Supplemental material to 'Detection of conditional dependence between multiple variables using multiinformation'

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

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

1 Proof of Theorem 1

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

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

(ii)

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

where $MI(X_i; X_1, \dots, X_{i-1}|Y)$ denotes conditional mutual information between X_i and $(X_1, \dots, X_{i-1})$ given Y.
(iii) We have

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

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

$$CMI(X|Y) = D_{KL}(P_{Y|X} || P^{ind}_{Y|X}) + D_{KL}(P_X || P^{ind}_X) \tag{1}$$

(v) We have

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

where χ^2 index is defined as

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

LHS and RHS equal 0 for conditional independence case.

Proof. In order to prove (i) note that in the view of equation (4) in the main body of the paper it is enough to check that

$$\sum_{k=1}^{i+1} H(X_k|Y) - H(X_1,\ldots,X_{i+1}|Y) \geq \sum_{k=1}^{i} H(X_k|Y) - H(X_1,\ldots,X_i|Y).$$

However, as

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

the inequality follows from the fact that conditioning decreases entropy and thus

$$H(X_{i+1}|X_1,\ldots,X_i,Y) \leq H(X_{i+1}|Y).$$

To see (ii) note that the RHS of the equality in question in view of definition of the conditional mutual information is

$$\sum_{i=2}^{d}(H(X_i|Y) + H(X_1,\ldots,X_{i-1}|Y) - H(X_1,\ldots,X_i|Y)).$$

As the sum of the two last terms equals $H(X_1|Y) - H(X_1,\ldots,X_d|Y)$ the result follows from (4) in the main body of the paper.
Note that in order to prove (iii) it is enough to check that for any conditional distributions $q(x_i|y)$ and for any y we have

$$\sum_{x_1\ldots,x_d} p(x_1,\ldots,x_d|y)\log(p(x_1|y)\cdots p(x_d|y)) \geq \sum_{x_1\ldots,x_d} p(x_1,\ldots,x_p|y)\log(q(x_1|y)\cdots q(x_d|y)).$$

But this, after simplification, follows from

$$\sum_{x_i} p(x_i|y)\log p(x_i|y) \geq \sum_{x_i} p(x_i|y)\log q(x_i|y).$$

which is a consequence of basic property of K-L divergence that $D_{KL}(p||q) \geq 0$.
Note that (1) in (iv) follows from general property that if two distributions $P_{Y,X}$ and $Q_{Y,X}$ are such that $P_Y = Q_Y$ we have

$$D_{KL}(P_{X|Y}||Q_{X|Y}) = D_{KL}(P_{Y|X}||Q_{Y|X}) + D_{KL}(P_X||Q_X) \tag{2}$$

Indeed, we have

$$\begin{aligned} D_{KL}(P_{X|Y}||Q_{X|Y}) &= \sum_{x,y} p(x,y)\log\frac{p(x|y)}{q(x|y)} = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{q(x,y)} \\ = \sum_{x,y} p(x,y)\log\frac{p(y|x)p(x)}{q(y|x)q(x)} &= \sum_{x,y} p(x,y)\log\frac{p(y|x)}{q(y|x)} + \sum_{x,y} p(x,y)\log\frac{p(x)}{q(x)} \\ = D_{KL}(P_{Y|X}||Q_{Y|X}) &+ D_{KL}(P_X||Q_X), \end{aligned} \tag{3}$$

where the second equality used $P_Y = Q_Y$.

2 Proof of Theorem 2

Proof. Part (i). Let $\hat{p}(x_1,\ldots,x_d,y) = \#\{i : (X_{i1},\ldots X_{id},Y_i) = (x_1,\ldots,x_d,y)\}/n$ be plug-in estimator for $p(x_1,\ldots,x_d,y)$ and $\mathbf{p} = (p(x_1,\ldots,x_d,y))_{(x,y,z)\in\mathcal{X}_1,\times\mathcal{X}_d\times\mathcal{Y}}$ be the corresponding vector of probabilities. We write $CMI(X|Y)$ as a function of $\mathbf{p}$, namely

$$CMI(X|Y) = \sum_{x_1,\ldots,x_d,y} p(x_1,\ldots,x_d,y)\log\left(\frac{p(x_1,\ldots,x_d,y)[p(y)]^{d-1}}{p(x_1,y)\cdots p(x_d,y)}\right) =: f(\mathbf{p}).$$

