

Multivariate Data Analysis

Dr Nick Fieller

Department of Probability & Statistics

University of Sheffield

MAS6011/MAS465

2011/2012



Contents

Preliminaries

0: Introduction

1: Graphical Displays

2: Reduction of Dimensionality

3: Multidimensional Scaling Techniques

4: Discriminant Analysis

5: Multivariate Regression Analysis

6: Canonical Correlation Analysis

7: Partial Least Squares

8: Statistical Analysis of Multivariate Data

9: Statistical Discriminant Analysis

© NRJF, University of Sheffield, 2011/12, Semester 1

Multivariate Data Analysis

2

Contents

Preliminaries

0: Introduction

1: Graphical Displays

2: Reduction of Dimensionality

3: Multidimensional Scaling Techniques

4: Discriminant Analysis

5: Multivariate Regression Analysis

6: Canonical Correlation Analysis

7: Partial Least Squares

8: Statistical Analysis of Multivariate Data

9: Statistical Discriminant Analysis

© NRJF, University of Sheffield, 2011/12, Semester 1

Multivariate Data Analysis

3

Books used to prepare course

◆ Gnanadesikan, R. (1997)

Methods for Statistical Data Analysis of Multivariate Observations. Wiley

◆ Mardia, K., Kent, J. & Bibby, J. (1981)

Multivariate Analysis. Wiley

• Venables, W. N. & Ripley, B. D. (2002)

Modern Applied Statistics with S-PLUS, (4th Edition). Springer. Support available from

<http://www.stats.ox.ac.uk/pub/MASS3>

© NRJF, University of Sheffield, 2011/12, Semester 1

Multivariate Data Analysis

4

Books to buy

◆ I ADVISE AGAINST BUYING ANY BOOKS

◆ The books by Gnanadesikan and Mardia *et al* are extremely expensive and contain far too much detail. **Do not buy**

- The course notes are intended to contain all that you need to know

◆ Venables & Ripley is useful for many areas of Applied Statistics

- Those by Trevor Cox, Bryan Manly and Brian Everitt are at a similar level to this course

© NRJF, University of Sheffield, 2011/12, Semester 1

Multivariate Data Analysis

5

Objectives

- ◆ Provide a guide to the **practical** & theoretical analysis of multivariate data
 - MV-data:- measurements made on each of several variables on each experimental unit
- ◆ Includes extensions of univariate methods & new problems that only exist in multidimensions
- ◆ Some emphasis on R (& S-plus & Minitab) implementation of techniques

© NRJF, University of Sheffield, 2011/12, Semester 1

Multivariate Data Analysis

6



Organization of course material

- ◆ Chapters 1 – 9 main part of course
 - Based on Gnanadesikan and Mardia et al.
 - (note chapters 5*– 7* are 'starred')
- ◆ Appendix 0: background maths
 - Lagrange multipliers, eigenvalues etc
 - Guide to matrix algebra on my teaching webpage
- ◆ Appendices 1 – 8 extra material to use in other project courses
 - Based on Venables & Ripley
- ◆ Exercises & Task Sheets are in Course Booklet
 - Solutions follow later as appropriate

Task Sheets & Exercises

- ◆ Task sheets:–
 - ~ each week
 - simple quick short exercises / reading
 - reinforce / consolidate lecture material
- ◆ Exercises:–
 - 3 sets during semester in weeks 3,6,8
 - work submitted within 2 weeks will be marked and returned
- ◆ See Study Guide
 - **recommendations on time to spend**

Task Sheets & Exercises

- ◆ Task sheets:–
 - are designed for you to test **your own understanding** of the course material
 - *you are responsible* for your own learning on the course — task sheets help you in self-assessment
- ◆ Exercises:–
 - **Prime route for individual feedback**
 - Task sheets often provide guide for exercises
- ◆ Unacceptable reasons for not submitting anything
 - I did not have enough time
 - I knew I could do them so I did not need to submit
 - I could not do anything so did not think it was worth it

Solutions to Task Sheets & Exercises

- ◆ Exercises:–
 - Solutions available on web soon after submission
 - Printed solutions will be provided to those who submit
- ◆ Task sheets:–
 - are designed for you to test your own understanding of the course material
 - if necessary go back to lecture notes (etc) & re-read relevant sections
 - (and if necessary re-read again &)
 - Solutions will be provided on web pages in due course (for revision etc)
 - but *deliberately* these will not appear very quickly

Course web page

<http://nickfieller.staff.shef.ac.uk/>

- Click on  & then on

[MAS6011/MAS465 Multivariate Data Analysis](#)

- ◆ Lecture notes, task sheets, solutions & data sets available here (or on MOLE) after distribution in lectures
 - (I don't keep back copies)

Purpose of Lectures

- ◆ There are 'complete' printed notes
 - ◆ These are **not** a textbook
 - some explanations are omitted
 - ◆ They are intended to allow you to concentrate on understanding & for me to cover some material very quickly
- ◆ Some lectures will be very close to the printed notes
 - ◆ This is **intended**
- ◆ Other lectures will fill in details & provide examples & R demos



Computing

- Computer package for the course is **R**
 - ◆ See course web pages for introduction to **R**
 - ◆ See [CRAN home page](#) for more information
 - ◆ See [Basics of Matrix Algebra with R](#) for using **R** for matrix calculations
- Some illustrative examples may be in S-Plus or Minitab &c.
 - ◆ **You need to be familiar with R output for the examinations**
 - The best way to become familiar is to attempt the computer tasks & exercises

Contents

Preliminaries

0: Introduction

- 1: Graphical Displays
- 2: Reduction of Dimensionality
- 3: Multidimensional Scaling Techniques
- 4: Discriminant Analysis
- 5: Multivariate Regression Analysis
- 6: Canonical Correlation Analysis
- 7: Partial Least Squares
- 8: Statistical Analysis of Multivariate Data
- 9: Statistical Discriminant Analysis

(My) Uses of Multivariate Analysis

- **Facial Identification**
 - ◆ Can we match two faces?
 - (statistically)?
- **High Throughput Screening**
 - ◆ ~1,000,000 compounds as Candidate Drugs with High Content Biology Measurements
- **Gene Expression Data from Microarrays**
 - ◆ Which genes are active in cancer samples?
- **Proteomic & Metabonomic Measurements**
 - ◆ Which proteins distinguish MRSA ('Superbug') from SA
 - Can they be used to screen for MRSA carriers?
- **Financial Statistics**
 - (Credit card transactions, customer segmentation)
- **Archaeostatistics**
 - ◆ See later

