Lecture 1

Introduction and Multivariate Data Exploration

Readings: Zelterman 2015 Chapter 1; Chapter 3

DSA 8070 Multivariate Analysis

Whitney Huang
Clemson University

Notes

Agenda

1. Course Overview
2. Multivariate Data Exploration: Numerical Summary
3. Multivariate Data Exploration: Graphical Summary

Notes

Course Overview

Notes

Introduction

- In many observational or experimental studies, measurements are collected simultaneously on more than one variable on each unit

```
> head(Boston)
     crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat medv
1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98 24.0
2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14 21.6
3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03 34.7
4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94 33.4
5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33 36.2
6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21 28.7
```

- Multivariate analysis is the collection of statistical methods that can be used to (jointly) analyze these multiple measurements
  ⇒ *some are extensions of familiar methods (t-test, ANOVA, Linear Regression, ...) while others are unique to multivariate analysis (PCA, CCA, Factor Analysis, ...)*

> To exploit potential "correlations" among the multiple measurements to improve inference

Notes

Introduction and Multivariate Data Exploration
CLEMSON UNIVERSITY
Course Overview
Multivariate Data Exploration: Numerical Summary
Multivariate Data Exploration: Graphical Summary

Why Univariate Analysis Falls Short

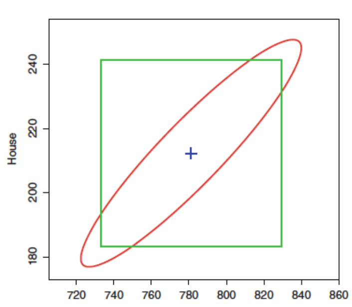
We've learned many tools for handling a single response variable (aka univariate analysis). Why bother with multivariate analysis?

- If variables are truly independent, analyzing each separately with histograms, box plots, or basic summary statistics is sufficient
- When variables influence each other, analyzing them in isolation misses important patterns and interactions
- Considering all variables together reveals connections that are invisible in separate analyses

Notes

Introduction and Multivariate Data Exploration
CLEMSON UNIVERSITY
Course Overview
Multivariate Data Exploration: Numerical Summary
Multivariate Data Exploration: Graphical Summary

Using Multivariate Methods Could Lead to Sharper Inference



Source: Fig. 1.1 of Zelterman 2015

> Joint analysis captures relationships that improve accuracy!

Notes

Dimensionality Reduction or Structural Simplification

- Goal: Reduce the “dimensionality” by representing many variables with a small number of (linear) combinations, without losing important information
- Example:
  A single index of a patient’s reaction to radiotherapy can be constructed from several related response measurements
  This is dimensionality reduction because it replaces multiple variables with one composite measure, *reducing the number of dimensions from many to one while still summarizing the essential information*
- Techniques:
  - **Principal Component Analysis** (Week 8)
  - **Factor Analysis** (Week 9)
  - **Multidimensional Scaling** (Week 13)

Notes

Grouping or Classification

- Goal: **Identify** groups of “similar” units or **classify** units into previously defined groups.
- Example:
  Using the concentration of elements (copper, silver, tin, antimony) in bullet lead, the FBI can **identify** bullets from the same production batch
  This is relevant because it groups items with similar characteristics from multiple measurements - the aim of grouping and classification
- Techniques:
  - **Classification Analysis** (Week 11)
  - **Cluster Analysis** (Week 12)

Notes

Dependence among Variables and Prediction

- Goal: Estimate relationships among variables and predict the value of some variables from others.
- Example:
  The association between test scores and several college performance variables can be used to predict a student’s likelihood of success in college
  This illustrates the goal because it models how variables are related and uses those relationships for prediction.
- Techniques:
  - **Multivariate Regression** (Weeks 5-6)
  - **Canonical Correlation Analysis** (Week 10)

