TEST (2023) 32:1459–1478
https://doi.org/10.1007/s11749-023-00878-7

ORIGINAL RESEARCH PAPER

Analysis of conditional randomisation and permutation schemes with application to conditional independence testing

Małgorzata Łazęcka[1,2,3] · Bartosz Kołodziejek[1] · Jan Mielniczuk[1,2]

Received: 3 October 2022 / Accepted: 1 August 2023 / Published online: 12 September 2023

Abstract

We study properties of two resampling scenarios: Conditional Randomisation and Conditional Permutation schemes, which are relevant for testing conditional independence of discrete random variables X and Y given a random variable Z. Namely, we investigate asymptotic behaviour of estimates of a vector of probabilities in such settings, establish their asymptotic normality and ordering between asymptotic covariance matrices. The results are used to derive asymptotic distributions of the empirical Conditional Mutual Information in those set-ups. Somewhat unexpectedly, the distributions coincide for the two scenarios, despite differences in the asymptotic distributions of the estimates of probabilities. We also prove validity of permutation p-values for the Conditional Permutation scheme. The above results justify consideration of conditional independence tests based on resampled p-values and on the asymptotic chi-square distribution with an adjusted number of degrees of freedom. We show in numerical experiments that when the ratio of the sample size to the number of possible values of the triple exceeds 0.5, the test based on the asymptotic distribution with the adjustment made on a limited number of permutations is a viable alternative to the exact test for both the Conditional Permutation and the Conditional Randomisation scenarios. Moreover, there is no significant difference between the performance of

✉ Małgorzata Łazęcka
m.lazecka@ipipan.waw.pl

Bartosz Kołodziejek
bartosz.kolodziejek@pw.edu.pl

Jan Mielniczuk
jan.mielniczuk@ipipan.waw.pl

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

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

[3] Present Address: Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Banacha 2, 02-097, Warsaw, Poland

exact tests for Conditional Permutation and Randomisation schemes, the latter requiring knowledge of conditional distribution of X given Z, and the same conclusion is true for both adaptive tests.

Keywords Conditional independence · Conditional mutual information · Kullback–Leibler divergence · Conditional randomisation and permutation

Mathematics Subject Classification 62E20 · 62G10 · 62G09

1 Introduction

Checking for conditional independence is a crucial ingredient of many Machine Learning algorithms, such as those designed to learn structure of graphical models or select active predictors for the response in a regression task, see e.g. Koller and Sahami (1995); Aliferis et al. (2003); Fu and Desormais (2017). In a greedy approach to the variable selection for the response, one needs to verify whether predictor X is conditionally independent of the response, say, Y, given Z (denoted by $X \perp\!\!\!\perp Y|Z$), where Z is a vector of predictors already chosen as active ones and X is any of the remaining candidates. When conditional independence holds, then X is deemed irrelevant; when the test fails, the candidate that 'most strongly' contradicts it, is chosen.

Verification of conditional independence of discrete-valued random variables uses a specially designed test statistic, say, T, such as Pearson χ^2 chi-square statistic or Conditional Mutual information CMI. The value of the statistic, calculated for the data considered, is compared with a benchmark distribution. Usually, as a benchmark distribution one either uses the asymptotic distribution of T under conditional independence or its distribution (or approximation thereof) obtained for resampled samples which conform to conditional independence. More often than not, the asymptotic test is too liberal, especially for small sample sizes, what leads to acceptance of too many false positive predictors. That is why resampling methods are of interest in this context (for other approaches see e.g. Candès et al. (2018); Watson and Wright (2021); Kubkowski et al. (2021) and references therein). The resampling is commonly performed by either permuting values of X on each strata of Z, see e.g. Tsamardinos and Borboudakis (2010), or by replacing original values of X by values generated according to conditional distribution $P_{X|Z}$ if the distribution is known (we will refer to the former as Conditional Permutation and to the latter as Conditional Randomisation, Candès et al. (2018)). Although the validity of resampling approach in the latter case can be established fairly easily (see ibidem), it was previously unknown for the conditional permutation approach as well as for the asymptotic approach in both settings. Based on the proved asymptotic results, we propose a modified asymptotic test that uses a χ^2 distribution with an adjusted number of degrees of freedom as the benchmark distribution. The major contributions of the paper are thus as follows: we (i) establish validity of the resampling method for conditional permutation approach; (ii) derive the asymptotic distributions of the estimated vector of probabilities and of the estimator of CMI under both resampling scenarios; (iii) compare asymptotic and resampled p-values approach in numerical experiments. In numerical experiments, we show that for the models considered and a ratio of the sample size to the size of the support of (X, Y, Z) larger than 0.5, the test based on the asymptotic distribution with adjustments based on a limited number of permutations performs equally well or better than the exact test for both the Conditional Permutation and the Conditional Randomisation scenarios. Moreover, there is no significant difference in the performance of the exact tests for Conditional Permutation and Conditional Randomisation scheme, the latter requiring knowledge of the conditional distribution of X given Z. The same is true for both adaptive tests.

As the null hypothesis of conditional independence is composite, an important question arises: how to control the type I error by choosing adequate conditionally independent probability structures. In the paper, we adopt a novel approach to address this issue, which involves investigating those null distributions that are Kullback–Leibler projections of probability distributions for which power is investigated.

An important by-product of the investigation in (i) is that we establish asymptotic normality of the normalised and centred vector having a multivariate hyper-geometric or generalised hyper-geometric distribution for the conditional permutation scheme.

2 Preliminiaries

We consider a discrete-valued triple (X, Y, Z), where $X \in \mathcal{X}$, $Y \in \mathcal{Y}$, $Z \in \mathcal{Z}$, and all variables are possibly multivariate. Assume that $P(X = x, Y = y, Z = z) = p(x, y, z) > 0$ holds for any $(x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. Moreover, we let $p(x, y|z) = P(X = x, Y = y|Z = z)$, where $p(z) = P(Z = z)$ and define $p(x|z)$ and $p(y|z)$ analogously. We will denote by I, J, K the respective sizes of supports of X, Y and Z: $|\mathcal{X}|= I$, $|\mathcal{Y}|= J$, $|\mathcal{Z}|= K$. As our aim is to check conditional independence, we will use Conditional Mutual Information (CMI) as a measure of conditional dependence (we refer to Cover and Thomas (2006) for basic information-theoretic concepts such as entropy and mutual information). Conditional Mutual Information is a non-negative number defined as

$$\begin{aligned}\mathrm{CMI} = I(Y; X|Z) &= \sum_{z\in\mathcal{Z}} p(z) I(X; Y|Z = z)\\ &= \sum_{z\in\mathcal{Z}} p(z) \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p(x, y|z) \log \frac{p(x, y|z)}{p(x|z)p(y|z)}\\ &= \sum_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} p(x, y, z) \log \frac{p(x, y|z)}{p(x|z)p(y|z)}, \end{aligned} \tag{1}$$

where $I(Y; X|Z = z)$ is defined as the inner sum in the middle line above.
We note that $I(Y; X|Z = z)$ is defined as the mutual information (MI) between $P_{Y,X|Z=z}$ and the product of $P_{Y|Z=z}$ and $P_{X|Z=z}$, and $I(X; Y|Z)$ is its probabilistic average over the values of Z. The non-negativity of $I(Y; X|Z)$ is due to the non-negativity of $I(X; Y|Z = z)$ and is a consequence of Jensen's inequality (see formula (2.92) in Cover and Thomas (2006)). As MI is the Kullback–Leibler diver-

gence between the joint and the product distribution, it follows from the properties of Kullback–Leibler divergence that

$$I(Y; X|Z) = 0 \iff X \text{ and } Y \text{ are conditionally independent given } Z.$$

