Conditional Independence test for Categorical Data towards Causal Discovery and Application in Bibliometrics

Sagar Nagaraj Simha

Master's Computer Science for Digital Media
Bauhaus University Weimar, Germany

27.10.2022

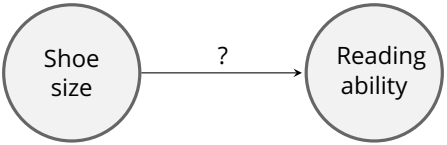
Overview

Motivation

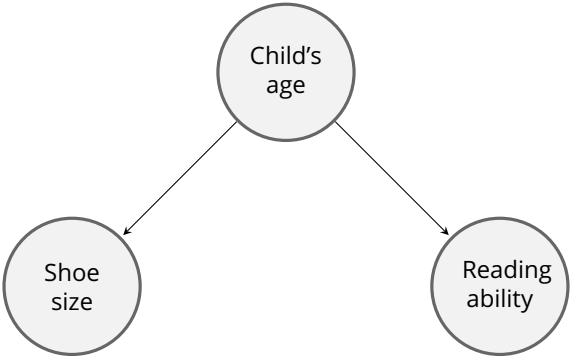
Correlation is not causation

Correlation is not causation

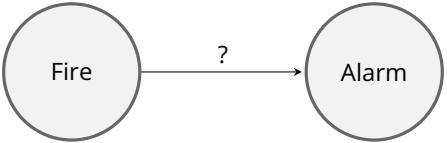
In a study on child development ...



Confounder

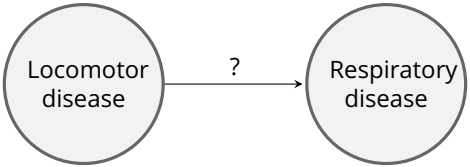


Correlation is not causation



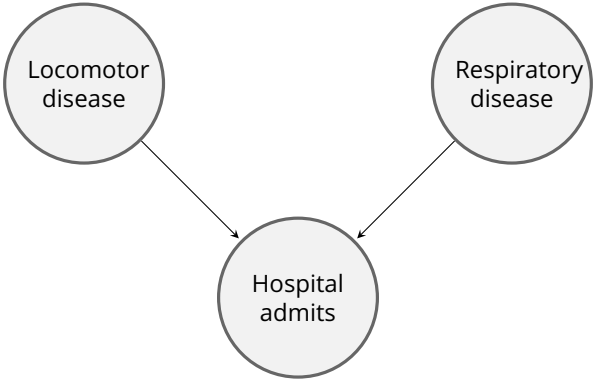
Mediator

Correlation is not causation



[Lee et al, 2019]

Selection bias (Collider bias)



PC Algorithm [Spirtes et al, 2000]

PC Algorithm

Under the assumptions of faithfulness, causal markov condition and causal sufficiency,