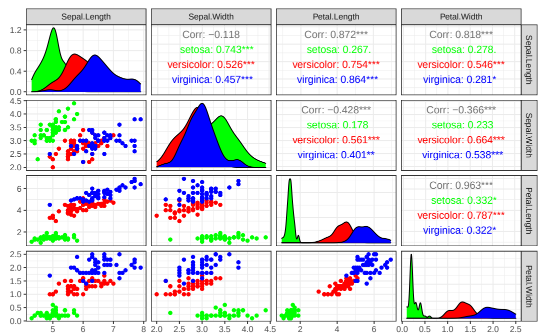
Notes

Hypothesis Testing

- Goal: Test differences in mean vectors across groups and test hypotheses about covariance matrices.
- Examples:
  - Compare mean gasoline mileage, repair costs, and downtime for different truck models
  - Check if the covariance structure of asset returns is the same before and after a market event
- Techniques:
  - **Hotelling's T^2 and MANOVA** (Week 4)
  - **Covariance Matrix Inference** (Week 7)

Notes

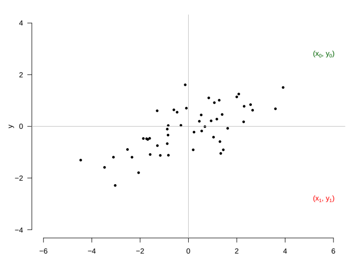
Exploratory Data Analysis: A Powerful First Step in Multivariate Analysis [EDA, Tukey 1977]



Notes

Statistical Distance: A Core Concept in Multivariate Analysis

Multivariate methods rely on "distances" between data points: clustering (group units that are "close"); classification (allocate each unit to the "closest" group)



Question: which one ((x_0, y_0) or (x_1, y_1)) is closer the center of the observations? ⇒ We will learn **Mahalanobis distance** to formally answer this question

Notes

Matrix Algebra (Week 2)

```
     crim zn indus chas   nox    rm
1 0.00632 18  2.31    0 0.538 6.575
2 0.02731  0  7.07    0 0.469 6.421
3 0.02729  0  7.07    0 0.469 7.185
4 0.03237  0  2.18    0 0.458 6.998
5 0.06905  0  2.18    0 0.458 7.147
6 0.02985  0  2.18    0 0.458 6.430
```

$$X_{n\times p} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

- Many operations performed on multivariate data are presented using vector/matrix notation, e.g., $\boldsymbol{X}_{n\times p}$ (Data matrix); $\hat{\boldsymbol{\mu}}_{p\times 1}$ (estimated mean vector); $\hat{\boldsymbol{\Sigma}}_{p\times p}$ (estimated covaraince matrix)
- The computation of eigenvalues and eigenvectors (i.e., the **spectral decomposition**) plays an important role in multivariate analysis $\Rightarrow$ reveals key directions and magnitudes of variation, forming the basis of many multivariate methods

Notes

Multivariate Normal Distribution (Week 3)

- We will often assume that the joint distribution of $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^{\mathrm{T}}$ follows a multivariate normal distribution with probability density function:

$$f(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{d}{2}} \det(\boldsymbol{\Sigma})^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right]$$

- The multivariate normal assumption is often useful:
  - Variables may be approximately multivariate normal, possibly after transformation.
  - By the Central Limit Theorem, many **multivariate sample statistics** have an approximately normal distribution, regardless of the population distribution.
  - We will also briefly cover copula models and nonparametric methods when the assumption is not appropriate.

