Tutorial Complex Systems

Freie Universität Berlin, Summer 2021
Martin Vingron · Persia Akbari Omgba

Assignment 2
Due date: 25.06.2021 10:00 AM before the lecture

You should upload a single PDF report (with all the important steps of your calculations), code files (with comments and clear function naming) and any additional files requested in the assignment on Whiteboard (KVV). Optimally you should generate your PDF report in Latex, we will also accept *legible* photos/scans grouped into a single PDF file. Form groups of 2 students to solve the problems but remember: everyone has to be able to explain the solutions.

Problem 1 (*60 Points; Mathematic Marks*)**.** In this exercise, you will work with the 'Mathematic Marks' data set that was introduced in the lecture. You can use R or Python for this exercise, but Python is recommended.

(A) Load the data set (available on the Whiteboard).

(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 $cor(X)$ (with Pearson's correlation coefficients for all pairs of variables) and plot a heatmap of correlations. Which variables are highly correlated? Use the viridis colormap with dark blue for -1, yellow for +1 and green (middle of the colormap) for 0.

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

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

(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?

(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:

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

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).

(J) Draw a network for the math marks data.

Problem 2 (*25 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. Which variables would you consider independent when looking at the scatterplots?

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

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

(D) Compute partial correlation based on the inverse of the covariance matrix and rescaling (see Problem 1) 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 conditionally independent?

Problem 3 (*15 Points; Gaussian distribution*)**.** Analyze the following two cases of the Gaussian distribution.

(A) Consider a two-dimesnional random variable (X, Y) from a bivariate Gaussian distribution:

$$(X, Y) \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}\right).$$

Draw a random sample from the bivariate Gaussian distribution with the size $n_1 = 10$ and calculate the correlation coefficient of the two vectors. Repeat the step for sample size $n_2 = 100$ and $n_3 = 1000$. What do you observe? What is the relation of the correlation coefficient and the covariance matrix? Take a look at the definition of the correlation coefficient and explain your observation.

(B) Draw a random sample of size $n = 100$ from a univariate Gaussian distribution with mean $\mu = 0$ and $\sigma^2 = 4$. Draw another sample of the same size from a univariate Gaussian distribution with mean $\mu = 1$ and $\sigma^2 = 3$. Calculate the correlation coefficient of these two vectors. Can you use the correlation coefficient to assess the independence of the two random variables? Why? Are the two variables independent? Justify your answer.

Remember to set a seed for your simulations. Report it.