Constrained biplots

Uluslararası Katılımlı 15. Ulusal Biyoistatistik Kongresi
19-23 Ağustos 2013 **Didim, Aydın**

Michael Greenacre
Universitat Pompeu Fabra
Barcelona

`michael.greenacre@upf.edu`
`www.econ.upf.edu/~michael`

`www.globalsong.net`

`www.multivariatestatistics.org`

`www.youtube.com/StatisticalSongs`
`.../ArcticFrontiers`
`.../CARMEnetwork`
`.../YouWomanGroup`
`.../GlobalSongProject`

SUMMARY:

- **MULTIVARIATE DATA & DISTANCE**
- **REGRESSION BIPLOT**
- **CORRESPONDENCE ANALYSIS BIPLOTS**
- **CONSTRAINED BIPLOTS**

(e.g. Canonical Correspondence Analysis, CCA)

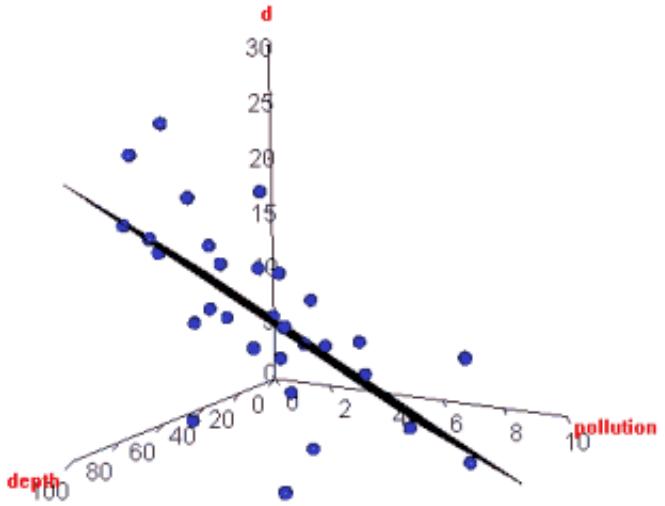
The general topic of this course is the *interpretation and quantification of relationships between sets of variables.*

I use mostly ecological data sets from marine biology, but have two medical examples at the end.

Regression of a response onto predictors

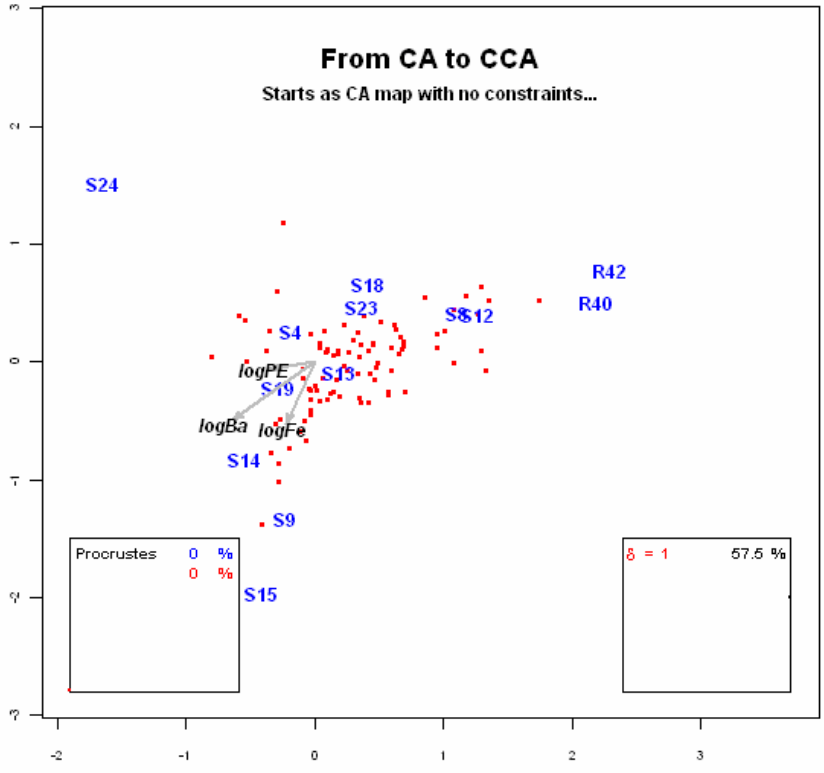
d	Depth	Pollution
14	72	4.8
11	75	2.8
8	59	5.4
3	64	8.2
10	61	3.9
16	94	2.6
11	53	4.6
0	61	5.1
14	68	3.9
9	69	10.0
6	57	6.5
15	84	3.8
0	53	9.4
9	83	4.7
12	100	6.7
3	84	2.8
0	96	6.4
20	74	4.4
16	79	3.1
9	73	5.6
4	59	4.3
9	54	1.9
17	95	2.4
7	64	4.3
23	97	2.0
10	78	2.5
25	85	2.1
20	92	3.4
8	51	6.0
18	99	1.9

$d = 6.14 - 1.39\ Pollution + 0.148\ Depth$

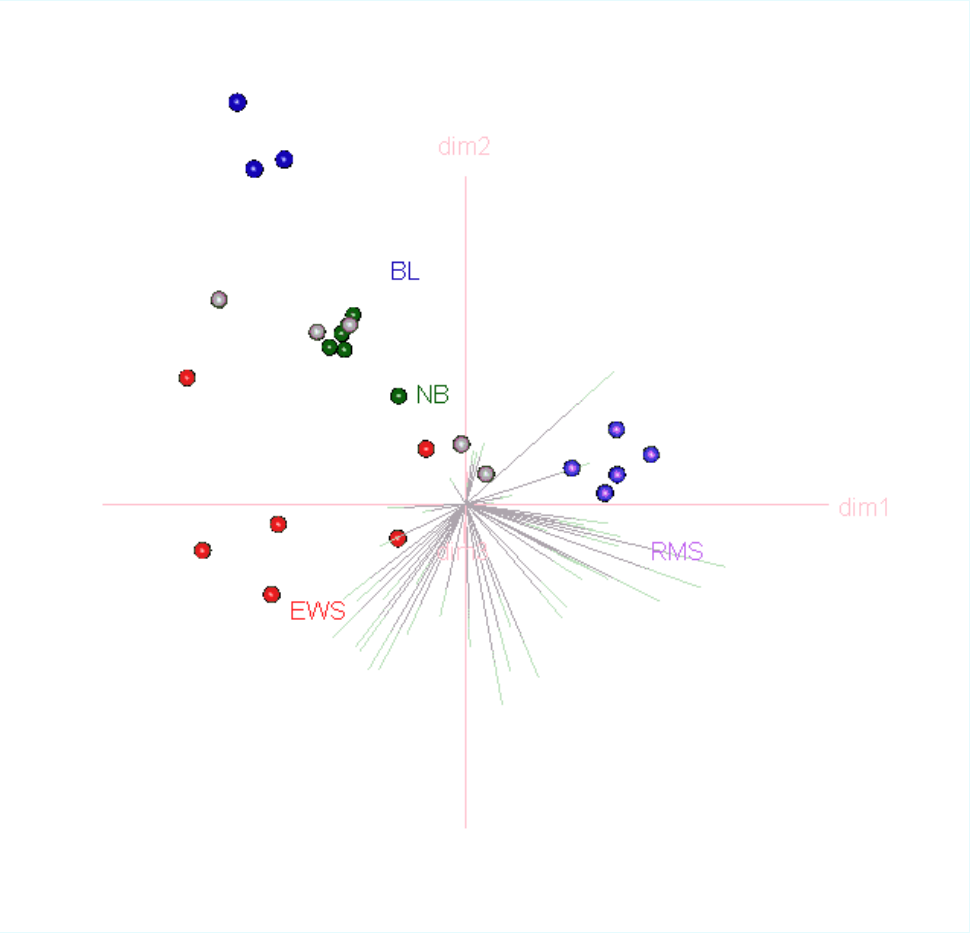


www.multivariatestatistics.org

Dynamic transition from CA to CCA



Rotation of 3-dimensional solution space where prediction takes place; test samples are classified to the closest centroid of the training set.

