WARSAW UNIVERSITY OF TECHNOLOGY

DISCIPLINE OF SCIENCE - MATHEMATICS
FIELD OF SCIENCE - NATURAL SCIENCES

Ph.D. Thesis

Małgorzata Łazęcka, M.Sc.

Properties of information-theoretic measures of conditional dependence

Supervisor
prof. Jan Mielniczuk

WARSAW, 2022

Podziękowania

Dziękuję profesorowi Mielniczukowi za pomoc w przygotowaniu tej pracy, cenne uwagi oraz za życzliwość.

Dziękuję Przemkowi za codzienne wpieranie mnie oraz rodzinie i znajomym za to, że zawsze mogłam na nich liczyć.

Abstract

In the thesis we investigate properties of information-theoretic measures for discrete distributions and show how most of them can be related to Möbius expansion of conditional mutual information (CMI). Moreover, we study asymptotic distributions of such measures (Chapter 1). In Chapter 2 we consider four resampling scenarios, which can be used for conditional independence (CI) testing: *CI bootstrap*, *conditional randomisation*, *bootstrap X* and *conditional permutation* schemes. We study asymptotic distributions of introduced measures evaluated for resampled data, as well as properties of the schemes themselves, which are useful in CI testing. Chapter 3 covers numerical experiments conducted in order to investigate performance of $\widehat{CMI}$ and $\widehat{CMI}$-related measures as test statistics in CI testing. Moreover, the related problem of testing global null hypothesis corresponding to individual hypothesis of conditional independence is investigated.

Keywords: mutual information, conditional mutual information, interaction information, information-theoretic criteria, resampling, bootstrap, asymptotic distribution, conditional independence

Streszczenie

W pracy zbadano własności miar teorioinformacyjnych dla zmiennych o rozkładach dyskretnych i pokazano, w jaki sposób większość z nich jest związana z rozwinięciem Möbiusa dla warunkowej informacji wzajemnej (CMI). W rozdziale 1 zbadano rozkłady asymptotyczne takich miar. W rozdziale 2 wprowadzono cztery schematy ponownego próbkowania, mające zastosowanie w testowaniu warunkowej niezależności: *CI bootstrap*, *conditional randomisation* (warunkowa randomizacja), *bootstrap X* i *conditional permutation* (warunkowe permutacje). Zbadano również zachowanie estymatorów miar obliczonych na podstawie prób resamplingowych oraz własności samych schematów przydatnych w testowaniu warunkowej niezależności. W rozdziale 3 przeprowadzono eksperymenty numeryczne pokazujące zastosowanie $\widehat{CMI}$ i kryteriów opartych na $\widehat{CMI}$ jako statystyk testowych w testowaniu warunkowej niezależności. Przerowadzono również symulacje dotyczące testowanie hipotezy globalnej z indywidualnymi hipotezami zerowymi będącymi hipotezami o warunkowej niezależności zmiennych.

Słowa kluczowe: informacja wzajemna, warunkowa informacja wzajemna, informacja interakcyjna, kryteria teorioinformacyjne, ponowne próbkowanie, bootstrap, rozkład asymptotyczny, warunkowa niezależność

Contents

Introduction 11

Chapter 1. Information-theoretic measures 13

1.1. Basic measures 13

1.1.1. Conditional independence and conditional mutual information 15

1.1.2. Feature selection based on mutual information 18

1.2. Interaction information 19

1.2.1. 3-way interaction information 19

1.2.2. Generalisation: k-way interaction information 20

1.3. Feature selection criteria 28

1.3.1. The generalized feature selection criterion based on Möbius expansion . . 29

1.3.2. Second and third order criteria 32

1.4. Asymptotic distributions of information-theoretic empirical measures 34

1.4.1. Asymptotic behaviour of the empirical conditional mutual information . . 35

1.4.2. Asymptotic behaviour of plug-in estimators of feature selection criteria . . 40

Chapter 2. Resampling schemes and asymptotic distributions of information-theoretic measures 49

2.1. Resampling schemes 49

2.1.1. CI bootstrap scenario 49

2.1.2. Conditional randomisation scenario (CR) 53

2.1.3. Bootstrap X scenario 55

2.1.4. Conditional permutation scenario 56

2.1.5. Comparison of covariance matrices 63

2.2. Asymptotic behaviour of $\widehat{CMI}^*$ 65

2.2.1. Distribution of $\widehat{CMI}^*$ 66

2.2.2. Validity of asymptotic convergence . . . 72
2.2.3. Non-asymptotic approach . . . 74
2.3. Asymptotic behaviour of $\widehat{JMI}^*$ . . . 78

Chapter 3. Simulations . . . 89

3.1. Conditional independence testing . . . 89
3.1.1. Models M1 and M2: description . . . 89
3.1.2. Asymptotic convergence: numerical analysis . . . 92
3.1.3. Comparison of testing procedures . . . 96
3.1.4. Significance level . . . 99
3.1.5. Power . . . 106
3.1.6. Analysis of high-order interaction model . . . 110
3.2. Global null for individual hypotheses of conditional independence . . . 113
3.2.1. Test based on $\widehat{JMI}$ . . . 113
3.2.2. Generic methods . . . 114
3.2.3. Simulations . . . 115
3.3. Summary of experiments . . . 118

Appendix A. Theorems . . . 121

A.1. Lemma used in Chapter 1 . . . 121
A.2. Theorems used in Chapter 2 . . . 123

Appendix B. List of Symbols . . . 125

B.1. Information-theoretic measures . . . 126

Bibliography . . . 127

Introduction

Conditional independence (CI) is one of the central concepts in statistics, which plays a fundamental role in such areas as e.g. casual inference, dependence analysis and regression modelling. Checking CI is a building block for feature selection. In particular, CI testing is a crucial part of algorithms used for discovering Markov Blanket (MB) such as e.g. GS (*Grow and Shrink* [27]) or IAMB (*Incremental Association Markov Blanket* [40]). Such algorithms typically apply a series of conditional independence tests in order to learn a structure from observational data and recover MB if the CI condition can be verified without errors.

Testing for conditional independence is much more challenging than for unconditional independence. In discrete case, conditional independence of two variables given the third one holds if for every layer of conditioning variable (i.e. within each subset for which the value of conditioning variable is fixed) the two variables are independent. Thus curse of dimensionality is a significant obstacle when the dimensionality of the conditioning variable is high. Conditional independence testing in a continuous case is even more difficult [37].

Conditional mutual information (CMI) is a measure of conditional dependence studied in information theory, which possesses many attractive properties. This explains its frequent use. Also it is a source of many CMI-based measures of dependence ([6]). The CMI-based measures are usually called *criteria*, as they are often used as scoring criteria to measure how potentially useful a feature may be when used to explain a dependent variable. The criteria are designed to cope better with the problem of multidimensionality, but on the other hand they miss many useful properties of CMI.

In the thesis we investigate properties of such CMI-based measures for discrete dis-

tributions and show how most of them can be related to Möbius expansion of CMI. Moreover, we study asymptotic distributions of such measures (Chapter 1). In Chapter 2 we consider four resampling scenarios, which can be used for CI testing: CI bootstrap, conditional randomisation, bootstrap X and conditional permutation schemes. We study asymptotic distributions of empirical conditional mutual information and introduced measures evaluated for resampled data, as well as properties of the schemes themselves, which are useful in CI testing. Chapter 3 covers numerical experiments conducted in order to investigate performance of $\widehat{CMI}$ and $\widehat{CMI}$-related measures as test statistics in CI testing. Moreover, the related problem of testing global null hypothesis corresponding to individual hypothesis of independence is investigated.

Similar concepts as in Chapter 2 have been studied earlier. Conditional randomisation test and its permutational counterpart were introduced in [8] and [5], respectively. They both rely on the availability of the distribution of the potential explanatory variable given a vector of other variables, which may contain confounding factors. Other schemes of resampling applying conditional permutations used for CI testing are proposed in [41], in which also a semi-parametric approach included in simulations in Chapter 3 is proposed. Conditional permutation scheme for $\widehat{CMI}$-related measures was also used in [20] and [18], in which the asymptotic distributions of the criteria estimators were obtained (these results have been extended in Chapter 1), although theoretical analysis of the behaviour of $\widehat{CMI}$ and criteria based on resampled samples was limited.

Chapter 1

Information-theoretic measures

In this chapter we first recall the definitions of measures such as entropy and mutual information used in information theory, and their basic properties. Next we give the definition of 3 and k-way interaction information and state its representations in terms of entropy and mutual information and their inverse formulas. In particular, we establish Möbius formula, which gives a representation of conditional mutual information in terms of interaction informations. Then we focus on feature selection criteria based on the previously presented measures and Möbius formula.

Let X, Y and Z be discrete random variables with values in $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$, respectively. We assume that $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ are finite. We denote the probability mass function by p and instead of using e.g. p_X for the probability mass function of the random variable X, we will write simply $p(x)$. By $p(x|z)$ we will denote probability of obtaining $X = x$ given $Z = z$, and $p(x, y, z)$ is a joint probability mass function of (X, Y, Z) etc.

X, Y and Z might be multivariate and in some cases we will use the notation $Z = Z_S = (Z_1, Z_2, \ldots, Z_{|S|})$ to underline that Z is multivariate. S is a set of indices and for convenience we assume that $S = \{1, 2, \ldots, k\}$ and $|S| = k$. By Z_T we denote the subvector of variables $(Z_{t_1}, Z_{t_2}, \ldots, Z_{t_i})$ with elements such that their indices are from a set $T = \{t_1, t_2, \ldots, t_i\} \subseteq S$.

Logarithm in definitions below denotes the natural logarithm.

1.1. Basic measures

We introduce basic information theoretic measures for discrete random variables. We refer to [9] for more information and properties.

Definition 1.1.1 (Entropy)**.** The entropy of X is defined as

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

Similarly, the joint entropy of X and Y equals

$$H(X,Y) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log p(x,y).$$

In the definition above the convention $0\log 0 = 0$ is used. Note that from the definition we also have that $H(X) \geq 0$ and the equality holds if and only if X is constant.

Below we give the definition of conditional entropy of X given Y, which averages entropies of X over all layers $Y = y$ with respect to distribution $p(y)$.

Definition 1.1.2 (Conditional entropy)**.** The conditional entropy is defined in the following way

$$\begin{aligned} H(X|Y) &= \sum_{y\in\mathcal{Y}} p(y)H(X|Y=y) = -\sum_{y\in\mathcal{Y}} p(y)\sum_{x\in\mathcal{X}} p(x|y)\log p(x|y) \\ &= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log p(x|y) = -\mathbb{E}\log p(X|Y), \end{aligned}$$

where expected value is computed with respect to joint distribution $p(x,y)$ of (X,Y).

Next, we give a definition of Kullback-Leibler divergence, which is a measure of how much two probability distributions p and q differ. That definition will be useful for interpreting measures of dependence called mutual information and conditional mutual information, and also it will give a different view on interaction information.

Definition 1.1.3 (Kullback-Leibler divergence)**.** We define Kullback-Leibler divergence between two probability mass functions p and q of a random variable X as

$$D_{KL}(p||q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}.$$

Kullback-Leibler divergence is defined if for all $x \in \mathcal{X}$ we have $q(x) = 0 \Rightarrow p(x) = 0$ and then we use a convention $0\log\frac{0}{0} = 0$. If $p(x) = 0$, then $0\log 0 = 0$ as previously.

Kullback-Leibler divergence is non-negative and equals zero if and only if $p = q$ (see Theorem 2.6.3 in [9]).

Definition 1.1.4 (Mutual information)**.** We define mutual information (MI) as

$$I(X,Y) = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} = D_{KL}(p(x,y)||p(x)p(y)) = \mathbb{E}\log\frac{p(X,Y)}{p(X)p(Y)}.$$

Intuitively, the mutual information measures how much uncertainty of X is reduced given the variable Y (or alternatively, how much uncertainty of Y is reduced due to knowing X) as

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y), \quad (1.1)$$

where entropy is understood as a measure of uncertainty. Another way to interpret mutual information is to understand it as a measure of distance in terms of Kullback-Leibler divergence between the joint distribution $p(x,y)$ and its factorised counterpart $p(x)p(y)$. Equality of these two distributions is equivalent to the divergence between them being equal to 0 and their independence.

1.1.1. Conditional independence and conditional mutual information

The definition of conditional independence and conditional mutual information and its properties presented below play an important role in the following sections, especially in Chapter 2, where we focus on distribution of plug-in estimators of conditional mutual information based on resampled samples.

Two variables X and Y are conditionally independent given Z (we assume $p(z) > 0$) if for all $(x,y,z) \in \mathcal{X}\times\mathcal{Y}\times\mathcal{Z}$ holds

$$p(x,y,z) = p(x|z)p(y|z)p(z).$$

Below we give the definition of conditional mutual information, which averages mutual informations over all layers of conditioning variable.

Definition 1.1.5 (Conditional mutual information)**.** Conditional mutual information (CMI) is defined as

$$I(X,Y|Z) = \sum_{z\in\mathcal{Z}} p(z) \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y|z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)}$$
$$= \sum_{x\in\mathcal{X},y\in\mathcal{Y},z\in\mathcal{Z}} p(x,y,z) \log \frac{p(x,y,z)}{p(x|z)p(y|z)p(z)} = D_{KL}(p(x,y,z)||p(x|z)p(y|z)p(z)).$$

We note that both mutual information and conditional mutual information are non-negative [9] and might be considered as measures of strength of unconditional and conditional dependence, respectively. In case of conditional mutual information the strength of dependence is measured given the conditioning variable Z. Both measures equal zero if and only if $X \perp\!\!\!\perp Y$ (MI) or $X \perp\!\!\!\perp Y|Z$ (CMI). What is more, we have that

$$I(X,Y|Z) = \sum_{z\in\mathcal{Z}} p(z) I(X,Y|Z=z),$$

where $I(X,Y|Z=z)$ denotes mutual information for a pair of variables (X,Y) given $Z=z$ with a probability mass function $p(x,y|z)$, thus as $p(z)>0$ and $I(X,Y|Z=z)>0$ for all z, we have that

$$I(X,Y|Z)=0 \Rightarrow I(X,Y|Z=z)=0 \text{ for all } z\in\mathcal{Z}.$$

Hence, if X and Y are independent given Z, the variables are also independent on each layer of Z.

We denote by p_{ci} probability mass function corresponding to $X \perp\!\!\!\perp Y|Z$ (ci stands for *conditional independence*) equal $p_{ci}(x,y,z) = p(x|z)p(y|z)p(z)$, where

$$p(x|z) = \frac{\sum_y p(x,y,z)}{p(z)}$$
$$p(y|z) = \frac{\sum_x p(x,y,z)}{p(z)}$$

and $p(z) = \sum_{x,y} p(x,y,z)$. We will show that p_{ci} minimises Kullback-Leibler divergence of p and any probability mass function satisfying conditional independence. Namely, the following lemma holds

Lemma 1.1.1. *Probability mass function p_{ci} defined above minimises $D_{KL}(p(x,y,z)||q(x,y,z))$ over $q \in \mathcal{C}$ such that*

$$\mathcal{C} = \{q(x,y,z) : q(x,y,z) = q(x|z)q(y|z)q(z)\}.$$

Proof. Indeed,

$$\begin{aligned} &D_{KL}(p(x,y,z)||q(x,y,z)) - D_{KL}(p(x,y,z)||p(x|z)p(y|z)p(z)) \qquad (1.2)\\ &\quad = \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y,z)}{q(x,y,z)} - \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y,z)}{p(x|z)p(y|z)p(z)} \\ &\quad = \sum_{x,y,z} p(x,y,z) \log \frac{p(x|z)p(y|z)p(z)}{q(x|z)q(y|z)q(z)}. \end{aligned}$$

Next, by breaking the above expression into three sums, we obtain

$$\sum_z p(z) \sum_x p(x|z) \log \frac{p(x|z)}{q(x|z)} + \sum_z p(z) \sum_y p(y|z) \log \frac{p(y|z)}{q(y|z)} + \sum_z p(z) \log \frac{p(z)}{q(z)}.$$

The expression $\sum_x p(x|z) \log \frac{p(x|z)}{q(x|z)}$ is equal to Kullback-Leibler divergence of $p(x|z)$ and $q(x|z)$ for a fixed value of Z (similarly for $\sum_y p(y|z) \log \frac{p(y|z)}{q(y|z)}$ the expression under the sum denotes $D_{KL}(p(y|z)||q(y|z))$ and $\sum_z p(z) \log \frac{p(z)}{q(z)} = D_{KL}(p(z)||q(z))$). Thus (1.2) is non-negative and equal to 0 if and only if $q(x|z) = p(x|z)$, $q(y|z) = p(y|z)$ and $q(z) = p(z)$. □

Lemma 1.1.2 provides a way to decompose the entropy of a vector of variables $(Z_1, Z_2, \ldots, Z_{|S|})$ and the mutual information of a variable X and a vector of variables $(Z_1, Z_2, \ldots, Z_{|S|})$.

Lemma 1.1.2 (Chain rules [9])**.** *(i) Chain rule for entropy has the following form*

$$H(Z_1, Z_2, \ldots, Z_{|S|}) = H(Z_1) + \sum_{i=2}^{|S|} H(Z_i|Z_{i-1}, \ldots, Z_1).$$

(ii) Chain rule for mutual information is given by

$$I(X, (Z_1, Z_2, \ldots, Z_{|S|})) = I(X, Z_1) + \sum_{i=2}^{|S|} I(X, Z_i|Z_{i-1}, \ldots, Z_1).$$

In particular, for mutual information of Y and $(X, Z_1, Z_2, \ldots, Z_{|S|})$ we have that

$$I(Y, (X, Z_1, Z_2, \ldots, Z_{|S|})) = I(Y, X) + I(Y, (Z_1, Z_2, \ldots, Z_{|S|})|X) \tag{1.3}$$

and

$$I(Y, (X, Z_1, Z_2, \ldots, Z_{|S|})) = I(Y, (Z_1, Z_2, \ldots, Z_{|S|})) + I(Y, X|(Z_1, Z_2, \ldots, Z_{|S|})). \tag{1.4}$$

1.1.2. Feature selection based on mutual information

Mutual information and conditional mutual information are used in feature selection. First, we consider a problem in which we want to identify a subset of the most significant variables of the size k, where $1 \leq k \leq |S|$. Thus, we want to maximize the mutual information between Y and a set of chosen variables, the size of which equals k, namely

$$\arg\max_{\substack{T \subseteq S \\ |T|=k}} I(Y, Z_T),$$

where $Z_T = (Z_{t_1}, Z_{t_2}, \ldots, Z_{t_k})$. To search through all possible subsets T of S such that $|T| = k$, we need to compute $\binom{|S|}{k}$ mutual informations, thus this task usually is not feasible for large number of features. In practise, instead of searching for the globally optimal solution, some greedy algorithms are applied e.g. stepwise forward selection, in which in the first step we choose the feature that maximizes mutual information and then in kth step we choose such $i \in S_{k-1}^c$, that

$$\arg\max_{i \in S_{k-1}^c} I(Y, Z_{S_{k-1} \cup \{i\}}), \tag{1.5}$$

where S_k is a set of variables chosen in k steps and $S_k^c = S \setminus S_k$. Thus in each step the variable that gives the most information about Y given already chosen variables is added to the set of selected variables. Note that from (1.4) we have

$$I(Y, Z_{S_{k-1} \cup \{i\}}) = I(Y, Z_{S_{k-1}}) + I(Y, Z_i|Z_{S_{k-1}}),$$

hence instead of (1.5) we can optimize $\arg\max_{i \in S_{k-1}^c} I(Y, Z_i|Z_{S_{k-1}})$, as $I(Y, Z_{S_{k-1}})$ depends only on the already chosen features. In the following we will usually denote the candidate

variable as X and the set of selected features as Z_S, thus in the next sections we will be interested in $I(X, Y|Z_S)$ and its approximations in order to examine, whether X is significant in explaining Y in the presence of Z_S.

1.2. Interaction information

Below we state definitions of 3-way and then k-way interaction information with notes on differences in notation, signs and other details appearing in literature. We also provide basic properties of the interaction information.

1.2.1. 3-way interaction information

The measure called interaction information was introduced by McGill [28] and the definition was given for three random variables to quantify the gain or the loss in sample information transmitted between any two of the variables, due to additional knowledge of the third variable. The definition stated in [28] is based on entries in contingency table instead of probabilities, but the main idea remains unchanged.

Definition 1.2.1 (Interaction information [28])**.** The interaction information of three discrete random variables X, Y and Z (3-way interaction information) is defined as

$$II(X, Y, Z) = I(X, Y|Z) - I(X, Y). \tag{1.6}$$

Although it is not clear from (1.6) itself, interaction information in Definition 1.2.1 is symmetric with respect to X, Y and Z as we have

$$\begin{aligned} II(X, Y, Z) = I(X, Y|Z) - I(X, Y) = &-H(X) - H(Y) - H(Z) \\ &+ H(X, Y) + H(X, Z) + H(Y, Z) - H(X, Y, Z), \end{aligned}$$

which follows from basic properties of entropy and mutual information (cf. (1.1) and Lemma 1.1.2).

Interaction information is used to measure the strength of interaction between variables i.e. from (1.6) we see that $II(X, Y, Z)$ measures how much information of X about Y we gain due to knowing additional variable Z. In contrast to MI or CMI, interaction information can be negative. We give an example below showing that.

Example 1.2.1. *We will construct an example, in which* $I(X,Y|Z) = 0$ *and* $I(X,Y) > 0$ *as in such a case we obtain negative interaction information as*

$$II(X,Y,Z) = I(X,Y|Z) - I(X,Y) < 0.$$

First, we notice that the condition $I(X,Y|Z) = 0$ *is equivalent to conditional independence of* X *and* Y *given* Z*, thus the joint probability should factorise in the following way*

$$p(x,y,z) = p(x|z)p(y|z)p(z).$$

Assume that all variables are binary. Next, let $Z \sim Bern(1/2)$ *(*$P(Z=0) = P(Z=1) = 1/2$*),* $X \perp\!\!\!\perp Y|Z$ *and*

$$1 - P(X = 1-z|Z=z) = P(X=z|Z=z) = \alpha,$$
$$1 - P(Y = 1-z|Z=z) = P(Y=z|Z=z) = \alpha,$$

where $z \in \{0,1\}$ *and* $\alpha \in [0,1]$*. The joint distribution of* (X,Y) *can be calculated and the corresponding probabilities are given in the table below:*

(X,Y)	$p(\cdot,0)$	$p(\cdot,1)$
$p(0,\cdot)$	$1/2 - \alpha + \alpha^2$	$\alpha - \alpha^2$
$p(1,\cdot)$	$\alpha - \alpha^2$	$1/2 - \alpha + \alpha^2$

.

One can check, that for $\alpha \neq 1/2$ *the variables* X *and* Y *are not independent, hence* $I(X,Y) > 0$ *and thus* $II(X,Y,Z) < 0$ *for* $\alpha \neq 1/2$*.*

1.2.2. Generalisation: k-way interaction information

The Definition 1.2.2 of k-way interaction information comes from series of Fano's lectures [13] and in the book it is called as *a general definition of the mutual information between an arbitrary number of points.* We note that the sign in (1.7) below is changed compared to original definition in the case of an odd number of variables as we want interaction information to satisfy the recursive formula stated in Lemma 1.2.5 for any number of variables k (this property also appears in [13]). We note that in [13] the definition is

given without averaging over the distribution of the variables i.e. the analogous formula to (1.7) misses the averaging term $\sum_{z_1,z_2,\ldots,z_{|S|}} p(z_1, z_2, \ldots, z_{|S|})$ at the beginning.

Definition 1.2.2 (Interaction information [13])**.** The k-way interaction information (the interaction information of $k = |S|$ discrete random variables) of Z_1, Z_2, ..., $Z_{|S|}$ for $|S| \geq 1$ is defined as

$$\begin{aligned} &II(Z_1, Z_2, \ldots, Z_{|S|}) \\ &= (-1)^{|S|} \sum_{z_1,z_2,\ldots,z_{|S|}} p(z_1, z_2, \ldots, z_{|S|}) \log \left(\prod_{i=1}^{|S|} \left(\prod_{\substack{T \subseteq S \\ |T|=i}} p(z_{t_1}, \ldots, z_{t_i}) \right)^{(-1)^i} \right) \\ &= (-1)^{|S|} \sum_{z_1,z_2,\ldots,z_{|S|}} p(z_1, z_2, \ldots, z_{|S|}) \log \frac{\prod\limits_{i_1<i_2} p(z_{i_1}, z_{i_2}) \cdots}{\prod\limits_{i_1} p(z_{i_1}) \prod\limits_{i_1<i_2<i_3} p(z_{i_1} z_{i_2}, z_{i_3}) \cdots}. \end{aligned} \tag{1.7}$$

Note that logarithmic expression involves probabilities of subvectors of Z_S of even dimension in the numerator and of odd dimension in the denominator.

Definition 1.2.2 for $|S| = 2$ yields $II(Z_1, Z_2) = I(Z_1, Z_2)$ and for $|S| = 1$ we have that $II(Z_1) = H(Z_1)$. We define $II(\emptyset) = 0$. For $|S| = 3$ we obtain

$$\begin{aligned} II(Z_1, Z_2, Z_3) &= -\sum_{z_1,z_2,z_3} p(z_1, z_2, z_3) \log \frac{p(z_1, z_2)p(z_1, z_3)p(z_2, z_3)}{p(z_1)p(z_2)p(z_3)p(z_1, z_2, z_3)} \\ &= -H(Z_1) - H(Z_2) - H(Z_3) + H(Z_1, Z_2) + H(Z_1, Z_3) + H(Z_2, Z_2) - H(Z_1, Z_2, Z_3), \end{aligned} \tag{1.8}$$

thus Definition 1.2.2 of k-way interaction information is consistent with Definition 1.2.1 of 3-way interaction information. We also have the following generalisation of (1.8) for k-way interaction information in terms of entropies for $k > 3$.

Lemma 1.2.2. *The interaction information defined in* (1.6) *satisfies the equation*

$$II(Z_1, Z_2, \ldots, Z_{|S|}) = -\sum_{i=1}^{|S|} \sum_{\substack{T \subseteq S \\ |T|=i}} (-1)^{|S|-|T|} H(Z_{t_1}, Z_{t_2}, \ldots, Z_{t_i}).$$

Proof. It follows in straightforward manner from Definition 1.2.2. □

Below we give the interpretation of interaction information using the Kirkwood superposition approximation.

Definition 1.2.3 (Kirkwood superposition approximation)**.** The Kirkwood approximation P^K of a multivariate discrete random variable $(Z_1, Z_2, \ldots, Z_{|S|})$ with a probability mass function $p(z_1, z_2, \ldots, z_{|S|})$ is given by

$$p^K(z_1, z_2, \ldots, z_{|S|}) = \prod_{i=1}^{|S|-1} \left(\prod_{\substack{T \subseteq S \\ |T|=i}} p(z_{t_1}, \ldots, z_{t_i}) \right)^{(-1)^{|S|-1-i}}$$

$$= \left(\frac{\prod_{i=1}^{\lfloor |S|/2 \rfloor} \prod_{\substack{T \subseteq S \\ |T|=2i-1}} p(z_{t_1}, \ldots, z_{t_{2i-1}})}{\prod_{i=1}^{\lfloor (|S|-1)/2 \rfloor} \prod_{\substack{T \subseteq S \\ |T|=2i}} p(z_{t_1}, \ldots, z_{t_{2i}})} \right)^{(-1)^{|S|}} = \left[\frac{\prod_{i_1<i_2} p(z_{i_1}, z_{i_2}) \cdots}{\prod_{i_1} p(z_{i_1}) \prod_{i_1<i_2<i_3} p(z_{i_1} z_{i_2}, z_{i_3}) \cdots} \right]^{(-1)^{|S|-1}}. \tag{1.9}$$

P^K is not necessarily probability distribution as it might not sum up to 1. We also note that in (1.9) the probability of full vector Z_S, $p(z_1, z_2, \ldots, z_{|S|})$, is not present neither in the denominator nor in the numerator. We denote the norming constant by

$$\eta = \sum_{z_1, z_2, \ldots, z_{|S|}} p^K(z_1, z_2, \ldots, z_{|S|})$$

and the normalized Kirkwood distribution by $\tilde{P}^K = P^K/\eta$.

Remark 1.2.3. *Alternative view of Definition 1.2.2 is that the interaction information is equal to the Kullback–Leibler divergence between joint probability of* $(Z_1, Z_2, \ldots, Z_{|S|})$ *and its Kirkwood approximation* $p^K(z_1, z_2, \ldots, z_{|S|})$

$$II(Z_1, Z_2, \ldots, Z_{|S|}) = D_{KL}(P||P^K).$$

Lemma 1.2.4. *If* $\eta \leq 1$ *defined in Definition 1.2.3 of the Kirkwood superposition approximation, the interaction information is non-negative and, analogously, if* $\eta < 1$ *then the interaction information is positive.*

Proof. We have that

$$D_{KL}(P||\tilde{P}^K) = D_{KL}(P||P^K) + \log(\eta) \tag{1.10}$$

and as the Kullback-Leibler divergence is non-negative for probability distributions P and $\tilde{P}^K$ we obtain that for $\eta \leq 1$

$$D_{KL}(P||P^K) \geq -\log(\eta) \geq 0. \tag{1.11}$$

Thus when $\eta \leq 1$ the interaction information is non-negative and analogously when $\eta < 1$ the interaction information is positive. We note that Kullback-Leibler divergence $D_{KL}(P||P^K)$ might be negative as P^K is not a proper probability distribution (it does not sum up to 1). □

Lemma 1.2.5 (Recursive formula for interaction information)**.** *The interaction information defined in* (1.6) *satisfies the recursive formula (or chain rule):*

$$II(Z_1, Z_2, \ldots, Z_{|S|}) = II(Z_1, Z_2, \ldots, Z_{(|S|-1)}|Z_{|S|}) - II(Z_1, Z_2, \ldots, Z_{(|S|-1)}),$$

where $II(Z_1, Z_2, \ldots, Z_{(|S|-1)}|Z_{|S|})$ *is defined in the following way (cf. Definition 1.1.2 of conditional entropy)*

$$II(Z_1, Z_2, \ldots, Z_{|S|-1}|Z_{|S|}) = \sum_{z_{|S|}} p(z_{|S|}) II(Z_1, Z_2, \ldots, Z_{|S|-1}|Z_{|S|} = z_{|S|}). \tag{1.12}$$

Proof. In view of Definition 1.7:

$$II(Z_1, Z_2, \ldots, Z_{|S|}) = (-1)^{|S|} \sum_{z_1, z_2, \ldots, z_{|S|}} p(z_1, z_2, \ldots, z_{|S|}) \log \frac{\prod\limits_{i_1<i_2} p(z_{i_1}, z_{i_2}) \cdots}{\prod\limits_{i_1} p(z_{i_1}) \prod\limits_{i_1<i_2<i_3} p(z_{i_1} z_{i_2}, z_{i_3}) \cdots}$$

$$= (-1)^{|S|} \sum_{z_1, z_2, \ldots, z_{|S|}} p(z_1, z_2, \ldots, z_{|S|}) \log \frac{\prod\limits_{i_1<|S|} p(z_{i_1}, z_{|S|}) \cdots}{p(z_{|S|}) \prod\limits_{i_1<i_2<|S|} p(z_{i_1}, z_{i_2}, z_{|S|}) \cdots}$$

$$+ (-1)^{|S|} \sum_{z_1, z_2, \ldots, z_{|S|}} p(z_1, z_2, \ldots, z_{|S|}) \log \frac{\prod\limits_{i_1<i_2<|S|} p(z_{i_1}, z_{i_2}) \cdots}{\prod\limits_{i_1<|S|} p(z_{i_1}) \prod\limits_{i_1<i_2<i_3<|S|} p(z_{i_1}, z_{i_2}, z_{i_3}) \cdots}$$

$$= (-1)^{|S|-1} \sum_{z_{|S|}} p(z_{|S|}) \sum_{z_1,z_2,\ldots,z_{|S|-1}} p(z_1, z_2, \ldots, z_{|S|-1}|z_{|S|}) \log \frac{\prod\limits_{i_1<i_2<|S|} p(z_{i_1} z_{i_2}|z_{|S|}) \cdots}{\prod\limits_{i_1<|S|} p(z_{i_1}|z_{|S|}) \cdots}$$

$$- (-1)^{|S|-1} \sum_{z_1,z_2,\ldots,z_{|S|-1}} p(z_1, z_2, \ldots, z_{|S|-1}) \log \frac{\prod\limits_{i_1<i_2<|S|} p(z_{i_1}, z_{i_2}) \cdots}{\prod\limits_{i_1<|S|} p(z_{i_1}) \prod\limits_{i_1<i_2<i_3<|S|} p(z_{i_1} z_{i_2}, z_{i_3}) \cdots}$$

$$= II(Z_1, Z_2, \ldots, Z_{|S|-1}|Z_{|S|}) - II(Z_1, Z_2, \ldots, Z_{|S|-1}).$$

The third equation follows from the fact that for any number of variables $|S| > 1$ we have $\sum_{k=0}^{|S|-1} (-1)^k \binom{|S|-1}{k} = (1-1)^{|S|-1} = 0$, thus we divide the numerator and the denominator of the first expression by the same number of $p(z_{|S|})$. □

Usually the formula stated in Lemma 1.2.5 is used as an alternative recursive definition of k-way interaction information. Namely, we define $II(Z_1, Z_2) = I(Z_1, Z_2)$ and $II(Z_1, Z_2|Z_3) = I(Z_1, Z_2|Z_3)$, then for $|S| \geq 3$ interaction information is defined in the following way

$$II(Z_1, Z_2, \ldots, Z_{|S|}) = II(Z_1, Z_2, \ldots, Z_{|S|-1}|Z_{|S|}) - II(Z_1, Z_2, \ldots, Z_{|S|-1}),$$

where in each step conditional $(k-1)$-way interaction information is defined in (1.12). We note that in formula (1.12) we average interaction informations of $Z_1, \ldots, Z_{|S|-1}$ computed on layers of $Z_{|S|} = z_{|S|}$ with respect to probabilities of the layer $Z_{|S|} = z_{|S|}$ equal to $p(z_{|S|})$.

Now we state some results showing the relation between k-way interaction information and other information theoretic measures as e.g. entropy. First we prove Theorem 1.2.6 and Corollary 1.2.6.1, which will be useful in proving statements below.

Theorem 1.2.6. *Let f and g be real functions on the power set of $\{Z_1, Z_2, \ldots, Z_{|S|}\}$ such that for any $T \subseteq \{Z_1, Z_2, \ldots, Z_{|S|}\}$ we have*

$$f(T) = \sum_{i=0}^{|T|} \sum_{\substack{W \subseteq T \\ |W|=i}} g(W). \tag{1.13}$$

Then

$$g(T) = \sum_{i=0}^{|T|} \sum_{\substack{W \subseteq T \\ |W|=i}} (-1)^{|T|-i} f(W). \tag{1.14}$$

Proof. We start with the right hand side of (1.14). We have that

$$\sum_{i=0}^{|T|}\sum_{\substack{W\subseteq T\\|W|=i}}(-1)^{|T|-i}f(W)=\sum_{i=0}^{|T|}\sum_{\substack{W\subseteq T\\|W|=i}}(-1)^{|T|-i}\left(\sum_{j=0}^{i}\sum_{\substack{Z\subseteq W\\|Z|=j}}g(Z)\right)$$
$$=\sum_{j=0}^{|T|}\sum_{\substack{Z\subseteq T\\|Z|=j}}\sum_{k=j}^{|T|}\binom{|T|-j}{k-j}(-1)^{|T|-k}g(Z)$$
$$=\sum_{j=0}^{|T|}\sum_{\substack{Z\subseteq T\\|Z|=j}}(-1)^{|T|-j}\sum_{k=j}^{|T|}\binom{|T|-j}{k-j}(-1)^{k-j}g(Z)=:(\star).$$

We have that

$$\sum_{k=j}^{|T|}\binom{|T|-j}{k-j}(-1)^{k-j}=\begin{cases}0 & \text{if } |T|-j>0\\ 1 & \text{if } j=|T|\end{cases}$$

as the sum equals $(1-1)^{|T|-j}$ for $|T|\neq j$. That completes the proof as $(\star)=g(T)$. □

Corollary 1.2.6.1. *Using the notation from Theorem 1.2.6 we have that if* (1.14) *is true then* (1.13) *holds.*

Proof. This is a simple consequence of Theorem 1.2.6 applied to the functions $\tilde{f}$ and $\tilde{g}$ defined as $\tilde{f}(T)=(-1)^{|T|}f(T)$ and $\tilde{g}(W)=(-1)^{|W|}g(W)$. □

We recall that from Lemma 1.2.2 we have

$$II(Z_1,Z_2,\ldots,Z_{|S|})=-\sum_{i=1}^{|S|}\sum_{\substack{T\subseteq S\\|T|=i}}(-1)^{|S|-|T|}H(Z_{t_1},Z_{t_2},\ldots,Z_{t_i}).$$

Below we assume that $H(\emptyset)=0$ and $II(\emptyset)=0$.

Corollary 1.2.6.2. *The entropy* $H(Z_1,\ldots,Z_{|S|})$ *satisfies*

$$H(Z_1,Z_2,\ldots,Z_{|S|})=-\sum_{i=1}^{|S|}\sum_{\substack{T\subseteq S\\|T|=i}}II(Z_{t_1},Z_{t_2},\ldots,Z_{t_i}).$$

Proof. We obtain the conclusion from Lemma 1.2.2 and Corollary 1.2.6.1 defining $f(T)=II(Z_{t_1},Z_{t_2},\ldots,Z_{t_i})$ and $g(T)=-H(Z_{t_1},Z_{t_2},\ldots,Z_{t_i})$, where $T=\{t_1,t_2,\ldots,t_i\}\subseteq S$. □

Lemma 1.2.7. *The conditional interaction information* $II(Z_1, Z_2, \ldots, Z_{|S|}|Y)$ *satisfies the equation*

$$II(Z_1, Z_2, \ldots, Z_{|S|}|Y) = -\sum_{i=1}^{|S|} \sum_{\substack{T \subseteq S \\ |T|=i}} (-1)^{|S|-|T|} H(Z_{t_1}, Z_{t_2}, \ldots, Z_{t_i}|Y).$$

Proof. The proof follows from Lemma 1.2.2 and (1.12) (the formula for conditional information interaction). □

Similarly as in Corollary 1.2.6.2, we assume here that $H(\emptyset|Y) = 0$ and $II(\emptyset|Y) = 0$.

Corollary 1.2.7.1. *The conditional entropy* $H(Z_1, \ldots, Z_{|S|}|Y)$ *satisfies the equation*

$$H(Z_1, Z_2, \ldots, Z_{|S|}|Y) = -\sum_{i=1}^{|S|} \sum_{\substack{T \subseteq S \\ |T|=i}} II(Z_{t_1}, Z_{t_2}, \ldots, Z_{t_i}|Y).$$

Proof. We obtain the thesis from Lemma 1.2.7 and Corollary 1.2.6.1. □

Lemma 1.2.8. *The interaction information satisfies the equation*

$$II(Y, Z_1, Z_2, \ldots, Z_{|S|}) = \sum_{i=1}^{|S|} \sum_{\substack{T \subseteq S \\ |T|=i}} (-1)^{|S|-|T|} I(Y, (Z_{t_1}, Z_{t_2}, \ldots, Z_{t_i})).$$

Proof. The proof follows from

$$I(Y, (Z_{t_1}, Z_{t_2}, \ldots, Z_{t_i})) = H(Z_{t_1}, Z_{t_2}, \ldots, Z_{t_i}) - H(Z_{t_1}, Z_{t_2}, \ldots, Z_{t_i}|Y)$$

and Lemmas 1.2.2 and 1.2.7, which represent the interaction information by a sum of the entropies and Lemma 1.2.5 (a chain rule for interaction information). □

Corollary 1.2.8.1. *The mutual information satisfies the equation*

$$I(Y, (Z_1, Z_2, \ldots, Z_{|S|})) = \sum_{i=1}^{|S|} \sum_{\substack{T \subseteq S \\ |T|=i}} II(Y, Z_{t_1}, Z_{t_2}, \ldots, Z_{t_i})$$

Proof. The proof follows from Lemma 1.2.8 and Corollary 1.2.6.1. □

Lemma 1.2.9. *The interaction information satisfies the equation*

$$II(X,Y,Z_1,Z_2,\ldots,Z_{|S|}) = \sum_{i=0}^{|S|}\sum_{\substack{T\subseteq S\\|T|=i}}(-1)^{|S|-|T|}I(X,Y|Z_{t_1},Z_{t_2},\ldots,Z_{t_i}).$$

Proof. For the first equation below we use Lemma 1.2.8 and we split the mutual information terms into two groups - one with terms that contain the variable X and the second that do not:

$$\begin{aligned}
II(X,Y,Z_1,Z_2,\ldots,Z_{|S|}) &= \sum_{i=0}^{|S|}\sum_{\substack{T\subseteq S\\|T|=i}}(-1)^{|S|+1-|T|-1}I(Y,(X,Z_{t_1},Z_{t_2},\ldots,Z_{t_i}))\\
&+\sum_{i=1}^{|S|}\sum_{\substack{T\subseteq S\\|T|=i}}(-1)^{|S|+1-|T|}I(Y,(Z_{t_1},Z_{t_2},\ldots,Z_{t_i})) = (-1)^{|S|}I(X,Y)\\
&+\sum_{i=1}^{|S|}\sum_{\substack{T\subseteq S\\|T|=i}}(-1)^{|S|-|T|}\left(I(Y,(Z_{t_1},Z_{t_2},\ldots,Z_{t_i})) + I(X,Y|Z_{t_1},Z_{t_2},\ldots,Z_{t_i})\right)\\
&+\sum_{i=1}^{|S|}\sum_{\substack{T\subseteq S\\|T|=i}}(-1)^{|S|+1-|T|}I(Y,(Z_{t_1},Z_{t_2},\ldots,Z_{t_i}))\\
&= (-1)^{|S|}I(X,Y) + \sum_{i=1}^{|S|}\sum_{\substack{T\subseteq S\\|T|=i}}(-1)^{|S|-|T|}I(X,Y|Z_{t_1},Z_{t_2},\ldots,Z_{t_i}).
\end{aligned}$$

For the second equality a chain rule for mutual information is used (Lemma 1.1.2). □

Corollary 1.2.9.1. *If $X \perp\!\!\!\perp Y|W$, where W is any subset of $\{Z_1,Z_2,\ldots,Z_{|S|}\}$ (including $\emptyset$) then $II(X,Y,Z_1,Z_2,\ldots,Z_{|S|}) = 0$.*

Proof. The proof follows in straightforward manner from Lemma 1.2.9. □

Theorem 1.2.10 (Möbius expansion of conditional mutual information)**.** *The conditional mutual information satisfies the equation*

$$I(X,Y|Z_1,Z_2,\ldots,Z_{|S|}) = \sum_{i=0}^{|S|}\sum_{\substack{T\subseteq S\\|T|=i}}II(X,Y,Z_{t_1},Z_{t_2},\ldots,Z_{t_i}). \tag{1.15}$$

Proof. The proof follows from Corollary 1.2.6.1 and Lemma 1.2.9. □

Remark 1.2.11. *The definition of k-way interaction information is also given in [16]. The formulation is lattice-theoretic ([32]) and it uses difference operator (for odd number of variables the sign in [16] is switched). All subsets of a set of random variables* $\{Z_1, Z_2, \ldots, Z_{|S|}\}$ *form Boolean lattice with set union and intersection operations. It also contains the maximum* ($\{Z_1, Z_2, \ldots, Z_{|S|}\}$) *and the minimum element* ($\emptyset$) *and the set-inclusion relation induces the partial order. Let* $T = \{t_1, \ldots, t_k\}$, $|T| = k$, *be any subset of* S. *We define the entropy on a subset of random variables as*

$$h(T) = H(Z_{t_1}, Z_{t_2}, \ldots, Z_{t_k}) = -\sum_{z_{t_1}, z_{t_2}, \ldots, z_{t_k}} p(z_{t_1}, z_{t_2}, \ldots, z_{t_k}) \log p(z_{t_1}, z_{t_2}, \ldots, z_{t_k})$$

and $h(\emptyset) = H(\emptyset) = 0$. *The difference operator for a function on a Boolean lattice is defined as*

$$\Delta f(T) = \sum_{i=0}^{|T|} \sum_{\substack{W \subseteq T \\ |W|=i}} (-1)^{|T|-|W|} f(W).$$

Using Möbius inversion formula (Theorem 1.2.10) and noting, that for a Boolean lattice the principle of inclusion-exclusion holds, we have

$$f(T) = \sum_{i=0}^{|T|} \sum_{\substack{W \subseteq T \\ |W|=i}} \Delta f(W).$$

The difference operator for entropy is given by

$$\Delta h(T) = \sum_{i=0}^{|T|} \sum_{\substack{W \subseteq T \\ |W|=i}} (-1)^{|T|-|W|} h(W), \tag{1.16}$$

and if we multiply both sides by -1, *we obtain the definition of interaction information (or as called in [16] McGill's (or Fano's) multiple mutual information) as* $-\Delta h(T) = II(Z_{t_1}, Z_{t_2}, \ldots, Z_{t_k})$.

