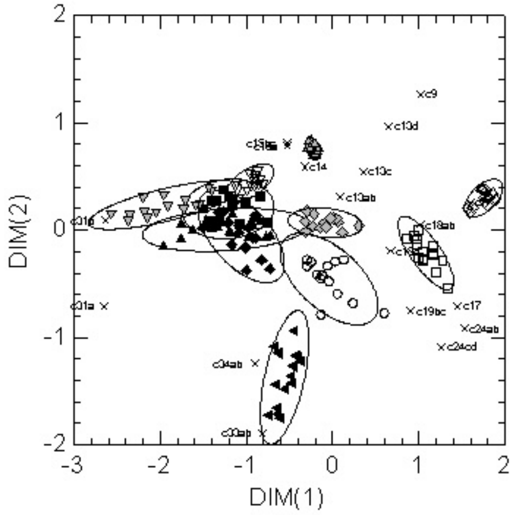
MULTIVARIATE STATISTICAL ANALYSIS FOR FOOD SCIENCE AND AGRICULTURE: AN INTRODUCTION
5. CLUSTER ANALYSIS

Prof. Eugenio Parente
Scuola di Scienze Agrarie - Università della Basilicata



Outline

- objectives of Cluster analysis
- what is a cluster?
- more on multivariate displays
- similarity/dissimilarity measures for categorical and continuous data
- hierarchical cluster analysis (agglomerative techniques)
- optimization clustering techniques (k-means)

Objectives of cluster analysis

Problem: given a set of *n* objects or individuals for each of which *p* variables (characters, attributes) have been measured, find a classification scheme to group the objects in classes , find the number of classes (*g*) and their characteristics. The objectives of the analysis may be:

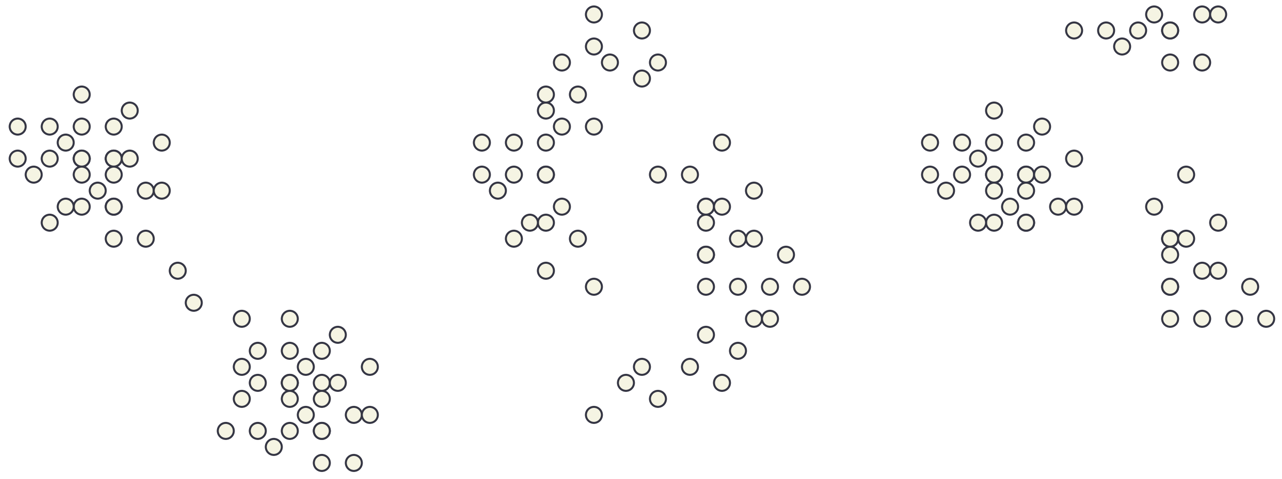
1. explorative data analysis
2. data reduction
3. finding a “true” (natural) classification
2. fitting a model
3. make predictions based on groups
4. generating and testing hypotheses on groups

What is a cluster?

According to Kendal and Buckland **a cluster is a contiguous grop of elements in a statistical populations**.

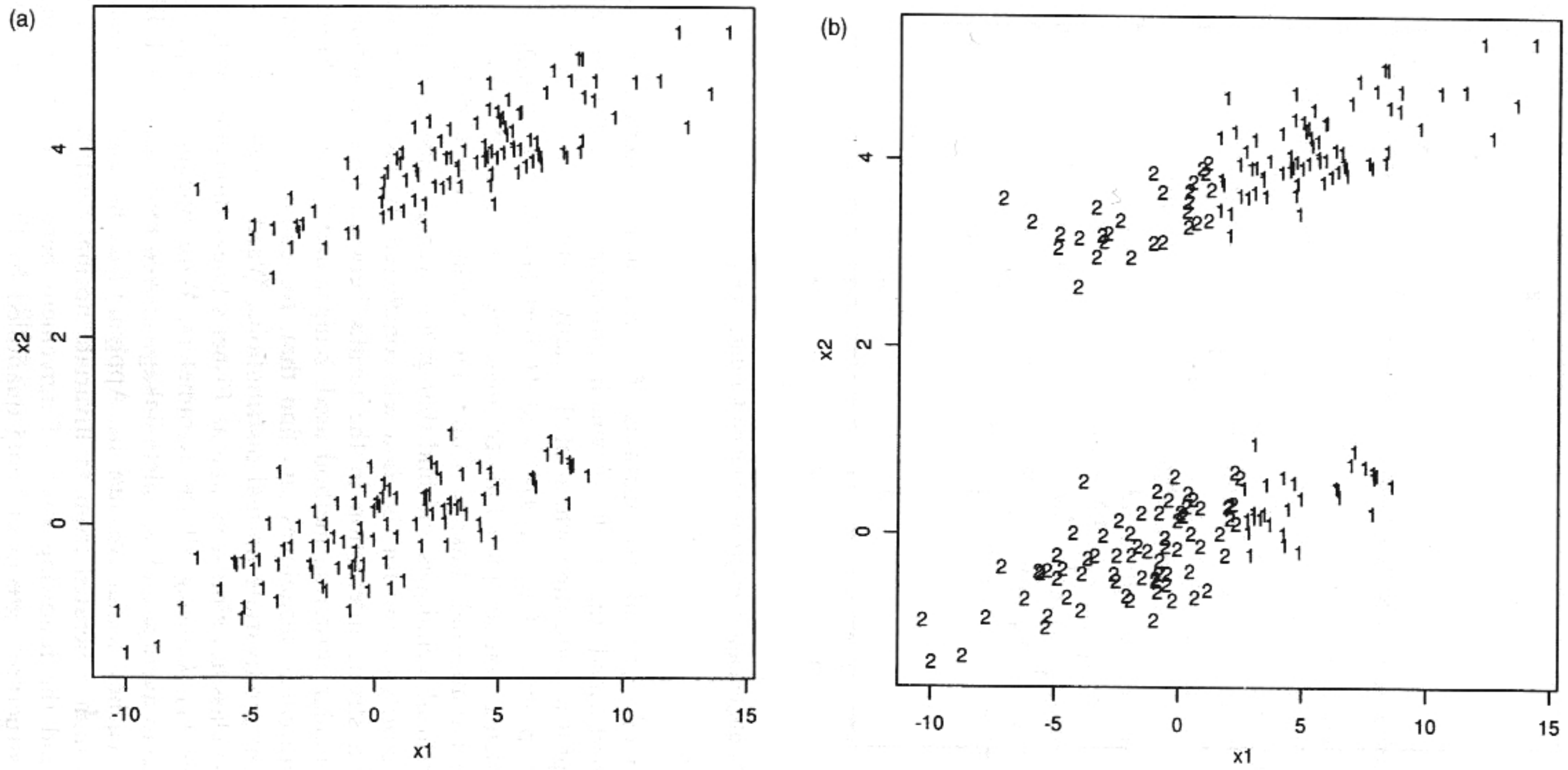
Another more operational definition is based on **internal cohesion (homogeneity)** and **external isolation (separation)**: in a *p*-dimensional space a (natural) cluster may be defined as a continuous portion containing a relatively high density of points separated from other clusters by regions of space containing a relatively low density of points

What is a cluster?

