Multivariate Statistical Analysis

Nusrat Jahan
Mathematics & Statistics
James Madison University

Multivariate Statistical Analysis

- "Multivariate data" means multiple outcome variables measured on the same individual or object.
- "Multivariate analysis" is the analysis where multiple variables are statistically analyzed simultaneously.
- Multivariate analysis is needed to get an overall picture.

Example of Multivariate data

Vocational interests of 231 undergraduate students. The 22 interest areas are:

- X_1: Public speaking
- X_2: Law and politics
- X_3: Business management
- X_4: Sales
- X_5: Merchandising
- X_6: Office practice
- X_7: Military activities
- X_8: Technical supervision
- X_9: Mathematics
- X_{10}: Science
- X_{11}: Mechanical

X_{12}: Nature
X_{13}: Agriculture
X_{14}: Adventure
X_{15}: Recreational leadership
X_{16}: Medical service
X_{17}: Social service
X_{18}: Religious activities
X_{19}: Teaching
X_{20}: Music
X_{21}: Art
X_{22}: Writing

Why, When, and How?

- Why use multivariate methods?
  - ❑ Simultaneous analysis of multiple response variables
- When to use multivariate methods?
  - ❑ Response variables are related, an overall picture needs input from most of the variables
- How to use multivariate methods?
  - ❑ Computer packages: SAS, SPSS, R, etc.

Goals of multivariate analysis:

- Description
- Summarization
- Simplification
- Grouping – both cases and variables
- Relationships – among variables, among cases
- To detect strange or unusual aspects of the data
- To create new variables (dimension reduction)
- To suggest additional analyses

Multivariate Procedures

Variable directed

- ❑ Principal Component Analysis
- ❑ Factor Analysis
- ❑ Canonical Correlation Analysis

Individual directed

Discriminant Analysis

Cluster Analysis

Hotelling's T^2 Tests

Multivariate Analysis of Variance (MANOVA)

Variable Directed Methods: Analysis of Variables

Principal component analysis
Factor analysis
Canonical correlation analysis*

- Creates a new set of synthetic variables (factors) that can summarize the original set
- Mathematics is used to create these factors so that only a few can capture the maximum variance of the original variables
- How many 'new' variables to consider in the analysis – a trade off between % of variance explained and complexity of analysis
- Multiple dependent variables are predicted by independent variables*

Principal Component Analysis (PCA) of Vocation data

- Factor1: Business management, Merchandizing, Sales, Office practice, Technical supervision.

  Label: *Business?*

- Factor2: Music, Art, Writing.

  Label: *Creativity?*

- Factor3: Social service, Religious activities, service

  Label: *Service?*

Individual Directed Methods:
Groupings of Individuals, Objects

Discriminant analysis
Cluster analysis
Hotelling's T^2 tests
MANOVA

- ❑ Separates distinct sets of objects, allocates new objects to pre-defined groups
- ❑ Discriminatory procedure, classification rules are used for separation & allocation
- ❑ Objects within a group are similar, between groups they are dissimilar
- ❑ Misclassification could be a potential problem

Example: Fisher’s Iris data

- 150 sets of measurements are available on 3 varieties of *Iris*:
  - ❑ Variety 1: *Iris setosa*
  - ❑ Variety 2: *Iris versicolor*
  - ❑ Variety 3: *Iris virginica*