Notes

Data Mining, Machine Learning, and Multivariate Analysis

- Data Mining is the process of extracting and discovering patterns (e.g., unexpected structures or relationships, trends, clusters, and outliers) in **massive data sets**
- Supervised learning and unsupervised learning are two most common problems in machine learning
- Data mining/machine learning applications usually involve many variables, often related in complex ways, hence techniques from multivariate analysis play an important role

Notes

Multivariate Data Exploration: Numerical Summary

Notes

Organization of Data and Notation

- n: number of units; p: number of variables per unit $\Rightarrow$ $p = 1$ is univariate, $p > 1$ is multivariate.
- x_{ik}: k-th measurement on unit i; unit i: $(x_{i1}, x_{i2}, \ldots, x_{ip})$. Measurements from n units can be arranged in matrix form:

$$X_{n\times p} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

where rows correspond to units and columns to variables

Notes

Descriptive Statistics: Sample Mean & Variance

- The sample mean of the k-th variable ($k = 1, \cdots, p$) is computed as

$$\bar{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik}$$

- The sample variance of the k-th variable is usually computed as

$$s_k^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2$$

and the sample standard deviation is given by

$$s_k = \sqrt{s_k^2}$$

Notes

Descriptive Statistics: Sample Covariance

- We often use s_{kk} to denote the sample variance for the k-th variable. Thus,

$$s_k^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2 = s_{kk}$$

- The sample covariance between variable k and variable j is computed as

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

- If variables k and j are independent, the population covariance is exactly zero, but the sample covariance will fluctuate around zero

```
```{r}
dat <- mvrnorm(n = 50, mu = c(0, 0), Sigma = matrix(c(1, 0, 0, 1), 2))
cov(dat[, 1], dat[, 2])
```
```

```
[1] -0.1508848
```


Notes

Descriptive Statistics: Sample Correlation

- The sample correlation between variables k and j is defined as

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}}$$

- r_{jk} is symmetric ($r_{jk} = r_{kj}$) and lies between -1 and 1

- The sample correlation is equal to the sample covariance if measurements are standardized (i.e., $s_{kk} = s_{jj} = 1$)

- Covariance and correlation measure linear association. Other non-linear dependencies may exist among variables even if $r_{jk} = 0$

- The sample correlation (r_{ij}) will vary about the value of the population correlation (ρ_{ij})

Notes

Matrix Representation of Sample Statistics

Sample statistics of a p-dimnesional multivariate data can be organized as vectors and matrices:

- $\bar{\boldsymbol{x}} = [\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_p]^{\mathrm{T}}$ is the $p \times 1$ vector of sample means

- $\boldsymbol{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \cdots & \cdots & \cdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$ is the $p \times p$ symmetric matrix of variance (on the diagonal) and covariances (the off-diagonal elements)

- $\boldsymbol{R} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \cdots & \cdots & \cdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}$ is the $p \times p$ symmetric matrix of sample correlations. Diagonal elements are all equal to 1

Notes

Generalized Variance

- The generalized variance is a scalar value which generalizes variance for multivariate random variables
- The generalized variance is defined as the determinant of the (sample) covariance matrix $\boldsymbol{S}$, $\det(\boldsymbol{S})$
- **Example:**

```{r}
data(mtcars)
vars <- which(names(mtcars) %in% c("mpg", "disp", "hp", "drat", "wt"))
car <- mtcars[, vars]; S <- cov(car)
(genVar <- det(S))
```

```
[1] 3951786
```


Notes

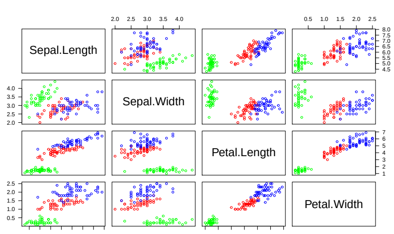
Multivariate Data Exploration: Graphical Summary

Introduction and Multivariate Data Exploration
CLEMSON UNIVERSITY
Course Overview
Multivariate Data Exploration: Numerical Summary
Multivariate Data Exploration: Graphical Summary

Notes

Graphs and Visualization

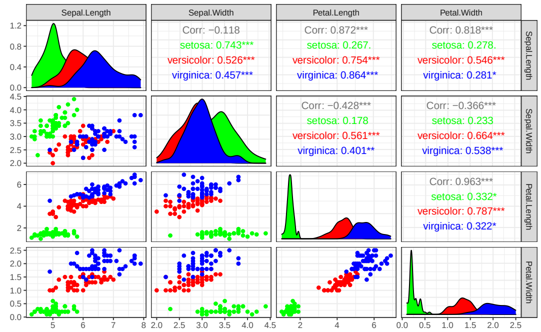
- Graphs reveal variable associations and unusual observations.
- Visualizing multivariate data is challenging, especially for $p > 3$.
- At minimum, use pairwise scatter plots.



