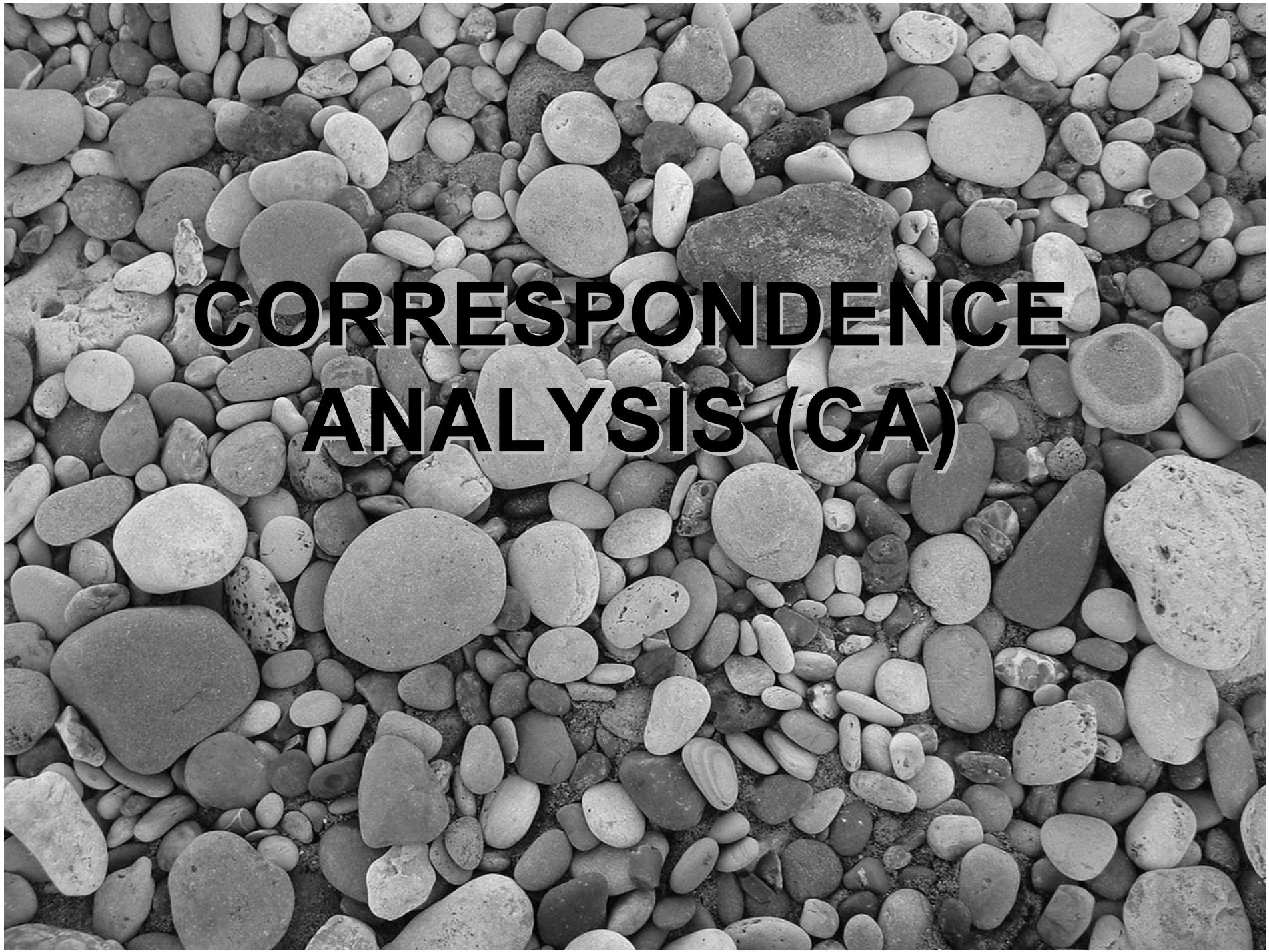
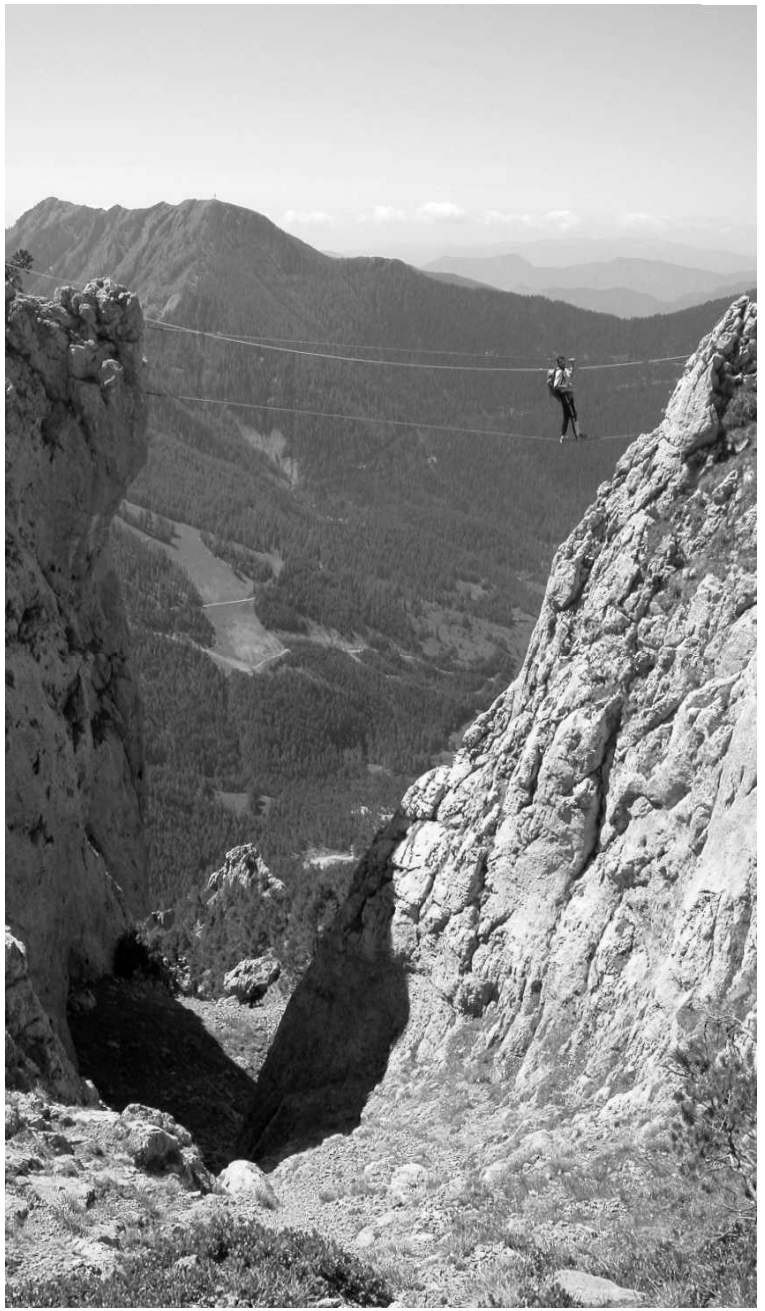

CORRESPONDENCE
ANALYSIS (CA)

What is it?

- CA (= reciprocal averaging of Hill, 1973) is a ‘form’ of PCA which uses a different association coefficient (X^2);
- Like PCA, eigenvector ordination method to reduce the dimensionality of a multivariate data set;
- Unlike PCA however data can be qualitative, semi-quantitative and quantitative data = VERY useful in microfacies analysis;
- Data need to be dimensionally homogeneous (same units) and > or = 0 !



Remember…

In PCA eigenvectors and their eigenvalues are calculated on the covariance/correlation matrix = association matrix summarizing relationships between the variables (components).

$$S = \begin{pmatrix} \sigma 11 & \sigma 12 & \dots & \sigma 1p \\ \sigma 21 & \sigma 22 & \dots & \sigma 2p \\ \dots & \dots & \dots & \dots \\ \sigma p1 & \sigma p2 & \dots & \sigma pp \end{pmatrix}$$

variances

$$\sigma_{pp} = \frac{1}{n-1} \sum_{i=1}^{i=n} (X_i - \bar{X})\ (Y_i - \bar{Y})$$

Eigenvectors give the dimensions of strongest correlation in the data set (PC's) and because the covariance/correlation matrix is **square** and **symmetric**, they are perpendicular (= linearly independent) to each other and explain each a fraction of the total variance.

Association coefficients !

The choice of the association coefficient is critical because all subsequent analysis is done on the resulting association matrix!

In PCA the use of the covariance or the correlation coefficient *r* has strong implications:

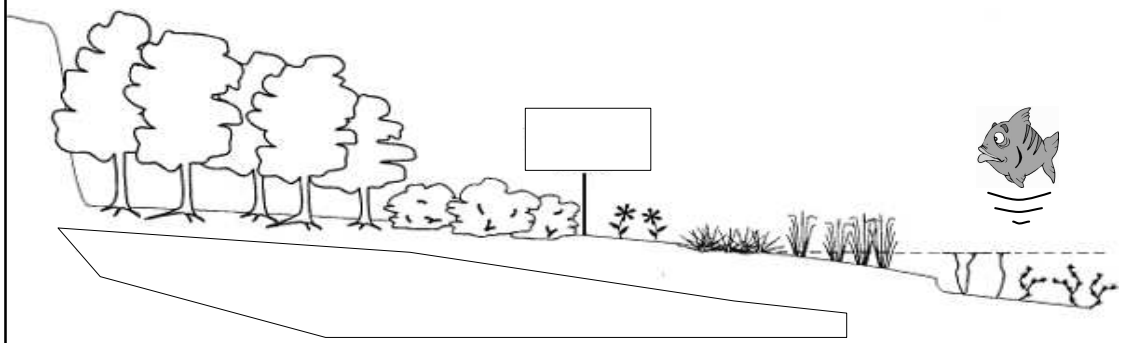
1) 2 samples with absence of a component will be considered as similar = the zero's problem, is that ok for us?
2) Make no sense for presence/absence or semi-quantitative data (ranks);
3) Work only to characterize LINEAR relationships between variables! Components always found together but whose abundances are not in linear relationship would not be considered as an association by these coefficients…

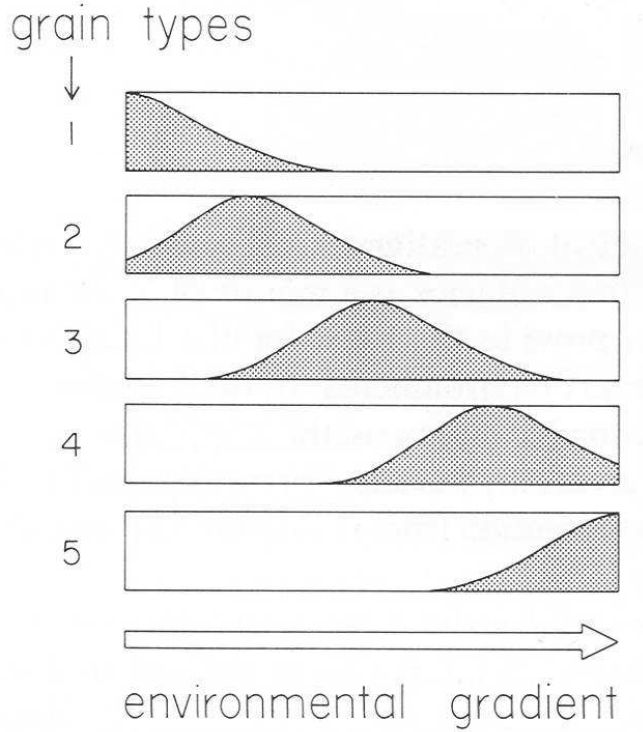
In our case these are all major problems.