Number of variables: 4

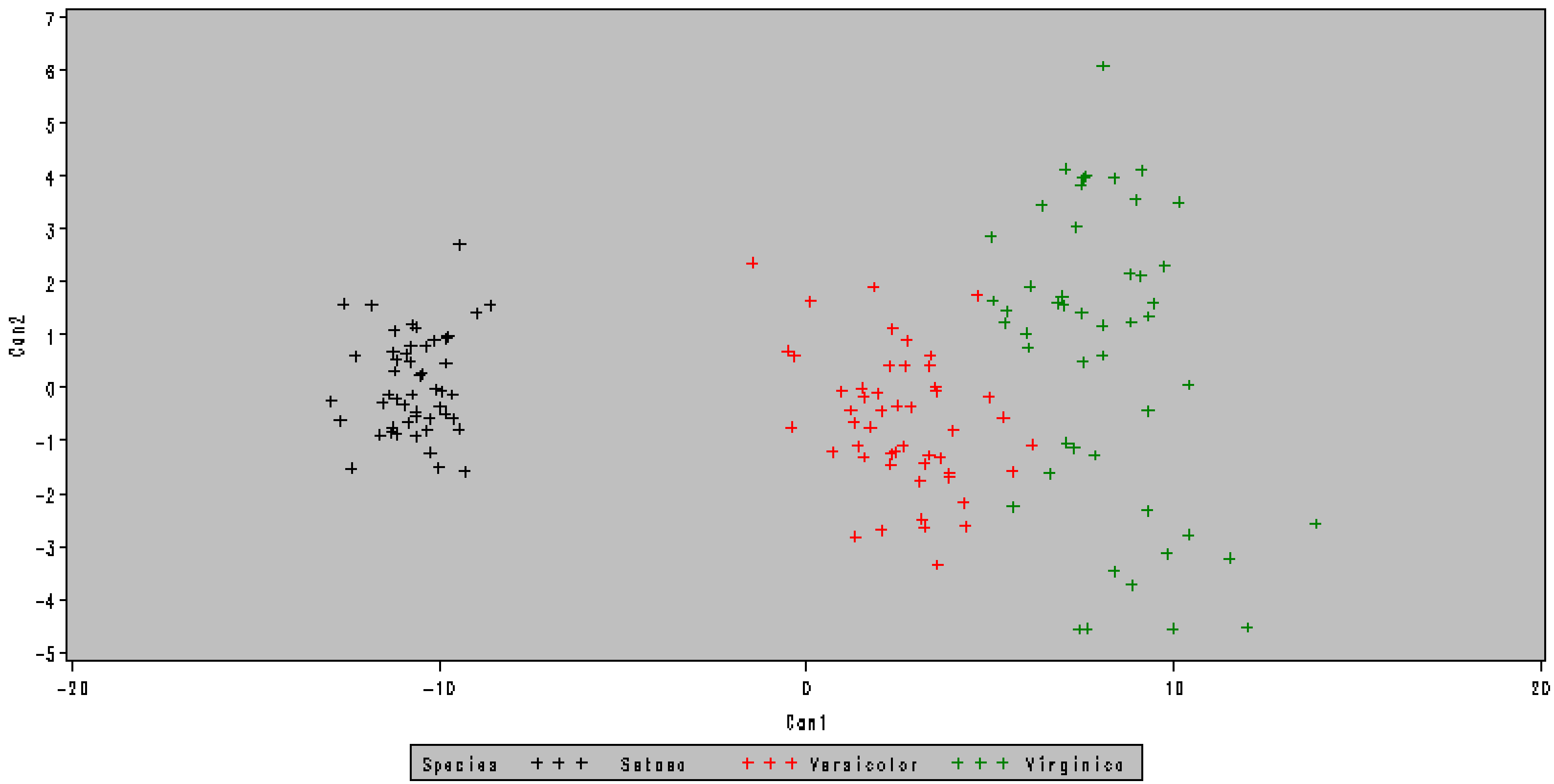
Number of objects: 150

Goal: Can we classify these 150 obs. Into 3 major groups

Discrimination/classification rules based on :

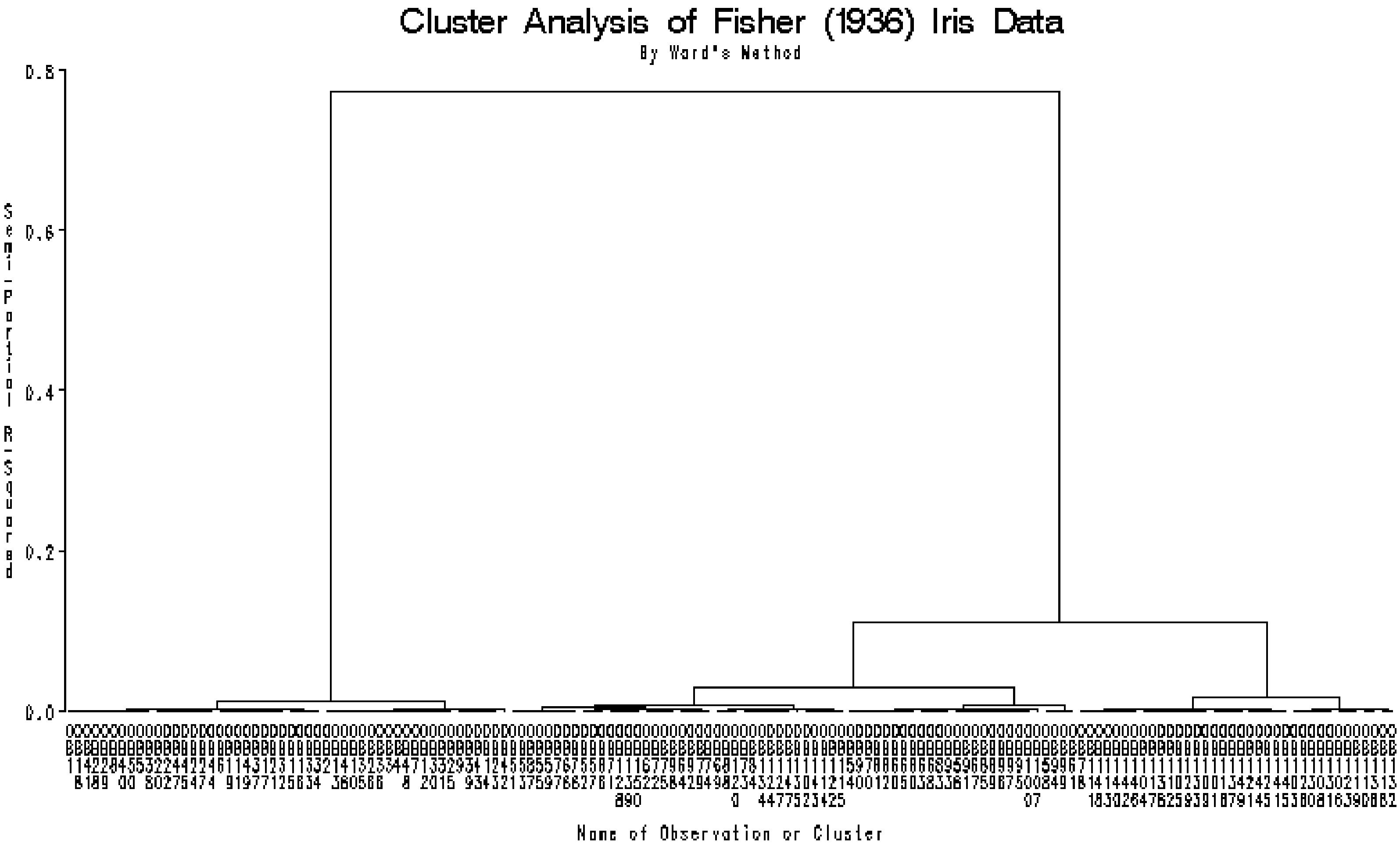
Sepal length, Sepal width, Petal length, Petal width