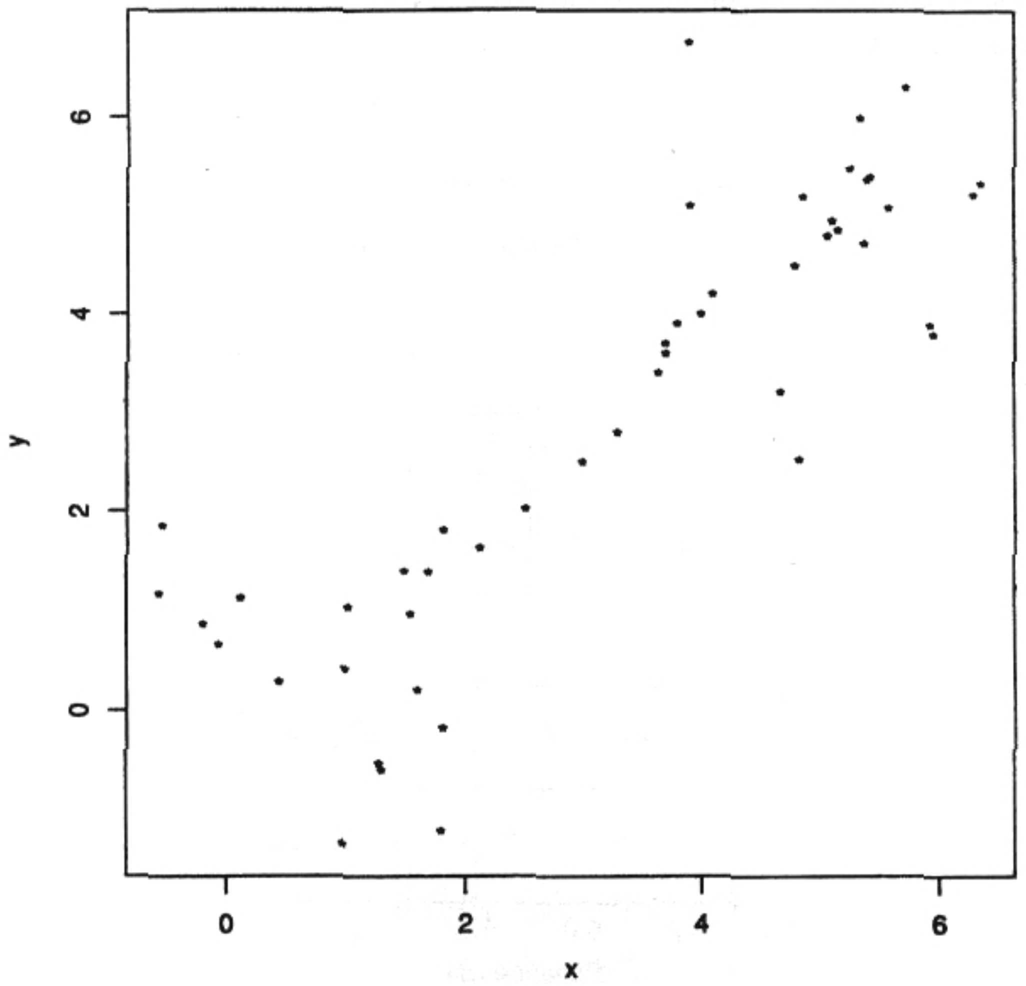


open file Clusters\clusters.syo for some simple examples of clusters and for the effect of clustering technique and standardization on cluster structure

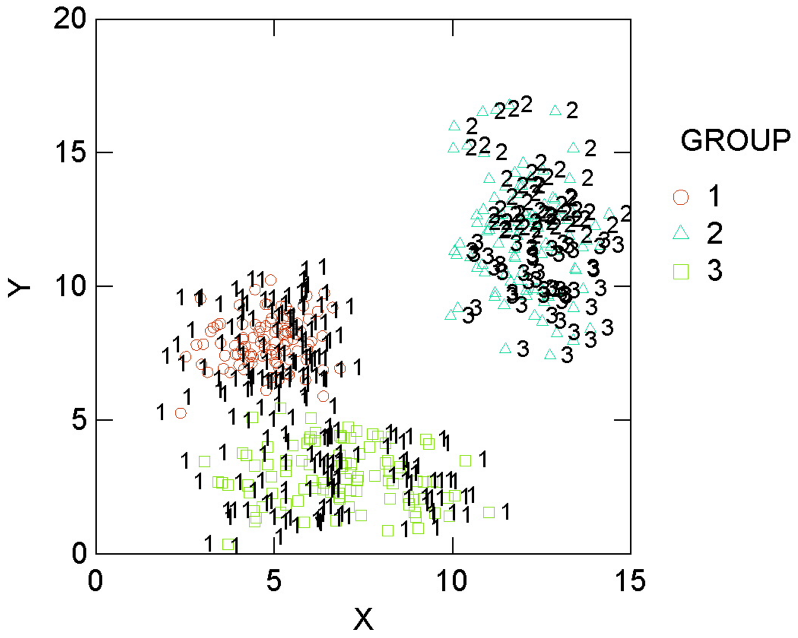
What is a cluster?

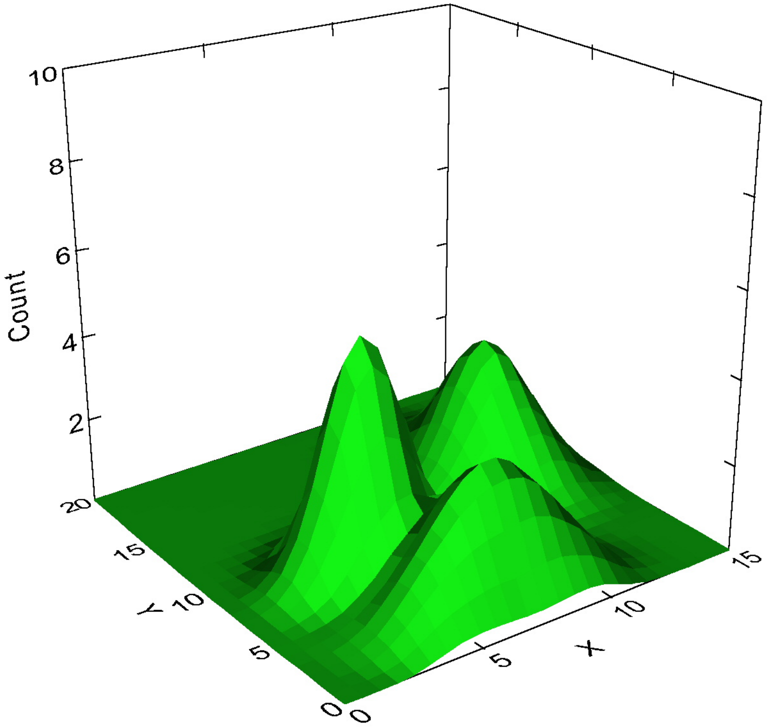


What is a cluster?

