error slightly better. As expected, the power increases with sample size. In general, the performance of the tests strongly depends on 'm to n' ratio. For large 'm to n' ratio, CMI-based test work poorly, even when theoretical value of CMI is positive (e.g. for model P4). Recall that in the considered models each explanatory variable admits three values, thus the number of possible values of vector $(Z_1, \ldots, Z_m)$ equals 3^m and depends exponentially on m. Note also that in model P1, for small γ and small m all methods work similarly, whereas for larger m and larger γ, $SECMI$-based methods are evidently superior (see Figure 11).

Next, we studied observed type I errors for simulation models when the assumed type I error is $\alpha = 0.05$. Recall that we consider two situations: $X \perp Y|Z$ (models E1, E1bis, E3 and E3bis) and $X \perp (Y, Z)$ (models E2 and E4). The proposed methods control the type I error, for larger sample size $n = 5000$, see Figures 13, 14 and 15. We also observe rather unstable behaviour of CMI, for all considered sample sizes. For model E1bis we observe that the empirical type I error exceeds assumed level 0.05 for $m = 3$ for $SECMI$-based methods but in general they control the assumed type I error satisfactorily. In the most challenging situation (large m and small n) it may happen that there is only one observation for each combination of conditioning variables. We observe such effect for $m = 10$ and $n = 1000$. Note that for $m = 10$ we have $3^m = 59049$ possible values of $(Z_1, \ldots, Z_m)$. This is analogous to the case of having only one observation per cell in the contingency table. In such situations, permutation scheme fails as the number of degrees of freedom is estimated by the value of the statistics and the hypothesis is never rejected. This explains why the curves in Figures 13 and 15 approach 0 for $n = 1000$ and $m = 10$. We also analysed the type I errors as a function of α. Figures 16, 17 and 18 show the results for $n = 5000$. Observe that for small $m = 2$, all methods control the type I error fairly well although CMI behaves conservatively in model E3bis (left hand side panels). For larger $m = 6$, CMI does not control the type I error (right hand side panels); the curve corresponding to CMI oscillates around zero, which means that this test does not reject the null hypothesis at all.


Model P1, n=1000

Model P1, n=5000

Model P2, n=1000

Model P2, n=5000

Figure 9: Power wrt to m for models P1 and P2 described in Section 6.1.1. Left-hand side panels correspond to sample size $n = 1000$, right-hand side panels correspond to sample size $n = 5000$.

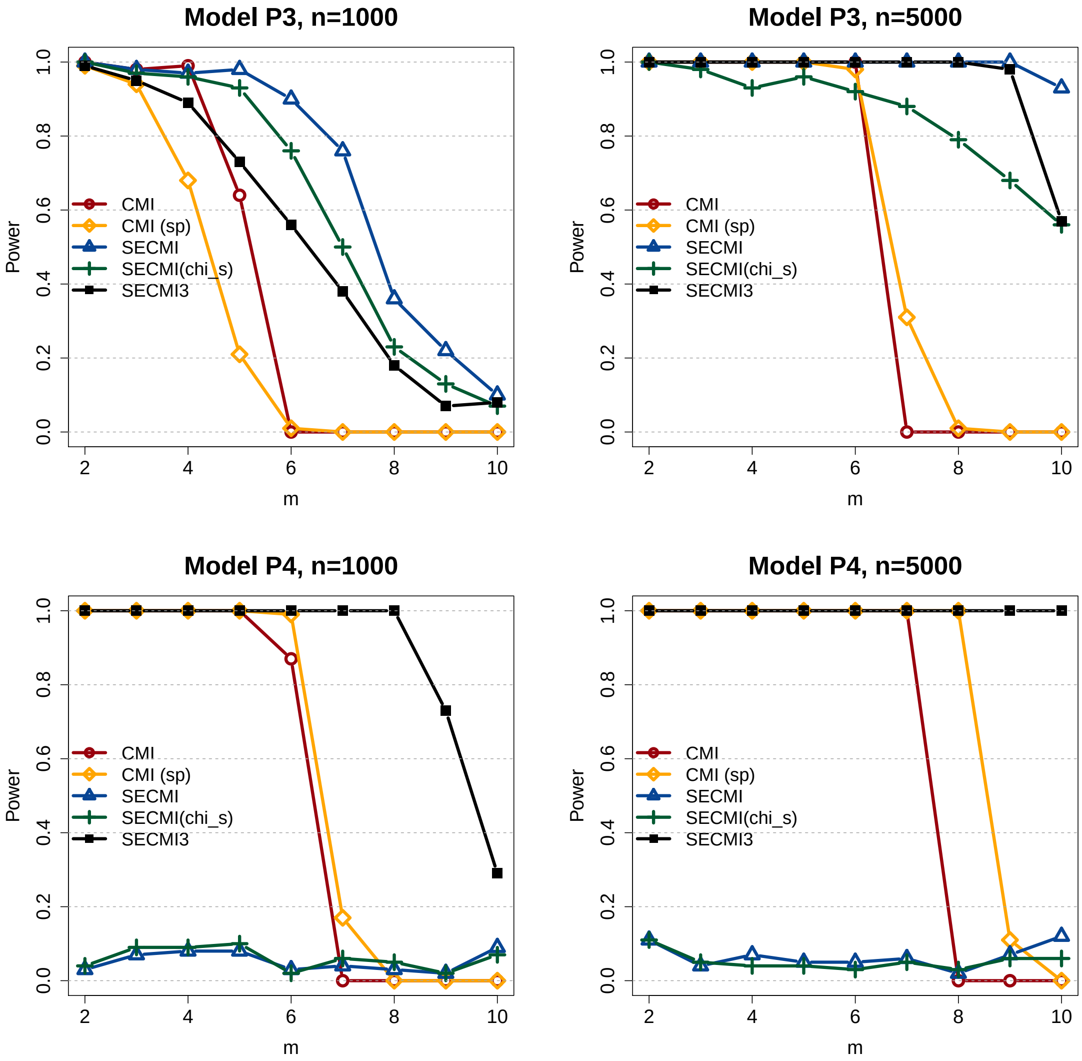


Figure 10: Power wrt to m for models P3 and P4 described in Section 6.1.1. Left-hand side panels correspond to sample size $n = 1000$, right-hand side panels correspond to sample size $n = 5000$.

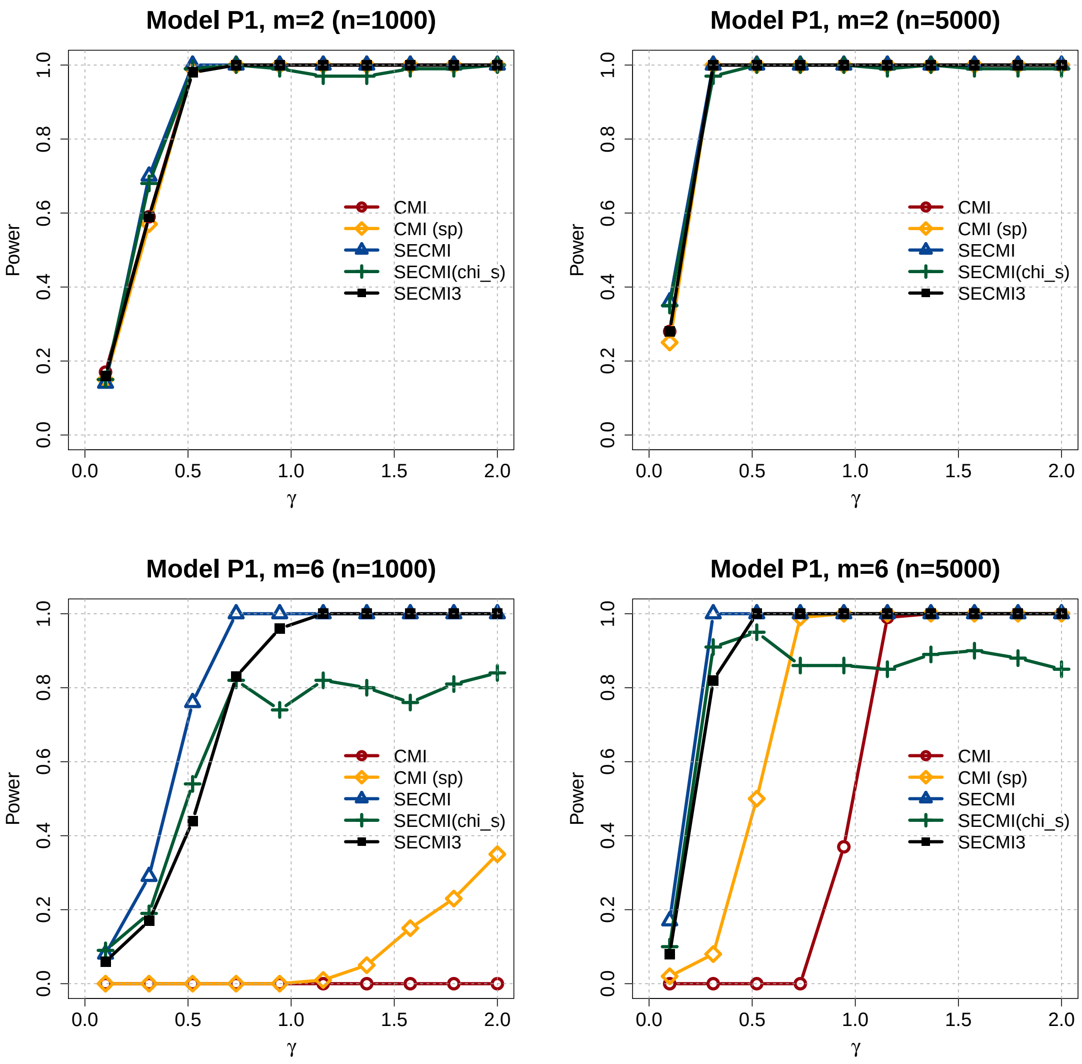


Figure 11: Power wrt to γ for model P1 described in Section 6.1.1, for sample sizes $n = 1000$ (left panels) and $n = 5000$ (right panels). Parameter γ controls the strength of conditional dependence between Y and X given $Z_1, \ldots, Z_m$. The first row corresponds to $m = 2$ and the second row corresponds to $m = 6$.

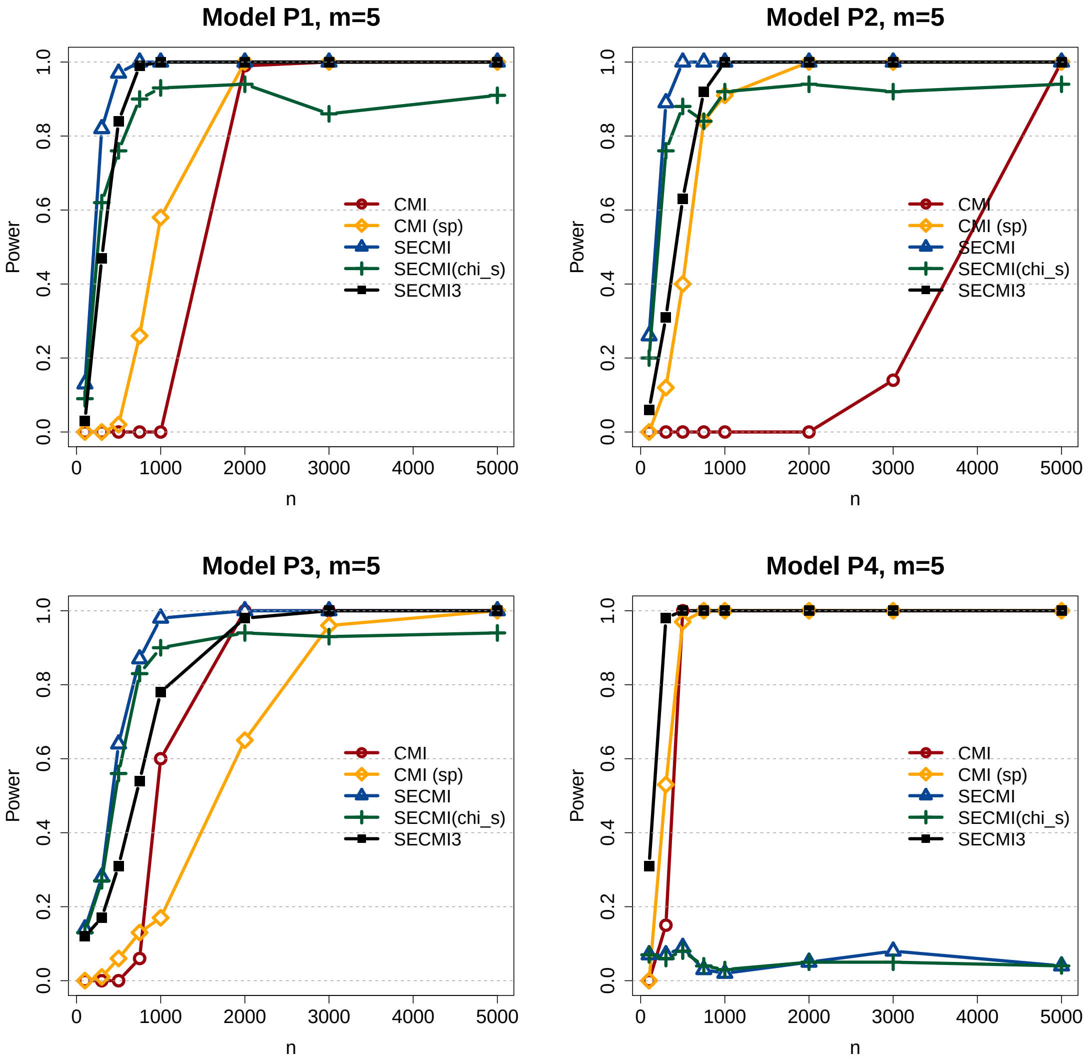


Figure 12: Power wrt to sample size n, for $m = 5$, for models P1, P2, P3 and P4 described in Section 6.1.1.

