Multivariate Analysis

Zhaoxia Yu
Professor, Department of Statistics

2023-04-04

Section 1

Intro

Subsection 1

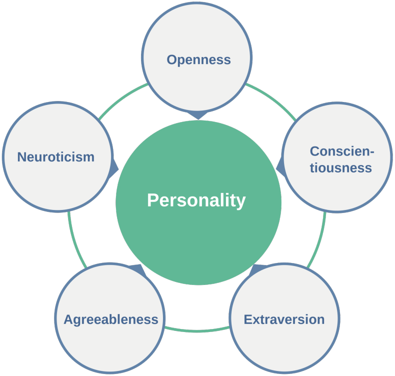
Introduction

Course Information

- Please use the Canvas website for course materials, important updates, and deadlines.
- Announcements will be sent to the mailing list or posted in Canvas.
- Assignment submission: GradeScope on Canvas.

Multivriate Data

- “multi” means more than one
- Multivariate data: the data with simultaneous measurements on many variables



More Examples of Multivariate Data

- A basketball player: points, rebounds, steals, assists, turnovers, free throws, fouls, etc
- A person's well-being: social, economic, psychological, medical, physical, etc
- A person's annual physical exam report

What is Multivariate Analysis

- The term “multivariate analysis” implies a broader scope than univariate analysis.
- Certain approaches like simple linear regression and multiple regression are typically not considered as multivariate analysis as they tend to focus on the conditional distribution of one univariate variable rather than multiple variables.
- Multivariate analysis focuses on the joint behavior of several variables simultaneously to identify patterns and relationships

Learning Objectives

- Matrix algebra, distributions
- Visualization
- Inference about a mean vector or multiple mean vectors
- Multivariate analysis of variance (MANOVA) and multivariate regression
- Linear discriminant analysis (LDA)
- Principal component analysis (PCA)
- Cluster analysis
- Factor analysis

Milestones in the history of multivariate analysis

- 1901: PCA was invented by Karl Pearson; independently developed by Harold Hotelling in the 1930s.
- 1904: Charles Spearman introduced factor analysis to identify underlying factors that explain the correlation between multiple variables.
- 1928: Wishart presented the distribution of the covariance matrix of a random sample from a multivariate normal distribution.
- 1936: Ronald Fisher developed discriminant analysis.
- ????: Cluster analysis.
- 1936: Canonical analysis by Harold Hotelling.
- 1960s: Multidimensional scaling.
- 1970s: Multivariate regression.
- 1980s: Structural equation modeling; the idea dated back to (1920-1921) by Sewall Wright.

Section 2

Matrix Algebra

Subsection 1

Vectors: We begin with a little bit matrix algebra

Vectors in R

- There are many ways to create or define a vector

```
x=rep(0.3, 4)
x
```

```
## [1] 0.3 0.3 0.3 0.3
```

```
x=seq(1, 4, by=0.2)
x
```

```
##  [1] 1.0 1.2 1.4 1.6 1.8 2.0 2.2 2.4 2.6 2.8 3.0 3.2 3.
```

```
c("a1", "a2", "a3")
```

```
## [1] "a1" "a2" "a3"
```

Vectors in R

```r
x=c(0.4, 0.2, 0.5)
x
```

```
## [1] 0.4 0.2 0.5
```

```r
length(x)
```

```
## [1] 3
```

```r
dim(x) #note that there is no dimension information
```

```
## NULL
```

A row or column of a matrix is also a vector

```
x=rbind(c(0.4,0.2,0.5), rep(1,3))
dim(x)
```

```
## [1] 2 3
```

```
x[1,]
```

```
## [1] 0.4 0.2 0.5
```

```
x[,1]
```

```
## [1] 0.4 1.0
```

