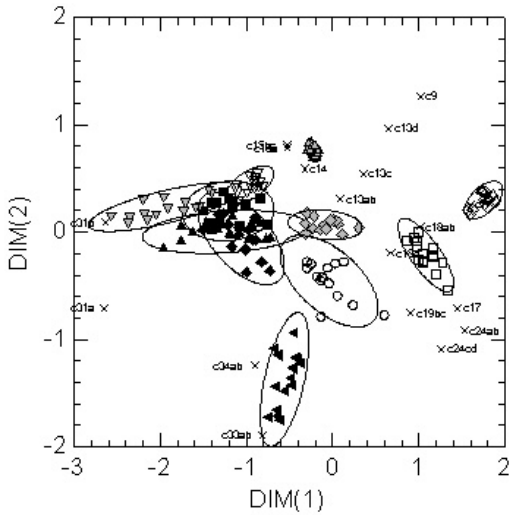
MULTIVARIATE STATISTICAL ANALYSIS FOR FOOD SCIENCE AND AGRICULTURE: AN INTRODUCTION
4. MULTIDIMENSIONAL SCALING

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



Outline

- **Multidimensional scaling (MDS)**
  - objectives of MDS
  - metric and monotonic MDS
  - MDS output
  - examples

Multidimensional scaling

Multidimensional scaling is a group of techniques used to fit a set of points in a q-dimensional space such as the distance between the points (δ_{ij}) matches as closely as possible the dissimilarity (d_{ij}) between the original objects in the p-dimensional space **in order to obtain a simple spatial model** (map).

The model does not need statistical distribution assumptions, but the data should satisfy metric conditions:

- distance from an object with itself is 0
- distance from object A to object B is the same as the distance of B from A (the dissimilarity matrix should be simmetrical)
- distance from A to C is less or equal than the sum of distances between A to B and B to C (triangle inequality)

MDS vs. PCA

- Advantages
  - MDS will usually find a solution with less dimensions compared to PCA
  - If the purpose is just finding or visualizing natural groups of objects, a map is easier to explain than a score plot
  - Cluster of objects are easier to visualize / highlight in a MDS map
- Disadvantages
  - It is usually more difficult to find relationships between dimensions and original variables
  - MDS is usually effective only when observation are fairly well spread out in the space

Multidimensional scaling

The model can be written as:

$$\delta_{ij} = f(d_{ij})$$

$$d_{ij} = h(\mathbf{x}_i, \mathbf{x}_j)$$

where:

$\mathbf{x}_i$ and $\mathbf{x}_j$ are the vectors of the coordinates of objects i and j in the q-dimensional space (q<p)

$f(d_{ij})$ is the assumed functional relationship between the dissimilarities and the distances

h is the distance function (usually Euclidean, but Minkowski metrics can be used)

Metric multidimensional scaling

A direct numerical comparison between fitted distances and dissimilarities (usually based on a least squares criterion) is used. The coordinates are iteratively calculated to minimize a goodness of fit statistics (**stress**). In linear metric scaling a linear model is used to relate distances to dissimilarities:

$$d_{ij} = \alpha + \beta\delta_{ij} + \varepsilon_{ij}$$

Metric multidimensional scaling

A fit criterion which is invariant both under rigid transformations (rotations, reflections, translations) and under non-rigid transformations (stretching and shrinking obtained by multiplications of the coordinate by a factor k) is:

$$S_3 = \frac{\sum_{i<j} \left(d_{ij} - f\left(\delta_{ij}\right)\right)^2}{\sum_{i<j} d_{ij}^2}$$

whose square root is known as **stress**

Non-metric (monotonic) MDS

When observed proximities contain information on rank order rather than on real distances, assuming a linear relationship between observed and fitted distances may be inappropriate and monotonic regression should be used

$$d_{ij} = \hat{d}_{ij} + \varepsilon_{ij}$$

$$\delta_{i_1,j_1} < \delta_{i_2,j_2} < \ldots \delta_{i_N,j_N}$$

$$\hat{d}_{i_1 j_1} \leq \hat{d}_{i_2 j_2} \leq \ldots \leq \hat{d}_{i_N j_N}$$

the fitted distances are chosen to represent a weak monotonicity condition

MDS input data

- Dissimilarity matrices obtained in a direct way (ask assessors to state how different two objects are, take measurement from a map, etc.)
- Dissimilarity matrices calculated from rectangular (*n* x *p*) data matrices
  - Euclidean distance (on standardized or unstandardized data)
  - Negative correlation (beware, high positive correlation = most similar, high negative correlation = most dissimilar)
  - Other distance or correlation measures (Spearman, Guttman)

