MATH5855: Multivariate Analysis

Dr Pavel Krivitsky
based on notes by A/Prof Spiridon Penev
University of New South Wales
School of Mathematics
Department of Statistics

2021 Term 3

Since the notes will be uploaded in parts, these page numbers are indicative.

Foreword

These notes

These notes do **not** substitute the lectures in Multivariate Analysis for Masters students. You are strongly recommended to attend each and every lecture and laboratory hour because the conceptual bases of the discussed modelling methods, as well as some additional derivations and explanations will then be focused on, as will be important portions of pertinent computer output. This volume is therefore not meant to be a substitute for a textbook, computer package manual, or lecture attendance.

We rely on the widely spread and powerful statistical suites R and SAS to perform the actual calculations during the course. These notes are a compilation from several sources and other notes. Some of the sources are listed in your handout. As the closest reference book the following source could be mentioned:

> Johnson, R. & Wichern, D. (2007) *Applied Multivariate Statistical Analysis.* Sixth Edition, Prentice Hall.

By no means can this book be a substitute for the whole set of notes, though.

It is assumed that you are familiar with some basic concepts of linear algebra. These will be summarised at the beginning and will be used essentially in the rest of the course. These concepts include matrix and vector operations, determinants, traces, ranks, projectors, linear equations, inverses, eigenvectors and eigenvalues etc.

I would appreciate it if you would let me know about any ways these notes could be further improved.

Overview

First we shall discuss some general aspects of Multivariate Analysis. Usually, when studying complex phenomena, **many** variables are required. Besides, the process of studying is usually an iterative one with many variables often added or deleted from the study. Multivariate analysis deals with developing methods for better understanding the relationships between the many variables included in the analysis of such complex phenomena.

What makes Multivariate Analysis different?

In your other classes, you have learned about a variety of methods for analysing many variables. For example, you have probably learned about *multiple regression* linear model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i,\ i = 1, \ldots, n$$

where Y_i is the ith observation of the response variable, $x_{i,k}$ ith observation of the kth predictor variable, and ϵ_i the ith error. However, in this regression, we designate the p predictors as fixed (conditioned on) and only *one* variable per observation is random. Typically, we assume that the ϵ_is and therefore Y_is are independent (conditional on the xs) or at least uncorrelated.

Contrast this with a *multivariate* linear model,

$$\begin{aligned} Y_{i1} &= \beta_{01} + \beta_{11} x_{i1} + \beta_{21} x_{i2} + \cdots + \beta_{p1} x_{ip} + \epsilon_{i1}, \\ Y_{i2} &= \beta_{02} + \beta_{12} x_{i1} + \beta_{22} x_{i2} + \cdots + \beta_{p2} x_{ip} + \epsilon_{i2}, \end{aligned}$$

where Y_{i1} and Y_{i2} are the ith observations of two distinct response variables, and ϵ_{i1} and ϵ_{i2} may be correlated. The multivariate linear model can be used when multiple observations are taken

on each individual in the sample, and it can allow us to model the relationships among these measurements.

Difficulties in such a process:

- More data to analyse
- More involved mathematics necessary
- Computer intensive methods involved in the process

Objectives of multivariate methods:

Data reduction: presenting the phenomenon as simply as possible **but** without sacrificing valuable information. Typical *representative method:* Principal components analysis. Sometimes, this reduction is achieved by introducing a small number of unobservable (latent) variables when trying to explain a large number of observable output variables. Representative methods: *factor analysis* and *covariance structure analysis.*

Sorting or grouping: creating groups of "similar" objects or variables that in a sense are more closer to each other than to objects outside the group; and finding reasonable explanation for the existing grouping. *Representative methods:* Factor Analysis, Cluster Analysis, Discriminant Analysis.

Investigation of dependence among variables: finding which sets of variables can be considered as independent and which are "more dependent"; and "measuring" the dependence.*Representative Methods:* Correlation Analysis, Partial Correlations, Canonical Correlations.

Prediction: predicting values of one or more variables on the basis of observations of other variables that have been found to influence the former variables: a basic but important goal. *Representative:* Multivariate Regression.

Hypothesis testing: either validating assumptions (e.g., normality) on the basis of which certain analysis is being done or to reinforce some prior modelling convictions (e.g., equality of parameters).Hypothesis testing is relevant to the applications of all multivariate methods we will be dealing with.

As a basic **mathematical model** for our analyses in this course the **multivariate normal distribution** will be used. Reasons are: our limited time and the complexity of other approaches. Although in practice also other distributions are relevant, modelling based on the multivariate normal distribution can still be a very good approximation.

0 Preliminaries

0.1 Matrix algebra

0.1.1 Vectors and matrices

As a shorthand notation, we shall be using $X \in \mathcal{M}_{p,n}$ to indicate that X is a matrix with p rows and n columns. A notation $\boldsymbol{x} \in \mathbb{R}^n$ will be used to indicate that $\boldsymbol{x}$ is a n-dimensional *column* vector. Of course, if $\boldsymbol{x} \in \mathbb{R}^n$, it also means that $\boldsymbol{x} \in \mathcal{M}_{n,1}$. *Transposition* will be denoted by $^\top$. After a transposition, from a matrix $X \in \mathcal{M}_{p,n}$ we get a new matrix $X^\top \in \mathcal{M}_{n,p}$. In particular, from a *column* vector $\boldsymbol{x} \in \mathbb{R}^n$ we arrive, after a transposition, to a *row* a vector $\boldsymbol{x}^\top \in \mathcal{M}_{1,n}$. It is well known that multiplication of a matrix (vector) with a scalar means multiplication of each of the elements of the matrix (vector) with that scalar. Also, two matrices (vectors) of the same dimension can be added (subtracted) and the result is a new matrix (vector) of the same dimension and elements which are the element wise sum (difference) of the elements of the matrices (vectors) to be added (subtracted). The *Euclidean norm* of a vector $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \in \mathbb{R}^p$ is denoted by $\|\boldsymbol{x}\|$ and is defined as $\|\boldsymbol{x}\| = \sqrt{\sum_{i=1}^p x_i^2}$.

The *inner product* or, equivalently, the *scalar product* of two p-dimensional vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is denoted and defined in the following way:

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^\top \boldsymbol{y} = \sum_{i=1}^p x_i y_i \tag{0.1}$$

Obviously, the relation $\|\boldsymbol{x}\|^2 = \langle \boldsymbol{x}, \boldsymbol{x} \rangle$ holds. It is well known that if θ is the angle between two p-dimensional vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ then it also holds

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \|\boldsymbol{x}\| \|\boldsymbol{y}\| \cos(\theta) \tag{0.2}$$

Since $|\cos(\theta)| \leq 1$, we have the inequality

$$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \leq \|\boldsymbol{x}\| \|\boldsymbol{y}\|$$

which is one variant of the *Cauchy–Bunyakovsky–Schwartz* Inequality. Further, if we want to *orthogonally project* the vector $\boldsymbol{x} \in \mathbb{R}^p$ on the vector $\boldsymbol{y} \in \mathbb{R}^p$ then (having in mind the geometric interpretation of orthogonal projection) the result will be: $\frac{\boldsymbol{x}^\top \boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{y}} \boldsymbol{y}$.

Finally, the rules for matrix multiplication are recalled: if $X \in \mathcal{M}_{p,k}$ and $Y \in \mathcal{M}_{k,n}$ (i.e. the number of columns in X is equal to the number of rows in Y) then the multiplication XY is possible and the result is a matrix $Z = XY \in \mathcal{M}_{p,n}$ with elements

$$z_{i,j},\ i = 1, 2, \ldots, p,\ j = 1, 2, \ldots, n :\ z_{i,j} = \sum_{m=1}^{k} x_{i,m} y_{m,j} \tag{0.3}$$

i.e. the element in the ith row and jth column of Z is a scalar product of the ith row of X and the jth column of Y. Note that the multiplication of matrices is **not commutative** and in general, it is not necessary for YX to even exist when XY exists. When the matrices are both square (quadratic) of the same dimension p (i.e. both $X \in \mathcal{M}_{p,p}$ and $Y \in \mathcal{M}_{p,p}$) then both XY and YX will be defined but would in general **not** give rise to the same result.

The following transposition rule is important to be mentioned (and easy to check): if $X \in \mathcal{M}_{p,k}$ and $Y \in \mathcal{M}_{k,n}$ then the product XY exists and it holds:

$$(XY)^\top = Y^\top X^\top \tag{0.4}$$

One should be very careful with transposition though in order to avoid silly mistakes. If $\boldsymbol{x} \in \mathbb{R}^p$, for example, both $\boldsymbol{x}^\top \boldsymbol{x}$ and $\boldsymbol{x}\boldsymbol{x}^\top$ exist. While the former is a scalar, the latter belongs to $\mathcal{M}_{p,p}$!

A square matrix $X \in \mathcal{M}_{p,p}$ is called *symmetric* if $x_{i,j} = x_{j,i}$ for $i = 1, 2, \ldots, p$ and $j = 1, 2, \ldots, p$ holds. For such a matrix, we have $X^\top = X$.

The square matrix $\overset{p \times p}{\boldsymbol{I}} = \delta_{ij}$ for $i = 1, 2, \ldots, p$ and $j = 1, 2, \ldots, p$ holds (i.e., ones on the diagonal and zeros outside the diagonal) is called the *identity matrix* (of dimension p). Obviously, when the multiplication is possible then always $X\boldsymbol{I} = X$ and $\boldsymbol{I}X = X$ holds.

The trace of a square matrix $X \in \mathcal{M}_{p,p}$ is denoted by $\mathrm{tr}(X) = \sum_{i=1}^{p} x_{ii}$. The following properties of traces are easy to obtain:

i) $\mathrm{tr}(X + Y) = \mathrm{tr}(X) + \mathrm{tr}(Y)$

ii) $\mathrm{tr}(XY) = \mathrm{tr}(YX)$

iii) $\mathrm{tr}(X^{-1}YX) = \mathrm{tr}(Y)$

iv) If $\boldsymbol{a} \in \mathbb{R}^p$ and $X \in \mathcal{M}_{p,p}$ then $\boldsymbol{a}^\top X \boldsymbol{a} = \mathrm{tr}(X\boldsymbol{a}\boldsymbol{a}^\top)$

0.1.2 Inverse matrices

To any **square** matrix $X \in \mathcal{M}_{p,p}$ one can attach a number $|X| \equiv \det(X)$ called a *determinant* of the matrix. It is defined as

$$|X| = \sum \pm x_{1i} x_{2j} \ldots x_{pm}$$

where the summation is over **all** permutations $(i, j, \ldots, m)$ of the numbers $(1, 2, \ldots, p)$ by taking into account the **sign rule**: summands with an even permutation get a $(+)$ whereas the ones with an odd permutation get a $(-)$ sign.

It can be seen that this is equivalent to another recursive definition, namely:

- when $p = 1$ (scalar case) $X = a$ is just a number and $|X| = a$ in this case

- when $p = 2$ then $\begin{vmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{vmatrix} = x_{11}x_{22} - x_{12}x_{21}$
- when $p = 3$ then the following rule applies:

$$\begin{vmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{vmatrix} = x_{11}x_{22}x_{33} + x_{12}x_{23}x_{31} + x_{21}x_{32}x_{13} - x_{31}x_{22}x_{13} - x_{11}x_{23}x_{32} - x_{12}x_{21}x_{33} \tag{0.5}$$

- recursively, for $X \in \mathcal{M}_{(p,p)}$,

$$|X| = \sum_i (-1)^{i+j} x_{ij} |X_{ij}| = \sum_j (-1)^{i+j} x_{ij} |X_{ij}|$$

  where X_{ij} denotes the matrix we get by deleting the ith row and jth column of X, and $|X_{ij}|$ is therefore the (i, j)th *minor* of X.

Here we list some elementary properties of determinants that follow directly from the definition:

i) If one row or one column of the matrix contains zeros only, then the value of the determinant is zero.

ii) $|X^\top| = |X|$

iii) If one row (or one column) of the matrix is modified by multiplying with a scalar c then so is the value of the determinant.

iv) $|cX| = c^p |X|$

v) If $X, Y \in \mathcal{M}_{p,p}$ then $|XY| = |X||Y|$

vi) If the matrix X is *diagonal* (i.e. all non-diagonal elements are zero) then $|X| = \prod_{i=1}^p x_{ii}$. In particular, *the determinant of the identity matrix is always equal to one.*

Given that $|X| \neq 0$ (or equivalently, if the matrix $X \in \mathcal{M}_{p,p}$ is *nonsingular* then an **inverse** matrix $X^{-1} \in \mathcal{M}_{p,p}$ can be defined that has to satisfy $XX^{-1} = \boldsymbol{I}_{p,p}$. It is easy to check that the inverse X^{-1} has as its (j, i)th entry $\frac{|X_{ij}|}{|X|}(-1)^{i+j}$, where $|X_{ij}|$ is, as before, the (i, j)th *minor* of X.

Some elementary properties of inverses follow:

i) $XX^{-1} = X^{-1}X = \boldsymbol{I}$

ii) $(X^{-1})^\top = (X^\top)^{-1}$

iii) $(XY)^{-1} = Y^{-1}X^{-1}$ when both X and Y are nonsingular square matrices of the same dimension.

iv) $|X^{-1}| = |X|^{-1}$

v) If X is diagonal and nonsingular then all its diagonal elements are nonzero and X^{-1} is again diagonal with diagonal elements equal to $\frac{1}{x_{ii}}, i = 1, 2, \ldots, p$.

0.1.3 Rank

A set of vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k \in \mathbb{R}^n$ is *linearly dependent* if there exist k numbers $a_1, a_2, \ldots, a_k$ **not all zero** such that

$$a_1\boldsymbol{x}_1 + a_2\boldsymbol{x}_2 + \cdots + a_k\boldsymbol{x}_k = \boldsymbol{0} \tag{0.6}$$

holds. Otherwise the vectors are *linearly independent*. In particular, for k linearly independent vectors the equality (0.6) would only be possible if **all** numbers $a_1, a_2, \ldots, a_k$ were zero.

The *row rank* of a matrix is the maximum number of linearly independent row vectors. The *column rank* is the rank of its set of column vectors. It turns out that the row rank and the column rank of a matrix are always equal. Thus the rank of a matrix X (denoted $\mathrm{rk}(X)$) is either the row or the column rank. If $X \in \mathcal{M}_{p,n}$ and $\mathrm{rk}(X) = \min(p, n)$ we say that the matrix is of full rank. In particular, a square matrix $A \in \mathcal{M}_{p,p}$ is of full rank if $\mathrm{rk}(A) = p$. As is well known from the basic theorem of linear algebra *Kronecker–Capelli* or *Rouché–Capelli Theorem* this means also that $|A| \neq 0$ when A is of full rank. Then the inverse of A will also exist. Let $\boldsymbol{b} \in \mathbb{R}^p$ be a given vector. Then the linear equation system $A\boldsymbol{x} = \boldsymbol{b}$ has a unique solution $\boldsymbol{x} = A^{-1}\boldsymbol{b} \in \mathbb{R}^p$.

0.1.4 Orthogonal matrices

A square matrix $X \in \mathcal{M}_{p,p}$ is *orthogonal* if $XX^\top = X^\top X = \boldsymbol{I}_{p,p}$ holds. The following properties of orthogonal matrices are obvious:

i) X is of full rank ($\mathrm{rk}(X) = p$) and $X^{-1} = X^\top$

ii) The name *orthogonal* of the matrix originates from the fact that the scalar product of each two different column vectors equals zero. The same holds for the scalar product of each two different row vectors of the matrix. The norm of each column vector (or each row vector) is equal to one. These properties are equivalent to the definition.

iii) $|X| = \pm 1$

0.1.5 Eigenvalues and eigenvectors

For **any** square matrix $X \in \mathcal{M}_{p,p}$ we can define the *characteristic polynomial* equation of degree p,

$$f(\lambda) = |X - \lambda\boldsymbol{I}| = 0. \tag{0.7}$$

Equation (0.7) is a polynomial equation of power p so it has exactly p roots. In general, some of them may be complex and some may coincide. Since the coefficients are real, if there is a complex root of 0.7 then also its complex conjugate must be a root of the same equation. Denote **any** such eigenvalue by λ^*. In addition, $\mathrm{tr}(X) = \sum_{i=1}^p \lambda_i$ and $|X| = \prod_{i=1}^p \lambda_i$.

Obviously, the matrix $X - \lambda^*\boldsymbol{I}$ is singular (its determinant is zero). Then, according to the Kronecker theorem, there exists a non-zero vector $\boldsymbol{y} \in \mathbb{R}^p$ such that $(X - \lambda^*\boldsymbol{I})\boldsymbol{y} = \boldsymbol{0}, \boldsymbol{0} \in \mathbb{R}^p$. We call $\boldsymbol{y}$ an *eigenvector* of X that corresponds to the eigenvalue λ^*. Note that the eigenvector is not uniquely defined: $\mu\boldsymbol{y}$ for any real non-zero μ would also be an eigenvector corresponding to the same eigenvalue.

Sparing some details of the derivation, we shall formulate the following basic result:

Theorem 0.1. *When the matrix X is real symmetric then **all** of its p eigenvalues are **real**. If the eigenvalues are all different then all the p eigenvectors that correspond to them, are orthogonal (and hence form a basis in $\mathbb{R}^p$). These eigenvectors are also unique (up to the norming constant μ above). If some of the eigenvalues coincide then the eigenvectors corresponding to them are not necessarily unique but even in this case they can be chosen to be mutually orthogonal.*

For each of the p eigenvalues λ_i, $i = 1, 2, \ldots, p$, of X, denote its corresponding set of mutually orthogonal eigenvectors of *unit length* by $\boldsymbol{e}_i$, $i = 1, 2, \ldots, p$, i.e.

$$X\boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i,\ i = 1, 2, \ldots, p,\ \|\boldsymbol{e}_i\| = 1,\ \boldsymbol{e}_i^\top \boldsymbol{e}_j = 0,\ i \neq j$$

holds. Then is can be shown that the following decomposition *(spectral decomposition)* of any symmetric matrix $X \in \mathcal{M}_{p,p}$ holds:

$$X = \lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1^\top + \lambda_2 \boldsymbol{e}_2 \boldsymbol{e}_2^\top + \ldots \lambda_p \boldsymbol{e}_p \boldsymbol{e}_p^\top. \tag{0.8}$$

Equivalently, $X = P\Lambda P^\top$ where $\Lambda = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p \end{pmatrix}$ is diagonal and $P \in \mathcal{M}_{p,p}$ is an *orthogonal matrix* containing the p orthogonal eigenvectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_p$.

The above decomposition is a very important analytical tool. One of its most widely used applications is for defining a square root of a symmetric positive definite matrix.

A symmetric matrix $X \in \mathcal{M}_{p,p}$ is *positive definite* if all of its eigenvalues are positive. (It is called *non-negative definite* if all eigenvalues are ≥ 0.) For a symmetric positive definite matrix we have all λ_i, $i = 1, 2, \ldots, p$, to be positive in the spectral decomposition (0.8).

But then

$$X^{-1} = (P^\top)^{-1}\Lambda^{-1}P^{-1} = P\Lambda^{-1}P^\top = \sum_{i=1}^{p} \frac{1}{\lambda_i} \boldsymbol{e}_i \boldsymbol{e}_i^\top \tag{0.9}$$

(i.e. inverting X is very easy if the spectral decomposition of X is known).

Moreover we can define the *square root* of the symmetric non-negative definite matrix X in a natural way:

$$X^{\frac{1}{2}} = \sum_{i=1}^{p} \sqrt{\lambda_i} \boldsymbol{e}_i \boldsymbol{e}_i^\top \tag{0.10}$$

The definition (0.10) makes sense since $X^{\frac{1}{2}} X^{\frac{1}{2}} = X$ holds. Note that $X^{\frac{1}{2}}$ is also symmetric and non-negative definite. Also $X^{-\frac{1}{2}} = \sum_{i=1}^{p} \lambda_i^{-\frac{1}{2}} \boldsymbol{e}_i \boldsymbol{e}_i^\top = P\Lambda^{-\frac{1}{2}}P^\top$ can be defined where $\Lambda^{-\frac{1}{2}}$ is a diagonal matrix with $\lambda_i^{-1/2}, i = 1, 2, \ldots, p$ being its diagonal elements. These facts will be used essentially in the subsequent sections.

As an illustration of the usefulness of the spectral decomposition approach we shall show the following statement:

Example 0.2. Let $X \in \mathcal{M}_{p,p}$ be symmetric *positive definite* matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ and associated eigenvectors of unit length $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots \boldsymbol{e}_p$. Show that

- $\max_{\boldsymbol{y} \neq \boldsymbol{0}} \frac{\boldsymbol{y}^\top X \boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{y}} = \lambda_1$ attained when $\boldsymbol{y} = \boldsymbol{e}_1$.
- $\min_{\boldsymbol{y} \neq \boldsymbol{0}} \frac{\boldsymbol{y}^\top X \boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{y}} = \lambda_p$ attained when $\boldsymbol{y} = \boldsymbol{e}_p$.

Let $X = P\Lambda P^\top$ be the decomposition (0.8) for X. Denote $\boldsymbol{z} = P^\top \boldsymbol{y}$. Note that $\boldsymbol{y} \neq 0$ implies $\boldsymbol{z} \neq 0$. Thus

$$\frac{\boldsymbol{y}^\top X \boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{y}} = \frac{\boldsymbol{y}^\top P \Lambda P^\top \boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{y}} = \frac{\boldsymbol{z}^\top \Lambda \boldsymbol{z}}{\boldsymbol{z}^\top \boldsymbol{z}} = \frac{\sum_{i=1}^{p} \lambda_i z_i^2}{\sum_{i=1}^{p} z_i^2} \leq \lambda_1 \frac{\sum_{i=1}^{p} z_i^2}{\sum_{i=1}^{p} z_i^2} = \lambda_1$$

If we take $\boldsymbol{y} = \boldsymbol{e}_1$ then having in mind the structure of the matrix P we have $\boldsymbol{z} = P^\top \boldsymbol{e}_1 = (1 \quad 0 \quad \cdots \quad 0)^\top$ and for this choice of $\boldsymbol{y}$ also $\frac{\boldsymbol{z}^\top \Lambda \boldsymbol{z}}{\boldsymbol{z}^\top \boldsymbol{z}} = \frac{\lambda_1}{1} = \lambda_1$. The first part of the exercise is

shown. Similar arguments (just changing the sign of the inequality) apply to show the second part.

In addition, you can try to show that $\max_{\boldsymbol{y}\neq 0, \boldsymbol{y}\perp \boldsymbol{e}_1} \frac{\boldsymbol{y}^\top X \boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{y}} = \lambda_2$ holds. How?

0.1.6 Cholesky Decomposition

Computers perform arithmetic to a finite precision, typically around 16 decimal significant figures. Furthermore, the numbers are expressed internally in scientific notation, and so the absolute magnitude of the number typically has little effect on precision, but certain operations on numbers with very different magnitudes can sometimes produce severe rounding errors. For example, to a computer $1 \times 10^{18} + 1 \times 10^0 = 1{,}000{,}000{,}000{,}000{,}000{,}000 + 1 = 1{,}000{,}000{,}000{,}000{,}000{,}000$: the 1 gets lost to a rounding error.

When it comes to matrix inversion in particular, the key number is the *condition number*, $|\lambda_1/\lambda_p|$ of a positive definite matrix X, where λ_1 is the largest eigenvalue of X and λ_p is the smallest. (The definition for non-positive-definite matrices can be different.) The higher this number is, the less numerically stable the inversion is likely to be. (Notice that if the matrix is singular, this number is infinite.)

We generally try to avoid asking the computer to invert matrices in ways that lose precision.

An alternative, more numerically stable definition of a "matrix square root" is the *Cholesky decomposition*. For a symmetric positive definite matrix $X \in \mathcal{M}_{p,p}$, there exists a unique upper-triangular matrix $U \in \mathcal{M}_{p,p}$ such that $U^\top U = X$ holds. Note that many sources use a lower-triangular matrix L such that $LL^\top = X$ instead. It is easy to see that $L \equiv U^\top$, and which definition is used is arbitrary, provided it is used consistently, since $UU^\top \neq X$ and neither do $L^\top L$. For example, the Wikipedia article uses L, whereas the R builtin function is `chol()` and SAS/IML's `root(x)` both return U. This decomposition is particularly useful for generating correlated variables.

0.1.7 Orthogonal Projection

Orthogonal projection of any vector $\boldsymbol{y} \in \mathbb{R}^n$ on the space $\mathcal{L}(X)$ spanned by the columns of the matrix $X \in \mathcal{M}_{n,p}$ is a linear operation. Hence the result is a vector $\boldsymbol{z} \in \mathbb{R}^n$ that has the representation $\boldsymbol{z} = P\boldsymbol{y}$ where the matrix $P \in \mathcal{M}_{n,n}$ is called (orthogonal) **projector**. Since $\boldsymbol{z} \in \mathcal{L}(X)$ (being a projection in this space), the projection of $\boldsymbol{z}$ on $\mathcal{L}(X)$ is $\boldsymbol{z}$ itself. Hence $P\boldsymbol{y} = \boldsymbol{z} = P\boldsymbol{z} = PP\boldsymbol{y} = P^2\boldsymbol{y}$ or $(P - P^2)\boldsymbol{y} = \boldsymbol{0} \to P^2 = P$ (since $\boldsymbol{y} \in \mathbb{R}^n$ is arbitrary). Therefore, P should be *idempotent*. Further $(\boldsymbol{y} - \boldsymbol{z})^\top \boldsymbol{z} = \boldsymbol{0}$ or $\boldsymbol{y}^\top (P^\top - \boldsymbol{I})P\boldsymbol{y} = \boldsymbol{0}$ for all $\boldsymbol{y} \to (P^\top - \boldsymbol{I})P = 0$ or $P^\top P = P$. Taking transposes, $P^\top P = P^\top$ or $P = P^\top$ that is, P is symmetrical. So, the orthogonal projector is a symmetric and idempotent matrix.

Vice versa, consider a symmetric and idempotent matrix P. Then if we take any $\boldsymbol{y} \in \mathbb{R}^n$ then for $\boldsymbol{z} = P\boldsymbol{y} \to P\boldsymbol{z} = P^2\boldsymbol{y} = P\boldsymbol{y} \to P(\boldsymbol{y} - \boldsymbol{z}) = \boldsymbol{0}$ (and also $P^\top(\boldsymbol{y} - \boldsymbol{z}) = \boldsymbol{0}$ since $P = P^\top$). Consider $\mathcal{L}(P)$ (the space generated by the rows/columns of P). Now: $\boldsymbol{z} = P\boldsymbol{y} \to \boldsymbol{z} \in \mathcal{L}(P)$ and $P^\top(\boldsymbol{y} - \boldsymbol{z}) = \boldsymbol{0}$ means that $\boldsymbol{y} - \boldsymbol{z}$ is perpendicular to $\mathcal{L}(P)$. Hence $P\boldsymbol{y}$ is the projection of $\boldsymbol{y}$ on $\mathcal{L}(P)$.

Hence, we have seen that $P \in \mathcal{M}_{n,n}$ is an orthogonal projection matrix if and only if it is a symmetric and idempotent matrix.

Also, if P is an orthogonal projection on a given linear space $\mathcal{M}$ of dimension $\dim(\mathcal{M})$ then $\boldsymbol{I} - P$ an orthogonal projection on the orthocomplement of $\mathcal{M}$. It holds $\mathrm{rk}(P) = \dim(\mathcal{M})$.

Further, it can be seen that the rank of an orthogonal projector is equal to the sum of its diagonal elements.

Finally, it can be shown that if the matrix X above has a full rank then the projector $P_{\mathcal{L}(X)} = X(X^\top X)^{-1}X^\top$. If the matrix X is not of full rank then the *generalised inverse*

$(X^\top X)^-$ of $X^\top X$ can be defined instead. Note that the generalised inverse may not be uniquely defined but no matter which version of it has been chosen, the matrix $X(X^\top X)^- X^\top$ is uniquely defined and is the orthogonal projector on the space $\mathcal{L}(X)$ spanned by the columns of X also in cases when the rank of X is not full.

0.2 Standard facts about multivariate distributions

0.2.1 Random samples in multivariate analysis

In order to study the sampling variability of statistics, with the ultimate goal of making inferences, one needs to make some assumptions about the random variables whose values constitute the data set $X \in \mathcal{M}_{p,n}$ in (1.1). Suppose the data has not been observed yet but we *intend* to collect n sets of measurements on p variables. Since the actual observations can not be predicted before the measurements are made, we treat them as random variables. Each set of p measurements can be considered as a realisation of p-dimensional *random vector* and we have n independent realisations of such random vectors $\boldsymbol{X}_i, i = 1, 2, \ldots, n$, so we have the *random matrix* $\boldsymbol{X} \in \mathcal{M}_{p,n}$:

$$\boldsymbol{X} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1j} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2j} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{i1} & X_{i2} & \cdots & X_{ij} & \cdots & X_{in} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{p1} & X_{p2} & \cdots & X_{pj} & \cdots & X_{pn} \end{pmatrix} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n] \tag{0.11}$$

The vectors $\boldsymbol{X}_i, i = 1, 2, \ldots, n$ are considered as independent observations of a p-dimensional random vector. We start discussing the distribution of such a vector.

0.2.2 Joint, marginal, conditional distributions

A random vector $\boldsymbol{X} = (X_1 \quad X_2 \quad \cdots \quad X_p)^\top \in \mathbb{R}^p$, $p \geq 2$ has a *joint cdf*

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_p \leq x_p) = F_{\boldsymbol{X}}(x_1, x_2, \ldots, x_p).$$

In case of a *discrete* vector of observations $\boldsymbol{X}$ the *probability mass function* is defined as

$$P_{\boldsymbol{X}}(\boldsymbol{x}) = P(X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p).$$

If a *density* $f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{\boldsymbol{X}}(x_1, x_2, \ldots, x_p)$ exists such that

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_p} f_{\boldsymbol{X}}(\boldsymbol{t}) dt_1 \ldots dt_p \tag{0.12}$$

then $\boldsymbol{X}$ is a *continuous* random vector with a joint density function of p arguments $f_{\boldsymbol{X}}(\boldsymbol{x})$. From (0.12) we see that in this case $f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{\partial^p F_{\boldsymbol{X}}(\boldsymbol{x})}{\partial x_1 \partial x_2 .. \partial x_p}$ holds.
The *marginal cdf of the first* $k < p$ components of the vector $\boldsymbol{X}$ is defined in a natural way as follows:

$$\begin{aligned} P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_k \leq x_k) &= P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_k \leq x_k, X_{k+1} \leq \infty, ..., X_p \leq \infty) \\ &= F_{\boldsymbol{X}}(x_1, x_2, \ldots, x_k, \infty, \infty, \ldots, \infty) \end{aligned} \tag{0.13}$$

The *marginal density* of the first k components can be obtained by partial differentiation in (0.13) and we arrive at

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\boldsymbol{X}}(x_1, x_2, \ldots, x_p) dx_{k+1} \ldots dx_p$$

For **any** other subset of $k < p$ components of the vector $\boldsymbol{X}$, their marginal cdf and density can be obtained along the same lines.

In particular, each component X_i has marginal cdf $F_{X_i}(x_i), i = 1, 2, \ldots, p$.
The *conditional density* $\boldsymbol{X}$ **when** $X_{r+1} = x_{r+1}, \ldots, X_p = x_p$ is defined by

$$f_{(X_1,\ldots,X_r|X_{r+1},\ldots,X_p)}(x_1, \ldots, x_r|x_{r+1}, \ldots, x_p) = \frac{f_{\boldsymbol{X}}(\boldsymbol{x})}{f_{X_{r+1},\ldots,X_p}(x_{r+1}, \ldots, x_p)} \tag{0.14}$$

The above conditional density is interpreted as the joint density of $X_1, \ldots, X_r$ when $X_{r+1} = x_{r+1}, \ldots, X_p = x_p$ and is only defined when $f_{X_{r+1},\ldots,X_p}(x_{r+1}, \ldots, x_p) \neq 0$.
In case $\boldsymbol{X}$ has p independent components then

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = F_{X_1}(x_1) F_{X_2}(x_2) \cdots F_{X_p}(x_p) \tag{0.15}$$

holds and, equivalently, also

$$P_{\boldsymbol{X}}(\boldsymbol{x}) = P_{X_1}(x_1) P_{X_2}(x_2) \cdots P_{X_p}(x_p), \;\; f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_p}(x_p) \tag{0.16}$$

holds. We note that in case of mutual independence the p components, all conditional distributions do **not** depend on the conditions and the factorisations

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{p} F_{X_i}(x_i), \;\; f_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{p} f_{X_i}(x_i)$$

hold.

0.2.3 Moments

Given the density $f_{\boldsymbol{X}}(\boldsymbol{x})$ of the random vector $\boldsymbol{X}$ the *joint moments of order* $s_1, s_2 \ldots, s_p$ are defined, in analogy to the univariate case, as

$$\mathrm{E}(X_1^{s_1} \cdots X_p^{s_p}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_1^{s_1} \cdots x_p^{s_p} f_{\boldsymbol{X}}(x_1, \ldots, x_p) dx_1 \ldots dx_p \tag{0.17}$$

Note that if some of the s_i in (0.17) are equal to zero then in effect we are calculating the joint moment of a subset of the p random variables.

Now, let $\boldsymbol{X} \in \mathbb{R}^p$ and $\boldsymbol{Y} \in \mathbb{R}^q$ with densities as above. The following moments are commonly used:

Expectation:

$$\boldsymbol{\mu}_{\boldsymbol{X}} = \mathrm{E}(\boldsymbol{X}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \boldsymbol{x} f_{\boldsymbol{X}}(x_1, \ldots, x_p) dx_1 \ldots dx_p \in \mathbb{R}^p.$$

Variance–covariance matrix: (a.k.a. variance or covariance matrix)

$$\begin{aligned}\Sigma_{\boldsymbol{X}} = \mathrm{Var}(\boldsymbol{X}) = \mathrm{Cov}(\boldsymbol{X}) &= \mathrm{E}(\boldsymbol{X}-\boldsymbol{\mu}_{\boldsymbol{X}})(\boldsymbol{X}-\boldsymbol{\mu}_{\boldsymbol{X}})^\top \\ &= \mathrm{E}\,\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{\mu}_{\boldsymbol{X}}\boldsymbol{\mu}_{\boldsymbol{X}}^\top = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix} \in \mathcal{M}_{p,p}.\end{aligned}$$

Covariance matrix:

$$\begin{aligned}\Sigma_{\boldsymbol{X},\boldsymbol{Y}} = \mathrm{Cov}(\boldsymbol{X},\boldsymbol{Y}) &= \mathrm{E}(\boldsymbol{X}-\boldsymbol{\mu}_{\boldsymbol{X}})(\boldsymbol{Y}-\boldsymbol{\mu}_{\boldsymbol{Y}})^\top \\ &= \mathrm{E}\,\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{\mu}_{\boldsymbol{X}}\boldsymbol{\mu}_{\boldsymbol{Y}}^\top = \begin{pmatrix} \sigma_{X_1Y_1} & \sigma_{X_1Y_2} & \cdots & \sigma_{X_1Y_q} \\ \sigma_{X_2Y_1} & \sigma_{X_2Y_2} & \cdots & \sigma_{X_2Y_q} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_pY_1} & \sigma_{X_pY_2} & \cdots & \sigma_{X_pY_q} \end{pmatrix} \in \mathcal{M}_{p,q}.\end{aligned}$$

Let $A \in \mathcal{M}_{p',p}$ and $B \in \mathcal{M}_{q',q}$ fixed and known. Then,

- $\boldsymbol{\mu}_{A\boldsymbol{X}} = A\boldsymbol{\mu}_{\boldsymbol{X}} \in \mathbb{R}^{p'}$
- $\Sigma_{A\boldsymbol{X}} = A\Sigma_{\boldsymbol{X}}A^\top \in \mathcal{M}_{p',p'}$
- $\Sigma_{A\boldsymbol{X},B\boldsymbol{Y}} = A\Sigma_{\boldsymbol{X},\boldsymbol{Y}}B^\top \in \mathcal{M}_{p',q'}$

As a corollary, if $\boldsymbol{X}'$, $\boldsymbol{Y}'$, A' and B' are variables and matrices with the same dimensions as originals (but possibly distributions and values),

- $\mathrm{E}(A\boldsymbol{X} + A'\boldsymbol{X}') = A\boldsymbol{\mu}_{\boldsymbol{X}} + A'\boldsymbol{\mu}_{\boldsymbol{X}'}$
- $\mathrm{Var}(A\boldsymbol{X} + A'\boldsymbol{X}') = A\Sigma_{\boldsymbol{X}}A^\top + A\Sigma_{\boldsymbol{X},\boldsymbol{X}'}(A')^\top + A'\Sigma_{\boldsymbol{X}',\boldsymbol{X}}A^\top + A'\Sigma_{\boldsymbol{X}'}(A')^\top$
- $\mathrm{Cov}(A\boldsymbol{X}+A'\boldsymbol{X}', B\boldsymbol{Y}+B'\boldsymbol{Y}') = A\Sigma_{\boldsymbol{X},\boldsymbol{Y}}B^\top + A\Sigma_{\boldsymbol{X},\boldsymbol{Y}'}(B')^\top + A'\Sigma_{\boldsymbol{X}',\boldsymbol{Y}}B^\top + A'\Sigma_{\boldsymbol{X}',\boldsymbol{Y}'}(B')^\top$

These identities are also useful when $p = p' = q = q' = 1$ (i.e., scalars).

0.2.4 Density transformation formula

Assume, the p existing random variables $X_1, X_2, \ldots, X_p$ with given density $f_{\boldsymbol{X}}(\boldsymbol{x})$ have been transformed by a smooth (i.e. differentiable) one-to-one transformation into p new random variables $Y_1, Y_2 \ldots, Y_p$, i.e. a new random vector $\boldsymbol{Y} \in \mathbb{R}^p$ has been created by calculating

$$Y_i = y_i(X_1, X_2 \ldots, X_p), i = 1, 2, \ldots, p \tag{0.18}$$

The question is how to calculate the density $g_{\boldsymbol{Y}}(\boldsymbol{y})$ of $\boldsymbol{Y}$ by knowing the transformation functions $y_i(X_1, X_2 \ldots, X_p), i = 1, 2, \ldots, p$ and the density $f_{\boldsymbol{X}}(\boldsymbol{x})$ of the original random vector. Naturally, since the transformation (0.18) is assumed to be one-to-one, its inverse transformation $X_i = x_i(Y_1, Y_2 \ldots, Y_p), i = 1, 2, \ldots, p$ also exists and then the following density transformation formula applies:

$$f_{\boldsymbol{Y}}(y_1, \ldots, y_p) = f_{\boldsymbol{X}}[x_1(y_1, \ldots, y_p), \ldots, x_p(y_1, \ldots, y_p)]|J(y_1, \ldots, y_p)| \tag{0.19}$$

where $J(y_1, \ldots, y_p)$ is the *Jacobian* of the transformation:

$$J(y_1, \ldots, y_p) = \left|\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}}\right| \equiv \left| \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_p} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_p}{\partial y_1} & \frac{\partial x_p}{\partial y_2} & \cdots & \frac{\partial x_p}{\partial y_p} \end{pmatrix} \right| \equiv \left|\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right|^{-1} \tag{0.20}$$

Note that in (0.19) the *absolute value* of the Jacobian is substituted.

0.2.5 Characteristic and moment generating functions

The *characteristic function (cf)* $\varphi(\boldsymbol{t})$ of the random vector $\boldsymbol{X} \in \mathbb{R}^p$ is a function of a p-dimensional argument. For any real vector $\boldsymbol{t} = (t_1 \quad t_2 \quad \cdots \quad t_p)^\top \in \mathbb{R}^p$ it is defined as $\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = \mathrm{E}(\mathrm{e}^{i\boldsymbol{t}^\top \boldsymbol{X}})$ where $i = \sqrt{-1}$. Note that the cf always exists since $|\varphi_{\boldsymbol{X}}(\boldsymbol{t})| \le \mathrm{E}(|\mathrm{e}^{i\boldsymbol{t}^\top \boldsymbol{X}}|) = 1 < \infty$. Maybe more simple (since it does not involve complex numbers) is the notion of *moment generating function (mgf)*. It is defined as $M_{\boldsymbol{X}}(\boldsymbol{t}) = \mathrm{E}(\mathrm{e}^{\boldsymbol{t}^\top \boldsymbol{X}})$. Note however that in some cases the mgf may not exist for values of $\boldsymbol{t}$ further away from the zero vector.

Characteristic functions are in one-to-one correspondence with distributions and this is the reason to use them as a machinery to operate with in cases where direct operation with the distribution is not very convenient. In fact, when the density exists, under mild conditions the following inversion formulas hold for one-dimensional random variables and random vectors, respectively:

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \mathrm{e}^{-itx}\, \varphi_X(t) dt$$

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = (2\pi)^{-p} \int_{\mathbb{R}^p} \mathrm{e}^{-i\boldsymbol{t}^\top \boldsymbol{x}}\, \varphi_{\boldsymbol{X}}(\boldsymbol{t}) d\boldsymbol{t}.$$

One important property of *cf* is the following:

Theorem 0.3. *If the* cf $\varphi_{\boldsymbol{X}}(\boldsymbol{t})$ *of the random vector* $\boldsymbol{X} \in \mathbb{R}^p$ *is given and* $\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{b}, \boldsymbol{b} \in \mathbb{R}^q, A \in \mathcal{M}_{q,p}$ *is a linear transformation of* $\boldsymbol{X} \in \mathbb{R}^p$ *into a new random vector* $\boldsymbol{Y} \in \mathbb{R}^q$ *then it holds for all* $\boldsymbol{s} \in \mathbb{R}^q$ *that*

$$\varphi_{\boldsymbol{Y}}(\boldsymbol{s}) = \mathrm{e}^{i\boldsymbol{s}^\top \boldsymbol{b}}\, \varphi_{\boldsymbol{X}}(A^\top \boldsymbol{s}).$$

Proof. at lecture. □

0.3 Additional resources

An alternative presentation of these concepts can be found in JW Ch. 2–3.

0.4 Exercises

Exercise 0.1

In an ecological experiment, colonies of 2 different species of insect are confined to the same habitat. The survival times of the two species (in days) are random variables X_1 and X_2 respectively. It is thought that X_1 and X_2 have a joint density of the form

$$f_{\boldsymbol{X}}(x_1, x_2) = \theta x_1\, \mathrm{e}^{-x_1(\theta + x_2)} \qquad (0 < x_1, x_2)$$

for some constant $\theta > 0$.

(a) Show that $f_{\boldsymbol{X}}(x_1, x_2)$ is a valid density.

(b) Find the probability that both species die out within t days of the start of the experiment.

(c) Derive the marginal density of X_1. Identify this distribution and write down $\mathrm{E}(X_1)$ and $\mathrm{Var}(X_1)$.

(d) Derive the marginal density of X_2, and the conditional density of X_2 **given** $X_1 = x_1$.

(e) What evidence do you now have that X_1 and X_2 are not independent?

Exercise 0.2

Let $\boldsymbol{X} = [X_1, X_2]^\top$ a random vector with $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ and

$$\mathrm{Var}(\boldsymbol{X}) = \Sigma = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

(a) Find $\mathrm{Cov}(X_1 - X_2, X_1 + X_2)$.

(b) Find $\mathrm{Cov}(X_1, X_2 - \rho X_1)$.

(c) Choose b to minimise $\mathrm{Var}(X_2 - bX_1)$.

Exercise 0.3

Suppose $\boldsymbol{X}$ is a p-dimensional random vector with cf $\varphi_{\boldsymbol{X}}(\boldsymbol{t})$. If $\boldsymbol{X}$ is partitioned as $\begin{bmatrix} \boldsymbol{X}_{(1)} \\ \boldsymbol{X}_{(2)} \end{bmatrix}$, where $\boldsymbol{X}_{(1)}$ is a p_1-dimensional subvector, then show that

(a) $\boldsymbol{X}_{(1)}$ has cf $\varphi^{(1)}_{\boldsymbol{X}}(\boldsymbol{t}_{(1)}) = \varphi_{\boldsymbol{X}}\left\{ \begin{bmatrix} \boldsymbol{t}_{(1)} \\ \boldsymbol{0} \end{bmatrix} \right\}$, $\boldsymbol{t}_{(1)} \in \mathbb{R}^{p_1}$.

(b) $\boldsymbol{X}_{(1)}$ and $\boldsymbol{X}_{(2)}$ are independent if and only if

$$\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = \varphi_{\boldsymbol{X}}\left\{ \begin{bmatrix} \boldsymbol{t}_{(1)} \\ \boldsymbol{0} \end{bmatrix} \right\} \varphi_{\boldsymbol{X}}\left\{ \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{t}_{(2)} \end{bmatrix} \right\},$$

$\forall \boldsymbol{t}_{(1)} \in \mathbb{R}^{p_1}$, $\forall \boldsymbol{t}_{(2)} \epsilon \mathbb{R}^{p-p_1}$.

Exercise 0.4

Let $\boldsymbol{X} \in \mathcal{M}_{p,p}$ is a symmetric positive definite matrix with eigenvalues $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p > 0$ and associated eigenvectors of unit length $\boldsymbol{e}_i, i = 1, 2, \ldots, p$ that give rise to the following spectral decomposition:

$$X = \lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1^\top + \lambda_2 \boldsymbol{e}_2 \boldsymbol{e}_2^\top + \ldots \lambda_p \boldsymbol{e}_p \boldsymbol{e}_p^\top$$

It is known that $\max_{\boldsymbol{y} \neq 0} \frac{\boldsymbol{y}^\top X \boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{y}} = \lambda_1$. Now, you show that $\max_{\boldsymbol{y} \neq 0, \langle \boldsymbol{y}, \boldsymbol{e}_1 \rangle = 0} \frac{\boldsymbol{y}^\top X \boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{y}} = \lambda_2$. Can you find further generalisations of this claim?

Exercise 0.5

We know that an orthogonal projection matrix has only 0 or 1 as possible eigenvalues. Using this property or otherwise, show that the rank of an orthogonal projector is equal to the sum of its diagonal elements.

1 Exploratory Data Analysis of Multivariate Data

1.1 Data organisation

Assume, we are dealing with $p \geq 1$ *variables*. The values of these variables are all recorded for each distinct *item, individual*, or *experimental trial*. Each of these three words will be substituted sometimes by the word "case". We will use the notation x_{ij} to indicate a particular value of the ith variable that is observed on the jth case. Consequently, n measurements on p variables can be represented in a form of a matrix

$$\overset{p\times n}{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{in} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pj} & \cdots & x_{pn} \end{pmatrix} \tag{1.1}$$

The matrix X above contains the data consisting of all the observations on all the variables. This way of representing the data allows easy manipulations to be performed in order to obtain some easy descriptive statistics for each of the variables.

1.2 Basic summaries

For example, the *sample mean* of the second variable is just $\bar{x}_2 = \frac{1}{n}\sum_{j=1}^{n} x_{2j}$, the *sample variance* of the second variable is just $s_2^2 = \frac{1}{n}\sum_{j=1}^{n}(x_{2j} - \bar{x}_2)^2$ (Note that for the sample variance we shall sometimes use the divisor of $n-1$ rather than n and each time this will be differentiated by displaying the appropriate expression).

The *sample covariance* (the simple measure of linear association between variables 1 and 2) is given by $s_{12} = \frac{1}{n}\sum_{j=1}^{n}(x_{1j} - \bar{x}_1)(x_{2j} - \bar{x}_2)$ and one can understand easily how $s_{ik}, i = 1, 2, \ldots, p$, $k = 1, 2, \ldots, p$ can be defined. Finally, the *sample correlation coefficient* (the measure of linear association between two variables that does not depend on the units of measurement) can be defined. The sample correlation coefficient of the ith and kth variables is defined by $r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}}$. Because of the well-known Cauchy–Bunyakovsky–Schwartz Inequality, $|r_{ik}| \leq 1$ holds. Note also that $r_{ik} = r_{ki}$ for all $i = 1, 2, \ldots, p$ and $k = 1, 2, \ldots, p$ holds.

It should be repeatedly noted that the sample correlations and covariance are useful only when trying to measure the *linear* association between two variables. Their value is less informative and is misleading in cases of *nonlinear* association. In this case one needs to invoke the *quotient correlation* instead:

> Zhang, Zhengjun. Quotient correlation: A sample based alternative to Pearson's correlation. *Annals of Statistics* 36 (2008), no. 2, 1007--1030. doi:10.1214/009053607000000866

But because of the fact that covariance and correlation coefficients are routinely calculated and analysed they are very widely used and provide nice numerical summaries of association when the data do not exhibit obvious nonlinear patterns of association.

The descriptive statistics that we discussed until now are usually organised into arrays, namely:

Vector of sample means $\bar{\boldsymbol{x}} = \begin{pmatrix} \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \end{pmatrix}^\top$

Matrix of sample variances and covariances

$$\underset{p\times p}{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix} \tag{1.2}$$

Matrix of sample correlations

$$\underset{p\times p}{R} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix} \tag{1.3}$$

1.3 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.

1.4 Software

SAS 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.