1.3. Feature selection criteria

In this section we introduce feature selection criteria, which can be used as substitutes for CMI-based selection in choosing a subset of significant variables. First we introduce

generalised feature selection criterion $I_{\beta,\gamma}$, which uses the Möbius expansion of CMI (Theorem 1.2.10) and assigns weights to interaction informations and conditional interaction informations terms appearing in the formula

$$I(X,Y|Z) = I(X,Y) - \sum_{k=1}^{|S|} \sum_{\substack{T\subseteq S \\ |T|=k}} \left(II(X, Z_{t_1}, Z_{t_2}, \ldots, Z_{t_k}) - II(X, Z_{t_1}, Z_{t_2}, \ldots, Z_{t_k}|Y)\right),$$

which is rearranged version of the equation (1.15) from Theorem 1.2.10. This allows us to obtain feature selection criteria used in literature, in particular, deleting all expansion terms of order higher than 2 leads to the criterion $J_{\beta,\gamma}$ with two parameters introduced in [6]. Assigning specific values to parameters in $J_{\beta,\gamma}$ yields the criterion called JMI ([43]). In the next section we will state lemmas about the asymptotic behaviour of empirical versions of these three criteria.

1.3.1. The generalized feature selection criterion based on Möbius expansion

We start by defining the most general form of the criterion $I_{\alpha,\beta}$ and then we introduce two parameter reduced version of it, namely $J_{\beta,\gamma}$ which appears in [6].

Definition 1.3.1 (Generalized feature selection criterion)**.** The generalised feature selection criterion $I_{\beta,\gamma}$, where β and γ are vectors of parameters in $\mathbb{R}^{|S|}$, is defined in the following way

$$I_{\beta,\gamma}(X,Y|Z_S) = I(X,Y) - \sum_{k=1}^{|S|} \sum_{\substack{T\subseteq S \\ |T|=k}} \big(\beta(k) II(X, Z_{t_1}, Z_{t_2}, \ldots, Z_{t_k}) - \gamma(k) II(X, Z_{t_1}, Z_{t_2}, \ldots, Z_{t_k}|Y)\big). \quad (1.17)$$

For convenience we will use two additional parameters $\beta(0) = 1$ and $\gamma(0) = 1$ and write mutual information in Definition 1.3.1 as

$$I(X,Y) = \beta(0)H(X) - \gamma(0)H(X|Y).$$

Frequently, many parameters among $\beta(i)$ and $\gamma(i)$ are equal to 0, as higher order terms are deleted in order to make estimation of feature selection criterion feasi-

ble. Moreover, the family of criteria $I_{\beta,\gamma}$ includes conditional mutual information CMI, as $I_{\beta,\gamma}(X,Y|Z_S) = I(X,Y|Z_S)$ for $\beta(i) = 1$ and $\gamma(i) = 1$ for $i = 1, 2, \ldots, |S|$.

Below we establish another representation of $I_{\beta,\gamma}$ in terms of entropies, which is particularly useful for convergence lemma and differentiation performed in the next section (cf. Theorem 1.4.6 and its proof).

Theorem 1.3.1. *We have the following representation of* $I_{\beta,\gamma}$

$$I_{\beta,\gamma}(X,Y|Z_S) = \sum_{k=1}^{|S|} \sum_{\substack{T\subseteq S\\|T|=k}} \tilde{\beta}_k^{|S|} H(Z_{t_1}, Z_{t_2}, \ldots, Z_{t_k}) - \sum_{k=0}^{|S|} \sum_{\substack{T\subseteq S\\|T|=k}} \tilde{\beta}_k^{|S|} H(X, Z_{t_1}, Z_{t_2}, \ldots, Z_{t_k})$$
$$- \sum_{k=0}^{|S|} \sum_{\substack{T\subseteq S\\|T|=k}} \tilde{\gamma}_k^{|S|} H(Y, Z_{t_1}, Z_{t_2}, \ldots, Z_{t_k}) + \sum_{k=0}^{|S|} \sum_{\substack{T\subseteq S\\|T|=k}} \tilde{\gamma}_k^{|S|} H(X, Y, Z_{t_1}, Z_{t_2}, \ldots, Z_{t_k}), \quad (1.18)$$

where

$$\tilde{\beta}_k^{|S|} = \sum_{j=k}^{|S|} (-1)^{j-k+1} \binom{|S|-k}{j-k} \beta(j) \quad \text{and} \quad \tilde{\gamma}_k^{|S|} = \sum_{j=k}^{|S|} (-1)^{j-k+1} \binom{|S|-k}{j-k} \gamma(j).$$

Proof. We write

$$I_{\beta,\gamma}(X,Y|Z_S) = \beta(0)H(X) - \sum_{k=1}^{|S|} \sum_{\substack{T\subseteq S\\|T|=k}} \beta(k) II(X, Z_{t_1}, Z_{t_2}, \ldots, Z_{t_k})$$
$$- \gamma(0)H(X|Y) + \sum_{k=1}^{|S|} \sum_{\substack{T\subseteq S\\|T|=k}} \gamma(k) II(X, Z_{t_1}, Z_{t_2}, \ldots, Z_{t_k}|Y)$$

and deal first with the two first summands. We have from Lemma 1.2.2

$$\beta(0)H(X) - \sum_{k=1}^{|S|} \sum_{\substack{T\subseteq S\\|T|=k}} \beta(k) II(X, Z_{t_1}, Z_{t_2}, \ldots, Z_{t_k}) = \beta(0)H(X)$$
$$+ \sum_{k=1}^{|S|} \sum_{\substack{T\subseteq S\\|T|=k}} \beta(k) \Bigg[\sum_{l=1}^{k} \sum_{\substack{R\subseteq T\\|R|=l}} (-1)^{k+1-l} H(Z_{r_1}, Z_{r_2}, \ldots, Z_{r_l})$$
$$- \sum_{l=0}^{k} \sum_{\substack{R\subseteq T\\|R|=l}} (-1)^{k+1-l} H(X, Z_{r_1}, Z_{r_2}, \ldots, Z_{r_l}) \Bigg] =: (*),$$

thus we obtain

$$
\begin{aligned}
(*) &= \sum_{l=1}^{|S|} \sum_{\substack{R\subseteq S\\|R|=k}} \underbrace{\left[\sum_{j=l}^{|S|}(-1)^{j-l+1}\binom{|S|-l}{j-l}\beta(j)\right]}_{\tilde{\beta}_l^{|S|}} H(Z_{t_1}, Z_{t_2}, \ldots, Z_{t_l}) \\
&\quad - \sum_{l=0}^{|S|} \sum_{\substack{R\subseteq S\\|R|=l}} \underbrace{\left[\sum_{j=l}^{|S|}(-1)^{j-l+1}\binom{|S|-l}{j-l}\beta(j)\right]}_{\tilde{\beta}_l^{|S|}} H(X, Z_{t_1}, Z_{t_2}, \ldots, Z_{t_l}) \\
&= \sum_{l=1}^{|S|} \sum_{\substack{R\subseteq S\\|R|=l}} \tilde{\beta}_l^{|S|} H(Z_{t_1}, Z_{t_2}, \ldots, Z_{t_l}) - \sum_{l=0}^{|S|} \sum_{\substack{R\subseteq S\\|R|=l}} \tilde{\beta}_l^{|S|} H(X, Z_{t_1}, Z_{t_2}, \ldots, Z_{t_l}) \\
&= \sum_{k=1}^{|S|} \sum_{\substack{T\subseteq S\\|T|=k}} \tilde{\beta}_k^{|S|} H(Z_{t_1}, Z_{t_2}, \ldots, Z_{t_k}) - \sum_{k=0}^{|S|} \sum_{\substack{T\subseteq S\\|T|=k}} \tilde{\beta}_k^{|S|} H(X, Z_{t_1}, Z_{t_2}, \ldots, Z_{t_k}).
\end{aligned}
$$

Note that $\beta(0)H(X)$ has been included as the first term of the sum corresponding to $k = 0$ in the second term above. Similarly as above we obtain the expression for γ terms using Lemma 1.2.7:

$$
\begin{aligned}
&- \gamma(0)H(X|Y) + \sum_{k=1}^{|S|} \sum_{\substack{T\subseteq S\\|T|=k}} \gamma(k) II(X, Z_{t_1}, Z_{t_2}, \ldots, Z_{t_k}|Y) \\
&= \sum_{k=0}^{|S|} \sum_{\substack{T\subseteq S\\|T|=k}} \tilde{\gamma}_k^{|S|} H(X, Z_{t_1}, Z_{t_2}, \ldots, Z_{t_k}|Y) - \sum_{k=1}^{|S|} \sum_{\substack{T\subseteq S\\|T|=k}} \tilde{\gamma}_k^{|S|} H(Z_{t_1}, Z_{t_2}, \ldots, Z_{t_k}|Y) =: (\triangle).
\end{aligned}
$$

Now we use the chain rule for entropy and as terms including $H(Y)$ reduce for $k \geq 1$, after rearranging, $(\triangle)$ is equal to

$$
\begin{aligned}
(\triangle) &= \tilde{\gamma}_0^{|S|}(H(Y) - H(X,Y)) \\
&\quad + \sum_{k=1}^{|S|} \sum_{\substack{T\subseteq S\\|T|=k}} \tilde{\gamma}_k^{|S|} H(X, Y, Z_{t_1}, Z_{t_2}, \ldots, Z_{t_k}) - \sum_{k=1}^{|S|} \sum_{\substack{T\subseteq S\\|T|=k}} \tilde{\gamma}_k^{|S|} H(Y, Z_{t_1}, Z_{t_2}, \ldots, Z_{t_k}) \\
&= \sum_{k=0}^{|S|} \sum_{\substack{T\subseteq S\\|T|=k}} \tilde{\gamma}_k^{|S|} H(X, Y, Z_{t_1}, Z_{t_2}, \ldots, Z_{t_k}) - \sum_{k=0}^{|S|} \sum_{\substack{T\subseteq S\\|T|=k}} \tilde{\gamma}_k^{|S|} H(Y, Z_{t_1}, Z_{t_2}, \ldots, Z_{t_k}),
\end{aligned}
$$

where the last equality follows as we include both $-\tilde{\gamma}_0^{|S|}H(X,Y)$ and $\tilde{\gamma}_0^{|S|}H(Y)$ in the first and second sum, respectively. As $I_{\beta,\gamma} = (*) + (\triangle)$, the theorem follows. □

1.3.2. Second and third order criteria

We show that the family of second order criteria introduced in [6] are special cases of (1.17). The criterion introduced in [6] is parametrized by two parameters β and γ in the following way

$$J_{\beta,\gamma}(X,Y|Z_S) = I(X,Y) - \beta\sum_{i\in S} I(X,Z_i) + \gamma\sum_{i\in S} I(X,Z_i|Y). \tag{1.19}$$

We use the notation $J_{\beta,\gamma}$ to distinguish two-parameter criterion from generalised criterion $I_{\beta,\gamma}$ introduced in (1.17). We use the same letters for parameters, but in $I_{\beta,\gamma}$ the parameters are vectors and we denote their i-th element by $\beta(i)$ and $\gamma(i)$, thus they are easily distinguishable from β and γ related to $J_{\beta,\gamma}$. We obtain $J_{\beta,\gamma}$ by replacing parameters of order higher than two in $I_{\beta,\gamma}$ by 0, namely we have $\beta(1) = \beta$, $\gamma(1) = \gamma$ and $\beta(i) = \gamma(i) = 0$ for $i > 1$. In terms of $\tilde{\beta}$ and $\tilde{\gamma}$ from Theorem 1.3.1 the $I_{\beta,\gamma}$ and $J_{\beta,\gamma}$ criteria are equivalent for

$$\begin{array}{lll} \tilde{\beta}_0^{|S|} = -\beta(0) + |S|\beta(1) = \beta|S| - 1 & & \tilde{\gamma}_0^{|S|} = -\gamma(0) + |S|\gamma(1) = \gamma|S| - 1 \\ \tilde{\beta}_1^{|S|} = -\beta(1) = -\beta & \text{and} & \tilde{\gamma}_1^{|S|} = -\gamma(1) = -\gamma \\ \tilde{\beta}_i^{|S|} = 0 \text{ for } i > 1 & & \tilde{\gamma}_i^{|S|} = 0 \text{ for } i > 1 \end{array} . \tag{1.20}$$

We will frequently use the criterion called JMI (*Joint Mutual Information*) introduced in [43] and defined as

$$JMI(X,Y|Z_S) = I(X,Y) - \frac{1}{|S|}\sum_{i\in S}(I(X,Z_i) - I(X,Z_i|Y)). \tag{1.21}$$

From the equation above we see that JMI is a special case of $J_{\beta,\gamma}$, as we obtain the criterion by choosing $\beta = \gamma = 1/|S|$. JMI is an approximation of CMI under certain dependence assumptions (cf. [42]). The criterion averages conditional mutual informations of X and Y given individual variables Z_i over $i \in S$. Indeed, JMI can be represented in the following way

$$JMI(X,Y|Z_S) = \frac{1}{|S|}\sum_{i\in S} I(X,Y|Z_i). \tag{1.22}$$

which follows from using a chain rule twice for $I(X,(Y,Z_i))$:

$$I(X,Z_i)+I(X,Y|Z_i)=I(X,Y)+I(X,Z_i|Y).$$

Thus we also have that

$$JMI(X,Y|Z_S)=0 \Leftrightarrow X \perp\!\!\!\perp Y|Z_i \text{ for all } i\in S.$$

The third-order version of JMI is introduced in [36] and equals

$$\sum_{\{i,j\}\subseteq S} I((X,Z_i,Z_j),Y). \tag{1.23}$$

We note, that up to a constant $c=\sum_{\{i,j\}\subseteq S} I((Z_i,Z_j),Y)$ which does not depend on X and scaling factor $2/(|S|(|S|-1))$, (1.23) equals to (cf. Theorem 1.2.10)

$$\begin{aligned} JMI3(X,Y|Z_S) &= \frac{2}{|S|(|S|-1)}\sum_{\{i,j\}\subseteq S} I(X,Y|Z_i,Z_j) \\ &= I(X,Y)+\frac{2}{|S|}\sum_{\{i\}\subseteq S} II(X,Y,Z_i)+\frac{1}{\binom{|S|}{2}}\sum_{\{i,j\}\subseteq S} II(X,Y,Z_i,Z_j). \end{aligned} \tag{1.24}$$

Thus $JMI3(X,Y|Z)=I_{\beta,\gamma}(X,Y|Z)$ for $\beta(1)=\gamma(1)=2/|S|$, $\beta(2)=\gamma(2)=1/\binom{|S|}{2}$ and $\beta(i)=\gamma(i)=0$ for $i>2$.

Another popular criterion is called $CIFE$ (*Conditional Infomax Feature Extraction* [26]) or $SECMI$ (*Short Expansion of Conditional Mutual Information* [20]) which is truncated Möbius expansion (1.15) of order two and thus equals

$$CIFE(X,Y|Z_S)=I(X,Y)+\sum_{\{i\}\subseteq S} II(X,Y,Z_i). \tag{1.25}$$

If we use the first three components of the expansion we obtain a third order expansion

$$SECMI3(X,Y|Z_S)=I(X,Y)+\sum_{\{i\}\subseteq S} II(X,Y,Z_i)+\sum_{\{i,j\}\subseteq S} II(X,Y,Z_i,Z_j). \tag{1.26}$$

It is easy to generalise it to $SECMI$-k (then in terms of $I_{\beta,\gamma}$ we have $\beta(i)=\gamma(i)=1$ for $i<k$ and $\beta(i)=\gamma(i)=0$ for $i\geq k$).

Among other second order selection criteria are e.g. (we give the corresponding parameters in brackets) MIFS (*Mutual Information Feature Selection* [3], $(\beta, \gamma) = (\beta, 0)$) and mRMR (*Minimal Redundancy Maximal Relevance* [30], $(\beta, \gamma) = (1/|S|, 0)$).

In the next section we will establish the behaviour of a sample counterparts of conditional mutual information and feature selection criteria. Comparison of properties of theoretical measures JMI, $CIFE$ and CMI under specific dependence structure was studied in [23].

1.4. Asymptotic distributions of information-theoretic empirical measures

Let $(X_1, Y_1, Z_1), (X_2, Y_2, Z_2), \ldots, (X_n, Y_n, Z_n)$ be independent random variables (Z_i might be multivariate) with common discrete distribution

$$P(X_i = x, Y_i = y, Z_i = z) = p(x, y, z) > 0.$$

We recall that p_{ci} denotes probability mass function corresponding to $X \perp\!\!\!\perp Y|Z$ equal $p_{ci}(x, y, z) = p(x|z)p(y|z)p(z)$. We begin this section by stating how unconstrained maximum likelihood estimator of a vector of probabilities $(p(x, y, z))_{x,y,z}$, namely a vector of fractions $(\hat{p}(x, y, z))_{x,y,z}$, where $\hat{p}(x, y, z) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i = x, Y_i = y, Z_i = z) = n(x, y, z)/n$, behaves asymptotically. Then using that result and the delta method ([1]) we will establish the behaviour of the plug-in conditional mutual information estimator and of the plug-in estimator of a generalised feature selection criterion based on Möbius expansion introduced in the previous section. As a special case we will consider JMI criterion. We also note that asymptotic behaviour of sample version of interaction information $II(X, Y, Z)$ was analysed in [19].

From now on $\xrightarrow{d}$ denotes convergence in distribution, $\Sigma_{x,y,z}^{x',y',z'}$ an element of the matrix Σ with row index (x, y, z) and column index (x', y', z'), where $(x, y, z), (x', y', z') \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ and $(p(x, y, z))_{x,y,z}$ denotes a vector of probabilities p for all the triples (x, y, z) such that $(x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$.

Lemma 1.4.1. *We have that*

$$\sqrt{n}\left(\hat{p}(x,y,z) - p(x,y,z)\right)_{x,y,z} \xrightarrow{d} \mathcal{N}(0,\Sigma)$$

as $n \to \infty$*, where*

$$\Sigma_{x,y,z}^{x',y',z'} = \mathbb{I}(x = x', y = y', z = z')p(x,y,z) - p(x,y,z)p(x',y',z') \tag{1.27}$$

and $\hat{p}(x,y,z) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i = x, Y_i = y, Z_i = z)$.

A proof can be found in [1], Subsection 16.1.4.

1.4.1. Asymptotic behaviour of the empirical conditional mutual information

One of the aims of establishing asymptotic behaviour of a sample version of conditional mutual information is to test conditional independence of two variables X and Y given the third one Z, using a plug-in estimator of CMI, namely

$$\widehat{CMI} = CMI(\hat{p}) = \sum_{x,y,z} \hat{p}(x,y,z) \log \frac{\hat{p}(x,y,z)\hat{p}(z)}{\hat{p}(x,z)\hat{p}(y,z)},$$

where

$$\hat{p}(x,y,z) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i = x, Y_i = y, Z_i = z)$$

and analogously for $\hat{p}(x,z)$, $\hat{p}(y,z)$ and $\hat{p}(z)$ (we replace the probabilities with corresponding fractions). In literature, the test based on asymptotic approximation of $\widehat{CMI}$ stated in Lemma 1.4.2 is known as G^2 or G-test and can be derived as the log-likelihood ratio test [1].

Lemma 1.4.2. *If* $X \perp\!\!\!\perp Y|Z$ *we have that*

$$2nCMI(\hat{p}) \xrightarrow{d} \chi^2_{(|\mathcal{X}|-1)(|\mathcal{Y}|-1)|\mathcal{Z}|},$$

where $CMI(\hat{p}) = \sum_{x,y,z} \hat{p}(x,y,z) \log \frac{\hat{p}(x,y,z)\hat{p}(z)}{\hat{p}(x,z)\hat{p}(y,z)}$.

Below we present a straightforward proof of Lemma 1.4.2, the second part of which is based on calculation of eigenvalues of a matrix $M := H\Sigma$, where H is a Hessian matrix of CMI as a function of p, and Σ is defined in Lemma 1.4.1. We will use its form in Section 2

to prove analogous convergence theorems for resampling scenarios. Different proof based on showing that $CMI(\hat{p})$ can be approximated by the chi-square statistics as n tends to infinity can be found e.g. in [29].

Proof. First, we compute the gradient and Hessian matrix of conditional mutual information considered as a function of $(p(x,y,z))_{x,y,z}$

$$CMI(p) = \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} = \sum_{x,y,z} p(x,y,z) \log p(x,y,z) \\ - \sum_{x,z} p(x,z) \log p(x,z) - \sum_{y,z} p(y,z) \log p(y,z) + \sum_{z} p(z) \log p(z).$$

The derivative of a function $f(x) = x\log(x)$ equals $f'(x) = \log(x) + 1$ and the derivative of $f(x,z) = \sum_y p(x,y,z) = p(x,z)$ with respect to $p(x',y',z')$ equals $\frac{\partial f}{\partial p(x',y',z')} = \mathbb{I}(x = x', z = z')$. Analogously, if $f(y,z) = \sum_x p(x,y,z) = p(y,z)$, we have $\frac{\partial f}{\partial p(x',y',z')} = \mathbb{I}(y = y', z = z')$ and if $f(z) = \sum_{x,y} p(x,y,z) = p(z)$, we have $\frac{\partial f}{\partial p(x',y',z')} = \mathbb{I}(z = z')$. Thus we obtain

$$(D_{CMI}(p))(x,y,z) = \frac{\partial CMI(p)}{\partial p(x,y,z)} = \log \frac{p(x,y,z)p(z)}{p(x,z)p(y,z)}, \tag{1.28}$$

where $(D_{CMI}(p))(x,y,z)$ denotes the element of the vector $D_{CMI}(p)$ with an index corresponding to (x,y,z). Similarly, the Hessian equals

$$(H_{CMI}(p))^{x',y',z'}_{x,y,z} = \frac{\partial^2 CMI(p)}{\partial p(x,y,z)\partial p(x',y',z')} = \frac{\mathbb{I}(x = x', y = y', z = z')}{p(x,y,z)} \\ - \frac{\mathbb{I}(x = x', z = z')}{p(x,z)} - \frac{\mathbb{I}(y = y', z = z')}{p(y,z)} + \frac{\mathbb{I}(z = z')}{p(z)}. \tag{1.29}$$

The proof now follows now from the expansion

$$CMI(\hat{p}) = CMI(p) + (\hat{p} - p)' D_{CMI}(p) + \frac{1}{2}(\hat{p} - p)' H_{CMI}(\xi)(\hat{p} - p), \tag{1.30}$$

where $\xi = (\xi_{x,y,z})_{x,y,z}$ is a point between $\hat{p}$ and p. As $\hat{p} \to p$ a.s. and H_{CMI} is continuous if $p(x,y,z) > 0$ for all (x,y,z), then $H_{CMI}(\xi) \to H_{CMI}(p)$ a.s. Hence we have

$$CMI(\hat{p}) = CMI(p) + (\hat{p} - p)' D_{CMI}(p) + \frac{1}{2}(\hat{p} - p)' H_{CMI}(p)(\hat{p} - p) + o_p(\|\hat{p} - p\|^2).$$

Using the assumption $X \perp\!\!\!\perp Y|Z$, we have that $p(x,y,z) = p_{ci}(x,y,z) =: p(x|z)p(y|z)p(z)$, thus the gradient of CMI equals 0 as

$$D_{CMI}(p) = \log \frac{p(x,y,z)p(z)}{p(x,z)p(y,z)} = \log \frac{p(x,y|z)}{p(x|z)p(y|z)} = 0.$$

Similarly, we have that $CMI(p_{ci}) = 0$, hence from (1.30) we obtain

$$2nCMI(\hat{p}) = \sqrt{n}(\hat{p}-p)' H_{CMI}(p)\sqrt{n}(\hat{p}-p) + o_p(n\left\|\hat{p}-p\right\|^2).$$

Then, as $o_p(n\left\|\hat{p}-p\right\|^2) = o_p(1)$ in view of Lemma 1.4.1, we have that

$$2nCMI(\hat{p}) \xrightarrow{d} W' H_{CMI}(p_{ci}) W,$$

where $W \sim \mathcal{N}(0,\Sigma)$ and Σ is defined in (1.27). Thus,

$$2nCMI(\hat{p}) \xrightarrow{d} \sum_{x,y,z} \lambda_{x,y,z} Z^2_{x,y,z}, \tag{1.31}$$

where $Z = (Z_{x,y,z})_{x,y,z} \sim \mathcal{N}(0,I)$ and $\lambda_{x,y,z}$ are eigenvalues of a matrix $M = H_{CMI}(p_{ci})\Sigma$. This can be justified as follows: in view of spectral decomposition of $\Sigma^{1/2} H_{CMI}(p_{ci})\Sigma^{1/2}$, $W' H_{CMI}(p_{ci}) W$ can be represented as

$$W' H_{CMI}(p_{ci}) W = \tilde{Z}' \Sigma^{1/2} H_{CMI}(p_{ci}) \Sigma^{1/2} \tilde{Z} = \sum_{x,y,z} \lambda_{x,y,z} \tilde{Z}' v_{x,y,z} v'_{x,y,z} \tilde{Z} = \sum_{x,y,z} \lambda_{x,y,z} Z^2_{x,y,z},$$

where $\tilde{Z} \sim \mathcal{N}(0,I)$, $v_{x,y,z}$ are normalised eigenvectors and $\lambda_{x,y,z}$ are eigenvalues of the matrix $\Sigma^{1/2} H_{CMI} \Sigma^{1/2}$. Moreover, $Z_{x,y,z} = v'_{x,y,z}\tilde{Z}$. As $v_{x,y,z}$ are orthonormal, $Z \sim \mathcal{N}(0,I)$. The matrices $\Sigma^{1/2} H_{CMI}(p_{ci})\Sigma^{1/2}$ and $H_{CMI}(p_{ci})\Sigma$ have the same eigenvalues, thus we obtain (1.31). In Lemma 1.4.3 below we obtain an explicit formula of the matrix M. Then from Lemma 1.4.5 we have that $M = M^2$, thus $\lambda_i = 0$ or $\lambda_i = 1$ and from Lemma 1.4.4 it follows that $\sum_i \lambda_i = (|\mathcal{X}|-1)(|\mathcal{Y}|-1)|\mathcal{Z}|$. Thus

$$2nCMI(\hat{p}) \xrightarrow{d} \chi^2_{(|\mathcal{X}|-1)(|\mathcal{Y}|-1)|\mathcal{Z}|}. \qquad \square$$

Now in Lemma 1.4.3 we show the explicit formula for the matrix M defined in the

proof of Lemma 1.4.2, and in Lemmas 1.4.4 and 1.4.5 the properties of M used to prove Lemma 1.4.2.

Lemma 1.4.3. *Matrix $M = H\Sigma = H_{CMI}(p_{ci})\Sigma$ has the following form*

$$M_{x,y,z}^{x'',y'',z''} = \mathbb{I}(z = z'')\big(\mathbb{I}(x = x'', y = y'') - \mathbb{I}(x = x'')p(y''|z'') - \mathbb{I}(y = y'')p(x''|z'') + p(x''|z'')p(y''|z'')\big). \quad (1.32)$$

We note that M is a sparse matrix such that all its non-zero elements $M_{x,y,z}^{x'',y'',z''}$ satisfy $z = z''$. Moreover, it is not symmetric.

Proof. Multiplication of matrices H and Σ yields:

$$\begin{aligned}
M_{x,y,z}^{x'',y'',z''} &= \sum_{x',y',z'} H_{x,y,z}^{x',y',z'}\Sigma_{x',y',z'}^{x'',y'',z''} = \sum_{x',y',z'}\Bigg(\underbrace{\frac{\mathbb{I}(x = x', y = y', z = z')}{p(x,y,z)}}_{a} - \underbrace{\frac{\mathbb{I}(x = x', z = z')}{p(x,z)}}_{b} \\
&- \underbrace{\frac{\mathbb{I}(y = y', z = z')}{p(y,z)}}_{c} + \underbrace{\frac{\mathbb{I}(z = z')}{p(z)}}_{d}\Bigg)\Bigg(\underbrace{\mathbb{I}(x' = x'', y' = y'', z' = z'')p(x'|z')p(y',z')}_{e} \\
&- \underbrace{p(x'|z')p(y',z')p(x''|z'')p(y'',z'')}_{f}\Bigg) = \underbrace{\mathbb{I}(x = x'', y = y'', z = z'')}_{a\cdot e} \\
&- \underbrace{\mathbb{I}(x = x'', z = z'')p(y''|z'')}_{b\cdot e} - \underbrace{\mathbb{I}(x = x'', z = z'')p(x''|z'')}_{c\cdot e} + \underbrace{\mathbb{I}(z = z'')p(x''|z'')p(y''|z'')}_{d\cdot e} \\
&- \underbrace{p(x''|z'')p(y'',z'')}_{a\cdot f} + \underbrace{p(x''|z'')p(y'',z'')}_{b\cdot f} + \underbrace{p(x''|z'')p(y'',z'')}_{c\cdot f} \\
&- \underbrace{p(x''|z'')p(y'',z'')}_{d\cdot f} = \mathbb{I}(x = x'', y = y'', z = z'') - \mathbb{I}(x = x'', z = z'')p(y''|z'') \\
&- \mathbb{I}(y = y'', z = z'')p(x''|z'') + \mathbb{I}(z = z'')p(x''|z'')p(y''|z'').
\end{aligned}$$

Below we present detailed calculations for the terms $c \cdot f$ and $d \cdot e$ (the calculations for other terms are analogous):

$$\begin{aligned}
c \cdot f &= \sum_{x',y',z'} \mathbb{I}(y = y', z = z')\frac{p(x'|z')p(y',z')p(x''|z'')p(y'',z'')}{p(y,z)} \\
&= p(x''|z'')p(y'',z'')\sum_{x'} p(x'|z) = p(x''|z'')p(y'',z''), \\
b \cdot e &= \sum_{x',y',z'} \mathbb{I}(x = x', z = z')\mathbb{I}(x' = x'', y' = y'', z' = z'')\frac{p(x'|z')p(y',z')}{p(x,z)}
\end{aligned}$$

$$= \mathbb{I}(x = x'', z = z'')p(y''|z''). \qquad \square$$

Lemma 1.4.4. *The trace of M defined in Lemma 1.4.3 equals*

$$tr(M) = (|\mathcal{X}| - 1)(|\mathcal{Y}| - 1)|\mathcal{Z}|,$$

where $|\mathcal{Z}| = \prod_i |\mathcal{Z}_i|$.

Proof.

$$\begin{aligned}\sum_{x,y,z} M_{x,y,z}^{x,y,z} &= \sum_{x,y,z}(1 - p(y|z) - p(x|z) + p(x|z)p(y|z)) \\ &= |\mathcal{X}| \cdot |\mathcal{Y}| \cdot |\mathcal{Z}| - |\mathcal{X}| \cdot |\mathcal{Z}| - |\mathcal{Y}| \cdot |\mathcal{Z}| + |\mathcal{Z}| = (|\mathcal{X}| - 1)(|\mathcal{Y}| - 1)|\mathcal{Z}| \qquad \square\end{aligned}$$

Lemma 1.4.5. *We have that*

$$M^2 = M,$$

where M has the form (1.32)*.*

Proof. We compute $(M^2)_{x,y,z}^{x'',y'',z''}$. The first term in the first bracket is multiplied by the consecutive terms in the second bracket, then the second term in the first bracket and so on:

$$\begin{aligned}&\sum_{x',y',z'} M_{x,y,z}^{x',y',z'} M_{x',y',z'}^{x'',y'',z''} = \big(\mathbb{I}(x = x', y = y', z = z') - \mathbb{I}(x = x', z = z')p(y'|z') \\ &- \mathbb{I}(y = y', z = z')p(x'|z') + \mathbb{I}(z = z')p(x'|z')p(y'|z')\big) \cdot \big(\mathbb{I}(x' = x'', y' = y'', z' = z'') \\ &- \mathbb{I}(x' = x'', z' = z'')p(y''|z'') - \mathbb{I}(y' = y'', z' = z'')p(x''|z'') + \mathbb{I}(z' = z'')p(x''|z'')p(y''|z'')\big) \\ &= (\mathbb{I}(x = x'', y = y'', z = z'') - \mathbb{I}(x = x'', z = z'')p(y''|z'') - \mathbb{I}(y = y'', z = z'')p(x''|z'') \\ &+ \mathbb{I}(z = z'')p(x''|z'')p(y''|z'')) - (\mathbb{I}(x = x'', z = z'')p(y''|z'') - \mathbb{I}(x = x'', z = z'')p(y''|z'') \\ &- \mathbb{I}(z = z'')p(x''|z'')p(y''|z'') + \mathbb{I}(z = z'')p(x''|z'')p(y''|z'')) - (\mathbb{I}(y = y'', z = z'')p(x''|z'') \\ &- \mathbb{I}(z = z'')p(x''|z'')p(y''|z'') - \mathbb{I}(y = y'', z = z'')p(x''|z'') + \mathbb{I}(z = z'')p(x''|z'')p(y''|z'')) \\ &+ (\mathbb{I}(z = z'')p(x''|z'')p(y''|z'') - \mathbb{I}(z = z'')p(x''|z'')p(y''|z'') - \mathbb{I}(z = z'')p(x''|z'')p(y''|z'') \\ &+ \mathbb{I}(z = z'')p(x''|z'')p(y''|z'')) = \mathbb{I}(x = x'', y = y'', z = z'') - \mathbb{I}(x = x'', z = z'')p(y''|z'') \\ &- \mathbb{I}(y = y'', z = z'')p(x''|z'') + \mathbb{I}(z = z'')p(x''|z'')p(y''|z'') = M_{x,y,z}^{x'',y'',z''}. \qquad \square\end{aligned}$$

1.4.2. Asymptotic behaviour of plug-in estimators of feature selection criteria

We start with stating the lemma on asymptotic distribution of a plug-in estimator of the generalised feature selection measure defined as

$$\hat{I}_{\beta,\gamma}(X,Y|Z) = I_{\beta,\gamma}(\hat{p}).$$

The measure $I_{\beta,\gamma}(X,Y|Z)$ is defined in (1.17). In order to estimate that measure, for each interaction information appearing in the formula (1.17) we replace unknown probabilities with sample fractions according to the formula (1.7).

Theorem 1.4.6. *Let $\sigma^2 = D_{I_{\beta,\gamma}}(p)'\Sigma D_{I_{\beta,\gamma}}(p)$, where $D_{I_{\beta,\gamma}}(p)$ is a gradient of a function $I_{\beta,\gamma}$ at p and $H_{I_{\beta,\gamma}}(p)$ is its Hessian and Σ is defined in Lemma 1.4.1. Then*

i) if $\sigma^2 > 0$, we have

$$\sqrt{n}(I_{\beta,\gamma}(\hat{p}) - I_{\beta,\gamma}(p)) \xrightarrow{d} \mathcal{N}(0,\sigma^2),$$

ii) if $\sigma^2 = 0$, we have

$$2n(I_{\beta,\gamma}(\hat{p}) - I_{\beta,\gamma}(p)) \xrightarrow{d} \sum_{i=1}^{K} \lambda_i(M) Z_i^2,$$

where $K = |\mathcal{X}| \cdot |\mathcal{Y}| \cdot |\mathcal{Z}|$, Z_i are independent $\mathcal{N}(0,1)$ variables, $\lambda_i(M)$ for $i = 1, 2, \ldots, K$ are eigenvalues of matrix M and $M = H_{I_{\beta,\gamma}}\Sigma$. The explicit formula for M is given in Lemma 1.4.8.

The proof generalises the one given for $J_{\beta,\gamma}$ in [18].

Proof. As in the proof of Lemma 1.4.2, first we compute the gradient and Hessian of $\hat{I}_{\beta,\gamma}(X,Y|Z)$. We consider $I_{\beta,\gamma}(X,Y|Z)$ as a function of $(p(x,y,z))_{x,y,z}$ and we use its representation from Theorem 1.3.1, as we will use the fact that we know the gradient and the Hessian for entropy functional H, i.e.

$$D_H\left((p(v_T))_{v_T}\right)_v := \left(\frac{\partial H\left((p(v_{t_1}, v_{t_2}, \ldots, v_{t_k}))_{v_{t_1}, v_{t_2}, \ldots, v_{t_k}}\right)}{\partial p(v_1, v_2, \ldots, v_p)}\right)_v = \log(p(v_{t_1}, v_{t_2}, \ldots, v_{t_k})) + 1$$

and

$$H_H\left((p(v_T))_{v_T}\right)_v^{v'} := \left(\frac{\partial^2 H\left((p(v_{t_1}, v_{t_2}, \ldots, v_{t_k}))_{v_{t_1}, v_{t_2}, \ldots, v_{t_k}}\right)}{\partial p(v_1, v_2, \ldots, v_p)\partial p(v'_1, v'_2, \ldots, v'_p)}\right)_v^{v'} = \frac{\mathbb{I}(v_{t_1} = v'_{t_1}, v_{t_2} = v'_{t_2}, \ldots, v_{t_k} = v'_{t_k})}{p(v_{t_1}, v_{t_2}, \ldots, v_{t_k})}$$

where $p = (p(v))_v$, $v = (v_1, v_2, \ldots, v_p)$ and $T = \{t_1, t_2, \ldots, t_k\} \subseteq \{1, 2, \ldots, p\}$. We note that although some variables of the vector v do not appear in explicitly in $H\left((p(v_T))_{v_T}\right)$, it is actually a function of $(p(v))_v$ as $p(v_{t_1}, v_{t_2}, \ldots, v_{t_k}) = \sum_{v_i : i \in S \setminus T} p(v_1, v_2, \ldots, v_p)$. Using the equalities above for the generalised measure, in view of Theorem 1.3.1 we obtain

$$\begin{aligned}
D_{I_{\beta,\gamma}}(p)_{x,y,z} &= \sum_{k=1}^{|S|} \sum_{\substack{T \subseteq S \\ |T|=k}} \tilde{\beta}_k^{|S|}(\log(p(z_{t_1}, \ldots, z_{t_k})) + 1) - \sum_{k=0}^{|S|} \sum_{\substack{T \subseteq S \\ |T|=k}} \tilde{\beta}_k^{|S|}(\log(p(z_{t_1}, \ldots, z_{t_k}, x)) + 1) \\
&\quad - \sum_{k=0}^{|S|} \sum_{\substack{T \subseteq S \\ |T|=k}} \tilde{\gamma}_k^{|S|}(\log(p(z_{t_1}, \ldots, z_{t_k}, y)) + 1) + \sum_{k=0}^{|S|} \sum_{\substack{T \subseteq S \\ |T|=k}} \tilde{\gamma}_k^{|S|}(\log(p(z_{t_1}, \ldots, z_{t_k}, x, y)) + 1) \\
&= \sum_{k=1}^{|S|} \sum_{\substack{T \subseteq S \\ |T|=k}} \tilde{\beta}_k^{|S|} \log(p(z_{t_1}, \ldots, z_{t_k})) - \sum_{k=0}^{|S|} \sum_{\substack{T \subseteq S \\ |T|=k}} \tilde{\beta}_k^{|S|} \log(p(z_{t_1}, \ldots, z_{t_k}, x) - \tilde{\beta}_0^{|S|} \\
&\quad - \sum_{k=0}^{|S|} \sum_{\substack{T \subseteq S \\ |T|=k}} \tilde{\gamma}_k^{|S|} \log(p(z_{t_1}, \ldots, z_{t_k}, y) + \sum_{k=0}^{|S|} \sum_{\substack{T \subseteq S \\ |T|=k}} \tilde{\gamma}_k^{|S|} \log(p(z_{t_1}, \ldots, z_{t_k}, x, y))
\end{aligned} \tag{1.33}$$

and

$$\begin{aligned}
H_{I_{\beta,\gamma}}(p)_{x,y,z}^{x',y',z'} &= \sum_{k=1}^{|S|} \sum_{\substack{T \subseteq S \\ |T|=k}} \tilde{\beta}_k^{|S|} \frac{\mathbb{I}(z_{t_1} = z'_{t_1}, \ldots, z_{t_k} = z'_{t_k})}{p(z_{t_1}, \ldots, z_{t_k})} \\
&\quad - \sum_{k=0}^{|S|} \sum_{\substack{T \subseteq S \\ |T|=k}} \tilde{\beta}_k^{|S|} \frac{\mathbb{I}(z_{t_1} = z'_{t_1}, \ldots, z_{t_k} = z'_{t_k}, x = x')}{p(z_{t_1}, \ldots, z_{t_k}, x)} \\
&\quad - \sum_{k=0}^{|S|} \sum_{\substack{T \subseteq S \\ |T|=k}} \tilde{\gamma}_k^{|S|} \frac{\mathbb{I}(z_{t_1} = z'_{t_1}, \ldots, z_{t_k} = z'_{t_k}, y = y')}{p(z_{t_1}, \ldots, z_{t_k}, y)} \\
&\quad + \sum_{k=0}^{|S|} \sum_{\substack{T \subseteq S \\ |T|=k}} \tilde{\gamma}_k^{|S|} \frac{\mathbb{I}(z_{t_1} = z'_{t_1}, \ldots, z_{t_k} = z'_{t_k}, x = x', y = y')}{p(z_{t_1}, \ldots, z_{t_k}, x, y)}.
\end{aligned} \tag{1.34}$$

If $p(x, y, z_1, z_2, \ldots, z_p) > 0$ for all (x, y, z), then each element of $H(\hat{p})$ converges to the corresponding element of $H(p)$ almost surely, hence we have that

$$I_{\beta,\gamma}(\hat{p}) = I_{\beta,\gamma}(p) + (\hat{p} - p)' D_{I_{\beta,\gamma}}(\hat{p}) + \frac{1}{2}(\hat{p} - p)' H_{I_{\beta,\gamma}}(p)(\hat{p} - p) + o_p(\|\hat{p} - p\|^2). \quad (1.35)$$

After rearranging terms and multiplying both sides by $\sqrt{n}$, we obtain

$$\sqrt{n}\,((I_{\beta,\gamma}(\hat{p}) - I_{\beta,\gamma}(p)) = \sqrt{n}(\hat{p}-p)' D_{I_{\beta,\gamma}}(p) + \frac{\sqrt{n}}{2}(\hat{p}-p)' H_{I_{\beta,\gamma}}(p)(\hat{p}-p) + o_p(\sqrt{n}\,\|\hat{p} - p\|^2) \quad (1.36)$$

and if $\sigma^2 > 0$, then

$$\sqrt{n}\,((I_{\beta,\gamma}(\hat{p}) - I_{\beta,\gamma}(p)) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

as both the second and the third terms tend to 0. Moreover, we have that $\sigma^2 = 0$ if and only if $\mathrm{Var}(D_{I_{\beta,\gamma}}(p)'\hat{p}) = 0$ and that means that $(D_{I_{\beta,\gamma}}(p)\hat{p})_{x,y,z} = C$ for all (x, y, z) a.s. As probability of getting $\hat{p}(x, y, z) = 1$ equals $p(x, y, z)^n > 0$ for the sample size n for any (x, y, z), we obtain

$$(D_{I_{\beta,\gamma}}(p))_{x,y,z} = C \text{ for all } (x, y, z). \quad (1.37)$$

Thus the second term in (1.35) equals 0, as $(\hat{p} - p)' D_{I_{\beta,\gamma}}(\hat{p}) = C - C = 0$. Then

$$2n(I_{\beta,\gamma}(\hat{p}) - I_{\beta,\gamma}(p)) = n(\hat{p} - p)' H_{I_{\beta,\gamma}}(p)(\hat{p} - p) + o_p(n\,\|\hat{p} - p\|^2)$$

and that converges to $W' H_{I_{\beta,\gamma}}(p) W$, where $W \sim \mathcal{N}(0, \Sigma)$ as in Lemma 1.4.1. Similarly as in the proof of Lemma 1.4.2, we represent $W' H_{I_{\beta,\gamma}}(p) W$ as

$$W' H_{I_{\beta,\gamma}}(p) W = Z' \Sigma^{1/2} H_{I_{\beta,\gamma}}(p) \Sigma^{1/2} Z = \sum_{x,y,z} \lambda_{x,y,z} Z^2_{x,y,z},$$

where $Z \sim \mathcal{N}(0, I)$ and $\lambda_{x,y,z}$ are eigenvalues of the matrix $\Sigma^{1/2} H_{I_{\beta,\gamma}}(p) \Sigma^{1/2}$. As the matrices $\Sigma^{1/2} H_{I_{\beta,\gamma}}(p) \Sigma^{1/2}$ and $M = H_{I_{\beta,\gamma}}(p)\Sigma$ have the same eigenvalues, we obtain *ii).*

□

Remark 1.4.7. *If σ_a^2 is defined as $\sigma_a^2 = a'\Sigma a$, where a is a vector $a = (a(x,y,z))_{x,y,z}$ and Σ is defined in Lemma 1.4.1, then*

$$\sigma_a^2 = \mathrm{Var}(a(X,Y,Z)).$$

We obtain that in a straightforward manner:

$$\begin{aligned}\sigma_a^2 &= \sum_{x,y,z}\sum_{x',y',z'} a(x,y,z)a(x',y',z')\left(\mathbb{I}(x=x',y=y',z=z')p(x,y,z) - p(x,y,z)p(x',y',z')\right)\\ &= \sum_{x,y,z} p(x,y,z)a^2(x,y,z) - \left(\sum_{x,y,z} p(x,y,z)a(x,y,z)\right)^2 = \mathbb{E}a^2(X,Y,Z) - \left(\mathbb{E}a(X,Y,Z)\right)^2\\ &= \mathrm{Var}(a(X,Y,Z)).\end{aligned}$$

Using that, in view of (1.33)*, we obtain that σ^2 defined in Theorem 1.4.6 equals* $\mathrm{Var}(D_{I_{\beta,\gamma}}(p)_{X,Y,Z}) = \mathrm{Var}(D_{I_{\beta,\gamma}}(p)_{X,Y,Z} + \tilde{\beta}_0^{|S|})$*:*

$$\begin{aligned}\sigma^2 = \mathrm{Var}\Big(&\sum_{k=1}^{|S|}\sum_{\substack{T\subseteq S\\|T|=k}} \tilde{\beta}_k^{|S|}\log(p(Z_{t_1},\ldots,Z_{t_k})) - \sum_{k=0}^{|S|}\sum_{\substack{T\subseteq S\\|T|=k}} \tilde{\beta}_k^{|S|}\log(p(Z_{t_1},\ldots,Z_{t_k},X)\\ &- \sum_{k=0}^{|S|}\sum_{\substack{T\subseteq S\\|T|=k}} \tilde{\gamma}_k^{|S|}\log(p(Z_{t_1},\ldots,Z_{t_k},Y) + \sum_{k=0}^{|S|}\sum_{\substack{T\subseteq S\\|T|=k}} \tilde{\gamma}_k^{|S|}\log(p(Z_{t_1},\ldots,Z_{t_k},X,Y))\Big). \quad (1.38)\end{aligned}$$

Note the the expression for which variance is calculated under variance operator is similar to the representation of $I_{\beta,\gamma}$ given in Theorem 1.3.1 (the difference being that for the entropies in (1.18) *averaging over probabilities is missing, in contrast to* (1.38)*).*

Lemma 1.4.8. *Matrix $M = H_{I_{\beta,\gamma}}(p)\Sigma$ defined in Theorem 1.4.6 equals*

$$\begin{aligned}M_{x,y,z}^{x'',y'',z''} &= p(x'',y'',z'')\cdot\\ &\Bigg[\sum_{k=0}^{|S|}\sum_{\substack{T\subseteq S\\|T|=k}}\left(\tilde{\beta}_k^{|S|}\frac{\mathbb{I}(z_{t_1}=z''_{t_1},\ldots,z_{t_k}=z''_{t_k})}{p(z_{t_1},\ldots,z_{t_k})} - \tilde{\beta}_k^{|S|}\frac{\mathbb{I}(z_{t_1}=z''_{t_1},\ldots,z_{t_k}=z''_{t_k},x=x'')}{p(z_{t_1},\ldots,z_{t_k},x)}\right)\\ &\qquad - \sum_{k=0}^{|S|}\sum_{\substack{T\subseteq S\\|T|=k}}\left(\tilde{\gamma}_k^{|S|}\frac{\mathbb{I}(z_{t_1}=z''_{t_1},\ldots,z_{t_k}=z''_{t_k},y=y'')}{p(z_{t_1},\ldots,z_{t_k},y)}\right.\end{aligned}$$

$$+ \tilde{\gamma}_k^{|S|} \frac{\mathbb{I}(z_{t_1} = z''_{t_1}, \ldots, z_{t_k} = z''_{t_k}, x = x'', y = y'')}{p(z_{t_1}, \ldots, z_{t_k}, x, y)} \Bigg) \Bigg], \quad (1.39)$$

where we define $\mathbb{I}(\emptyset)/p(\emptyset) = 1$.

