Introduction to Multivariate Statistics
Lecture 1: Overview

Lexin Li

University of California, Berkeley

Course information

- Instructor: Lexin Li
  - Professor, Department of Biostatistics and Epidemiology
  - Email: lexinli@berkeley.edu
  - Office: Berkeley Way West, Room 5330
  - Office phone: 510-664-4584
  - Lecture: M 2-5pm, 145 Moffitt
  - Office hour: M 1-2pm
- GSIs:
  - Yang Li: liyangc@berkeley.edu
  - Weijie Yuan: weijie_yuan@berkeley.edu
- Lab: computing and hands-on data analysis examples
  - Section 1: Th 2-4pm, 136 Barrows
  - Section 2: F 2-4pm, 56 Barrows
  - Section 3: W 4-6pm, 234 Dwinelle
  - Section 4: W 2-4pm, 110 Barker

Course information

- Suggested books for reading:
  - *Applied Multivariate Statistical Analysis*, Richard Johnson and Dean Wichern, 2007, 6th edition, Pearson Prentice Hall
  - *Statistics for Epidemiology*, Nicholas Jewell, 2003, Chapman & Hall
  - *Applied Regression Including Computing and Graphics*, R Dennis Cook and Sanford Weisberg, 1999, Wiley
  - *The Elements of Statistical Learning – Data Mining, Inference and Prediction*, Trevor Hastie, Robert Tibshirani, and Jerome Friedman, 2009, 2nd edition, Springer
- Instructor's lecture notes
- Web:
  - All course materials can be found on bCourses

Course information

- Topics and tentative schedule:
  - Course overview: 09/09
  - Review of basic concepts and matrix algebra: 09/09, 09/16
  - Comparison of multivariate means (chapter 6): 09/16, 09/23
  - Linear regression model (chapter 7, Cook and Weisberg's book): 09/30, 10/07, 10/14
  - Logistic regression model (Jewell's book, chapters 12-15): 10/21, 10/28
  - Principal components analysis (chapter 8): 11/04
  - Factor analysis (chapter 9): 11/18
  - Classification (chapter 11): 11/25
  - Clustering (chapter 12): 12/02
- Computing:
  - **R, R, R**
  - Lecture: command, output, and interpretations
  - Lab: everything else

Course information

- Homework:
  - There will be **4 homework assignments** + a suggested problem set
  - You must do the homework **on your own**
  - Due time: **two weeks** from the assignment
  - A **tentative** homework schedule:
    - Homework 1: 09/23, comparison of multivariate means
    - Homework 2: 10/14, linear regression model
    - Homework 3: 10/28, liner and logistic regression model
    - Homework 4: 11/18, principal components analysis and factor analysis
    - Suggestion problem set: 12/02, classification and clustering
- Exam:
  - There is **NO in-class final exam**.
- Grading:
  - class and lab attendance 10% + homework 40% + project 50%

Course information

- Final project:
  - There is a final project on real data analysis.
  - You can **pair up** to do the project, and each team is **no more than 2 persons**.
  - Data:
    - Analysis of a data set of moderate complexity, using one or more of the techniques covered in the course.
    - It would be best if this data is from your own research / thesis / dissertation, while I can recommend some datasets as well.
  - Schedule:
    - Please **email** the instructor a **one-page project proposal** (problem to address, key info about the data, the method to use, etc) **between November 11 and 15**.
    - The **final project report** is **due on December 16**.

Course information

- Final project report:
  - No more than **5 pages, including** figures and tables.
  - Please divide the report in the following **sections**:
    - Executive summary (using bullet points)
    - Background
    - Problem (the question you wish to address)
    - Data (summary of the data, the study design, data collection)
    - Method (your choice of model / analytic method, and why)
    - Results (summary of numerical analysis, interpretation, assumptions check)
    - Conclusion.

Course information

- Objectives: by the end of the semester, you can
  - have an appreciation of a range of multivariate methods and their use and limitations in a research context
  - examine critically other researchers' use of methods of analysis for multivariate data
  - select (**know what to search and why to choose**), carry out and interpret appropriate statistical methods for describing and analyzing multivariate data sets, in the context of your own research
  - If you have any specific objective in mind, feel free to let me know!
- A note about math:
  - Will I use a fair amount of math (such as matrix algebra)?
    The answer is

Course information

- Objectives: by the end of the semester, you can
  - have an appreciation of a range of multivariate methods and their use and limitations in a research context
  - examine critically other researchers' use of methods of analysis for multivariate data
  - select (**know what to search and why to choose**), carry out and interpret appropriate statistical methods for describing and analyzing multivariate data sets, in the context of your own research
  - If you have any specific objective in mind, feel free to let me know!
- A note about math:
  - Will I use a fair amount of math (such as matrix algebra)?
    The answer is YES!

Course information

- Objectives: by the end of the semester, you can
  - have an appreciation of a range of multivariate methods and their use and limitations in a research context
  - examine critically other researchers' use of methods of analysis for multivariate data
  - select (**know what to search and why to choose**), carry out and interpret appropriate statistical methods for describing and analyzing multivariate data sets, in the context of your own research
  - If you have any specific objective in mind, feel free to let me know!
- A note about math:
  - Will I use a fair amount of math (such as matrix algebra)? The answer is YES!
  - How to deal with it?

Course information

Given data $x_1, x_2, \ldots, x_n$. i.i.d $\sim N_p(\mu, \Sigma)$, the likelihood function is

$$
\begin{aligned}
L(\mu, \Sigma) &= \prod_{i=1}^{n} \left\{ \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x_i - \mu)'\Sigma^{-1}(x_i - \mu) \right] \right\} \\
&= \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)'\Sigma^{-1}(x_i - \mu) \right] \\
&= \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n} tr\left\{(x_i - \mu)'\Sigma^{-1}(x_i - \mu)\right\} \right] \\
&= \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} \exp\left[-\frac{1}{2} tr\left\{ \Sigma^{-1}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)' \right\} \right] \\
&= \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} \exp\left[-\frac{1}{2}\left(tr\left\{ \Sigma^{-1}\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})' \right\} \right.\right. \\
&\quad \left.\left. + n(\overline{x} - \mu)'\Sigma^{-1}(\overline{x} - \mu) \right) \right]
\end{aligned}
$$

Course information

- What I plan to do:
  - **Review** basic matrix algebra and concepts
  - Emphasize **intuition** behind each method
  - Emphasize **assumptions** of each method and their consequences
  - Emphasize characterization of **uncertainty**
  - Connect to some typical **real world examples**
  - Connect with other related methods
  - Introduce briefly some more recent extensions

- It would be helpful to always keep in mind:
  - What is the method about? / What kind of question does this method try to address? – Can you explain it in no more than 3 min / 3 sentences?
  - What is the intuition behind this method?

Comparison of multivariate means

- Motivating example: Anesthetizing effect of CO_2 ad halothane
  - $n = 19$ dogs, **each** of which was treated with 4 treatments
  - the response variable is milliseconds between heartbeats
  - 4 different treatments: 2 CO_2 pressures $\times$ halothane

Treatment	CO_2 pressure	Halothane
1	high	present
2	low	absent
3	high	present
4	low	absent

- What will change to answer the same question?
  - Suppose we select 5 dogs for each of 4 treatments
  - Suppose there are 20 different levels of CO_2
  - Suppose the CO_2 level should really be treated as continuous rather than discrete

Comparison of multivariate means

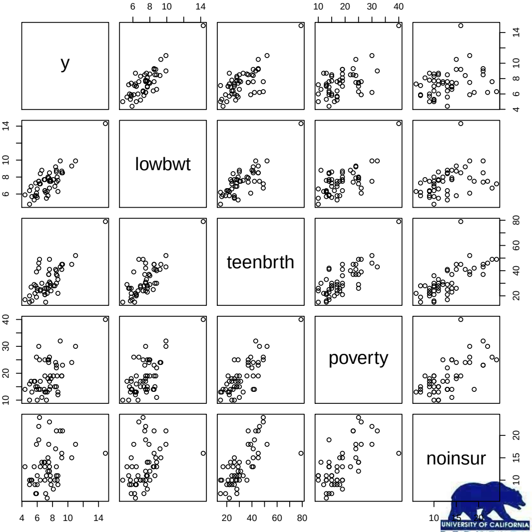
- What is it about:
  - Compare **quantitative** measures of **subjects** between groups that are defined by **factor(s)** with two or more **levels**

- Topics to cover:
  - Same subjects – within-subject comparison (Section 6.2)
    - multiple variables – paired comparison
    - multiple measurements – repeated measures design
  - Different subjects – between-subject comparison
    - one factor: two populations (Section 6.3) – two sample T^2 test
    - one factor: more than two populations (Section 6.4) – one-way MANOVA
    - two factors – two-way MANOVA (Section 6.7)
  - Multiple testing

- What to pay special attention:
  - One-to-one correspondence to the **univariate** comparison
  - Assumptions! (because they lead to different choices of test)

Linear regression model

- Motivating example: U.S. infant mortality rate from Annie E. Casey Kids Count Data Center
  - Y: infant mortality rate
  - X_1: low birthweight rate
  - X_2: teen birth rate
  - X_3: poverty rate
  - X_4: no insurance rate
  - 50 states + D.C.
  - many other variables



