Topic 1: Partial correlations

Welcome to Week 2

Dr Pavel Krivitsky gives you a brief overview of topics and concepts we'll be covering in this week.

Transcript

Weekly learning outcomes

- Calculate and interpret a partial correlation between two variables controlling for one or more variables.
- Perform a hypothesis test for a partial correlation and interpret the conclusion in the context of the problem.
- Explain the relationship between multiple correlation coefficient and R^2 in multiple regression.
- Explain the linear algebra foundations of a principal component analysis.
- Interpret the results of a principal component analysis, relating the principal components to underlying variables.
- Create and interpret a principal component biplot.
- Select the optimal number of principal components according to several techniques.
- Utilise principal components as a data reduction technique.

Topics we will cover are:

- Topic 1: Partial correlations
    - *H&S* does not have a dedicated section for this topic, but there is some discussion in Sections 4.2 (under **Conditional Expectations**) and 5.1 (under **Conditional Approximations**).
    - *J&W* discusses this topic in Section 7.8 (under **Partial Correlation Coefficient**).
    - *M* discusses this topic in Section 5.3.
- Topic 2: Multiple correlation coefficients
    - *H&S* does not have a dedicated section for this topic, but there is some discussion in Sections 4.2 (under **Conditional Expectations**) and 5.1 (under **Conditional Approximations**).
    - *J&W* discusses this topic in Section 7.8.
    - *M* discusses this topic in Section 5.2.
- Topic 3: Testing correlation coefficients
    - *M* discusses this topic in Sections 5.2 and 5.3. (Note that it is more theoretical than is required for this course.)
- Topic 4: Principal component analysis
    - *H&S* discusses this topic in Chapter 11.
    - *J&W* discusses this topic in Chapter 8.
    - *M* discusses this topic in Chapter 9.

H&S: Härdle, Wolfgang K. and Simar, Léopold (2015) *Applied Multivariate Statistical Analysis*. 4th ed. Springer.

J&W: Johnson, Richard A. and Wichern, Dean W. (2007) *Applied Multivariate Statistical Analysis*. 6th ed. Prentice Hall.

M: Muirhead, Robb J. (1982) *Aspects of Multivariate Statistical Theory*. Wiley.

Questions about this week's topics?

This week's topics were prepared by Dr P. Krivitsky. If you have any questions or comments, please post them in the Forum.

Partial correlation

Introduction

To begin Week 2, we will make some general comments on similarities and differences between correlations and dependencies.

Very often we are interested in correlations (dependencies) between a number of random variables and are trying to describe the “strength” of the (mutual) dependencies. For example, we would like to know if there is a correlation (mutual non-directed dependence) between the length of the arm and of the leg. But, if we would like to get information about (or to predict) the length of the arm by measuring the length of the leg, we are dealing with the dependence of the arm’s length on the leg’s length. Both problems described in this example make sense.

On the other hand, there are other examples/situations in which only one of the problems is interesting or makes sense. If we study the dependence between rain and crops, this makes a perfect sense but there is no sense at all to study the (directed) influence of crops on rain.

In a nutshell, we can say that when studying the mutual (linear) dependence, we are dealing with correlation theory whereas when studying directed influence of one (input) variable on another (output) variable, we are dealing with regression theory.

It should be clearly pointed out though that correlation alone, no matter how strong, can not help us identify the direction of influence and can not help us in regression modelling. Our reasoning about direction of influence should come outside of statistical theory, from another theory.

Another important point to always bear in mind is that, as already discussed in The Multivariate Normal Distribution, uncorrelated does not necessarily mean independent if the multivariate data happens to fail the multivariate normality test. Nonetheless, for multivariate normal data, the notions of "uncorrelated" and "independent" coincide.

In general, there are 3 types of correlation coefficients:

- The usual correlation coefficient between 2 variables
- *Partial correlation* coefficient between 2 variables after adjusting for the effect (regression, association) of a set of other variables
- *Multiple correlation* between a single random variable and a set of p other variables.

Partial correlation

For $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$ we defined the correlation coefficient $\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}}, i, j = 1, 2, \ldots, p$ and discussed the MLE $\hat{\rho}_{ij}$ in (1.12). It turned out that they coincide with the sample correlations r_{ij} we introduced in the Exploratory Data Analysis of Multivariate Data slide (formula (1.3)).