This is a powerful property, not satisfied for other measures of dependence, such as the partial correlation coefficient in the case of continuous random variables. The conditional independence of X and Y given Z will be denoted by $X \perp\!\!\!\perp Y|Z$ and referred to as CI. We note that since $I(Y; X|Z)$ is defined as a probabilistic average of $I(Y; X|Z = z)$ over $Z = z$, it follows that

$$I(Y; X|Z) = 0 \iff I(Y; X|Z = z) = 0 \;\text{ for any } z \text{ in the support of } Z.$$

This is due to (1) as $I(Y; X|Z = z)$ is non-negative. Let $(X_i, Y_i, Z_i)_{i=1}^n$ be an independent sample of copies of (X, Y, Z) and consider the unconstrained maximum likelihood estimator of the probability mass function (p.m.f.) $((p(x, y, z))_{x,y,z}$ based on this sample being simply a vector of fractions $((\hat{p}(x, y, z))_{x,y,z} = (n(x, y, z)/n)_{x,y,z}$, where $n(x, y, z) = \sum_{i=1}^n \mathbb{I}(X_i = x, Y_i = y, Z_i = z)$. In the following, we will examine several resampling schemes that involve generating new data such that they satisfy CI hypothesis for the *fixed* original sample. Extending the observed data to an infinite sequence, we will denote by P^* the conditional probability related to the resampling schemes considered, given the sequence $(X_i, Y_i, Z_i)_{i=1}^{\infty}$.

3 Resampling scenarios

We first discuss the Conditional Permutation scheme, which can be applied to conditional independence testing. We then establish validity of the *p*-values based on this scheme, and the form of asymptotic distribution for the sample proportions, which is used later to derive asymptotic distribution of empirical CMI.

3.1 Conditional Permutation (CP) scenario

We assume that the sample $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = (X_i, Y_i, Z_i)_{i=1}^n$ is given and we consider CI hypothesis $H_0 : X \perp\!\!\!\perp Y|Z$. The Conditional Permutation (CP) scheme, used e.g. in Tsamardinos and Borboudakis (2010), is a generalisation of a usual permutation scenario applied to test unconditional independence of X and Y. It consists in the following: for every value z_k of Z appearing in the sample, we consider the strata corresponding to this value, namely $P_k = \{j : Z_j = z_k\}$. CP sample is obtained from the original sample by replacing (X_i, Y_i, Z_i) for $i \in P_k$ by $(X_{\pi^k(i)}, Y_i, Z_i)$, where π^k is a randomly and uniformly chosen permutation of P_k and π^k are independent (see Algorithm 1). Thus on every strata $Z = z$, we randomly permute values of corresponding X independently of values of Y.It is, in fact, sufficient to permute only the values of X to ensure conditional independence, which follows from the fact that for any discrete random variable (X, Y) we have that X is independent of $\sigma(Y)$,

where σ is a randomly and uniformly chosen permutation of the values of Y such that $\sigma \perp\!\!\!\perp (X, Y)$. The pseudo-code of the algorithm is given below.
We consider the family of all permutations Π of all permutations π of $\{1, \ldots, n\}$ which preserve each of P_k i.e. π is composed of π^k's, i.e. such that their restriction to every P_k is a permutation of P_k. The number of such permutations is $\prod_z n(z)!$, where $n(z) = \sum_{i=1}^n \mathbb{I}(Z_i = z)$.

Algorithm 1: Conditional Permutation algorithm

```
Input: $(X_i, Y_i, Z_i)_{i=1}^n$
Output: $(X_i^*, Y_i, Z_i)_{i=1}^n$
for $k \in \{1, 2, \ldots, K\}$ do
    $\pi^k \leftarrow$ a random permutation of $P_k$ ;
    for $i \in P_k$ do
        $X_i^* \leftarrow X_{\pi^k(i)}$ ;
```

3.1.1 Validity of *p*-values for CP scenario

We first prove the result which establishes validity of resampled p-values for any statistic for the Conditional Permutation scheme. Let $X_i^* = X_{\pi^k(i)}$ for $i \in P_k$ and denote by $(\mathbf{X}^*, \mathbf{Y}, \mathbf{Z})$ the sample (X_i^*, Y_i, Z_i), $i = 1, \ldots, n$. Let $T(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{Z}_n)$ be any statistic defined on the underlying sample $(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{Z}_n) = (X_i, Y_i, Z_i)_{i=1}^n$ which is used for CI testing. We choose B independent permutations in Π, construct B corresponding resampled samples by CP scenario $(\mathbf{X}_{n,b}^*, \mathbf{Y}_{n,b}, \mathbf{Z}_{n,b})$ for $b = 1, 2, \ldots, B$ and calculate the values of statistic $T_b^* = T(\mathbf{X}_{n,b}^*, \mathbf{Y}_{n,b}, \mathbf{Z}_{n,b})$. The pertaining p-value based on CP resampling is defined as

$$\frac{1 + \sum_{b=1}^B \mathbb{I}(T \le T_b^*)}{1 + B}.$$

Thus, up to ones added to the numerator and the denominator, the resampling p-value is defined as the fraction of T_b^* not smaller than T (ones are added to avoid null p-values). Although p-values based on CP scheme have been used in practice (see e.g. Tsamardinos and Borboudakis (2010)) to the best of our knowledge, their validity has not been established previously, to the best of our knowledge.

Theorem 1 *(Validity of p-values for CP scheme)*
If the null hypothesis $H_0 : X \perp\!\!\!\perp Y|Z$ holds, then

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

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

The result implies that if the testing procedure rejects H_0 when the resampling p-value does not exceed α its level of significance is also controlled at α. The proof is based on exchangeability of $T, T_1^*, \ldots, T_B^*$ and is given in the online Appendix.

3.1.2 Asymptotic distribution of sample proportions for Conditional Permutation method

We define $\hat{p}^*$ to be an empirical p.m.f. based on sample $(\mathbf{X}^*, \mathbf{Y}, \mathbf{Z})$: $\hat{p}^*(x, y, z) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}(X_{\pi(i)} = x, Y_i = y, Z_i = z)$, where $\pi \in \Pi$ is randomly and uniformly chosen from Π. Similarly to $n(x, y, z)$ we let $n(y, z) = \sum_{i=1}^n \mathbb{I}\{Y_i = y, Z_i = z\}$ and $n(x, z)$ is defined analogously. We first prove

Theorem 2 *(i) Joint distribution of the vector $(n\hat{p}^*(x, y, z))_{x,y,z}$ given $(X_i, Y_i, Z_i)_{i=1}^n$ is as follows:*

$$P\big(n\hat{p}^*(x, y, z) = k(x, y, z),\ (x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \mid (X_i, Y_i, Z_i)_{i=1}^n = (x_i, y_i, z_i)_{i=1}^n\big)$$
$$= \prod_{z \in \mathcal{Z}} \left(\frac{\prod_{x \in \mathcal{X}} n(x, z)! \prod_{y \in \mathcal{Y}} n(y, z)!}{n(z)! \prod_{(x,y) \in \mathcal{X} \times \mathcal{Y}} k(x, y, z)!} \right), \tag{2}$$

where $(k(x, y, z))_{x,y,z}$ is a sequence taking values in nonnegative integers such that $\sum_x k(x, y, z) = n(y, z)$ and $\sum_y k(x, y, z) = n(x, z)$, otherwise $P\big(n\hat{p}^*(x, y, z) = k(x, y, z) \mid (X_i, Y_i, Z_i)_{i=1}^n = (x_i, y_i, z_i)_{i=1}^n\big) = 0$.
(ii) Asymptotic behaviour of the vector $(\hat{p}^*(x, y, z))_{x,y,z}$ conditionally on $(X_i, Y_i, Z_i)_{i=1}^\infty$ is given by the following weak convergence

$$\sqrt{n}\left(\hat{p}^*(x, y, z) - \hat{p}(x|z)\hat{p}(y|z)\hat{p}(z)\right)_{x,y,z} \xrightarrow{d} N(0, \Sigma), \tag{3}$$

for almost all $(X_i, Y_i, Z_i)_{i=1}^\infty$, where $\Sigma_{x,y,z}^{x',y',z'}$, element of Σ corresponding to row index x, y, z and column index x', y', z', is defined by

$$\Sigma_{x,y,z}^{x',y',z'} = \mathbb{I}(z = z')p(z)\Big(p(x|z)p(y|z)p(x'|z)p(y'|z) - \mathbb{I}(x = x')p(x|z)p(y|z)p(y'|z)$$
$$- \mathbb{I}(y = y')p(x|z)p(x'|z)p(y|z) + \mathbb{I}(x = x', y = y')p(x|z)p(y|z)\Big). \tag{4}$$