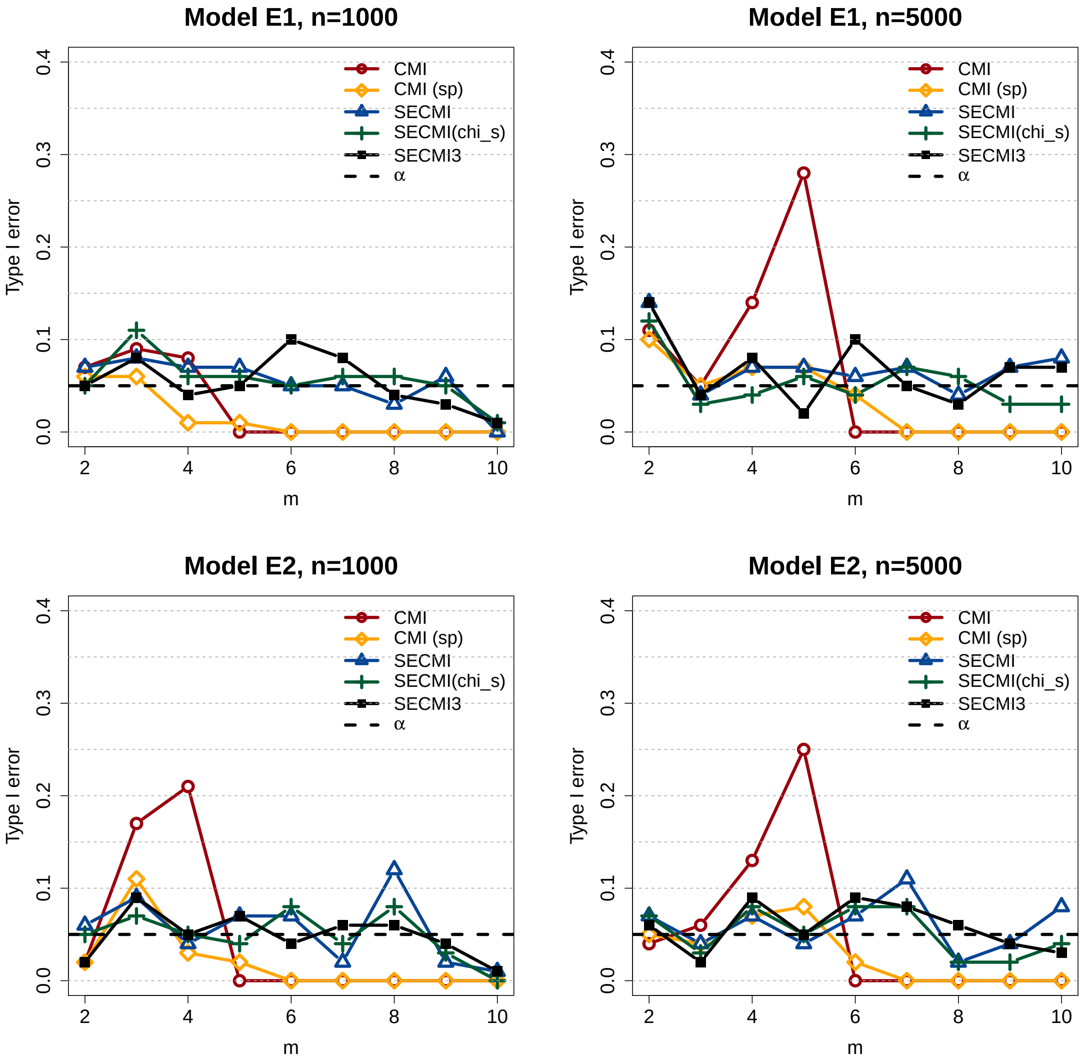


Figure 13: Type I error wrt to m for models E1 and E2 described in Section 6.1.1, for $\alpha = 0.05$. Left-hand side panels correspond to sample size $n = 1000$, right-hand side panels correspond to sample size $n = 5000$.

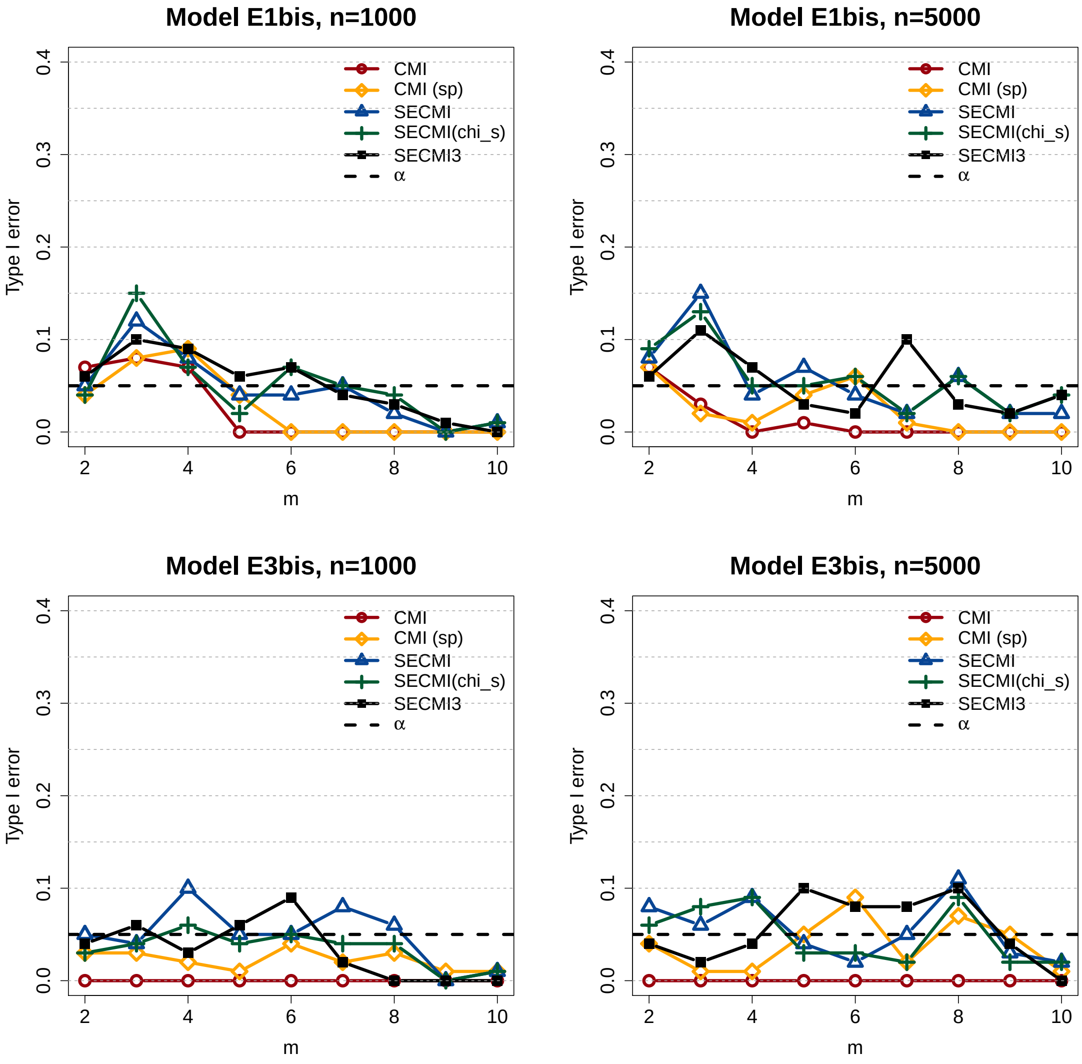


Figure 14: Type I error wrt to m for models E1bis and E3bis described in Section 6.1.1, for $\alpha = 0.05$. Left-hand side panels correspond to sample size $n = 1000$, right-hand side panels correspond to sample size $n = 5000$.

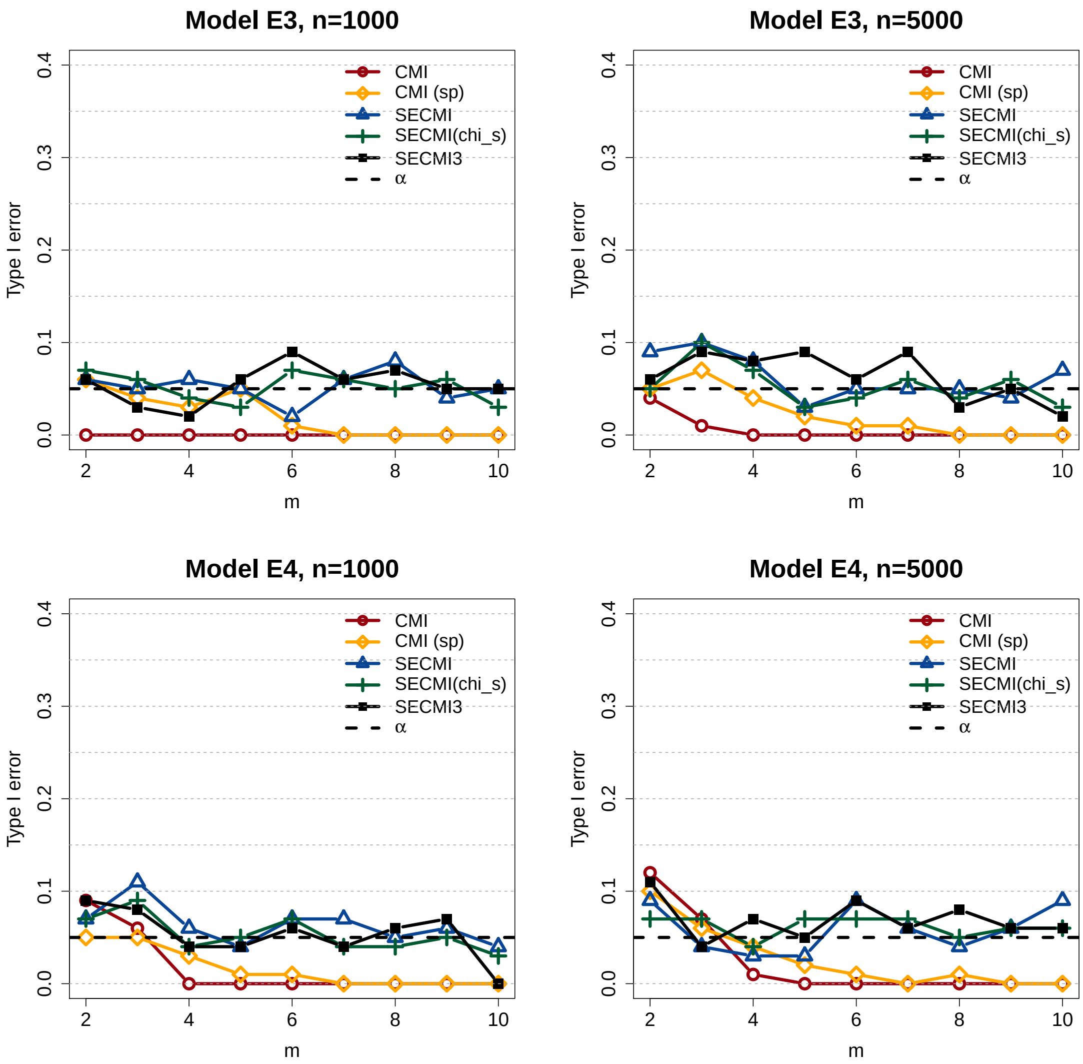


Figure 15: Type I error wrt to m for models E3 and E4 described in Section 6.1.1, for $\alpha = 0.05$. Left-hand side panels correspond to sample size $n = 1000$, right-hand side panels correspond to sample size $n = 5000$.