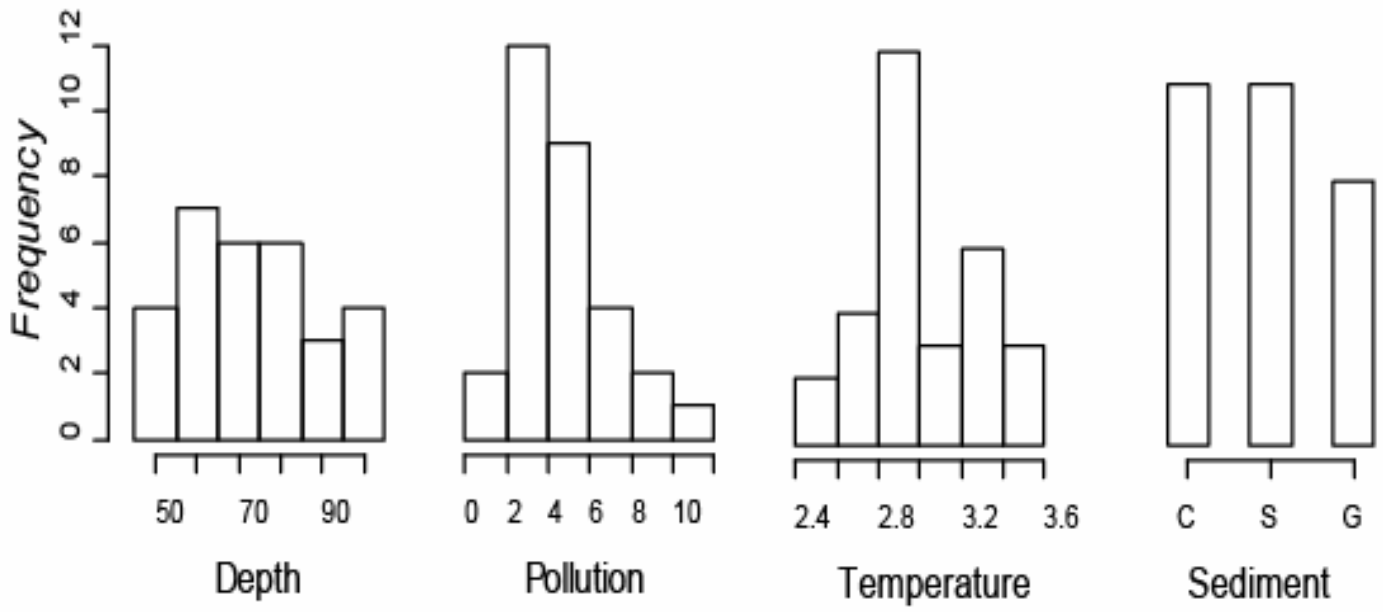


Five biological and four environmental variables (n=30 sites)

biological variables are abundance counts, three environmental variables are measured on a continuous scale, sediment is categorical: = C (clay), S (sandy) or G (gravel)

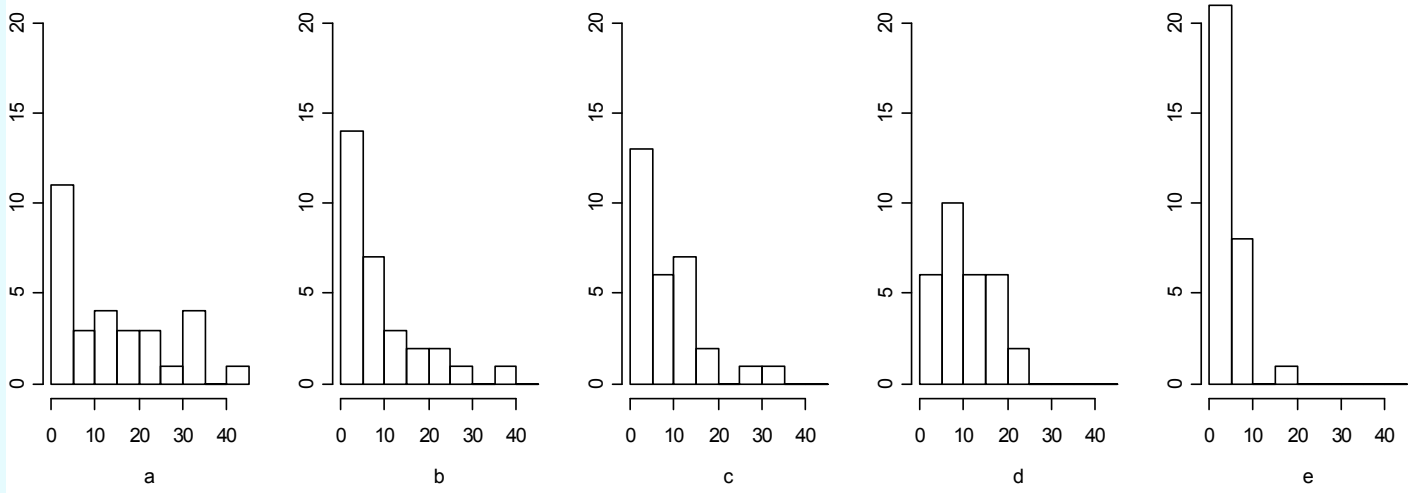
Samples	SPECIES COUNTS					ENVIRONMENTAL PREDICTORS			
	a	b	c	d	e	Depth	Pollution	Temperature	Sediment
s1	0	2	9	14	2	72	4.8	3.5	S
s2	26	4	13	11	0	75	2.8	2.5	C
s3	0	10	9	8	0	59	5.4	2.7	C
s4	0	0	15	3	0	64	8.2	2.9	S
s5	13	5	3	10	7	61	3.9	3.1	C
s6	31	21	13	16	5	94	2.6	3.5	G
s7	9	6	0	11	2	53	4.6	2.9	S
s8	2	0	0	0	1	61	5.1	3.3	C
s9	17	7	10	14	6	68	3.9	3.4	C
s10	0	5	26	9	0	69	10.0	3.0	S
s11	0	8	8	6	7	57	6.5	3.3	C
s12	14	11	13	15	0	84	3.8	3.1	S
s13	0	0	19	0	6	53	9.4	3.0	S
s14	13	0	0	9	0	83	4.7	2.5	C
s15	4	0	10	12	0	100	6.7	2.8	C
s16	42	20	0	3	6	84	2.8	3.0	G
s17	4	0	0	0	0	96	6.4	3.1	C
s18	21	15	33	20	0	74	4.4	2.8	G
s19	2	5	12	16	3	79	3.1	3.6	S
s20	0	10	14	9	0	73	5.6	3.0	S
s21	8	0	0	4	6	59	4.3	3.4	C
s22	35	10	0	9	17	54	1.9	2.8	S
s23	6	7	1	17	10	95	2.4	2.9	G
s24	18	12	20	7	0	64	4.3	3.0	C
s25	32	26	0	23	0	97	2.0	3.0	G
s26	32	21	0	10	2	78	2.5	3.4	S
s27	24	17	0	25	6	85	2.1	3.0	G
s28	16	3	12	20	2	92	3.4	3.3	G
s29	11	0	7	8	0	51	6.0	3.0	S
s30	24	37	5	18	1	99	1.9	2.9	G

Distributions



- Histograms of continuous variables: Bars are shown connected. Depth distribution looks uniform; Pollution and Temperature look skew.
- Bar-chart of categorical variable: Bars are shown separated for three categories of Sediment. Distribution approximately uniform.