Microfacies versus Relays:

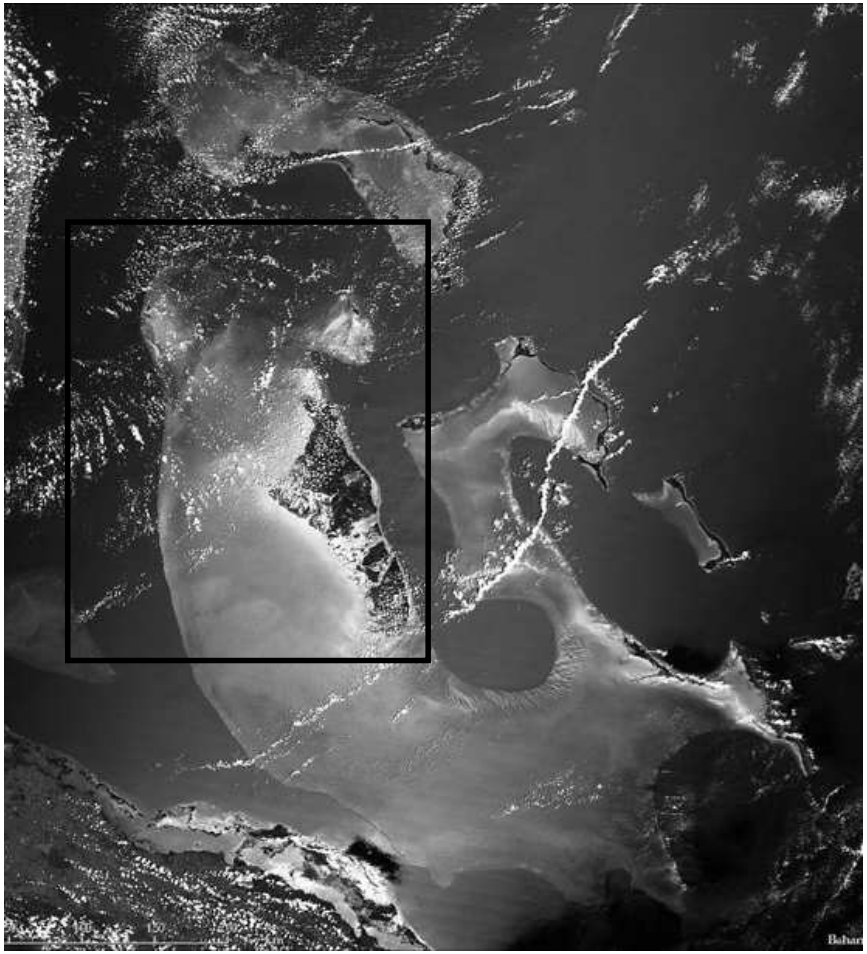
- Sedimentary environments can be relatively well partitioned (1) or on the contrary grade progressively into each other (2).
- In the first case a microfacies (cluster analysis) approach is appropriate, in the second it will be rather artificial and a crude representation of the reality.

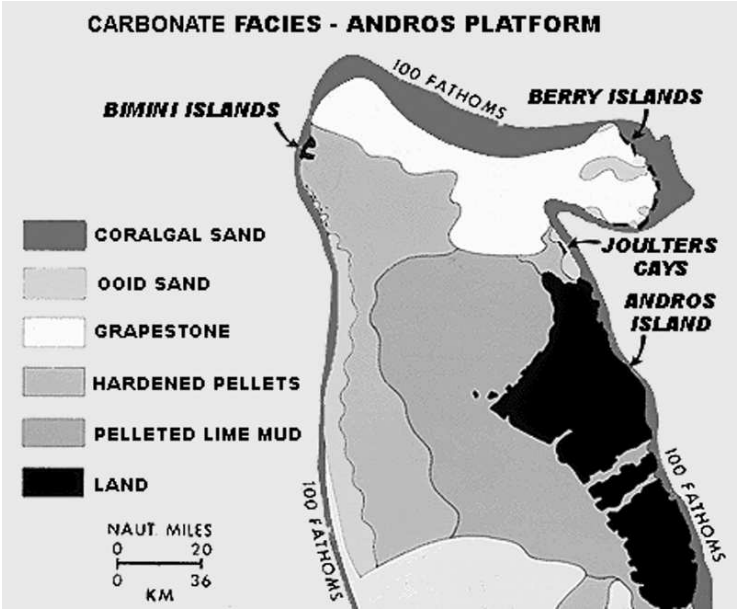
A progressive gradient (whatever it's nature) typically induces a systematic shift of the relative importance of the various components (or species). Such a systematic and progressive change is called a relay. CA offers a powerful tool for gradient analysis (used extensively in ecology).

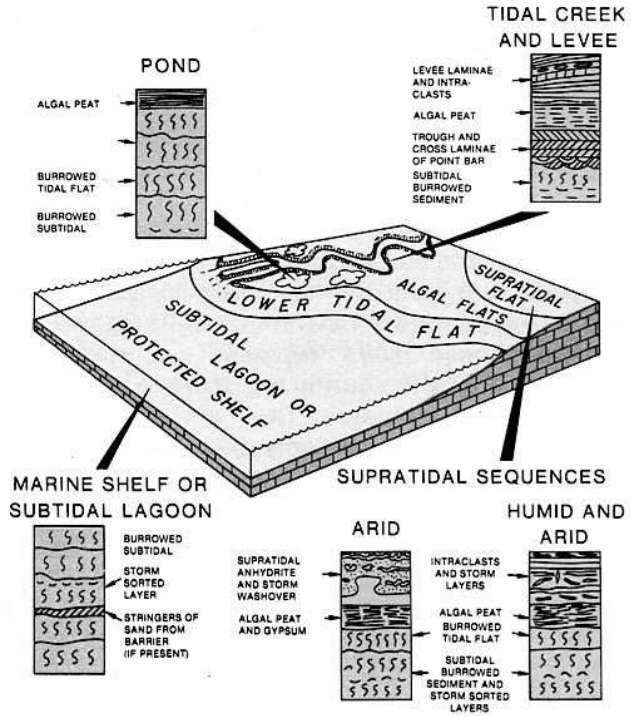




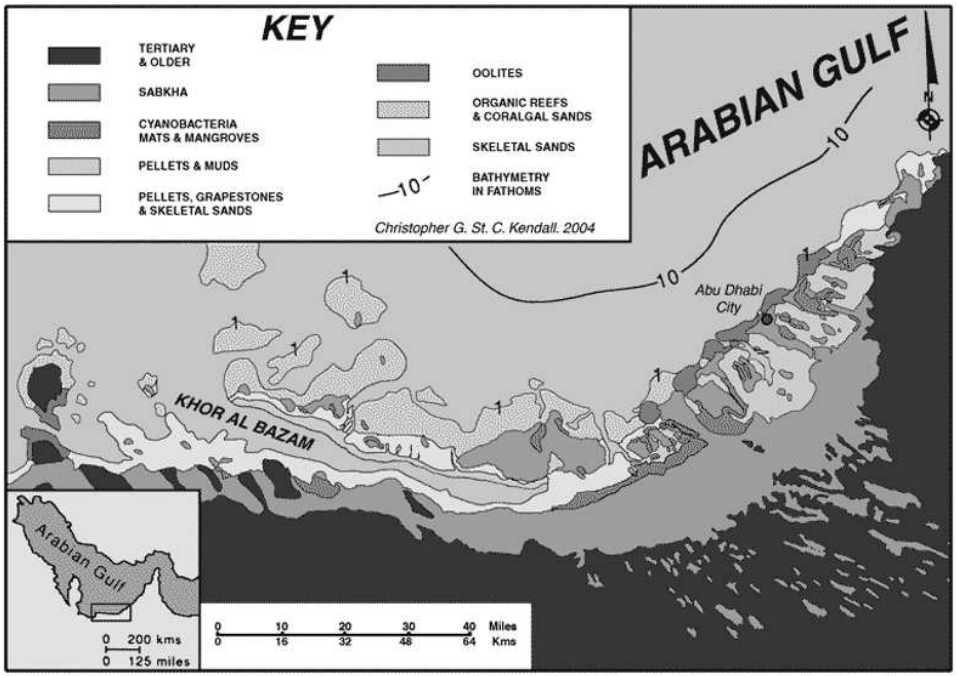
1 **Modern flat-topped platform**: Bahamas - Andros



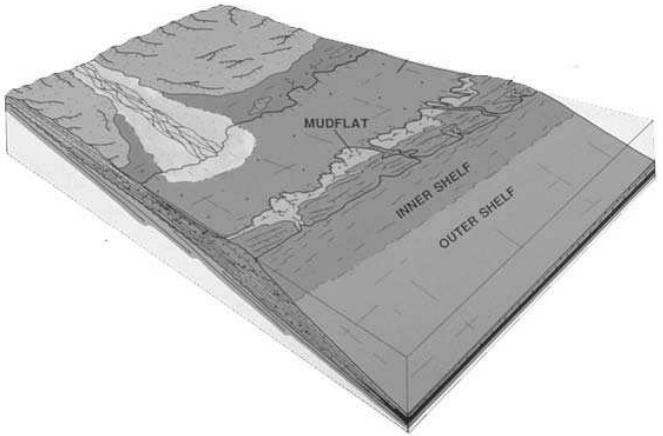




2



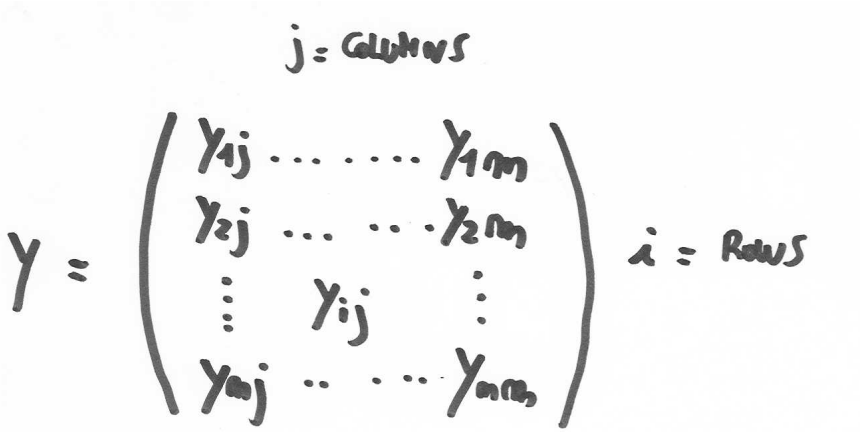
Modern carbonate ramp: Southern coast of the Persian Gulf



Why is CA adapted ?