We stress that (2) is a deterministic equality describing the distribution of $n\hat{p}^*$: for $k(x, y, z)_{x,y,z}$ such that $\sum_x k(x, y, z) = n(y, z)$ and $\sum_y k(x, y, z) = n(x, z)$ (where $n(x, z)$ and $n(y, z)$ are based on the original sample) corresponding value of p.m.f. is given by the left-hand side, otherwise it is 0.

Proof (i) The proof is a simple generalisation of the result of Halton (1969) who established the form of the conditional distribution of a bivariate contingency table given its marginals and we omit it.

(ii) In view of (2) subvectors

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

are independent given $(X_i, Y_i, Z_i)_{i=1}^n$, thus in order to prove (3) it is sufficient to prove analogous result when the stratum $Z = z$, i.e. for the unconditional permutation scenario. Note that since we consider conditional result given $(X_i, Y_i, Z_i)_{i=1}^\infty$,the strata sample sizes $n(z_i)$ are deterministic and such that $n(z_i)/n \to P(Z = z_i)$ for almost every such sequence. The needed result is stated below. □

Theorem 3 *Assume that n_{ij}, $i = 1, \dots, I$, $j = 1, \dots, J$ are elements of $I \times J$ contingency table based on iid sample of n observations pertaining to a discrete distribution (p_{ij}) satisfying $p_{ij} = p_{i.}p_{.j}$. Then we have provided $p_{ij} > 0$ for all i, j that*

$$\frac{1}{\sqrt{n}}\Big(n_{ij} - \frac{n_{i.}n_{.j}}{n}\Big)_{i,j} \mid (n_{.i}, n_{j.})_{i,j} \xrightarrow{d} N(0, \Sigma), \tag{5}$$

where $\Sigma = (\Sigma_{i,j}^{k,l})$ and $\Sigma_{i,j}^{k,l} = p_{i.}(\delta_{ik} - p_{.k})p_{.j}(\delta_{jl} - p_{l.})$.

Remark 1 Let $(X_i^*, Y_i)_{i=1}^n$ be a sample obtained from $(X_i, Y_i)_{i=1}^n$ by a random (unconditional) permutation of values of X_i and $\hat{p}^*(x, y)$ be an empirical p.m.f. corresponding to $(X_i^*, Y_i)_{i=1}^n$. Then obviously (n_{ij}/n) and $(\hat{p}^*(x, y))$ follow the same distribution and (5) is equivalent to

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

Moreover, the elements of Σ can be written as (compare (4))

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

Remark 2 Matrix Σ introduced above has the rank $(I-1) \times (J-1)$ and can be written using the tensor products as $(\mathrm{diag}(\alpha) - \alpha \otimes \alpha) \otimes (\mathrm{diag}(\beta) - \beta \otimes \beta)$, where $\alpha = (p_{i.})_i$ and $\beta = (p_{.j})_j$.

The proof of Theorem 3 follows from a weak convergence result for table-valued hypergeometric distributions and is important in its own right.

Let R denote the range of indices (i, j): $R = \{1, \dots, I\} \times \{1, \dots, J\}$. For $x = (x_i, \dots, x_d)^\top \in \mathbb{R}^d$ we write $|x| = \sum_{i=1}^d x_i$. Let $T_d = \{x \in (0, 1)^d : |x| = 1\}$ denote the simplex in $\mathbb{R}^d$.

Lemma 4 *Let $a_r = (a_1^{(r)}, \dots, a_I^{(r)})^\top$ and $b_r = (b_1^{(r)}, \dots, b_J^{(r)})^\top$ be two vectors with coordinates being natural numbers such that*

$$n_r := |a_r| = |b_r|.$$

Suppose that the law of $W_r = (W_{ij}^{(r)})_{(i,j)\in R}$ is given by

$$P(W_r = k) = \frac{\prod_{i=1}^I a_i^{(r)}! \prod_{j=1}^J b_j^{(r)}!}{n_r! \prod_{(i,j)\in R} k_{ij}!} \tag{6}$$

for $k = (k_{ij})_{(i,j)\in R}$ *such that* $k_{ij} \in \{0, 1, \ldots\}$,

$$\sum_{j=1}^{J} k_{ij} = a_i^{(r)} \quad and \quad \sum_{i=1}^{I} k_{ij} = b_j^{(r)}, \qquad (i, j) \in R.$$

Assume that as $r \to \infty$,

$$n_r \to \infty, \quad a_r/n_r \to \alpha = (\alpha_1, \ldots, \alpha_I) \in T_I, \quad b_r/n_r \to \beta = (\beta_1, \ldots, \beta_J) \in T_J.$$

Then,

$$\frac{1}{\sqrt{n_r}} \left(W_r - \frac{1}{n_r} a_r b_r^\top \right) \xrightarrow{d} N(0, \Sigma),$$

where $\Sigma = (\Sigma_{i,j}^{k,l})$ *and*

$$\Sigma_{i,j}^{k,l} = \alpha_i(\delta_{ik} - \alpha_k)\beta_j\left(\delta_{jl} - \beta_l\right). \tag{7}$$

The proof of Lemma 4 is relegated to the online Appendix. Theorem 3 is a special case of Lemma 4 with $a_r = (n_{i.})_i$, $b_r = (n_{j.})_j$, $r = n$ on a probability space $(\Omega, \mathcal{F}, P_{\mathbf{n}})$, where $P_{\mathbf{n}} = P(\cdot \mid (n_{.i}, n_{j.})_{i,j})$ is a regular conditional probability.

3.2 Conditional Randomisation scenario

We now consider the Conditional Randomisation (CR) scheme, popularised in Candès et al. (2018). This scheme assumes that the conditional distribution $P_{X|Z}$ is known, and the resampled sample is $(X_i^*, Y_i, Z_i)_{i=1}^n$, where X_i^* is independently generated according to the conditional distribution $P_{X|Z=z_i}$ and independently of $(\mathbf{X}, \mathbf{Y})$. The assumption that $P_{X|Z}$ is known is frequently considered (see e.g. Candès et al. (2018) or Berrett et al. (2020)) and is realistic in the situations when a large database containing observations of unlabelled data (X, Z) is available, upon which an accurate approximation of $P_{X|Z}$ is based. Theorem 4 in Berrett et al. (2020) justifies the robustness of the type I error for the corresponding testing procedure.

We note that the conclusion of Theorem 1 is also valid for CR scenario (cf. Candès et al. (2018), Lemma 4.1).

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

Theorem 5 *Conditionally on* $(X_i, Y_i, Z_i)_{i=1}^\infty$, *we have almost surely that*

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

where

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

The proof which is based on multivariate Berry-Esseen theorem is moved to the online Appendix.