Facial Identification

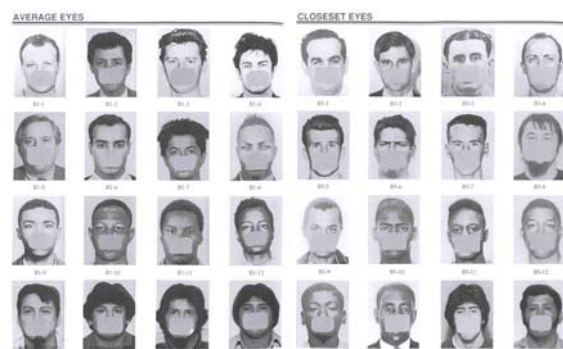
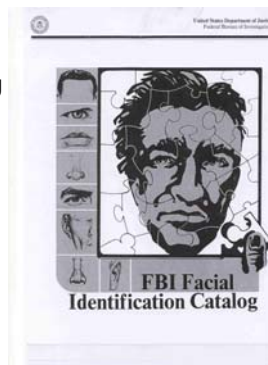


sponsored by FBI

- **Objectives**
 - ◆ To provide **quantifiable** measures of quality of match of faces
 - ◆ Use as **evidence** in a Court of Law
 - e.g. Face captured on CCTV camera
 - Suspect is arrested and photographed
 - What is **probability** of the two pictures of faces being of the **same** person?

Pilot Study

- ◆ Based on measuring faces from the FBI facial identification catalogue
- ◆ Faces scanned & landmarks captured manually



SQUARE CHIN THIN EYE BROWS

(it was not Al Capone)

Which face has appeared twice?

©NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 19

Traditional anthropometric landmarks

coordinates captured 'manually' with public domain software

©NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 20

Four subjects from the 2002 pilot study

Different sections of the face are censored so only use landmarks common to a pair of faces

©NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 21

Cluster analysis of Procrustes coordinates of mean shapes of 48 images from 2 sections of catalogue: matches identified

©NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 22

AVERAGE EYES THIN EYE BROWS

©NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 23


3D scan taken from eight 2D digital photos

Lucy with Geomatrix© 3D Scanner


©NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 24





■ **The eight 2D views**



Obtain landmarks by triangulation from two 2D view



Generates 3D pictures in any orientation to match view from crime scene (based on internally constructed model)

© NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 25

■ **Delimar Vera Cuevas**

Mother finds kidnapped child six years on

A 10-day-old girl thought to have died in a fire in 1997 was actually kidnapped by a woman who started the blaze to cover her tracks, police said yesterday.

The girl's real mother saw the girl, now six, at a birthday party and recognised her as her own.

Delimar Vera was thought to have died in her family's home in Philadelphia, consumed by the heat and flames of a fire blamed on an extension cord for a heater. No body was ever found.

Captain John Darby of the Philadelphia police said the mother, Luz Cuevas, contacted the authorities after spotting the child in January. An investigation prompted DNA tests that confirmed her suspicion.

The mother "didn't know whether to cry, to yell or to scream", a police officer, Manuel Gonzalez, said. "She was just in total shock."

Police have issued a warrant for the arrest of Carolyn Correa, 41, of Willingboro, New Jersey, on charges of arson, kidnapping and conspiracy. Her whereabouts were unknown.

"This child, now six years old, who has been raised by Carolyn Correa as her own, is not her own," Capt Darby said.

Ms Cuevas told a local television station that she recognised the child from a dimple on her face.

"I said to my sister, look, she's my daughter," Ms Cuevas said.

It was unclear what brought the child and her mother to the same party.

State representative Angel Cruz, who helped Ms Cuevas to contact the police after she spotted the girl, credited "motherly instinct" for connecting the parent and child.

Ever since the blaze, Ms Cuevas had held on to the belief that her child was alive, partly because it did not make sense that a window in the infant's room was found to have been open even though it was the middle of December, Mr Cruz said.

The girl was placed in the custody of New Jersey division of youth and family services.

It was not clear when she would be reunited with her mother. AP, Philadelphia



Delimar Vera: recognised by mother at birthday party

Garden Walk 3/3/4

© NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 26

■ **The case of Delimar Vera:-**

- ◆ 1997:- 10-day old baby Delimar believed burnt in house fire
- ◆ 2004:- Mother recognises child at a birthday party in January
- ◆ 2004:- February; DNA tests on sample of hair taken by mother prove identity
- ◆ 2004:- March; Carolyn Correa, who had raised Delimar for six years, charged with arson + kidnapping + conspiracy

■ **Unanswered question:-**

- ◆ **Both at same party only a coincidence?**

© NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 27

A mother's instinct



December 1997 January 2004

© NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 28



© NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 29

A mother's instinct is impressive, — but see later



© NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 30



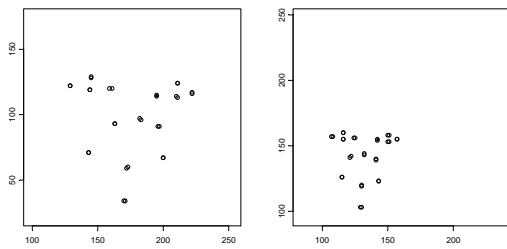
■ Back to Delimar Vera



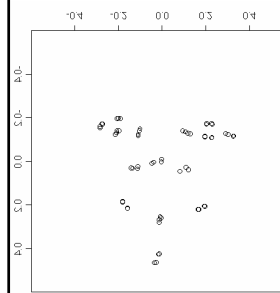
Delimar as a baby:
15 anthropometrical landmarks (2D) have been located on this image (2 repetitions of each).



Delimar at six years old:
The same fifteen landmarks (2D) have been located on this image (2 repetitions of each).



Plots of the raw landmark coordinates (left is the baby picture, right is the picture from six years old). Quite clearly the scales for the two images are not the same.



Plots of the Procrustes rotated coordinates from both images (baby & six years old)

.....a fairly close match?

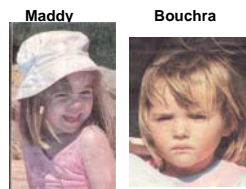
Assessment based on modelling coordinates as MV normal, estimated covariance from sample data. Clearly important to include gender/age group etc if possible.

Madeleine McCann: is this child in Morocco the missing Maddy?



