Lecture 1

Course Infromation and Overview

DSA 8070 Multivariate Analysis

Whitney Huang
Clemson University

Instructor Background
Class Policies
Course Overview

Agenda

Instructor Background
Class Policies
Course Overview

1. Instructor Background
2. Class Policies
3. Course Overview

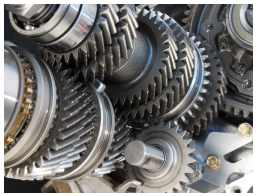
Instructor Background
Class Policies
Course Overview

About the Instructor

About the Instructor

- Assistant Professor of Applied Statistics and Data Science
- Born in Laramie, WY, and grew up in Taiwan

- Obtained a B.S. in Mechanical Engineering and switched to Statistics in graduate school



- Earned a Ph.D. in Statistics in 2017 from Purdue University.

How to Reach Me?

- **Email** ✉: wkhuang@clemson.edu

  Please include [DSA 8070] in your email subject line

- **Office:** O-221 Martin Hall

- **Office Hours:** Wednesday 8-9 pm and by appointment

Instructor Background
Class Policies
Course Overview

Class Policies

Logistics

- There will be two projects. The due dates are:
  - **Project I**: Oct. 17, Thursday
  - **Project II**: Dec. 12, Thursday
- There will be biweekly R Labs:
  - To be uploaded to Canvas by 11:59 pm ET on the due dates
  - Worst grade will be dropped
- No lectures during Thanksgiving week (Nov. 25-29)

Course Materials at CANVAS

- Course syllabus / Announcements
- Lecture slides/notes/videos
- R Labs/Projects
- Data sets for lectures and labs

Reference Books

- *Applied Multivariate Statistics with* R, **Daniel Zelterman, 2015** [Link]
- *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*, **Alan Izenman, 2008**, [Link]
- *Methods of Multivariate Analysis*, 3_{rd} Edition, **Alvin Rencher and William Christensen, 2012** [Link]
- *Applied Multivariate Statistical Methods,* 6_{th} Edition, **Richard Johnson and Dean Wichern, 2008** [Link]

Evaluation

Grades will be weighted as follows:

R Labs	20%
Project I	40%
Project II	40%

Final course grades will be assigned using the following grading scheme:

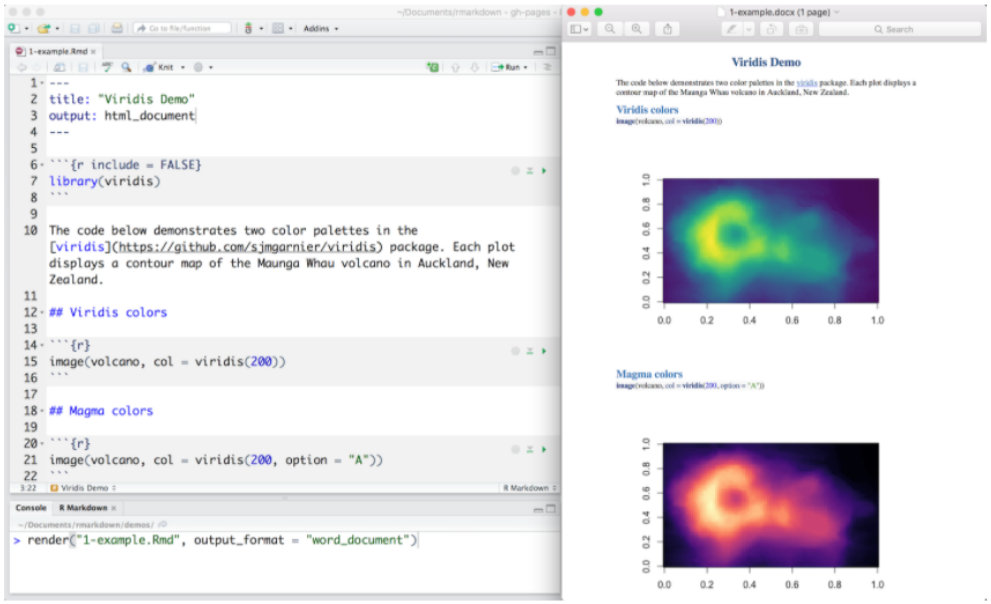
>= 90.00	A
88.00 ~ 89.99	A-
85.00 ~ 87.99	B+
80.00 ~ 84.99	B
78.00 ~ 79.99	B-
75.00 ~ 77.99	C+
70.00 ~ 74.99	C
68.00 ~ 69.99	C-
<= 67.99	F

