EYSM - example from the article

Table of contents

Introduction

In this repo there is an example presented in the EYSM short paper "Resampling Methods in Conditional Independence Testing". The details are described in the article.

- `resampling.R` - implementation of the resampling methods
- `simulation_function.R`, `simulations.R` - simulation-related files
- `model.R` - implementation of the model described in the section Model description
- `cpp_functions.cpp` - helper functions

Short description

Conditional mutual information is defined as

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

The estimator of a vector of probabilities $(p(x, y, z))_{x,y,z}$ is a vector of fractions $(\hat{p}(x, y, z))_{x,y,z} = (n(x, y, z)/n)_{x,y,z}$, where $n(x, y, z) = \sum_{i=1}^{n} \mathbb{I}(X_i = x, Y_i = y, Z_i = z)$. We estimate conditional mutual information using a plug-in estimator, namely

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

Lemma 1

If $X \perp\!\!\!\perp Y|Z$ we have that $2nCMI(\hat{p}) \xrightarrow{d} \chi^2_{(|\mathcal{X}|-1)(|\mathcal{Y}|-1)|\mathcal{Z}|}$.

Theorem 2

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 $T_b^* = T(\mathbf{X}_{n,b}^*, \mathbf{Y}_{n,b}^*, \mathbf{Z}_{n,b}^*)$, $(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{Z}_n)$ is a sample and $(\mathbf{X}_{n,b}^*, \mathbf{Y}_{n,b}^*, \mathbf{Z}_{n,b}^*)$ is a resampled sample (CR or CP scenarios).

Tests

We consider three tests for testing conditional independence $H_0 : X \perp\!\!\!\perp Y|Z$:

- `asymptotic` - a test based on Lemma 1,
- `exact` - a test based on Theorem 2,
- `df estimation` - a test that uses the chi-squared distribution as a benchmark distribution, but adjusts the number of degrees of freedom d based on resampled samples, $\hat{d} = \frac{1}{B} \sum_b \widehat{CMI}_b^*$.

Model description

Joint probability in the model is given as

$$p(x, y, z_1, z_2, z_3, z_4) = p(y)p(x|y) \prod_{s=1}^{4} p(z_i|y),$$

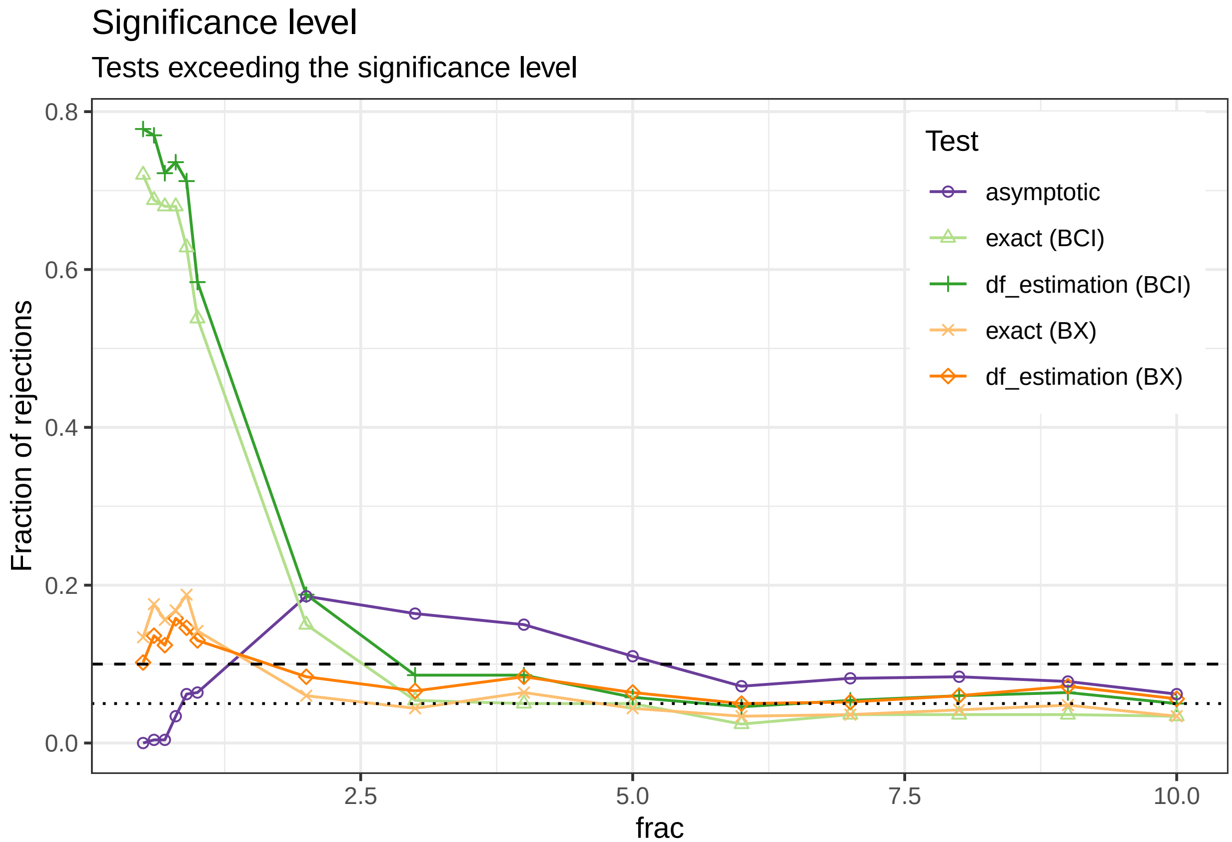
and is presented in Figure 1. Y is a Bernoulli random variable with probability of success equal to 0.5 and conditional distribution of $\tilde{X}$ and $\tilde{Z}_i$ for $i = 1, 2, 3, 4$ given $Y = y$ follows a normal distribution: $\tilde{X}|Y = y \sim N(y, 1)$, $\tilde{Z}_i|Y = y \sim N(\gamma^i y, 1)$ and $\gamma \in [0, 1]$ is a parameter. In this example, $\gamma = 0.5$. In order to obtain discrete variables from continuous $(\tilde{X}, \tilde{Z}_1, \tilde{Z}_2, \tilde{Z}_3, \tilde{Z}_4)$ we define