Adjustable parameters in the analysis

- Type of scaling (monotonic, metric: linear, log, power)
- Number of dimensions
- Stress function
- Iteration and convergence parameters

MDS output

In the MDS output look for:

- **final configuration** in the q dimensions (q<p), coordinates and plots
- **final stress and proportion of the variance**; according to Kruskal (1964)
  - stress 0.20 -> poor fit
  - stress 0.10 -> fair fit
  - stress 0.05 -> good fit
  - stress 0.025 -> excellent fit
- **Shepard diagram** (plot of observed vs fitted distances)

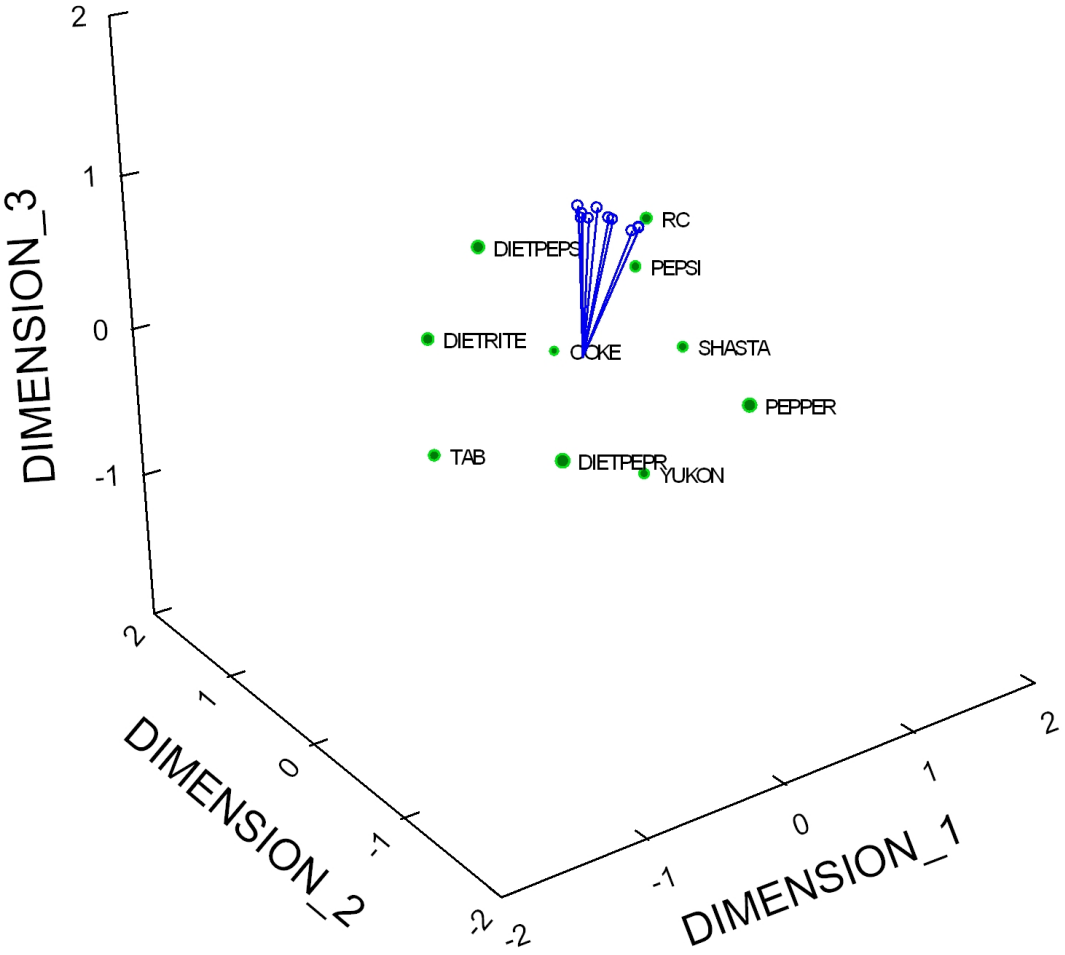
Individual differences MDS

- Uses multiple dissimilarity matrices (for example different judges evaluating a common set of ptoducts)
- The input is a rectangular matrix containing stacked triangular dissimilarity matrices
- Scales both objects in a common space in order to
  - Find a common configuration for all objects
  - Calculate weights for judges in the common space
  - Assess goodness of fit for both objects and judges

The data file

	DIETPEPS	RC	YUKON	PEPPER	SHASTA	COKE	DIETPEPR	TAB	PEPSI	DIETRITE
1	0.000	.	.	.	.	.	.	.	.	.
2	16.000	0.000	.	.	.	.	.	.	.	.
3	81.000	47.000	0.000	.	.	.	.	.	.	.
4	56.000	32.000	71.000	0.000	.	.	.	.	.	.
5	87.000	68.000	44.000	71.000	0.000	.	.	.	.	.
6	60.000	35.000	21.000	98.000	34.000	0.000	.	.	.	.
7	84.000	94.000	98.000	57.000	99.000	99.000	0.000	.	.	.
8	50.000	87.000	79.000	73.000	19.000	92.000	45.000	0.000	.	.
9	99.000	25.000	53.000	98.000	52.000	17.000	99.000	84.000	0.000	.
10	16.000	92.000	90.000	83.000	79.000	44.000	24.000	18.000	98.000	0.000
11	0.000	.	.	.	.	.	.	.	.	.
12	9.000	0.000	.	.	.	.	.	.	.	.