Subsection 2

Special Matrices

Row or Column Vectors

- A vector (column vector) is a special matrix consisting of a single column of elements. e.g.,

$$a = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}$$

- A row vector is a special matrix consisting of a single row of elements

$$b = (b_1, b_2, b_3, b_4)$$

- In this class, a vector means a column vector
- A row or column vector is also a matrix
- The transpose of a row vector is a column vector; the transpose of a column vector is row vector. e.g.,

$$a' = (a_1, a_2, a_3)$$

Row or Column Vectors

- In vector/matrix operations, it is helpful to define row or column vectors
- A row vector

```
matrix(rep(0.5,3), 1, 3)
```

```
##      [,1] [,2] [,3]
## [1,]  0.5  0.5  0.5
```

```
dim(matrix(rep(0.5,3), 1, 3))
```

```
## [1] 1 3
```

```
#A neater way is to use the pipe "%>%"
matrix(rep(0.5,3), 1, 3) %>% dim
```

Row or Column Vectors

- A column vector

```
x= matrix(rep(0.5,3), 3, 1)
dim(x)
```

```
## [1] 3 1
```

```
# use pipe
x %>% dim
```

```
## [1] 3 1
```

Transposes

- The transpose of a column vector is a row vector
- The transpose of a row vector is a column vector

```
x= matrix(rep(0.5,3), 3, 1)
x
```

```
##      [,1]
## [1,]  0.5
## [2,]  0.5
## [3,]  0.5
```

```
t(x)
```

```
##      [,1] [,2] [,3]
## [1,]  0.5  0.5  0.5
```

Types of Special Matrices

- Identity matrix
- Diagonal matrix
- All-ones matrix
- Random matrix: a matrix whose entries are random variables. I will introduce matrix normal distributions.

Indentity Matrix

```
#diag(1, 2)
diag(5, 3)
```

```
##      [,1] [,2] [,3]
## [1,]    5    0    0
## [2,]    0    5    0
## [3,]    0    0    5
```

```
diag(1, 2, 3)
```

```
##      [,1] [,2] [,3]
## [1,]    1    0    0
## [2,]    0    1    0
```

Diagonal Matrix

```
diag(1:3)
```

```
##      [,1] [,2] [,3]
## [1,]    1    0    0
## [2,]    0    2    0
## [3,]    0    0    3
```

```
seq(1,2, by=0.5) %>% diag
```

```
##      [,1] [,2] [,3]
## [1,]    1  0.0    0
## [2,]    0  1.5    0
## [3,]    0  0.0    2
```

All-ones

```r
matrix(1, 3, 2)
```

```
##      [,1] [,2]
## [1,]    1    1
## [2,]    1    1
## [3,]    1    1
```

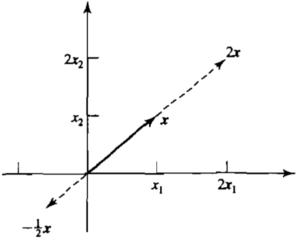
Subsection 3

Common Vector Operations

Scalar Multiplication

$$c\mathbf{x} = \begin{bmatrix} cx_1 \\ cx_2 \\ \vdots \\ cx_n \end{bmatrix} \quad \text{where} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

e.g.,

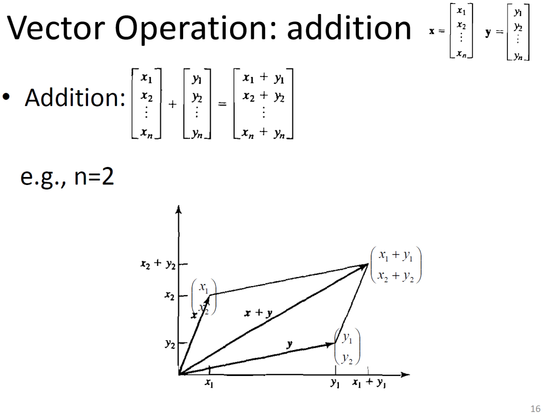


Examples of Scalar Multiplication

```r
x=matrix(c(0.4,0.2,0.5), 3, 1)
10*x
```

```
##      [,1]
## [1,]    4
## [2,]    2
## [3,]    5
```

Addition and Substraction



Vector Operation: addition

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

- Addition: $\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{bmatrix}$

e.g., n=2

$x_2 + y_2$
x_2
y_2
$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$
$\mathbf{x}$
$\mathbf{x} + \mathbf{y}$
$\mathbf{y}$
$\begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \end{pmatrix}$
$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$
x_1
y_1
$x_1 + y_1$