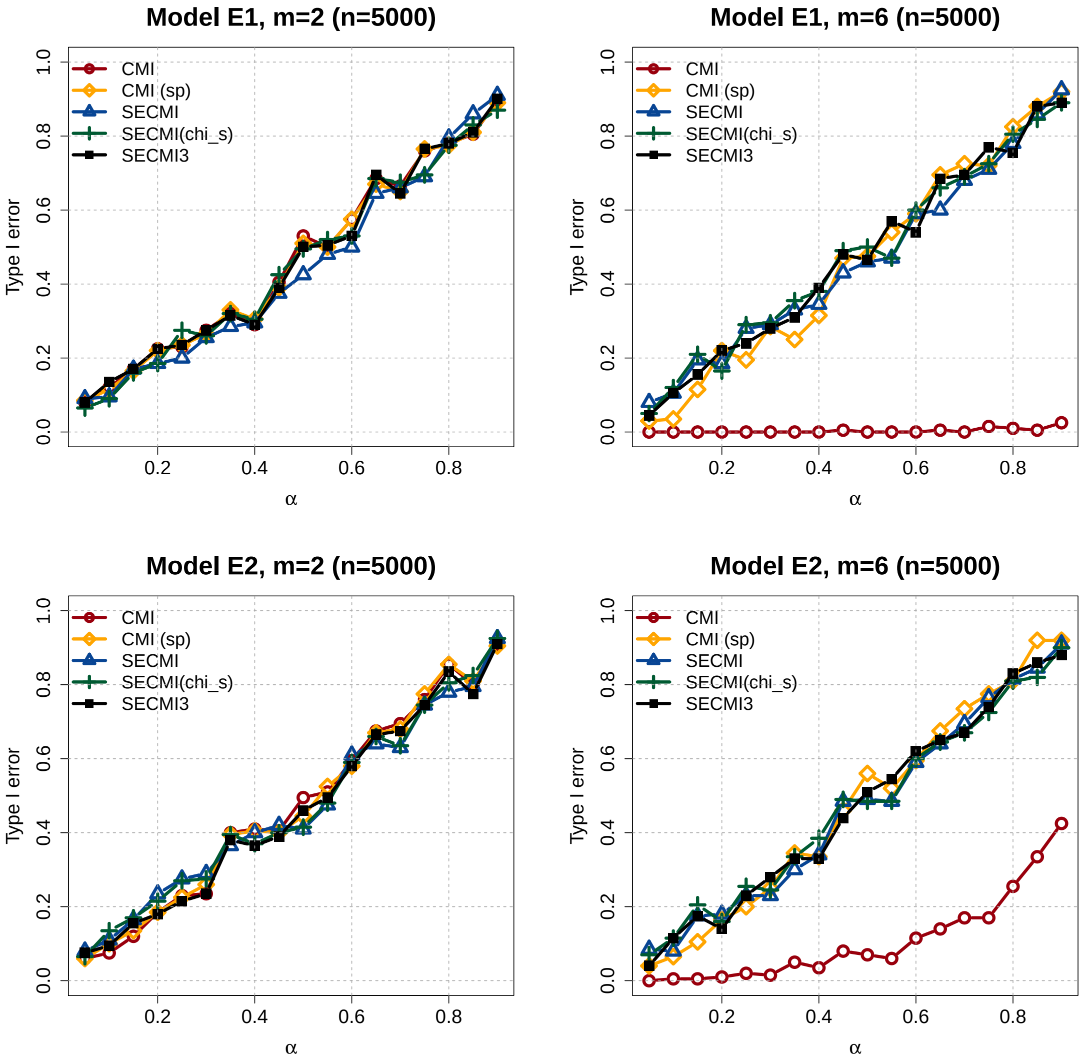


Figure 16: Type I error wrt to α for models E1 and E2 described in Section 6.1.1, for $n = 5000$. Left-hand side panels correspond to $m = 2$, right-hand side panels correspond to $m = 6$.

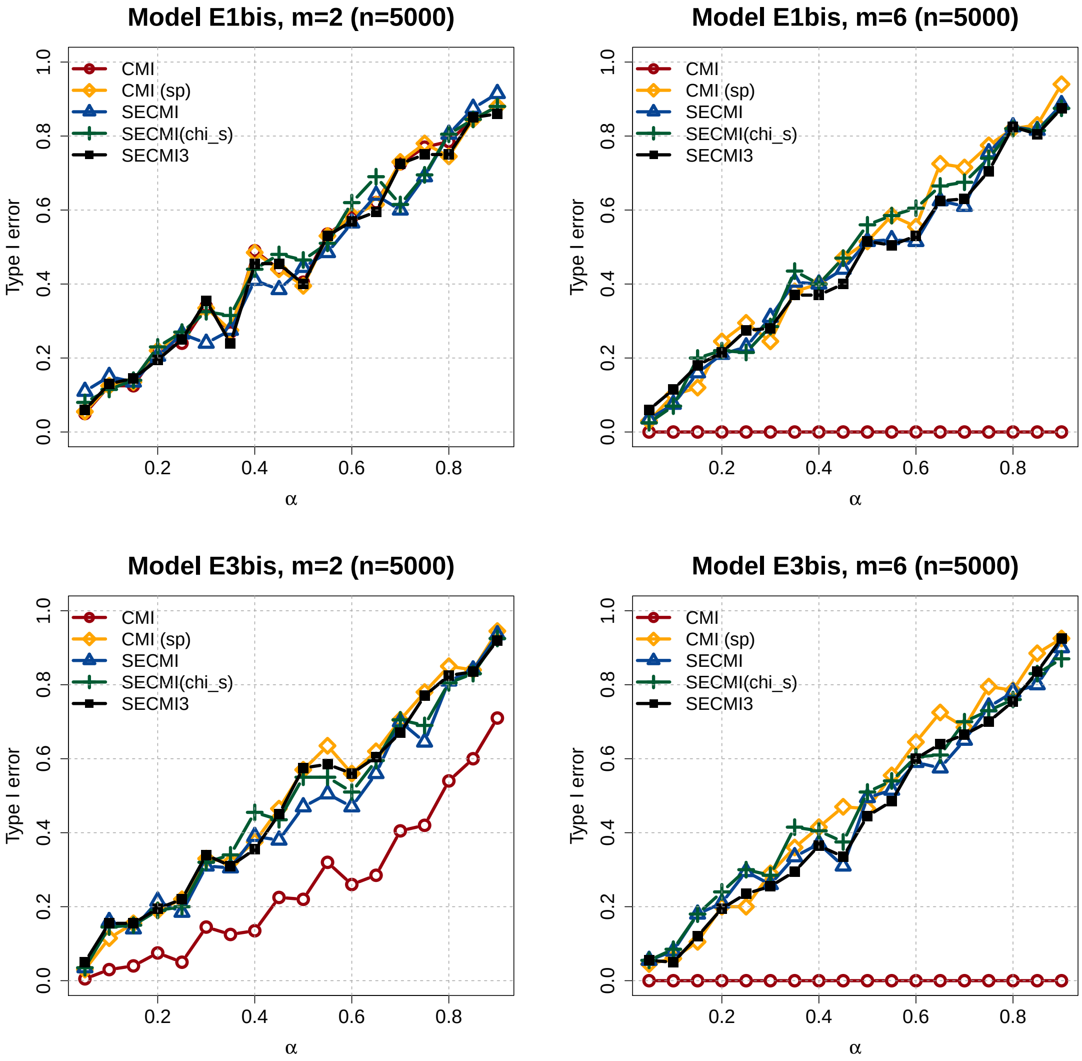


Figure 17: Type I error wrt to α for models E1bis and E3bis described in Section 6.1.1, for $n = 5000$. Left-hand side panels correspond to $m = 2$, right-hand side panels correspond to $m = 6$.

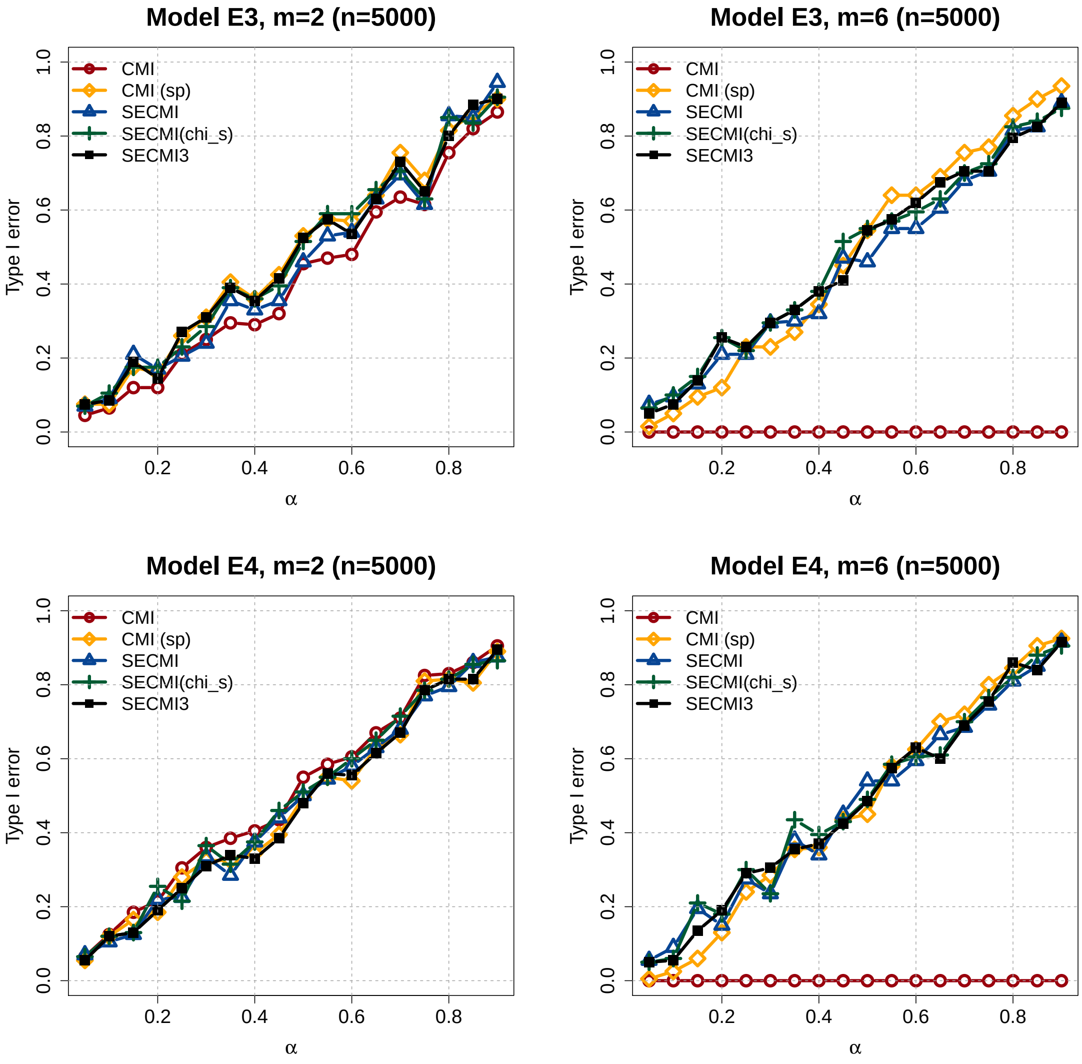


Figure 18: Type I error wrt to α for models E3 and E4 described in Section 6.1.1, for $n = 5000$. Left-hand side panels correspond to $m = 2$, right-hand side panels correspond to $m = 6$.

