Practical for "Intro to Causal Discovery" – July 2022

Vanessa Didelez, Christine W Bang, Ronja Foraita

Note: *You find the data and R code (including an installation guide) on:*
`https://github.com/bips-hb/IntroductionCausalDiscovery`

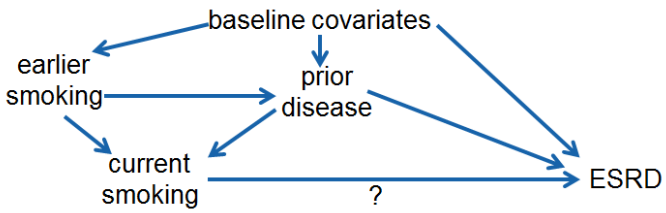
Note that this was originally set-up for a different (longer) course and may therefore contain much more material than was covered at the Reisensburg.

Causal DAGs – Basic understanding

Consider the causal DAG below (Staplin et al, 2016). ESRD = 'end-stage renal disease' is the binary outcome and all other nodes represent binary covariates with a set of baseline variables not further detailed.
Use the web-based version of `DAGitty` (or if you prefer, there is some example R code in `datatrain1`).

Enter the DAG below in DAGitty and investigate:

- (i) With / (ii) without data on 'early smoking' are there any testable implications of the causal model?
- If we want to identify the causal effect of each (i) prior disease, or (ii) current smoking, or (iii) earlier smoking on ESRD, what do we need to adjust for (and for what not), respectively?
- Specifically: can the causal effect of current smoking on ESRD be identified without data on 'early smoking'? Convince yourself that your chosen adjustment set blocks all back-door paths.
- Add a node for a further, unobservable, variable that directly affects prior disease and ESRD. Can the effect of current smoking on ESRD still be identified from the measured data? Explain your conclusion.



please turn over

Causal Discovery

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. Example R code can be found in `Causal-Discovery`.

- 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:
    - Kernel conditional independence testing
    - Use the discretised variables and the G^2-test
    - *This takes 30min to run (see if you have the time):* the generalised covariance measure

  Run these with 5% and 10% nominal levels. Alternatively, grid-search for the level that will result in 12 edges.
- 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