16

Example of Addition and Substraction

```
x1=matrix(c(0.4,0.2,0.5), 3, 1)
x2=rep(1, 3)
x1+x2
```

```
##      [,1]
## [1,]  1.4
## [2,]  1.2
## [3,]  1.5
```

```
x1-x2
```

```
##      [,1]
## [1,] -0.6
## [2,] -0.8
## [3,] -0.5
```

Outer Product

- The outer product of two vectors $x = (x_1, \cdots, x_m)'$ and $y = (y_1, \cdots, y_n)'$ is

$$x \otimes y = \begin{pmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_n \\ \cdots & \cdots & \cdots & \cdots \\ x_m y_1 & x_m y_2 & \cdots & x_m y_n \end{pmatrix}$$

- A similar operation for matrices is called Kronecker product.

Example: outer product

```
x1=matrix(c(0.4,0.2,0.5), 3, 1)
x2=rep(1, 3)
x1%*%x2
```

```
##      [,1] [,2] [,3]
## [1,]  0.4  0.4  0.4
## [2,]  0.2  0.2  0.2
## [3,]  0.5  0.5  0.5
```

Inner product

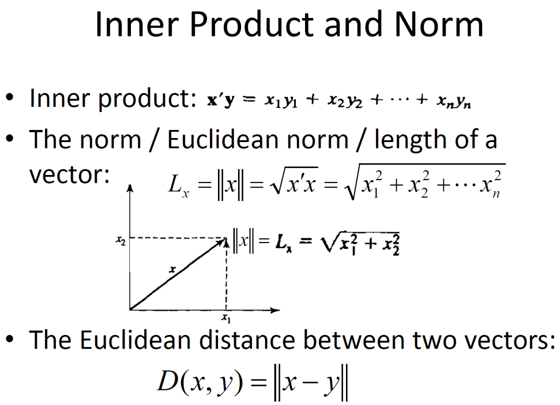
- Let $x = \begin{pmatrix} x_1 \\ \cdots \\ x_k \end{pmatrix}, y = \begin{pmatrix} y_1 \\ \cdots \\ y_k \end{pmatrix}$ The inner product of x and y is

$$< x, y >= x_1 y_1 + \cdots x_k y_k$$

- Note, the two vectors must have the same length
- The norm / Euclidean norm / length of x is $||x|| = \sqrt{< x, x >}$
- The Euclidean distance between x and y is

$$D(x, y) = ||x - y|| = \sqrt{(x_1 - y_1)^2 + \cdots (x_k - y_k)^2}$$

Inner Product and Norm



Inner Product and Norm

- Inner product: $\mathbf{x}'\mathbf{y} = x_1y_1 + x_2y_2 + \cdots + x_ny_n$
- The norm / Euclidean norm / length of a vector: $L_x = \|x\| = \sqrt{x'x} = \sqrt{x_1^2 + x_2^2 + \cdots x_n^2}$

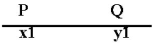
$$\|x\| = L_x = \sqrt{x_1^2 + x_2^2}$$

- The Euclidean distance between two vectors:

$$D(x, y) = \|x - y\|$$

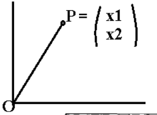
Distance: 1d and 2d

1 d

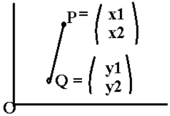


$$d(P,Q) = |x_1 - y_1| = \sqrt{(x_1 - y_1)^2}$$

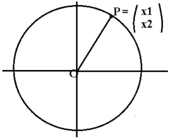
2d



$$d(O,P) = \sqrt{x_1^2 + x_2^2 + \cdots + x_p^2}$$



$$d(P,Q) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

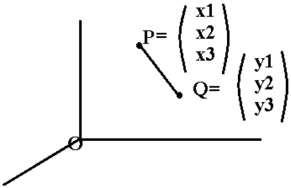


$$\{(x_1, x_2) : x_1^2 + x_2^2 = c^2\}$$

All the points on the circle have the same distance to the origin

Distance: 3d

3d or higher



$$d(P,Q) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2}$$

Example: Norm

```
x1=matrix(c(0.4,0.2,0.5), 3, 1)
#the norm/length of x1
sqrt(sum(x1^2))
```

```
## [1] 0.6708204
```

```
#or use pipe
x1^2 %>% sum %>% sqrt
```

```
## [1] 0.6708204
```