6.2 Application to Markov Blanket Discovery

Markov Blanket (MB) of the target variable Y is defined as the minimal subset $MB(Y)$ of variables $\{1,\ldots,p\}$ conditioned on which all other variables are independent of Y, i.e.

$$Y \perp (MB(Y))^c | MB(Y),$$

where A^c denotes the complement of a set A in $\{1,\ldots,p\}$. Note that Y is excluded from $\{1,\ldots,p\}$. In this Section we combine popular MB-discovery algorithms with the proposed conditional independence tests. In order to measure the performance of MB discovery algorithm, we need to know the true Markov Blanket to use it as a ground truth, which in practice is possible only for the simulated data. For this purpose, we consider databases sampled from known Bayesian Networks available at BN repository `http://www.bnlearn.com/bnrepository/`. The performance of each algorithm is assessed using the following evaluation measures. Let T be the true MB for the given variable and $\hat{T}$ be MB returned by the considered algorithm. We define three measures

$$\text{Recall} := \frac{|T \cap \hat{T}|}{|T|}, \quad \text{Precision} := \frac{|T \cap \hat{T}|}{|\hat{T}|}, \quad \text{F measure} := 2 \cdot \frac{\text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}}.$$

Recall is thus the fraction of the chosen relevant variables and all relevant variables. Note that Recall = 1 indicates that all relevant variables were selected. Precision is the ratio of the chosen relevant variables to all variables selected as relevant. A large value of Precision indicates that only few false positive variables are included in the chosen MB. F-measure is an aggregate measure of Recall and Precision defined as their harmonic mean. A large value of F-measure indicates that the chosen MB has relatively many ’true positives’ and at the same time few ’false positives’. Since we are mainly interested in the performance of proposed methods for large Markov Blankets, for each network we have chosen a node having the largest MB. Then we run the MB discovery algorithm, combined with the proposed test $SECMI$. As a reference we use two methods: the standard asymptotic test for CMI (denoted as CMI) and the semi-parametric test for CMI (denoted as $CMI(sp)$), proposed in Tsamardinos and Borboudakis (2010). We tested three popular MB discovery algorithms: GS, IAMB and MMPC. Since the performance of GS was slightly better than for IAMB and MMPC, we only present the results for GS. We consider 10 networks. Their basic characteristics are given in Table 1. Number of nodes varies from 5 to 76, whereas the size of the MB, for the chosen node ranges from 4 to 29.

Network	nodes	true MB size
asia	8	5
sachs	11	7
survey	6	4
alarm	37	8
hepar2	70	26

Network	nodes	true MB size
cancer	5	4
child	20	8
insurance	27	10
earthquake	5	4
win95pts	76	29

Table 1: Basic characteristics of networks: the total number of nodes and the size of true Markov Blanket for a chosen node.

Tables 2-5 show values of F measure (averaged over 200 repetitions), for different sample sizes: $n = 500, 1000, 3000, 5000$. The winner method for each data set is marked in bold, the last row contains averaged ranks (the lower the rank the better). Test $CMI(sp)$ combined with GS has the smallest averaged rank for $n = 500$, whereas $SECMI$ has the smallest averaged rank for $n = 1000, 3000, 5000$. As expected, CMI works much worse than $SECMI$ and $CMI(sp)$; $SECMI3$ works worse than $SECMI$ and $CMI(sp)$, but it is usually clearly better than CMI. This is consistent with the results from the previous section. Generally, for the considered data sets $CMI(sp)$ and $SECMI$ work on par: for most data sets the values of F measure corresponding to these two methods are very close.

To analyse in detail the results presented in the tables, we followed the two-step statistical procedure recommended by Demšar (2006). In the first step we use the Friedman test (based on averaged ranks) (Friedman, 1940) to assess the null hypothesis that all methods have equal performance. When the null hypothesis is rejected, a Nemenyi post-hoc test (Nemenyi, 1963) is used to compare the methods in a pairwise way. Figure 19 shows the results. For all considered sample sizes, the null hypothesis of the Friedman test is rejected, when the standard significance level 0.05 is assumed. The blue line denotes the Nemenyi critical region defined as mean rank ± Nemenyi critical distance. When two intervals intersect, then we conclude that there is no significant difference in performances between the methods. The critical region for the winner method is highlighted. The analysis confirms that there is no significant difference between the winner method $SECMI$ and the second best method $CMI(sp)$. However, both methods work significantly better than CMI.

To gain a deeper insight into behaviour of the methods, we analyse simple and popular network `cancer` in more detail. The left panel of Figure 20 visualizes the structure of the network. Variable *cancer*, denoting occurrence of the disease, is chosen as the target variable, whereas the remaining variables (*Pollution, Smoke, Xray, Dyspnoea*) constitute its MB. Right panel of Figure 20 shows the selection probabilities for the variables. Note that two variables (*Pollution, Dyspnoea*) are more often chosen by $SECMI$ than for CMI and $CMI(sp)$. Variable *Xray* is selected by all methods in almost all simulations, whereas variable *Smoker* is more often chosen by $SECMI$ and $CMI(sp)$ than by CMI.

	p	t	GS+CMI	GS+CMI(sp)	GS+SECMI	GS+SECMI3
asia	8	5	**0.571**	0.571	0.505	0.495
sachs	11	7	0.594	0.885	**0.895**	0.812
survey	6	4	**0.543**	0.482	0.520	0.459
alarm	37	8	0.280	0.505	**0.515**	0.392
hepar2	70	26	0.204	**0.283**	0.222	0.139
earthquake	5	4	0.680	**0.816**	0.766	0.773
cancer	5	4	0.589	**0.707**	0.688	0.703
insurance	27	10	0.462	**0.778**	0.766	0.611
mildew	27	10	0.462	**0.754**	0.747	0.606
win95pts	76	29	0.160	0.178	**0.198**	0.181
avg rank			3.4	**1.6**	2.0	3.0

Table 2: F-measure for n=500 and GS algorithm. Winner method is in bold. The last row shows the averaged ranks. Parameters p and t denote the total number of nodes and the size of true Markov Blanket, respectively.

	p	t	GS+CMI	GS+CMI(sp)	GS+SECMI	GS+SECMI3
asia	8	5	**0.571**	0.571	0.552	0.552
sachs	11	7	0.651	**0.964**	0.932	0.896
survey	6	4	0.641	0.683	**0.684**	0.650
alarm	37	8	0.402	0.585	**0.585**	0.428
hepar2	70	26	0.243	**0.348**	0.276	0.186
earthquake	5	4	0.842	0.893	**0.909**	0.907
cancer	5	4	0.771	0.789	**0.859**	0.830
insurance	27	10	0.479	**0.839**	0.819	0.654
mildew	27	10	0.492	0.829	**0.838**	0.695
win95pts	76	29	0.192	0.208	0.220	**0.222**
avg rank			3.6	2.0	**1.6**	2.8

Table 3: F-measure for n=1000 and GS algorithm. Winner method is in bold. The last row shows the averaged ranks. Parameters p and t denote the total number of nodes and the size of true Markov Blanket, respectively.

	p	t	GS+CMI	GS+CMI(sp)	GS+SECMI	GS+SECMI3
asia	8	5	**0.571**	0.571	0.571	0.571
sachs	11	7	0.727	**0.995**	0.963	0.959
survey	6	4	0.864	0.829	0.895	**0.909**
alarm	37	8	0.364	0.638	**0.687**	0.385
hepar2	70	26	0.311	**0.466**	0.425	0.327
earthquake	5	4	0.954	0.989	**0.994**	0.994
cancer	5	4	0.935	0.943	0.971	**0.977**
insurance	27	10	0.571	**0.881**	0.868	0.833
mildew	27	10	0.571	**0.880**	0.845	0.840
win95pts	76	29	0.236	0.287	**0.321**	0.266
avg rank			3.8	2.0	**1.8**	2.4

Table 4: F-measure for n=3000 and GS algorithm. Winner method is in bold. The last row shows the averaged ranks. Parameters p and t denote the total number of nodes and the size of true Markov Blanket, respectively.

	p	t	GS+CMI	GS+CMI(sp)	GS+SECMI	GS+SECMI3
asia	8	5	**0.571**	0.571	0.571	0.571
sachs	11	7	0.833	**0.997**	0.977	0.982
survey	6	4	0.944	0.927	**0.977**	0.938
alarm	37	8	0.364	0.687	**0.727**	0.395
hepar2	70	26	0.337	**0.508**	0.508	0.327
earthquake	5	4	**1.000**	1.000	1.000	1.000
cancer	5	4	0.977	0.989	0.983	**0.994**
insurance	27	10	0.632	**0.871**	0.862	0.861
mildew	27	10	0.648	0.875	**0.876**	0.863
win95pts	76	29	0.250	0.353	**0.390**	0.344
avg rank			3.4	2.0	**1.9**	2.7

Table 5: F-measure for n=5000 and GS algorithm. Winner method is in bold. The last row contains the averaged ranks. Parameters p and t denote the total number of nodes and the size of true Markov Blanket, respectively.