Five species



- Histograms of count variables: all the distribution are 'skew to the right'

Scatterplots: continuous variables

Showing smoothed relationship between the variables ("scatterplot smoother").

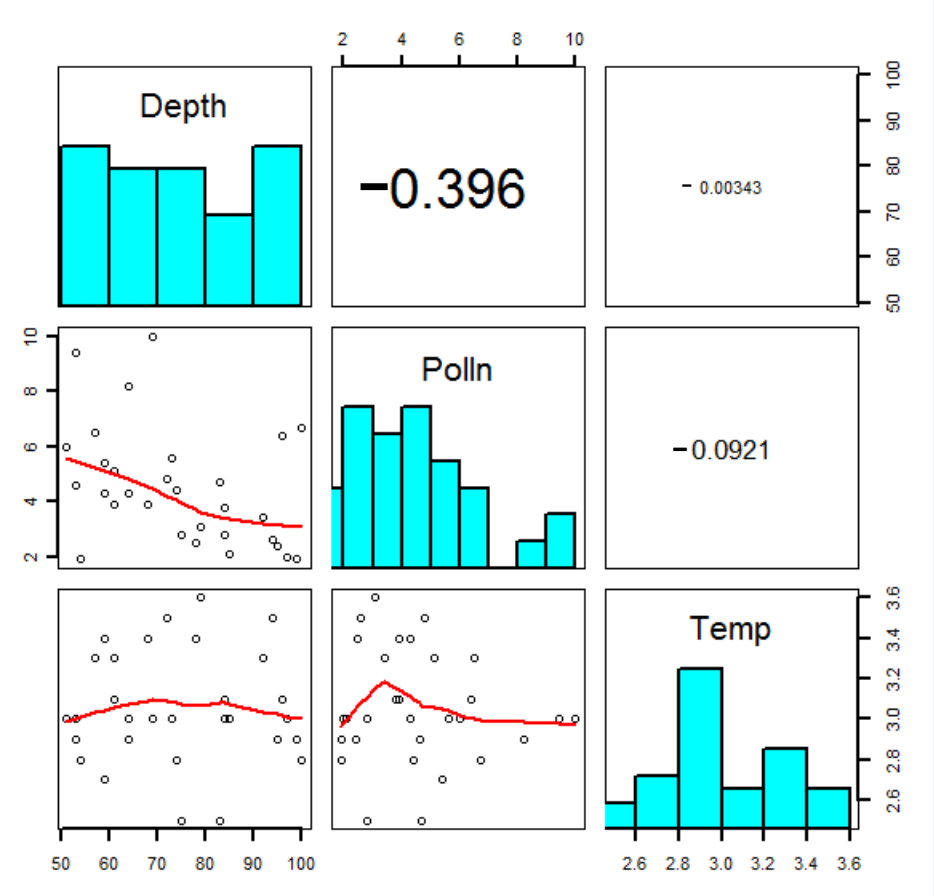
The correlations are shown opposite their scatterplot. Assuming normality, the critical value at 5% level (two-sided test) is ±0.361

Pearson's r = −0.396 ($P = 0.030$)

Permutation test on r ($P = 0.027$)

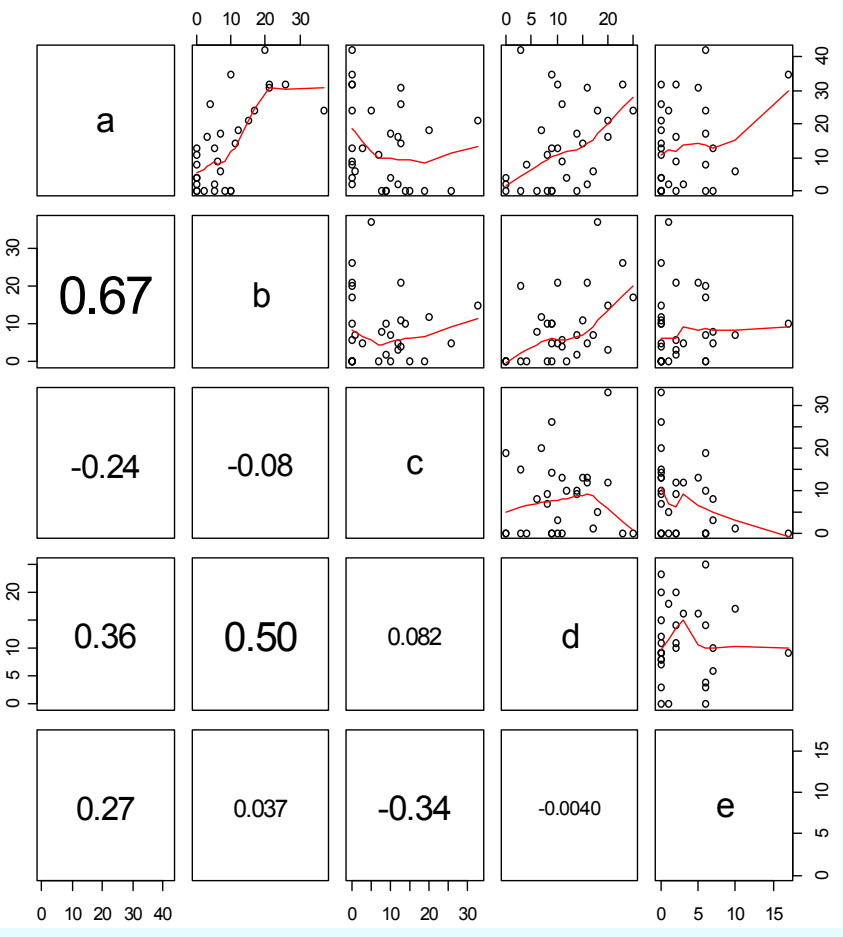
Spearman ρ = −0.423 ($P = 0.020$)