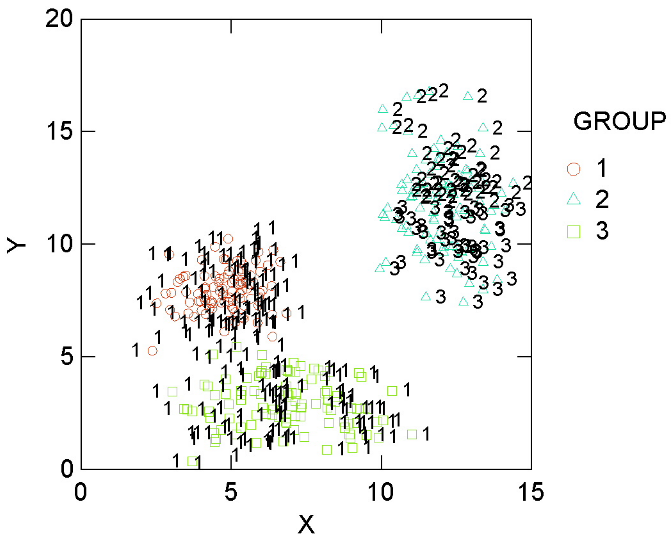


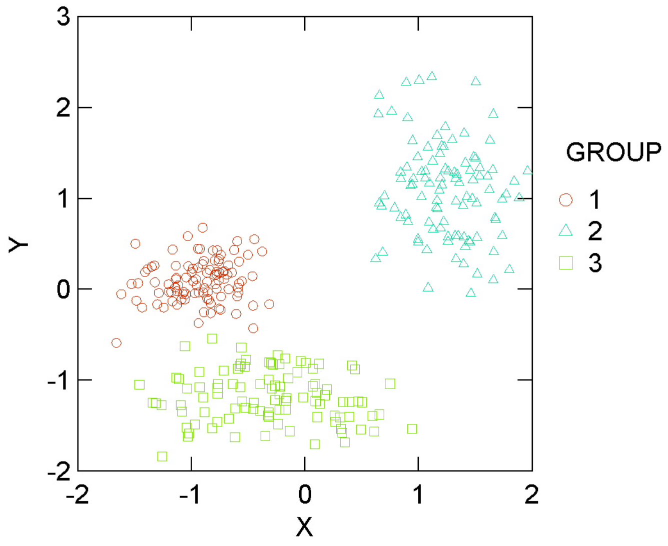
An artificial dataset





An artificial dataset



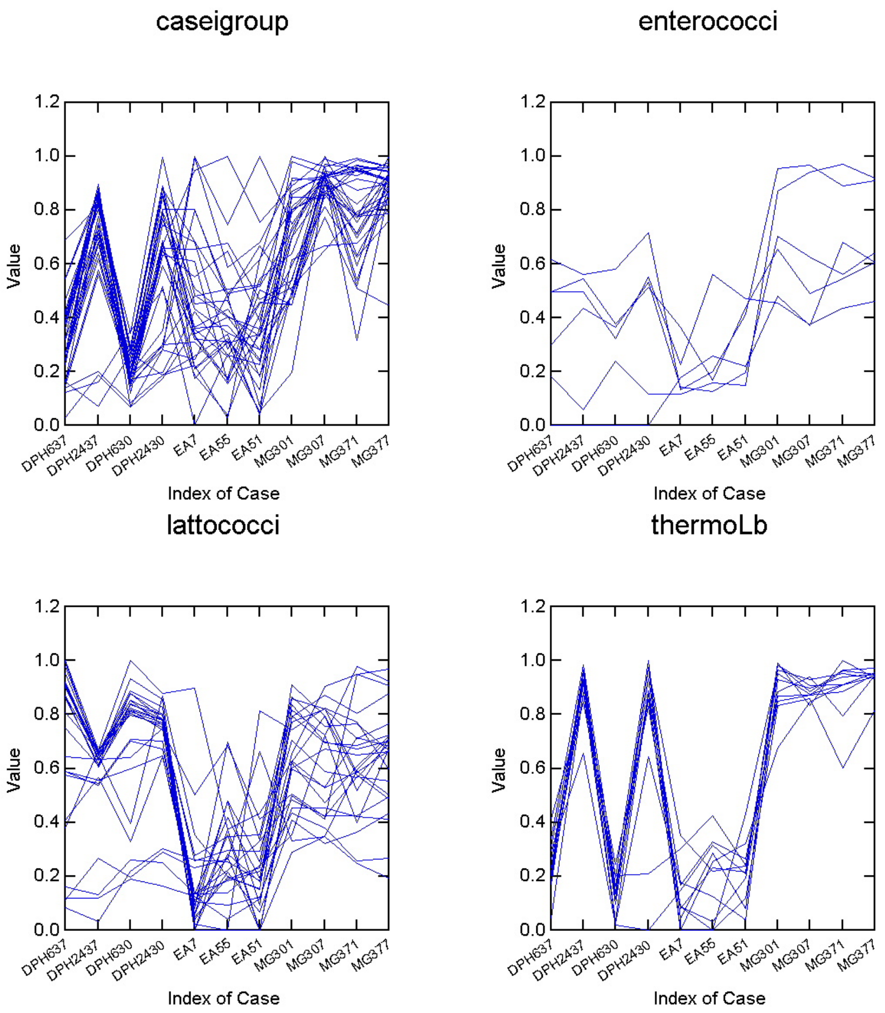


More on multivariate displays

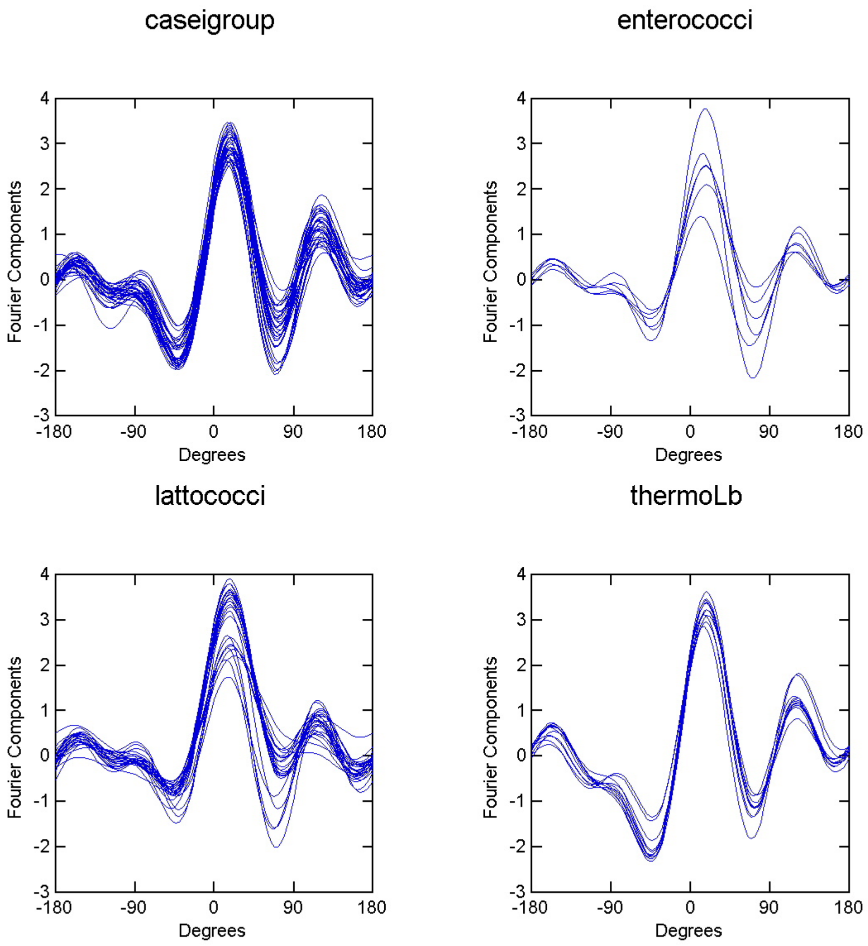
Bivariate or 3-D graphs with density displays on original data may not be of much assistance in exploring the data with large numbers of variables. You can:

- try the same graphs on PCA score plots and on MDS plots
- use a variety of multivariate displays
    - Andrew’ s Fourier plot
    - Parallel coordinates displays
    - Icon plots

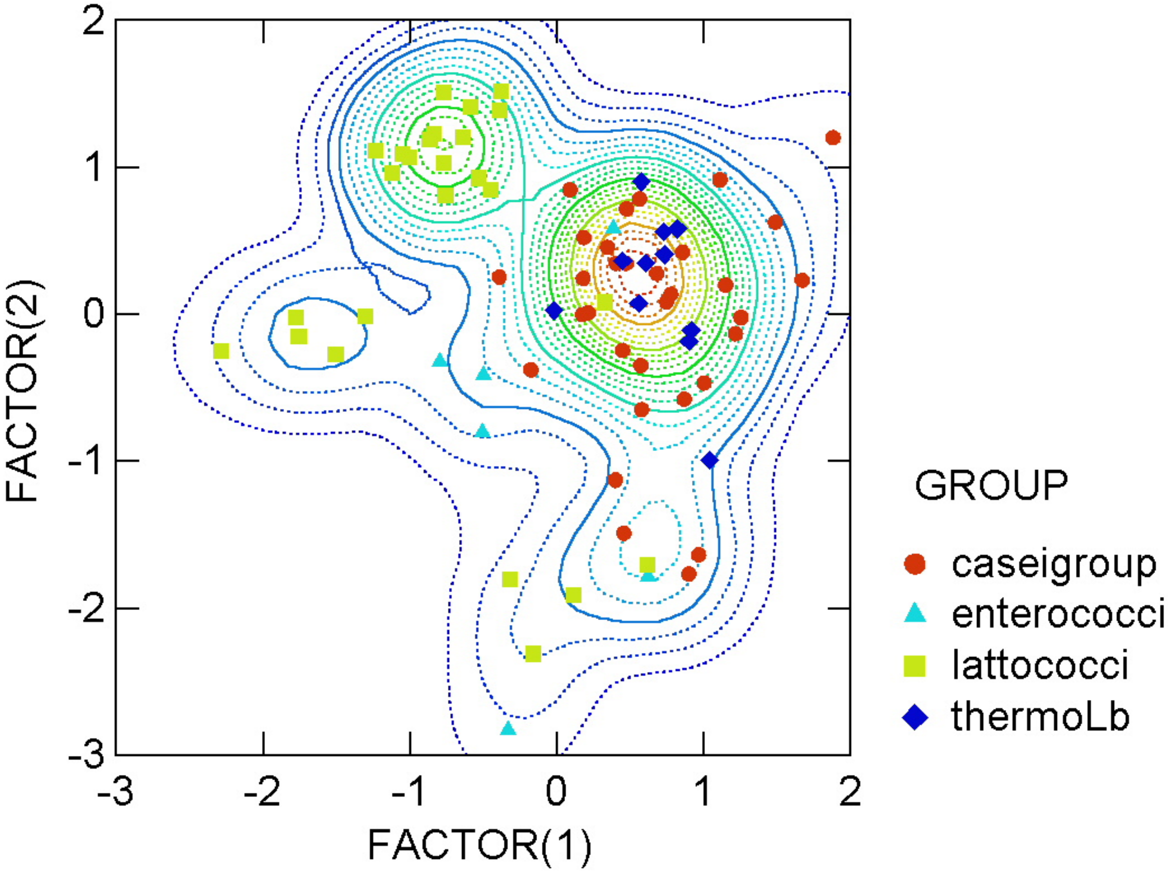
More on multivariate displays: parallel plot



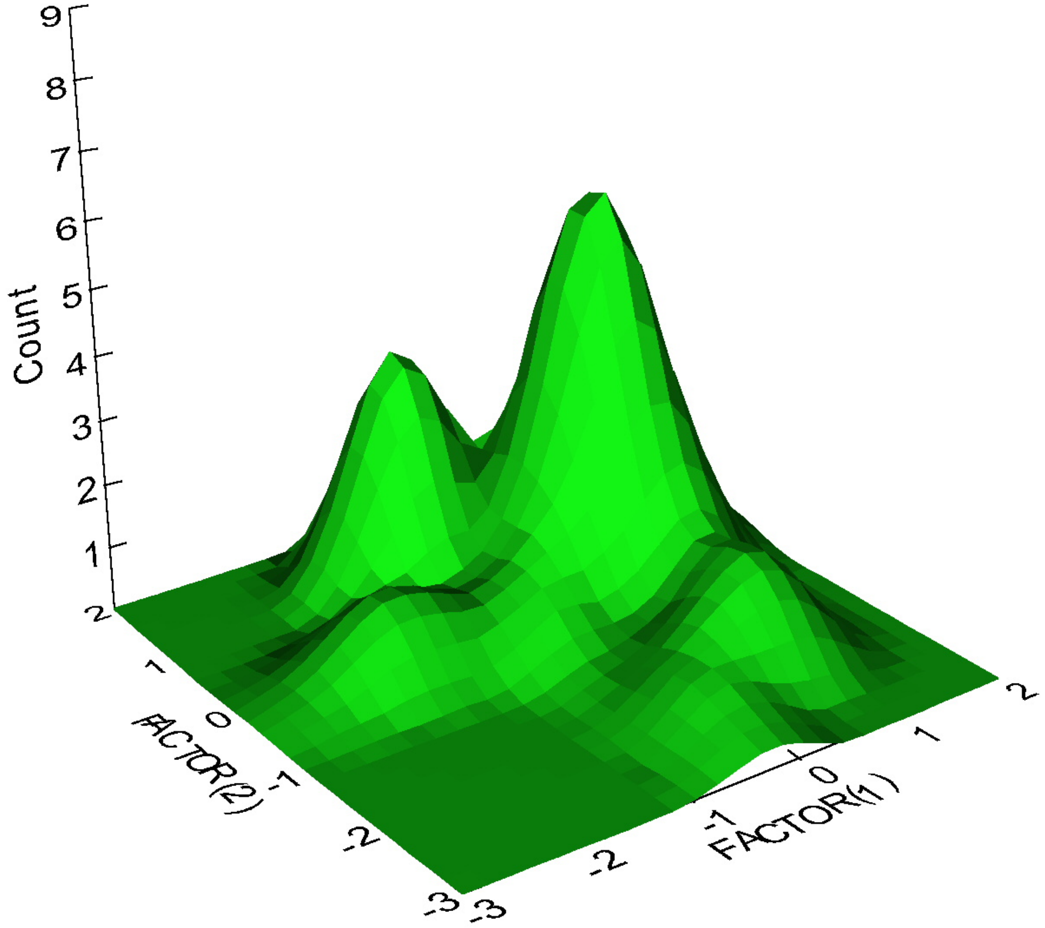
More on multivariate displays: Andrew’ s Fourier transform