Remark 3 Recall that Σ and $\tilde{\Sigma}$ are the asymptotic covariance matrices for Conditional Permutation and Conditional Randomisation scenarios, respectively. Intuitively, the amount of variability introduced by resampling should be smaller in the case of the Conditional Permutation scheme as it retains the empirical conditional distribution of X given Z. This is indeed the case and is reflected in the covariance matrix ordering. Namely, we have that $\Sigma \leq \tilde{\Sigma}$, where $A \leq B$ means that $B - A$ is a nonnegative definite matrix (see Lemma 6 in the online Appendix). The inequalities between the covariance matrices can be strict. In view of this, it is somewhat surprising that the asymptotic distributions of $\widehat{\mathrm{CMI}}$ based on $\hat{p}^*$ in all resampling scenarios coincide. This is investigated in the next Section.

4 Asymptotic distribution of $\widehat{\mathrm{CMI}}$ for considered resampling schemes

We consider CMI as a functional of probability vector $(p(x, y, z))_{x,y,z}$ defined as (compare (1))

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

We prove that despite differences in asymptotic behaviour of $n^{1/2}(\hat{p}^* - \hat{p})$ for both resampling schemes considered, the asymptotic distributions of

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

based on them coincide. Moreover, the common limit coincides with asymptotic distribution of $\widehat{\mathrm{CMI}}$, namely χ^2 distribution with $(|\mathcal{X}|-1) \times (|\mathcal{Y}|-1) \times |\mathcal{Z}|$ degrees of freedom. Thus in this case the general bootstrap principle holds as the asymptotic distributions of $\widehat{\mathrm{CMI}}$ and $\widehat{\mathrm{CMI}}^*$ are the same.

Theorem 6 *For almost all sequences $(X_i, Y_i, Z_i), i = 1, \ldots$ and conditionally on $(X_i, Y_i, Z_i)_{i=1}^{\infty}$ we have*

$$2n \times CMI(\hat{p}^*) \xrightarrow{d} \chi^2_{(|\mathcal{X}|-1)\times(|\mathcal{Y}|-1)\times|\mathcal{Z}|}, \tag{8}$$

a.e., where $\hat{p}^*$ is based on CP or CR scheme.

Proof We will prove the result for the Conditional Permutation scheme and indicate the differences in the proof in the case of CR scheme at the end. The approach is based on delta method as in the case of $\widehat{\mathrm{CMI}}$ (see e.g. Kubkowski et al. (2021)). The gradient and Hessian of CMI(p) considered as a function of p are equal to, respectively,

$$(D_{\mathrm{CMI}}(p))(x,y,z) = \frac{\partial \mathrm{CMI}(p)}{\partial p(x,y,z)} = \log \frac{p(x,y,z)p(z)}{p(x,z)p(y,z)}, \tag{9}$$

and

$$(H_{\mathrm{CMI}}(p))_{x,y,z}^{x',y',z'} = \frac{\mathbb{I}(x=x', y=y', z=z')}{p(x,y,z)} - \frac{\mathbb{I}(x=x', z=z')}{p(x,z)} - \frac{\mathbb{I}(y=y', z=z')}{p(y,z)} + \frac{\mathbb{I}(z=z')}{p(z)}, \tag{10}$$

where $(H_{\mathrm{CMI}}(p))_{x,y,z}^{x',y',z'}$ denotes element of Hessian with row column index x, y, z and column index x', y', z'. In order to check it, it is necessary to note that e.g. the term $p(x', y') = \sum_{z'} p(x', y', z')$ contains the summand $p(x, y, z)$ if $x = x'$ and $y = y'$, and thus $\frac{\partial p(x',y')}{\partial p(x,y,z)} = I(x = x', y = y')$. The proof follows now from expanding CMI$(\hat{p}^*)$ around $\hat{p}_{ci} := \hat{p}(x|z)\hat{p}(y|z)\hat{p}(z)$:

$$\mathrm{CMI}(\hat{p}^*) = \mathrm{CMI}(\hat{p}_{ci}) + (\hat{p}^* - \hat{p}_{ci})^\top D_{\mathrm{CMI}}(\hat{p}_{ci}) + \frac{1}{2}(\hat{p}^* - \hat{p}_{ci})^\top H_{\mathrm{CMI}}(\xi)(\hat{p}^* - \hat{p}_{ci}), \tag{11}$$

where $\xi = (\xi_{x,y,z})_{x,y,z}$ and $\xi_{x,y,z}$ is a point in-between $\hat{p}^*(x, y, z)$ and $\hat{p}_{ci}(x, y, z)$. We note that CMI$(\hat{p}_{ci}) = 0$ as $\hat{p}_{ci}$ is a distribution satisfying CI and, moreover, the gradient of conditional mutual information D_{CMI} at $\hat{p}_{ci}$ is also 0 as

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

Thus two first terms on RHS of (11) are 0. Moreover, using continuity of $H_{\mathrm{CMI}}(\cdot)$ following from $p(x, y, z) > 0$ for all (x, y, z) and (3) it is easy to see that

$$n(\hat{p}^* - \hat{p}_{ci})^\top (H_{\mathrm{CMI}}(\xi) - H_{\mathrm{CMI}}(p_{ci}))(\hat{p}^* - \hat{p}_{ci}) \to 0$$

a.e. Thus the asymptotic distribution of $2n \times$ CMI$(\hat{p}^*)$ coincides with that of $n^{1/2}(\hat{p}^* - \hat{p}_{ci})^\top H_{\mathrm{CMI}}(p_{ci}) n^{1/2}(\hat{p}^* - \hat{p}_{ci})$. Using (3) again we see that the asymptotic distribution is that of quadratic form $Z^\top H(p_{ci})Z$, where $Z \sim N(0, \Sigma)$. Alternatively, in view of the spectral decomposition, we have that

$$2nCMI(\hat{p}^*) \xrightarrow{d} \sum_{x,y,z} \lambda_{x,y,z} Z_{x,y,z}^2, \tag{12}$$

where $Z = (Z_{x,y,z})_{x,y,z} \sim N(0, I)$ and $\lambda_{x,y,z}$ are eigenvalues of a matrix $M = H_{\mathrm{CMI}}(p_{ci})\Sigma$. To finish the proof it is enough to check that M is idempotent, thus all

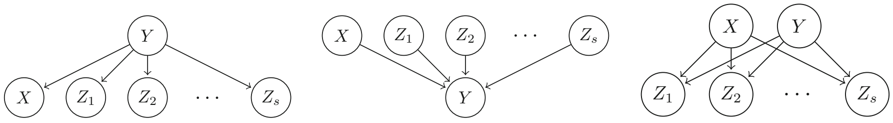


Fig. 1 Considered models

its eigenvalues are 0 or 1, and verify that the trace of M equals $(|\mathcal{X}|-1)\times(|\mathcal{Y}|-1)\times|\mathcal{Z}|$. This is proved in Lemma 3 in the online Appendix.

The proof for CR scheme is analogous and differs only in that in the final part of the proof matrix M is replaced by matrix $\tilde{M} = H_{\text{CMI}}\tilde{\Sigma}$ where $\tilde{\Sigma}$ is defined in Theorem 5. However, its shown in Lemma 3 in the online Appendix that $\tilde{M} = M$ thus the conclusion of the Theorem holds also for CR scheme. □

Remark 4 We note that two additional resampling scenarios can be defined. The first one, which we call bootstrap.X, is a variant of CR scenario in which, instead of sampling on the strata $Z = z_i$ from the distribution $P_{X|Z=z_i}$ the pseudo-observations are sampled from the empirical distribution of $\hat{P}(x|z_i)$. In order to introduce the second proposal, Conditional Independence Bootstrap (CIB), consider first empirical distribution $\hat{p}_{ci} = \hat{p}(x|z)\hat{p}(y|z)\hat{p}(z)$. We note that probability mass function $(\hat{p}_{ci}(x, y, z))_{x,y,z}$ is the maximum likelihood estimator of p.m.f. $(p(x, y, z))_{x,y,z}$ when conditional independence of X and Y given Z holds. Then $(X_i^*, Y_i, Z_i)_{i=1}^n$ is defined as iid sample given $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ drawn from $\hat{p}_{ci}$. Note that there is a substantial difference between this and previous scenarios as in contrast to them X and Z observations are also sampled. For the both scenarios convergence established in Theorem 6 holds (see Łazęcka 2022). However, we conjecture that validity of p-values does not hold for these schemes. As we did not establish substantial advantages of using either bootstrap.X or CIB over neither CP or CR scheme we have not pursued discussing them here in detail.

5 Numerical experiments

In the experiments, we will consider the following modification of a classical asymptotic test based on χ^2 distribution as the reference distribution. Namely, since it is established in Theorem 6 that $2n \times \widehat{\text{CMI}}^*$ is approximately χ^2 distributed for both scenarios considered, we use the limited number of resampled samples to approximate the mean of the distribution of $2n \times \widehat{\text{CMI}}^*$ and use the obtained value as an estimate of the number of degrees of freedom of χ^2 distribution. The adjustment corresponds to the equality of the mean and the number of degrees of freedom in the case of χ^2 distribution. Thus, we still consider χ^2 distribution as the reference distribution for CI testing; however, we adjust its number of degrees of freedom. The idea appeared already in Tsamardinos and Borboudakis (2010). Here, the approach is supported by Theorem 6 and the behaviour of the resulting test is compared with the other tests considered in the paper.

We will thus investigate three tests in both resampling schemes CR and CP. The test which will be called `exact` is based on Theorem 1 in the case of CP scenario and the analogous result for CR scenario in Candès et al. (2018). The test `df estimation` uses χ^2 distribution with the degrees of freedom estimated in data-dependent way as just described. As a benchmark test we use the asymptotic test which uses the asymptotic χ^2 distribution established in Theorem 6 as a reference distribution.

Choice of number of resampled samples B. As in the case of `df estimation` test the reference distribution involves only the estimator of the mean and not the estimators of upper quantiles of high order, we use a moderate number of resampled samples $B = 50$ for this purpose. In order to have equal computational cost for all tests, $B = 50$ is also used in the case of `exact` test. Note that applying moderate B renders application of such tests in greedy feature selection (when such tests have to be performed many times) feasible.

The models considered are standard models to study various types of conditional dependence of X and Y given vector Z: e.g. in model 'Y to XZ' , Y conveys information to both X and Z whereas in model 'X and Y to Z' both X and Y convey information to Z. Model XOR is a standard model to investigate interactions of order 3. Below we will describe the considered models in detail by giving the formula for joint distribution of $(X, Y, Z_1, Z_2, \ldots, Z_s)$. Conditional independence case (the null hypothesis) will be investigated by projecting considered models on the family of conditionally independent distributions.

- **Model 'Y to XZ'** (the first panel of Fig. 1). Joint probability in the model is factorised as follows

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

  thus it is sufficient to define p.m.f. of Y and conditional p.m.f. of $(X, Z_1, \ldots, Z_s)$ given Y. First, Y is a Bernoulli random variable with probability of success equal to 0.5 and conditional distribution of $(\tilde{X}, \tilde{Z}_1, \ldots, \tilde{Z}_s)$ given $Y = y$ follows a multivariate normal distribution $N_{s+1}(y\gamma_s, \sigma^2 I_{s+1})$, where $\gamma_s = (1, \gamma, \ldots, \gamma^s)$, and $\gamma \in [0, 1]$ and $\sigma > 0$ are parameters in that model. In order to obtain discrete variables from continuous $(\tilde{X}, \tilde{Z}_1, \ldots, \tilde{Z}_s)$ we define the conditional distribution of $(X, Z_1, \ldots, Z_s)$ given $Y = y$ by assuming their conditional independence given Y and

$$P(X = x|Y = y) = P\Big((-1)^x \tilde{X} \leq \frac{(-1)^x}{2}|Y = y\Big),$$
$$P(Z_i = z_i|Y = y) = P\Big((-1)^{z_i} \tilde{Z}_i \leq \frac{(-1)^{z_i}\gamma^i}{2}|Y = y\Big)$$

  for $i = 1, 2, \ldots, s$, where $x, z_1, z_2, \ldots, z_s \in \{0, 1\}$. Thus $X|Y = y \sim \text{Bern}(\Phi((2y-1)/(2\sigma)))$ and $Z_i|Y = y \sim \text{Bern}(\Phi((2y-1)\gamma^i/(2\sigma)))$. Variables $X, Z_1, Z_2, \ldots, Z_s$ are conditionally independent given Y but X an Y are not conditionally independent given $Z_1, Z_2, \ldots, Z_s$.
- **Model 'XZ to Y'** This model is obtained by changing the direction of all arrows in the graph corresponding to the previous model; compare the first and the second

panel of Fig. 1. In the model the joint distribution is given by

$$p(x, y, z_1, z_2, \ldots, z_s) = p(x)\Big(\prod_{i=1}^{s} p(z_i)\Big)p(y|x, z_1, z_2, \ldots, z_s).$$

The variables X and Z_i all have Bern(0.5) distribution and conditional distribution of Y follows

$$Y|X = x, Z_1 = z_1, \ldots, Z_s = z_s \\ \sim Bern\left(1 - \Phi\left(\left(\frac{x + z_1 + z_2 + \ldots + z_s}{s+1} - 0.5\right)/\sigma\right)\right).$$

- **Model 'XY to Z'** (the third panel in Fig. 1) The joint probability factorises as follows

$$p(x, y, z_1, z_2, \ldots, z_s) = p(x)p(y)\prod_{i=1}^{s} p(z_i|x, y).$$

  X and Y are independent and both follow Bernoulli distribution Bern(0.5). The distribution of Z_i depends on the arithmetic mean of X and Y and the variables $Z_1, \ldots, Z_s$ are conditionally independent given (X, Y). They follow Bernoulli distribution $Z_i|(X + Y)/2 = w \sim \text{Bern}(1 - \Phi(\alpha(\frac{1}{2} - w))$ for $i \in \{1, 2, ..., s\}$, where $\alpha \geq 0$ controls the strength of dependence. For $\alpha = 0$, the variables Z_i do not depend on (X, Y).
- **Model XOR** The distribution of Y is defined as follows:

$$P(Y = 1|X + Z_1 + Z_2 =_2 1) = P(Y = 0|X + Z_1 + Z_2 =_2 0) = \beta,$$

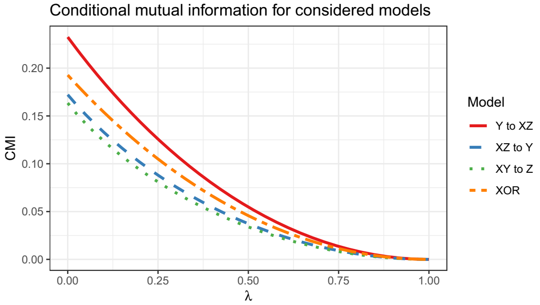
  where $0.5 < \beta < 1$ and $=_2$ denotes addition modulo 2. We also introduce variables $Z_3, Z_4, \ldots, Z_s$ independent of (X, Y, Z_1, Z_2) . All variables $X, Z_1, Z_2, \ldots, Z_s$ are independent and binary with the probability of success equal to 0.5.