Fisher (1936) Iris Data



Four Types of Research Questions:

- Degree of relationship between the variables
  - *Bivariate correlation & regression (multiple, & multivariate/canonical)*
- Measure significant differences between group means
  - *Hotelling's T^2 tests, MANOVA*
- Predicting membership into groups
  -
- Explaining underlying structure


Cluster Analysis of Fisher (1936) Iris Data
By Ward's Method
Semi-Partial R-Squared
0.0
0.2
0.4
0.6
0.8
Name of Observation or Cluster

After Multivariate Analysis …

- Multivariate analysis is an exploratory analysis:
  - ❑ suggests additional analysis – regression, ANOVA etc. of the factors
  - ❑ generates hypothesis regarding factors
  - ❑

Most research articles report multivariate analysis result very briefly, mainly reports the univariate results obtained on the factors.

Principal Component Analysis of the Vocation data

How many new variables to consider in a PCA?

Eigenvalues of the Correlation Matrix:

Total = 22 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	5.59637727	2.11784592	0.2544	0.2544
2	3.47853135	0.85459327	0.1581	0.4125
3	2.62393808	0.74258975	0.1193	0.5318
4	1.88134833	0.59339867	0.0855	0.6173
5	1.28794966	0.07117677	0.0585	0.6758
6	1.21677289	0.27185782	0.0553	0.7311
7	0.94491507	0.28304545	0.043	0.7741
	...	...	...	...

PCA SAS output of Vocation data

Rotated Factor Pattern (Method = Varimax)

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
x1	0.49593	0.26395	0.34243	-0.16506	-0.05825	0.57486
x2	0.4263	0.02462	0.31505	-0.11918	-0.22047	0.63547
x3	0.92687	-0.03452	0.07002	0.02739	0.00679	0.15176
x4	0.82476	0.07037	0.13096	-0.02887	0.03365	0.18179
x5	0.93212	0.052	0.06303	-0.0098	0.03826	0.15286
x6	0.82525	-0.18738	0.18875	0.13444	-0.05584	-0.13369
x7	0.27137	-0.45726	0.29634	0.19806	0.31886	0.07984
x8	0.82355	-0.12223	-0.00732	0.16405	0.12326	0.02939
x9	0.0963	-0.04386	-0.16168	0.74606	-0.16586	0.01498
x10	-0.09477	0.1135	0.13732	0.86711	0.11228	-0.02805
x11	0.26124	-0.00669	-0.02198	0.76353	0.29515	-0.02148
x12	-0.00083	0.28163	0.30364	0.20299	0.79111	-0.01815
x13	0.03647	-0.01018	0.03488	-0.03942	0.91429	0.12708
x14	0.04897	-0.05288	-0.21468	0.20696	0.33933	0.72413
x15	0.13505	-0.48547	0.2191	-0.03235	0.33238	0.44523
x16	-0.03802	0.08915	0.46815	0.45384	-0.01023	0.3177
x17	0.03511	0.12905	0.79281	-0.15898	0.00646	0.18413
x18	0.16134	0.11188	0.69522	0.05285	0.19387	-0.0634
x19	0.28981	0.28494	0.64497	0.07068	0.1356	-0.0486
x20	0.00316	0.82271	0.29615	0.15561	0.02057	-0.05548
x21	-0.04609	0.83575	0.18874	0.13052	0.24476	0.01156
x22	0.02158	0.68814	0.35576	-0.21847	0.09917	0.29232