Multivariate Analysis (MATH5855)

Dr. Atefeh Zamani

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

Term 3, 2022

Introduction

Dr. Atefeh Zamani

Education:

- Ph.D. in Mathematical Statistics
- Master of Data Science

Research interest:

- Functional data analysis
- Time series Analysis

Contact :

- Email: atefeh.zamani@unsw.edu.au
  Please make sure to mention "MATH5855" in the subject of the email.
- Consultation:

  Monday 14-15
  Friday 13-14

Why Multivariate Analysis??

- complex phenomena → simultaneous measurements on many variables
- full set of variables often not known a prior

Multivariate analysis: methods for understanding relationships among many variables

- more mathematics
- algebraic concepts

What makes Multivariate Analysis Different?

Multiple linear regression models:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, 2, \cdots, n.$$

Here,

Y_i : the i-th observation of the response variable

x_{ik} : the i-th observation of the k-th predictor variable

ε_i : the i-th error.

- p predictors fixed (conditioned on)
- one random variable per observation
- ε_is and therefore Y_is assumed to be independent (given xs) or at least uncorrelated

Matrix Notation

$$\boldsymbol{Y}_{n\times 1} = \mathrm{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$$

Cont. What makes Multivariate Analysis Different?

Multivariate linear regression model:

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

(Y_{i1}, Y_{i2}) : the i-th observations of two distinct response variables

- ε_{i1} and ε_{i2} may be correlated
- E.g., when multiple observations are taken on each individual.
- more than one response variable and and p predictors

Matrix Notation

$$\mathrm{Y}_{n\times 2} = \mathrm{X}_{n\times p}\boldsymbol{\beta}_{p\times 2} + \boldsymbol{\varepsilon}_{n\times 2}$$

Cont. What makes Multivariate Analysis Different?

Multivariate linear regression model:

Example

Suppose it is desired to predict the response variables

Y_1= height

Y_2= height at shoulder

of a person from partial skeletal remains, such as

x_1=femur length

x_2=ulna length

x_3=gender.

Difficulties

- Size and complexity of the data sets
- More mathematics necessary
- Computer intensive methods involved

Objectives of MV methods

Data reduction or structural simplification. representing the phenomenon as simply as possible without sacrificing valuable information to make interpretation easier.

Sorting and grouping. creating groups of “similar” objects or variables that in a sense are more closer to each other than to objects outside the group with reasonable explanation for the existing grouping.

Investigation of the dependence finding more about the nature of the relationships among variables: Are they mutually independent or are one or more variables dependent on the others? If so, how?

Cont. Objectives of MV methods

Prediction. Determining the the relationships between variables for predicting the values of one or more variables on the basis of observations on the other variables.

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).

Structure and Resources: Topics

Week 1 Background (L0); Exploratory Data Analysis (L1); The Multivariate Normal Distribution (L2–L3)

Week 2 Estimation and testing for the Multivariate Normal Mean (L3–L4)

Week 3 Catch-up and revision; Correlations, Partial Correlations, and Multiple Correlations (L5)

Week 4 Principal Component Analysis (L6); Canonical Correlation Analysis (L7)

Week 5 Multivariate linear models and MANOVA (L8); Catch-up and revision; Tests of the Covariance Matrix (L9)

Week 7 Factor Analysis (L10); Structural Equation Models (L11)

Week 8 Classification: Discriminant Analysis (L12) and Support Vector Machines (L13)

Week 9 Cluster Analysis (L14); Copulae (L15)

Week 10 Catch-up; Revision; Special topics (to be decided)

Structure and Resources: Lectures

3 Hours/Week

Mon 12:00 - 14:00 (Weeks:1-3,5,7-10)

Fri 12:00 - 13:00 (Weeks:1-5,7-10)

- Introduce topics; prove the major results.
- Demonstrate examples in R.
- Answer questions as they arise. (Please ask!)
- Revise topics and provide clarification.
- Provide general feedback on assessments.