Permutation test on ρ ($P = 0.022$)



function **`pairs`** in R

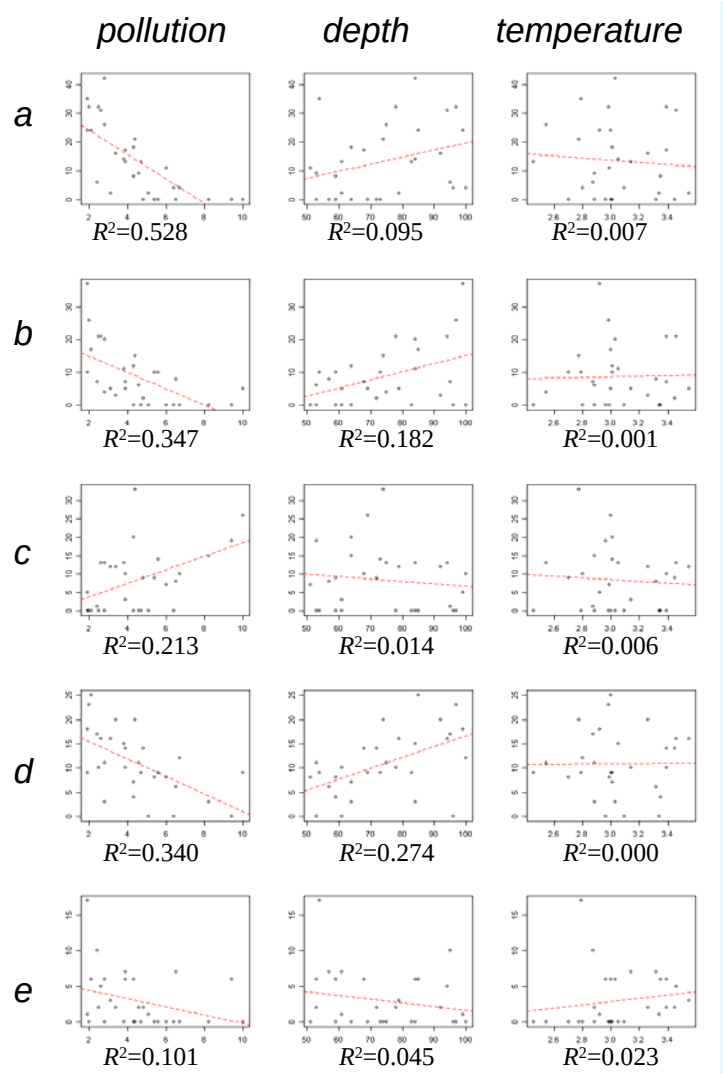
Scatterplots: count variables



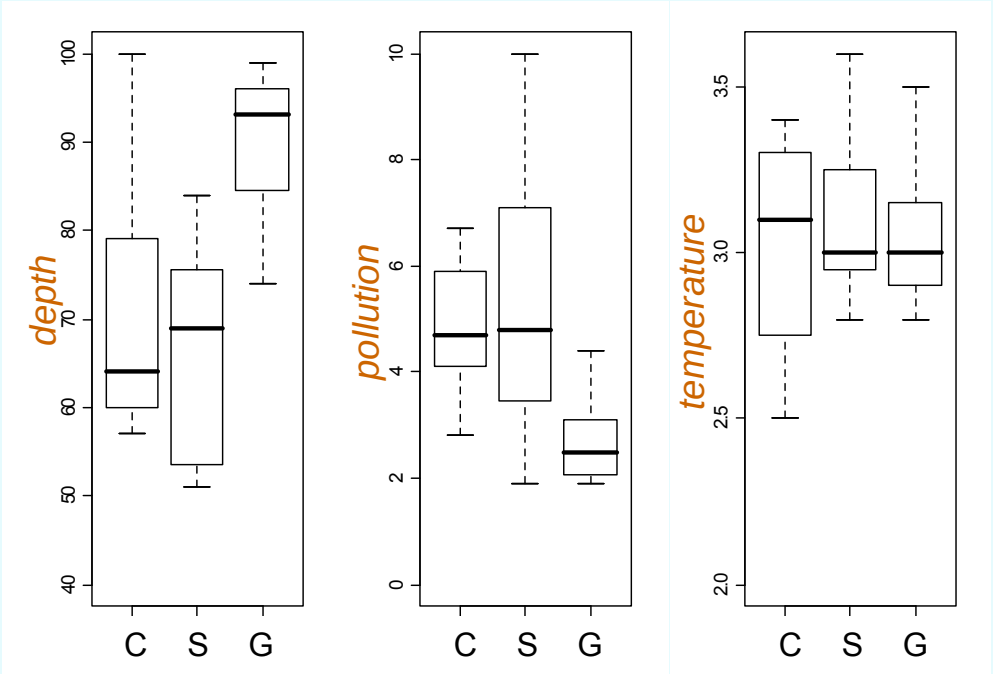
Scatterplots: count vs continuous

Showing coefficients of determination (measure of explained variance, usually expressed as a percentage) – in this simple linear regression case R^2 is the square of r, the correlation coefficient. The sign of r can be deduced by the slope of the regression line.

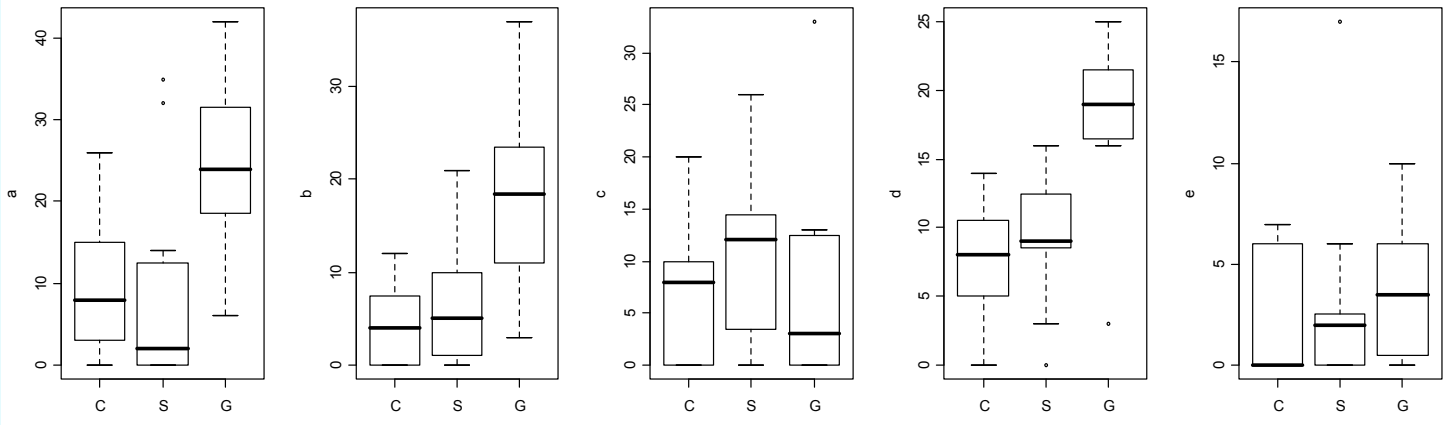
Often the count variable is transformed, for example by a power transformation.