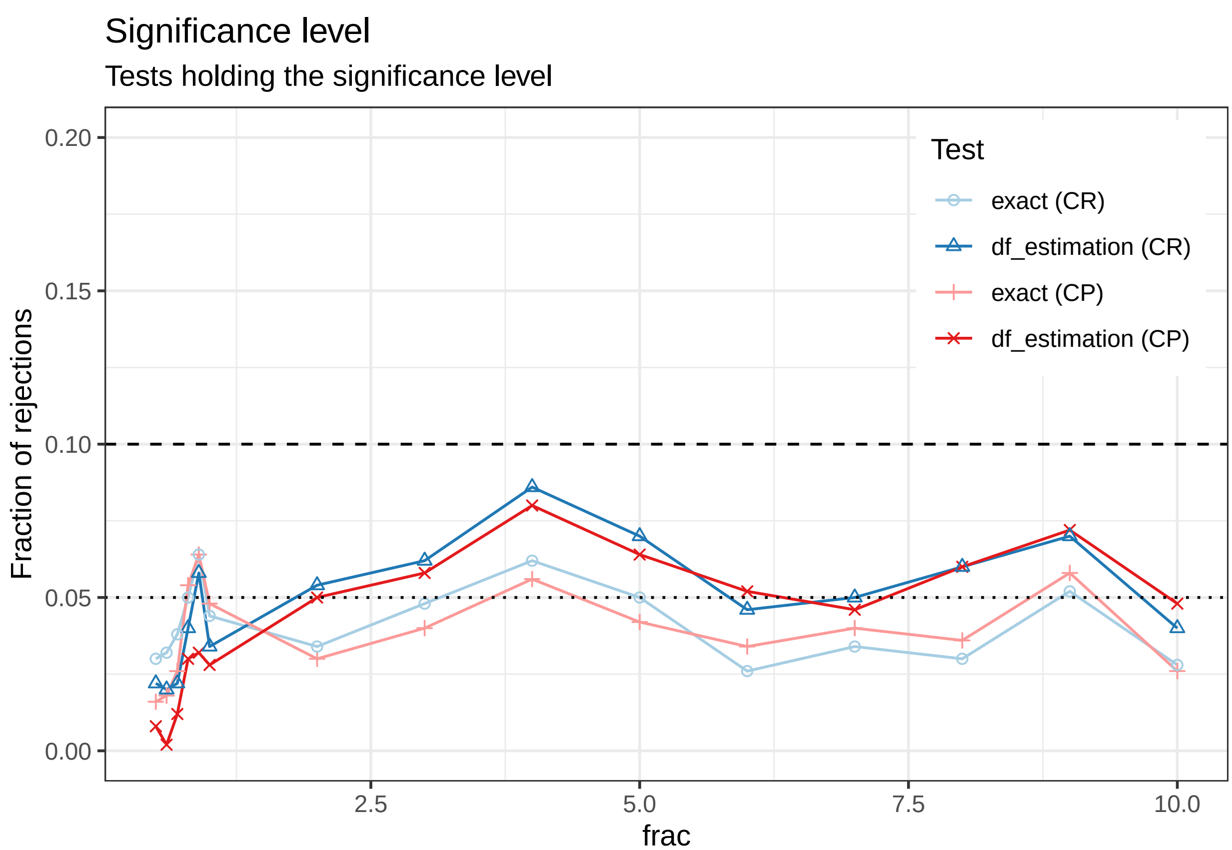
$$P(X = x|Y = y) = P\Big((-1)^x \tilde{X} \leq \frac{(-1)^x}{2} |Y = y\Big),$$