Structure and Resources: Moodle Site

- Post lecture slides in advance of the lecture.
- Post Exercises (as a part of lecture notes). (Not assessed.)
- Solutions posted as needed.
- Post R examples covered in lecture (and others), in advance of the lecture.
- Answer questions on the discussion forum.
- Discuss Exercises and Challenges problems on the discussion forum.

Structure and Resources: Tutorial

1 Hours/Week

- Work through Exercises, Challenge problems, and past assignment questions.
  - ⋆ Selected and prioritised based on your requests.
  - ⋆ A form for requesting topics to be covered will be provided.

Structure and Resources: Texts

- No required text.
- Useful alternative presentation in:
  Johnson, R. & Wichern, D. (2007) Applied Multivariate Statistical Analysis. Sixth Edition, Prentice Hall.
- Abbreviated as “JW”.
- Some examples drawn from there as well.

Expectations: Theory

- This is a theory-heavy course.
- Lectures will feature a lot of deep theory and proofs.
- Assessment will be on deep theory and proofs and practical questions.
- Understanding of theoretical concepts needed for application and straightforward derivations.

Expectations: Software

- This course is offered in R.
- Software demonstrations and (most) labs will be provided.

Expectations: Assessments

- 3 Assignments (Week 4, Week 7, Week 10).
- Equal weights: 15% each.
- Final exam, during the exam period: Weighted 55%.
- A mix of Moodle quizzes, handwritten or typeset derivations uploaded to Moodle assignments, and typed reports for Turnitin.
- Some assessments might have multiple modes.
- Read instructions carefully!

Section 0: Preliminaries

Matrix algebra, Standard facts about multivariate distributions,

Additional resources and Exercises

Vectors and Matrices

Matrix operations

$\mathcal{M}_{n,p}$: vector space of matrices with n rows and p columns

$$X \in \mathcal{M}_{n,p} \Rightarrow X \text{ is a } n \times p \text{ matrix.}$$

$\mathbb{R}^n := \mathcal{M}_{n,1}$: space of n-dimensional column vectors

$$\boldsymbol{x} \in \mathbb{R}^n$$

^{T} : Transposition: $X \in \mathcal{M}_{n,p} \Rightarrow X^T \in \mathcal{M}_{p,n}$

- $\boldsymbol{x} \in \mathbb{R}^n$ a column vector → $\boldsymbol{x}^T$ is a row vector in $\mathcal{M}_{1,n}$

Scalar operations:

- matrix (vector) multiplied by scalar
- two matrices (vectors) of the same dimension can be added or subtracted

Euclidean norm $\|\boldsymbol{x}\|$ of a vector $\boldsymbol{x} = (x_1\ x_2\ \cdots\ x_n)^T \in \mathbb{R}^n$:

$$\|\boldsymbol{x}\| = \sqrt{\sum_{i=1}^{n} x_i^2}$$

Inner Product

inner product (a.k.a scalar product) of $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ is denoted and defined in the following way:

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^T \boldsymbol{y} = \sum_{i=1}^{n} x_i y_i$$

- $\|\boldsymbol{x}\|^2 = \langle \boldsymbol{x}, \boldsymbol{x} \rangle$
- If θ is the angle between $\boldsymbol{x}$ and $\boldsymbol{y}$,

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \|\boldsymbol{x}\| \|\boldsymbol{y}\| \cos \theta$$

Cauchy-Bunyakovsky-Schwartz Inequality:

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

orthogonal projection of $\boldsymbol{x}$ on $\boldsymbol{y}$:

$$\frac{\boldsymbol{x}^T \boldsymbol{y}}{\boldsymbol{y}^T \boldsymbol{y}} \boldsymbol{y}$$

Matrix multiplication

matrix multiplication: if $X \in \mathcal{M}_{n,k}$ and $Y \in \mathcal{M}_{k,p}$ (# columns of $X=$ # rows in Y , a.k.a. **conformable**) then XY exists and $Z = XY \in \mathcal{M}_{n,p}$ with elements

$$z_{ij} = \sum_{m=1}^{k} x_{im} y_{mj}, \; i = 1, 2, \cdots, n, \; j = 1, 2, \cdots, p.$$