Observe that $\widehat{CMI}(X|Y) = f(\hat{\mathbf{p}})$. Denote the summand in the above decomposition of $CMI(X|Y)$ by $T(x_1,\ldots x_d,y)$. We note that

$$\begin{aligned}\frac{\partial T(x_1,\ldots x_d,y)}{\partial p(x_1,\ldots x_d,y)} &= \log\frac{p(x_1,\ldots,x_d,y)[p(y)]^{d-1}}{p(x_1,y)\cdots p(x_d,y)} + 1\\ &- p(x_1,\ldots x_d,y)\Big(\sum_{i=1}^{d}\frac{1}{p(x_i,y)} - \frac{1}{p(y)}\Big)\end{aligned} \tag{4}$$

and for $T = f(\mathbf{p}) - T(x_1,\ldots x_d,y)$ we have

$$\begin{aligned}&\frac{\partial T}{\partial p(x_1',\ldots x_d',y')} =\\ &\sum_{x_1',\ldots,x_d':(x_1',x_d',\ldots,x_d')\neq(x_1,\ldots,x_d)}\frac{(p-1)p(x_1,,\ldots,x_d',y)}{p(y)}\\ &-\sum_{i=1}^{d}\sum_{x'_{-i}\neq x_{-i}}\frac{p(x_1',\ldots,x_{i-1}',x_i,x_{i+1}',\ldots,x_d',y)}{p(x_i,y)}\end{aligned} \tag{5}$$

where x_{-i} denotes $=(x_1,\ldots,x_d)$ with ith coordinate omitted. It is easy to see that (5) equals

$$\frac{\partial T}{\partial p(x_1,\ldots x_d,y)} = \frac{(d-1)(p(y)-p(x_1,\ldots x_d,y))}{p(y)} - \sum_{i=1}^{d}\frac{p(x_i,y)-p(x_1,\ldots,x_d)}{p(x_i,y)} \tag{6}$$

Adding (4) and (6) we obtain

$$\begin{aligned}&\frac{\partial f(\mathbf{p})}{\partial p(x_1,\ldots,x_d,y)} = \log\left(\frac{p(x_1,\ldots,x_d,y)[p(y)]^{d-1}}{p(x_1,y)\cdots p(x_d,y)}\right) + 1 + (d-1) - d =\\ &\log\left(\frac{p(x_1,\ldots,x_d,y)[p(y)]^{d-1}}{p(x_1,y)\cdots p(x_d,y)}\right).\end{aligned} \tag{7}$$

Reasoning analogously we have

$$\begin{aligned}\frac{\partial^2 f(\mathbf{p})}{\partial p(x_1,\ldots,x_d,y)\partial p(x_1',\ldots,x_d',y')} &= \frac{I(x_1=x_1',\ldots x_d=x_d',y=y')}{p(x_1,\ldots,x_d,y)}\\ &-\sum_{i=1}^{d}\frac{I(x_i=x_i',y=y')}{p(x_i,y)} + \frac{(d-1)I(y=y')}{p(y)},\end{aligned}$$

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

$$f(\hat{\mathbf{p}}) - f(\mathbf{p}) = Df(\mathbf{p})^T(\hat{\mathbf{p}} - \mathbf{p}) + \frac{1}{2}(\hat{\mathbf{p}} - \mathbf{p})^T D^2 f(\mathbf{p})(\hat{\mathbf{p}} - \mathbf{p}) + O(||\hat{\mathbf{p}} - \mathbf{p}||_2^3). \quad (8)$$

Moreover, we have that an element of $\Sigma = n\,\mathrm{Var}(\hat{\mathbf{p}}-\mathbf{p})$ with row index $x_1, \ldots x_d y$ and column index $x_1 \ldots x_d' y'$ is

$$\Sigma_{x_1 \ldots x_d y}^{x_1' \ldots x_d' y'} = p(x_1', \ldots, x_d', y')(I(x_1 = x_1', \ldots x_d = x_d', y = y') - p(x_1, \ldots, x_d, y)).$$