Because it uses a different coefficient of association (X^2) and because it works on contingency tables.

The X^2 distance compares the conditional probabilities between rows (or columns) of the original frequency matrix (contingency table).

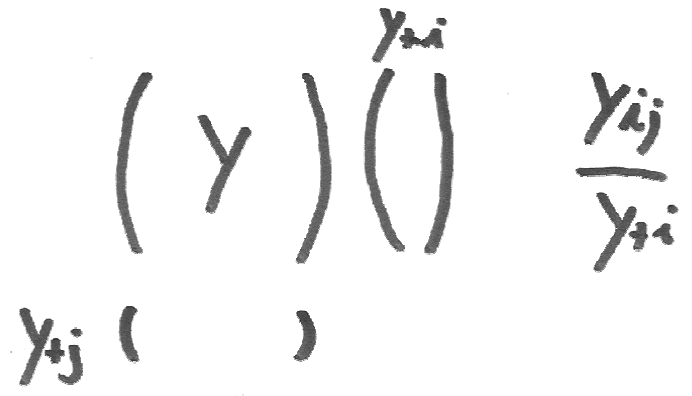


n rows (i) = samples

m columns (j) = components

Yij = count (absolute frequency) of component j in sample i (can be 0-1 or rankings)

Let's say we start by comparing rows (samples) in fct of their variables. First we calculate conditional probabilities (= relative frequencies) by rows:



Y_{ij}/Y_{+i} = probability that sample i contains component j knowing that sample i contains Y_{+i} components = marginal probability

The distance between the two first rows in fct of the variables (components) is calculated by:

X^2 metric

$$D(i_1, i_2) = \sqrt{\sum_{j=1}^{m} \frac{1}{y_{+j}} \left(\frac{y_{1j}}{y_{+1}} - \frac{y_{2j}}{y_{+2}} \right)^2}$$

Sum of frequencies for each column

p_{1j} p_{2j}

$1/Y_{+j}$ makes sure that distance does not increase for larger frequencies.

Operation is repeated for all rows and then for all columns. Two matrix are produced with the X^2 values, $m \times m$ between components and one $n \times n$ between samples. Eigenvectors/-values give the Principal Axes (PC's) and the contribution of each in R-mode (components) and Q-mode (samples).

Both matrix are square and symmetric!

Sum of all X^2 in each matrix = the total inertia. NOT variance because data are normally not continuous (0/1, ranks), but equivalent concept of spread of the data cloud. Each eigenvalue therefore gives the contribution of its associated eigenvector (Principal Axis) to the total **inertia**.

Because of Y_{+j}/Y_{++} relationships between rows and columns in the original frequency matrix (contingency table) are preserved > modes Q and R are equivalent. Allows to compare directly components and samples in the new reduced space.

Note also that if both Y_{1j} and Y_{2j} are 0 it does not increase or decrease the X^2 distance, the pair is neutral, it is not taken into account in the distance between rows 1 and 2!

Example?

$$\mathbf{Y} = \begin{bmatrix} 45 & 10 & 15 & 0 & 10 \\ 25 & 8 & 10 & 0 & 3 \\ 7 & 15 & 20 & 14 & 12 \end{bmatrix} \quad \begin{matrix} [y_{i+1}] \\ \begin{bmatrix} 80 \\ 46 \\ 68 \end{bmatrix} \end{matrix}$$

$$[y_{+j}] \quad \begin{bmatrix} 77 & 33 & 45 & 14 & 25 \end{bmatrix} \quad 194$$

$$\rightarrow \left[y_{ij} / y_{i+} \right] = \begin{bmatrix} 0.563 & 0.125 & 0.188 & 0.000 & 0.125 \\ 0.543 & 0.174 & 0.217 & 0.000 & 0.065 \\ 0.103 & 0.221 & 0.294 & 0.206 & 0.176 \end{bmatrix}$$

$$D_{15}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \frac{1}{y_{+j}} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2}$$

$$\mathrm{D}_{15}(\mathbf{x}_1, \mathbf{x}_2) = \left[\frac{(0.563-0.543)^2}{77} + \frac{(0.125-0.174)^2}{33} + \frac{(0.188-0.217)^2}{45} + \frac{(0-0)^2}{14} + \frac{(0.125-0.065)^2}{25} \right]^{1/2}$$

$= 0.015$

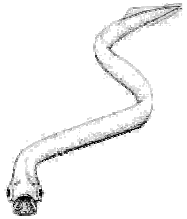
The component which is absent from the first two samples, cancels itself out; thus χ^2 metric deals with double-zeros.

Note however that if a component j is rare its column sum Y_{+j} is small and this species contributes a great deal to the X^2 … CA is therefore strongly influenced by rare components (variables).

Method

1) The data are put in the form of a contingency table with frequencies (counts ij) = matrix Y;
2) Absolute frequencies are transformed in relative frequencies (= probabilities);
3) 'X^2 distances' are calculated for each pair of columns (R-mode) and rows (Q-mode), this produce two new square matrix m x m and n x n;
4) Eigenvectors and eigenvalues are calculated for these matrix in order to find the orthogonal directions of maximum inertia (~'variance') in the new data cloud;
5) The sum of the eigenvalues = total inertia and each eigenvector (Factor Axis or Principal Axis) has its associated eigenvalue which gives its contribution to the total inertia.

Example 1



Total inertia = sum of eignevalues = 0.7938

m

n

TABLE 6.37 Counts of Conodont Tests Recovered from 10-kg Samples of Rock; Columns are Conodont Varieties; Rows are Stratigraphic Units that are Members in a Section of Missourian Age in Eastern Kansas; Megacyclothem Classifications are Outside Shale (O), Shoal Limestone (S), Upper Limestone (U), Middle Limestone (M), "Phantom Black Shale" (P), Black Shale (B)

			Counts of Conodonts										
N	Class	Rock Unit	*Adetognathus*	*Ozarkodina*	*Aethotaxis*	*Idiognathodus delicatus*	*I. elegantulus*	*Magnilaterella*	*Hindeodella*	*Idioprioniodus*	*Gondolella*	Others	TOTAL
			A	B	C	D	E	F	G	H	I	J	
1.	M	South Bend Ls.	13	10	0	0	37	0	0	0	0	0	60
2.	O	Rock Lake Sh.	0	0	0	0	11	0	0	0	0	0	11
3.	U	Stoner Ls.	4	2	1	51	26	1	0	0	0	0	85
4.	B	Eudora Sh.	0	7	1	207	350	0	0	34	14	3	606
5.	M	Captain Creek Ls.	8	28	6	0	60	0	0	0	0	0	102
6.	O	Vilas Sh.	145	20	5	0	10	0	0	0	0	0	180
7.	U	Spring Hill Ls.	5	134	8	0	353	1	0	4	0	0	505
8.	P	Hickory Creek Ls.	20	60	0	0	920	0	0	0	0	0	100
9.	M	Merriam Ls.	115	255	10	0	1140	0	0	0	0	0	1520
10.	S	Bonner Springs Sh.	1	0	0	0	3	0	0	0	0	0	4
11.	S	Farley Ls.	31	21	7	0	4	1	0	0	0	0	61
12.	—	Island Creek Sh.	100	5	0	0	5	0	0	0	0	0	110
13.	U	Argentine Ls.	0	39	1	0	80	0	1	0	0	0	121
14.	P	Quindaro Sh.	10	70	0	0	538	0	0	5	0	0	623
15.	M	Frisbee Ls.	3	78	5	0	450	0	0	3	0	0	539
16.	O	Lane Sh.	0	0	0	0	28	0	0	0	0	0	28
17.	U	Raytown Ls.	38	20	3	100	267	3	0	25	0	0	456
18.	B	Muncie Creek Sh.	15	8	0	243	515	0	10	85	55	13	946
19.	M	Paola Ls.	10	130	10	200	900	0	0	50	0	0	1300
20.	O	Chanute Sh.	117	20	0	63	57	0	0	7	0	0	264
		TOTAL	258	389	31	367	1929	4	5	82	32	7	3104

m x m

TABLE 6.38 χ^2 Similarity Matrix, Eigenvalues, and First Two Eigenvectors Calculated for Conodont Abundance Data; Also Given are *R*- and *Q*-Mode Correspondence Factor Loadings for First Two Factors

χ^2 SIMILARITY MATRIX

	A	B	C	D	E	F	G	H	I	J
A	.3843	.0037	.0257	−.0273	−.1136	.0056	−.0079	−.0239	−.0204	−.0098
B	.0037	.0568	.0196	−.0645	.0088	.0015	−.0076	−.0292	−.0268	−.0129
C	.0257	.0196	.0216	−.0119	−.0117	.0063	−.0026	−.0066	−.0070	−.0034
D	−.0273	−.0645	−.0119	.1655	−.0477	.0090	.0150	.0620	.0486	.0233
E	−.1136	.0088	−.0117	−.0477	.0592	−.0066	−.0052	−.0159	−.0136	−.0066
F	.0056	.0015	.0063	.0090	−.0066	.0075	−.0010	.0006	−.0024	−.0011
G	−.0079	−.0076	−.0026	.0150	−.0052	−.0010	.0091	.0129	.0179	.0088
H	−.0239	−.0292	−.0066	.0620	−.0159	.0006	.0129	.0365	.0330	.0160
I	−.0204	−.0268	−.0070	.0486	−.0136	−.0024	.0179	.0330	.0430	.0210
J	−.0098	−.0129	−.0034	.0233	−.0066	−.0011	.0088	.0160	.0210	.0102