13	90.000	70.000	0.000	.	.	.	.	.	.	.
14	87.000	65.000	6.000	0.000	.	.	.	.	.	.
15	87.000	77.000	83.000	83.000	0.000	.	.	.	.	.
16	33.000	79.000	25.000	89.000	39.000	0.000	.	.	.	.
17	86.000	86.000	99.000	22.000	90.000	40.000	0.000	.	.	.
18	81.000	30.000	57.000	88.000	69.000	39.000	97.000	0.000	.	.
19	74.000	20.000	94.000	78.000	5.000	81.000	92.000	88.000	0.000	.

The configuration



The output

Monotonic Multidimensional Scaling
Kruskal Method
The data are analyzed as dissimilarities
There are 10 replicated data matrices
Dimensions are weighted separately for each matrix
Fitting is split between data matrices
Minimizing Kruskal STRESS (form 1) in 3 dimensions

Iteration History

Iteration	STRESS
0	0.22090
1	0.18442
0	0.22131
1	0.18451

Stress of Final Configuration : 0.18451
Proportion of Variance (RSQ) : 0.53501

Coordinates in 3 Dimensions

Variable	Dimension		
	1	2	3
DIETPEPS	-0.60820	0.19557	0.77706
RC	0.52175	0.05235	0.75639
YUKON	0.41586	-0.08904	-0.86786
PEPPER	0.27187	-1.26587	0.05912
SHASTA	0.79785	0.02490	-0.14379
COKE	0.39073	0.83659	-0.34734
DIETPEPR	-0.74711	-0.84291	-0.17340
TAB	-0.79097	0.43843	-0.60917
PEPSI	0.57067	0.22100	0.38103
DIETRITE	-0.82245	0.42898	0.16795

Matrix Weights

Matrix	Stress	RSQ	Dimension		
			1	2	3
1	0.18837	0.54776	0.69776	0.43369	0.52679
2	0.19981	0.41627	0.45210	0.46536	0.72123
3	0.19643	0.46782	0.34785	0.52299	0.73932
4	0.17068	0.56431	0.59123	0.49248	0.60823
5	0.17816	0.59439	0.70448	0.37011	0.56384
6	0.17191	0.62137	0.70417	0.36761	0.57019
7	0.18107	0.55169	0.41948	0.58226	0.65912
8	0.18046	0.55973	0.48352	0.59725	0.60864
9	0.16326	0.62452	0.56269	0.49556	0.62580
10	0.21166	0.40227	0.43525	0.60937	0.61744

