Multivariate Analysis (MATH5855)

Dr. Atefeh Zamani

Based on notes by **Prof. Spiridon Penev** and **Dr. Pavel Krivitsky**
University of New South Wales
School of Mathematics & Statistics
Department of Statistics

Term 3, 2022

Section 1:
Exploratory Data Analysis of Multivariate Data

Data organisation, Basic summaries, Visualisation, Software

Data organisation

Representation

case (a.k.a. item, individual, or experimental trial) $p \geq 1$ variables recorded on each unit of analysis

x_{ij} is the i-th (of p) variable observed on j-th (of n) case

data matrix:

$$X_{p\times n} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{in} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pj} & \cdots & x_{pn} \end{pmatrix} \tag{1}$$

Basic summaries

Univariate summaries

sample mean (of variable i) $\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij}$

sample variance (of variable i) $s_i^2 = \frac{1}{n}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2$

- Sometimes, we will use divisor of $n - 1$ instead.

Bivariate summaries

sample covariance (of variables i and k)

$$s_{ik} = \frac{1}{n}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)(x_{kj} - \bar{x}_k)$$

- Linear association only!
- Symmetric: $s_{ik} = s_{ki}$

sample correlation (of variables i and k)

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} \equiv \frac{s_{ik}}{s_i s_k}$$

- A unitless measure.
- Also symmetric.
- By Cauchy–Bunyakovsky–Schwartz Inequality $|r_{ik}| \leq 1$.
- Also linear; can use quotient correlation instead for nonlinear.

Calculations on matrix data

The descriptive statistics that we discussed until now are usually organised into arrays, namely:
Vector of sample means $\bar{\boldsymbol{x}} = (\bar{x}_1 \ \bar{x}_2 \ \cdots \ \bar{x}_p)^T$
Matrix of sample variances and covariances

$$S_{p\times p} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix} \tag{2}$$

Matrix of sample correlations

$$R_{p\times p} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix} \tag{3}$$

Visualisation

Some simple characteristics of the data are worth studying before the actual multivariate analysis would begin:

- drawing scatterplot of the data;
- calculating simple univariate descriptive statistics for each variable;
- calculating sample correlation and covariance coefficients; and
- linking multiple two-dimensional scatterplots.

Software

- In SAS, the procedures that are used for this purpose are called proc means, proc plot and proc corr. Please study their short description in the included SAS handout.
- In R, these are implemented in base::rowMeans, base::colMeans, stats::cor, graphics::plot, graphics::pairs, GGally::ggpairs. Here, the format is PACKAGE::FUNCTION, and you can learn more by running
  library(PACKAGE)
  ? FUNCTION