We run simulations for fixed model parameters (Model **'Y to XZ'**: $\gamma = 0.5$, $\sigma = 0.5$, Model **'XZ to Y'**: $\sigma = 0.07$, model **'XY to Z'**: $\alpha = 3$, model **XOR**: $\beta = 0.8$. In all the models the same number of conditioning variables $s = 4$ was considered. The parameters are chosen in such a way that in all four models values of conditional mutual information CMI$(X, Y|Z)$ are similar and contained in the interval [0.16, 0.24] (see Fig. 2 for $\lambda = 0$ which corresponds to the chosen p.m.f. $p(x, y, z)$). We define a family of distributions parameterised by parameter $\lambda \in [0, 1]$ in the following way:

$$p_\lambda(x, y, z) = \lambda p_{ci}(x, y, z) + (1 - \lambda)p(x, y, z),$$

where p denotes the joint distribution pertaining to the model with the chosen parameters and $p_{ci}(x, y, z) = p(x|z)p(y|z)p(z)$ is the Kullback–Leibler projection of p onto

Fig. 2 Conditional mutual information of random variables X and Y given $Z = (Z_1, Z_2, Z_3, Z_4)$, joint distribution of which equals $p_\lambda = \lambda p_{ci} + (1 - \lambda)p$, and p and p_{ci} are characterised by the chosen models and parameters (see text)



the family $\mathcal{P}_{ci}$ of p.m.fs satisfying conditional independence $X \perp\!\!\!\perp Y|Z$ (see Lemma 4 in online Appendix). Probability mass function $p_{ci}(x, y, z)$ can be explicitly calculated for the given $p(x, y, z)$. Note that λ is a parameter which controls the strength of shrinkage of p toward p_{ci}. We also underline that the Kullback–Leibler projection of p_λ onto $\mathcal{P}_{ci}$ is also equal to p_{ci} (see Lemma 5 in the online Appendix). Figure 2 shows how conditional mutual information of X and Y given $(Z_1, Z_2, \ldots, Z_s)$ changes with respect to λ. For $\lambda = 1$, $p_\lambda = p_{ci}$, thus X and Y are conditionally independent and $\mathrm{CMI}(X, Y|Z) = 0$.

The simulations, besides standard analysis of attained levels of significance and power, are focused on the following issues. Firstly, we analyse levels of significance of $\widehat{\mathrm{CMI}}$-based tests for small sample sizes. It is known that for small sample sizes problems with control of significance levels arise, as the probability of obtaining the samples which result in empty cells (i.e. some values of $(x, y, z_1, \ldots, z_s)$ are not represented in the sample) is high. This issue obviously can not be solved by increasing the number of resampled samples as it is due the original sample itself. However, we would like to check whether using χ^2 distribution with estimated number of degrees of freedom as a benchmark distribution provides a solution to this problem. Moreover, the power of such tests in comparison with `exact` tests is of interest. Secondly, it is of importance to verify whether the knowledge of the conditional distribution of X given Z which is needed for CR scheme, actually translates into better performance of the resulting test over the performance of the same test in CP scenario.
The conditional independence hypothesis is a composite hypothesis, thus an important question is how to choose representative null examples on which control of significance level should be checked. Here we adapt a natural, and to our knowledge, novel approach which consists in considering as the nulls the projections p_{ci} of p.m.fs p for which power is investigated.

In Fig. 3 histograms of $p_{ci}(x, y, z)$ for the considered models are shown. Although all $2^{s+2} = 64$ probabilities $p(x, y, z)$ are larger than 0 in all the models, some probabilities may be very close to 0 (as it happens in **’XZ to Y’** model). For model **XOR** all triples are equally likely and thus for all (x, y, z) $p_{ci}(x, y, z) = 1/2^6 = 0.015625$. If there are many values of $p_{ci}(x, y, z)$ that are close to 0, the probability of obtaining a sample without some triples (x, y, z) for which $p_{ci}(x, y, z) > 0$ is high. In particular, this happens in **’XZ to Y’** model. In the following the performance of the procedures is studied with respect to the parameter `frac` $= n/2^{s+2}$ instead of sample size n. As

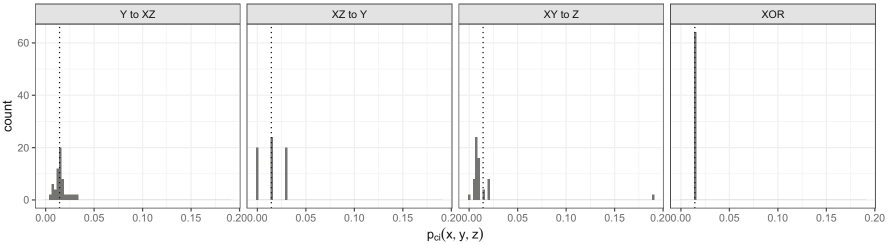


Fig. 3 Histograms of values of probabilities p_{ci} for the four considered models. The vertical dotted line shows the value of probability p_{ci} when all triples (x, y, z) are equally probable

Table 1 Values of $np_{\min}$, where $p_{\min} = \min_{(x,y,z)} p_{ci}(x, y, z)$ or $p_{\min} = \min_{(x,y,z)} p(x, y, z)$ with respect to n. `frac` values correspond to $s = 4$

`frac`	0.5	1	3	5	20		0.5	1	3	5	20	
n	32	64	192	320	1280		2	64	192	320	1280	
	$n\min_{(x,y,z)} p_{ci}(x, y, z)$						$n\min_{(x,y,z)} p(x, y, z)$					
Y to XZ	0.2	0.4	1.1	1.9	7.5	$\cdot 1$	0.1	0.2	0.5	0.9	3.5	$\cdot 1$
XZ to Y	0.5	0.9	2.7	4.6	18.2	$\cdot 10^{-5}$	0.5	0.9	2.7	4.6	18.3	$\cdot 10^{-12}$
XY to Z	0.0	0.1	0.2	0.4	1.4	$\cdot 1$	0.2	0.3	1.0	1.6	6.4	$\cdot 10^{-3}$
XOR	0.5	1.0	3.0	5.0	20.0	$\cdot 1$	0.2	0.4	1.2	2.0	8.0	$\cdot 1$

the number of unique values of triples (x, y, z) equals 2^{s+2}, thus `frac` is the average number of observations per cell in the uniform case and roughly corresponds to this index for a general binary discrete distribution.

In Table 1 we provide the values of sample sizes corresponding to changing `frac` as well as the value of $np_{\min}$ for $s = 4$, where $p_{\min}$ is the minimal value of either probability mass function $p(x, y, z)$ or $p_{ci}(x, y, z)$. As $np_{\min}$ is the expected value of observations for the least likely triple it indicates that occurrence of empty cells is typical for `frac` as large as 20.

In Fig. 4 the estimated fraction of rejections for the tests based on resampling in case when the null hypothesis is true ($\lambda = 1$) is shown when the assumed level of significance equals 0.05. The attained levels of significance for asymptotic test are given separately in Fig. 5. Overall, for all the procedures based on resampling the attained level of significance is approximately equal to the assumed one. The `df estimation` methods both for CP and CR do not exceed assumed significance level for the considered range of `frac` $\in [0.5, 5]$. Figure 4 indicates that distribution of $\widehat{\text{CMI}}$ is adequately represented by χ^2 distribution with estimated number of degrees of freedom. This will be further analysed below (see discussion of Figs. 5 and 6).
In Fig. 5 in the top row the attained values of significance levels for the asymptotic test are shown. That test significantly exceeds the assumed level $\alpha = 0.05$. The reason for that is shown in the bottom panel of Fig. 5. The red dots represent the mean of $2n\widehat{\text{CMI}}$ based on $n = 10^5$ samples for each value of `frac` and the solid line indicates the number of degrees of freedom of the asymptotic distribution of $2n\widehat{\text{CMI}}$, which for $s = 4$ equals $(|\mathcal{X}|-1)(|\mathcal{Y}|-1)|\mathcal{Z}|= 2^4$. For all the models except **'XZ to Y'** for small