Linear regression model

- What is it about:
  - **Association/relation** between **response/output/dependent variable** (Y) and **predictor/input/feature variable** $\boldsymbol{X}$; how the value of Y changes as a function of $\boldsymbol{X}$

- Topics to cover:
  - Data visualization
  - Model, interpretation, estimation, prediction
  - Characterization of uncertainty
  - Categorical explanatory variables
  - Goodness-of-fit, model diagnosis, and remedies
  - Extensions: multivariate responses, nonlinear models, variable selection

- What to pay special attention:
  - Interpretation, interpretation, interpretation!
  - Is this a good model?

Logistic regression model

- Motivating example: western collaborative group study
  - Samples: 3154 men, ages 39 to 59, free of coronary heart disease at the beginning of the study
  - Response: a **binary** indicator whether a CHD event occurred (about 8%) within the 8.5-year follow-up period
  - Predictors: the type A/B behavior pattern, height and weight, total cholesterol levels, systolic and diastolic blood pressure, and smoking history (number of cigarettes smoked per day)
  - One of main goals was to was to explore the relationship between behavior pattern, so-called Type A behavior, and the risk of coronary heart disease (CHD)

Logistic regression model

- What is it about:
  - Model the **probability of disease** as a function of a number of explanatory variables
  - Explanatory variables include: **exposure/risk factors**, or **treatment variables** + covariates to adjust for

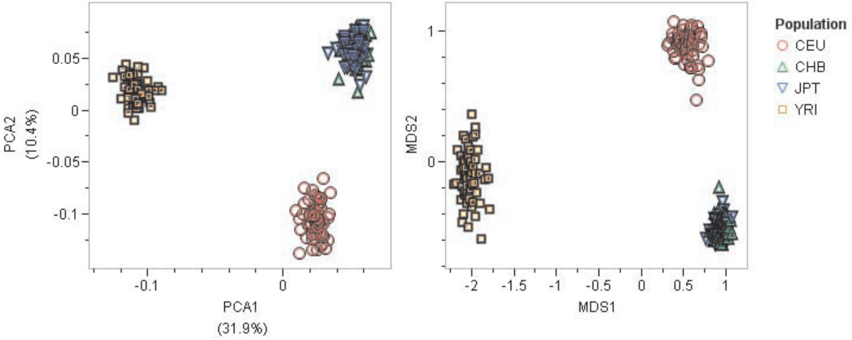
- Topics to cover:
  - Model, interpretation, estimation
  - Extensions: parallel to linear regression

- What to pay special attention:
  - Interpretation: relative risks, odds ratio
  - Study design and how it affects the model

Principal components analysis

- Motivating example: HapMap data
  - HapMap: an international organization that aims to develop a haplotype (collection of specific alleles) map of the human genome, which will describe the common patterns of human genetic variation
  - 2918 SNPs on chromosome 21 (smallest chromosome; associated with diseases such as Down syndrome)
  - 208 Yoruban (YRI, an ethnic group in west Africa), Japanese (JPT), Han Chinese (CHB), and CEPH (CEU, Utah residents with ancestry from northern and western Europe) individuals

Principal components analysis



- Figure reproduced from Miclaus, K.J., Wolfinger, R., and Czika, W. (2009). SNP selection and multidimensional scaling to quantify population structure. *Genetic Epidemiology*, 33, 488-496.

Principal components analysis

- What is it about:
  - Dimension reduction / data compression / data visualization ...
  - Find a few (linear) combinations of the variables to "best summarize" those variables
  - Represent data in a low dimensional space, and preserve data variability by exploiting the **covariance** structure of the set of variables

- Topics to cover:
  - Population solution, sample version (Sections 8.2, 8.3)
  - Extensions: sparse pca, principal components regression

- What to pay special attention:
  - Applications!

Factor analysis

- Motivating example: consumer-preference study
  - In a consumer-preference study, a random sample of customers were asked to rate 5 attributes of a new product
  - The response is on a 7-point semantic differential scale, and the correlation matrix is

	T	G	F	S	P
Taste	1.00	0.02	0.96	0.42	0.01
Good buy for money	0.02	1.00	0.13	0.71	0.85
Flavor	0.96	0.13	1.00	0.50	0.11
Suitable for snack	0.42	0.71	0.50	1.00	0.79
Provides lots of energy	0.01	0.85	0.11	0.79	1.00

  - Question of interest: any underlying patterns / grouping of those attributes?