Box-and-whisker plots: continuous vs categorical



Box-and-whisker plots: count vs categorical

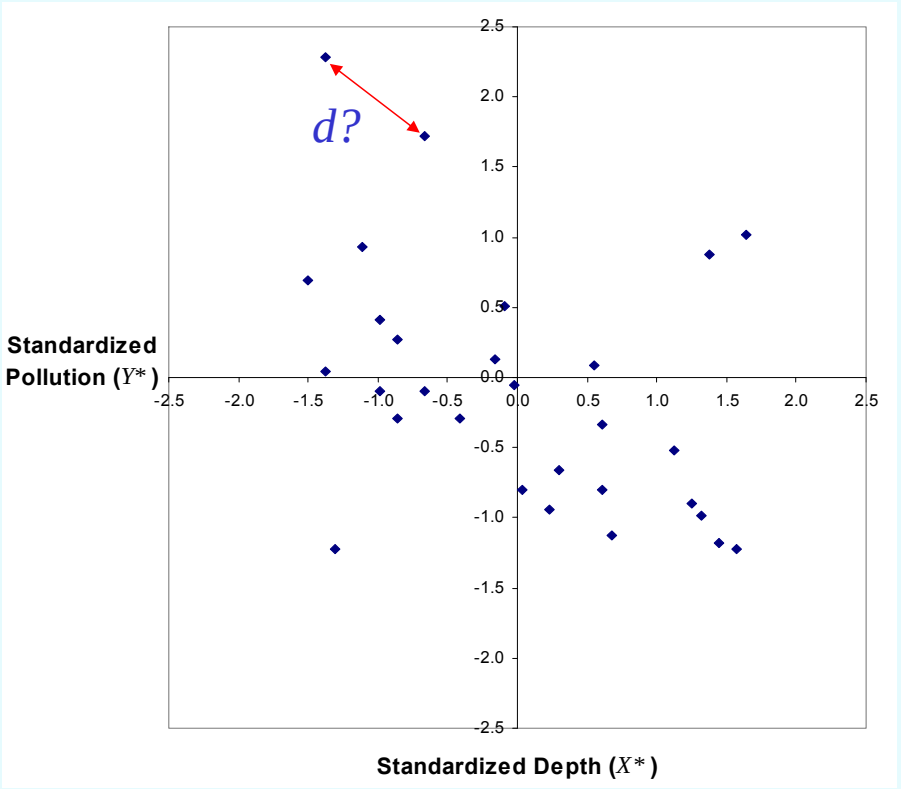


$F = 5.53$ $F = 9.74$ $F = 0.54$ $F = 9.61$ $F = 0.24$

Showing F statistics for the analysis of variance (ANOVA), which tests the difference in group means. In this case the critical point of the test at the 1% level is 5.49.

Standardized Euclidean distance

y	x	y^*	x^*
4.8	72	0.132	-0.156
2.8	75	-0.802	0.036
5.4	59	0.413	-0.988
8.2	64	1.720	-0.668
3.9	61	-0.288	-0.860
2.6	94	-0.895	1.253
4.6	53	0.039	-1.373
5.1	61	0.272	-0.860
3.9	68	-0.288	-0.412
10.0	69	2.561	-0.348
6.5	57	0.926	-1.116
3.8	84	-0.335	0.613
9.4	53	2.281	-1.373
4.7	83	0.086	0.549
6.7	100	1.020	1.637
2.8	84	-0.802	0.613
6.4	96	0.880	1.381
4.4	74	-0.054	-0.028
3.1	79	-0.662	0.292
5.6	73	0.506	-0.092
4.3	59	-0.101	-0.988
1.9	54	-1.222	-1.309
2.4	95	-0.989	1.317
4.3	64	-0.101	-0.668
2.0	97	-1.175	1.445
2.5	78	-0.942	0.228
2.1	85	-1.129	0.677
3.4	92	-0.522	1.125
6.0	51	0.693	-1.501
1.9	99	-1.222	1.573



$$x^* = [x - \text{mean}(x)]/\text{sd}(x) = [x - 4.52]/2.14$$
$$y^* = [y - \text{mean}(y)]/\text{sd}(y) = [y - 74.4]/15.6$$

(Weighted) Euclidean distance

y	x
4.8	72
2.8	75
5.4	59
8.2	64
3.9	61
2.6	94
4.6	53
5.1	61
3.9	68
10.0	69
6.5	57
3.8	84
9.4	53
4.7	83
6.7	100
2.8	84
6.4	96
4.4	74
3.1	79
5.6	73
4.3	59
1.9	54
2.4	95
4.3	64
2.0	97
2.5	78
2.1	85
3.4	92
6.0	51
1.9	99

distance between rows (sites) i and i'

$$d(i,i') = \sqrt{(x_i^* - x_{i'}^*)^2 + (y_i^* - y_{i'}^*)^2}$$

$$= \sqrt{\left(\frac{x_i - \bar{x}}{s_x} - \frac{x_{i'} - \bar{x}}{s_x}\right)^2 + \left(\frac{y_i - \bar{y}}{s_y} - \frac{y_{i'} - \bar{y}}{s_y}\right)^2}$$

$$= \sqrt{\left(\frac{x_i}{s_x} - \frac{x_{i'}}{s_x}\right)^2 + \left(\frac{y_i}{s_y} - \frac{y_{i'}}{s_y}\right)^2}$$

$$= \sqrt{\frac{1}{s_x^2}(x_i - x_{i'})^2 + \frac{1}{s_y^2}(y_i - y_{i'})^2}$$

"weights", inverses of variances

For more than two variables, just add similar terms for the other variables

y^*	x^*
0.132	-0.156
-0.802	0.036
0.413	-0.988
1.720	-0.668
-0.288	-0.860
-0.895	1.253
0.039	-1.373
0.272	-0.860
-0.288	-0.412
2.561	-0.348
0.926	-1.116
-0.335	0.613
2.281	-1.373
0.086	0.549
1.020	1.637
-0.802	0.613
0.880	1.381
-0.054	-0.028
-0.662	0.292
0.506	-0.092
-0.101	-0.988
-1.222	-1.309
-0.989	1.317
-0.101	-0.668
-1.175	1.445
-0.942	0.228
-1.129	0.677
-0.522	1.125
0.693	-1.501
-1.222	1.573

Inter-site distances

for the three continuous variables depth, pollution and temperature

$$d(i,i') = \sqrt{\frac{1}{s_x^2}(x_i - x_{i'})^2 + \frac{1}{s_y^2}(y_i - y_{i'})^2 + \frac{1}{s_z^2}(z_i - z_{i'})^2}$$

	s1	s2	s3	s4	s5	s6	· · ·	s24	s25	s26	s27	s28	s29
s2	3.681												
s3	2.977	1.741											
s4	2.708	2.980	1.523										
s5	1.642	2.371	1.591	2.139									
s6	1.744	3.759	3.850	3.884	2.619								
s7	2.458	2.171	0.890	1.823	0.935	3.510							
.	.	.	.	.	.	.	.	.					
.	.	.	.	.	.	.	.	.					
.	.	.	.	.	.	.	.	.					
s25	2.727	2.299	3.095	3.602	2.496	1.810	· · ·	2.371					
s26	1.195	3.209	3.084	3.324	1.658	1.086	· · ·	1.880	1.886				
s27	2.333	1.918	2.507	3.170	1.788	1.884	· · ·	1.692	0.770	1.503			
s28	1.604	3.059	3.145	3.204	2.122	0.813	· · ·	2.128	1.291	1.052	1.307		
s29	2.299	2.785	1.216	1.369	1.224	3.642	· · ·	1.150	3.488	2.772	2.839	3.083	
s30	3.062	2.136	3.121	3.699	2.702	2.182	· · ·	2.531	0.381	2.247	0.969	1.648	3.639