R 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
```

2 The Multivariate Normal Distribution

2.1 Definition

The *multivariate normal* (*MVN*) density is a generalisation of the univariate normal for $p \geq 2$ dimensions. Looking at the term $(\frac{x-\mu}{\sigma})^2 = (x-\mu)(\sigma^2)^{-1}(x-\mu)$ in the exponent of the well known formula

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, \mathrm{e}^{-[(x-\mu)/\sigma]^2/2}, -\infty < x < \infty \tag{2.1}$$

for the univariate density function, a natural way to generalise this term in higher dimensions is to *replace* it by $(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$. Here $\boldsymbol{\mu} = \mathrm{E}\,\boldsymbol{X} \in \mathbb{R}^p$ is the expected value of the random vector $\boldsymbol{X} \in \mathbb{R}^p$ and the matrix

$$\Sigma = \mathrm{E}(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^\top = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix} \in \mathcal{M}_{p,p}$$

is the *covariance matrix*. Note that on the diagonals of Σ we get the *variances* of each of the p random variables whereas $\sigma_{ij} = \mathrm{E}[(X_i - \mathrm{E}(X_i))(X_j - \mathrm{E}(X_j))], i \neq j$ are the *covariances* between the ith and jth random variable. Sometimes, we will also denote σ_{ii} by σ_i^2.

Of course, the above replacement would only make sense if Σ was positive definite. In general, however, we can only claim that Σ is (as any covariance matrix) non-negative definite (try to prove this claim e.g. using Example 0.2 from Section 0.1.5 or some other argument).

If Σ was positive definite then the density of the random vector $\boldsymbol{X}$ can be written as

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{\frac{1}{2}}}\, \mathrm{e}^{-(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/2},\ -\infty < x_i < \infty,\ i = 1, 2 \ldots, p. \tag{2.2}$$

It can be directly checked that the random vector $\boldsymbol{X} \in \mathbb{R}^p$ has $\mathrm{E}\,\boldsymbol{X} = \boldsymbol{\mu}$ and

$$\mathrm{E}[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^\top] = \Sigma.$$

Since the density is uniquely defined by the *mean vector and the covariance matrix* we will denote it by $N_p(\boldsymbol{\mu}, \Sigma)$.

In these notes, however, we will introduce the multivariate normal distribution not through its density formula but through more general reasoning that also allows to cover the case of singular Σ. We will utilise the famous **Cramer–Wold argument** according to which the distribution of a p-dimensional random vector $\boldsymbol{X}$ is completely characterised by the one-dimensional distributions of **all** linear transformations $\boldsymbol{t}^\top \boldsymbol{X}, \boldsymbol{t} \in \mathbb{R}^p$. Indeed, if we consider $\mathrm{E}[\mathrm{e}^{(i t \boldsymbol{t}^\top \boldsymbol{X})}]$ (which is assumed to be known for every $t \in \mathbb{R}^1, \boldsymbol{t} \in \mathbb{R}^p$) then we see that by substituting $t = 1$ we can get $\mathrm{E}[\mathrm{e}^{(i \boldsymbol{t}^\top \boldsymbol{X})}]$ which is the cf of the vector $\boldsymbol{X}$ (and the latter uniquely specifies the distribution of $\boldsymbol{X}$). Hence the following definition will be adopted here:

Definition 2.1. The random vector $\boldsymbol{X} \in \mathbb{R}^p$ has a multivariate normal distribution if and only if (iff) any linear transformation $\boldsymbol{t}^\top \boldsymbol{X}, \boldsymbol{t} \in \mathbb{R}^p$ has a univariate normal distribution.

Lemma 2.2. *The characteristic function of the (univariate) standard normal random variable* $X \sim N(0,1)$ *is*

$$\psi_X(t) = \exp(-t^2/2).$$

Proof. (optional, not examinable)

$$\begin{aligned}
\psi_X(t) = \mathrm{E}\exp(itX) &= \int_{-\infty}^{+\infty} \exp(itx)\frac{1}{\sqrt{2\pi}}\exp(-x^2/2)dx \\
&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}\exp(itx - x^2/2)dx \\
&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}\exp(-(x^2 - 2itx)/2)dx \\
&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}\exp(-(x^2 - 2itx + (it)^2)/2 + (it)^2/2)dx \\
&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}\exp(-(x - it)^2/2 + (it)^2/2)dx \\
&= \exp(-t^2/2)\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}\exp(-(x - it)^2/2)dx \\
&= \exp(-t^2/2)\lim_{h\to\infty}\int_{-h}^{+h} \frac{1}{\sqrt{2\pi}}\exp(-(x - it)^2/2)dx.
\end{aligned}$$

Change of variable:

$$\begin{aligned}
z &= x - it \\
x &= z + it \\
dx &= dz
\end{aligned}$$

results in

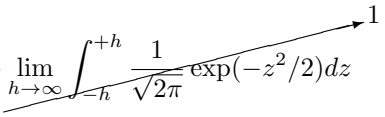
$$\psi_X(t) = \exp(-t^2/2)\lim_{h\to\infty}\int_{-h+it}^{+h+it} \frac{1}{\sqrt{2\pi}}\exp(-z^2/2)dz.$$

The remaining integral is over a complex domain, so we must use Cauchy's Theorem: contour integration over the contour $+h + it \to +h \to -h \to -h + it \to +h + it$ should result in 0, so

$$\begin{aligned}
\int_{+h+it}^{+h} \frac{1}{\sqrt{2\pi}}\exp(-z^2/2)dz + \int_{+h}^{-h} \frac{1}{\sqrt{2\pi}}\exp(-z^2/2)dz + \\
\int_{-h}^{-h+it} \frac{1}{\sqrt{2\pi}}\exp(-z^2/2)dz + \int_{-h+it}^{+h+it} \frac{1}{\sqrt{2\pi}}\exp(-z^2/2)dz = 0
\end{aligned}$$

for any real h and t. Solving for the integral of interest and taking the limit,

$$\begin{aligned}
&\lim_{h\to\infty}\int_{-h+it}^{+h+it}\frac{1}{\sqrt{2\pi}}\exp(-z^2/2)dz \\
&\quad= -\lim_{h\to\infty}\int_{+h+it}^{+h}\frac{1}{\sqrt{2\pi}}\exp(-z^2/2)dz - \lim_{h\to\infty}\int_{+h}^{-h}\frac{1}{\sqrt{2\pi}}\exp(-z^2/2)dz \\
&\qquad-\lim_{h\to\infty}\int_{-h}^{-h+it}\frac{1}{\sqrt{2\pi}}\exp(-z^2/2)dz \\
&\quad= -\lim_{h\to\infty}\int_{+h+it}^{+h}\frac{1}{\sqrt{2\pi}}\exp(-z^2/2)dz + \lim_{h\to\infty}\int_{-h}^{+h}\frac{1}{\sqrt{2\pi}}\exp(-z^2/2)dz \\
&\qquad-\lim_{h\to\infty}\int_{-h}^{-h+it}\frac{1}{\sqrt{2\pi}}\exp(-z^2/2)dz,
\end{aligned}$$



since the standard normal density integrates to 1. □

Lastly, consider $\lim_{h\to\infty}\int_{+h+it}^{+h}\exp(-z^2/2)dz$: change of variable

$$\begin{aligned}
y &= (z-h)/i \\
z &= h+iy \\
dz &= idy,
\end{aligned}$$

then

$$\begin{aligned}
\lim_{h\to\infty}\int_{+h+it}^{+h}\exp(-z^2/2)dz &= \lim_{h\to\infty}\int_1^0\exp(-(h+iy)^2/2)idy \\
&= \int_1^0\lim_{h\to\infty}\exp(-(h^2+2ihy-y^2)/2)idy \\
&= \int_1^0\lim_{h\to\infty}\exp(-h^2/2)\exp(-ihy)\exp(-y^2/2)idy \\
&= \int_1^0 0dy = 0,
\end{aligned}$$

and, analogously, $\lim_{h\to\infty}\int_{-h}^{-h+it}\frac{1}{\sqrt{2\pi}}\exp(-z^2/2)dz = 0$, leaving

$$\lim_{h\to\infty}\int_{-h+it}^{+h+it}\frac{1}{\sqrt{2\pi}}\exp(-z^2/2)dz = 1$$

and

$$\psi_X(t) = \exp(-t^2/2).$$

Aside: The *mgf* $M_X(t) = \mathrm{E}\exp(tX)$ can also be derived and used in the argument below; however, *cf*s are more general so are preferred when possible. We show the (optional, not examinable) derivation here.

We begin by completing the square:

$$
\begin{aligned}
M_X(t) = \mathrm{E}\exp(tX) &= \int_{-\infty}^{+\infty} \exp(tx)\frac{1}{\sqrt{2\pi}}\exp(-x^2/2)dx \\
&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}\exp(tx - x^2/2)dx \\
&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}\exp(-(x^2 - 2tx)/2)dx \\
&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}\exp(-(x^2 - 2tx + t^2)/2 + t^2/2)dx \\
&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}\exp(-(x-t)^2/2 + t^2/2)dx \\
&= \exp(t^2/2)\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}\exp(-(x-t)^2/2)dx.
\end{aligned}
$$

Change of variable:

$$
\begin{aligned}
z &= x - t \\
x &= z + t \\
dx &= dz
\end{aligned}
$$

results in

$$
M_X(t) = \exp(t^2/2)\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}\exp(-z^2/2)dz = \exp(t^2/2),
$$

since the integrand is just a standard normal density.

Theorem 2.3.

Suppose that for a random vector $\boldsymbol{X} \in \mathbb{R}^p$ with a normal distribution according to Definition 2.1 we have $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ and $D(\boldsymbol{X}) = \mathrm{E}[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^\top] = \Sigma$. Then:

i) For any fixed $\boldsymbol{t} \in \mathbb{R}^p$, $\boldsymbol{t}^\top\boldsymbol{X} \sim N(\boldsymbol{t}^\top\boldsymbol{\mu}, \boldsymbol{t}^\top\Sigma\boldsymbol{t})$ i.e. $\boldsymbol{t}^\top\boldsymbol{X}$ has an one dimensional normal distribution with expected value $\boldsymbol{t}^\top\boldsymbol{\mu}$ and variance $\boldsymbol{t}^\top\Sigma\boldsymbol{t}$.

ii) The cf *of $\boldsymbol{X} \in \mathbb{R}^p$ is*

$$
\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = \mathrm{e}^{(i\boldsymbol{t}^\top\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^\top\Sigma\boldsymbol{t})}. \tag{2.3}
$$

Proof. Part i) is obvious. For part ii) we recall from Lemma 2.2 that the cf of the standard univariate normal random variable Z is $\mathrm{e}^{-t^2/2}$. Since any $U \sim N_1(\mu_1, \sigma_1^2)$ has a distribution that coincides with the distribution of $\mu_1 + \sigma_1 Z$ we have:

$$
\varphi_U(t) = \mathrm{e}^{it\mu_1}\,\varphi_{\sigma_1 Z}(t) = \mathrm{e}^{it\mu_1}\,\mathrm{E}(\mathrm{e}^{it\sigma_1 Z}) = \mathrm{e}^{it\mu_1}\,\varphi_Z(t\sigma_1) = \mathrm{e}^{(it\mu_1 - \frac{1}{2}t^2\sigma_1^2)}
$$

But then, for the *univariate random variable* $\boldsymbol{t}^\top\boldsymbol{X} \sim N_1(\boldsymbol{t}^\top\boldsymbol{\mu}, \boldsymbol{t}^\top\Sigma\boldsymbol{t})$ we would have as a characteristic function $\varphi_{\boldsymbol{t}^\top\boldsymbol{X}}(t) = \mathrm{e}^{(it\boldsymbol{t}^\top\boldsymbol{\mu} - \frac{1}{2}t^2\boldsymbol{t}^\top\Sigma\boldsymbol{t})}$. Substituting $t = 1$ in the latter formula we find that

$$
\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = \mathrm{e}^{(i\boldsymbol{t}^\top\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^\top\Sigma\boldsymbol{t})}.
$$

□

As an upshot, we see that given the expected value vector $\boldsymbol{\mu}$ and the covariance matrix Σ we can use the *cf* formula (2.3) rather than the density formula (2.2) to define the p dimensional multivariate normal distribution. The advantage of the former in comparison to the latter is that in (2.3) only Σ is used, i.e. this definition makes also sense in cases of singular (i.e. non-invertible) Σ. We still want to know that in case of non-singular Σ the more general definition would give raise to the density (2.2). This is the content of the next theorem.

Theorem 2.4. *Assume the matrix Σ in* (2.3) *is nonsingular. Then the density of the random vector $\boldsymbol{X} \in \mathbb{R}^p$ with cf as in* (2.3) *is given by* (2.2).

Proof. Consider the vector $\boldsymbol{Y} \in \mathbb{R}^p$ such that $\boldsymbol{Y} = \Sigma^{-\frac{1}{2}}(\boldsymbol{X}-\boldsymbol{\mu})$ (compare (0.10) in Section 0.1.5). Since obviously $\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{0}$ and $D(\boldsymbol{Y}) = \mathrm{E}(\boldsymbol{Y}\boldsymbol{Y}^\top) = \Sigma^{-\frac{1}{2}}\,\mathrm{E}[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^\top]\Sigma^{-\frac{1}{2}} = I_p$ holds we can substitute to get the cf of $\boldsymbol{Y} \in \mathbb{R}^p$: $\varphi_{\boldsymbol{Y}}(\boldsymbol{t}) = \mathrm{e}^{-\frac{1}{2}\sum_{i=1}^p t_i^2}$. But the latter can be seen directly to be the characteristic function of the vector of p independent standard normal variables. Hence, from the relation $\boldsymbol{Y} = \Sigma^{-\frac{1}{2}}(\boldsymbol{X}-\boldsymbol{\mu})$ we can also conclude that $\boldsymbol{X} = \boldsymbol{\mu} + \Sigma^{\frac{1}{2}}\boldsymbol{Y}$ where the density $f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{(2\pi)^{p/2}}\,\mathrm{e}^{-\frac{1}{2}\sum_{i=1}^p y_i^2}$. With other words, $\boldsymbol{X}$ is a *linear transformation* of $\boldsymbol{Y}$ where the density of $\boldsymbol{Y}$ is *known*. We can therefore apply the density transformation approach (Section 0.2.4 of this lecture) to obtain: $f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{\boldsymbol{Y}}(\Sigma^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\mu}))|J(x_1,\ldots,x_p)|$. It is easy to see (because of the linearity of the transformation) that $|J(x_1,\ldots,x_p)| = |\Sigma^{-\frac{1}{2}}| = |\Sigma^{\frac{1}{2}}|^{-1}$. Taking into account that $\sum_{i=1}^p y_i^2 = \boldsymbol{y}^\top\boldsymbol{y} = (\boldsymbol{x}-\boldsymbol{\mu})^\top\Sigma^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\mu}) = (\boldsymbol{x}-\boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ we finally arrive at the density formula (2.2) for $f_{\boldsymbol{X}}(\boldsymbol{x})$. □

2.2 Properties of multivariate normal

The following *properties* of multivariate normal can be easily derived using the machinery developed so far:

Property 1

If $\Sigma = D(\boldsymbol{X}) = \Lambda$ is a diagonal matrix then the p components of $\boldsymbol{X}$ are independent.

(Indeed, in this case $\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = \mathrm{e}^{i\sum_{j=1}^p t_j\mu_j - \frac{1}{2}t_j^2\sigma_j^2}$ which can be seen to be the *cf* of the vector of p independent components each distributed according to $N(\mu_j,\sigma_j^2), j = 1,\ldots,p$).

The above property can be paraphrased as "for a multivariate normal, if its components are uncorrelated they are also independent". On the other hand, it is well known that *always, i.e. not only for normal* from the fact that certain components are independent we can conclude that they are also uncorrelated. Therefore, for the **multivariate normal distribution** we can conclude that its components are **independent if and only if they are uncorrelated!**

Example 2.5 (Random variables that are marginally normal and uncorrelated but not independent). Consider two variables $Z_1 = (2W-1)Y$ and $Z_2 = Y$, where $Y \sim N_1(0,1)$ and, independently, $W \sim \mathrm{Binomial}(1,1/2)$ (so $2W-1$ takes -1 and $+1$ with equal probability).

Property 2

If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu},\Sigma)$ and $C \in \mathcal{M}_{q,p}$ is an arbitrary matrix of real numbers then

$$\boldsymbol{Y} = C\boldsymbol{X} \sim N_q(C\boldsymbol{\mu}, C\Sigma C^\top).$$

To prove this property note that (see Section 0.2.5) for any $\boldsymbol{s} \in \mathbb{R}^q$ we have:

$$\varphi_{\boldsymbol{Y}}(\boldsymbol{s}) = \varphi_{\boldsymbol{X}}(C^\top\boldsymbol{s}) = \mathrm{e}^{i\boldsymbol{s}^\top C\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{s}^\top C\Sigma C^\top\boldsymbol{s}}$$

which means that $\boldsymbol{Y} = C\boldsymbol{X} \sim N_q(C\boldsymbol{\mu}, C\Sigma C^\top)$.

Note also that if it happens that the rank of C is full and if $\mathrm{rk}(\Sigma) = p$ then the rank of $C\Sigma C^\top$ is also full, i.e. the distribution of $\boldsymbol{Y}$ would not be degenerate in this case.

Property 3

(This is a finer version of Property 1). Assume the vector $\boldsymbol{X} \in \mathbb{R}^p$ is divided into subvectors $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_{(1)} \\ \boldsymbol{X}_{(2)} \end{pmatrix}$ and according to this subdivision the vector means are $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_{(1)} \\ \boldsymbol{\mu}_{(2)} \end{pmatrix}$ and the covariance matrix Σ has been subdivided into $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. Then the vectors $\boldsymbol{X}_{(1)}$ and $\boldsymbol{X}_{(2)}$ are independent iff $\Sigma_{12} = 0$.

Proof. (Exercise (see lecture)). □

Property 4

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

$$N_r(\boldsymbol{\mu}_{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_{(2)} - \boldsymbol{\mu}_{(2)}), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) \tag{2.4}$$

Proof. Perhaps the easiest way to proceed is the following. Note that the expression $\boldsymbol{\mu}_{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_{(2)} - \boldsymbol{\mu}_{(2)})$ (for which we want to show that it equals the conditional mean), is a function of $\boldsymbol{x}_{(2)}$. Denote is as $g(\boldsymbol{x}_{(2)})$ for short. Let us construct the random vectors $\boldsymbol{Z} = \boldsymbol{X}_{(1)} - g(\boldsymbol{X}_{(2)})$ and $\boldsymbol{Y} = \boldsymbol{X}_{(2)} - \boldsymbol{\mu}_{(2)}$. Obviously $\mathrm{E}\,\boldsymbol{Z} = \boldsymbol{0}$ and $\mathrm{E}\,\boldsymbol{Y} = \boldsymbol{0}$ holds. The vector $\begin{pmatrix} \boldsymbol{Z} \\ \boldsymbol{Y} \end{pmatrix}$ is a *linear transformation of a normal vector* ($\begin{pmatrix} \boldsymbol{Z} \\ \boldsymbol{Y} \end{pmatrix} = A(\boldsymbol{X} - \boldsymbol{\mu})$, $A = \begin{pmatrix} I_r & -\Sigma_{12}\Sigma_{22}^{-1} \\ \boldsymbol{0} & I_{p-r} \end{pmatrix}$) and hence, its distribution is normal (Property 2). Calculating therefore covariance matrix of the vector $\begin{pmatrix} \boldsymbol{Z} \\ \boldsymbol{Y} \end{pmatrix}$ we find that

$$\mathrm{Var}\begin{pmatrix} \boldsymbol{Z} \\ \boldsymbol{Y} \end{pmatrix} = A\Sigma A^\top = \begin{pmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & \boldsymbol{0} \\ \boldsymbol{0} & \Sigma_{22} \end{pmatrix}$$

after a simple exercise in block multiplication of matrices.

Hence the two vectors $\boldsymbol{Z}$ and $\boldsymbol{Y}$ are uncorrelated normal vectors and therefore are independent (Property 3). But $\boldsymbol{Y}$ is a linear transformation of $\boldsymbol{X}_{(2)}$ and this means that $\boldsymbol{Z}$ and $\boldsymbol{X}_{(2)}$ are independent. Hence the conditional density of $\boldsymbol{Z}$ given $\boldsymbol{X}_{(2)} = \boldsymbol{x}_{(2)}$ will **not** depend on $\boldsymbol{x}_{(2)}$ and coincides with the unconditional density of $\boldsymbol{Z}$. This means, it is normal with zero mean vector and its covariance matrix is

$$\mathrm{Cov}(\boldsymbol{Z}) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \Sigma_{1|2}$$

Hence we can state that $\boldsymbol{X}_{(1)} - g(\boldsymbol{x}_{(2)}) \sim N(\boldsymbol{0}, \Sigma_{1|2})$ and correspondingly, the conditional distribution of $\boldsymbol{X}_{(1)}$ given that $\boldsymbol{X}_{(2)} = \boldsymbol{x}_{(2)}$ is (2.4). □

Example 2.6. As an immediate consequence of Property 4 we see that if $p = 2, r = 1$ then for a two-dimensional normal vector $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left\{ \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right\}$ its conditional density $f(x_1|x_2)$ is $N(\mu_1 + \frac{\sigma_{12}}{\sigma_2^2}(x_2 - \mu_2), \sigma_1^2 - \frac{\sigma_{12}^2}{\sigma_2^2})$.

As an exercise, try to derive the above result by direct calculations starting from the joint density $f(x_1, x_2)$, going over to the marginal $f(x_2)$ by integration and finally getting $f(x_1|x_2) = \frac{f(x_1,x_2)}{f(x_2)}$.

Property 5

If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$ and Σ is nonsingular then $(\boldsymbol{X} - \boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \sim \chi_p^2$ where χ_p^2 denotes the chi-square distribution with p degrees of freedom.

Proof. It suffices to use the fact that (see also Theorem 2.4) the vector $\boldsymbol{Y} \in \mathbb{R}^p : \boldsymbol{Y} = \Sigma^{-\frac{1}{2}}(\boldsymbol{X} - \boldsymbol{\mu}) \sim N(\boldsymbol{0}, I_p)$ i.e. it has p independent standard normal components. Then

$$(\boldsymbol{X} - \boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) = \boldsymbol{Y}^\top \boldsymbol{Y} = \sum_{i=1}^{p} Y_i^2 \sim \chi_p^2$$

according to the definition of χ_p^2 as a distribution of the sum of squares of p independent standard normals. □

Finally, one more interpretation of the result in Property 4 will be given. Assume we want, as is a typical situation in statistics, to predict a random variable Y that is correlated with some p random variables (predictors) $\boldsymbol{X} = (X_1 \quad X_2 \quad \cdots \quad X_p)$. Trying to find the *best predictor* of Y we would like to minimise the expected value $\mathrm{E}_Y[\{Y - g(\boldsymbol{X})\}^2 | \boldsymbol{X} = \boldsymbol{x}]$ over all possible choices of the function g such that $\mathrm{E}\, g(\boldsymbol{X})^2 < \infty$. A little careful work and use of basic properties of conditional expectations leads us (see lecture) to the conclusion that the optimal solution to the above minimisation problem is $g^*(\boldsymbol{x}) = \mathrm{E}(Y|\boldsymbol{X} = \boldsymbol{x})$. This optimal solution is also called the *regression function*. Thus given a particular realisation $\boldsymbol{x}$ of the random vector $\boldsymbol{X}$ the regression function is just the conditional expected value of Y given $\boldsymbol{X} = \boldsymbol{x}$.

In general, the conditional expected value may be a complicated nonlinear function of the predictors. However, if we assume *in addition* that the joint $(p+1)$-dimensional distribution of Y and $\boldsymbol{X}$ is **normal** then by applying Property 4 we see that given the realisation $\boldsymbol{x}$ of $\boldsymbol{X}$, the best prediction of the Y value is given by $b + \sigma_0^\top C^{-1}\boldsymbol{x}$ where $b = \mathrm{E}(Y) - \sigma_0^\top C^{-1}\,\mathrm{E}(\boldsymbol{X}), C$ is the covariance matrix of the vector $\boldsymbol{X}$, σ_0 is the vector of Covariances of Y with $X_i, i = 1, \ldots, p$.

Indeed, we know that when the joint $(p+1)$-dimensional distribution of Y and $\boldsymbol{X}$ is **normal** the regression function is given by

$$\mathrm{E}(Y) + \sigma_0^\top C^{-1}(\boldsymbol{x} - \mathrm{E}(\boldsymbol{X})).$$

By introducing the notation $b = \mathrm{E}(Y) - \sigma_0^\top C^{-1}\,\mathrm{E}(\boldsymbol{X})$ we can write this as $b + \sigma_0^\top C^{-1}\boldsymbol{x}$.

That is, **in case of normality, the optimal predictor of Y in the least squares sense turns out to be a very simple linear function of the predictors**. The vector $C^{-1}\sigma_0 \in \mathbb{R}^p$ is the *vector of the regression coefficients*. Substituting the optimal values we get the minimal value of the sum of squares which is equal to $\mathrm{Var}(Y) - \sigma_0^\top C^{-1}\sigma_0$.

2.3 Tests for Multivariate Normality

We have seen that the assumption of multivariate normality may bring essential simplifications in analysing data. But applying inference methods based on the multivariate normality assumption in cases where it is grossly violated may introduce serious defects in the quality of the analysis. It is therefore important to be able to check the multivariate normality assumption. Based on the properties of normal distributions discussed in this lecture, we know that all linear combinations of normal variables are normal and the contours of the multivariate normal density are ellipsoids. Therefore we can (to some extent) check the multivariate normality hypothesis by:

1. checking if the marginal distributions of each component appear to be normal (by using Q–Q plots and the Shapiro–Wilk test, for example);
2. checking if the scatterplots of pairs of observations give the elliptical appearance expected from normal populations;
3. are there any outlying observations that should be checked for accuracy.

All this can be done by applying univariate techniques and by drawing scatterplots which are well developed in SAS and R. To some extent, however, there is a price to be paid for concentrating on univariate and bivariate examinations of normality.

There is a need to construct a "good" overall test of multivariate normality. One of the simple and tractable ways to verify the multivariate normality assumption is by using tests based on **Mardia's multivariate skewness and kurtosis measures**. For any general multivariate distribution we define these respectively as

$$\beta_{1,p} = \mathrm{E}[(\boldsymbol{Y}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{X}-\boldsymbol{\mu})]^3 \tag{2.5}$$

provided that $\boldsymbol{X}$ is independent of $\boldsymbol{Y}$ but has the same distribution and

$$\beta_{2,p} = \mathrm{E}[(\boldsymbol{X}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{X}-\boldsymbol{\mu})]^2 \tag{2.6}$$

(if the expectations in (2.5) and (2.6) exist). For the $N_p(\boldsymbol{\mu}, \Sigma)$ distribution: $\beta_{1,p} = 0$ and $\beta_{2,p} = p(p+2)$.

(Note that when $p = 1$, the quantity $\beta_{1,1}$ is the square of the skewness coefficient $\frac{\mathrm{E}(X-\mu)^3}{\sigma^3}$ whereas $\beta_{2,1}$ coincides with the kurtosis coefficient $\frac{\mathrm{E}(X-\mu)^4}{\sigma^4}$.)

For a sample of size n *consistent estimates* of $\beta_{1,p}$ and $\beta_{2,p}$ can be obtained as

$$\hat{\beta}_{1,p} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} g_{ij}^3$$

$$\hat{\beta}_{2,p} = \frac{1}{n}\sum_{i=1}^{n} g_{ii}^2$$

where $g_{ij} = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top \boldsymbol{S}_n^{-1}(\boldsymbol{x}_j - \bar{\boldsymbol{x}})$. Notice that for $\hat{\beta}_{1,p}$, we take advantage of our sample being independent and use observations $\boldsymbol{x}_j$ for $j \neq i$ as the "$\boldsymbol{Y}$" values for $\boldsymbol{x}_i$.

Both quantities $\hat{\beta}_{1,p}$ and $\hat{\beta}_{2,p}$ are nonnegative and for multivariate data, one would expect them to be around zero and $p(p+2)$, respectively. Both quantities can be utilised to detect departures from multivariate normality.

Mardia has shown that asymptotically, $k_1 = n\hat{\beta}_{1,p}/6 \sim \chi^2_{p(p+1)(p+2)/6}$, and $k_2 = [\hat{\beta}_{2,p} - p(p+2)]/[8p(p+2)/n]^{\frac{1}{2}}$ is standard normal. Thus we can use k_1 and k_2 to test the null hypothesis of

multivariate normality. If neither hypothesis is rejected, the multivariate normality assumption is in reasonable agreement with the data. It also has been observed that Mardia's multivariate kurtosis can be used as a measure to detect outliers from the data that are supposedly distributed as multivariate normal.

Shapiro–Wilk, Mardia, and other distribution tests have, as their null hypothesis, that the true population distribution is (multivariate) normal. This means that if the population distribution deviates from normality even a little, then as the sample size increases, the power of the test (the probability of rejecting the null hypothesis of normality) approaches 1.

At the same time, as the sample size increases, the Central Limit Theorem tells us that many statistics, including sample means and (much more slowly) sample variances and covariances, approach normality—and multivariate statistics generally approach multivariate normality. This means that regardless of the underlying distribution, the statistical procedures depending on the normality assumption become valid—even as the chances that a statistical hypothesis test will detect what non-normality there is approaches 1.

This means that we must not rely on hypothesis testing blindly but consider the situation on a case-by-case basis, particularly when dealing with large datasets. For a decent sample size, the "symmetric, bell-shaped" heuristic may indicate an adequate distribution, even if a hypothesis test reports a small p-value.

2.4 Software

SAS Use `CALIS` procedure. The quantity k_2 is called *Normalised Multivariate Kurtosis* there, whereas $\hat{\beta}_{2,p} - p(p+2)$ bears the name *Mardia's Multivariate Kurtosis*.

R `MVN::mvn`, `psych::mardia`

2.5 Examples

Example 2.7. Testing multivariate normality of microwave oven radioactivity measurements (JW).

2.6 Additional resources

An alternative presentation of these concepts can be found in JW Sec. 4.1–4.2, 4.6.

2.7 Exercises

Exercise 2.1

Let X_1 and X_2 denote i.i.d. $N(0,1)$ r.v.'s.

(a) Show that the r.v.'s $Y_1 = X_1 - X_2$ and $Y_2 = X_1 + X_2$ are **independent**, and find their marginal densities.

(b) Find $P(X_1^2 + X_2^2 < 2.41)$.

Exercise 2.2

Let $\boldsymbol{X} \sim N_3(\boldsymbol{\mu}, \Sigma)$ where

$$\boldsymbol{\mu} = \begin{pmatrix} 3 \\ -1 \\ 2 \end{pmatrix} \quad \Sigma = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 3 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

(a) For $A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -2 & 1 \end{pmatrix}$ find the distribution of $\boldsymbol{Z} = A\boldsymbol{X}$ and find the correlation between the two components of $\boldsymbol{Z}$.

(b) Find the conditional distribution of $[X_1, X_3]^\top$ given $X_2 = 0$.

Exercise 2.3

Suppose that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independent random vectors, with each $\boldsymbol{X}_i \sim N_p(\boldsymbol{\mu}_i, \Sigma_i)$. Let $a_1, \ldots, a_n$ be real constants. Using characteristic functions, show that

$$a_1\boldsymbol{X}_1 + \cdots + a_n\boldsymbol{X}_n \sim N_p(a_1\boldsymbol{\mu}_1 + \cdots + a_n\boldsymbol{\mu}_n, a_1^2\Sigma_1 + \cdots + a_n^2\Sigma_n)$$

Therefore, deduce that, if $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ form a random sample from the $N_p(\boldsymbol{\mu}, \Sigma)$ distribution, then the sample mean vector, $\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i$, has distribution

$$\bar{\boldsymbol{X}} \sim N_p(\boldsymbol{\mu}, \frac{1}{n}\Sigma) \ .$$

Exercise 2.4

Prove that if $\boldsymbol{X}_1 \sim N_r(\boldsymbol{\mu}_1, \Sigma_{11})$ and $(\boldsymbol{X}_2|\boldsymbol{X}_1 = \boldsymbol{x}_1) \sim N_{p-r}(A\boldsymbol{x}_1 + \boldsymbol{b}, \Omega)$ where Ω does not depend on $\boldsymbol{x}_1$ then $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} \sim N_p(\boldsymbol{\mu}, \Sigma)$ where

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ A\boldsymbol{\mu}_1 + \boldsymbol{b} \end{pmatrix}, \ \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{11}A^\top \\ A\Sigma_{11} & \Omega + A\Sigma_{11}A^\top \end{pmatrix}.$$

Exercise 2.5

Knowing that,

i) $Z \sim N_1(0, 1)$

ii) $Y|Z = z \sim N_1(1 + z, 1)$

iii) $X|(Y, Z) = (y, z) \sim N_1(1 - y, 1)$

(a) Find the distribution of $\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$ and of $Y|(X, Z)$.

(b) Find the distribution of $\begin{pmatrix} U \\ V \end{pmatrix} = \begin{pmatrix} 1 + Z \\ 1 - Y \end{pmatrix}$.

(c) Compute $\mathrm{E}(Y|U = 2)$.

3 Estimation of the Mean Vector and Covariance Matrix of Multivariate Normal Distribution

3.1 Maximum Likelihood Estimation

3.1.1 Likelihood function

Suppose we have observed n independent realisations of p-dimensional random vectors from $N_p(\boldsymbol{\mu}, \Sigma)$. Suppose for simplicity that Σ is non-singular. The data matrix has the form

$$\boldsymbol{X} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1j} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2j} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{i1} & X_{i2} & \cdots & X_{ij} & \cdots & X_{in} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{p1} & X_{p2} & \cdots & X_{pj} & \cdots & X_{pn} \end{pmatrix} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n] \tag{3.1}$$

The goal to estimate the unknown mean vector and the covariance matrix of the multivariate normal distribution by the Maximum Likelihood Estimation (MLE) method.

Based on our knowledge from Lecture 2 we can write down the *Likelihood function*

$$L(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = (2\pi)^{-\frac{np}{2}} |\Sigma|^{-\frac{n}{2}} \, \mathrm{e}^{-\frac{1}{2}\sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})} \tag{3.2}$$

(Note that we have substituted the observations in (3.2) and consider L as a function of the unknown parameters $\boldsymbol{\mu}, \Sigma$ only.) Correspondingly, we get the *log-likelihood function* in the form

$$\log L(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = -\frac{np}{2} \log(2\pi) - \frac{n}{2} \log(|\Sigma|) - \frac{1}{2} \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) \tag{3.3}$$

It is well known that maximising either (3.2) or (3.3) will give the same solution for the MLE.

We start deriving the MLE by trying to maximise (3.3). To this end, first note that by

utilising properties of traces from Section 0.1.1, we can transform:

$$\sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})=\sum_{i=1}^{n}\mathrm{tr}[\Sigma^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^\top]=$$
$$\mathrm{tr}[\Sigma^{-1}(\sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^\top)]=$$

(by adding $\pm\bar{\boldsymbol{x}}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$ to each term $(\boldsymbol{x}_i-\boldsymbol{\mu})$ in $\sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^\top$)

$$\mathrm{tr}[\Sigma^{-1}(\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^\top+n(\bar{\boldsymbol{x}}-\boldsymbol{\mu})(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^\top)]$$
$$=\mathrm{tr}[\Sigma^{-1}(\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^\top)]+n(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^\top\Sigma^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}).$$

Thus

$$\log L(\boldsymbol{x};\boldsymbol{\mu},\Sigma)=-\frac{np}{2}\log(2\pi)-\frac{n}{2}\log(|\Sigma|)-\frac{1}{2}\mathrm{tr}[\Sigma^{-1}(\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^\top)]-\frac{1}{2}n(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^\top\Sigma^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}) \tag{3.4}$$

3.1.2 Maximum Likelihood Estimators

The MLE are the ones that maximise (3.4). Looking at (3.4) we realise that (since Σ is non-negative definite) the minimal value for $\frac{1}{2}n(\bar{\boldsymbol{x}}-\hat{\boldsymbol{\mu}})^\top\Sigma^{-1}(\bar{\boldsymbol{x}}-\hat{\boldsymbol{\mu}})$ is zero and is attained when $\hat{\boldsymbol{\mu}}=\bar{\boldsymbol{x}}$. It remains to find the optimal value for Σ. We will use the following

Theorem 3.1 (Anderson's lemma)**.** *If $A\in\mathcal{M}_{p,p}$ is symmetric positive definite, then the maximum of the function $h(G)=-n\log(|G|)-\mathrm{tr}(G^{-1}A)$ (defined over the set of symmetric positive definite matrices $G\in\mathcal{M}_{p,p}$) exists, occurs at $G=\frac{1}{n}A$ and has the maximal value of $np\log(n)-n\log(|A|)-np$.*

Proof. (sketch, details at lecture): Indeed, (see properties of traces):

$$\mathrm{tr}(G^{-1}A)=\mathrm{tr}((G^{-1}A^{\frac{1}{2}})A^{\frac{1}{2}})=\mathrm{tr}(A^{\frac{1}{2}}G^{-1}A^{\frac{1}{2}})$$

Let $\eta_i, i=1,\dots,p$ be the eigenvalues of $A^{\frac{1}{2}}G^{-1}A^{\frac{1}{2}}$. Then (since the matrix $A^{\frac{1}{2}}G^{-1}A^{\frac{1}{2}}$ is positive definite) $\eta_i>0, i=1,\dots,p$. Also, $\mathrm{tr}(A^{\frac{1}{2}}G^{-1}A^{\frac{1}{2}})=\sum_{i=1}^{p}\eta_i$ and $|A^{\frac{1}{2}}G^{-1}A^{\frac{1}{2}}|=\prod_{i=1}^{p}\eta_i$ holds. Hence

$$-n\log|G|-\mathrm{tr}(G^{-1}A)=n\sum_{i=1}^{p}\log\eta_i-n\log|A|-\sum_{i=1}^{p}\eta_i \tag{3.5}$$

Considering the expression $n\sum_{i=1}^{p}\log\eta_i-n\log|A|-\sum_{i=1}^{p}\eta_i$ as a function of the eigenvalues $\eta_i, i=1,\dots,p$ we realise that it has a **maximum** which is attained when all $\eta_i=n, i=1,\dots,p$. Indeed, the first partial derivatives with respect to $\eta_i, i=1,\dots,p$ are equal to $\frac{n}{\eta_i}-1$ and hence the stationary points are $\eta_i^*=n, i=1,\dots,p$. The matrix of second derivatives calculated at $\eta_i^*=n, i=1,\dots,p$ is equal to $-I_p$ and hence the stationary points give rise to a maximum of the function. Now, we can check directly by substituting the η^* values that the maximal value of the function is $np\log(n)-n\log(|A|)-np$. But a direct substitution in the formula $h(G)=-n\log(|G|)-\mathrm{tr}(G^{-1}A)$ with $G=\frac{1}{n}A$ also gives rise to $np\log(n)-n\log(|A|)-np$, i.e. the maximum is attained at $G=\frac{1}{n}A$. □

Using the structure of the log-likelihood function in (3.4) and Theorem 3.1 (applied for the case $A = \sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top$ (!)) it is now easy to formulate following:

Theorem 3.2. *Suppose $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ is a random sample from $N_p(\boldsymbol{\mu}, \Sigma), p < n$. Then $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}}$ and $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top$ are the* maximum likelihood estimators *of $\boldsymbol{\mu}$ and Σ, respectively.*

3.1.3 Alternative proofs

Alternative proofs of Theorem 3.2 are also available that utilise some formal rules for vector and matrix differentiation that have been developed as a standard machinery in multivariate analysis (recall that according to the folklore, in order to find the maximum of the log-likelihood, we need to differentiate it with respect to its arguments, i.e. with respect to the *vector* $\boldsymbol{\mu}$ and to the *matrix* Σ), set the derivatives equal to zero and solve the corresponding equation system. If time permits, these matrix differentiation rules will also be discussed later in this course.

3.1.4 Application in correlation matrix estimation

The correlation matrix can be defined in terms of the elements of the covariance matrix Σ. The correlation coefficients $\rho_{ij}, i = 1, \ldots, p, j = 1, \ldots, p$ are defined as $\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}}$ where $\Sigma = (\sigma_{ij}, i = 1 \ldots, p; j = 1, \ldots, p)$ is the covariance matrix. Note that $\rho_{ii} = 1, i = 1, \ldots, p$. To derive the *MLE* of $\rho_{ij}, i = 1, \ldots, p, j = 1, \ldots, p$ we note that these are continuous transformations of the covariances whose maximum likelihood estimators have already been derived. Then we can claim (*according to the transformation invariance properties of MLE*) that

$$\hat{\rho}_{ij} = \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}}\sqrt{\hat{\sigma}_{jj}}}, i = 1, \ldots, p, j = 1, \ldots, p \tag{3.6}$$

3.1.5 Sufficiency of $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$

Back from (3.4) we can write the likelihood function as

$$L(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{np}{2}}|\Sigma|^{\frac{n}{2}}} \mathrm{e}^{-\frac{1}{2}\,\mathrm{tr}[\Sigma^{-1}(\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top + n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^\top)]}$$

which means that $L(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)$ can be factorised into $L(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = g_1(\boldsymbol{x})g_2(\boldsymbol{\mu}, \Sigma; \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$, i.e. the likelihood function depends on the observations **only** through the values of $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}}$ and $\hat{\boldsymbol{\Sigma}}$. Hence the pair $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are *sufficient statistics* for $\boldsymbol{\mu}$ and Σ in the case of a sample from $N_p(\boldsymbol{\mu}, \Sigma)$. Note that the structure of the multivariate normal density was essentially used here thus underlying the importance of the check on adequacy of multivariate normality assumptions in practice. If testing indicates significant departures from multivariate normality then inferences that are based solely on $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ may not be very reliable.

3.2 Distributions of MLE of mean vector and covariance matrix of multivariate normal distribution

Inference is not restricted to only find point estimators but also to construct confidence regions, test hypotheses etc. To this end we need the distribution of the estimators (or of suitably chosen functions of them).

3.2.1 Sampling distribution of $\bar{\boldsymbol{X}}$

In the univariate case ($p = 1$) it is well known that for a sample of n observations from *normal distribution* $N(\mu, \sigma^2)$ the sample mean is normally distributed: $N(\mu, \frac{\sigma^2}{n})$. Moreover, the sample mean and the sample variance are *independent* in the case of sampling from a univariate normal population (Basu's Lemma). This fact was very useful in developing t-statistics for testing the mean vector. Do we have similar statements about the sample mean and sample variance in the multivariate ($p > 1$) case?

Let the random vector $\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \in \mathbb{R}^p$. For any $\boldsymbol{l} \in \mathbb{R}^p : \boldsymbol{l}^\top \bar{\boldsymbol{X}}$ is a linear combination of normals and hence is normal (see Definition 2.1). Since taking expected value is a linear operation, we have $\mathrm{E}\,\bar{\boldsymbol{X}} = \frac{1}{n} n\boldsymbol{\mu} = \boldsymbol{\mu}$; In analogy with the univariate case we could formally write $\mathrm{Cov}\,\bar{\boldsymbol{X}} = \frac{1}{n^2} n \,\mathrm{Cov}\,\boldsymbol{X}_1 = \frac{1}{n}\Sigma$ and hence $\bar{\boldsymbol{X}} \sim N_p(\boldsymbol{\mu}, \frac{1}{n}\Sigma)$. But we would like to develop a more appropriate machinery for the multivariate case that would help us to more rigorously prove statements like the last one. It is based on operations with *Kronecker products*.

Kronecker product of two matrices $A \in \mathcal{M}_{m,n}$ and $B \in \mathcal{M}_{p,q}$ is denoted by $A \otimes B$ and is defined (in block matrix notation) as

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{pmatrix} \tag{3.7}$$

The following basic properties of Kronecker products will be used:

$$\begin{aligned} (A \otimes B) \otimes C &= A \otimes (B \otimes C) \\ (A + B) \otimes C &= A \otimes C + B \otimes C \\ (A \otimes B)^\top &= A^\top \otimes B^\top \\ (A \otimes B)^{-1} &= A^{-1} \otimes B^{-1} \\ (A \otimes B)(C \otimes D) &= AC \otimes BD \end{aligned}$$

(whenever the corresponding matrix products and inverses exist)

$$\begin{aligned} \mathrm{tr}(A \otimes B) &= \mathrm{tr}(A)\,\mathrm{tr}(B) \\ |A \otimes B| &= |A|^p |B|^m \end{aligned}$$

(in case $A \in \mathcal{M}_{m,m}, B \in \mathcal{M}_{p,p}$).

In addition, the $\vec{\square}$ operation on a matrix $A \in \mathcal{M}_{m,n}$ will be defined. This operation creates a vector $\vec{A} \in \mathbb{R}^{mn}$ which is composed by stacking the n columns of the matrix $A \in \mathcal{M}_{m,n}$ under each other (the second below the first etc). For matrices A, B and C (of suitable dimensions) it holds:

$$\overrightarrow{ABC} = (C^\top \otimes A)\vec{B}$$

Let us see how we could utilise the above to derive the distribution of $\bar{\boldsymbol{X}}$. Denote by $\mathbf{1}_n$ the vector of n ones. Note that if $\boldsymbol{X}$ is the random data matrix (see (0.11) in Lecture 0.2) then $\vec{\boldsymbol{X}} \sim N(\mathbf{1}_n \otimes \boldsymbol{\mu}, I_n \otimes \Sigma)$ and $\bar{\boldsymbol{X}} = \frac{1}{n}(\mathbf{1}_n^\top \otimes I_p)\vec{\boldsymbol{X}}$. Hence $\bar{\boldsymbol{X}}$ is multivariate normal with

$$\mathrm{E}\,\bar{\boldsymbol{X}} = \frac{1}{n}(\mathbf{1}_n^\top \otimes I_p)(\mathbf{1}_n \otimes \boldsymbol{\mu}) = \frac{1}{n}(\mathbf{1}_n^\top \mathbf{1}_n \otimes \boldsymbol{\mu}) = \frac{1}{n} n\boldsymbol{\mu} = \boldsymbol{\mu},$$

$$\mathrm{Cov}\,\bar{\boldsymbol{X}} = n^{-2}(\mathbf{1}_n^\top \otimes I_p)(I_n \otimes \Sigma)(\mathbf{1}_n \otimes I_p) = n^{-2}(\mathbf{1}_n^\top \mathbf{1}_n \otimes \Sigma) = n^{-1}\Sigma.$$

Independence of $\bar{\boldsymbol{X}}$ and $\hat{\boldsymbol{\Sigma}}$

How can we show that $\bar{\boldsymbol{X}}$ and $\hat{\boldsymbol{\Sigma}}$ are independent? Recall the likelihood function

$$L(\boldsymbol{x};\boldsymbol{\mu},\Sigma)=\frac{1}{(2\pi)^{\frac{np}{2}}|\Sigma|^{\frac{n}{2}}}\,\mathrm{e}^{-\frac{1}{2}\,\mathrm{tr}[\Sigma^{-1}(\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^\top+n(\bar{\boldsymbol{x}}-\boldsymbol{\mu})(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^\top)]}$$

We have two summands in the exponent from which one is a function of the observations through $n\hat{\boldsymbol{\Sigma}}=\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^\top$ only and the other one depends on the observations through $\bar{\boldsymbol{x}}$ only. The idea is now to transform the original data matrix $\boldsymbol{X}\in\mathcal{M}_{p,n}$ into a new matrix $\boldsymbol{Z}\in\mathcal{M}_{p,n}$ whose columns are independent normal and in such a way that $\bar{\boldsymbol{X}}$ would only be a function of the first column $\boldsymbol{Z}_1$, whereas $\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^\top$ would only be a function of $\boldsymbol{Z}_2,\ldots,\boldsymbol{Z}_n$. If we succeed, then clearly $\bar{\boldsymbol{X}}$ and $\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^\top=n\hat{\boldsymbol{\Sigma}}$ would be independent.

Now the claim is that the sought after transformation is given by $\boldsymbol{Z}=\boldsymbol{X}A$ with $A\in\mathcal{M}_{n,n}$ being an *orthogonal matrix* with a first column equal to $\frac{1}{\sqrt{n}}\mathbf{1}_n$. Note that the first column of $\boldsymbol{Z}$ would be then $\sqrt{n}\bar{\boldsymbol{X}}$. (An explicit form of the matrix A can be obtained using the Gramm–Schmidt Process discussed later.) Since $\vec{\boldsymbol{Z}}=\overrightarrow{I_p\boldsymbol{X}A}=(A^\top\otimes I_p)\vec{\boldsymbol{X}}$, the Jacobian of the transformation ($\vec{\boldsymbol{X}}$ into $\vec{\boldsymbol{Z}}$) is $|A^\top\otimes I_p|=|A|^p=\pm1$ (note that A is orthogonal). Therefore, the absolute value of the Jacobian is equal to one. For $\vec{\boldsymbol{Z}}$ we have:

$$\mathrm{E}(\vec{\boldsymbol{Z}})=(A^\top\otimes I_p)(\mathbf{1}_n\otimes\boldsymbol{\mu})=A^\top\mathbf{1}_n\otimes\boldsymbol{\mu}=\begin{pmatrix}\sqrt{n}\\0\\\vdots\\0\end{pmatrix}\otimes\boldsymbol{\mu}$$

Further,

$$\mathrm{Cov}(\vec{\boldsymbol{Z}})=(A^\top\otimes I_p)(I_n\otimes\Sigma)(A\otimes I_p)=A^\top A\otimes I_p\Sigma I_p=I_n\otimes\Sigma$$

which means that the $\boldsymbol{Z}_i, i=1,\ldots,n$ are *independent*. Note $\boldsymbol{Z}_1=\sqrt{n}\bar{\boldsymbol{X}}$ holds (because of the choice of the first column of the orthogonal matrix A). Further

$$\sum_{i=1}^{n}(\boldsymbol{X}_i-\bar{\boldsymbol{X}})(\boldsymbol{X}_i-\bar{\boldsymbol{X}})^\top=\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^\top-\frac{1}{n}(\sum_{i=1}^{n}\boldsymbol{X}_i)(\sum_{i=1}^{n}\boldsymbol{X}_i^\top)=$$

$$\boldsymbol{Z}A^\top A\boldsymbol{Z}^\top-\boldsymbol{Z}_1\boldsymbol{Z}_1^\top=\sum_{i=1}^{n}\boldsymbol{Z}_i\boldsymbol{Z}_i^\top-\boldsymbol{Z}_1\boldsymbol{Z}_1^\top=\sum_{i=2}^{n}\boldsymbol{Z}_i\boldsymbol{Z}_i^\top$$

Hence we proved the following

Theorem 3.3. *For a sample of size n from $N_p(\boldsymbol{\mu},\Sigma), p<n$ the sample average $\bar{\boldsymbol{X}}\sim N_p(\mu,\frac{1}{n}\Sigma)$. Moreover, the MLE $\hat{\boldsymbol{\mu}}=\bar{\boldsymbol{X}}$ and $\hat{\boldsymbol{\Sigma}}$ are independent.*

3.2.2 Sampling distribution of the MLE of Σ

Definition 3.4. A random matrix $\boldsymbol{U}\in\mathcal{M}_{p,p}$ has a **Wishart distribution** with parameters Σ,p,n (denoting this by $\boldsymbol{U}\sim W_p(\Sigma,n)$) if there exist n independent random vectors $\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n$ each with $N_p(0,\Sigma)$ distribution such that the distribution of $\sum_{i=1}^{n}\boldsymbol{Y}_i\boldsymbol{Y}_i^\top$ coincides with the distribution of $\boldsymbol{U}$.

Note that we *require* that $p < n$ and that $\boldsymbol{U}$ be non-negative definite.

Having in mind the proof of Theorem 3.3 we can claim that the distribution of the matrix $n\hat{\boldsymbol{\Sigma}} = \sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\top$ is the same as that of $\sum_{i=2}^{n} \boldsymbol{Z}_i\boldsymbol{Z}_i^\top$ and therefore is Wishart with parameters $\Sigma, p, n-1$. That is, we can denote:

$$n\hat{\boldsymbol{\Sigma}} \sim W_p(\Sigma, n-1).$$

The density formula for the Wishart distribution is given in several sources but we will not deal with it in this course. Some properties of Wishart distribution will be mentioned though since we will make use of them later in the course:

1. If $p = 1$ and if we denote the "matrix" Σ by σ^2 (as usual) then $W_1(\Sigma, n)/\sigma^2 = \chi^2_n$. In particular, when $\sigma^2 = 1$ we see that $W_1(1, n)$ is exactly the χ^2_n random variable. In that sense we can state that the Wishart distribution is a generalisation (with respect to the dimension p) of the chi-squared distribution.

2. For an arbitrary fixed matrix $H \in \mathcal{M}_{k,p}, k \leq p$ one has:

$$nH\hat{\boldsymbol{\Sigma}}H^\top \sim W_k(H\Sigma H^\top, n-1).$$

   (Why? Show it!)

3. Refer to the previous case for the particular value of $k = 1$. The matrix $H \in \mathcal{M}_{1,p}$ is just a p-dimensional row vector that we could denote by $\boldsymbol{c}^\top$. Then:

   i) $n\frac{\boldsymbol{c}^\top\hat{\boldsymbol{\Sigma}}\boldsymbol{c}}{\boldsymbol{c}^\top\Sigma\boldsymbol{c}} \sim \chi^2_{n-1}$

   ii) $n\frac{\boldsymbol{c}^\top\Sigma^{-1}\boldsymbol{c}}{\boldsymbol{c}^\top\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{c}} \sim \chi^2_{n-p}$

4. Let us partition $\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\top \in \mathcal{M}_{p,p}$ into

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \boldsymbol{S}_{21} & \boldsymbol{S}_{22} \end{pmatrix}, \boldsymbol{S}_{11} \in \mathcal{M}_{r,r}, r < p$$
$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \Sigma_{11} \in \mathcal{M}_{r,r}, r < p.$$

   Further, denote

$$\boldsymbol{S}_{1|2} = \boldsymbol{S}_{11} - \boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{21},\ \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

   Then it holds

$$(n-1)\boldsymbol{S}_{11} \sim W_r(\Sigma_{11}, n-1)$$
$$(n-1)\boldsymbol{S}_{1|2} \sim W_r(\Sigma_{1|2}, n-p+r-1)$$

3.2.3 Aside: The Gramm–Schmidt Process (not examinable)

Let $A = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n] \in \mathcal{M}_{n,n}$ be an arbitrary full-rank matrix whose first column must be preserved (up to a constant multiple) but which must otherwise be made into an orthogonal matrix. The idea of the the Gram–Schmidt Orthogonalisation (and Orthonormalisation) is to first make $\boldsymbol{a}_2$ orthogonal to $\boldsymbol{a}_1$, then $\boldsymbol{a}_3$ orthogonal to $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$, all the way to making $\boldsymbol{a}_n$ orthogonal to all the previous vectors. This is accomplished by the following procedure:

1. For each $i = 2, \ldots, n$,
2. For each $j = 1, \ldots, i-1$,
3. Update $\boldsymbol{a}_i = \boldsymbol{a}_i - \frac{\langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle}{\langle \boldsymbol{a}_j, \boldsymbol{a}_j \rangle} \boldsymbol{a}_j$.
4. For each $k = 1, \ldots, n$,
5. Update $\boldsymbol{a}_k = \frac{\boldsymbol{a}_k}{\|\boldsymbol{a}_k\|}$.

Then, after Step 3 for a given i and j,

$$\langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle = \langle \boldsymbol{a}_i - \frac{\langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle}{\langle \boldsymbol{a}_j, \boldsymbol{a}_j \rangle} \boldsymbol{a}_j, \boldsymbol{a}_j \rangle = \langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle - \frac{\langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle}{\langle \boldsymbol{a}_j, \boldsymbol{a}_j \rangle} \langle \boldsymbol{a}_j, \boldsymbol{a}_j \rangle = 0.$$

We can use induction to show that by the time we reach Step 4, $\langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle = 0$ for any i and j. Observe that after Step 3 completes with $i = 2$ (and therefore $j = 1$ only), $\langle \boldsymbol{a}_1, \boldsymbol{a}_2 \rangle = 0$.

Now, suppose that $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{i-1}$ are orthogonal. Then, after Step 3 for some j, for an arbitrary $l < i$, $l \neq j$,

$$\langle \boldsymbol{a}_i - \frac{\langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle}{\langle \boldsymbol{a}_j, \boldsymbol{a}_j \rangle} \boldsymbol{a}_j, \boldsymbol{a}_l \rangle = \langle \boldsymbol{a}_i, \boldsymbol{a}_l \rangle - \frac{\langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle}{\langle \boldsymbol{a}_j, \boldsymbol{a}_j \rangle} \cancelto{0}{\langle \boldsymbol{a}_j, \boldsymbol{a}_l \rangle} = \langle \boldsymbol{a}_i, \boldsymbol{a}_l \rangle,$$

since $l, j \leq i - 1$ and are therefore orthogonal. This means that Step 3 only affects $\langle \boldsymbol{a}_i, \boldsymbol{a}_l \rangle$ for $l = j$: Step 3 cannot make $\boldsymbol{a}_i$ no longer orthogonal to any of the vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{i-1}$ to which it was previously orthogonal, so by the time the loop increments i, $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_i$ will be orthogonal, completing the proof by induction.

Lastly, Steps 4 and 5 simply ensure that $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ are normal. At no point is $\boldsymbol{a}_1$ changed except for being normalised.

Example 3.5. Gram–Schmidt process implemented in R.

3.3 Additional resources

An alternative presentation of these concepts can be found in JW Sec. 4.3–4.5.

3.4 Exercises

Exercise 3.1

Find the product $A \otimes B$ if $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, B = \begin{pmatrix} 5 & 0 \\ 2 & 1 \end{pmatrix}$.

4 Confidence Intervals and Hypothesis Tests for the Mean Vector

4.1 Hypothesis tests for the multivariate normal mean

4.1.1 Hotelling's T^2

Suppose again that, like in Lecture 3, we have observed n independent realisations of p-dimensional random vectors from $N_p(\boldsymbol{\mu}, \Sigma)$. Suppose for simplicity that Σ is non-singular. The data matrix has the form

$$\boldsymbol{x} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{in} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pj} & \cdots & x_{pn} \end{pmatrix} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]$$

Based on our knowledge from Section 3.2 we can claim that $\bar{\boldsymbol{X}} \sim N_p(\boldsymbol{\mu}, \frac{1}{n}\Sigma)$ and $n\hat{\boldsymbol{\Sigma}} \sim W_p(\Sigma, n-1)$.

Consequently, any linear combination $\boldsymbol{c}^\top \bar{\boldsymbol{X}}, \boldsymbol{c} \neq 0 \in \mathbb{R}^p$ follows $N(\boldsymbol{c}^\top \boldsymbol{\mu}, \frac{1}{n}\boldsymbol{c}^\top \Sigma \boldsymbol{c})$ and the quadratic form $n\boldsymbol{c}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{c} / \boldsymbol{c}^\top \Sigma \boldsymbol{c} \sim \chi^2_{n-1}$. Further, we have shown that $\bar{\boldsymbol{X}}$ and $\hat{\boldsymbol{\Sigma}}$ are independently distributed and hence

$$T = \sqrt{n}\boldsymbol{c}^\top (\bar{\boldsymbol{X}} - \boldsymbol{\mu}) / \sqrt{\boldsymbol{c}^\top \frac{n}{n-1} \hat{\boldsymbol{\Sigma}} \boldsymbol{c}} \sim t_{n-1},$$

i.e. follows the t distribution with $n-1$ degrees of freedom. This result has useful applications in testing for contrasts.

Indeed, if we would like to test $H_0 : \boldsymbol{c}^\top \boldsymbol{\mu} = \sum_{i=1}^p c_i \mu_i = 0$, we note that under H_0, T becomes simply

$$T = \sqrt{n}\boldsymbol{c}^\top \bar{\boldsymbol{X}} / \sqrt{\boldsymbol{c}^\top \boldsymbol{S} \boldsymbol{c}},$$

that is, does not involve the unknown $\boldsymbol{\mu}$ and can be used as a test-statistic whose distribution under H_0 is known. If $|T| > t_{1-\alpha/2,n-1}$ we should reject H_0 in favour of $H_1 : \boldsymbol{c}^\top \mu = \sum_{i=1}^p c_i \mu_i \neq 0$.

The formulation of the test for other (one-sided) alternatives is left for you as an exercise.

More often we are interested in testing the mean vector of a multivariate normal. First consider the case of *known* covariance matrix Σ (variance σ^2 in the univariate case). The standard univariate ($p = 1$) test for this purpose is the following: to test $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ at level of significance α, we look at $U = \sqrt{n}\frac{\bar{X}-\mu_0}{\sigma}$ and reject H_0 if $|U|$ exceeds the upper $\frac{\alpha}{2} \cdot 100\%$ point of the standard normal distribution. Checking if $|U|$ is large enough is equivalent to checking if $U^2 = n(\bar{X} - \mu_0)(\sigma^2)^{-1}(\bar{X} - \mu_0)$ is large enough. We can now easily generalise the above test statistic in a natural way for the multivariate ($p > 1$) case: calculate $U^2 = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^\top \Sigma^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)$ and reject the null hypothesis of $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ when U^2 is large enough. Similarly to the proof of *Property 5* of the multivariate normal distribution (Section 2.2) and by using Theorem 3.3 of Section 3.2 you can convince yourself (do it (!)) that $U^2 \sim \chi^2_p$ under the null hypothesis. Hence, tables of the χ^2-distribution will suffice to perform the above test in the multivariate case.

Now let us turn to the (practically more relevant) case of unknown covariance matrix Σ. The standard univariate ($p = 1$) test for this purpose is the t-test. Let us recall it: to test $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ at level of significance α, we look at

$$T = \sqrt{n}\frac{\bar{X} - \mu_0}{S},\ S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$$

and reject H_0 if $|T|$ exceeds the upper $\frac{\alpha}{2} \cdot 100\%$ point of the t-distribution with $n - 1$ degrees of freedom. We note that checking if $|T|$ is large enough is equivalent to checking if $T^2 = n(\bar{X} - \mu_0)(s^2)^{-1}(\bar{X} - \mu_0)$ is large enough. Of course, under H_0, the statistic T^2 is F-distributed: $T^2 \sim F_{1,n-1}$ which means that H_0 would be rejected at level α when $T^2 > F_{1-\alpha;1,n-1}$. We can now easily generalise the above test statistic in a natural way for the multivariate ($p > 1$) case:

Definition 4.1 (Hotelling's T^2)**.** The statistic

$$T^2 = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^\top \boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0) \tag{4.1}$$

where $\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i$, $\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^n (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\top$, $\boldsymbol{\mu}_0 \in \mathbb{R}^p$, $\boldsymbol{X}_i \in \mathbb{R}^p, i = 1, \ldots, n$ is named after Harold Hotelling.

4.1.2 Sampling distribution of T^2

Obviously, the test procedure based on Hotelling's statistic will reject the null hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ if the value of T^2 is sufficiently high. It turns out we do not need special tables for the distribution of T^2 under the null hypothesis because of the following basic result (that represents a true generalisation of the univariate ($p = 1$) case:

Theorem 4.2. *Under the null hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$, Hotelling's T^2 is distributed as $\frac{(n-1)p}{n-p}F_{p,n-p}$ where $F_{p,n-p}$ denotes the F-distribution with p and $n - p$ degrees of freedom.*

Proof. Indeed, we can write the T^2 statistic in the form:

$$T^2 = \frac{n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^\top \boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)}{n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^\top \Sigma^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)} n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^\top \Sigma^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0).$$

Denote by $\boldsymbol{C} = \sqrt{n}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)$. Conditionally on $\boldsymbol{C} = \boldsymbol{c}$ we have:

$$\frac{n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^\top \boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)}{n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^\top \Sigma^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)} = \frac{\boldsymbol{c}^\top \boldsymbol{S}^{-1}\boldsymbol{c}}{\boldsymbol{c}^\top \Sigma^{-1}\boldsymbol{c}},$$

has a distribution that only depends on the data through $\boldsymbol{S}^{-1}$. Noting that $n\hat{\boldsymbol{\Sigma}} = (n-1)\boldsymbol{S}$ and having in mind the third property of Wishart distributions from Section 3.2.2, we can claim that this distribution is the same as of $(n-1)/\chi^2_{n-p}$. Note also that the distribution does not depend on the particular $\boldsymbol{c}$. The second factor $n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)\Sigma^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0) \sim \chi^2_p$ and its distribution depends on the data through $\bar{\boldsymbol{X}}$ only. Because of the independence of the mean and covariance estimators, we have that the distribution of T^2 is the same as the distribution of $\frac{\chi^2_p(n-1)}{\chi^2_{n-p}}$ where the two chi-squares are *independent*. But this means that $\frac{T^2(n-p)}{p(n-1)} \sim F_{p,n-p}$ and hence $T^2 \sim \frac{p(n-1)}{n-p}F_{p,n-p}$. □

4.1.3 Noncentral Wishart

It is possible to extend the definition of the Wishart distribution in Section 3.2.2 by allowing the random vectors $\boldsymbol{Y}_i, i = 1, \ldots, n$ there to be independent with $N_p(\boldsymbol{\mu}_i, \Sigma)$ (instead of just having all $\boldsymbol{\mu}_i = \boldsymbol{0}$). One arrives at the *noncentral* Wishart distribution with parameters $\Sigma, p, n-1, \Gamma$ in that way (denoted also as $W_p(\Sigma, n-1, \Gamma)$. Here $\Gamma = MM^\top \in \mathcal{M}_{p,p}$, $M = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_n]$ is called a *noncentrality parameter*. When all columns of $M \in \mathcal{M}_{p,n}$ are zero, this is the usual (*central*) Wishart distribution. Theorem 4.2 can be extended to derive the distribution of the T^2 statistic under alternatives, i.e. the distribution of $T^2 = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^\top \boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})$ for $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. This distribution turns out to be related to *noncentral F-distribution*. It is helpful in studying power of the test of $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ versus $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. We shall spare the details here.

4.1.4 T^2 as a likelihood ratio statistic

It is worth mentioning that Hotelling's T^2 that we introduced by *analogy* with the univariate squared t statistic can in fact also be derived as the *likelihood ratio test statistic* for testing $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ versus $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. This safeguards the asymptotic optimality of the test suggested in Sections 4.1.1–4.1.2. To see this, first recall the likelihood function (3.2). Its unconstrained maximisation gives as a maximum value:

$$L(\boldsymbol{x}; \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = \frac{1}{(2\pi)^{\frac{np}{2}}|\hat{\boldsymbol{\Sigma}}|^{\frac{n}{2}}}\,\mathrm{e}^{-\frac{np}{2}}$$

On the other hand, under H_0 :

$$\max_{\Sigma} L(x; \boldsymbol{\mu}_0, \Sigma) = \max_{\Sigma} \frac{1}{(2\pi)^{\frac{np}{2}}|\Sigma|^{\frac{n}{2}}}\,\mathrm{e}^{-\frac{1}{2}\sum_{i=1}^n (\boldsymbol{x}_i - \boldsymbol{\mu}_0)^\top \Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_0)}$$

Since $\log L(\boldsymbol{x}; \boldsymbol{\mu}_0, \Sigma) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\,\mathrm{tr}[\Sigma^{-1}(\sum_{i=1}^n (\boldsymbol{x}_i - \boldsymbol{\mu}_0)(\boldsymbol{x}_i - \boldsymbol{\mu}_0)^\top)]$, on applying Anderson's lemma (see Theorem 3.1 in Section 3.1.2) we find that maximum of $\log L(\boldsymbol{x}; \boldsymbol{\mu}_0, \Sigma)$ (whence also of $L(\boldsymbol{x}; \boldsymbol{\mu}_0, \Sigma)$) is obtained when $\hat{\boldsymbol{\Sigma}}_0 = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i - \boldsymbol{\mu}_0)(\boldsymbol{x}_i - \boldsymbol{\mu}_0)^\top$ and the maximal value is

$$\frac{1}{(2\pi)^{\frac{np}{2}}|\hat{\boldsymbol{\Sigma}}_0|^{\frac{n}{2}}}\,\mathrm{e}^{-\frac{np}{2}}.$$

Hence the likelihood ratio is:

$$\Lambda = \frac{\max_\Sigma L(x; \boldsymbol{\mu}_0, \Sigma)}{\max_{\boldsymbol{\mu}, \Sigma} L(x; \boldsymbol{\mu}, \Sigma)} = \left(\frac{|\hat{\boldsymbol{\Sigma}}|}{|\hat{\boldsymbol{\Sigma}}_0|}\right)^{\frac{n}{2}} \tag{4.2}$$

The equivalent statistic $\Lambda^{\frac{2}{n}} = \frac{|\hat{\boldsymbol{\Sigma}}|}{|\hat{\boldsymbol{\Sigma}}_0|}$ is called *Wilks' lambda*. Small values of Wilks' lambda lead to rejecting $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$.

4.1.5 Wilks' lambda and T^2

The following theorem shows the relation between Wilks' lambda and T^2:

Theorem 4.3. *The likelihood ratio test is equivalent to the test based on T^2 since $\Lambda^{\frac{2}{n}} = (1 + \frac{T^2}{n-1})^{-1}$ holds.*

Proof. Consider the matrix $A \in \mathcal{M}_{p+1,p+1}$:

$$A = \begin{pmatrix} \sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top & \sqrt{n}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0) \\ \sqrt{n}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)^\top & -1 \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

It is easy to check that

$$|A| = |A_{22}||A_{11} - A_{12}A_{22}^{-1}A_{21}| = |A_{11}||A_{22} - A_{21}A_{11}^{-1}A_{12}| \tag{4.3}$$

holds from which we get:

$$(-1)|\sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top + n(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)^\top| =$$
$$|\sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top||-1 - n(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)^\top (\sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top)^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)|$$

Hence $(-1)|\sum_{i=1}^n (\boldsymbol{x}_i - \boldsymbol{\mu}_0)(\boldsymbol{x}_i - \boldsymbol{\mu}_0)^\top| = |\sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top|(-1)(1 + \frac{T^2}{n-1})$. Thus $|\hat{\boldsymbol{\Sigma}}_0| = |\hat{\boldsymbol{\Sigma}}|(1 + \frac{T^2}{n-1})$, i.e.

$$\Lambda^{\frac{2}{n}} = (1 + \frac{T^2}{n-1})^{-1} \tag{4.4}$$

□

4.1.6 Numerical calculation of T^2

Hence H_0 is rejected for small values of $\Lambda^{\frac{2}{n}}$ or equivalently, for large values of T^2. The critical values for T^2 are determined from Theorem 4.2. Relation (4.4) can be used to calculate T^2 from $\Lambda^{\frac{2}{n}} = \frac{|\hat{\boldsymbol{\Sigma}}|}{|\hat{\boldsymbol{\Sigma}}_0|}$ thus avoiding the need to invert the matrix $\boldsymbol{S}$ when calculating T^2!

4.1.7 Asymptotic distribution of T^2

Since $\boldsymbol{S}^{-1}$ is a consistent estimator of Σ^{-1}, the limiting distribution of T^2 will coincide with the one of $n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^\top \Sigma^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})$ which, as we know already, is χ_p^2. This coincides with a general claim of asymptotic theory which states that $-2\log\Lambda$ is asymptotically distributed as χ_p^2. Indeed:

$$-2\log\Lambda = n\log(1 + \frac{T^2}{n-1}) \approx \frac{n}{n-1}T^2 \approx T^2$$

(by using the fact that $\log(1 + x) \approx x$ for small x).

4.2 Confidence regions for the mean vector and for its components

4.2.1 Confidence region for the mean vector

For a given confidence level $(1-\alpha)$ it can be constructed in the form

$$\{\boldsymbol{\mu} | n(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^\top \boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}) \leq F_{1-\alpha,p,n-p}\frac{p}{n-p}(n-1)\}$$

where $F_{1-\alpha,p,n-p}$ is the upper $\alpha\cdot 100\%$ percentage point of the F distribution with $(p, n-p)$ df. This confidence region has the form of an *ellipsoid* in $\mathbb{R}^p$ centred at $\bar{\boldsymbol{x}}$. The axes of this *confidence ellipsoid* are directed along the eigenvectors $\boldsymbol{e}_i$ of the matrix $\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^\top$. The half-lengths of the axes are given by the expression $\sqrt{\lambda_i}\sqrt{\frac{p(n-1)F_{1-\alpha,p,n-p}}{n(n-p)}}$, with $\lambda_i, i=1,\ldots,p$ being the corresponding eigenvalues, i.e.

$$\boldsymbol{S}\boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i, i = 1,\ldots,p$$

Example 4.4. Microwave ovens (Example 5.3., pages 221–223, JW).

4.2.2 Simultaneous confidence statements

For a given confidence level $(1-\alpha)$ the confidence ellipsoids in Section 4.2.1 correctly reflect the joint (multivariate) knowledge about plausible values of $\boldsymbol{\mu} \in \mathbb{R}^p$ but nevertheless one is often interested in confidence intervals for means of each individual component. We would like to formulate these statements in such a form that *all of the separate confidence statements should hold* simultaneously with a prespecified probability. This is why we are speaking about *simultaneous confidence intervals*.

First, note that if the vector $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$ then for any $\boldsymbol{l} \in \mathbb{R}^p : \boldsymbol{l}^\top \boldsymbol{X} \sim N_1(\boldsymbol{l}^\top\boldsymbol{\mu}, \boldsymbol{l}^\top\Sigma\boldsymbol{l})$ and, hence, for any fixed $\boldsymbol{l}$ we can construct an $(1-\alpha)\cdot 100\%$ confidence interval of $\boldsymbol{l}^\top\boldsymbol{\mu}$ in the following simple way:

$$\left(\boldsymbol{l}^\top\bar{\boldsymbol{x}} - t_{1-\alpha/2,n-1}\frac{\sqrt{\boldsymbol{l}^\top\boldsymbol{S}\boldsymbol{l}}}{\sqrt{n}},\ \boldsymbol{l}^\top\bar{\boldsymbol{x}} + t_{1-\alpha/2,n-1}\frac{\sqrt{\boldsymbol{l}^\top\boldsymbol{S}\boldsymbol{l}}}{\sqrt{n}}\right) \tag{4.5}$$

By taking $\boldsymbol{l}^\top = [1,0,\ldots,0]$ or $\boldsymbol{l}^\top = [0,1,0,\ldots,0]$ etc. we obtain from (4.5) the usual confidence interval for each separate component of the mean. Note however that the confidence level for all these statements taken together is not $(1-\alpha)$. To make it $(1-\alpha)$ for all possible choices simultaneously we need to take a larger constant than $t_{1-\alpha/2,n-1}$ in the right hand side of the inequality $|\frac{\sqrt{n}(\boldsymbol{l}^\top\bar{\boldsymbol{x}}-\boldsymbol{l}^\top\bar{\boldsymbol{\mu}})}{\sqrt{\boldsymbol{l}^\top\boldsymbol{S}\boldsymbol{l}}}| \leq t_{1-\alpha/2,n-1}$ (or equivalently $\frac{n(\boldsymbol{l}^\top\bar{\boldsymbol{x}}-\boldsymbol{l}^\top\bar{\boldsymbol{\mu}})^2}{\boldsymbol{l}^\top\boldsymbol{S}\boldsymbol{l}} \leq t^2_{1-\alpha/2,n-1}$).

4.2.3 Simultaneous confidence ellipsoid

Theorem 4.5. *Simultaneously for all $\boldsymbol{l} \in \mathbb{R}^p$, the interval*

$$\left(\boldsymbol{l}^\top\bar{\boldsymbol{x}} - \sqrt{\frac{p(n-1)}{n(n-p)}F_{1-\alpha,p,n-p}\boldsymbol{l}^\top\boldsymbol{S}\boldsymbol{l}},\ \boldsymbol{l}^\top\bar{\boldsymbol{x}} + \sqrt{\frac{p(n-1)}{n(n-p)}F_{1-\alpha,p,n-p}\boldsymbol{l}^\top\boldsymbol{S}\boldsymbol{l}}\right)$$

will contain $\boldsymbol{l}^\top\bar{\boldsymbol{\mu}}$ with a probability at least $(1-\alpha)$.

Example 4.6. Microwave Ovens (Example 5.4, p. 226 in JW).

Proof. Note that according to Cauchy–Bunyakovski–Schwartz Inequality:

$$[\boldsymbol{l}^\top(\bar{\boldsymbol{x}}-\boldsymbol{\mu})]^2 = [(\boldsymbol{S}^{1/2}\boldsymbol{l})^\top \boldsymbol{S}^{-1/2}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})]^2 \le \|\boldsymbol{S}^{1/2}\boldsymbol{l}\|^2 \|\boldsymbol{S}^{-1/2}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})\|^2 = (\boldsymbol{l}^\top \boldsymbol{S}\boldsymbol{l})(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^\top \boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}).$$

Therefore,

$$\max_{\boldsymbol{l}} \frac{n(\boldsymbol{l}^\top(\bar{\boldsymbol{x}}-\boldsymbol{\mu}))^2}{\boldsymbol{l}^\top \boldsymbol{S}\boldsymbol{l}} \le n(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^\top \boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}) = T^2 \tag{4.6}$$

Inequality (4.6) helps us to claim that whenever a constant c has been such that $T^2 \le c^2$ then also $\frac{n(\boldsymbol{l}^\top \bar{\boldsymbol{x}} - \boldsymbol{l}^\top \bar{\boldsymbol{\mu}})^2}{\boldsymbol{l}^\top \boldsymbol{S}\boldsymbol{l}} \le c^2$ holds for any $\boldsymbol{l} \in \mathbb{R}^p, \boldsymbol{l} \ne 0$. Equivalently,

$$\boldsymbol{l}^\top \bar{\boldsymbol{x}} - c\sqrt{\frac{\boldsymbol{l}^\top \boldsymbol{S}\boldsymbol{l}}{n}} \le \boldsymbol{l}^\top \bar{\boldsymbol{\mu}} \le \boldsymbol{l}^\top \bar{\boldsymbol{x}} + c\sqrt{\frac{\boldsymbol{l}^\top \boldsymbol{S}\boldsymbol{l}}{n}} \tag{4.7}$$

for every $\boldsymbol{l}$. Now it remains to choose $c^2 = p(n-1)F_{1-\alpha,p,n-p}/(n-p)$ to make sure that $1-\alpha = P(T^2 \le c^2)$ holds and this will automatically ensure that (4.7) will contain $\boldsymbol{l}^\top \bar{\boldsymbol{\mu}}$ with probability $1-\alpha$. □

Bonferroni Method

The simultaneous confidence intervals when applied for the vectors $\boldsymbol{l}^\top = [1, 0, \ldots, 0], \boldsymbol{l}^\top = [0, 1, 0, \ldots, 0]$ etc. are much more reliable at a given confidence level than the one-at-a-time intervals. Note that the former also utilise the covariance structure of all p variables in their construction. However, sometimes we can do better in cases where one is interested in a small number of individual confidence statements.

In this latter case, the simultaneous confidence intervals may give too large a region and the Bonferroni method may prove more efficient instead. The idea of the Bonferroni approach is based on a simple probabilistic inequality. Assume that simultaneous confidence statements about m linear combinations $\boldsymbol{l}_1^\top \boldsymbol{\mu}, \boldsymbol{l}_2^\top \boldsymbol{\mu}, \ldots, \boldsymbol{l}_m^\top \boldsymbol{\mu}$ are required. If $C_i, i = 1, 2, \ldots, m$ denotes the ith confidence statement and $P(C_i \text{ true}) = 1 - \alpha_i$ then

$$P(\text{all } C_i \text{ true}) = 1 - P(\text{at least one } C_i \text{ false}) \ge$$
$$1 - \sum_{i=1}^{m} P(C_i \text{ false}) = 1 - \sum_{i=1}^{m}(1 - P(C_i \text{ true})) = 1 - (\alpha_1 + \alpha_2 + \cdots + \alpha_m)$$

Hence, if we choose $\alpha_i = \frac{\alpha}{m}, i = 1, 2, \ldots, m$ (that is, if calculate each statement at confidence level $(1 - \frac{\alpha}{m}) \cdot 100\%$ instead of $(1-\alpha)\cdot 100\%$) then the probability of any statement being false will not exceed α.

Example 4.7. Microwave Ovens (based on JW Example 5.4, p. 226).

4.3 Comparison of two or more mean vectors

Finally, let us note that comparison of the mean vectors of two or more than two different multivariate populations when there are independent observations from each of the populations is an important, practically relevant problem. For the purposes of this section, suppose that we observe two samples, $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_{n_{\boldsymbol{X}}} \in \mathbb{R}^p$ and $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_{n_{\boldsymbol{Y}}} \in \mathbb{R}^p$, with means $\boldsymbol{\mu}_{\boldsymbol{X}} \in \mathbb{R}^p$ and $\boldsymbol{\mu}_{\boldsymbol{Y}} \in \mathbb{R}^p$ respectively and variances $\Sigma_{\boldsymbol{X}} \in \mathcal{M}_{p,p}$ and $\Sigma_{\boldsymbol{Y}} \in \mathcal{M}_{p,p}$, respectively. Typically, we wish to test $H_0 : \boldsymbol{\mu}_{\boldsymbol{X}} - \boldsymbol{\mu}_{\boldsymbol{Y}} = \boldsymbol{\delta}_0$.

Multivariate ANOVA for comparing more than two populations is discussed in Lecture 8.

4.3.1 Reducing to a single population

As with the univariate t-test, under some scenarios the test of a difference between two populations in fact reduces to a one-sample test. For example, if the samples are paired and $n_{\boldsymbol{X}} = n_{\boldsymbol{Y}} = n$, we may proceed analogously to the paired t-test: we take $\boldsymbol{D}_i = \boldsymbol{X}_i - \boldsymbol{Y}_i$ for $i = 1, \ldots, n$ and proceed as if with a 1-sample T^2 test:

$$T^2 = n(\bar{\boldsymbol{D}} - \boldsymbol{\delta}_0)^\top \boldsymbol{S}_{\boldsymbol{D}}^{-1}(\bar{\boldsymbol{D}} - \boldsymbol{\delta}_0) \sim \frac{(n-1)p}{n-p} F_{p,n-p}, \tag{4.8}$$

where $\bar{\boldsymbol{D}} \in \mathbb{R}^p$ and $\boldsymbol{S}_{\boldsymbol{D}} \in \mathcal{M}_{p,p}$ are the sample mean and variance of $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_n$, respectively, assuming $\boldsymbol{D}_i$ are normally distributed. (It is important to note that any diagnostics for this test should be performed on the differences, not on the original values.)

We can also formulate this is in a "multivariate" form: let the *contrast matrix* $C \in \mathcal{M}_{p,p+p}$ be

$$C = \begin{pmatrix} +1 & & & -1 & & \\ & +1 & & & -1 & \\ & & +1 & & & -1 \end{pmatrix}.$$

Then, we can express $\boldsymbol{D}_i = C \begin{pmatrix} \boldsymbol{X}_i \\ \boldsymbol{Y}_i \end{pmatrix}$ and the test as $H_0 : C \begin{pmatrix} \boldsymbol{\mu}_{\boldsymbol{X}} \\ \boldsymbol{\mu}_{\boldsymbol{Y}} \end{pmatrix} = \boldsymbol{\delta}_0$. It is easy to show that the test statistic reduces to (4.8).

C can have more complex forms. For example, in a *repeated measures design*, we may measure the results of a series of p treatment outcomes on each sampling unit. If we then collect each individual i's measurements into a vector $\boldsymbol{X}_i$, we may test whether all outcomes are the same in expectation by forming

$$C = \begin{pmatrix} 1 & -1 & & \\ \vdots & & \ddots & \\ 1 & & & -1 \end{pmatrix} \in \mathcal{M}_{p-1,p}$$

and testing $H_0 : C\boldsymbol{\mu}_{\boldsymbol{X}} = \boldsymbol{0}_{p-1}$. It is easy to show that $C\boldsymbol{\mu}_{\boldsymbol{X}} = \boldsymbol{0}_{p-1}$ holds if and only if all elements of $\boldsymbol{\mu}_{\boldsymbol{X}}$ are equal.

4.3.2 The two-sample T^2-test

We now turn to the scenario where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are, in fact, independent samples. As with the univariate test, we must decide whether we are prepared to assume that $\Sigma_{\boldsymbol{X}} = \Sigma_{\boldsymbol{Y}} = \Sigma$ in the population and therefore use the pooled test. If so—and necessarily if the sample sizes are small—we evaluate

$$\boldsymbol{S}_{\text{pooled}} = \frac{(n_{\boldsymbol{X}} - 1)\boldsymbol{S}_{\boldsymbol{X}} + (n_{\boldsymbol{Y}} - 1)\boldsymbol{S}_{\boldsymbol{Y}}}{n_{\boldsymbol{X}} + n_{\boldsymbol{Y}} - 2}.$$

Since $\boldsymbol{S}_{\text{pooled}}$ estimates Σ,

$$\text{Var}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}) = \frac{\Sigma}{n_{\boldsymbol{X}}} + \frac{\Sigma}{n_{\boldsymbol{Y}}} \approx \frac{\boldsymbol{S}_{\text{pooled}}}{n_{\boldsymbol{X}}} + \frac{\boldsymbol{S}_{\text{pooled}}}{n_{\boldsymbol{Y}}} = \boldsymbol{S}_{\text{pooled}} \left(\frac{1}{n_{\boldsymbol{X}}} + \frac{1}{n_{\boldsymbol{Y}}} \right).$$

And, since $\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}} \sim N_p(\boldsymbol{\mu}_{\boldsymbol{X}} - \boldsymbol{\mu}_{\boldsymbol{Y}}, \Sigma(n_{\boldsymbol{X}}^{-1} + n_{\boldsymbol{Y}}^{-1}))$, we write

$$T^2 = (\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}} - \boldsymbol{\delta}_0)^\top \left\{ \boldsymbol{S}_{\text{pooled}} \left(\frac{1}{n_{\boldsymbol{X}}} + \frac{1}{n_{\boldsymbol{Y}}} \right) \right\}^{-1} (\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}} - \boldsymbol{\delta}_0) \sim \frac{(n_{\boldsymbol{X}} + n_{\boldsymbol{Y}} - 2)p}{n_{\boldsymbol{X}} + n_{\boldsymbol{Y}} - p - 1} F_{p, n_{\boldsymbol{X}} + n_{\boldsymbol{Y}} - p - 1}. \tag{4.9}$$

We would thus reject H_0 if T^2 falls above the F critical value in (4.9), construct a confidence region based on

$$\left\{\boldsymbol{\delta}\middle|(\bar{\boldsymbol{x}}-\bar{\boldsymbol{y}}-\boldsymbol{\delta})^\top \bar{\boldsymbol{S}}_{\mathrm{p}}^{-1}(\bar{\boldsymbol{x}}-\bar{\boldsymbol{y}}-\boldsymbol{\delta}) \le \frac{(n_{\boldsymbol{X}}+n_{\boldsymbol{Y}}-2)p}{n_{\boldsymbol{X}}+n_{\boldsymbol{Y}}-p-1}F_{1-\alpha,p,n_{\boldsymbol{X}}+n_{\boldsymbol{Y}}-p-1}\right\}$$

and simultaneous contrast confidence intervals

$$\boldsymbol{l}^\top(\bar{\boldsymbol{x}}-\bar{\boldsymbol{y}}) \pm \sqrt{\frac{(n_{\boldsymbol{X}}+n_{\boldsymbol{Y}}-2)p}{n_{\boldsymbol{X}}+n_{\boldsymbol{Y}}-p-1}F_{1-\alpha,p,n_{\boldsymbol{X}}+n_{\boldsymbol{Y}}-p-1}\boldsymbol{l}^\top \boldsymbol{S}_{\text{pooled}}\left(\frac{1}{n_{\boldsymbol{X}}}+\frac{1}{n_{\boldsymbol{Y}}}\right)\boldsymbol{l}}.$$

If we are not prepared to make the pooling assumption, our test statistic is instead

$$T^2 = (\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}}-\boldsymbol{\delta}_0)^\top\left(\frac{\boldsymbol{S}_{\boldsymbol{X}}}{n_{\boldsymbol{X}}}+\frac{\boldsymbol{S}_{\boldsymbol{Y}}}{n_{\boldsymbol{Y}}}\right)^{-1}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}}-\boldsymbol{\delta}_0).$$

Even for modest sample sizes, under multivariate normality, the distribution of this T^2 is reasonably well approximated by $\frac{\nu p}{\nu-p+1}F_{p,\nu-p+1}$, where

$$\nu = \frac{p+p^2}{\sum_{i=1}^2 \frac{1}{n_i}\left(\operatorname{tr}\left[\left\{\frac{1}{n_i}\boldsymbol{S}_i\left(\frac{1}{n_1}\boldsymbol{S}_1+\frac{1}{n_2}\boldsymbol{S}_2\right)^{-1}\right\}^2\right]+\left[\operatorname{tr}\left\{\frac{1}{n_i}\boldsymbol{S}_i\left(\frac{1}{n_1}\boldsymbol{S}_1+\frac{1}{n_2}\boldsymbol{S}_2\right)^{-1}\right\}\right]^2\right)}.$$

The confidence regions are then produced by

$$\left\{\boldsymbol{\delta}\middle|(\bar{\boldsymbol{x}}-\bar{\boldsymbol{y}}-\boldsymbol{\delta})^\top\left(\frac{\boldsymbol{S}_{\boldsymbol{X}}}{n_{\boldsymbol{X}}}+\frac{\boldsymbol{S}_{\boldsymbol{Y}}}{n_{\boldsymbol{Y}}}\right)^{-1}(\bar{\boldsymbol{x}}-\bar{\boldsymbol{y}}-\boldsymbol{\delta}) \le \frac{\nu p}{\nu-p+1}F_{p,\nu-p+1}\right\}$$

and simultaneous contrast confidence intervals

$$\boldsymbol{l}^\top(\bar{\boldsymbol{x}}-\bar{\boldsymbol{y}}) \pm \sqrt{\frac{\nu p}{\nu-p+1}F_{p,\nu-p+1}\boldsymbol{l}^\top\left(\frac{\boldsymbol{S}_{\boldsymbol{X}}}{n_{\boldsymbol{X}}}+\frac{\boldsymbol{S}_{\boldsymbol{Y}}}{n_{\boldsymbol{Y}}}\right)\boldsymbol{l}}.$$

4.4 Software

R: `car::confidenceEllipse`, package `Hotelling`, `rrcov::T2.test`, `ergm::approx.hotelling.diff.test`, `MVTests::TwoSamplesHT2`

SAS: See IML implementations.

4.5 Additional resources

An alternative presentation of these concepts can be found in JW Sec. 5.1–5.5 and 6.

4.6 Exercises

Exercise 4.1

Suppose $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ are independent $N_p(\boldsymbol{\mu}, \Sigma)$ random vectors with sample mean vector $\bar{\boldsymbol{X}}$ and sample covariance matrix $\boldsymbol{S}$. We wish to test the hypothesis

$$H_0 : \mu_2 - \mu_1 = \mu_3 - \mu_2 = \cdots = \mu_p - \mu_{p-1} = 1$$

where $\mu_1, \mu_2, \ldots, \mu_p$ are the elements of $\boldsymbol{\mu}$.

(a) Determine a $(p-1) \times p$ matrix C so that H_0 may be written equivalently as $H_0 : C\boldsymbol{\mu} = \mathbf{1}$ where $\mathbf{1}$ is a $(p-1) \times 1$ vector of ones.

(b) Make an appropriate transformation of the vectors $\boldsymbol{X}_i, i = 1, 2, \ldots, n$ and hence find the rejection region of a size α test of H_0 in terms of $\bar{\boldsymbol{X}}$, $\boldsymbol{S}$, and C.

Exercise 4.2

A sample of 50 vector observations, each containing three components, is drawn from a normal distribution having covariance matrix

$$\Sigma = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 4 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

The components of the sample mean are 0.8, 1.1 and 0.6. Can you reject the null hypothesis of zero distribution mean against a general alternative?

Exercise 4.3

Evaluate Hotelling's statistic T^2 for testing $H_0 : \boldsymbol{\mu} = \begin{pmatrix} 7 \\ 11 \end{pmatrix}$ using the data matrix $\boldsymbol{X} = \begin{pmatrix} 2 & 8 & 6 & 8 \\ 12 & 9 & 9 & 10 \end{pmatrix}$. Test the hypothesis H_0 at level $\alpha = 0.05$. What conclusion is reached?

Exercise 4.4

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_1}$, i.i.d. $N_p(\boldsymbol{\mu}_1, \Sigma)$ independently of $\boldsymbol{Y}_1, \ldots \boldsymbol{Y}_{n_2}$ i.i.d. $N_p(\boldsymbol{\mu}_2, \Sigma)$, Σ known. Prove that $\bar{\boldsymbol{X}} \sim N_p(\boldsymbol{\mu}_1, \frac{1}{n_1}\Sigma)$ and $\bar{\boldsymbol{Y}} \sim N_p(\boldsymbol{\mu}_2, \frac{1}{n_2}\Sigma)$. Hence $\boldsymbol{W} = \bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}} \sim N(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\Sigma)$ so that $\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}} - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \sim N(\mathbf{0}, \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\Sigma)$. Construct a test of $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$.

Exercise 4.5

Let $\bar{\boldsymbol{X}}$ and $\boldsymbol{S}$ be based on n observations from $N_p(\boldsymbol{\mu}, \Sigma)$ and let $\boldsymbol{X}$ be an additional observation from $N_p(\boldsymbol{\mu}, \Sigma)$. Show that $\boldsymbol{X} - \bar{\boldsymbol{X}} \sim N_p(0, (1 + \frac{1}{n})\Sigma)$. Find the distribution of $\frac{n}{n+1}(\boldsymbol{X} - \bar{\boldsymbol{X}})^\top S^{-1}(\boldsymbol{X} - \bar{\boldsymbol{X}})$ and suggest how to use this result to give a $(1-\alpha)$ prediction region for $\boldsymbol{X}$ based on $\bar{\boldsymbol{X}}$ and $\boldsymbol{S}$ (i.e., a region in $\mathbb{R}^p$ such that one has a given confidence $(1-\alpha)$ that the next observation will fall into it).

5 Correlation, Partial Correlation, Multiple Correlation

First of all, we would like to 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 an information about (or to predict) the length of the arm by measuring the length of the leg, we are dealing with 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 Lecture 2, 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 set of other variables.
- *Multiple correlation* between a single random variable and a set of p other variables

5.1 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 (3.6). It turned out that they coincide with the sample correlations r_{ij} we introduced in the first lecture (formula (1.3)).

To define *partial correlation coefficients*, recall the Property 4 of the multivariate normal distribution from Section 2.2:

> If vector $\boldsymbol{X} \in \mathbb{R}^p$ is divided into $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_{(1)} \\ \boldsymbol{X}_{(2)} \end{pmatrix}$, $\boldsymbol{X}_{(1)} \in R^r, r < p, \boldsymbol{X}_{(2)} \in R^{p-r}$ and according to this subdivision the vector means are $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_{(1)} \\ \boldsymbol{\mu}_{(2)} \end{pmatrix}$ 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
>
> $$N_r(\boldsymbol{\mu}_{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_{(2)} - \boldsymbol{\mu}_{(2)}), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}).$$

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.

$$\rho_{ij.(r+1),(r+2),\ldots,p} = \frac{\sigma_{ij.(r+1),(r+2),\ldots,p}}{\sqrt{\sigma_{ii.(r+1),(r+2),\ldots,p}}\sqrt{\sigma_{jj.(r+1),(r+2),\ldots,p}}}. \tag{5.1}$$

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 transformation 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 $\rho_{ij.(r+1),(r+2)\ldots,p}, i, j = 1, 2, \ldots, r$.

5.1.1 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:

i) partial correlation between first and second variable by adjusting for the effect of the third:

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

ii) 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.

5.1.2 Software

SAS: `PROC CORR`

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

5.1.3 Examples

Example 5.1. Three variables have been measured for a set of schoolchildren:

i) X_1: Intelligence

ii) X_2: Weight

iii) X_3: Age

The number of observations was large enough so that one can assume the empirical correlation matrix $\hat{\rho} \in \mathcal{M}_{3,3}$ to be the true correlation matrix: $\hat{\rho} = \begin{pmatrix} 1 & 0.6162 & 0.8267 \\ 0.6162 & 1 & 0.7321 \\ 0.8267 & 0.7321 & 1 \end{pmatrix}$. This suggests there is a high degree of positive dependence between weight and intelligence. But (**do the calculation (!)**) $\hat{\rho}_{12.3} = 0.0286$ so that, after the effect of age is adjusted for, there is virtually no correlation between weight and intelligence, i.e. weight obviously plays little part in explaining intelligence.

5.2 Multiple correlation

Recall our discussion in the end of Section 2.2 for the best prediction in mean squares sense in case of multivariate normality: If we want to predict a random variable Y that is correlated with p random variables (predictors) $\boldsymbol{X} = \begin{pmatrix} X_1 & X_2 & \cdots & X_p \end{pmatrix}^\top$ by trying to minimise the expected value $\mathrm{E}[\{Y - g(\boldsymbol{X})\}^2 | \boldsymbol{X} = \boldsymbol{x}]$ the optimal solution (i.e. the regression function) was $g^*(\boldsymbol{X}) = \mathrm{E}(Y|\boldsymbol{X})$. When the joint $(p+1)$-dimensional distribution of Y and $\boldsymbol{X}$ is **normal** this function was **linear** in $\boldsymbol{X}$. Given a specific realisation $\boldsymbol{x}$ of $\boldsymbol{X}$ it was given by $b + \boldsymbol{\sigma}_0^\top C^{-1}\boldsymbol{x}$ where $b = \mathrm{E}(Y) - \boldsymbol{\sigma}_0^\top C^{-1}\,\mathrm{E}(\boldsymbol{X})$, C is the covariance matrix of the vector $\boldsymbol{X}$, $\boldsymbol{\sigma}_0$ is the vector of Covariances of Y with $X_i, i = 1, \ldots, p$. The vector $C^{-1}\boldsymbol{\sigma}_0 \in \mathbb{R}^p$ was the *vector of the regression coefficients*.

Now, let us **define** the multiple correlation coefficient between the random variable Y and the random vector $\boldsymbol{X} \in \mathbb{R}^p$ to be the maximum correlation between Y and *any linear combination* $\boldsymbol{\alpha}^\top \boldsymbol{X}$, $\boldsymbol{\alpha} \in \mathbb{R}^p$. This makes sense: to look at the maximal correlation that we can get by trying to predict Y as a linear function of the predictors. The solution to this which also gives us an algorithm to calculate (and estimate) the multiple correlation coefficient is given in the next lemma.

5.2.1 Multiple correlation coefficient as ordinary correlation coefficient of transformed data

Lemma 5.2. *The multiple correlation coefficient is the ordinary correlation coefficient between* Y *and* $\boldsymbol{\sigma}_0^\top C^{-1}\boldsymbol{X} \equiv \boldsymbol{\beta}^{*\top}\boldsymbol{X}$. *(I.e.,* $\boldsymbol{\beta}^* \equiv C^{-1}\boldsymbol{\sigma}_0$.*)*

Proof. Note that for any $\boldsymbol{\alpha} \in \mathbb{R}^p : \mathrm{Cov}(Y, \boldsymbol{\alpha}^\top\boldsymbol{X}) = \boldsymbol{\alpha}^\top C\boldsymbol{\beta}^*$ and, in particular, $\mathrm{Cov}(Y, \boldsymbol{\beta}^{*\top}\boldsymbol{X}) = \boldsymbol{\beta}^{*\top} C\boldsymbol{\beta}^*$ holds.

Using *Cauchy–Bunyakovsky–Schwartz* inequality we have:

$$[\mathrm{Cov}(\boldsymbol{\alpha}^\top\boldsymbol{X}, \boldsymbol{\beta}^{*\top}\boldsymbol{X})]^2 \le \mathrm{Var}(\boldsymbol{\alpha}^\top\boldsymbol{X})\,\mathrm{Var}(\boldsymbol{\beta}^{*\top}\boldsymbol{X})$$

and therefore:

$$\sigma_Y^2\rho^2(Y, \boldsymbol{\alpha}^\top\boldsymbol{X}) = \frac{(\boldsymbol{\alpha}^\top\boldsymbol{\sigma}_0)^2}{\boldsymbol{\alpha}^\top C\boldsymbol{\alpha}} = \frac{(\boldsymbol{\alpha}^\top C\boldsymbol{\beta}^*)^2}{\boldsymbol{\alpha}^\top C\boldsymbol{\alpha}} \le \boldsymbol{\beta}^{*\top} C\boldsymbol{\beta}^*$$

holds, σ_Y^2 denoting the variance of Y. In this last equality we can get the equality sign by choosing $\boldsymbol{\alpha} = \boldsymbol{\beta}^*$, i.e. the squared correlation coefficient $\rho^2(Y, \boldsymbol{\alpha}^\top\boldsymbol{X})$ of Y and $\boldsymbol{\alpha}^\top\boldsymbol{X}$ is maximised over $\boldsymbol{\alpha}$ when $\boldsymbol{\alpha} = \boldsymbol{\beta}^*$. □

Coefficient of Determination

From Lemma 5.2 we see that the maximum correlation between Y and any linear combination $\boldsymbol{\alpha}^\top\boldsymbol{X}$, $\boldsymbol{\alpha} \in \mathbb{R}^p$, is $R = \sqrt{\frac{\boldsymbol{\beta}^{*\top} C\boldsymbol{\beta}^*}{\sigma_Y^2}}$. This is the multiple correlation coefficient. Its square R^2 is called *coefficient of determination*. Having in mind that $\boldsymbol{\beta}^* = C^{-1}\boldsymbol{\sigma}_0$ we see that $R = \sqrt{\frac{\boldsymbol{\sigma}_0^\top C^{-1}\boldsymbol{\sigma}_0}{\sigma_Y^2}}$. If $\Sigma = \begin{pmatrix} \sigma_Y^2 & \boldsymbol{\sigma}_0^\top \\ \boldsymbol{\sigma}_0 & C \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ is the partitioned covariance matrix of the $(p+1)$-dimensional vector $(Y, \boldsymbol{X})^\top$ then we know how to calculate the MLE of Σ by $\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} \hat{\boldsymbol{\Sigma}}_{11} & \hat{\boldsymbol{\Sigma}}_{12} \\ \hat{\boldsymbol{\Sigma}}_{21} & \hat{\boldsymbol{\Sigma}}_{22} \end{pmatrix}$ so the MLE of R would be $\hat{R} = \sqrt{\frac{\hat{\boldsymbol{\Sigma}}_{12}\hat{\boldsymbol{\Sigma}}_{22}^{-1}\hat{\boldsymbol{\Sigma}}_{21}}{\hat{\boldsymbol{\Sigma}}_{11}}}$.

5.2.2 Interpretation of R

At the end of Section 2.2 we derived the minimal value of the mean squared error when trying to predict Y by a linear function of the vector $\boldsymbol{X}$. It is achieved when using the regression function and the value itself was $\sigma_Y^2 - \boldsymbol{\sigma}_0^\top C^{-1}\boldsymbol{\sigma}_0$. The latter value can also be expressed by using the value of R. It is equal to $\sigma_Y^2(1 - R^2)$. Thus, our conclusion is that when $R^2 = 0$ there is no predictive power at all. In the opposite extreme case, if $R^2 = 1$, it turns out that Y can be predicted without any error at all (it is a true linear function of $\boldsymbol{X}$).

5.2.3 Remark about the calculation of R^2

Sometimes, the *correlation matrix only* may be available. It can be shown that in that case the relation

$$1 - R^2 = \frac{1}{\rho^{YY}} \tag{5.2}$$

holds. In (5.2), $\rho^{YY} \equiv (\boldsymbol{\rho}^{-1})_{11}$ is the upper left-hand corner of the inverse of the *correlation matrix* $\boldsymbol{\rho} \in \mathcal{M}_{p+1,p+1}$ determined from Σ. We note that the relation $\boldsymbol{\rho} = V^{-\frac{1}{2}}\Sigma V^{-\frac{1}{2}}$ holds with

$$V = \begin{pmatrix} \sigma_y^2 & 0 & \cdots & 0 \\ 0 & c_{11} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & c_{pp} \end{pmatrix}$$

One can use (5.2) to calculate R^2 by first calculating the right hand side in (5.2). To show Equality (5.2) we note that

$$1 - R^2 = \frac{\sigma_Y^2 - \boldsymbol{\sigma}_0^\top C^{-1}\boldsymbol{\sigma}_0}{\sigma_Y^2} = \frac{|C|}{|C|}\frac{\sigma_Y^2 - \boldsymbol{\sigma}_0^\top C^{-1}\boldsymbol{\sigma}_0}{\sigma_Y^2} = \frac{|\Sigma|}{|C|\sigma_Y^2},$$

with the last equality in the numerator holding because of (4.3). But $\frac{|C|}{|\Sigma|} = \sigma^{YY} \equiv (\Sigma^{-1})_{11}$, the entry in the first row and column of Σ^{-1}. (Recall from Section 0.1.2: $(X^{-1})_{ji} = \frac{|X_{ij}|}{|X|}(-1)^{i+j}$.) Since $\boldsymbol{\rho}^{-1} = V^{\frac{1}{2}}\Sigma^{-1}V^{\frac{1}{2}}$, we see that $\rho^{YY} = \sigma^{YY}\sigma_Y^2$ holds. Therefore $1 - R^2 = \frac{1}{\rho^{YY}}$.

5.2.4 Examples

Example 5.3. Let $\mu = \begin{pmatrix} \mu_Y \\ \mu_{X_1} \\ \mu_{X_2} \end{pmatrix} = \begin{pmatrix} 5 \\ 2 \\ 0 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 10 & 1 & -1 \\ 1 & 7 & 3 \\ -1 & 3 & 2 \end{pmatrix} = \begin{pmatrix} \sigma_{YY} & \boldsymbol{\sigma}_0^\top \\ \boldsymbol{\sigma}_0 & \Sigma_{XX} \end{pmatrix}$. Calculate:

(a) The best linear prediction of Y using X_1 and X_2.

(b) The multiple correlation coefficient $R^2_{Y.(X_1,X_2)}$.

(c) The mean squared error of the best linear predictor.

Solution

$$\boldsymbol{\beta}^* = \Sigma_{XX}^{-1}\boldsymbol{\sigma}_0 = \begin{pmatrix} 7 & 3 \\ 3 & 2 \end{pmatrix}^{-1}\begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} .4 & -.6 \\ -.6 & 1.4 \end{pmatrix}\begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 1 \\ -2 \end{pmatrix}$$

and

$$b = \mu_Y - \boldsymbol{\beta}^{*\top}\mu_X = 5 - (1,-2)\begin{pmatrix} 2 \\ 0 \end{pmatrix} = 3.$$

Hence the best linear predictor is given by $3 + X_1 - 2X_2$. The value of:

$$R_{Y.(X_1,X_2)} = \sqrt{\frac{(1,-1)\begin{pmatrix} .4 & -.6 \\ -.6 & 1.4 \end{pmatrix}\begin{pmatrix} 1 \\ -1 \end{pmatrix}}{10}} = \sqrt{\frac{3}{10}} = .548$$

The mean squared error of prediction is: $\sigma_Y^2(1 - R^2_{Y.(X_1,X_2)}) = 10(1 - \frac{3}{10}) = 7$.

Example 5.4. Relationship between multiple correlation and regression, and equivalent ways of computing it.

5.3 Testing of correlation coefficients

5.3.1 Usual correlation coefficients

When considering the distribution of a particular correlation coefficient $\hat{\rho}_{ij} = r_{ij}$ the problem becomes bivariate because only the variables X_i and X_j are involved. Direct transformations with the bivariate normal can be utilised to derive the **exact** distribution of r_{ij} under the hypothesis $H_0 : \rho_{ij} = 0$. It turns out that in this case the statistic $T = r_{ij}\sqrt{\frac{n-2}{1-r_{ij}^2}} \sim t_{n-2}$ and tests can be performed by using the t-distribution. For other hypothetical values the derivations are more painful. There is one most frequently used **approximation** that holds no matter what the true value of ρ_{ij} is. We shall discuss it here. Consider **Fisher's Z transformation** $Z = \frac{1}{2}\log[\frac{1+r_{ij}}{1-r_{ij}}]$. Under the hypothesis $H_0 : \rho_{ij} = \rho_0$ it holds:

$$Z \approx N(\frac{1}{2}\log[\frac{1+\rho_0}{1-\rho_0}], \frac{1}{n-3})$$

In particular, in the most common situation, when one would like to test $H_0 : \rho_{ij} = 0$ versus $H_1 : \rho_{ij} \neq 0$ one would reject H_0 at 5% significance level if $|Z|\sqrt{n-3} \geq 1.96$.

Based on the above, now you suggest how to test the hypothesis of equality of two correlation coefficients from two different populations(!).

5.3.2 Partial correlation coefficients

Coming over to testing *partial correlations*, not much has to be changed. Fisher's Z approximation can be used again in the following way: to test $H_0 : \rho_{ij.r+1,r+2,\ldots,r+k} = \rho_0$ versus $H_1 : \rho_{ij.r+1,r+2,\ldots,r+k} \neq \rho_0$ (i.e., conditioning on k variables) we construct $Z = \frac{1}{2}\log[\frac{1+r_{ij.r+1,r+2,\ldots,r+k}}{1-r_{ij.r+1,r+2,\ldots,r+k}}]$ and $a = \frac{1}{2}\log[\frac{1+\rho_0}{1-\rho_0}]$. Asymptotically $Z \sim N(a, \frac{1}{n-k-3})$ holds. Hence, test statistic to be compared with significance points of the standard normal is now : $\sqrt{n-k-3}|Z-a|$. If $\rho_0 = 0$, the t-test can be used, with "$n-2$" replaced by "$n-k-2$" in both the statistic and the degrees of freedom.

5.3.3 Multiple correlation coefficients

It turns out that under the hypothesis $H_0 : R = 0$ the statistic $F = \frac{\hat{R}^2}{1-\hat{R}^2} \times \frac{n-p}{p-1} \sim F_{p-1,n-p}$. Hence, when testing significance of the multiple correlation, the rejection region would be $\{\frac{\hat{R}^2}{1-\hat{R}^2} \times \frac{n-p}{p-1} > F_{1-\alpha,p-1,n-p}\}$ for a given significance level α.

It should be stressed that the value of p in Section 5.3.3 refers to the **total** number of all variables (the output Y and all of the input variables in the input vector $\boldsymbol{X}$). This is different from the value of p that was used in Section 5.2. In other words, the p in Section 5.3.3 is the $p+1$ in Section 5.2.

5.3.4 Software

SAS: `PROC CORR`

R: `ggm::pcor.test`

5.3.5 Examples

Example 5.5. Testing ordinary correlations: age, height, and intelligence.

Example 5.6. Testing partial correlations: age, height, and intelligence.

5.4 Additional resources

An alternative presentation of these concepts can be found in JW Sec. 7.8.

5.5 Exercises

Exercise 5.1

Suppose $\boldsymbol{X} \sim N_4(\mu, \Sigma)$ where $\mu = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 3 & 1 & 0 & 1 \\ 1 & 4 & 0 & 0 \\ 0 & 0 & 1 & 4 \\ 1 & 0 & 4 & 20 \end{pmatrix}$. Determine:

(a) the distribution of $\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_1 + X_2 + X_4 \end{pmatrix}$;

(b) the conditional mean and variance of X_1 given x_2, x_3, and x_4;

(c) the partial correlation coefficients $\rho_{12.3}, \rho_{12.4}$;

(d) the multiple correlation between X_1 and (X_2, X_3, X_4). Compare it to ρ_{12} and comment.

(e) Justify that $\begin{pmatrix} X_2 \\ X_3 \\ X_4 \end{pmatrix}$ is independent of $X_1 - (1 \quad 0 \quad 1) \begin{pmatrix} 4 & 0 & 0 \\ 0 & 1 & 4 \\ 0 & 4 & 20 \end{pmatrix}^{-1} \begin{pmatrix} X_2 \\ X_3 \\ X_4 \end{pmatrix}$.

Exercise 5.2

A random vector $\boldsymbol{X} \sim N_3(\boldsymbol{\mu}, \Sigma)$ with $\boldsymbol{\mu} = \begin{pmatrix} 2 \\ -3 \\ 1 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 2 \end{pmatrix}$.

(a) Find the distribution of $3X_1 - 2X_2 + X_3$.

(b) Find a vector $\boldsymbol{a} \in \mathbb{R}^2$ such that X_2 and $X_2 - \boldsymbol{a}^\top \begin{pmatrix} X_1 \\ X_3 \end{pmatrix}$ are independent.

6 Principal Components Analysis

6.1 Introduction

Principal components analysis is applied mainly as a **variable reduction procedure**. It is usually applied in cases when data is obtained from a possibly **large number** of variables which are possibly **highly correlated**. The goal is to try to "condense" the information. This is done by summarising the data in a (small) number of transformations of the original variables. Our motivation to do that is that we believe there is some redundancy in the presentation of the information by the original set of variables since e.g. many of these variables are measuring the same construct. In that case we try to reduce the observed variables into a smaller number of **principal components** (artificial variables) that would account for most of the variability in the observed variables. For simplicity, these artificial new variables are presented as a **linear combinations** of the (**optimally weighted**) observed variables. If one linear combination is not enough, we can choose to construct two, three, etc. such combinations. Note also that principal components analysis may be just an intermediate step in much larger investigations. The principal components obtained can be used for example as inputs in a regression analysis or in a cluster analysis procedure. They are also a basic method in extracting factors in factor analysis.

6.2 Precise mathematical formulation

Let $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$ where p is assumed to be relatively large. To perform a reduction, we are looking for a linear combination $\boldsymbol{\alpha}_1^\top \boldsymbol{X}$ with $\boldsymbol{\alpha}_1 \in \mathbb{R}^p$ suitably chosen such that it maximises the variance of $\boldsymbol{\alpha}_1^\top \boldsymbol{X}$ subject to the reasonable normalising constraint $\|\boldsymbol{\alpha}_1\|^2 = \boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1 = 1$. Since $\mathrm{Var}(\boldsymbol{\alpha}_1^\top \boldsymbol{X}) = \boldsymbol{\alpha}_1^\top \Sigma \boldsymbol{\alpha}_1$ we need to choose $\boldsymbol{\alpha}_1$ to maximise $\boldsymbol{\alpha}_1^\top \Sigma \boldsymbol{\alpha}_1$ subject to $\boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1 = 1$.

This requires us to apply Lagrange's optimisation under constraint procedure:

1. construct the Lagrangian function

$$\mathrm{Lag}(\boldsymbol{\alpha}_1, \lambda) = \boldsymbol{\alpha}_1^\top \Sigma \boldsymbol{\alpha}_1 + \lambda(1 - \boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1)$$

   where $\lambda \in \mathbb{R}^1$ is the Lagrange multiplier;

2. take the partial derivative with respect to $\boldsymbol{\alpha}_1$ and equate it to zero:

$$2\Sigma\boldsymbol{\alpha}_1 - 2\lambda\boldsymbol{\alpha}_1 = \mathbf{0} \implies (\Sigma - \lambda I_p)\boldsymbol{\alpha}_1 = \mathbf{0}. \tag{6.1}$$

From (6.1), we see that $\boldsymbol{\alpha}_1$ must be an eigenvector of Σ and since we know from Example 0.2 what the maximal value of $\frac{\boldsymbol{\alpha}^\top \Sigma \boldsymbol{\alpha}}{\boldsymbol{\alpha}^\top \boldsymbol{\alpha}}$ is, we conclude that $\boldsymbol{\alpha}_1$ should be the **eigenvector that**

corresponds to the largest eigenvalue $\bar{\lambda}_1$ of Σ. The random variable $\boldsymbol{\alpha}_1^\top \boldsymbol{X}$ is called the **first principal component**.

For the **second** principal component $\boldsymbol{\alpha}_2^\top \boldsymbol{X}$ we want it to be normalised according to $\boldsymbol{\alpha}_2^\top \boldsymbol{\alpha}_2 = 1$, uncorrelated with the first component and to give maximal variance of a linear combination of the components of $\boldsymbol{X}$ under these constraints. To find it, we construct the Lagrange function:

$$\text{Lag}_1(\boldsymbol{\alpha}_2, \lambda_1, \lambda_2) = \boldsymbol{\alpha}_2^\top \Sigma \boldsymbol{\alpha}_2 + \lambda_1(1 - \boldsymbol{\alpha}_2^\top \boldsymbol{\alpha}_2) + \lambda_2 \boldsymbol{\alpha}_1^\top \Sigma \boldsymbol{\alpha}_2$$

Its partial derivative w.r.t. $\boldsymbol{\alpha}_2$ gives

$$2\Sigma\boldsymbol{\alpha}_2 - 2\lambda_1\boldsymbol{\alpha}_2 + \lambda_2\Sigma\boldsymbol{\alpha}_1 = \mathbf{0} \tag{6.2}$$

Multiplying (6.2) by $\boldsymbol{\alpha}_1^\top$ from left and using the two constraints $\boldsymbol{\alpha}_2^\top \boldsymbol{\alpha}_2 = 1$ and $\boldsymbol{\alpha}_2^\top \Sigma \boldsymbol{\alpha}_1 = 0$ gives:

$$-2\lambda_1 \boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_2 + \lambda_2 \boldsymbol{\alpha}_1^\top \Sigma \boldsymbol{\alpha}_1 = 0 \implies \lambda_2 = 0$$

(WHY? Have in mind that $\boldsymbol{\alpha}_1$ was an eigenvector of Σ.) But then (6.2) also implies that $\boldsymbol{\alpha}_2 \in \mathbb{R}^p$ must be an eigenvector of Σ (has to satisfy $(\Sigma - \lambda_1 I_p)\boldsymbol{\alpha}_2 = \mathbf{0}$). Since it has to be different from $\boldsymbol{\alpha}_1$, having in mind that we aim at variance maximisation, we see that $\boldsymbol{\alpha}_2$ has to be the normalised eigenvector that corresponds to the second largest eigenvalue $\bar{\lambda}_2$ of Σ. The process can be continued further. The third principal component should be uncorrelated with the first two, should be normalised and should give maximal variance of a linear combination of the components of $\boldsymbol{X}$ under these constraints. One can easily realise then that the vector $\boldsymbol{\alpha}_3 \in \mathbb{R}^p$ in the formula $\boldsymbol{\alpha}_3^\top \boldsymbol{X}$ should be the normalised eigenvector that corresponds to the third largest eigenvalue $\bar{\lambda}_3$ of the matrix Σ etc..

Note that if we extract **all possible** p principal components then $\sum_{i=1}^p \text{Var}(\boldsymbol{\alpha}_i^\top \boldsymbol{X})$ will just equal the sum of all eigenvalues of Σ and hence

$$\sum_{i=1}^{p} \text{Var}(\boldsymbol{\alpha}_i^\top \boldsymbol{X}) = \text{tr}(\Sigma) = \Sigma_{11} + \cdots + \Sigma_{pp}.$$

Therefore, if we only take a small number of k principal components instead of the total possible number p we can interpret their inclusion as one that explains a $\frac{\text{Var}(\boldsymbol{\alpha}_1^\top \boldsymbol{X}) + \cdots + \text{Var}(\boldsymbol{\alpha}_k^\top \boldsymbol{X})}{\Sigma_{11} + \cdots + \Sigma_{pp}} \times 100\% = \frac{\bar{\lambda}_1 + \cdots + \bar{\lambda}_k}{\Sigma_{11} + \cdots + \Sigma_{pp}} \times 100\%$ of the total population variance $\Sigma_{11} + \cdots + \Sigma_{pp}$.

6.3 Estimation of the Principal Components

In practice, Σ is unknown and has to be estimated. The principal components are derived from the normalised eigenvectors of the estimated covariance matrix.

Note also that extracting principal components from the (estimated) covariance matrix has the drawback that it is influenced by the scale of measurement of each variable X_i, $i = 1, \ldots, p$. A variable with large variance will necessarily be a large component in the first principal component (note the goal of explaining **the bulk** of variability by using the first principal component). Yet the large variance of the variable may be just an artefact of the measurement scale used for this variable. Therefore, an alternative practice is adopted sometimes to extract principal components from the correlation matrix $\boldsymbol{\rho}$ instead of the covariance matrix Σ.

Example 6.1 (Eigenvalues obtained from Covariance and Correlation Matrices: see JW p. 437). It demonstrates the great effect standardisation may have on the principal components. The relative magnitudes of the weights after standardisation (i.e. from $\boldsymbol{\rho}$ may become in direct opposition to the weights attached to the same variables in the principal component obtained from Σ).

For the reasons mentioned above, variables are often **standardised** before sample principal components are extracted. Standardisation is accomplished by calculating the vectors $\boldsymbol{Z}_i = \left(\frac{X_{1i}-\bar{X}_1}{\sqrt{s_{11}}} \quad \frac{X_{2i}-\bar{X}_2}{\sqrt{s_{22}}} \quad \cdots \quad \frac{X_{pi}-\bar{X}_p}{\sqrt{s_{pp}}}\right)^\top$, $i = 1, \ldots, n$. The standardised observations matrix $\boldsymbol{Z} = [\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_n] \in \mathcal{M}_{p,n}$ gives the sample mean vector $\bar{\boldsymbol{Z}} = \frac{1}{n}\boldsymbol{Z}\boldsymbol{1}_n = 0$ and a sample covariance matrix $\boldsymbol{S}_{\boldsymbol{Z}} = \frac{1}{n-1}\boldsymbol{Z}\boldsymbol{Z}^\top = \boldsymbol{R}$ (the correlation matrix of the original observations). The principal components are extracted in the usual way from $\boldsymbol{R}$ now.

6.4 Deciding how many principal components to include

To reduce the dimensionality (which is the motivating goal), we should restrict attention to the first k principal components and ideally, k should be kept much less than p but there is a trade-off to be made here since we would also like the proportion $\psi_k = \frac{\bar{\lambda}_1+\ldots\bar{\lambda}_k}{\bar{\lambda}_1+\ldots\bar{\lambda}_p}$ be close to one. How could a reasonable trade-off be made? Three methods are most widely used:

- The "scree plot": basically, it is a graphical method of plotting the ordered $\bar{\lambda}_k$ against k and deciding visually when the plot has flattened out. Typically, the initial part of the plot is like the side of the mountain, while the flat portion where each $\bar{\lambda}_k$ is just slightly smaller than $\bar{\lambda}_{k-1}$, is like the rough scree at the bottom. This motivates the name of the plot. The task here is to find where "the scree begins".

- Choose an arbitrary constant $c \in (0, 1)$ and choose k to be the smallest one with the property $\psi_k \geq c$. Usually, $c = 0.9$ is used, but please, note the arbitrariness of the choice here.

- **Kaiser's rule:** it suggests that from all p principal components only the ones should be retained whose variances (after standardisation) are greater than unity, or, equivalently, only those components which, individually, explain at least $\frac{1}{p}100\%$ of the total variance. (This is the same as excluding all principal components with eigenvalues less than the overall average). This criterion has a number of positive features that have contributed to its popularity but can not be defended on a safe theoretical ground.

- Formal tests of significance. Note that it actually **does not make sense** to test whether $\bar{\lambda}_{k+1} = \cdots = \bar{\lambda}_p = 0$ since if such a hypothesis were true then the population distribution would be contained **entirely** within a k-dimensional subspace and the same would be true for any **sample** from this distribution, hence we would have the **estimated** $\bar{\lambda}$ values for indices $k+1, \ldots, p$ being also equal to zero with probability one! What seems to be reasonable to do instead, is to test $H_0 : \bar{\lambda}_{k+1} = \cdots = \bar{\lambda}_p$ (without asking the common value to be zero). This is a more quantitative variant of the scree test. A test for this hypothesis is to form the arithmetic and geometric means a_0 = arithmetic mean of the last $p-k$ estimated eigenvalues; g_0 = geometric mean of the last $p-k$ estimated eigenvalues, and then construct $-2\log\lambda = n(p-k)\log\frac{a_0}{g_0}$. The asymptotic distribution of this statistic under the null hypothesis is χ^2_ν where $\nu = \frac{(p-k+2)(p-k-1)}{2}$. The interested student can find more details about this test in the monograph of Mardia, Kent and Bibby. We should note, however, that the last result holds under multivariate normality assumption and is only valid as stated for the **covariance-based** (**not** the correlation-based) version of the principal component analysis. In practice, many data analysts are reluctant to make a multivariate normality assumption at the early stage of the descriptive data analysis and hence distrust the above quantitative test but prefer the simple Kaiser criterion.

6.5 Software

Principal components analysis can be performed in SAS by using either the `PRINCOMP` or the `FACTOR` procedures and in R using `stats::prcomp`, `stats::princomp`, or about half-dozen other implementations.

6.6 Examples

Example 6.2. **The Crime Rates** example will be discussed at the lecture. The data gives crime rates per 100,000 people in seven categories for each of the 50 states in USA in 1997. Principal components are used to summarise the 7-dimensional data in 2 or 3 dimensions only and help to visualise and interpret the data.

6.7 PCA and Factor Analysis

Principal components can serve as a method for initial factor extraction in exploratory factor analysis. But one should mention here that *Principal component analysis is* ***not*** *Factor analysis.* The main difference is that in factor analysis (to be studied later in this course) one assumes that the covariation in the observed variables is due to the presence of one or more latent variables (factors) that exert casual influence on the observed variables. Factor analysis is being used when it is believed that certain latent factors exist and it is hoped to explore the nature and number of these factors. In contrast, in principal component analysis there is no prior assumption about an underlying casual model. The goal here is just variable reduction.

6.8 Application to finance: Portfolio optimisation

Many other problems in Multivariate Statistics lead to formulating optimisation problems that are similar in spirit to the Principal Component Analysis problem. Hereby, we shall illustrate the Efficient portfolio choice problem.

Assume that a p-dimensional vector $\boldsymbol{X}$ of returns of the p assets is given. Then the return of a **portfolio** that has these assets with weights $(c_1, c_2, \ldots, c_p)$ (with $\sum_{i=1}^{p} c_i = 1$) is $Q = \boldsymbol{c}^\top \boldsymbol{X}$ and the mean return is $\boldsymbol{c}^\top \boldsymbol{\mu}$. (Here we assume that $\mathrm{E}\,\boldsymbol{X} = \boldsymbol{\mu}$, $\mathrm{Var}(\boldsymbol{X}) = \Sigma$.) The *risk* of the portfolio is $\boldsymbol{c}^\top \Sigma \boldsymbol{c}$. Further, assume that a prespecified mean return $\bar{\mu}$ is to be achieved. The question is *how to choose the weights* $\boldsymbol{c}$ so that the risk of a portfolio that achieves the prespecified mean return, is as small as possible.

Mathematically, this is equivalent to the requirement to find the solution of an optimisation problem under two constraints. The Lagrangian function is:

$$\mathrm{Lag}(\lambda_1, \lambda_2) = \boldsymbol{c}^\top \Sigma \boldsymbol{c} + \lambda_1(\bar{\mu} - \boldsymbol{c}^\top \boldsymbol{\mu}) + \lambda_2(1 - \boldsymbol{c}^\top \mathbf{1}_p) \tag{6.3}$$

where $\mathbf{1}_p$ is a p-dimensional vector of ones. Differentiating (6.3) with respect to $\boldsymbol{c}$ we get the first order conditions for a minimum:

$$2\Sigma \boldsymbol{c} - \lambda_1 \boldsymbol{\mu} - \lambda_2 \mathbf{1}_p = 0. \tag{6.4}$$

To simplify derivations, we shall consider the so-called case of non-existence of a riskless asset with a fixed (non-random) return. Then it makes sense to assume that Σ is positive definite and hence Σ^{-1} exists. We get from (6.4) then:

$$\boldsymbol{c} = \frac{1}{2}\Sigma^{-1}(\lambda_1 \boldsymbol{\mu} + \lambda_2 \mathbf{1}_p). \tag{6.5}$$

After multiplying by $\mathbf{1}_p^\top$ from left both sides of the equality, we get:

$$1 = \frac{1}{2}\mathbf{1}_p^\top \Sigma^{-1}(\lambda_1 \boldsymbol{\mu} + \lambda_2 \mathbf{1}_p) \tag{6.6}$$

We can get λ_2 from (6.6) as $\lambda_2 = \frac{2-\lambda_1 \mathbf{1}_p^\top \Sigma^{-1}\boldsymbol{\mu}}{\mathbf{1}_p^\top \Sigma^{-1}\mathbf{1}_p}$ and then substitute it in the formula for $\boldsymbol{c}$ to end up with:

$$\boldsymbol{c} = \frac{1}{2}\lambda_1(\Sigma^{-1}\boldsymbol{\mu} - \frac{\mathbf{1}_p^\top \Sigma^{-1}\boldsymbol{\mu}}{\mathbf{1}_p^\top \Sigma^{-1}\mathbf{1}_p}\Sigma^{-1}\mathbf{1}_p) + \frac{\Sigma^{-1}\mathbf{1}_p}{\mathbf{1}_p^\top \Sigma^{-1}\mathbf{1}_p}. \tag{6.7}$$

In a similar way, if we multiply both sides of (6.5) by $\boldsymbol{\mu}^\top$ from left and use the restriction $\boldsymbol{\mu}^\top \boldsymbol{c} = \bar{\boldsymbol{\mu}}$ we can get one more relationship between λ_1 and λ_2 : $\lambda_1 = \frac{2\bar{\boldsymbol{\mu}} - \lambda_2 \boldsymbol{\mu}^\top \Sigma^{-1}\mathbf{1}_p}{\boldsymbol{\mu}^\top \Sigma^{-1}\boldsymbol{\mu}}$ The linear system of 2 equations with respect to λ_1 and λ_2 can be solved then and the values substituted in (6.7) to get the final expression for $\boldsymbol{c}$ using $\boldsymbol{\mu}$, $\bar{\boldsymbol{\mu}}$ and Σ. (Do it (!))

One special case is of particular interest. This is the so-called variance-efficient portfolio (as opposed to the *mean–variance-efficient portfolio* considered above). For the variance-efficient portfolio, there is *no prespecified mean return, that is, there is no restriction on the mean.* It is only required to minimise the variance. Obviously, we have $\lambda_1 = 0$ then and from (6.7) we get the *optimal weights for the variance efficient portfolio:* $\boldsymbol{c}_{\text{opt}} = \frac{\Sigma^{-1}\mathbf{1}_p}{\mathbf{1}_p^\top \Sigma^{-1}\mathbf{1}_p}$.

6.9 Additional resources

An alternative presentation of these concepts can be found in JW Ch. 8.

6.10 Exercises

Exercise 6.1

A random vector $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}$ is normally distributed with zero mean vector and $\Sigma = \begin{pmatrix} 1 & \rho/2 & 0 \\ \rho/2 & 1 & \rho \\ 0 & \rho & 1 \end{pmatrix}$ where ρ is positive.