Wednesday, 26th September 2007

By Thursday reporters had found the 'fake Madeleine' in Jebala, north Morocco.



Madeleine has a very 'normal face' so there are many children who look quite like her. She doesn't have an unusual face shape

Face to face

- Before it became clear that Bouchra Ben Aisa was the girl photographed by a Spanish tourist, a team in London had begun to analyse the image
- The Child Exploitation and Online Protection Centre uses software designed to identify children in pornographic images
- The system projects a map on to the image that picks out key features
- Experts said that the image purporting to show Madeleine was extremely poor quality and could not have yielded a conclusive match
- David McIntosh, of OmniPerception, a firm specialising in face recognition analysis, said: "This is a large picture but at its full size it has very few pixels. It is a very low-resolution image"
- Investigators were also hampered by Madeleine's lack of a distinctive face shape. Nick Fisher, of the University of Sheffield, said: "Despite the fact that everybody now recognises Madeleine's face it is actually a very normal face"



The Times, 27th September 2007



Examples (Cont^d)

- ◆ Digitization of a spectrum
 - ($p=10000$, $n=100$ is typical)
- ◆ Activation levels of all genes on a genome
 - ($p=30000$ genes, $n=10$ microarrays is typical)
- ◆ Credit Card Transactions of 100,000 customers in 1 year
 - ($p=500$, $n=50000000$ is typical)

here we have $n \ll p$ and $n \ll \ll p$ and $n \gg \gg p$
 typical situations needing **data mining**
 (either or both of n and p are **BIG**)



© NRJF, University of Sheffield, 2011/12, Semester 1

Multivariate Data Analysis

43

NOTE:-

- ◆ typically the variables are **correlated**
but
- ◆ individual sets of observations are **independent**
- ◆ c.f. Time Series Analysis
 - **correlated** observations, (& independent variables)



© NRJF, University of Sheffield, 2011/12, Semester 1

Multivariate Data Analysis

44

- ◆ Blood pressure and body mass correlated in the same individual
- ◆ Measurements on different people are independent
- ◆ Long petals are also likely to be wide and also have large sepals
- ◆ Sizes of different flowers are independent



© NRJF, University of Sheffield, 2011/12, Semester 1

Multivariate Data Analysis

45

Some Multivariate Problems

- ◆ Reduction of dimensionality for
 - exploratory analysis
 - simplification (MVA is easier if $p=1$ or $p=2$)
 - methods of **data mining**
e.g. principal component analysis, factor analysis, non-metric scaling



© NRJF, University of Sheffield, 2011/12, Semester 1

Multivariate Data Analysis

46

Some Multivariate Problems (Cont^d)

- ◆ Cluster Analysis/Classification
 - Do data arise from a homogeneous source or do they come from a variety of sources
 - e.g. does a medical condition have sub-variants

This is a problem of **unsupervised learning**



© NRJF, University of Sheffield, 2011/12, Semester 1

Multivariate Data Analysis

47

Some Multivariate Problems (Cont^d)

- ◆ Canonical Correlation Analysis
 - Of use when looking at relationships between sets of variables
 - e.g. questionnaire analysis
2 groups of questions
first group investigate *expectations*
second group their *evaluations*
 - How do **evaluations** relate to **expectations**?
(both measured by multiple questions)



© NRJF, University of Sheffield, 2011/12, Semester 1

Multivariate Data Analysis

48



- **Course sequence**
 - ◆ multivariate graphics
 - ◆ dimensionality reduction
 - ◆ estimation and testing problems
 - straightforward generalizations of univariate statistical techniques
 - new techniques special for multivariate problems
 - ◆ statistical approach to discrimination

© NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 49

- **Quotation:–**
 - ◆ Much classical and formal theoretical work in Multivariate Analysis rests on assumptions of underlying *multivariate normality* — resulting in techniques of very limited value
(Gnanadesikan, page 2)

© NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 50

- ◆ The most useful techniques in Multivariate Data Analysis are *data mining* techniques of
 - graphical display
 - dimensionality reduction
 - looking at data in different ways
 - discovering structure
- ◆ Interpretation of formal statistical tests needs to be supplemented by these techniques

© NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 51

- ◆ The most useful techniques in Multivariate Data Analysis are *data mining* techniques of
 - graphical display
 - dimensionality reduction
 - looking at data in different ways
 - discovering structure
- ◆ Interpretation of formal statistical tests needs to be supplemented by these techniques
- ◆ most formal statistical tests require more observations than dimensions
 - (i.e. $n > p$)
- ◆ the most **useful** multivariate procedures are those which ‘work’ when $p > n$
 - i.e. principal component analysis (PCA) & partial least squares (PLS)

© NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 52

- **Basic Notation**
 - See Course Notes P14
 - ◆ Observations are column vectors

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$$

(a $p \times 1$ vector)

© NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 53

- ◆ **Transpose:**
 - a dash ' denotes transpose: $x' = (x_1, x_2, \dots, x_p)$
- ◆ The $n \times p$ **data matrix** X' is

$$X' = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix} = \begin{pmatrix} X'_1 \\ X'_2 \\ \vdots \\ X'_n \end{pmatrix}$$
- ◆ **NOTE:–** X is $p \times n$; X' is $n \times p$

© NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 54



- ◆ Define the sample mean vector

$$\bar{X}' = (\bar{X}_1, \bar{X}_2, \dots, \bar{X}_p) = \frac{1}{n} \mathbf{1}' X'$$
 - where $\mathbf{1}$ is the column vector of n 1s
- ◆ sample variance $S = \text{var}(X')$ by

$$S = \frac{1}{n-1} (X - \bar{X})(X - \bar{X})'$$
- ◆ S can also be written as

$$S = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})' = \frac{1}{n-1} \left(\sum_{i=1}^n x_i x_i' - n \bar{x} \bar{x}' \right)$$

© NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 55

- **Notes**
 - ◆ \bar{X} is a p -vector
 - ◆ S is a $p \times p$ matrix, the diagonals give **variances** of the p variables
off-diagonals give **covariances**
 - ◆ S is non-singular and positive definite, provided all measured variables are 'distinct'
 - (i.e. none is a linear combination of any of the others).
 - non-singular and positive definite \Rightarrow **invertible**

© NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 56

- **Notes (Cont'd)**
 - ◆ If w is any vector then

$$\text{var}(X'w) = w' \text{var}(X') w = w' S w$$
 - c.f. in 1-dim: $\text{var}(\lambda X) = \lambda^2 \text{var}(X)$
 - ◆ If A is any $p \times q$ matrix then

$$\text{var}(X'A) = A' \text{var}(X') A = A' S A$$

(1×p)×(p×p)×(p×1) = 1×1
i.e. a scalar

(q×p)×(p×p)×(p×q) = q×q

Check these results — see Task Sheet 1
(if in doubt try $w' = (w_1, w_2)$ explicitly)

© NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 57

- **Notes (Cont'd)**
 - ◆ use `apply(. , . , .)` to obtain \bar{X} in R
 - ◆ use `var(.)` to obtain S

```
> airmean <- apply(airpoll, 2, mean)
> airvar <- var(airpoll)
```

© NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 58

© NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 59

© NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 60



Contents

- Preliminaries
- 0: Introduction
- 1: Graphical Displays**
- 2: Reduction of Dimensionality
- 3: Multidimensional Scaling Techniques
- 4: Discriminant Analysis
- 5: Multivariate Regression Analysis
- 6: Canonical Correlation Analysis
- 7: Partial Least Squares
- 8: Statistical Analysis of Multivariate Data
- 9: Statistical Discriminant Analysis

© NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 61

- ### Graphical Displays
- Exploratory analysis aims:
 - ◆ Detect structure
 - Outliers
 - Subgroups
 - Relationships
 - ◆ Suggest models
 - Normal?
 - Skewness.....
- © NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 62

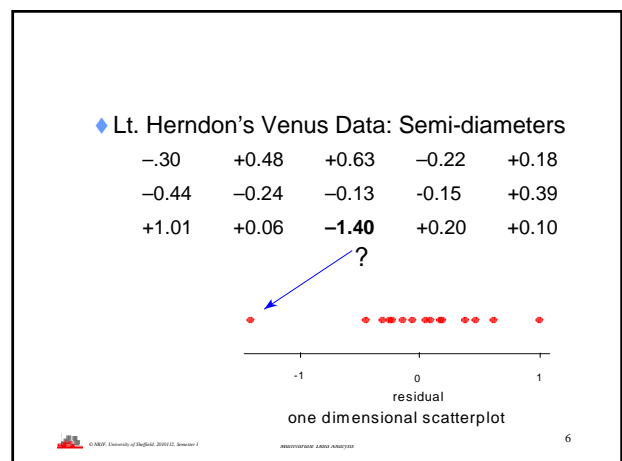
- Presentational aims (after analysis)
 - ◆ Summarize & illustrate results
 - Highlight key conclusions
 - Capture interest of reader
 - ◆ Explain analysis
 - Shew validity / limitations of analysis
 - e.g. how scattered are points around a regression line
- © NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 63

- **1 Dimension**
 - ◆ Small # points
 - 1-dimensional scatter plots
 - dot plots
 - ◆ Large # points
 - stem & leaf plots
 - histograms
 - box plots
 - + other special data dependent techniques
 - e.g. circular data need circular histograms
- © NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 64

- ◆ Lt. Herndon's Venus Data: Semi-diameters

-.30	+0.48	+0.63	-0.22	+0.18
-0.44	-0.24	-0.13	-0.15	+0.39
+1.01	+0.06	-1.40	+0.20	+0.10

© NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 65

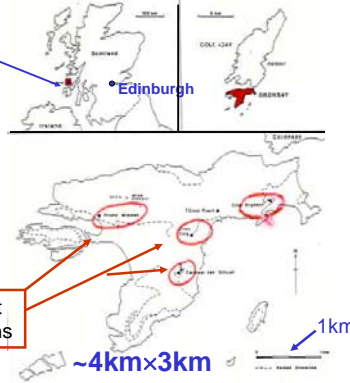


Examples of histograms:

Data:

- ◆ lengths of ancient otoliths at four sites
- ◆ contemporary fish caught on known dates
- ◆ fishbones grow larger with age and the pictures suggest that the four archaeological sites had bones from fish caught on different dates

Location of Oronsay



4 best middens

~4km x 3km

1km

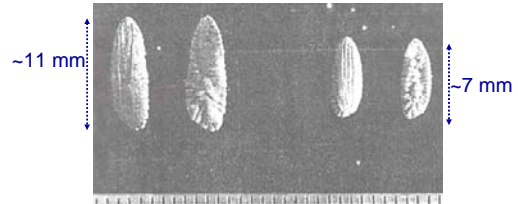
Otoliths

- ◆ Fossilized ear-bones from fish
- ◆ Size of otolith depends on age of fish
- ◆ Fish are born at same time each year