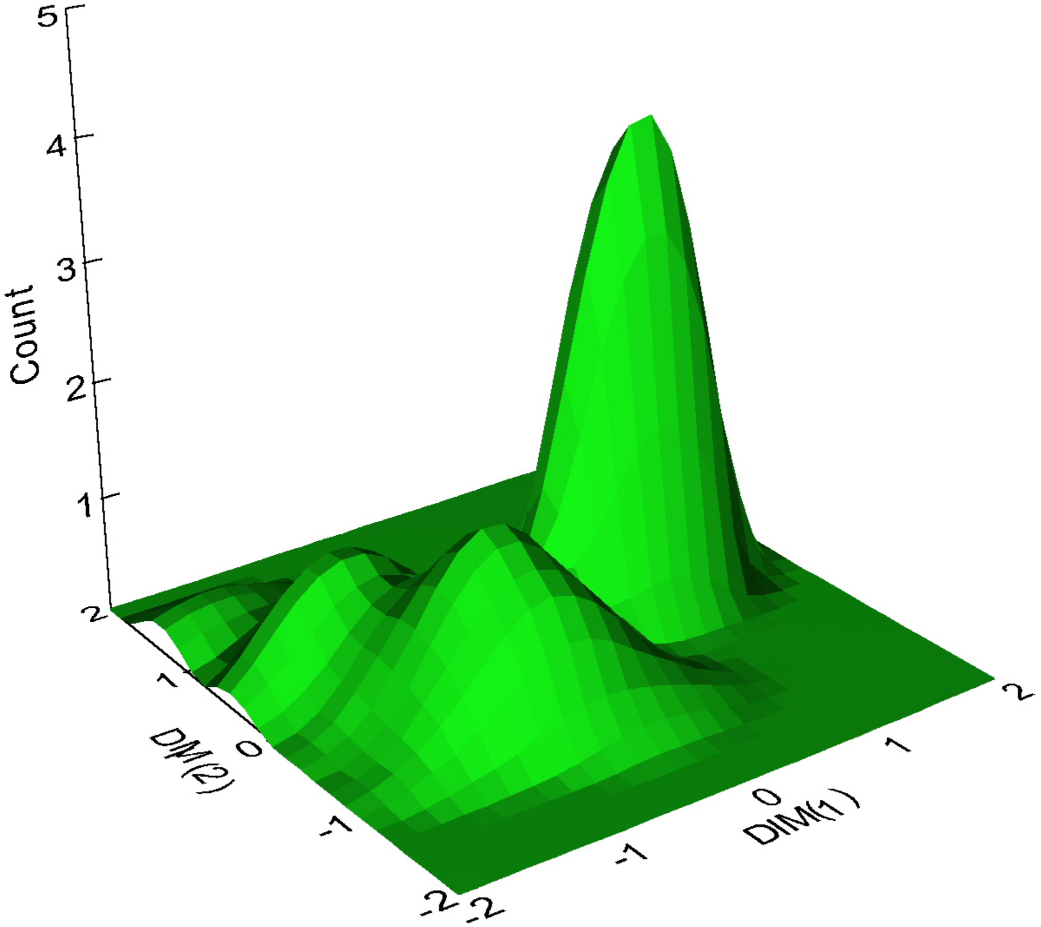
More on multivariate displays: density search on PCA score plots



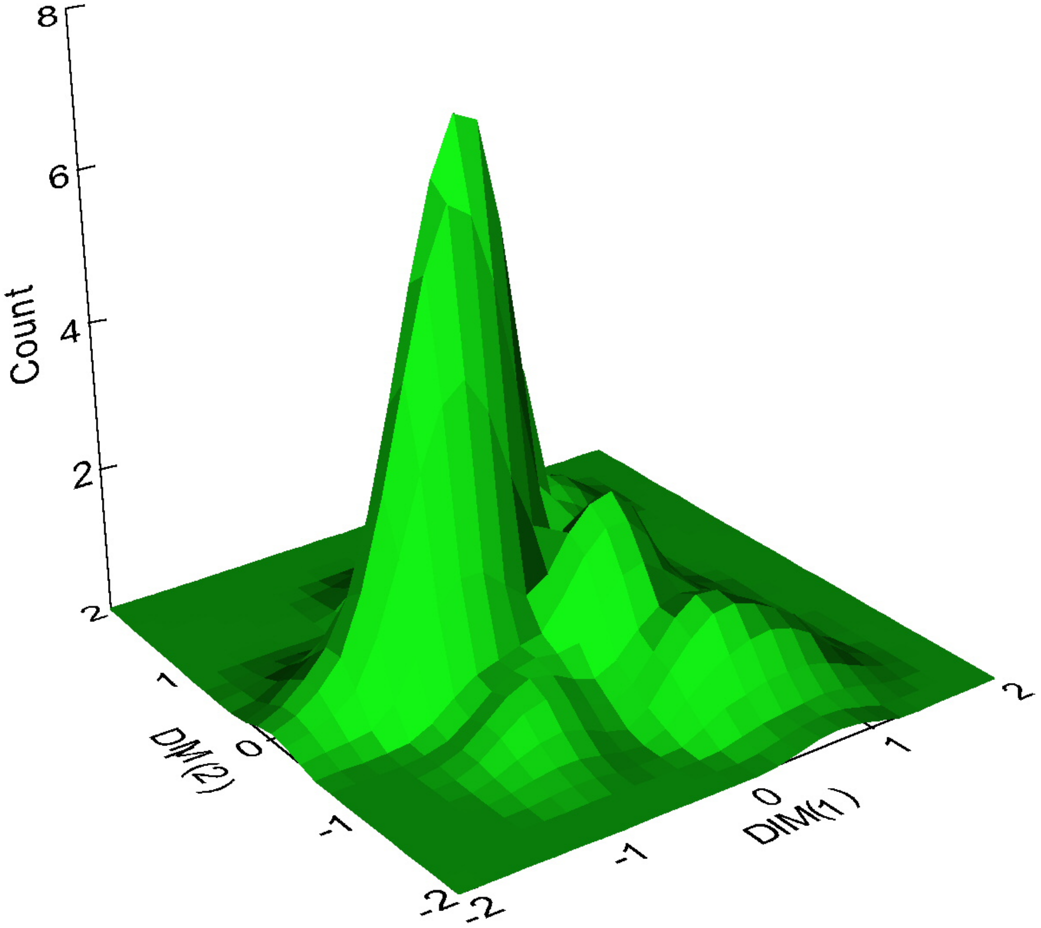
More on multivariate displays: density search on PCA score plots the technolab example