Computing

We will use software to perform statistical analyses.

Specifically, we will be using `R/Rstudio`

- a **free/open-source** programming language for statistical analysis
- available at `https://www.r-project.org/` (`R`); `https://rstudio.com/` (`Rstudio`)
- I strongly encourage you to use **`R` Markdown** for homework assignments



Topics

Week	Dates	Topic
1	8/19 - 8/23	Introduction
2	8/26 - 8/30	Characterizing and Displaying Multivariate Data
3	9/2 - 9/6	A Short Review of Matrix Algebra
4	9/9 - 9/13	Multivariate Normal Distribution and Copula
5	9/16 - 9/20	Inferences about a Mean Vector
6	9/23 - 9/27	Comparisons of Several Mean Vectors
7	9/30 - 10/4	Multivariate Linear Regression
8	10/7 - 10/11	Repeated Measures Analysis
9	10/14 - 10/18	Principal Components Analysis
10	10/21 - 10/25	Factor Analysis
11	10/28 - 11/1	Canonical Correlation Analysis
12	11/4 - 11/8	Discrimination and Classification
13	11/11 - 11/15	Cluster Analysis
14	11/18 - 11/22	Multidimensional Scaling
15	11/25 - 11/29	**No Class–Thanksgiving**
16	12/2 - 12/6	Review

Instructor Background
Class Policies
Course Overview

Course Overview

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

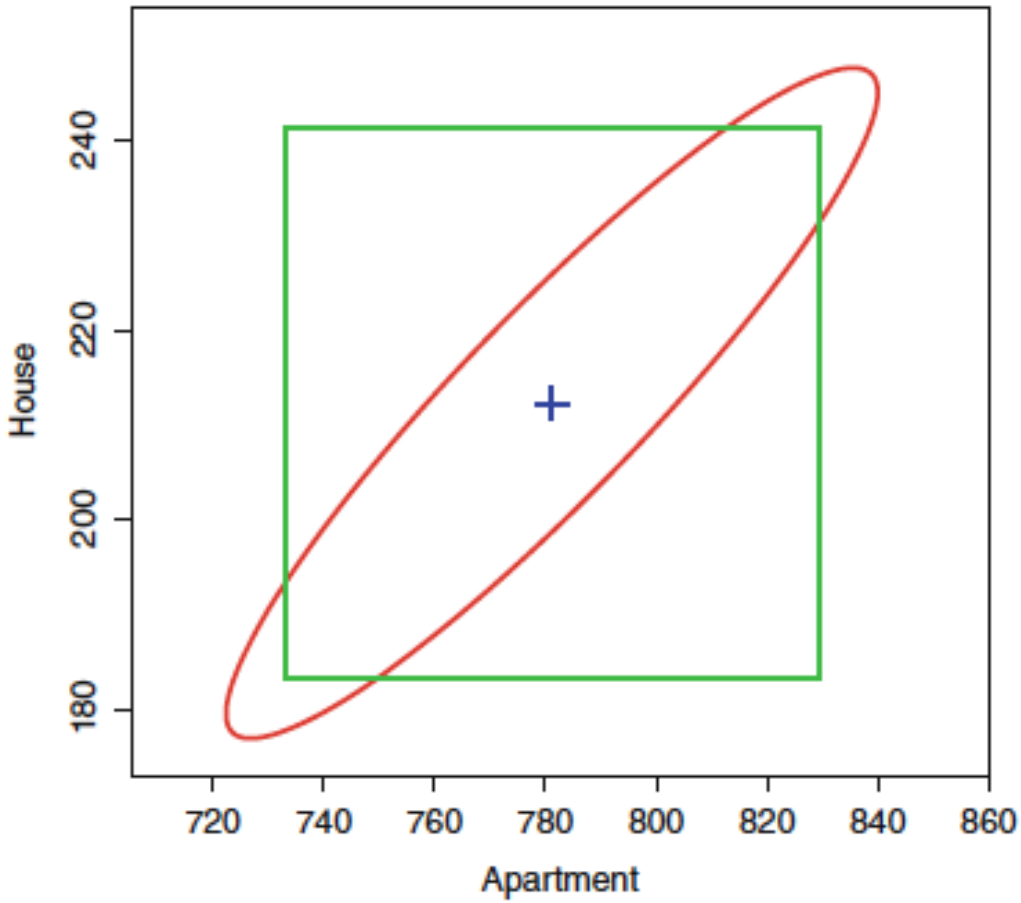
  $\Rightarrow$ *some are extensions of familiar methods (t-test, ANOVA, Linear Regression, ...) while others are unique to multivariate analysis (PCA, CCA, Factor Analysis, ...)*