(a) Find the coefficients of the first principal component and the variance of that component. What percentage of the overall variability does it explain?

(b) Find the joint distribution of Y_1, Y_2 and $Y_1 + Y_2 + Y_3$.

(c) Find the conditional distribution of Y_1, Y_2 given $Y_3 = y_3$.

(d) Find the multiple correlation of Y_3 with Y_1, Y_2.

7 Canonical Correlation Analysis

7.1 Introduction

Assume we are interested in the association between two **sets** of random variables. Typical examples include: relation between set of governmental policy variables and a set of economic goal variables; relation between college "performance" variables (like grades in courses in five different subject matter areas) and pre-college "achievement" variables (like high-school grade-point averages for junior and senior years, number of high-school extracurricular activities) etc.

The way the above problem of measuring association is solved in Canonical Correlation Analysis, is to consider the largest possible correlation between *linear combination of the variables in the first set* and a *linear combination of the variables in the second set*. The pair of linear combinations obtained through this maximisation process is called **first canonical variables** and their correlation is called **first canonical correlation**. The process can be continued (similarly to the principal components procedure) to find a second pair of linear combinations having the largest correlation among all pairs that are uncorrelated with the initially selected pair. This would give us the second set of canonical variables with their second canonical correlation etc. The maximisation process that we are performing at each step reflects our wish (again like in principal components analysis) to concentrate the initially high dimensional relationship between the 2 sets of variables into a few pairs of canonical variables only. Often, even only **one** pair is considered. The rationale in canonical correlation analysis is that when the number of variables is large, interpreting the **whole set** of correlation coefficients between pairs of variables from each set is hopeless and in that case one should concentrate on a *few* carefully chosen representative correlations. Finally, we should note that the traditional (simple) correlation coefficient and the multiple correlation coefficient (Lecture 5) are *special cases* of canonical correlation in which one or both sets contain a single variable.

7.2 Application in testing for independence of sets of variables

Besides being interesting in its own right (see Section 7.1), calculating canonical correlations turns out to be important for the sake of **testing independence of sets of random variables**. Let us remember that testing for independence and for uncorrelatedness in the case of multivariate normal are equivalent problems. Assume now that that $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$. Furthermore, let $\boldsymbol{X}$ be partitioned into r, q components $(r+q=p)$ with $\boldsymbol{X}^{(1)} \in \mathbb{R}^r, \boldsymbol{X}^{(2)} \in \mathbb{R}^q$ and correspondingly, the covariance matrix

$$\Sigma = \mathrm{E}(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^\top = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix} \in \mathcal{M}_{p,p}$$

has been also partitioned into $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$, accordingly. We shall assume for simplicity that the matrices Σ, Σ_{11}, and Σ_{22} are nonsingular. To test $H_0 : \Sigma_{12} = \mathbf{0}$ against a general alternative, a sensible way to go would be the following: for fixed vectors $\boldsymbol{a} \in \mathbb{R}^r, \boldsymbol{b} \in \mathbb{R}^q$ let $Z_1 = \boldsymbol{a}^\top \boldsymbol{X}^{(1)}$ and $Z_2 = \boldsymbol{b}^\top \boldsymbol{X}^{(2)}$ giving $\rho_{\boldsymbol{a},\boldsymbol{b}} = \mathrm{Cor}(Z_1, Z_2) = \frac{\boldsymbol{a}^\top \Sigma_{12} \boldsymbol{b}}{\sqrt{\boldsymbol{a}^\top \Sigma_{11} \boldsymbol{a} \boldsymbol{b}^\top \Sigma_{22} \boldsymbol{b}}}$. H_0 is equivalent to $H_0 : \rho_{\boldsymbol{a},\boldsymbol{b}} = 0$ for all $\boldsymbol{a} \in \mathbb{R}^r, \boldsymbol{b} \in \mathbb{R}^q$. For a particular pair $\boldsymbol{a}, \boldsymbol{b}$, H_0 would be accepted if $|r_{\boldsymbol{a},\boldsymbol{b}}| = \frac{|\boldsymbol{a}^\top \boldsymbol{S}_{12} \boldsymbol{b}|}{\sqrt{\boldsymbol{a}^\top \boldsymbol{S}_{11} \boldsymbol{a} \boldsymbol{b}^\top \boldsymbol{S}_{22} \boldsymbol{b}}} \leq k$ for certain positive constant k. (Here $\boldsymbol{S}_{ij}$ are the corresponding data based estimators of Σ_{ij}.) Hence an appropriate acceptance region for H_0 would be given in the form $\{\boldsymbol{X} \in \mathcal{M}_{p,n} : \max_{\boldsymbol{a},\boldsymbol{b}} r^2_{\boldsymbol{a},\boldsymbol{b}} \leq k^2\}$. But maximising $r^2_{\boldsymbol{a},\boldsymbol{b}}$ means to find the maximum of $(\boldsymbol{a}^\top \boldsymbol{S}_{12} \boldsymbol{b})^2$ under constraints $\boldsymbol{a}^\top \boldsymbol{S}_{11} \boldsymbol{a} = 1$ and $\boldsymbol{b}^\top \boldsymbol{S}_{22} \boldsymbol{b} = 1$, and this is exactly the data-based version of the optimisation problem to be solved in Section 7.1. For the goals in Sections 7.1 and 7.2 to be achieved, we need to solve problems of the following type.

7.3 Precise mathematical formulation and solution to the problem

Canonical variables are the variables $Z_1 = \boldsymbol{a}^\top \boldsymbol{X}^{(1)}$ and $Z_2 = \boldsymbol{b}^\top \boldsymbol{X}^{(2)}$ where $\boldsymbol{a} \in \mathbb{R}^r, \boldsymbol{b} \in \mathbb{R}^q$ are obtained by maximising $(\boldsymbol{a}^\top \Sigma_{12} \boldsymbol{b})^2$ under the constraints $\boldsymbol{a}^\top \Sigma_{11} \boldsymbol{a} = \boldsymbol{b}^\top \Sigma_{22} \boldsymbol{b} = 1$. To solve the above maximisation problem, we construct

$$\mathrm{Lag}(\boldsymbol{a}, \boldsymbol{b}, \lambda_1, \lambda_2) = (\boldsymbol{a}^\top \Sigma_{12} \boldsymbol{b})^2 + \lambda_1 (\boldsymbol{a}^\top \Sigma_{11} \boldsymbol{a} - 1) + \lambda_2 (\boldsymbol{b}^\top \Sigma_{22} \boldsymbol{b} - 1).$$

Partial differentiation with respect to the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ gives:

$$2(\boldsymbol{a}^\top \Sigma_{12} \boldsymbol{b}) \Sigma_{12} \boldsymbol{b} + 2\lambda_1 \Sigma_{11} \boldsymbol{a} = \mathbf{0} \in \mathbb{R}^r, \tag{7.1}$$

$$2(\boldsymbol{a}^\top \Sigma_{12} \boldsymbol{b}) \Sigma_{21} \boldsymbol{a} + 2\lambda_2 \Sigma_{22} \boldsymbol{b} = \mathbf{0} \in \mathbb{R}^q. \tag{7.2}$$

We multiply (7.1) by the vector $\boldsymbol{a}^\top$ from left and equation (7.2) by $\boldsymbol{b}^\top$ from left and after subtracting the two equations obtained we get $\lambda_1 = \lambda_2 = -(\boldsymbol{a}^\top \Sigma_{12} \boldsymbol{b})^2 = -\mu^2$. Hence:

$$\Sigma_{12} \boldsymbol{b} = \mu \Sigma_{11} \boldsymbol{a} \tag{7.3}$$

and

$$\Sigma_{21} \boldsymbol{a} = \mu \Sigma_{22} \boldsymbol{b} \tag{7.4}$$

hold.

Now we first multiply (7.3) by $\Sigma_{21} \Sigma_{11}^{-1}$ from left, then both sides of (7.4) by the scalar μ and after finally adding the two equations we get:

$$(\Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} - \mu^2 \Sigma_{22}) \boldsymbol{b} = \mathbf{0}. \tag{7.5}$$

The homogeneous equation system (7.5) having a non-trivial solution w.r.t. $\boldsymbol{b}$ means that

$$|\Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} - \mu^2 \Sigma_{22}| = 0 \tag{7.6}$$

must hold. Then, of course,

$$|\Sigma_{22}^{-\frac{1}{2}}||\Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} - \mu^2 \Sigma_{22}||\Sigma_{22}^{-\frac{1}{2}}| = |\Sigma_{22}^{-\frac{1}{2}} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-\frac{1}{2}} - \mu^2 I_q| = 0$$

must hold. This means that μ^2 has to be an eigenvalue of the matrix $\Sigma_{22}^{-\frac{1}{2}} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-\frac{1}{2}}$. Also, $\boldsymbol{b} = \Sigma_{22}^{-\frac{1}{2}} \hat{\boldsymbol{b}}$ where $\hat{\boldsymbol{b}}$ is the eigenvector of $\Sigma_{22}^{-\frac{1}{2}} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-\frac{1}{2}}$ corresponding to this eigenvalue (WHY?!).

(Note, however, that this representation is good mainly for theoretical purposes, the main advantage being that one is dealing with eigenvalues of a symmetric matrix. If doing calculations by hand, it is usually easier to calculate $\boldsymbol{b}$ directly as the solution of the linear equation (7.5), i.e., find the largest eigenvalue of the (non-symmetric) matrix $\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ and then find the eigenvector $\boldsymbol{b}$ that corresponds to it. Besides, we also see from the definition of μ that $\mu^2 = (\boldsymbol{a}^\top\Sigma_{12}\boldsymbol{b})^2$ holds.)

Since we wanted to **maximise** the right hand side, it is obvious that μ^2 must be chosen to be the **largest eigenvalue** of the matrix $\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-\frac{1}{2}}$ (or, which is the same thing, the largest eigenvalue of the matrix $\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}$). Finally, we can obtain the vector $\boldsymbol{a}$ from (7.3): $\boldsymbol{a} = \frac{1}{\mu}\Sigma_{11}^{-1}\Sigma_{12}\boldsymbol{b}$. That way, the **first** canonical variables $Z_1 = \boldsymbol{a}^\top\boldsymbol{X}^{(1)}$ and $Z_2 = \boldsymbol{b}^\top\boldsymbol{X}^{(2)}$ are determined and the value of the first canonical correlation is just μ. The orientation of the vector $\boldsymbol{b}$ is chosen such that the sign of μ should be positive.

Now, it is easy to see that if we want to extract a second pair of canonical variables we need to repeat the same process by starting with the **second largest** eigenvalue μ^2 of the matrix $\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-\frac{1}{2}}$ (or of the matrix $\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$). This will automatically ensure that the second pair of canonical variables is uncorrelated with the first pair. The process can theoretically be continued until the number of pairs of canonical variables equals the number of variables in the smaller group. But in practice, much fewer canonical variables will be needed. Each canonical variable is uncorrelated with all the other canonical variables of either set except for the one corresponding canonical variable in the opposite set.

Note. It is important to point out that already by definition the canonical correlation is at least as large as the multiple correlation between any variable and the opposite set of variables. It is in fact possible for the first canonical correlation to be *very large* while all the multiple correlations of each separate variable with the opposite set of canonical variables are small. This once again underlines the importance of Canonical Correlation analysis.

7.4 Estimating and testing canonical correlations

The way to estimate the canonical variables and canonical correlation coefficients is based on the plug-in technique: one follows the steps outlined in Section 7.3, by each time substituting $\boldsymbol{S}_{ij}$ in place of Σ_{ij}.

Let us now discuss the independence testing issue outlined in Section 7.2. The acceptance region of the independence test of H_0 in Section 7.2. would be $\{\boldsymbol{X} \in \mathcal{M}_{p,n}$: largest eigenvalue of $\boldsymbol{S}_{22}^{-\frac{1}{2}}\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-\frac{1}{2}} \leq k_\alpha\}$ where k_α has been worked out and is given in the so called **Hecks charts**. This distribution depends on three parameters: $s = \min(r, q)$, $m = \frac{|r-q|-1}{2}$, and $N = \frac{n-r-q-2}{2}$, n being the sample size. Besides using the charts, one can also use good F-distribution-based approximations for a (transformations of) this distribution like Wilk's lambda, Pillai's trace, Hotelling trace, and Roy's greatest root.

7.5 Software

Here we shall only mention that all these statistics and their P-values (using suitable F-distribution-based approximations) are readily available as an output in the SAS program `CANCORR` so that performing the test is really easy-one can read out directly the p-value from the SAS output. In R, see `stats::cancor` and package `CCA` for computing and visualisation, and package `CCP` for testing canonical correlations.

7.6 Some important computational issues

Note that calculating $X^{-\frac{1}{2}}$ and $X^{\frac{1}{2}}$ for a symmetric positive definite matrix X according to the theoretically attractive spectral decomposition method may be numerically unstable. This is especially the case when some of the eigenvalues are close to zero (or, more precisely, when the the ratio of the greatest eigenvalue and the least eigenvalue—the *condition number*—is high). We can use the **Cholesky decomposition** described in Section 0.1.6 instead. Looking back at (7.5), we see that if $U^\top U = \Sigma_{22}^{-1}$ gives the Cholesky decomposition of the matrix Σ_{22}^{-1} then μ^2 is an eigenvalue of the matrix $A = U\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}U^\top$. Indeed, by multiplying from left by U and from right by $U^\top$ in (7.6) we get:

$$|A - \mu^2 U\Sigma_{22}U^\top| = 0.$$

But $U\Sigma_{22}U^\top = U(U^\top U)^{-1}U^\top = UU^{-1}(U^\top)^{-1}U^\top = I$ holds.

7.7 Examples

Example 7.1. **Canonical Correlation Analysis of the Fitness Club Data**. Three physiological and three exercise variables were measured on twenty middle aged men in a fitness club. Canonical correlation is used to determine if the physiological variables are related in any way to the exercise variables.

Example 7.2. **JW Example 10.4, p. 552** Studying canonical correlations between leg and head bone measurements: X_1, X_2 are skull length and skull breadth, respectively; X_3, X_4 are leg bone measurements: femur and tibia length, respectively. Observations have been taken on $n = 276$ White Leghorn chicken. The example is chosen to also illustrate how a canonical correlation analysis can be performed when the original data is not given but the empirical correlation matrix (or empirical covariance matrix) is available.

7.8 Additional resources

An alternative presentation of these concepts can be found in JW Ch. 10.

7.9 Exercises

Exercise 7.1

Let the components of X correspond to scores on tests in arithmetic speed (X_1), arithmetic power (X_2), memory for words (X_3), memory for meaningful symbols (X_4), and memory for meaningless symbols (X_5). The observed correlations in a sample of 140 are

$$\begin{bmatrix} 1.0000 & 0.4248 & 0.0420 & 0.0215 & 0.0573 \\ & 1.0000 & 0.1487 & 0.2489 & 0.2843 \\ & & 1.0000 & 0.6693 & 0.4662 \\ & & & 1.0000 & 0.6915 \\ & & & & 1.0000 \end{bmatrix}.$$

Find the canonical correlations and canonical variates between the first two variates and the last three variates. Comment. Write a SAS-IML or R code to implement the required calculations.

Exercise 7.2

Students sit 5 different papers, two of which are closed book and the rest open book. For the 88 students who sat these exams the sample covariance matrix is

$$S = \begin{bmatrix} 302.3 & 125.8 & 100.4 & 105.1 & 116.1 \\ & 170.9 & 84.2 & 93.6 & 97.9 \\ & & 111.6 & 110.8 & 120.5 \\ & & & 217.9 & 153.8 \\ & & & & 294.4 \end{bmatrix}.$$

Find the canonical correlations and canonical variates between the first two variates (closed book exams) and the last three variates (open book exams). Comment.

Exercise 7.3

A random vector $\boldsymbol{X} \sim N_4(\boldsymbol{\mu}, \Sigma)$ with $\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$ and $\begin{pmatrix} 1 & 2\rho & \rho & \rho \\ 2\rho & 1 & \rho & \rho \\ \rho & \rho & 1 & 2\rho \\ \rho & \rho & 2\rho & 1 \end{pmatrix}$ where ρ is a small enough positive constant.

(a) Find the two canonical correlations between $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ and $\begin{pmatrix} X_3 \\ X_4 \end{pmatrix}$. Comment.

(b) Find the first pair of canonical variables.

Exercise 7.4

Consider the following covariance matrix Σ of a four dimensional normal vector: $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \left(\begin{array}{cc|cc} 100 & 0 & 0 & 0 \\ 0 & 1 & 0.95 & 0 \\ \hline 0 & 0.95 & 1 & 0 \\ 0 & 0 & 0 & 100 \end{array}\right)$. Verify that the first pair of canonical variates are just the second and the third component of the vector and the canonical correlation equals .95.

8 Multivariate Linear Models and Multivariate ANOVA

8.1 Univariate linear models and ANOVA

Recall the univariate linear model: for observations $i = 1, 2, \ldots, n$, let the response variable $Y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i$, for predictor row vector $\boldsymbol{x}_i^\top \in \mathbb{R}^k$ assumed fixed and known, coefficient vector $\boldsymbol{\beta} \in \mathbb{R}^p$ fixed and unknown, and $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. In matrix form, $\boldsymbol{Y} = \begin{pmatrix} Y_1 & Y_2 & \cdots & Y_n \end{pmatrix}^\top$ and $X = \begin{pmatrix} \boldsymbol{x}_1^\top & \boldsymbol{x}_2^\top & \cdots & \boldsymbol{x}_n^\top \end{pmatrix}^\top \in \mathcal{M}_{n,k}$. We will assume that X contains an intercept. Then,

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, I_n\sigma^2)$. The MLE for $\boldsymbol{\beta}$ requires us to minimise

$$\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2 = \|\boldsymbol{Y} - X\boldsymbol{\beta}\|^2 = (\boldsymbol{Y} - X\boldsymbol{\beta})^\top(\boldsymbol{Y} - X\boldsymbol{\beta}),$$

and, after some vector calculus, we get

$$\hat{\boldsymbol{\beta}} = (X^\top X)^{-1}X^\top\boldsymbol{Y}$$

with

$$\operatorname{Var}(\hat{\boldsymbol{\beta}}) = (X^\top X)^{-1}X^\top \operatorname{Var}(\boldsymbol{Y})X(X^\top X)^{-1} = (X^\top X)^{-1}\sigma^2.$$

Furthermore, we can consider projection matrices $A = I_n - X(X^\top X)^{-1}X^\top$ and $B = X(X^\top X)^{-1}X^\top - \boldsymbol{1}_n(\boldsymbol{1}_n^\top\boldsymbol{1}_n)^{-1}\boldsymbol{1}_n^\top$, with

$$A\boldsymbol{Y} = \boldsymbol{Y} - X\{(X^\top X)^{-1}X^\top\boldsymbol{Y}\} = \boldsymbol{Y} - \hat{\boldsymbol{Y}},$$

the residual vector and

$$B\boldsymbol{Y} = X\{(X^\top X)^{-1}X^\top\}\boldsymbol{Y} - \boldsymbol{1}_n(\boldsymbol{1}_n^\top\boldsymbol{1}_n)^{-1}\boldsymbol{1}_n^\top\boldsymbol{Y} = \hat{\boldsymbol{Y}} - \boldsymbol{1}_n\bar{\boldsymbol{Y}}$$

the vector of fitted values over and above the mean, and observe that

$$\begin{aligned}
\operatorname{Cov}(A\boldsymbol{Y}, B\boldsymbol{Y}) &= A\operatorname{Var}(\boldsymbol{Y})B^\top = \sigma^2 AB^\top \\
&= X(X^\top X)^{-1}X^\top - X(X^\top X)^{-1}X^\top X(X^\top X)^{-1}X^\top \\
&\quad - \boldsymbol{1}_n(\boldsymbol{1}_n^\top\boldsymbol{1}_n)^{-1}\boldsymbol{1}_n^\top + X(X^\top X)^{-1}X^\top\boldsymbol{1}_n(\boldsymbol{1}_n^\top\boldsymbol{1}_n)^{-1}\boldsymbol{1}_n^\top \\
&= \frac{1}{n}(X(X^\top X)^{-1}X^\top\boldsymbol{1}_n - \boldsymbol{1}_n)\boldsymbol{1}_n^\top = \boldsymbol{0}
\end{aligned}$$

if X contains an intercept effect. Then, $\text{SSE} = \boldsymbol{Y}^\top A\boldsymbol{Y} \sim \sigma^2\chi^2_{n-k}$ and $\text{SSA} = \boldsymbol{Y}^\top B\boldsymbol{Y} \sim \sigma^2\chi^2_{k-1}$, independent, letting us set up $F = \frac{\text{SSA}/(k-1)}{\text{SSE}/(n-k)} \sim F_{k-1,n-k}$, etc..

8.2 Multivariate Linear Model and MANOVA

How do we generalise it to multivariate response? That is, suppose that we observe the following response matrix:

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1^\top \\ \boldsymbol{Y}_2^\top \\ \vdots \\ \boldsymbol{Y}_n^\top \end{pmatrix} = \begin{pmatrix} Y_{11} & Y_{12} & \cdots & Y_{1p} \\ Y_{21} & Y_{22} & \cdots & Y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{np} \end{pmatrix} \in \mathcal{M}_{n,p}$$

with $\boldsymbol{x}_i$ and X as before, and

$$\boldsymbol{Y}_i^\top = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i^\top$$

where $\boldsymbol{\beta} \in \mathcal{M}_{k,p}$, and $\boldsymbol{\epsilon}_i \sim N_p(\boldsymbol{0}, \Sigma)$, $\Sigma \in \mathcal{M}_{p,p}$ symmetric positive definite. In matrix form,

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{E},$$

where

$$\boldsymbol{E} = \begin{pmatrix} \boldsymbol{\epsilon}_1 & \boldsymbol{\epsilon}_2 & \cdots & \boldsymbol{\epsilon}_n \end{pmatrix}^\top \in \mathcal{M}_{n,p}.$$

Then, we can write $\vec{\boldsymbol{E}} \sim N_{np}(\boldsymbol{0}, \Sigma \otimes I_n)$ or $\overrightarrow{\boldsymbol{E}^\top} \sim N_{np}(\boldsymbol{0}, I_n \otimes \Sigma)$, and

$$\vec{\boldsymbol{Y}} \sim N_{np}(\{\boldsymbol{\beta}^\top \otimes I_n\}\vec{X}, \Sigma \otimes I_n)$$

or

$$\overrightarrow{\boldsymbol{Y}^\top} \sim N_{np}(\{I_n \otimes \boldsymbol{\beta}^\top\}\overrightarrow{X^\top}, I_n \otimes \Sigma).$$

MLE is equivalent to the OLS problem minimising $\sum_{i=1}^n \mathrm{tr}\{(\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta})(\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta})^\top\} = \mathrm{tr}\{(\boldsymbol{Y} - X\boldsymbol{\beta})^\top(\boldsymbol{Y} - X\boldsymbol{\beta})\}$, leading to

$$\hat{\boldsymbol{\beta}} = (X^\top X)^{-1}X^\top \boldsymbol{Y}$$

again, with

$$\begin{aligned} \mathrm{Var}(\overrightarrow{\hat{\boldsymbol{\beta}}^\top}) &= \mathrm{Var}(\overrightarrow{\boldsymbol{Y}^\top X(X^\top X)^{-1}}) = \mathrm{Var}\{((X^\top X)^{-1}X^\top \otimes I_p)\overrightarrow{\boldsymbol{Y}^\top}\} \\ &= ((X^\top X)^{-1}X^\top \otimes I_p)(I_p \otimes \Sigma)((X^\top X)^{-1}X^\top \otimes I_p)^\top \\ &= ((X^\top X)^{-1}X^\top \otimes I_p)((X^\top X)^{-1}X^\top \otimes \Sigma)^\top \\ &= (X^\top X)^{-1} \otimes \Sigma, \end{aligned}$$

or

$$\mathrm{Var}(\overrightarrow{\hat{\boldsymbol{\beta}}}) = \Sigma \otimes (X^\top X)^{-1}.$$

Projection matrices A and B still work (check it!), and we can write $\mathrm{SSE} = \boldsymbol{Y}^\top A\boldsymbol{Y} \sim W_p(\Sigma, p(n-k-1))$ and $\mathrm{SSA} = \boldsymbol{Y}^\top B\boldsymbol{Y} \sim W_p(\Sigma, p(k-1))$. Notice that they are now matrices.

8.3 Computations used in the MANOVA tests

In standard (univariate) Analysis of Variance, with usual normality assumptions on the errors, testing about effects of the factors involved in the model description is based on the F test. The F tests are derived from the ANOVA decomposition $\mathrm{SST} = \mathrm{SSA} + \mathrm{SSE}$. The argument goes as follows:

i) SSE and SSA are independent, (up to constant factors involving the variance σ^2 of the errors) χ^2 distributed;

ii) By proper norming to account for degrees of freedom, from SSE and SSA one gets statistics that have the following behaviour: the normed SSE always delivers an unbiased estimator of σ^2 no matter if the null hypothesis or alternative is true; the normed SSA delivers an unbiased estimator of σ^2 under the null hypothesis but delivers an unbiased estimator of a "larger" quantity under the alternative.

The above observation is crucial and motivates the F-testing: F statistics are (**suitably normed to account for degrees of freedom**) ratios of SSA/SSE. When taking the ratio, the factors involving σ^2 **cancel out** and σ^2 does not play any role in the distribution of the ratio. Under H_0 their distribution is F. When the null hypothesis is violated, then the same statistics will tend to have "larger" values as compared to the case when H_0 is true. Hence significant (w.r.t. the corresponding F-distribution) values of the statistic lead to rejection of H_0.

Aiming at generalising these ideas to the Multivariate ANOVA (MANOVA) case, we should note that instead of χ^2 distributions we now have to deal with **Wishart** distributions and we need to properly define (a proper functional of) the SSA/SSE ratio which would be a "ratio" of matrices now. Obviously, there are more ways to define suitable statistics in this context! It turns out that such functionals are related to the eigenvalues of the (properly normed) Wishart-distributed matrices that enter the decomposition SST = SSA + SSE in the multivariate case.

8.3.1 Roots distributions

Let $\boldsymbol{Y}_i, i = 1, 2, \ldots, n \overset{\text{ind.}}{\sim} N_p(\boldsymbol{\mu}_i, \Sigma)$. Then the following data matrix:

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1^\top \\ \boldsymbol{Y}_2^\top \\ \vdots \\ \boldsymbol{Y}_n^\top \end{pmatrix} = \begin{pmatrix} Y_{11} & Y_{12} & \cdots & Y_{1p} \\ Y_{21} & Y_{22} & \cdots & Y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{np} \end{pmatrix} \in \mathcal{M}_{n,p}$$

is a $n \times p$ matrix containing n p-dimensional (transposed) vectors. Denote: $\mathrm{E}(\boldsymbol{Y}) = M$, $\mathrm{Var}(\vec{\boldsymbol{Y}}) = \Sigma \otimes I_n$. Let A and B be projectors such that $\boldsymbol{Q}_1 = \boldsymbol{Y}^\top A \boldsymbol{Y}$ and $\boldsymbol{Q}_2 = \boldsymbol{Y}^\top B \boldsymbol{Y}$ are two **independent** $W_p(\Sigma, v)$ and $W_p(\Sigma, q)$ matrices, respectively. Although the theory is general, to keep you on track, you could always think about a multivariate linear model example:

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{E},\ \hat{\boldsymbol{Y}} = X\hat{\boldsymbol{\beta}}$$

$$A = I_n - X(X^\top X)^- X^\top,\ B = X(X^\top X)^- X^\top - \mathbf{1}_n(\mathbf{1}_n^\top \mathbf{1}_n)^{-1}\mathbf{1}_n^\top$$

and the corresponding decomposition

$$\boldsymbol{Y}[I_n - \mathbf{1}_n(\mathbf{1}_n^\top \mathbf{1}_n)^{-1}\mathbf{1}_n^\top]\boldsymbol{Y} = \boldsymbol{Y}^\top B \boldsymbol{Y} + \boldsymbol{Y}^\top A \boldsymbol{Y} = \boldsymbol{Q}_2 + \boldsymbol{Q}_1$$

of SST = SSA + SSE = $\boldsymbol{Q}_2 + \boldsymbol{Q}_1$ where $\boldsymbol{Q}_2$ is the "hypothesis matrix" and $\boldsymbol{Q}_1$ is the "error matrix".

Lemma 8.1. *Let $\boldsymbol{Q}_1, \boldsymbol{Q}_2 \in \mathcal{M}_{p,p}$ be two positive definite symmetric matrices . Then the roots of the determinant equation $|\boldsymbol{Q}_2 - \theta(\boldsymbol{Q}_1 + \boldsymbol{Q}_2)| = 0$ are related to the roots of the equation $|\boldsymbol{Q}_2 - \lambda \boldsymbol{Q}_1| = 0$ by: $\lambda_i = \frac{\theta_i}{1-\theta_i}$ (or $\theta_i = \frac{\lambda_i}{1+\lambda_i}$).*

Lemma 8.2. *Let $\boldsymbol{Q}_1, \boldsymbol{Q}_2 \in \mathcal{M}_{p,p}$ be two positive definite symmetric matrices . Then the roots of the determinant equation $|\boldsymbol{Q}_1 - v(\boldsymbol{Q}_1 + \boldsymbol{Q}_2)| = 0$ are related to the roots of the equation $|\boldsymbol{Q}_2 - \lambda \boldsymbol{Q}_1| = 0$ by: $\lambda_i = \frac{1-v_i}{v_i}$ (or $v_i = \frac{1}{1+\lambda_i}$).*

We can employ the above two lemmas to see that if λ_i, v_i, θ_i are the roots of

$$|\boldsymbol{Q}_2 - \lambda \boldsymbol{Q}_1| = 0,\ |\boldsymbol{Q}_1 - v(\boldsymbol{Q}_1 + \boldsymbol{Q}_2)| = 0,\ |\boldsymbol{Q}_2 - \theta(\boldsymbol{Q}_1 + \boldsymbol{Q}_2)| = 0$$

then:

$$\Lambda = |\boldsymbol{Q}_1(\boldsymbol{Q}_1 + \boldsymbol{Q}_2)^{-1}| = \prod_{i=1}^{p}(1+\lambda_i)^{-1}$$

(Wilks' Criterion statistic) or

$$|\boldsymbol{Q}_2\boldsymbol{Q}_1^{-1}| = \prod_{i=1}^{p}\lambda_i = \prod_{i=1}^{p}\frac{1-v_i}{v_i} = \prod_{i=1}^{p}\frac{\theta_i}{1-\theta_i}$$

or

$$|\boldsymbol{Q}_2(\boldsymbol{Q}_1 + \boldsymbol{Q}_2)^{-1}| = \prod_{i=1}^{p}\theta_i = \prod_{i=1}^{p}\frac{\lambda_i}{1+\lambda_i} = \prod_{i=1}^{p}(1-v_i)$$

and other functional transformations of these products of (random) roots would have a distribution that would only depend on p (the dimension of $\boldsymbol{Y}_i$), v (the Wishart degrees of freedom for $\boldsymbol{Q}_1$), and q (same for $\boldsymbol{Q}_2$).

There are various ways to choose such functional transformations (statistics) and many have been suggested like:

- Λ (Wilks's Lambda)
- $\mathrm{tr}(\boldsymbol{Q}_2\boldsymbol{Q}_1^{-1}) = \mathrm{tr}(\boldsymbol{Q}_1^{-1}\boldsymbol{Q}_2) = \sum_{i=1}^{p}\lambda_i$ (Lawley–Hotelling trace)
- $\max_i \lambda_i$ (Roy's criterion)
- $V = \mathrm{tr}[\boldsymbol{Q}_2(\boldsymbol{Q}_1 + \boldsymbol{Q}_2)^{-1}] = \sum_{i=1}^{p}\frac{\lambda_i}{1+\lambda_i}$ (Pillai statistic / Pillai's trace)

Tables and charts for their exact or approximate distributions are available. Also, P-values for these statistics are readily calculated in statistical packages. In these applications, the meaning of $\boldsymbol{Q}_1$ is of the "error matrix" (also denoted by $\boldsymbol{E}$ sometimes) and the meaning of $\boldsymbol{Q}_2$ is that of a "hypothesis matrix" (also denoted by $\boldsymbol{H}$ sometimes).

The distribution of the statistics defined above depends on the following three parameters:

- p = the number of responses
- $q = \nu_h$ = degrees of freedom for the hypothesis
- $v = \nu_e$ = degrees of freedom for the error

Based on these, the following quantities are calculated: $s = \min(p, q)$, $m = 0.5(|p-q|-1)$, $n = 0.5(v-p-1)$, $r = v - 0.5(p-q+1)$, $u = 0.25(pq-2)$. Moreover, we define: $t = \sqrt{\frac{p^2q^2-4}{p^2+q^2-5}}$ if $p^2+q^2-5 > 0$ and $t = 1$ otherwise. Let us order the eigenvalues of $\boldsymbol{E}^{-1}\boldsymbol{H} = \boldsymbol{Q}_1^{-1}\boldsymbol{Q}_2$ according to: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$.

Then the following distribution results are **exact** if $s = 1$ or 2, otherwise **approximate**:

- Wilks's test. The test statistics, Wilks's lambda, is $\Lambda = \frac{|\boldsymbol{E}|}{|\boldsymbol{E}+\boldsymbol{H}|} = \prod_{i=1}^{p}\frac{1}{1+\lambda_i}$ Then it holds: $F = \frac{1-\Lambda^{1/t}}{\Lambda^{1/t}}.\frac{rt-2u}{pq} \sim F_{pq,rt-2u}$ df (**Rao's F**).
- Lawley–Hotelling trace Test. The Lawley–Hotelling statistic is $U = \mathrm{tr}(\boldsymbol{E}^{-1}\boldsymbol{H}) = \lambda_1 + \cdots + \lambda_p$, and $F = 2(sn+1)\frac{U}{s^2(2m+s+1)} \sim F_{s(2m+s+1),2(sn+1)}$ df.

- Pillai's test. The test statistic, Pillai trace, is $V = \text{tr}(\boldsymbol{H}(\boldsymbol{H}+\boldsymbol{E})^{-1}) = \frac{\lambda_1}{1+\lambda_1} + \cdots + \frac{\lambda_p}{1+\lambda_p}$ and $F = \frac{2n+s+1}{2m+s+1} \times \frac{V}{s-V} \sim F_{s(2m+s+1),s(2n+s+1)}$ df.
- Roy's maximum root criterion. The test statistic is just the largest eigenvalue λ_1.

Finally, we shall mention one historically older and very universal approximation to the distribution of the Λ statistic due to Bartlett (1927):

It holds: level of $-[\nu_e - \frac{p-\nu_h+1}{2}] \log \Lambda = c(p, \nu_h, M) \times$ level of $\chi^2_{p\nu_h}$, where the constant $c(p, \nu_h, M = \nu_e - p + 1)$ is given in tables. Such tables are prepared for levels $\alpha = 0.10, 0.05, 0.025$ etc..

In the context of testing the hypothesis about significance of the first canonical correlation, we have:

$$\boldsymbol{E} = \boldsymbol{S}_{22} - \boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}, \ \boldsymbol{H} = \boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}.$$

The Wilks's statistic becomes $\frac{|\boldsymbol{S}|}{|\boldsymbol{S}_{11}||\boldsymbol{S}_{22}|}$. (Recall (4.3)!) We also see that in this case, if μ_i^2 were the squared canonical correlations then μ_1^2 was defined as the maximal eigenvalue to $\boldsymbol{S}_{22}^{-1}\boldsymbol{H}$, that is, it is a solution to $|(\boldsymbol{E}+\boldsymbol{H})^{-1}\boldsymbol{H} - \mu_1^2 I| = 0$ However, setting $\lambda_1 = \frac{\mu_1^2}{1-\mu_1^2}$ we see that:

$$|(\boldsymbol{E}+\boldsymbol{H})^{-1}\boldsymbol{H} - \mu_1^2 I| = 0 \implies |\boldsymbol{H} - \mu_1^2(\boldsymbol{E}+\boldsymbol{H})| = 0 \implies |\boldsymbol{H} - \frac{\mu_1^2}{1-\mu_1^2}\boldsymbol{E}| = 0 \implies |\boldsymbol{E}^{-1}\boldsymbol{H} - \lambda_1 I| = 0$$

holds and λ_1 is an eigenvalue of $\boldsymbol{E}^{-1}\boldsymbol{H}$. Similarly you can argue for the remaining $\lambda_i = \frac{\mu_i^2}{1-\mu_i^2}$ values. What are the degrees of freedom of $\boldsymbol{E}$ and $\boldsymbol{H}$?

8.3.2 Comparisons

From all statistics discussed, Wilks's lambda has been most widely applied. One important reason for this is that this statistic has the virtue of being convenient to use and, more importantly, being related to the Likelihood Ratio Test! Despite the above, the fact that so many different statistics exist for the same hypothesis testing problem, indicates that there is no universally best test. Power comparisons of the above tests are almost lacking since the distribution of the statistic under alternatives is hardly known.

8.4 Software

In SAS, both `PROC GLM` and `PROC REG` can conduct analysis and perform hypothesis tests. In R, use `stats::lm`.

8.5 Examples

Example 8.3. Multivariate linear modelling of the Fitness dataset.

8.6 Additional resources

An alternative presentation of these concepts can be found in JW Ch. 7.

9 Tests of a Covariance Matrix

Previously, we developed a number of techniques for decomposing and analysing covariance matrices and their properties. Here, we develop a general family of tests for their structure, which will let you specify almost arbitrary tests for the covariance structure of a multivariate normal population.

9.1 Test of $\Sigma = \Sigma_0$

We start with this simpler case since ideas are more transparent. The practically more relevant cases are about comparing covariance matrices of two or more multivariate normal populations but the derivations of the latter tests is more subtle. For these we will only formulate the final results.

Assume now that we have the sample $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ from a $N_p(\boldsymbol{\mu}, \Sigma)$ distribution and we would like to test $H_0 : \Sigma = \Sigma_0$ against the alternative $H_1 : \Sigma \neq \Sigma_0$. Obviously the problem can be easily transformed into testing $\bar{H}_0 : \Sigma = I_p$ since otherwise we can consider the modified observations $\boldsymbol{Y}_i = \Sigma_0^{-\frac{1}{2}} \boldsymbol{X}_i$ which under H_0 will be multivariate normal with a covariance matrix being equal to I_p. Therefore we can assume that $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ is a sample from a $N_p(\boldsymbol{\mu}, \Sigma)$ and we want to test $H_0 : \Sigma = I_p$ versus $H_1 : \Sigma \neq I_p$.

We will derive the likelihood ratio test for this problem. The likelihood function is

$$\begin{aligned} L(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) &= (2\pi)^{-\frac{np}{2}} |\Sigma|^{-\frac{n}{2}} \, \mathrm{e}^{-\frac{1}{2} \sum_{i=1}^n (\boldsymbol{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})} \\ &= (2\pi)^{-\frac{np}{2}} |\Sigma|^{-\frac{n}{2}} \, \mathrm{e}^{-\frac{1}{2} \operatorname{tr}[\Sigma^{-1} \sum_{i=1}^n (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top]} . \end{aligned}$$

Under the hypothesis H_0, the maximum of the likelihood function is obtained when $\bar{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}$. Under the alternative we have to maximise with respect to both $\boldsymbol{\mu}$ and Σ and we know from Section 3.1.2 that the maximum of the likelihood function is obtained for $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}$ and $\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top$. Then we obtain easily the likelihood ratio

$$\Lambda = \frac{\max_{\boldsymbol{\mu}} L(\boldsymbol{x}; \boldsymbol{\mu}, I_p)}{\max_{\boldsymbol{\mu}, \Sigma} L(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)} = \frac{\mathrm{e}^{[-\frac{1}{2} \operatorname{tr} \boldsymbol{V}]}}{|\boldsymbol{V}|^{-\frac{n}{2}} n^{\frac{np}{2}} \, \mathrm{e}^{-\frac{np}{2}}}$$

where $\boldsymbol{V} = \sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top$. Therefore

$$-2 \log \Lambda = np \log n - n \log |\boldsymbol{V}| + \operatorname{tr} \boldsymbol{V} - np, \tag{9.1}$$

and according to the asymptotic theory the quantity in (9.1) is asymptotically distributed as $\chi^2_{p(p+1)/2}$ (the degrees of freedom being the difference of the number of free parameters under the alternative and under the hypothesis). This test would reject H_0 if the value of the $-2 \log \Lambda$ statistic is significantly large.

9.2 Sphericity test

Further, it is more realistic to assume that the structure of the covariance matrix is only known up to some constant. Having in mind the discussion in the beginning of Section 9.1, we can

assume without loss of generality that $H_0 : \Sigma = \sigma^2 I_p$ against a general alternative. This test has the name "sphericity test". The likelihood ratio test can be developed in a manner similar to the previous case (do it (!)) and the final result is that

$$-2\log\Lambda = np\log(n\hat{\sigma}^2) - n\log|\boldsymbol{V}|.$$

Here, $\hat{\sigma}^2 = \frac{1}{np}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$. The asymptotic distribution of $np\log(n\hat{\sigma}^2) - n\log|\boldsymbol{V}|$ under the null hypothesis will be again χ^2 but the degrees of freedom are this time $\frac{p(p+1)}{2} - 1 = \frac{(p-1)(p+2)}{2}$ (WHY (?!)). Again, the hypothesis will be rejected for large values.

9.3 General situation

Testing equality of covariance matrices from k different multivariate normal populations $N_p(\boldsymbol{\mu}_i, \Sigma_i), i = 1, 2, \ldots, k$ is a very important problem especially in discriminant analysis and multivariate analysis of variance. Let,

k be the number of populations;

p the dimension of vector;

n the total sample size $n = n_1 + n_2 + \ldots n_k$,

n_i being the sample size for each population.

The analysis of deviance test statistic that results is

$$-2\log\frac{\prod_{i=1}^{k}|\hat{\boldsymbol{\Sigma}}_i|^{\frac{n_i}{2}}}{|\hat{\boldsymbol{\Sigma}}_{\text{pooled}}|^{\frac{n}{2}}},$$

with $\hat{\boldsymbol{\Sigma}}_i$ the MLE sample variance (with denominator n_i as opposed to $n_i - 1$) of population i, and $\hat{\boldsymbol{\Sigma}}_{\text{pooled}} = \frac{1}{n}\sum_{i=1}^{k} n_i\hat{\boldsymbol{\Sigma}}_i$, asymptotically distributed $\chi^2_{(k-1)p(p+1)/2}$.

It has been noticed that this test has the defect that it is (asymptotically) biased: that is, the probability of rejecting H_0 when H_0 is false can be smaller than the probability of rejecting H_0 when H_0 is true (i.e., it may happen that in some points of the parameter space the probability of a correct decision is smaller than the probability for a wrong decision). Hence it is desirable to modify it to make it asymptotically unbiased.

Further let $N = n - k$ and $N_i = n_i - 1$. Under the null hypothesis of equality of all k covariance matrices, it holds:

$$-2\rho\log\frac{\prod_{i=1}^{k}|\boldsymbol{S}_i|^{\frac{N_i}{2}}}{|\boldsymbol{S}_{\text{pooled}}|^{\frac{N}{2}}}, \tag{9.2}$$

for $\rho = 1 - [(\sum_{i=1}^{k}\frac{1}{N_i}) - \frac{1}{N}]\frac{2p^2+3p-1}{6(p+1)(k-1)}$, $\boldsymbol{S}_i$ the sample variance (with $n-1$ denominator) of population i, and $\boldsymbol{S}_{\text{pooled}} = \frac{1}{N}\sum_{i=1}^{k} N_k\boldsymbol{S}_i$, is asymptotically distributed as $\chi^2_{(k-1)p(p+1)/2}$. Large values of the statistic are significant and lead to the rejection of the hypothesis about equality of the k covariance matrices.

In the following, we will avoid the subtle details and refer to Chapter 8 of the monograph

Muirhead, R. (1982) *Aspects of Multivariate Statistical Theory.* Wiley, New York.

The *modified* LR is achieved by replacing n_i and n by N_i and N (that is, by the correct degrees of freedom). We note that indeed $\rho = 1 - [(\sum_{i=1}^{k} \frac{1}{N_i}) - \frac{1}{N}]\frac{2p^2+3p-1}{6(p+1)(k-1)}$ is close to 1 anyway if all sample sizes n_i were very large. Finally, the scaling of the test statistic by $\rho = 1 - [(\sum_{i=1}^{k} \frac{1}{N_i}) - \frac{1}{N}]\frac{2p^2+3p-1}{6(p+1)(k-1)}$ that is made in (9.2) serves to improve the quality of the asymptotic approximation of the statistic by the limiting $\chi^2_{\frac{1}{2}(k-1)p(p+1)}$ distribution. Such (asymptotically negligible) scalar transformations of the LR statistic that yield improved test statistic with a chi-squared null distribution of order $O(1/n)$ instead of the ordinary $O(1)$ for the standard LR, are known in the literature under the common name **Bartlett corrections**. Thus (9.2) is a Bartlett corrected version of the modified LR statistic.

9.4 Software

SAS: `PROC CALIS`, `PROC DISCRIM` (option)

R: `heplots::boxM`, `MVTests::BoxM`

The statistic (9.2) is the one that is implemented in software packages.

9.5 Exercises

Exercise 9.1

Follow the discussion about the sphericity test. Argue that if $\hat{\lambda}_i, i = 1, 2, \ldots, p$ denote the eigenvalues of the empirical covariance matrix S then

$$-2 \log \Lambda = np \log \frac{\text{arithm. mean } \hat{\lambda}_i}{\text{geom. mean } \hat{\lambda}_i}.$$

Of course, the above statistic is asymptotically $\chi^2_{(p+2)(p-1)/2}$ distributed under H_0 since it only represents the sphericity test in a different form.

Exercise 9.2

Show that the likelihood ratio test of

$$H_0 : \ \Sigma \text{ is a diagonal matrix}$$

rejects H_0 when $-n \log |\boldsymbol{R}|$ is larger than $\chi^2_{1-\alpha,p(p-1)/2}$. (Here $\boldsymbol{R}$ is the empirical correlation matrix, p is the dimension of the multivariate normal and n is the sample size.)

10 Factor Analysis

Let $\boldsymbol{Y}_i, i = 1, 2, .., n$ be independent $N_p(\boldsymbol{\mu}, \Sigma)$ variables (think of the $\boldsymbol{Y}_i$s as a results of a battery of p tests applied to the ith individual). Fundamental assumption in factor analysis:

$$\boldsymbol{Y}_i = \Lambda \boldsymbol{f}_i + \boldsymbol{e}_i \tag{10.1}$$

$\Lambda \in \mathcal{M}_{p,k}$ factor loading matrix (full rank);

$\boldsymbol{f}_i \in \mathbb{R}^k$ $(k < p)$ factor variable. The components of $\boldsymbol{f}_i$ are thought to be the (latent) factors. Usually $\boldsymbol{f}_i$ are taken to be independent $N(\boldsymbol{\alpha}, I_k)$ (i.e., "orthogonal") but also "oblique" factors are considered sometimes with a covariance matrix $\neq I_k$.

$\boldsymbol{e}_i$ independent $N_p(\boldsymbol{\theta}, \Sigma_e)$ with Σ_e **diagonal**, i.e., $\Sigma_e = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2)$.

Also, the $\boldsymbol{e}$s are independent of the $\boldsymbol{f}$s.

Then,

$$\boldsymbol{\mu} = \Lambda \boldsymbol{\alpha} + \boldsymbol{\theta};\ \Sigma = \Lambda \Lambda^\top + \Sigma_e,$$

or, componentwise:

$$\text{Var}(Y_{ir}) = \sum_{j=1}^{k} \lambda_{rj}^2 + \sigma_r^2 = \text{communality} + \text{uniqueness}.$$

$$\text{Cov}(Y_{ir}, Y_{is}) = \sum_{j=1}^{k} \lambda_{rj} \lambda_{sj}.$$

The fundamental idea of factor analysis is to describe the **covariance relationships** among **many** variables (p "large") in terms of few (k "small") underlying, not observable (latent) random quantities (the **factors**). The model is motivated by the following argument: suppose variables can be grouped by their correlations. That is, all variables in a particular group are highly correlated among themselves but have relatively small correlations with variables in a different group. It is then quite reasonable to assume that each group of variables represents a single underlying construct (**factor**) that is "responsible" for the observed correlations.

Important notes

- The model (10.1) is similar to a linear regression model but the key differences are that $\boldsymbol{f}_i$ are **random and are not observable**.

- If we knew the Λ (or have found estimates of them), then using properties of orthogonal projections on the linear space spanned by the columns of Λ, we would get:

$$\hat{\boldsymbol{\alpha}} = (\Lambda^\top \Lambda)^{-1}\Lambda^\top \bar{\boldsymbol{Y}};\ \hat{\boldsymbol{\theta}} = \bar{\boldsymbol{Y}} - \Lambda\hat{\boldsymbol{\alpha}}.$$

  Because of the above observation, we can consider only $\boldsymbol{\mu}$, Λ, and σ_i^2, $i = 1, 2, \ldots, p$ as unknown parameters when parameterising the factor analysis model. Note also that primary interest in factor analysis is focused on estimating Λ.

- **There is a fundamental indeterminacy** in this model even when we require that $\mathrm{Var}(\boldsymbol{f}) = I_k$ since, if $P \in \mathcal{M}_{k,k}$ **is any** orthogonal matrix then obviously

$$\Lambda\Lambda^\top = \Lambda P(\Lambda P)^\top;\ \Lambda \boldsymbol{f}_i = (\Lambda P)(P^\top \boldsymbol{f}_i).$$

  Hence replacing Λ by ΛP and $\boldsymbol{f}_i$ by $P^\top \boldsymbol{f}_i$ leads to the same equations.

10.1 ML Estimation

The likelihood function for the n observations $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n \in \mathbb{R}^p$ is

$$\begin{aligned} L(\boldsymbol{Y}; \boldsymbol{\mu}, \Lambda, \sigma_1^2, \sigma_2^2, .., \sigma_p^2) &= (2\pi)^{-np/2}|\Sigma|^{-n/2}\exp[-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{Y}_i - \boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu})] \\ &= (2\pi)^{-np/2}|\Sigma|^{-n/2}\exp[-\frac{n}{2}(\mathrm{tr}(\Sigma^{-1}\boldsymbol{S}) + (\bar{\boldsymbol{Y}} - \boldsymbol{\mu})^\top\Sigma^{-1}(\bar{\boldsymbol{Y}} - \boldsymbol{\mu}))] \end{aligned}$$

with $\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})^\top$. Taking $\log L$, we get:

$$\log L(\boldsymbol{Y}; \boldsymbol{\mu}, \Lambda, \sigma_1^2, \sigma_2^2, .., \sigma_p^2) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(|\Sigma|) - \frac{n}{2}[\mathrm{tr}(\Sigma^{-1}\boldsymbol{S}) + (\bar{\boldsymbol{Y}} - \boldsymbol{\mu})^\top\Sigma^{-1}(\bar{\boldsymbol{Y}} - \boldsymbol{\mu}))].$$

After differentiating w.r.t. $\boldsymbol{\mu}$,

$$\frac{\partial \log L}{\partial \boldsymbol{\mu}} = n\Sigma^{-1}(\bar{\boldsymbol{Y}} - \boldsymbol{\mu}) = \boldsymbol{0} \implies \hat{\boldsymbol{\mu}} = \bar{\boldsymbol{Y}}.$$

It remains to estimate Λ and Σ_e by minimising:

$$Q = \frac{1}{2}\log|\Lambda\Lambda^\top + \Sigma_e| + \frac{1}{2}\mathrm{tr}(\Lambda\Lambda^\top + \Sigma_e)^{-1}\boldsymbol{S}.$$

To implement the minimisation of Q we use the following rules for matrix differentiation:

$$\frac{\partial}{\partial \Lambda}\log|\Lambda\Lambda^\top + \Sigma_e| = 2(\Lambda\Lambda^\top + \Sigma_e)^{-1}\Lambda \tag{10.2}$$

$$\frac{\partial}{\partial A}\mathrm{tr}(A^{-1}B) = -(A^{-1}BA^{-1})^\top. \tag{10.3}$$

Applying (10.3) and the chain rule we get:

$$\frac{\partial}{\partial \Lambda}\mathrm{tr}[(\Lambda\Lambda^\top + \Sigma_e)^{-1}\boldsymbol{S}] = -2(\Lambda\Lambda^\top + \Sigma_e)^{-1}\boldsymbol{S}(\Lambda\Lambda^\top + \Sigma_e)^{-1}\Lambda.$$

Hence after substitution:

$$\frac{\partial}{\partial \Lambda} Q = (\Lambda\Lambda^\top + \Sigma_e)^{-1}\Lambda - (\Lambda\Lambda^\top + \Sigma_e)^{-1}\boldsymbol{S}(\Lambda\Lambda^\top + \Sigma_e)^{-1}\Lambda = (\Lambda\Lambda^\top + \Sigma_e)^{-1}[\Lambda\Lambda^\top + \Sigma_e - \boldsymbol{S}](\Lambda\Lambda^\top + \Sigma_e)^{-1}\Lambda = \mathbf{0}. \quad (10.4)$$

Woodbury Matrix Identity gives

$$(\Lambda\Lambda^\top + \Sigma_e)^{-1} = \Sigma_e^{-1} - \Sigma_e^{-1}\Lambda(I + \Lambda^\top\Sigma_e^{-1}\Lambda)^{-1}\Lambda^\top\Sigma_e^{-1}. \quad (10.5)$$

Hence form (10.4) and (10.5) we get

$$[\Lambda\Lambda^\top + \Sigma_e - \boldsymbol{S}]\Sigma_e^{-1}\Lambda\{I - (I + \Lambda^\top\Sigma_e^{-1}\Lambda)^{-1}\Lambda^\top\Sigma_e^{-1}\Lambda\} = \mathbf{0}. \quad (10.6)$$

Since the rank of the matrix in the curly brackets in (10.6) is full we get

$$[\Lambda\Lambda^\top + \Sigma_e - \boldsymbol{S}]\Sigma_e^{-1}\Lambda = \mathbf{0},$$

or, equivalently,

$$\boldsymbol{S}\Sigma_e^{-1}\Lambda = \Lambda(I + \Lambda^\top\Sigma_e^{-1}\Lambda).$$

The latter can also be written as

$$(\Sigma_e^{-1/2}\boldsymbol{S}\Sigma_e^{-1/2})\Sigma_e^{-1/2}\Lambda = \Sigma_e^{-1/2}\Lambda(I + \Lambda^\top\Sigma_e^{-1}\Lambda). \quad (10.7)$$

To find a particular solution, **we require $\Lambda^\top\Sigma_e^{-1}\Lambda$ to be diagonal**. Then (10.7) implies that the matrix $\Sigma_e^{-1/2}\Lambda$ has as its columns k eigenvectors that correspond to the k eigenvalues of $\Sigma_e^{-1/2}\boldsymbol{S}\Sigma_e^{-1/2}$. More subtle analysis shows that to obtain the minimum value of Q these have to be the eigenvectors that correspond to the **largest** eigenvalues of $\Sigma_e^{-1/2}\boldsymbol{S}\Sigma_e^{-1/2}$.

Based on this fact, the following iterative solution (due to Lawley) has been proposed that can be described algorithmically as follows:

1. With an initial guess $\tilde{\Sigma}_e$, calculate $\tilde{\Sigma}_e^{-1/2}\tilde{\Lambda}$ by using the eigenvectors of the k largest eigenvalues of $\tilde{\Sigma}_e^{-1/2}\boldsymbol{S}\tilde{\Sigma}_e^{-1/2}$.
2. Then from $\tilde{\Sigma}_e^{-1/2}\tilde{\Lambda}$, get a (first iteration) value for $\tilde{\Lambda}$.
3. With this value of $\tilde{\Lambda}$ we can calculate the value of $\tilde{Q}(\tilde{\Sigma}_e) = \frac{1}{2}\log|\tilde{\Lambda}\tilde{\Lambda}^\top + \tilde{\Sigma}_e| + \frac{1}{2}\operatorname{tr}(\tilde{\Lambda}\tilde{\Lambda}^\top + \tilde{\Sigma}_e)^{-1}\boldsymbol{S}$ (which is the value of the functional). This functional only depends on the p nonzero values of $\tilde{\Sigma}_e$ and there are several powerful numerical procedures to find its minimum.
4. If it is achieved at Σ_e^*, then update $\tilde{\Sigma}_e$ with the new guess Σ_e^* and repeat from Step 1 to convergence.

10.2 Hypothesis testing under multivariate normality assumption

The most interesting hypothesis is $H_0 : k$ factors against $H_1 : \neq k$ factors.

$$\log L_1 = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{S}| - \frac{np}{2}$$

$$\log L_0 = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\hat{\Sigma}| - \frac{n}{2}\operatorname{tr}(\hat{\Sigma}^{-1}\boldsymbol{S})$$

(where $\hat{\Sigma} = \hat{\Lambda}\hat{\Lambda}^\top + \hat{\Sigma}_e$). Hence $-2\log\frac{L_0}{L_1} = n[\log|\hat{\Sigma}| - \log|\boldsymbol{S}| + \operatorname{tr}(\hat{\Sigma}^{-1}\boldsymbol{S}) - p]$. The asymptotic distribution of this statistic is χ^2 with df $= \frac{p(p+1)}{2} - [pk + p - \frac{k(k-1)}{2}] = \frac{1}{2}[(p-k)^2 - p - k]$. Why?

10.3 Varimax method of rotating the factors

If $\hat{\Lambda}_0$ is the estimated factor loading matrix obtained by the ML method, we know that $\hat{\Lambda} = \hat{\Lambda}_0 P$ with any orthogonal $P \in \mathcal{M}_{k,k}$ can be used instead. How to choose a particular P such that $\hat{\Lambda}$ has some desirable properties?

Let $d_r = \sum_{i=1}^{p} \lambda_{ir}^2$, then the **varimax method of rotating the factors** consists in choosing P to maximise

$$\boldsymbol{S}_d = \sum_{r=1}^{k} \{ \sum_{i=1}^{p} (\lambda_{ir}^2 - \frac{d_r}{p})^2 \} = \sum_{r=1}^{k} \{ \sum_{i=1}^{p} \lambda_{ir}^4 - \frac{(\sum_{i=1}^{p} \lambda_{ir}^2)^2}{p} \}.$$

This corresponds to the wish to make, for each column of factor loadings, some of the coordinates to be "very large" and the rest to be "very small" (in absolute value). Iterative solution to the above rotation problem exists.

Note: Rotation of factor loadings is **particularly recommended** for loadings obtained by ML method since the initial values of $\hat{\Lambda}_0$ are constrained to satisfy the condition that $\hat{\Lambda}_0^\top \Sigma_e^{-1} \hat{\Lambda}_0$ be diagonal. This is convenient for computational purposes but may not lead to easily interpretable factors.

10.4 Relationship to Principal Component Analysis

There are different ways in which you can relate Factor analysis to Principal Component analysis. We will discuss two of them here.

10.4.1 The principal component solution of the factor model

Starting with the matrix

$$\boldsymbol{S} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})^\top$$

we can write down its spectral decomposition by using **all** of its p eigenvalues and eigenvectors. In such a way we would derive a perfect reconstruction of $\boldsymbol{S}$ but since it has been achieved by using p factors, it does not deliver any dimension reduction and is useless. We would prefer to employ a smaller number k of eigenvalues and eigenvectors of $\boldsymbol{S}$ and to get only an approximate reconstruction of $\boldsymbol{S}$

$$\boldsymbol{S} \approx \sum_{i=1}^{k} \tau_i \vec{\boldsymbol{a}}_i \vec{\boldsymbol{a}}_i^\top = \Lambda \Lambda^\top$$

whereby τ_i are the characteristic roots of $\boldsymbol{S}$, taking the k biggest ones (w.o.l.g. $\tau_1, \tau_2, .., \tau_k$) and $\boldsymbol{a}_i$ being their corresponding eigenvectors. Since the understanding is that (if k is the right number of factors) all communalities have been taken into account then $s_{ii} - \sum_{j=1}^{k} \lambda_{ij}^2$ would be the estimators of the uniquenesses. This approach shows the k factors have been extracted from $\boldsymbol{S}$ in the same way like the principal components are calculated. The method is called **the principal component solution of the factor model.**

10.4.2 The Principal Factor Solution

This is yet another method that uses similar ideas from principal components analysis. It is similar to the principal component solution, but the factor extraction is not performed directly

on $\boldsymbol{S}$. To describe it, let us assume for a moment that the uniquenesses are known (or can be estimated reasonably well) and we can decompose

$$\boldsymbol{S} = \boldsymbol{S}_r + \Sigma_e$$

whereby the number k of factors is known and Σ_e is the diagonal matrix containing the uniquenesses. Then the factor analysis model states that (an estimate of) Λ should satisfy

$$\boldsymbol{S}_r = \boldsymbol{S} - \Sigma_e = \Lambda\Lambda^\top$$

Hence Λ estimate can be found by performing principal component analysis on $\boldsymbol{S}_r$:

If $\boldsymbol{S}_r = \sum_{i=1}^{p} t_i \vec{\boldsymbol{b}}_i \vec{\boldsymbol{b}}_i^\top$, t_i being the characteristic roots of $\boldsymbol{S}_r$, take the k biggest ones (w.o.l.g. $t_1, t_2, .., t_k$). Denote

$$B = \begin{pmatrix} \boldsymbol{b}_1 & \boldsymbol{b}_2 & \cdots & \boldsymbol{b}_k \end{pmatrix};\ \Delta = \operatorname{diag}(t_1, t_2, \ldots, t_k).$$

Then $\hat{\Lambda} = B\Delta^{1/2}$. Can do it also iteratively!

This approach has some problems:

i) There is no reliable estimate of Σ_e available. (The most commonly used one in the case where $\boldsymbol{S}$ is the **correlation** matrix $\boldsymbol{R}$ is $\sigma_{ei}^2 = 1/r^{ii}$ where r^{ii} is the ith diagonal element of $\boldsymbol{R}^{-1}$.)

ii) How to select k?

Note: The methods in Section 10.4.2 are not efficient as compared to the ML method and in general, the ML method is the preferred one. However, for the ML method one has to assume normality and the alternative approaches described here are used in cases where multivariate normality in a serious doubt. Most often in practice the choice of k is done by combining subject matter knowledge, "reasonableness" of results and by looking at proportion variance explained.

10.5 Software

SAS

As you might expect, factor analysis is implemented in `PROC FACTOR`. Some remarks:

- if you want to extract different numbers of factors (the example below shows how to extract $n = 2$ factors), you should run the procedure once for each number of factors;
- the communalities need a preliminary estimate. If one considers the **correlation matrix** instead of Σ, then the communalities can be estimated by the squared multiple correlations of each of the variables with the rest (these communality estimates are used to get preliminary estimates of the uniquenesses to start the iteration process). If in the **iteration process** it happens that a communality estimate exceeds 1-the case is referred to as an **ultra-Heywood** case and the `Heywood` option sets such communality to one thus allowing iterations to continue;
- the `scree` option can be used to produce a plot of the eigenvalues Σ that is helpful in deciding how many factors to use;
- besides `method=ml` you can use `method=principal`;
- with the ML method option, the Akaike's Information criterion (and Schwarz's Bayesian Criterion) are included. These can be used to estimate the "best" number of parameters to include in a model (in case more than one model is acceptable). The number of factors that yields the smallest AIC is considered "best".

R

Function `stats::factanal()` is the built-in implementation. Package `psych` contains additional functions and utilities, as well as its own implementation, `psych::fa()`, with a number of model selection tools. Package `nFactors` contains utilities for determining the number of factors (e.g., scree plots).

10.6 Examples

Example 10.1. Data about five socioeconomic variables for 12 census data in the Los Angeles area. The five variables represent total population, median school years, total unemployment, miscellaneous professional services, and median house value. Use ML method and varimax rotation.

- Try to run the above model with $n = 3$ factors. The message "WARNING: Too many factors for a unique solution" appears. This is not surprising as the number of parameters in the model will exceed the number of elements in Σ ($\frac{1}{2}[(p-k)^2 - p - k] = -2$). In this example you can run the procedure for $n = 1$ and for $n = 2$ only (do it!) and you will see that $n = 2$ gives the adequate representation.
- Try using `psych::fa.parallel()` to search for optimal number of factors.

10.7 Additional resources

An alternative presentation of these concepts can be found in JW Ch. 9.

11 Structural Equation Modelling

Factor analysis (FA) is only one example of a new approach to data analysis which is **not based on the individual observations**. We were not able to use the regression approach since the input factors were **latent** (not observable). There were too many unknowns. We went to analyse the covariance matrix Σ (and its estimator $\boldsymbol{S}$) which involved the actual parameters of interest—σ_i^2 and Λ. That is, we switched **from the level of individual observations** to analyse covariance matrices instead. There are a **series** of methods which are based on analysis of **covariances** rather than individual cases. Instead of minimising functions of observed and predicted **individual values**, we minimise the differences between **sample covariances and covariances predicted by the model**.

The fundamental hypothesis in these analyses is

$$H_0 : \Sigma = \Sigma(\boldsymbol{\theta}) \quad \text{against} \quad H_1 : \Sigma \neq \Sigma(\boldsymbol{\theta}).$$

Here Σ has $p(p+1)/2$ unknown elements (estimated by $\boldsymbol{S}$) **but these are assumed to be reproducible by just** $k = \dim(\boldsymbol{\theta}) < p(p+1)/2$ **parameters**. Note that more generally we could consider fitting **means and covariances, or means and covariances and higher moments** to a given structure. **Regression analysis with random inputs, simultaneous equations systems, confirmatory factor analysis, canonical correlations, (M)ANOVA** can be considered special cases.