- 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
- not commutative, i.e., XY and YX are not necessarily equal, even if both exist.
  - Important example: if $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{x}^T\boldsymbol{x} \in \mathbb{R}$ but $\boldsymbol{x}\boldsymbol{x}^T \in \mathcal{M}_{n,n}$
- $(XY)^T = Y^T X^T$

Symmetric and identity matrices

symmetric matrix: a **square** matrix $X \in \mathcal{M}_{p,p}$ for which $x_{i,j} = x_{j,i}$, $i = 1, 2, \cdots, p$, $j = 1, 2, \cdots, p$ holds, i.e.,

$$X^T = X$$

identity matrix: $\boldsymbol{I}_{p \times p}$ is the identity matrix if it has ones on the diagonal and zeros outside the diagonal,

$$\boldsymbol{I} = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}$$

if conformable, $X\boldsymbol{I} = X$ and $\boldsymbol{I}X = X$

Matrix trace

The trace of a square matrix $X \in \mathcal{M}_{p,p}$ is denoted by

$$tr(X) = \sum_{i=1}^{p} x_{ii}$$

- Properties of the trace
  - $tr(X + Y) = tr(X) + tr(Y)$
  - $tr(XY) = tr(YX)$
  - $tr(X^{-1}YX) = tr(Y)$
  - if $\boldsymbol{a} \in \mathbb{R}^p$ and $X \in \mathcal{M}_{p,p}$ then $\boldsymbol{a}^T X \boldsymbol{a} = tr(X\boldsymbol{a}\boldsymbol{a}^T)$

inverse Matrices

Matrix determinant

For a square matrix $X \in \mathcal{M}_{p,p}$, a number $|X| \equiv det(X)$ is defined as

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

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

Matrix determinant calculations

- When $p = 1$ (scalar) and $X = a$, then $|X| = a$.
- 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

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

$$- x_{31}x_{22}x_{13} - x_{11}x_{23}x_{32} - x_{12}x_{21}x_{33}$$

Matrix determinant calculations: General rule

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

$$\begin{aligned} |X| &= \sum_i (-1)^{i+j} x_{ij} |X_{ij}| \qquad \text{(for any given } j\text{)} \\ &= \sum_i (-1)^{i+j} x_{ij} |X_{ij}| \qquad \text{(for any given } i\text{)} \end{aligned}$$

X_{ij} matrix constructed by deleting i-th row and j-th column of X.

Matrix determinant: Properties

- If one row or one column of the matrix contains zeros only, then the value of the determinant is zero. (Why??)
- 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.
- $|cX| = c^p|X|$
- If $X, Y \in \mathcal{M}_{p,p}$, then $|XY| = |X||Y|$
- If the matrix X is diagonal (i.e. all non-diagonal elements are zero) then

$$|X| = \prod_{i=1}^{p} x_{ii}$$

- the determinant of the identity matrix is always equal to one, i.e., $|\boldsymbol{I}| = 1$.

Matrix inverse

If

- $|X| \neq 0$, i.e., $X \in \mathcal{M}_{p,p}$ is **non-singular**, then,
  - An inverse matrix $X^{-1} \in \mathcal{M}_{p,p}$ exists s.t. $XX^{-1} = \boldsymbol{I}_{p,p}$.
  - if $|X_{ij}|$ stands for (i,j)-th minor of X, then

$$(X^{-1})_{ji} = \frac{|Xij|}{|X|}(-1)^{i+j}.$$

Matrix inverse: properties

- $XX^{-1} = X^{-1}X = \boldsymbol{I}$
- $(X^{-1})^T = (X^T)^{-1}$
- $(XY)^{-1} = Y^{-1}X^{-1}$ when both X and Y are non-singular square matrices of the same dimension.
- $|X^{-1}| = |X|^{-1}$
- If X is diagonal and non-singular then all its diagonal elements are nonzero and X^{-1} is again diagonal with diagonal elements equal to $\frac{1}{x_{ii}}$, $i = 1, 2, \cdots, p$.

Rank

Linear independence

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

$$a_1 \boldsymbol{x}_1 + a_2 \boldsymbol{x}_2 + \cdots + a_k \boldsymbol{x}_k = 0 \tag{1}$$

holds.

