A Bayesian Nonparametric Conditional Two-sample Test with an Application to Local Causal Discovery

UNIVERSITY OF AMSTERDAM

Philip A. Boeken, Joris M. Mooij

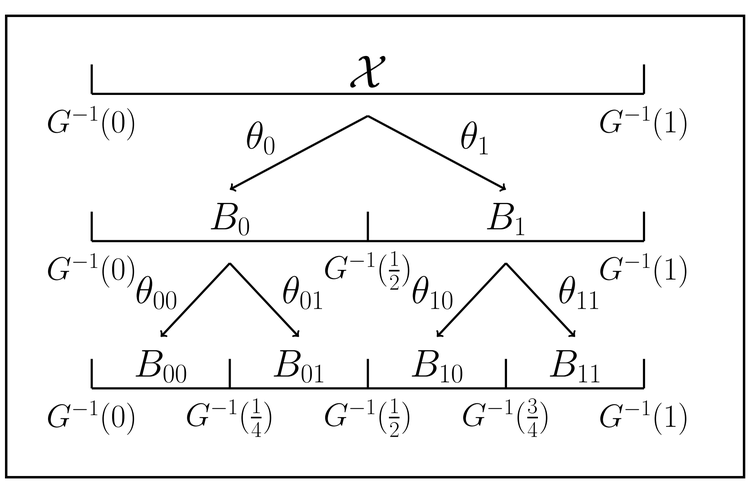
Korteweg-de Vries Institute for Mathematics, University of Amsterdam

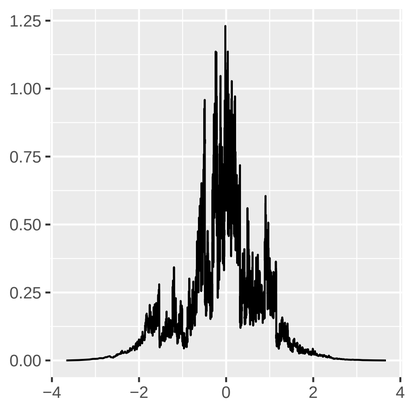
Abstract

Conditional independence (CI) testing is paramount to many constraint-based causal discovery algorithms. Many applications require 'mixed' CI testing: $C \perp\!\!\!\perp X|Z$, where C is binary (discrete) and X, Z are continuous. To our knowledge, only parametric mixed tests are available. We propose a nonparametric conditional two-sample test by combining the works of Holmes et al. (2015) and Teymur and Filippi (2020), and analyse its performance when used in the Local Causal Discovery algorithm.

Pólya Tree (Lavine, 1992)

Random probability measure $\mathcal{P}$ on $\mathcal{X}$.





- $(\theta_{\kappa 0}, \theta_{\kappa 1}) \sim \mathrm{Dir}(\alpha_{\kappa 0}\alpha_{\kappa 1})$
- $\mathcal{P}(B_\kappa) = \prod_{i=1}^{|\kappa|} \theta_{\kappa_1 \dots \kappa_i}$
- Pólya tree: $\mathcal{P} \sim \mathrm{PT}(\mathcal{A}, \mathcal{T})$
- $\mathcal{A} := \{\alpha_0, \alpha_1, \alpha_{00}, \alpha_{01}, \dots\}$
- $\mathcal{T} := \{\mathcal{X}, \{B_0, B_1\}, \{B_{00}, B_{01}, B_{10}, B_{11}\}, \dots\}$

$$X \sim \mathcal{P} \implies p(X_{1:n}) = \prod_\kappa \frac{\mathrm{B}(\alpha_{\kappa 0} + n_{\kappa 0}, \alpha_{\kappa 1} + n_{\kappa 1})}{\mathrm{B}(\alpha_{\kappa 0}, \alpha_{\kappa 1})}$$

Conditional Optional Pólya Tree (Ma, 2017)

Conditional random probability measure Φ on $\mathcal{X}|\mathcal{Z}$.