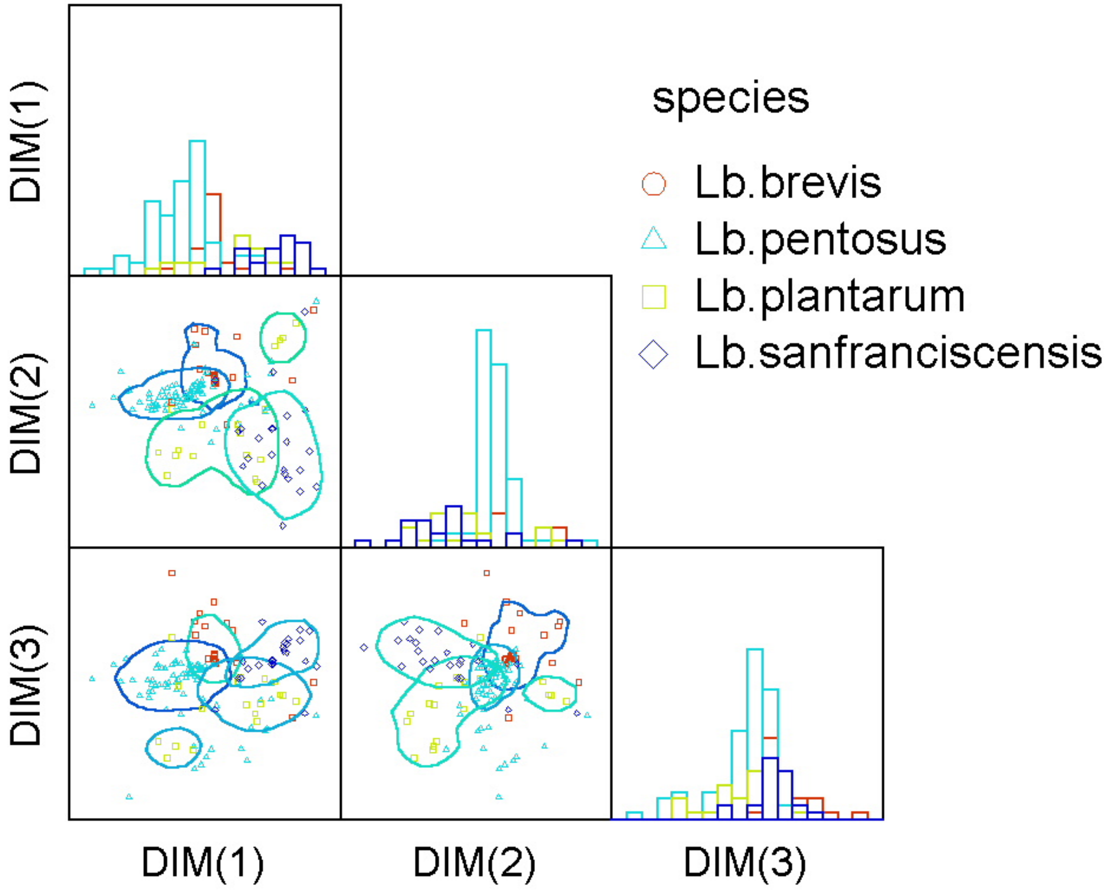
More on multivariate displays: density search on MDS maps (the RP-HPLC example)



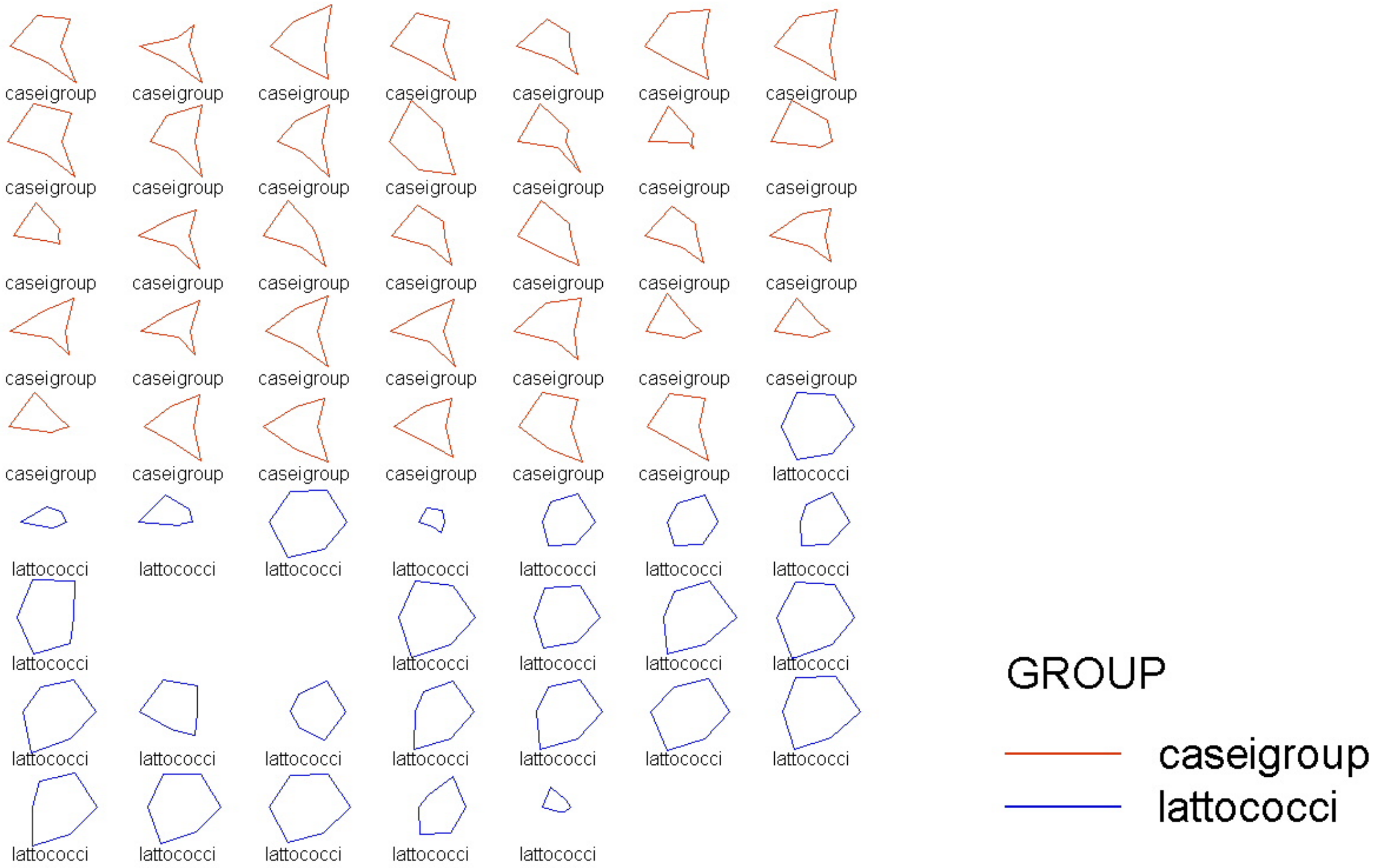
More on multivariate displays: density search on MDS maps (the RAPD-PCR example)