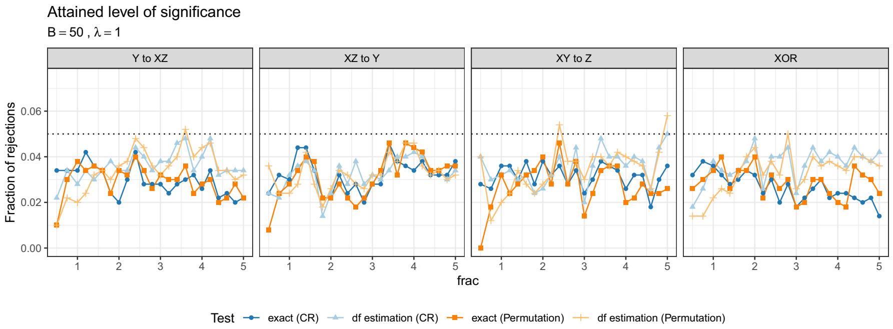


Fig. 4 Attained significance level of the tests based on resampled samples for the considered model p_{ci} corresponding to $\lambda = 1$, $B = 50$ with respect to `frac`

number of observations per cell we underestimate the mean of $2n\widehat{\text{CMI}}$ by using the asymptotic number of degrees of freedom and in these cases the significance level is exceeded. This effect is apparent even for `frac` equal to 5. On the other hand in the model **'XZ to Y'** the situation is opposite and in this case the test rarely rejects the null hypothesis. This is due to the overestimation of the mean of $2n\widehat{\text{CMI}}$ by asymptotic number of degrees of freedom in the case when many empty cells occur. Note that the estimation of the mean based on resampled samples is much more accurate (in Fig. 5 we present the results for Conditional Permutation only; the mean of $B = 50$ values of $\widehat{\text{CMI}}^*$ is computed 500 times and its mean and the mean ± standard error of obtained results is marked in blue). We also note that the condition $np_{\min} \geq 5$ is frequently cited as the condition under which test based on asymptotic χ^2 distribution can be applied. Note, however, that in the considered examples and for `frac` ≥ 20, asymptotic test controls fairly well level of significance, whereas $np_{\min}$ can be of order 10^{-11} (Table 1). Moreover, for `frac`=20 and $\lambda = 0.5$ the power of asymptotic test is 1.

In Fig. 6 we compare the distributions of $\widehat{\text{CMI}}$ with those of resampling distributions of $\widehat{\text{CMI}}^*$ and χ^2 distribution with the estimated number of degrees of freedom by means of QQ plots. For each of 500 original samples 50 resampled samples are generated by the Conditional Permutation method and quantiles of resampling distributions of $\widehat{\text{CMI}}^*$ are calculated, resulting in 500 quantiles, medians of which which are shown in the plot. Medians of quantiles for χ^2 distribution with an estimated number of degrees of freedom are obtained in the similar manner. Quantiles of the asymptotic distribution are also shown. Besides the fact that the distribution of $\widehat{\text{CMI}}$ is better approximated by the distribution of $\widehat{\text{CMI}}^*$, what confirms the known property of bootstrap in the case of CMI estimation (compare Section 2.6.1 in Davison and Hinkley (1997)), it also follows from the figure that the distribution of $\widehat{\text{CMI}}$ is even better approximated by χ^2 distribution with estimated number of degrees of freedom.

Figure 7 shows the results for the power of testing procedures for $\lambda = 0.25, 0.5, 0.75$ with respect to `frac`. Since asymptotic test does not control significance level for these models for $\lambda = 1$, the pertaining power is omitted from the figure. As for increasing λ, p.m.f. of p_λ approaches the null hypothesis described by p_{ci} the power becomes

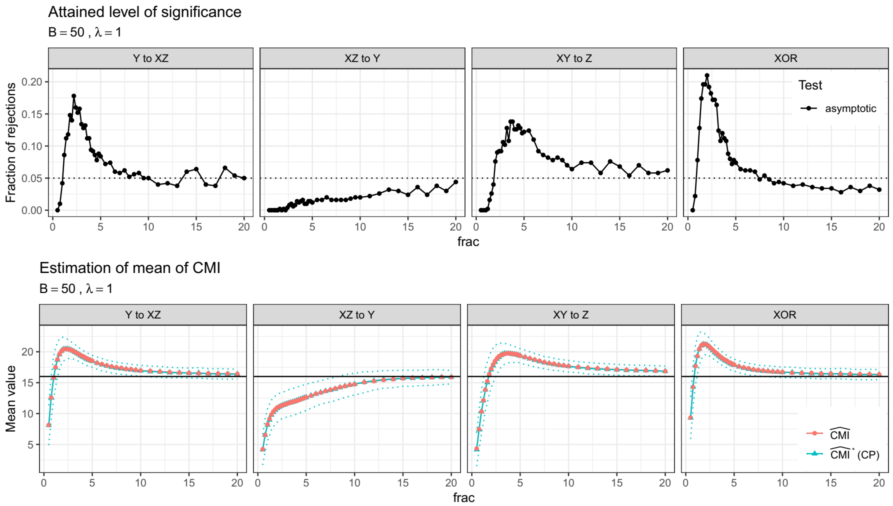


Fig. 5 Top panels: Levels of significance for asymptotic test. Bottom panels: comparison of the estimated and assumed number of degrees of freedom in testing procedures: mean of $2n\widehat{\mathrm{CMI}}$ based on 10^5 samples generated according to p_{ci}, mean of $2n\widehat{\mathrm{CMI}}^*$ (each estimated mean is based on $B = 50$ resampled samples and the simulation is repeated 500 times; the average of the obtained means and mean$\pm SE$ is shown in blue. The number of degrees of freedom for asymptotic χ^2 distribution is a solid horizontal line

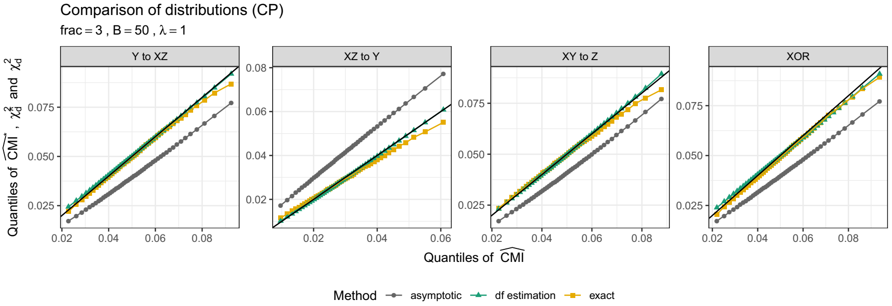


Fig. 6 Q–Q plots of distribution of $\widehat{\mathrm{CMI}}$ versus asymptotic distribution (grey), exact resampling distribution (yellow) based on permutations and χ^2 distribution with an estimated number of degrees of freedom (green) under conditional independence for p_{ci}. For the two last distributions medians of 500 quantiles for resampling distributions each based on 50 resampled samples are shown. Straight black line corresponds to $y = x$

smaller in rows. As `frac` gets smaller, the power of the tests also decreases and this is due to the increased probability of obtaining empty cells $(x, y, z_1, \ldots, z_s)$ in the sample, and because of that such observations are also absent in the resampled samples for Conditional Permutation scheme. CR is more robust in this respect as such occurs only when not all values of $(z_1, \ldots, z_s)$ are represented in the sample. This results in better performance of the tests for CR scheme than for CP scheme for small values of `frac` (see also Fig. 8). It follows that the procedures based on χ^2 distribution with the estimated number for of degrees of freedom are more powerful than `exact` tests,