- The idea is to exploit potential "correlations" among the multiple measurements to improve inference

Importance of Multivariate Analysis

- If all the variables are independent, one can't do better than analyze each variable's behavior by using histograms or box plots, looking at the means, medians, variances and other 'one dimensional statistics'
- However if some of the variables are acting together, either that they are positively correlated or that they inhibit each other, one will miss a lot of important information by slicing the data up into those column vectors and studying them separately
- Thus important connections between variables are only available to us if we consider the data as a whole.

Using Multivariate Methods Could Lead to Sharper Inference



Source: Fig. 1.1 of Applied Multivariate Statistics with R by Zelterman

Dimensionality Reduction or Structural Simplification

- Goal: to reduce the "dimensionality" by considering a small number of (linear) combinations of a large number of measurements without losing important information

- Examples:
  - A single index of patient reaction to radiotherapy can be constructed from measurements on several response variables
  - Wildlife ecologists can construct a few indices of habitat preference from measurements of dozens of features of nesting sites selected by a certain bird species

- Techniques:
  - **Principal Component Analysis** (Week 9)
  - **Factor Analysis** (Week 10)
  - **Multidimensional Scaling** (Week 14)

Grouping or Classification

- Goal: to **identify** groups of "similar" units or to **classify** units into previously defined groups

- Examples:
  - Using the concentration of elements (copper, silver, tin, antimony) in the lead alloy used in bullets, the FBI **identifies** 'similar' bullets that may be used to infer whether bullets were produced from the same batch of lead
  - The US IRS uses data collected from tax returns (income, amount withheld, deductions, ...) to **classify** taxpayers into two groups: those who will be audited and those who will not

- Techniques:
  - **Classification Analysis** (Week 12)
  - **Cluster Analysis** (Week 13)

Dependence among Variables and Prediction

- Goal: to estimate the relationship among variables and to predict the value of some of them given information on the others

- Examples:
  - The associations between measures of risk-taking propensity and measures of socioeconomic characteristics for top-level business executives were used to assess the relation between risk-taking behavior and performance
  - The association between test scores, and several college performance variables were used to develop predictors of success in college

- Techniques:
  - **Multivariate Regression** (Week 7)
  - **Repeated Measures Analysis** (Week 8)
  - **Canonical Correlation Analysis** (Week 11)

Hypothesis Testing

- Goal: to test if differences in sets of response mean vectors for two or more groups large enough to be distinguished from sampling variation