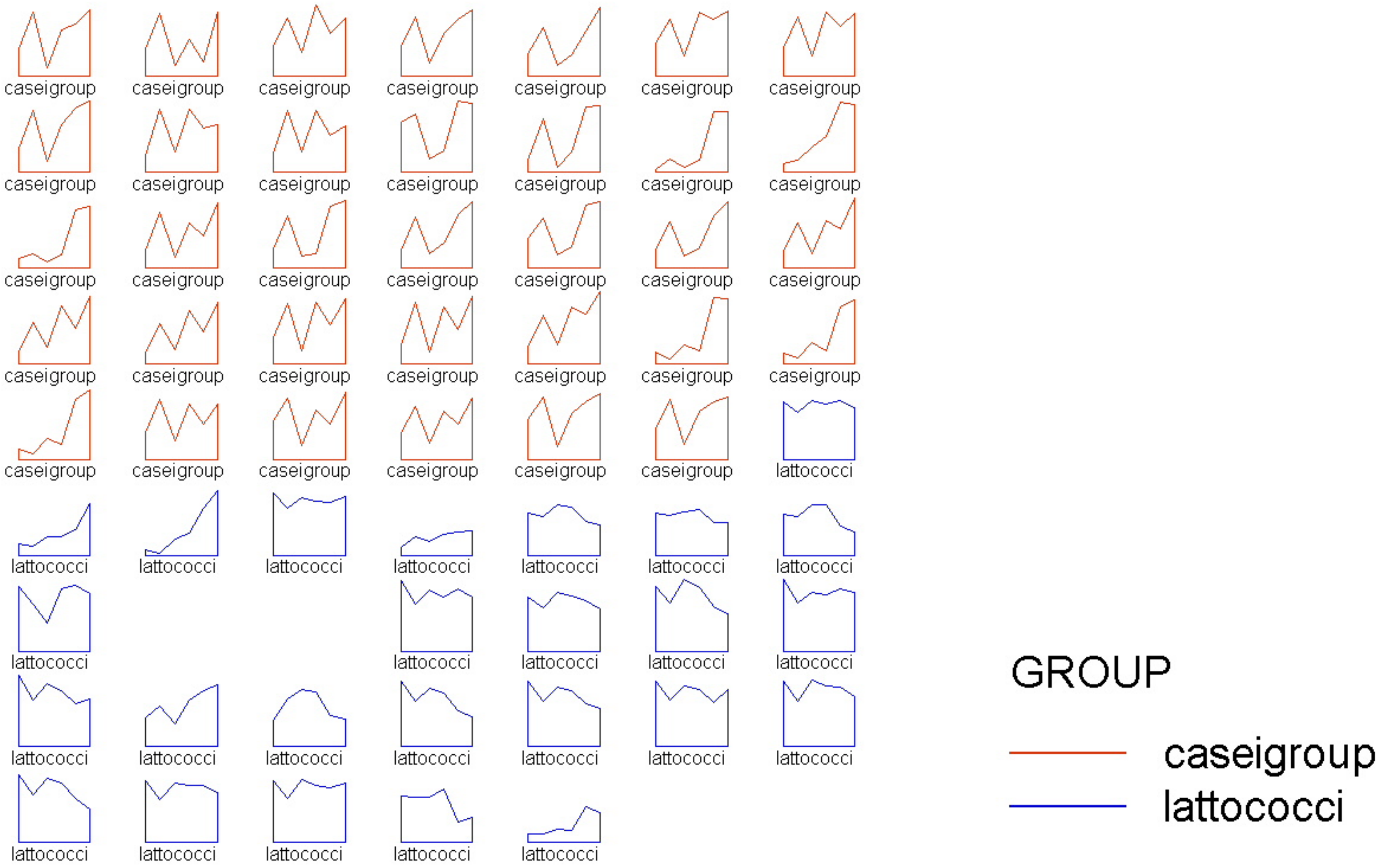
More on multivariate displays: SPLOM on MDS score plots (RAPD-PCR example)



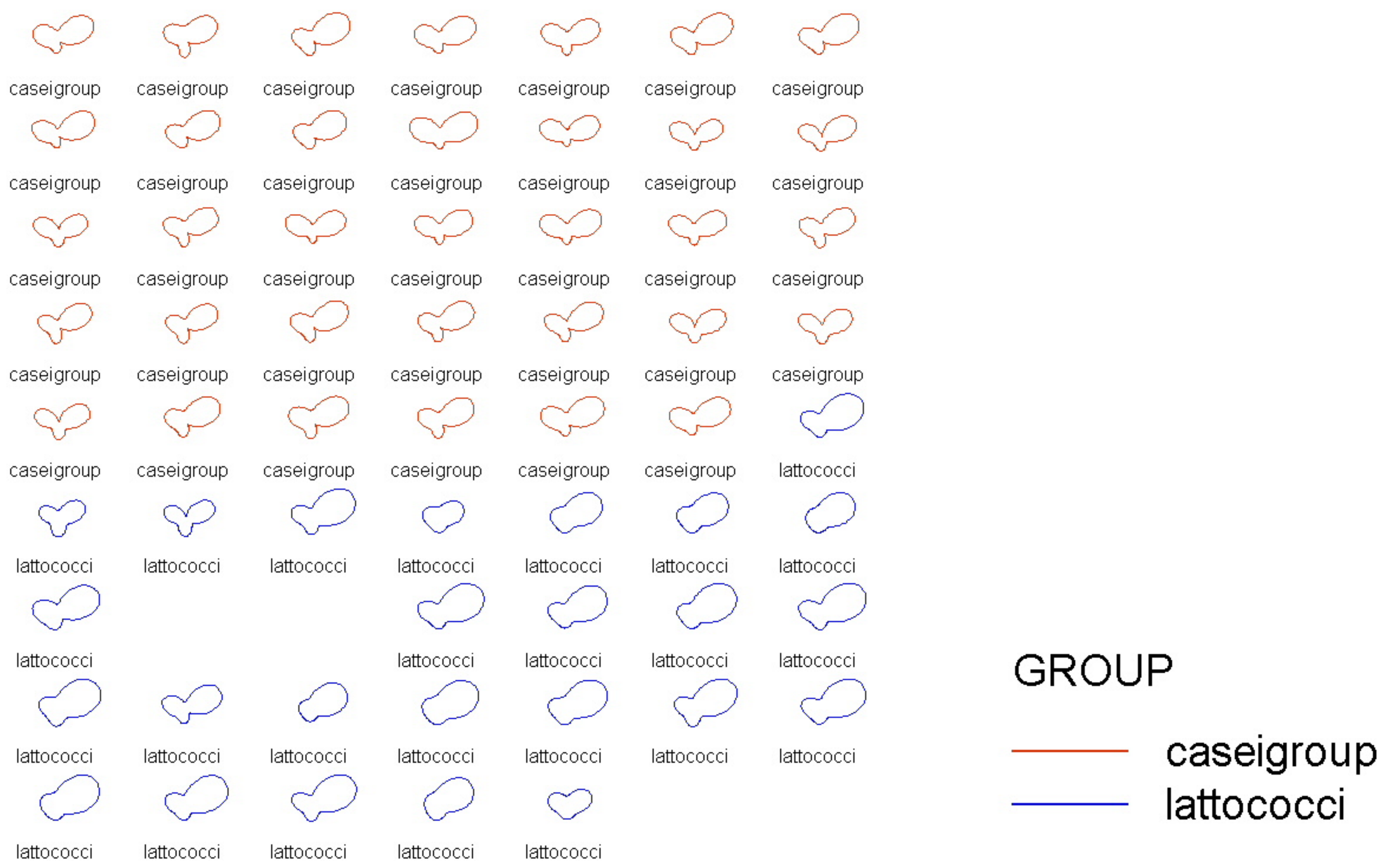
More on multivariate displays: icon (star) plot (technolab example)