Proof. First we note that for given k and $T \subset S$ such that $|T| = k$, for any (x, y, z) and (x'', y'', z'') we have for the first term of $H_{I_{\beta,\gamma}}(p)$ in 1.34

$$\sum_{x',y',z'} \tilde{\beta}_k^{|S|} \frac{\mathbb{I}(z_{t_1} = z'_{t_1}, \ldots, z_{t_k} = z'_{t_k})}{p(z_{t_1}, \ldots, z_{t_k})} p(x', y', z') p(x'', y'', z'') = p(x'', y'', z'') \tilde{\beta}_k^{|S|}$$

and

$$\sum_{x',y',z'} \tilde{\beta}_k^{|S|} \frac{\mathbb{I}(z_{t_1} = z'_{t_1}, \ldots, z_{t_k} = z'_{t_k})}{p(z_{t_1}, \ldots, z_{t_k})} \mathbb{I}(x' = x'', y' = y'', z' = z'') p(x', y', z')$$
$$= p(x'', y'', z'') \tilde{\beta}_k^{|S|} \frac{\mathbb{I}(z_{t_1} = z''_{t_1}, \ldots, z_{t_k} = z''_{t_k})}{p(z_{t_1}, \ldots, z_{t_k})}.$$

For the three remaining terms of $H_{I_{\beta,\gamma}}(p)$ the formulas are analogous. We notice that

$$\sum_{x',y',z'} (H_{I_{\beta,\gamma}}(p))_{x,y,z}^{x',y',z'} p(x', y', z') p(x'', y'', z'') = \tilde{\beta}_0^{|S|},$$

thus defining $\mathbb{I}(\emptyset)/p(\emptyset) = 1$ we obtain the formula (1.39). □

Remark 1.4.9. *We note that Theorem 1.4.6 in particular describes the behaviour of the measures introduced in Section 1.3.2 i.e.* $J_{\beta,\gamma}$, JMI, $JMI3$, $CIFE$ *and* $SECMI3$, *and others. The asymptotic behaviour of the empirical generalised measure and* $\widehat{JMI}$ *was analyzed in [18], and* $\widehat{CIFE}$ *and* $\widehat{SECMI3}$ *in [20].*

Detailed characterization of distribution of $J_{\beta,\gamma}(\hat{p})$

In this section we present in Lemma 1.4.10 the results from [18] with a slight correction of Theorem 2 in [18].

In the proof of Theorem 1.4.6 we showed that the condition $\sigma^2 = D_{J_{\beta,\gamma}}(p)' \Sigma D_{J_{\beta,\gamma}}(p) = 0$ is equivalent to

$$(D_{I_{\beta,\gamma}}(p))_{x,y,z} = C \text{ for all } (x, y, z). \quad (1.40)$$

In this section we will further analyze that case for a *binary* random variable Y and for certain values of β and γ we will determine the dependence structure of (X, Y, Z).

We note that the gradient $D_{J_{\beta,\gamma}}$ of $J_{\beta,\gamma}$ equals

$$\log\left(\frac{p(x,y)}{p(x)p(y)}\right) - \beta\sum_{i=1}^{|S|}\log\left(\frac{p(x,z_i)}{p(x)p(z_i)} - 1\right) + \gamma\sum_{i=1}^{|S|}\log\left(\frac{p(x,y,z_i)p(y)}{p(x,y)p(y,z_i)}\right) - 1.$$

This can be obtained from the formula for the gradient computed for $D_{I_{\beta,\gamma}}$ in the proof of Theorem 1.4.6 using the fact that the parameters $\tilde{\beta}_k^{|S|}$ and $\tilde{\gamma}_k^{|S|}$ for $J_{\beta,\gamma}$ equal 0 for $k \geq 2$ and the formula for $k = 0, 1$ is given in (1.20).

To characterize the case when $\sigma^2 = 0$ we define two scenarios:

— Scenario 1 (S1): $X \perp\!\!\!\perp Y|Z_i$ for any $i \in S$ and $X \perp\!\!\!\perp Y$,

— Scenario 2 (S2): $\exists W \subset S$ such that $W \neq \emptyset$ and for $i \in W$ $Z_i \perp\!\!\!\perp Y|X$, $X \not\perp\!\!\!\perp Y|Z_i$ and for $i \in W^c$ we have $X \perp\!\!\!\perp Y|Z_i$,

where we define W as

$$W = \left\{i \in S : \exists_{x,y,z_i} \frac{p(x,y,z_i)p(z_i)}{p(x,z_i)p(y,z_i)} \neq 1\right\}. \tag{1.41}$$

Lemma 1.4.10. *Assume that $\beta = \gamma \neq 0$ and $\sigma^2 = D_{J_{\beta,\gamma}}(p)'\Sigma D_{J_{\beta,\gamma}}(p) = 0$. Then*

(i) if $|S| > 1$ and $\beta^{-1} \in \{1, 2, \ldots, |S| - 1\}$ then one of the above scenarios holds with W defined in (1.41)*,*

(ii) if $\beta^{-1} = |S|$ then $X \perp\!\!\!\perp Y|Z_i$ for all $i \in S$. If $\beta^{-1} \notin \{1, 2, \ldots, |S|\}$ then Scenario 1 is valid.

Proof. As $\sigma^2 = 0$, we have that $(D_{I_{\beta,\gamma}}(p))_{x,y,z} = C$ for all (x, y, z):

$$\log\left(\frac{p(x,y)}{p(x)p(y)}\right) - \beta\sum_{i=1}^{|S|}\left(\log\left(\frac{p(x,z_i)}{p(x)p(z_i)}\right) - 1\right) + \gamma\sum_{i=1}^{|S|}\log\left(\frac{p(x,y,z_i)p(y)}{p(x,y)p(y,z_i)}\right) - 1 = C. \tag{1.42}$$

Fix $j \in S$, then

$$\begin{aligned}\beta\left(\log\left(\frac{p(x,z_j)}{p(x)p(z_j)}\right) - 1\right) - \gamma\log\left(\frac{p(x,y,z_j)}{p(x,y)p(y,z_j)}\right) &= -C - 1 \\ + \log\left(\frac{p(x,y)}{p(x)p(y)}\right) - \beta\sum_{i\neq j}\left(\log\left(\frac{p(x,z_i)}{p(x)p(z_i)}\right) - 1\right) &+ \gamma\sum_{i\neq j}\log\left(\frac{p(x,y,z_i)p(y)}{p(x,y)p(y,z_i)}\right).\end{aligned}$$

The left hand side does not depend on z_j, as the right hand side does not, thus we have for all (x, y, z_j)

$$\beta\left(\log\left(\frac{p(x,z_j)}{p(x)p(z_j)}\right)-1\right)-\gamma\log\left(\frac{p(x,y,z_j)p(y)}{p(x,y)p(y,z_j)}\right)=a_{xy}.$$

Hence

$$\beta\log\left(\frac{p(x,z_j)}{p(z_j)}\right)-\gamma\log\left(\frac{p(x,y,z_j)}{p(y,z_j)}\right)=b_{xy},$$

where $b_{xy}=a_{xy}+\beta+\beta\log(p(x))+\gamma\log\left(p(y)/p(x,y)\right)$. For $\beta=\gamma\neq 0$ we obtain

$$\log\left(\frac{p(x,z_j)p(y,z_j)}{p(x,y,z_j)p(z_j)}\right)=\frac{b_{xy}}{\beta}=c_{xy}. \tag{1.43}$$

If $j\in W$ then in view of Lemma A.1.1

$$c_{xy}=\log\left(\frac{p(x)p(y)}{p(x,y)}\right), \tag{1.44}$$

otherwise, if $j\in W^c$, then $c_{xy}=0$. Thus (1.42) can be written as

$$\begin{aligned}
&\log\left(\frac{p(x,y)}{p(x)p(y)}\right)-\beta\sum_{i=1}^{|S|}\log\left(\frac{p(x,z_i)}{p(x)p(z_i)}-1\right)+\beta\sum_{i=1}^{|S|}\log\left(\frac{p(x,y,z_i)p(y)}{p(x,y)p(y,z_i)}\right)-1\\
&=\log\left(\frac{p(x,y)}{p(x)p(y)}\right)+\beta\sum_{i=1}^{|S|}\log\left(\frac{p(x,y,z_i)p(z_i)}{p(x,z_i)p(y,z_i)}\right)+\beta\sum_{i=1}^{|S|}\log\left(\frac{p(x)p(y)}{p(x,y)}\right)+\beta|S|-1\\
&=(1+\beta|W|-\beta|S|)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)+\beta|S|-1\\
&=(1-\beta|W^c|)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)+\beta|S|-1.
\end{aligned}$$

We observe that as the gradient does not depend on (x, y), at least one of the two cases should occur: $X \perp\!\!\!\perp Y$ or $1+\beta|W|-\beta|S|=0$. Now we proceed to prove (i) and (ii).

(ii) If $\beta^{-1}=|S|$ then either $W=\emptyset$ as $|W|=|S|-\beta^{-1}=0$ or $X \perp\!\!\!\perp Y$. In both cases we have that $c_{xy}=0$ and thus we obtain also that $X \perp\!\!\!\perp Y|Z_i$ for all $i\in S$ (compare (1.43)). Note that although we can prove that if $X \perp\!\!\!\perp Y$ then $X \perp\!\!\!\perp Y|Z_i$ as $c_{xy}=0$ for all x and y, the opposite does not hold. In the case when $|W|=0$ we do not obtain that $X \perp\!\!\!\perp Y$ as we cannot infer that $p(x,y)/p(x)p(y)=1$. This corrects the mistake in [18] p. 696. If $\beta^{-1}\notin\{1,\ldots,|S|-1\}$, then $1+\beta|W|-\beta|S|\neq 0$ as

in that case $|W| \notin \mathbb{N}$, $|W| > |S|$ or $|W| \leq 0$, a contradiction. Hence $X \perp\!\!\!\perp Y$ and it follows that $c_{xy} = 0$ and $X \perp\!\!\!\perp Y|Z_i$ thus Scenario 1 is valid.

(i) We consider now the case when $1 + \beta|W| - \beta|S| = 0$. Thus $|W| = |S| - \beta^{-1}$ and as $|W| \in \{0, 1, \ldots, |S|\}$, we obtain $\beta^{-1} \in \{1, \ldots, |S|\}$. We already considered the case of $\beta^{-1} = |S|$. In other cases and as $|W| > 0$ from Lemma A.1.1 we obtain (i). □

Special case: $\widehat{JMI}$

In the following we explicitly state the asymptotic distribution of $JMI(\hat{p})$. $JMI(\hat{p})$ is a plug-in estimator of JMI defined in (1.21), namely

$$\widehat{JMI} = JMI(\hat{p}) = \frac{1}{|S|}\sum_{i=1}^{|S|}\hat{I}(X, Y|Z_i) = \frac{1}{|S|}\sum_{i=1}^{|S|}\sum_{x,y,z_i}\hat{p}(x, y, z_i)\log\frac{\hat{p}(x, y|z_i)}{\hat{p}(x|z_i)\hat{p}(y|z_i)}.$$

In Theorem 1.4.11 we distinguish two cases - the first one, in which for all $i \in S$ we have $X \perp\!\!\!\perp Y|Z_i$ and the second, which is the complement of the first case. The two cases can be also described in a different way. First, we define $H_{0,i}$ in the following way

$$H_{0,i} : X \perp\!\!\!\perp Y|Z_i, \tag{1.45}$$

for $i = 1, \ldots, |S|$. We want to construct a test which controls type I error under so called global null $H_0 = \cap_{i=1}^{|S|} H_{0,i}$ when all null hypotheses are true. Thus the first case occurs, when the global null H_0 is true, and the second, when the alternative hypothesis holds. The following result has been proved in [24].

Theorem 1.4.11. *(i) Assume that the global null H_0 holds. Then*

$$2nJMI(\hat{p}) \xrightarrow{d} \sum_{i=1}^{K}\lambda_i(M)Z_i^2, \tag{1.46}$$

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

$$M_{x,y,z}^{x',y',z'} = \frac{1}{|S|}p(x', y', z')\sum_{i=1}^{|S|}\left[\frac{\mathbb{I}(z_i = z_i')}{p(z_i)} - \frac{\mathbb{I}(x = x', z_i = z_i')}{p(x, z_i)}\right.$$
$$\left. - \frac{\mathbb{I}(y = y', z_i = z_i')}{p(y, z_i)} + \frac{\mathbb{I}(x = x', y = y', z_i = z_i')}{p(x, y, z_i)}\right], \tag{1.47}$$

where $z = (z_1, \ldots, z_{|S|})$ *and* $z' = (z'_1, \ldots, z'_{|S|})$ *and* $K = |\mathcal{X}| \cdot |\mathcal{Y}| \cdot |\mathcal{Z}|$. *Moreover, the trace of* M *equals* $|S|^{-1}(|\mathcal{X}| - 1)(|\mathcal{Y}| - 1) \sum_i |\mathcal{Z}_i|$.

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

$$\sigma^2 = \mathrm{Var}\left(\frac{1}{|S|} \log \prod_{i=1}^{|S|} \frac{p(X, Y, Z_i)p(Z_i)}{p(X, Z_i)p(Y, Z_i)} \right) > 0$$

and

$$n^{1/2}(JMI(\hat{p}) - JMI(p)) \xrightarrow{d} \mathcal{N}(0, \sigma^2). \tag{1.48}$$

Proof. From Theorem 1.4.6 follows $i)$ and the formula for M can be derived from (1.39) ($\tilde{\beta}_1^{|S|} = \tilde{\gamma}_1^{|S|} = 1/|S|$ and $\tilde{\beta}_i^{|S|} = \tilde{\gamma}_i^{|S|} = 0$ for $i \neq 1$). $ii)$ follows from Lemma 1.4.10. □

Note that for $|S| = 1$ the form of matrix M in (1.47) coincides with that in (1.32) as in that case $JMI(\hat{p}) = CMI(\hat{p})$.

The result in Theorem 1.4.11 shows that there is an exact dichotomy of asymptotic behaviour which makes the construction of the test for testing the global null H_0 possible: the asymptotic distribution of $\widehat{JMI}$ is either that of quadratic form in normal variables as in (1.46) or normal (cf. (1.48)) depending on whether H_0 is satisfied or not.

Chapter 2

Resampling schemes and asymptotic distributions of information-theoretic measures

The main goal of this section is to introduce various resampling scenarios mimicking the situation, in which X is conditionally independent of Y given Z. The aim is to obtain asymptotic distributions of the vectors of estimated probabilities based on resampled samples and to show asymptotic behaviour of plug-in estimators of CMI and JMI, in which instead of using the usual vector of fractions $\hat{p}$, the estimators $\hat{p}^*$ based on resampled samples are used.

The topic of bootstrap and related resampling methods is much broader than the problems discussed in this thesis. For more information we refer to [10].

2.1. Resampling schemes

In this section we introduce resampling scenarios and we obtain asymptotic distributions of vectors of estimated probabilities $(\hat{p}^*(x,y,z))_{x,y,z}$ based on resampled sample conditionally given the sample.

2.1.1. CI bootstrap scenario

Let $(X_1,Y_1,Z_1),(X_2,Y_2,Z_2),\ldots,(X_n,Y_n,Z_n)$ be independent random variables with common discrete distribution $P(X_i=x,Y_i=y,Z_i=z)=p(x,y,z)>0$, where the last condition holds for all (x,y,z). The unconstrained maximum likelihood estimator of $(p(x,y,z))_{x,y,z}$ is a vector of fractions $(\hat{p}(x,y,z))_{x,y,z}=(n(x,y,z)/n)_{x,y,z}$, where $n(x,y,z)=\sum_{i=1}^{n}\mathbb{I}(X_i=x,Y_i=y,Z_i=z)$. This is an equivalent to estimating a cumulative distribution function F by its empirical counterpart $\hat{F}_n$, which assigns mass $1/n$ to each of the sampled points. The most common approach in bootstrap resampling is to sample from empirical distribution $\hat{F}_n$ and thus each point (x,y,z) is chosen with

a probability equal to $\hat{p}(x,y,z)$. In order to investigate conditional independence problems we replace $\hat{p}(x,y,z)$ by

$$\hat{p}_{ci}(x,y,z) = \hat{p}(x|z)\hat{p}(y|z)\hat{p}(z) = \frac{n(x,z)}{n(z)}\frac{n(y,z)}{n(z)}\frac{n(z)}{n},$$

which is the maximum likelihood estimator of $p(x,y,z)$ in a model with assumed conditional independence of X and Y given Z. Note that we used the notation $p_{ci}(x,y,z) = p(x|z)p(y|z)p(z)$ previously. We define a CI bootstrap sample in the following way: let $(X_1^*, Y_1^*, Z_1^*), (X_2^*, Y_2^*, Z_2^*), \ldots, (X_n^*, Y_n^*, Z_n^*)$ be random variables conditionally independent given original sample, with common distribution $(\hat{p}_{ci}(x,y,z))_{x,y,z}$. By $\hat{p}^*$ we denote an estimator of probabilities based on the bootstrap sample, namely

$$\hat{p}^*(x,y,z) = \frac{n^*(x,y,z)}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i^* = x, Y_i^* = y, Z_i^* = z).$$

We use a superscript $*$ to indicate that the probabilities or the expected values are calculated given the sample $(X_1,Y_1,Z_1),(X_2,Y_2,Z_2),\ldots,(X_n,Y_n,Z_n)$, e.g.

$$P^*(X_1^* = x) = P(X_1^* = x|(X_1,Y_1,Z_1),(X_2,Y_2,Z_2),\ldots,(X_n,Y_n,Z_n)). \tag{2.1}$$

To avoid burdensome notation, while conditioning on the sequence $(X_1,Y_1,Z_1),(X_2,Y_2,Z_2),\ldots,(X_n,Y_n,Z_n)$, we will write

$$P(\cdot|(X_1,Y_1,Z_1),(X_2,Y_2,Z_2),\ldots,(X_n,Y_n,Z_n))$$

meaning that

$$P(\cdot|(X_1,Y_1,Z_1) = (x_1,y_1,z_1),(X_2,Y_2,Z_2) = (x_2,y_2,z_2),\ldots,(X_n,Y_n,Z_n) = (x_n,y_n,z_n)),$$

where (x_i,y_i,z_i) denotes generic but fixed value of (X_i,Y_i,Z_i).

In Lemma 2.1.1 we prove that the vector $\hat{p}^*$ has asymptotically normal distribution almost surely given $(X_i,Y_i,Z_i)_{i=1}^{\infty}$ and we state an explicit formula for its covariance matrix Σ. Note that the limiting law in Lemma 2.1.1 below coincides with the law of sample fractions for $p(x,y,z) = p_{ci}(x,y,z)$ as $\Sigma_{x,y,z}^{x',y',z'} = \mathbb{I}(x=x',y=y',z=z')p_{ci}(x,y,z) - p_{ci}(x,y,z)p_{ci}(x',y',z')$. In particular, Σ is

equal to asymptotic covariance matrix of a vector $\sqrt{n}(\hat{p}(x,y,z) - p(x,y,z))_{x,y,z}$ in case when $X \perp\!\!\!\perp Y|Z$ (cf. Lemma 1.4.1).

We recall that $\Sigma_{x,y,z}^{x',y',z'}$ denotes an element of the matrix Σ with row index (x,y,z) and column index (x',y',z'), where $(x,y,z),(x',y',z') \in \mathcal{X}\times\mathcal{Y}\times\mathcal{Z}$. Σ will be generic notation of a covariance matrix, which will change in the results to follow.

Lemma 2.1.1. *For almost all sequences* $(X_1,Y_1,Z_1),(X_2,Y_2,Z_2),\ldots$ *and conditionally on* $(X_i,Y_i,Z_i)_{i=1}^{\infty}$*, we have that*

$$\sqrt{n}\left(\hat{p}^*(x,y,z) - \hat{p}(x|z)\hat{p}(y|z)\hat{p}(z)\right)_{x,y,z} \xrightarrow{d} \mathcal{N}(0,\Sigma),$$

where

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

and $\hat{p}^*(x,y,z) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(X_i^*=x,Y_i^*=y,Z_i^*=z)$*, where* $(X_i^*,Y_i^*,Z_i^*)_{i=1}^n$ *is a CI bootstrap sample.*

The technique of the proof presented here follows from [39] and it is based on the Berry-Esseen theorem (cf. Theorem A.2.1).

Without loss of generality, to simplify the notation we assume that $\mathcal{X}=\{1,2,\ldots,|\mathcal{X}|\}$, $\mathcal{Y}=\{1,2,\ldots,|\mathcal{Y}|\}$ and $\mathcal{Z}=\{1,2,\ldots,|\mathcal{Z}|\}$. We define a function $k(\cdot)$, which assigns a triple $(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}$ to each index $i=1,2,\ldots,K$, where $K=|\mathcal{X}|\cdot|\mathcal{Y}|\cdot|\mathcal{Z}|$ in the following way

$$k(i) = (x,y,z),$$

and

$$i = x + |\mathcal{X}|(y-1) + |\mathcal{X}||\mathcal{Y}|(z-1).$$

Thus, in the notation using the function k, we write e.g. a vector of all probabilities $(p(x,y,z))_{x,y,z}$ as $(p(k(i)))_{i=1}^K$.

Proof. We recall that (see (2.1))

$$\hat{p}^*(x,y,z) = \frac{n^*(x,y,z)}{n} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(X_i^*=x,Y_i^*=y,Z_i^*=z)$$

and

$$\hat{p}_{ci}(x,y,z) := \hat{p}(x|z)\hat{p}(y|z)\hat{p}(z) = \frac{n(x,z)}{n(z)}\frac{n(y,z)}{n(z)}\frac{n(z)}{n}.$$

Thus, since $\hat{p}^*$ follows the multinomial distribution with an observation (x,y,z) having a probability equal to $\hat{p}_{ci}(x,y,z)$, conditionally on the original sample we have that

$$\mathbb{E}^*\hat{p}^*(x,y,z) = \hat{p}(x|z)\hat{p}(y|z)\hat{p}(z)$$

and

$$\left(\mathrm{Cov}^*\left(\left(\hat{p}^*(x,y,z)\right)_{x,y,z}\right)\right)_{x,y,z}^{x',y',z'} = \begin{cases} \frac{1}{n}\hat{p}_{ci}(x,y,z)(1-\hat{p}_{ci}(x,y,z)) & \text{if } (x,y,z) = (x',y',z') \\ -\frac{1}{n}\hat{p}_{ci}(x,y,z)\hat{p}_{ci}(x',y',z') & \text{if } (x,y,z) \neq (x',y',z') \end{cases}.$$

We define

$$\hat{\Sigma}_{x,y,z}^{x',y',z'} = n\left(\mathrm{Cov}^*\left(\left(\hat{p}^*(x,y,z)\right)_{x,y,z}\right)\right)_{x,y,z}^{x',y',z'}.$$

Then we define Q_i and W in Theorem A.2.1 (Appendix) in the following way

$$Q_j^* := \frac{1}{\sqrt{n}}\hat{\Sigma}_{-K}^{-1/2}\left(\mathbb{I}((X_j^*,Y_j^*,Z_j^*) = k(i)) - \hat{p}_{ci}(k(i))\right)_{i=1}^{K-1},$$

$$W^* = \sum_{j=1}^{n} Q_j^* = \sqrt{n}\hat{\Sigma}_{-K}^{-1/2}\left(\hat{p}^*(k(i)) - \hat{p}_{ci}(k(i))\right)_{i=1}^{K-1},$$

where $\hat{\Sigma}_{-K} = \mathrm{Cov}^*\left((\hat{p}^*(k(i)))_{i=1}^{K-1}\right)$. If $p(x,y,z) > 0$ for all (x,y,z), the matrix $\hat{\Sigma}_{-K}$ is invertible, cf. e.g. [35]. One element of the vector $\hat{p}^*$ is omitted to ensure that the covariance matrix is a full rank matrix. As we have $\sum_{x,y,z}\hat{p}^*(x,y,z) = 1$, the full size matrix $\hat{\Sigma}$ is singular. Then we apply Theorem A.2.1

$$\begin{aligned} &|P^*(W^* \in A) - P(Z \in A)| \\ &\qquad \leq K_d \sum_{j=1}^{n} \mathbb{E}^* \left\| \frac{1}{\sqrt{n}}\hat{\Sigma}_{-K}^{-1/2}\left(\mathbb{I}((X_j^*,Y_j^*,Z_j^*) = k(i)) - \hat{p}_{ci}(k(i))\right)_{i=1}^{K-1} \right\|^3 \end{aligned}$$

and $d = K-1$. We notice that as (conditionally on $(X_i,Y_i,Z_i)_{i=1}^{\infty}$)

$$\hat{p}_{ci} \to p_{ci} \text{ and } \hat{\Sigma}_{-K} \to \Sigma_{-K} \quad a.s.,$$

where Σ_{-K} denotes the matrix Σ without the last row and the last column, and for all $j = 1, 2, \ldots, K-1$

$$-1 \leq \mathbb{I}(X_j^* = x, Y_j^* = y, Z_j^* = z) - \hat{p}_{ci}(x, y, z) \leq 1,$$

we have that $\mathbb{E}^* \left\| \hat{\Sigma}_{-K}^{-1/2} \left(\mathbb{I}((X_j^*, Y_j^*, Z_j^*) = k(i)) - \hat{p}_{ci}(k(i)) \right)_{i=1}^{K-1} \right\|^3$ is bounded for almost all sequences. Thus in view of Theorem A.2.1, conditionally, $W^* \to \mathcal{N}(0, I)$ and as $\hat{\Sigma}_{-K}^{-1/2}$ converges to $\Sigma_{-K}^{-1/2}$ a.s., from Slutsky's theorem we have that

$$\sqrt{n} \left(\hat{p}^*(k(i)) - \hat{p}_{ci}(k(i)) \right)_{i=1}^{K-1} \xrightarrow{d} \mathcal{N}(0, \Sigma_{-K}).$$

Now, by the continuous mapping theorem, using a function

$$f\left(v(k(i))_{i=1}^{K} \right) = \left(v(k(i))_{i=1}^{K-1}, -\sum_{i=1}^{K-1} v(k(i)) \right),$$

where $v(k(i)) = \hat{p}^*(x, y, z) - \hat{p}(x|z)\hat{p}(y|z)\hat{p}(z)$ we obtain the conclusion, as $0 = \sum_{i=1}^{K} (\hat{p}^*(k(i)) - \hat{p}_{ci}(k(i))$, and thus $\hat{p}^*(k(K)) - \hat{p}_{ci}(k(K)) = -\sum_{i=1}^{K-1} (\hat{p}^*(k(i)) - \hat{p}_{ci}(k(i))$. □

2.1.2. Conditional randomisation scenario (CR)

Now, we consider another resampling scenario, which was used for example in [8] and is called CRT (*conditional randomisation test*) there. We focus here on sampling and as we omit testing for the time being, we refer to that method as conditional randomisation using an abbreviation CR. Importantly, we assume that conditional mass function $P(X = x|Z = z)$ is known.

Let $(X_1, Y_1, Z_1), (X_2, Y_2, Z_2), \ldots, (X_n, Y_n, Z_n)$ be independent random variables and $(X_1^*, Y_1, Z_1), (X_2^*, Y_2, Z_2), \ldots, (X_n^*, Y_n, Z_n)$ be the CR sample with $X_i^* \sim p(x|z_i)$. As previously, p^* denotes estimator of probabilities based on a new sample, namely

$$\hat{p}^*(x, y, z) = \frac{n^*(x, y, z)}{n} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i^* = x, Y_i = y, Z_i = z).$$

Thus we use the pairs (Y_i, Z_i) from the original sample and we independently draw new observations X_i^* only. Moreover, we assume that X_i^* is independent of (X_i, Y_i) given Z_i,

hence

$$P(X_i^* = x' | X_i = x, Y_i = y, Z_i = z) = P(X_i^* = x' | Z_i = z) = p(x'|z)$$

and X_i^* are conditionally independent given the original sample. We have

Lemma 2.1.2. *For almost all sequences* $(X_1, Y_1, Z_1), (X_2, Y_2, Z_2), \ldots$ *and conditionally on* $(X_i, Y_i, Z_i)_{i=1}^{\infty}$*, we have that*

$$\sqrt{n}\left(\hat{p}^*(x,y,z) - p(x|z)\hat{p}(y|z)\hat{p}(z)\right)_{x,y,z} \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

where

$$\Sigma_{x,y,z}^{x',y',z'} = \mathbb{I}(y = y', z = z')\left(\mathbb{I}(x = x')p(x|z)p(y|z)p(z) - p(x|z)p(x'|z')p(y|z)p(z)\right)$$

and $\hat{p}^*(x,y,z) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}(X_i^* = x, Y_i = y, Z_i = z)$*, where* $(X_i^*, Y_i, Z_i)_{i=1}^n$ *is a sample obtained using conditional randomisation scenario.*

Proof. First, we calculate the expected value and covariance matrix of the vector of estimated probabilities $(\hat{p}^*(x,y,z))_{x,y,z}$:

$$\begin{aligned}
\mathbb{E}^*\hat{p}^*(x,y,z) &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}^*\mathbb{I}(X_i^* = x, Y_i = y, Z_i = z) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}(Y_i = y, Z_i = z)\mathbb{E}^*\mathbb{I}(X_i^* = x) \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{I}(Y_i = y, Z_i = z)P(X_i^* = x | (X_1, Y_1, Z_1)\ldots, (X_n, Y_n, Z_n)) \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{I}(Y_i = y, Z_i = z)P(X_i^* = x | (X_i, Y_i, Z_i)) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}(Y_i = y, Z_i = z)P(X_i^* = x | Z_i) \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{I}(Y_i = y, Z_i = z)p(x|Z_i) = p(x|z)\hat{p}(y,z)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}^*\hat{p}^*(x,y,z)\hat{p}^*(x',y',z') &= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathbb{E}^*\mathbb{I}(X_i^* = x, Y_i = y, Z_i = z)\mathbb{I}(X_j^* = x', Y_j = y', Z_j = z') \\
&= \frac{1}{n^2}\sum_{i\neq j} p(x|z)p(x'|z')\mathbb{I}(Y_i = y, Z_i = z)\mathbb{I}(Y_j = y', Z_j = z') \\
&\quad + \mathbb{I}(x = x', y = y', z = z')\frac{1}{n^2}\sum_i \mathbb{E}^*\mathbb{I}(X_i^* = x, Y_i = y, Z_i = z)
\end{aligned}$$

$$= p(x|z)p(x'|z')\Big(\frac{1}{n^2}\sum_{i,j}\mathbb{I}(Y_i = y, Z_i = z)\mathbb{I}(Y_j = y', Z_j = z')$$
$$- \frac{1}{n^2}\sum_{i}\mathbb{I}(Y_i = y, Z_i = z)\mathbb{I}(Y_i = y', Z_i = z')\Big) + \frac{1}{n}\mathbb{I}(x = x', y = y', z = z')p(x|z)\hat{p}(y, z)$$
$$= p(x|z)p(x'|z')\hat{p}(y, z)\hat{p}(y', z') - \frac{1}{n}\mathbb{I}(y = y', z = z')p(x|z)p(x'|z')\hat{p}(y, z)$$
$$+ \frac{1}{n}\mathbb{I}(x = x', y = y', z = z')p(x|z)\hat{p}(y, z),$$

hence

$$\left(\mathrm{Cov}^*\left((\hat{p}^*(x, y, z))_{x,y,z}\right)\right)_{x,y,z}^{x',y',z'} = -\frac{1}{n}\mathbb{I}(y = y', z = z')p(x|z)p(x'|z')\hat{p}(y, z)$$
$$+ \frac{1}{n}\mathbb{I}(x = x', y = y', z = z')p(x|z)\hat{p}(y, z).$$

As in the proof of Lemma 2.1.1, we define Q_j^* and W^* and thus by Berry-Essen theorem (Theorem A.2.1) we obtain

$$|P^*(W^* \in A) - P(Z \in A)|$$
$$\leq K_d \sum_{j=1}^{n} \mathbb{E}^* \left\| \frac{1}{\sqrt{n}}\hat{\Sigma}_{-K}^{-1/2}\left(\mathbb{I}((X_j^*, Y_j, Z_j) = k(i)) - \hat{p}_{tci}(k(i))\right)_{i=1}^{K-1} \right\|^3,$$

where we use the notation $k(i)$ introduced before the proof of the Lemma 2.1.1, $\hat{\Sigma}_{-K} = \mathrm{Cov}^*\left((\hat{p}^*(k(i)))_{i=1}^{K-1}\right)$ and $\hat{p}_{tci}(x, y, z) = p(x|z)\hat{p}(y, z)$ (t in tci stands for 'true', not estimated distribution $p(x|z)$ in the formula, ci for *conditional independence*). We have that conditionally on the sample

$$\hat{p}_{tci} \to p_{ci} \text{ and } \hat{\Sigma}_{-K} \to \Sigma_{-K} \quad a.s.,$$

thus similarly to the proof of Lemma 2.1.1 we obtain the lemma. □

2.1.3. Bootstrap X scenario

We use the same scenario as in previous section, but now we avoid the assumption that the distribution $p(x|z)$ is known. Thus in the bootstrap X sample $(X_1^*, Y_1, Z_1), (X_2^*, Y_2, Z_2), \ldots, (X_n^*, Y_n, Z_n)$ we use the pairs (Y_i, Z_i) from the original sample and X_i^* drawn from the distribution $\hat{p}(x|z_i)$. We state the lemma about the asymptotic conditional distribution of the vector $(\hat{p}^*(x, y, z))_{x,y,z}$.

Lemma 2.1.3. *For almost all sequences* $(X_1, Y_1, Z_1), (X_2, Y_2, Z_2), \ldots$ *and conditionally on* $(X_i, Y_i, Z_i)_{i=1}^{\infty}$*, we have that*

$$\sqrt{n}\left(\hat{p}^*(x,y,z) - \hat{p}(x|z)\hat{p}(y|z)\hat{p}(z)\right)_{x,y,z} \xrightarrow{d} \mathcal{N}(0,\Sigma)$$

where Σ *is defined in Lemma 2.1.2 and* $\hat{p}^*(x,y,z) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i^* = x, Y_i = y, Z_i = z)$*, where* $(X_i^*, Y_i, Z_i)_{i=1}^n$ *is a bootstrap X sample.*

Proof. The proof is analogous to the proof of Lemma 2.1.2. The expected value and the covariance matrix of $(\hat{p}^*(x,y,z))_{x,y,z}$ is computed in the same way as for the CR scheme, only $p(x|z)$ should be replaced with $\hat{p}(x|z)$, which is deterministic given $(X_i, Y_i, Z_i)_{i=1}^n$. □

2.1.4. Conditional permutation scenario

We also consider a resampling scenario based on permutations on strata corresponding to the conditioning variable $Z = z$. Such a scheme was used e.g. in [41] to perform a test of conditional independence. We obtain permuted sample $(X_{\pi(1)}, Y_1, Z_1), (X_{\pi(2)}, Y_2, Z_2), \ldots, (X_{\pi(n)}, Y_n, Z_n)$ in the following way: for each value z of Z we permute all observations X_i for which $Z_i = z$ and we keep the pairs (Y_i, Z_i) unchanged. Thus the number of permissible permutations equals $\prod_{z' \in \mathcal{Z}} n(z')!$ as in each group $\{i : Z_i = z\}$ the number of permutation equals $n(z)!$ and we apply permutations for all distinct values of Z. By Π we denote a set of all permissible conditional permutations given the sample and by π we denote an element of the set Π.

For this method of resampling we cannot apply Berry-Esseen theorem as we did in Lemmas 2.1.1 - 2.1.3, because the observations $\mathbb{I}(X_{\pi(i)} = x, Y_i = y, Z_i = z)$ and $\mathbb{I}(X_{\pi(j)} = x, Y_j = y, Z_j = z)$ for $i \neq j$ are no longer conditionally independent and thus W^* is not a sum of conditionally independent random variables Q_i^* (cf. the proof of Lemma 2.1.1).

However, we calculate the expectation and the covariance matrix of the vector of probabilities in order to obtain parameters of the asymptotic distribution.