Structural equation modelling is an important statistical tool in economics and behavioural sciences. Structural equations express relationships among several variables that can be either directly observed variables (manifest variables) or unobserved hypothetical variables (latent variables). In **structural models**, as opposed to **functional models**, all variables are taken to be **random** rather than having fixed levels. In addition, for maximum likelihood estimation and generalised least squares estimation (see below), the random variables are assumed to have an approximately multivariate normal distribution. Hence you are advised to remove outliers and consider transformations to normality before fitting.

11.1 General form of the model

$$\boldsymbol{\eta} = B\boldsymbol{\eta} + \Gamma\boldsymbol{\xi} + \boldsymbol{\zeta}. \tag{11.1}$$

Here,

$\boldsymbol{\eta} \in \mathbb{R}^m$ vector of output latent variables;

$\boldsymbol{\xi} \in \mathbb{R}^{n'}$ vector of input latent variables;

$B \in \mathcal{M}_{m,m}$, $\Gamma \in \mathcal{M}_{m,n'}$ coefficient matrices;

Note: $(I - B)$ is assumed to be nonsingular.

$\boldsymbol{\zeta} \in \mathbb{R}^m$ disturbance vector with $\mathrm{E}\,\boldsymbol{\zeta} = 0$.

To this **modelling equation** (11.1) we attach 2 **measurement equations:**

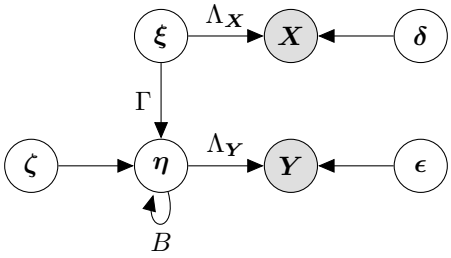
$$\boldsymbol{Y} = \Lambda_{\boldsymbol{Y}}\boldsymbol{\eta} + \boldsymbol{\epsilon}; \tag{11.2}$$

$$\boldsymbol{X} = \Lambda_{\boldsymbol{X}}\boldsymbol{\xi} + \boldsymbol{\delta}; \tag{11.3}$$

$$\boldsymbol{Y} \in \mathbb{R}^p,\ \boldsymbol{X} \in \mathbb{R}^q;\ \Lambda_{\boldsymbol{Y}} \in m_{p\times m}, \Lambda_{\boldsymbol{X}} \in m_{q\times n'}$$

with $\boldsymbol{\epsilon} \in \mathbb{R}^p,\ \boldsymbol{\delta} \in \mathbb{R}^q$ zero-mean measurement errors. These errors are assumed to be uncorrelated with $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ and with each other.

Generative model for $\boldsymbol{X}$ and $\boldsymbol{Y}$



The above quite general model (11.1)–(11.2)–(11.3) is called **Keesling–Wiley–Jöreskog** model. Its interpretation is that the input and output latent variables $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ are connected by a system of linear equations (the structural model (11.1)) with coefficient matrices B and Γ and an error vector $\boldsymbol{\zeta}$. The random vectors $\boldsymbol{Y}$ and $\boldsymbol{X}$ represent the observable vectors (measurements).

The implied covariance matrix for this model can be obtained. Let

$$\mathrm{Var}(\boldsymbol{\xi}) = \Phi;\ \mathrm{Var}(\boldsymbol{\zeta}) = \Psi;\ \mathrm{Var}(\boldsymbol{\epsilon}) = \boldsymbol{\theta}_{\boldsymbol{\epsilon}};\ \mathrm{Var}(\boldsymbol{\delta}) = \boldsymbol{\theta}_{\boldsymbol{\delta}}.$$

Then,

$$\begin{aligned}\Sigma = \Sigma(\boldsymbol{\theta}) &= \begin{pmatrix}\Sigma_{\boldsymbol{YY}}(\boldsymbol{\theta}) & \Sigma_{\boldsymbol{YX}}(\boldsymbol{\theta})\\ \Sigma_{\boldsymbol{XY}}(\boldsymbol{\theta}) & \Sigma_{\boldsymbol{XX}}(\boldsymbol{\theta})\end{pmatrix}\\ &= \begin{pmatrix}\Lambda_{\boldsymbol{Y}}(I-B)^{-1}(\Gamma\Phi\Gamma^\top + \Psi)[(I-B)^{-1}]^\top\Lambda_{\boldsymbol{Y}}^\top + \boldsymbol{\theta}_{\boldsymbol{\epsilon}} & \Lambda_{\boldsymbol{Y}}(I-B)^{-1}\Gamma\Phi\Lambda_{\boldsymbol{X}}^\top\\ \Lambda_{\boldsymbol{X}}\Phi\Gamma^\top[(I-B)^{-1}]^\top\Lambda_{\boldsymbol{Y}}^\top & \Lambda_{\boldsymbol{X}}\Phi\Lambda_{\boldsymbol{X}}^\top + \boldsymbol{\theta}_{\boldsymbol{\delta}}\end{pmatrix}.\end{aligned} \tag{11.4}$$

11.2 Estimation

Under the normality assumption, we can use the MLE. Since the "data" is the estimated covariance matrix $\boldsymbol{S} = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}\left\{\begin{pmatrix}\boldsymbol{Y}_i - \bar{\boldsymbol{Y}}\\ \boldsymbol{X}_i - \bar{\boldsymbol{X}}\end{pmatrix}\begin{pmatrix}\boldsymbol{Y}_i - \bar{\boldsymbol{Y}}\\ \boldsymbol{X}_i - \bar{\boldsymbol{X}}\end{pmatrix}^\top\right\}$, and since it is known that $(n-1)\boldsymbol{S} \sim W_{p+q}(n-1, \Sigma)$, we can utilise the form of the Wishart density to derive that

$$\log L(\boldsymbol{S}, \Sigma(\boldsymbol{\theta})) = \text{constant} - \frac{n-1}{2}\{\log|\Sigma(\boldsymbol{\theta})| + \mathrm{tr}[\boldsymbol{S}\Sigma^{-1}(\boldsymbol{\theta})]\}.$$

This is the function that has to be maximised. Hence, to find MLE, we minimise

$$F_{\mathrm{ML}}(\boldsymbol{\theta}) = \log|\Sigma(\boldsymbol{\theta})| + \mathrm{tr}[\boldsymbol{S}\Sigma^{-1}(\boldsymbol{\theta})] - \log|\boldsymbol{S}| - (p+q). \tag{11.5}$$

The function (11.5) has the advantage that F_{ML} would be zero for the "saturated model" (with $\hat{\Sigma} = \boldsymbol{S}$). I.e., a perfect fit is indicated by zero (and any non-perfect fit gives rise to > 0 value of F_{ML}).

11.3 Model evaluation

Under normality, model adequacy is mostly tested by an asymptotic χ^2-test. Under $H_0 : \Sigma = \Sigma(\boldsymbol{\theta})$ versus $H_1 : \Sigma \neq \Sigma(\boldsymbol{\theta})$, the statistic to be used is $T = (n-1)F_{\mathrm{ML}}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}})$ and under H_0, its asymptotic distribution is χ^2 with $\mathrm{df} = \frac{(p+q)(p+q+1)}{2} - \dim(\boldsymbol{\theta})$.

Reason:

$$\begin{aligned}\log L_0 &= \log L(\boldsymbol{S}, \hat{\Sigma}_{\mathrm{MLE}}) = \log L(\boldsymbol{S}, \Sigma(\hat{\boldsymbol{\theta}}_{ML})) \\ &= -\frac{n-1}{2}\{\log|\hat{\Sigma}_{\mathrm{MLE}}| + \mathrm{tr}[\boldsymbol{S}\hat{\Sigma}_{\mathrm{MLE}}^{-1}]\} + \text{constant};\end{aligned}$$

$$\log L_1 = \log L(\boldsymbol{S}, \boldsymbol{S}) = -\frac{n-1}{2}\{\log|\boldsymbol{S}| + (p+q)\} + \text{constant}.$$

Then,

$$\begin{aligned}-2\log\frac{L_0}{L_1} &= (n-1)\{\log|\hat{\Sigma}_{\mathrm{MLE}}| + \mathrm{tr}(\boldsymbol{S}\hat{\Sigma}_{\mathrm{MLE}}^{-1}) - \log|\boldsymbol{S}| - (p+q)\} \\ &= (n-1)F_{\mathrm{ML}}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}).\end{aligned}$$

11.4 Some particular SEM

From the general model (11.1)–(11.2)–(11.3), we can obtain following particular models:

A) $\Lambda_{\boldsymbol{Y}} = I_m$, $\Lambda_{\boldsymbol{X}} = I_{n'}$; $p = m$; $q = n'$; $\boldsymbol{\theta}_{\boldsymbol{\epsilon}} = 0$; $\boldsymbol{\theta}_{\boldsymbol{\delta}} = 0 \implies \boldsymbol{Y} = B\boldsymbol{Y} + \Gamma\boldsymbol{X} + \boldsymbol{\zeta}$ (the classical econometric model).

B) $\Lambda_{\boldsymbol{Y}} = I_p$, $\Lambda_{\boldsymbol{X}} = I_q \implies$ The measurement error model:

- $\boldsymbol{\eta} = B\boldsymbol{\eta} + \Gamma\boldsymbol{\xi} + \boldsymbol{\zeta}$
- $\boldsymbol{Y} = \boldsymbol{\eta} + \boldsymbol{\epsilon}$
- $\boldsymbol{X} = \boldsymbol{\xi} + \boldsymbol{\delta}$

C) Factor Analysis Models: Just take the measurement part $\boldsymbol{X} = \Lambda_{\boldsymbol{X}}\boldsymbol{\xi} + \boldsymbol{\delta}$.

11.5 Relationship between exploratory and confirmatory FA

In EFA the number of latent variables is not determined in advance; further, the measurement errors are assumed to be uncorrelated. In CFA a model is constructed to a great extent **in advance**, the number of latent variables $\boldsymbol{\xi}$ is set by the analyst, whether a latent variable influences an observed variable is specified, some direct effects of latent on observed values are fixed to zero or some other constant (e.g., one), measurement errors $\boldsymbol{\delta}$ may correlate, the covariance of latent variables can be either estimated or set to any value. In practice, distinction between EFA and CFA is more blurred. For instance, researchers using traditional EFA procedures may restrict their analysis to a group of indicators that they believe are influenced by one factor. Or, researchers with poorly fitting models in CFA often modify their model in an exploratory way with the goal of improving fit.

11.6 Software

SAS

In SAS, the standard `PROC CALIS` is used for fitting Structural Equation Models, and it has been significantly upgraded in SAS 9.3. In particular, now you can analyse **means and covariance (or even higher order)** structures (instead of just covariance structures like in the classical SEM).

R

There are two packages for SEM in R: `lavaan` and `sem`. `sem` is an older package, whereas `lavaan` aims to provide an extensible framework for SEMs and their extensions:

- can mimic commercial packages (including those below)
- provides convenience functions for specifying simple special cases (such as CFA) but also a more flexible interface for advanced users
- mean structures and multiple groups
- different estimators and standard errors (including robust)
- handling of missing data
- linear and nonlinear equality and inequality constraints
- categorical data support
- multilevel SEMs
- package `blavaan` for Bayesian estimation
- etc.

Others

Note that the general form of the SEM model given here is only one possible description due to Karl Jöreskog. His paradigm has been first implemented in the software called `LISREL` (**Li**near **S**tructural **Rel**ationships).

There are other equivalent descriptions due to Bentler and Weeks, to McDonald and some other prominent researchers in the field. Some of them also have proposed their own software for fitting SEM models according to their model specification. The `EQS` program for PC that deals with the Bentler/Weeks model, was very popular for a while. The latest "hit" in the area is the program `MPLUS` (M is for Bengt Muthén). Mutén is a former PhD student of Jöreskog and has been the developer of LISREL. During the last 15 years or so however, he has developed his own program `MPLUS`. Its latest version 6 represents a fully integrated framework and is the premier software in the area of general latent variable modelling specifically in the behavioural sciences. `MPLUS` capabilities include:

- Exploratory factor analysis
- Structural equation modelling
- Item response theory analysis
- Growth curve modelling

- Mixture modelling (latent class analysis)
- Longitudinal mixture modelling (hidden Markov, latent transition analysis, latent class growth analysis, growth mixture analysis)
- Survival analysis (continuous- and discrete-time)
- Multilevel analysis
- Bayesian analysis
- etc.

11.7 Examples

Example 11.1. Wheaton, Muthen, Alwin, and Summers (1977) Anomie example.

12 Discrimination and Classification

12.1 Separation and Classification for two populations

Discriminant analysis and classification are widely used multivariate techniques. The goal is either *separating sets of objects* (in discriminant analysis terminology) or *allocating new objects to given groups* (in classification theory terminology).

Basically, discriminant analysis is more exploratory in nature than classification. However, the difference is not significant especially because very often a function that separates may sometimes serve as an allocator, and, conversely, a rule of allocation may suggest a discriminatory procedure. In practice, the goals in the two procedures often overlap. We will consider the case of two populations (classes of objects) first.

Typical examples include: an anthropologist wants to classify a skull as a male or female; a patient needs to be classified as needing surgery or not needing surgery etc..

Denote the two classes by π_1 and π_2. The separation is to be performed on the basis of measurements of p associated random variables that form a vector $\boldsymbol{X} \in \mathbb{R}^p$. The observed values of $\boldsymbol{X}$ belong to different distributions when taken from π_1 and π_2 and we shall denote the densities of these two distributions by $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$, respectively.

Allocation or classification is possible due to the fact that one has a *learning sample* at hand, i.e., there are some measurement vectors that are known to have been generated from each of the two populations. These measurements have been generated in earlier similar experiments. The goal is to partition the sample space into 2 mutually exclusive regions, say R_1 and R_2, such that if a *new* observation falls in R_1, it is allocated to π_1 and if it falls in R_2, it is allocated to π_2.

12.2 Classification errors

There is always a chance of an erroneous classification (misclassification). Our goal will be to develop such classification methods that in a suitably defined sense minimise the chances of misclassification.

It should be noted that one of the two classes may have a greater likelihood of occurrence because one of the two populations might be much larger than the other. For example, there tend

to be a lot more financially sound companies than bankrupt companies. These *prior probabilities* of occurrence should also be taken into account when constructing an optimal classification rule if we want to perform optimally.

In a more detailed study of optimal classification rules, cost is also important. If classifying a π_1 object to the class π_2 represents a much more serious error than classifying a π_2 object to the class π_1 then these cost differences should also be taken into account when designing the optimal rule.

The **conditional** probabilities for misclassification are defined naturally as:

$$\Pr(2|1) = \Pr(\boldsymbol{X} \in R_2|\pi_1) = \int_{R_2} f_1(\boldsymbol{x})d\boldsymbol{x} \tag{12.1}$$

$$\Pr(1|2) = \Pr(\boldsymbol{X} \in R_1|\pi_2) = \int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x} \tag{12.2}$$

12.3 Summarising

We turn briefly to the question of how to summarise a classifier's performance. Each object has a true class membership and the one predicted by the classifier, and for a given dataset for which true memberships are known, we may summarise the counts of the four resulting possibilities in a contingency table called a *confusion matrix*, i.e.,

		Predicted class	
		1	2
Actual class	1	Members of 1 correctly classified	Members of 1 misclassified as 2
	2	Members of 2 misclassified as 1	Members of 2 correctly classified

A confusion matrix can be produced when there are more than two classes as well.
In the special case where there are two classes that can be meaningfully labelled as Negative/Positive, False/True, No/Yes, Null/Alternative, or similar, it is common to use the following terminology for them:

		Predicted class	
		Negative	Positive
Actual class	Negative	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (TP)

One can then define various performance metrics such as

sensitivity (a.k.a. recall, true positive rate (TPR)): $\Pr(\text{Pred. pos.}|\text{Act. pos.}) = \frac{\text{TP}}{\text{TP+FN}}$

specificity (a.k.a. selectivity, true negative rate (TNR)): $\Pr(\text{Pred. neg.}|\text{Act. neg.}) = \frac{\text{TN}}{\text{TN+FP}}$

false positive rate (a.k.a. FPR, fall-out): $\Pr(\text{Pred. pos.}|\text{Act. neg.}) = \frac{\text{FP}}{\text{TN+FP}} = 1 - \text{TNR}$

accuracy: $\frac{\text{TP+TN}}{\text{TP+FP+TN+FN}}$

total probability of misclassification (a.k.a. TPM): $1 - \text{accuracy}$

precision (a.k.a. positive predictive value): $\Pr(\text{Act. pos.}|\text{Pred. pos.}) = \frac{\text{TP}}{\text{TP+FP}}$

negative predictive value: $\Pr(\text{Act. neg.}|\text{Pred. neg.}) = \frac{\text{TN}}{\text{TN+FN}}$

F1 score: $\frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}}$

Many classifiers return a continuous score that needs to be thresholded to produce a binary decision (e.g., predict "Yes" if the score exceeds some constant k and "No" otherwise), it is a common practice to plot a *receiver operating characteristic* (ROC) curve by varying the threshold and then plotting the TPR (on the vertical axis) against FPR (on the horizontal axis) that result. Both of which decrease as k increases. A perfect classifier would have a threshold for which the curve achieves the $(0, 1)$ point, whereas classifier close to the $y = x$ line is no better than chance.

12.4 Optimal classification rules

12.4.1 Rules that minimise the expected cost of misclassification (ECM)

Lemma 12.1. *Denote by p_i the **prior** probability of $\pi_i, i = 1, 2, p_1 + p_2 = 1$. Then the **overall** probabilities of incorrectly classifying objects will be:* $\Pr(\textit{misclassified as } \pi_1) = \Pr(1|2)p_2$ *and* $\Pr(\textit{misclassified as } \pi_2) = \Pr(2|1)p_1$. *Further, let $c(i|j), i \neq j, i, j = 1, 2$ be the misclassification costs. Then the **expected cost of misclassification** is*

$$ECM = c(2|1)\Pr(2|1)p_1 + c(1|2)\Pr(1|2)p_2 \tag{12.3}$$

The regions R_1 and R_2 that minimise ECM are given by

$$R_1 = \{\boldsymbol{x} : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}\} \tag{12.4}$$

and

$$R_2 = \{\boldsymbol{x} : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}\}. \tag{12.5}$$

Proof. It is easy to see that $\text{ECM} = \int_{R_1}[c(1|2)p_2 f_2(\boldsymbol{x}) - c(2|1)p_1 f_1(\boldsymbol{x})]d\boldsymbol{x} + c(2|1)p_1$. Hence, the ECM will be minimised if R_1 includes those values of $\boldsymbol{x}$ for which the integrand $[c(1|2)p_2 f_2(\boldsymbol{x}) - c(2|1)p_1 f_1(\boldsymbol{x})] \leq 0$ and excludes all the complementary values.

Note the significance of the fact that in Lemma 12.1 **only ratios** are involved. Often in practice, one would have a much clearer idea about the cost ratio rather than for the actual costs themselves. □

For your own exercise, consider the partial cases of Lemma 12.1 when $p_2 = p_1$, $c(1|2) = c(2|1)$ and when both these equalities hold. Comment on the soundness of the classification regions in these cases.

12.4.2 Rules that minimise the total probability of misclassification (TPM)

If we ignore the cost of misclassification, we can define the total probability of misclassification as

$$\text{TPM} = p_1 \int_{R_2} f_1(\boldsymbol{x})d\boldsymbol{x} + p_2 \int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x}$$

Mathematically, this is a particular case of Lemma 12.1 when the costs of misclassification are equal—so nothing new here.

12.4.3 Bayesian approach

Here, we try to allocate a new observation $\boldsymbol{x}_0$ to the population with the larger posterior probability $\Pr(\pi_i|\boldsymbol{x}_0), i = 1, 2$. According to Bayes's formula we have

$$\Pr(\pi_1|\boldsymbol{x}_0) = \frac{p_1 f_1(\boldsymbol{x}_0)}{p_1 f_1(\boldsymbol{x}_0) + p_2 f_2(\boldsymbol{x}_0)},\ \Pr(\pi_2|\boldsymbol{x}_0) = \frac{p_2 f_2(\boldsymbol{x}_0)}{p_1 f_1(\boldsymbol{x}_0) + p_2 f_2(\boldsymbol{x}_0)}$$

Mathematically, the strategy of classifying an observation $\boldsymbol{x}_0$ as π_1 if $\Pr(\pi_1|\boldsymbol{x}_0) > \Pr(\pi_2|\boldsymbol{x}_0)$ is again a particular case of Lemma 12.1 when the costs of misclassification are equal. (**Why?**) But note that the calculation of the posterior probabilities themselves is in itself a useful and informative operation.

12.5 Classification with two multivariate normal populations

Until now we did not specify any particular form of the densities $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$. Essential simplification occurs under normality assumption and we are going over to a more detailed discussion of this particular case now. Two different cases will be considered- of equal and of non-equal covariance matrices.

12.5.1 Case of equal covariance matrices $\Sigma_1 = \Sigma_2 = \Sigma$

Now we assume that the two populations π_1 and π_2 are $N_p(\boldsymbol{\mu}_1, \Sigma)$ and $N_p(\boldsymbol{\mu}_2, \Sigma)$, respectively. Then, (12.4) becomes

$$R_1 = \{\boldsymbol{x} : \exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)] \geq \frac{c(1|2)}{c(2|1)} \times \frac{p_2}{p_1}\}.$$

Similarly, from (12.5) we get

$$R_2 = \{\boldsymbol{x} : \exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)] < \frac{c(1|2)}{c(2|1)} \times \frac{p_2}{p_1}\},$$

and we arrive at the following result:

Theorem 12.2. *Under the above assumptions, the allocation rule that minimises the ECM is given by:*

1. allocate $\boldsymbol{x}_0$ to π_1 if

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}\boldsymbol{x}_0 - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq \log[\frac{c(1|2)}{c(2|1)} \times \frac{p_2}{p_1}].$$

2. Otherwise, allocate $\boldsymbol{x}_0$ to π_2.

Proof. Simple exercise (to be discussed at lectures). □

Note also that it is unrealistic to assume in most situations that the parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and Σ are known. They will need to be estimated by the data instead. Assume, n_1 and n_2 observations are available from the first and from the second population, respectively. If $\bar{\boldsymbol{x}}_1$ and $\bar{\boldsymbol{x}}_2$ are the sample mean vectors and $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ the corresponding sample covariance matrices, then under the assumption of $\Sigma_1 = \Sigma_2 = \Sigma$ we can derive the pooled covariance matrix estimator $\boldsymbol{S}_{\text{pooled}} = \frac{(n_1-1)\boldsymbol{S}_1 + (n_2-1)\boldsymbol{S}_2}{n_1+n_2-2}$ (This is an unbiased estimator of Σ (!)).

Hence the *sample classification rule* becomes:

1. allocate $\boldsymbol{x}_0$ to π_1 if

$$(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{\text{pooled}}^{-1} \boldsymbol{x}_0 - \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{\text{pooled}}^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) \geq \log[\frac{c(1|2)}{c(2|1)} \times \frac{p_2}{p_1}] \tag{12.6}$$

2. Otherwise, allocate $\boldsymbol{x}_0$ to π_2.

This empirical classification rule is called **an allocation rule based on Fisher's discriminant function**. The function

$$(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{\text{pooled}}^{-1} \boldsymbol{x}_0 - \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{\text{pooled}}^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2)$$

itself (which is linear in the vector observation $\boldsymbol{x}_0$) is called **Fisher's linear discriminant function.**

Of course, the latter rule is only an *estimate* of the optimal rule since the parameters in the latter have been replaced by estimated quantities. But we are expecting this rule to perform well when n_1 and n_2 are large. It is to be pointed out that the allocation rule in (12.6) is **linear** in the new observation $\boldsymbol{x}_0$. The simplicity of its form is a consequence of the multivariate normality assumption.

- Allocation rule based on *Fisher's discriminant function*:

$$(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{\text{pooled}}^{-1} \boldsymbol{x}_0 - \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{\text{pooled}}^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2)$$

- Function itself called *Fisher's linear discriminant function.*
- Only an *estimate* of the optimal rule.
  - **linear** in the new observation $\boldsymbol{x}_0$

12.5.2 Case of different covariance matrices ($\Sigma_1 \neq \Sigma_2$)

Theorem 12.3. *Now we assume that the two populations π_1 and π_2 are $N_p(\boldsymbol{\mu}_1, \Sigma_1)$ and $N_p(\boldsymbol{\mu}_2, \Sigma_2)$, respectively. Repeating the same steps as in Theorem 12.2 we get*

$$R_1 = \{\boldsymbol{x} : -\frac{1}{2}\boldsymbol{x}^\top(\Sigma_1^{-1} - \Sigma_2^{-1})\boldsymbol{x} + (\boldsymbol{\mu}_1^\top \Sigma_1^{-1} - \boldsymbol{\mu}_2^\top \Sigma_2^{-1})\boldsymbol{x} - k \geq \log[\frac{c(1|2)}{c(2|1)} \times \frac{p_2}{p_1}]\}$$

$$R_2 = \{\boldsymbol{x} : -\frac{1}{2}\boldsymbol{x}^\top(\Sigma_1^{-1} - \Sigma_2^{-1})\boldsymbol{x} + (\boldsymbol{\mu}_1^\top \Sigma_1^{-1} - \boldsymbol{\mu}_2^\top \Sigma_2^{-1})\boldsymbol{x} - k < \log[\frac{c(1|2)}{c(2|1)} \times \frac{p_2}{p_1}]\}$$

*where $k = \frac{1}{2}\log(\frac{|\Sigma_1|}{|\Sigma_2|}) + \frac{1}{2}(\boldsymbol{\mu}_1^\top \Sigma_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top \Sigma_2^{-1}\boldsymbol{\mu}_2)$ and we see that the classification regions are **quadratic** functions of the new observation in this case.*

One obtains the following rule:

1. allocate $\boldsymbol{x}_0$ to π_1 if

$$-\frac{1}{2}\boldsymbol{x}_0^\top(\boldsymbol{S}_1^{-1} - \boldsymbol{S}_2^{-1})\boldsymbol{x}_0 + (\bar{\boldsymbol{x}}_1^\top \boldsymbol{S}_1^{-1} - \bar{\boldsymbol{x}}_2^\top \boldsymbol{S}_2^{-1})\boldsymbol{x}_0 - \hat{k} \geq \log[\frac{c(1|2)}{c(2|1)} \times \frac{p_2}{p_1}]$$

where $\hat{k}$ is the empirical analog of k.

2. Allocate $\boldsymbol{x}_0$ to π_2 otherwise.

When $\Sigma_1 = \Sigma_2$, the quadratic term disappears and we can easily see that the classification regions from Theorem 12.2 are obtained. Of course, the case considered in Theorem 12.3 is more general but we should be cautious when applying it in practice. It turns out that in more than two dimensions, classification rules based on quadratic functions do not always perform nicely and can lead to strange results. This is especially true when the data are not quite normal and when the differences in the covariance matrices are significant. The rule is very sensitive (non-robust) towards departures from normality. Therefore, it is advisable to try to first transform the data to more nearly normal by using some classical normality transformations. A detailed discussion of these effects will be provided during the lecture. Also, tests discussed in Lecture 9 can be used to check if equal variance assumption is valid.

12.5.3 Optimum error rate and Mahalanobis distance

We defined the TPM quantity in general terms for any classification rule (12.3). When the regions R_1 and R_2 are selected in an optimal way, one obtains the minimal value of TPM which is called **optimum error rate (OER)** and is being used to characterise the difficulty of the classification problem at hand. Hereby we shall illustrate the calculation of the OER for the simple case of two normal populations with $\Sigma_1 = \Sigma_2 = \Sigma$ and prior probabilities $p_1 = p_2 = \frac{1}{2}$. In this case

$$\text{TPM} = \frac{1}{2}\int_{R_2} f_1(\boldsymbol{x})d\boldsymbol{x} + \frac{1}{2}\int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x},$$

and OER is obtained by choosing

$$R_1 = \{\boldsymbol{x} : (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq 0\}$$

and

$$R_2 = \{\boldsymbol{x} : (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) < 0\}.$$

If we introduce the random variable $Y = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}\boldsymbol{X} = \boldsymbol{l}^\top \boldsymbol{X}$ then $Y|i \sim N_1(\mu_{iY}, \Delta^2), i = 1, 2$ for the two populations π_1 and π_2 where $\mu_{iY} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}\boldsymbol{\mu}_i, i = 1, 2$. The quantity $\Delta = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}$ is the **Mahalanobis distance** between the two normal populations and it has an important role in many applications of Multivariate Analysis. Now

$$\Pr(2|1) = \Pr(Y < \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)) = \Pr(\frac{Y - \mu_{1Y}}{\Delta} < -\frac{\Delta}{2}) = \Phi(-\frac{\Delta}{2}),$$

$\Phi(\cdot)$ denoting the cumulative distribution function of the standard normal. Along the same lines we can get (**do it (!)**) : $\Pr(1|2) = \Phi(-\frac{\Delta}{2})$ to that finally OER = minimum TPM = $\Phi(-\frac{\Delta}{2})$.

In practice, Δ is replaced by its estimated value $\hat{\Delta} = \sqrt{(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{\text{pooled}}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)}$.

12.6 Classification with more than 2 normal populations

Formal generalisation of the theory for the case of $g > 2$ groups $\pi_1, \pi_2, \ldots, \pi_g$ is straightforward but optimal error rate analysis is difficult when $g > 2$. It is easy to see that the ECM classification rule with **equal** misclassification costs becomes (compare to (12.4) and (12.5)) now:

1. Allocate $\boldsymbol{x}_0$ to π_k if $p_k f_k > p_i f_i$ for all $i \neq k$.

Equivalently, one can check if $\log p_k f_k > \log p_i f_i$ for all $i \neq k$.

When applying this classification rule to g normal populations $f_i(\boldsymbol{x}) \sim N_p(\boldsymbol{\mu}_i, \Sigma_i), i = 1, 2, \ldots, g$ it becomes:

1. Allocate $\boldsymbol{x}_0$ to π_k if

$$\log p_k f_k(\boldsymbol{x}_0) = \log p_k - \frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(\boldsymbol{x}_0 - \boldsymbol{\mu}_k)^\top \Sigma_k^{-1}(\boldsymbol{x}_0 - \boldsymbol{\mu}_k) = \max_i \log p_i f_i(\boldsymbol{x}_0).$$

Ignoring the constant $\frac{p}{2}\log(2\pi)$ we get the **quadratic discriminant score for the ith population**:

$$d_i^Q(\boldsymbol{x}) = -\frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^\top \Sigma_i^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i) + \log p_i \tag{12.7}$$

and the rule advocates to allocate $\boldsymbol{x}$ to the population with a largest quadratic discriminant score. It is obvious how one would estimate from the data the unknown quantities involved in (12.7) in order to obtain the *estimated* minimum total probability of misclassification rule. (You formulate the precise statement (!)).

In the case we are justified to assume that **all covariance matrices** for the g populations are equal, a simplification is possible (like in the case $g = 2$). Looking only at the terms that vary with $i = 1, 2, \ldots, g$ in (12.7) we can define the **linear discriminant score**: $d_i(\boldsymbol{x}) = \boldsymbol{\mu}_i^\top \Sigma^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_i^\top \Sigma^{-1}\boldsymbol{\mu}_i + \log p_i$. Correspondingly, a **sample version** of the linear discriminant score is obtained by substituting the arithmetic means $\bar{\boldsymbol{x}}_i$ instead of $\boldsymbol{\mu}_i$ and $\boldsymbol{S}_{\text{pooled}} = \frac{n_1 - 1}{n_1 + n_2 + \ldots n_g - g}\boldsymbol{S}_1 + \cdots + \frac{n_g - 1}{n_1 + n_2 + \cdots + n_g - g}\boldsymbol{S}_g$ instead of Σ thus arriving at

$$\hat{d}_i(\boldsymbol{x}) = \bar{\boldsymbol{x}}_i^\top \boldsymbol{S}_{\text{pooled}}^{-1}\boldsymbol{x} - \frac{1}{2}\bar{\boldsymbol{x}}_i^\top \boldsymbol{S}_{\text{pooled}}^{-1}\bar{\boldsymbol{x}}_i + \log p_i$$

Therefore the **Estimated Minimum TPM Rule for Equal Covariance Normal Populations** is the following:

1. Allocate $\boldsymbol{x}$ to π_k if $\hat{d}_k(\boldsymbol{x})$ is the largest of the g values $\hat{d}_i(\boldsymbol{x}), i = 1, 2, \ldots, g$.

In this form, the classification rule has been implemented in many computer packages.

12.7 Software

SAS: `PROC DISCRIM`

R: `MASS:lda`, `MASS:qda`

12.8 Examples

Example 12.4. Linear and quadratic discriminant analysis for the Edgar Anderson's Iris data, and using cross-validation to assess classifiers.

12.9 Additional resources

An alternative presentation of these concepts can be found in JW Sec. 11.1–11.6.

12.10 Exercises

Exercise 12.1

Three bivariate normal populations, labelled $i = 1, 2, 3$ have same covariance matrix given by $\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$ and means $\boldsymbol{\mu}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $\boldsymbol{\mu}_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \boldsymbol{\mu}_3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, respectively.

(a) Suggest a classification rule for an observation $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ that corresponds to one of the three populations. You may assume equal priors for the three populations and equal misclassification costs.

(b) Classify the following observations to one of the three distributions: $\begin{pmatrix} 0.2 \\ 0.6 \end{pmatrix}, \begin{pmatrix} 2 \\ 0.8 \end{pmatrix}, \begin{pmatrix} 0.75 \\ 1 \end{pmatrix}$.

(c) Show that in $\mathbb{R}^2$, the 3 classification regions are bounded by straight lines and draw a graph of these three regions.

13 Support Vector Machines

13.1 Introduction and motivation

As seen in Lecture 12, when classifying into one of two p-dimensional multivariate normal populations, the scores are either linear (when the same covariance matrices are used) or quadratic (when the covariance matrices are different). Even optimality for such simple classifiers could be shown due to the multivariate normality assumption. However, when the two populations are **not** multivariate normal, the situation is more difficult, the bounds between the populations may be more blurry and significantly more non-linear classification techniques may be necessary to achieve a good classification. Support vector machines (SVM) are an example of such non-linear statistical classification techniques. They usually achieve superior results in comparison to more traditional non-linear parametric classification techniques such as *logit analysis* or non-parametric techniques such as *neural networks*. Mathematically, when using SVM, we try to formulate the classification as an empirical risk minimisation problem and to solve the problem under additional restrictions on the allowed (nonlinear) classifier functions.

13.2 Expected versus Empirical Risk minimisation

Let Y be and "indicator" with values $+1$ and -1 that indicate if certain p dimensional observation belongs to one of two groups of interest. We want to find a "best" classifier in a class $\mathcal{F}$ of functions f. Each classifier function $f(\boldsymbol{x})$ is meant to deliver a value of $+1$ or -1 for a given observation vector $\boldsymbol{x}$. To this end, we consider the *expected risk*

$$R(f) = \int \frac{1}{2}|f(\boldsymbol{x}) - y| dP(\boldsymbol{x}, y)$$

Since the joint distribution $P(\boldsymbol{x}, y)$ is unknown in practice, we consider the *empirical risk* over a training set $(\boldsymbol{x}_i, y_i), i = 1, 2, \ldots, n$ of observations instead:

$$\hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2}|f(\boldsymbol{x}_i) - y_i|$$

The loss in the risk's definition is the "zero-one loss" given by

$$L(\boldsymbol{x}, y) = \frac{1}{2}|f(\boldsymbol{x}) - y|$$

and, thanks to the chosen labels ± 1 for Y obviously has the values 0 (if classification is correct) and 1 (if classification is wrong).

Minimising the empirical (instead of the unknown expected) risk means to find $f_n = \arg\min_{f\in\mathcal{F}} \hat{R}(f)$ as an approximation to $f_{\text{opt}} = \arg\min_{f\in\mathcal{F}} R(f)$. Generally speaking the two solutions f_n and f_{opt} do not coincide and without further assumptions may be quite different. However, thanks to some ground breaking work by V. Vapnik there are theoretical results which, loosely speaking, state that if $\mathcal{F}$ is not too large and $n \to \infty$, there is an upper bound on their difference with probability $(1-\eta)$:

$$R(f) \le \hat{R}(f) + \phi(\frac{h}{n}, \frac{\log\eta}{n})$$

The above inequality can be interpreted as stating that the test error is bounded from above by the sum of the training error and the complexity of the set of models under consideration. We can then try to minimise the bound from above and hope that in that way we keep under control to a minimum the (unknown) test error.

The function ϕ above is monotone increasing in h (at least for large enough sample sizes n). Here h denotes the Vapnik–Chervonenkis **(VC) dimension** (i.e., a measure of the complexity of the class $\mathcal{F}$).

For a linear classification rule $f(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^\top\boldsymbol{x} + b)$ with a p dimensional predictor $\boldsymbol{x}$ it is known that

$$\phi(\frac{h}{n}, \frac{\log\eta}{n}) = \sqrt{\frac{h(\log(\frac{2n}{h})+1) - \log(\frac{\eta}{4})}{n}}.$$

and that the VC dimension is $h = p + 1$. You can now **directly check** that

$$\frac{\partial}{\partial h}[\frac{h(\log(\frac{2n}{h})+1) - \log(\frac{\eta}{4})}{n}] = \frac{1}{n}\log(\frac{2n}{h}) > 0$$

as long as $h < 2n$ which confirms the monotone increasing property stated above.

In general, the VC dimension of a given set of functions is equal to the maximal number of points that can be separated *in all possible ways* by that set of functions.

At first glance, the "more rich" the function class $\mathcal{F}$ the better the classification rule would be. Indeed you can construct a classifier that has zero classification error on the training set. However, this classifier will be too specialised for the given training set with no ability to generalise for other sets. Hence such a classifier would be undesirable.

At first glance, the "more rich" the function class $\mathcal{F}$ the better the classification rule would be. Indeed you can construct a classifier that has zero classification error on the training set. However, this classifier will be too specialised for the given training set with no ability to generalise for other sets. Hence such a classifier would be undesirable. "More rich" is tantamount to require bigger complexity of $\mathcal{F}$ or equivalently higher value of h (and therefore of ϕ). The term $\phi(\frac{h}{n}, \frac{\log\eta}{n})$ can be considered penalty for the excessive complexity of the classifier function. You can see directly that the derivative $\frac{\partial\phi(\frac{h}{n}, \frac{\log(\eta)}{n})}{\partial h} \ge 0$ if and only if $2n \ge h$. For large enough n this means that the function ϕ is increasing with the complexity of the model. Hence the sum of the two terms: $\hat{R}(f)$ (precision) and $\phi(\frac{h}{n}, \frac{\log\eta}{n})$ (complexity) represents the compromise between precision in the risk estimation and the complexity of the classifier. Therefore minimising this sum is the sensible thing to do in order to perform "optimally".

The rest of the lecture focuses on ways to solve (or solve approximately) this minimisation problem for some classes $\mathcal{F}$.

For additional information, see Section 19.4 of

> Härdle, W. and Simar, L., *Applied Multivariate Statistical Analysis*, Third Edition, Springer, 2012.

A treatment along similar lines can also be found in the e-book (available from the library)

Hastie, T., Friedman, J. and Tibshirani, R. *The Elements of Statistical Learning: Data Mining, Inference and Prediction,* Second Edition, Springer 2009.

13.3 Basic idea of SVMs

A *linear classifier* is one that given feature vector $\boldsymbol{x}_{\text{new}}$ and weights $\boldsymbol{w}$, classifies y_{new} based on the value of $\boldsymbol{w}^\top \boldsymbol{x}_{\text{new}}$; for example,

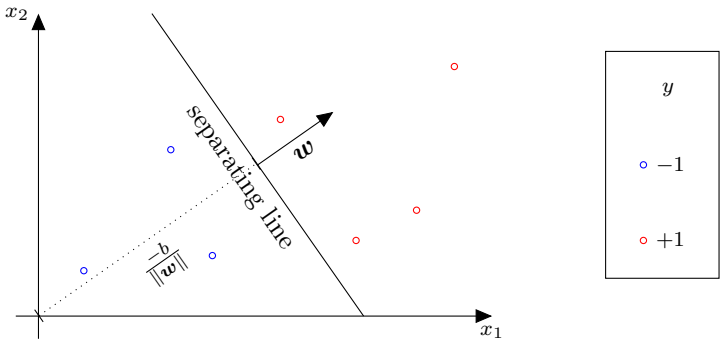
$$\hat{y}_{\text{new}} = \begin{cases} +1 & \text{if } \boldsymbol{w}^\top \boldsymbol{x}_{\text{new}} + b > 0 \\ -1 & \text{if } \boldsymbol{w}^\top \boldsymbol{x}_{\text{new}} + b < 0 \end{cases}$$

for a threshold $-b$. Here, we see that every element of $\boldsymbol{x}$, x_i, gets a weight w_i:

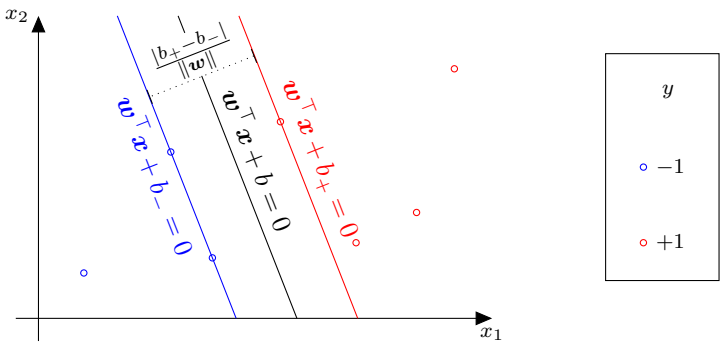
Sign of w_i determines whether increasing x_i pushes the prediction toward $y_i = -1$ or $y_i = +1$.

Magnitude of w_i determines how strongly.

The regions of $\boldsymbol{x}$ for which the model predicts $+1$ as opposed to -1 are defined by $\boldsymbol{w}^\top \boldsymbol{x} + b = 0$. Points $\boldsymbol{x}$ that satisfy that equation exactly form a line (if $d = 2$), a plane (if $d = 3$), or a hyperplane (if $d \geq 3$). We call the data *linearly separable* if a hyperplane that separates them exists. Let us focus on this linearly separable case (and consider the nonseparable case later.) The following diagram illustrates one such line:



Now, usually, there are infinitely many different hyperplanes which could be used to separate a linearly separable dataset. We therefore have to define the "best" one. The "best" choice can be regarded as the middle of the widest empty strip (or higher dimensional analogue) between the two classes, one that maximises the *margin* $\frac{|b_+ - b_-|}{\|\boldsymbol{w}\|}$ in the following illustration:



$\implies$ We want to make the *margin* $\frac{|b_+ - b_-|}{\|\boldsymbol{w}\|}$ as big as possible.

The scale of $\boldsymbol{w}$ and b is arbitrary: for arbitrary $\alpha \neq 0$, any $\boldsymbol{x}$ that satisfies $\boldsymbol{w}^\top \boldsymbol{x} + b = 0$ also satisfies $(\alpha\boldsymbol{w})^\top \boldsymbol{x} + (\alpha b) = \alpha(\boldsymbol{w}^\top \boldsymbol{x} + b) = 0$, so $(\boldsymbol{w}, b)$ and $(\alpha\boldsymbol{w}, \alpha b)$ define the same plane. We fix $|b_+ - b| = |b_- - b| = 1$, and only vary $\boldsymbol{w}$: our "outer" hyperplanes become

$$\boldsymbol{w}^\top \boldsymbol{x} + (b-1) = 0$$
$$\boldsymbol{w}^\top \boldsymbol{x} + (b+1) = 0.$$

Then, the margin of $\frac{|b_+ - b_-|}{\|\boldsymbol{w}\|} = \frac{2}{\|\boldsymbol{w}\|}$ is maximised by minimising $\|\boldsymbol{w}\|$. Therefore, a *Linear Support Vector Machine* minimises $\|\boldsymbol{w}\|^2$ subject to separating -1s and $+1$s.

13.4 Estimation

13.4.1 Linear SVM: Separable Case

We write the boundaries of the empty region as

$$\boldsymbol{w}^\top \boldsymbol{x} + (b-1) = 0 \implies \boldsymbol{w}^\top \boldsymbol{x} + b = +1$$
$$\boldsymbol{w}^\top \boldsymbol{x} + (b+1) = 0 \implies \boldsymbol{w}^\top \boldsymbol{x} + b = -1,$$

and observe that

$$\hat{y}_i = \begin{cases} +1 & \text{if } \boldsymbol{w}^\top \boldsymbol{x}_i + b > 0 \\ -1 & \text{if } \boldsymbol{w}^\top \boldsymbol{x}_i + b < 0 \end{cases} = \text{sign}(\boldsymbol{w}^\top \boldsymbol{x}_i + b).$$

This means that if $\boldsymbol{w}^\top \boldsymbol{x} + b = 0$ separates -1s and $+1$s (i.e., $y_i = \hat{y}_i$ for all $i = 1, \ldots, n.$),

$$y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1.$$

Therefore, a linear SVM learning task for can be expressed as a constrained optimisation problem:

$$\arg\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|^2 \text{ subject to } y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1, \; i = 1, \ldots, n.$$

(Here and elsewhere, $\arg\min_a h(a)$ is that a which minimises the value of $h(a)$.)

The objective is *quadratic* (convex) and the constraints are *linear*. This problem can be solved by Lagrange multipliers. The following outlines the steps and the key results.

1. Rewrite the objective function as the *Lagrangian*: (note the use of α_is instead of λ_is):

$$\text{Lag}(\boldsymbol{w}, b; \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{n} \alpha_i \left[y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - 1 \right].$$

2. As the constraints are inequalities rather than equalities, apply the so-called KKT (Karush–Kuhn–Tucker) conditions: the *saddle point* $(\boldsymbol{w}, b, \boldsymbol{\alpha})$: $\text{Lag}'(\boldsymbol{w}, b; \boldsymbol{\alpha}) = \mathbf{0}$ will be the constrained optimum if $\alpha_i \geq 0$, $i = 1, \ldots, n$. Thus, our goal becomes to solve for $\text{Lag}'(\boldsymbol{w}, b; \boldsymbol{\alpha}) = \mathbf{0}$ subject to $\alpha_i \geq 0$.

3. Set derivatives of Lag with respect to $\boldsymbol{w}$ and b equal to zero:

$$\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i = \mathbf{0} \quad \implies \quad \boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i,$$
$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0 \quad \implies \quad \sum_{i=1}^{n} \alpha_i y_i = 0.$$

4. Note, also, that

$$y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - 1 \geq 0,\ i = 1, \ldots, n,$$
$$\alpha_i \left(y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - 1\right) = 0,\ i = 1, \ldots, n.$$

for some $\alpha_i \geq 0,\ i = 1, \ldots, n$. Notice that the second equation implies that *either* $\alpha_i = 0$ *or* $y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) = 1$ (or both). But that means that if $\alpha_i \neq 0$, the observation lies on a corresponding hyperplane and is known as a *support vector*.

Dual Optimisation Problem

Substituting the expression of $\boldsymbol{w}$ in terms of $\boldsymbol{\alpha}$ and expanding $\|\boldsymbol{w}\|^2$, we get the *dual problem*:

$$\text{Lag}_D(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j,$$

to be maximised subject to

$$\alpha_i \geq 0,\ i = 1, \ldots, n$$
$$\sum_{i=1}^{n} \alpha_i y_i = 0.$$

This is a *quadratic programming* problem, for which many software tools are available.

13.4.2 Linear SVM: Nonseparable Case

Of course, in real-world problems, it is not possible to find hyperplanes which perfectly separate the target classes. The *soft margin* approach considers a trade-off between margin width and number of training misclassifications. *Slack* variables $\xi_i \geq 0$ are included in the constraints: we insist that

$$y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 - \xi_i. \tag{13.1}$$

The optimisation then becomes

$$\arg\min_{\boldsymbol{w}, \boldsymbol{\xi}} \left(\frac{1}{2}\|\boldsymbol{w}\|^2 + C \sum_{i=1}^{n} \xi_i \right) \text{ subject to } y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 - \xi_i,\ i = 1, \ldots, n,$$

or a tuning constant C. Small C means a lot of slack, whereas a large C means little slack. In particular, if we set $C = \infty$, we require separation to be perfect, a *hard margin*.

Now, taking (13.1) and solving for ξ_i gives $\xi_i \geq 1 - y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)$. We want to make ξ_i as small as possible, so we can set $\xi_i = 1 - y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)$.

Dual Optimisation Problem

The Lagrangian is now (with additional multipliers $\boldsymbol{\mu}$),

$$\text{Lag}(\boldsymbol{w}, b, \boldsymbol{\xi}; \boldsymbol{\alpha}, \boldsymbol{\mu}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i \left[y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - 1 + \xi_i \right] - \sum_{i=1}^{n} \mu_i \xi_i.$$

Now,

$$\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i = \boldsymbol{0} \quad &\Longrightarrow \quad \boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i \\
\frac{\partial L}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0 \quad &\Longrightarrow \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \\
\frac{\partial L}{\partial \boldsymbol{\xi}} = C\mathbf{1}_n - \boldsymbol{\alpha} - \boldsymbol{\mu} = \boldsymbol{0} \quad &\Longrightarrow \quad C - \alpha_i - \mu_i = 0,\ i = 1, \ldots, n.
\end{aligned}$$

with additional KKT conditions for $i = 1, \ldots, n$:

$$\begin{aligned}
\alpha_i &\geq 0 \\
\mu_i &\geq 0 \\
\alpha_i \left(y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - 1 + \xi_i \right) &= 0.
\end{aligned}$$

Substituting into the Lagrangian leads to

$$\text{Lag}_D(\boldsymbol{\alpha}, \boldsymbol{\mu}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{n} \alpha_j \alpha_k y_j y_k (\boldsymbol{x}_j^\top \boldsymbol{x}_k) + \sum_{i=1}^{n} \xi_i (C - \alpha_i - \mu_i).$$

But $C - \alpha_i - \mu_i = 0$, so as long as $\alpha_i \leq C$, $\mu_i \geq 0$ is completely determined by α_i, and we get a dual problem

$$\arg\max_{\boldsymbol{\alpha}} \left(\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{n} \alpha_j \alpha_k y_j y_k (\boldsymbol{x}_j^\top \boldsymbol{x}_k) \right)$$

$$\text{subject to } \sum_{i=1}^{n} \alpha_i y_i = 0 \text{ and } 0 \leq \alpha_i \leq C,\ i = 1, \ldots, n.$$

We can also express the prediction in two ways:

$$\text{Primal: } \hat{y}(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^\top \boldsymbol{x} + b), \tag{13.2}$$

$$\text{Dual: } \hat{y}(\boldsymbol{x}) = \text{sign}\{\sum_{j=1}^{n} \alpha_j y_j (\boldsymbol{x}_j^\top \boldsymbol{x}) + b\}. \tag{13.3}$$

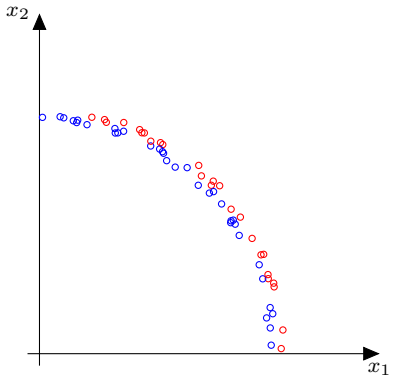
Primal ($\boldsymbol{w}$) form requires d parameters, while *dual* ($\boldsymbol{\alpha}$) form requires n parameters. This means that for high-dimensional problems—those with $d \gg n$, a huge number of predictors—the dual representation can be more efficient.

But it gets better! Notice that only the $\boldsymbol{x}_i$s closest to the separating hyperplane—those with $\alpha_j > 0$—matter in determining $\hat{y}(\boldsymbol{x})$, so most of them will have no effect. Thus, computationally, effective "n" will actually much smaller than the sample size, so the above condition can be met far more often than one might expect. Again, those $\boldsymbol{x}_i$s that "support" the hyperplane are called *support vectors*.

In addition, notice that the dual form only depends on $(\boldsymbol{x}_j^\top \boldsymbol{x}_k)$s. This opens the door to nonlinear SVMs.

13.5 Nonlinear SVMs

Consider:



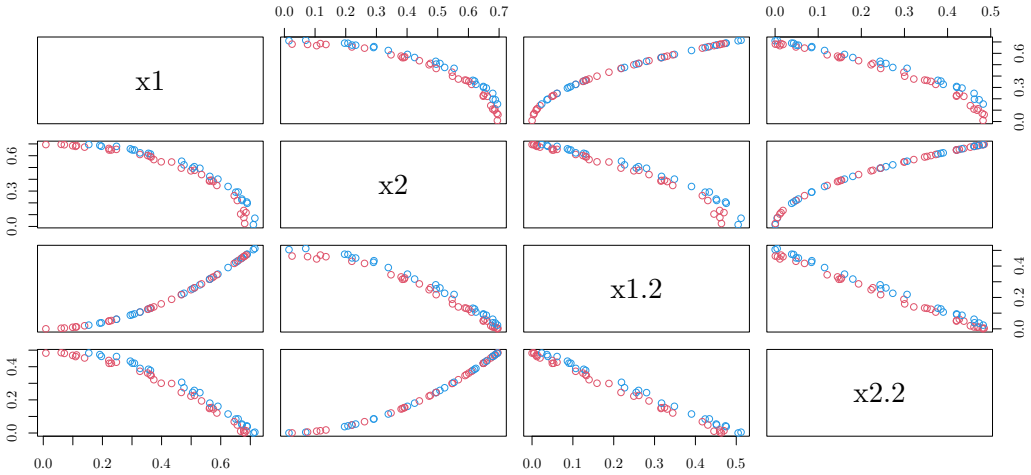
The true classification for these points is

$$y = \begin{cases} +1 & \text{if } {x_1}^2 + {x_2}^2 > 0.75^2 \\ -1 & \text{if } {x_1}^2 + {x_2}^2 < 0.75^2 \end{cases},$$

but one can hardly draw a line separating them.

What we can do is transform $\boldsymbol{x}$ so that a linear decision boundary can separate them. In this case, suppose we augmented our $\boldsymbol{x}$ with squared terms:

$$(x_1, x_2) \to (x_1, x_2, x_1^2, x_2^2) :$$



Now, a linear separator exists! Better yet, recall that the dual form (13.3) depends only on dot products $\boldsymbol{x}_i^\top \boldsymbol{x}_j$. However, we can specify other *kernels* $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. For example, a "kernel" function of the form $k(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u}^\top \boldsymbol{v} + 1)^2$ can be regarded as a dot product

$$u_1^2 v_1^2 + u_2 v_2^2 + 2u_1 v_1 + 2u_2 v_2 + 1$$

$$= (u_1^2, u_2^2, \sqrt{2}u_1, \sqrt{2}u_2, 1)^\top (v_1^2, v_2^2, \sqrt{2}v_1, \sqrt{2}v_2, 1).$$

which reconstructs the above augmentation. In general, kernel functions can be expressed in terms of high dimensional dot products. Computing dot products via kernel functions is computationally "cheaper" than using transformed attributes directly.

A common type of kernel is a *radial basis function*: a function of distance from the origin, or from another fixed point $\boldsymbol{v}$. Usually, the distance is *Euclidean*, *i.e.*

$$\|\boldsymbol{u} - \boldsymbol{v}\| = \sqrt{(u_1 - v_1)^2 + \cdots + (u_n - v_n)^2}.$$

A common radial basis function is *Gaussian*:

$$\phi(\boldsymbol{u}, \boldsymbol{v}) = \exp\left(-\gamma \|\boldsymbol{u} - \boldsymbol{v}\|^2\right).$$

We can use $\phi(\cdot, \cdot)$ as our SVM kernel.

13.6 Multiple classes

Finally, we briefly consider the problem when there are more than two classes. Suppose that there are $K > 2$ categories.

Recall that $\boldsymbol{w}^\top \boldsymbol{x}_i$ gives us a "score" that we normally compare to b. However, we do not have to do so. Instead, for each $k = 1, \ldots, K$, we can fit a separate SVM (i.e., $\boldsymbol{w}_k$ and b_k) for whether an observation is in k vs. not. We can then predict $\hat{y}_{\text{new}}$ by evaluating $\boldsymbol{w}_k^\top \boldsymbol{x}_{\text{new}} + b_k$ for each k and taking highest biggest one. This is called the *One-against-rest* approach.

A computationally more expensive approach that tends to perform better is the *One-against-one*: an SVM is fit for every distinct pair $k_1, k_2 = 1, \ldots, K$, fit an SVM for k_1 vs. k_2, and predict the "winner" of all the rounds (if any). This requires fitting $K(K-1)/2$ binary classifiers, but to smaller datasets.

13.7 SVM specification and tuning

Categorical data can be handled by introducing binary *dummy* variables to indicate each possible value.

When fitting an SVM, the user must specify some control parameters, these include cost constant C for slack variables, the type of kernel function, and its parameters. Unlike the more probabilistic forms of classification, it is difficult to predict the out-of-sample classification error for SVMs, so cross-validation is used.

The following kernel functions available via the R `e1071` package:

linear: $\boldsymbol{u}^\top \boldsymbol{v}$

polynomial: $(\gamma \boldsymbol{u}^\top \boldsymbol{v} + c_0)^p$

radial basis: $\exp(-\gamma \|\boldsymbol{u} - \boldsymbol{v}\|^2)$

sigmoid: $\tanh(-\gamma \boldsymbol{u}^\top \boldsymbol{v} + c_0)$

for constants γ, p, and c_0.

13.8 Examples

Example 13.1. SVM classification for the Edgar Anderson's Iris data, and using ROC curves.

13.9 Conclusion

We conclude with a brief discussion of the advantages and disadvantages of SVMs. SVM training can be formulated as a convex optimisation problem, with efficient algorithms for finding the global minimum, and the final result involves support vectors rather than the whole training set. This is both a computational benefit, but also one to robustness: outliers have less effect than for other methods.

On the other hand, they are much more difficult to interpret than model-based classification techniques like the linear discriminant analysis. Furthermore, SVMs do not actually provide class probability estimates. These can be estimated by cross-validation, however.

14 Cluster Analysis

The goal of cluster analysis is to identify groups in data. In contrast to SVMs and discriminant analysis, no preexisting group labels are provided. This makes it an example of *unsupervised learning*.

The input of cluster analysis is therefore an *unlabelled* sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$, and the output is a *grouping* of observations such that more similar (in some sense) observations are placed in the group. That is, cluster analysis assigns to each $\boldsymbol{x}_i$ a group index $G_i \in \{1, \ldots, K\}$ such that if $G_i = G_j$, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are "on average" more similar in some sense than if $G_i \neq G_j$.

Throughout this lecture, we will use the following additional notation.

$\boldsymbol{G} = (G_1, \ldots, G_n)^\top$**:** a column vector of cluster memberships.

$S_1, \ldots, S_K$**:** a *partitioning* of the observations $\{1, \ldots, n\}$ into K non-overlapping sets such that for every $i \in S_k$, $G_i = k$.

$\boldsymbol{S} = (S_1, \ldots, S_K)$**:** a shorthand for the clustering expressed in terms of sets.

We will consider a taxonomy of approaches to clustering. The "classical" approach is to specify an *algorithm* that assigns observations to clusters. (Often, but not always, an objective function may be defined that is optimised by the algorithm.) Classical approaches can be further subdivided into *hierarchical* clustering, which produces a hierarchy of nested clusterings in a tree which has observations as leaves; and *non-hierarchical*, which merely assigns a label to each point.

The *model-based* approach to clustering is to postulate a *mixture model*—a model consisting of a mixture of probability distributions with different location parameters. The parameters of this model embody information about the clusters (e.g., their means and frequencies), and estimating them enables probabilistic, or *soft* clusterings.

We discuss these approaches in turn.

14.1 "Classical"

14.1.1 Components

In order to cluster data—particularly multivariate data—we must first define a *proximity measure*: some function $d(\boldsymbol{x}_1, \boldsymbol{x}_2)$ that determines difference between two observations. (Equivalently

we can define a similarity score and negate or invert it.) Here are some common metrics measures:

Euclidean: $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| = \sqrt{\sum_{j=1}^{p}(x_{1j} - x_{2j})^2}$, the "ordinary" straight-line distance.

taxicab/Manhattan: $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_1 = \sum_{j=1}^{p}|x_{1j} - x_{2j}|$, distance if one is only allowed to travel parallel to the axes (like a taxicab on the Manhattan city grid).

Gower: $p^{-1}\sum_{j=1}^{p}\mathbb{I}(x_{1j} \neq x_{2j})$: for binary measures.

A metric should be substantively meaningful and appropriate for the data. It is also common to scale all of the dimensions (say, to have variance of 1 or to be between 0 and 1) before clustering.

Given these distances, we specify the algorithm that minimises within-cluster and maximises between-cluster distances in some sense—that sense often operationalised in an *objective function*.

14.1.2 Example: K-means

Perhaps the best known clustering algorithm is the K-means. It has the advantage of being simple and intuitive. The objective function that it ultimately minimises (over the partitioning $\boldsymbol{S} = (S_1, \ldots, S_K)$) is

$$\sum_{k=1}^{K}\frac{1}{2|S_k|}\sum_{i,j\in S_k}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2,$$

the sum of squared Euclidean distances between every distinct pair of observations within each cluster (appropriately scaled). It can be shown (using a decomposition similar to that of ANOVA) that this is equivalent to minimising

$$\sum_{k=1}^{K}\sum_{i\in S_k}\|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{S_k}\|^2,\ \bar{\boldsymbol{x}}_{S_k} = \frac{1}{|S_k|}\sum_{i\in S_k}\boldsymbol{x}_i,$$

which is simply the sum of the squared Euclidean distances between each data point and the mean of its cluster.

The following algorithm often does a good job finding such a clustering:

1. Randomly assign a cluster index to each element of $\boldsymbol{G}^{(0)}$.
2. Calculate cluster means (centroids):

$$\bar{\boldsymbol{x}}_{S_k^{(t-1)}} = \frac{1}{|S_k^{(t-1)}|}\sum_{i\in S_k^{(t-1)}}\boldsymbol{x}_i,\ k = 1, \ldots, K.$$

3. Calculate distances of each data point from each mean:

$$d_{ik} = \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{S_k^{(t-1)}}\|,\ i = 1, \ldots, n,\ k = 1, \ldots, K.$$

4. Reassign each point to its nearest mean:

$$G_i^{(t)} = \arg\min_k d_{ik}.$$

(Here and elsewhere, $\arg\min_a h(a)$ is that a which minimises the value of $h(a)$.)

5. Repeat from Step 2 until $\boldsymbol{G}^{(t)} = \boldsymbol{G}^{(t-1)}$.

14.1.3 Extension: K-medioids

A generalisation of K-means is the K-medioids technique. We define a *medioid* $\tilde{\boldsymbol{x}}_{S_k}$ *of cluster* k to be a *specific observation* that has the closest summed distance (however defined) to all other observations in S_k:

$$\tilde{\boldsymbol{x}}_{S_k} = \arg\min_{\boldsymbol{x}_j} \sum_{i \in S_k} d(\boldsymbol{x}_j, \boldsymbol{x}_i).$$

The *Method of K-medioids* or *partitioning around medioids (PAM)* minimises the sum of these distances:

$$\arg\min_{\boldsymbol{S}} \sum_{k=1}^{K} \sum_{i \in S_k} d(\boldsymbol{x}_i, \tilde{\boldsymbol{x}}_{S_k}).$$

This method is much more expensive computationally than K-means, but it is also more robust to outliers.

It is typically fit as follows:

1. Randomly assign a cluster index to each element of $\boldsymbol{G}^{(0)}$.
2. Calculate cluster medioids:

$$\tilde{\boldsymbol{x}}_{S_k^{(t-1)}} = \arg\min_{\boldsymbol{x}_j} \sum_{i \in S_k^{(t-1)}} d(\boldsymbol{x}_j, \boldsymbol{x}_i),\ k = 1, \ldots, K.$$

3. Calculate distances of each data point from each medioid:

$$d_{ik} = d(\boldsymbol{x}_i, \tilde{\boldsymbol{x}}_{S_k^{(t-1)}}),\ i = 1, \ldots, n,\ k = 1, \ldots, K.$$

4. Reassign each point to its nearest medioid:

$$G_i^{(t)} = \arg\min_k d_{ik}.$$

5. Repeat from Step 2 until $\boldsymbol{G}^{(t)} = \boldsymbol{G}^{(t-1)}$.