- Otherwise the vectors are **linearly independent**.
- For k linearly independent vectors, (1) would only be possible if **all** numbers $a_1, a_2, \cdots, a_k$ were zero.

Matrix rank

row rank: the maximum number of linearly independent row vectors.
column rank: the rank of its set of column vectors

- Always row rank=column rank
- denoted as $rk(X)$

full rank: If $X \in \mathcal{M}_{n,p}$ and $rk(X) = \min(n, p)$

- square matrix $X \in \mathcal{M}_{p,p}$ is full rank if $rk(X) = p$
- being full rank implies that $|A| \neq 0$ (Rouche–Capelli theorem a.k.a Kronecker–Capelli theorem)
    - 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 $\boldsymbol{x} = A^{-1}\boldsymbol{b} \in \mathbb{R}^p$.

Orthogonal matrices

Orthogonal matrix

A square matrix $X \in \mathcal{M}_{p,p}$ is orthogonal if $XX^T = X^T X = \boldsymbol{I}_{p,p}$ holds.

Properties of orthogonal matrices:

- X is of full rank ($rk(X) = p$) and $X^{-1} = X^T$
- 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.
- $|X| = \pm 1$ (why??)

Eigenvalues and eigenvectors

Eigenvalues

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

$$f(\lambda) = |X - \lambda \boldsymbol{I}| = 0 \tag{2}$$

- Has exactly p roots.
- Some may be complex and some may coincide.
- Since the coefficients are real, if there is a complex root of (2) then also its complex conjugate must be a root of the same equation.
- Each of the above p roots is called eigenvalue of the matrix X.
- $tr(X) = \sum_{i=1}^{p} \lambda_i$
- $|X| = \prod_{i=1}^{p} \lambda_i$

Eigenvectors

- For any such eigenvalue λ^*, $X - \lambda^* \boldsymbol{I}$ is singular.
- There exists a non-zero vector $\boldsymbol{y} \in \mathbb{R}^p$ s.t. $(X - \lambda^* \boldsymbol{I})\boldsymbol{y} = 0$
    - An eigenvector of X that corresponds to the eigenvalue λ^*
    - Not unique: $\mu \boldsymbol{y}$ for any real non-zero μ also an eigenvector for the same eigenvalue.

Uniqueness of eigenvectors

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

Theorem

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.*

Spectral decomposition

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

$$X\boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i, \qquad i = 1, 2, \cdots, p, \qquad \|\boldsymbol{e}_i\| = 1, \qquad \boldsymbol{e}_i^T \boldsymbol{e}_i = 0, \qquad i \neq j.$$

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

$$X = \lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1^T + \lambda_2 \boldsymbol{e}_2 \boldsymbol{e}_2^T + \cdots + \lambda_p \boldsymbol{e}_p \boldsymbol{e}_p 1^T \tag{3}$$

Equivalently, $X = P \Lambda P^T$ where $\Lambda = diag(\lambda_, \lambda_2, \cdots, \lambda_p)$ and $P\ in \mathcal{M}_{p,p}$is an orthogonal matrix containing the p orthogonal eigenvectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_p$.

Powers of a matrix: inverse

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, $i = 1, 2, \cdots, p$, to be positive in the spectral decomposition (3). But then

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

Powers of a matrix: square root

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

$$X^{1/2} = \sum_{i=1}^{p} \sqrt{\lambda_i} \boldsymbol{e}_i \boldsymbol{e}_i^T \tag{4}$$

- makes sense since $X^{1/2} X^{1/2} = X$
- $X^{1/2}$ is also symmetric and non-negative definite
- $X^{-1/2} = \sum_{i=1}^{p} \lambda_i{}^{-1/2} \boldsymbol{e}_i \boldsymbol{e}_i^T = P \Lambda^{-1/2} P^T$

Example

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

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

Solution

Let $X = P\Lambda P^T$ be the decomposition (3) for X. Denote $\boldsymbol{z} = P^T\boldsymbol{y}$. Note that $\boldsymbol{y} \neq 0$ implies $\boldsymbol{z} \neq 0$. Thus

