# Probabilistic Graphical Models: Problem Set 4

Svante Linusson, Liam Solus
KTH Royal Institute of Technology

15 September 2023

1. Run the PC-algorithm by hand for the following set of CI-relations on 6 variables.

$$\mathcal{CI} = \{M \perp\!\!\!\perp S \mid F, \quad M \perp\!\!\!\perp H \mid F, \quad M \perp\!\!\!\perp C \mid F, \quad M \perp\!\!\!\perp L \mid F,$$
$$F \perp\!\!\!\perp H \mid S, \quad F \perp\!\!\!\perp L \mid C, \quad H \perp\!\!\!\perp L \mid C, \quad S \perp\!\!\!\perp C \mid F, H, \quad S \perp\!\!\!\perp L \mid C\}$$

   Assume that we know the CI relations $\mathcal{CI}$ hold in the data-generating distribution $\mathbf{X} = [S, M, F, C, L, H]^T$ and that $\mathbf{X}$ is faithful to its true causal structure $\mathcal{G}$. Can you identify $\mathcal{G}$ based on the output of the algorithm you just ran?

2. Suppose we have a random sample $\mathbb{D}$ (a set of independent and identically distributed observations) from $\mathbf{X} = [X_1, X_2, X_3, X_4]^T$. Based on conditional independence tests performed using the data $\mathbb{D}$ we estimate that $\mathbf{X}$ entails exactly two CI relations

$$\mathcal{CI} = \{X_1 \perp\!\!\!\perp X_3 \mid X_4, \quad X_2 \perp\!\!\!\perp X_4 \mid X_3\}.$$

   Can you use the PC algorithm to learn a DAG representation of $\mathbf{X}$?

3. We say that a distribution $\mathbf{X}$ satisfies **restricted faithfulness** with respect to a DAG $\mathcal{G} = ([m], E)$ if $\mathbf{X}$ satisfies the global Markov property with respect to $\mathcal{G}$ and

   - For all edges $i \to j \in E$, $X_i \not\perp\!\!\!\perp X_j \mid X_S$ for all $S \subset [m] \setminus \{i, j\}$, and
   - For all paths $\langle i, j, k \rangle$ in $\mathcal{G}$ with $i$ and $k$ not adjacent in $\mathcal{G}$ and for all subsets $C \subset [m] \setminus \{i, k\}$ such that $i$ and $k$ are d-connected given $C$ in $\mathcal{G}$, $X_i \not\perp\!\!\!\perp X_k \mid X_C$.

   (a) Show that if $\mathbf{X}$ is faithful to $\mathcal{G}$ then it also satisfies restricted faithfulness with respect to $\mathcal{G}$.

   (b) Suppose $\mathbf{X}$ satisfies restricted faithfulness with respect to $\mathcal{G}$. Show that, when given a conditional independence test that perfectly answers any query for $\mathbf{X}$, any acyclic orientation of the output of the PC algorithm that does not introduce new v-structures will be Markov equivalent to $\mathcal{G}$.

4. Let $\mathbf{X} = [X_1, \ldots, X_m]^T$, and suppose that $\mathbf{X}$ is faithful to $\mathcal{G}$ and $\mathcal{H}$ is a minimal I-MAP of $\mathbf{X}$.

   (a) Show that $\mathcal{G} \leq \mathcal{H}$.

   (b) Are $\mathcal{G}$ and $\mathcal{H}$ are Markov equivalent?

5. In this problem, we will set up and use an R package that allows us to apply the PC algorithm and GES to real data. The package is called `pcalg`, and you should take the following steps to install it:

   - Download R and R Studio if you have not done so already. This can be done by following the download links at `https://www.rstudio.com/`.
   - Open R Studio and type:
     ```
     install.packages("pcalg")
     ```
   - The `pcalg` package requires a number of packages that are only available through `bioconductor`. To download and install these packages, type the following into the R terminal:
     ```
     if (!requireNamespace("BiocManager", quietly = TRUE))
     install.packages("BiocManager")
     BiocManager::install(c("graph", "RBGL", "ggm", "Rgraphviz"))
     ```

- We can then start experimenting with the `pcalg` package! It has some built in data sets we will experiment with. In the data set we will use, we have $m = 8$ continuous variables with Gaussian noise and $n = 5000$ observations (samples). To analyze this data set we first load the `pcalg` package and the data set by typing the following:

```
library("pcalg")
library("Rgraphviz")
data("gmG")
```

The object gmG contains two components, gmG$x, which contains the samples, and gmG$g, which contains the true graph used to generate the data. To see the true graph type the following. The first command sets up the viewing window so that we can display four different figures. The second command plots the true graph:

```
par(mfrow=c(2,2))
plot(gmG$g, main="True Graph")
```

To start, we will use only the first 1000 samples in the data set. To take this subset type:

```
dataSub<-gmG$x[c(1:1000),]
```

To use the PC algorithm to try and learn this graph from the data `dataSub`, we first generate a sufficient statistic from the data that will be used by our conditional independence test:

```
suffStat <- list(C = cor(dataSub), n = nrow(dataSub))
```

We then run the PC algorithm on the sufficient statistic. Aside from the sufficient statistic, we also give a choice for the conditional independence test, the number of variables, and a significance level $\alpha$ for the test. The package pcalg includes several conditional independence tests (Gaussian, discrete and binary) for standard data types. We use the Gaussian test as we know our data is drawn from a multivariate normal model:

```
pc.fit <- pc(suffStat, indepTest = gaussCItest, p=8, alpha = 0.01)
```

To see the graph learned by the PC algorithm we type:

```
plot(pc.fit, main="PC 1000 Samples")
```

How does the result compare with the true graph? (Notice that R represents a undirected edge with a bidirect edge in its plots.) Alternatively, we can use to GES to estimate the causal structure. To do so, we first define the score (BIC) object associated to the data, and then apply the function `ges`. (Note that the character after the capital "L" when defining the object `score` is a zero and not a capital O, whereas the character after "pen" is a capital O.)

```
score <- new("GaussL0penObsScore", dataSub)
ges.fit <- ges(score)
plot(ges.fit$essgraph, main="GES 1000 Samples")
```

How do the two results compare? If they are different, what do you think lead to the differences? How does each compare with the true graph? Now try running the PC algorithm on the full data set `gmG$x`. How does the result compare with the other graphs? How might you explain the similarities/differences?

Now experiment with your own data! (real or simulated!) This intro is drawn from the paper *Causal Inference Using Graphical Models with the R Package pcalg* (2012) by Kalisch et. al. For a more detailed exploration of the package, please see this paper or the R package documentation.