First, we derive the expected value of $\hat{p}^*(x,y,z)$, where

$$\hat{p}^*(x,y,z) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_{\pi(i)} = x, Y_i = y, Z_i = z)$$

and $\pi \in \Pi$ is a random permissible permutation. We use $\hat{p}^*$ to denote an estimator of probabilities based on sample obtained by permutations, and thus we will also use the notation $(X_i^*, Y_i, Z_i)_{i=1}^n := (X_{\pi(i)}, Y_i, Z_i)_{i=1}^n$. Recall that $\mathbb{E}^*$ and Cov^* are conditional expectation and conditional covariance respectively and the sum $\sum_\pi$ over π denotes the sum over all permissible permutations $\sum_{\pi\in\Pi}$.

$$\begin{aligned}
\mathbb{E}^*\hat{p}^*(x,y,z) &= \frac{1}{\prod_{z''} n(z'')!}\sum_\pi\left(\frac{1}{n}\sum_{i=1}^n \mathbb{I}(X_{\pi(i)}=x, Y_i=y, Z_i=z)\right)\\
&= \frac{1}{\prod_{z''} n(z'')!}\frac{1}{n}\sum_{i=1}^n \mathbb{I}(Y_i=y, Z_i=z)\left(\sum_\pi \mathbb{I}(X_{\pi(i)}=x)\right) \qquad (2.2)\\
&= \frac{1}{\prod_{z''} n(z'')!}\frac{1}{n}\sum_{i=1}^n \mathbb{I}(Y_i=y, Z_i=z)n(x,z)(n(z)-1)!\prod_{z'\neq z} n(z')! = \hat{p}(x|z)\hat{p}(y,z).
\end{aligned}$$

Then, we compute the expected value of $\hat{p}^*(x,y,z)\hat{p}^*(x',y',z')$

$$\begin{aligned}
&\mathbb{E}^*\hat{p}^*(x,y,z)\hat{p}^*(x',y',z')\\
&= \frac{1}{\prod_{z''} n(z'')!}\sum_\pi\Big(\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathbb{I}(X_{\pi(i)}=x, Y_i=y, Z_i=z)\mathbb{I}(X_{\pi(j)}=x', Y_j=y', Z_j=z')\Big)\\
&= \frac{1}{\prod_{z''} n(z'')!}\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathbb{I}(Y_i=y, Z_i=z)\mathbb{I}(Y_j=y', Z_j=z')\sum_\pi \mathbb{I}(X_{\pi(i)}=x)\mathbb{I}(X_{\pi(j)}=x')\\
&= \frac{1}{\prod_{z''} n(z'')!}\frac{1}{n^2}\sum_{i\neq j} \mathbb{I}(Y_i=y, Z_i=z)\mathbb{I}(Y_j=y', Z_j=z')\sum_\pi \mathbb{I}(X_{\pi(i)}=x)\mathbb{I}(X_{\pi(j)}=x')\\
&\quad + \frac{1}{\prod_{z''} n(z'')!}\frac{1}{n^2}\mathbb{I}(x=x', y=y', z=z')\sum_i \mathbb{I}(Y_i=y, Z_i=z)\sum_\pi \mathbb{I}(X_{\pi(i)}=x)\\
&\overset{(\star)}{=} \left(\hat{p}(x|z)\hat{p}(x'|z')\left(1+\frac{\mathbb{I}(z=z')}{n(z)-1}-\frac{\mathbb{I}(x=x', z=z')}{(n(z)-1)\hat{p}(x|z)}\right)\right.\\
&\qquad \left.\cdot\sum_{i\neq j}\mathbb{I}(Y_i=y, Z_i=z)\mathbb{I}(Y_j=y', Z_j=z')\right) + \frac{1}{n}\mathbb{I}(x=x', y=y', z=z')\hat{p}(x|z)\hat{p}(y,z),
\end{aligned}$$

where $(\star)$ follows from ($i \neq j$ and $Z_i = z, Z_j = z'$)

$$\begin{aligned}
&\sum_\pi \mathbb{I}(X_{\pi(i)}=x)\mathbb{I}(X_{\pi(j)}=x')\\
&\qquad = \begin{cases}
n(x,z)(n(x,z)-1)(n(z)-2)!\prod_{z''\neq z} n(z'')! & \text{if } x=x', z=z'\\
n(x,z)n(x',z)(n(z)-2)!\prod_{z''\neq z} n(z'')! & \text{if } x\neq x', z=z'\\
n(x,z)n(x',z')(n(z)-1)!(n(z')-1)!\prod_{z''\neq z,z'} n(z'')! & \text{if } z\neq z'
\end{cases}
\end{aligned}$$

and as $\hat{p}(x|z) = n(x,z)/n$, we have

$$\frac{\sum_\pi \mathbb{I}(X_{\pi(i)} = x)\mathbb{I}(X_{\pi(j)} = x')}{\prod_{z''} n(z'')!} = \begin{cases} \hat{p}^2(x|z)\frac{n(z)}{n(z)-1} - \hat{p}(x|z)\frac{1}{n(z)-1} & \text{if } x = x', z = z' \\ \hat{p}(x|z)\hat{p}(x'|z')\frac{n(z)}{n(z)-1} & \text{if } x \neq x', z = z' \\ \hat{p}(x|z)\hat{p}(x'|z') & \text{if } z \neq z' \end{cases}.$$

We note that the sum $\sum_\pi \mathbb{I}(X_{\pi(i)} = x)\mathbb{I}(X_{\pi(j)} = x')$ does not depend on i and j. Next, writing the sum below as the difference of two sums, we have

$$\frac{1}{n^2}\sum_{i\neq j} \mathbb{I}(Y_i = y, Z_i = z)\mathbb{I}(Y_j = y', Z_j = z') = \frac{1}{n^2}\sum_{i,j} \mathbb{I}(Y_i = y, Z_i = z)\mathbb{I}(Y_j = y', Z_j = z')$$
$$-\frac{1}{n^2}\sum_{i} \mathbb{I}(Y_i = y, Z_i = z)\mathbb{I}(Y_i = y', Z_i = z') = \hat{p}(y,z)\hat{p}(y',z') - \frac{1}{n}\mathbb{I}(y = y', z = z')\hat{p}(y,z)$$

and hence

$$\begin{aligned}
&\mathbb{E}^*\hat{p}^*(x,y,z)\hat{p}^*(x',y',z') \\
&\quad= \hat{p}(y,z)\hat{p}(y',z')\hat{p}(x|z)\hat{p}(x'|z')\left(1 + \frac{\mathbb{I}(z=z')}{n(z)-1} - \frac{\mathbb{I}(x=x', z=z')}{(n(z)-1)\hat{p}(x|z)}\right) \\
&\qquad- \frac{1}{n}\hat{p}(y,z)\hat{p}(x|z)\hat{p}(x'|z')\Bigg(\mathbb{I}(y=y', z=z') + \frac{\mathbb{I}(y=y', z=z')}{n(z)-1} \\
&\qquad\quad- \frac{\mathbb{I}(x=x', y=y', z=z')}{(n(z)-1)\hat{p}(x|z)}\Bigg) + \frac{1}{n}\mathbb{I}(x=x', y=y', z=z')\hat{p}(x|z)\hat{p}(y,z) \\
&\quad= \hat{p}(y,z)\hat{p}(y',z')\hat{p}(x|z)\hat{p}(x'|z') + \frac{w(z)}{n}\mathbb{I}(z=z')\hat{p}(y,z)\hat{p}(y',z')\hat{p}(x|z)\hat{p}(x'|z')/\hat{p}(z) \\
&\qquad- \frac{w(z)}{n}\mathbb{I}(x=x', z=z')\hat{p}(y,z)\hat{p}(y',z')\hat{p}(x|z)/\hat{p}(z) \\
&\qquad- \frac{w(z)}{n}\mathbb{I}(y=y', z=z')\hat{p}(y,z)\hat{p}(x|z)\hat{p}(x'|z') \\
&\qquad+ \frac{w(z)}{n}\mathbb{I}(x=x', y=y', z=z')\hat{p}(y,z)\hat{p}(x|z),
\end{aligned}$$

where $w(z) = \frac{n\hat{p}(z)}{n\hat{p}(z)-1} = \frac{n(z)}{n(z)-1}$. Hence, we obtain

$$\begin{aligned}
(\mathrm{Cov}^*((\hat{p}^*(x,y,z))_{x,y,z}))_{x,y,z}^{x',y',z'} &= \frac{w(z)}{n}\mathbb{I}(z=z')\Big(\hat{p}(x|z)\hat{p}(x'|z)\hat{p}(y,z)\hat{p}(y',z)/\hat{p}(z) \\
&\quad- \mathbb{I}(x=x')\hat{p}(x|z)\hat{p}(y,z)\hat{p}(y',z)/\hat{p}(z) - \mathbb{I}(y=y')\hat{p}(x|z)\hat{p}(x'|z)\hat{p}(y,z) \\
&\quad+ \mathbb{I}(x=x', y=y')\hat{p}(x|z)\hat{p}(y,z)\Big). \quad (2.3)
\end{aligned}$$

Lemma 2.1.4. *Let* $\hat{p}^*(x,y,z) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}(X_i^* = x, Y_i = y, Z_i = z)$, *where*

$(X_i^*, Y_i, Z_i)_{i=1}^n = (X_{\pi(i)}, Y_i, Z_i)_{i=1}^n$, *where* $\pi \in \Pi$ *denotes a random permutation, is a sample obtained using conditional permutation. We have that*

i) for almost all sequences $(X_i, Y_i, Z_i)_{i=1}^\infty$

$$\mathbb{E}^* \hat{p}^*(x, y, z) \to p(x|z)p(y|z)p(z),$$

ii) for almost all sequences $(X_i, Y_i, Z_i)_{i=1}^\infty$

$$n(\mathrm{Cov}^*((\hat{p}^*(x, y, z))_{x,y,z})) \to \Sigma,$$

where

$$\begin{aligned}\Sigma_{x,y,z}^{x',y',z'} = \mathbb{I}(z = z')\Big(&p(x|z)p(y|z)p(x'|z)p(y'|z)p(z) - \mathbb{I}(x = x')p(x|z)p(y|z)p(y'|z)p(z) \\ &-\mathbb{I}(y = y')p(x|z)p(x'|z)p(y|z)p(z) + \mathbb{I}(x = x', y = y')p(x|z)p(y|z)p(z)\Big),\end{aligned} \tag{2.4}$$

iii) for any $(x_0, y_0, z_0) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ *conditionally on* $(X_i, Y_i, Z_i)_{i=1}^\infty$ *we have*

$$\sqrt{n}(\hat{p}^*(x_0, y_0, z_0) - \hat{p}_{ci}(x_0, y_0, z_0)) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{x_0,y_0,z_0}),$$

where

$$\sigma^2_{x_0,y_0,z_0} = \Sigma^{x_0,y_0,z_0}_{x_0,y_0,z_0} = p(z_0)p(x_0|z_0)(1 - p(x_0|z_0))p(y_0|z_0)(1 - p(y_0|z_0)).$$

iv) joint distribution of the vector $(n\hat{p}^*(x, y, z))_{x,y,z}$ *is given by*

$$P^*(n\hat{p}^*(x, y, z) = k(x, y, z)) = \prod_z \left(\frac{\prod_x n(x, z)! \prod_y n(y, z)!}{n(z)! \prod_{x,y} k(x, y, z)!} \right), \tag{2.5}$$

where $(k(x, y, z))_{x,y,z}$ *are sequences such that* $\sum_x k(x, y, z) = n(y, z)$ *and* $\sum_y k(x, y, z) = n(x, z)$, *otherwise* $P^*(n\hat{p}^*(x, y, z) = k(x, y, z)) = 0$.

Proof. The proof of *i*) and *ii*) follows from (2.2) and (2.3) and from the fact that $w(z) \to 1$ a.s. as $n \to \infty$.

For part *iii*), we prove first that $n\hat{p}^*(x_0, y_0, z_0)$ follows hypergeometric distribution, namely

$$P^*(n\hat{p}^*(x_0, y_0, z_0) = k) = \frac{\binom{n(y_0,z_0)}{k}\binom{n(z_0)-n(y_0,z_0)}{n(x_0,z_0)-k}}{\binom{n(z_0)}{n(x_0,z_0)}} \tag{2.6}$$

and k is an integer satisfying

$$\max\{0, n(x_0, z_0) + n(y_0, z_0) - n(z_0)\} \leq k \leq \min\{n(x_0, z_0), n(y_0, z_0)\}.$$

We note that $n\hat{p}^*(x_0, y_0, z_0) = n^*(x_0, y_0, z_0)$ and we have that $n^*(z_0) = n(z_0)$ as well as $n^*(x_0, z_0) = n(x_0, z_0)$ and $n^*(y_0, z_0) = n(y_0, z_0)$, but not necessarily $n^*(x_0, y_0, z_0) = n(x_0, y_0, z_0)$. Therefore, knowing $n^*(z_0)$, $n^*(x_0, z_0)$ and $n^*(y_0, z_0)$, we want to calculate a fraction of permutations, for which $n^*(x_0, y_0, z_0) = k$, which means that on the layer $Z = z_0$ we want to have exactly k pairs (X_i^*, Y_i), which are equal to (x_0, y_0). We have $n(x_0, z_0)$ and $n(y_0, z_0)$ observations on the layer $Z = z_0$, for which $X_i = x_0$ and $Y_i = y_0$ respectively, thus, to obtain exactly k pairs such that $(X_i^*, Y_i) = (x_0, y_0)$, we choose k out of $n(y_0, z_0)$ observations, which will also satisfy $X_i^* = x_0$ and $n(x_0, z_0) - k$ out of $n - n(y_0, z_0)$ to ensure, that all other observations such that $(Y_i, Z_i) = (y_0, z_0)$ have $X_i^* \neq x_0$. Now, we multiply it by the number of possible permutations in such a scenario, namely we separately permute observations with $X_i^* = x_0$ and $X_i^* \neq x_0$, and we obtain

$$\binom{n(y_0, z_0)}{k}\binom{n(z_0) - n(y_0, z_0)}{n(x_0, z_0) - k} n(x_0, z_0)!(n(z_0) - n(x_0, z_0)))!.$$

On the layer $Z = z_0$ we have $n(z_0)!$ permutations in total, thus we obtain (2.6). The rest of the proof follows from Theorem 2.1 in [21] (Theorem A.2.2), where using their notation we let $M_r = n(y_0, z_0)$, $N_r = n(z_0)$, $n = n(x_0, z_0)$ and $r = n$, thus $p_r = \hat{p}(y_0|z_0)$, $f_r = \hat{p}(x_0|z_0)$, $\sigma_r^2 = n\hat{p}(z_0)\hat{p}(x_0|z_0)(1 - \hat{p}(x_0|z_0))\hat{p}(y_0|z_0)(1 - \hat{p}(y_0|z_0))$ and $n_r p_r = n\hat{p}(z_0)\hat{p}(x_0|z_0)\hat{p}(y_0|z_0)$. We obtain that

$$\sup_{x\in\mathbb{R}} \left| P^* \left(\frac{n\hat{p}^*(x_0, y_0, z_0) - n\hat{p}(z_0)\hat{p}(x_0|z_0)\hat{p}(y_0|z_0)}{\sqrt{n\hat{p}(z_0)\hat{p}(x_0|z_0)(1 - \hat{p}(x_0|z_0))\hat{p}(y_0|z_0)(1 - \hat{p}(y_0|z_0))}} \right) - P(W \leq x) \right| \to 0$$

as $n \to \infty$ and $W \sim \mathcal{N}(0, 1)$ as $\sigma_r^2 \to \infty$. Hence, in view of Slutsky's theorem we have

$$\sqrt{n}(\hat{p}^*(x_0, y_0, z_0) - \hat{p}_{ci}(x_0, y_0, z_0)) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{x_0,y_0,z_0}),$$

as $n\sigma^2_{x_0,y_0,z_0}/\sigma_r \to 1$ a.s.

To prove *iv*) lets first focus on one layer of Z, namely $Z = z$. We have on that layer $n(z)$ observations and $n(x,z)$, $n(y,z)$ observations such that $X_i = x$ and $Y_i = y$, respectively. For each layer of Y we know exactly how many times each value x of X has to appear and that equals $k(x,y,z)$ for the layer $Y = y, Z = z$. We want to calculate how many permutations of X_i satisfy that condition. First, we will assign a layer of Y to each observation such that $X = x$. The number of such assignments is

$$\binom{n(x,z)}{k(y_1,x,z)} \cdot \binom{n(x,z) - k(y_1,x,z)}{k(y_2,x,z)} \cdot \binom{n(x,z) - \sum_{i=1}^{2} k(y_i,x,z)}{k(y_3,x,z)} \cdot \ldots \cdot \binom{n(x,z) - \sum_{i=1}^{r-2} k(y_i,x,z)}{k(y_{r-1},x,z)} \cdot \binom{n(x,z) - \sum_{i=1}^{r-1} k(y_i,x,z)}{k(y_r,x,z)}, \quad (2.7)$$

where $|\mathcal{Y}| = r$ and $\mathcal{Y} = \{y_1, y_2, \ldots, y_r\}$. Namely, we choose which X_i such that $X_i = x$ will be in the layer $Y = y_1$, then from the remaining X we choose those that will be in the layer $Y = y_2$ and so on. Finally, we have $n(x,z) - \sum_{i=1}^{r-1} k(y_i,x,z) = k(y_r,x,z)$ of X left and we assign them to the last layer of Y. The term (2.7) can be expressed as

$$\frac{n(x,z)!}{\prod_y k(x,y,z)!}$$

and as we repeat that for all $x \in \mathcal{X}$, we obtain

$$\frac{\prod_x n(x,z)!}{\prod_{x,y} k(x,y,z)!}.$$

We now have to take into account the number of permutations of Y on each layer $Y = y, Z = z$ which equals $n(y,z)!$, as we omitted that above. Repeating this for every layer $Z = z$ we finally obtain

$$\prod_z \left(\frac{\prod_x n(x,z)! \prod_y n(y,z)!}{\prod_{x,y} k(x,y,z)!} \right).$$

This follows from the fact, that $n\hat{p}^*(x,y,z_1)$ and $n\hat{p}^*(x',y',z_2)$ are conditionally independent given the original sample for $z_1 \neq z_2$. The number of unconstrained permutations on the layers of Z equals $\prod_z n(z)!$, hence we get (2.5). □

We note that *iv*) was proven more rigorously for unconditional case in [15].

Lemma 2.1.5. *Let $(\hat{p}^*(x,y,z))_{x,y,z}$ be a vector of probabilities estimated based on sample obtained through conditional permutation. For almost all $(X_i,Y_i,Z_i)_{i=1}^{\infty}$ and conditionally on $(X_i,Y_i,Z_i)_{i=1}^{\infty}$ we have that*

$$\sqrt{n}\left(\hat{p}^*(x,y,z) - \hat{p}(x|z)\hat{p}(y|z)\hat{p}(z)\right)_{x,y,z} \xrightarrow{d} \mathcal{N}(0,\Sigma), \tag{2.8}$$

where Σ is defined in (2.4).

Note that the above lemma is consistent with results from Lemma 2.1.4, but it does not follow from them directly. In *iii*) we prove convergence just for one fixed triple (x_0,y_0,z_0), whereas in Lemma 2.1.5 we assert that the convergence holds for the whole vector of probabilities $\hat{p}^*$ and the asymptotic covariance matrix is equal to the one given in *ii*). Also note that in view of (2.5) subvectors

$$\left(\hat{p}^*(\cdot,\cdot,z_1),\hat{p}^*(\cdot,\cdot,z_2),\ldots,\hat{p}^*(\cdot,\cdot,z_{|S|})\right)$$

are independent given (X_i,Y_i,Z_i), thus in order to prove (2.8) it is sufficient to prove analogous result for unconditional permutation scenario on one fixed layer of Z. Namely, having

$$\sqrt{n(z_0)}(\hat{p}^*(x,y|z_0) - \hat{p}(x|z_0)\hat{p}(y|z_0))_{x,y} \xrightarrow{d} \mathcal{N}(0,\Sigma_{z_0}) \tag{2.9}$$

conditionally on the sample, we obtain that also the following convergence holds

$$\begin{aligned}\sqrt{n}\hat{p}^*(z_0)&(\hat{p}^*(x,y|z_0) - \hat{p}(x|z_0)\hat{p}(y|z_0))_{x,y}\\ &= \sqrt{\frac{n(z_0)}{n}}\cdot\sqrt{n(z_0)}(\hat{p}^*(x,y|z_0) - \hat{p}(x|z_0)\hat{p}(y|z_0))_{x,y}\\ &\xrightarrow{d} \sqrt{p(z_0)}\mathcal{N}(0,\Sigma_{z_0})\end{aligned}$$

for fixed z_0. The vector of probabilities can be written as

$$\begin{aligned}\sqrt{n}(\hat{p}^*(x,y,z) - \hat{p}(x|z)\hat{p}(y|z)\hat{p}(z))_{x,y,z} =&\\ \sqrt{n}\Big(&\hat{p}^*(z_1)(\hat{p}^*(x,y|z_1) - \hat{p}(x|z_1)\hat{p}(y|z_1))_{x,y},\ldots,\\ &\hat{p}^*(z_{|S|})(\hat{p}^*(x,y|z_{|S|}) - \hat{p}(x|z_{|S|})\hat{p}(y|z_{|S|}))_{x,y}\Big)\end{aligned}$$

and as its components $\hat{p}^*(z_i)(\hat{p}^*(x,y|z_i)-\hat{p}(x|z_i)\hat{p}(y|z_i))_{x,y}$ are independent given the sample, we obtain Lemma 2.1.5. Showing convergence (2.9) for permutation scenario is thus crucial for showing the convergence of vector $\hat{p}^*$. The proof of that based on asymptotic normality of standardised generalised multivariate hypergeometric distribution is given in [22].

2.1.5. Comparison of covariance matrices

We compare now asymptotic covariance matrices of the resampling scenarios presented before, namely CI bootstrap sampling, CR/bootstrap X and permutation. We denote their covariance matrices by $\Sigma_{(1)}$, $\Sigma_{(2)}$, $\Sigma_{(3)}$, respectively (see Lemma 2.1.1, 2.1.3 and 2.1.4)).

Remark 2.1.6. *Note that $\Sigma_{(3)}$ for permutation scenario defined in* (2.4) *can be written as*

$$\Sigma_{x,y,z}^{x',y',z'} = \mathbb{I}(z=z')\cdot p(x|z)\big(\mathbb{I}(x=x')-p(x'|z)\big)\cdot p(y|z)\big(\mathbb{I}(y=y')-p(y'|z)\big)$$

and $\Sigma_{(2)}$ for CR and Bootstrap X equals

$$\Sigma_{x,y,z}^{x',y',z'} = \mathbb{I}(z=z')\cdot p(x|z)\big(\mathbb{I}(x=x')-p(x'|z)\big)\cdot p(y|z)\mathbb{I}(y=y').$$

Sum of two terms of elements of the matrix $\Sigma_{(3)}$ coincides with corresponding elements of the asymptotic covariance matrix $\Sigma_{(2)}$, namely $\mathbb{I}(x=x',y=y',z=z')p(x|z)p(y|z)p(z) - \mathbb{I}(y=y',z=z')p(x|z)p(x'|z)p(y|z)p(z)$.

Lemma 2.1.7. *The covariance matrix for CI bootstrap scenario dominates the covariance matrix for CR/bootstrap X scenarios, whereas the covariance matrix for CR/bootstrap X scenarios dominates the covariance matrix for permutation scenario*

$$\Sigma_{(1)} \geq \Sigma_{(2)} \geq \Sigma_{(3)},$$

i.e. matrices $\Sigma_{(1)}-\Sigma_{(2)}$ and $\Sigma_{(2)}-\Sigma_{(3)}$ are positive semi-definite.

Proof. We start by proving $\Sigma_{(1)} \geq \Sigma_{(2)}$. We define

$$(R)_{x,y,z}^{x',y',z'} = (\Sigma_{(1)} - \Sigma_{(2)})_{x,y,z}^{x',y',z'} = \mathbb{I}(y = y', z = z')p(x|z)p(x'|z')p(y,z) \\ - p(x|z)p(x'|z')p(y,z)p(y',z')$$

and

$$(\tilde{R})_{y,z}^{y',z'} = r_{y,z}^{y',z'} = \mathbb{I}(y = y', z = z')p(y,z) - p(y,z)p(y',z').$$

The matrix $\tilde{R}$ is positive semi-definite as it is a covariance matrix of multinomial distribution on $\mathcal{Y} \times \mathcal{Z}$ with probabilities given by the vector $(p(y,z))_{y,z}$. We have that

$$(R)_{x,y,z}^{x',y',z'} = r_{y,z}^{y',z'} p(x|z)p(x'|z').$$

For any non-zero vector $a = (a(x,y,z))_{x,y,z}$ we have that

$$a'Ra = \sum_{x,y,z} \sum_{x',y',z'} a_{x,y,z} r_{y,z}^{y',z'} p(x|z)p(x'|z')a_{x',y',z'} \\ = \sum_{y,z} \sum_{y',z'} \left(\sum_{x} a_{x,y,z} p(x|z) \right) r_{y,z}^{y',z'} \left(\sum_{x'} a_{x',y',z'} p(x'|z') \right) \geq 0,$$

thus R is a positive semi-definite matrix.

Now we prove $\Sigma_{(2)} \geq \Sigma_{(3)}$. Lets define (cf. Remark 2.1.6)

$$(R)_{x,y,z}^{x',y',z'} = (\Sigma_{(2)} - \Sigma_{(3)})_{x,y,z}^{x',y',z'} = \mathbb{I}(z = z') \big[\mathbb{I}(x = x')p(x|z)p(y,z)p(y',z)/p(z) \\ - p(x|z)p(x'|z)p(y,z)p(y',z)/p(z) \big].$$

We notice that for any z the matrix $\tilde{R}(z)$ defined in the following way

$$(\tilde{R}(z))_{x}^{x'} = r_{x}^{x'}(z) = \mathbb{I}(x = x')p(x|z) - p(x|z)p(x'|z)$$

is positive semi-definite. Now we define $\bar{R}(z)$

$$\big(\bar{R}(z)\big)_{x,y}^{x',y'} = r_{x,y}^{x',y'}(z) = r_{x}^{x'}(z)p(y,z)p(y',z)$$

and we show, that $\bar{R}(z) \geq 0$. Namely, analogously to previous reasoning, for any non-zero vector $a = (a(x,y))_{x,y}$ we have that

$$a'\bar{R}(z)a = \sum_{x,y}\sum_{x',y'} a_{x,y} r_{x,y}^{x',y'}(z) a_{x',y'} = \sum_{x,y}\sum_{x',y'} a_{x,y} r_{x}^{x'}(z) p(y,z) p(y',z) a_{x',y'}$$
$$= \sum_{x,x'} \left(\sum_{y} a_{x,y} p(y,z) \right) r_{x}^{x'}(z) \left(\sum_{y'} a_{x',y'} p(y',z) \right) \geq 0,$$

where the last inequality follows as $\tilde{R}(z) \geq 0$. We have that

$$(R)_{x,y,z}^{x',y',z'} = r_{x,y,z}^{x',y',z'} = r_{x,y}^{x',y'} \mathbb{I}(z = z')/p(z),$$

thus for any non-zero vector $a = (a(x,y,z))_{x,y,z}$ we have that

$$a'Ra = \sum_{x,y,z}\sum_{x',y',z'} a_{x,y,z} r_{x,y,z}^{x',y',z'} a_{x',y',z'} = \sum_{x,y,z}\sum_{x',y',z'} a_{x,y,z} r_{x,y}^{x',y'}(z) \mathbb{I}(z = z')/p(z) a_{x',y',z'}$$
$$= \sum_{z} \left(\sum_{x,y}\sum_{x',y'} a_{x,y,z} r_{x,y}^{x',y'}(z) a_{x',y',z} \right) /p(z) \geq 0. \quad \square$$

Intuitively, the ordering of matrices in Lemma 2.1.7 is due to the fact that permutation scenario introduces the smallest amount of external variability (conditional empirical distribution of X given $Z = z$ remains unchanged in the permuted sample), whereas CI bootstrap introduces the largest amount.

2.2. Asymptotic behaviour of $\widehat{CMI}^*$

In this section we obtain an asymptotic distribution of conditional mutual information computed at estimated probabilities for the resampling scenarios presented previously, namely

$$\widehat{CMI}^* := CMI(\hat{p}^*) = \sum_{x,y,z} \hat{p}^*(x,y,z) \log \frac{\hat{p}^*(x,y|z)}{\hat{p}^*(x|z)\hat{p}^*(y|z)},$$

then in Sections 2.2.2 and 2.2.3 we show that we can use quantiles of $\widehat{CMI}^*$ for conditional independence testing.

2.2.1. Distribution of $\widehat{CMI}^*$

Below we state asymptotic distributions of $\widehat{CMI}^*$ for bootstrap CI, CR/bootstrap X and permutation scenario.

CI bootstrap

We start by showing that the asymptotic distributions of $2n\widehat{CMI}^*$ for CI bootstrap is the same as for $2n\widehat{CMI}$.

Lemma 2.2.1. *For almost all sequences $(X_1, Y_1, Z_1), (X_2, Y_2, Z_2), \ldots$ and conditionally on $(X_i, Y_i, Z_i)_{i=1}^{\infty}$, we have that*

$$2nCMI(\hat{p}^*) \xrightarrow{d} \chi^2_{(|\mathcal{X}|-1)(|\mathcal{Y}|-1)|\mathcal{Z}|}$$

where $\hat{p}^*$ denotes estimated probabilities based on CI bootstrap sample.

Proof. First, we recall that the gradient and Hessian matrix of conditional mutual information considered as a function of $(p(x, y, z))_{x,y,z}$ equal

$$(D_{CMI}(p))(x, y, z) = \frac{\partial CMI(p)}{\partial p(x, y, z)} = \log \frac{p(x, y, z)p(z)}{p(x, z)p(y, z)}, \tag{2.10}$$

and

$$(H_{CMI}(p))^{x',y',z'}_{x,y,z} = \frac{\mathbb{I}(x = x', y = y', z = z')}{p(x, y, z)} - \frac{\mathbb{I}(x = x', z = z')}{p(x, z)} - \frac{\mathbb{I}(y = y', z = z')}{p(y, z)} + \frac{\mathbb{I}(z = z')}{p(z)}, \tag{2.11}$$

see the proof of Lemma 1.4.2. The proof now follows from the expansion

$$CMI(\hat{p}^*) \;=\; CMI(\hat{p}_{ci}) \;+\; (\hat{p}^* \;-\; \hat{p}_{ci})' D_{CMI}(\hat{p}_{ci}) \;+\; \frac{1}{2}(\hat{p}^* \;-\; \hat{p}_{ci})' H_{CMI}(\xi)(\hat{p}^* \;-\; \hat{p}_{ci}),$$

where $\xi = (\xi_{x,y,z})_{x,y,z}$ and $\xi_{x,y,z}$ is a point between $\hat{p}^*(x, y, z)$ and $\hat{p}_{ci}(x, y, z)$. We have that

$$CMI(\hat{p}^*) = CMI(\hat{p}_{ci}) + (\hat{p}^* - \hat{p}_{ci})' D_{CMI}(\hat{p}_{ci}) + \frac{1}{2}(\hat{p}^* - \hat{p}_{ci})' H_{CMI}(p_{ci})(\hat{p}^* - \hat{p}_{ci}) + \frac{1}{2}(\hat{p}^* - \hat{p}_{ci})' (H_{CMI}(\xi) - H_{CMI}(p_{ci}))(\hat{p}^* - \hat{p}_{ci}).$$

For almost all sequences $|\hat{p}^*(x,y,z) - \hat{p}_{ci}(x,y,z)| \to 0$ conditionally on the sample and $|\hat{p}_{ci}(x,y,z) - p_{ci}(x,y,z)| \to 0$ a.s., thus for almost all sequences and conditionally on the sample $\xi_{x,y,z} \to p_{ci}(x,y,z)$ as

$$\begin{aligned}|\xi_{x,y,z} - p_{ci}(x,y,z)| &\leq |\xi_{x,y,z} - \hat{p}_{ci}(x,y,z)| + |\hat{p}_{ci}(x,y,z) - p_{ci}(x,y,z)| \\ &\leq |\hat{p}^*(x,y,z) - \hat{p}_{ci}(x,y,z)| + |\hat{p}_{ci}(x,y,z) - p_{ci}(x,y,z)|.\end{aligned}$$

For all triples (x,y,z) we have $p(x,y,z) > 0$ (and consequently $p(x,z) > 0$, $p(y,z) > 0$, $p(z) > 0$) and using continuity at $p_{ci}(x,y,z)$ of the matrix H_{CMI}, each element of $H_{CMI}(\xi)$ converges to the element of $H_{CMI}(p_{ci})$. Hence,

$$\begin{aligned}CMI(\hat{p}^*) = CMI(\hat{p}_{ci}) &+ (\hat{p}^* - \hat{p}_{ci})' D_{CMI}(\hat{p}_{ci}) \\ &+ \frac{1}{2}(\hat{p}^* - \hat{p}_{ci})' H_{CMI}(p_{ci})(\hat{p}^* - \hat{p}_{ci}) + o_{p^*}(\|\hat{p}^* - \hat{p}_{ci}\|^2).\end{aligned}$$

The gradient of conditional mutual information D_{CMI} at $\hat{p}_{ci}$, where $\hat{p}_{ci}(x,y,z) = \hat{p}(x|z)\hat{p}(y,z)$, equals

$$\begin{aligned}(D_{CMI}(\hat{p}_{ci}))(x,y,z) = \log \frac{\hat{p}_{ci}(x,y,z)\hat{p}_{ci}(z)}{\hat{p}_{ci}(x,z)\hat{p}_{ci}(y,z)} &= \log \frac{\hat{p}_{ci}(x,y,z)\hat{p}(z)}{\hat{p}(x,z)\hat{p}(y,z)} \\ &= \log \frac{\hat{p}(x|z)\hat{p}(y,z)\hat{p}(z)}{\hat{p}(x,z)\hat{p}(y,z)} = 0,\end{aligned}$$

as

$$\hat{p}_{ci}(x,z) = \sum_y \hat{p}_{ci}(x,y,z) = \sum_y \hat{p}(x|z)\hat{p}(y,z) = \hat{p}(x,z)$$

and analogously $\hat{p}_{ci}(y,z) = \hat{p}(y,z)$ and $\hat{p}_{ci}(z) = \hat{p}(z)$. We also have that $CMI(\hat{p}_{ci}) = 0$, thus

$$2nCMI(\hat{p}^*) = \sqrt{n}(\hat{p}^* - \hat{p}_{ci})' H_{CMI}(p_{ci}) \sqrt{n}(\hat{p}^* - \hat{p}_{ci}) + o_{p^*}(n\,\|\hat{p}^* - \hat{p}_{ci}\|^2).$$

Then, as $o_{p^*}(\sqrt{n}\,\|\hat{p}^* - \hat{p}_{ci}\|) = o_{p^*}(1)$, we obtain that

$$2nCMI(\hat{p}^*) \to W' H_{CMI}(p_{ci}) W,$$

where $W \sim \mathcal{N}(0, \Sigma)$ and Σ is defined in Lemma 2.1.1. Thus,

$$2nCMI(\hat{p}^*) \xrightarrow{d} \sum_{x,y,z} \lambda_{x,y,z} Z^2_{x,y,z}, \tag{2.12}$$

where $Z = (Z_{x,y,z})_{x,y,z} \sim \mathcal{N}(0, I)$ and in view of spectral decomposition $\lambda_{x,y,z}$ are eigenvalues of a matrix $M = H_{CMI}(p_{ci})\Sigma$ (justification is analogous to that in the proof of Lemma 1.4.2). Using the fact that the matrix M is the same as in Lemma 1.4.3, we obtain

$$2nCMI(\hat{p}^*) \xrightarrow{d} \chi^2_{(|\mathcal{X}|-1)(|\mathcal{Y}|-1)|\mathcal{Z}|}. \qquad \square$$

CR/Bootstrap X

The result below states that the asymptotic distribution of conditional mutual information at estimated probabilities for CR and bootstrap X is the chi-square distribution with the same number of degrees of freedom as for CI bootstrap (the analogous result is also shown for permutation scenario). Thus we show that different resampling scenarios lead to the same asymptotic distribution of $\widehat{CMI}^*$. This is interesting as CR scenario requires additional information not used in the remaining scenarios. Moreover, although the matrices $\Sigma_{(i)}$ corresponding to different resampling scenarios are non-identical (cf. Lemma 2.1.7 in which we showed that covariance matrix corresponding to bootstrap CI scenario dominates covariance matrix of CR/bootstrap X scenarios as well as permutation scenario), they also lead to the same asymptotic result.

Lemma 2.2.2. *If $\hat{p}^*$ denotes estimated probabilities based on CR or bootstrap X, namely*

$$\hat{p}^*(x, y, z) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i^* = x, Y_i = y, Z_i = z),$$

where $(X_i^*, Y_i, Z_i)_{i=1}^n$ is a CR or a bootstrap X sample, then for almost all sequences $(X_1, Y_1, Z_1), (X_2, Y_2, Z_2), \ldots$ and conditionally on $(X_i, Y_i, Z_i)_{i=1}^\infty$, we have that

$$2nCMI(\hat{p}^*) \xrightarrow{d} \chi^2_{(|\mathcal{X}|-1)(|\mathcal{Y}|-1)|\mathcal{Z}|}.$$

Proof. As in the proof of Lemma 2.2.1, we have that

$$CMI(\hat{p}^*) = CMI(\hat{p}_{ci}) + (\hat{p}^* - \hat{p}_{ci}))' D_{CMI}(\hat{p}_{ci})$$

$$+\frac{1}{2}(\hat{p}^*-\hat{p}_{ci})'H_{CMI}(p_{ci})(\hat{p}^*-\hat{p}_{ci})+o_{p^*}(\|\hat{p}^*-\hat{p}_{ci}\|^2),$$

where the formula for D_{CMI} and H_{CMI} is given in (2.10) and (2.11). The gradient of conditional mutual information D_{CMI} at $\hat{p}_{ci}$, where $\hat{p}_{ci}=\hat{p}(x|z)\hat{p}(y,z)$, equals 0 (cf. the proof of Lemma 2.2.1). Similarly,

$$(D_{CMI}(\hat{p}_{tci}))(x,y,z)=\log\frac{\hat{p}_{tci}(x,y,z)\hat{p}_{tci}(z)}{\hat{p}_{tci}(x,z)\hat{p}_{tci}(y,z)}=\log\frac{\hat{p}_{tci}(x,y,z)\hat{p}(z)}{p(x|z)\hat{p}(z)\hat{p}(y,z)}$$
$$=\log\frac{p(x|z)\hat{p}(y,z)\hat{p}(z)}{p(x|z)\hat{p}(z)\hat{p}(y,z)}=0,$$

where $\hat{p}_{tci}=p(x|z)\hat{p}(y,z)$ as

$$\hat{p}_{tci}(x,z)=\sum_y\hat{p}_{tci}(x,y,z)=\sum_y p(x|z)\hat{p}(y,z)=p(x|z)\hat{p}(z),$$

$$\hat{p}_{tci}(z)=\sum_x p(x|z)\hat{p}(z)=\hat{p}(z)$$

and $\hat{p}_{tci}(y,z)=\hat{p}(y,z)$. Thus in both cases the second order expansion is needed. We note that $CMI(\hat{p}_{ci})=CMI(\hat{p}_{tci})=0$. Hence, reasoning similarly as in the proof of Lemma 2.2.1, we have

$$2nCMI(\hat{p}^*)=n(\hat{p}^*-\hat{p}_{ci})'H_{CMI}(p_{ci})(\hat{p}^*-\hat{p}_{ci})+o_{p^*}(n\|\hat{p}^*-\hat{p}_{ci}\|^2).$$

As $o_{p^*}(n\|\hat{p}^*-\hat{p}_{ci}\|^2)\to 0$, we obtain that

$$2nCMI(\hat{p}^*)\xrightarrow{d}W'H_{CMI}(p_{ci})W,$$

where $W\sim\mathcal{N}(0,\Sigma)$ and Σ is defined in Lemma 2.1.2. It follows that

$$2nCMI(\hat{p}^*)\xrightarrow{d}\sum_{x,y,z}\lambda_{x,y,z}Z^2_{x,y,z},$$

where $Z=(Z_{x,y,z})_{x,y,z}\sim\mathcal{N}(0,I)$ and $\lambda_{x,y,z}$ are eigenvalues of a matrix $M=H_{CMI}(p_{ci})\Sigma$ (justification is analogous to that in the proof of Lemma 1.4.2). From Lemma 2.2.3 below

we see that the form of the matrix M is the same as in the proof of Lemma 1.4.2, hence

$$2nCMI(\hat{p}^*) \xrightarrow{d} \chi^2_{(|\mathcal{X}|-1)(|\mathcal{Y}|-1)|\mathcal{Z}|}. \qquad \square$$

Now in the lemma below we give an explicit formula for the matrix M used in the proof of Lemma 2.2.2. We recall that the covariance matrix Σ of asymptotic normal distribution of $\hat{p}^*$, where $\hat{p}^*$ is based on bootstrap X or CR sample, equals

$$\Sigma_{x,y,z}^{x',y',z'} = -\mathbb{I}(y=y',z=z')p(x|z)p(x'|z')p(y,z) + \mathbb{I}(x=x',y=y',z=z')p(x|z)p(y,z)$$

and the Hessian matrix of conditional mutual information equals

$$H_{x,y,z}^{x',y',z'} = \frac{\mathbb{I}(x=x',y=y',z=z')}{p(x,y,z)} - \frac{\mathbb{I}(x=x',z=z')}{p(x,z)} - \frac{\mathbb{I}(y=y',z=z')}{p(y,z)} + \frac{\mathbb{I}(z=z')}{p(z)}.$$

The Hessian in Lemma 2.2.2 is computed at point p_{ci} and we use that fact in the proof of Lemma 2.2.3, as we are interested in M defined at that point, namely $M := H\Sigma = H_{CMI}(p_{ci})\Sigma$.

Lemma 2.2.3. *Matrix $M = H\Sigma = H_{CMI}(p_{ci})\Sigma$, where Σ is an asymptotic covariance matrix for CR/bootstrap X scenarios, has the following form*

$$\begin{aligned} M_{x,y,z}^{x'',y'',z''} &= \mathbb{I}(x=x'',y=y'',z=z'') - \mathbb{I}(x=x'',z=z'')p(y''|z'') \\ &\quad - \mathbb{I}(y=y'',z=z'')p(x''|z'') + \mathbb{I}(z=z'')p(x''|z'')p(y''|z'') \end{aligned} \tag{2.13}$$

and coincides with M given in (1.32).

Proof. Multiplication of matrices H and Σ yields:

$$\begin{aligned} M_{x,y,z}^{x'',y'',z''} &= \sum_{x',y',z'} H_{x,y,z}^{x',y',z'}\Sigma_{x',y',z'}^{x'',y'',z''} = \sum_{x',y',z'} \Bigg(\underbrace{\frac{\mathbb{I}(x=x',y=y',z=z')}{p(x,y,z)}}_{a} - \underbrace{\frac{\mathbb{I}(x=x',z=z')}{p(x,z)}}_{b} \\ &\quad - \underbrace{\frac{\mathbb{I}(y=y',z=z')}{p(y,z)}}_{c} + \underbrace{\frac{\mathbb{I}(z=z')}{p(z)}}_{d} \Bigg) \Bigg(- \underbrace{\mathbb{I}(y'=y'',z'=z'')p(x'|z')p(x''|z'')p(y',z')}_{e} \\ &\quad + \underbrace{\mathbb{I}(x'=x'',y'=y'',z'=z'')p(x'|z')p(y',z')}_{f} \Bigg) = - \underbrace{\mathbb{I}(y=y'',z=z'')p(x''|z'')}_{a\cdot e} \end{aligned}$$

$$+\underbrace{\mathbb{I}(z=z'')p(x''|z'')p(y''|z'')}_{b\cdot e}+\underbrace{\mathbb{I}(y=y'',z=z'')p(x''|z'')}_{c\cdot e}-\underbrace{\mathbb{I}(z=z'')p(x''|z'')p(y''|z'')}_{d\cdot e}$$
$$+\underbrace{\mathbb{I}(x=x'',y=y'',z=z'')}_{a\cdot f}-\underbrace{\mathbb{I}(x=x'',z=z'')p(y''|z'')}_{b\cdot f}-\underbrace{\mathbb{I}(x=x'',z=z'')p(x''|z'')}_{c\cdot f}$$
$$+\underbrace{\mathbb{I}(z=z'')p(x''|z'')p(y''|z'')}_{d\cdot f}=\mathbb{I}(x=x'',y=y'',z=z'')-\mathbb{I}(x=x'',z=z'')p(y''|z'')$$
$$-\mathbb{I}(y=y'',z=z'')p(x''|z'')+\mathbb{I}(z=z'')p(x''|z'')p(y''|z'').$$

Below we present detailed calculations for the terms $c\cdot e$ and $d\cdot f$ (the calculations for other terms are analogous):

$$\begin{aligned}
c\cdot e &= \sum_{x',y',z'}\mathbb{I}(y=y',z=z')\mathbb{I}(y'=y'',z'=z'')\frac{p(x'|z')p(x''|z'')p(y',z')}{p(y,z)}\\
&= \mathbb{I}(y=y'',z=z'')\sum_{x'}\frac{p(x'|z)p(x''|z'')p(y,z)}{p(y,z)} = \mathbb{I}(y=y'',z=z'')p(x''|z'')\sum_{x'}p(x'|z)\\
&= \mathbb{I}(y=y'',z=z'')p(x''|z''),\\
d\cdot f &= \sum_{x',y',z'}\mathbb{I}(z=z')\mathbb{I}(x'=x'',y'=y'',z'=z'')\frac{p(x'|z')p(y',z')}{p(z)}\\
&= \mathbb{I}(z=z'')\frac{p(x''|z'')p(y'',z'')}{p(z)} = \mathbb{I}(z=z'')p(x''|z'')p(y''|z''). \qquad \square
\end{aligned}$$

Permutation

The lemma analogous to Lemmas 2.2.1 and 2.2.2 also holds for conditional permutation scenario:

Lemma 2.2.4. *If $\hat{p}^*$ denotes estimated probabilities based on conditional permutations, namely*

$$\hat{p}^*(x,y,z)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(X_i^*=x,Y_i=y,Z_i=z),$$

where $(X_i^*,Y_i,Z_i)_{i=1}^n$ is a sample obtained by permuting X on Z layers, then for almost all sequences $(X_1,Y_1,Z_1),(X_2,Y_2,Z_2),\ldots$ and conditionally on $(X_i,Y_i,Z_i)_{i=1}^\infty$, we have that

$$2nCMI(\hat{p}^*)\xrightarrow{d}\chi^2_{(|\mathcal{X}|-1)(|\mathcal{Y}|-1)|\mathcal{Z}|}.$$

Proof. The proof is analogous to the proof of Lemma 2.2.2. In this case matrix $M=H_{CMI}(p_{ci})\Sigma$, where Σ is asymptotic covariance of vector of probabilities for per-

mutation scenario (cf. Lemma 2.1.4 for the formula), is equal to the matrix M defined in the proof of Lemma 2.2.2. This is shown below in Lemma 2.2.5. □

Lemma 2.2.5. *Matrix $M = H\Sigma = H_{CMI}(p_{ci})\Sigma$, where Σ is an asymptotic covariance matrix for permutation scenario, is equal to* (2.13).