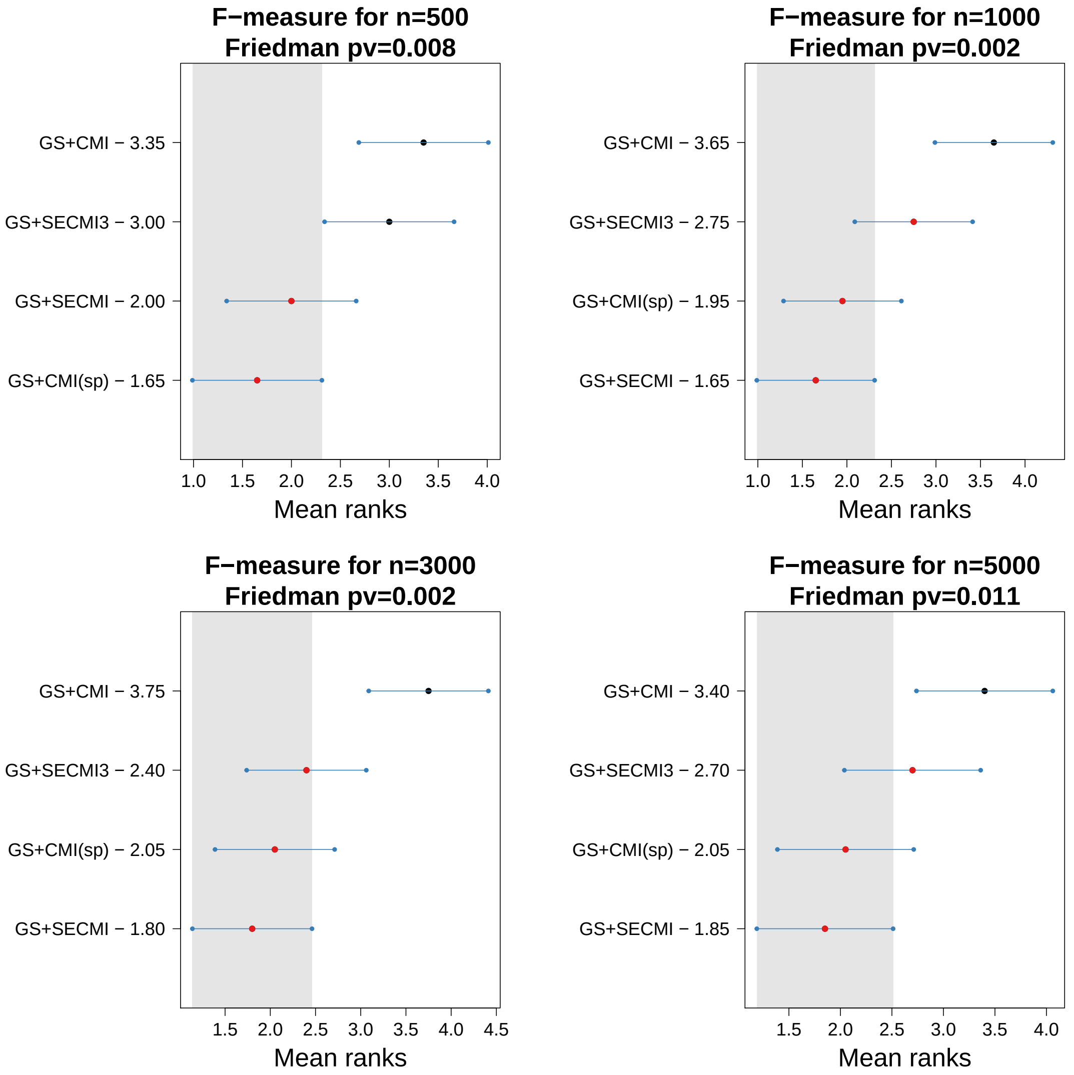
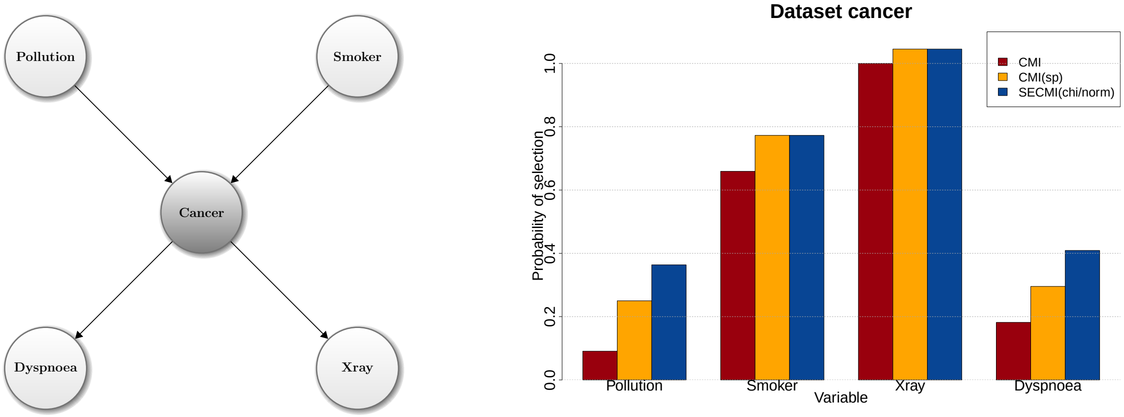


Figure 19: Results of Friedman and pairwise tests for F-measure and different sample sizes. The blue line denotes the Nemenyi critical region. The critical region for the winner method is highlighted.



(a) Dependence structure

(b) Selection probabilities

Figure 20: Data set cancer. LHS Figure: network visualizing dependency structure (target variable is marked in grey, MB is marked in white). RHS Figure: selection probability of true MB variables.

6.3 Application to Feature Selection and Classification

The discussed conditional independence tests can be also used for the feature selection in supervised classification. Algorithms of Markov Blanket discovery (such as GS, IAMB or others) combined with the conditional independence tests allow to choose a subset of relevant features in supervised learning. As in the previous section, we use GS algorithm combined with 4 tests: CMI, $CMI(sp)$, $SECMI$ and $SECMI3$.

We consider 12 data sets from UCI machine learning repository (Dheeru and Taniskidou, 2017). These data sets are chosen to represent various characteristics. Most of them were already used in the related studies on the feature selection, see e.g. Brown et al. (2012). Table 6 shows the basic statistics of the data sets: the number of observations n, features p, ratio p/n and the number of classes. The quantitative features are discretized into 2 bins, whereas the discrete features are left intact.

In order to assess the performance of the considered conditional independence tests we build a classifier using features selected by the considered methods and then report its classification performance. We use two popular evaluation measures: the accuracy (i.e. the fraction of correctly classified elements) and the balanced accuracy (the average accuracy obtained on all classes), which is a more appropriate measure for imbalanced data sets. We used kNN classifier for classification task. The kNN classifier is a generic method which avoids making any assumptions about the data and moreover it requires tuning only one parameter (the number of nearest neighbours k). For this reason it was used by several authors to compare the classification performance of feature selection methods (Brown et al., 2012). We also experimented with other classifiers (e.g. decision trees, random forests and multinomial/logistic regression) but their accuracies were similar for all considered methods and therefore they are not presented.

To estimate the classification accuracy for classifiers with optimally chosen parameters we perform the following steps. First we split data into three parts: a training set (50%), a validation set (25%) and a testing set (25%). The feature selection methods are launched on the training set and then the classifier is built on the training set using selected features. The above step is repeated for different values of hyper-parameters: we optimize over $k = 1, 2, \ldots, 10$ (number of nearest neighbours in kNN) and $\alpha = 0.001, 0.01, 0.05, 0.1$ (significance level of CMI tests). Validation set is used to choose the optimal values of hyper-parameters with respect to the considered evaluation measure. Finally, the evaluation measures (the accuracy and the balanced accuracy) are calculated on the testing set. The above steps are repeated for 50 random data splits.

Tables 7 and 8 show respectively the accuracy and the balanced accuracy (averaged over 50 data splits) for the considered data sets. The last row contains the averaged ranks (the lower the rank the better). Observe that $SECMI$ and $CMI(sp)$ work similarly, however $SECMI$ is the winner for most of the data sets, followed by $CMI(sp)$ and $SECMI3$. The CMI works clearly worse than the competitors for some data sets. Indeed, for madelon data set, the accuracy for CMI is 46% whereas the accuracy for $SECMI$ and $SECMI3$ oscillates around 84%. To analyse in detail the results presented in the tables, we followed the two-step statistical procedure based on Friedman and Nemenyi tests, already described in Section 6.2. Figure 21 shows the results. For both evaluation measures, the null hypothesis of the Friedman test is rejected, when a standard significance level 0.05 is assumed. The critical region for winner method $SECMI$ is highlighted. For the accuracy, there is no significant difference between $SECMI$ and the second and the third best method ($CMI(sp)$ and $SECMI3$). However, $SECMI$ works significantly better than CMI. For the balanced accuracy, there is no significant difference between the winner method $SECMI$ and the second best $CMI(sp)$. Moreover, $SECMI$ performs significantly better than $SECMI3$ and CMI. The worse performance of $SECMI3$ when compared to $SECMI$ may be associated with the lack of higher-order interactions in the considered data sets.

	n	p	p/n	classes
glass	214	9	0.04	6
wdbc	569	31	0.05	2
credit-a	690	38	0.06	2
sonar	208	60	0.29	2
diabetes	768	8	0.01	2
heart-c	303	19	0.06	2
waveform-5000	5000	40	0.01	3
vehicle	846	18	0.02	4
ionosphere	351	34	0.10	2
credit-g	1000	48	0.05	2
prostate	102	6033	59.15	2
madelon	2600	500	0.19	2

Table 6: Summary statistics of real data sets used in Section 6.3.

	CMI	CMI(SP)	SECMI	SECMI3
credit-g	0.678	**0.692**	0.676	0.670
diabetes	0.728	0.726	**0.732**	0.729
glass	0.416	0.580	**0.604**	0.604
heart-c	0.762	0.777	**0.793**	0.760
ionosphere	0.868	**0.889**	0.862	0.856
madelon	0.464	0.559	**0.841**	0.845
waveform-5000	0.656	0.780	**0.782**	0.781
vehicle	0.587	0.650	**0.664**	0.583
wdbc	0.898	**0.919**	0.915	0.919
credit-a	0.838	0.840	**0.846**	0.849
prostate	0.868	**0.880**	0.880	0.880
sonar	0.587	0.603	**0.615**	0.548
avg. rank	3.3	2.3	**1.8**	2.5

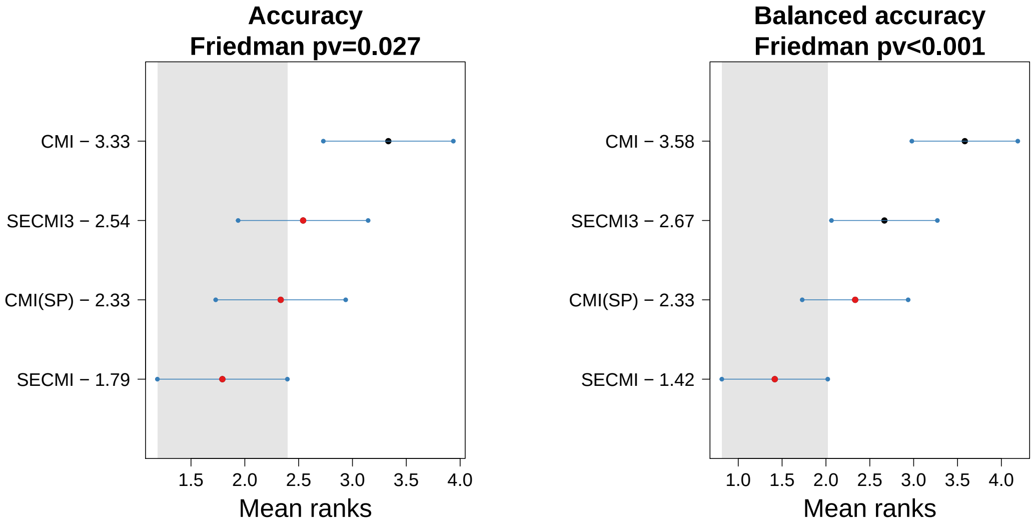
Table 7: Accuracy averaged over 50 random splits of data. The last row shows the averaged ranks (the lower the rank the better).

	CMI	CMI(SP)	SECMI	SECMI3
credit-g	0.579	0.592	**0.596**	0.589
diabetes	0.683	0.690	**0.696**	0.688
glass	0.484	**0.672**	0.638	0.694
heart-c	0.756	0.774	**0.796**	0.760
ionosphere	0.840	**0.863**	0.838	0.814
madelon	0.463	0.559	**0.841**	0.844
waveform-5000	0.598	0.771	**0.772**	0.758
vehicle	0.568	0.652	**0.658**	0.568
wdbc	0.888	0.905	**0.908**	0.906
credit-a	0.841	0.838	**0.845**	0.842
prostate	0.871	0.884	**0.890**	0.882
sonar	0.603	0.595	**0.613**	0.568
avg. rank	3.6	2.3	**1.4**	2.7

Table 8: Balanced accuracy averaged over 50 random splits of data. The last row shows the averaged ranks (the lower the rank the better).

6.4 Summary of Experiments

In the following we summarize the results of experiments and give guidelines which methods to use in specific situations. In the experiments we compared the tests ($SECMI$, $SECMI3$, $CMI(sp)$ and CMI) of the null hypothesis $H_0 : X \perp Y|Z_1, \ldots, Z_m$. We studied how the power and the type I error depend on various parameters such as the size of the conditioning set m, the sample size n, among others. All methods, except CMI, control the type I error well. The proposed method $SECMI$ achieves the highest power when the high-order



(a) (b)

Figure 21: Results of Friedman and pairwise tests for accuracy (a) and balanced accuracy (b). The blue line denotes the Nemenyi critical region. The critical region for $SECMI$ is highlighted.

interactions (of order larger than 2) are not present and the number of possible values of conditioning variables b^m (assuming that all variables are discretized using b bins and m is the number of conditioning variables), relative to sample size n, is large. When the high-order interactions exist, $SECMI$ can be modified in order to take the interactions into account ($SECMI3$ is a solution which takes into account the 3-rd order interactions). As expected, when the higher order interactions exist, CMI and $CMI(sp)$ work better than $SECMI$, but, importantly, they also fail to detect the true conditional dependence when m grows. In view of the above, we recommend to use the $SECMI$ criterion when the ratio b^m/n is large (as in this case CMI won't work anyway whereas $CMI(sp)$ has usually a smaller power than $SECMI$) and $CMI(sp)$ or $SECMI3$ when b^m/n is small. Obviously, if we do not expect higher-order interactions in the analysed data, we recommend to use $SECMI$, regardless of the size of conditioning set. A simple method of choosing a threshold for the ratio b^m/n can be based on the minimal averaged number of observations required for each combination of conditioning variables. For example, if the required number of observations is 10 (i.e. $n/b^m \geq 10$) then we use $CMI(sp)$ for $n/b^m > 10$ ($b^m/n < 0.1$) and $SECMI$ otherwise. The experiments on synthetic data sets indicate that threshold 10 is a reasonable choice. For example, Figures 9 and 10 show that, for $n = 1000$ and $b = 3$, performance of $CMI(sp)$ usually deteriorates for $m > 4$. For $m > 4$, the average number of observations for each combination of conditioning variables is smaller than 10 (we have $n/b^m \approx 12$ for $m = 4$ and $n/b^m \approx 4$ for $m = 5$).