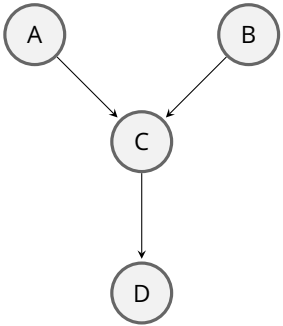


Figure: True causal graph

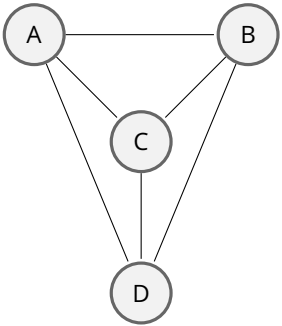


Figure: PC : Bi-partite graph

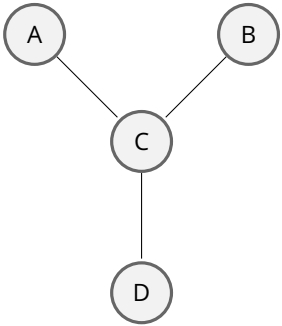


Figure: PC 1: Skeleton discovery. $X \perp\!\!\!\perp Y \mid Z$

PC Algorithm

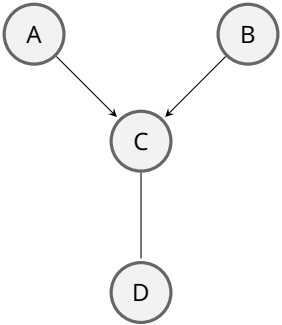


Figure: PC 2: Orienting colliders

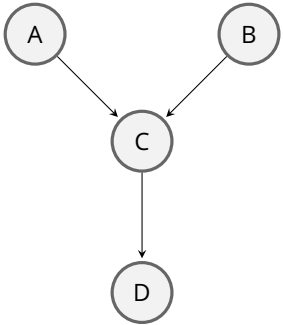


Figure: PC 3: Orientation rules

Motivation for design of CI test

- CI testing is crucial for constraint-based causal discovery algorithms.
- Focus on categorical/symbolic variables. E.g. Blood type of a person, number of sides in dice.
- Calibration of CI test (type I error <= α, maximize power).
- Evaluation using Bayesian Network (DGP).
- Comparison of exact and asymptotic test.
- Work heavily based on CMIknn, a non-parametric test for continuous variables [Runge, 2018]

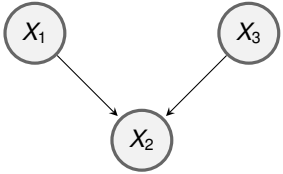
Data Generating Process

Structural Causal Model [Peters et al, 2017]

A structural causal model $\mathfrak{C} := (\mathrm{S}, \mathrm{P}_N)$ consists of a collection S of d structural assignments over random variables $X = \{X_1, ..., X_j\}$, where $X_j := f_j(PA_j, N_j), j = 1, 2, ...d$

- X_j is caused by parents $PA_j \subseteq \{X_1, ..., X_d\}$ through mechanism f_j
- Entails a joint distribution over all X
- N_j noise corresponding to X_j

A causal graph (DAG) is then $G(\mathfrak{C})$



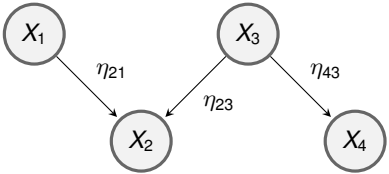
Bayesian Network

Categorical data generating process

A Bayesian Network is a DAG,

$$P(X_1, X_2, ..., X_n) = \prod_i P(X_i | PA_i)$$

- $X_1, ..., X_n$ are categorical random variables with m states each.
- Modeling dependencies using conditional probability tables (CPTs)
- Link strenth $\eta \in [0, 1]$ -> CPTs [Kokkonen et al, 2005]
- Assume that directed arrows denote causal directions.



Conditional independence testing

Exact vs Asymptotic tests

Given categorical random variables X, Y and Z,

$$H_0 : X \perp\!\!\!\perp Y \mid Z \tag{1}$$

$$H_1 : X \not\perp\!\!\!\perp Y \mid Z \tag{2}$$

Comparison [Tsamardinos et al, 2010] :

- Exact test (CMISymbPerm)
- Asymptotic test (G^2)
- For large sample sizes, the exact and asymptotic p-values are very similar.

CMISymbPerm

Test statistic

Conditional mutual information

$$
\begin{aligned}
I_{(X;Y|Z)} &= \sum_{x,y,z} p(x,y,z)\, log \left(\frac{p(x,y|z)}{p(x|z)p(y|z)} \right) \\
&= H(X|Z) - H(X|YZ) \\
&= H(XZ) + H(YZ) - H(XYZ) - H(Z)
\end{aligned}
\tag{3}
$$

- H denotes the Shannon entropy assuming that $p(\cdot)$ exist.
- The CMI $I_{X;Y|Z} = 0$ iff $X \perp\!\!\!\perp Y \mid Z$.

CMISymbPerm

Null distribution generation - Permutation scheme

Algorithm 1 Algorithm for permutation scheme

1. Compute CMI t_{obs} of data $\{x_i, y_i, z_i\}_{i=1}^{N}$.
2. Compute neighbors for each sample in z.
3. Randomly permute x within neighbors. Compute CMI T.
4. Repeat B times : $T_1, T_2, \ldots T_B$.
5. The approximate p-value is

$$\frac{1}{B}\sum_{j=1}^{B} I(T_j > t_{obs})$$

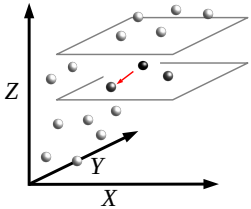


Figure: Permuting within z

G^2 test of conditional independence

$$G^2 = 2\sum_{xyz} N_{xyz} \ln\left(\frac{N_{xyz}}{E_{xyz}}\right) \tag{4}$$

N_{xyz} = observed frequencies of $X = x, Y = y, Z = z$, where $z = (z_1, ..., z_k)$.
N_{xz} = marginal total of $X = x, Z = z$.
N_{yz} = marginal total of $Y = y, Z = z$.
N = Total sample size.
E_{xyz} = $\frac{N_{xz}N_{yz}}{N_z}$ = Expected frequencies under the assumption of independence.