Proof. We have that

$$M_{x,y,z}^{x'',y'',z''} = \sum_{x',y',z'} H_{x,y,z}^{x',y',z'} \Sigma_{x',y',z'}^{x'',y'',z''} = \sum_{x',y',z'} \Bigg(\underbrace{\frac{\mathbb{I}(x=x',y=y',z=z')}{p(x,y,z)}}_{a} - \underbrace{\frac{\mathbb{I}(x=x',z=z')}{p(x,z)}}_{b}$$
$$- \underbrace{\frac{\mathbb{I}(y=y',z=z')}{p(y,z)}}_{c} + \underbrace{\frac{\mathbb{I}(z=z')}{p(z)}}_{d} \Bigg) \Big(\underbrace{\mathbb{I}(z'=z'')p(x'|z')p(x''|z')p(y',z')p(y'',z')/p(z)}_{g}$$
$$- \underbrace{\mathbb{I}(x'=x'',z'=z'')p(x'|z')p(y'',z'')p(y',z')/p(z')}_{h}$$
$$- \underbrace{\mathbb{I}(y'=y'',z'=z'')p(x'|z')p(x''|z'')p(y',z')}_{e} + \underbrace{\mathbb{I}(x'=x'',y'=y'',z'=z'')p(x'|z')p(y',z')}_{f} \Big)$$

and we notice, that $(a-b-c+d)(-e+f)$ was computed in the proof of Lemma 1.4.3, the term $(a-b-c+d)h = 0$ similarly to $(a-b-c+d)e$, thus we just need to compute $(a-b-c+d)g$. As $ag = bg = cg = dg = \mathbb{I}(z=z'')p(x''|z)p(y'',z)/p(z)$, we obtain the same matrix M as in Lemma 1.4.3. □

In this section we have shown that regardless of the resampling scenario, the asymptotic distribution of $2n\widehat{CMI}^*$ is the same as of $2n\widehat{CMI}$. In the next section this result is used to justify the use of quantiles of distribution of $\widehat{CMI}^*$ and comparing them with those of $\widehat{CMI}$ in testing conditional independence.

2.2.2. Validity of asymptotic convergence

First, assume that for almost all sequences $(X_1,Y_1,Z_1),(X_2,Y_2,Z_2),\ldots$ and conditionally on $(X_i,Y_i,Z_i)_{i=1}^{\infty}$, we have that

$$F_n^*(t) = P^*(T_n(\mathbf{X}_n^*,\mathbf{Y}_n^*,\mathbf{Z}_n^*) \leq t) \to F(t) \tag{2.14}$$

for all continuity points of F and that (unconditionally)

$$F_n(t) = P(T_n(\mathbf{X}_n,\mathbf{Y}_n,\mathbf{Z}_n) \leq t) \to F(t), \tag{2.15}$$

where $(\mathbf{X}_n^*, \mathbf{Y}_n^*, \mathbf{Z}_n^*) = (X_i^*, Y_i^*, Z_i^*)_{i=1}^n$ and $(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{Z}_n) = (X_i, Y_i, Z_i)_{i=1}^n$, respectively. We define the quantile function in the following way

$$F^{-1}(1-\alpha) = \inf\{t : F(t) \geq 1-\alpha\},$$

where F is a cumulative distribution function. The following Lemma is based on Theorem 1.2.1 in [31].

Lemma 2.2.6. *If the convergences* (2.14) *and* (2.15) *are satisfied,* $\alpha \in (0,1)$ *and* F *is strictly increasing at* $F^{-1}(1-\alpha)$*, then*

i) $(F_n^*)^{-1}(1-\alpha) \to F^{-1}(1-\alpha)$ *a.s.*

If additionally F *is continuous at* $F^{-1}(1-\alpha)$*, then*

ii) $P(T_n(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{Z}_n) \leq (F_n^*)^{-1}(1-\alpha)) \to 1-\alpha.$

Proof. We start with the proof of part *i*). Denote $F^{-1}(1-\alpha)$ by t. Fix $\varepsilon > 0$ and let F be continuous at $t - \varepsilon_0 = F^{-1}(1-\alpha) - \varepsilon_0$ and $t + \varepsilon_0 = F^{-1}(1-\alpha) + \varepsilon_0$ for some $\varepsilon_0 \in (0, \varepsilon]$. From assumptions

$$F(t-\varepsilon_0) < F(t) = 1-\alpha < F(t+\varepsilon_0),$$

thus

$$F_n^*(t-\varepsilon_0) \to F(t-\varepsilon_0) < 1-\alpha$$

and

$$F_n^*(t+\varepsilon_0) \to F(t+\varepsilon_0) > 1-\alpha,$$

thus for sufficiently large n

$$F_n^*(t-\varepsilon_0) \leq 1-\alpha \text{ and } F_n^*(t+\varepsilon_0) \geq 1-\alpha$$

and hence $|(F_n^*)^{-1}(1-\alpha) - F^{-1}(1-\alpha)| \leq \varepsilon$. As ε is arbitrary, the first part is proved.

The second part *ii*) follows from *i*). We have that

$$\begin{aligned} &P(T_n(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{Z}_n) \leq (F_n^*)^{-1}(1-\alpha)) \\ &\quad = P(T_n(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{Z}_n) - (F_n^*)^{-1}(1-\alpha) + F^{-1}(1-\alpha) \leq F^{-1}(1-\alpha)) \to 1-\alpha \end{aligned}$$

as

$$P(T_n(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{Z}_n) - (F_n^*)^{-1}(1-\alpha) + F^{-1}(1-\alpha) \leq t) \to F(t)$$

from Slutsky's theorem. □

Remark 2.2.7. *Similarly, a two-sided bootstrap confidence interval takes the form*

$$P(L \leq T_n(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{Z}_n) \leq R) \to 1-\alpha,$$

where $L = (F_n^*)^{-1}(\alpha_L)$, $R = (F_n^*)^{-1}(1-\alpha_R)$ *and* $\alpha_L + \alpha_R = \alpha$.

Lets set $T_n(\mathbf{X}_n^*, \mathbf{Y}_n^*, \mathbf{Z}_n^*) = 2nCMI(\hat{p}^*)$. From Lemmas 1.4.2 and 2.2.1 or 2.2.2 we have that $2nCMI(\hat{p})$ and $2nCMI(\hat{p}^*)$, where $\hat{p}^*$ is based on bootstrap, CR/bootstrap X or permutation sample, tend to χ^2_d, where $d = (|\mathcal{X}|-1)(|\mathcal{Y}|-1)|\mathcal{Z}|$ (the first statistic converges in distribution if $X \perp\!\!\!\perp Y|Z$, for the second distributional convergence is conditional given $(X_i, Y_i, Z_i)_{i=1}^n$ for almost all sequences of observations). Thus, from Lemma 2.2.6 we obtain that

$$P(2nCMI(\hat{p}) \leq (F_n^*)^{-1}(1-\alpha)) \to 1-\alpha,$$

where $(F_n^*)^{-1}(1-\alpha)$ is a random variable, which for a given sequence of observations is a quantile of $2nCMI(\hat{p}^*)$ at $1-\alpha$. This justifies asymptotic validity of the conditional independence test based on CMI statistics with a critical value constructed using resampled data. We reject the null hypothesis $H_0 : X \perp\!\!\!\perp Y|Z$ if the test statistic $2nCMI(\hat{p})$ is greater than $(F_n^*)^{-1}(1-\alpha)$ and if the null holds, the probability of rejecting H_0 asymptotically is equal to α. The constructed above test asymptotically behaves as the test with the same test statistics and the critical value equal to the proper quantile of the χ^2_d distribution. Note that the quantile $(F_n^*)^{-1}(1-\alpha)$ can be arbitrarily well approximated by choosing sufficiently many resampling samples and using empirical distribution corresponding to $F_n^*(\cdot)$.

2.2.3. Non-asymptotic approach

Below we discuss non-asymptotic approach which yields an exact level of confidence for tests based on it. This approach is valid for CR and permutation scenario as for these methods the conditional distribution of X^* is the same as of X given Z under hypothesis

of independence. For bootstrap CI and bootstrap X the distribution of resampled variable X^* is approximated by $\hat{p}(x|z)$ and thus it differs from $p(x|z)$.

Theorem 2.2.8. *Let $(X_i, Y_i, Z_i)_{i=1}^n$ be a sample and $(X_i^*, Y_i^*, Z_i^*)_{i=1}^n$ be a resampled sample obtained using CR scenario. If the null hypothesis $H_0 : X \perp\!\!\!\perp Y|Z$ holds, then*

$$P(T_n(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{Z}_n) > (F_n^*)^{-1}(1-\alpha)) \leq \alpha.$$

Here we provide a proof given in [8].

Proof. Using CR scenario we resample X in a way which provides $\mathbf{X}_n^*$ such that

$$\mathbf{X}_n^*|\mathbf{Y}_n = \mathbf{y}_n, \mathbf{Z}_n = \mathbf{z}_n \sim \mathbf{X}_n|\mathbf{Z}_n = \mathbf{z}_n$$

and thus as the null holds we have

$$\mathbf{X}_n^*|\mathbf{Y}_n = \mathbf{y}_n, \mathbf{Z}_n = \mathbf{z}_n \sim \mathbf{X}_n|\mathbf{Y}_n = \mathbf{y}_n, \mathbf{Z}_n = \mathbf{z}_n.$$

Using that we obtain that given $(\mathbf{Y}_n, \mathbf{Z}_n)$

$$T_n(\mathbf{X}_n^*, \mathbf{Y}_n^*, \mathbf{Z}_n^*) \sim T_n(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{Z}_n)$$

and thus with F_n^* defined in (2.14)

$$P(T_n(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{Z}_n) > (F_n^*)^{-1}(1-\alpha)|(\mathbf{Y}_n, \mathbf{Z}_n)) = (\mathbf{y}_n, \mathbf{z}_n)) \leq \alpha$$

as

$$P(T_n(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{Z}_n) \leq (F_n)^{-1}(1-\alpha)) = (\mathbf{y}_n, \mathbf{z}_n)) \geq 1-\alpha.$$

The inequality above holds also for unconditioned probability

$$P(T_n(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{Z}_n) > (F_n^*)^{-1}(1-\alpha)) \leq \alpha. \qquad \square$$

As the above theorem is given for a theoretical quantile $(F_n^*)^{-1}(1-\alpha)$, below we provide a similar inequality considering the finite number of resampled samples equal

to B. We omit the proof of the following theorem, as it is analogous (but easier) to the proof of Theorem 2.2.10 for permutations.

Theorem 2.2.9. *Let* $(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{Z}_n) = (X_i, Y_i, Z_i)_{i=1}^n$ *be a sample and* $(\mathbf{X}^*_{n,b}, \mathbf{Y}^*_{n,b}, \mathbf{Z}^*_{n,b}) = (X_i^*, Y_i^*, Z_i^*)_{i=1}^n$ *for* $b = 1, 2, \ldots, B$ *be resampled samples obtained using CR scenario. If the null hypothesis* $H_0 : X \perp\!\!\!\perp Y|Z$ *holds, then*

$$P\left(\frac{1 + \sum_{b=1}^B \mathbb{I}(T \leq T_b^*)}{1+B} \leq \alpha\right) \leq \alpha,$$

where $T = T(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{Z}_n)$ *and* $T_b^* = T(\mathbf{X}^*_{n,b}, \mathbf{Y}^*_{n,b}, \mathbf{Z}^*_{n,b})$.

Definition 2.2.1. The random variables $T_1, T_2, \ldots, T_s$ are exchangeable if their joint distribution is invariant under permutations of the components.

Theorem 2.2.10. *Let* $(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{Z}_n) = (X_i, Y_i, Z_i)_{i=1}^n$ *be a sample and* $(\mathbf{X}^*_{n,b}, \mathbf{Y}^*_{n,b}, \mathbf{Z}^*_{n,b}) = (X_i^*, Y_i^*, Z_i^*)_{i=1}^n$ *for* $b = 1, 2, \ldots, B$ *be resampled samples obtained using conditional permutation scenario. If the null hypothesis* $H_0 : X \perp\!\!\!\perp Y|Z$ *holds, then*

$$P\left(\frac{1 + \sum_{b=1}^B \mathbb{I}(T \leq T_b^*)}{1+B} \leq \alpha\right) \leq \alpha,$$

where $T = T(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{Z}_n)$ *and* $T_b^* = T(\mathbf{X}^*_{n,b}, \mathbf{Y}^*_{n,b}, \mathbf{Z}^*_{n,b})$.

Proof. We prove that $\mathbf{X}_n$ and $\mathbf{X}_n^*$ are exchangeable given $\mathbf{Z}_n = \mathbf{z}_n$. The proof that $\mathbf{X}_n, \mathbf{X}^*_{n,1}, \mathbf{X}^*_{n,2}, \ldots, \mathbf{X}^*_{n,B}$ are exchangeable is a straightforward extension as well as the fact that $(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{Z}_n), (\mathbf{X}^*_{n,1}, \mathbf{Y}_n, \mathbf{Z}_n), (\mathbf{X}^*_{n,2}, \mathbf{Y}_n, \mathbf{Z}_n), \ldots, (\mathbf{X}^*_{n,B}, \mathbf{Y}_n, \mathbf{Z}_n)$ are exchangeable.

We denote by π a function describing conditional permutation given $\mathbf{Z}_n$ applied to $\mathbf{X}_n$ resulting in $\mathbf{X}_n^*$. That transformation consists of permutations on the layers $\mathbf{Z}_n = z$ denoted by π_z for $z \in \mathcal{Z}$ and we use a notation $i_z \in \{i : Z_i = z\}$ to denote the indices of subsequent observations on the layer $\mathbf{Z}_n = z$. Consider $P(\mathbf{X}_n = \mathbf{x}_n, \mathbf{X}_n^* = \mathbf{x}_n^* | \mathbf{Z}_n = \mathbf{z}_n, \Pi = \pi)$. Note that this probability equals $P(\mathbf{X}_n = \mathbf{x}_n | \mathbf{Z}_n = \mathbf{z}_n, \Pi = \pi)$ if $\mathbf{x}_n^*$ is an image of $\mathbf{x}_n$ under transformation π and 0 otherwise. Note that if $\mathbf{x}_n^*$ is an image of $\mathbf{x}_n$ then for all $z \in \mathcal{Z}$ and for all $i_z \in \{i : Z_i = z\}$

$$x^*_{i_z} = x_{\pi_z(i_z)}.$$

In case when $\pi(\mathbf{x}_n) = \mathbf{x}_n^*$ we have

$$P(\mathbf{X}_n = \mathbf{x}_n, \mathbf{X}_n^* = \mathbf{x}_n^* | \mathbf{Z}_n = \mathbf{z}_n, \Pi = \pi) = P(\mathbf{X}_n = \mathbf{x}_n | \mathbf{Z}_n = \mathbf{z}_n, \Pi = \pi) \tag{2.16}$$

and

$$\begin{aligned} P(\mathbf{X}_n = \mathbf{x}_n | \mathbf{Z}_n = \mathbf{z}_n, \Pi = \pi) &= \prod_z \prod_{i_z} P(X_{i_z} = x_{i_z} | Z_{i_z} = z, \Pi = \pi) \\ &= \prod_z \prod_{i_z} P(X_{\pi_z(i_z)} = x_{i_z} | Z_{i_z} = z, \Pi = \pi) = P(\mathbf{X}_n = \mathbf{x}_n^* | \mathbf{Z}_n = \mathbf{z}_n, \Pi = \pi). \end{aligned}$$

We also have that

$$P(\mathbf{X}_n = \mathbf{x}_n^*, \mathbf{X}_n^* = \mathbf{x}_n | \mathbf{Z}_n = \mathbf{z}_n, \Pi = \pi) = P(\mathbf{X}_n = \mathbf{x}_n^* | \mathbf{Z}_n = \mathbf{z}_n, \Pi = \pi),$$

where the above equation follows from analogous reasoning as in (2.16) applied to π^{-1}. When $\pi(\mathbf{x}_n) \neq \mathbf{x}_n^*$, then

$$P(\mathbf{X}_n = \mathbf{x}_n, \mathbf{X}_n^* = \mathbf{x}_n^* | \mathbf{Z}_n = \mathbf{z}_n, \Pi = \pi) = P(\mathbf{X}_n = \mathbf{x}_n^*, \mathbf{X}_n^* = \mathbf{x}_n | \mathbf{Z}_n = \mathbf{z}_n, \Pi = \pi) = 0.$$

Thus

$$P(\mathbf{X}_n = \mathbf{x}_n, \mathbf{X}_n^* = \mathbf{x}_n^* | \mathbf{Z}_n = \mathbf{z}_n, \Pi = \pi) = P(\mathbf{X}_n = \mathbf{x}_n^*, \mathbf{X}_n^* = \mathbf{x}_n | \mathbf{Z}_n = \mathbf{z}_n, \Pi = \pi).$$

and as the above equation holds for all $\pi \in \Pi$, we obtain

$$P(\mathbf{X}_n = \mathbf{x}_n, \mathbf{X}_n^* = \mathbf{x}_n^* | \mathbf{Z}_n = \mathbf{z}_n) = P(\mathbf{X}_n = \mathbf{x}_n^*, \mathbf{X}_n^* = \mathbf{x}_n | \mathbf{Z}_n = \mathbf{z}_n).$$

As we have proven the exchangeability of the sample and resampled samples given $\mathbf{Z}_n$, the test statistics based on them are also exchangeable given $\mathbf{Z}_n$. That property also holds marginally without conditioning.

For exchangeable random variables $T, T_1^*, T_2^*, \ldots, T_B^*$ we have that for $i \in \{1, \ldots, B, B+1\}$

$$P\left(1 + \sum_{b=1}^{B} \mathbb{I}(T \leq T_b^*) = i\right) = \frac{1}{1+B}$$

as any order of $T, T_1^*, T_2^*, \ldots, T_B^*$ is equally probable. Thus

$$P\left(1 + \sum_{b=1}^{B} \mathbb{I}(T \leq T_b^*) \leq i\right) = \frac{i}{1+B}$$

and from that we obtain

$$P\left(\frac{1 + \sum_{b=1}^{B} \mathbb{I}(T \leq T_b^*)}{1+B} \leq \frac{i}{1+B}\right) = \frac{i}{1+B}.$$

For any $\alpha \in \left[\frac{i}{B+1}, \frac{i+1}{B+1}\right)$ and $\alpha \leq 1$ we obtain

$$P\left(\frac{1 + \sum_{b=1}^{B} \mathbb{I}(T \leq T_b^*)}{1+B} \leq \alpha\right) \leq \alpha. \tag{2.17}$$

In the considered case of conditional independence the exchangeability of $T, T_1^*, T_2^*, \ldots, T_B^*$ holds given $\mathbf{Z}_n$=$\mathbf{z}_n$, thus the last inequality (2.17) holds given $\mathbf{Z}_n = \mathbf{z}_n$. But as (2.17) holds conditionally, it also holds marginally. □

Remark 2.2.11. *Note that for observations which indices satisfy $i_z \in \{i : Z_i = z\}$, vectors $(X_{i_z}, Y_{i_z})_{i_z}$ and $(X_{i_z}^*, Y_{i_z})_{i_z}$ have the same distribution for any permutation of X on $Z = z$ as the variables X and Y are independent on $Z = z$ for all $z \in \mathcal{Z}$. Thus, the problem considered on a fixed layer of Z is analogous to unconditional permutations and unconditional independence of X and Y ([25], see Example 15.2.3 there). Permutations of X are independent on each layer of Z, thus Theorem 2.2.10 applies the result from [25] to all the layer of Z.*

2.3. Asymptotic behaviour of $\widehat{JMI}^*$

Analogously to $\widehat{CMI}^*$, we investigate now asymptotic behaviour of $\widehat{JMI}^*$. We show that it differs substantially from that of $\widehat{CMI}^*$.

We recall that JMI for probability vector p equals (cf. (1.22))

$$JMI(p) = \frac{1}{|S|} \sum_{i=1}^{|S|} \sum_{x,y,z_i} p(x,y,z_i) \log \frac{p(x,y|z_i)}{p(x|z_i)p(y|z_i)},$$

where z_i are coordinates of z, thus plug-in estimator based on resampled sample equals

$$\widehat{JMI}^* := JMI(\hat{p}^*),$$

where $\hat{p}^*$ is a probability vector based on a resampled sample.

Lemma 2.3.1. *Let $\hat{p}^*$ be a probability vector based on a CI bootstrap sample. For almost all sequences $(X_1, Y_1, Z_1), (X_2, Y_2, Z_2), \ldots$ and conditionally on $(X_i, Y_i, Z_i)_{i=1}^{\infty}$ we have*

$$\sqrt{n}\,(JMI(\hat{p}^*) - JMI(\hat{p}_{ci})) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

if $\sigma^2 > 0$, where $\sigma^2 = \mathrm{Var}_{p_{ci}}(D_{JMI}(p_{ci})(X, Y, Z))$.

Proof. First, using the formula for gradient of conditional mutual information D_{CMI} given in (2.10), we obtain that the gradient D_{JMI} of JMI equals

$$D_{JMI}(p)(x, y, z) = \frac{1}{|S|}\sum_{i=1}^{|S|} \log \frac{p(x, y|z_i)}{p(x|z_i)p(y|z_i)}$$

and the Hessian matrix H_{JMI} equals

$$(H_{JMI}(p))_{x,y,z}^{x',y',z'} = \frac{1}{|S|}\sum_{i=1}^{|S|}\Big(\frac{\mathbb{I}(x = x', y = y', z_i = z_i')}{p(x, y, z_i)} - \frac{\mathbb{I}(x = x', z_i = z_i')}{p(x, z_i)} - \frac{\mathbb{I}(y = y', z_i = z_i')}{p(y, z_i)} + \frac{\mathbb{I}(z_i = z_i')}{p(z_i)}\Big).$$

We use the expansion of JMI at $\hat{p}_{ci}$

$$JMI(\hat{p}^*) \;=\; JMI(\hat{p}_{ci}) \;+\; (\hat{p}^* \;-\; \hat{p}_{ci})' D_{JMI}(\hat{p}_{ci}) \;+\; \frac{1}{2}(\hat{p}^* \;-\; \hat{p}_{ci})' H_{JMI}(\xi)(\hat{p}^* \;-\; \hat{p}_{ci}),$$

where $\xi = (\xi_{x,y,z})_{x,y,z}$ and $\xi_{x,y,z}$ is a point between $\hat{p}^*(x, y, z)$ and $\hat{p}_{ci}(x, y, z)$. As for almost all sequences $\hat{p}^*(x, y, z) - \hat{p}_{ci}(x, y, z) \to 0$ in probability conditionally on the sample and $\hat{p}_{ci}(x, y, z) \to p_{ci}(x, y, z)$ a.s., we have $H_{JMI}(\xi) \to H_{JMI}(p_{ci})$ in probability, and we obtain, analogously as before

$$JMI(\hat{p}^*) = JMI(\hat{p}_{ci}) + (\hat{p}^* - \hat{p}_{ci})' D_{JMI}(\hat{p}_{ci}) + \frac{1}{2}(\hat{p}^* - \hat{p}_{ci})' H_{JMI}(p_{ci})(\hat{p}^* - \hat{p}_{ci}) + o_{p^*}(\|\hat{p}^* - \hat{p}_{ci}\|^2). \quad (2.18)$$

We note that $\sqrt{n}(\hat{p}^* - \hat{p}_{ci})' H_{JMI}(p_{ci})(\hat{p}^* - \hat{p}_{ci}) \to 0$ in probability from Slutsky's theorem, thus we have that conditionally

$$\sqrt{n}(JMI(\hat{p}^*) - JMI(\hat{p}_{ci})) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

if $\sigma^2 > 0$, where $\sigma^2 = \mathrm{Var}(D_{JMI}(p_{ci})(X, Y, Z))$ and $(X, Y, Z) \sim p_{ci}$. This claim can be justified as follows: from Lemma 2.1.1 we have that for almost all sequences of observations $(X_i, Y_i, Z_i)_{i=1}^{\infty}$ and conditionally on $(X_i, Y_i, Z_i)_{i=1}^{n}$ $\sqrt{n}(\hat{p}^* - \hat{p}_{ci}) \to \mathcal{N}(0, \Sigma)$, where

$$\Sigma_{x,y,z}^{x',y',z'} = \mathbb{I}(x = x', y = y', z = z') p_{ci}(x, y, z) - p_{ci}(x, y, z) p_{ci}(x', y', z').$$

Moreover, $D_{JMI}(\hat{p}_{ci}) \to D_{JMI}(p_{ci})$, hence

$$\sigma^2 = D_{JMI}(p_{ci})' \Sigma D_{JMI}(p_{ci}) = \mathrm{Var}_{p_{ci}}(D_{JMI}(p_{ci})(X, Y, Z)). \qquad \square$$

Remark 2.3.2. *If the CI bootstrap sampling scenario of $\hat{p}^*$ in Lemma 2.2.1 is replaced by CR/bootstrap X, then the asymptotic variance also changes and equals*

$$\sigma^2 = D_{JMI}(p_{ci})' \Sigma D_{JMI}(p_{ci}),$$

where Σ is the asymptotic covariance matrix in Lemma 2.1.2. Moreover,

$$\begin{aligned}\sigma^2 &= \mathbb{E}(\mathbb{E}(D_{JMI}(p_{ci})(X, Y, Z)|X, Y, Z))^2 - \mathbb{E}(\mathbb{E}(D_{JMI}(p_{ci})(X, Y, Z)|Y, Z))^2 \\ &= \mathbb{E}(D_{JMI}(p_{ci})(X, Y, Z))^2 - \mathbb{E}(\mathbb{E}(D_{JMI}(p_{ci})(X, Y, Z)|Y, Z))^2.\end{aligned}$$

The formula above follows from the fact that

$$\sigma^2 = \sum_{x,y,z} \sum_{x',y',z'} (D_{JMI}(p_{ci}))_{x,y,z} (D_{JMI}(p_{ci}))_{x',y',z'} \mathbb{I}(z = z') \Big(\mathbb{I}(x = x', y = y') p(x|z) p(y, z) - \mathbb{I}(y = y') p(x|z) p(x'|z') p(y, z) \Big)$$

and e.g. second term equals

$$\sum_{x,y,z} \sum_{x',y',z'} \mathbb{I}(y = y', z = z') p(x|z) p(x'|z') p(y, z) (D_{JMI}(p_{ci}))_{x,y,z} (D_{JMI}(p_{ci}))_{x',y',z'}$$

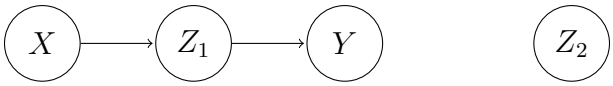


Figure 2.1: Graphical representation of a dependence structure used in Example 2.3.3. An arrow represents dependence. The variables X, Z_1 and Y form a Markov chain and Z_2 is independent of (X, Z_1, Y).

$$= \sum_{y,z} p(y,z) \sum_{x,x'} p(x|z)(D_{JMI}(p_{ci}))_{x,y,z} p(x'|z)(D_{JMI}(p_{ci}))_{x',y,z}$$
$$= \sum_{y,z} p(y,z) \left(\sum_{x} p(x|y,z)(D_{JMI}(p_{ci}))_{x,y,z} \right)^2 .$$

Similarly for the situation, in which $\hat{p}^*$ is obtained by permutations (see Lemma 2.1.5) we have

$$\sigma^2 = \mathbb{E}(\mathbb{E}(D_{JMI}(p_{ci})(X,Y,Z)|X,Y,Z))^2 - \mathbb{E}(\mathbb{E}(D_{JMI}(p_{ci})(X,Y,Z)|X,Z))^2$$
$$- \mathbb{E}(\mathbb{E}(D_{JMI}(p_{ci})(X,Y,Z)|Y,Z))^2 + \mathbb{E}(\mathbb{E}(D_{JMI}(p_{ci})(X,Y,Z)|Z))^2.$$

Note that in view of Lemma 2.1.7 the asymptotic variances of $JMI(\hat{p}^*)$ in the three resampling scenarios are also ordered.

Below we give an example of a distribution such that asymptotic variances σ^2 for $\widehat{JMI}^*$ differ according to the method of subsampling.

Example 2.3.3. *Similarly as in Lemma 2.1.7, we will denote by $\sigma^2_{(1)}$, $\sigma^2_{(2)}$, $\sigma^2_{(3)}$ asymptotic variances obtained through using bootstrap CI, CR/bootstrap X and permutation method, respectively. Let consider a distribution with a dependence structure as in Figure 2.1. Thus the following independencies holds: $X \perp\!\!\!\perp Y|Z_1$, $X \perp\!\!\!\perp Y|(Z_1, Z_2)$, $(X, Y, Z_1) \perp\!\!\!\perp Z_2$ and the joint distribution can be written as $p(x, y, z_1, z_2) = p_{ci}(x, y, z_1, z_2) = p(x)p(z_1|x)p(y|z_1)p(z_2)$. Hence the gradient of JMI equals*

$$D_{JMI}(p_{ci})(x,y,z) = \frac{1}{2}\sum_{i=1}^{2} \log \frac{p(x,y|z_i)}{p(x|z_i)p(y|z_i)} = \frac{1}{2}\log \frac{p(x,y|z_2)}{p(x|z_2)p(y|z_2)} = \frac{1}{2}\log \frac{p(x,y)}{p(x)p(y)},$$

where the second equality holds as $X \perp\!\!\!\perp Y$ given Z_1 and the third as $(X, Y) \perp\!\!\!\perp Z_2$. For simplicity of calculations we assume that all variables are binary and their marginal

distributions are $Bern(1/2)$*, and* $p(x, y, z_1, z_2) > 0$ *for all possible values of* (x, y, z_1, z_2)*. Thus*

$$D_{JMI}(p_{ci})(x, y, z) = \frac{1}{2}\log(4p(x, y)).$$

Next we compute the variance $\sigma^2_{(1)} = D_{JMI}(p_{ci})'\Sigma_{(1)}D_{JMI}(p_{ci})$*:*

$$\begin{aligned}\sigma^2_{(1)} = \frac{1}{4}\sum_{x,y,z}\sum_{x',y',z'}\log(4p(x, y))\log(4p(x', y'))\big(\mathbb{I}(x = x', y = y', z = z')p(x|z)p(y|z)p(z)\\ - p(x|z)p(y|z)p(z)p(x'|z')p(y'|z')p(z')\big) = \frac{1}{4}\sum_{x,y}p(x, y)\log^2(4p(x, y))\\ - \frac{1}{4}\left(\sum_{x,y}p(x, y)\log(4p(x, y))\right)^2\end{aligned}$$

and $\sigma^2_{(2)} = D_{JMI}(p_{ci})'\Sigma_{(2)}D_{JMI}(p_{ci})$*:*

$$\begin{aligned}\sigma^2_{(2)} = \frac{1}{4}\sum_{x,y,z}\sum_{x',y',z'}\log(4p(x, y))\log(4p(x', y'))\big(\mathbb{I}(x = x', y = y', z = z')p(x|z)p(y|z)p(z)\\ - \mathbb{I}(y = y', z = z')p(x|z)p(y|z)p(z)p(x'|z')p(y'|z')p(z')\big) = \frac{1}{4}\sum_{x,y}p(x, y)\log^2(4p(x, y))\\ - \frac{1}{4}\sum_{y,z_1}p(y, z_1)\left(\sum_{x}p(x|z_1)\log(4p(x, y))\right)^2.\end{aligned}$$

From Jensen's inequality we obtain

$$\begin{aligned}\sum_{y,z_1}p(y, z_1)\left(\sum_{x}p(x|z_1)\log(4p(x, y))\right)^2 \geq \left(\sum_{y,z_1}p(y, z_1)\sum_{x}p(x|z_1)\log(4p(x, y))\right)^2\\ = \left(\sum_{x,y,z_1}p(x, y, z_1)\log(4p(x, y))\right)^2 = \left(\sum_{x,y}p(x, y)\log(4p(x, y))\right)^2\end{aligned}$$

and equality hols if and only if $\sum_x p(x|z_1)\log(4p(x, y))$ *does not depend on* y *or* z_1*. Thus* $\sigma^2_{(2)} \leq \sigma^2_{(1)}$ *and equality holds if and only if (using* $p(z_1) = 1/2$*)*

$$\sum_{x}p(x, z_1)\log p(x, y) \equiv C.$$

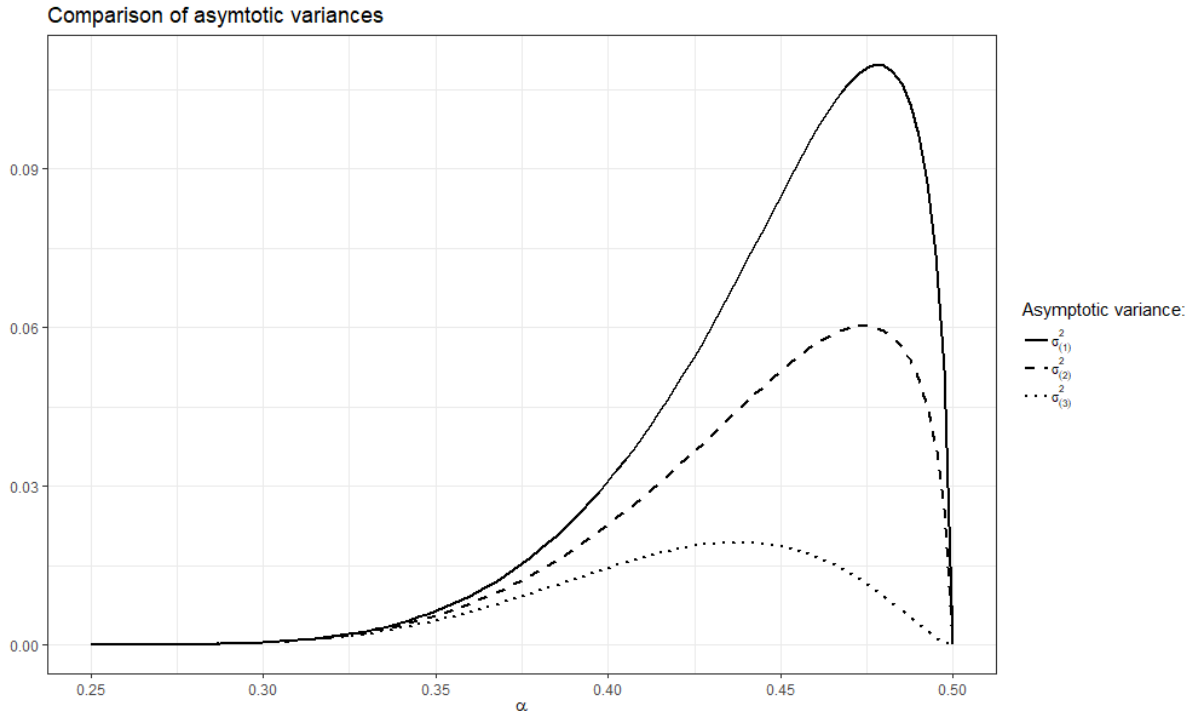


Figure 2.2: Comparison of asymptotic variances of $\widehat{JMI}^*$ for the distribution described in Example 2.3.3. The parameter α controls the strength of dependence (if $\alpha = 1/4$ then the variables are independent).

Writing the above equation for $z_1 = 0$ *and* $z_1 = 1$ *and subtracting the resulting equations we have*

$$0 = \sum_x (P(X = x, Z_1 = 0) - P(X = x, Z_1 = 1)) \log p(x, y). \tag{2.19}$$

Using $P(X = 0) = P(X = 1) = 1/2$ *and* $P(Z_1 = 0) = P(Z_1 = 1) = 1/2$*, we have that joint distribution of* (X, Z_1) *equals*

(X, Z_1)	$p(\cdot, 0)$	$p(\cdot, 1)$
$p(0, \cdot)$	α	$1/2 - \alpha$
$p(1, \cdot)$	$1/2 - \alpha$	α

for a certain $\alpha > 0$*. Then from* (2.19) *we obtain*

$$0 = \sum_x \left(2\alpha - \frac{1}{2}\right)(-1)^x \log p(x, y) = \left(2\alpha - \frac{1}{2}\right) \sum_x (-1)^x \log p(x, y),$$

hence $\alpha = 1/4$ *and* $X \perp\!\!\!\perp Z_1$ *or* $P(X = 0, Y = y) = P(X = 1, Y = y)$ *for* $y = 0, 1$*. Thus*

if X and Z_1, and X and Y (which follows from the second condition) are not independent in the considered model, then $\sigma^2_{(1)} > \sigma^2_{(2)}$ and as $\sigma^2_{(2)} \geq \sigma^2_{(3)}$, we also have $\sigma^2_{(1)} > \sigma^2_{(3)}$.

In Figure 2.2 we show how the values of $\sigma^2_{(1)}$, $\sigma^2_{(2)}$ and $\sigma^2_{(3)}$ vary with respect to α. In that example we additionally assume that (X, Z_1) and (Z_1, Y) have the same distribution. When $\alpha = 1/4$, all variables are independent and in that case $\sigma^2_{(1)} = \sigma^2_{(2)} = \sigma^2_{(3)} = 0$. For $\alpha \in (1/4, 1/2)$ the dependence between variables is positive and for that scenario we see that $\sigma^2_{(1)} > \sigma^2_{(2)} > \sigma^2_{(3)}$. For $\alpha = 0.5$, Z_1 and Y depend on X and Z_1 respectively in a deterministic way, thus some $p(x, y, z_1, z_2) = 0$ and thus this case is not covered by the calculations above.

Remark 2.3.4. *In Lemma 2.3.1 if $\sigma^2 = 0$, we consider a second-order term in the expansion* (2.18)

$$2n(JMI(\hat{p}^*) - JMI(\hat{p}_{ci})) = 2n(\hat{p}^* - \hat{p}_{ci})' D_{JMI}(\hat{p}_{ci}) + \sqrt{n}(\hat{p}^* - \hat{p}_{ci})' H_{JMI}(p_{ci}) \sqrt{n}(\hat{p}^* - \hat{p}_{ci}) + o_{p^*}(n \left\| \hat{p}^* - \hat{p}_{ci} \right\|^2).$$

In that case, in contrast to CMI in Lemma 2.2.1, where $CMI(\hat{p}_{ci}) = 0$ and $D_{CMI}(\hat{p}_{ci}) = 0$, we show numerically in Example 2.3.5 that it does not necessarily hold that $JMI(\hat{p}_{ci}) = 0$ or $2n(\hat{p}^* - \hat{p}_{ci})' D_{JMI}(\hat{p}_{ci}) \to 0$.

Example 2.3.5. *We use simulations involving two models to show that indeed as stated in Remark 2.3.4 the convergence*

$$2n(\hat{p}^* - \hat{p}_{ci})' D_{JMI}(\hat{p}_{ci}) \to 0$$

in probability does not hold in general. We recall that from the expansion (2.18) *for the case when $\sigma^2 = \mathrm{Var}(D_{JMI}(p_{ci})(X, Y, Z)) = 0$ we have that*

$$2n(JMI(\hat{p}^*) - JMI(\hat{p}_{ci})) \approx 2n(\hat{p}^* - \hat{p}_{ci})' D_{JMI}(\hat{p}_{ci}) + n(\hat{p}^* - \hat{p}_{ci})' H_{JMI}(p_{ci})(\hat{p}^* - \hat{p}_{ci}), \tag{2.20}$$

where $\approx$ means that both sides differ by $o_{p^*}(1)$. By $L(\hat{p}_{ci})$ and $R(\hat{p}_{ci})$ we will denote the left-hand side and the right-hand side of the approximate equation above and by $T_1(\hat{p}_{ci})$ and $T_2(\hat{p}_{ci})$ we will denote two consecutive terms of $R(\hat{p}_{ci})$.

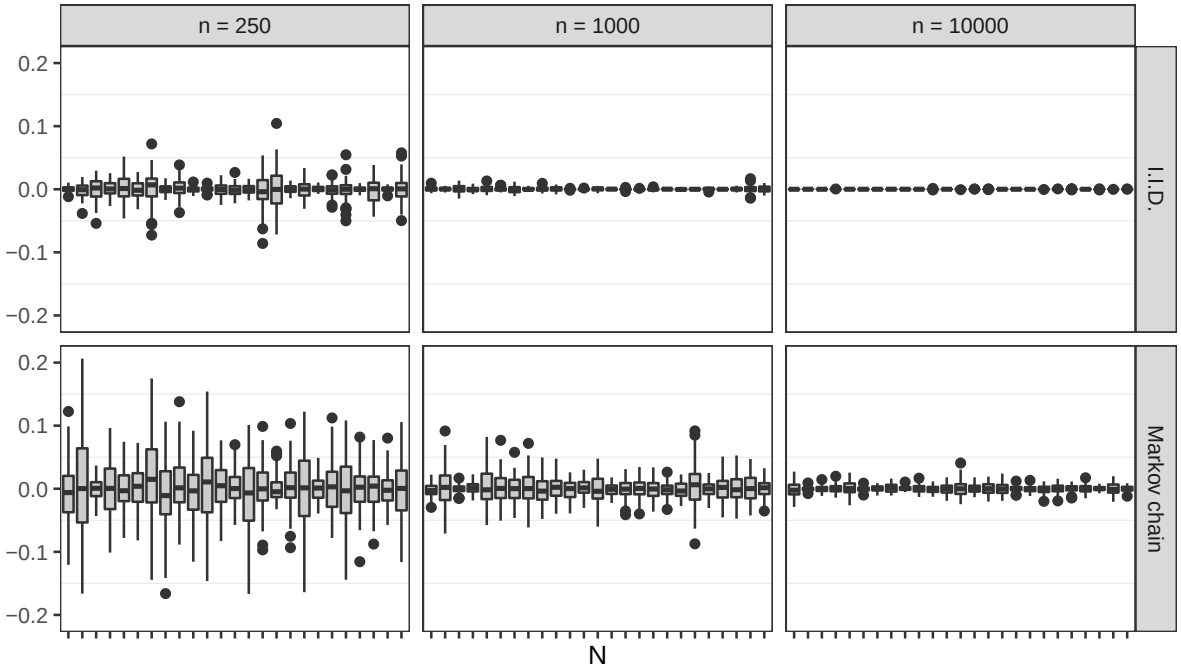


Figure 2.3: Boxplots of the statistic $T_1(\hat{p}_{ci})/(2\sqrt{n}) = \sqrt{n}(\hat{p}^* - \hat{p}_{ci})' D_{JMI}(\hat{p}_{ci})$ based on $B = 100$ subsamples for $N = 25$ samples (x-axis) for increasing sample sizes n and for the models I.I.D. and Markov chain.

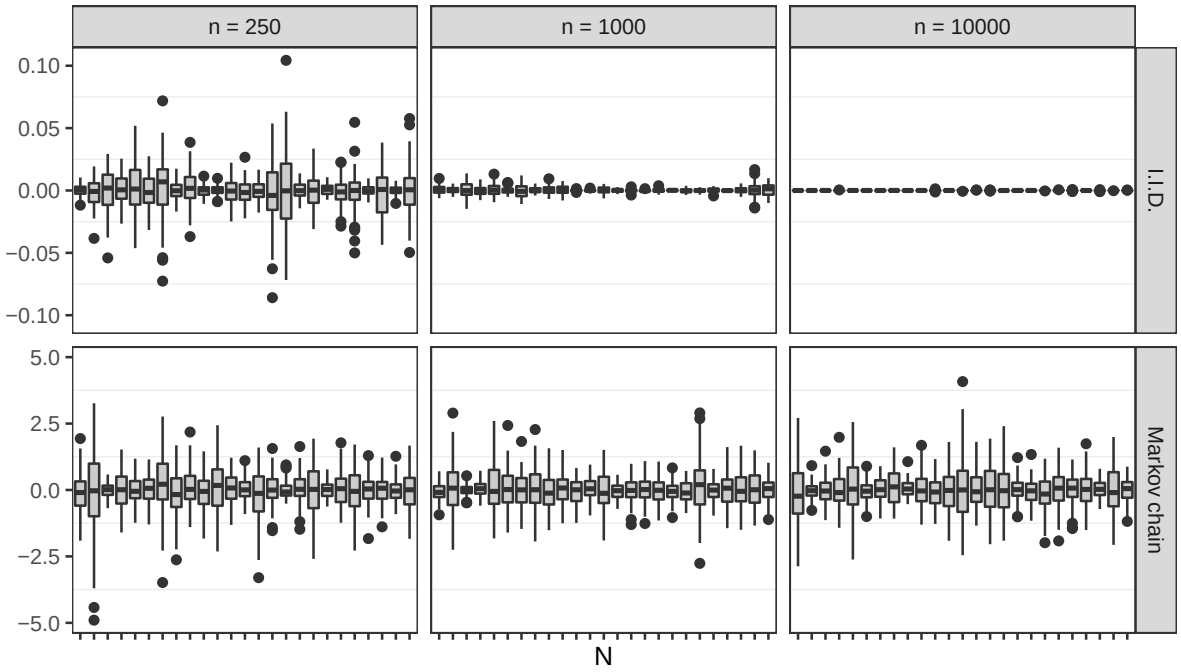


Figure 2.4: Boxplots of the statistic $T_1(\hat{p}_{ci})/2 = n(\hat{p}^* - \hat{p}_{ci})' D_{JMI}(\hat{p}_{ci})$ based on $B = 100$ subsamples for $N = 25$ samples (x-axis) for increasing sample sizes n and for the models I.I.D. and Markov chain.