14.1.4 Hierarchical clustering

Hierarchical clustering, instead of partitioning the data into K groups, produces a hierarchy of clusterings whose sizes range from 1 (no splits) to as high as n (every observation its own cluster). This clustering is typically visualised in a *dendrogram*, a tree diagram whose branching represents subdivisions of the data into clusters and whose height represents the distances between points or clusters.

The algorithms for producing these clusterings are either *agglomerative*, in that they start with each observation in its own cluster, then combine nearest observations into clusters, nearest clusters into bigger clusters, etc.; or *divisive*, starting with the whole dataset, then splitting it into a small number of clusters, those clusters into smaller clusters, etc..

The former require defining a notion of a *distance between clusters*. The latter require to defining a criterion based on which a cluster is split.

Some common examples of distances are provided in the following table:

Single linkage	$d(S_1, S_2) = \min\{d(\boldsymbol{x}_i, \boldsymbol{x}_j) : i \in S_1, j \in S_2\}$
Complete linkage	$d(S_1, S_2) = \max\{d(\boldsymbol{x}_i, \boldsymbol{x}_j) : i \in S_1, j \in S_2\}$
Average linkage (unweighted)	$d(S_1, S_2) = \frac{1}{\lvert S_1\rvert\lvert S_2\rvert} \sum_{i \in S_1} \sum_{j \in S_2} d(\boldsymbol{x}_i, \boldsymbol{x}_j)$
Average linkage (weighted)	$d(S_1 \cup S_2, S_3) = \frac{d(S_1,S_3)+d(S_2,S_3)}{2}$
Centroid	$d(S_1, S_2) = \lVert\bar{\boldsymbol{x}}_{S_1} - \bar{\boldsymbol{x}}_{S_2}\rVert$
Ward	$d(S_1, S_2) = \sum_{i \in S_1 \cup S_2} \lVert\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{S_1 \cup S_2}\rVert^2 - \sum_{i \in S_1} \lVert\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{S_1}\rVert^2 - \sum_{i \in S_2} \lVert\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{S_2}\rVert^2 = \frac{\lvert S_1\rvert\lvert S_2\rvert}{\lvert S_1\rvert+\lvert S_2\rvert} \lVert\bar{\boldsymbol{x}}_{S_1} - \bar{\boldsymbol{x}}_{S_2}\rVert^2$

A framework that is useful for expressing different between-cluster distances is the *Lance–Williams* framework. Given three clusters, S_1, S_2, and S_3, and suppose that we have some metric for evaluating pairwise distances between them, i.e., $d(S_1, S_2)$, $d(S_1, S_3)$, and $d(S_2, S_3)$. Then, we define the distance resulting from combining S_1 and S_2 in terms of these pairwise distances and coefficients α_1, α_2, β, and γ:

$$d(S_1 \cup S_2, S_3) = \alpha_1 d(S_1, S_3) + \alpha_2 d(S_2, S_3) + \beta d(S_1, S_2) + \gamma |d(S_1, S_3) - d(S_2, S_3)|.$$

This, plus the distance metric between individual points (which applies when the clusters have only one observation in them), allows us to define and efficiently calculate distances between clusters.

For example, the unweighted average linkage can be expressed in this framework as follows:

$$\begin{aligned} d(S_1 \cup S_2, S_3) &= \frac{1}{|S_1 \cup S_2||S_3|} \sum_{i \in S_1 \cup S_2} \sum_{j \in S_3} d(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ &= \frac{1}{(|S_1| + |S_2|)|S_3|} \left(\sum_{i \in S_1} \sum_{j \in S_3} d(\boldsymbol{x}_i, \boldsymbol{x}_j) + \sum_{i \in S_2} \sum_{j \in S_3} d(\boldsymbol{x}_i, \boldsymbol{x}_j) \right) \\ &= \frac{|S_1||S_3|d(S_1, S_3) + |S_2||S_3|d(S_2, S_3)}{(|S_1| + |S_2|)|S_3|} \\ &\implies \alpha_1 = \frac{|S_1|}{|S_1| + |S_2|},\ \alpha_2 = \frac{|S_2|}{|S_1| + |S_2|},\ \beta = \gamma = 0. \end{aligned}$$

Ward's method—the most popular hierarchical clustering criterion—similarly, uses the squared Euclidean distances $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2$ between points and then

$$\begin{aligned} \alpha_1 &= \frac{|S_1| + |S_3|}{|S_1| + |S_2| + |S_3|},\ \alpha_2 = \frac{|S_2| + |S_3|}{|S_1| + |S_2| + |S_3|}, \\ \beta &= \frac{-|S_3|}{|S_1| + |S_2| + |S_3|},\ \gamma = 0. \end{aligned}$$

Ward's method joins the groups that will increase the within-group variance least.

14.1.5 Software

SAS:

Hierarchical: `PROC CLUSTER` (`PROC TREE` to visualise, `PROC DISTANCE` to preprocess), `PROC VARCLUS`

Non-hierarchical: `PROC FASTCLUS`, `PROC MODECLUS`, `PROC FASTKNN`

R:

Hierarchical: `stats::hclust`, `cluster::agnes`

Non-hierarchical: `stats::kmeans`, `cluster::pam`

- Many others

14.1.6 Assessing

Lastly, we briefly discuss how a clustering $\boldsymbol{G}$ may be assessed. Ideally, this measurement should be "fair" to the number of clusters K. For example, in K-means clustering, splitting a cluster will *always* reduce the within-cluster variances, and so those cannot be used as a criterion.

- Given a clustering $\boldsymbol{G}$, how good is it?
- Ideally, measurement should be invariant to K.
  - I.e., not within-cluster variances.

A popular method, inspired by K-medioid clustering, is the *silhouettes*. For each $i = 1, \dots, n$, let

$$a(i) = \frac{1}{|S_{G_i}| - 1} \sum_{j \in S_{G_i}} d(\boldsymbol{x}_i, \boldsymbol{x}_j)$$
$$b(i) = \min_{k \neq G_i} \frac{1}{|S_k|} \sum_{j \in S_k} d(\boldsymbol{x}_i, \boldsymbol{x}_j).$$

Observe that $a(i)$ is the distance between i and other observations its own cluster and $b(i)$ is the distance between i and observations in the cluster nearest to i to which i does not belong. In a good clustering each observation will be much closer to its own cluster than to its neighbouring cluster, so $b(i) \gg a(i)$.

Then, *silhouette of* i is a value between -1 and $+1$ calculated as follows:

$$s(i) = \begin{cases} \frac{b(i)-a(i)}{\max(a(i),b(i))} & \text{if } |S_{G_i}| > 1 \\ 0 & \text{otherwise} \end{cases}.$$

That is $s(i)$ evaluates how much closer is i to the rest of its cluster than it is to its nearest cluster, and a higher silhouette indicates a better clustering for point i. Mean silhouette $n^{-1} \sum_{i=1}^{n} s(i)$ then measures the overall quality of clustering.

14.1.7 Examples

Example 14.1. Hierarchical, non-hierarchical clustering and assessment illustrated on the Edgar Anderson's Iris data.

14.2 Model-based clustering

14.2.1 Mixture Models

Lastly, we turn to model-based clustering. We will discuss the theoretical underpinnings of this approach—mixture models—and an important special case of Gaussian clustering and its parametrisation. The Expectation–Maximisation algorithm, often used to estimate these models will also be described, as it is useful in a wide variety of circumstances, but it is not examinable.

A *finite mixture model* is a probability model under which each observation comes from one of several distributions, but we do not observe from which one. (Infinite mixture models exist as well, but they are outside of the scope of this class.)

A mixture model is specified as follows. We set K to be the number of distributions (clusters), and a collection of K density functions on the support of $\boldsymbol{x}_i$, $f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k)$ (for $k = 1, \ldots, K$) each having a parameter vectors $\boldsymbol{\theta}_k$ (e.g., its expectation), which we do not know and must estimate. We also postulate K (unknown) probabilities π_k that an observation (any observation) comes from cluster k. (Standard restrictions apply: $0 \leq \pi_k \leq 1$, $\sum_{k=1}^{K} \pi_k = 1$.)

For brevity, we define $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^\top$, a vector of these probabilities; and $\Psi = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, \boldsymbol{\pi}\}$, the collection of all model parameters. Then, we assume the following data-generating process: for each $i = 1, \ldots, n$,

1. Sample $G_i \in \{1, \ldots, K\}$ with $\Pr(G_i = k; \boldsymbol{\pi}) = \pi_k$.
2. Sample $\boldsymbol{X}_i | G_i \sim f_{G_i}(\cdot; \boldsymbol{\theta}_{G_i})$.
3. Observe $\boldsymbol{X}_i$, and "forget" G_i.

The pdf of this *mixture density* is

$$f_{\boldsymbol{X}_i}(\boldsymbol{x}_i; \Psi) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k). \tag{14.1}$$

We wish to estimate the parameters Ψ from the sample of $\boldsymbol{x} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$. This leads to the likelihood

$$L_{\boldsymbol{x}}(\Psi) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k). \tag{14.2}$$

This formulation is convenient for a number of reasons. It is a probability model for the $\boldsymbol{X}_i$s, and therefore we can use it to obtain a *soft clustering*: rather than a *hard clustering* that assigns a point to a single cluster, we can apportion an observation's membership by how likely it to have come from each cluster. An application of Bayes's rule and (14.1) gives

$$\Pr(G_i = k | \boldsymbol{x}_i; \Psi) = \frac{\pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k)}{\sum_{k'=1}^{K} \pi_{k'} f_{k'}(\boldsymbol{x}_i; \boldsymbol{\theta}_{k'})}.$$

We can also embed it into a *hierarchical model* (a meaning distinct from the hierarchical clustering above), in which either $\boldsymbol{x}_i$s are parameters for some model for the data or for the observation process or $\boldsymbol{\theta}$s are functions of some *hyper-parameters*. Lastly, the fact that we have a well-defined likelihood facilitates model selection.

14.2.2 Multivariate normal clusters

As with other analysis scenarios discussed in this course, the multivariate normal distribution provides a useful formulation for the clusters. Consider the following parametrisation:

$$f_k(\boldsymbol{x}_i;\boldsymbol{\theta}_k) = \frac{1}{(2\pi)^{p/2}|\Sigma(\boldsymbol{\theta}_k)|^{1/2}}\,\mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}(\boldsymbol{\theta}_k))^\top\Sigma(\boldsymbol{\theta}_k)^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}(\boldsymbol{\theta}_k))}.$$

Here, $\boldsymbol{\mu}(\boldsymbol{\theta}_k)$ is the mean vector of cluster k (e.g., first p elements of $\boldsymbol{\theta}_k$), and $\Sigma(\boldsymbol{\theta}_k)$ is the model for the variances. We may also have different clusters "share" elements of $\boldsymbol{\theta}$, and a more general case is

$$f_k(\boldsymbol{x}_i;\boldsymbol{\theta}) = \frac{1}{(2\pi)^{p/2}|\Sigma_k(\boldsymbol{\theta})|^{1/2}}\,\mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k(\boldsymbol{\theta}))^\top\Sigma_k(\boldsymbol{\theta})^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k(\boldsymbol{\theta}))}, \tag{14.3}$$

where $\boldsymbol{\mu}_k(\boldsymbol{\theta})$ and $\Sigma_k(\boldsymbol{\theta})$ "extract" the appropriate elements from $\boldsymbol{\theta}$.

One advantage of multivariate normal clusters is in its flexibility in specifying cluster size and shape. (Recall your exercises from Week 1.) Recall the eigendecomposition of the covariance matrix $\Sigma = P\Lambda P^\top$, with P orthogonal and Λ diagonal and nonnegative. Let us further parametrise it as

$$\Sigma = \lambda PAP^\top,$$

with $P \in \mathcal{M}_{p,p}$ orthogonal, $A \in \mathcal{M}_{p,p}$ diagonal and nonnegative *with* $|A| = 1$ *(unimodular)*, and scalar $\lambda > 0$. This allows us to interpret the structure of the matrix in simple, substantive terms.

Starting with λ, recall recalling that the determinant of a matrix can be viewed as its volume. Then,

$$|\Sigma| = \lambda^p|P||A||P^\top| = \lambda^p,$$

which makes λ is the "spread", "size", or "volume" of the cluster.

To interpret the diagonal, unimodular matrix A, observe that if $A = I_p$, then

$$\Sigma = \lambda PAP^\top = \lambda PP^\top = \lambda I_p,$$

making the cluster spherical—equal variances on all dimensions. Similarly, if some diagonal elements of A are much larger than others, then the cluster will be an ellipsoid more stretched in one direction than in others.

Lastly, observe that if $P = I_p$, then

$$\Sigma = \lambda PAP^\top = \lambda A,$$

an ellipsoid whose axes are parallel to coordinate axes, implying the elements of $\boldsymbol{X}_i$ within each cluster are uncorrelated with unequal variances. More generally, P controls the rotation of ellipsoid—the correlation between the dimensions and the orientation of the cluster.

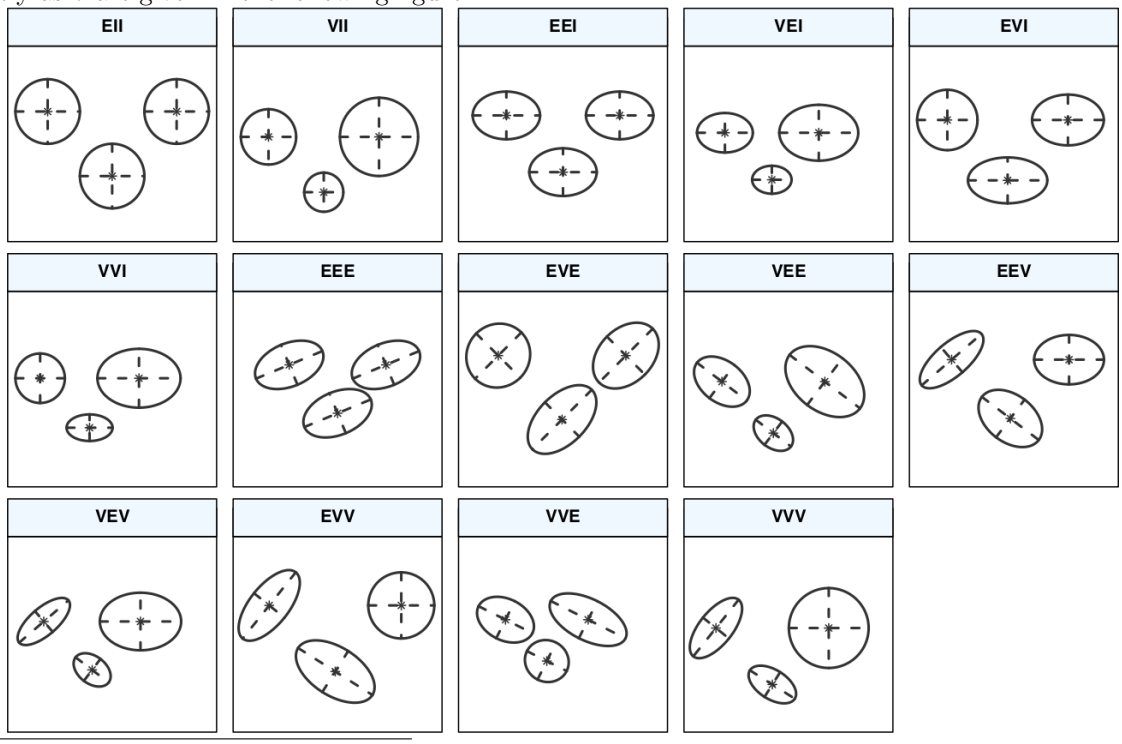
When it comes to estimating K clusters, we can permit the λs, the As, and the Ps to vary between the clusters, be constant between the clusters, or, for A and P, be fixed at the identity. Each combination embodies different assumption about the shape and the relationship between clusters; and, in general, the more we permit to vary, the more parameters we must estimate and the more data we therefore require. Generally,

1. For a mixture of K clusters, we must, invariably, estimate the cluster membership probabilities $\pi_1, \ldots, \pi_K$ ($K-1$ parameters) and cluster means $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$ (Kp parameters).
2. Then, λ can be constrained $\lambda_1 = \lambda_2 = \cdots = \lambda_K$ (1 parameter) or allowed to vary (K parameters).

3. Then, A can be fixed $A_1 = A_2 = \cdots = A_K = I_d$ (0 parameters), constrained $A_1 = A_2 = \cdots = A_K$ ($p-1$ parameters), or allowed to ($K(p-1)$ parameters).

4. Lastly, if A is not fixed at the identity matrix, P can either be fixed $P_1 = P_2 = \cdots = P_K = I_d$ (0 parameters), constrained $P_1 = P_2 = \cdots = P_K$ ($\binom{p}{2}$ parameters), or allowed to vary ($K\binom{p}{2}$ parameters).

The different cluster shapes identified by their constraint triple (λ, A, P) encoding being fixed at identity as `I`, being constrained to equality between clusters as `E`, and being allowed to vary freely as `V` are given in the following figure:



14.2.3 Model selection

As mentioned before, model-based clustering requires one to specify both the number of clusters K and the within-cluster models $f_k(\boldsymbol{x}_i; \Psi)$. In the case of multivariate normal clustering, we have a large number of possible specifications for the Σ_ks, and the number of parameters can grow quickly for "`XXV`" models in particular.

At the same time, because it is likelihood-based, a variety of standard model-selection techniques can be used. For example, BIC is recommended:

$$\mathrm{BIC}_\nu = -2 \log L_{\boldsymbol{x}}(\hat{\Psi}) + \nu \log n,$$

where ν the number of parameters estimated. (Here, lower BIC is better, but some authors and software packages use $2 \log L_{\boldsymbol{x}}(\hat{\Psi}) - \nu \log n$, with higher BIC being better.)

Substantive considerations also matter. For example, how many clusters does our research hypothesis predict? Do we expect correlations between dimensions to vary between clusters?

14.2.4 Software

SAS: `PROC MBC`

R: package `mclust` and others

14.2.5 Examples

Example 14.2. Model-based clustering and model selection illustrated on the Edgar Anderson's Iris data.

14.2.6 Expectation–Maximisation Algorithm

Lastly, we discuss the typical computational approach for estimating these mixture models. The $\log L(\Psi)$ in (14.2) is computationally tractable, but it does not simplify or decompose much, because while the logarithm of a product is a sum of the logarithms, the logarithm of a sum does not, in general, simplify further. Thus, we introduce the *Expectation–Maximisation (EM)* algorithm:

1. Introduce an unobserved (latent) variable G_i, $i = 1, \ldots, n$ giving the cluster membership of i.
2. Suppose that $G_1, \ldots, G_n$ *are* observed; then, this *complete-data likelihood*,
$$L_{\boldsymbol{x},G_1,\ldots,G_n}(\Psi) = \prod_{i=1}^{n} \pi_{G_i} f_{G_i}(\boldsymbol{x}_i; \boldsymbol{\theta}_{G_i}) :$$
we "know" the exact cluster from which each observation came, so we no longer have to sum over the possible clusters. Then, the log-likelihood decomposes into two summations:
$$\log L_{\boldsymbol{x},G_1,\ldots,G_n}(\Psi) = \sum_{i=1}^{n} \log \pi_{G_i} + \sum_{i=1}^{n} \log f_{G_i}(\boldsymbol{x}_i; \boldsymbol{\theta}_{G_i}), \tag{14.4}$$
one that depends only on the π_ks and the other only on the $\boldsymbol{\theta}_k$s.
3. Start with an initial guess $\Psi^{(0)}$.
4. Iterate **E-step** and **M-step** described below to convergence.

E-step

The *Expectation step* consists of starting with a parameter guess $\Psi^{(t-1)}$ and evaluating
$$Q(\Psi|\Psi^{(t-1)}) = \mathrm{E}_{G_1,\ldots,G_n|\boldsymbol{x};\Psi^{(t-1)}}(\log L_{\boldsymbol{x},G_1,\ldots,G_n}(\Psi)) :$$
the expected value of the complete-data log-likelihood. We can evaluate it by calculating (using the Bayes's rule)
$$q_{ik}^{(t-1)} = \Pr(G_i = k|\boldsymbol{x}; \Psi^{(t-1)}) = \frac{\pi_k^{(t-1)} f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k^{(t-1)})}{\sum_{k'=1}^{K} \pi_{k'}^{(t-1)} f_{k'}(\boldsymbol{x}_i; \boldsymbol{\theta}_{k'}^{(t-1)})}, \; i = 1, \ldots, n, \; k = 1, \ldots, K,$$
then substituting them in as
$$Q(\Psi|\Psi^{(t-1)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} q_{ik}^{(t-1)} \log \pi_k + \sum_{i=1}^{n} \sum_{k=1}^{K} q_{ik}^{(t-1)} \log f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k). \tag{14.5}$$
Observe that, like (14.4), (14.5) decomposes into a summation that depends only on the π_ks and a summation that depends only on the $\boldsymbol{\theta}_k$s.

M-step

The *Maximisation step* then consists of maximising the $Q(\Psi|\Psi^{(t-1)})$ with respect to Ψ to obtain the next parameter guess:

$$\Psi^{(t)} = \arg\max_{\Psi} Q(\Psi|\Psi^{(t-1)}), \text{ s.t. } \sum_{k=1}^{K} \pi_k = 1.$$

Conveniently, the form (14.5) separates the π_ks from the $\boldsymbol{\theta}_k$s, and so we can maximise them separately (i.e., if we differentiate with respect to one, the summation involving the other will vanish).

Maximising (14.5) with respect to $\boldsymbol{\theta}_k$s, we take the derivative

$$\frac{\partial Q(\Psi|\Psi^{(t-1)})}{\partial \boldsymbol{\theta}_k} = \sum_{i=1}^{n} q_{ik}^{(t-1)} \frac{\partial \log f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}_k},$$

and set to 0. This is a *weighted* maximum likelihood estimator.

Maximising (14.5) with respect to to π_ks is also straightforward. We will use Lagrange Multipliers to do so:

$$\text{Lag}(\boldsymbol{\pi}) = \sum_{i=1}^{n} \sum_{k=1}^{K} q_{ik}^{(t-1)} \log \pi_k - \alpha(\sum_{k=1}^{K} \pi_k - 1).$$

Differentiating,

$$\text{Lag}'_k(\boldsymbol{\pi}) = \sum_{i=1}^{n} q_{ik}^{(t-1)} \pi_k^{-1} - \alpha.$$

Setting to 0,

$$\pi_k = \sum_{i=1}^{n} q_{ik}^{(t-1)} / \alpha.$$

Summing and solving for α,

$$\sum_{k=1}^{K} \pi_k = \frac{1}{\alpha} \sum_{k=1}^{K} \sum_{i=1}^{n} q_{ik}^{(t-1)} = 1,$$

$$\alpha = \sum_{k=1}^{K} \sum_{i=1}^{n} q_{ik}^{(t-1)}.$$

Therefore,

$$\pi_k^{(t)} = \frac{\sum_{i=1}^{n} q_{ik}^{(t-1)}}{\sum_{k=1}^{K} \sum_{i=1}^{n} q_{ik}^{(t-1)}}.$$

"Sharing" $\boldsymbol{\theta}$s

Lastly, recall that when we select one of the "E" models and (14.3) in Section 14.2.2, we no longer have a separate $\boldsymbol{\theta}_k$ for every f_k. We may then need to redefine $\boldsymbol{\theta} \in \mathbb{R}^{Kp+1 \text{ or more}}$ to contain parameters for all groups (separate means, distinct variance parameters, etc.), and $f_k(\boldsymbol{x}_i; \boldsymbol{\theta})$ to "extract" those elements of $\boldsymbol{\theta}$ that it needs, with $\Psi = (\boldsymbol{\theta}, \boldsymbol{\pi})$.

Inferentially, $\boldsymbol{\theta}$ replaces $\boldsymbol{\theta}_k$ in all derivations above. In particular,

$$Q(\Psi|\Psi^{(t-1)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} q_{ik}^{(t-1)} \log \pi_k + \sum_{i=1}^{n} \sum_{k=1}^{K} q_{ik}^{(t-1)} \log f_k(\boldsymbol{x}_i; \boldsymbol{\theta}),$$

so

$$\frac{\partial Q(\Psi|\Psi^{(t-1)})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n}\sum_{k=1}^{K} q_{ik}^{(t-1)} \frac{\partial \log f_k(\boldsymbol{x}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}},$$

which is still a weighted MLE, but now it is joint for all groups, and without simplification.

14.3 Additional resources

An alternative presentation of these concepts can be found in JW Sec. 12.1–12.5. Additional software demonstration of model-based clustering can be found in

> Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1), 289.

15 Copulae

15.1 Formulation

For the multivariate normal, independence is equivalent to absence of correlation between any two components. In this case the joint cdf is a product of the marginals. When the independence is violated, the relation between the joint multivariate distribution and the marginals is more involved. An interesting concept that can be used to describe this more involved relation is the concept of *copula*. We focus on the two-dimensional case for simplicity. Then the copula is a function $C : [0,1]^2 \to [0,1]$ with the properties:

i) $C(0,u) = C(u,0) = 0$ for all $u \in [0,1]$.

ii) $C(u,1) = C(1,u) = u$ for all $u \in [0,1]$.

iii) For all pairs $(u_1,u_2),(v_1,v_2) \in [0,1]\times[0,1]$ with $u_1 \le v_1, u_2 \le v_2$:

$$C(v_1,v_2) - C(v_1,u_2) - C(u_1,v_2) + C(u_1,u_2) \ge 0.$$

The name is due to the implication that the copula links the multivariate distribution to its marginals. This is explicated in the following theorem:

Theorem 15.1 (Sklar's Theorem)**.** *Let $F(\cdot,\cdot)$ be a joint cdf with marginal cdf's $F_{X_1}(.)$ and $F_{X_2}(.)$. Then there exists a copula $C(\cdot,\cdot)$ with the property*

$$F(x_1,x_2) = C(F_{X_1}(x_1), F_{X_2}(x_2))$$

for every pair $(x_1,x_2) \in \mathbb{R}^2$. When $F_{X_1}(\cdot)$ and $F_{X_2}(\cdot)$ are continuous the above copula is unique. Vice versa, if $C(\cdot,\cdot)$ is a copula and $F_{X_1}(\cdot), F_{X_2}(\cdot)$ are cdf then the function $F(x_1,x_2) = C(F_{X_1}(x_1), F_{X_2}(x_2))$ is a joint cdf with marginals $F_{X_1}(\cdot)$ and $F_{X_2}(\cdot)$.

Taking derivatives we also get:

$$f(x_1,x_2) = c(F_{X_1}(x_1), F_{X_2}(x_2)) f_{X_1}(x_1) f_{X_2}(x_2) \tag{15.1}$$

where

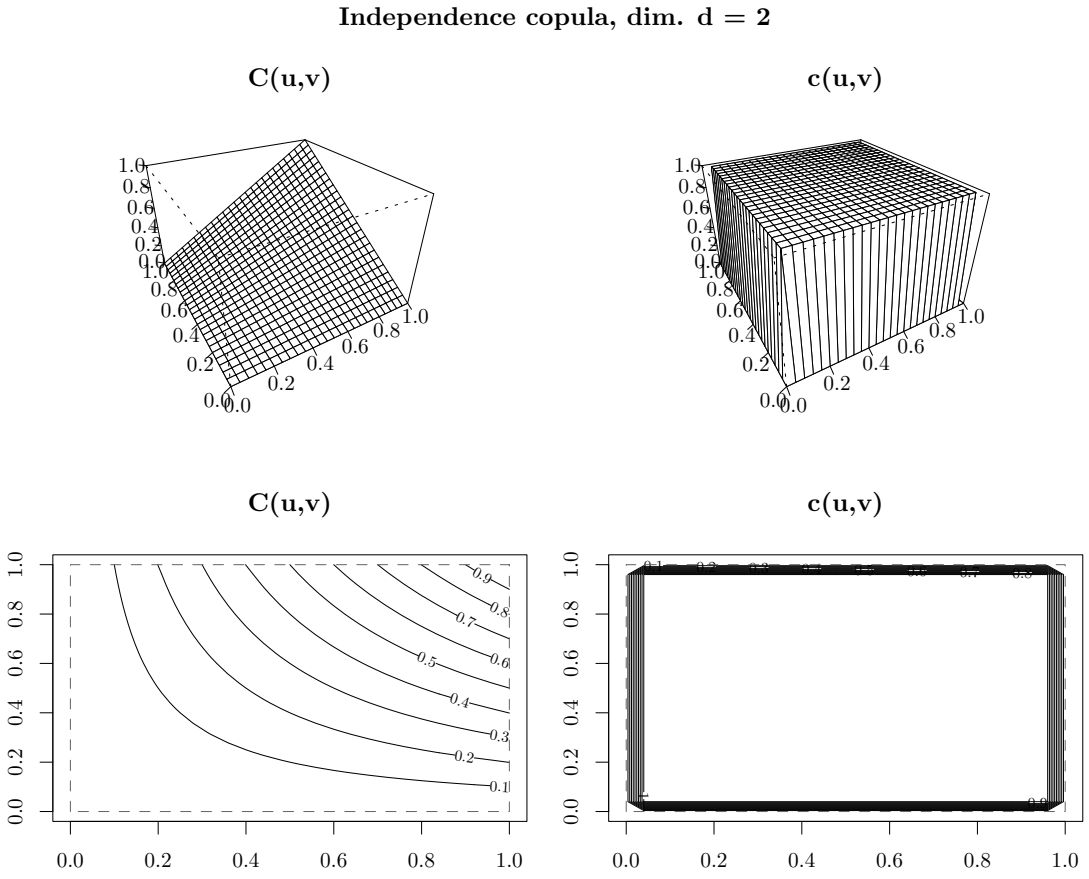
$$c(u,v) = \frac{\partial^2}{\partial u \partial v} C(u,v)$$

is the *density* of the copula. This relation clearly shows that the contribution to the joint density of X_1, X_2 comes from two parts: one that comes from the copula and is "responsible" for the dependence ($c(u,v) = \frac{\partial^2}{\partial u \partial v} C(u,v)$) and another one which takes into account marginal information only ($f_{X_1}(x_1) f_{X_2}(x_2)$).

It is also clear that the independence implies that the corresponding copula is $\Pi(u, v) = uv$ (this is called the independence copula).

These concepts are generalised also to p dimensions with $p > 2$.

The following figure illustrates an independence copula:



15.2 Common copula types

15.2.1 Elliptical copulae

An interesting example is the *Gaussian copula*. For $p = 2$ it is equal to:

$$\begin{aligned} C_\rho(u, v) &= \Phi_\rho(\Phi^{-1}(u), \Phi^{-1}(v)) \\ &= \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} f_\rho(x_1, x_2) dx_2 dx_1. \end{aligned}$$

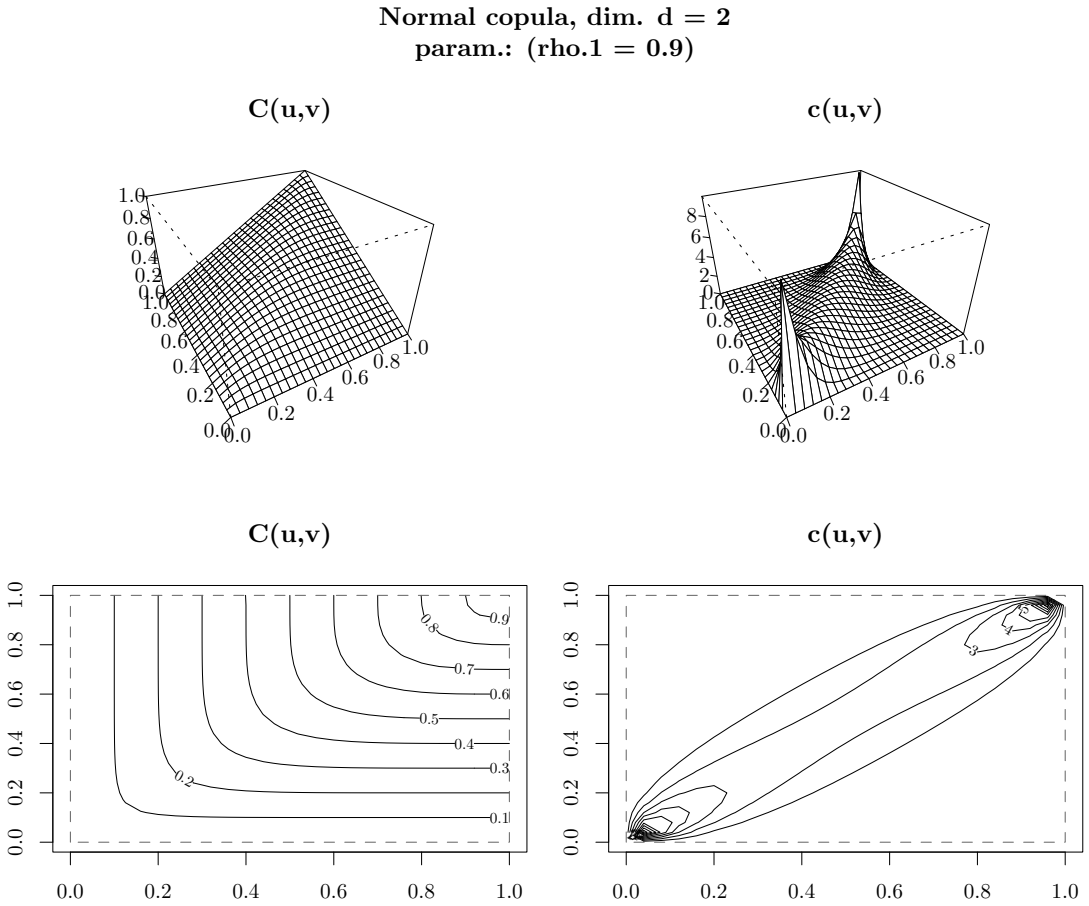
Here $f_\rho(\cdot, \cdot)$ is the joint bivariate normal density with zero mean, unit variances and a correlation ρ, $\Phi_\rho(\cdot, \cdot)$ is its cdf, and $\Phi^{-1}(\cdot)$ is the inverse of the cdf of the standard normal. (This is "The formula that killed Wall street".) When $\rho = 0$ we see that we get $C_0(u, v) = uv$ (as is to be expected).

Non-Gaussian copulae are much more important in practice and inference methods about copulae are a hot topic in Statistics. The reason for importance of non-Gaussian copulae is that

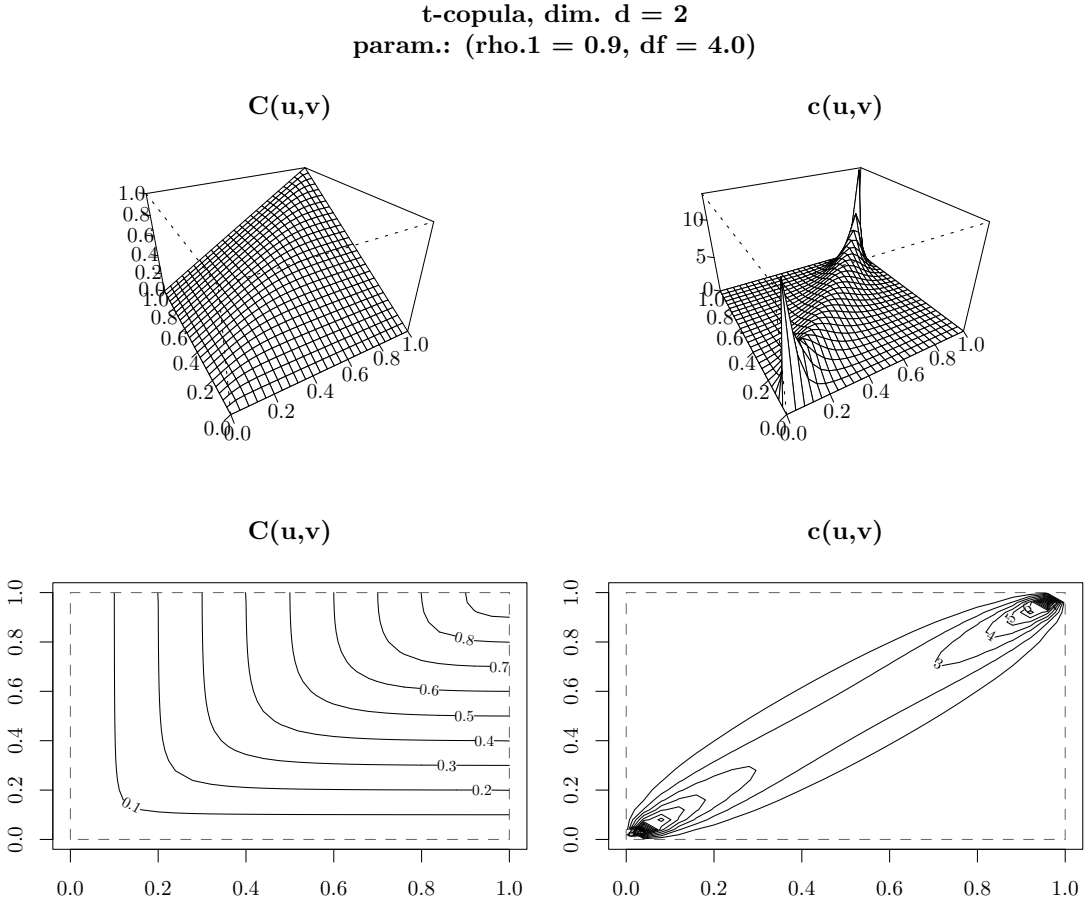
Gaussian copulae do not allow us to model reasonably well the tail dependence, that is, joint *extreme* events have virtually a zero probability. Especially in financial applications, it is very important to be able to model dependence in the tails.

The t-copula, based on the multivariate t-distribution does a slightly better job in tail behaviour. The multivariate t-distribution with variance parameter Σ and ν degrees of freedom is defined as $\boldsymbol{T} = \boldsymbol{Z}/X$, where $\boldsymbol{Z} \sim N(\boldsymbol{0}, \Sigma)$ and, independently, $X \sim \chi^2_\nu$. Note that $\mathrm{Var}(\boldsymbol{T}) \neq \Sigma$.

The following figure illustrates an Gaussian copula:



The following figure illustrates a multivariate t-copula copula:



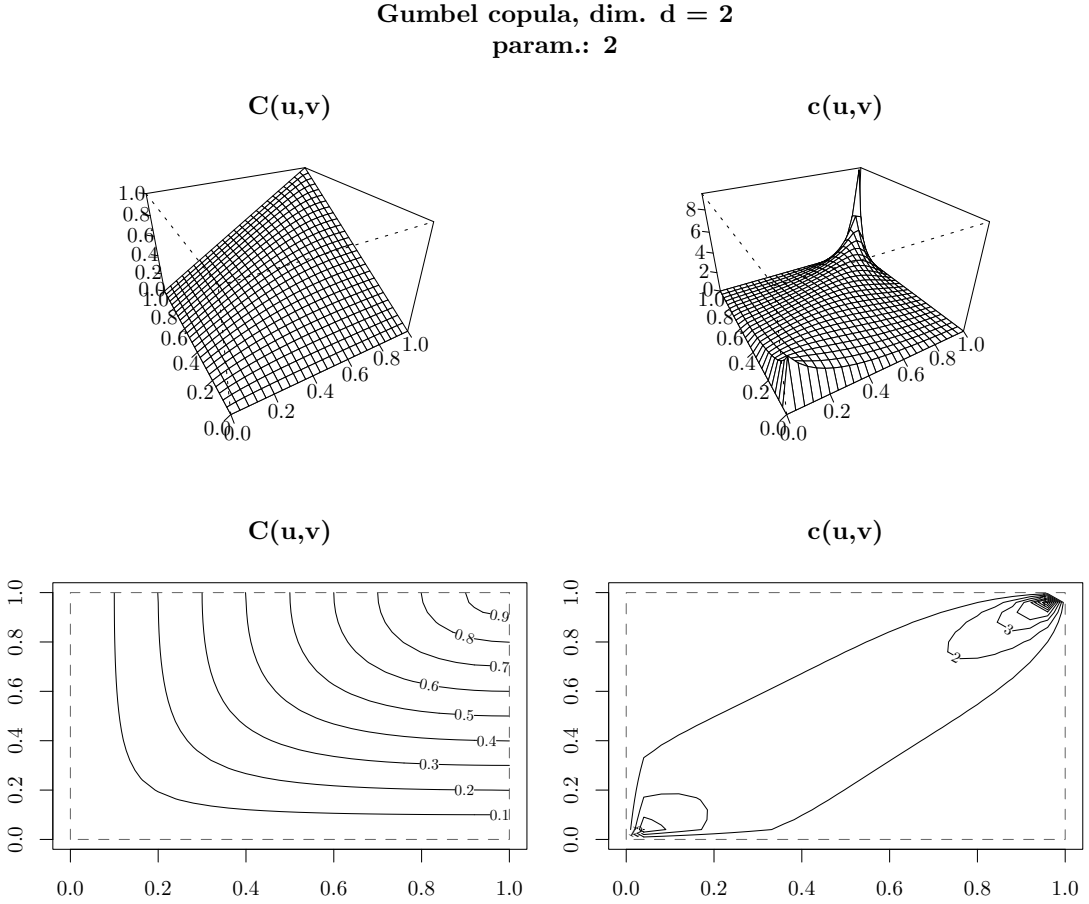
15.2.2 Archimedean copulae

The Gumbel–Hougaard copula is much more flexible in modeling dependence in the upper tails. For an arbitrary dimension p is is defined as

$$C_\theta^{\mathrm{GH}}(u_1, u_2, \ldots, u_p) = \exp\{-[\sum_{j=1}^{p}(-\log u_j)^\theta]^{1/\theta}\},$$

where $\theta \in [1, \infty)$ is a parameter that governs the strength of the dependence. You can easily see that the Gumbell-Hougaard copula reduces to the independence copula when $\theta = 1$ and to the Fréchet–Hoeffding upper bound copula $\min(u_1, \ldots, u_p)$ when $\theta \to \infty$.

The following figure illustrates a Gumbel–Hougaard copula:



The Gumbel–Hougaard copula is also an example of the so-called *Archimedean* copulae. The latter are characterised by their *generator* $\phi(\cdot)$: a continuous, strictly decreasing, convex function from $[0,1]$ to $[0,\infty)$ such that $\phi(1)=0$. Then the Archimedean copula is defined via the generator as follows:

$$C(u_1,u_2,\ldots,u_p)=\phi^{-1}(\phi(u_1)+\cdots+\phi(u_p)).$$

Here, $\phi^{-1}(t)$ is defined to be 0 if t is not in the image of $\phi(\cdot)$.

Example 15.2. Show that the Gumbell–Hougaard copula is an Archimeden copula with a generator $\phi(t)=(-\log t)^\theta$.

The benefit of using the Archimedean copulae is that they allow for simple description of the p-dim dependence by using a function of one argument only (the generator). However it is seen immediately that the Archimedean copula is symmetric in its arguments and this limits its applicability for modelling dependencies that are not symmetric in their arguments. The so-called *Liouville* copulae are an extension of the Archiemedean copulae and can be used also to model dependencies that are not symmetric in their arguments.

15.3 Margins, estimation, and simulation

So far, we have discussed the copula functions $C(\cdot,\cdot)$ and copula density $c(\cdot,\cdot)$, but using copulae also requires marginal cdfs $F_{X_1}(\cdot)$ and $F_{X_2}(\cdot)$ and pdfs $f_{X_1}(\cdot)$ and $f_{X_2}(\cdot)$ (and so on, for more

than two variables). We can, in fact, specify arbitrary univariate continuous distributions (e.g, normal, gamma, beta, Laplace, etc.) for them. This choice is driven by substantive considerations (E.g., is the distribution positive?)

Then, the density (15.1), appropriately parametrised, provides the likelihood, e.g.,

$$L(\boldsymbol{\rho}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = f_{\boldsymbol{\rho},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2}(x_1, x_2) = c_{\boldsymbol{\rho}}(F_{X_1|\boldsymbol{\theta}_1}(x_1), F_{X_2|\boldsymbol{\theta}_2}(x_2)) f_{X_1|\boldsymbol{\theta}_1}(x_1) f_{X_2|\boldsymbol{\theta}_2}(x_2),$$

which we can maximise in terms of the parameters of the copula and of the marginal distributions to obtain their estimates. A closed form for these estimators is rarely available, and so it is typically done numerically.

However, we might not want to specify margins in the first place. What can we do then? The *empirical distribution function* (edf) $\hat{F}(\cdot)$ is an unbiased estimator for the true cdf $F(\cdot)$. Given X_{ij}, $i = 1, 2$, $j = 1, \ldots, n$ observations we can obtain one for each of the 2 variables:

$$\hat{F}_{X_i}(x) = n^{-1} \sum_{j=1}^{n} \mathbb{I}(X_{ij} \leq x).$$

We can then use it in the copula cdf, i.e.,

$$F(x_1, x_2) = C(\hat{F}_{X_1}(x_1), \hat{F}_{X_2}(x_2)).$$

How do we estimate the parameters of the copula? Although $\hat{F}(\cdot)$ is straightforward, $\hat{f}(\cdot)$ is not and requires further assumptions and tuning parameters (e.g., kernel bandwidth). This means that likelihood $L(\boldsymbol{\rho}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is no longer available to maximise. However, other methods are possible. Typically, we convert the data into empirical quantiles $P_{ij} = \frac{n}{n+1} \hat{F}_{X_i}(X_{ij})$, with denominator $n+1$ used to ensure that P_{ij} run from $\frac{1}{n+1}$ to $\frac{n}{n+1}$. The resulting empirical quantiles will be uniform but maintain their correlations (approximately). Then, we can tune our copula function's parameters until the correlations it induces among the empirical quantiles matches their observed correlations.

Lastly, simulating copulae with parametric margins is straightforward, and simulating copulae with empirical margins is possible as well. $C(\cdot, \cdot)$ and $c(\cdot, \cdot)$ themselves represent a valid distribution with uniform margins can therefore be used to make random dependent draws of marginally uniform quantiles $\boldsymbol{P}_\star = [P_{1\star}, P_{2\star}]^\top$. The variables on the original scale can be obtained using inverse-transform sampling as $X_{i\star} = F^{-1}_{X_i}(P_{i\star})$, $i = 1, 2$ for parametric margins and $X_{i\star} = \hat{F}^{-1}_{X_i}(P_{i\star})$, $i = 1, 2$ for empirical margins. Here, $\hat{F}^{-1}_{X_i}(\cdot)$ is the inverse of the $\hat{F}_{X_i}(\cdot)$, typically smoothed in some way, since $\hat{F}_{X_i}(\cdot)$ represents a discrete distribution.

15.4 Software

SAS: `PROC COPULA`

R: Packages `copula`, `VineCopula`, and others.

15.5 Examples

Example 15.3. Microwave Ovens example (with empirical and gamma margins).

Example 15.4. Stock and portfolio modelling.

15.6 Exercises

Exercise 15.1

The (p-dimensional) Clayton copula is defined for a given parameter $\theta > 0$ as

$$C_\theta(u_1, u_2, \ldots, u_p) = \left[\sum_{i=1}^{p} u_i^{-\theta} - p + 1\right]^{-1/\theta}.$$

Show that it is an Archimedean copula and that its generator is $\phi(x) = \theta^{-1}(x^{-\theta} - 1)$.

A Exercise Solutions

Note that these solutions omit the steps of differentiation and integration, as well as arithmetic, as those can be performed by a computer.

0.1

(a)

1. $\theta x_1 \, \mathrm{e}^{-x_1(\theta+x_2)} \geq 0$ as long as θ, x_1, and $x_2 > 0$.
2. $\int_0^\infty \int_0^\infty \theta x_1 \, \mathrm{e}^{-x_1(\theta+x_2)} \, dx_2 dx_1 = 1$.

(b)

$$\begin{aligned}\Pr(X_1 < t, X_2 < t) = F(t,t) &= \int_0^t \int_0^t \theta x_1 \, \mathrm{e}^{-x_1(\theta+x_2)} \, dx_2 dx_1 \\ &= \frac{t}{\theta+t} + \mathrm{e}^{-t\theta}\left(\frac{\theta \, \mathrm{e}^{-t^2}}{\theta+t} - 1\right).\end{aligned}$$

(c)

$$f_{X_1}(x_1) = \int_0^\infty \theta x_1 \, \mathrm{e}^{-x_1(\theta+x_2)} \, dx_2 = \theta \, \mathrm{e}^{-x_1\theta} \, 1_{x_1>0} \sim \text{Exponential}(\theta).$$

Then $\mathrm{E}(X_1) = \theta^{-1}$ and $\mathrm{Var}(X_1) = \theta^{-2}$.

(d)

$$\begin{aligned}f_{X_2}(x_2) &= \int_0^\infty \theta x_1 \, \mathrm{e}^{-x_1(\theta+x_2)} \, dx_1 \\ &= \frac{\theta}{(\theta+x_2)^2} 1_{x_2>0},\end{aligned}$$

so

$$\begin{aligned}f_{X_2|X_1}(x_2|x_1) &= \frac{f_{\boldsymbol{X}}(x_1,x_2)}{f_{X_1}(x_1)} \\ &= \frac{\theta x_1 \, \mathrm{e}^{-x_1(\theta+x_2)}}{\theta \, \mathrm{e}^{-x_1\theta}} \\ &= x_1 \, \mathrm{e}^{-x_1 x_2} \, 1_{x_2>0} \sim \text{Exponential}(x_1).\end{aligned}$$

(e)

$f_{X_2|X_1}(x_2|x_1) = x_1 \, \mathrm{e}^{-x_1 x_2} \neq \frac{\theta}{(\theta+x_2)^2} = f_{X_2}(x_2)$. More simply, the conditional distribution of $X_2|X_1$ depends on X_1.

0.2

(a)

Let

$$\boldsymbol{Y} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \boldsymbol{X} = \begin{pmatrix} X_1 - X_2 \\ X_1 + X_2 \end{pmatrix}.$$

Then

$$\mathrm{Cov}(\boldsymbol{Y}) = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}^\top = \sigma^2 \begin{pmatrix} 2-2\rho & 0 \\ 0 & 2\rho+2 \end{pmatrix},$$

so $\mathrm{Cov}(X_1 - X_2, X_1 + X_2) = 0$. Note that we only actually require

$$\mathrm{Cov}(X_1 - X_2, X_1 + X_2) = \begin{pmatrix} 1 & -1 \end{pmatrix} \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 0.$$

(b)

$$\mathrm{Cov}(X_1, X_2 - \rho X_1) = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \begin{pmatrix} -\rho \\ 1 \end{pmatrix} = 0.$$

(c)

$$\mathrm{Var}(X_2 - bX_1) = \begin{pmatrix} -b & 1 \end{pmatrix} \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \begin{pmatrix} -b \\ 1 \end{pmatrix} = \sigma^2(b^2 - 2b\rho + 1).$$

$$\frac{\partial b^2 - 2b\rho + 1}{\partial b} = 2b - 2\rho \overset{\text{set}}{=} 0 \implies b = \rho,$$

and $\frac{\partial^2 b^2 - 2b\rho + 1}{\partial b^2} = 2 > 0 \implies b = \rho$ is a minimum.

0.3

(a)

This is trivial, but for additional rigour, we can use Theorem 0.3 letting $A = \begin{pmatrix} \boldsymbol{I}_{p_1} & \boldsymbol{0}_{p_1,p_2} \end{pmatrix} \in \mathcal{M}_{p_1,p}$ and $\boldsymbol{b} = \boldsymbol{0}$. Then $\boldsymbol{X}_{(1)} = A\boldsymbol{X} = \boldsymbol{b}$, and

$$\varphi^{(1)}_{\boldsymbol{X}}(\boldsymbol{s}) = \varphi_{\boldsymbol{X}}\left\{A^\top \boldsymbol{s}\right\} = \varphi_{\boldsymbol{X}}\left\{A^\top \boldsymbol{s}\right\} = \varphi_{\boldsymbol{X}}\left\{ \begin{pmatrix} \boldsymbol{I}_{p_1} \\ \boldsymbol{0}_{p_1,p_2} \end{pmatrix} \boldsymbol{s} \right\} = \varphi_{\boldsymbol{X}}\left\{ \begin{bmatrix} \boldsymbol{s} \\ \boldsymbol{0} \end{bmatrix} \right\}.$$

(b)

If $\boldsymbol{X}_{(1)}$ and $\boldsymbol{X}_{(2)}$ are independent, then $f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{\boldsymbol{X}_{(1)}}(\boldsymbol{x}_{(1)}) f_{\boldsymbol{X}_{(2)}}(\boldsymbol{x}_{(2)})$. Then for

$$\begin{aligned}
\varphi_{\boldsymbol{X}}(\boldsymbol{t}) &= \mathrm{E}(\mathrm{e}^{i\boldsymbol{t}^\top \boldsymbol{X}}) = \int_{\mathbb{R}^p} \mathrm{e}^{i\boldsymbol{t}^\top \boldsymbol{x}} f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} \\
&= \int_{\mathbb{R}^{p_2}} \int_{\mathbb{R}^{p_1}} \mathrm{e}^{i\boldsymbol{t}_{(1)}^\top \boldsymbol{x}_{(1)}} \mathrm{e}^{i\boldsymbol{t}_{(2)}^\top \boldsymbol{x}_{(2)}} f_{\boldsymbol{X}_{(1)}}(\boldsymbol{x}_{(1)}) f_{\boldsymbol{X}_{(2)}}(\boldsymbol{x}_{(2)}) d\boldsymbol{x}_{(1)} d\boldsymbol{x}_{(2)} \\
&= \int_{\mathbb{R}^{p_1}} \mathrm{e}^{i\boldsymbol{t}_{(1)}^\top \boldsymbol{x}_{(1)}} f_{\boldsymbol{X}_{(1)}}(\boldsymbol{x}_{(1)}) d\boldsymbol{x}_{(1)} \int_{\mathbb{R}^{p_2}} \mathrm{e}^{i\boldsymbol{t}_{(2)}^\top \boldsymbol{x}_{(2)}} f_{\boldsymbol{X}_{(2)}}(\boldsymbol{x}_{(2)}) d\boldsymbol{x}_{(2)} \\
&= \varphi_{\boldsymbol{X}}\left\{ \begin{bmatrix} \boldsymbol{t}_{(1)} \\ \boldsymbol{0} \end{bmatrix} \right\} \varphi_{\boldsymbol{X}}\left\{ \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{t}_{(2)} \end{bmatrix} \right\}.
\end{aligned}$$

Conversely, if

$$\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = \varphi_{\boldsymbol{X}}\left\{\begin{bmatrix}\boldsymbol{t}_{(1)}\\ \boldsymbol{0}\end{bmatrix}\right\}\varphi_{\boldsymbol{X}}\left\{\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{t}_{(2)}\end{bmatrix}\right\} = \varphi_{\boldsymbol{X}_{(1)}}(\boldsymbol{t}_{(1)})\varphi_{\boldsymbol{X}_{(2)}}(\boldsymbol{t}_{(2)}),$$

since always,

$$\mathrm{e}^{-i\boldsymbol{t}^\top\boldsymbol{x}} = \mathrm{e}^{-i\boldsymbol{t}_{(1)}^\top\boldsymbol{x}_{(1)}}\,\mathrm{e}^{-i\boldsymbol{t}_{(2)}^\top\boldsymbol{x}_{(2)}},$$

we can take the inverse of the Fourier transform (which cf is),

$$\begin{aligned}
f_{\boldsymbol{X}}(\boldsymbol{x}) &= (2\pi)^{-p}\int_{\mathbb{R}^p}\varphi_{\boldsymbol{X}}(\boldsymbol{t})\,\mathrm{e}^{-i\boldsymbol{t}^\top\boldsymbol{x}}\,d\boldsymbol{t}\\
&= (2\pi)^{-p_1}(2\pi)^{-p_2}\int_{\mathbb{R}^{p_2}}\int_{\mathbb{R}^{p_1}}\varphi_{\boldsymbol{X}_{(1)}}(\boldsymbol{x}_{(1)})\varphi_{\boldsymbol{X}_{(2)}}(\boldsymbol{x}_{(2)})\,\mathrm{e}^{-i\boldsymbol{t}_{(1)}^\top\boldsymbol{x}_{(1)}}\,\mathrm{e}^{-i\boldsymbol{t}_{(2)}^\top\boldsymbol{x}_{(2)}}\,d\boldsymbol{t}_{(1)}d\boldsymbol{t}_{(2)}\\
&= (2\pi)^{-p_1}\int_{\mathbb{R}^{p_1}}\varphi_{\boldsymbol{X}_{(1)}}(\boldsymbol{x}_{(1)})\,\mathrm{e}^{-i\boldsymbol{t}_{(1)}^\top\boldsymbol{x}_{(1)}}\,d\boldsymbol{t}_{(1)}(2\pi)^{-p_2}\int_{\mathbb{R}^{p_2}}\varphi_{\boldsymbol{X}_{(2)}}(\boldsymbol{x}_{(2)})\,\mathrm{e}^{-i\boldsymbol{t}_{(2)}^\top\boldsymbol{x}_{(2)}}\,d\boldsymbol{t}_{(2)}\\
&= f_{\boldsymbol{X}_{(1)}}(\boldsymbol{x}_{(1)})f_{\boldsymbol{X}_{(2)}}(\boldsymbol{x}_{(2)}).
\end{aligned}$$

0.4

Using the notation from Example 0.2, write $X = P\Lambda P^\top$, and denote $\boldsymbol{z} = P^\top\boldsymbol{y}$. Now, since we constrain $\langle\boldsymbol{y},\boldsymbol{e}_1\rangle = 0$, then $z_1 = \langle\boldsymbol{y},\boldsymbol{e}_1\rangle = 0$, so

$$\frac{\boldsymbol{y}^\top X\boldsymbol{y}}{\boldsymbol{y}^\top\boldsymbol{y}} = \frac{\boldsymbol{y}^\top P\Lambda P^\top\boldsymbol{y}}{\boldsymbol{y}^\top\boldsymbol{y}} = \frac{\boldsymbol{z}^\top\Lambda\boldsymbol{z}}{\boldsymbol{z}^\top\boldsymbol{z}} = \frac{\sum_{i=2}^p\lambda_i z_i^2}{\sum_{i=2}^p z_i^2},$$

which we maximise by setting $\boldsymbol{z} = (0 \quad 1 \quad \cdots \quad 0)^\top$ resulting in $\frac{\boldsymbol{z}^\top\Lambda\boldsymbol{z}}{\boldsymbol{z}^\top\boldsymbol{z}} = \lambda_2$.

0.5

First, let us show that an orthogonal projection matrix P has only 0 or 1 as possible eigenvalues. This stems directly from its idempotency: let λ be an eigenvalue of P and $\boldsymbol{y}$ the corresponding eigenvector. Then,

$$P^2\boldsymbol{y} = PP\boldsymbol{y} = \lambda P\boldsymbol{y} = \lambda^2\boldsymbol{y},$$

but idempotency implies that

$$P^2\boldsymbol{y} = P\boldsymbol{y} = \lambda\boldsymbol{y},$$

and so $\lambda = \lambda^2$, forcing it to be either 0 or 1.

Now, spectral decomposition implies that $P = \sum_{i=1}^n\lambda_i\boldsymbol{e}_i\boldsymbol{e}_i^\top$, and so $\mathrm{rk}(P)$ is the number of its nonzero eigenvalues. Meanwhile,

$$\mathrm{tr}(P) = \mathrm{tr}(\sum_{i=1}^n\lambda_i\boldsymbol{e}_i\boldsymbol{e}_i^\top) = \sum_{i=1}^n\lambda_i\,\mathrm{tr}(\boldsymbol{e}_i^\top\boldsymbol{e}_i) = \sum_{i=1}^n\lambda_i 1 = \mathrm{rk}(P).$$

2.1

(a)

Write the joint distribution of Y_1 and Y_2 as

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

then,

$$\operatorname{Var} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \boldsymbol{I}_2 \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix},$$

and Y_1 and Y_2 are independent (being multivariate normal and uncorrelated) and identically distributed $N(0, 2)$.

(b)

$P(\chi^2_2 < 2.41) = 0.7$ (i.e., `pchisq(2.41,2)`).

2.2

(a)

$$\boldsymbol{Z} \sim N\left(\begin{pmatrix} 4 \\ 7 \end{pmatrix}, \begin{pmatrix} 16 & -2 \\ -2 & 7 \end{pmatrix} \right).$$

Hence $\operatorname{Cor}(Z_1, Z_2) = -\frac{2}{\sqrt{16 \times 7}}$.

(b)

Take $\begin{pmatrix} \tilde{\boldsymbol{X}}_{(1)} \\ \hline \tilde{X}_{(2)} \end{pmatrix} = \left(\begin{array}{c} X_1 \\ X_3 \\ \hline X_2 \end{array}\right)$, and rearrange to get distribution is $N_3\left(\left(\begin{array}{c} 3 \\ 2 \\ \hline -1 \end{array}\right), \left(\begin{array}{cc|c} 3 & 1 & 2 \\ 1 & 2 & 1 \\ \hline 2 & 1 & 3 \end{array}\right) \right)$.
Call its mean and variance $\tilde{\mu}$ and $\tilde{\Sigma}$. Then, $X_1, X_3 \mid X_2 \sim N\left(\tilde{\boldsymbol{\mu}}_{(1)|(2)}, \quad \tilde{\Sigma}_{(1)|(2)}\right)$ where

$$\tilde{\boldsymbol{\mu}}_{(1)|(2)} = \tilde{\boldsymbol{\mu}}_{(1)} + \tilde{\Sigma}_{(1)(2)} \tilde{\Sigma}^{-1}_{(2)(2)} \left(\tilde{X}_{(2)} - \tilde{\mu}_{(2)} \right) = \begin{pmatrix} 3 \\ 2 \end{pmatrix} + \begin{pmatrix} 2 \\ 1 \end{pmatrix} \frac{1}{3} (X_2 + 1) = \begin{pmatrix} 3 \\ 2 \end{pmatrix} + \begin{pmatrix} 2/3 \\ 1/3 \end{pmatrix} (X_2 + 1)$$

$$\tilde{\Sigma}_{(1)|(2)} = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix} - \begin{pmatrix} 2 \\ 1 \end{pmatrix} \frac{1}{3} \begin{pmatrix} 2 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix} - \begin{pmatrix} 4/3 & 2/3 \\ 2/3 & 1/3 \end{pmatrix} = \frac{1}{3} \begin{pmatrix} 5 & 1 \\ 1 & 5 \end{pmatrix}$$

In particular, for $x_2 = 0$ we get,

$$X_1, X_3 \mid X_2 \sim N\left(\begin{pmatrix} 3\frac{2}{3} \\ 2\frac{1}{3} \end{pmatrix}, \frac{1}{3} \begin{pmatrix} 5 & 1 \\ 1 & 5 \end{pmatrix} \right).$$

2.3

Take $\boldsymbol{t} \in \mathbb{R}^p$. Observe that $a_1\boldsymbol{X}_1+\cdots+a_n\boldsymbol{X}_n = \boldsymbol{X}A$ for $A = [a_1, \ldots, a_n]$ and $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n]$. Then, along the lines of Theorem 0.3,

$$\begin{aligned}\varphi_{a_1\boldsymbol{X}_1+\cdots+a_n\boldsymbol{X}_n}(\boldsymbol{t}) &= \prod_{j=1}^{n} \varphi_{a_j\boldsymbol{X}_j}(\boldsymbol{t}) = \prod_{j=1}^{n} \mathrm{e}^{ia_j\boldsymbol{t}^\top\boldsymbol{\mu}_j - \frac{a_j^2}{2}\boldsymbol{t}^\top\Sigma_j\boldsymbol{t}} \\ &= \mathrm{e}^{i\boldsymbol{t}^\top(\sum_{j=1}^n a_j\boldsymbol{\mu}_j) - \frac{1}{2}\boldsymbol{t}^\top(\sum_{j=1}^n a_j^2\Sigma_j)\boldsymbol{t}} = \varphi_{N(\sum_{j=1}^n a_j\boldsymbol{\mu}_j, \sum_{j=1}^n a_j^2\Sigma_j)}(\boldsymbol{t})\end{aligned}$$

(by definition).

Then, substitute $\boldsymbol{\mu}_i = \boldsymbol{\mu}$, $\Sigma_i = \Sigma$, and $a_i = \frac{1}{n}$ for all $i = 1, \ldots, n$ to obtain the distribution of $\bar{\boldsymbol{X}}$.

2.4

By Property 4, the conditional distribution $\boldsymbol{X}_2 \mid \boldsymbol{X}_1 = \boldsymbol{x}_1$ must have the form $A\boldsymbol{x}_1 + \boldsymbol{b} + \boldsymbol{X}_3$ (i.e., a linear combination of $\boldsymbol{x}_1$, a constant, and some noise $\boldsymbol{X}_3 \sim N(\boldsymbol{0}, \Omega)$ independent of $\boldsymbol{X}_1$). Hence, the marginal distribution of $\boldsymbol{X}_2$ is the same as the distribution of $A\boldsymbol{X}_1 + \boldsymbol{b} + \boldsymbol{X}_3$. But then,

$$\boldsymbol{X} = \begin{pmatrix}\boldsymbol{X}_1 \\ \boldsymbol{X}_2\end{pmatrix} = \begin{pmatrix}\boldsymbol{I}_r & \boldsymbol{0} \\ A & \boldsymbol{I}_{p-r}\end{pmatrix}\begin{pmatrix}\boldsymbol{X}_1 \\ \boldsymbol{X}_3\end{pmatrix} + \begin{pmatrix}\boldsymbol{0} \\ \boldsymbol{b}\end{pmatrix}$$

and will be multivariate normal. We only need the mean and the covariance matrix.

Now, $\mathrm{E}(\boldsymbol{X}_1) = \boldsymbol{\mu}_1$ and $\mathrm{E}(\boldsymbol{X}_2) = \mathrm{E}_{\boldsymbol{X}_1}[\mathrm{E}_{\boldsymbol{X}_2}(\boldsymbol{X}_2 \mid \boldsymbol{X}_1)] = \mathrm{E}_{\boldsymbol{X}_1}[\mathrm{E}_{\boldsymbol{X}_3}(A\boldsymbol{X}_1+\boldsymbol{b}+\boldsymbol{X}_3)] = A\boldsymbol{\mu}_1+\boldsymbol{b}$, and

$$\mathrm{Var}(\boldsymbol{X}_2) = \mathrm{Var}(A\boldsymbol{X}_1 + \boldsymbol{X}_3) = A\Sigma_{11}A^\top + \Omega,$$

with

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{X}_1, \boldsymbol{X}_2) &= \mathrm{E}[(\boldsymbol{X}_1 - \boldsymbol{\mu}_1)(A\boldsymbol{X}_1 + \boldsymbol{b} - A\boldsymbol{\mu}_1 - \boldsymbol{b})^\top] \\ &= \mathrm{E}[(\boldsymbol{X}_1 - \boldsymbol{\mu}_1)(\boldsymbol{X}_1 - \boldsymbol{\mu}_1)^\top A^\top] \\ &= \mathrm{E}[(\boldsymbol{X}_1 - \mu_1)(\boldsymbol{X}_1 - \mu_1)^\top]A^\top = \Sigma_{11}A^\top,\end{aligned}$$

hence

$$\boldsymbol{X} \sim N_p\left(\begin{pmatrix}\boldsymbol{\mu}_1 \\ A\boldsymbol{\mu}_1 + \boldsymbol{b}\end{pmatrix}, \quad \begin{pmatrix}\Sigma_{11} & \Sigma_{11}A^\top \\ A\Sigma_{11} & \Omega + A\Sigma_{11}A^\top\end{pmatrix}\right).$$

2.5

(a)

Using Exercise 2.4, we can get the joint distribution of $\begin{pmatrix}Z \\ Y\end{pmatrix} \sim N_2\left(\begin{pmatrix}0 \\ 1\end{pmatrix}, \quad \begin{pmatrix}1 & 1 \\ 1 & 2\end{pmatrix}\right)$ (or, equivalently $\begin{pmatrix}Y \\ Z\end{pmatrix} \sim N_2\left(\begin{pmatrix}1 \\ 0\end{pmatrix}, \quad \begin{pmatrix}2 & 1 \\ 1 & 1\end{pmatrix}\right)$. Applying the same procedure again, we can get (with $\Omega = 1$, $b = 1$, and $A = (-1, 0)$)

$$\begin{pmatrix}Y \\ Z \\ X\end{pmatrix} \sim N_3\left(\begin{pmatrix}1 \\ 0 \\ 0\end{pmatrix}, \quad \begin{pmatrix}2 & 1 & -2 \\ 1 & 1 & -1 \\ -2 & -1 & 3\end{pmatrix}\right)$$

or, equivalently,

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 3 & -2 & -1 \\ -2 & 2 & 1 \\ -1 & 1 & 1 \end{pmatrix} \right).$$

Then, $Y \mid (X, Z)$ is normal with

$$\begin{aligned} \mu_{Y|(X,Z)} &= 1 + \begin{pmatrix} -2 & 1 \end{pmatrix} \begin{pmatrix} 3 & -1 \\ -1 & 1 \end{pmatrix}^{-1} \left(\begin{pmatrix} X \\ Z \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right) \\ &= 1 + \begin{pmatrix} -2 & 1 \end{pmatrix} \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} X \\ Z \end{pmatrix} = 1 + \frac{1}{2}(Z - X) \end{aligned}$$

and

$$\sigma^2_{Y|(X,Z)} = 2 - \begin{pmatrix} -2 & 1 \end{pmatrix} \begin{pmatrix} 3 & -1 \\ -1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} -2 \\ 1 \end{pmatrix} = \frac{1}{2}:$$

$$Y \mid (X, Z) \sim N(1 + \frac{1}{2}(Z - X), \frac{1}{2})$$

(b)

$$\begin{pmatrix} U \\ V \end{pmatrix} = \begin{pmatrix} 1 + Z \\ 1 - Y \end{pmatrix}$$

is obviously normal. Moreover, $\mu_U = 1 + \mathrm{E}(Z) = 1$, $\mu_V = \mathrm{E}(1 - Y) = 0$, $\sigma^2_U = \sigma^2_Z = 1$, $\sigma^2_V = \sigma^2_Y = 2$, $\sigma_{U,V} = -\sigma_{Z,Y} = -1$. Hence,

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim N_2\left(\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \right).$$

(c)

$Y \mid (U = 2)$ has the same distribution as $Y \mid Z + 1 = 2$, that is, $Y \mid Z = 1$. Using (b), we get $Y \mid U = 2 \sim N_1(2, 1)$.

3.1

$$\begin{pmatrix} 5 & 0 & 10 & 0 \\ 2 & 1 & 4 & 2 \\ 15 & 0 & 20 & 0 \\ 6 & 3 & 8 & 4 \end{pmatrix}$$

4.1

(a)

$$C = \begin{pmatrix} -1 & 1 & 0 & 0 & \cdots & \cdots \\ 0 & -1 & 1 & 0 & \cdots & \cdots \\ 0 & 0 & -1 & 1 & \ddots & \cdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & -1 & 1 \end{pmatrix} \in \mathcal{M}_{p-1,p}$$

is the required matrix.

(b)

$\boldsymbol{Y}_j = C\boldsymbol{X}_j \to \boldsymbol{Y}_j$ are i.i.d. $N(\ C\boldsymbol{\mu},\quad C\Sigma C^\top\ )$, $S_Y = CSC^\top$, $\bar{\boldsymbol{Y}} = C\bar{\boldsymbol{X}}$, $\boldsymbol{\mu}_Y = C\boldsymbol{\mu}$

$$n\left(\bar{\boldsymbol{Y}} - \boldsymbol{\mu}_Y\right)^\top S_Y^{-1}\left(\bar{\boldsymbol{Y}} - \boldsymbol{\mu}_Y\right) = n\left(\bar{\boldsymbol{X}} - \boldsymbol{\mu}\right)^\top C^\top (CSC^\top)^{-1} C(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \sim \frac{(n-1)(p-1)}{n-p+1} F_{p-1,n-p+1}$$

Hence, the rejection region would be

$$\left\{\boldsymbol{X} : n\left(C\bar{\boldsymbol{X}} - \mathbf{1}\right)^\top (CSC^\top)^{-1}(C\bar{\boldsymbol{X}} - \mathbf{1}) > \frac{(n-1)(p-1)}{n-p+1} F_{1-\alpha,p-1,n-p+1}\right\},$$

where $\mathbf{1}_{p-1} \in \mathbb{R}^{p-1}$ is a $(p-1)$ vector of ones.

4.2

Use the fact that $\boldsymbol{Y} = n\left(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0\right)^\top \Sigma^{-1}\left(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0\right) \sim \chi^2_3$: plug-in $n = 50$, $\bar{\boldsymbol{X}} = \begin{pmatrix} 0.8 \\ 1.1 \\ 0.6 \end{pmatrix}$, $\boldsymbol{\mu}_0 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 4 & 1 \\ 1 & 1 & 2 \end{pmatrix}$, find Σ^{-1}, and hence reject if $\boldsymbol{Y} > \chi^2_{1-\alpha,3}$.

4.3

From the data, $\bar{\boldsymbol{X}} = [6, 10]^\top$ and $\boldsymbol{S} = \begin{pmatrix} 24 & -10 \\ -10 & 6 \end{pmatrix}/3$ and $\boldsymbol{S}^{-1} = \begin{pmatrix} 18 & 30 \\ 30 & 72 \end{pmatrix}/44$. Then,

$$T^2 = n\left(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0\right)^\top \boldsymbol{S}^{-1}\left(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0\right) = 13.636.$$

To compute the P-value, evaluate $F = \frac{n-p}{(n-1)p} T^2 = 4.545$ and P-value $= \Pr(F \geq F_{p,n-p}) = 0.180 > 0.05$ (i.e., `pf(4.545455, 2, 2, lower.tail=FALSE)`). Do not reject H_0: there is not sufficient evidence to believe that the population mean differs from $[7, 11]^\top$.

4.4

Use Exercise 2.3 on the two samples, then the property of the difference of means. Observe that the variance does not depend on the means, and so we can the pooled T^2 test (4.9).

4.5

For a difference of independent variables, means subtract and variances add, so $\boldsymbol{X} - \bar{\boldsymbol{X}} \sim N_p\left(0, \quad \left(1 + \frac{1}{n}\right)\Sigma\right)$ and $(n-1)\boldsymbol{S} \sim W_p\left(\Sigma, n-1\right)$ by definition, and they are independent. Call $\boldsymbol{C} = \boldsymbol{X} - \bar{\boldsymbol{X}}$. Then,

$$\boldsymbol{C}\frac{n}{n+1}(\boldsymbol{X} - \bar{\boldsymbol{X}})^\top \boldsymbol{S}^{-1}(\boldsymbol{X} - \bar{\boldsymbol{X}}) = \frac{\boldsymbol{C}^\top \boldsymbol{S}^{-1}\boldsymbol{C}}{\boldsymbol{C}^\top \Sigma^{-1}\boldsymbol{C}}\left(\frac{n}{n+1}\right)(\boldsymbol{C}^\top \Sigma^{-1}\boldsymbol{C}).$$

Now,

$$\frac{\boldsymbol{C}^\top \boldsymbol{S}^{-1}\boldsymbol{C}}{\boldsymbol{C}^\top \Sigma^{-1}\boldsymbol{C}} = \frac{n-1}{\chi^2_{n-p}},$$

independent of $\boldsymbol{X}$ or $\bar{\boldsymbol{X}}$, so

$$\frac{n}{n+1}\boldsymbol{C}^\top \Sigma^{-1}\boldsymbol{C} = (\boldsymbol{X} - \bar{\boldsymbol{X}})^\top \left(\left(1 + \frac{1}{n}\right)\Sigma\right)(\boldsymbol{X} - \bar{\boldsymbol{X}}) \sim \chi^2_p$$

and independent of $\boldsymbol{S}$. Hence

$$\frac{n}{n+1}(\boldsymbol{X} - \bar{\boldsymbol{X}})^\top \boldsymbol{S}^{-1}(\boldsymbol{X} - \bar{\boldsymbol{X}}) \sim \frac{(n-1)\chi^2_p}{\chi^2_{n-p}},$$

i.e., the distribution asked: $\sim \frac{p(n-1)}{(n-p)}F_{p,n-p}$ (same as distribution of T^2). Then, $(1-\alpha)100\%$ prediction region would be:

$$\left\{\boldsymbol{X} : \frac{n}{n+1}(\boldsymbol{X} - \bar{\boldsymbol{X}})^\top \boldsymbol{S}^{-1}(\boldsymbol{X} - \bar{\boldsymbol{X}}) < \frac{p(n-1)}{(n-p)}F_{1-\alpha,p,n-p}\right\}.$$

5.1

(a)

Let $\tilde{X}_4 = X_1 + X_2 + X_4$. We can obtain what we are looking for as a linear combination:

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \tilde{X}_4 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_1 + X_2 + X_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix},$$

Then,

$$\mathrm{E} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \tilde{X}_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 7 \end{pmatrix}$$

and

$$\mathrm{Var} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \tilde{X}_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 3 & 1 & 0 & 1 \\ 1 & 4 & 0 & 0 \\ 0 & 0 & 1 & 4 \\ 1 & 0 & 4 & 20 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 1 & 0 & 5 \\ 1 & 4 & 0 & 5 \\ 0 & 0 & 1 & 4 \\ 5 & 5 & 4 & 31 \end{pmatrix},$$

so

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \tilde{X}_4 \end{pmatrix} \sim N \left(\begin{pmatrix} 1 \\ 2 \\ 3 \\ 7 \end{pmatrix}, \begin{pmatrix} 3 & 1 & 0 & 5 \\ 1 & 4 & 0 & 5 \\ 0 & 0 & 1 & 4 \\ 5 & 5 & 4 & 31 \end{pmatrix} \right).$$

(b)

Using the expression for the conditional distribution of a normal distribution,

$$\begin{aligned} \mathrm{E} \left(X_1 \middle| \begin{pmatrix} X_2 \\ X_3 \\ X_4 \end{pmatrix} \right) &= 1 + \begin{pmatrix} 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 4 & 0 & 0 \\ 0 & 1 & 4 \\ 0 & 4 & 20 \end{pmatrix}^{-1} \begin{pmatrix} x_2 - 2 \\ x_3 - 3 \\ x_4 - 4 \end{pmatrix} \\ &= 1 + \begin{pmatrix} 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1/4 & 0 & 0 \\ 0 & 5 & -1 \\ 0 & -1 & 1/4 \end{pmatrix} \begin{pmatrix} x_2 - 2 \\ x_3 - 3 \\ x_4 - 4 \end{pmatrix} \\ &= 1 + \begin{pmatrix} 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} -0.5 + \frac{x_2}{4} \\ 11 + 5x_3 - x_4 \\ 1 - x_3 + \frac{x_4}{4} \end{pmatrix} \\ &= 1 - 0.5 + \frac{x_2}{4} + 2 - x_3 + \frac{x_4}{4} \\ &= 2.5 + \frac{x_2}{4} - x_3 + \frac{x_4}{4}. \end{aligned}$$

And,

$$\text{Var}\left(X_1 \middle| \begin{pmatrix} X_2 \\ X_3 \\ X_4 \end{pmatrix}\right) = 3 - \begin{pmatrix} 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{4} & 0 & 0 \\ 0 & 5 & -1 \\ 0 & -1 & \frac{1}{4} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$$
$$= 3 - \begin{pmatrix} 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{4} \\ -1 \\ \frac{1}{4} \end{pmatrix} = 2.5.$$

(c)

Looking at the upper part $\begin{pmatrix} 3 & 1 & 0 \\ 1 & 4 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ of the covariance matrix, we see that X_3 is independent of $\begin{pmatrix} X_1, & X_2 \end{pmatrix}$. Hence x_3 does not influence the correlation of X_1 and $X_2 \implies \rho_{12.3} = \rho_{12} = \frac{\sqrt{3}}{6} = 0.2887$.

For $\rho_{12.4}$,

$$\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \begin{pmatrix} 3 & 1 \\ 1 & 4 \end{pmatrix} - \frac{1}{20}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \frac{59}{20} & 1 \\ 1 & 4 \end{pmatrix}.$$

Hence, $\rho_{12.4} = \sqrt{\frac{5}{59}} = 0.291$.

(d)

$$R_{1.234} = \sqrt{\frac{\begin{pmatrix} 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 4 & 0 & 0 \\ 0 & 1 & 4 \\ 0 & 4 & 20 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}}{3}}$$
$$= \sqrt{\frac{1}{3}\begin{pmatrix} 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{4} & 0 & 0 \\ 0 & 5 & -1 \\ 0 & -1 & \frac{1}{4} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}}$$
$$= \sqrt{\frac{1}{6}} = 0.408 > \rho_{12}.$$

Of course $R_{1.234}$ should be larger than ρ_{12} (or at least no smaller), and this is supported numerically ($0.408 > 0.2887$).

(e)

Consider $\begin{pmatrix} X_2 \\ X_3 \\ X_4 \end{pmatrix}$ and

$$X_1 - \begin{pmatrix} 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{X_2}{4} \\ 5X_3 - X_4 \\ \frac{X_4}{4} - X_3 \end{pmatrix} = X_1 - \frac{X_2}{4} - \frac{X_4}{4} + X_3.$$

Then directly you can check:

$$\begin{aligned}\mathrm{Cov}(X_2, X_1 - \frac{X_2}{4} - \frac{X_4}{4} + X_3) &= 1 - 1 = 0\\ \mathrm{Cov}(X_{3,} X_1 - \frac{X_2}{4} - \frac{X_4}{4} + X_3) &= -1 + 1 = 0\\ \mathrm{Cov}(X_{4,} X_1 - \frac{X_2}{4} - \frac{X_4}{4} + X_3) &= 1 - 5 + 4 = 0.\end{aligned}$$

But more clever is to say: $X_1 - \mathrm{E}\left(X_1 \middle| \begin{pmatrix} x_2 \\ x_3 \\ x_4 \end{pmatrix}\right)$ and $\begin{pmatrix} x_2 \\ x_3 \\ x_4 \end{pmatrix}$ are uncorrelated. This general argument was put forward and proved as a part of the proof of Property 4 of the Multivariate Normal Distribution in Section 2.2.

5.2

(a)

$$\begin{aligned}\begin{pmatrix} 3 \\ -2 \\ 1 \end{pmatrix}^\top \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} &\sim N\left(\begin{pmatrix} 3 & -2 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ -3 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 & -2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 2 \end{pmatrix} \begin{pmatrix} 3 \\ -2 \\ 1 \end{pmatrix} \right)\\ &\sim N(13, 9).\end{aligned}$$

(b)

Let vector $\boldsymbol{a} = \begin{pmatrix} U \\ V \end{pmatrix}$.

$$\begin{aligned}\mathrm{Cov}(X_2, X_2 - UX_1 - VX_3) &= \mathrm{Var}(X_2) - U\,\mathrm{Cov}(X_2, X_1) - V\,\mathrm{Cov}(X_2, X_3)\\ &= 3 - U - 2V = 0.\end{aligned}$$

Then, if, say, $U = 1$, then $V = 1$, so $\boldsymbol{a} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

6.1

First, let us note that not all $\rho > 0$ are allowed since Σ must be non-negative definite. It must hold that

$$\begin{vmatrix} 1 & \rho/2 \\ \rho/2 & 1 \end{vmatrix} = 1 - \rho^2/4 \geq 0$$

(since otherwise, for some $a \in \mathbb{R}$ and $b \in \mathbb{R}$,

$$\begin{pmatrix} a \\ b \\ 0 \end{pmatrix}^\top \begin{pmatrix} 1 & \rho/2 & 0 \\ \rho/2 & 1 & \rho \\ 0 & \rho & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ 0 \end{pmatrix} = a^2 + a\rho/2 + b\rho/2 + b^2 < 0,$$

making the whole matrix no longer non-negative definite) and

$$\begin{vmatrix} 1 & \rho/2 & 0 \\ \rho/2 & 1 & \rho \\ 0 & \rho & 1 \end{vmatrix} = 1 - \frac{5}{4}\rho^2 \geq 0.$$

This means that $0 < \rho \leq \frac{2}{\sqrt{5}}$.

(a)

First, let us find the 3 eigenvalues of Σ:

$$\begin{vmatrix} 1-\lambda & \rho/2 & 0 \\ \rho/2 & 1-\lambda & \rho \\ 0 & \rho & 1-\lambda \end{vmatrix} = 1 - 3\lambda + 3\lambda^2 - \lambda^3 - \frac{5}{4}\rho^2(1-\lambda) = 0$$

and

$$(1-\lambda)\left[\lambda^2 - 2\lambda + 1 - \frac{5}{4}\rho^2\right] = 0.$$

Solving this equation, we obtain three roots: $\lambda_1 = 1$, $\lambda_2 = 1 - \frac{\sqrt{5}}{2}\rho$, and $\lambda_3 = 1 + \frac{\sqrt{5}}{2}\rho$. The larges eigenvalue is $\lambda_3 = 1 + \frac{\sqrt{5}}{2}\rho$.

By definition, its corresponding eigenvector $\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}$ satisfies,

$$a_1 + \frac{\rho}{2}a_2 = a_1 + \frac{\sqrt{5}}{2}\rho a_1$$

$$\frac{\rho}{2}a_1 + a_2 + \rho a_3 = a_2 + \frac{\sqrt{5}}{2}\rho a_2$$

$$\rho a_2 + a_3 = a_3 + \frac{\sqrt{5}}{2}\rho a_3.$$

Solving (up to a constant), $a_2 = \sqrt{5}a_1$, $a_3 = \frac{2}{\sqrt{5}}a_2 = 2a_1$.

So $a_1 \begin{pmatrix} 1 \\ \sqrt{5} \\ 2 \end{pmatrix}$ is an eigenvector. To normalise it, choose $a_1 = \frac{1}{\sqrt{10}}$. Thus, the first principal component is

$$\frac{1}{\sqrt{10}}Y_1 + \sqrt{\frac{1}{2}}Y_2 + \frac{2}{\sqrt{10}}Y_3.$$

It explains $\frac{1+\frac{\sqrt{5}}{2}\rho}{3} \cdot 100\%$ of the overall variability.

(b)

$$\begin{aligned}
\begin{pmatrix} Y_1 \\ Y_2 \\ Y_1+Y_2+Y_3 \end{pmatrix} &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} \\
&\sim N\left(\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & \frac{\rho}{2} & 0 \\ \frac{\rho}{2} & 1 & \rho \\ 0 & \rho & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \right) \\
&\sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \frac{\rho}{2} & 1+\frac{\rho}{2} \\ \frac{\rho}{2} & 1 & 1+\frac{3}{2}\rho \\ 1+\frac{\rho}{2} & 1+\frac{3}{2}\rho & 3(1+\rho) \end{pmatrix} \right).
\end{aligned}$$

(c)

$$N\left(\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \rho \end{pmatrix} y_3 \right), \begin{pmatrix} 1 & \frac{\rho}{2} \\ \frac{\rho}{2} & 1 \end{pmatrix} - \begin{pmatrix} 0 \\ \rho \end{pmatrix} \begin{pmatrix} 0 & \rho \end{pmatrix} \right) = N\left(\begin{pmatrix} 0 \\ \rho y_3 \end{pmatrix}, \begin{pmatrix} 1 & \frac{\rho}{2} \\ \frac{\rho}{2} & 1-\rho^2 \end{pmatrix} \right).$$

(d)

$$\operatorname{Cov} \begin{pmatrix} Y_3 \\ Y_2 \\ Y_1 \end{pmatrix} = \begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & \frac{\rho}{2} \\ 0 & \frac{\rho}{2} & 1 \end{pmatrix},$$

so

$$\begin{aligned}
R &= \sqrt{\frac{\begin{pmatrix} \rho & 0 \end{pmatrix} \begin{pmatrix} 1 & \frac{\rho}{2} \\ \frac{\rho}{2} & 1 \end{pmatrix}^{-1} \begin{pmatrix} \rho \\ 0 \end{pmatrix}}{1}} \\
&= \frac{1}{\sqrt{1-\frac{\rho^2}{4}}} \sqrt{\begin{pmatrix} \rho & 0 \end{pmatrix} \begin{pmatrix} 1 & -\frac{\rho}{2} \\ -\frac{\rho}{2} & 1 \end{pmatrix} \begin{pmatrix} \rho \\ 0 \end{pmatrix}} = \frac{\rho}{\sqrt{1-\frac{\rho^2}{4}}}.
\end{aligned}$$

7.1

Split up the matrix: $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ into

$$\Sigma_{11} = \begin{pmatrix} 1 & 0.4248 \\ 0.4248 & 1 \end{pmatrix},$$

$$\Sigma_{12} = \begin{pmatrix} 0.0420 & 0.0215 & 0.0573 \\ 0.1487 & 0.2489 & 0.2843 \end{pmatrix},$$

$$\Sigma_{22} = \begin{pmatrix} 1 & 0.6693 & 0.4662 \\ 0.6693 & 1 & 0.6915 \\ 0.4662 & 0.6915 & 1 \end{pmatrix},$$

$$\Sigma_{21} = \Sigma_{12}.$$

We need to find eigenvalues for $\Sigma_{12}^\top \Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1}$ if calculating by hand (this would be easier than finding eigenvalues of $\Sigma_{22}^{-\frac{1}{2}} \Sigma_{12}^\top \Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-\frac{1}{2}}$). If using SAS, we would use the following statements:

```
proc iml;
S_11 = {1  0.4248 , 0.4248  1} ;
S_12 = {0.0420  0.0215  0.0573 , 0.1487  0.2489  0.2843} ;
S_22 = {1 0.6693  0.4662 , 0.6693  1  0.6915 , 0.4662  0.6915  1};
S_22inv = inv(S_22);
S_r = root(S_22inv);
a = S_r*S_12'*inv(S_11)*S_12*S_r';
call eigen(c, d, a);
print c; print d;
```

The result, $C = \begin{pmatrix} 0.0946455 \\ 0.0035185 \\ 2.252 \times 10^{-18} \end{pmatrix}$, $D = \begin{pmatrix} -0.1281 & 0.7192 & 0.6829 \\ 0.2840 & -0.6331 & 0.7201 \\ 0.9502 & 0.2862 & -0.1232 \end{pmatrix}$. Further,

```
b=S_r'*d[,1];
a=1/sqrt(0.09464557)*inv(S_11)*(S_12)*b;
```

gives $\boldsymbol{a} = \begin{pmatrix} 0.3262 \\ -1.0940 \end{pmatrix}$ and $\boldsymbol{b} = \begin{pmatrix} 0.1724 \\ -0.5079 \\ -0.6794 \end{pmatrix}$ with $\boldsymbol{a}^\top \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ and $\boldsymbol{b}^\top \begin{pmatrix} X_3 \\ X_4 \\ X_5 \end{pmatrix}$ the canonical variates, and relevant eigenvalues $\lambda = 0.0946, 0.0035$.

In R,