In the experiments we also the combined conditional independence tests with MB discovery algorithms. We focused on two problems. First, we analysed how accurately we can recover the true MB based on the data sets sampled from known Bayesian Networks.

Secondly, we analysed the classification performance based on the real benchmark data sets. Although in the two experiments it is not possible to measure differences in performance between conditional independence tests directly, they are important from the practical point of view. The proposed method $SECMI$ works usually on par or slightly better than $CMI(sp)$ and clearly better than CMI. The other proposed method $SECMI3$ works usually worse than $SECMI$ which may indicate that the high-order interactions are not very strong in the considered problems.

7. Conclusions and Future Work

In this paper we proposed the novel method for testing the conditional independence, based on $SECMI$, which is an empirical version of the truncated Möbius expansion of CMI. We derived its asymptotic distribution and analysed in detail the types of dependence structures which lead to a non-normal asymptotic distribution. Importantly, we have shown in numerical analysis that the $SECMI$-based tests achieve significantly larger power than tests using conditional mutual information, while controlling the type I error, especially when the size of the conditioning set is large. It also follows that in this case CMI will fail to discover important active features. As a consequence, we witnessed superior behaviour of GS methods for Markov Blanket discovery for some Bayesian networks in the terms of F-measure when CMI is replaced by $SECMI$. Compared with $CMI(sp)$, the $SECMI$ has in general larger power when testing the conditional independence but its superiority decreases when the methods are used for the feature selection and Markov Blanket discovery.

A drawback of the proposed method is that it fails to discover the interactions of order larger than 2, although the experiments suggest that the method can be extended to account for higher order interactions. Note however that finding a trade-off between including the higher-order interactions and controlling the size of conditioning set is a challenging and unresolved issue.

The $SECMI$-based method could be also possibly improved by considering alternative switches between normal and non-normal case in Algorithm 1 (such as the weighted supremum norm instead of Kolmogorov-Smirnov distance) and investigating different approximations to the distribution of the quadratic form other than those considered here. Lastly, it is of interest to extend the results of the paper to the case of time series data (for conditional independence tests designed for time series data and continuous variables we refer to Su and White 2007, Su and White 2014 and Wang and Hong 2018). We conjecture that the analogues of Theorems 2 and 3 still hold (with e.g. $\sigma^2_{\widehat{CMI}}$ changed to $\kappa(0) + 2\sum_{k=1}^{\infty}\kappa_k$, where $\kappa_k = \mathrm{Cov}(I_i, I_{i+k})$ and

$$I_i = \log\Big(\frac{p(X_i, Y_i, Z_i)p(Z_i)}{p(X_i, Y_i)p(Z_i)}\Big)$$

for strictly stationary data as suggested by the asymptotic distribution of the sample mean. However, it is an open problem how to construct appropriate permutation scheme for such a case.

Acknowledgement

We acknowledge insightful comments of three referees and the Associate Editor which led to the substantial improvement of the original manuscript.

Appendix A. Sample Size Calculations (Inequality (12))

Re-expressing the condition defining critical region $\mathcal{C}$ we have

$$\begin{aligned} P(\hat{I}(X,Y|Z) \geq t_\alpha) &= P\left(\sqrt{n}\frac{\hat{I}(X,Y|Z) - I(X,Y|Z)}{\sigma_{CMI}} \geq \sqrt{n}\frac{t_\alpha - I(X,Y|Z)}{\sigma_{CMI}}\right) \\ &\approx 1 - \Phi\left(\sqrt{n}\frac{(t_\alpha - I(X,Y|Z))}{\sigma_{CMI}}\right). \end{aligned}$$

Thus the approximate power of CMI test is:

$$\tilde{P}(\alpha) = 1 - \Phi\left(\sqrt{n}\frac{(t_\alpha - I(X,Y|Z))}{\sigma_{CMI}}\right). \tag{34}$$

We want to find minimal sample size n, under which $\tilde{P}(\alpha) \geq \beta$. This inequality is equivalent to:

$$\sqrt{n}\frac{(t_\alpha - I(X,Y|Z))}{\sigma_{CMI}} \leq \Phi^{-1}(1-\beta). \tag{35}$$

After rearranging terms we obtain:

$$2nI(X,Y|Z) + 2\sigma_{CMI}\sqrt{n}\Phi^{-1}(1-\beta) - F^{-1}_{\chi^2_d}(1-\alpha) \geq 0. \tag{36}$$

and solving this quadratic inequality we obtain (12):

$$n \geq \left(\frac{\sqrt{4\sigma^2_{CMI}\left(\Phi^{-1}(1-\beta)\right)^2 + 8I(X,Y|Z)F^{-1}_{\chi^2_d}(1-\alpha)} - 2\sigma_{CMI}\Phi^{-1}(1-\beta)}{4I(X,Y|Z)}\right)^2. \tag{37}$$

Appendix B. Proof of Theorem 1

Proof Let $\hat{p}(x,y,z) = \#\{i : (X_i,Y_i,Z_i) = (x,y,z)\}/n$ be plug-in estimator for $p(x,y,z)$ and $\mathbf{p} = (p(x,y,z))_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}}$ be $I \times J \times K$ vector of probabilities. We write $I(X,Y|Z)$ as a function of $\mathbf{p}$, namely

$$I(X,Y|Z) = \sum_{x,y,z} p(x,y,z)\log\left(\frac{p(x,y,z)p(z)}{p(x,z)p(y,z)}\right) =: f(\mathbf{p}).$$

Observe that $\hat{I}(X,Y|Z) = f(\hat{\mathbf{p}})$. We have

$$\frac{\partial f(\mathbf{p})}{\partial p(x,y,z)} = \log\left(\frac{p(x,y,z)p(z)}{p(x,z)p(y,z)}\right), \tag{38}$$

$$\begin{aligned} \frac{\partial^2 f(\mathbf{p})}{\partial p(x,y,z)\partial p(x',y',z')} = \frac{I(x=x',y=y',z=z')}{p(x,y,z)} - \frac{I(x=x',z=z')}{p(x,z)} \\ - \frac{I(y=y',z=z')}{p(y,z)} + \frac{I(z=z')}{p(z)}, \end{aligned}$$

where $I(A)$ is an indicator of set A. We use delta method (see e.g. Agresti 2002) which relies on second order Taylor expansion:

$$f(\hat{\mathbf{p}}) - f(\mathbf{p}) = Df(\mathbf{p})^T(\hat{\mathbf{p}} - \mathbf{p}) + \frac{1}{2}(\hat{\mathbf{p}} - \mathbf{p})^T D^2 f(\mathbf{p})(\hat{\mathbf{p}} - \mathbf{p}) + O(||\hat{\mathbf{p}} - \mathbf{p}||_2^3). \tag{39}$$

Note that the remainder term equals to the value of trilinear form pertaining to the third derivative of $f(p)$ calculated at $\tilde{p}$, where $\tilde{p}$ is some point in-between p and $\hat{p}$. As the form is a bounded operator it is easily seen that the remainder term is of the given order. Moreover, we have that an element of $\Sigma = n\,\mathrm{Var}(\hat{\mathbf{p}} - \mathbf{p})$ with row index xyz and column index $x'y'z'$ is

$$\Sigma_{xyz}^{x'y'z'} = p(x', y', z')(I(x = x', y = y', z = z') - p(x, y, z)).$$

It is easy to check (see Agresti 2002, Section 14.1.4 for the case of general f) that

$$\begin{aligned}
&n\mathrm{Var}(Df(\mathbf{p})^T(\hat{\mathbf{p}} - \mathbf{p})) = \\
&\sum_{x,y,z} p(x,y,z)\log^2\Big(\frac{p(x,y,z)p(z)}{p(x,z)p(y,z)}\Big) - \Big(\sum_{x,y,z} p(x,y,z)\log\Big(\frac{p(x,y,z)p(z)}{p(x,z)p(y,z)}\Big)\Big)^2 = \\
&\mathrm{Var}\Big(\log\Big(\frac{p(X,Y,Z)p(Z)}{p(X,Z)p(Y,Z)}\Big)\Big).
\end{aligned} \tag{40}$$

This ends the proof of part (i) as $I(X,Y|Z) \neq 0$ implies that $p(x,y,z)p(z)/p(x,z)p(y,z)$ is not constant and the variance above is not zero and thus the first term on RHS of (39) dominates.

In order to prove (ii) note that from assumption $I(X,Y|Z) = 0$ it follows that $Df(\mathbf{p}) = 0$. As Central Limit Theorem Implies $\sqrt{n}(\hat{\mathbf{p}} - \mathbf{p}) \xrightarrow{d} N(0,\Sigma)$ we have from (39) that

$$2nf(\hat{\mathbf{p}}) \xrightarrow{d} N(0,\Sigma)^T D^2 f(\mathbf{p}) N(0,\Sigma) = N(0,I)^T \Sigma^{1/2} D^2 f(\mathbf{p}) \Sigma^{1/2} N(0,I). \tag{41}$$

Since eigenvalues of $\Sigma^{1/2} D^2 f(\mathbf{p}) \Sigma^{1/2}$ coincide with those of $D^2 f(\mathbf{p})\Sigma =: M$ it follows that

$$2nf(\hat{\mathbf{p}}) \xrightarrow{d} \sum_{x,y,z} \lambda_{xyz}(M) Z_i, \tag{42}$$

where $\lambda_{xyz}(M)$ are eigenvalues of M and Z_i are independent χ_1^2-distributed random variables. Straightforward calculations yield:

$$\begin{aligned}
M_{xyz}^{x'y'z'} &= \sum_{x'',y'',z''} \Bigg(\frac{I(x = x'', y = y'', z = z'')}{p(x,y,z)} - \frac{I(x = x'', z = z'')}{p(x,z)} \\
&\quad - \frac{I(y = y'', z = z'')}{p(y,z)} + \frac{I(z = z'')}{p(z)}\Bigg) \\
&\quad \times p(x'', y'', z'')(I(x' = x'', y' = y'', z' = z'') - p(x', y', z')) \\
&= I(x = x', y = y', z = z') - I(x = x', z = z')\frac{p(x, y', z)}{p(x,z)} \\
&\quad - I(y = y', z = z')\frac{p(x', y, z)}{p(y,z)} + I(z = z')\frac{p(x', y', z)}{p(z)}.
\end{aligned}$$

As X and Y are independent given Z, the above formula reduces to:

$$M_{xyz}^{x'y'z'} = I(z=z')\left(I(x=x') - \frac{p(x',z)}{p(z)}\right)\left(I(y=y') - \frac{p(y',z)}{p(z)}\right).$$

We note now that in this case matrix M is idempotent, as we have:

$$\begin{aligned}(M^2)_{xyz}^{x'y'z'} &= \sum_{x'',y'',z''} I(z=z'=z'')\left(I(x=x'') - \frac{p(x'',z)}{p(z)}\right)\left(I(x'=x'') - \frac{p(x',z)}{p(z)}\right)\\ &\quad\times\left(I(y=y'') - \frac{p(y'',z)}{p(z)}\right)\left(I(y'=y'') - \frac{p(y',z)}{p(z)}\right)\\ &= \sum_{z''} I(z=z'=z'')\sum_{x''}\left(I(x=x'') - \frac{p(x'',z)}{p(z)}\right)\left(I(x'=x'') - \frac{p(x',z)}{p(z)}\right)\\ &\quad\times\sum_{y''}\left(I(y=y'') - \frac{p(y'',z)}{p(z)}\right)\left(I(y'=y'') - \frac{p(y',z)}{p(z)}\right)\\ &= I(z=z')\left(I(x=x') - \frac{p(x',z)}{p(z)}\right)\left(I(y=y') - \frac{p(y',z)}{p(z)}\right) = M_{xyz}^{x'y'z'}.\end{aligned}$$