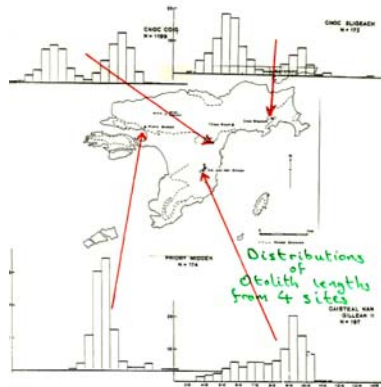
SO

- ◆ Size of otolith
 - age of fish
 - time of year fish was caught

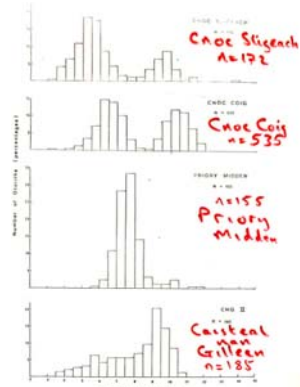
2nd year fish 1st year fish

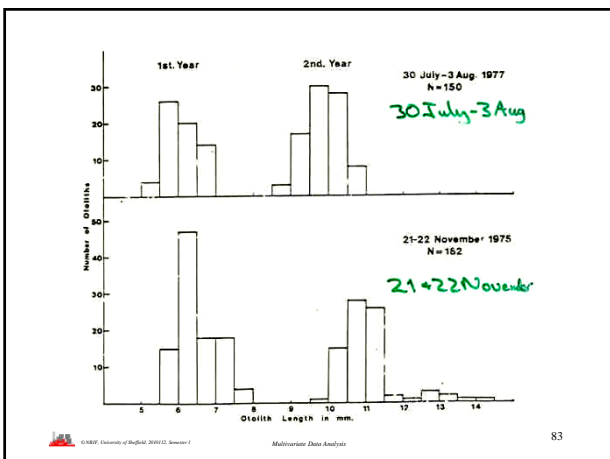
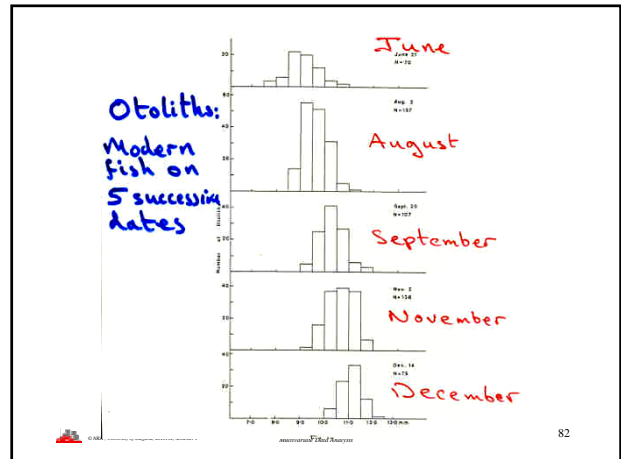
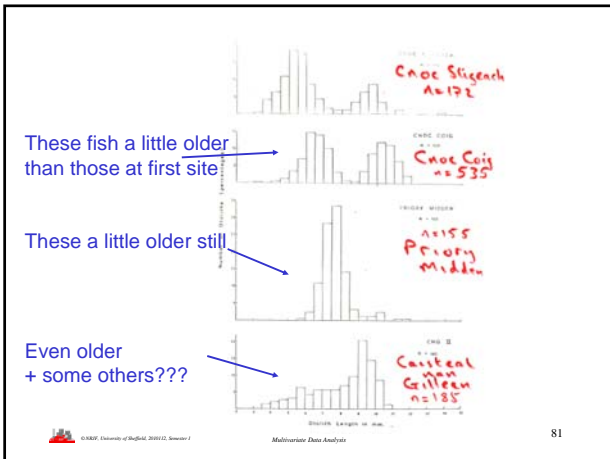
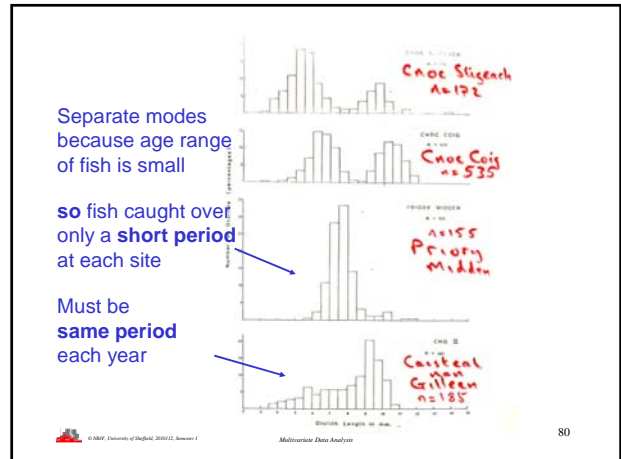
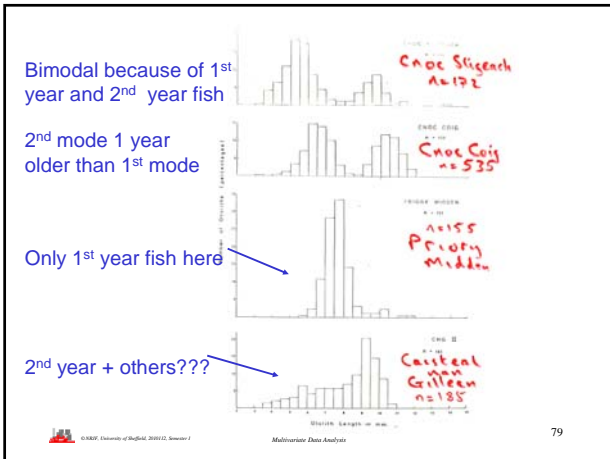


Otoliths: petrified ear-bones from fish
(coley or saithe)



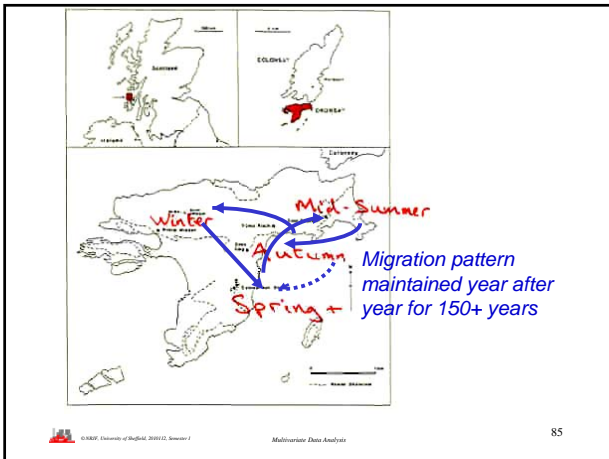
Histograms of otolith lengths at the 4 sites





- Simple conclusions on average time of catch easy from histograms
 - More sophisticated modelling gives estimates of length of period of occupation
 - ◆ Sample variance of lengths increase with length of time-span of catch
- 84

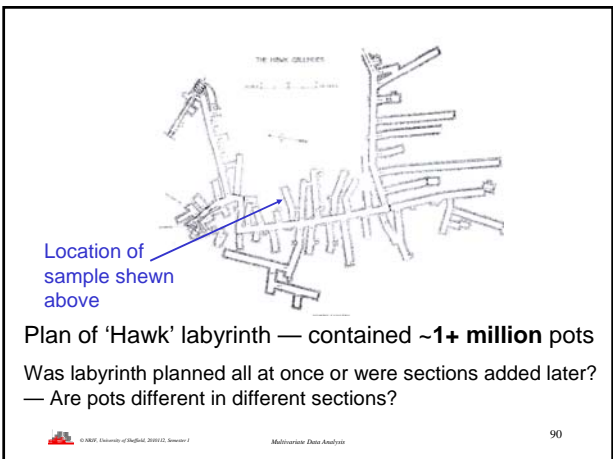




- **Explanation of pattern**
 - ◆ Big fish only an occasional opportunity
 - ◆ Main diet was shell-fish from shore
 - ◆ Move on when local stocks low
 - ◆ Allow maximum time for recovery, i.e. 12 months if possible
 - ◆ Once pattern is established it must be repeated each year

- ◆ Comparisons of small number samples using histograms is possible.
- ◆ With a larger number of samples boxplots are preferable
 - e.g. too large for a separate histogram of each to fit on a page


- **Example of Boxplots:**
 - ◆ Data are rim-circumferences of mummy-pots (containing mummified birds) found at various different galleries in the Sacred Animal Necropolis in Saqqara, Egypt.
 - ◆ The boxplots illustrate marked variation in sizes of pots between galleries






At top of 10m shaft leading to labyrinth

© NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 91



© NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 92



Intact pots *in situ* in labyrinth gallery
Now need a random sample of size 20.....

© NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 93



Pots contained mummified birds
(religious cult in ~200BC)

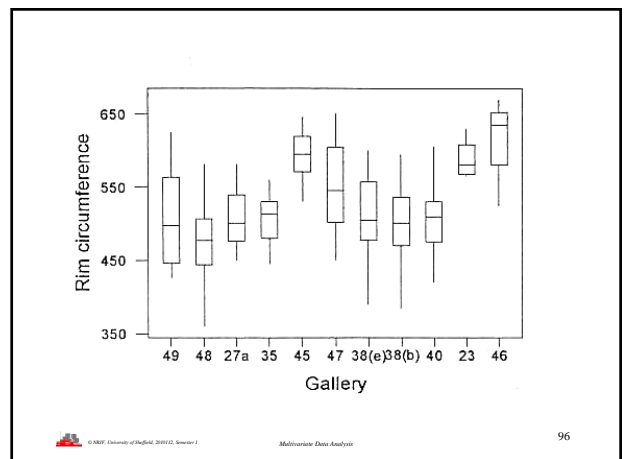
© NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 94

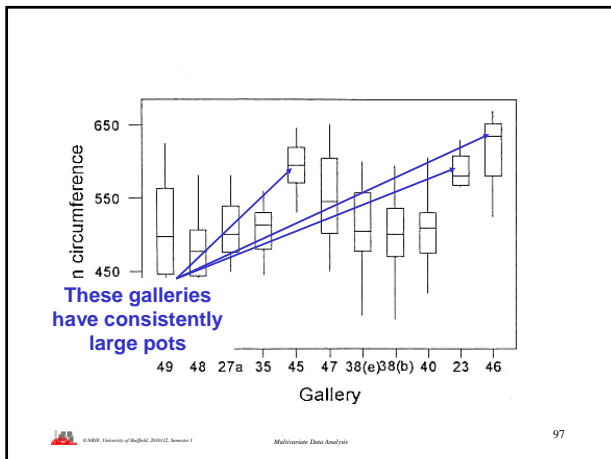
Mummy of apparent exceptional quality:

A later x-ray shewed it to be a fake (only sticks inside)

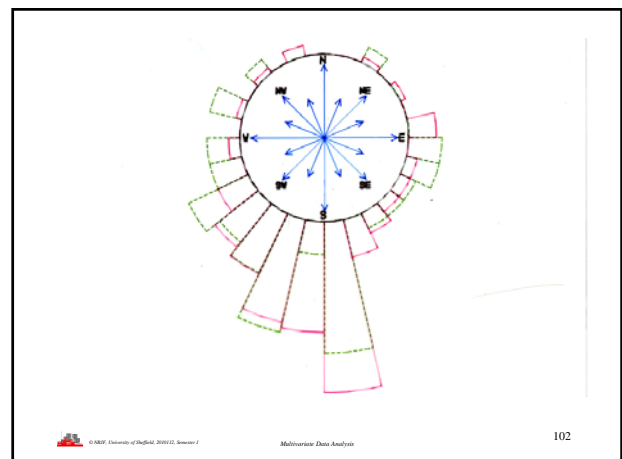
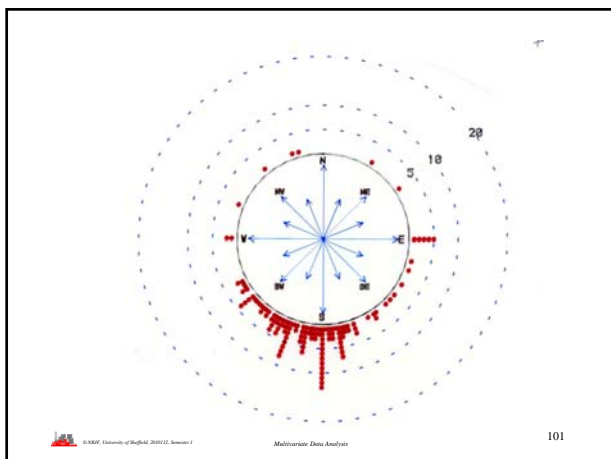
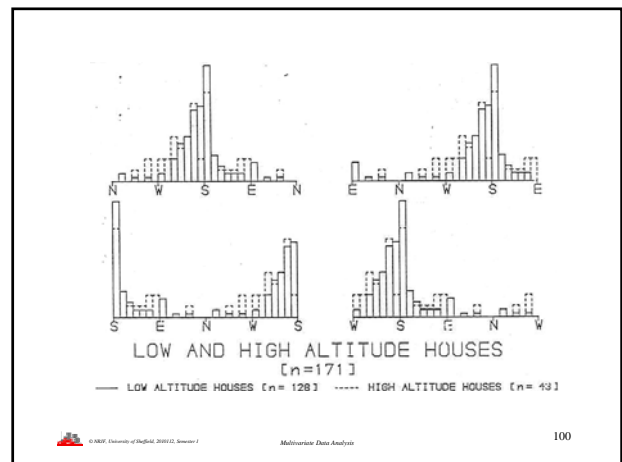
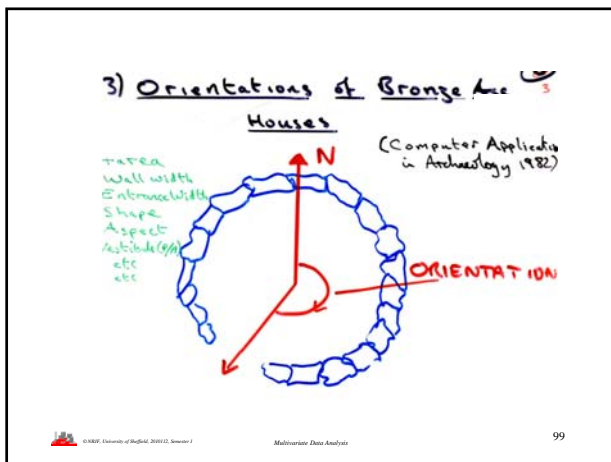


© NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 95



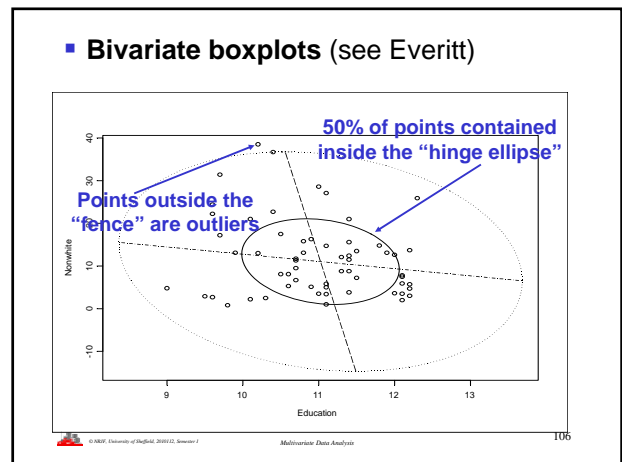
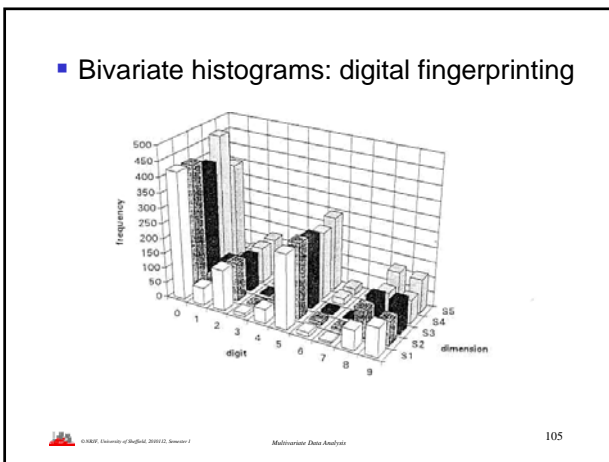
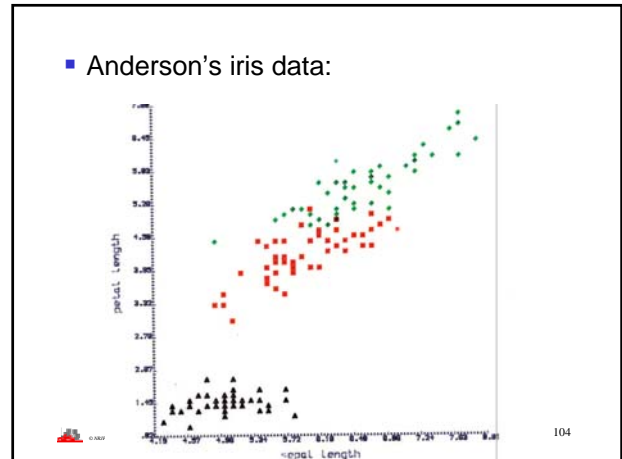


- **Example of circular data:**
 - ♦ orientations Bronze Age houses in Dartmoor
 - ♦ two groups of houses
 - ♦ are they equally consistently orientated?
 - ♦ ordinary histograms misleading
 - ♦ *circular dotplots & histograms* capture features of the data.



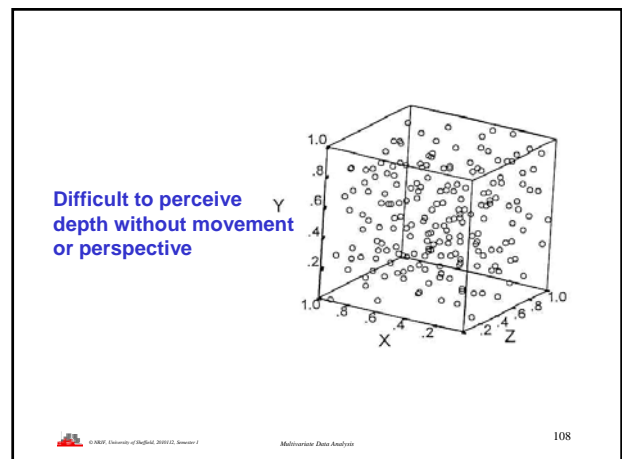
- **2 Dimensions**
 - ◆ Small # points
 - scatter plots
 - ◆ Large # points
 - bivariate histograms drawn in perspective
 - 2-dim frequency plots
 - bivariate boxplots
 - augmented scatter plots

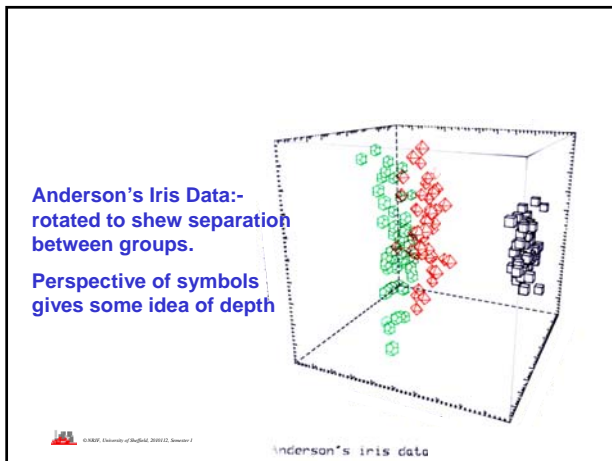
© NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 103



- **3 Dimensions**
 - ◆ 2-dim scatter plots of marginal components
 - perhaps joined sensibly
 - ◆ 3-dim scatter plots drawn in perspective
 - rotate interactively, using S-plus or SAS-insight