Null distribution - χ^2 distributed with degrees of freedom $(|X| - 1)(|Y| - 1)\prod_{i=1}^{k} |Z_i|$.

Experimental setup

CMISymbPerm and G^2 are run over,

- Model for DGP, c $\in$ [0, 1]

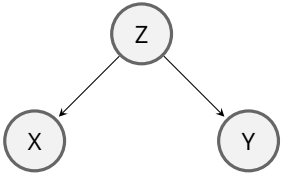


Figure: H_0 : $X \perp\!\!\!\perp Y \mid Z$

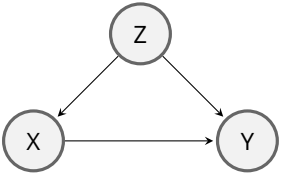


Figure: H_1 : $X \not\perp\!\!\!\perp Y \mid Z$

- Number of samples $N \in$ [50, 100, 150, 200, 250, 500, 1000, 2000]
- Number of categories/symbols $n_{symbs} \in$ [2, 3, 4, 5, 6]
- Number of dimensions of Z, $D_z = |Z| \in$ [1, 2, 3, 4]

Results

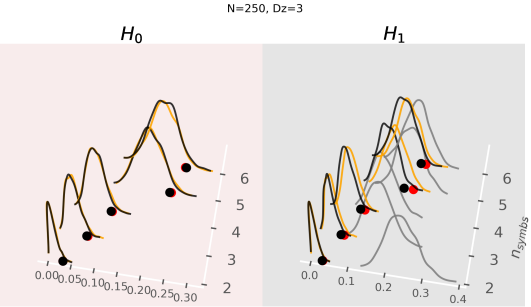


Figure: CMISymbPerm - null approximation

Results

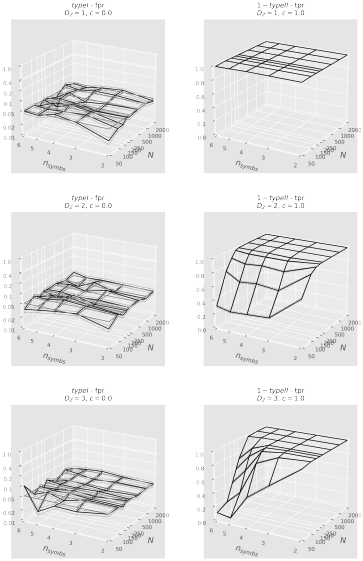


Figure: CMISymbPerm over n_{symbs} and N

Results

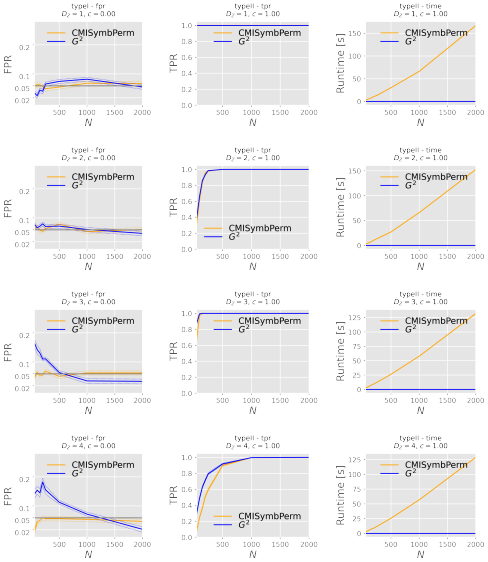


Figure: CMISymbPerm vs G^2. Over N and D_z with $n_{symbs} = 3$

Causal Discovery - Bibliometrics

Goal : Discover causal relationships amongst bibliometric features that influence a researcher's position in academia.

Dataset : Open Academic Graph 2.1 [Zhang et al, 2019]

- 240 million papers
- 243 million authors
- 53 thousand venues
- 25 thousand affiliation

Sample data :

- 10 features - continuous valued.
- Stratified sampling across 'position' (professor, associate professor, assistant professor)
- 1600 samples in each category. Total samples = 4800.