Factor analysis

- What is it about:
  - Find **unobservable** latent variables, called **factors**, which are responsible for groups of strongly correlated variables in the data
  - Widely used in the quality of life (QoL) studies. Among many questions in a questionnaire, it is of common interest to learn how they are grouped as well as characteristics of each group. Characteristics of each group may be represented by a (latent) factor, which reflects, say, physical ability or mental health
- Topics to cover:
  - Model, estimation, rotation, factor scores (Sections 9.2–9.5)
  - Extension: latent regression models
- What to pay special attention:
  - Identifiability, then interpretation

Discriminant analysis and classification

- Motivating example: Cleveland heart disease data
  - 303 patients, a diagnosis of heart disease (0=absence, 1=presence, 160 absence and 137 presence)
  - 13 attributes, including age, gender, chest pain type (1-4), resting blood pressure, serum cholesterol, fasting blood sugar (1=true,0=false), resting electrocardiographic results, maximum heart rate achieved, exercise induced angina (1=yes; 0=no), ...
- Many classification examples:
  - Classify tumor samples as benign or malignant
  - Classify patients as having Alzheimer's disease or general population
  - . . .

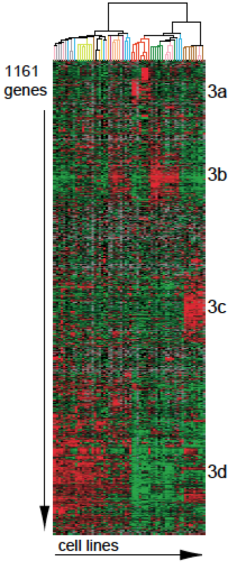
Discriminant analysis and classification

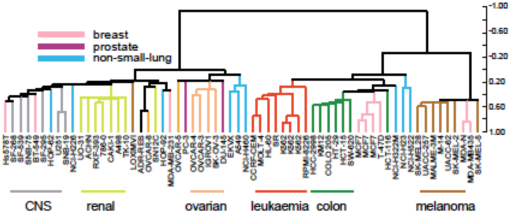
- What is it about:
  - Separation and allocation: to describe graphically or algebraically the **differential features** (e.g. biomarkers, patient's demographics etc) of data from several **known** populations (e.g. progressive and non-progressive); to develop a rule to allocate data cases (e.g. patients) into two or more **known** classes.
  - **Supervised learning**

- Topics to cover:
  - How to evaluate a classifier?
  - Two groups
    - Linear, quadratic and mixture discriminant analysis (Section 11.3)
    - Logistic regression revisited (Section 11.7)
  - More than two groups (Section 11.5)
  - Extensions: k-nearest-neighbor (kNN), classification and regression tree (CART), support vector machines (SVM)

- What to pay special attention:
  - Get the key ideas, then just try and see which performs the best!

Clustering





- Motivating example: NCI60
  - 60 cell lines derived from human tumors + gene expressions of about 8,000 genes
  - Cell lines from leukaemia, melanoma, central nervous system, colon, renal and ovarian tissue were clustered into branches specific to the respective organ types with few exceptions
  - Cell lines derived non-small lung carcinoma and breast tumors were distributed in different terminal branches suggesting that their gene expression patterns were more heterogeneous
  - Figure reproduced from Ross et al. (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*, 24, 227-235

Clustering

- What is it about:
  - Discover natural, **unknown** grouping patterns in data
  - **Unsupervised learning**
- Topics to cover:
  - Similarity / distance measures (Sections 12.1 and 12.2) – Define what you feel sound and legitimate
  - Clustering methods
    - Hierarchical clustering (Section 12.3)
    - K-means (Section 12.4)
    - Model-based clustering (Section 12.5)
  - Extensions: multidimensional scaling
- What to pay special attention:
  - Robustness
  - Exploratory nature: starting points for future research

A super example

- Body fat example:
  - Body fat, a measure of health, is estimated through an underwater weighing technique. Fitting body fat to the other measurements using multiple regression provides a convenient way of estimating body fat for men using only a scale and a measuring tape.
  - Percentage of body fat, age, weight, height, and ten body circumference measurements are recorded for 252 men.
    - Percent body fat using Brozek's equation, 457/Density - 414.2
    - Percent body fat using Siri's equation, 495/Density - 450
    - Density (gm/cm^3); Age (yrs); Weight (lbs); Height (inches); Adiposity index = Weight/Height^2 (kg/m^2); Fat Free Weight = (1 - fraction of body fat) * Weight, using Brozek's formula (lbs)
    - Circumference (cm): Neck; Chest; Abdomen; Hip; Thigh; Knee; Ankle; Extended biceps; Forearm; Wrist.
  - Dichotomized body fat groups: Obese ($\geq$ 25%), Normal ($<$ 25%).