Vector	Eigenvalue	Total Similarity (%)	Total Similarity (Cumulative %)
1	.4262	53.7003	53.7003
2	.2634	33.1837	86.8841
3	.0468	5.8936	92.7776
4	.0385	4.8532	97.6308
5	.0101	1.2691	98.8999
6	.0044	.5488	99.4487
7	.0036	.4523	99.9010
8	.0008	.0988	99.9997
9	.0000	.0003	100.0000
10	.0000	.0000	100.0000

So all factors explain 79.38% of the variance of the original contingency table and…

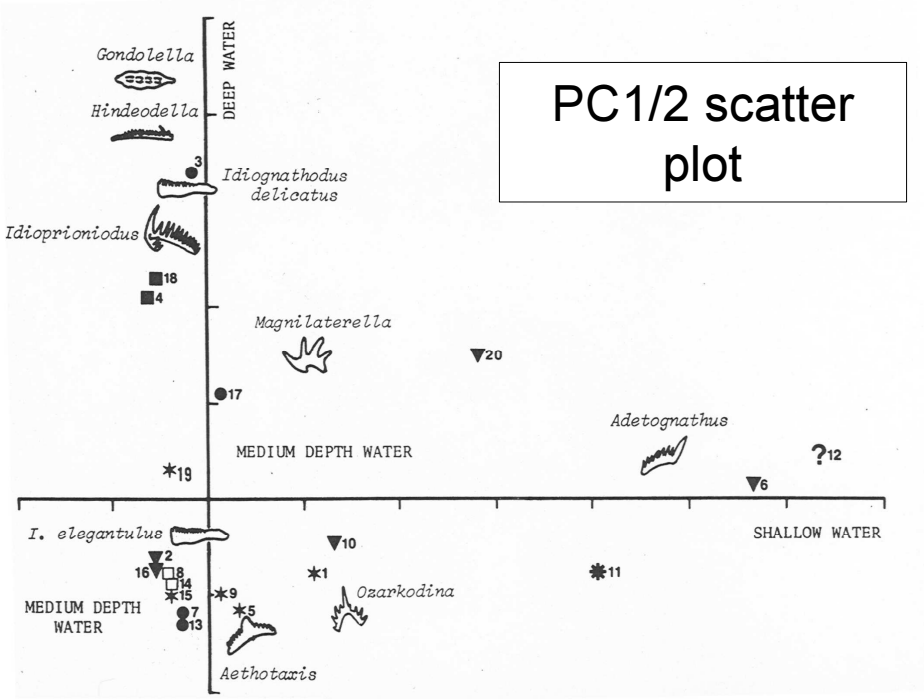
Factor 1 explains 0.4262/0.7938 = 53.7% of the inertia of the X^2 matrix, **NOT** of the original contingency table!

Loadings

Correspondence Axis Loadings

	R Mode			Q Mode	
Conodont	I	II	Unit	I	II
A	2.2655	.1229	1	.5628	−.3344
B	.0715	−.5492	2	−.3362	−.3372
C	.6038	−.4064	3	−.0968	1.3119
D	−.1938	1.2193	4	−.3338	.7964
E	−.2195	−.1730	5	.1589	−.5199
		.5546	6	2.8147	.0333
G	−.4512	1.4456	7	−.1593	−.5114
H	−.3037	1.0531	8	−.2333	−.3696
I	−.4768	1.6680	9	.0349	−.4195
J	−.4761	1.6703	10	.6154	−.1930
			11	1.8068	−.3260
			12	3.1445	.1537
			13	−.1850	−.5511
			14	−.2260	−.3911
			15	−.2395	−.4309
			16	−.3362	−.3372
			17	.0168	.4110
			18	−.3055	.8712
			19	−.2515	.0998
			20	1.3905	.5737

Interpretation



PC1/2 scatter plot

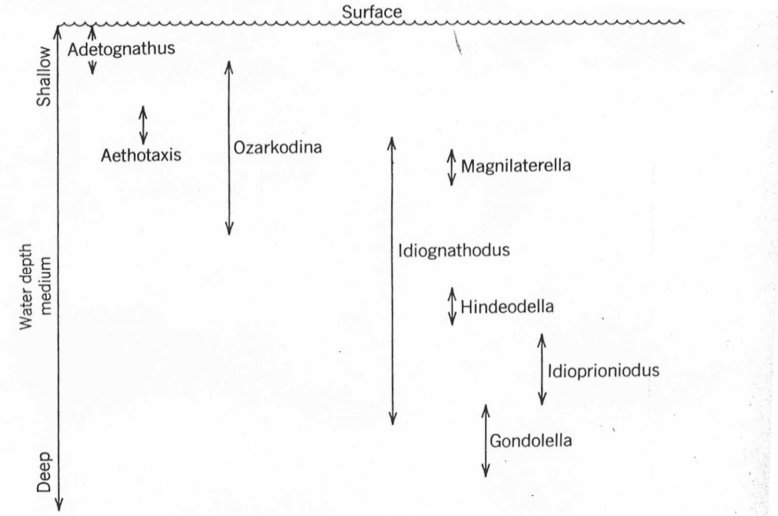
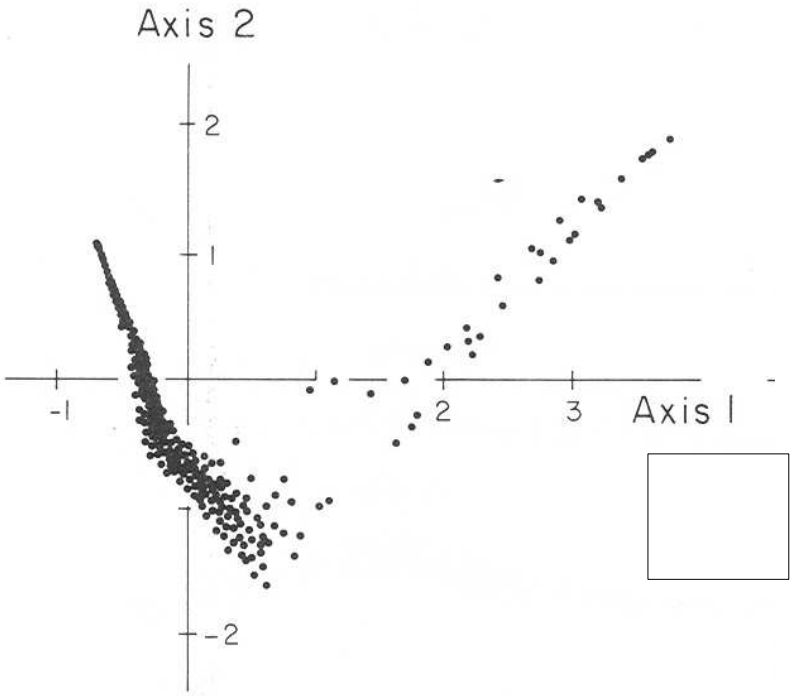


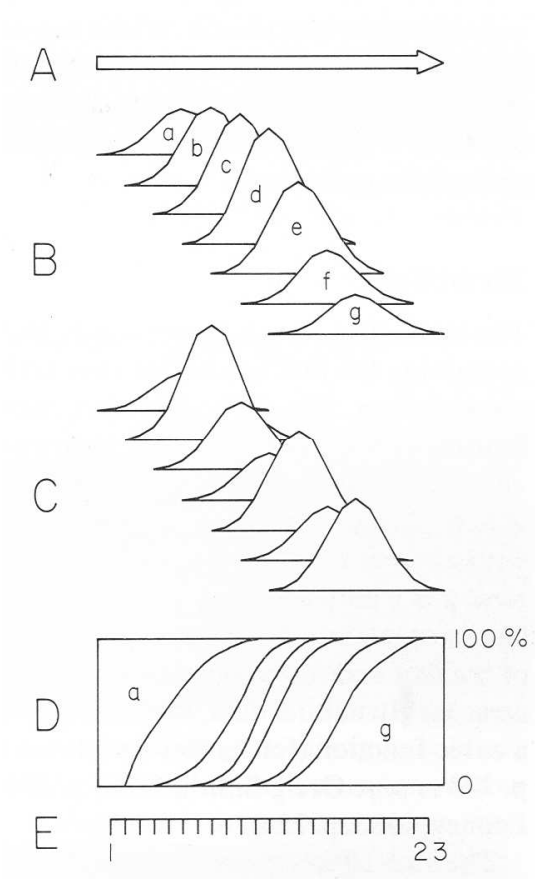
FIGURE 6.45 Relative depth ranges for conodonts collected from

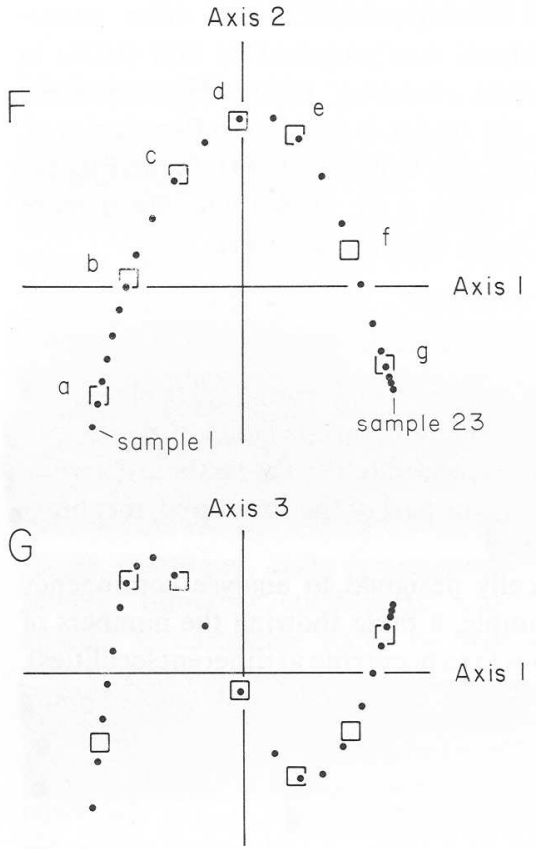
Arch effect

Occurs when, although not LINEARLY correlated, factor axes are all linked with a power function of the first factor. Uni-dimensional phenomenon controls most of the structure of the data set.