- Examples:

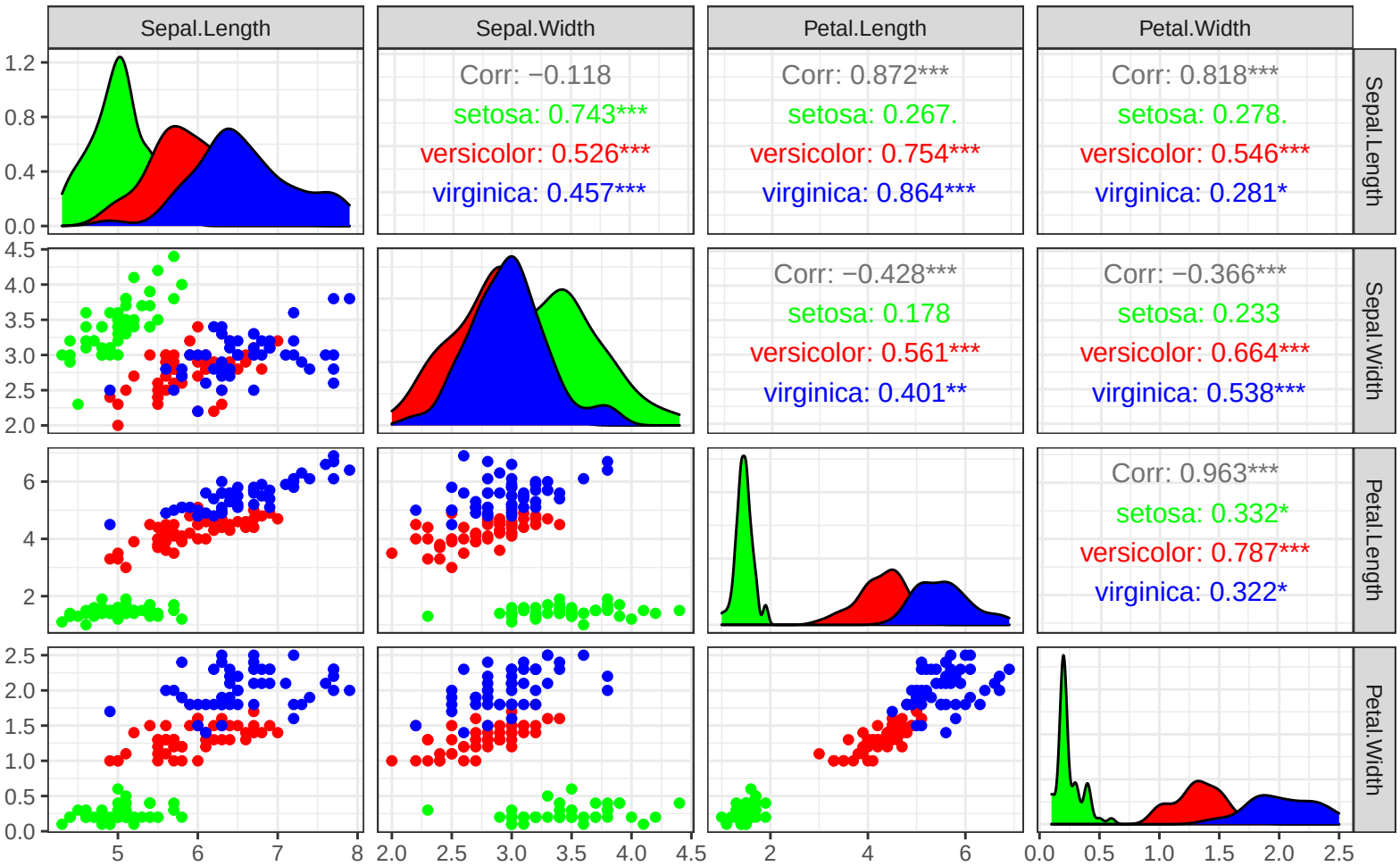
  - A transportation company wants to know if means for gasoline mileage, repair costs, downtime due to repairs differ for different truck models