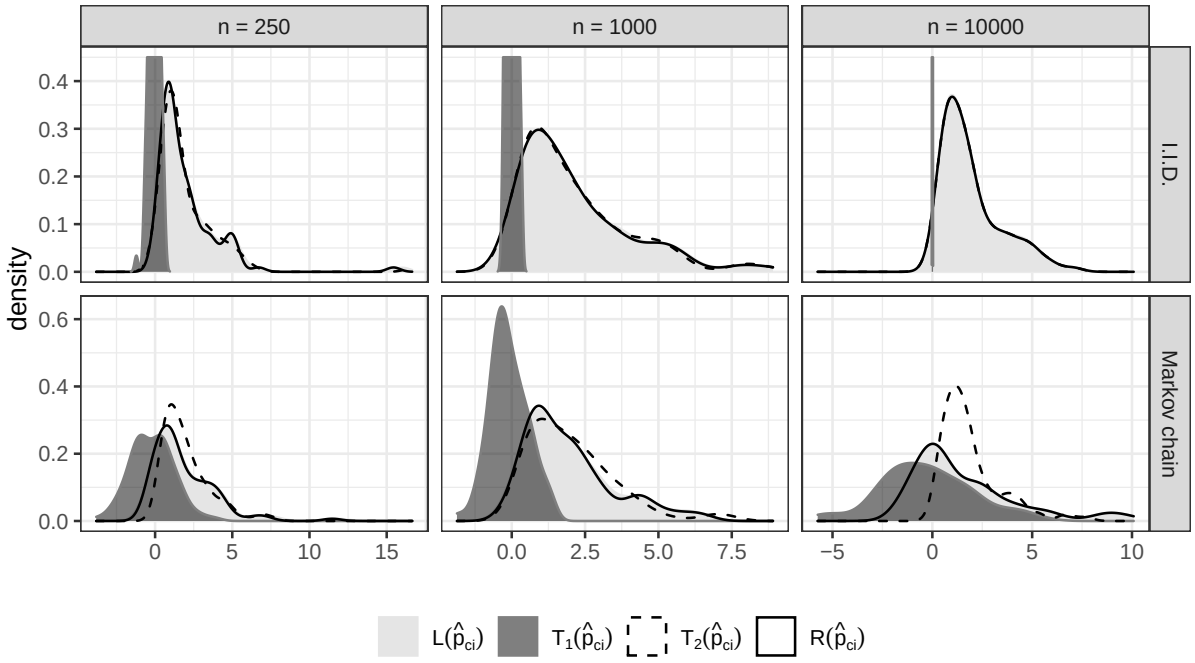


Figure 2.5: Comparison of distributions of $L(\hat{p}_{ci})$, $R(\hat{p}_{ci})$, $T_1(\hat{p}_{ci})$ and $T_2(\hat{p}_{ci})$ (cf. equation (2.20) and the notation introduced below the equation) based on $B = 100$ subsamples for one sample from I.I.D. and Markov chain models for increasing sample sizes n. The top plots are cut at the level $y = 0.045$.

The first considered model consists of five i.i.d. random variables X, Y and Z_i for $i = 1, 2, 3$ with Bernoulli distribution with probability of success equal to 0.5 *(model **I.I.D.**). It can be proved (we omit the proof) that in this model we have $T_1(\hat{p}_{ci}) \to 0$. In the second model the variables form a Markov chain $X = Z_0$, Z_1, Z_2, Z_3, $Y = Z_4$, hence they satisfy the Markov property (model **Markov chain**). We have that $Z_0 \sim Bern(0.5)$ and then for $i = 1, 2, 3, 4$*

$$P(Z_i = 1|Z_{i-1} = 1) = 1 - P(Z_i = 0|Z_{i-1} = 1) = 0.8,$$
$$P(Z_i = 1|Z_{i-1} = 0) = 1 - P(Z_i = 0|Z_{i-1} = 0) = 0.2.$$

We provide numerical evidence that the term $T_1(\hat{p}_{ci})$ behaves differently in these two models. The simulations are carried out as follows: for a given sample size n and a model with a vector of probabilities $(p(x, y, z))_{x,y,z}$ we sample $N = 25$ times $(X_i, Y_i, Z_i)_{i=1}^n$ and obtain $\hat{p}$. Then for each $\hat{p}$ we use CI bootstrap to get $B = 100$ subsamples $(X_i^*, Y_i^*, Z_i^*)_{i=1}^n$ and consequently $\hat{p}^*$. In Figure 2.3 the behaviour of $T_1(\hat{p}_{ci})/(2\sqrt{n}) = \sqrt{n}(\hat{p}^* - \hat{p}_{ci})' D_{JMI}(\hat{p}_{ci})$ is compared for the two models for $n = 250, 1000, 10000$. We see that in the both models

$T_1(\hat{p}_{ci})/(2\sqrt{n})$ seems to converge to 0 as $n \to \infty$ (and thus $\sqrt{n}(JMI(\hat{p}^*) - JMI(\hat{p}_{ci}))$ also converges to 0), thus we need the second order expansion to obtain limiting non-degenerate distribution. In Figure 2.4 we check whether the term $T_1(\hat{p}_{ci})/2 = n(\hat{p}^* - \hat{p}_{ci})' D_{JMI}(\hat{p}_{ci})$ vanishes as n grows. We see that although for the I.I.D. model it seems to be the case, for the Markov chain model we do not observe such a behaviour. Note that in Figure 2.4 the range of values of $T_1(\hat{p}_{ci})/2$ differs for the considered models and is about 50 times smaller for I.I.D. In Figure 2.5 we compare distributions of $L(\hat{p}_{ci})$, $R(\hat{p}_{ci})$, $T_1(\hat{p}_{ci})$ and $T_2(\hat{p}_{ci})$ for one sample from each model. The approximation of L by R is satisfactory even for small sample sizes e.g. $n = 250$ as the light gray filling (L) and the solid line (R) match. In the top three panels we see that as n grows, the distribution of T_2 (dashed line) becomes more similar to the distributions of L and R and the distribution of T_1 (dark gray filling) tends to $P(T_1 = 0) = 1$. In the bottom three panels the behaviour of T_1 differs, as it seems to have non-degenerate distribution at the limit and thus distribution at T_2 differs from that of L or R as $R = T_1 + T_2$.

Chapter 3

Simulations

3.1. Conditional independence testing

In order to depict the asymptotic behaviour of the criteria presented in previous chapters and to show their application in testing hypotheses of conditional independence, we will run simulations described below.

3.1.1. Models M1 and M2: description

The first type of a model we investigate is generative tree model shown in the left and the right panel of Figure 3.1. The models will be called M1 and M2, respectively. To ease the notation, the number of covariates $Z = Z_S = (Z_1, Z_2, \ldots, Z_{|S|})$ will be denoted by $s = |S|$. We will describe the models in detail by giving the formula for joint distribution of $(X, Y, Z_1, Z_2, \ldots, Z_s)$.

Figure 3.1: Generative tree models under consideration. The models in the left and right panel are called M1 and M2, respectively.

Joint probability mass function $p(x, y, z_1, z_2, \ldots, z_s)$ in the model M1 can be written as follows:

$$p(x, y, z_1, z_2, \ldots, z_s) = p(y)p(z_1, z_2, \ldots, z_s|y)p(x|z_1),$$

thus it is sufficient to define p.m.f. of Y and conditional p.m.f. of Z given Y and X given Z_1. First, Y is a Bernoulli random variable with probability of success $P(Y = 1) = 0.5$.

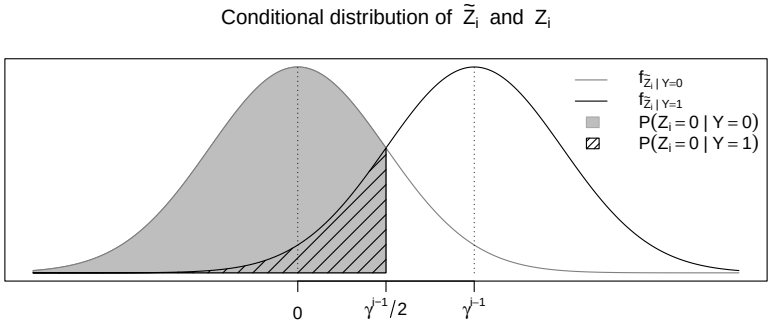


Figure 3.2: The conditional distribution of $\tilde{Z}_i|Y = 0$ and $\tilde{Z}_i|Y = 1$, and the conditional probability of $Z_i = 0$ given Y.

We define auxiliary continuous variable $\tilde{Z}$ such that $\tilde{Z}|Y = y$ follows multivariate normal distribution $\mathcal{N}_s(y\gamma_s, \Sigma)$, where $\gamma_s = (1, \gamma, \ldots, \gamma^{s-1})$ and $\gamma \in (0, 1]$, and covariance matrix Σ has an autoregression structure with elements equal to $\sigma_{i,j} = \rho^{|i-j|}$, $\rho \in [0, 1)$, $i, j \in \{1, 2, \ldots, s\}$ (if $\rho = 0$ then Σ is the identity matrix). In particular $\tilde{Z}_i|Y = y \sim \mathcal{N}(y\gamma^{i-1}, 1)$ for $i \in \{1, 2, \ldots, s\}$. Model $(X, Y, \tilde{Z}_1, \tilde{Z}_1, \ldots, \tilde{Z}_s)$ was considered in [23].

In order to obtain discrete variables Z from continuous $\tilde{Z}$ we define the distribution of Z given $Y = y$ in the following way

$$\begin{aligned} P(Z_1 = z_1, Z_2 = z_2, \ldots, Z_s = z_s \mid Y = y) \\ = P\bigg((-1)^{z_1}\tilde{Z}_1 \le \frac{(-1)^{z_1}}{2}, (-1)^{z_2}\tilde{Z}_2 \le \frac{(-1)^{z_2}\gamma}{2}, \ldots, \\ (-1)^{z_s}\tilde{Z}_s \le \frac{(-1)^{z_s}\gamma^{s-1}}{2}\bigg| Y = y\bigg), \end{aligned}$$

where variables $Z_1, Z_2, \ldots, Z_s$ take values in $\{0, 1\}$. Thus Z_i takes value 0 or 1 depending on whether corresponding $\tilde{Z}_i$ is smaller or larger then the mean of $\tilde{Z}_i$ equal to $\gamma^{i-1}/2$. Namely, for $i \in \{1, 2, \ldots, s\}$ we have

$$\begin{aligned} P(Z_i = 0) &= P(Z_i = 0|Y = 0)P(Y = 0) + P(Z_i = 0|Y = 1)P(Y = 1) \\ &= P\left(\tilde{Z}_i \le \frac{\gamma^{i-1}}{2}\middle| Y = 0\right) P(Y = 0) + P\left(\tilde{Z}_i \le \frac{\gamma^{i-1}}{2}\middle| Y = 1\right) P(Y = 1) \\ &= P\left(\tilde{Z}_i \le \frac{\gamma^{i-1}}{2}\right) = \frac{1}{2}. \end{aligned}$$

This is illustrated in Figure 3.2. Note that for smaller values of γ, the densities of

$\tilde{Z}_i|Y=0$ and $\tilde{Z}_i|Y=1$ become closer to each other and thus the influence of Y on the distribution of $\tilde{Z}_i$ is weaker. Analogously, the distribution of $\tilde{X}$ given Z_1 follows normal distribution $\tilde{X}|Z_1=z_1 \sim \mathcal{N}(z_1,1)$ and X is determined by $\tilde{X}$ in the following way

$$P(X=0|Z_1=z_1) = P\left(\tilde{X} \leq \frac{1}{2}\Big| Z_1=z_1\right)$$
$$= \begin{cases} \Phi_{\mathcal{N}(0,1)}(\frac{1}{2}) & \text{if } z_1=0 \\ \Phi_{\mathcal{N}(0,1)}(-\frac{1}{2}) & \text{if } z_1=1 \end{cases} \approx \begin{cases} 0.69 & \text{if } z_1=0 \\ 0.31 & \text{if } z_1=1 \end{cases}$$

and $P(X=1|Z_1=z_1) = 1 - P(X=0|Z_1=z_1)$. Note, that the dependence structure of M1 implies that $X \perp\!\!\!\perp Y|Z$, but not $X \perp\!\!\!\perp Y$. Moreover, when $\gamma < 1$, then for larger i the variable Z_i is less influenced by Y, whereas for $\gamma = 1$ the dependence between Y and Z_i is the same regardless of i for $i \in \{1,2,\ldots,s\}$. In general Z_i are dependent given Y, except for $\rho = 0$, as in that case Σ is diagonal. The difference between M2 and M1 is that the term $p(x|z_1)$ is replaced by $p(x)$. There, $X \perp\!\!\!\perp (Y,Z)$, thus $p(x|z_1) = p(x)$ and joint distribution equals $p(y)p(z_1,z_2,\ldots,z_s|y)p(x)$, where $P(X=0) = P(X=1) = 0.5$.

We would like to underline the link between the models we consider and graphical models (see e.g. [7] Chapter 13 Graphical modeling, [33]). Separability in the presented graphs is equivalent to separability with respect to its skeleton (the undirected graph formed by removing directions of all the edges of the directed graph), thus we treat these models as undirected graphs. A distribution satisfies global Markov condition with respect to graph G (in which the nodes correspond to considered variables) if for any triple of disjoint sets of nodes such that the two first are separated by the third, one has conditional independence of two first sets of corresponding variables given the third set. For $\rho = 0$ this property holds for the distribution we described for graphs presented in Figure 3.1. If besides the global Markov property the other implication holds as well, we say that the distribution is faithful to the graph G, which means that all conditional independencies can be read from the graphical separation. The faithfulness implies that if we enlarge the conditioning set for conditionally independent variables, the conditional independence is preserved for all triples of disjoint subsets of variables, thus e.g.

$$X \perp\!\!\!\perp Y|Z_{S_1} \Rightarrow X \perp\!\!\!\perp Y|Z_{S_2} \text{ for all } S_1 \subseteq S_2 \subseteq S. \tag{3.1}$$

In simulations we consider distributions both faithful to some graph and such which are

Table 3.1: Numerically obtained values of asymptotic variance $\sigma^2 = D_{Crit}(p_{ci})'\Sigma D_{Crit}(p_{ci})$ with respect to resampling scenarios and criteria in models M1 and M2 for chosen parameters. $D_{Crit}(p_{ci})$ is determined by the criterion and Σ is determined by the resampling scenario.

Model	$Crit$	Boot. CI	CR/Boot. X	Perm.
Model M1 ($\gamma = 1$, $\rho = 0$, $s = 3$)	$JMI2$	0.01040	0.00906	0.00613
	$JMI3$	0.00501	0.00435	0.00345
	$SECMI2$	0.01804	0.01553	0
	$SECMI3$	0.00269	0.00222	0
Model M2 ($\gamma = 1$, $\rho = 0$, $s = 3$)	$JMI2$	0	0	0
	$JMI3$	0	0	0
	$SECMI2$	0	0	0
	$SECMI3$	0	0	0
Model M1 ($\gamma = 1$, $\rho = 0.7$, $s = 3$)	$JMI2$	0.00294	0.00258	0.00245
	$JMI3$	0.00044	0.00039	0.00037
	$SECMI2$	0.02297	0.01904	0.00860
	$SECMI3$	0.00336	0.00270	0.00090

not (in Section 3.1.6 we give an example, for which (3.1) does not hold). In Figure 3.1 the models for $\rho = 0$ and $\gamma > 0$ are shown and in that case the joint probability distributions described above are faithful. For $\rho \neq 0$ variables Z_i are dependent given Y (which is not indicated in Figure 3.1, as arrows between Z_i are absent).

3.1.2. Asymptotic convergence: numerical analysis

We compute σ^2 for each criterion and each resampling method numerically using the formula (see Section 2.3)

$$\sigma^2 = D_{Crit}(p_{ci})'\Sigma D_{Crit}(p_{ci}),$$

where Σ is an asymptotic covariance matrix for one of the resampling scenarios and D_{Crit} is a gradient of one of the criteria: $JMI2$, $JMI3$, $SECMI2$ ($CIFE$) and $SECMI3$. We use here notation $JMI2$ and $SECMI2$ instead of JMI and $SECMI$ to make the order of the expansion explicit. By $Crit = Crit(X, Y, Z_S)$ we denote one of the listed criteria and by $\widehat{Crit}$, $\widehat{Crit}^*$ and $\widehat{Crit}_{ci}$ we denote plug-in estimators of $Crit$ based on $\hat{p}$, $\hat{p}^*$ and $\hat{p}_{ci}$, respectively. The formula for σ^2 in case of $JMI2$ and bootstrap CI or CR/bootstrap X scenarios is given in the proof of Lemma 2.3.1 and Remark 2.3.2. In Table 3.1 we list numerically obtained values of σ^2 rounded to five decimal places for three distributions

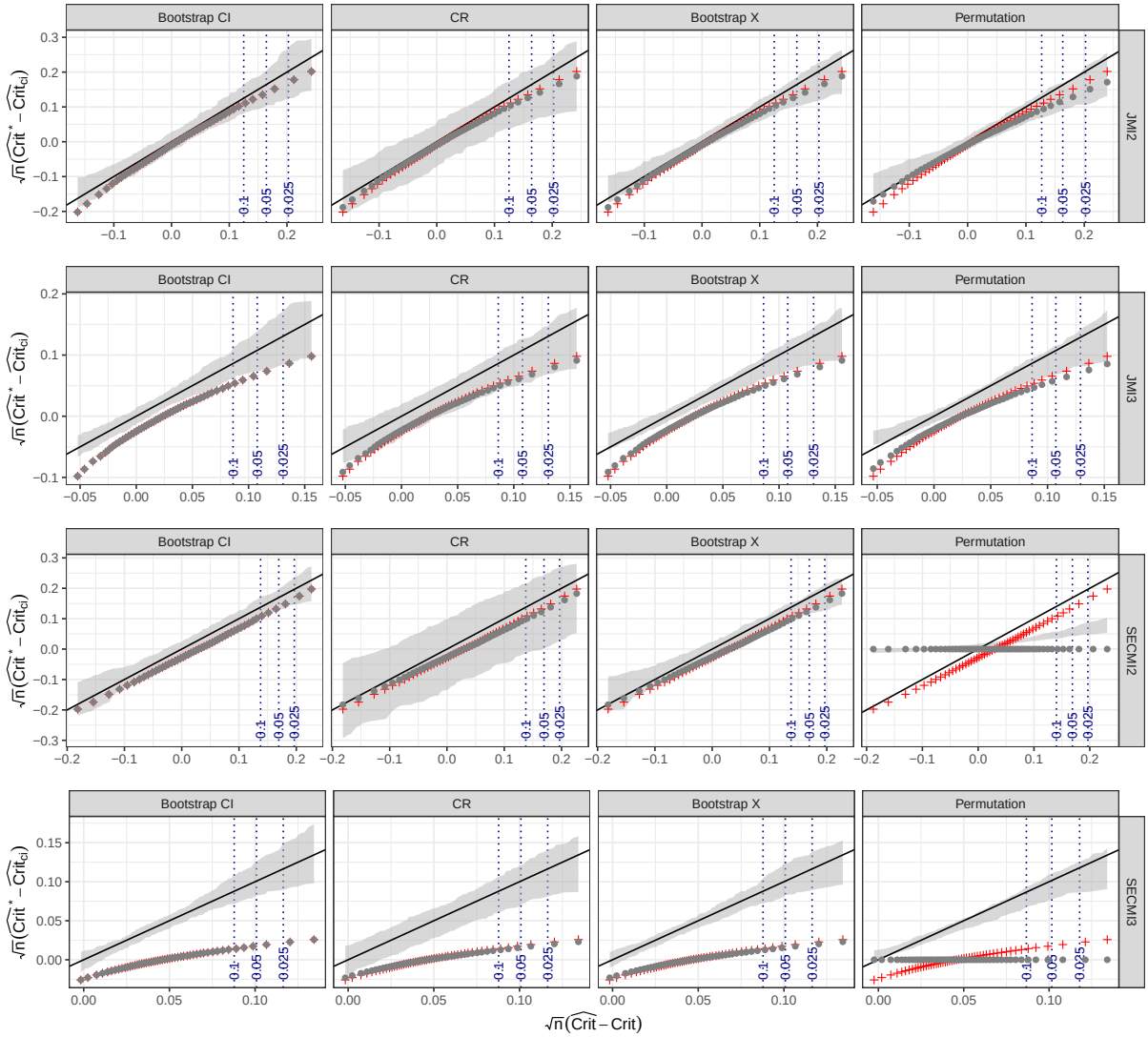


Figure 3.3: The comparison of the quantiles of $\sqrt{n}(\widehat{Crit} - Crit)$ and $\sqrt{n}(\widehat{Crit}^* - \widehat{Crit}_{ci})$ in model M1 ($\gamma = 1$, $\rho = 0$, $s = 3$). The gray ribbon denotes 90% interval of quantiles of $\sqrt{n}(\widehat{Crit}^* - \widehat{Crit}_{ci})$ and the gray dots denote the quantiles of its asymptotic distribution. The red crosses denote quantiles of the asymptotic distribution of $\sqrt{n}(\widehat{Crit} - Crit)$. A solid line correspods to $x = y$.

satisfying conditional independence of X and Y given Z, determined for models under consideration. In the first model for considered parameters (M1: $\gamma = 1$, $\rho = 0$, $s = 3$ and M1: $\gamma = 1$, $\rho = 0.7$, $s = 3$) random variables $\sqrt{n}(\widehat{Crit} - Crit)$ converge to normal distribution for all criteria, whereas in the second (M2: $\gamma = 1$, $\rho = 0$, $s = 3$) they converge to 0 and $2n(\widehat{Crit} - Crit)$ converges to quadratic form of a vector of normal variables, with accordance to the values of σ^2 shown in Table 3.1. Note that since asymptotic distributions of the criterion based on original samples and resampled samples for bootstrap CI coincide, the former asymptotic distribution is determined by the value of σ^2 given in

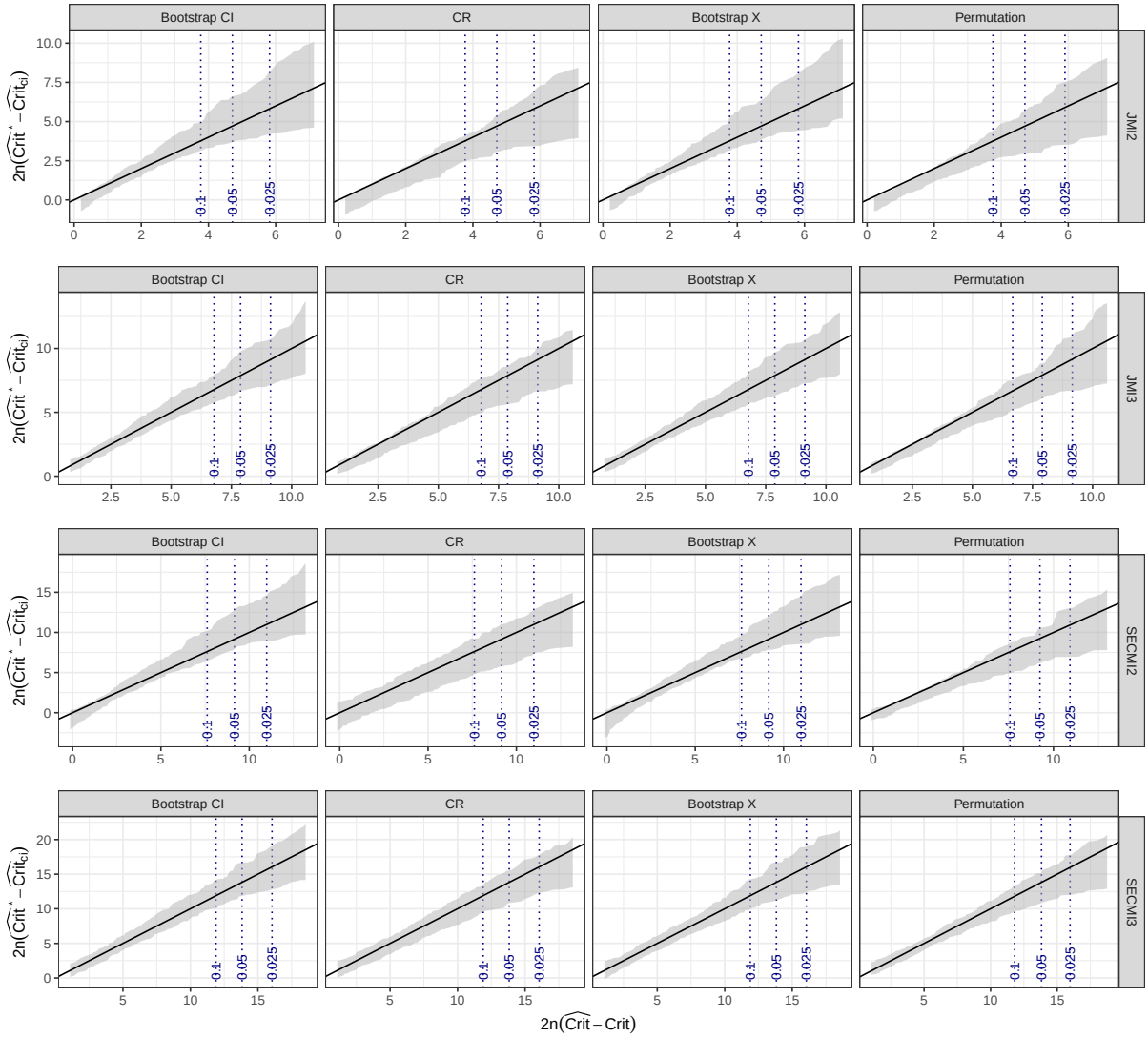


Figure 3.4: The comparison of the quantiles of $2n(\widehat{Crit} - Crit)$ and $2n(\widehat{Crit}^* - \widehat{Crit}_{ci})$ in model M2 ($\gamma = 1$, $\rho = 0$, $s = 3$). The gray ribbon denotes 90% interval of quantiles of $2n(\widehat{Crit}^* - \widehat{Crit}_{ci})$. A solid line correspods to $x = y$.

Boot. CI column (equal or not equal to 0). In almost all cases the asymptotic distribution of $\sqrt{n}(\widehat{Crit}^* - \widehat{Crit}_{ci})$ computed based on resampling scenarios is consistent with the asymptotic distribution of criteria estimated based on sample, as both variances equal 0 or are greater than 0 simultaneously. The exception is the result for permutation scenario for $SECMI2$ and $SECMI3$ in the model for the first set of parameters presented in Figure 3.3, in which the distribution of the criteria based on sample is normal, whereas for permuted samples apparently is not normal. *Thus, importantly, asymptotic law of criterion based on resampled samples may depend on resampling scheme applied.* In Figures 3.3 and 3.5 we show quantile-quantile plots comparing the distributions of $\sqrt{n}(\widehat{Crit}^* - \widehat{Crit}_{ci})$ and

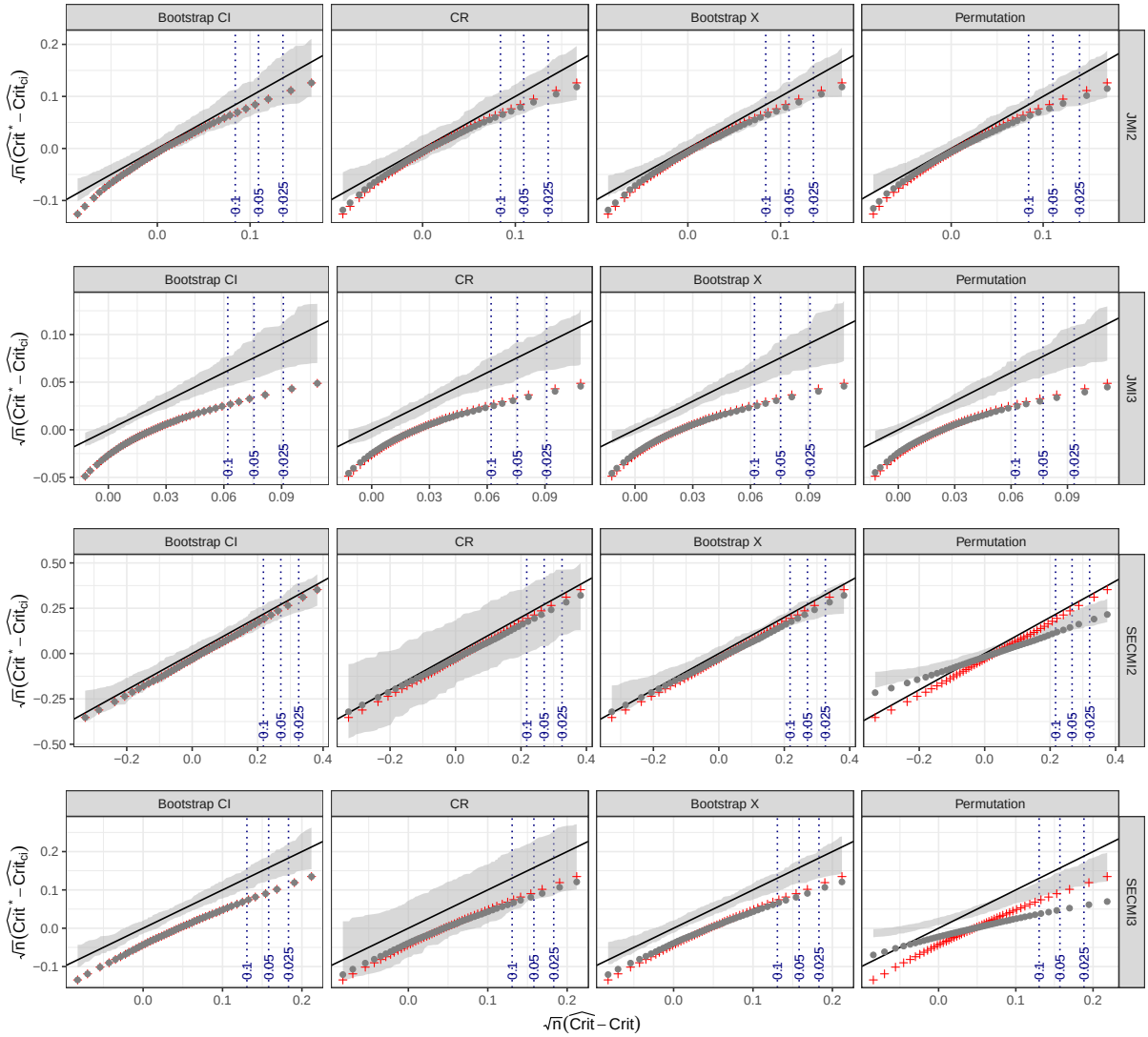


Figure 3.5: The comparison of the quantiles of $\sqrt{n}(\widehat{Crit} - Crit)$ and $\sqrt{n}(\widehat{Crit}^* - \widehat{Crit}_{ci})$ in model M1 ($\gamma = 1$, $\rho = 0.7$, $s = 3$) . The gray ribbon denotes 90% interval of quantiles of $\sqrt{n}(\widehat{Crit}^* - \widehat{Crit}_{ci})$ and the gray dots denote the quantiles of its asymptotic distribution. The red crosses denote quantiles of the asymptotic distribution of $\sqrt{n}(\widehat{Crit} - Crit)$. A solid line correspods to $x = y$.

$\sqrt{n}(\widehat{Crit} - Crit)$ (in Figure 3.4 we compare $2n(\widehat{Crit}^* - \widehat{Crit}_{ci})$ and $2n(\widehat{Crit} - Crit)$). The quantiles of $\sqrt{n}(\widehat{Crit} - Crit)$ are estimated based on 10000 samples consisting of $n = 5000$ observations and plotted on x-axis. In order to show the variability of estimation based on each sample obtained through resampling, the following experiment was run $N = 100$ times: we draw a sample of $n = 5000$ observations from the given distribution, then we draw $B = 50$ resampled samples from the original data and based on that we estimate its quantiles. The 5th and 95th percentiles of the estimated distribution of quantiles for resampling form a 'ribbon' plotted against the quantiles of sample distribution. Addi-

tionally, wherever possible, the quantiles of the theoretic asymptotic distribution of both $\sqrt{n}(\widehat{Crit}^* - \widehat{Crit}_{ci})$ (gray points) and $\sqrt{n}(\widehat{Crit} - Crit)$ (red crosses) are shown. In Figure 3.4 instead of multiplying centered criterion by $\sqrt{n}$, we multiply $\widehat{Crit}$ by $2n$ to obtain non-degenerate distribution. Vertical lines indicate 0.1, 0.05 and 0.025 quantiles. Figures 3.3-3.5 show that the accuracy of asymptotic approximation differs between criteria e.g. in Figure 3.5 for $SECMI2$ the quantiles of asymptotic distribution of the sample and based on resampling have similar values, whereas for the other criteria the approximation is not accurate. On the other hand, the resampling scenarios behave in a similar way, although the effect of smaller or vanishing asymptotic variance for permutation scenario can be observed (in presented examples the effect of a smaller variance for CR and bootstrap X scenarios is minimal). We note also that variability of quantiles of resampling distributions increases with the order of the criteria.

3.1.3. Comparison of testing procedures

In this section we describe testing procedures based on conditional mutual information and criteria and using asymptotic and exact approaches introduced in the previous chapters.

Tests based on $\widehat{CMI}$

The first test we consider is a standard asymptotic test of conditional independence based on conditional mutual information and on Theorem 1.4.2, which gives the asymptotic distribution of $\widehat{CMI}$ under the null hypothesis. In this approach the test statistic $\widehat{CMI}$ is compared with the quantiles of the χ^2_d distribution, where $d = (|\mathcal{X}| - 1)(|\mathcal{Y}| - 1)|\mathcal{Z}_S|$. We will call that test `asymptotic`. In the second test (`estimated df`) we estimate the number of degrees of freedom d, $d \in \mathbb{R}^+$ (equal to the expectation in case of χ^2 distribution) based on samples of $\widehat{CMI}^*$ in the following way: $d = \frac{1}{B}\sum_{b=1}^{B} \widehat{CMI}^*_b$, where $\widehat{CMI}^*_b$ denotes $\widehat{CMI}$ computed on bth resampled sample out of $B = 50$. This test for permutations was described e.g. in [41]. In the third test (`exact`) we compare the value of $\widehat{CMI}$ with the quantiles of $\widehat{CMI}^*$ i.e. the p-value is computed in the following way

$$\text{p-value} = \frac{1 + \sum_{b=1}^{B} \mathbb{I}(\widehat{CMI}^*_b > \widehat{CMI})}{B + 1}.$$

This approach is justified in Section 2.2.2, as for all presented resampling methods $\widehat{CMI}^*$ converges to the same distribution as $\widehat{CMI}$. Also, for CR another justification is given in [8] and for permutation scheme the justification is given in Section 2.2.3.

Tests based on criteria

We will use also other criteria as test statistics. By 'criteria' we refer to $JMI2$, $JMI3$, $SECMI2$ and $SECMI3$ (the term *criteria* will not include CMI in the following). As their asymptotic behaviour under the null hypothesis is dichotomous (the asymptotic distribution is either normal or a distribution of the quadratic form in normal variables, cf. Theorem 1.4.6), one of the considered testing procedures based on the criteria switches between two parametric distributions in order to account for this dichotomy (test `switch`). The switch is based on numerically obtained theoretical values of $\sigma^2 = D'_{Crit}\Sigma D_{Crit}$, where D_{Crit} is a gradient of $Crit$ computed at p and Σ is a covariance matrix defined in Theorem 1.4.6. Thus the test is not applicable in practice, as we do not know the joint distribution. We include this test in our analysis for comparative purposes only. If $\sigma^2 > 0$, then we compare $\widehat{Crit}$ with a quantile of $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$, where $\hat{\mu} = \frac{1}{B}\sum_{b=1}^{B}\widehat{Crit}_b^*$ and $\hat{\sigma}^2 = \frac{1}{B-1}\sum_{b=1}^{B}(\widehat{Crit}_b^* - \hat{\mu})^2$ and $\widehat{Crit}_b^*$ denotes $Crit$ computed for bth resampled sample. In case when $\sigma^2 = 0$, $\widehat{Crit}$ has distribution of a quadratic form and we use a quantile of a distribution $\alpha\chi^2_d + \beta$, where the parameters $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}$ and $d \in \mathbb{R}^+$ are estimated based on $\widehat{Crit}^*$ values using the method of moments as in [44]. In case of the second test called `approximation` we consider simplified procedure. In that case regardless of the true asymptotic distribution we estimate the resampled distribution by fitting distribution $\alpha\chi^2_d+\beta$. It is justified by noting that, as we use tests for finite samples and for small to moderate number of observations, an asymptotic approximation might be not accurate. We choose the distribution of the quadratic form, because the finite sample distribution is skewed and thus fitted normal distribution might not approximate large quantiles properly. The last test is also called `exact` as for CMI, because the test is basically the same with CMI replaced by the criteria. Namely, the p-value is computed in the following way

$$\text{p-value} = \frac{1 + \sum_{b=1}^{B}\mathbb{I}(\widehat{Crit}_b^* > \widehat{Crit})}{B+1}.$$

Note that in test `exact` if the significance level and the number of resampled samples are fixed, then we can compute the empirical quantile of $\widehat{Crit}_b^*$ and use it to construct

rejection region, e.q. for $\alpha = 0.05$ and $B = 50$ we reject the null if

$$\frac{1 + \sum_{b=1}^{50} \mathbb{I}(\widehat{Crit}_b^* > \widehat{Crit})}{51} < 0.05.$$

By transforming the equation we obtain

$$\sum_{b=1}^{50} \mathbb{I}(\widehat{Crit}_b^* > \widehat{Crit}) < 1.55,$$

and hence we reject the null if

$$\sum_{b=1}^{50} \mathbb{I}(\widehat{Crit}_b^* \leq \widehat{Crit}) > 48.45.$$

Thus if $\widehat{Crit} \geq \widehat{Crit}_{(49)}^*$ then the null is rejected, where $\widehat{Crit}_{(k)}^*$ denotes kth order statistic. In order to compare it with previous tests we note that to estimate the number of degrees of freedom in `estimated df` for CMI, or three parameters of $\alpha\chi_d^2 + \beta$ in `approximation`, all values of $\widehat{Crit}_b^*$ are used and the quantile is computed based on parametric distribution, not just on one extreme order statistic. Therefore, the test based on semi-parametric approach might be more stable than the test `exact`, especially for small number of resampled samples B.

We note that test `exact` is justified for any criterion considered in case of CR and permutation scenario by results of Section 2.2.3. For bootstrap X and CI bootstrap scenarios this test has heuristic justification only and its properties will be checked numerically. The same concerns `approximation` test. We discuss this problem in the next section when distributions of pvalues of the tests based on $\widehat{Crit}^* - \widehat{Crit}_{ci}$ and $\widehat{Crit} - Crit$ are compared.

The experiments in Sections 3.1.4, 3.1.5 and 3.1.6 were run $N = 200$ times, thus for each given joint distribution $p(x, y, z_S)$, we sampled 200 samples and each sample was resampled $B = 50$ times. Thus each test, for which the size or power is estimated, was run 200 times. The number of resampled samples is purposefully chosen to be $B = 50$ to compare efficacy of test `exact` with proposed procedures in such a case. In view of [17] it seems appropriate when the significance level α equals 0.05. The authors analyse the number of permutations needed to properly estimate p-values for testing procedures

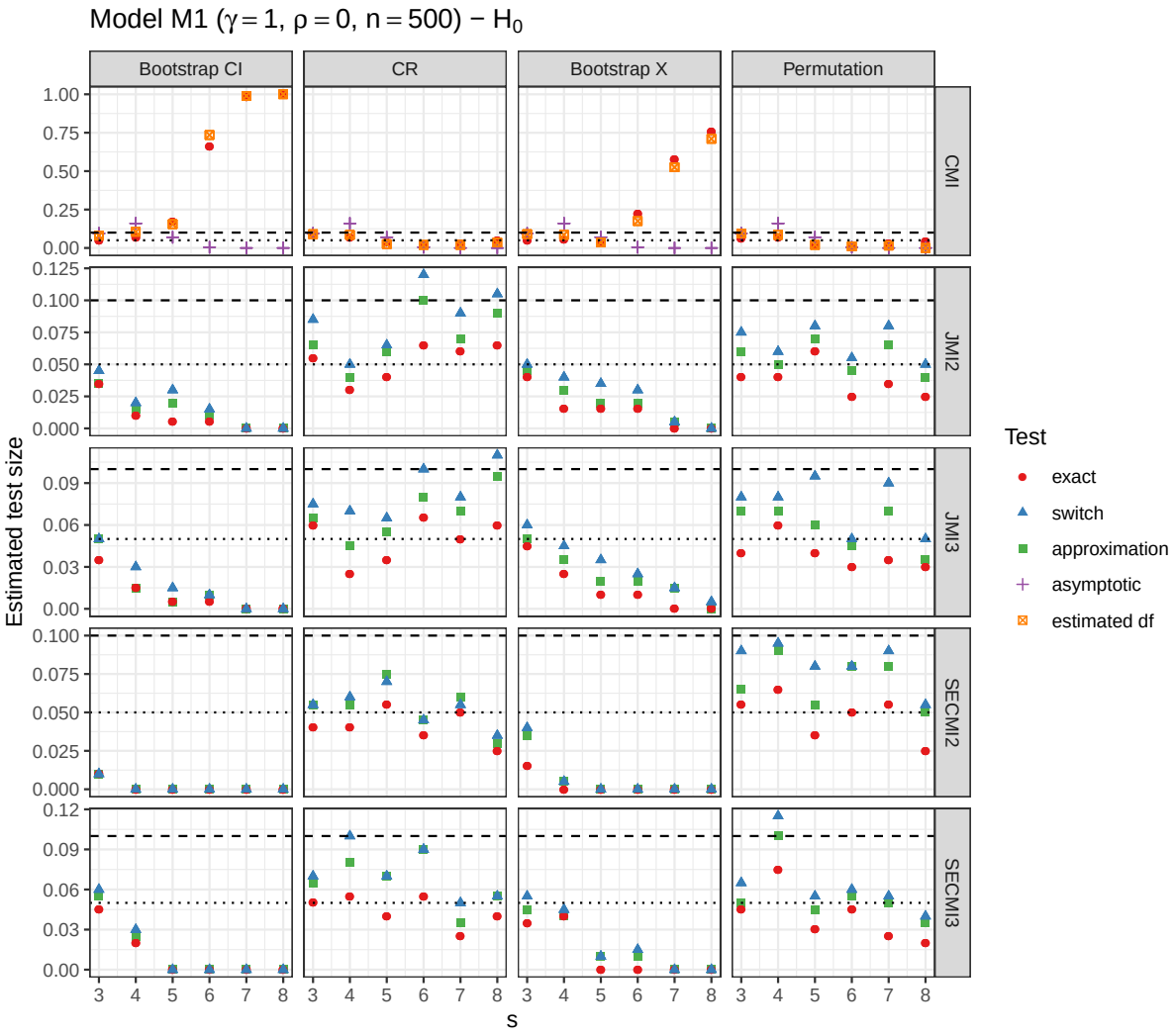


Figure 3.6: Attained significance level compared with the assumed one ($\alpha = 0.05$) in testing $H_0 : X \perp\!\!\!\perp Y|(Z_1, Z_2, \ldots, Z_s)$ in model M1 for $\gamma = 1$, $\rho = 0$, $n = 500$ versus s.

similar to the tests `estimated df` and `approximation` (they use generalized Pareto distribution as a benchmark), and `exact`. Their conclusion is that for the procedure analogous to `estimated df` and `approximation` usually requires fewer than $1/P_{perm}$ of resamples, whereas for procedure analogous to test `exact` usually more than $1/P_{perm}$ permutations is needed, where P_{perm} denotes p-value. Our aim is to obtain tests controlling type I errors, not precise estimation of p-values, thus chosen number of resamples $B = 50$ seems reasonable, as $1/\alpha = 20$ for $\alpha = 0.05$.