Hence $M^2 = M$ and the only possible eigenvalues for M are 0 and 1. Now we compute $trM = \sum_i M_{ii}$:

$$\begin{aligned}\operatorname{tr} M &= \sum_{x,y,z}\left(1 - \frac{p(x,y,z)}{p(x,z)} - \frac{p(x,y,z)}{p(y,z)} + \frac{p(x,y,z)}{p(z)}\right)\\ &= |\mathcal{X}||\mathcal{Y}||\mathcal{Z}| - |\mathcal{X}||\mathcal{Z}| - |\mathcal{Y}||\mathcal{Z}| + |\mathcal{Z}| = (I-1)(J-1)K.\end{aligned}$$

This means that the number of eigenvalues equal 1 is $(I-1)(J-1)K$. Thus it follows from (42) that:

$$2nf(\hat{\mathbf{p}}) = 2n\widehat{I}(X,Y|Z) \xrightarrow{d} \chi^2_{(I-1)(J-1)K}.$$

■

Appendix C. Proof of Theorem 2

Proof We introduce more general notation: $Z = (Z_{s_1},\ldots,Z_{s_{|S|}})$, $S = \{s_1,\ldots,s_{|S|}\}$ and $|S|$ is dimensionality of S. We put $z = (z_{s_1},\ldots,z_{s_{|S|}})$. Let $p(x,y,z) = P(X=x,Y=y,Z=z)$, $p(x) = P(X=x)$, $p(y,z) = P(Y=y,Z=z), p(y,z_s) = P(Y=y,Z_s=z_s)$ and $\hat{p}(x,y,z) = n(x,y,z)/n$ be plug-in estimator for $p(x,y,z)$. Then we have:

$$SECMI(X,Y|Z) = \sum_{x,y,z} p(x,y,z)\log\left(\left(\frac{p(x,y)}{p(x)p(y)}\right)^{1-|S|}\prod_{s\in S}\frac{p(x,y,z_s)p(z_s)}{p(x,z_s)p(y,z_s)}\right) =: f(\mathbf{p}),$$

where $\mathbf{p} = (p(x,y,z))_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}_S}$. Then $\widehat{SECMI}(X,Y|Z) = f(\hat{\mathbf{p}})$. Moreover,

$$\frac{\partial f(\mathbf{p})}{\partial p(x,y,z)} = (1-|S|)\left(\log\left(\frac{p(x,y)}{p(x)p(y)}\right) - 1\right) + \sum_{s\in S}\log\left(\frac{p(x,y,z_s)p(z_s)}{p(x,z_s)p(y,z_s)}\right). \tag{43}$$

$$\frac{\partial^2 f}{\partial p(x,y,z)\partial p(x'y'z')} = (1-|S|)\left(\frac{I(x=x',y=y')}{p(x,y)} - \frac{I(x=x')}{p(x)} - \frac{I(y=y')}{p(y)}\right)$$
$$+\sum_{s\in S} I(z_s = z'_s)\left(\frac{I(x=x',y=y')}{p(x,y,z_s)} - \frac{I(x=x')}{p(x,z_s)} - \frac{I(y=y')}{p(y,z_s)} + \frac{1}{p(z_s)}\right). \quad (44)$$

It is easy to see that

$$\sum_{x,y,z}\Big(\frac{\partial f}{\partial p(x,y,z)}\Big)^2 p(x,y,z) - \Big(\sum_{x,y,z}\frac{\partial f}{\partial p(x,y,z)}p(x,y,z)\Big)^2 = \sigma^2_{\widehat{SECMI}}.$$

As in proof of Theorem 1 we have that when $\sigma^2_{SECMI} > 0$ then

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

and in the opposite case we have

$$2n(\widehat{SECMI} - SECMI) \xrightarrow{d} Z'HZ = \tilde{Z}'M\tilde{Z},$$

where Z has $N(0,\Sigma)$ distribution and H is a Hessian defined in (44), $\tilde{Z}$ has $N(0,I)$ distribution and $M = \Sigma^{1/2}H\Sigma^{1/2}$.

■

We prove the following Lemma which is instrumental in establishing Theorem 4.3. The following remark is in order. Note that the condition $P(X=x,Y=y)/P(X=x)P(Y=y)$ is constant implies that X and Y are independent, however, the conditional version of this statement is not true as the fact that $P(X=x,Y=y|Z=z)/P(X=x|Z=z)P(Y=y|Z=z)$ does not depend on z does not necessarily imply that X and Y are not necessarily independent given Z. Indeed, the second possibility exists namely that Y and Z are conditionally independent given X and Y and Z are unconditionally independent.

Lemma 1 *Let $Y \in \{0,1\}$ be a binary random variable and $X, Z \in N_+ = N\setminus\{0\}$ be discrete variables. (i) If for all $y \in \{0,1\}$ and $x, z \in N_+$ we have:*

$$\frac{P(X=x,Y=y|Z=z)}{P(X=x|Z=z)P(Y=y|Z=z)} = a_{xy}, \quad (45)$$

where $a_{xy} > 0$ does not depend on z, then at least one of the following possibilities holds:

1. *X and Y are conditionally independent given Z and $a_{xy} = 1$ for all x, y .*

2. *Y and Z are independent and Y and Z are conditionally independent given X, for all x, y:*

$$a_{xy} = \frac{P(X=x,Y=y)}{P(X=x)P(Y=y)},$$

where $a_{xy} \neq 1$ for some x, y (hence X and Y are dependent).
(ii) Conversely, if (i)1. or (i)2. holds than (45) is valid.

Proof First we observe that for all $x, z \in N_+$ we have:

$$\sum_{y=0}^{1} a_{xy} P(Y=y, Z=z) = P(Z=z) \sum_{y=0}^{1} a_{xy} P(Y=y|Z=z)$$
$$= P(Z=z) \sum_{y=0}^{1} \frac{P(X=x, Y=y|Z=z)}{P(X=x|Z=z)} = P(Z=z). \tag{46}$$

Thus for all x we have:

$$\sum_{y=0}^{1} a_{xy} P(Y=y) = \sum_{z \in N_+} \sum_{y=0}^{1} a_{xy} P(Y=y, Z=z)$$
$$= \sum_{z \in N_+} P(Z=z) = 1. \tag{47}$$

Hence:

$$a_{x1} = \frac{1 - a_{x0} P(Y=0)}{P(Y=1)}. \tag{48}$$

From (46) it follows that for all x we have:

$$\begin{cases} P(Z=z) = P(Y=0, Z=z) a_{x0} + P(Y=1, Z=z) a_{x1}, \\ P(Z=z) = P(Y=0, Z=z) + P(Y=1, Z=z). \end{cases} \tag{49}$$

Subtracting second equation from first and using (48) yields:

$$0 = P(Y=0, Z=z)(a_{x0} - 1) + P(Y=1, Z=z)\left(\frac{1 - a_{x0} P(Y=0)}{P(Y=1)} - 1\right) \tag{50}$$
$$= P(Y=0, Z=z)(a_{x0} - 1) + P(Y=1, Z=z)(1 - a_{x0}) \frac{P(Y=0)}{P(Y=1)}. \tag{51}$$

We have two cases:

1) If $a_{x0} \neq 1$ for some x (note that $a_{x0} = 1$ is equivalent to $a_{x1} = 1$ in view of (48)), then the above equation reduces to:

$$P(Y=0, Z=z) = P(Y=1, Z=z) \frac{P(Y=0)}{P(Y=1)}. \tag{52}$$

This means that:

$$P(Z=z) = P(Y=0, Z=z) + P(Y=1, Z=z)$$
$$= P(Y=1, Z=z) \left(1 + \frac{P(Y=0)}{P(Y=1)}\right) = \frac{P(Y=1, Z=z)}{P(Y=1)}.$$

Analogously, we obtain:

$$P(Z=z) = \frac{P(Y=0, Z=z)}{P(Y=0)}. \tag{53}$$

Thus Y and Z are independent i.e. $P(Y=y,Z=z)=P(Y=y)P(Z=z)$. Inserting this equation into (45) yields:

$$a_{xy}=\frac{P(X=x,Y=y,Z=z)}{P(X=x,Z=z)P(Y=y)} \tag{54}$$

and

$$a_{xy}P(X=x,Z=z)=\frac{P(X=x,Y=y,Z=z)}{P(Y=y)}. \tag{55}$$

Thus

$$a_{xy}P(X=x)=\sum_z a_{xy}P(X=x,Z=z)=\sum_z \frac{P(X=x,Y=y,Z=z)}{P(Y=y)}=\frac{P(X=x,Y=y)}{P(Y=y)}. \tag{56}$$

Consequently,

$$a_{xy}=\frac{P(X=x,Y=y)}{P(X=x)P(Y=y)}. \tag{57}$$

Whence, inserting the last equality into (54), we obtain:

$$\frac{P(X=x,Y=y,Z=z)}{P(X=x,Z=z)P(Y=y)}=\frac{P(X=x,Y=y)}{P(X=x)P(Y=y)}, \tag{58}$$

what is equivalent to:

$$P(Y=y,Z=z|X=x)=P(Y=y|X=x)P(Z=z|X=x). \tag{59}$$

Hence Y and Z are conditionally independent given X.

2) If $a_{x0}=1$ for all x, then in view of (48) we obtain $a_{x1}=1$ for all x. This implies conditional independence of (X,Y) given Z.

In order to prove (ii) note that it is obvious when (i) 1. holds and in the case of (ii) 2. it is easy to see that the expression in (45) equals $P(X=x,Y=y)/P(X=x)P(Y=y)$ and thus it does not depend on Z. ■

Appendix D. Proof of Theorem 4.3

Proof (i) First we observe that

$$\sigma^2_{SECMI}=\sum_{x,y,z}p(x,y,z)\log^2\left(\left(\frac{p(x,y)}{p(x)p(y)}\right)^{1-|S|}\prod_{s\in S}\frac{p(x,y,z_s)p(z_s)}{p(x,z_s)p(y,z_s)}\right)$$
$$-\left(\sum_{x,y,z}p(x,y,z)\log\left(\left(\frac{p(x,y)}{p(x)p(y)}\right)^{1-|S|}\prod_{s\in S}\frac{p(x,y,z_s)p(z_s)}{p(x,z_s)p(y,z_s)}\right)\right)^2=0$$

if and only if

$$\left(\frac{p(x,y)}{p(x)p(y)}\right)^{1-|S|}\prod_{s\in S}\frac{p(x,y,z_s)p(z_s)}{p(x,z_s)p(y,z_s)}=C \tag{60}$$

for some $C \geq 0$ and all x, y, z. Let $s_0 \in S$. We rewrite the above equation as:

$$C\left(\left(\frac{p(x,y)}{p(x)p(y)}\right)^{1-|S|}\prod_{s\in S\setminus\{s_0\}}\frac{p(x,y,z_s)p(z_s)}{p(x,z_s)p(y,z_s)}\right)^{-1} = \frac{p(x,y,z_{s_0})p(z_{s_0})}{p(x,z_{s_0})p(y,z_{s_0})}. \tag{61}$$

The left side of above equation does not depend on z_{s_0}. Define

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

Then in view of Lemma 1 we have one of the following possibilities:
1) $s_0 \notin W$. In this case (X, Y) are conditionally independent given Z_{s_0} and for all x, y, z_{s_0}:

$$\frac{p(x,y,z_{s_0})p(z_{s_0})}{p(x,z_{s_0})p(y,z_{s_0})} = 1. \tag{62}$$

2) $s_0 \in W$. In this case Lemma 45 implies that (Y, Z_{s_0}) are independent, (Y, Z_{s_0}) are conditionally independent given X,

$$\frac{p(x,y,z_{s_0})p(z_{s_0})}{p(x,z_{s_0})p(y,z_{s_0})} = \frac{p(x,y)}{p(x)p(y)} \tag{63}$$

and X and Y are dependent.