Correlation

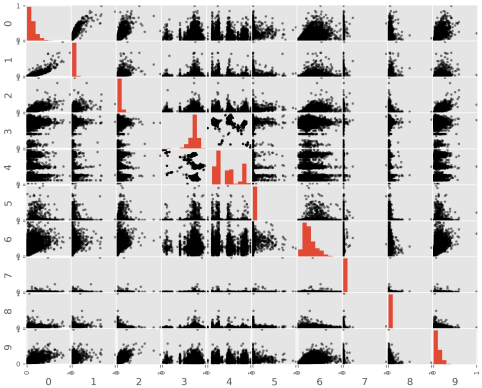


Figure: Scatter plot of features

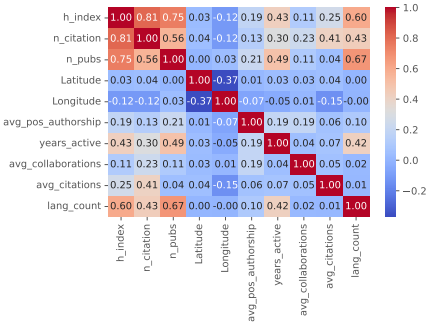


Figure: Correlation matrix

Multinomial logistic regression on *position*

k-fold cross validation
Mean accuracy = 0.582. Standard deviation = 0.02
Permutation feature importance (Dominant features)

Feature	Mean score	Standard deviation
years_active	0.132	0.006
n_pubs	0.036	0.005
h_index	0.014	0.003
n_citation	0.011	0.004
longitude	0.010	0.003
avg_citations	0.006	0.002

Causal graph

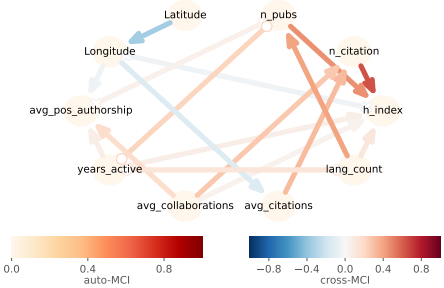


Figure: Causal graph. pc_alpha=0.005

- Confounded, no direct link
  - n_pubs and n_citations (corr = 0.56)
  - years_active and n_citations (corr = 0.3)
- lang_count and n_citations has no direct link but corr = 0.43.
- Mediated path amongst dominant features. Longitude of the affiliation –> average citations –> citations –> h-index

Conclusion

CI test :

- G^2 test and permutation test converge in type I error for larger samples.
- Time complexity - Permutation test is $\mathcal{O}(c^n)$, G^2 is $\mathcal{O}(1)$ for large samples.
- Prefer permutation test for lower sample size and larger dimensions, while G^2 for larger sample sizes.
- Prefer G^2 when time is a constraint.

Bibliometrics:

- Identified confounded and mediated paths amongst dominant features.

Future Work

- Parallelizing permutation scheme to improve time complexity.
- Evaulate using more models (chains, colliders ...)
- Selection bias in location and position
- Violation of causal sufficiency
- Location non-linearly related to other variables.
- pc_alpha=0.005 very conservative. Varied pc_alpha.
- Mixed variable CI test

References I

- Catalogue of bias collaboration, Lee H, Aronson JK, Nunan D. Collider bias. In Catalogue Of Bias. 2019. https://catalogofbias.org/biases/collider-bias/
- Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. Causation, prediction, and search. MIT press, 2000.
- Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Elements of Causal Inference: Foundations and Learning Algorithms. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2017. ISBN 978-0-262-03731-0. URL https://mitpress.mit.edu/books/elements-causal-inference.
- Teemu Kokkonen, Harri Koivusalo, Hanne Laine, Ari Jolma, and Olli Varis. A method for defining conditional probability tables with link strength parameters for a bayesian network.2005.
- I. Tsamardinos and Giorgos Borboudakis. Permutation testing improves bayesian network learning. In ECML/PKDD, 2010a.
- Larry Wasserman. All of statistics : a concise course in statistical inference. Springer, New York, 2010.
- Yvonne M Bishop, Stephen E Fienberg, and Paul W Holland. Discrete multivariate analysis: theory and practice. Springer Science Business Media, 2007
- Jakob Runge. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In Amos Storkey and Fernando Perez-Cruz, editors, Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, volume 84 of Proceedings of Machine Learning Research, pages 938–947. PMLR, 09–11 Apr 2018. URL https://proceedings.mlr.press/v84/runge18a.html.

References II

Fanjin Zhang, Xiao Liu, Jie Tang, Yuxiao Dong, Peiran Yao, Jie Zhang, Xi-aotao Gu, Yan Wang, Bin Shao, Rui Li, and Kuansan Wang. Oag: Toward linking large-scale heterogeneous entity graphs. In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), August 2019.

