Practical for "Causal Learning" on the UBRA Data Train – December 2025

Vanessa Didelez, Luca Bergen, Ronja Foraita

Part 3

We consider a (much simplified subset of the) genetic regulatory system for head-and-neck squamous cell carcinoma (HNSCC) – the original system has 73 nodes. The causal role of the high mobility group AT-Hook 2 gene (HMGA2) in the protein 53 signalling pathway is of particular interest as this could be a target for drug development (Foraita et al., 2020).

The `tcgas` dataset contains (log-transformed) gene expressions of the genes "BAX", "CDKN2A", "CDKN1A", "HMGA2", "MDM2", "SERPINE1", "THBS1", "CDK6" in tumour tissues from n=392 HNSCC patients. There is also the additional dataset `tcgadisc` where all variables have been discretised into three categories.

- Investigate descriptively if the data look normally distributed with linear relations.
- Using R-package `pcalg`, apply the PC algorithm with the Z-test and with 1%, 5%, 10% nominal significance level and choose the options maj.rule = T, solve.confl = T, u2pd = "relaxed" for a fully order independent output. Compare the three graphs wrt the causal role of HMGA2.
- Discuss why the FCI algorithm may be more appropriate, and apply it to the same data (with 5% level).
- Discuss how Greedy equivalence search with Gaussian-BIC score is expected to related to the above PC and FCI. Apply GES to the same data.
- Explain why the LiNGAM approach is likely to be unsuitable for the current data. Apply it anyway.
- Consider now the following approaches to relax the (implicit) linearity assumption of the above methods:
    - randomised kernel conditional independence testing (with function RCIT with `suffStat = list(data = tcgas)` in pcalg)
    - use the discretised variables and the G^2-test
    - the generalised covariance measure (with function GCM with `suffStat = list(data = tcgas)` in pcalg)

  Run these with 5% and 10% nominal levels.
- Compare all the above with regard to the inferred causal role of HMGA2 (i.e. (possible) descendants and ancestors).
- Based on the output from GES, use the IDA function to estimate (i) the causal effect of HMGA2 on CDK6 and (ii) the causal effect of MDM2 on CDK6. Compare these with a linear regression of CDK6 on everything.

Foraita, R., Friemel, J., Günther, K., Behrens, T., Bullerdiek, J., Nimzyk, R., Ahrens, W. and Didelez, V. (2020), Causal discovery of gene regulation with incomplete data. J. R. Stat. Soc. A, 183: 1747-1775. https://doi.org/10.1111/rssa.12565

You find example R code by typing the following in R:

```
> library(DataTrainCausalLearning)
> openCode("Causal3_code.html")
```