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

If we take $\boldsymbol{y} = \boldsymbol{e}_1$ then having in mind the structure of the matrix P we have $\boldsymbol{z} = P^T\boldsymbol{e}_1 = (1\ 0\ \cdots\ 0)^T$ and for this choice of $\boldsymbol{y}$ also $\frac{\boldsymbol{z}^T\Lambda\boldsymbol{z}}{\boldsymbol{y}^T\boldsymbol{y}} = \lambda_1$. 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 $\max_{\boldsymbol{y}\neq 0,\, \boldsymbol{y}\perp\boldsymbol{e}_1} \frac{\boldsymbol{y}^T X\boldsymbol{y}}{\boldsymbol{y}^T\boldsymbol{y}} = \lambda_1$ attained when $\boldsymbol{y} = \boldsymbol{e}_2$. holds. How?

Cholesky Decomposition

Numerical stability

- Computers have finite precision
  - around 16 decimal significant figures
  - scientific notation ⇒ absolute magnitude has little effect on precision, different magnitudes produce rounding
  - E.g., $1 \times 10^{18} + 1 \times 10^{0} = 1,000,000,000,000,000,000 + 1 = 1,000,000,000,000,000,000$
- For matrices, condition number $|\lambda_1/\lambda_p|$ (for pos. def. matrix) to assess the potential error ⇒ higher = worse

Cholesky decomposition

For a symm. pos.def. matrix $X \in \mathcal{M}_{p,p}$, a unique matrix $U \in \mathcal{M}_{p,p}$ exists that is:

- upper triangular
- $U^T U = X$
  - Many authors use $LL^T = X$ for a lower-triangular matrix instead.
  - $L \equiv U^T$
- In SAS/IML, root(x) gives this.
- In R, the function is chol().

Useful for generating correlated variables.

Orthogonal Projection

Orthogonal projection matrix: necessary conditions

- Let $\mathcal{L}(X)$ be space spanned by the columns of the matrix $X \in \mathcal{M}_{p,p}$.
- Project a vector $\boldsymbol{y} \in \mathbb{R}^n$ to it with matrix $P \in \mathcal{M}_{n,n}$ (an orthogonal **projector**):
  - Let $\boldsymbol{z} = P\boldsymbol{y} \in \mathbb{R}^n$ be the projection.
  - $\boldsymbol{z} \in \mathcal{L}(X) \Rightarrow$ projection of $\boldsymbol{z}$ on $\mathcal{L}(X)$ is $\boldsymbol{z}$ itself:

$$\begin{aligned} P\boldsymbol{y} = \boldsymbol{z} = P\boldsymbol{z} = PP\boldsymbol{y} = P^2\boldsymbol{y} \\ \Rightarrow (P - P^2)\boldsymbol{y} = 0 \\ \Rightarrow P^2 = P \\ \Rightarrow P \text{ should be idempotent.} \end{aligned}$$

Cont. Orthogonal projection matrix: necessary conditions

Besides,

$$
\begin{aligned}
&\forall \boldsymbol{y} : \ (\boldsymbol{y} - \boldsymbol{z})^T \boldsymbol{z} = 0 \\
&\Rightarrow \forall \boldsymbol{y} : \ \boldsymbol{y}^T (P^T - \boldsymbol{I}) P \boldsymbol{y} = 0 \\
&\Rightarrow (P^T - \boldsymbol{I}) P = 0 \\
&\Rightarrow P^T P = P \\
&\Rightarrow P^T P = P^T \\
&\Rightarrow P = P^T \Rightarrow P \text{ is symmetrical.}
\end{aligned}
$$

Therefore, the orthogonal projector is a **symmetric** and **idempotent** matrix.

