

Overview

General topics for data analysis	2
Main topics in multivariate statistics.	3
Main topics in multivariate statistics.	4
Exploratory methods	5
Graphics for multivariate data	6
Principal component analysis (PCA)	7
Possible uses of PCA	8
Possible uses of PCA	9
Factor analysis: idea	10
Factor analysis: model	11
Factor analysis	12
Linear discriminant analysis.	13
Linear discriminant analysis.	14
Linear discriminant analysis.	15
Cluster analysis.	16
Multidimensional scaling.	17
More formal methods	18
Normal distribution theory	19
Tests of significance for multivariate data	20
Remaining topics we did not cover.	21

General topics for data analysis

- Discussed in class:
 - ◆ Study design
 - ◆ Sampling methods
- Not discussed in this class:
 - ◆ Missing data
 - ◆ Outliers

2 / 21

Main topics in multivariate statistics

- We have data on several variables, there is some interdependence between the variables, and none of them is clearly the main variable of interest
- Methods that are mostly of exploratory nature:
 - ◆ Graphics for multivariate data
 - ◆ Principal component analysis (PCA)
 - ◆ Factor analysis
 - ◆ Linear discriminant analysis (LDA)
 - ◆ Cluster analysis
 - ◆ Multidimensional scaling
 - ◆ ...

3 / 21

Main topics in multivariate statistics

- More 'formal' topics:
 - ◆ Normal distribution theory
 - ◆ Tests of significance for multivariate data
 - ◆ Multivariate analysis of variance (MANOVA)
 - ◆ Multivariate regression analysis
 - ◆ Canonical correlation analysis
 - ◆ ...

4 / 21

Graphics for multivariate data

- Goal: visualize multivariate data
- We covered:
 - ◆ Scatterplot matrix: `pairs()`
 - ◆ Star plots and segment plots: `stars()`
 - ◆ Conditioning plots: `coplot()`
 - ◆ Bi-plot of first two principal components: `biplot()`
- Other techniques:
 - ◆ Interactive 3 dimensional plots
 - ◆ Plots based on multidimensional scaling (more about this later)
 - ◆ ...

6 / 21

Principal component analysis (PCA)

- Main idea:
 - ◆ Start with variables X_1, \dots, X_p
 - ◆ Find a *rotation* of these variables, say Y_1, \dots, Y_p (called principal components), so that:
 - Y_1, \dots, Y_p are uncorrelated. Idea: they measure different dimensions of the data.
 - $\text{Var}(Y_1) \geq \text{Var}(Y_2) \geq \dots \text{Var}(Y_p)$. Idea: Y_1 is most important, then Y_2 , etc.
- Method is based on spectral decomposition of the covariance matrix
- No need to make distributional assumptions

7 / 21

Possible uses of PCA

- Interest in first principal component:
 - ◆ Example: How to combine the scores on 5 different examinations to a total score? Since the first principal component maximizes the variance, it spreads out the scores as much as possible.
- Interest in 2nd - pth principal components:
 - ◆ When all measurements are positively correlated, the first principal component is often some kind of average of the measurements (e.g., size of birds, severity index of psychiatric symptoms).
 - ◆ Then the other principal components give important information about the remaining pattern (e.g., shape of birds, pattern of psychiatric symptoms)

8 / 21

Factor analysis

- Assumptions:
 - ◆ $E(x) = 0$ (if this is not the case, simply subtract the mean vector)
 - ◆ $E(f) = 0$, $\text{Cov}(f) = I$
 - ◆ $E(u) = 0$, $\text{Cov}(u_i, u_j) = 0$ for $i \neq j$
 - ◆ $\text{Cov}(f, u) = 0$
- Estimation:
 - ◆ Under the above assumptions, $\text{Cov}(x) = \Sigma = \Lambda\Lambda' + \Psi$
 - ◆ Two estimation methods: principal factor analysis and maximum likelihood
- Factor loadings are non-unique; factor rotation can be used to ease interpretation

12 / 21

Linear discriminant analysis

- Goal: Suppose that we have an $n \times p$ data matrix consisting of g different groups. How can we classify new observations into one of these groups? This is sometimes called 'supervised learning'
- Fisher:
 - ◆ Look for the linear function Xa which maximizes the ratio of the between-groups sum of squares to the within-groups sum of squares.
 - ◆ Compute average score $(\bar{x}_i)'a$ for each group $i = 1, \dots, g$.
 - ◆ Compute the score $x_{new}a$ for the new observation.
 - ◆ Classify the new observation in group j if $|x_{new}a - (\bar{x}_j)'a| < |x_{new}a - (\bar{x}_i)'a|$ for all $i \neq j$.

13 / 21

Linear discriminant analysis

- Maximum likelihood:
 - ◆ Suppose the exact distributions of the populations Π_1, \dots, Π_g are known.
 - ◆ Then the maximum likelihood discriminant rule for allocating a new observation is to allocate it to the population which gives the largest likelihood to x , i.e., to the population with the highest density at the point x . See picture on overhead.
 - ◆ If the exact distributions are unknown, but we know for example that the populations are all multivariate normal, then we can first estimate their parameters, and then use the above rule. This is the sample maximum likelihood discriminant rule.

14 / 21

Linear discriminant analysis

- If we have two groups from two multivariate normal distributions with the same covariance matrix, then Fisher's linear discriminant analysis corresponds exactly to the maximum likelihood rule for classification.

15 / 21

Cluster analysis

- We have multivariate data without group labels.
- We want to see if there are clusters in the data, i.e., groups of observations that are homogeneous and separated from the other groups. This is sometimes called 'unsupervised learning'.
- Methods we discussed:
 - ◆ Hierarchical clustering
 - ◆ k -means clustering
 - ◆ Model based clustering
- Possible applications:
 - ◆ Marketing: find groups of customers with similar behavior
 - ◆ Biology: classify plants or animals
 - ◆ Internet: cluster text documents

16 / 21

Multidimensional scaling

- Not discussed in class
- Goal: Construct a 'map' from a distance matrix, where the map should represent the distances between the objects as accurate as possible.
- Possible applications:
 - ◆ Psychology/sociology: subjects say how similar/different pairs of objects are. Multidimensional scaling then creates a pictures showing the overall relationships between the subjects.
- Can be used to aid clustering
- See overhead slides and R-code

17 / 21

More formal methods

18 / 21

Normal distribution theory

- Multivariate normal distribution
- Wishart distribution (for sample covariance matrix)
- Hotelling's T^2 distribution (for Mahalanobis distance, closely related to F -distribution)

19 / 21

Tests of significance for multivariate data

- Discussed in class:
 - ◆ Comparison of mean values for two samples, when covariance matrices are assumed to be identical: multivariate T^2 -test
- Other tests:
 - ◆ Comparison of mean values for several samples
 - ◆ Comparison of mean values for several samples when covariance matrices are not the same
 - ◆ Comparison of variation for two samples
 - ◆ Comparison of variation for several samples

20 / 21

Remaining topics we did not cover

- MANOVA: multivariate version of ANOVA (analysis of variance)
- Multivariate regression: multivariate version of multiple regression (when doing least squares, estimates are the same as when doing multiple regression for each dependent variable separately)
- Canonical correlation analysis: variables (not observations) are divided into groups, and we want to find out the relationship between these groups

21 / 21