  - An insurance company wants to know if changing case management practices leads to changes in mean length of hospital stay, mean infection rates, and mean costs

- Techniques:

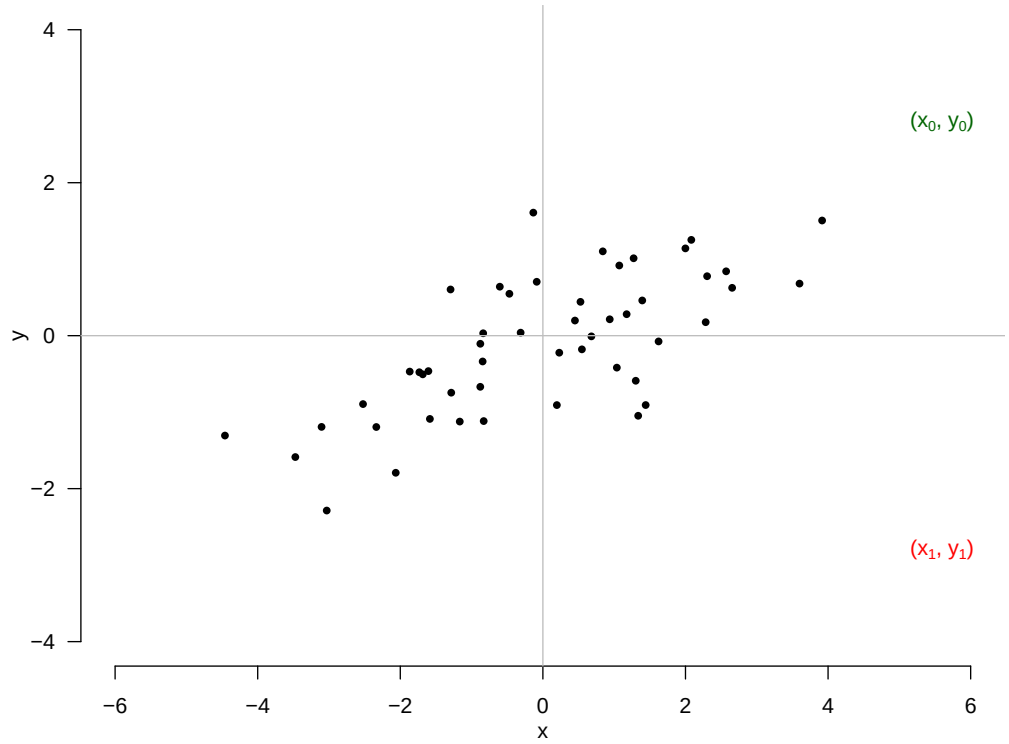
  - **Hotelling's T^2 and MAVONA** (Week 5 and Week 6)

Exploratory Data Analysis [EDA, Tukey 1977]



Statistical Distance

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? $\Rightarrow$ We will learn **Mahalanobis distance** to formally answer this question

Instructor Background
Class Policies
Course Overview

The study of multivariate methods is greatly facilitated by the use of matrix algebra

- 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
- We will use `R` to perform the needed matrix operations

Multivariate Normal Distribution (Week 4)

- We will often assume the joint distribution of $\boldsymbol{X} = (X_1, X_2, \cdots, X_p)^{\mathrm{T}}$ follows a multivariate normal distribution with the 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 appropriate:
  - Variables can sometimes be assumed to be multivariate normal (perhaps after transformation)
  - Central limit theorem tells us that distribution of many **multivariate sample statistics** is approximately normal, regardless of the form of the population distribution

Data Mining, Machine Learning, and Multivariate Analysis

CLEMSON UNIVERSITY

Instructor Background
Class Policies
Course Overview

- 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