Distance between rows of matrix $\mathbf{X} = [x_{ij}]$

$$d(i,i') = \sqrt{\sum_{j=1}^{m} \frac{1}{s_j^2}(x_{ij} - x_{i'j})^2}$$

Chi-square distance

SITE	a	b	c	d	e	sum
s1	0	2	9	14	2	27
s2	26	4	13	11	0	54
s3	0	10	9	8	0	27
s4	0	0	15	3	0	18
s5	13	5	3	10	7	38
s6	31	21	13	16	5	86
s7	9	6	0	11	2	28
s8	2	0	0	0	1	3
s9	17	7	10	14	6	54
s10	0	5	26	9	0	40
s11	0	8	8	6	7	29
s12	14	11	13	15	0	53
s13	0	0	19	0	6	25
s14	13	0	0	9	0	22
s15	4	0	10	12	0	26
s16	42	20	0	3	6	71
s17	4	0	0	0	0	4
s18	21	15	33	20	0	89
s19	2	5	12	16	3	38
s20	0	10	14	9	0	33
s21	8	0	0	4	6	18
s22	35	10	0	9	17	71
s23	6	7	1	17	10	41
s24	18	12	20	7	0	57
s25	32	26	0	23	0	81
s26	32	21	0	10	2	65
s27	24	17	0	25	6	72
s28	16	3	12	20	2	53
s29	11	0	7	8	0	26
s30	24	37	5	18	1	85
sum	404	262	252	327	89	1334

SITE	a	b	c	d	e
s1	0.000	0.074	0.333	0.519	0.074
s2	0.481	0.074	0.241	0.204	0.000
s3	0.000	0.370	0.333	0.296	0.000
s4	0.000	0.000	0.833	0.167	0.000
s5	0.342	0.132	0.079	0.263	0.184
s6	0.360	0.244	0.151	0.186	0.058
s7	0.321	0.214	0.000	0.393	0.071
s8	0.667	0.000	0.000	0.000	0.333
s9	0.315	0.130	0.185	0.259	0.111
s10	0.000	0.125	0.650	0.225	0.000
s11	0.000	0.276	0.276	0.207	0.241
s12	0.264	0.208	0.245	0.283	0.000
s13	0.000	0.000	0.760	0.000	0.240
s14	0.591	0.000	0.000	0.409	0.000
s15	0.154	0.000	0.385	0.462	0.000
s16	0.592	0.282	0.000	0.042	0.085
s17	1.000	0.000	0.000	0.000	0.000
s18	0.236	0.169	0.371	0.225	0.000
s19	0.053	0.132	0.316	0.421	0.079
s20	0.000	0.303	0.424	0.273	0.000
s21	0.444	0.000	0.000	0.222	0.333
s22	0.493	0.141	0.000	0.127	0.239
s23	0.146	0.171	0.024	0.415	0.244
s24	0.316	0.211	0.351	0.123	0.000
s25	0.395	0.321	0.000	0.284	0.000
s26	0.492	0.323	0.000	0.154	0.031
s27	0.333	0.236	0.000	0.347	0.083
s28	0.302	0.057	0.226	0.377	0.038
s29	0.423	0.000	0.269	0.308	0.000
s30	0.282	0.435	0.059	0.212	0.012
ave.	0.303	0.196	0.189	0.245	0.067

The **chi-square distance** is an alternative measure of dissimilarity that is applicable to frequencies and which is the basis of correspondence analysis (CA).

First convert the rows of frequencies into **profiles**, that is the frequencies relative to their totals. The row of column sums is also expressed as proportions, called the **average profile**.

Chi-square distance

SITE	a	b	c	d	e
s1	0.000	0.074	0.333	0.519	0.074
s2	0.481	0.074	0.241	0.204	0.000
s3	0.000	0.370	0.333	0.296	0.000
s4	0.000	0.000	0.833	0.167	0.000
s5	0.342	0.132	0.079	0.263	0.184
s6	0.360	0.244	0.151	0.186	0.058
s7	0.321	0.214	0.000	0.393	0.071
s8	0.667	0.000	0.000	0.000	0.333
s9	0.315	0.130	0.185	0.259	0.111
s10	0.000	0.125	0.650	0.225	0.000
s11	0.000	0.276	0.276	0.207	0.241
s12	0.264	0.208	0.245	0.283	0.000
s13	0.000	0.000	0.760	0.000	0.240
s14	0.591	0.000	0.000	0.409	0.000
s15	0.154	0.000	0.385	0.462	0.000
s16	0.592	0.282	0.000	0.042	0.085
s17	1.000	0.000	0.000	0.000	0.000
s18	0.236	0.169	0.371	0.225	0.000
s19	0.053	0.132	0.316	0.421	0.079
s20	0.000	0.303	0.424	0.273	0.000
s21	0.444	0.000	0.000	0.222	0.333
s22	0.493	0.141	0.000	0.127	0.239
s23	0.146	0.171	0.024	0.415	0.244
s24	0.316	0.211	0.351	0.123	0.000
s25	0.395	0.321	0.000	0.284	0.000
s26	0.492	0.323	0.000	0.154	0.031
s27	0.333	0.236	0.000	0.347	0.083
s28	0.302	0.057	0.226	0.377	0.038
s29	0.423	0.000	0.269	0.308	0.000
s30	0.282	0.435	0.059	0.212	0.012
ave.	0.303	0.196	0.189	0.245	0.067

Calculate the chi-square distance between the first two sites as follows:

$$d(1,2)^2 = \frac{(0-0.481)^2}{0.303} + \frac{(0.074-0.074)^2}{0.196} + \frac{(0.333-0.241)^2}{0.189} + \frac{(0.519-0.204)^2}{0.245} + \frac{(0.074-0)^2}{0.067}$$

$$= 1.297$$

Therefore

$$d(1,2) = 1.139$$

In general, for profiles x_{ij} (sample i, species j) with average profile $x_{.j}$ (for species j), the chi-square distance between samples is:

$$d(i,i') = \sqrt{\sum_j \frac{(x_{ij} - x_{i'j})^2}{x_{.j}}}$$

standardization using average

Chi-square distance

	s1	s2	s3	s4	s5	s6	· · ·	s24	s25	s26	s27	s28	s29
s2	1.139												
s3	0.855	1.137											
s4	1.392	1.630	1.446										
s5	1.093	0.862	1.238	2.008									
s6	1.099	0.539	0.887	1.802	0.597								
s7	1.046	0.845	1.081	2.130	0.573	0.555							
.	.	.	.	.	.	.	.						
.	.	.	.	.	.	.	.						
.	.	.	.	.	.	.	.						
s25	1.312	0.817	1.057	2.185	0.858	0.495	· · ·	0.917					
s26	1.508	0.805	1.224	2.241	0.834	0.475	· · ·	0.915	0.338				
s27	1.100	0.837	1.078	2.136	0.520	0.489	· · ·	0.983	0.412	0.562			
s28	0.681	0.504	0.954	1.572	0.724	0.613	· · ·	0.699	0.844	0.978	0.688		
s29	0.951	0.296	1.145	1.535	0.905	0.708	· · ·	0.662	0.956	1.021	0.897	0.340	
s30	1.330	0.986	0.846	2.101	0.970	0.535	· · ·	0.864	0.388	0.497	0.617	1.001	1.142

For chi-square distances on “non-relativized”, “raw” abundances, see

Greenacre, M (2010) Correspondence analysis of raw data. *Ecology* 91, 958–963

Presence/absence data

SITES	SPECIES S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
A	1	1	1	0	1	0	0	1	1	1
B	1	1	0	1	1	0	0	0	0	1
C	0	1	1	0	1	0	0	1	0	0
D	0	0	0	1	0	1	0	0	0	0
E	1	1	1	0	1	0	1	1	1	0
F	0	1	0	1	1	0	0	0	0	1
G	0	1	1	0	1	1	0	1	1	0