To define *partial correlation coefficients*, recall the Property 4 of the multivariate normal distribution from Week 1 (Properties of multivariate normal slide):

If vector $\boldsymbol{X} \in \mathbb{R}^p$ is divided into $X = \begin{pmatrix} X_{(1)} \\ X_{(2)} \end{pmatrix}, X_{(1)} \in R^r, r < p, \boldsymbol{X}_{(2)} \in R^{p-r}$ and according to this subdivision the vector means are $\boldsymbol{u} = \begin{pmatrix} \boldsymbol{\mu}_{(1)} \\ \boldsymbol{\mu}_{(2)} \end{pmatrix}$ and the covariance matrix Σ has been subdivided into $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ and the rank of Σ_{22} is full then the conditional density of $\boldsymbol{X}_{(1)}$ given that $\boldsymbol{X}_{(2)} = \boldsymbol{x}_{(2)}$ is

$$N_r\left(\boldsymbol{\mu}_{(1)} + \Sigma_{12}\Sigma_{22}^{-1}\left(\boldsymbol{x}_{(2)} - \boldsymbol{\mu}_{(2)}\right), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$$

We **define** the partial correlations of $\boldsymbol{X}_{(1)}$ given $\boldsymbol{X}_{(2)} = \boldsymbol{x}_{(2)}$ as the usual correlation coefficients calculated from the elements $\sigma_{ij.(r+1),(r+2)\ldots,p}$ of the matrix $\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, i.e.

$$\rho_{ij.(r+1),(r+2),\ldots,p} = \frac{\sigma_{ij.(r+1),(r+2),\ldots,p}}{\sqrt{\sigma_{ii.(r+1),(r+2),\ldots,p}}\sqrt{\sigma_{jj.(r+1),(r+2),\ldots,p}}} \tag{2.1}$$

We call $\rho_{ij.(r+1),(r+2),\ldots,p}$ the correlation of the ith and jth component when the components $(r+1), (r+2)$, etc. up to the pth (i.e. the last $p-r$ components) have been held fixed. The interpretation is that we are looking for the association (correlation) between the ith and jth component after eliminating the effect that the last $p-r$ components might have had on this association.

To find ML estimates for these, we use the translation invariance property of the MLE to claim that if $\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} \hat{\boldsymbol{\Sigma}}_{11} & \hat{\boldsymbol{\Sigma}}_{12} \\ \hat{\boldsymbol{\Sigma}}_{21} & \hat{\boldsymbol{\Sigma}}_{22} \end{pmatrix}$ is the usual MLE of the covariance matrix then $\hat{\boldsymbol{\Sigma}}_{1|2} = \hat{\boldsymbol{\Sigma}}_{11} - \hat{\boldsymbol{\Sigma}}_{12}\hat{\boldsymbol{\Sigma}}_{22}^{-1}\hat{\boldsymbol{\Sigma}}_{21}$ with elements $\hat{\sigma}_{ij.(r+1),(r+2),\ldots,p}, i, j = 1, 2, \ldots, r$ is the MLE of $\Sigma_{1|2}$ and correspondingly,

$$\hat{\rho}_{ij.(r+1),(r+2),\ldots,p} = \frac{\hat{\sigma}_{ij.(r+1),(r+2),\ldots,p}}{\sqrt{\hat{\sigma}_{ii.(r+1),(r+2),\ldots,p}}\sqrt{\hat{\sigma}_{jj.(r+1),(r+2),\ldots,p}}}, i, j = 1, 2, \ldots, r$$

will be the ML estimators of $\rho_{ij.(r+1),(r+2)\ldots,p}, i, j = 1, 2, \ldots, r$.

Simple formulae

For situations when p is not large, as a partial case of the above general result, simple plug-in formulae are derived that express the partial correlation coefficients by the usual correlation coefficients. We shall discuss such formulae now. The formulae are given below:

1. Partial correlation between first and second variable by adjusting for the effect of the third:

$$\rho_{12.3} = \frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{(1-\rho_{13}^2)(1-\rho_{23}^2)}}$$

2. Partial correlation between first and second variable by adjusting for the effects of third and fourth variable:

$$\rho_{12.3,4} = \frac{\rho_{12.4} - \rho_{13.4}\rho_{23.4}}{\sqrt{(1-\rho_{13.4}^2)(1-\rho_{23.4}^2)}}.$$

For higher dimensional cases computers need to be utilised:

```
R: ggm::pcor, ggm::parcor
```

Demonstration: Partial correlations

Start by watching the following demonstration by Dr Krivitsky.

Transcript

This demonstration can be completed using the provided RStudio or your own RStudio.

To complete this task select the 'Parcor_Example.demo.Rmd' in the 'Files' section of RStudio. Follow the demonstration contained within the RMD file.

If you choose to complete the example in your own RStudio, upload the following file:

Parcor_Example.demo.Rmd

The output of the RMD file is also displayed below:

Another example

Optional viewing: Partial Correlation Practice Problem

Statistics at Nevada State College. (2014). Partial Correlation Practice Problem. Retrieved from: https://youtu.be/8i0h98chSHU.

Challenge: Partial correlations

If you choose to complete this task in your own RStudio, upload the following file:

Parcor_Example.challenge.Rmd

Click on the 'Parcor_Example.challenge.Rmd' in the 'Files' section to begin. Enter your response to the tasks in the 'Enter your code here' section.

This activity and the solution will be discussed at the Collaborate session this week. In the meantime, share and discuss your results in the 'Tutorials' discussion forum.

The solution will also be available here on Friday of this week by clicking on the 'Solution' tab in the top right corner.

The output of the RMD file is also displayed below: