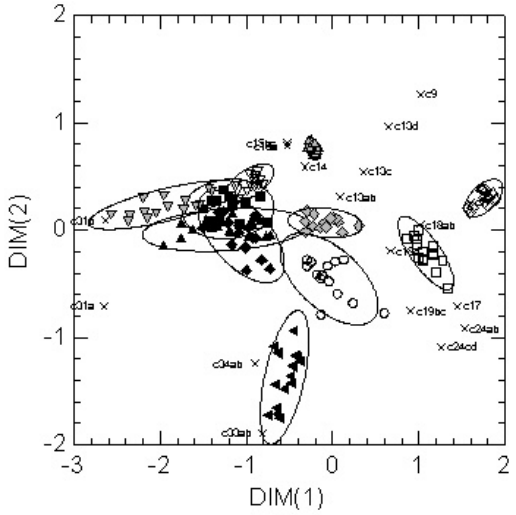
MULTIVARIATE STATISTICAL ANALYSIS FOR FOOD SCIENCE AND AGRICULTURE: AN INTRODUCTION

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)

Issues in cluster analysis

- choice of objects (observations)
- choice of variables
  - weighting of variables
  - standardization of variables
- choice of the dissimilarity/similarity measurement
- choice of the clustering method and inter-group proximity measures
- representation of the results of the analysis
- determination of the number of clusters
- comparison of dendrograms and measurement of distortion

Clustering techniques

- **hierarchical cluster analysis**: observations are partitioned in a series of nested clusters with hierarchical relationships, represented by a dendrogram.The procedure can be agglomerative or divisive (examples: single linkage analysis, UPGMA)
- **optimization methods**: individuals are partitioned into a predefined number of (mutually exclusive) clusters by optimization of a predefined criterion; the structure obtained is not usually hierarchical (examples: k-means)

Clustering techniques

- **finite mixture models**: the data are assumed to come from a mixture of density functions and the objective is to estimate the parameters of the mixture and to determine the posterior probabilities of cluster membership
- **density search clustering techniques**
- **clumping techniques**
- **fuzzy clustering**
- **Kohonen Self-Organizing Maps** (unsupervised artificial neural networks)

Similarity/dissimilarity measures for categorical (binary) data

	Individual i			
Individual j	Outcome	1	0	Total
	1	a	b	a+b
	0	c	d	c+d
	Total	a+c	b+d	p=a+b+c+d

Similarity/dissimilarity measures for categorical (binary) data: dichotomy coefficients

Measure	**Formula**
Simple matching coefficient	$S_{ij}=(a+d)/(a+b+c+d)$ **(S4)**
Jaccard coefficient	$S_{ij}=a/(a+b+c)$ **(S3)**
Rogers and Tanimoto (1960)	$S_{ij}=(a+d)/[a+2(b+c)+d]$ **(S6)**
Sokal and Sneath (1963)	$S_{ij}=a/[a+2(b+c)]$ **(S5)**
Gower and Legendre (1986)	$S_{ij}=(a+d)/[a+(b+c)/2+d]$
Gower and Legendre (1986)	$S_{ij}=a/[a+(b+c)/2]$ (Dice)

Dissimilarity and distance measures for continuous data

Minkowski distance (r=1, City block; r=2 Euclidean)

$$d = \sum_{i}^{p} \left[(x_i - y_i)^r \right]^{\frac{1}{r}}$$

Pearson correlation, with corresponding dissimilarity (1-r)/2, or other correlation measures for rank-order data (Spearman ρ, Goodman-Kruskal γ, Kendall τ)

Weighting variables

a. 0 or 1 weights to exclude/include variables
b. weights meant to represent the relative importance of variables
c. weights based on measures of dispersions of the variable
    - total variance (which included within and between group variance)
    - within-group variance

Examples

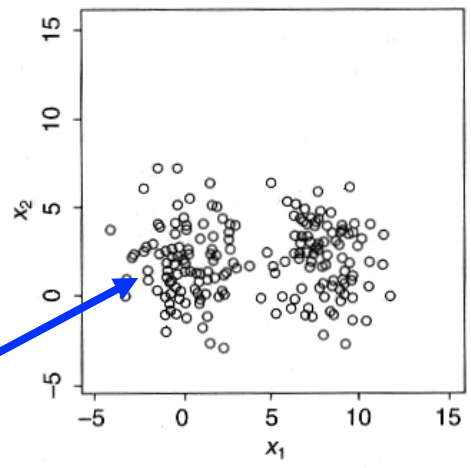
Some clustering procedures calculate the distances directly from the rectangular data matrix. However, some procedures (centroid, Ward) need original data and some similarity measures are not implemented in cluster modules.

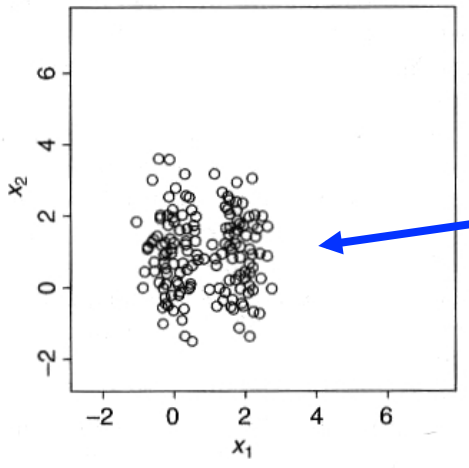
Open file feno4: it is an example of phenotypic characterization of LAB. The file is transposed into feno4t and then Jaccard similarity index can be calculated (saved in feno4S3) and used for clustering.

Other examples have been presented for the file simpledata.

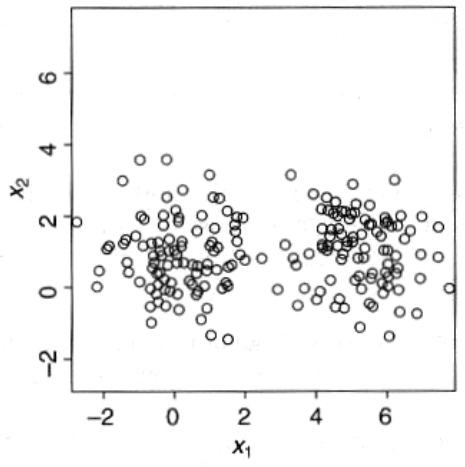
Effect of standardization and weighting

data on the original scale





weights based on total standard deviation



weights based on within-group standard deviations

Standardization

$$x_{ij} = \frac{x_{ij}}{\sum_{j=1}^{p} x_{ij}}$$

standardization based on total information or on range: robust, used for chromatographic data

$$x_{ij} = \frac{x_{ij}}{x_{ik}}$$

standardization based on an attribute k present for all observations which has relatively large values: may decrease significantly the contribution of the variables which have low values for all observations

Standardization

$$x_{ij} = \frac{x_{ij}}{s_j}$$

standardization based on standard deviation of each attribute: downweights variables with high standard deviation

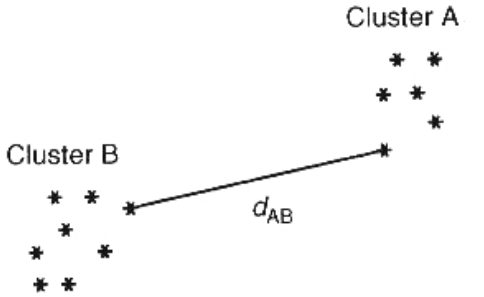
$$x_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

standardization based on z-scores: choice of mean and standard deviation is critical, see discussion on weighting

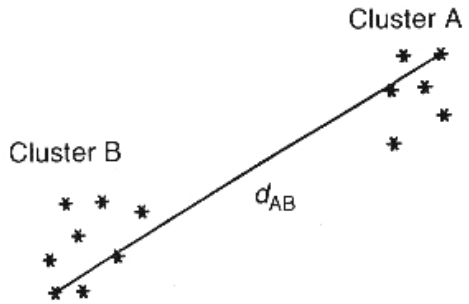
Hierarchical methods

Observations are partitioned in a series of nested clusters with hierarchical relationships, represented by a dendrogram.The procedure can be agglomerative or divisive (examples: single linkage analysis, UPGMA).
The choice made at each stage (agglomeration of two objects in a cluster, division of a cluster in two clusters or objects) is irreversible.
Eventually all objects will be aggregated in a single cluster and a decision needs to be made on the *optimal* number of clusters

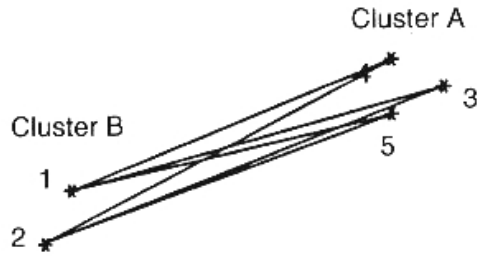
Examples of intergroup distances



single linkage
(nearest neighbour)



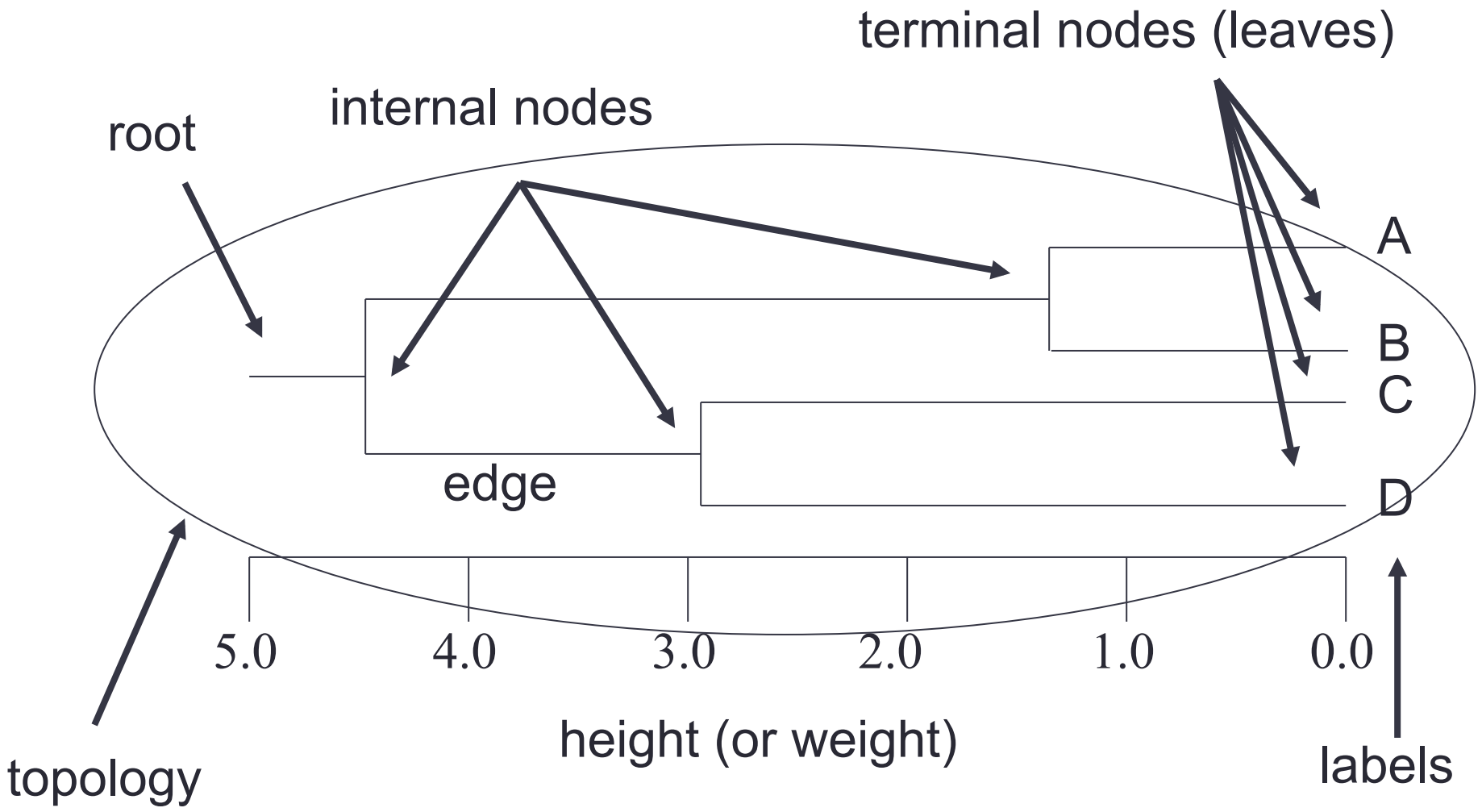
complete linkage
(furthest neighbour)



group average linkage
(unweighted pair-group method using the average approach, UPGMA)

$d_{AB}=(d_{13}+d_{14}+d_{15}+d_{23}+d_{24}+d_{25})/6$

Anatomy of a dendrogram



Intergroup distances (from Everitt et al. 2001)

Method	Alternative name	Usually used with	Distance between clusters defined as	Remarks
Single linkage	Nearest neighbour	Similarity or distance	Minimum distance between pair or objects, one in one cluster and one in another	Tendence to chaining, does not take account of cluster structure
Complete linkage	Furthest neighbour	Similarity or distance	Maximum distance between pair or objects, one in one cluster and one in another	Tends to find compact clusters with equal diameter, does not take account of cluster structure

Intergroup distances (from Everitt et al. 2001)

Method	Alternative name	Usually used with	Distance between clusters defined as	Remarks
(Group) Average Linkage	UPGMA	Similarity or distance	Average distance between pair or objects, one in one cluster and one in another	Tends to join clusters with small variances, intermediate between SL and CL, relatively robust, takes into account cluster structure
Centroid linkage	UPGMC	Distance (requires raw data)	Squared Euclidean distance between mean vectors (centroids)	Assumes points can be represented in Euclidean space. The more numerous of the two groups dominates the merged cluster

Intergroup distances (from Everitt et al. 2001)

Method	Alternative name	Usually used with	Distance between clusters defined as	Remarks
Median linkage	WPGMC (the same can be done with means)	Distance (requires raw data)	Squared Euclidean distance between weighted centroids (weighting is inversely proportional to n)	Assumes points can be represented in Euclidean space. New group intermediate in position between merged groups
Ward's method	Minimum sum of squares	Distance (requires raw data)	Increase in sum of squares within cluster, after fusion, summed for all variables	Assumes points can be represented in Euclidean space. Tends to find same size spherical clusters, sensitive to outliers

Ward's method

The objective at each stage is to minimize the increase in total within-cluster error sum of squares, E

$$E = \sum_{m=1}^{g} E_m$$

each of the g groups has n_m objects

$$E_m = \sum_{i=1}^{n_m} \sum_{k=1}^{p} (x_{ml,k} - \bar{x}_{m,k})^2$$

$$\bar{x}_{m,k} = (1/n) \sum_{l=1}^{n_m} x_{ml,k}$$

mean of the m^{th} cluster for the k^{th} object

Example of Single Linkage

$$\mathbf{D} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0.0 & & & & \\ 2.0 & 0.0 & & & \\ 6.0 & 5.0 & 0.0 & & \\ 10.0 & 9.0 & 4.0 & 0.0 & \\ 9.0 & 8.0 & 5.0 & 3.0 & 0.0 \end{pmatrix} \end{array}.$$

$$\mathbf{D}_1 = \begin{array}{c} \\ (12) \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} (12) & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0.0 & & & \\ 5.0 & 0.0 & & \\ 9.0 & 4.0 & 0.0 & \\ 8.0 & 5.0 & 3.0 & 0.0 \end{pmatrix} \end{array}.$$

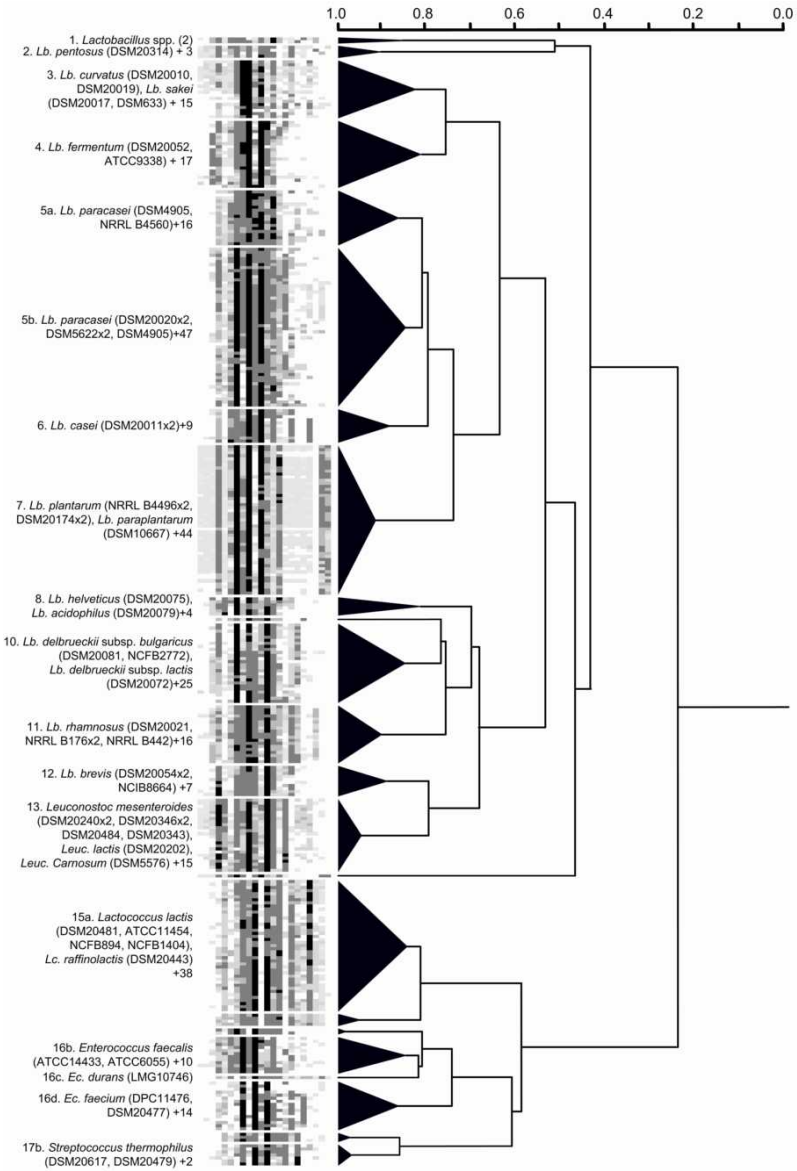
Example of Single Linkage

$$\mathbf{D}_2 = \begin{array}{c} \\ (12) \\ 3 \\ (45) \end{array} \begin{array}{c} \begin{array}{ccc} (12) & 3 & (45) \end{array} \\ \begin{pmatrix} 0.0 & & \\ 5.0 & 0.0 & \\ 8.0 & 4.0 & 0.0 \end{pmatrix} \end{array}.$$

In the following steps individual 3 and cluster 45 and then cluster 345 and 12 are merged

Examples

Open the file clusterLAB for examples of dendrograms calculated on the S3 matrix for the LAB data and for examples of dendrograms calculated on the Pearson correlation matrix of SDS-PAGE WCP patterns.



Cluster analysis of WCP patterns

ELSEVIER

Available online at www.sciencedirect.com
SCIENCE DIRECT

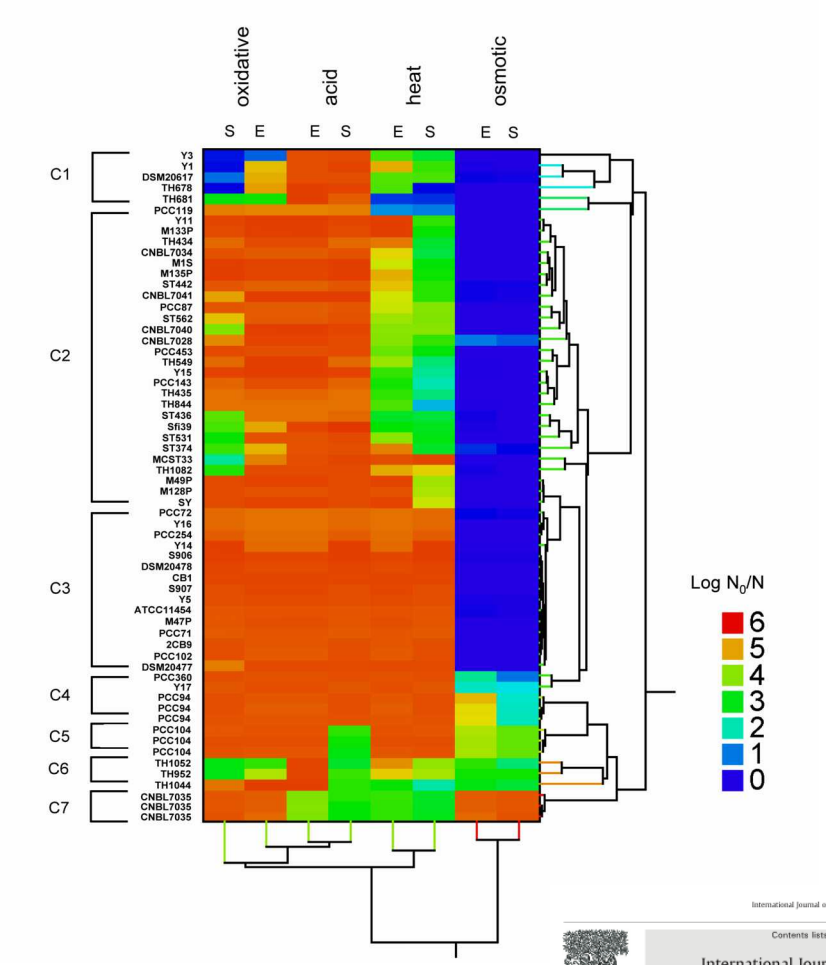
Journal of Microbiological Methods 66 (2006) 336 – 346

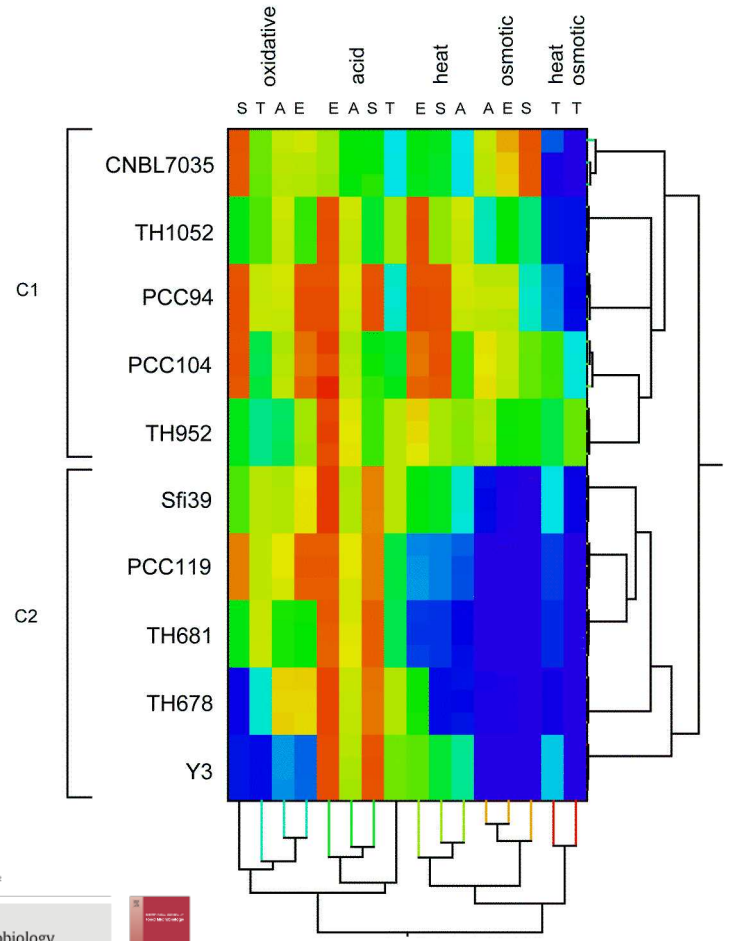
Journal of Microbiological Methods
www.elsevier.com/locate/jmicmeth

Use of unsupervised and supervised artificial neural networks for the identification of lactic acid bacteria on the basis of SDS-PAGE patterns of whole cell proteins

P. Piraino, A. Ricciardi, G. Salzano, T. Zotta, E. Parente *

Matrix cluster analysis of stress tolerance in streptococci






International Journal of Food Microbiology 124 (2008) 34–42

Contents lists available at ScienceDirect

International Journal of Food Microbiology

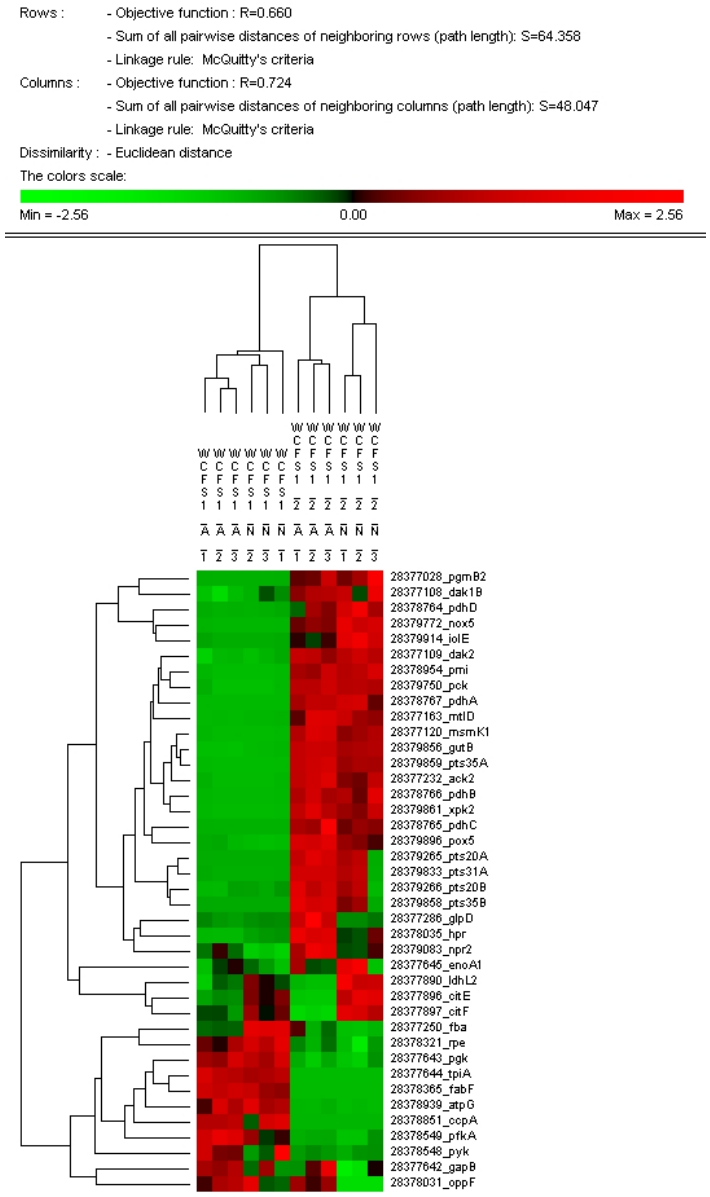
journal homepage: www.elsevier.com/locate/ijfoodmicro

ELSEVIER

Diversity of stress responses in dairy thermophilic streptococci

Teresa Zotta *, Annamaria Ricciardi, Felicia Ciocia, Rocco Rossano, Eugenio Parente

Dipartimento di Biologia, Difesa e Biotecnologie Agro-Forestali, Università degli Studi della Basilicata, Viale dell'Ateneo Lucano, 10, 85100 Potenza, Italy



Permutation analysis with Permutmatrix

- Permutmatrix is a Windows software for permutation analysis of data matrices
- Useful with proteomic and transcriptomic data
- Note how mutation and growth condition affects the pattern of protein expression to find group of proteins which are coregulated

Comparing dendrograms: use of the cophenetic matrix

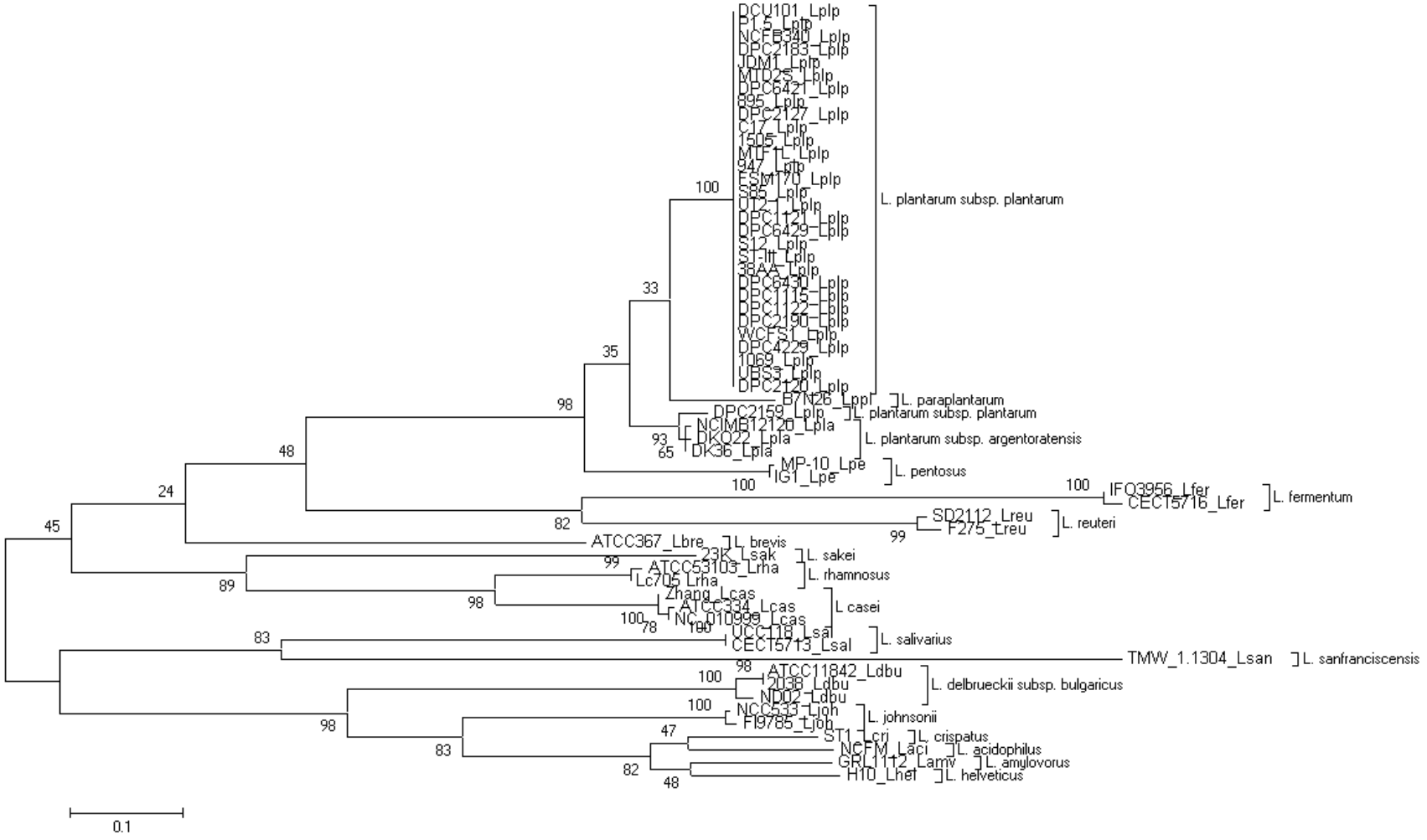
Entries in the cophenetic matrix are the heights h_{ij} at which objects i and j become members of the same cluster in the dendrogram. For the single linkage example:

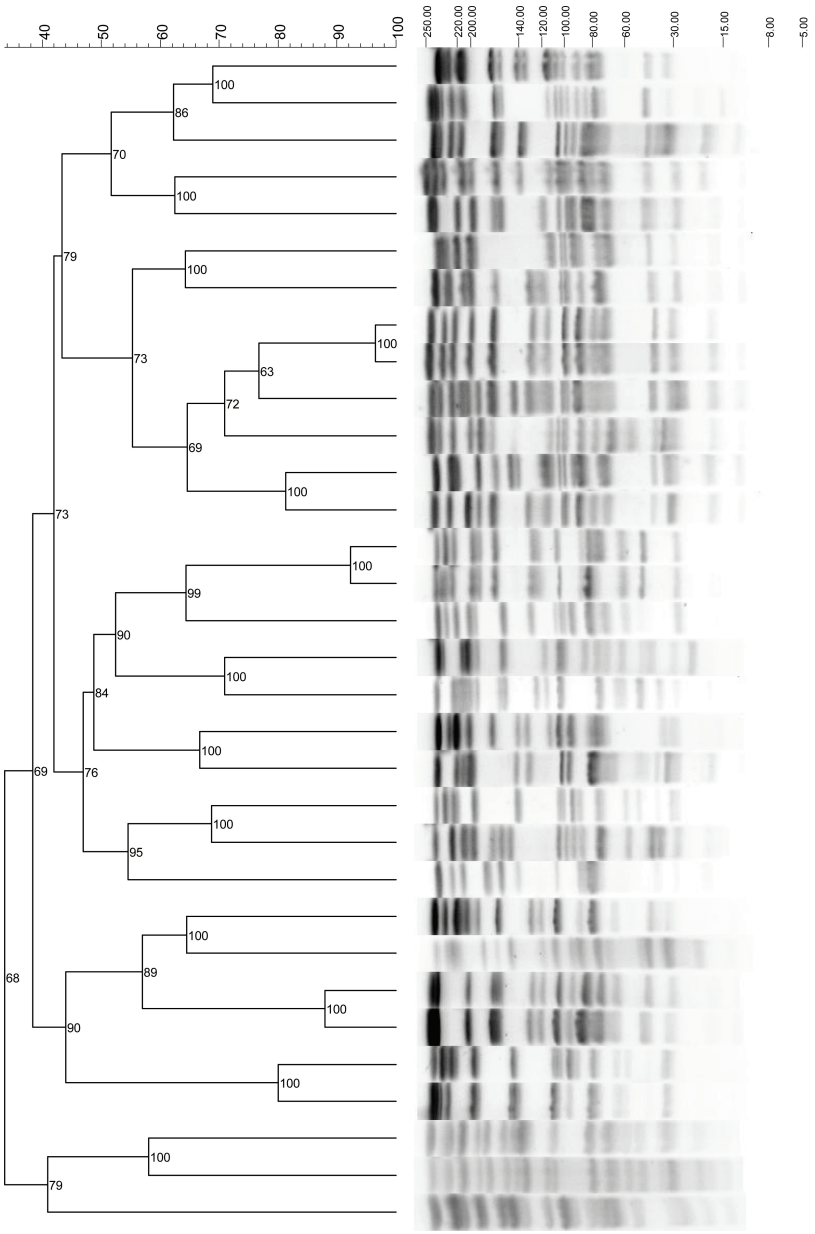
$$\mathbf{H} = \begin{pmatrix} 0.0 & & & & \\ 2.0 & 0.0 & & & \\ 5.0 & 5.0 & 0.0 & & \\ 5.0 & 5.0 & 4.0 & 0.0 & \\ 5.0 & 5.0 & 4.0 & 3.0 & 0.0 \end{pmatrix}$$

The cophenetic correlation is the Pearson correlation between the vector formed by the lower triangular matrix of the original dissimilarity matrix and the corresponding vector of the cophenetic matrix

An alternative is using Goodman-Kruskal's γ instead of r

Comparing dendrograms: bootstrapping. ML tree of HrcA sequences for lactobacilli produced by MEGA 5





SUBSPECIES	STRAIN
plantarum	MTF1L
plantarum	UT2.1
plantarum	DPC6430
plantarum	MTD2S
plantarum	P1.5
plantarum	DPC6429
plantarum	WCFS1
plantarum	DPC1122
plantarum	UBS3
plantarum	DPC1115
plantarum	1069
plantarum	DPC6421
plantarum	C17
plantarum	DPC2120
plantarum	DPC2190
plantarum	DPC4229
plantarum	S12
plantarum	DPC2183
plantarum	1505
plantarum	DPC2127
plantarum	DPC1121
plantarum	38AA
plantarum	DPC2159
plantarum	FSM170
plantarum	S85
plantarum	947
plantarum	895
plantarum	DCU101
plantarum	NCF340
argentoratensis	DKO22
argentoratensis	DK36
argentoratensis	NCIMB12120

Bootstrapped dendrogram of PFGE patterns of *L. plantarum* strains obtained with Gelcompar

Other properties of well-behaved dendrograms

Ultrametric property: for any three objects, the two largest distances between objects are equal (rarely holds for dissimilarity matrices); if this properties does not hold **inversions** or **reversals** (fusions do not follow a monotonic sequence, a later fusion takes place at a lower level of dissimilarity) are possible (can happen in centroid and median clustering.

Space distortion: with space contraction dissimilar objects are drawn in the same cluster (chaining in single linkage); with space dilation similar objects are drawn in different clusters (complete linkage)

Clump admissibility: "there exist a clustering such that all within-cluster distances are smaller than all between cluster distances"

Other properties of well-behaved dendrograms

Convex admissibility: if the objects can be represented in the Euclidean space, the convex hulls of the partitions never intersect (single and complete linkage do not have this property).

Point proportion admissibility: replication of points does not alter the boundaries of partitions (UPGMA and UPGMC do not have this property, while the corresponding methods using weights do)

Monotone admissibility: monotonic transformations of the elements in the proximity matrix does not alter the clustering; this is appropriate when the rank-order information is reliable while the distance information is not (for examples judges evaluating preferences on different scales)

How many clusters?

In hierarchical cluster analysis choice of the number of clusters is equivalent to finding the “best cut” of the height of the dendrogram.
An informal solution is finding the height in the dendrogram where large changes in fusion level occur. This is subjective, but uses previous knowledge on the data structure.
Several formal approaches have been proposed, usually based on the selection of the first stage of the dendrogram satisfying some numerical value.
Bootstrap analysis can also be used.
No method (formal or informal) works equally well for all dendrograms or for all datasets

Optimization clustering techniques

Individuals are allocated to a specified number of clusters by minimizing or maximizing a numerical criterium.

Methods differ in:

1. the criteria to be optimized, which can be derived from the dissimilarity matrix (lack of homogeneity, separation) or directly on the continuous data
2. the optimization algorithm (iterative procedures are used)

Individuals can be re-allocated during the process.

Criteria derived from continuous data

Given a *nxp* matrix **X** the *pxp* total dispersion matrix **T** can be decomposed in the within-group dispersion matrix **W** and in the between group dispersion matrix **B**

$$\mathbf{T} = \sum_{m=1}^{g}\sum_{l=1}^{n_m}\left(\mathbf{x}_{ml} - \bar{\mathbf{x}}\right)\left(\mathbf{x}_{ml} - \bar{\mathbf{x}}\right)$$

$$\mathbf{W} = \sum_{m=1}^{g}\sum_{l=1}^{n_m}\left(\mathbf{x}_{ml} - \bar{\mathbf{x}}_m\right)\left(\mathbf{x}_{ml} - \bar{\mathbf{x}}_m\right) \qquad \mathbf{T=W+B}$$

$$\mathbf{B} = \sum_{m=1}^{g} n_m\left(\bar{\mathbf{x}}_m - \bar{\mathbf{x}}\right)\left(\bar{\mathbf{x}}_m - \bar{\mathbf{x}}\right)'$$

Criteria derived from continuous data

1. minimization of trace(**W**): minimization of the within-cluster sum-of-squares, equivalent to maximization of trace(**B**) and to minimization of the Euclidean distances between individuals and their group mean
2. minimization of det(**W**): analogous to trying to find group mean vectors which differ significantly, to maximize det(**T**)/det(**W**) (a test criterion in MANOVA)
3. maximization of trace($\mathbf{BW}^{-1}$): in MANOVA large values of trace($\mathbf{BW}^{-1}$) indicate that group mean vectors differ

Criteria derived from continuous data

1. minimization of trace(**W**): scale sensitive, tends to find spherical clusters even if the shape of the natural clusters differ, tends to find groups with roughly the same number of objects (if clusters are close)
2. minimization of det(**W**): scale insensitive, not restricted to spherical clusters, tends to find groups of the same shape with roughly the same number of objects (if clusters are close)
3. maximization of trace($\mathbf{BW}^{-1}$): scale insensitive, not restricted to spherical clusters, tends to find groups with roughly the same number of objects (if clusters are close)

Optimization algorithms

The number of different partitions of *n* objects in *g* clusters is very large, even with relatively small values of *n* and *g*:

$$N(n,g) = \frac{1}{g!}\sum_{m=1}^{g}(-1)^{g-m}\binom{g}{m}m^{n}$$

N(5,2)=15
N(10,3)=9330
N(50,4)=5.3x10^{28}

Optimization algorithms

1. find an initial partitionof n objects in g groups

↓

2. calculate the change in clustering criterion obtained by moving each object to another group

↓

3. make the change which results in thegreatest improvement of the clustering criterion

↓

4. repeat steps 1. and 2. until no furtherimprovement is obtained

k-means

k-means (SYSTAT implementation) begins with a single cluster and starts a new cluster by selecting the object which is farthest from the group mean (centroid); cases are allocated between the two clusters on the basis of their distance from the group centre. A new cluster is started by dividing one of the clusters with the same approach and the process is repeated until the specified number of clusters (g) is formed. The reassignment continues until no further change in within-group sum of squares is obtained

k-means

k-means (SYSTAT implementation) provides a choice of a large number of distance measures:

1. continuous data: Euclidean, Minkowski, Pearson (1-r), Rsquared ($1-r^2$)
2. rank-order data: Goodman-Kruskal γ
3. counts of objects or events: Chi-square (χ^2 for independence of rows and columns of a 2xn table of pair of cases), Phi-square (χ^2/n)

Distance metric is Euclidean distance

k-means splitting cases into 6 groups
Summary statistics for all cases

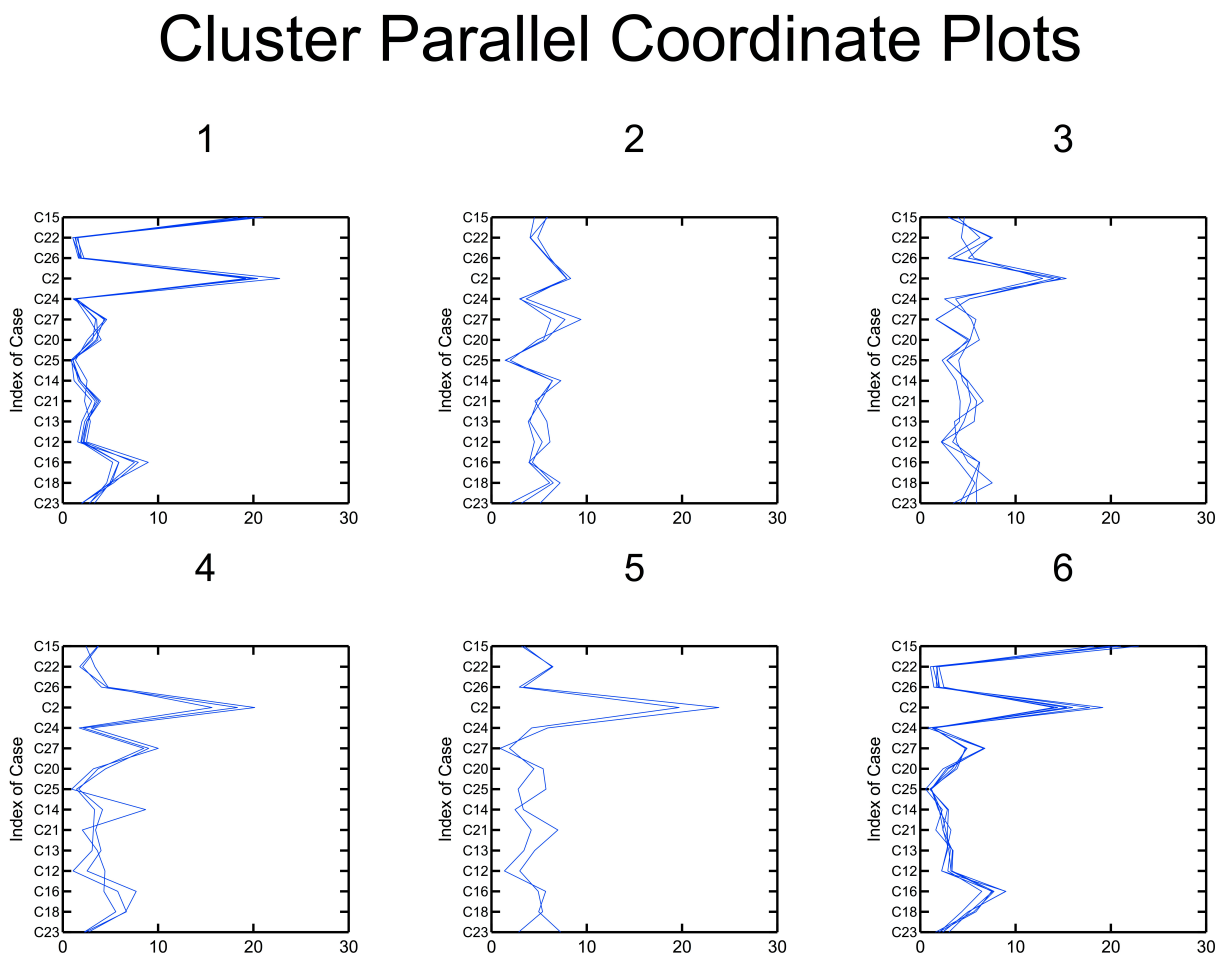
Variable	Between SS	df	Within SS	df	F-ratio
C2	395.815	5	49.563	18	28.750
C12	22.970	5	11.963	18	6.913
C13	14.767	5	6.615	18	8.037
C14	70.783	5	20.467	18	12.450
C15	1616.916	5	20.989	18	277.333
C16	37.979	5	24.600	18	5.558
C18	9.064	5	6.115	18	5.336
C20	21.987	5	5.609	18	14.113
C21	34.594	5	11.697	18	10.647
C22	103.641	5	9.406	18	39.668
C23	24.843	5	21.095	18	4.240
C24	43.267	5	7.829	18	19.895
C25	25.353	5	6.749	18	13.524
C26	54.119	5	6.253	18	31.159
C27	118.996	5	30.013	18	14.273
** TOTAL **	2595.097	75	238.962	270	

Cluster 1 of 6 contains 6 cases

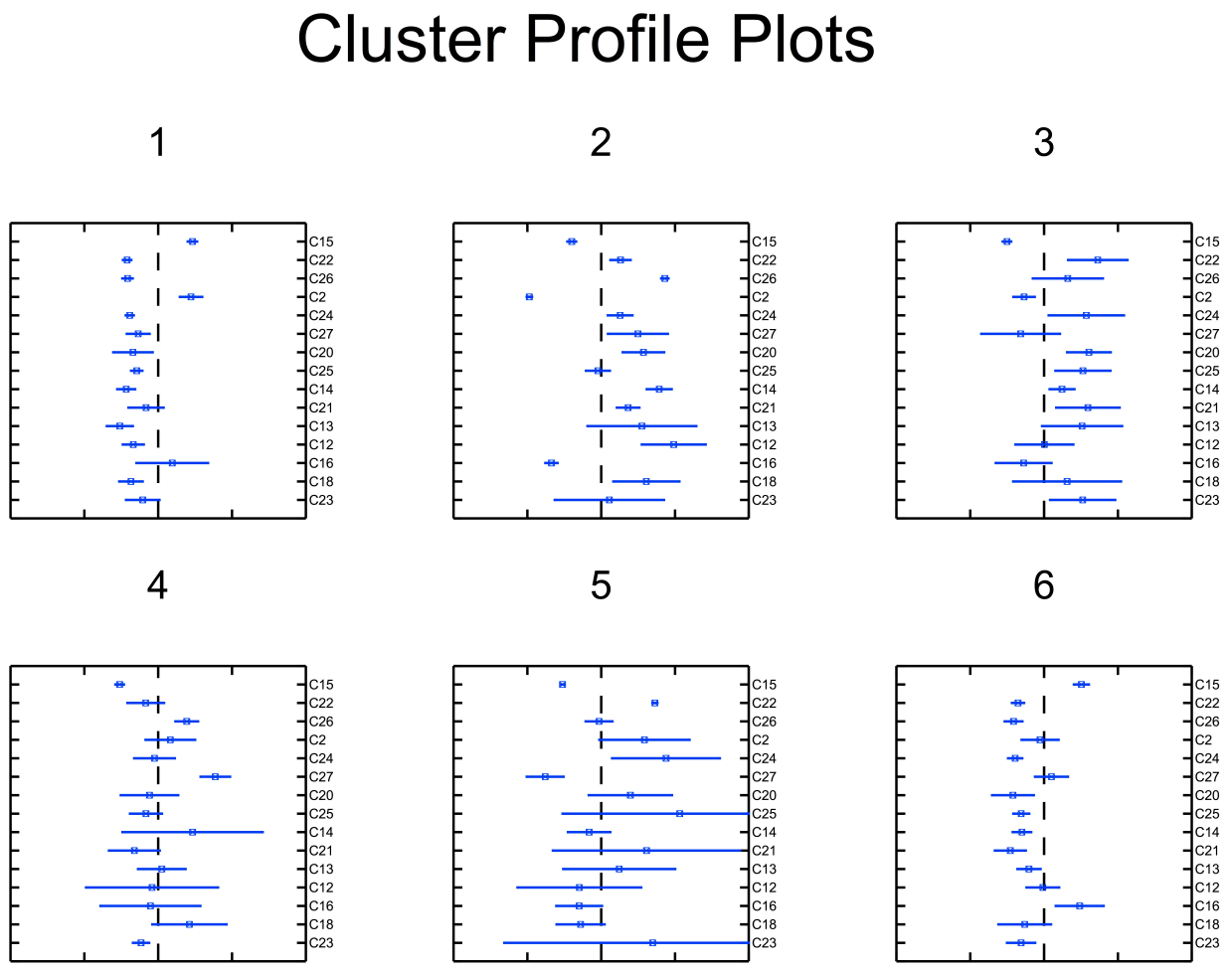
Members			Statistics				
Case	Distance	Variable	Minimum	Mean	Maximum	St.Dev.
DC1	0.54	C2	19.19	20.51	22.78	1.21
RC4	0.73	C12	1.57	2.06	2.50	0.32
RC2	0.35	C13	2.00	2.49	2.90	0.31
RC1	0.89	C14	1.20	1.82	2.55	0.44
CC2	0.58	C15	18.38	19.95	21.04	0.98
CC1	0.36	C16	5.25	6.91	8.97	1.45
		C18	4.62	5.00	5.33	0.23
		C20	2.59	3.32	4.04	0.53
		C21	2.29	3.35	3.96	0.61
		C22	1.03	1.30	1.59	0.22
		C23	1.96	2.48	3.33	0.58
		C24	1.09	1.29	1.43	0.15
		C25	0.86	1.05	1.33	0.16
		C26	1.64	1.89	2.23	0.20
		C27	2.80	3.87	4.61	0.71

k-means output (Surface ripened cheese example)

k-means output (Surface ripened cheese example)



k-means output (Surface ripened cheese example)



Examples

Open the file clusterLAB for examples of k-means clustering based on Euclidean distance for standardized (range) technological properties of LAB.
In the same file: examples of k-means clustering on RP-HPLC data of smear cheese.

How many clusters?

Informal criteria:

1. perform a hierarchical cluster analysis to identify number of clusters in the dendrogram at some specified height;
2. look at some sort of density display of the original data or of scores after PCA, or coordinates in MDS
3. use a sort of scree plot: plot the final value of the clustering criterion against the number of clusters and look for elbows in the plot

Formal criteria:

Choose the number of clusters on the basis of a numerical index. For example, choose the number of clusters which maximizes C(g)

$$C(g) = \frac{trace(\mathbf{B})}{g-1} \Big/ \frac{trace(\mathbf{W})}{n-g}$$

Some rights reserved