- $\tilde{\mathcal{T}}_Z$: random subset of $\mathcal{T}_Z = \{\mathcal{Z}, \{B_0, B_1\}, \{B_{00}, B_{01}, B_{10}, B_{11}\}, \dots\}$
  Add $B_{\kappa 0}, B_{\kappa 1}$ to $\tilde{\mathcal{T}}_Z$ if $S_\kappa = 0$ for $S_\kappa \sim \mathrm{Bernoulli}(\rho)$
- 'Local' Pólya trees $\mathcal{P}(\cdot|B_\kappa) \sim \mathrm{PT}(\mathcal{T}_X, \mathcal{A})$ for all $B_\kappa \in \tilde{\mathcal{T}}_Z$
- Conditional Optional Pólya Tree: $\Phi \sim \text{Cond-OPT}(\rho, \mathcal{T}_Z, \mathcal{T}_X, \mathcal{A})$

$$\Phi(X|Z \in B_\kappa) = \begin{cases} p(X|Z \in B_\kappa) & \text{if } |Z_{1:n} \cap B_\kappa| \leq 1 \\ \rho \cdot p(X|Z \in B_\kappa) & \\ \quad + (1-\rho)\Phi(X|Z \in B_{\kappa 0})\Phi(X|Z \in B_{\kappa 1}) & \text{otherwise} \end{cases}$$

Conditional Independence Test

- Split dataset: $X^{(i)} := X|\{C = i\}$ for $i = 0, 1$
- Hypotheses: $H_0 : C \perp\!\!\!\perp X|Z \iff X|Z \sim \Phi$

$$H_1 : C \not\perp\!\!\!\perp X|Z \iff \begin{cases} X^{(0)}|Z \sim \Phi^{(0)} \\ X^{(1)}|Z \sim \Phi^{(1)} \end{cases}$$

  with priors $\Phi, \Phi^{(0)}, \Phi^{(1)} \overset{\text{i.i.d.}}{\sim} \text{Cond-OPT}(\rho, \mathcal{T}_Z, \mathcal{T}_X, \mathcal{A})$

- Bayes Factor:

$$\mathrm{BF}_{01} = \frac{\Phi(X|Z)}{\Phi^{(0)}(X^{(0)}|Z)\Phi^{(1)}(X^{(1)}|Z)}$$

Local Causal Discovery (Cooper, 1997; Mooij, 2020)

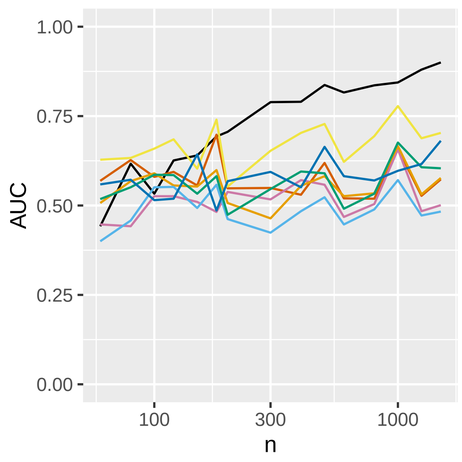
If the data generating process of the random variables (C, X, Y) has no selection bias, can be modelled by a faithful simple SCM, and X is not a cause of C, then the presence of (in)dependences

$$C \not\perp\!\!\!\perp X, \quad X \not\perp\!\!\!\perp Y, \quad C \perp\!\!\!\perp Y|X$$

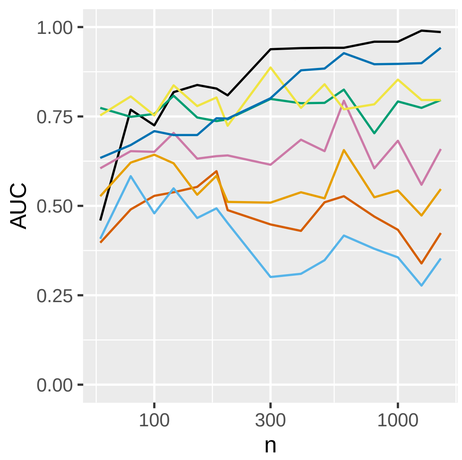
implies that X is a (possibly indirect) cause of Y.

Simulations

$C \not\perp\!\!\!\perp X|Z$



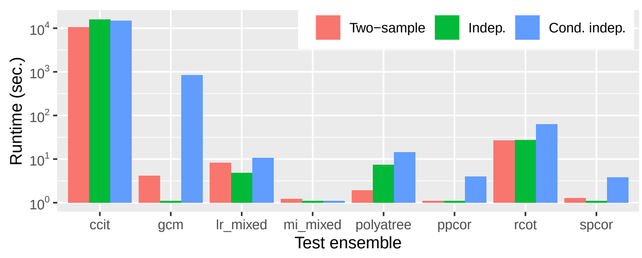
LCD Triple (C, X, Z)



- This work

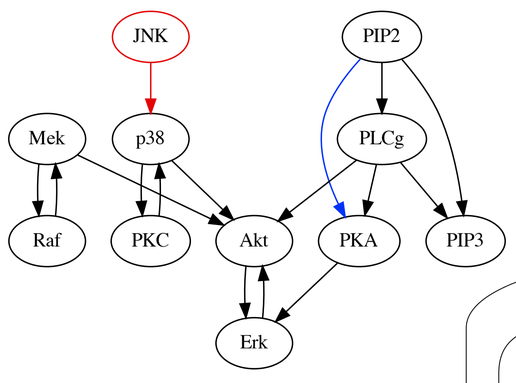
Parametric mixed tests:
- Mutual information (Scutari, 2010))
- Likelihood ratio (Sedgewick et al., 2019)

(Non)parametric continuous tests:
- Pearson Partial Correlation
- Spearman Partial Correlation
- GCM (Shah and Peters, 2020)
- RCoT (Strobl et al., 2019)
- Classifier CI Test (Sen et al., 2017)

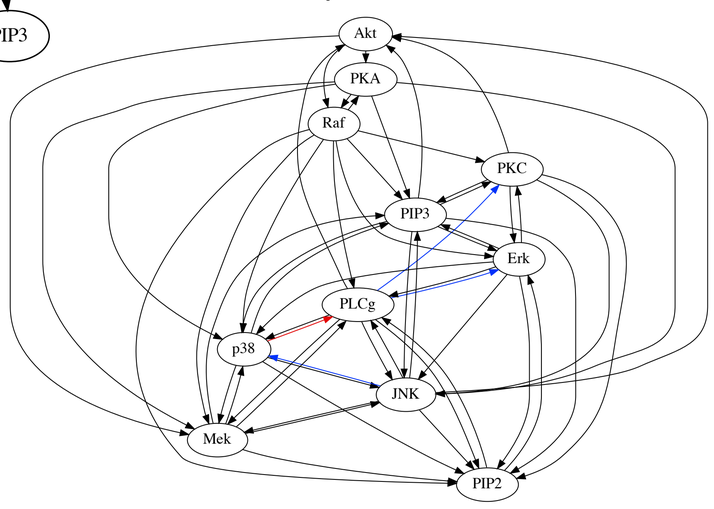


Protein Expression Data (Sachs et al., 2005)

LCD with Pólya tree tests



LCD with partial correlation



Conclusions

- The Pólya tree test can perform better than parametric mixed tests and nonparametric continuous tests, and the choice of CI test heavily influences the performance of LCD.
- Further research: consistency, extend to discrete C, multidimensional Z, choice of parameters.

JMM was supported by the ERC under the European Union's Horizon 2020 programme (grant agreement 639466).

37th Conference on Uncertainty in Artificial Intelligence (UAI) July 27 - 29, 2021