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

$$\begin{aligned}
&n\mathrm{Var}(Df(\mathbf{p})^T(\hat{\mathbf{p}} - \mathbf{p})) = \\
&\sum_{x_1, \ldots, x_d, y} p(x_1, \ldots, x_d, y) \log^2\left(\frac{p(x_1, \ldots, x_d|y)}{p(x_1|y) \cdots p(x_d|y)}\right) - \\
&\Big(\sum_{x_1, \ldots, x_d, y} p(x_1, \ldots, x_d, y) \log\left(\frac{p(x_1, \ldots, x_d|y)}{p(x_1|y) \cdots p(x_d|y)}\right)\Big)^2 = \\
&\mathrm{Var}\Big(\log\left(\frac{p(X_1, \ldots, X_d|Y)}{p(X_1|Y) \cdots p(X_d|Y)}\right)\Big). \qquad (9)
\end{aligned}$$

This ends the proof of part (i) as $CMI(X|Y) \neq 0$ implies that

$$p(x_1, \ldots, x_d|y)/p(x_1|y) \cdots p(x_d|y)$$

is not constant and the variance above is not zero and thus the first term on RHS of (8) dominates.
In order to prove (ii) note that from assumption $CMI(X|Y) = 0$ it follows that $Df(\mathbf{p})$ is constant and the first term on thr RHS of (8) equals 0. As Central Limit Theorem Implies $\sqrt{n}(\hat{\mathbf{p}} - \mathbf{p}) \xrightarrow{d} N(0, \Sigma)$ we have from (8) that

$$2nf(\hat{\mathbf{p}}) \xrightarrow{d} N(0, \Sigma)^T D^2 f(\mathbf{p}) N(0, \Sigma) = N(0, I)^T \Sigma^{1/2} D^2 f(\mathbf{p}) \Sigma^{1/2} N(0, I). \quad (10)$$

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

$$2nf(\hat{\mathbf{p}}) \xrightarrow{d} \sum_1^l \lambda_i(M) Z_i^2, \quad (11)$$

where $\lambda_i(M)$ are eigenvalues of M and Z_i are independent $N(0,1)$-distributed random variables. Some algebraic manipulations yield:

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

3 Results of additional experiments

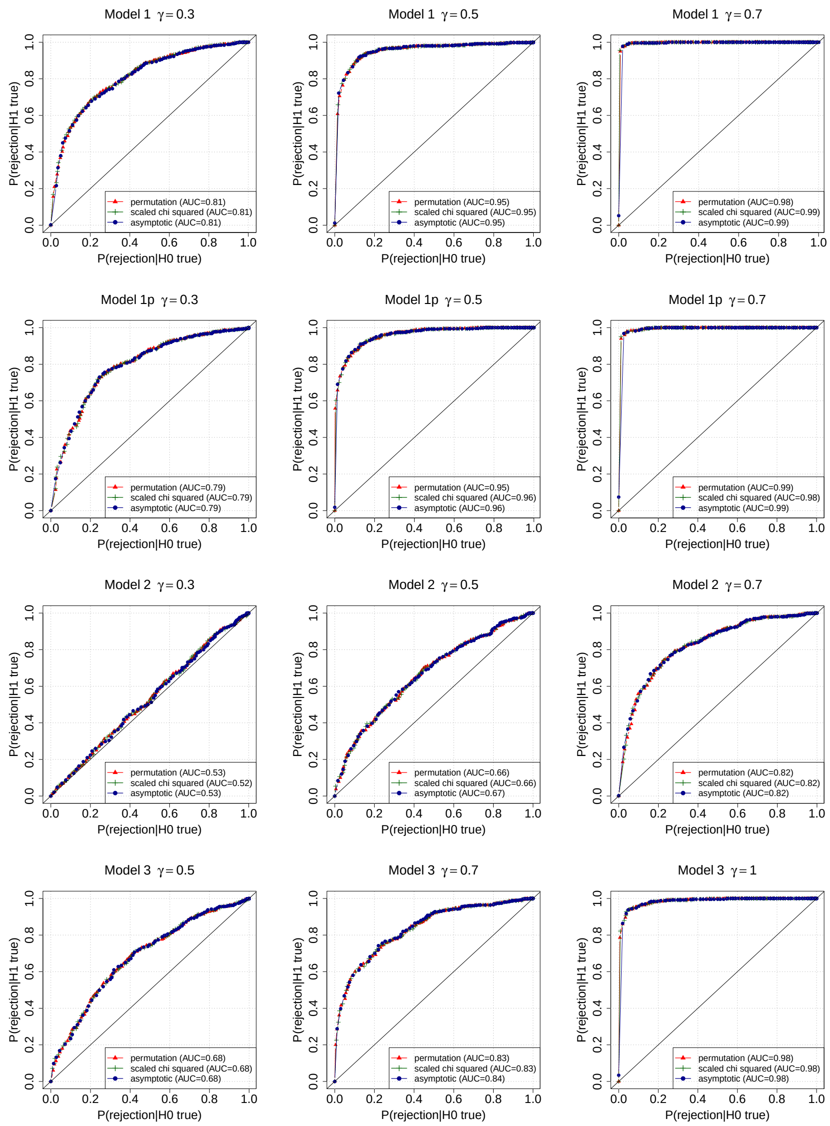

Fig. 1. ROC-type curves for simulation models 1, 1p, 2, 3 and permutation test (red), scaled chi-squared test (green) and asymptotic test (blue). Number of variables $d = 3$ and sample size $n = 500$.

