Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D-BSSE
Department of Biosystems
Science and Engineering

Statistical Models in Computational Biology

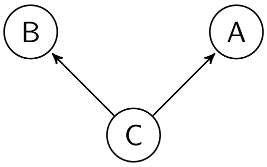
Jack Kuipers
David Dreifuss
Xiang Ge Luo
Rudolf Schill

Due date: 2th Mar 2023 before 12:00 pm noon

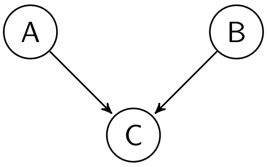
Problem 1: Conditional independence and BNs (3 points)

Consider the following graphical structures, corresponding to (different) Bayesian networks. For which network does the statement $A \perp B \mid C$ hold? For which does the statement $A \perp B$ hold? Prove your answers by the laws of probability.

a)

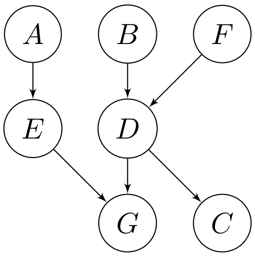


b)



Problem 2: Markov blanket (2 points)

Consider the following graphical structure of a Bayesian network:



Determine the Markov blanket $\mathrm{MB}(D)$ of node D and show that the conditional probability $P(D \mid A, B, C, E, F, G)$ is the same as $P(D \mid \mathrm{MB}(D))$.

Problem 3: Learning Bayesian networks from protein data (5 points)

In this exercise, we will use the R package `BiDAG`[1] to learn Bayesian networks from a dataset consisting of the measurements of 11 phosphorylated proteins and phospholipids derived from primary immune system cells, subjected to both general and specific molecular interventions [2].

(a) First, download the logged and standardized data file here. Read in the data in R. Report the number of variables n and the number of observations N. Visualize the transformed data using the ggpairs function in the R package `GGally`. Randomly split the data into 80% training data and 20% test data (run `set.seed(2023)` for reproducibility). Initialize the parameters using the function `BiDAG::scoreparameters` with the training data and the Bayesian Gaussian equivalent (BGe) score [3, 4]. (1 point)

[1] Run `install.packages("BiDAG")` in the R console. Then load the package by running `library(BiDAG)`. Read the help files of the package and use default parameters unless otherwise stated.

[Note: The BGe score is a fully-decomposable marginal likelihood function $P(\mathcal{D} \mid \mathcal{G})$ for scoring Bayesian networks. The main underlying assumption is that the data is normally distributed with $\mathcal{N}(\mu, W^{-1})$. The precision matrix W follows a Wishart prior $\mathcal{W}_n(T^{-1}, \alpha_w)$, where $\alpha_w > n - 1$ is the degrees of freedom and T is the positive definite parametric matrix. The mean vector μ follows a normal prior $\mathcal{N}(\nu, \alpha_\mu W)$ with $\alpha_\mu > 0$.]

(b) Learn a Bayesian network using the `BiDAG::iterativeMCMC` function. Plot the directed acyclic graph (DAG). Evaluate the log score of the test data against the estimated DAG using the function `BiDAG::scoreagainstDAG`. (1 point)

(c) One of the arguments in the `scoreparameters` function is `bgepar = list(am = 1, aw = NULL)`, which corresponds to the hyper-parameters α_μ and α_w for the BGe score. By default, $\alpha_\mu = 1$ and $\alpha_w = n + \alpha_\mu + 1$.

Now, consider the set of values $\{10^{-3}, 10^{-1}, 1, 10, 10^2\}$ for `am` and keep `aw = NULL` fixed. For each value, repeat the process of splitting the data, initializing the parameters, and learning the DAG for many times (*e.g.,* 100 times). Then, report the average number of edges in the DAGs and the average log score of the test data in a table as the one shown below. Remember to run `set.seed(2023)` for reproducibility. (*Hint: running the code parallelly with the function* `parallel::mclapply` *can help reducing the runtime. In this case, run* `RNGkind("L'Ecuyer-CMRG")` *for reproducibility.*)

Parameter `am`	10^{-3}	10^{-1}	1	10	10^2
Average number of edges					
Average log score of the test data					

What do you observe? Choose the value of `am` corresponding to the highest test score and plot the DAG re-learned from the whole dataset. (3 point)

References

[1] Suter, P., Kuipers, J., Moffa, G., & Beerenwinkel, N. (2023). Bayesian structure learning and sampling of bayesian networks with the r package BiDAG. *Journal of Statistical Software, 105, 1–31.*

[2] Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., & Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science, 308, 523-529.*

[3] Geiger, D., & Heckerman, D. (2002). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *The Annals of Statistics, 30, 1412-1440.*

[4] Kuipers, J., Moffa, G., & Heckerman, D. (2014). Addendum on the scoring of Gaussian directed acyclic graphical models. *The Annals of Statistics, 42, 1689-1691.*