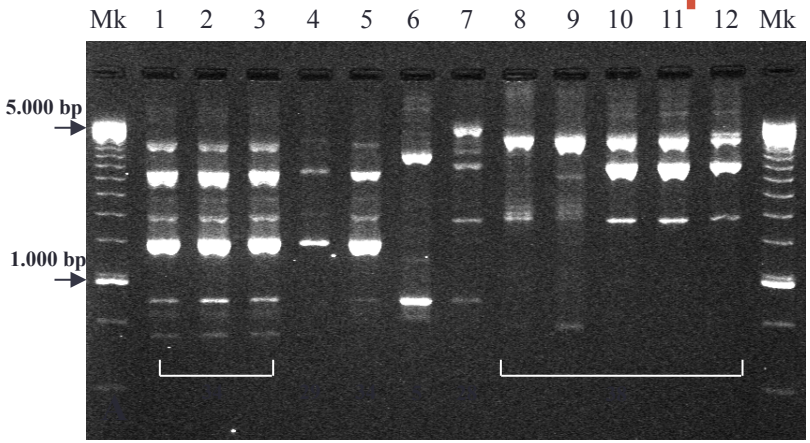
MDS examples

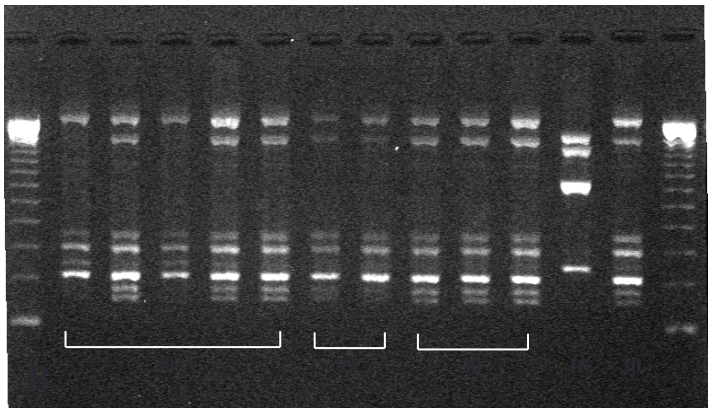
Open file MDSRAPD.syo for examples of MDS on RAPD-PCR data (including bootstrapping/MDS procedure, which can be found in file sardiniabread.syo);
look at the original data and final configuration for the MDS on RAPD data in file breadlab.xls;
look at file MDSboot.xls for the final bootstrapping/ MDS configuration;
look at the command files for details on the procedures of pretreatment (boot.syc, Bread\fornextloop.syc)