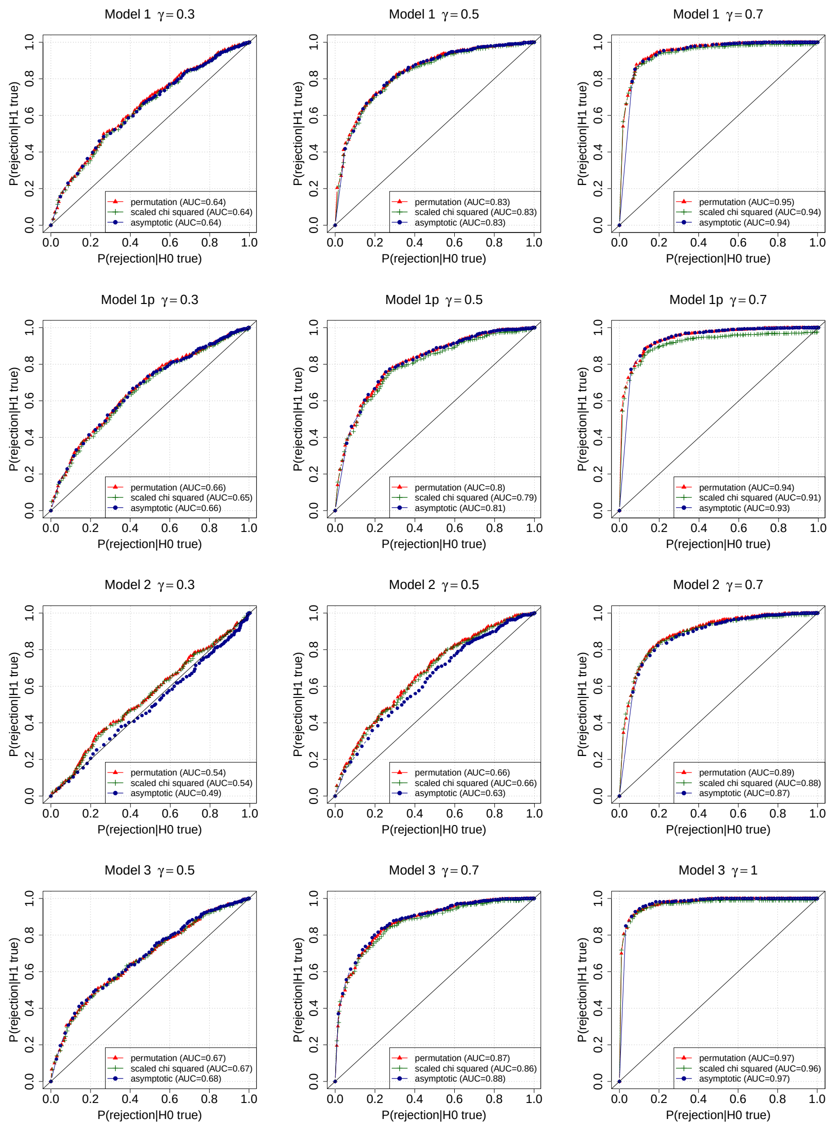

Fig. 2. ROC-type curves for simulation models 1, 1p, 2, 3 and permutation test (red), scaled chi-squared test (green) and asymptotic test (blue). Number of variables $d = 5$ and sample size $n = 500$.

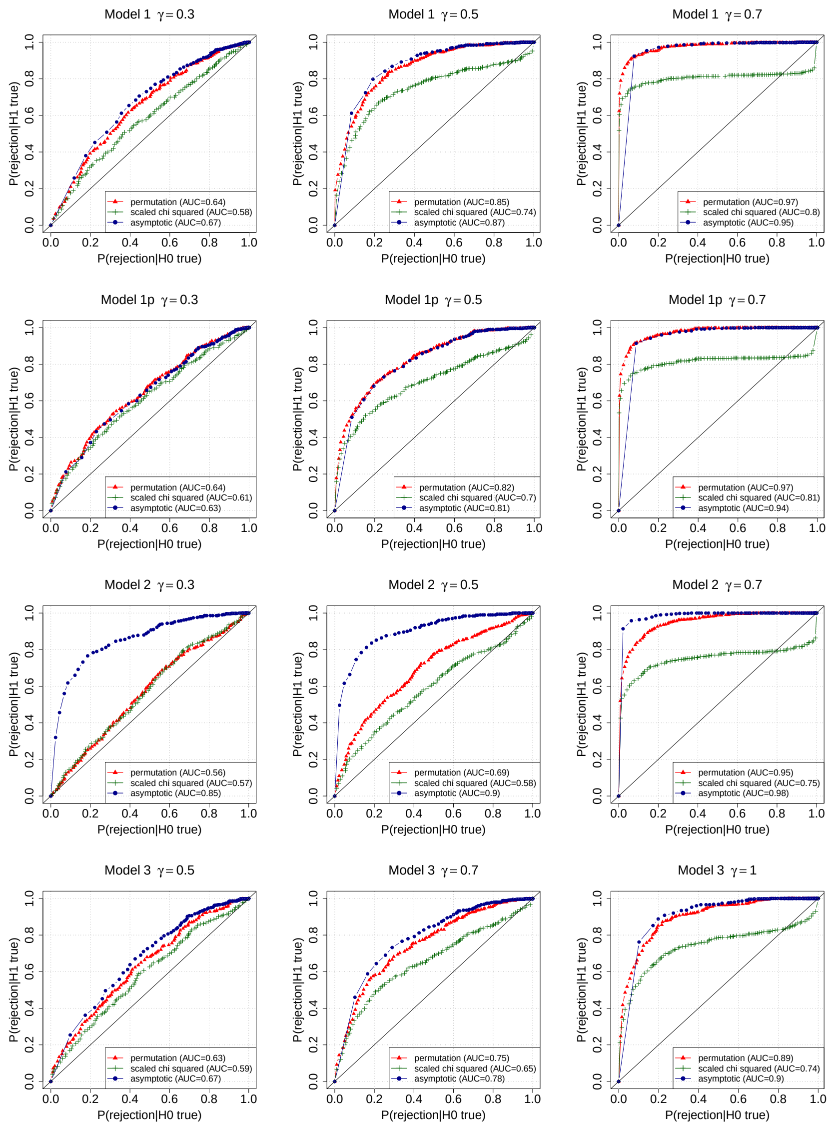

Fig. 3. ROC-type curves for simulation models 1, 1p, 2, 3 and permutation test (red), scaled chi-squared test (green) and asymptotic test (blue). Number of variables $d = 7$ and sample size $n = 1000$.