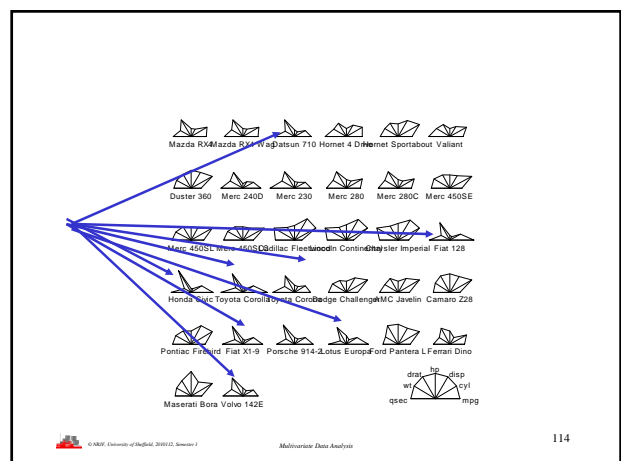
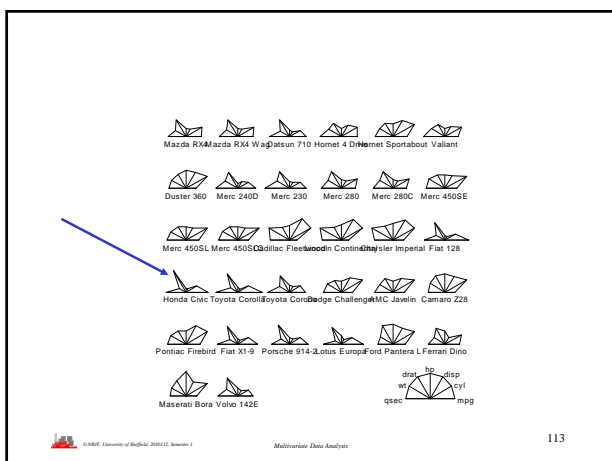
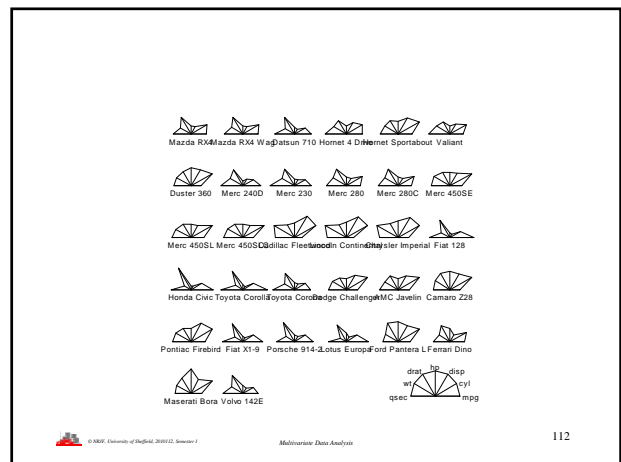
© NRJF, University of Sheffield, 2011/12, Semester 1 Multivariate Data Analysis 107





- **≥ 3 Dimensions**
 - ◆ **Sensible Methods**
 - **Matrix plots**
 - key tool in displaying multivariate data
 - pairwise scatterplots arranged in matrix
 - available in most packages
 - OK for medium numbers of points and dimensions
($3 \leq p \leq 6, 20 \leq n \leq 200$ say)

- **Star Plot**
 - ◆ each observation represented by a polygon
 - value of each variable corresponds to length of the vector to each vertex
 - OK for $p \leq \sim 12, n \leq \sim 50$
 - can compare individual points
 - may be able identify similar points

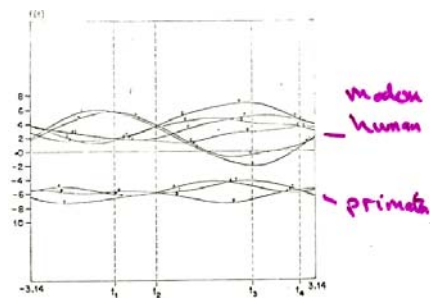


▪ **Andrews' Plots**

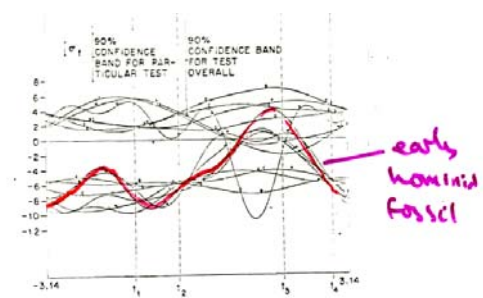
- ◆ maps p-dimensional data $\{x_i\}$ onto 1-dimensional $\{f_x(t)\}$ for any t.
- ◆ plot $\{f_x(t)\}$ over $-\pi < t < +\pi$ yields 1-dimensional representation
- ◆ preserves many statistical properties
 - but some drawbacks if p large
- ◆ Available in some packages and Excel add-in

▪ **Example**

- ◆ 9-dim data on teeth taken from
 - 6 modern humans
 - 3 various primates
 - 3 early hominids
- ◆ Do hominids 'look like' primates or humans?



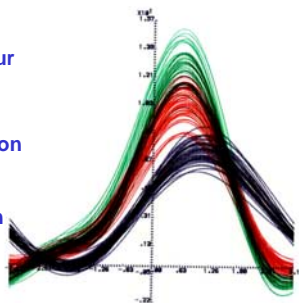
Humans quite different from primates



Hominid mostly like primates but with some characteristics of humans

Andrews' plot of all four components of the Iris Data:-

Can see clear separation of black from red/gree and some overlap between red and green from all points of view



▪ **Other methods**

- ◆ use of complicated symbols perhaps on scatterplots of 2 of the variables
 - (c.f. weather maps).
- ◆ e.g. Anderson [Florence Nightingale] Glyphs
- ◆ Kleiner-Hartigan trees
- ◆ Chernoff faces
 - code values of various variables as different facial features



Examples:-

- ◆ 24 hourly measurements of air pollution
 - rising ozone
 - sulphur dioxide
 - radiation, etc
- ◆ association of variables with features carefully chosen to give the impression the pollution 'gets worse' in afternoon

121

122

123

◆ Example: national stereotypes

	1	2	3	4	5	6	7	8	9	10	11	12	13
French	37
Spanish
Italian
British
Irish
Dutch
German	8

1. Stylish, 2. Arrogant, 3. Sexy, 4. Devious, 5. Easy-going, 6. Greedy, 7. Cowardly, 8. Boring, 9. Efficient, 10. Lazy, 11. Hard-working, 12. Clever, 13. Courageous.

124

125

Conclusions

- ◆ Scatter plots good for few variables
 - (and not to many observations, say < ~200)
- ◆ Star plots OK for medium data sets
- ◆ Special techniques Andrews' plots and Chernoff Faces etc OK for special data sets but not routine use

126



- Scatter plots are most useful routine tool
 - ◆ Need to pick the most informative 'few' dimensions
 - ◆ Either select the 'most interesting variables'
or
 - ◆ Combine variables to create a few new ones which are the 'most interesting'
 - ◆ need techniques of **dimensionality reduction**.

