Correlation & causation
(practice in R)

The dataframe **data.rds** contains multivariate normally distributed data with a dependency structure corresponding to the DAG in **Figure 1**. We will use the PC algorithm for structure learning, but first we will look at the steps involved in the inference procedure.

Problem 1: Testing for marginal correlation

The covariance between two random variables X and Y captures their linear relationship, and is defined as

$$\mathrm{Cov}(X, Y) := \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right].$$

Their correlation $\rho_{X,Y} := \dfrac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$

is merely their covariance scaled by the product of their respective standard deviations. Note that for a multivariate normal distribution, uncorrelated variables are independent. However, it is important to keep in mind that this implication does not hold in general.

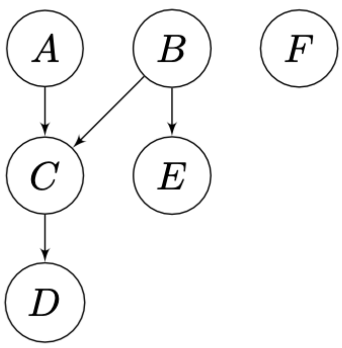

Figure 1

Using the data from **data.rds**, display the observations of A and B in a scatterplot. What does the plot suggest about their (marginal) correlation? Does it agree with **Figure 1**? Use the function `cor.test()` to test the null hypothesis of no correlation between A and B. What is your conclusion?

Problem 2: Testing for partial correlation

The partial correlation between two random variables X and Y given a random variable Z is

$$\rho_{X,Y|Z} := \frac{\rho_{X,Y} - \rho_{X,Z}\rho_{Y,Z}}{\sqrt{(1-\rho_{X,Z}^2)(1-\rho_{Y,Z}^2)}}$$

Alternatively, the partial correlation $\rho_{X,Y|Z}$ equals the correlation between residuals from the linear regressions of X on Z, and Y on Z, respectively. We will now compute the partial correlation $\rho_{A,B|C}$ to assess the association between A and B given C as follows:

- Linearly regress A on C (that is, with A as the response variable and C as the explanatory variable). Compute and store the residuals.
- Linearly regress B on C. Compute and store the residuals.
- Plot the residuals of A (regressed on C) against the residuals of B (regressed on C). What do you see?
- Use the function `cor.test()` to test the null hypothesis of no correlation between the residuals of A (regressed on C) and the residuals of B (regressed on C). What is your conclusion? Does this agree with your expectation based on the underlying DAG in **Figure 1**?

Problem 3: Running the PC algorithm

Install and load the R package **pcalg.** Use the function `pc()` to run the PC algorithm on the data in **data.rds**, and plot the result. Does the algorithm successfully learn the structure of the data-generating graph in **Figure 1**? How is the result affected by the significance level α for the conditional independence tests?

Hints: For the PC algorithm applied to normally distributed data, the sufficient statistics are the sample correlation matrix C of the data (see `cor()`), as well as the sample size n. Supply these as a list for the `suffStat` argument of the function `pc()`. Specify `indepTest = gaussCItest`, and set a reasonable significance level alpha for the independence tests. Supply the node names `colnames(data)` to the argument labels. Note that when plotting a pDAG, undirected edges are drawn as '↔' rather than '–'.