Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D-BSSE
Department of Biosystems
Science and Engineering

Statistical Models in Computational Biology

Niko Beerenwinkel
David Dreifuss
Pedro Ferreira
Xiang Ge Luo

Due 12th of May 2021
Please submit your project with the filename Lastname(s)_Project9.pdf.

Problem 23: d-separation **(3 points)**
For the following Bayesian network,

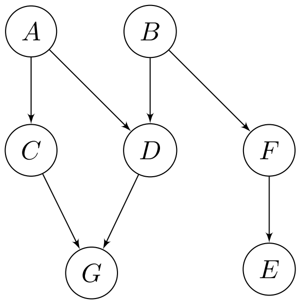

Figure 1

(i) Write down all the variables that are d-separated from A given {C,D}.

(ii) Indicate whether each statement is true or false and explain your choice.

(a) B is conditionally independent of C given D.
(b) G is conditionally independent of E given D.
(c) C is conditionally independent of F given A.
(d) C is conditionally independent of E given its Markov blanket (of C).

Programming exercises to be solved using R

The data frame MVN_DAG.rds contains multivariate normally distributed data with a dependency structure corresponding to the DAG in Figure 3. We will use the PC algorithm[1] for structure learning, but first we will look at the steps involved in the inference procedure.

Problem 24: Testing for marginal correlation **(1 point)**
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 $\rho_{X,Y} := \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}\, X\, \mathrm{Var}\, Y}}$ 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.

Using the data from MVN_DAG.rds[2], display the observations of A and B in a scatterplot. What does the plot suggest about their (marginal) correlation? Does it agree with Figure 3? Use the

[1] Section 5.4.2 in P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT press, Cambridge, MA, USA, 2nd edition, 2000.
[2] Please load the data frame using readRDS function.

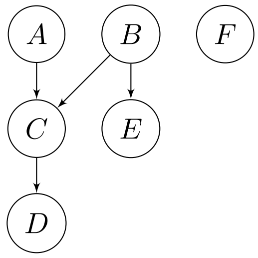

Figure 2

function cor.test() to test the null hypothesis of no correlation between A and B. What is your conclusion?

Problem 25: Testing for partial correlation **(2 points)**

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. For hints, see footnote[3].
- 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 3?

Problem 26: Running the PC algorithm **(2 points)**

Install and load the R package pcalg. Use the function pc() to run the PC algorithm on the data in MVN_DAG.rds, and plot the result. For hints, see footnote[4]. Does the algorithm successfully learn the structure of the data-generating graph in Figure 3? How is the result affected by the significance level α for the conditional independence tests?

[3]The R code lmFit <- lm(Y ~ X, data = Data) performs linear regression with Y as the response variable and X as the explanatory variable, where X and Y are columns of the data frame 'Data'. The function residuals(lmFit) computes the residuals of the fitted linear model lmFit.

[4]If you encounter trouble installing pcalg, make sure that you have a recent version of R, such as 3.4.4 or higher. 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 '–'.

Problem 27: Running the partition MCMC algorithm **(2 points)**

Install and load the R package `BiDAG`. Initialize the parameters using the function `scoreparameters()` on the data in `MVN_DAG.rds` with the Bayesian Gaussian equivalent (BGe) score. Run the partition MCMC algorithm using the function `partitionMCMC()` and plot the result. How is the result affected by the hyper-parameter α_μ[5]?

[5] See Project 1 Problem 3 for more details.