A \ B	1	0	
1	4	3	7
0	1	2	3
	5	5	10

Similarity $s_{12} = 6/10 = 0.6$

Dissimilarity $d_{12} = 4/10 = 0.4 = 1 - s_{12}$

In general there are n sites, p species, and we can count combinations for each pair of sites i and i' as:

i \ i'	1	0	
1	a	b	$a+b$
0	c	d	$c+d$
	$a+c$	$b+d$	p

Similarity $s_{ii'} = (a+d)/p$

Dissimilarity $d_{ii'} = 1 - s_{ii'} = (b+c)/p$

Dealing with heterogeneous data

	Continuous variables			*Sampled region*				*Substrate character*				
Station	*Depth*	*Temperature*	*Salinity*	*Tarehola*	*Skognes*	*Njosken*	*Storura*	*Clay*	*Silt*	*Sand*	*Gravel*	*Stone*
s3	30	3.15	33.52	1	0	0	0	0	1	0	0	1
s8	29	3.15	33.52	1	0	0	0	1	0	0	1	0
s25	30	3.00	33.45	0	1	0	0	1	0	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
s84	66	3.22	33.48	0	0	0	1	1	0	0	0	0
mean	58.15	3.086	33.50	0.242	0.273	0.242	0.242	0.606	0.152	0.364	0.182	0.061
s.d.	32.45	0.100	0.076	0.435	0.452	0.435	0.435	0.496	0.364	0.489	0.392	0.242

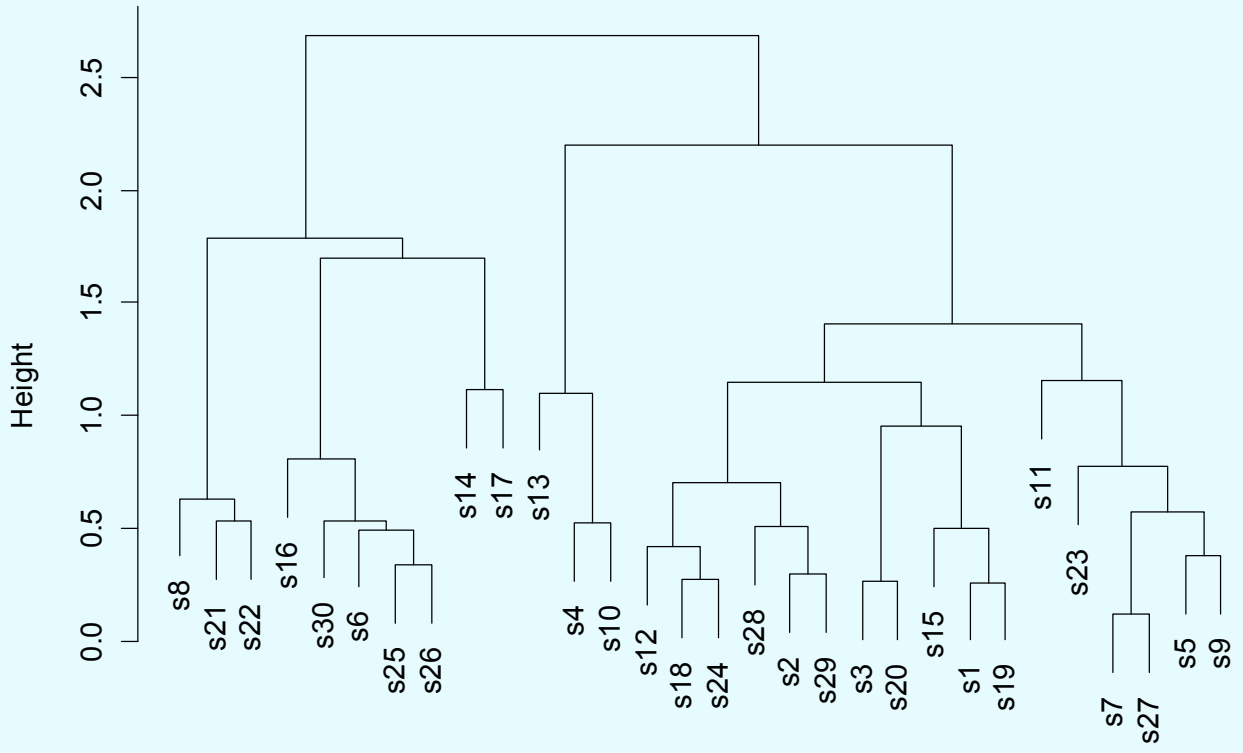
Gower's general coefficient of (dis)similarity: standardize each variable and multiply all the columns corresponding to dummy variables by $1/\sqrt{2} = 0.7071$***, a factor which compensates for their 0/1 coding:***

	Continuous variables			*Sampled region*				*Substrate character*				
Station	*Depth*	*Temperature*	*Salinity*	*Tarehola*	*Skognes*	*Njosken*	*Storura*	*Clay*	*Silt*	*Sand*	*Gravel*	*Stone*
s3	-0.868	0.615	0.260	1.231	-0.426	-0.394	-0.394	-0.864	1.648	-0.526	-0.328	2.741
s8	-0.898	0.615	0.260	1.231	-0.426	-0.394	-0.394	0.561	-0.294	-0.526	1.477	-0.177
s25	-0.868	-0.854	-0.676	-0.394	1.137	-0.394	-0.394	0.561	-0.294	0.921	-0.328	-0.177
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
s84	0.242	1.294	-0.294	-0.394	-0.426	-0.394	1.231	0.561	-0.294	-0.526	-0.328	-0.177

compute Euclidean distance between stations

Clustering of sites

based on chi-square distances calculated on species abundances

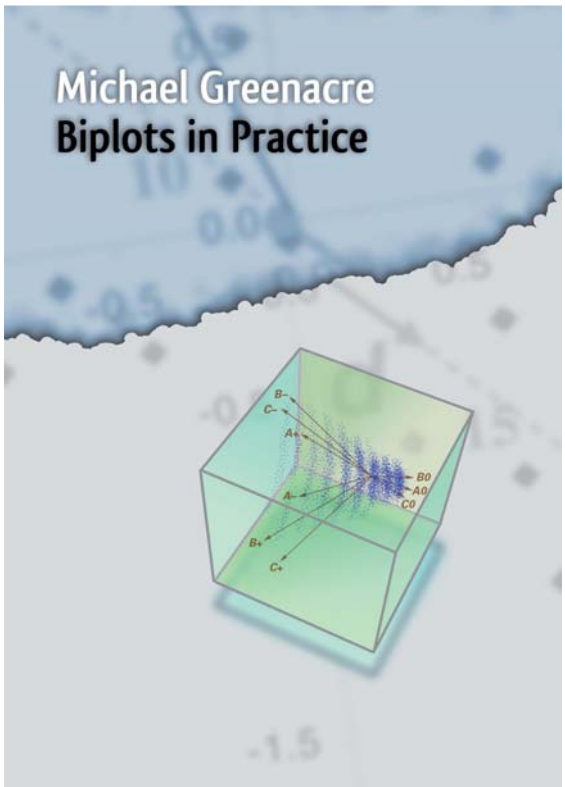


www.multivariatestatistics.org

Website of book

Biplots in Practice

published September 2010 by the BBVA Foundation in Madrid and available online for free download. Includes R code for all the analyses in the book, as well as data sets, videos and a searchable glossary in English and Spanish.