In view of (60) we obtain:

$$\left(\frac{p(x,y)}{p(x)p(y)}\right)^{1-|S|+|W|} = C. \tag{64}$$

We have three cases: 1) $|S| = 1$. In this case from (60) we have:

$$\frac{p(x,y,z_s)p(z_s)}{p(x,z_s)p(y,z_s)} = C. \tag{65}$$

Thus:

$$\frac{p(x,y,z_s)}{p(y,z_s)} = C\frac{p(x,z_s)}{p(z_s)}. \tag{66}$$

Summing over x, we obtain $C = 1$ and thus in view of (65) (X, Y) are conditionally independent given Z_s.

2) $|W| \neq |S| - 1$ and $|S| > 1$. In this case (64) implies

$$p(x,y) = p(x)p(y)C^{\frac{1}{1-|S|+|W|}}. \tag{67}$$

Summing over x and y yields that $C = 1$ and thus X and Y are independent (see (64)). However, independence of X and Y in view of (63) and definition of W implies that $W = \emptyset$. This means in view of (67) that for all $s \in S$ (X, Y) are conditionally independent given Z_s and (X, Y) are independent.
3) $|W| = |S| - 1$ and $|S| > 1$. In this case $W^c = \{s_0\}$ for some $s_0 \in S$. This means that (X, Y) are conditionally independent given X_{s_0} and Lemma 45 implies that for all

$s \in S \setminus \{s_0\}$ (Y, Z_s) are independent, (Y, Z_s) are conditionally independent given X and (X, Y) are dependent, because $W \neq \emptyset$.
Part (ii). Part (i) is obvious in the case when conditions 1) holds, in the case of 2) it follows form the proof of part (i) of the Lemma above that for any $s \neq s_0$ $p(x, y, z_s)p(z_s)/p(x, z_s)p(y, z_s)$ equals $p(x, y)/p(x)p(y)$ and for $s = s_0$ it equals 1. Thus

$$\left(\frac{p(x,y)}{p(x)p(y)}\right)^{1-|S|} \prod_{s\in S} \frac{p(x,y,z_s)p(z_s)}{p(x,z_s)p(y,z_s)} = 1 \tag{68}$$

and it follows that $\sigma^2_{\widehat{SECMI}} = 0$.
We prove now (iii) which states that if Y is binary then $SECMI \neq 0$ implies $\sigma^2_{\widehat{SECMI}} > 0$. In order to prove it note that for each case considered in proof of part (i) we obtain $C = 1$. This means that for all x, y, z we have that (68) holds (see 61). From the definition of $SECMI$ it follows that:

$$\begin{aligned} SECMI(X,Y|Z) &= \sum_{x,y,z} p(x,y,z) \log\left(\left(\frac{p(x,y)}{p(x)p(y)}\right)^{1-|S|} \prod_{s\in S} \frac{p(x,y,z_s)p(z_s)}{p(x,z_s)p(y,z_s)}\right) \\ &= \sum_{x,y,z} p(x,y,z) \log 1 = 0. \end{aligned}$$

■

Appendix E. Proof of Theorem 4

Proof It follows from (4) and (8) that

$$II(X,Y,Z_s,Z_{s'}) = I(X,Y|Z_s,Z_{s'}) - I(X,Y|Z_s) - I(X,Y|Z_{s'}) + I(X,Y). \tag{69}$$

Using this and definition of $SECMI3$ we have that

$$SECMI3 = A \times I(X,Y) + B \sum_{s\in S} I(X,Y|Z_s) + \sum_{s<s',s,s'\in S} I(X,Y|Z_s,Z_{s'}),$$

where A and B are defined in (27). Thus $SECMI3 = \tilde{f}(\mathbf{p})$, where $\tilde{f}(\mathbf{p})$ is given by $(z = (z_1, \ldots, z_{|S|}))$

$$\sum_{x,y,z} p(x,y,z) \log\Big(\Big[\frac{p(x,y)}{p(x)p(y)}\Big]^A \Big[\prod_{s\in S} \frac{p(x,y,z_s)}{p(x,z_s)p(y,z_s)}\Big]^B \prod_{s<s',s,s'\in S} \frac{p(x,y,z_s,z_{s'})p(z_s,z_{'_s})}{p(x,z_s,z_{s'}p(y,z_s,z_{s'})}\Big). \tag{70}$$

Some tedious but straightforward manipulations yield that

$$\begin{aligned} \frac{\partial \tilde{f}}{p(x,y,z)} \quad &= \quad A(\Big(\log\Big(\frac{p(x,y)}{p(x)p(y)}\Big) - 1) + B\sum_{s\in S} \log\Big(\frac{p(x,y,z_s)}{p(x,z_s)p(y,z_s)}\Big) + \\ &\sum_{s<s',s,s'\in S} \log\Big(\frac{p(x,y,z_s,z_{s'})p(z_s,z_{'_s})}{p(x,z_s,z_{s'}p(y,z_s,z_{s'})}\Big)\Big) \end{aligned}$$

■

It is now easy to see that

$$\sum_{x,y,z}\Big(\frac{\partial \tilde{f}}{\partial p(x,y,z)}\Big)^2 p(x,y,z) - \Big(\sum_{x,y,z}\frac{\partial f}{\partial p(x,y,z)}p(x,y,z)\Big)^2 = \sigma^2_{\widehat{SECMI}3}.$$

The proof now follows as in the case of proof of Theorem 2 (i). The proof of part (ii) is analogous.

References

A. Agresti. *Categorical Data Analysis.* Wiley, 2002.

C. Aliferis, I. Tsamardinos, and A. Statnikov. Hiton, a novel Markov blanket algorithm for optimal variable selection. In *Proceedings of the American Medical Informatics Association*, 2003.

D. Barber. *Bayesian Approach and Machine Mearninig.* Cambridge University Press, 5th edition, 2014.

R. Barber and E. Candès. Controlling the false discovery rate via knockoffs. *Ann. Statist.*, 43(5):2055–2085, 10 2015.

G. Borbudakis and I. Tsamardinos. Forward-backward selection with early dropping. *Journal of Machine Learning Resesarch*, 20:1–39, 2019.

F. Bromberg, D. Margaritis, and V. Honavar. Efficient Markov network structure discovery using independence tests. *J. Artif. Int. Res.*, 35(1):449–484, 2009.

G. Brown, A. Pocock, M. J. Zhao, and M. Luján. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *Journal of Machine Learning Research*, 13(1):27–66, 2012.

M. Buckley and G. Eagleson. An approximation to the distribution of quadratic forms in normal random variables. *Australian Journal of Statistics*, 1:150–159, 1988.

P. Bühlmann and S. van de Geer. *Statistics for High-dimensional Data.* Springer, 1st edition, 2015.

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, 2018.

T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing).* Wiley-Interscience, 2006.

J. Demšar. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30, 2006.

D. Dheeru and K. Taniskidou. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

R. Fisher. On the interpretation of chi square from contingency tables and interpretation of p. *Journal of Royal Statisical Society*, 85:87–94, 1922.

M. Friedman. A comparison of alternative tests of significance for the problem of m rankings. *Annals of Mathematical Statistics*, 11:86–92, 1940.

T. Gao and J. Qiang. Efficient Markov blanket discovery and its application. *IEEE Transactions on Cybernetics*, 47(5):1169–1179, 2017.

I. Guyon and A. Elyseeff. An introduction to feature selection. *Feature Extraction, Foundations and Applications*, 207(Studies in Fuzziness and Soft Computing):1–25, 2006.

T. S. Han. Multiple mutual informations and multiple interactions in frequency data. *Information and Control*, 46(1):26 – 45, 1980.

T. Inglot. Inequalities of the quantiles of the chisquare distribution. *Probability and Mathematical Statistics*, 30(2):339–351, 2010.

D. Koller and M. Sahami. Toward optimal feature selection. In *ICML-1995*, pages 284–292, 1995.

M. Kubkowski and J. Mielniczuk. Asymptotic distributions of interaction information. *Methodology and Computing in Applied Probability*, 2020. URL `https://doi.org/10.1007/s11009-020-09783-0`.

S. Kullback. *Information Theory and Statistics*. Peter Smith, 1978.

J. Leppä-Aho, S. Räisänen, X. Yang, and T. Roos. Learning non-parametric Markov networks with mutual information. In *Proceedings of the Ninth International Conference on Probabilistic Graphical Models*, volume 72, pages 213–224, 2018.

D. Lin and X. Tang. Conditional infomax learning: An integrated framework for feature extraction and fusion. In *Proceedings of the 9th European Conference on Computer Vision - Volume Part I*, ECCV'06, pages 68–82, 2006.

D. Margaritis and S. Thrun. Bayesian network induction via local neighborhoods. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, pages 505–511, 1999.

W. J. McGill. Multivariate information transmission. *Psychometrika*, 19(2):97–116, 1954.

J. Mielniczuk and P. Teisseyre. A deeper look at two concepts of measuring gene-gene interactions: logistic regression and interaction information revisited. *Genetic Epidemiology*, 42:187–200, 2018.

J. Mielniczuk and P. Teisseyre. Stopping rules for mutual information-based feature selection. *Neurocomputing*, 358:255–271, 2019.

P. B. Nemenyi. *Distribution-free multiple comparisons.* PhD thesis, Princeton University, 1963.

M. Pawluk, P. Teisseyre, and J. Mielniczuk. Information-theoretic feature selection using high-order interactions. In *Machine Learning, Optimization, and Data Science*, pages 51–63. Springer International Publishing, 2019.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks for Plausible Inference.* Morgan Kaufmann, 1988.

J-P. Pellet and A. Eliseeff. Using Markov Blankets for causal structure learning. *Journal of Machine Learning Research*, 8:1295– 2042, 2008.

J. M. Pena, R. Nilsson, J. Bjoerkegren, and J Tegner. Towards scalable and data efficient learning of Markov boundaries. *International Journal of Approximate Reasoning*, 45(2): 211 – 232, 2007.

O.C. Rota. On the foundations of combinatorial theory i: Theory of Möbius functions. *Zeitschrift für Warscheinlichkeitstheorie und verwandene Gebiete*, 2:340–368, 1964.

J. Runge. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84, pages 938–947, 2018.

F. Schlüter. A survey on independence-based Markov networks learning. *Artificial Intelligence Review*, 42:1069–1093, 2012.

M. Scutari. Learning bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3):1–22, 2010.

R. Shah and J. Peters. The hardness of conditional independence testing and the generalised covariance measure. *Ann. Statist.*, 48(3):1514–1538, 2020.

I. Shao. *Mathematical Statistics.* Springer, 2003.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search.* MIT press, 2nd edition, 2000.

L. Su and H. White. A consistent characteristic function-based test for conditional independence. *Journal of Econometrics*, 141:807–834, 2007.

L. Su and H. White. Testing conditional independence via empirical likelihood. *Journal of Econometrics*, 182:27–44, 2014.

H. K. Ting. On the amount of information. *Theory Probab. Appl.*, 7(4):439–447, 1960.

I. Tsamardinos and G. Borboudakis. Permutation testing improves on Bayesian network learning. In *Proceedings of ECML PKDD 2010*, pages 322–337, 2010.

I. Tsamardinos, C. Aliferis, and A. Statnikov. Algorithms for large scale Markov Blanket discovery. In *FLAIRS Conference*, pages 376–381, 2003.

N. Vinh, S. Zhou, J. Chan, and J. Bailey. Can high-order dependencies improve mutual information based feature selection? *Pattern Recognition*, 53:45–58, 2016.

Q. Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2):307–333, 1989.

X. Wang and Y. Hong. Characteristic function based testing for conditional independence: A nonparametric regression approach. *Econometric Theory*, 34:815–849, 2018.

R. W. Yeung. *A First Course in Information Theory.* Kluwer, 2002.

K. Yu, L. Liu, and J. Li. A unfied view of causal and non-causal feature selection. manuscript, 2018. URL `ArXiv:1802.05844`.

J-T. Zhang. Approximate and asymptotic distributions of chi-squared type mixtures with applications. *Journal of the American Statistical Association*, 100(469):273–285, 2005.