Cont. Orthogonal projection matrix: necessary conditions

- Let P be symmetric and idempotent. Therefore, For any $\boldsymbol{y} \in \mathbb{R}^n$, we have:

  $$\begin{aligned} \boldsymbol{z} &= P\boldsymbol{y} \\ &\Rightarrow P\boldsymbol{z} = P^2\boldsymbol{z} = P\boldsymbol{y} \\ &\Rightarrow P(\boldsymbol{y} - \boldsymbol{z}) = 0 \end{aligned}$$

  Also $P^T(\boldsymbol{y} - \boldsymbol{z}) = 0$ since $P = P^T$.
- Consider $\mathcal{L}(P)$ (the space generated by the rows/columns of P).
  - $\boldsymbol{z} = P\boldsymbol{y} \Rightarrow \boldsymbol{z} \in \mathcal{L}(P)$
  - $P^T(\boldsymbol{y} - \boldsymbol{z}) = 0$ means that $\boldsymbol{y} - \boldsymbol{z}$ is perpendicular to $\mathcal{L}(P)$.
    $\Rightarrow P\boldsymbol{y}$ is the projection of $\boldsymbol{y}$ on $\mathcal{L}(P)$.
- $P \in \mathcal{M}_{n,n}$ is an orthogonal projection matrix if and only if it is a symmetric and idempotent matrix./

Orthogonal projection matrix: properties

- 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 $\mathcal{M}$.
    - $rk(P) = dim(\mathcal{M})$.
- The rank of an orthogonal projector is equal to the sum of its diagonal elements.

Orthogonal projection matrix: form

- If the matrix X has a full rank then the projector

$$P_{\mathcal{L}(X)} = X(X^T X)^{-1} X^T$$

- If the matrix X is not of full rank then the generalised inverse $(X^T X)^-$ of $X^T X$ can be defined instead.
  - Not unique
  - But $X(X^T X)^- X^T$ is unique
  - Is the orthogonal projector on the space $\mathcal{L}(X)$ spanned by the columns of X when X is not full rank.

Random samples in multivariate analysis

Random dataset

- Inference depends on variability of statistics.
- Some assumptions are required about the data matrix.
- n observations of p-variate random vectors
  - random matrix $X \in \mathcal{M}_{p,n}$:

$$X = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1j} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2j} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{i1} & X_{i2} & \cdots & X_{ij} & \cdots & X_{in} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{p1} & X_{p2} & \cdots & X_{pj} & \cdots & X_{p} \end{pmatrix} = (\boldsymbol{X}_1 \; \boldsymbol{X}_2 \; \cdots \; \boldsymbol{X}_n)$$

  - $\boldsymbol{X}_i$, $i = 1, 2, \cdots, n$ assumed to be independent.

Joint, marginal, conditional distributions

Random vector cdf, pmf, and/or density

- Random vector $\boldsymbol{X} = (X_1 \ \cdots \ X_p)^T \in \mathbb{R}^p, \ p \geq 2$ has joint Cumulative distribution function (cdf)

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = P(X_1 \leq x_1, \cdots, X_p \leq x_p) = F_{\boldsymbol{X}}(x_1, \cdots, x_p)$$

- If discrete the probability mass function

$$P_{\boldsymbol{X}}(\boldsymbol{x}) = P(X_1 = x_1, \ \cdots, \ X_p = x_p)$$

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

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_p} f_{\boldsymbol{X}}(\boldsymbol{t}) dt_1 \cdots dt_2$$

then $\boldsymbol{X}$ is continuous and $f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{\partial^p F_{\boldsymbol{X}}(\boldsymbol{x})}{\partial x_1 \cdots \partial x_p}$

Marginal distribution

- marginal cdf of the first $k < p$ components of the vector $\boldsymbol{X}$ is

$$\begin{aligned} & P(X_1 \leq x_1, X_2 \leq x_2, \cdots, X_k \leq x_k) = \\ & \qquad = P(X_1 \leq x_1, X_2 \leq x_2, \cdots, X_k \leq x_k, X_{k+1} \leq \infty, \cdots, X_p \leq \infty) \\ & \qquad = F_{\boldsymbol{X}}(x_1, x_2, \cdots, x_k,\ \infty,\ \cdots,\ \infty) \qquad (5) \end{aligned}$$

- marginal density can be obtained by partial differentiation in (5)

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

- Similarly for any other set of components.
- Each component X_i has marginal cdf $F_{X_i}(x_i)$, $i = 1, 2, \cdots, p$.