Introduction and Multivariate Data Exploration
CLEMSON UNIVERSITY
Course Overview
Multivariate Data Exploration: Numerical Summary
Multivariate Data Exploration: Graphical Summary

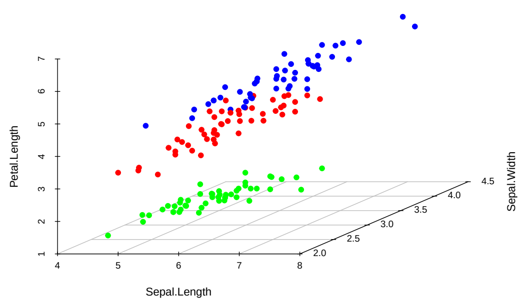
Notes

Visualizing Data with *ggpairs*: Combining Scatterplots and Numerical Summaries



Notes

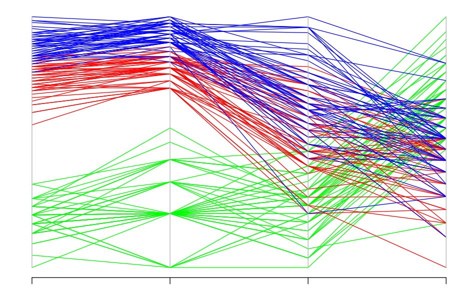
3D Scatter Plot



Notes

Parallel Coordinate Plot

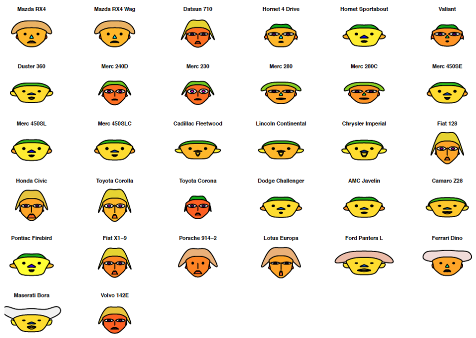
- Plots each observation across parallel axes (one per variable).
- Shows patterns, group differences, and outliers.
- Axis order and scaling affect interpretation.



Notes

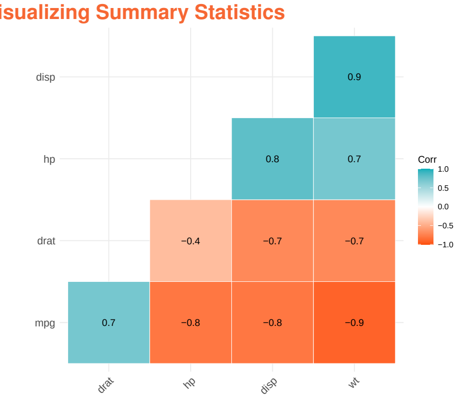
Chernoff Faces

```
> head(mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```



Notes

Visualizing Summary Statistics



Visualize summary measures (e.g., correlation matrix) with color showing strength and direction of relationships

Introduction and Multivariate Data Exploration
Course Overview
Multivariate Data Exploration: Numerical Summary
Multivariate Data Exploration: Graphical Summary

Notes

Summary

In this lecture, we covered:

- A high-level overview of the course
- Summarizing multivariate data numerically
- Summarizing multivariate data graphically

In the next lecture, I will give a short review of Matrix Algebra

Introduction and Multivariate Data Exploration
Course Overview
Multivariate Data Exploration: Numerical Summary
Multivariate Data Exploration: Graphical Summary

Notes