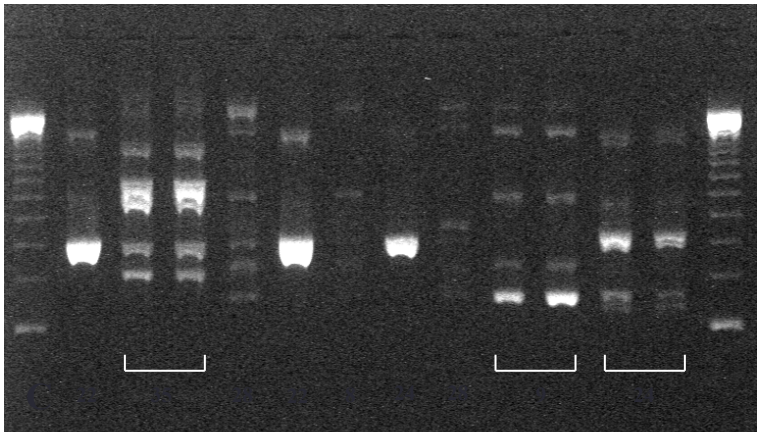
MDS examples

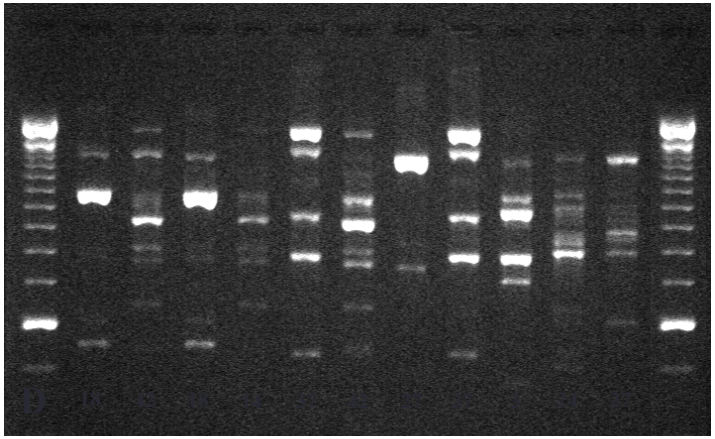
Open file mds.syo for MDS examples on the RP-HPLC dataset for smear cheese.

MDS of RAPD patterns of LAB









ELSEVIER

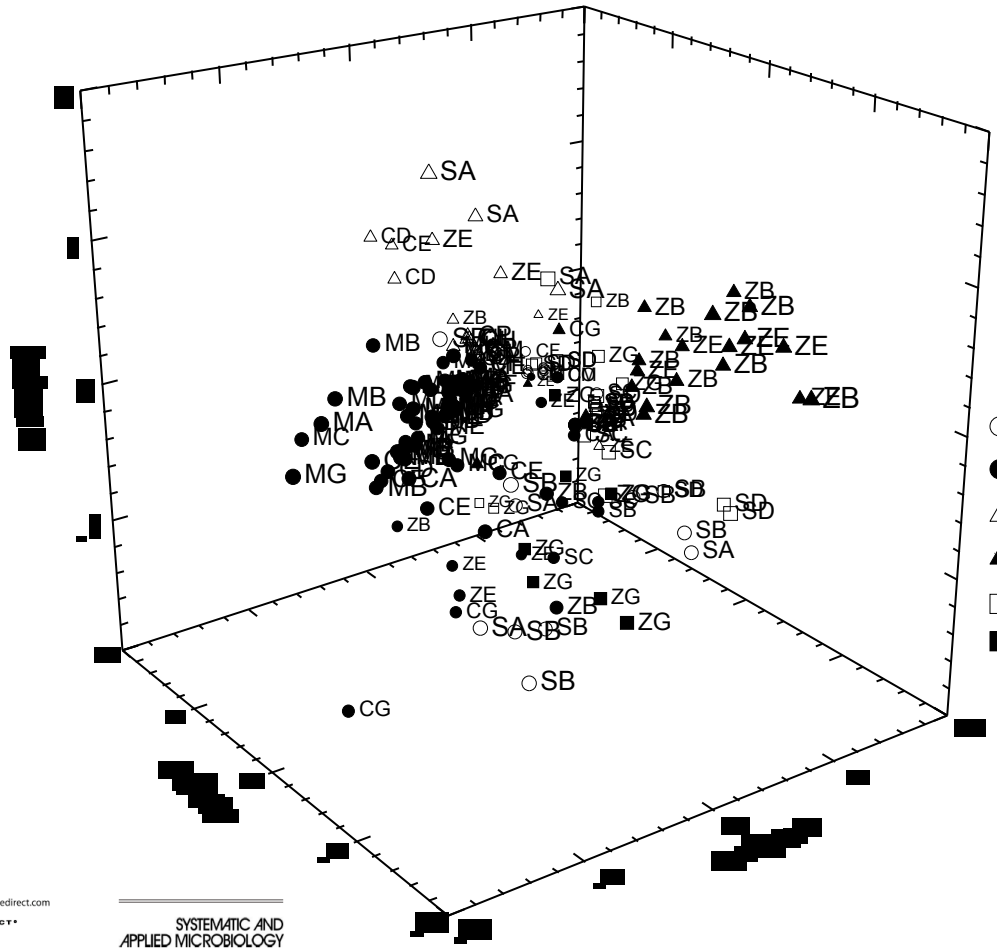
Available online at www.sciencedirect.com
SCIENCE DIRECT®
Systematic and Applied Microbiology 29 (2006) 138–144