Conditional distribution

- conditional density of $\boldsymbol{X}$ given $X_{r+1} = x_{r+1}, \cdots, X_p = x_p$ is given by

$$\begin{aligned} & f_{(X_1, \cdots, X_r | X_{r+1}, \cdots, X_p)}(x_1, \cdots, x_r | x_{r+1}, \cdots, x_p) \\ & \qquad = \frac{f_{\boldsymbol{X}}(\boldsymbol{x})}{f_{(X_{r+1}, \cdots, X_p)}(x_{r+1}, \cdots, x_p)} \end{aligned} \tag{6}$$

- joint density of $X_1, \cdots, X_r$ when $X_{r+1} = x_{r+1}, \cdots, X_p = x_p$ is only defined when $f_{(X_{r+1}, \cdots, X_p)}(x_{r+1}, \cdots, x_p) \neq 0$

Independence

- If $\boldsymbol{X}$ has p independent components, then

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = F_{X_1}(x_1)F_{X_2}(x_2)\cdots F_{X_p}(x_p) \tag{7}$$

- Equivalently,

$$P_{\boldsymbol{X}}(\boldsymbol{x}) = P_{X_1}(x_1)P_{X_2}(x_2)\cdots P_{X_p}(x_p) \tag{8}$$

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{X_1}(x_1)f_{X_2}(x_2)\cdots f_{X_p}(x_p) \tag{9}$$

- Conditional distributions do **not** depend on the conditions
- Functions factorise:

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{p} F_{X_i}(x_i), \; f_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{p} f_{X_i}(x_i) \tag{10}$$

Moments

Moments

- For density $f_{\boldsymbol{X}}(\boldsymbol{x})$ joint moments of order $s_1, s_2, \cdots, s_p$ are

$$E(X_1^{s_1} \cdots X_p^{s_p}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_1^{s_1} \cdots x_p^{s_p} f_{\boldsymbol{X}}(x_1, \cdots, x_p) dx_1 \cdots dx_p \tag{11}$$

- if some $s_i = 0$, we are calculating the joint moment of a subset of the random variables

Common multivariate moments

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

Expectation:

$$\boldsymbol{\mu}_{\boldsymbol{X}} = E(\boldsymbol{X}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_1 \cdots x_p f_{\boldsymbol{X}}(x_1, \cdots, x_p) dx_1 \cdots dx_p \in \mathbb{R}^p \tag{12}$$

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

$$\begin{aligned} \Sigma_{\boldsymbol{X}} &= Var(\boldsymbol{X}) = Cov(\boldsymbol{X}, \boldsymbol{X}) = E(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}})(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}})^T \\ &= E(\boldsymbol{X}\boldsymbol{X}^T) - \boldsymbol{\mu}_{\boldsymbol{X}}\boldsymbol{\mu}_{\boldsymbol{X}}^T = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix} \in \mathcal{M}_{p,p} \end{aligned} \tag{13}$$

Cont. Common multivariate moments

Covariance matrix:

$$\Sigma_{\boldsymbol{X},\boldsymbol{Y}} = Cov(\boldsymbol{X}, \boldsymbol{Y}) = E(\boldsymbol{X} - \boldsymbol{\mu_X})(\boldsymbol{Y} - \boldsymbol{\mu_Y})^T = E(\boldsymbol{X}\boldsymbol{Y}^T) - \boldsymbol{\mu_X}\boldsymbol{\mu_Y}^T$$

$$= \begin{pmatrix} \sigma_{X_1Y_1} & \sigma_{X_1Y_2} & \cdots & \sigma_{X_1Y_p} \\ \sigma_{X_2Y_1} & \sigma_{X_2Y_2} & \cdots & \sigma_{X_2Y_p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_pY_1} & \sigma_{X_pY_2} & \cdots & \sigma_{X_pY_q} \end{pmatrix} \in \mathcal{M}_{p,q} \qquad (14)$$

