Tutorial Network Analysis

Freie Universität Berlin, SS 2016/17
Martin Vingron · Alena van Bömmel

Assignment 2
Due date: 7.7.2017 9:00AM before the lecture

Include all important steps of your calculations/solutions. Give the important parts of your code or send the complete code to: `alena.vanboemmel@molgen.mpg.de`. Form groups of max. 2 students to solve the problems.

Name(s): Matrikelnr.:

Problem 1 (*70 Points; Mathematic Marks*)**.** In this exercise, you will work with the 'Mathematic Marks' data set that was introduced in the lecture.

(A) Load the data set. If using R, install the package `ggm` and load `data(marks)`. If you want to analyse the data with other software load the data first in R and then save them in your favourite format using `write.table()` or `write.csv()` functions.

(B) Create a scatterplot matrix with all 5 variables V =(`mechanics, vectors, algebra, analysis, statistics`). What do you observe in terms of dependencies?

(C) Calculate the correlation matrix $corr(X)$ (with Pearson's correlation coefficients for all pairs of variables) and plot a heatmap of correlations. Which variables are highly correlated?

Hint: For creating nice colors, use library `RColorBrewer` *and a color palette:*

```
hmcol = colorRampPalette(brewer.pal(9, "RdBu"))(20)[20:1]
```

To plot a heatmap use `heatmap.2` *from library* `gplots`*. Here is an example with some useful options:*

```
heatmap.2(C,dendrogram="none", Rowv=FALSE, Colv=FALSE,
          trace="none", cellnote=round(C,2), notecol="black",
          col=hmcol, breaks=seq(from=-1,to=1,length.out=21), symbreaks=TRUE)
```

(D) Take the diagonal entries of $D = cor(X)^{-1}$, denoted as D_{ii} and calculate the following scores:

$$S = \left(\frac{D_{ii} - 1}{D_{ii}}\right), a \in V$$

Hint: You need to calculate the inverse of a matrix, check this overview[1] *if you need help how to solve it with R.*

(E) Fit 5 different linear models, such that each variable is a linear combination of all other variables (e.g. `mechanics` $\sim$ `vectors+algebra+analysis+statistics`). Look at the R^2 values in the linear models, what do you notice? *Hint: Use function* `lm` *in R.*

[1] http://www.statmethods.net/advstats/matrix.html

(F) Now calculate the covariance matrix Σ and its inverse matrix (precision matrix) $P = \Sigma^{-1}$. Obtain matrix K which is a rescaled version of P such that all diagonal values equal 1 and all off-diagonal entries are calculated as follows:

$$K_{ij} = -\frac{p_{ij}}{\sqrt{p_{ii}p_{jj}}} \quad \forall i \neq j \; .$$

(G) Now fit the following two linear models:

$$\begin{aligned} Y_{mechanics} &\sim X_{algebra} + X_{analysis} + X_{statistics} \\ Y_{vectors} &\sim X_{algebra} + X_{analysis} + X_{statistics} \end{aligned} \tag{1}$$

and calculate the correlation between the *residuals* of these two models.

(H) Repeat the previous step for pairs of response variables $(Y_{mechanics}, Y_{algebra})$, $(Y_{mechanics}, Y_{analysis})$ and $(Y_{mechanics}, Y_{statistics})$ in Eq.(1). The three remaining variables are then the explanatory variables X. Look at the correlation values between all residuals and compare them with the values in matrix K. What do you observe? Do you know the explanation?

(I) Plot a heatmap of values in matrix K and compare it to the heatmap from (C).

Problem 2 (*30 Points; Correlations between Gaussian variables*)**.** Generate $n = 1000$ samples for the following three random variables(first parameter denotes mean μ, second parameter standard deviation σ):

$$\begin{aligned} &X \sim N(0,1) \\ &Y \sim N(2 * X + 1, 0.5) + \varepsilon \quad \text{with } \varepsilon \sim N(0, 0.5) \\ &Z \sim N(5 * X + 1, 1) + \varepsilon \end{aligned}$$

(A) Plot the data in a scatterplot matrix. What do you think, which variables are independent?

(B) Compute correlations between each pair of the variables and plot a heatmap of correlations. Would you change your independent assumptions?

(C) Compute partial correlations between each pair of the variables given the third one based on the regression residuals (e.g. $cor(Y - Y(\hat{X}), Z - Z(\hat{X}))$, see lecture slides)

(D) Compute partial correlation based on the inverse of the covariance matrix and rescaling and compare the result with (C).

(E) Plot a heatmap of partial correlations. Compare it with the heatmap of correlation. What do you observe? Which variables are conditional independent?