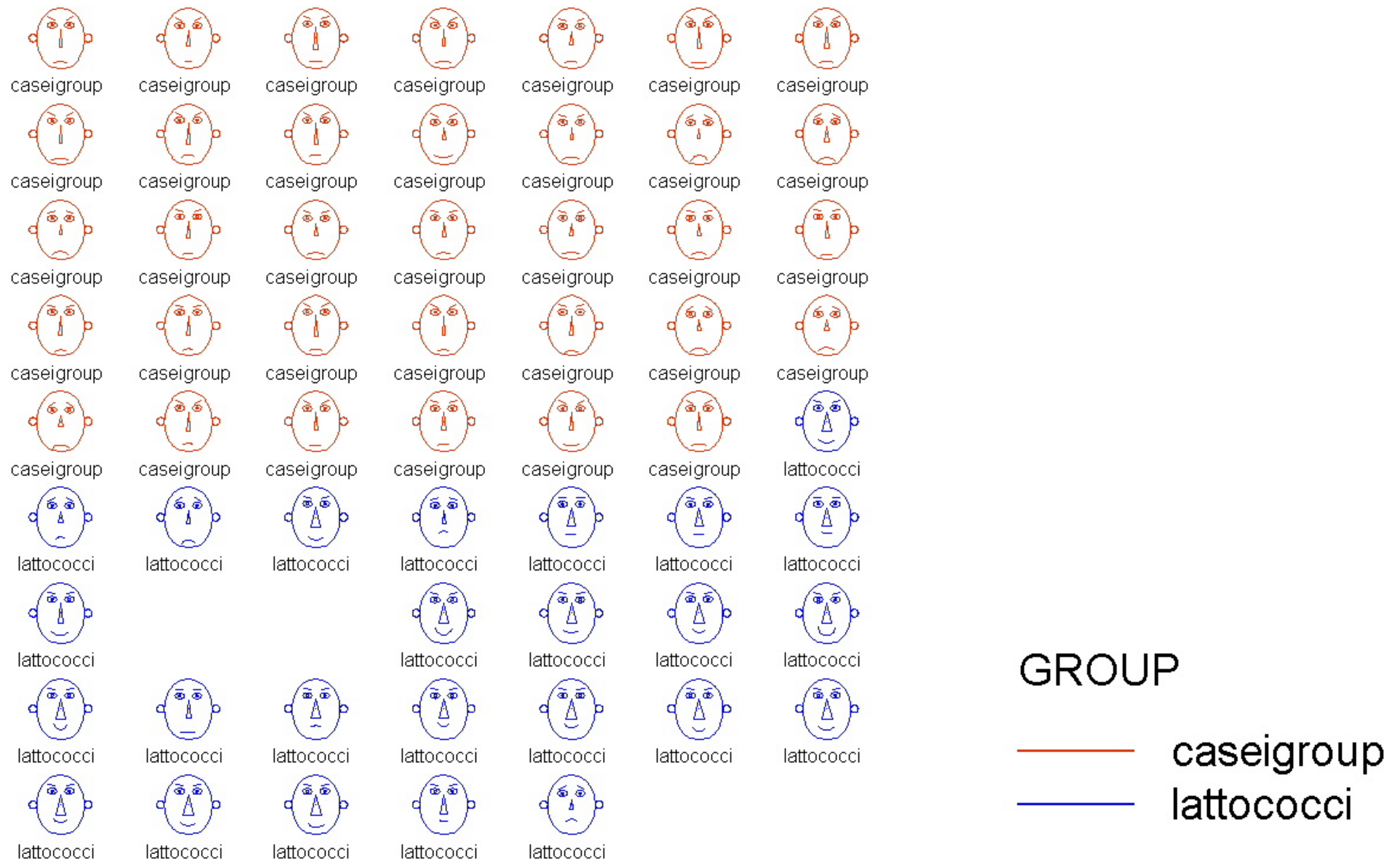
More on multivariate displays: icon (profile) plot (technolab example)



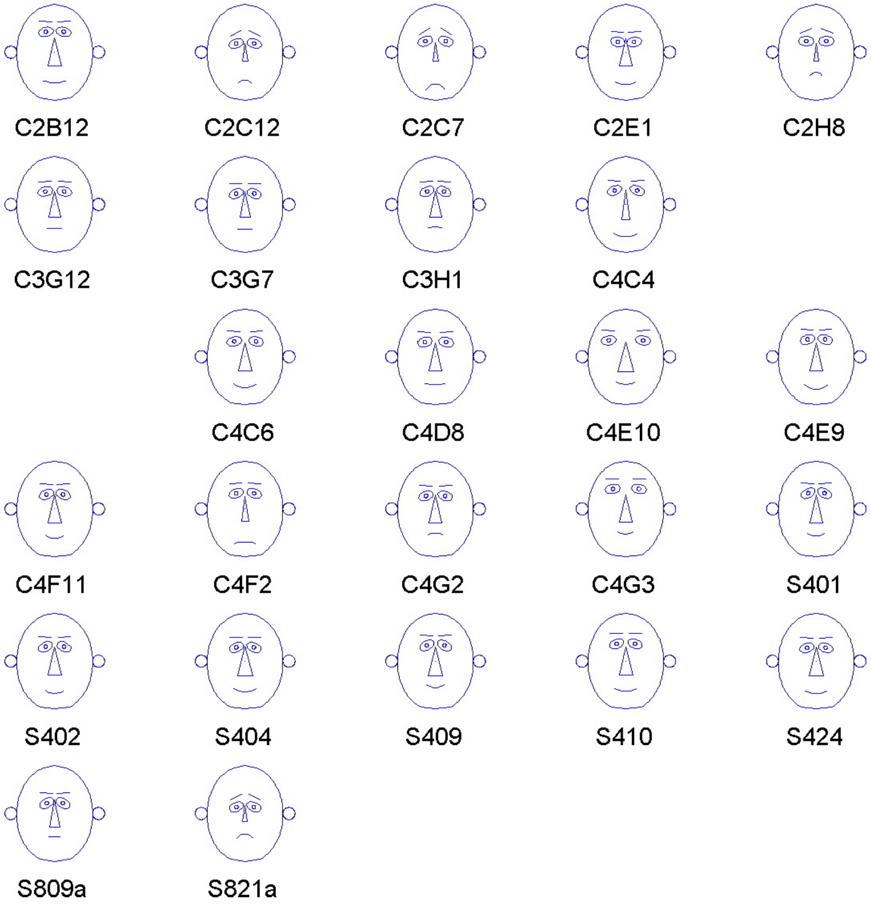
More on multivariate displays: icon (Fourier bubbles) plot (technolab example)



More on multivariate displays: icon (Chernoff' s faces) plot (technolab example)



More on multivariate displays: icon (Chernoff’ s faces) plot (technolab example)


C2B12
C2C12
C2C7
C2E1
C2H8
C3G12
C3G7
C3H1
C4C4
C4C6
C4D8
C4E10
C4E9
C4F11
C4F2
C4G2
C4G3
S401
S402
S404
S409
S410
S424
S809a
S821a