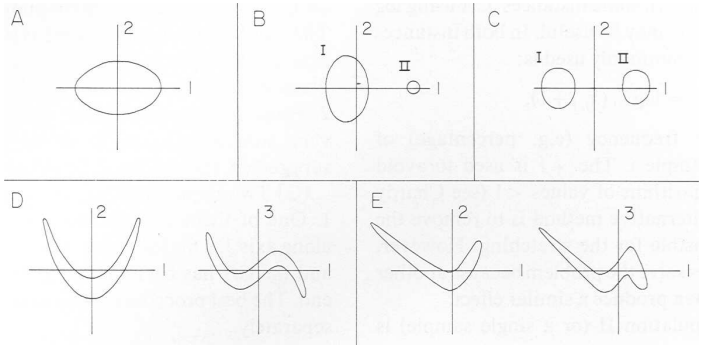


In a long ecological or sedimentological gradient there is usually a succession of components (species) with more or less unimodal distributions (reflecting optimal range of conditions). So if samples are compared on the basis of the presence/absence or abundance of these species/components the distance relationship is necessarily non linear.



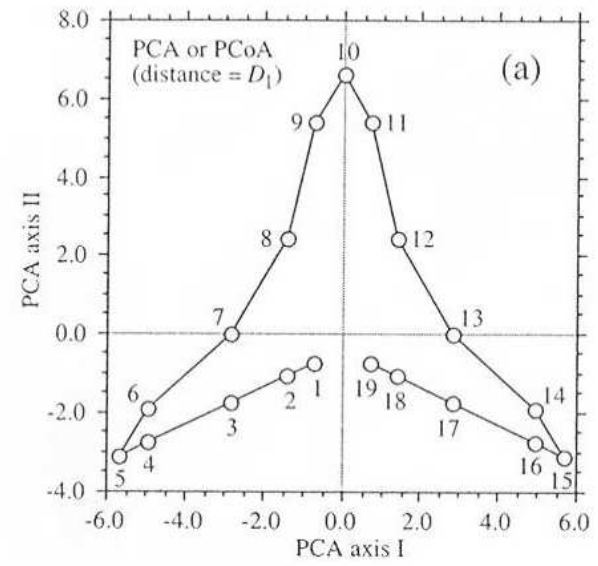


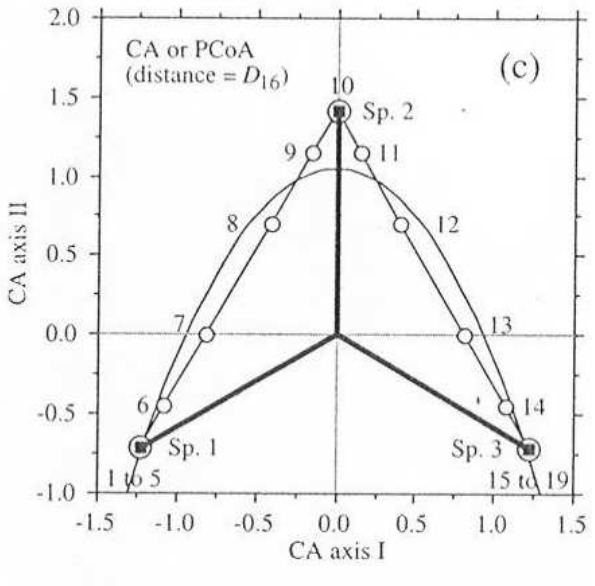
Ordination methods aim at rendering this non-linear relationships in an Euclidian space and two-dimensions plots.



Ordination axes (PC's) try to maximally separate the species/components while remaining uncorrelated with one another. If PC1 is enough to order the samples and species/components, an independent PC2 (no meaning) can only be obtained by folding the first axis in the middle and bringing the ends together.

In **PCA** this effect is especially strong because distance coefficient used considers extreme samples on both end of the gradient with increasing number of non-overlapping components (increasing number of double-zeros in the matrix) as increasingly similar (shorter distance). This results in an inwards folding along PC1.

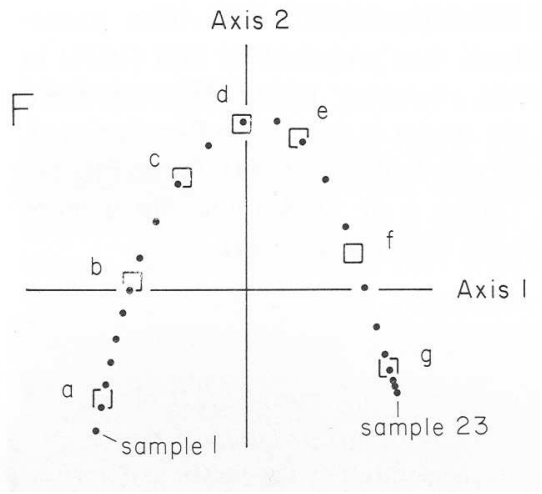




In **CA** samples at both ends of the gradient which have no components in common have a similarity 0 (maximum distance) and are at both ends of the arch in a scatter plot PC1/2. There is no closing effect.

The arch effect can therefore be used:

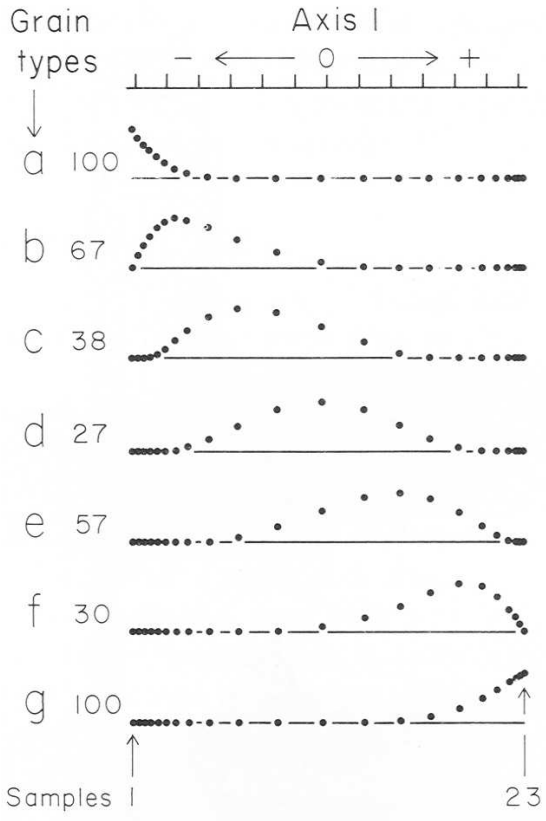
1) to detect environmental gradients and
2) to characterize them.



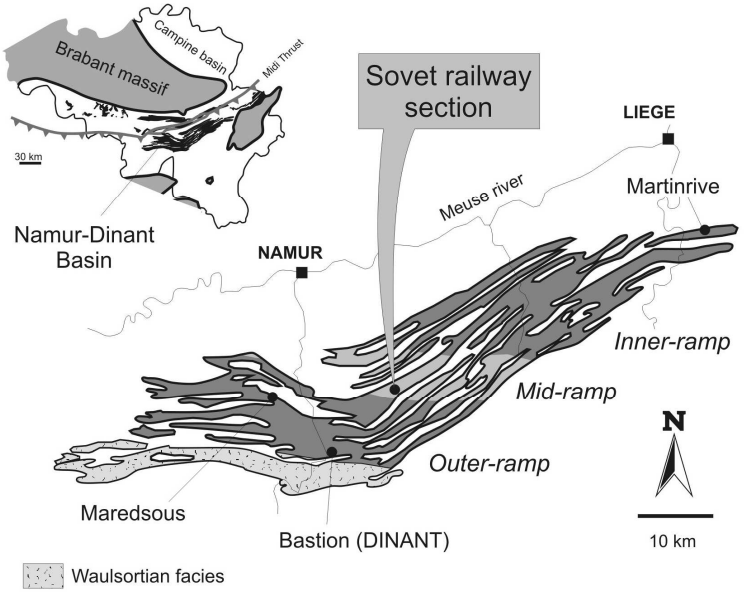
• Because there is no inwards folding in CA, the coordinates of the variables and samples on PC1 can be used to display how variables behave along the gradients and where samples are located on it = **Relay Index (RI)**.

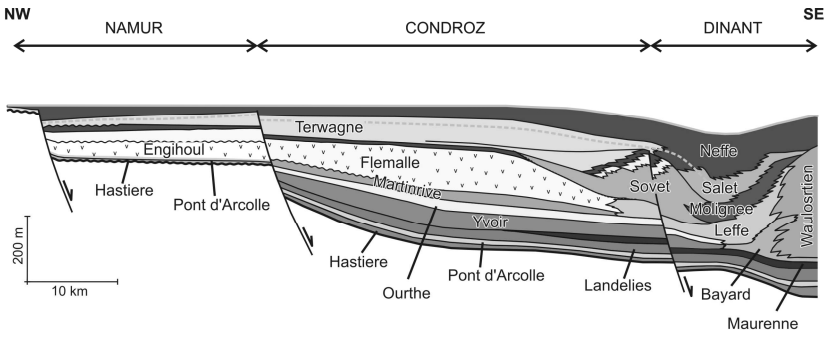
• The RI can be further used to plot the position of the samples in the gradient sratigraphically.

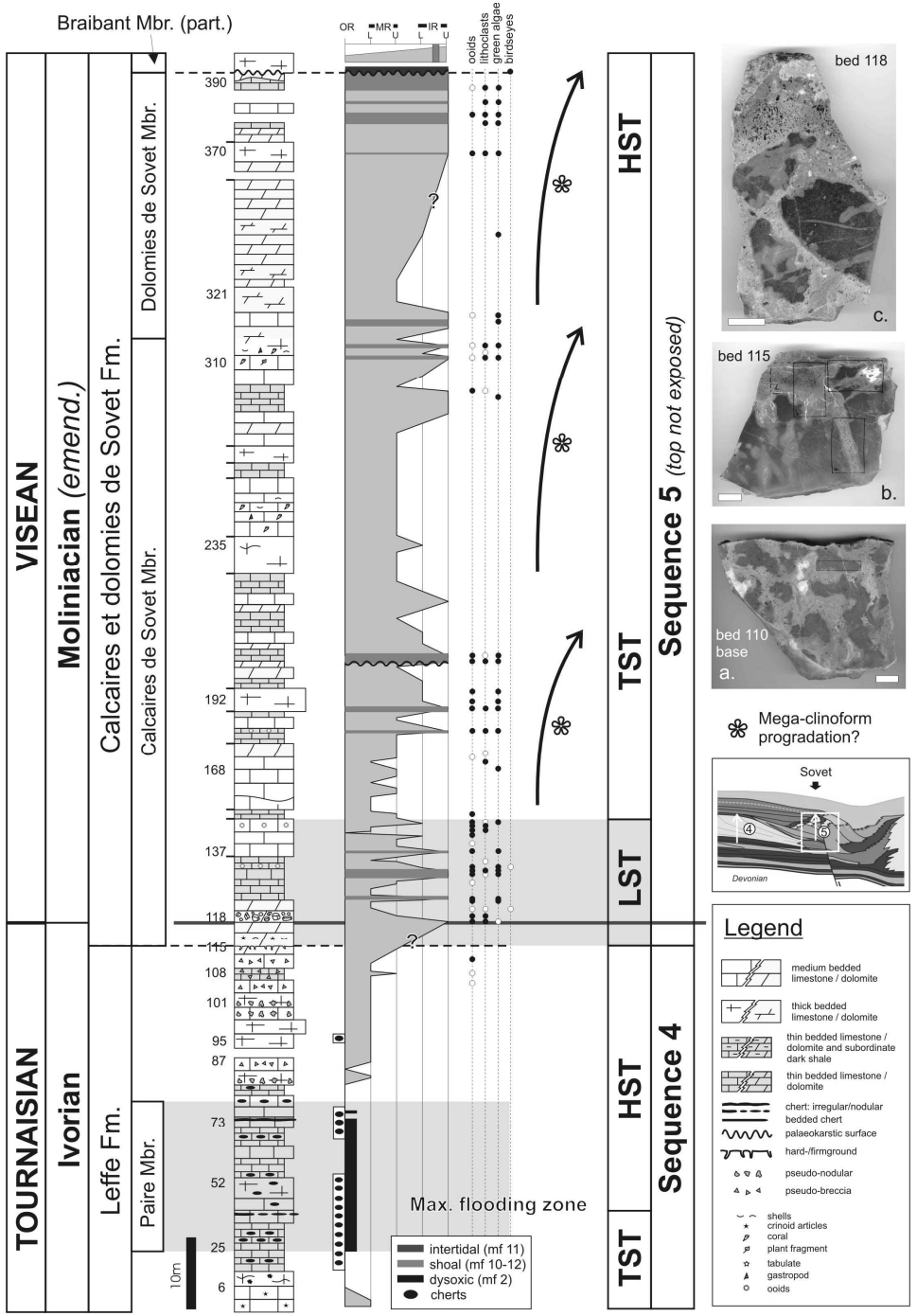
•This can reveal a very powerful method in carbonate sedimentology.

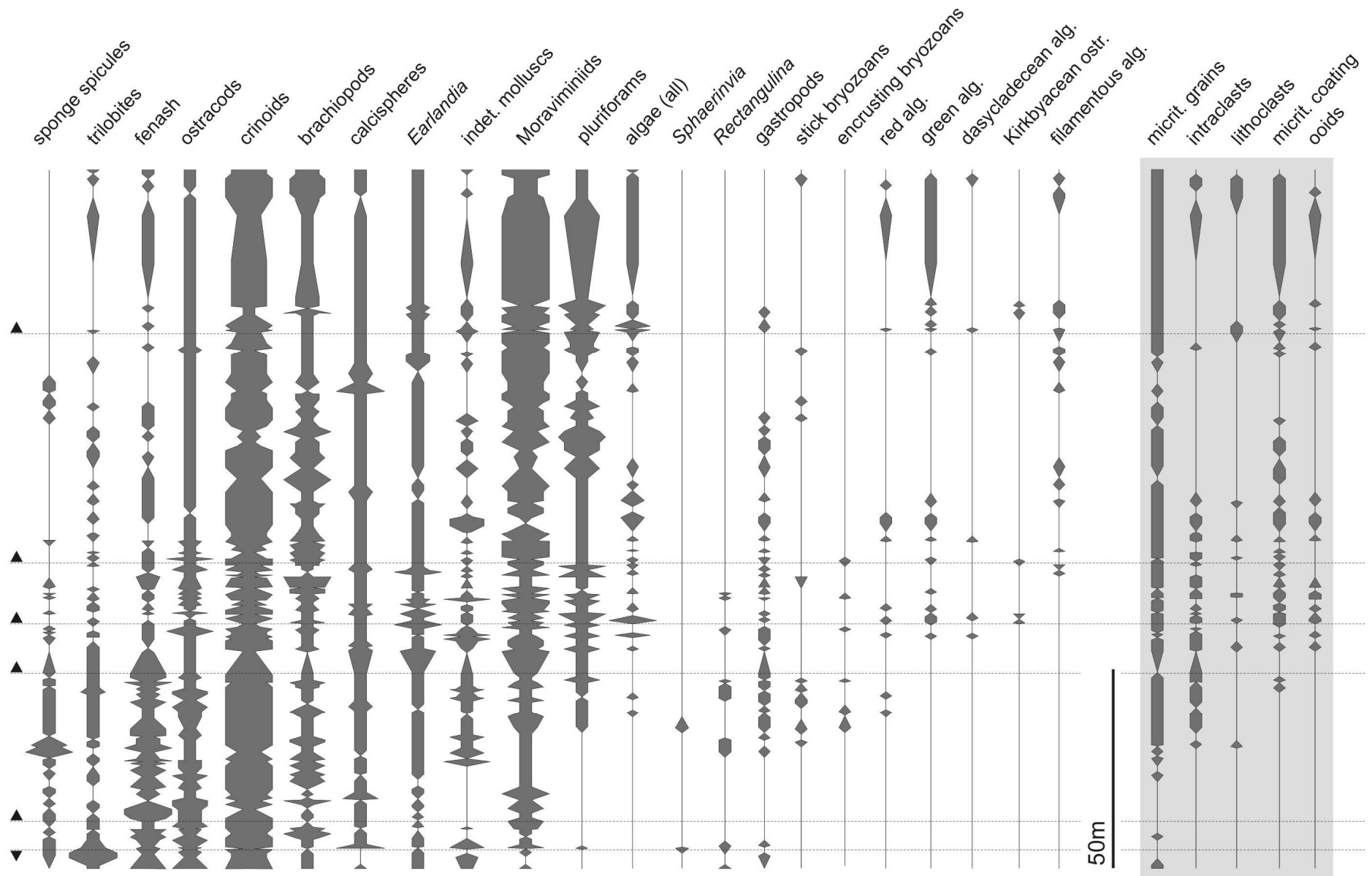


Example 2

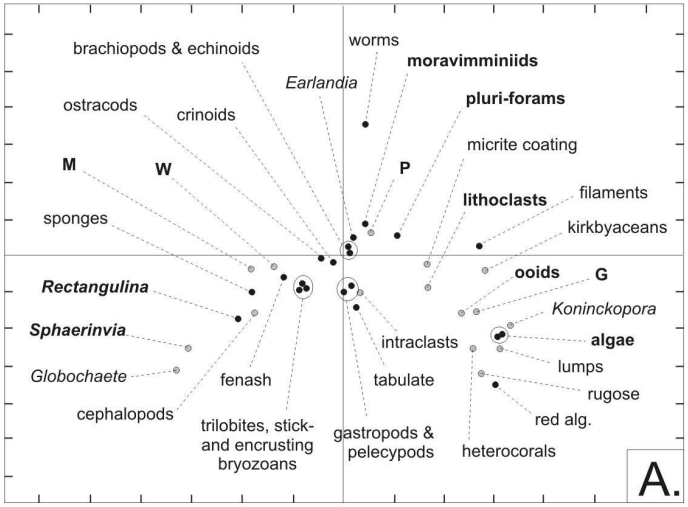

Brabant massif
Campine basin
Midi Thrust
30 km
Namur-Dinant Basin
Sovet railway section
LIEGE
Martinrive
Meuse river
NAMUR
Inner-ramp
Mid-ramp
Outer-ramp
Maredsous
Bastion (DINANT)
N
10 km
Waulsortian facies

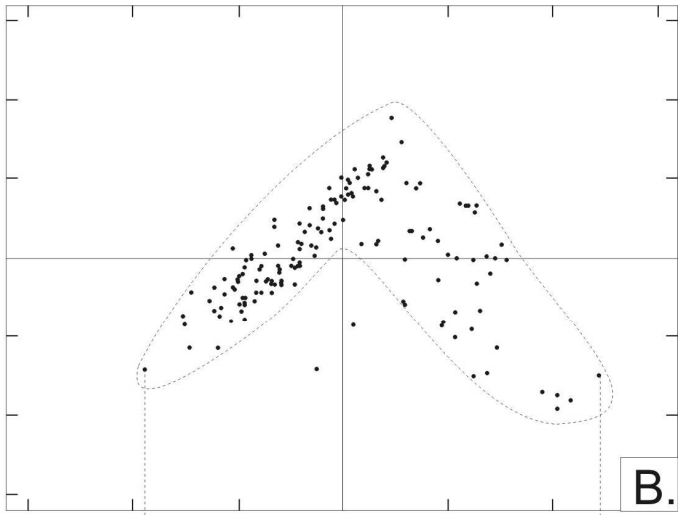

NW
NAMUR
CONDROZ
DINANT
SE
Engihoul
Hastiere
Pont d'Arcolle
Terwagne
Flemalle
Martinrive
Yvoir
Hastiere
Ourthe
Pont d'Arcolle
Landelies
Sovet
Neffe
Salet
Molignee
Leffe
Waulsortien
Bayard
Maurenne
200 m
10 km