Example: (Euclidean) Distance

```
x1=matrix(c(0.4,0.2,0.5), 3, 1)
x2=rep(1, 3)
sqrt(sum((x1-x2)^2))
```

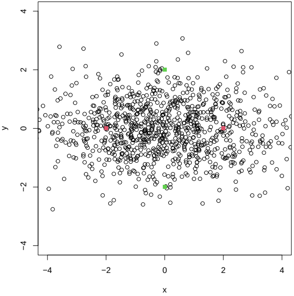
```
## [1] 1.118034
```

```
#or use pipe
(x1-x2)^2 %>% sum %>% sqrt
```

```
## [1] 1.118034
```

Example: (Euclidean) Distance

- Motivating example. Consider bivariate random vectors. The standard deviations are 2 and 1, respectively.
- What is the distance between (-2,0) and (2,0)? 4.
- What is the distance between (0, -2) and (0,2)? 4.



Example: (Euclidean) Distance

```
#The R code
set.seed(20230404)
mvrnorm(n=1000, c(0,0), matrix(c(4,0,0,1),2,2)) %>%
  plot(xlab="x", ylab="y", xlim=c(-4,4), ylim=c(-4,4))
points(x=c(-2, 0, 0, 2), y=c(0, -2, 2, 0), pch=15, col=c(2
```

- Both pairs have a distance of 4.
- The variation along x is greater than along y. Let X_1 and X_2 be two random points along the x direction, Y_1 and Y_2 be two random points along the y direction.

Euclidean Distance can be misleading

- Suppose X_1, X_2, Y_1, Y_2 are mutually independent.
  - X_1 and X_2 are iid from $N(\mu = 0, \sigma_x^2 = 2^2)$
  - Y_1 and Y_2 are iid from $N(\mu = 0, \sigma_y^2 = 1^2)$
- A homework problem. Calculate

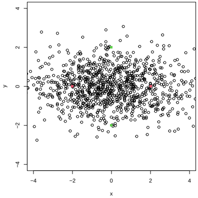
$$P(|X_1 - X_2| > 2), P(|Y_1 - Y_2| > 2)$$

  - First express each in terms of $\Phi(\cdot)$, the CDF of the standard normal distribution;
  - Then use the "pnorm" function in R to find the numerical values.
  - Last, estimate the two probabilities using simulations. The code in the previous page generates 1000 random samples. Change the sample size from n=1000 to n=10000 and then estimate the two probabilities. To do that, you need to examine all pairs of data points and then calculate the proportion of pairs satisfying a certain condition.

Statistical / Mahalanobis Distance

- The two probabilities are quite different, suggesting that the Euclidean distance might be misleading given the joint distribution of the variable.
- We will introduce a type of statistical distance, which is known as Mahalanobis distance.

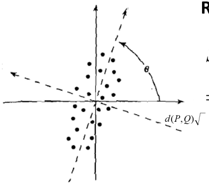
Statistical Distance



- The scatter plot suggests that we should make the two variables equally spread out. We can first standardize each of them.
- point (-2,0) becomes (-1,0), points (2, 0) becomes (1,0)
- Using the standardized points, the distance between the red pair is 2, the distance between the green pair is 4.

Statistical Distance

- In The example above X and Y are independent, as a result, the covariance is zero. Statistical distance can also be defined when the covriance matrix Σ is not diagonal;



Rotation and Standardization

$$SD(O,P) = \sqrt{x^T \Sigma^{-1} x}$$

$$= \sqrt{(\Gamma x)^T \Lambda^{-1} \Gamma x}$$

When the ellipse is not centered at the origin:

$$SD(O,P) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

It is also knowns as Mahalanobis distance.

–>