Linear transformations of moments

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

- $\boldsymbol{\mu}_{A\boldsymbol{X}} = A\boldsymbol{\mu}_{\boldsymbol{X}} \in \mathbb{R}^{p'}$
- $\Sigma_{A\boldsymbol{X}} = A\Sigma_{\boldsymbol{X}}A^T \in \mathcal{M}_{p',p'}$
- $\Sigma_{A\boldsymbol{X},B\boldsymbol{Y}} = A\Sigma_{\boldsymbol{X},\boldsymbol{Y}}B^T \in \mathcal{M}_{p',q'}$

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

- $E(A\boldsymbol{X} + A'\boldsymbol{X}') = A\boldsymbol{\mu}_{\boldsymbol{X}} + A'\boldsymbol{\mu}_{\boldsymbol{X}'}$
- $Var(A\boldsymbol{X} + A'\boldsymbol{X}') =$
  $A\Sigma_{\boldsymbol{X}}A^T + A\Sigma_{\boldsymbol{X},\boldsymbol{X}'}(A')^T + A'\Sigma_{\boldsymbol{X}',\boldsymbol{X}}A^T + A'\Sigma_{\boldsymbol{X}'}(A')^T$
- $Cov(A\boldsymbol{X} + A'\boldsymbol{X'}, B\boldsymbol{Y} + B'\boldsymbol{Y'} =$
  $A\Sigma_{\boldsymbol{X},\boldsymbol{Y}}B^T + A\Sigma_{\boldsymbol{X},\boldsymbol{Y'}}(B')^T + A'\Sigma_{\boldsymbol{X'},\boldsymbol{Y}}B^T + A'\Sigma_{\boldsymbol{X'},\boldsymbol{Y'}}(B')^T$

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

Density transformation formula

Density transformation

- p existing random variables $X_1, X_2, \cdots, X_p$ with density $f_{\boldsymbol{X}}(\boldsymbol{x})$ transformed into p new random variables $Y_1, X_2, \cdots, Y_p$ ($\boldsymbol{Y} \in \mathbb{R}^p$), via function

$$Y_i = y_i(X_1, X_2, \cdots, X_p),\ i = 1, 2, \cdots, p \tag{15}$$

- Must be smooth and one-to-one (invertible on codomain of $\boldsymbol{y}(\cdot)$)
- Call inverse $X_i = x_i(Y_1, Y_2, \cdots, Y_p),\ i = i = 1, 2, \cdots, p$
- Then

$$\begin{aligned} f_{\boldsymbol{Y}}(y_1, \cdots, y_p) =& f_{\boldsymbol{X}}(x_1(y_1, \cdots, y_p), \cdots, x_p(y_1, \cdots, y_p)) \\ & \times |J(y_1, \cdots, y_p)| \end{aligned} \tag{16}$$

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

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

Cont. Density transformation

- In (15) the absolute value of the Jacobian is substituted.

Characteristic and moment generating functions

Characteristic function

The characteristic function (cf) $\varphi_{\boldsymbol{X}}(\boldsymbol{t})$ of the random vector $\boldsymbol{X} \in \mathbb{R}^p$ is a function of a p-dimensional argument $\boldsymbol{t} = (t_1 \ t_2 \ \cdots t_p)^T \in \mathbb{R}^p$ is defined as

$$\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = E\left(e^{i\boldsymbol{t}^T\boldsymbol{X}}\right) \tag{17}$$

where $i = \sqrt{-1}$

- always exists (since $|\varphi_{\boldsymbol{X}}(\boldsymbol{t})| \leq E(\left|e^{i\boldsymbol{t}^T\boldsymbol{X}}\right| = 1 < \infty)$
- Related to the moment generating function (mgf):

  $$M_{\boldsymbol{X}}(\boldsymbol{t}) = E\left(e^{\boldsymbol{t}^T\boldsymbol{X}}\right)$$

  which may not exist for all $\boldsymbol{t}$.
- cf's have one-to-one correspondence with distributions
- under some conditions, can go the other way:

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

Characteristic function: linear transformation

Theorem

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

$$\varphi_{\boldsymbol{Y}}(\boldsymbol{s}) = e^{i\boldsymbol{s}^T\boldsymbol{b}}\varphi_{\boldsymbol{X}}(A^T\boldsymbol{s}) \tag{18}$$

Proof: at lecture.