3.1.4. Significance level

We estimated first test sizes for a predefined significance level $\alpha = 0.05$ and checked if they do not exceed α significantly in models M1 and M2. The analysed null hypothesis is

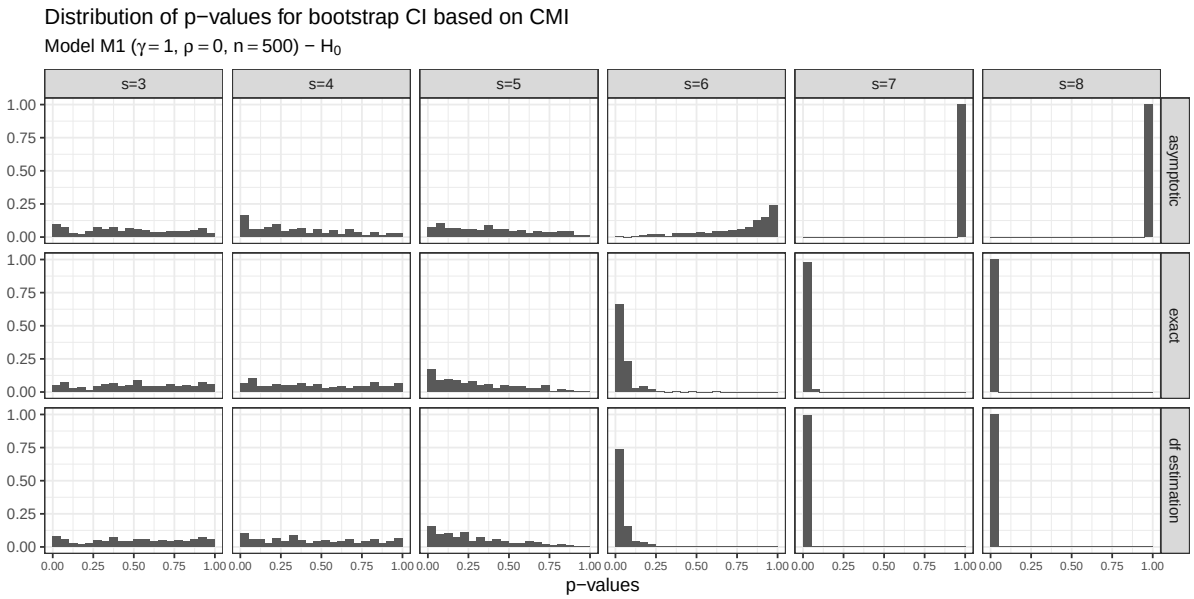


Figure 3.7: The histograms show the distribution of p-values for tests based on $\widehat{CMI}$ for the model in Figure 3.6 ($\gamma = 1$, $\rho = 0$). The bars are 0.05 wide and their height denotes the fraction of p-values in a given interval. Each sample consists of $n = 500$ observations and histograms are based on $N = 200$ values.

$$H_0 : X \perp\!\!\!\perp Y | Z_1, Z_2, \ldots, Z_s \tag{3.2}$$

and both for M1 and M2 the above property is satisfied. In Figures 3.6 and 3.13 below we present the results for all testing procedures and resampling scenarios for chosen set of parameters. Figures 3.7-3.12 focus on problems revealed by Figure 3.6.

Figure 3.6 shows the dependence of estimated test sizes on the number of conditioning variables. In general, for the tests based on criteria and $\widehat{CMI}$ (except for test `asymptotic`) for both CR and permutation scenario they do not exceed 0.1 and fluctuate around the level $\alpha = 0.05$. The most conservative approach based on criteria in the setting of Figure 3.6 is the test using quantiles of $\widehat{Crit}^*$ (test `exact`), whereas the most liberal is the procedure based on switch between the normal and the distribution of quadratic form (test `switch`). Although the test based on modified chi-square distribution (`approximate`) is asymptotically justified only in the situations, in which $\sigma^2 = 0$, it seems to have the appropriate size. The tests based on $\widehat{CMI}$ (except test `asymptotic`) work also for bootstrap CI and bootstrap X scenarios, but only for small s, and as s increases, the asymptotic approximation becomes inaccurate. The reason of that is that there is smaller average number of observations per cell and thus the approximation of the distribution

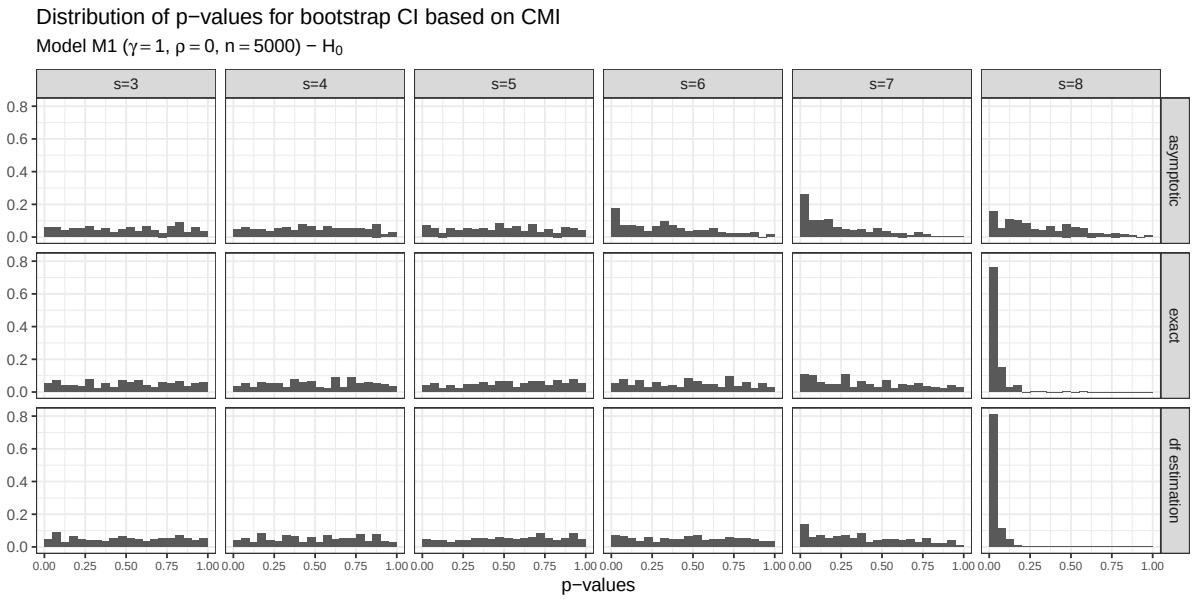


Figure 3.8: The histograms show the distribution of p-values for tests based on $\widehat{CMI}$ for the model in Figure 3.6 ($\gamma = 1$, $\rho = 0$). The bars are 0.05 wide and their height denotes the fraction of p-values in a given interval. Each sample consists of $n = 5000$ observations and histograms are based on $N = 200$ values.

$p(x|z_S)p(y|z_S)p(z_S)$ based on fractions is less accurate. For example, when $s = 8$ the number of distinct values of (x, y, z_S) equals 2^{10} and thus when $n = 500$ and for uniform $p(x, y, z_S)$ we have $500/2^{10} \approx 0.5$ observation per cell on average. When $n = 5000$, the average number of observations per cell equals 4.9. Compared to bootstrap CI, in bootstrap X we estimate only $\hat{p}(x|z_S)$, thus we only use (x, z_S) and not (x, y, z_S). Hence the number of distinct values decreases twice and for this reason the number of observations per cell in the example above equals approximately 1 and 9.8, respectively. Similar significance levels are attained for the same set of parameters as in Figure 3.6 for model M2 (the plot showing estimated test sizes with respect to s for M2 is omitted, but other results for this model are shown in Figure 3.13).

In Figure 3.7 we show the histograms of p-values for $\widehat{CMI}$ for all three testing procedures in model M1 ($\gamma = 1$, $\rho = 0$) for changing number of conditioning variables s. For small s the distributions seem to be close to uniform, whereas for larger conditioning sets all the p-values are in $[0, 0.05]$ (tests `exact` and `estimated df`) or $[0.95, 1]$ (test `asymptotic`). For test `asymptotic` as the number of conditioning variables increases, the fraction of rejections at first grows but then falls to 0. That effect can be seen in Figures 3.6 and 3.22 showing estimated test sizes, and in Figure 3.7 showing the distribution

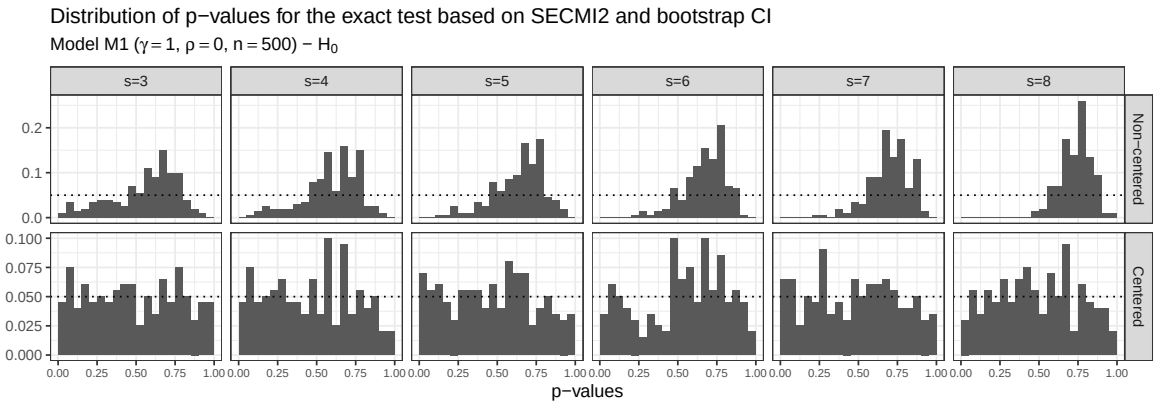


Figure 3.9: The histograms show the distribution of p-values for test `exact` and $SECMI2$ criterion for the model in Figure 3.6 ($\gamma = 1$, $\rho = 0$, $n = 500$) in bootstrap CI scenario for centered and not centered test statistics. The bars are 0.05 wide and their height denotes the fraction of p-values in a given interval.

of p-values. Discussing this issue, in [41] the authors assert that for small sample sizes some triples (x, y, z) are not represented in data and thus probability of obtaining them in resampling equals 0. In that cases the degrees of freedom should be adjusted. The erratic behaviour in terms of the estimated test sizes and the power for tests based on $\widehat{CMI}$ might be also caused be the problems with precise estimation of the corresponding criterion. When the number of observations per cell is small, many components of the sum in the formula for $\widehat{CMI}$ equal 0 as $\hat{p}(x, y, z_1, z_2, \ldots, z_s) = 0$. The estimation of the criteria requires less observations, as e.g. to compute $\widehat{JMI2}$ or $\widehat{SECMI2}$ we use $\hat{p}(x, y, z_i)$ instead of joint distribution for the whole vector of z_S. Thus instead of having $500/2^{10} \approx 0.5$ observation per cell on average as in previous example for binary variables and $s = 8$, we have $500/2^3 = 62.5$ observations on average. In Figures 3.9 and 3.10 the distribution of p-values for $\widehat{SECMI2}$ and test `exact` are shown for the same parameters as in Figure 3.7. In these plots the distribution for large s is not contained in the interval $[0, 0.05]$ and for centered statistics their distribution is closer to the uniform (for more details see the discussion on Figures 3.9-3.11 below). This is due to a better approximation of the criteria than for CMI. We also note that when we increase the number of observations, the distribution of p-values for $\widehat{CMI}$ is closer to uniform. In Figure 3.8 the number of observations $n = 5000$, thus is ten times larger than in Figure 3.7 and the presented histograms are indeed visually closer to the uniform.

In Figure 3.6 the sizes of tests based on criteria for bootstrap CI and bootstrap X fall to 0. The histograms of p-values for $SECMI2$ in bootstrap CI and bootstrap X scenarios,

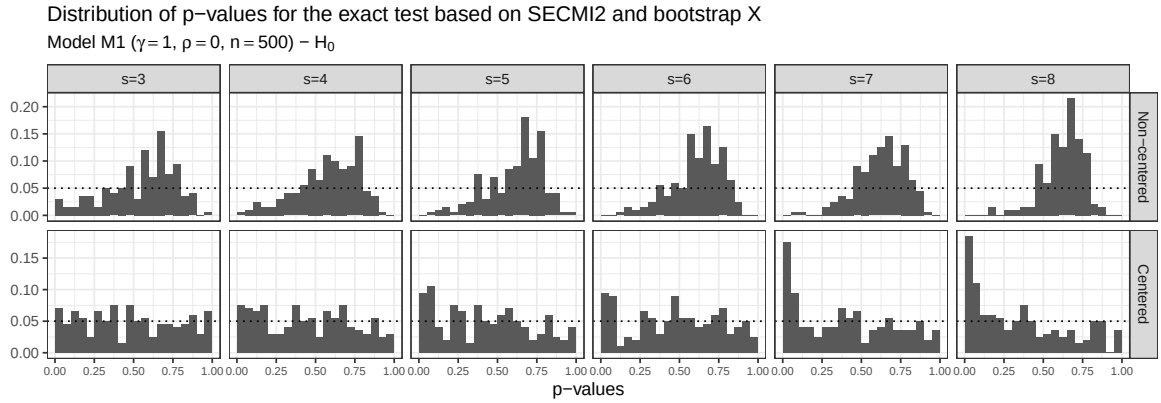


Figure 3.10: The histograms show the distribution of p-values for test `exact` and $SECMI2$ criterion for the model in Figure 3.6 ($\gamma = 1$, $\rho = 0$, $n = 500$) in bootstrap X scenario for centered and not centered test statistics. The bars are 0.05 wide and their height denotes the fraction of p-values in a given interval.

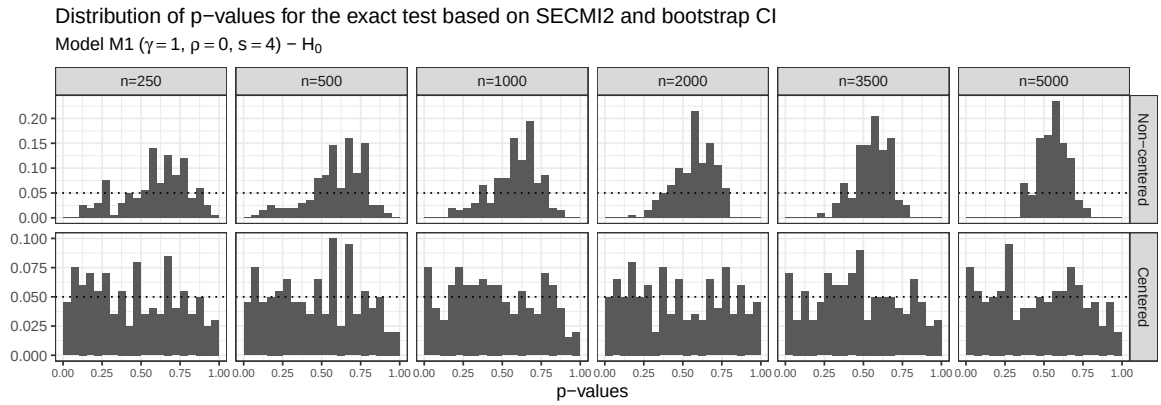


Figure 3.11: The histograms show the distribution of p-values for test `exact` and $SECMI2$ criterion in for the model in Figure 3.6 ($\gamma = 1$, $\rho = 0$ and $s = 4$) in bootstrap CI scenario for centered and not centered test statistics versus n. The bars are 0.05 wide and their height denotes the fraction of p-values in a given interval.

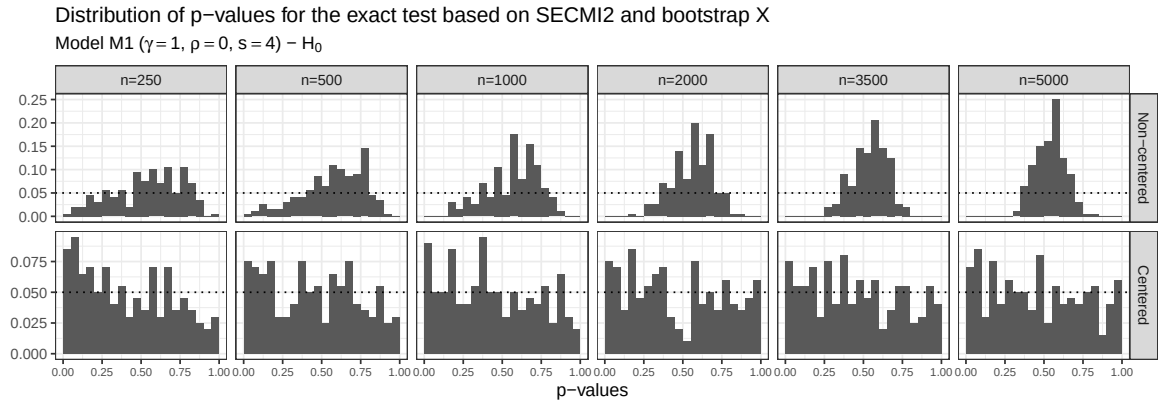


Figure 3.12: The histograms show the distribution of p-values for test `exact` and $SECMI2$ criterion for the model in Figure 3.6 ($\gamma = 1$, $\rho = 0$ and $s = 4$) in bootstrap X scenario for centered and not centered test statistics versus n. The bars are 0.05 wide and their height denotes the fraction of p-values in a given interval.

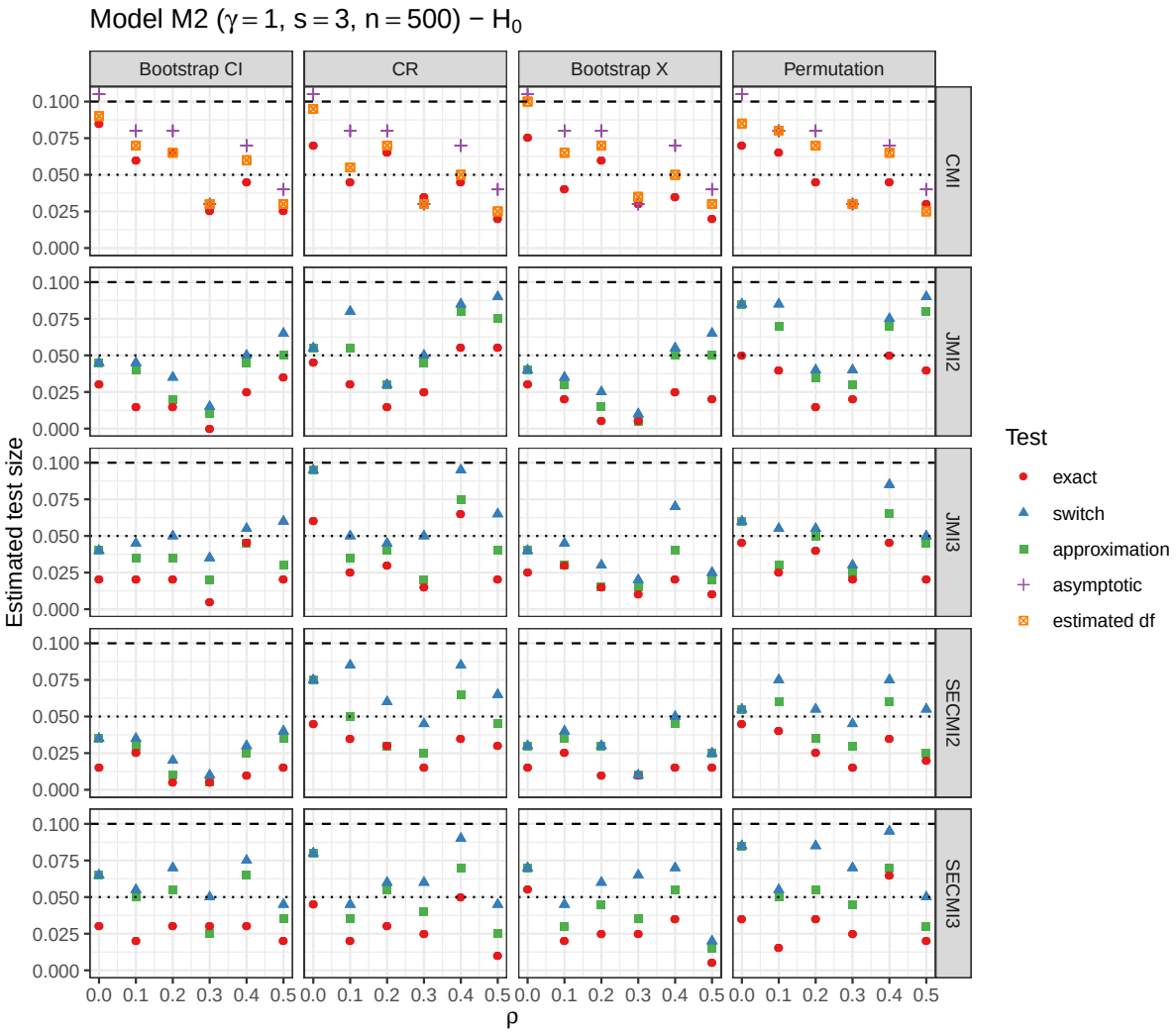


Figure 3.13: Attained significance level compared with the assumed one ($\alpha = 0.05$) in testing $H_0 : X \perp\!\!\!\perp Y|(Z_1, Z_2, \ldots, Z_s)$ in model M2 for $\gamma = 1$, $s = 3$ and $n = 500$ versus ρ.

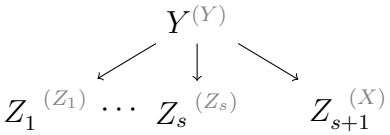
for which this effect is the most noticeable are shown in Figures 3.9 - 3.12 in the top panels. The histograms of p-values under the null hypothesis should approximate the density of the uniform distribution, but the p-values are more concentrated around the center of the interval. One of the reason for that effect was mentioned above: as s grows the ratio n/s became smaller and thus expected number of observations per cell is insufficient to apply asymptotic approximation as for $\widehat{CMI}$. But in this case, the main reason for that is that the asymptotic distributions hold for *centered* criteria. Thus comparing quantiles of $\widehat{Crit}^* - \widehat{Crit}_{ci}$ and $\widehat{Crit} - Crit$ is justified, whereas comparing quantiles of $\widehat{Crit}^*$ with $\widehat{Crit}$ is based on heuristic reasoning. In Figures 3.9 - 3.12 we show p-values computed for the `exact` test in which we use centered ($\widehat{Crit}^* - \widehat{Crit}_{ci}$ and $\widehat{Crit} - Crit$), and not centered ($\widehat{Crit}^*$ and $\widehat{Crit}$) test statistics. In case of bootstrap CI asymptotic distribution

of $\widehat{Crit}^* - \widehat{Crit}_{ci}$ in case of $\sigma^2 > 0$ is the same as $\widehat{Crit} - Crit$, but it is not the case for bootstrap X, as for that scenario σ^2 for the criterion computed on a resampled sample is smaller than σ^2 for the criterion based on a sample. This explains that even for centered statistic the distribution is not uniform in Figure 3.10.

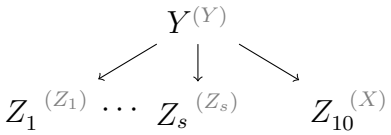
We stress that the approach based on centered statistics is not applicable in practise. Although we can approximate the distribution of $\widehat{Crit} - Crit$ using $\widehat{Crit}^* - \widehat{Crit}_{ci}$ as both $\widehat{Crit}^*$ and $\widehat{Crit}_{ci}$ can be computed based on a resampled sample, we do not know the value of $Crit$ under H_0 in general. In case of CMI the situation is easier, as assuming conditional independence of X and Y given Z, we have $I(X, Y|Z) = 0$. For criteria similar conclusions do not hold under $X \perp\!\!\!\perp Y|Z_S$, e.g. for JMI we have the equivalence $JMI = 0 \iff X \perp\!\!\!\perp Y|Z_i$ for all $i \in S$. Note, however, that for faithful distribution we have an implication (see (3.1))

$$X \perp\!\!\!\perp Y|Z_i \Rightarrow X \perp\!\!\!\perp Y|Z_S,$$

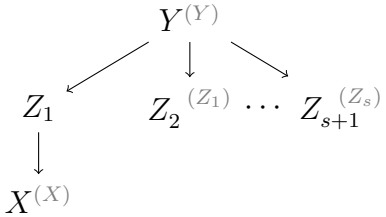
hence in that case $JMI = 0$ implies $X \perp\!\!\!\perp Y|Z_S$.



(a) Model used in testing $H_{0,1}$



(b) Model used in testing $H_{0,3}$



(c) Model used in testing $H_{0,2}$

Figure 3.14: The models used to estimate the power of the procedures. The models are generated as model M1 described in Section 3.1.1. The variables considered in the null hypotheses $H_{0,1}$ - $H_{0,3}$ in terms of (3.2) are in gray color.

In Figure 3.13 the number of conditioning variables equals 3 and attained level of significance does not exceed 0.05 significantly for a grid of values of the parameter ρ. The results presented in Figure 3.13 for model M1 instead of M2 are analogous. The same

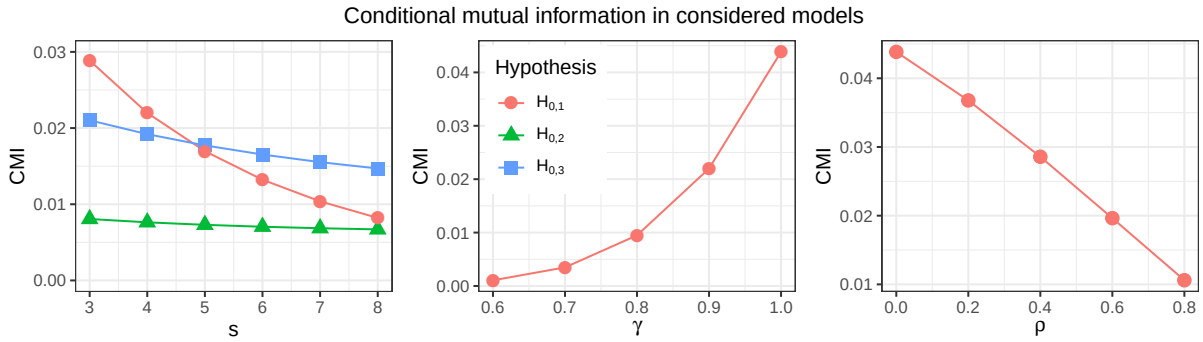


Figure 3.15: Conditional mutual information computed for models for which power is investigated (change in variables γ and ρ was considered for $H_{0,1}$ only). The changes of the CMI value reflects the difficulty of the problem - for lower CMI, the null hypothesis is harder to reject.

conclusion can be drawn for models in which γ changes, whereas ρ is fixed and s is fixed and small enough.

3.1.5. Power

In this section we show the results on power for testing various null hypotheses in models M1 and M2, when they do not hold. We consider three hypotheses $H_{0,1}$, $H_{0,2}$ and $H_{0,3}$ (we analyse $H_{0,2}$ only in model M1 as in model M2 we have $X \perp\!\!\!\perp (Y, Z_1, Z_2, \ldots, Z_s)$):

$$
\begin{aligned}
H_{0,1} &: Z_{s+1} \perp\!\!\!\perp Y | Z_1, Z_2, \ldots, Z_s, \\
H_{0,2} &: X \perp\!\!\!\perp Y | Z_2, Z_3, \ldots, Z_{s+1}, \\
H_{0,3} &: Z_{10} \perp\!\!\!\perp Y | Z_1, Z_2, \ldots, Z_s, \quad s < 10.
\end{aligned}
$$

Obviously, null hypotheses above are special cases of the general CI hypothesis (3.2) for specific choices of X and conditioning set. In Figure 3.14 the choices of the variables for the above hypotheses are shown in terms of (3.2) in brackets and in gray color. The values of conditional mutual information for corresponding parameters settings and models from the null hypotheses i.e. $I(Z_{s+1}, Y | Z_1, Z_2, \ldots, Z_s)$, $I(X, Y | Z_2, Z_3, \ldots, Z_{s+1})$ and $I(Z_{10}, Y | Z_1, Z_2, \ldots, Z_s)$ are shown in Figure 3.15. Rejection of the null hypothesis should be easier, the larger the values of conditional mutual information are.

When s and ρ are fixed and we change the value of γ, the dependence between Z_i and Y for $i > 1$ is weaker for smaller γ. In the boundary case, if $\gamma = 0$ and $\rho = 0$, Y and Z_i for $i > 1$ are independent. We recall that for $\rho = 0$ variables Z_i for $i = 1, 2, \ldots, s$

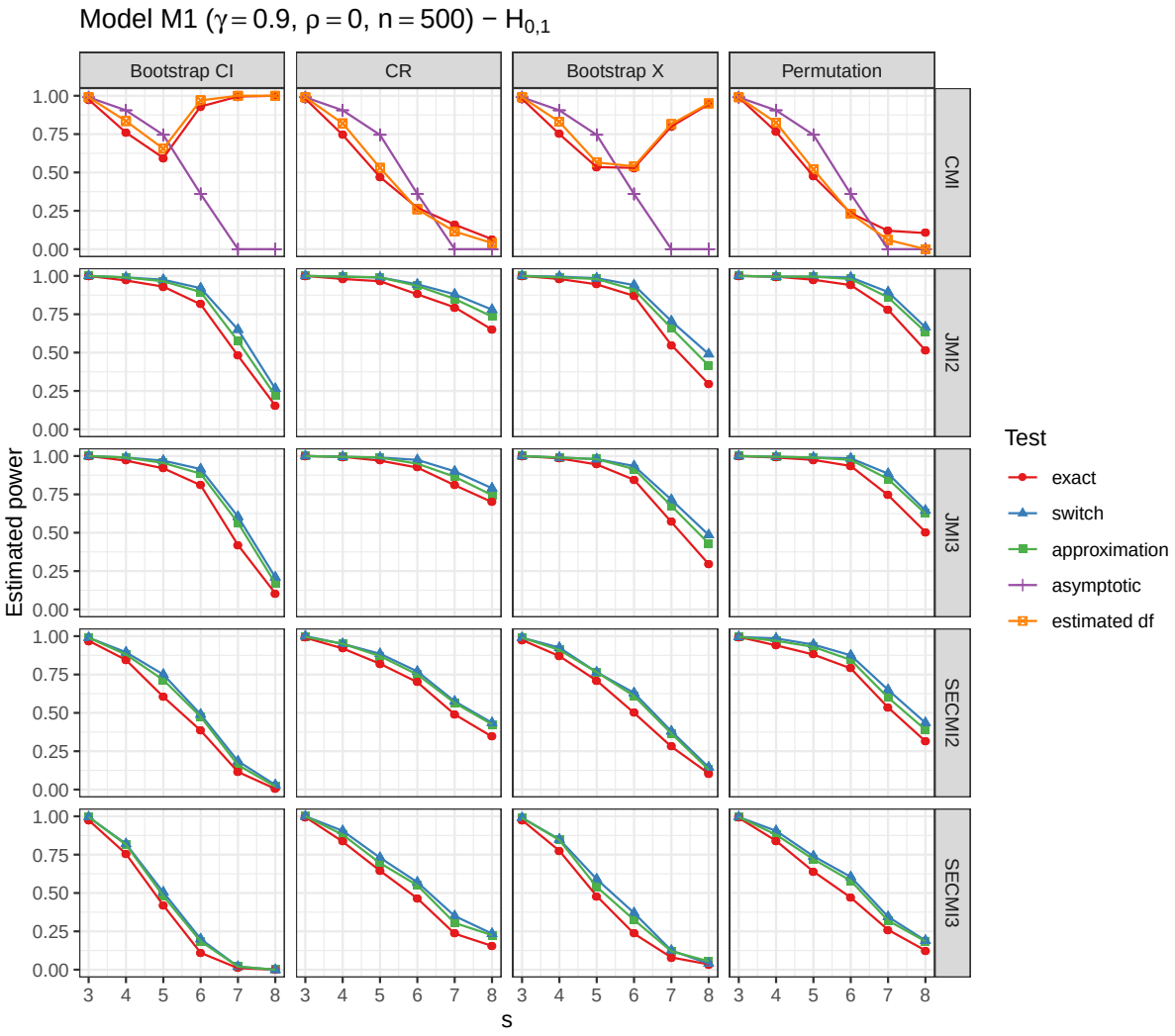


Figure 3.16: The power with respect to s for the testing procedures in testing $H_{0,1}$ in model M1 for $\gamma = 0.9$, $\rho = 0$ and $n = 500$.

are independent given Y. When the number of conditioning variables grows and the rest of parameters is unchanged, the problem becomes more difficult as we obtain more information from the conditioning set and thus the information about Y included in X decreases. In Figure 3.16, showing the power of procedures for testing $H_{0,1}$, the marginal dependence between Y and Z_s also changes. Thus in that example as s grows, the null becomes harder to reject for two reasons: the conditioning set grows and the marginal dependence between Y and Z_s decreases as s grows. For that reason we also consider the hypothesis $H_{0,2}$ and $H_{0,3}$ in which only the number of conditioning variables changes. When γ and s are fixed and ρ varies, it is the most easy to reject the null when variables $Z_1, Z_2, \ldots, Z_s$ are independent given Y. Indeed, in that case no extra information is provided about Z_1 through conditioning by $Z_2, Z_3, \ldots, Z_s$. As the value of ρ increases, the

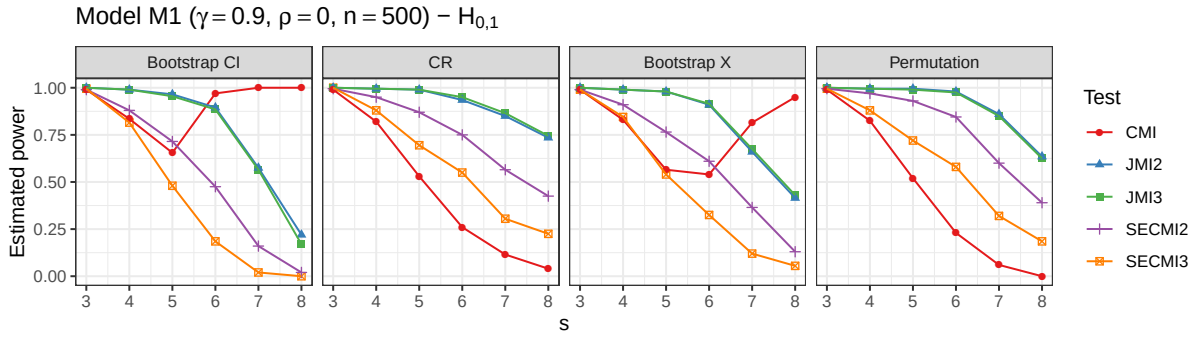


Figure 3.17: The power with respect to s for `df estimation` test based on CMI and `approximation` based on the criteria in testing $H_{0,1}$ in model M1 for $\gamma = 0.9$, $\rho = 0$ and $n = 500$. Results for all the procedures for the same parameter values and the null is presented in Figure 3.16.

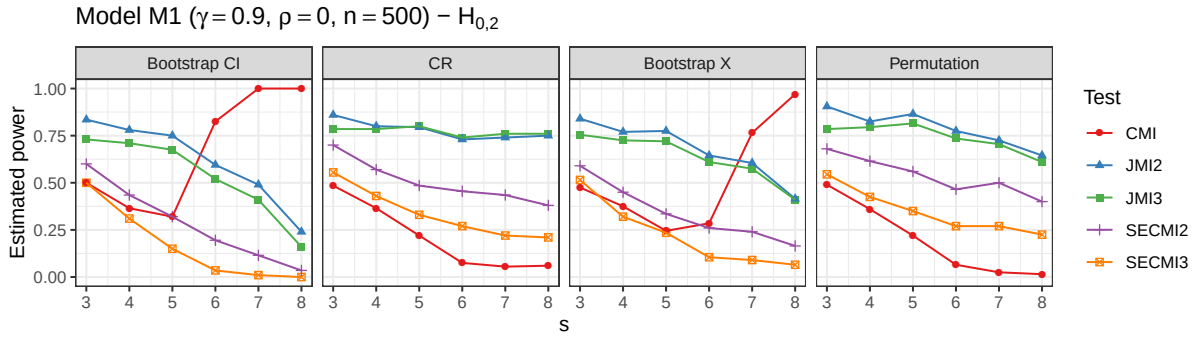


Figure 3.18: The power with respect to s for `df estimation` test based on CMI and `approximation` based on the criteria in testing $H_{0,2}$ in model M1 for $\gamma = 0.9$, $\rho = 0$ and $n = 500$.

more we know about Z_1 looking at $Z_2, Z_3, \ldots, Z_s$ and thus given $Z_2, Z_3, \ldots, Z_s$, variable Z_1 itself contains less information about Y.

The results for testing $H_{0,1}$ are presented in Figures 3.16-3.17 and 3.20-3.21. In each figure the dependence of power of testing procedures on s (two first figures), γ and ρ is shown. In Figures 3.18 and 3.19 the dependence on s is shown for testing $H_{0,2}$ and $H_{0,3}$, respectively. The parameter γ in Figures 3.16-3.17 and 3.18 has the same value $\gamma = 0.9$, whereas in Figure 3.19 we consider $\gamma = 0.95$. For $\gamma = 0.9$ the distributions of $\tilde{Z}_{10}|Y = 0$ and $\tilde{Z}_{10}|Y = 1$ are $\mathcal{N}(0, 1)$ and $\mathcal{N}((0.9)^9, 1)$ ($(0.9)^9 \approx 0.39$) respectively, thus the dependence between Z_{10} and Y is weak. In that case the differences between procedures are not pronounced. For $\gamma = 0.95$ distributions are $\mathcal{N}(0, 1)$ and $\mathcal{N}((0.95)^9, 1)$, respectively ($(0.95)^9 \approx 0.63$), thus the problem becomes easier which makes the comparison of methods possible.

In Figure 3.16 the behaviour of all procedures is compared. In the remaining plots we

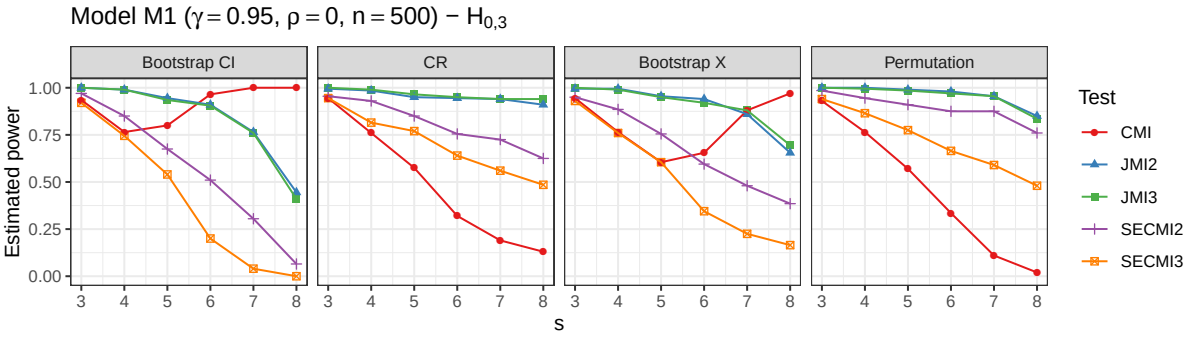


Figure 3.19: The power with respect to s for `df estimation` test based on CMI and `approximation` based on the criteria in testing $H_{0,3}$ in model M1 for $\gamma = 0.95$, $\rho = 0$ and $n = 500$.

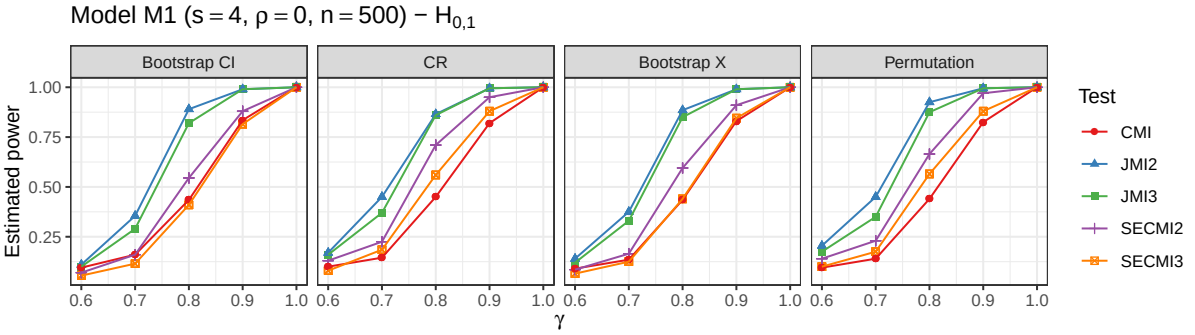


Figure 3.20: The power with respect to γ for `estimated df` test based on CMI and `approximation` based on the criteria in testing $H_{0,1}$ in model M1 for $s = 4$, $\rho = 0$ and $n = 500$.

compare only `estimated df` test for CMI and `approximation` test for the criteria as the semi-parametric procedures works better than the test `exact` and they are applicable in practice (we recall that test `switch` is not). In case of CMI the asymptotic distribution of $2n\widehat{CMI}$ is χ^2_d and thus in the semi-parametric approach in which we use chi-square distribution as a benchmark, just the estimation of the number of degrees of freedom is needed, whereas for the criteria the situation is more complicated as we need to center the test statistic in order to obtain convergence to the distribution of the quadratic form in normal variables and thus we use more flexible distribution proposed in [44]. Note that in Figure 3.16 the markings of the test procedures are the same as in plots showing significance level, but they differ from the markings in plots, where only `estimated df` test and `approximation` for four criteria are shown (both in colors and in shapes).

The results for the power of testing procedures shows similar pattern to those of the results for significance level. The power of `approximation` procedure is usually between

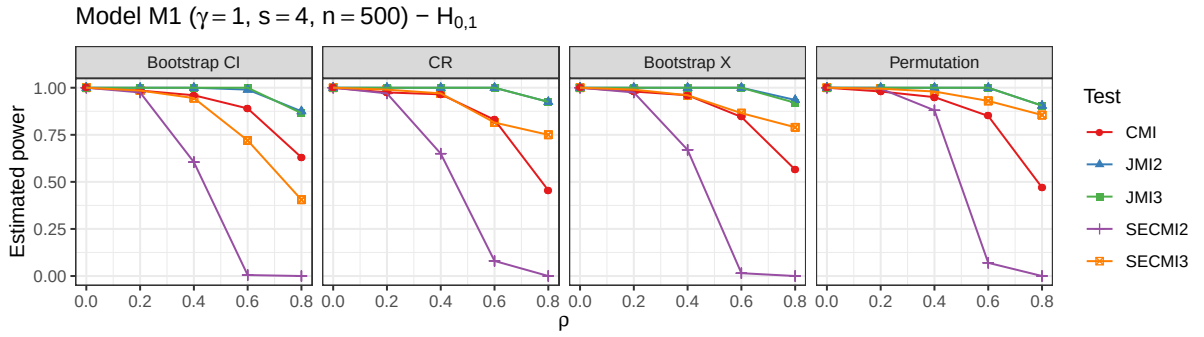


Figure 3.21: The power with respect to ρ for `df estimation` test based on CMI and `approximation` based on the criteria in testing $H_{0,1}$ in model M1 for $\gamma = 1$, $s = 4$ and $n = 500$.

the power of `switch`, which is the largest, and `exact` (cf. Figure 3.16). We also notice that as for large s the nominal significance level for CMI and bootstrap CI/bootstrap X is significantly exceeded (see Figure 3.13), the power of these procedures also grows, and the test is not reliable in that cases. For $s \geq 7$ the test `asymptotic` cannot reject the null hypothesis due to the problems with estimating joint probability $\hat{p}(x, y, z_s)$ discussed before, which can be also observed in Figure 3.13, where the attained significance level equals 0, although the assumed significance level equals $\alpha = 0.05$.

Figures 3.18 and 3.19 show how the estimated power changes when the conditioning set grows in testing $H_{0,2}$ and $H_{0,3}$ respectively. The results are similar to the one obtained in testing $H_{0,1}$ (cf. Figure 3.16 and 3.17), but there the true value of CMI decreases slower with increasing s than in testing $H_{0,1}$ (see left panel of Figure 3.15). In both settings for CR and permutation scenario, $JMI2$ and $JMI3$ work the best followed by $SECMI2$ and $SECMI3$. The power of CMI, although it starts approximately from the same point as the power of $SECMI3$ for $s = 3$, decreases faster and thus the test based on CMI is the weakest.

The conclusion from the results presented in Figures 3.20 and 3.21 is that in these cases the criteria $JMI2$ and $JMI3$ outperform other procedures. Test based on CMI usually works worse than all the criteria with an exception of that case in Figure 3.21, where $SECMI2$ has the lowest power.