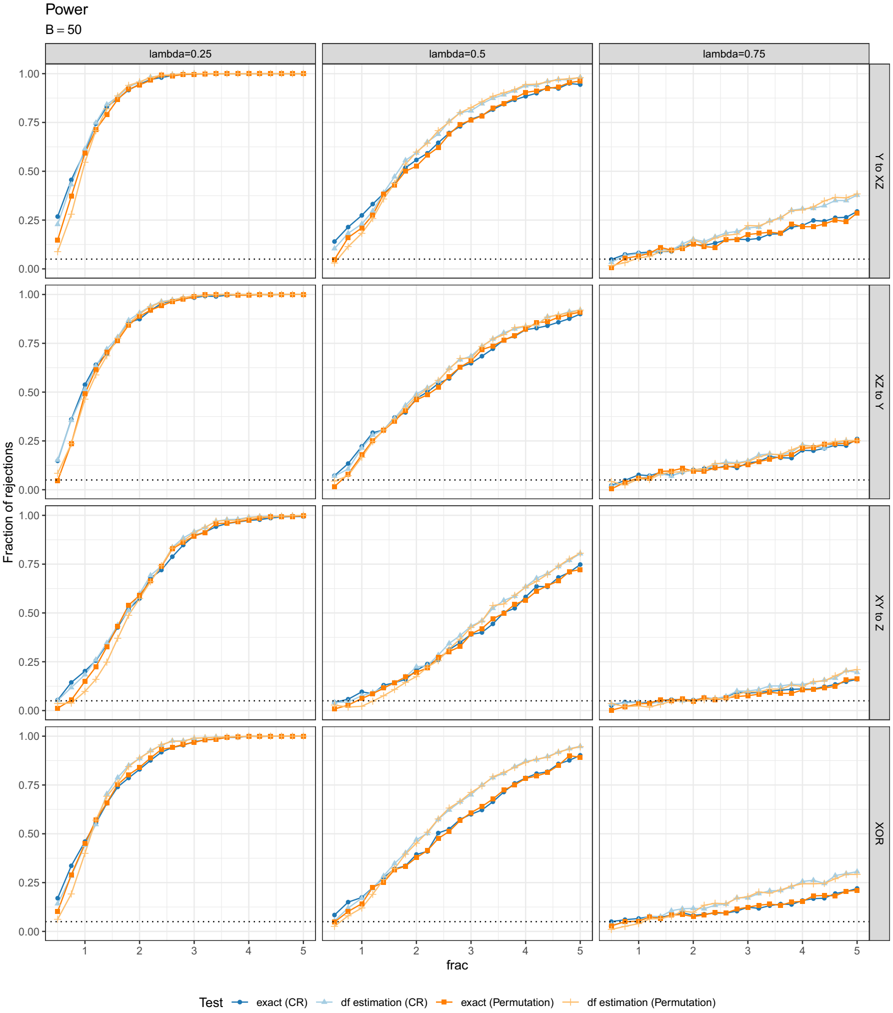


Fig. 7 Power of the tests based on resampled samples for the considered model for $\lambda = 0.25, 0.5, 0.75$ and $B = 50$ with respect to `frac`

regardless of the resampling scenario used. Although the advantage is small, it occurs in all cases considered. The plot also indicates that `exact` tests in both scenarios act similarly and are inferior to tests based on asymptotic distribution with estimated dfs which also exhibit similar behaviour.

We compare powers in CP and CR scenarios in Fig. 8 in which ratios of respective powers for exact tests and `df estimation` tests are depicted by orange and green lines, respectively. The values below 1 mean that the CR has greater power. The differences occur only for small `frac` values. Both `df estimation` and `exact` tests have larger power in CR scenario than in CP scenario for `frac` $\in [0.5, 2]$. The

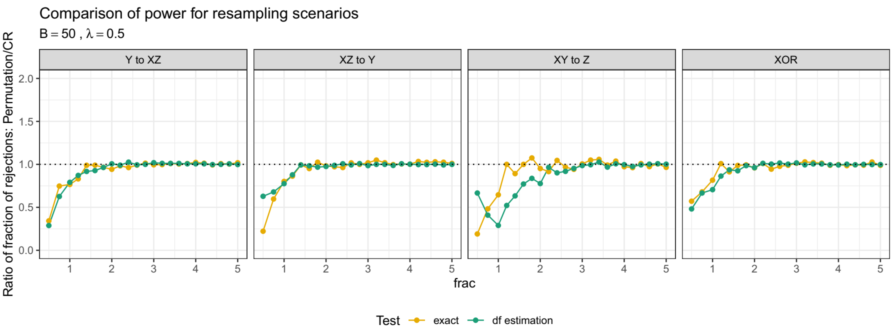


Fig. 8 Comparison of resampling scenarios. Fraction of rejections for CP divided by fraction of rejections for CR for both `exact` and `df estimation` tests for $\lambda = 0.5$ and $B = 50$

power for both methods is similar for `frac` ≥ 2, thus it follows that CP scenario might be used instead of CR, as it is as efficient as CR.

Our conclusions can be summarised as follows:

- The significance level is controlled by `df estimation` and `exact` tests both for CP and CR scenarios. It happens that asymptotic test does not control significance level even for `frac` larger than 10. Interestingly, although asymptotic case is usually significantly too liberal for small `frac` it also happens that it is very conservative (Fig. 4, model **'XZ to Y'**);
- The power of `estimated df` test is consistently larger than `exact` test, both for CR and CP scenarios. The advantage is usually more significant closer to null hypothesis (larger λ);
- There is no significant difference in power between `df estimation` tests in CR and CP scenarios apart from the region `frac` $\in [0.5, 2]$. The same holds for both `exact` tests excluding `frac` $\in [0.5, 1.5]$. Moreover, `df estimation` test for CP scenario has larger power than CR `exact` test.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11749-023-00878-7.

Acknowledgements B. Kołodziejek was partially supported by the NCN Grant UMO-2022/45/B/ST1/00545.

References

Aliferis C, Tsamardinos I, Statnikov A (2003) Hiton: a novel Markov Blanket algorithm for optimal variable selection. In: AMIA Annu. Symp. Proc., pp 21–25

Berrett T, Wang Y, Barber R, Samworth R (2020) The conditional permutation test for independence while controlling for confounders. J R Stat Soc B

Candès E, Fan Y, Janson L, Lv J (2018) Panning for gold: model-X knockoffs for high dimensional controlled variable selection. J R Stat Soc B Stat Methodol 80:551–577

Cover TM, Thomas JA (2006) Elements of information theory (Wiley series in telecommunications and signal processing). Wiley-Interscience, Hoboken, New Jersey

Davison A, Hinkley D (1997) Bootstrap methods and their applications. Cambridge University Press, Cambridge, United Kingdom

Fu S, Desormais M, (2017) Fast Markov Blanket discovery algorithm via local learning within single pass. In: CSCSI Conference, pp 96–107

Halton J (1969) A rigorous derivation of the exact contingency formula. Math Proc Camb Philos Soc 65:527–530

Koller D, Sahami M (1995) Toward optimal feature selection. In: ICML-1995, pp 284–292

Kubkowski M, Mielniczuk J, Teisseyre P (2021) How to gain on power: novel conditional independence tests based on short expansion of conditional mutual information. J Mach Learn Res 22:1–57

Łazęcka M (2022) Properties of information-theoretic measures of conditional dependence. PhD thesis. https://home.ipipan.waw.pl/m.lazecka/files/publications/phd_thesis_mlazecka.pdf

Tsamardinos I, Borboudakis G (2010) Permutation testing improves Bayesian network learning. In: Lecture notes in computer science vol. 6323 LNAI, pp 322–337

Watson D, Wright M (2021) Testing conditional independence in supervised testing algorithms. Mach Learn 110:2129–2129