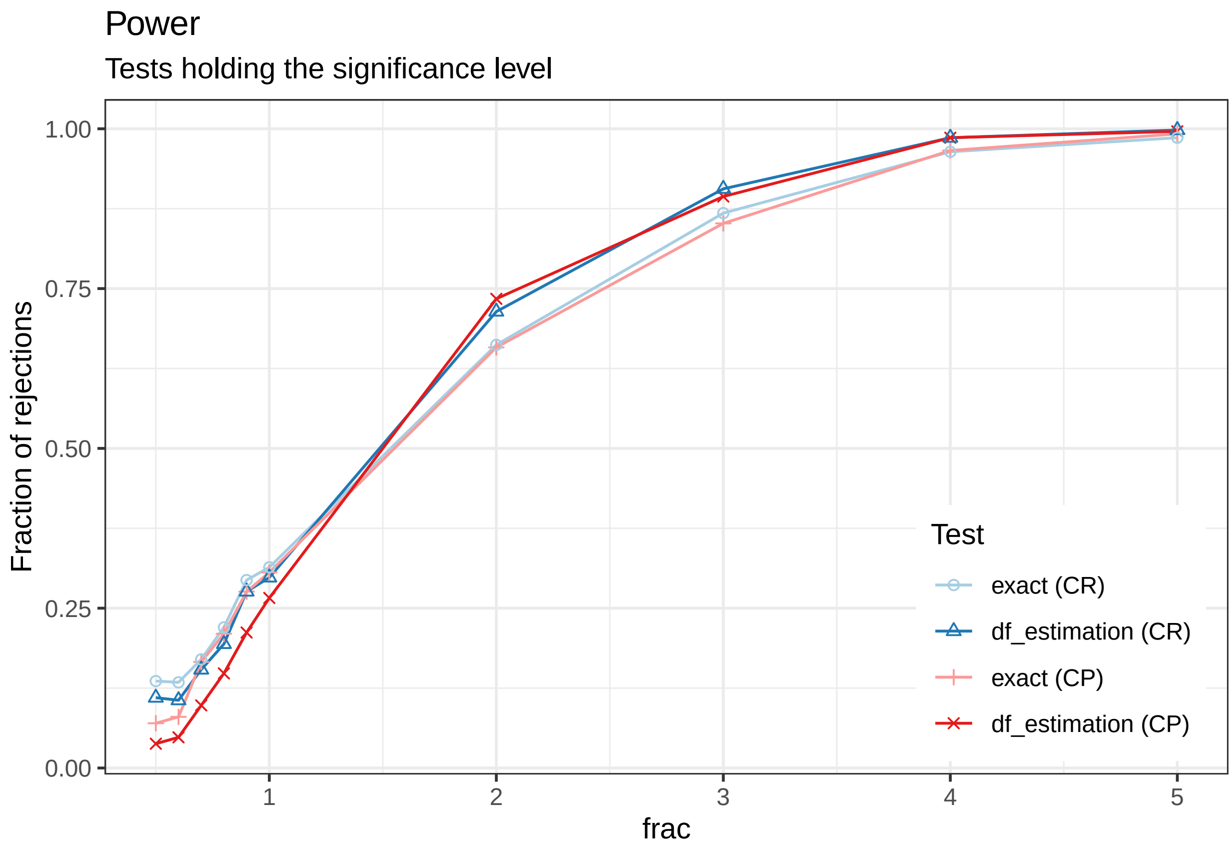
$$P(Z_i = z_i|Y = y) = P\Big((-1)^{z_i} \tilde{Z}_i \leq \frac{(-1)^{z_i}\gamma^i}{2} |Y = y\Big)$$

for $i = 1, 2, 3, , 4$, where $x, z_1, z_2, z_3, z_4 \in \{0, 1\}$. Thus $X|Y = y \sim Bern(\Phi((2y - 1)/2)$ and $Z_i|Y = y \sim Bern(\Phi(\gamma^i(2y - 1)/2))$. Variables X, Z_1, Z_2, Z_3, Z_4 are conditionally independent given Y but X an Y are not conditionally independent given Z_1, Z_2, Z_3, Z_4.

Plots


Significance level
Tests exceeding the significance level
Fraction of rejections
0.0
0.2
0.4
0.6
0.8
2.5
5.0
7.5
10.0
frac
Test
asymptotic
exact (BCI)
df_estimation (BCI)
exact (BX)
df_estimation (BX)


Significance level
Tests holding the significance level
Fraction of rejections
0.00
0.05
0.10
0.15
0.20
2.5
5.0
7.5
10.0
frac
Test
exact (CR)
df_estimation (CR)
exact (CP)
df_estimation (CP)


Power
Tests holding the significance level
1.00
0.75
0.50
0.25
0.00
Fraction of rejections
1
2
3
4
5
frac
Test
exact (CR)
df_estimation (CR)
exact (CP)
df_estimation (CP)