3.1.6. Analysis of high-order interaction model

In the previous examples the second order criteria $JMI2$ and $SECMI2$ had an advantage as one can detect conditional dependencies conditioning just on one variable e.g. in

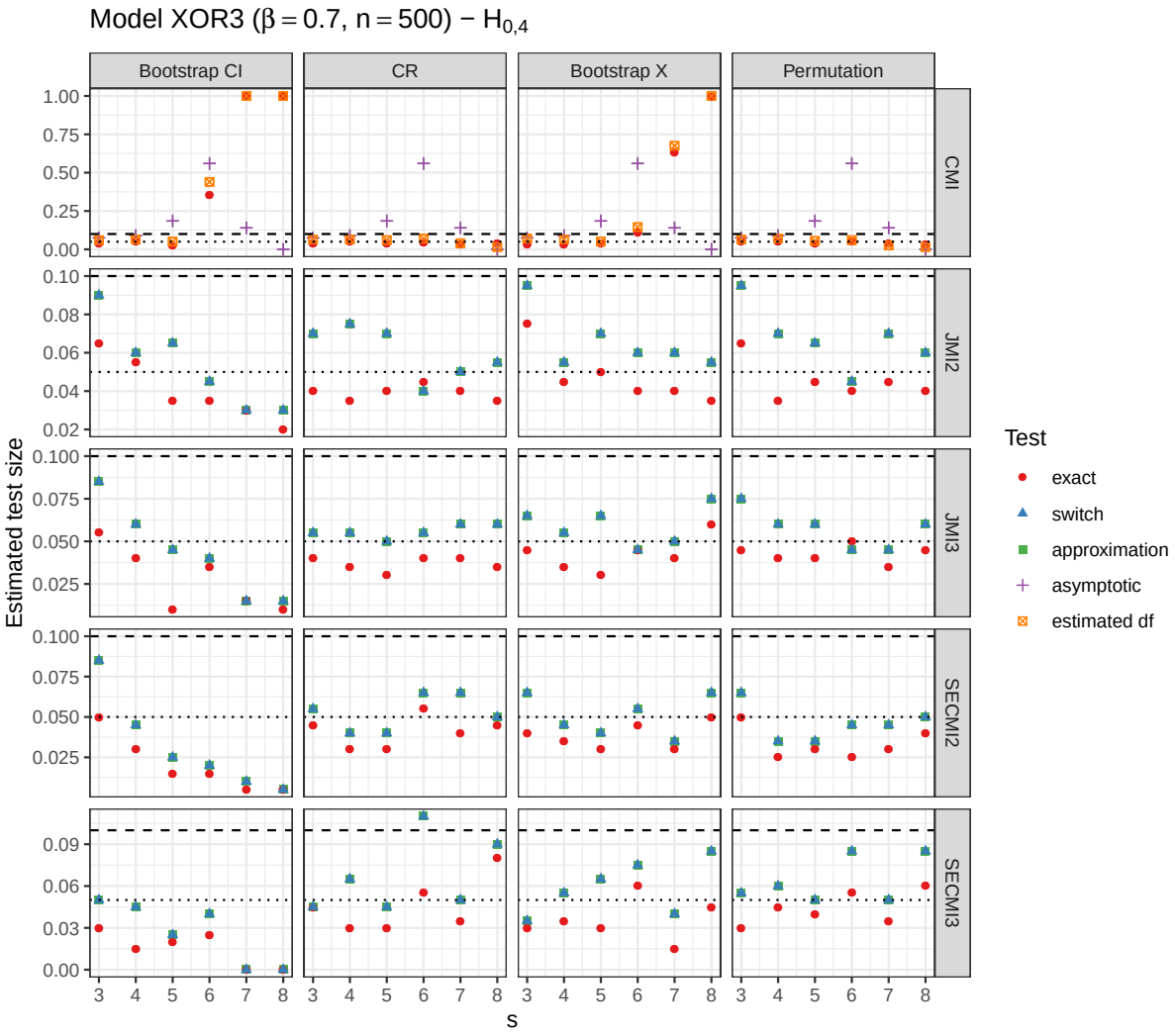


Figure 3.22: Attained significance level with respect to s compared with the assumed one ($\alpha = 0.05$) in testing $H_{0,4}$ in XOR3 model for $\beta = 0.7$ and $n = 500$.

model used in $H_{0,1}$ we have $I(Z_{s+1}, Y|Z_i) > 0$ for all $i \in \{1, 2, \ldots, s\}$. Therefore here we investigate how the criteria behave in the scenario in which only high-order interactions appear. The model is based on three-dimensional XOR of two conditioning variables and X. The distribution of Y is as follows:

$$P(Y = 1|X + Z_1 + Z_2 =_2 1) = P(Y = 0|X + Z_1 + Z_2 =_2 0) = \beta,$$

where $0.5 < \beta < 1$ and $=_2$ denotes addition modulo 2. For $\beta = 1$ we obtain deterministic XOR based on three variables. We also introduce variables $Z_3, Z_4, \ldots, Z_s$ independent of (X, Y, Z_1, Z_2) . All variables $X, Z_1, Z_2, \ldots, Z_s$ are binary with the probability of success equal to 0.5 and are independent. We will call this model XOR3.

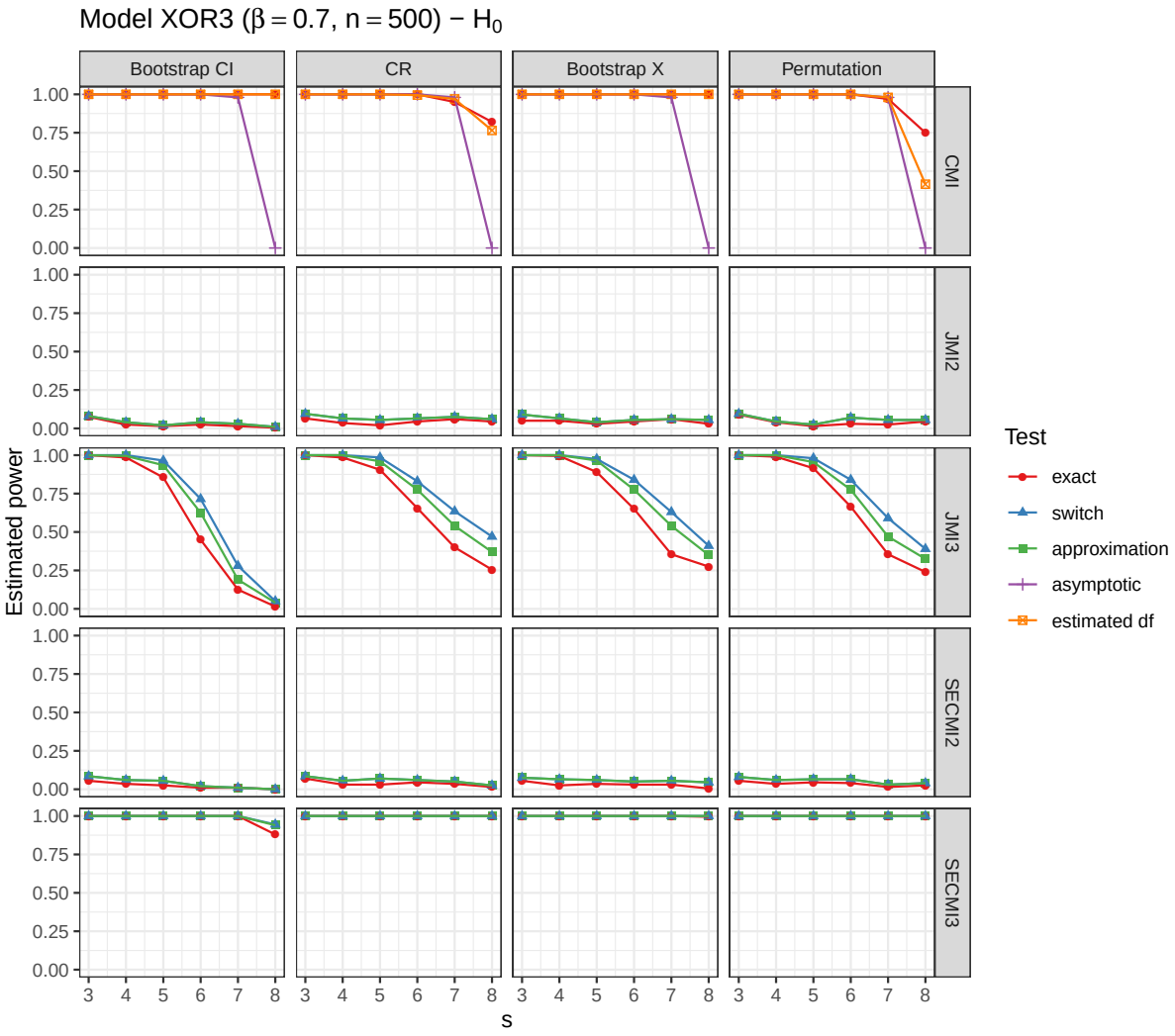


Figure 3.23: Power with respect to s of the testing procedures in testing H_0 in XOR3 model for $\beta = 0.7$ and $n = 500$.

We consider two hypotheses in XOR3 model:

$$H_{0,4} : X \perp\!\!\!\perp Y | Z_2, Z_3, \ldots, Z_{s+1} \tag{3.3}$$

and

$$H_0 : X \perp\!\!\!\perp Y | Z_1, Z_2, \ldots, Z_s. \tag{3.4}$$

The first null holds as without fixing Z_1, X does not provide any information about Y since the result of $X + Z_2$ modulo 2 does not determine the result of $X + Z_1 + Z_2$ modulo 2 in any way. On the other hand, the second hypothesis H_0 should be rejected as the triple (X, Z_1, Z_2) affects the distribution of Y and thus X and Y are not independent given (Z_2, Z_3).

In Figure 3.22 the attained level of significance for (3.3) is compared with the assumed one ($\alpha = 0.05$). All tests using CR and permutation scenario do not significantly exceed the assumed level. The behaviour of the test `asymptotic` is similar as in Figure 3.13 - as the conditioning set grows, at first the fraction of rejections increases and then it falls to zero. The problems described in the previous section (see Figure 3.13) for procedures based criteria, and `exact` and `estimated df` for CMI also occur in this case.

The power of testing H_0 in case of (3.4) for $\beta = 0.7$ is shown in Figure 3.23. The plots show that in the situations, in which it is impossible to detect the dependence based on conditioning by one variable, the criteria $JMI2$ and $SECMI2$ do not work at all, whereas their third-order counterparts and CMI detect the dependence in most of cases. The test based on $SECMI3$ outperforms other procedures, i.e. for $s = 7, 8$ test based on CMI rejects null less often than the one using $SECMI3$.

Note that also Figure 3.21 corresponds to occurrence of higher order interaction as when $\rho > 0$ the variables $(Z_1, Z_2, \ldots, Z_s)$ are dependent. In that case $SECMI3$ also outperforms $SECMI2$ in CR and permutation scenarios. $JMI2$ and $JMI3$ yield similar results.

3.2. Global null for individual hypotheses of conditional independence

We recall that the global null defined in Section 1.4.2 is a hypothesis $H_0 : \cap_{i=1}^{s} H_{0,i}$, where individual hypotheses assert that the conditional independence

$$H_{0,i} : X \perp\!\!\!\perp Y | Z_i, \tag{3.5}$$

holds for $i = 1, \ldots, s$. To test H_0 we use $\widehat{JMI}$ and its asymptotic distribution under the null hypothesis which is given in Theorem 1.4.11, and compare it with behaviour of generic tests introduced below.

3.2.1. Test based on $\widehat{JMI}$

For a given sample drawn from $p(x, y, z)$ we calculate $\widehat{JMI}$ and plug-in estimator $\widehat{M}$ of a matrix M defined in Theorem 1.4.11. We use now the fact that the asymptotic distribution W of $\widehat{JMI}$ under H_0 given in (1.46) is determined by the eigenvalues

$\lambda_i(M)$ and we approximate them by $\widehat{W}$ plugging in $\lambda_i(\widehat{M})$ for $\lambda_i(M)$, where $\lambda_i(\widehat{M})$ are numerically calculated. Then the rejection region for a given significance level α is given by $\{\widehat{JMI} \geq q_{\widehat{W},1-\alpha}\}$, where $q_{\widehat{W},1-\alpha}$ is quantile of the order $1-\alpha$ of $\widehat{W}$. A function `eigen` from R package `base` has been used to calculate the eigenvalues and package `CompQuadForm` ([11]) is used for quantiles of $\hat{W}$.

Note that this approach is also possible for CI testing discussed previously, but have not been attempted due to time constraints.

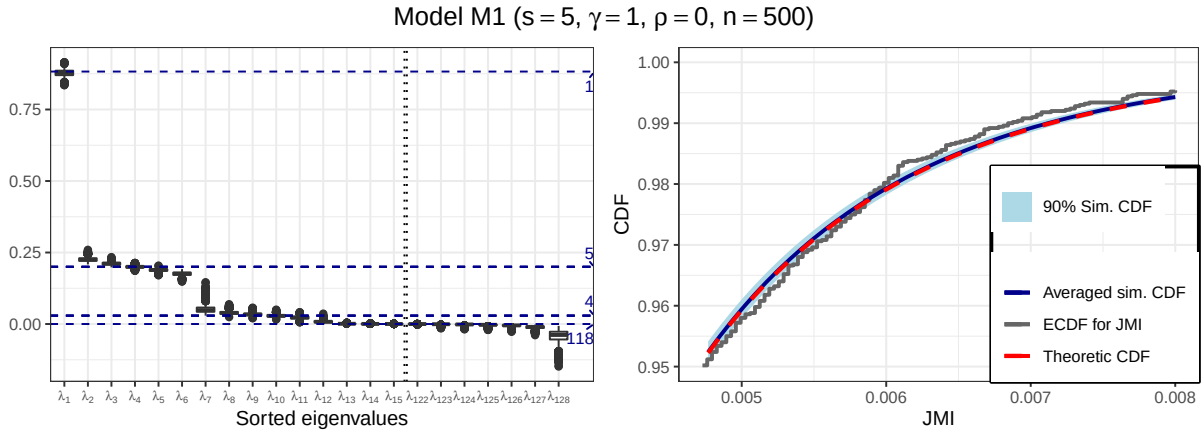


Figure 3.24: Left: Box-plots of the empirical values $\lambda_i(\widehat{M}), i = 1, \ldots, 128$ for model M1 ($n = 500, s = 5, \gamma = 1, \rho = 0$). Eigenvalues $\lambda_i(M)$ approximately equal to 0 (multiplicity 118), 0.029 (multiplicity 4), 0.2 (multiplicity 5) and 0.883 (multiplicity 1) are marked by the horizontal lines. Right: values of theoretical CDF, the empirical CDF of $\widehat{JMI}$ and the average of CDFs corresponding to $\lambda_i(\widehat{M})$ for the values of JMI greater than 0.95th quantile of $\widehat{JMI}$.

3.2.2. Generic methods

We use three generic methods which are designed to control type I error while performing multiple tests, namely Bonferroni correction, Simes method and Fisher test (see e.g. [12], [38] and [14]).

i) Bonferroni correction (`Bonferroni`). The null hypothesis H_0 is rejected when $\min_i p_i \leq \alpha/s$, where s is the number of tests performed. Probability of type I error is bounded by α and $p_1, p_2, \ldots, p_s$ are p-values of individual tests. The correction is known to work well when the test statistics used to test individual hypotheses are independent, but in a general case is conservative leading to the low power when H_0 fails. Bonferroni's method controls the Type I error rate for both independent

and dependent p-values. Individual tests applied here are $\widehat{CMI}$-based tests based on chi-square benchmark distribution (test `asymptotic` from the previous section).

ii) Simes method (`Simes`). P-values of individual test $p_1, p_2, \ldots, p_s$ are calculated and ordered: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(s)}$. H_0 is rejected when for certain $i \leq s$ we have $p_{(i)} \leq i\alpha/s$, or equivalently if $\min_i p_{(i)}/i \leq \alpha/s$. Individual tests considered are the same as for Bonferroni correction method. The Simes correction was proven to satisfy the following inequality

$$P_{H_0}(\min_i p_{(i)}/i \leq \alpha/s) \leq \alpha \tag{3.6}$$

for independent p-values having uniform distribution and then it was generalised to cases in which test statistics are dependent. Namely, in [34] it is shown that (3.6) holds for p-values that satisfy MTP_2 (multivariate totally positive of order two) condition and have common marginal distribution (Proposition 3.1 in [34]).

iii) Fisher's combination test (`Fisher`). P-values of individual test $p_1, p_2, \ldots, p_s$ are combined into one statistic $T = -\sum_{i=1}^{s} 2\log(p_i)$. If under the null p-values are uniformly distributed and independent, then $T \sim \chi^2_{2s}$. The global null is rejected when $T > \chi^2_{2s}(1-\alpha)$, where $\chi^2_{2s}(\alpha)$ denotes an α-quantile of χ^2_{2s} distribution.

Note that the individual tests in the procedures discussed above have assumed level of significance only approximately, as we do not know the exact distribution of the test statistic for finite sample and we use asymptotic distribution instead.

In simulations below we repeat each experiment $N = 5000$ times.

3.2.3. Simulations

We considered models M2 and M1 to study the actual type I error and power, respectively. In model M1 the null hypothesis is not satisfied as X is not independent of Y given Z_i, where $i \in \{2, \ldots, s\}$. Only $X \perp\!\!\!\perp Y|Z_1$ holds, thus in the global null $s-1$ individual hypotheses are violated. The situation is different in model M2, where $X \perp\!\!\!\perp (Y, Z_1, Z_2, \ldots, Z_s)$, thus in this case all individual conditional independence hypotheses hold.

The eigenvalues of the estimated matrix $\widehat{M}$ approximate very closely the eigenvalues of the theoretical matrix M. In Figure 3.24 true values of eigenvalues for M are marked

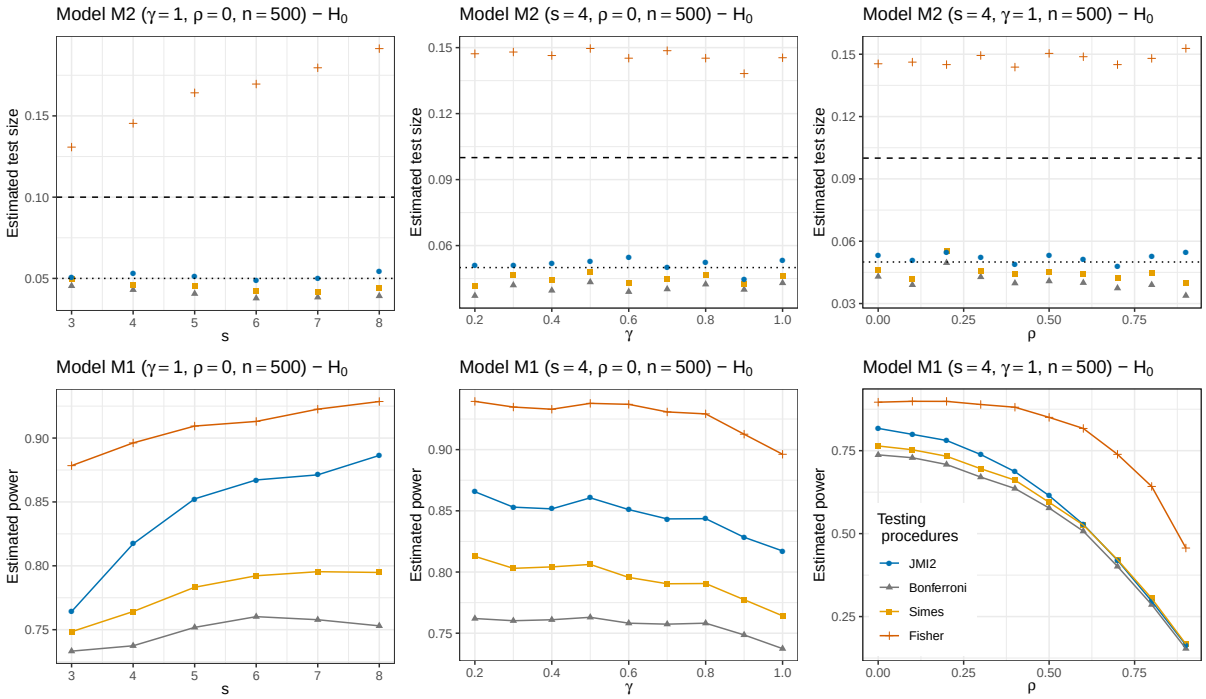


Figure 3.25: Actual test sizes and power of testing procedures for varying parameters in testing global null hypothesis.

by horizontal lines (their number is given on the right side of the plot) and their sorted sample counterparts by boxplots. The middle part of the plot (between λ_{15} and λ_{122}) is truncated, as the values in the middle are close to 0. Note that as $s = 5$, the number of the unique values of triples (x, y, z_S) equals $2^7 = 128$, thus the number of eigenvalues also equals 128. On the left panel of Figure 3.24 the plots of the averaged CDFs based on eigenvalues $\lambda_i(\widehat{M})$ (with 90% confidence interval), CDF using eigenvalues $\lambda_i(M)$ and empirical CDF of $\widehat{JMI}$ are shown for quantiles of orders larger than 0.95. They almost overlap for $n = 500$, hence such sample size in that case is sufficient to ensure the adequate approximation of the distribution of $\widehat{JMI}$ by its approximate asymptotic counterpart.

In Figure 3.25 both estimated test sizes (top plots) and power (bottom plots) are shown. In the top row the results for the model M2 are shown, as in that model the global null holds, whereas in the bottom row there are results in model M1 in which the alternative to the null is true. The results in columns are presented for the same sets of parameters. All procedures except for `Fisher` do not exceed significance level significantly. The size of Fisher test is around 3 times assumed level, which might be caused by the fact, that the individual tests and their p-values are not independent. Similar effects are discussed in [2] (see Theorem 1 there). Considering the power, the procedure based on

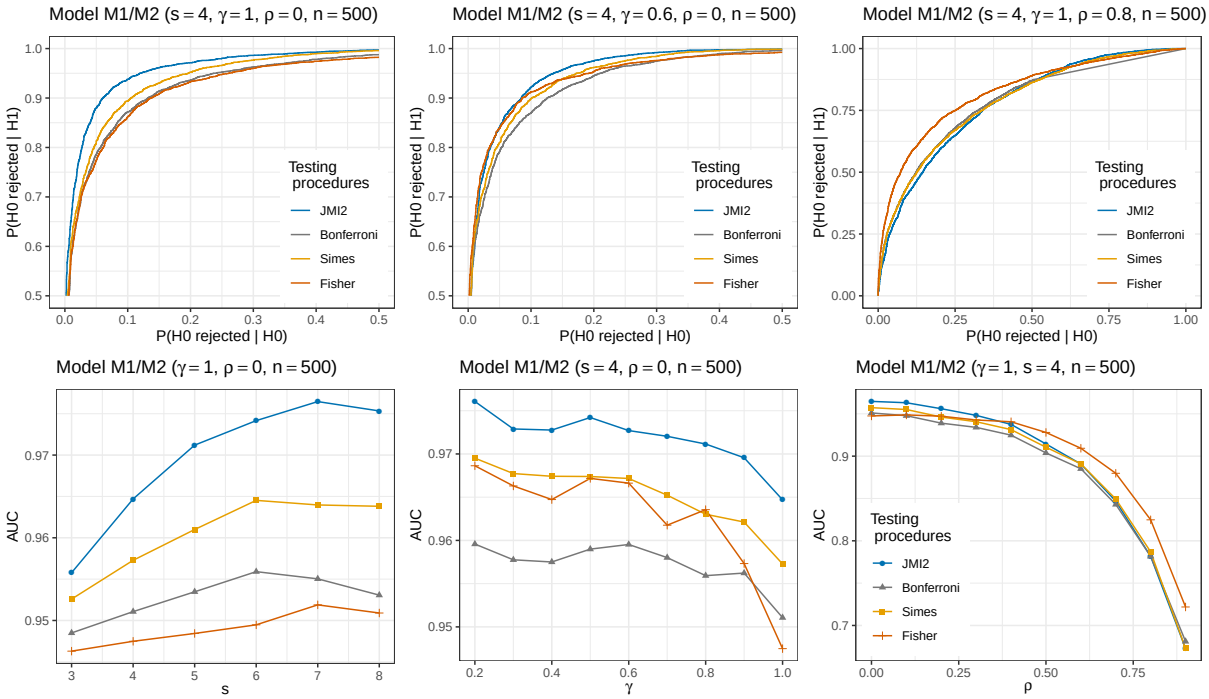


Figure 3.26: ROC-type curves for three chosen scenarios and AUC for varying parameters in testing global null hypothesis.

$\widehat{JMI}$ works the best, when Fisher test is omitted due to lack of control of significance level. The tests are analysed more objectively in Figure 3.26 using ROC-type curves. In the top row there are ROC-type curves for all the procedures. ROC-type curves are based on two models: the one for which H_0 holds (model M2) and the second for which H_1 is true (model M1), and they report *the actual* type I error and the power approximated by means of simulations for varying α. For each hypothesis the equal number of samples ($N = 5000$) are generated and for each sample p-values is calculated. Then all the p-values are sorted and the number of I type errors and valid rejections are computed. Thus in fact in this way we compare the ability of the procedures to distinguish between the null hypothesis and the alternative in a given two models scenario.

As the ROC-type curves can be plotted only for all parameters fixed for both considered models, in the bottom row the results for AUC (Area Under the Curve) are presented for parameter values corresponding to Figure 3.25. In most cases the test based on $\widehat{JMI}$ outperforms other methods even after significance level adjustment. Fisher test is better only in the case in which conditioning variables are highly correlated.

Bonferroni and Simes methods are known to perform well when there is strong evidence against a few individual null hypotheses, whereas Fisher test is designed to detect weak

violations of multiple null hypotheses. As the proposed method based on $\widehat{JMI}$ similarly to Fisher test averages evidence against the null, we expect it to perform the best in the same cases as the Fisher method. In examples presented in Figure 3.25 in the bottom panel where the power is shown, all the individual hypotheses are violated except for one, as for all $i \in \{2, 3, \ldots, s\}$ X is not independent of Y given Z_i. More examples including models, in which there are a few strong signals (and thus in these cases Bonferroni and Simes procedure should work better) are presented in [24].

3.3. Summary of experiments

In this chapter we applied the results obtained in the previous chapters for testing hypotheses. We considered two types of null hypotheses: conditional independence hypothesis

$$H_0 : X \perp\!\!\!\perp Y | Z_S$$

and globall null hypothesis

$$H_0 : \cap_{i=1}^{s} H_{0,i}, \text{ where } H_{0,i} : X \perp\!\!\!\perp Y | Z_i.$$

First we summarize results concerning conditional independence testing. In Section 3.1 resampling scenarios, criteria and conditional mutual information, and testing procedures were analysed. Testing based on the considered criteria differs from CMI-based testing. Test `exact` for CMI is asymptotically justified for all resampling scenarios and justified for finite samples for CR and permutation scenario. As for bootstrap CI and bootstrap X the justification for CMI-based test is only asymptotic, the problems with controlling type I error arise when the number of observations per cell is small and thus the precise estimation of $\hat{p}(x|z_S)\hat{p}(y|z_S)\hat{p}(z_S)$ or $\hat{p}(x|z_S)$ is difficult to achieve. The issue related to small n/s ratio (or, more exactly, ratio of n to the number of values of (X, Y, Z)) affects the estimation of CMI based on sample and resampled samples for all resampling scenarios and all testing procedures as they all require an estimator of joint probability $p(x, y, z_S)$. Thus there is a need to consider surrogate test statistics that are more adequately estimated. For this reason we also included the criteria in our study. The considered criteria have an advantage over CMI that they require only estimators of $p(x, y, z_i)$ (in the case of second

order criteria) or $p(x, y, z_i, z_j)$ (in the case of third order criteria) instead of an estimator of the joint distribution of (X, Y, Z_S). On the other hand, the asymptotic justification of the tests based on criteria fails in most cases. Firstly, if the asymptotic distribution of the resampled-based estimator of the criterion is normal than its variance might be different from its sample counterpart for all resampling scenarios except bootstrap CI. Secondly, when resampled-based statistics does not converge to normal distribution, its asymptotic law might differ significantly from the law of the statistics based on an original sample. What is more, the asymptotic distribution of criterion based on resampled samples may depend on resampling scheme applied, thus differences are not only between samples and resampled samples but also between resampled samples themselves. These issues are illustrated by the plots in Section 3.1.2. Thirdly, as the simulation study shows (see Section 3.1.4 in which controlling significance level is investigated), the centering in limit theorems is needed and cannot be omitted without consequences. The finite sample justification for criteria holds for CR and permutation scenario, thus we focus on these two approaches. *The criteria usually outperform tests based on CMI for these two resampling scenarios.* Usually lower-order criteria work better than higher order criteria (as predominantly, higher order interactions occur when lower order interactions are present usually the former are weaker than the latter), but in the models with only higher-order interactions between variables, the higher-order criteria are superior. Therefore, choosing the appropriate test statistics for dependence structure of given data is crucial. When we have prior knowledge about the data, the choice might be based on it, but in general it is an open question how to choose the test statistics adaptively. For the parametric models considered in this study $\widehat{JMI}$ based tests has shown overall good performance.

The problem of distinguishing between situations, in which the estimator based on criteria converges to normal or to quadratic form in normal variables, simplifies for $\widehat{JMI}$ when Y is a binary random variable. Also, in case when the asymptotic distribution is the distribution of quadratic form, the centering is not needed. We used these facts to construct a test for testing global null consisting of conditional independence tests. The test works better than other procedures in situations, in which weak evidence against multiple null hypotheses. We also note that in some cases the global null might be used as a proxy for testing conditional independence of X and Y given Z_S. E.g. for faithful distributions if the global null holds, then $X \perp\!\!\!\perp Y | Z_S$. In fact, if any of the individual

hypotheses holds, then $X \perp\!\!\!\perp Y | Z_S$. On the other hand, if the global null is false, then the conditional independence given Z_S still might happen, as one true individual hypothesis is sufficient to obtain the conditional independence of X and Y given Z_S.

Appendix A

Theorems

A.1. Lemma used in Chapter 1

Lemma A.1.1 below and its proof comes from [20].

Lemma A.1.1. *Let $Y \in \{0,1\}$ be a binary random variable and $X, Z \in \mathbb{N}_+$ be discrete variables. If for all $y \in \{0,1\}$ and $x, z \in \mathbb{N}_+$ we have:*

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

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

1) 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 not independent).

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

Conversely, if either of the above conditions is true then (A.1) holds.

Proof. First we observe that for all $x, z \in \mathbb{N}_+$ we have

$$\begin{aligned}\sum_{y=0}^{1} a_{xy} P(Y=y, Z=z) &= P(Z=z) \sum_{y=0}^{1} a_{xy} P(Y=y|Z=z) \\ &= P(Z=z) \sum_{y=0}^{1} \frac{P(X=x, Y=y|Z=z)}{P(X=x|Z=z)} = P(Z=z). \end{aligned} \tag{A.2}$$

This means that for all x we have

$$\begin{aligned}\sum_{y=0}^{1} a_{xy}P(Y=y) &= \sum_{z\in\mathbb{N}_+}\sum_{y=0}^{1} a_{xy}P(Y=y,Z=z)\\ &= \sum_{z\in\mathbb{N}_+} P(Z=z) = 1.\end{aligned} \tag{A.3}$$

Hence

$$a_{x1} = \frac{1-a_{x0}P(Y=0)}{P(Y=1)}. \tag{A.4}$$

From (A.2) it follows that for all x we have

$$\begin{cases} P(Z=z) = P(Y=0,Z=z)a_{x0} + P(Y=1,Z=z)a_{x1}, \\ P(Z=z) = P(Y=0,Z=z) + P(Y=1,Z=z). \end{cases} \tag{A.5}$$

Subtracting second equation from the first and using (A.4) yields

$$\begin{aligned}0 &= P(Y=0,Z=z)(a_{x0}-1) + P(Y=1,Z=z)\left(\frac{1-a_{x0}P(Y=0)}{P(Y=1)} - 1\right)\\ &= P(Y=0,Z=z)(a_{x0}-1) + P(Y=1,Z=z)(1-a_{x0})\frac{P(Y=0)}{P(Y=1)}.\end{aligned}$$

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 (A.4)), then the above equation reduces to:

$$P(Y=0,Z=z) = P(Y=1,Z=z)\frac{P(Y=0)}{P(Y=1)}. \tag{A.6}$$

This yields

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

Analogously, we obtain

$$P(Z=z) = \frac{P(Y=0,Z=z)}{P(Y=0)}. \tag{A.7}$$

Thus Y and Z are independent. This means that $P(Y = y, Z = z) = P(Y = y)P(Z = z)$. Inserting this equation into (A.1) yields

$$a_{xy} = \frac{P(X = x, Y = y, Z = z)}{P(X = x, Z = z)P(Y = y)}. \tag{A.8}$$

Equivalently,

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

Hence

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

It follows that

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

Thus, inserting this into (A.8), we obtain

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

what is equivalent to conditional independence of Y and Z given X.

2) If $a_{x0} = 1$ for all x, then in view of (A.4) we obtain $a_{x1} = 1$ for all x. This implies conditional independence of (X, Y) given Z. To see the converse note that a_{xy} in (A.1) equals 1 when 2) is true and $a_{xy} = p(x, y)/(p(x)p(y))$ when 1) holds. □

A.2. Theorems used in Chapter 2

Theorem A.2.1 (Multivariate Berry-Esseen theorem [4]). *Let $Q_1, Q_2, \ldots, Q_n \in \mathbb{R}^d$ be random independent vectors having zero mean. Let $W = \sum_{i=1}^n Q_i$, $\mathrm{Cov}(W) = I$ and $Z \sim \mathcal{N}(0, I)$. Then for all convex A there exist a positive constant K_d (which depends on the dimension d) such that*

$$|P(W \in A) - P(Z \in A)| \leq K_d \sum_{i=1}^n \mathbb{E}||Q_i||^3$$

and $K_d = cd^{1/4}$ (c is an absolute positive constant not depending on dimension).

Theorem A.2.2 (Normal approximation to the hypergeometric distribution [21])**.** *Let X_r be a random variable having the hypergeometric distribution with parameters (n_r, M_r, N_r), namely*

$$P(X_r = x) = \frac{\binom{M_r}{x}\binom{N_r - M_r}{n_r - x}}{\binom{N_r}{n_r}}$$

and $r \in \mathcal{N}$. We denote the sampling fractions by $p_r = \frac{M_r}{N_r}$ and $f_r = \frac{n_r}{N_r}$. We assume $1 \leq M_r < N_r$ and $1 \leq n_r < N_r$ and $N_r^{-1} = o(1)$ as $r \to \infty$. Let $\sigma_r^2 = N_r p_r (1-p_r) f_r (1-f_r)$. Then there exits a normal random variable $W \sim \mathcal{N}(\mu, \sigma^2)$ such that

$$\sup_{x \in \mathcal{R}} \left| P\left(\frac{X_r - n_r p_r}{\sigma_r} \leq x \right) - P(W \leq x) \right| \to 0 \quad as \quad r \to \infty$$

if and only if

$$\sigma_r^2 \to \infty \quad as \quad r \to \infty \tag{A.9}$$

When (A.9) *holds, one must have $\mu = 0$ and $\sigma = 1$.*

Appendix B

List of Symbols

X, Y, Z	discrete random variables
$\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$	supports of random variables X, Y and Z, respectively
Z_S	vector of random variables such that their indices are in the set S (when $S = \{1, 2, \ldots, \lvert S\rvert\}$, then $Z_S = (Z_1, Z_2, \ldots, Z_{\lvert S\rvert})$)
$p(x, y, z)$	joint distribution of the triple of random variables (X, Y, Z); also $P(X = x, Y = y, Z = z)$
$p_{ci}(x, y, z)$	joint distribution of the triple of random variables (X, Y, Z) factorised in the following way: $p(x\lvert z)p(y\lvert z)p(z)$ (*ci* stands for *conditional independence*)
$\hat{p}(x, y, z)$	estimator of the joint distribution of (X, Y, Z) based on fractions
$\hat{p}^*(x, y, z)$	estimator of the joint distribution of (X, Y, Z) computed for resampled sample based on fractions
$n(x, y, z)$	number of observed triples (x, y, z) in a sample, i.e. $n(x, y, z) = \sum_{i=1}^{n} \mathbb{I}(X_i = x, Y_i = y, Z_i = z)$
$n^*(x, y, z)$	number of observed triples (x, y, z) in a resampled sample, i.e. $n^*(x, y, z) = \sum_{i=1}^{n} \mathbb{I}(X_i^* = x, Y_i^* = y, Z_i^* = z)$
$(p(x, y, z))_{x,y,z}$	vector of probabilities for all the triples (x, y, z) such that $(x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$
P^K	Kirkwood superposition approximation of the distribution P (Definition 1.2.3)
$\xrightarrow{d}$	convergence in distribution
$o_p(z_n)$	random variable of smaller order than z_n: $P(\lvert o_p(z_n)/z_n\rvert \leq \varepsilon) \to 0$ as $n \to +\infty$
Σ	covariance matrix
$\Sigma_{x,y,z}^{x',y',z'}$	element of a matrix Σ with row index (x, y, z) and column index (x', y', z'), where $(x, y, z), (x', y', z') \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$
x', Σ'	transpose of a vector x and a matrix Σ, respectively

$D_f(x)$	gradient of a function f at x
$H_f(x)$	Hessian of a function f at x
H_0, H_1	null and alternative hypotheses, respectively

B.1. Information-theoretic measures

$H(X)$	entropy of random variable X (Definition 1.1.1)
$H(X\|Y)$	conditional entropy of random variable X given Y (Definition 1.1.2)
$I(X,Y)$, MI	mutual information of random variables X an Y (Definition 1.1.4)
$I(X,Y\|Z)$, CMI	conditional mutual information of random variables X and Y given Z (Definition 1.1.5)
$II(X,Y,Z)$	3-way interaction information of random variables X, Y and Z (Definition 1.2.1)
$II(Z_1,Z_2,\ldots,Z_k)$	k-way interaction information of random variables $Z_1,Z_2,\ldots,Z_k$ (Definition 1.2.2)
$D_{KL}(p\|\|q)$	Kullback-Leibler divergence (Definition 1.1.3)
$I_{\beta,\gamma}(X,Y\|Z)$	generalized feature selection criterion, where β and γ are vectors of parameters (Definition 1.3.1)
$J_{\beta,\gamma}(X,Y\|Z)$	generalized feature selection criterion, where β and γ are scalar parameters (definition in (1.19))
$JMI, JMI2$	Joint Mutual Information criterion (of order 2; definition in (1.22))
$JMI3$	Joint Mutual Information criterion (of order 3; definition in (1.24))
$CIFE$, $SECMI$, $SECMI2$	Conditional Infomax Feature Extraction/Short Expansion of Conditional Mutual Information criterion (of order 2; definition in (1.25))
$SECMI3$	Short Expansion of Conditional Mutual Information criterion (of order 3; definition in (1.26))

Bibliography

[1] Agresti, A. (2013). *Categorical Data Analysis*. Wiley, Chicester.

[2] Bates, S., Candès, E. J., Lei, L., Romano, Y., and Sesia, M. (2021). Testing for outliers with conformal p-values. `https://arxiv.org/pdf/2104.08279.pdf`.

[3] Battiti, R. (1994). Using mutual information for selecting features in supervised neural-net learning. *IEEE Transactions on Neural Networks*, 5(4):537–550.

[4] Bentkus, V. (2005). A Lyapunov-type bound in $\mathbb{R}^d$. *Theory of Probability and its Applications*, 49(2):311–371.

[5] Berrett, T. B., Wang, Y., Barber, R. F., and Samworth, R. J. (2020). The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 82(1):175–197.

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

[7] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer Berlin Heidelberg.

[8] Candès, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold: model-X knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 80(3):551–577.

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

[10] Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

[11] Duchesne, P. and Lafaye de Micheaux, P. (2010). Computing the distribution of quadratic forms: Further comparisons between the Liu-Tang-Zhang approximation and exact methods. *Computational Statistics and Data Analysis*, 54:858–862.

[12] Dudoit, S. and van der Laan, M. J. (2009). *Multiple Testing Procedures with Applications to Genomics*. Springer.

[13] Fano, R. M. (1961). *Transmission of Information: A Statistical Theory of Commu-*

nication. MIT Press Classics. MIT Press.

[14] Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver & Boyd, Edinburgh.

[15] Halton, J. H. (1969). A rigorous derivation of the exact contingency formula. *Mathematical Proceedings of the Cambridge Philosophical Society*, 65(2):527–530.

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

[17] Knijnenburg, T. A., Wessels, L. F. A., Reinders, M. J. T., and Shmulevich, I. (2009). Fewer permutations, more accurate P-values. *Bioinformatics*, 25(12):i161–i168.

[18] Kubkowski, M., Łazęcka, M., and Mielniczuk, J. (2020). Distributions of reduced-order general dependence measure and conditional independence testing. In *ICCS2020, LNCS 12143*, pages 51–63. Springer Nature.

[19] Kubkowski, M. and Mielniczuk, J. (2021). Asymptotic distributions of empirical interaction information. *Methodology and Computing in Applied Probability*, 23.

[20] Kubkowski, M., Mielniczuk, J., and Teisseyre, P. (2021). How to gain on power: novel conditional independence tests based on short expansion of conditional mutual information. *Journal of Machine Learning Research*, 22:1–57.

[21] Lahiri, S. and Chatterjee, A. (2007). A Berry-Esseen theorem for hypergeometric probabilities under minimal conditions. *Proceedings of the American Mathematical Society*, 135(5):1535–1545.

[22] Łazęcka, M., Kołodziejek, B., and Mielniczuk, J. (2022). Analysis of conditional randomisation and conditional permutation schemes with application to conditional independence testing. Manuscript in preparation.

[23] Łazęcka, M. and Mielniczuk, J. (2020). Analysis of information-based nonparametric variable selection criteria. *Entropy*, 22(9).

[24] Łazęcka, M. and Mielniczuk, J. (2021). Multiple testing of conditional independence using information theoretic-approach. In *Proceedings of Modelling Decisions for Artificial Intelligence, LNAI 12898*, pages 51–63. Springer Nature.

[25] Lehmann, E. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer New York, New York, NY.

[26] Lin, D. and Tang, X. (2006). 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.

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

[28] McGill, W. J. (1954). Multivariate Information Transmission. *Psychometrika*, 19(2):97–116.

[29] Mielniczuk, J. and Teisseyre, P. (2019). Stopping rules for mutual information-based

feature selection. *Neurocomputing*, 358:255–274.

[30] Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(1):1226–1238.

[31] Politis, D., Wolf, D., Romano, J., Wolf, M., Bickel, P., Diggle, P., and Fienberg, S. (1999). *Subsampling*. Springer Series in Statistics. Springer New York.

[32] Rota, G. C. (1964). On the foundations of combinatorial theory I. Theory of Möbius Functions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 2(4):340–368.

[33] Sadeghi, K. (2017). Faithfulness of probability distributions and graphs. *J. Mach. Learn. Res.*, 18(1):5429–5457.

[34] Sarkar, S. K. (1998). Some probability inequalities for ordered MTP_2 random variables: a proof of the Simes conjecture. *The Annals of Statistics*, 26(2):494 – 504.

[35] Seber, G. A. F. (2008). *A Matrix Handbook for Statisticians*, volume 15. John Wiley & Sons.

[36] Sechidis, K., Azzimonti, L., Pocock, A., Corani, G., Weatherall, J., and Brown, G. (2019). Efficient feature selection using shrinkage estimators. *Machine Learning*, 108:1261–1286.

[37] Shah, R. and Peters, J. (2018). The hardness of conditional independence testing and the generalised covariance measure. *Annals of Statistics*, 48:1514–1538.

[38] Simes, R. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73:751–754.

[39] Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *The Annals of Statistics*, 9(6):1403–1433.

[40] Tsamardinos, I., Aliferis, C. F., and Statnikov, A. R. (2003). Algorithms for large scale Markov blanket discovery. In *FLAIRS Conference*, pages 376–381.

[41] Tsamardinos, I. and Borboudakis, G. (2010). Permutation testing improves bayesian network learning. In *Machine Learning and Knowledge Discovery in Databases*, pages 322–337, Berlin, Heidelberg. Springer Berlin Heidelberg.

[42] Vergara, J. and Estevez, P. (2014). A review of feature selection methods based on mutual information. *Neural Computing and and Applications*, 24(1):175–16.

[43] Yang, H. and Moody, J. (1999). Data visualization and feature selection: new algorithms for nongaussian data. *Advances in Neural Information Processing Systems*, 12:687–693.

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