SYSTEMATIC AND APPLIED MICROBIOLOGY
www.elsevier.de/syapm

Molecular characterization of lactic acid bacteria from sourdough breads produced in Sardinia (Italy) and multivariate statistical analyses of results

Pasquale Catzeddu[a,*], Enrica Mura[a], Eugenio Parente[b], Manuela Sanna[a], Giovanni Antonio Farris[c]

MDS of RAPD patterns of LAB

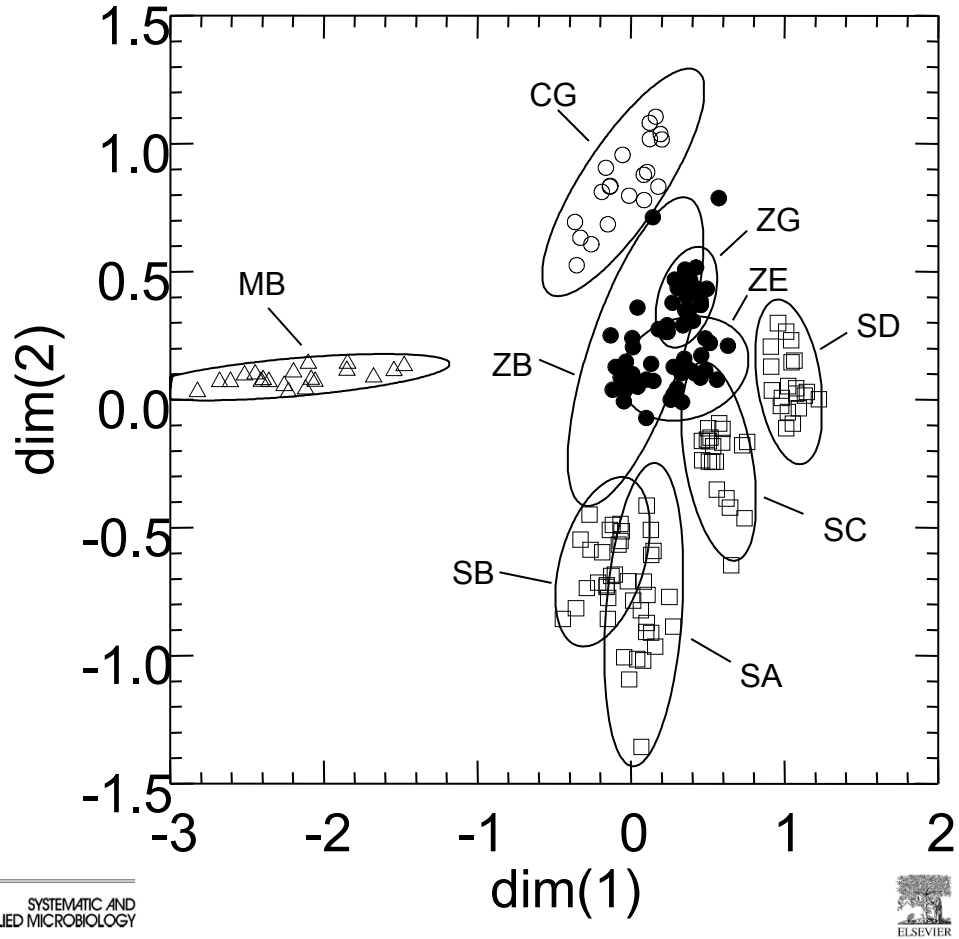


Available online at www.sciencedirect.com
SCIENCE DIRECT
Systematic and Applied Microbiology 29 (2006) 138–144
SYSTEMATIC AND APPLIED MICROBIOLOGY
www.elsevier.de/syapm

Molecular characterization of lactic acid bacteria from sourdough breads produced in Sardinia (Italy) and multivariate statistical analyses of results

Pasquale Catzeddu[a,*], Enrica Mura[a], Eugenio Parente[b], Manuela Sanna[a], Giovanni Antonio Farris[c]

MDS of RAPD patterns of LAB



ELSEVIER
Available online at www.sciencedirect.com
SCIENCE DIRECT
Systematic and Applied Microbiology 29 (2006) 138–144
SYSTEMATIC AND APPLIED MICROBIOLOGY
www.elsevier.de/syapm