```
s <- c(1, 0.4248, 0.0420, 0.0215, 0.0573,
               1, 0.1487, 0.2489, 0.2843,
                      1, 0.6693, 0.4662,
                             1, 0.6915,
                                    1)
S <- matrix(NA, 5, 5)
S[lower.tri(S,TRUE)] <- s
S[upper.tri(S)] <- t(S)[upper.tri(S)]
S_11 <- S[1:2,1:2]
```

```
S_12 <- S[1:2,3:5]
S_22 <- S[3:5,3:5]
S_22inv <- solve(S_22)
S_r <- chol(S_22inv) # Can use Cholesky instead of square root.
A <- S_r%*%t(S_12)%*%solve(S_11)%*%S_12%*%t(S_r)
(e <- eigen(A))
(b <- t(S_r)%*%e$vectors[,1])
(a <- 1/sqrt(e$values[1]) * solve(S_11)%*%S_12%*%b)
```

This suggests that the first canonical correlation is sufficient. The first canonical correlation represents primarily a positive association between arithmetic power and memory for symbols (both kinds).

7.2

R and SAS implementations are as in previous exercise, but with modified matrices give $\boldsymbol{a} = \begin{pmatrix} -0.0260 \\ -0.0518 \end{pmatrix}$, $\boldsymbol{b} = \begin{pmatrix} -0.0823 \\ -0.0081 \\ -0.0035 \end{pmatrix}$, with the relevant eigenvalues $\lambda = 0.4396, 0.0016$.

The eigenvalues suggest that there is little for the second canonical correlation left to explain. (I.e., a factor of over 200.)

The first canonical correlation appears to indicate a positive relationship (i.e., negative $\times$ negative) between the first open book exam and the two closed book exams, whereas the other two open book exams are weakly associated with the closed book exams. (Rerunning after converting to correlation matrix does not change this.)

7.3

(a)

The following is an outline of the solution:

1. Since this makes them easier to perform, work with the matrix $\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}$
2. Using the 2×2 matrix inversion formula, evaluate it.
3. Using the 2×2 martix determinant formula, find the expression for the characteristic polynomial, in terms of ρ and λ.

You should get $\lambda_1 = \frac{4\rho^2}{1+4\rho+4\rho^2}$ and $\lambda_2 = 0$, which means that one canonical variables pair is enough.

(b)

Similarly, solve for eigenvectors of $\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}$ and transform them.

7.4

(a)

Splitting this up, $\Sigma_{11} = \begin{pmatrix} 100 & 0 \\ 0 & 1 \end{pmatrix}$, $\Sigma_{22} = \begin{pmatrix} 1 & 0 \\ 0 & 100 \end{pmatrix}$, $\Sigma_{12} = \begin{pmatrix} 0 & 0 \\ 0.95 & 0 \end{pmatrix}$, $\Sigma_{22}^{-1} = \frac{1}{100}\begin{pmatrix} 100 & 0 \\ 0 & 1 \end{pmatrix}$,$\Sigma_{22}^{-\frac{1}{2}} = \frac{1}{10}\begin{pmatrix} 10 & 0 \\ 0 & 1 \end{pmatrix}$, so

$$\Sigma_{22}^{-\frac{1}{2}}\Sigma_{12}^{\top}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-\frac{1}{2}} = \begin{pmatrix} (0.95)^2 & 0 \\ 0 & 0 \end{pmatrix}.$$

$\mu^2 = (0.95)^2$ and eigenvector $\boldsymbol{b} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, hence $Z_2 = 1 \times X_3 + 0 \times X_4 = X_3$ and

$$\boldsymbol{a} = \frac{1}{0.95}\frac{1}{100}\begin{pmatrix} 1 & 0 \\ 0 & 100 \end{pmatrix}\begin{pmatrix} 0 & 0 \\ 0.95 & 0 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

so $Z_1 = 0 \times X_1 + 1 \times X_2 = X_2$. $\mu^2 = (0.95)^2$, and $\mu = 0.95$ is the first canonical correlation.

Can you give another argument for the canonical variables and canonical correlation in this problem that will help you to avoid all the calculations above?

9.1

Since $\boldsymbol{S} = \frac{1}{n}\boldsymbol{V} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top$ (using n instead of $n-1$ here to simplify notation—the factor cancels), observe that

$$\begin{aligned}
\text{arithm. mean } \hat{\lambda}_i &= \frac{1}{p}\sum_{i=1}^{p}\hat{\lambda}_i = \frac{1}{p}\operatorname{tr}(\boldsymbol{S}) \\
&= \frac{1}{pn}\operatorname{tr}\{\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top\} \\
&= \frac{1}{pn}\operatorname{tr}\{\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\} \\
&= \hat{\sigma}^2
\end{aligned}$$

and

$$\text{geom. mean } \hat{\lambda}_i \;= \left(\prod_{i=1}^{p}\hat{\lambda}_i\right)^{1/p} = |\boldsymbol{S}|^{1/p} = (\frac{1}{n^p}|\boldsymbol{V}|)^{1/p} = \frac{1}{n}|\boldsymbol{V}|^{1/p}.$$

Substituting,

$$\begin{aligned}
-2\log\Lambda = np\log\frac{\text{arithm. mean } \hat{\lambda}_i}{\text{geom. mean } \hat{\lambda}_i} &= np\log\frac{\hat{\sigma}^2}{\frac{1}{n}|\boldsymbol{V}|^{1/p}} \\
&= np\log n\hat{\sigma}^2 - np\log|\boldsymbol{V}|^{1/p} \\
&= np\log n\hat{\sigma}^2 - n\log|\boldsymbol{V}|,
\end{aligned}$$

the test statistic from Section 9.2.

9.2

Observe that we can write the sample correlation matrix as

$$\boldsymbol{R} = \operatorname{diag}(\hat{\boldsymbol{\Sigma}})^{-1/2}\hat{\boldsymbol{\Sigma}}\operatorname{diag}(\hat{\boldsymbol{\Sigma}})^{-1/2},$$

where

$$\operatorname{diag}(A)_{ij} = \begin{cases} A_{ii} & i = j \\ 0 & \text{otherwise} \end{cases},$$

a diagonal matrix whose diagonal elements are the diagonal elements of A. Recall that for a diagonal matrix, the matrix inverse, the matrix square root, etc. become simply elementwise operations on the diagonal and its determinant is a product of its diagonal values.

Then, let $\boldsymbol{V} = \sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top = n\hat{\boldsymbol{\Sigma}}$ as before. If Σ is diagonal, then elements of $\boldsymbol{X}$

are independent, so if $\sigma_j^2 = \operatorname{Var} \boldsymbol{X}_j$, $\hat{\sigma}_j^2 = n^{-1}\sum_{i=1}^n (x_{ji} - \bar{x}_j)^2 = n^{-1} V_{jj} = \hat{\boldsymbol{\Sigma}}_{jj}$. Then,

$$\begin{aligned}
\Lambda &= \frac{\prod_{j=1}^p (\hat{\boldsymbol{\Sigma}}_{jj})^{-\frac{n}{2}} \mathrm{e}^{-\frac{1}{2\hat{\boldsymbol{\Sigma}}_{jj}} \sum_{i=1}^n (x_{ji}-\bar{x}_j)^2}}{|\boldsymbol{V}|^{-\frac{n}{2}} n^{\frac{np}{2}} \mathrm{e}^{-\frac{np}{2}}} \\
&= \frac{\prod_{j=1}^p (\hat{\boldsymbol{\Sigma}}_{jj})^{-\frac{n}{2}} \cancel{\mathrm{e}^{-\frac{np}{2}}}}{|\hat{\boldsymbol{\Sigma}}|^{-\frac{n}{2}} \cancel{\mathrm{e}^{-\frac{np}{2}}}} \\
&= \frac{\prod_{j=1}^p (\hat{\boldsymbol{\Sigma}}_{jj})^{-\frac{n}{2}} (|\operatorname{diag}(\hat{\boldsymbol{\Sigma}})^{-1/2}|^{-\frac{n}{2}})^2}{|\operatorname{diag}(\hat{\boldsymbol{\Sigma}})^{-1/2}|^{-\frac{n}{2}} |\hat{\boldsymbol{\Sigma}}|^{-\frac{n}{2}} |\operatorname{diag}(\hat{\boldsymbol{\Sigma}})^{-1/2}|^{-\frac{n}{2}}} \\
&= \frac{\cancel{\{\prod_{j=1}^p (\hat{\boldsymbol{\Sigma}}_{jj})\}^{-\frac{n}{2}}} \cancel{\{(\prod_{j=1}^p (\hat{\boldsymbol{\Sigma}}_{jj})^{-1/2})^2\}^{-\frac{n}{2}}}}{|\operatorname{diag}(\hat{\boldsymbol{\Sigma}})^{-1/2} \hat{\boldsymbol{\Sigma}} \operatorname{diag}(\hat{\boldsymbol{\Sigma}})^{-1/2}|^{-\frac{n}{2}}} \\
&= |\operatorname{diag}(\hat{\boldsymbol{\Sigma}})^{-1/2} \hat{\boldsymbol{\Sigma}} \operatorname{diag}(\hat{\boldsymbol{\Sigma}})^{-1/2}|^{\frac{n}{2}} = |\boldsymbol{R}|^{\frac{n}{2}}.
\end{aligned}$$

so

$$-2 \log \Lambda = -n \log |\boldsymbol{R}|.$$

Lastly, the degrees of freedom for the χ^2 distribution are

$$\overbrace{\frac{p(p+1)}{2}}^{\text{\# param. SPD matrix}} - \overbrace{p}^{\text{\# param. diag. matrix}} = \frac{p(p-1)}{2}.$$

12.1

(a)

Normal populations with equal variances implies LDA, so use the the expression from Section 12.6, with $\boldsymbol{\mu}_i$ replacing $\bar{\boldsymbol{x}}_i$ and Σ replacing $\boldsymbol{S}_{\text{pooled}}$, since those are given to us rather than estimated from the sample. This leads to the following rule:

1. Evaluate
$$d_i(\boldsymbol{x}) = \boldsymbol{\mu}_i^\top \Sigma^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_i^\top \Sigma^{-1}\boldsymbol{\mu}_i + \log\frac{1}{3}$$
for $i = 1, 2, 3$.
2. Classify $\boldsymbol{x}$ into the category with the highest $d_i(\boldsymbol{x})$.

(b)

We shall illustrate the first case in detail, and only the results for the remainder. Let $\boldsymbol{x} = \begin{pmatrix}0.2\\0.6\end{pmatrix} = \begin{pmatrix}1/5\\3/5\end{pmatrix}$, and evaluate $\Sigma^{-1} = \begin{pmatrix}4/3 & -2/3\\-2/3 & 4/3\end{pmatrix}$. Then,

$$d_1(\boldsymbol{x}) = \begin{pmatrix}1\\1\end{pmatrix}^\top \begin{pmatrix}4/3 & -2/3\\-2/3 & 4/3\end{pmatrix}\begin{pmatrix}1/5\\3/5\end{pmatrix} - \frac{1}{2}\begin{pmatrix}1\\1\end{pmatrix}^\top \begin{pmatrix}4/3 & -2/3\\-2/3 & 4/3\end{pmatrix}\begin{pmatrix}1\\1\end{pmatrix} + \log\frac{1}{3} = -\frac{2}{15} + \log\frac{1}{3}$$
$$d_2(\boldsymbol{x}) = \begin{pmatrix}1\\0\end{pmatrix}^\top \begin{pmatrix}4/3 & -2/3\\-2/3 & 4/3\end{pmatrix}\begin{pmatrix}1/5\\3/5\end{pmatrix} - \frac{1}{2}\begin{pmatrix}1\\0\end{pmatrix}^\top \begin{pmatrix}4/3 & -2/3\\-2/3 & 4/3\end{pmatrix}\begin{pmatrix}1\\0\end{pmatrix} + \log\frac{1}{3} = -\frac{4}{5} + \log\frac{1}{3}$$
$$d_3(\boldsymbol{x}) = \begin{pmatrix}0\\1\end{pmatrix}^\top \begin{pmatrix}4/3 & -2/3\\-2/3 & 4/3\end{pmatrix}\begin{pmatrix}1/5\\3/5\end{pmatrix} - \frac{1}{2}\begin{pmatrix}0\\1\end{pmatrix}^\top \begin{pmatrix}4/3 & -2/3\\-2/3 & 4/3\end{pmatrix}\begin{pmatrix}0\\1\end{pmatrix} + \log\frac{1}{3} = 0 + \log\frac{1}{3}$$

Thus, we classify to Category 3.

For $\boldsymbol{x} = \begin{pmatrix}2\\0.8\end{pmatrix}$, $d_1(\boldsymbol{x}) = \frac{6}{5} + \log\frac{1}{3}$, $d_2(\boldsymbol{x}) = \frac{22}{15} + \log\frac{1}{3}$, $d_3(\boldsymbol{x}) = -\frac{14}{15} + \log\frac{1}{3}$. Classify into Category 2.

For $\boldsymbol{x} = \begin{pmatrix}0.75\\1\end{pmatrix}$, $d_1(\boldsymbol{x}) = \frac{1}{2} + \log\frac{1}{3}$, $d_2(\boldsymbol{x}) = -\frac{1}{3} + \log\frac{1}{3}$, $d_3(\boldsymbol{x}) = \frac{1}{6} + \log\frac{1}{3}$. Classify into Category 1.

(c)

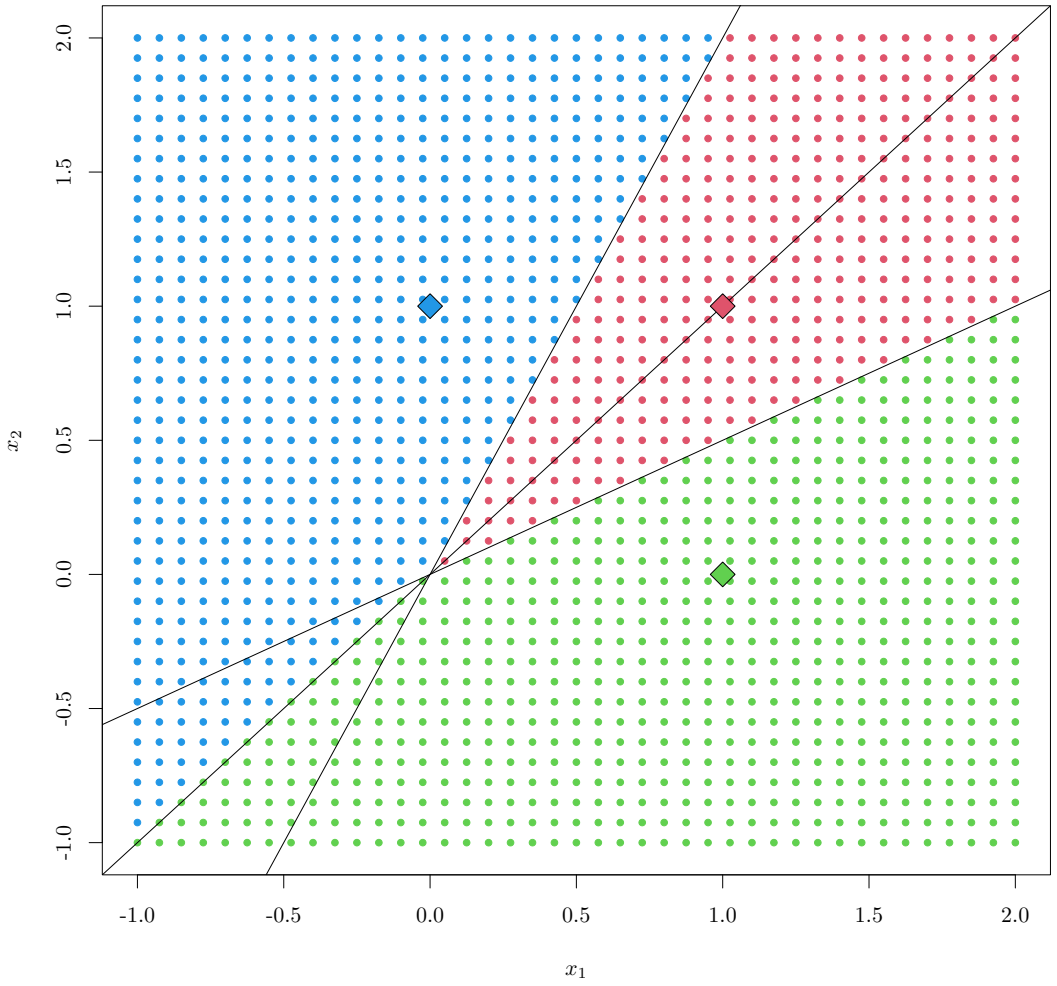
To be at the boundary between two regions, say, i and j, the point $\boldsymbol{x}$ must have $d_i(\boldsymbol{x}) = d_j(\boldsymbol{x})$. Then,

$$\boldsymbol{\mu}_i^\top \Sigma^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_i^\top \Sigma^{-1}\boldsymbol{\mu}_i + \log\pi_i = \boldsymbol{\mu}_j^\top \Sigma^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_j^\top \Sigma^{-1}\boldsymbol{\mu}_j + \log\pi_j$$
$$\boldsymbol{\mu}_i^\top \Sigma^{-1}\boldsymbol{x} - \boldsymbol{\mu}_j^\top \Sigma^{-1}\boldsymbol{x} = -\frac{1}{2}\boldsymbol{\mu}_j^\top \Sigma^{-1}\boldsymbol{\mu}_j + \frac{1}{2}\boldsymbol{\mu}_i^\top \Sigma^{-1}\boldsymbol{\mu}_i + \log\pi_j - \log\pi_i$$
$$(\boldsymbol{\mu}_i^\top \Sigma^{-1} - \boldsymbol{\mu}_j^\top \Sigma^{-1})\boldsymbol{x} = -\frac{1}{2}\boldsymbol{\mu}_j^\top \Sigma^{-1}\boldsymbol{\mu}_j + \frac{1}{2}\boldsymbol{\mu}_i^\top \Sigma^{-1}\boldsymbol{\mu}_i + \log\pi_j - \log\pi_i.$$

If we call $\boldsymbol{a} = (\boldsymbol{\mu}_i^\top \Sigma^{-1} - \boldsymbol{\mu}_j^\top \Sigma^{-1})^\top \in \mathbb{R}^2$ and $c = -\frac{1}{2}\boldsymbol{\mu}_j^\top \Sigma^{-1}\boldsymbol{\mu}_j + \frac{1}{2}\boldsymbol{\mu}_i^\top \Sigma^{-1}\boldsymbol{\mu}_i + \log\pi_j - \log\pi_i \in \mathbb{R}$, neither of them depending on $\boldsymbol{x}$, we can write

$$\boldsymbol{a}^\top \boldsymbol{x} = c \implies a_1x_1 + a_2x_2 = c \implies x_2 = \frac{c}{a_2} - \frac{a_1}{a_2}x_1,$$

an equation for a line with slope $-a_1/a_2$ and y-intercept c/a_2. Here's a sketch of region boundaries:



15.1

First, we solve for the inverse of the generator:

$$\phi^{-1}(x) = (\theta x + 1)^{-1/\theta}.$$

Then, substitute into the Archimedean form:

$$\begin{aligned} C_\theta(u_1, u_2, \ldots, u_p) &= \phi^{-1}\{\sum_{i=1}^{p} \phi(u_i)\} \\ &= \left[\theta\{\sum_{i=1}^{p} \theta^{-1}(u_i^{-\theta} - 1)\} + 1\right]^{-1/\theta} \\ &= \left[\theta\theta^{-1}\{(\sum_{i=1}^{p} u_i^{-\theta}) - p\} + 1\right]^{-1/\theta} \\ &= \left[\sum_{i=1}^{p} u_i^{-\theta} - p + 1\right]^{-1/\theta}. \end{aligned}$$