Braibant Mbr. (part.)
OR
MR
IR
ooids
lithoclasts
green algae
birdseyes
VISEAN
Moliniacian (emend.)
Calcaires et dolomies de Sovet Fm.
Dolomies de Sovet Mbr.
Calcaires de Sovet Mbr.
TOURNAISIAN
Ivorian
Leffe Fm.
Paire Mbr.
390
370
321
310
235
192
168
137
118
115
108
101
95
87
73
52
25
6
10m
Max. flooding zone
intertidal (mf 11)
shoal (mf 10-12)
dysoxic (mf 2)
cherts
HST
TST
LST
HST
TST
Sequence 5 (top not exposed)
Sequence 4
bed 118
c.
bed 115
b.
bed 110 base
a.
Mega-clinoform progradation?
Sovet
Devonian
Legend
medium bedded limestone / dolomite
thick bedded limestone / dolomite
thin bedded limestone / dolomite and subordinate dark shale
thin bedded limestone / dolomite
chert: irregular/nodular bedded chert
palaeokarstic surface
hard-/firmground
pseudo-nodular
pseudo-breccia
shells
crinoid articles
coral
plant fragment
tabulate
gastropod
ooids

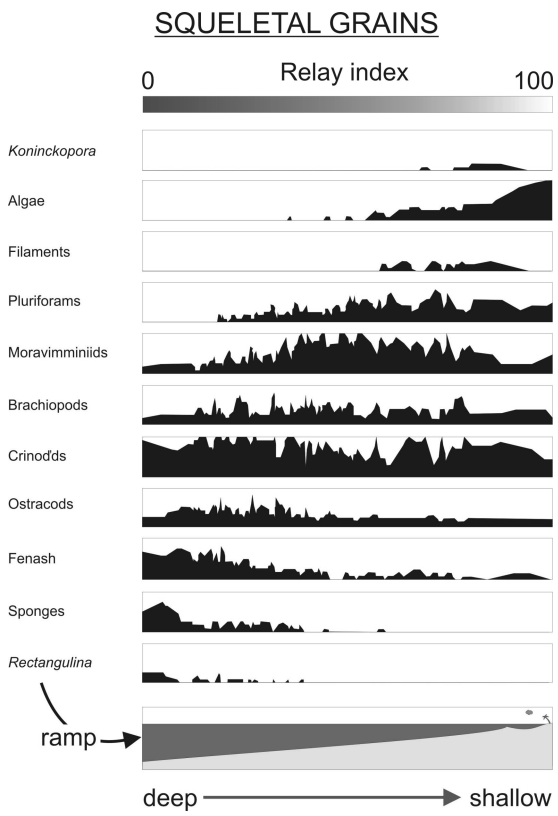


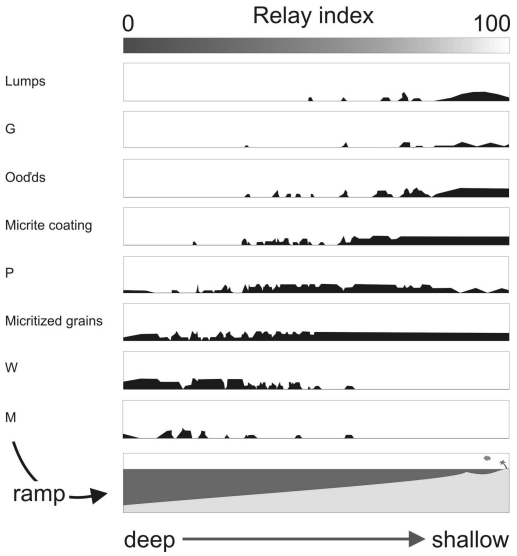
The data… do you see something?





Relay and interpretation





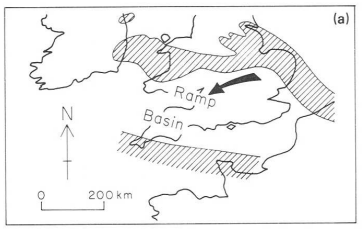
samples = 180	eigen values: axis 1 = 0.1943, axis 2 = 0.1344
total inertia = 0.9652653	contribution to total inertia: axis 1 = 20.1%, axis 2 = 13.9%

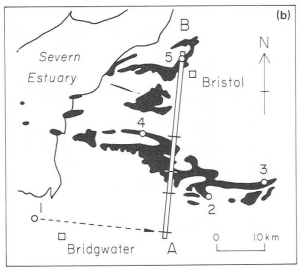
Relation between Relay Index and Microfacies

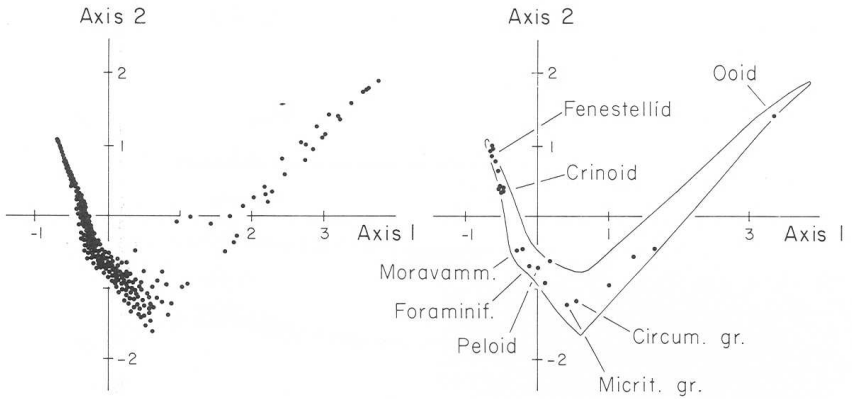
Samples in CA

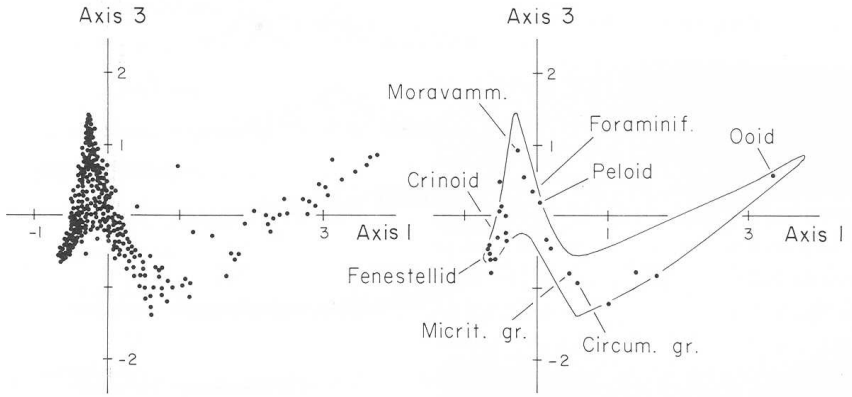
Microfacies

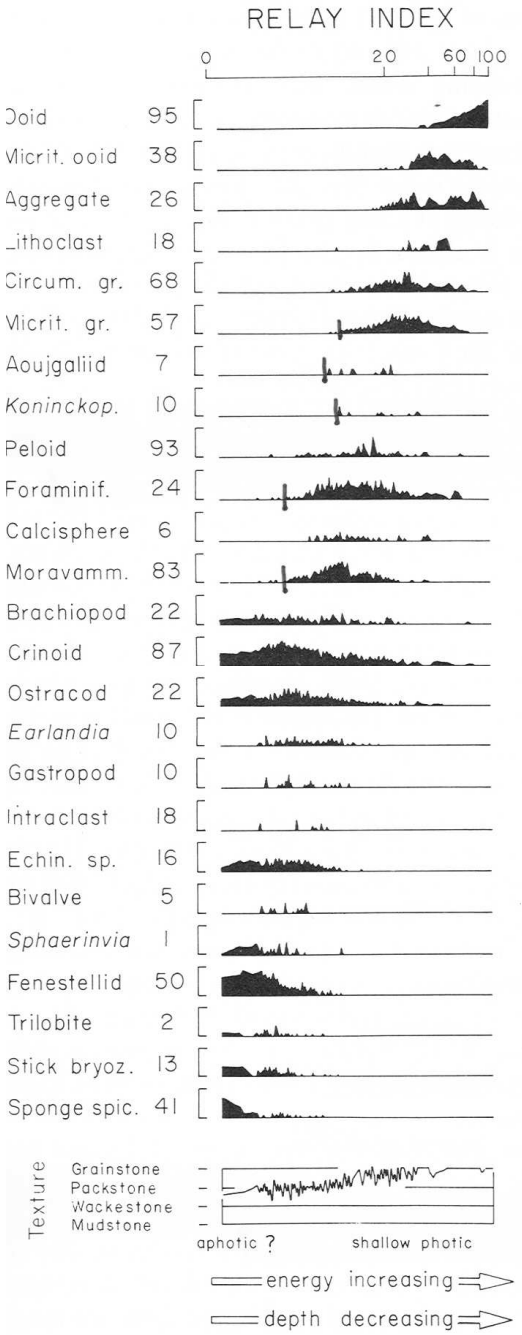
Example 3



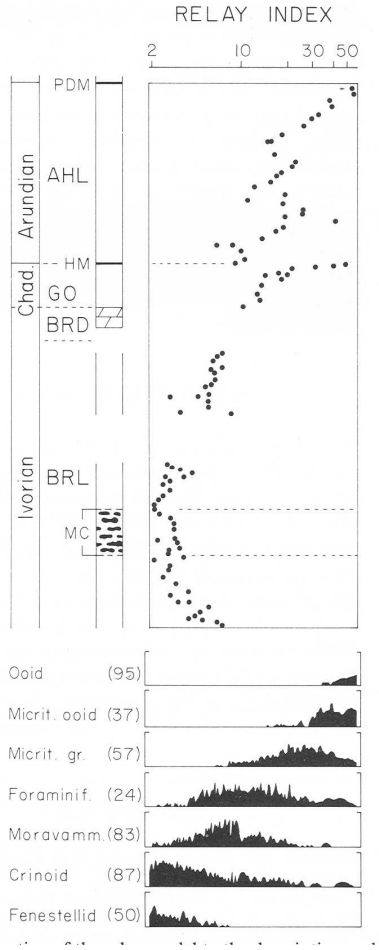








Avon Gorge



Detrended CA

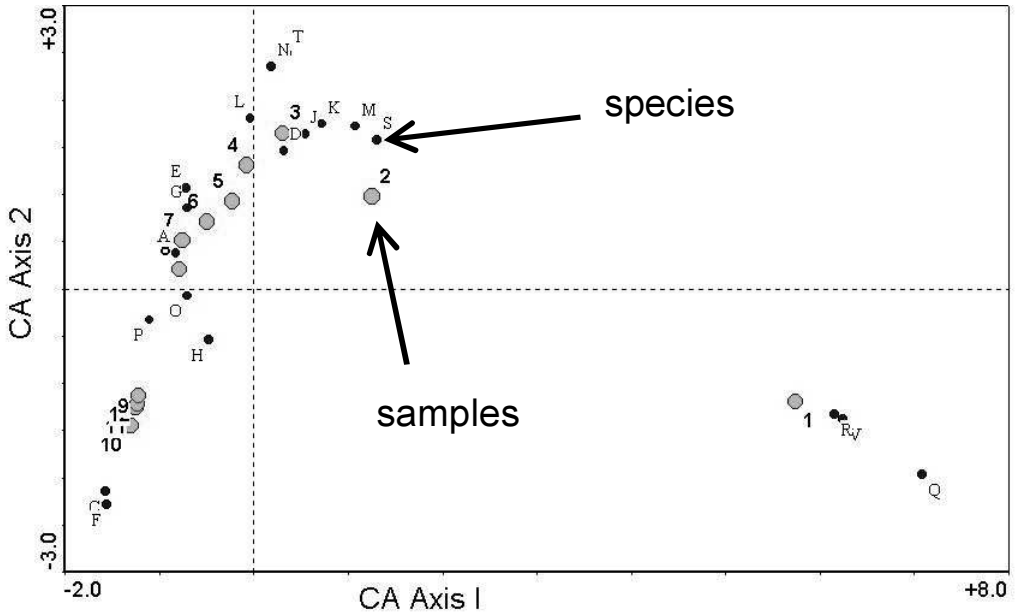
- Detrending aims at supressing the arch effect and spreading the data points along PC1.
- Two main methods:

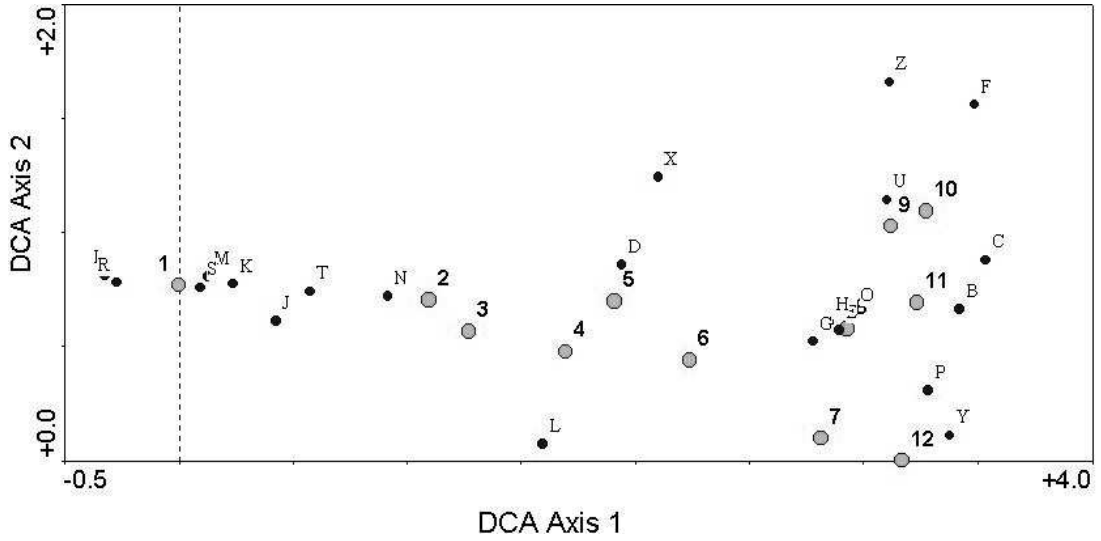
-Segments

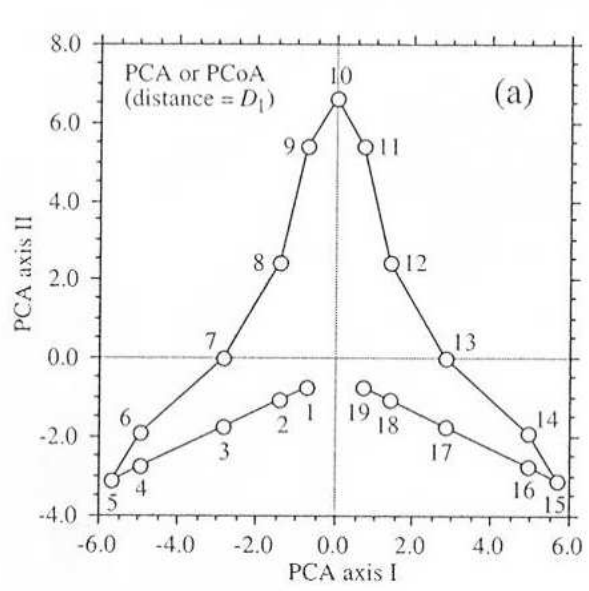
-Polynomials

- Segments do NOT preserve distances between points (arbitrary divisions). Polynomials don't solve the terminal gradient compression.

BOTH are very dodgy and better avoided!



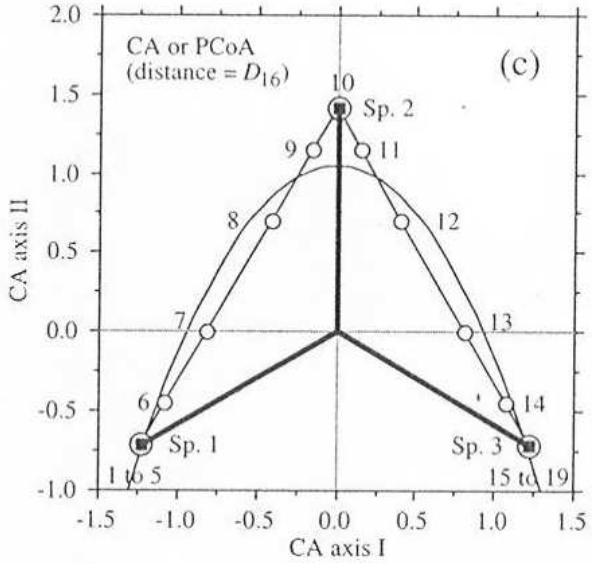




PCA

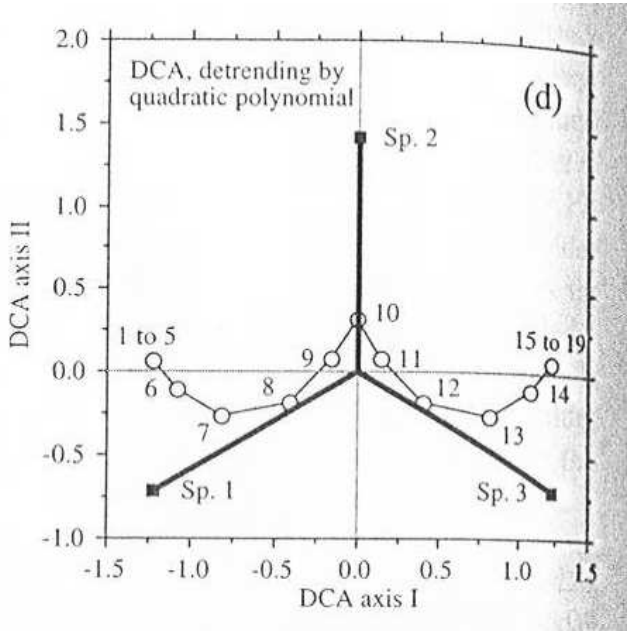
And the winner is…

CA



CA

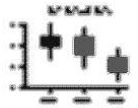
Comparison of ‘relay management’ between the main ordination methods on an artificial data set



DCA

In PAST 1.33

- The PCA routine finds the eigenvalues and eigenvectors of the matrix containing the X2 distances between all data points:
  - Scatter plot of samples (rows) in the CA coordinate system (main factor axes 1-3);
  - Variables can be plotted also in the same coordinate system;
  - A ‘Relay plot’ is also available with CA first PC as the vertical axis and the original data point value (abundances) on the horizontal axis (samples in rows, variables in columns). It shows the variables ordered along a potential gradient;
  - Detrended CA is also performed by PAST in two steps: straightening and spreading.



In AnalLog 1.82

Visual Basic program for the analysis of petrographical data developped by A. Lees (formerly at Catholic University of Louvain – UCL, Belgium, now retired but still active in carbonate sedimentology). Freeware available upon request (alanlees@gofree.indigo.ie).

Does CA (based on a routine published by T. Foucart, 1982) but serves also as customisable petrographical database and draws petrographical logs. Developped by carbonate sedimentologists for carbonate sedimentologists… ☺ Only problem is export of graphics.

The CA routine offers:

-Scatter plots of samples and components on the same scale;

- Plots of the components along the relay;

- Stratigraphical plot of the relay index;

- Various plotting options (log, moving average etc.).

Final words

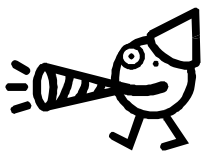
- CA is very powerful and largelly under-used in sedimentology and palaeoecology;
- Ordination (PCA-CA) can (should?) be combined with cluster analysis, not especially antagonistic:

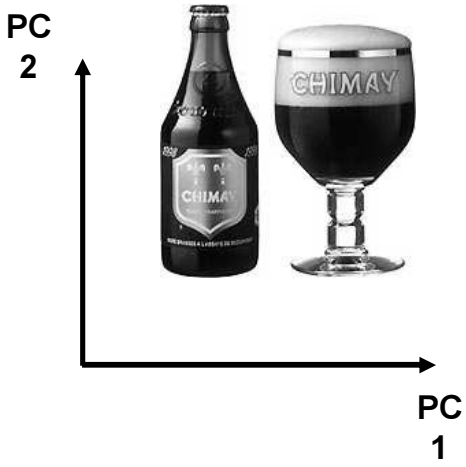
  - Cluster analysis looks at pairwise distances among samples = fine relationships,

  - Ordination (CA, PCA) considers the variability of the whole association matrix and thus brings out general gradients;

- Methods exist to superimpose the two approaches;
- Problem of extreme values in CA because of the X2. Very sensitive to rare species/components which tend to be located at the extremes in the ordination plot.

References





- PAST: http://folk.uio.no/ohammer/past/
- AnalLog: contact A. Lees at alanlees@gofree.indigo.ie
- (very) Good websites:

  - http://ordination.okstate.edu/
  - http://www.plantbio.ohiou.edu/epb/instruct/multivariate/multivariate.htm
  - http://www.okstate.edu/artsci/botany/ordinate/software.htm#method
  - http://149.170.199.144/multivar/intro.htm

The bible…:

Legendre, P. & Legendre, L. 1998. Numerical Ecology. Elsevier.