Multivariate Statistical Analysis (Math347)

Instructor: Kani Chen

An Overview

Math347 vs. Math341:

- Blue collar vs. white collar.

Math347 vs. Math341:

- Blue collar vs. white collar.
- Construction work vs. architecture.

Math347 vs. Math341:

- Blue collar vs. white collar.
- Construction work vs. architecture.
- Action vs. Drama.

Math347 vs. Math341:

- Blue collar vs. white collar.
- Construction work vs. architecture.
- Action vs. Drama.
- Messy vs. clean.

Math347 vs. Math341:

- Blue collar vs. white collar.
- Construction work vs. architecture.
- Action vs. Drama.
- Messy vs. clean.
- Expect lots of sweat throughout this course.

Structure of the course:

- Part 1: Conventional streamline of elementary statistics:
  about means;
  linear regression.

Structure of the course:

- Part 1: Conventional streamline of elementary statistics:
  about means;
  linear regression.
- Part 2: (Special) methodologies.

Part I. Conventional problems.

One sample problem:

- $$X_1, ..., X_n \quad iid \quad \sim MN(\mu, \Sigma),$$

  where X_i and μ are p-dimension, and Σ is $p \times p$.
  Concerning with μ. The objective is to estimate μ with accuracy justification.

Part I. Conventional problems.

One sample problem:

- $$X_1, ..., X_n \quad iid \quad \sim MN(\mu, \Sigma),$$
  where X_i and μ are p-dimension, and Σ is $p \times p$.
  Concerning with μ. The objective is to estimate μ with accuracy justification.
- X_i is the (IQ, EQ) of the i-th randomly selected student from our university. $p = 2$.

Paired sample problem:

- (still one sample problem):

$$\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix} \dots \begin{pmatrix} X_n \\ Y_n \end{pmatrix} \quad iid \quad \sim MN(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma),$$

  where X_i, Y_i, μ_1 and μ_2 are p-dimension, and Σ is $(2p) \times (2p)$ matrix.
  Concerning with the difference of μ_1 and μ_2, and objective: Estimating $\mu_1 - \mu_2$ with accuracy justification.

Paired sample problem:

- (still one sample problem):

$$\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix} \dots \begin{pmatrix} X_n \\ Y_n \end{pmatrix} \quad iid \quad \sim MN(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma),$$

  where X_i, Y_i, μ_1 and μ_2 are p-dimension, and Σ is $(2p) \times (2p)$ matrix.
  Concerning with the difference of μ_1 and μ_2, and objective: Estimating $\mu_1 - \mu_2$ with accuracy justification.
- (X_i, Y_i) are the (IQ, EQ) of the i-th randomly selected student from our university before and after a training program. ($p = 2$.) Does the training program make a difference?

Repeated measurement.

- (Exactly one sample, but with a special care.)

$$X_1, ..., X_n \quad iid \quad \sim MN(\mu, \Sigma),$$

Concerning the differences of the components of μ.

Repeated measurement.

- (Exactly one sample, but with a special care.)

$$X_1, ..., X_n \quad iid \quad \sim MN(\mu, \Sigma),$$

  Concerning the differences of the components of μ.
- Example: X_i is the returns of year 2006, 2007, 2008 of the i-th randomly selected stock in HKEX. ($p = 3$). Is there a difference between the three years in terms of stock returns?
  The components are of the same nature and there are numerically comparable.

Two sample problem:

- $$X_1, ..., X_n \quad iid \quad \sim MN(\mu_1, \Sigma_1),$$
  $$Y_1, ..., Y_m \quad iid \quad \sim MN(\mu_2, \Sigma_2),$$

  and $\{X_i\}$ are independent of $\{Y_j\}$.
  Concerning the difference of μ_1 and μ_2, and the objective is to estimate $\mu_1 - \mu_2$ with accuracy justification.

Two sample problem:

- $$X_1, ..., X_n \quad iid \quad \sim MN(\mu_1, \Sigma_1),$$
  $$Y_1, ..., Y_m \quad iid \quad \sim MN(\mu_2, \Sigma_2),$$
  and $\{X_i\}$ are independent of $\{Y_j\}$.
  Concerning the difference of μ_1 and μ_2, and the objective is to estimate $\mu_1 - \mu_2$ with accuracy justification.
- Example: X_i is the monthly $(income, spending)$ of the i-th randomly selected household in Hong Kong. Y_i is that in Shanghai. ($p = 2$). Any difference between HK and SH in terms income and spending?
  Special interest: $\Sigma_1 = \Sigma_2$.