Molecular characterization of lactic acid bacteria from sourdough breads produced in Sardinia (Italy) and multivariate statistical analyses of results

Pasquale Catzeddu[a,*], Enrica Mura[a], Eugenio Parente[b], Manuela Sanna[a], Giovanni Antonio Farris[c]

ELSEVIER
Journal of Microbiological Methods 49 (2002) 121–134
Journal of Microbiological Methods
www.elsevier.com/locate/jmicmeth

A statistical procedure for the analysis of microbial communities based on phenotypic properties of isolates

Eugenio Parente*, Annamaria Ricciardi
Dipartimento di Biologia, Difesa e Biotecnologie Agro-Forestali, Università della Basilicata, Campus di Macchia Romana, 85100 Potenza, Italy

Received 13 June 2001; received in revised form 4 October 2001; accepted 4 October 2001

MDS of RP-HPLC data from smear cheese

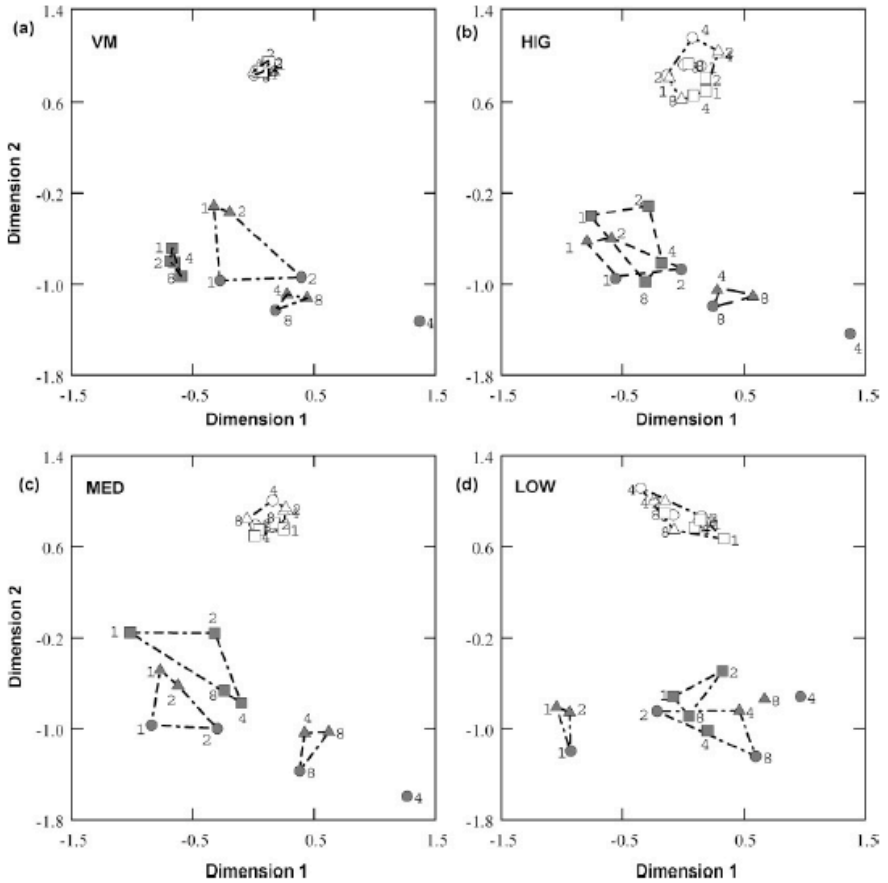
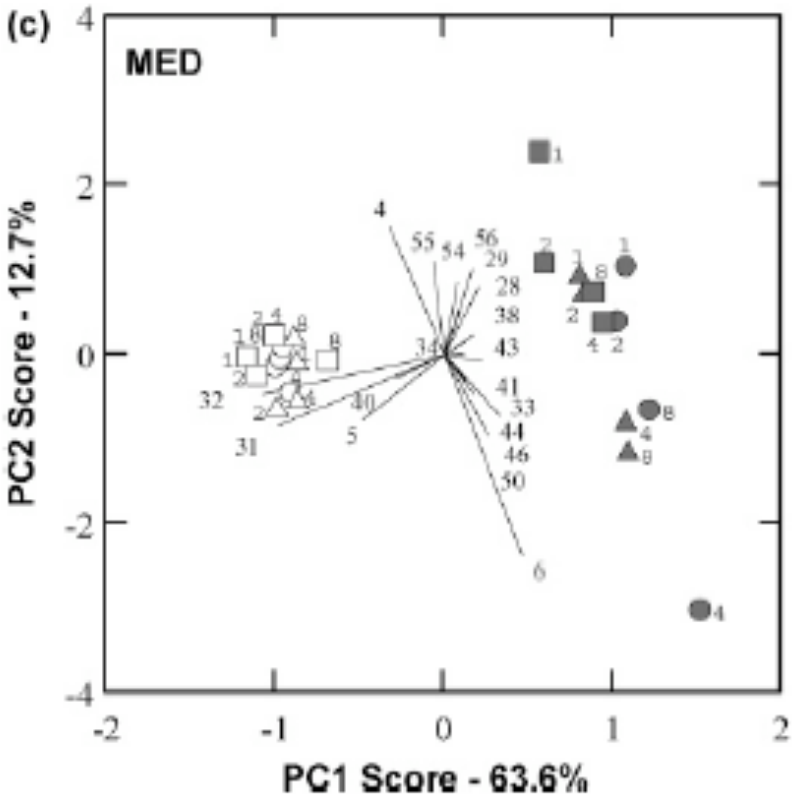
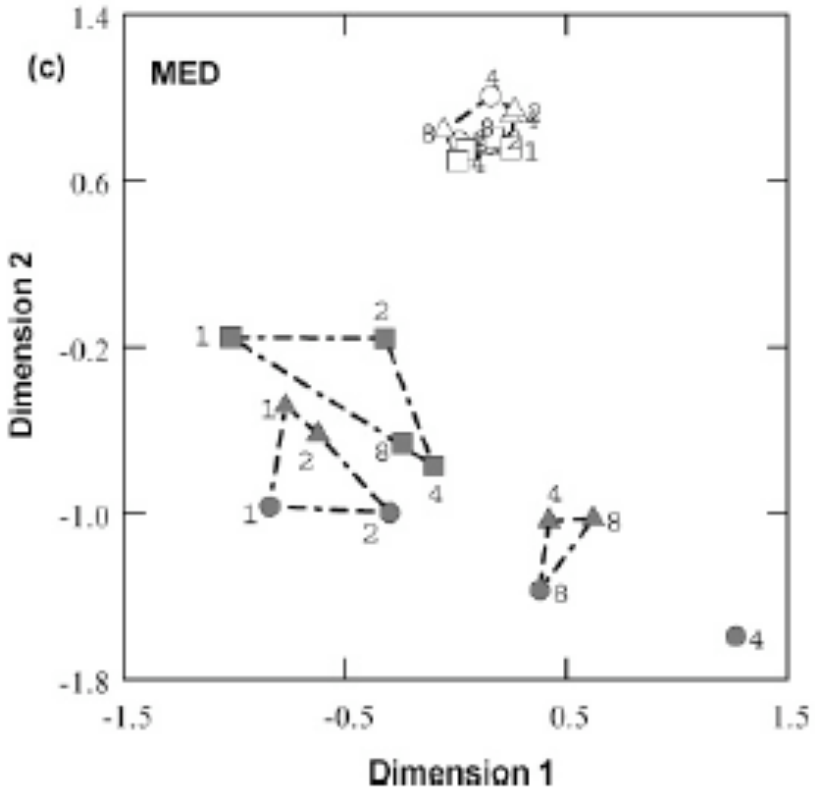


Figure 5. MDS configuration of the matrix of Pearson product-moment correlation between peptide profiles processed by visual matching (VM, **a**) or by fuzzy approach (HI, **b**; MED, **c**; and LOW, **d**). Symbols refer to reference smear R (■) and to defined-strain smear mix D (▲) or mix C (●); open and solid symbols are for core and surface samples, respectively. Convex hulls around samples show the five groups obtained by *K*-means clustering.

RP-HPLC data from smear cheese: PCA vs. MDS of Eulidean distance matrix





Some rights reserved