Thank you

Appendix

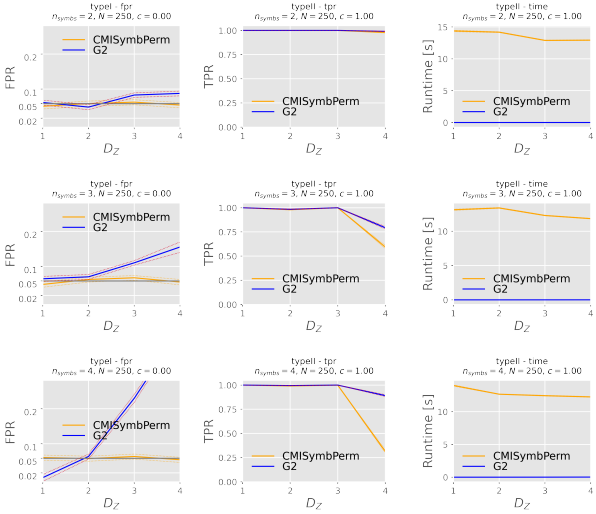


Figure: CMISymbPerm vs G^2. Over D_z with $N = 250$ and $n_{symbs} = 2, 3, 4$

Appendix

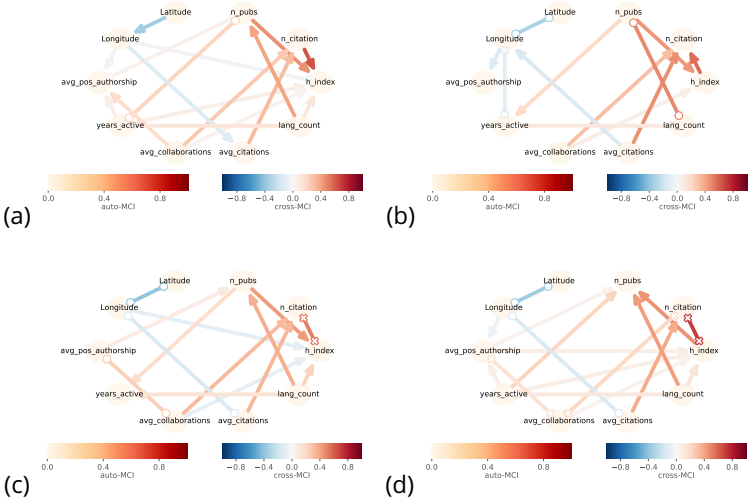


Figure: pc_alpha=0.005 (a) All authors (b) Professors (c) Associate professors (d) Assistant professors

Appendix

Link Strength-Encode association strength

Link strength [Kokkonen et al, 2005] $\eta = [0, 1]$ defines the strength of association between a parent node p and child x, all nodes having m states each.

1. $\eta = 0$, child is independent of the parent p.
2. $\eta_1 = \eta_2$, both parents p_1 and p_2 have equal effect on child x.
3. When all $\eta = 0$, the CPT is non-informative.

- Reduces CPT elicitation complexity from m^{p+1} to p.
- Helps elicit relative causal strengths.

Appendix

Generalized Noisy-Or model with link strength

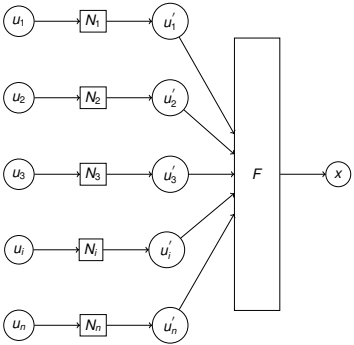


Figure: Schematic of the generalized Noisy-Or model.

$$N_i = P(u_i'(r)|u_i(c)) = \begin{cases} \frac{1}{m} + \eta_i(1 - \frac{1}{m}) & \text{if } r = c \\ \frac{1}{m-1}[1 - \frac{1}{m} - \eta_i(1 - \frac{1}{m})] & \text{if } r \neq c \end{cases} \tag{5}$$

$$F(u') = x\left(ceiling\left(\frac{1}{\sum_i \eta_i)}\sum_i[\eta_i I(u_i')]\right)\right) \tag{6}$$

$$P(x|u) = \sum_{u':x=F(u')} P(u'|u) = \sum_{u':x=F(u')} \prod_{u'} P(u_i'|u_i) \tag{7}$$