Several/Multiple sample problem:

- $$X_{11}, ..., X_{1n_1} \quad iid \quad \sim MN(\mu_1, \Sigma),$$
  $$\cdots$$
  $$X_{g1}, ..., X_{gn_g} \quad iid \quad \sim MN(\mu_g, \Sigma),$$
  Concerning the differences between $\mu_1, \mu_2, ..., \mu_g$, and the objective answer the question whether they any different with accuracy justification.

Several/Multiple sample problem:

- $$X_{11}, ..., X_{1n_1} \quad iid \quad \sim MN(\mu_1, \Sigma),$$
  $$\cdots$$
  $$X_{g1}, ..., X_{gn_g} \quad iid \quad \sim MN(\mu_g, \Sigma),$$
  Concerning the differences between $\mu_1, \mu_2, ..., \mu_g$, and the objective answer the question whether they any different with accuracy justification.
- Example: monthly $(income, spending)$ for randomly selected households in HK, SH and BJ. ($p = 2$ and $g = 3$).
  Any difference between HK, SH and BJ in terms of income and spending?
  MANOVA

Regression

- Ordinary/univariate linear regression:

$$Y = \beta' X + \epsilon.$$

where Y is of 1 dimension (*uni*);
ϵ: 1 dimension;
β and X are $r + 1$ dimension.

Regression

- Ordinary/univariate linear regression:

$$Y = \beta' X + \epsilon.$$

where Y is of 1 dimension (*uni*);
ϵ: 1 dimension;
β and X are $r + 1$ dimension.

- Example: Y_i is math exam score and X_i is (1, time spent) on studying for the i-th randomly selected student.
Here 1 is for the intercept component.

Regression

- Multivariate linear regression:

$$Y = \beta' X + \epsilon.$$

where Y as well as ϵ are of m dimension (*uni*);
β: $(r+1) \times m$ matrix;
m univariate linear models putting together.
Connected by the dependence of the components of ϵ.

Regression

- Multivariate linear regression:

$$Y = \beta' X + \epsilon.$$

where Y as well as ϵ are of m dimension (*uni*);
β: $(r+1) \times m$ matrix;
m univariate linear models putting together.
Connected by the dependence of the components of ϵ.

- Example: Y_i is $(math, physics)$ exam scores and X_i same as above.

Part II:

1. Principle component analysis and factor analysis.

- Linearly combine the variables in study to produce the most influential (hidden) variables.

Part II:

1. Principle component analysis and factor analysis.

- Linearly combine the variables in study to produce the most influential (hidden) variables.
- Factor analysis, a refined method.

Part II:

1. Principle component analysis and factor analysis.

- Linearly combine the variables in study to produce the most influential (hidden) variables.
- Factor analysis, a refined method.
- Example: We have the exam scores of history, English, Chinese, math, physics and chemistry for a number of randomly selected students. What are the most important factors that contribute to the exam performance?

Part II:

1. Principle component analysis and factor analysis.

- Linearly combine the variables in study to produce the most influential (hidden) variables.
- Factor analysis, a refined method.
- Example: We have the exam scores of history, English, Chinese, math, physics and chemistry for a number of randomly selected students. What are the most important factors that contribute to the exam performance?
- Example: With the observations of the daily returns of PetroChina, Sinopec, CNOOC, Tencent, Kingsoft and Alibaba in the past year, can we find out what are the underlying driving forces that affect the stock prices?

2. Canonical Analysis.

- Find out and sort out the connection between two sets of variables using linear combinations.

2. Canonical Analysis.

- Find out and sort out the connection between two sets of variables using linear combinations.
- Example: With the observations of the daily returns of PetroChina, Sinopec, CNOOC, Tencent, Kingsoft and Alibaba in the past year, can we find out and explain the relation, if any, between the stock performances the oil industry and the internet industry?

3. Classification/Separation/Discrimination.

- Given multivariate observations from two populations.

3. Classification/Separation/Discrimination.

- Given multivariate observations from two populations.
- Come up with a proper criterion to classify a new observation into one of the two populations, in other words, to determine properly which population this observation comes from.

3. Classification/Separation/Discrimination.

- Given multivariate observations from two populations.
- Come up with a proper criterion to classify a new observation into one of the two populations, in other words, to determine properly which population this observation comes from.
- Hypothetical example: initial judgment of whether patient is infected with, say, H1N1 is based on fever (degree of temperature) and number of white cells.

3. Classification/Separation/Discrimination.

- Given multivariate observations from two populations.
- Come up with a proper criterion to classify a new observation into one of the two populations, in other words, to determine properly which population this observation comes from.
- Hypothetical example: initial judgment of whether patient is infected with, say, H1N1 is based on fever (degree of temperature) and number of white cells.
- Minimize mistaken classifications.
  Minimize the cost of misclassifications.