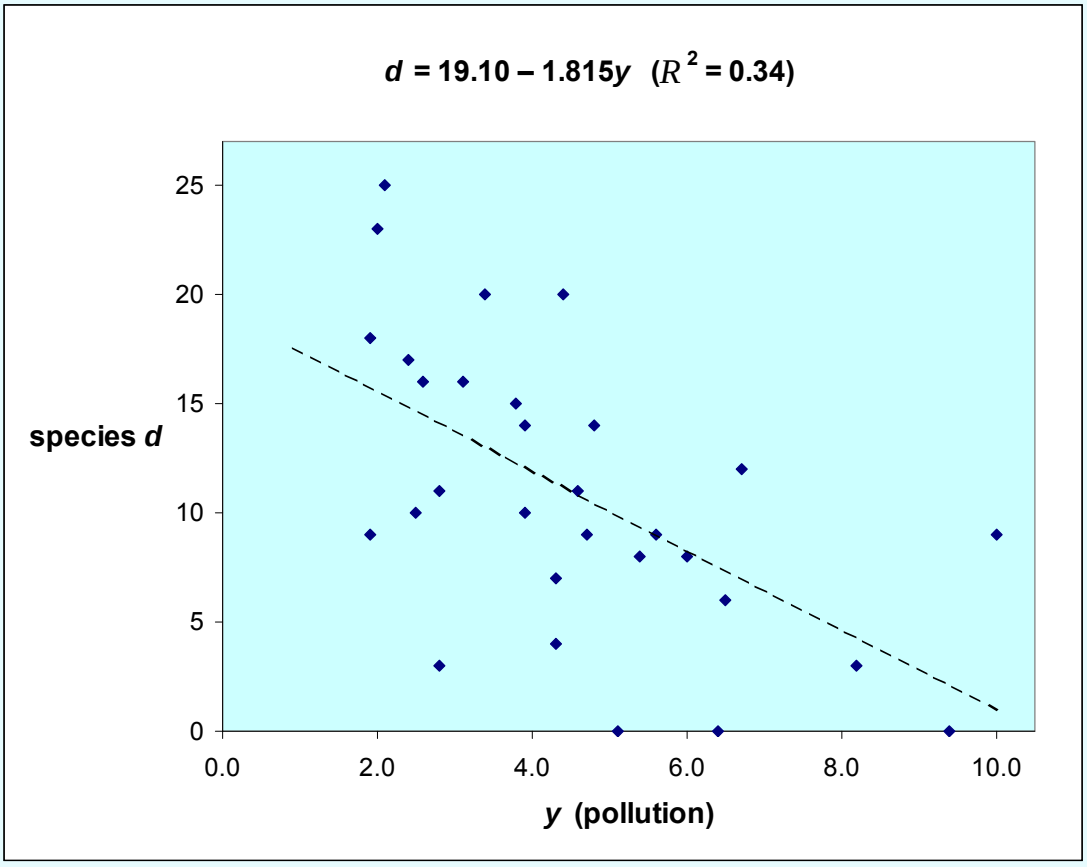


Coming soon in same series:

Multivariate Analysis of Ecological Data

Simple linear regression species *d* versus pollution (*y*)

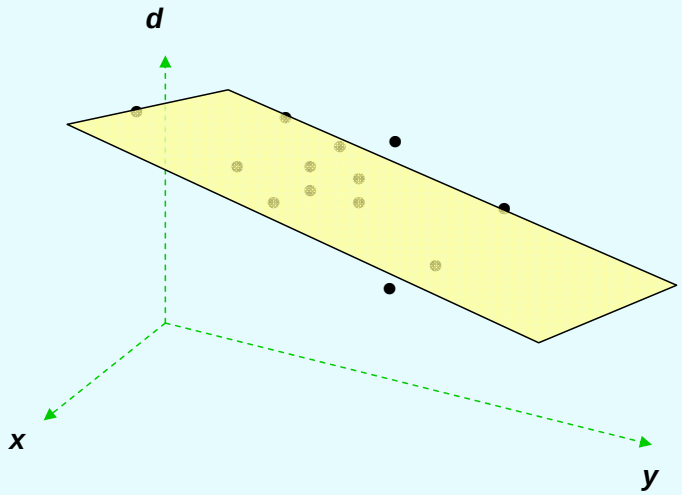
d	*y* (pollution)
14	4.8
11	2.8
8	5.4
3	8.2
10	3.9
16	2.6
11	4.6
0	5.1
14	3.9
9	10.0
6	6.5
15	3.8
0	9.4
9	4.7
12	6.7
3	2.8
0	6.4
20	4.4
16	3.1
9	5.6
4	4.3
9	1.9
17	2.4
7	4.3
23	2.0
10	2.5
25	2.1
20	3.4
8	6.0
18	1.9



d	*y*	*x*
14	4.8	72
11	2.8	75
8	5.4	59
3	8.2	64
10	3.9	61
16	2.6	94
11	4.6	53
0	5.1	61
14	3.9	68
9	10.0	69
6	6.5	57
15	3.8	84
0	9.4	53
9	4.7	83
12	6.7	100
3	2.8	84
0	6.4	96
20	4.4	74
16	3.1	79
9	5.6	73
4	4.3	59
9	1.9	54
17	2.4	95
7	4.3	64
23	2.0	97
10	2.5	78
25	2.1	85
20	3.4	92
8	6.0	51
18	1.9	99

Multiple linear regression

$$\boldsymbol{d} = 6.14 - 1.389\boldsymbol{y} + 0.148\boldsymbol{x}$$

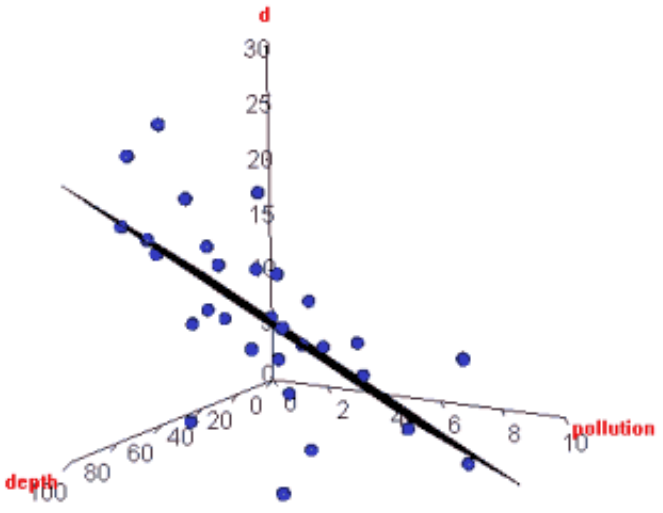


Regression model is a (hyper)plane

Regression of a response onto predictors

d	*Depth*	*Pollution*
14	72	4.8
11	75	2.8
8	59	5.4
3	64	8.2
10	61	3.9
16	94	2.6
11	53	4.6
0	61	5.1
14	68	3.9
9	69	10.0
6	57	6.5
15	84	3.8
0	53	9.4
9	83	4.7
12	100	6.7
3	84	2.8
0	96	6.4
20	74	4.4
16	79	3.1
9	73	5.6
4	59	4.3
9	54	1.9
17	95	2.4
7	64	4.3
23	97	2.0
10	78	2.5
25	85	2.1
20	92	3.4
8	51	6.0
18	99	1.9

$$d = 6.14 - 1.39\ Pollution + 0.148\ Depth$$

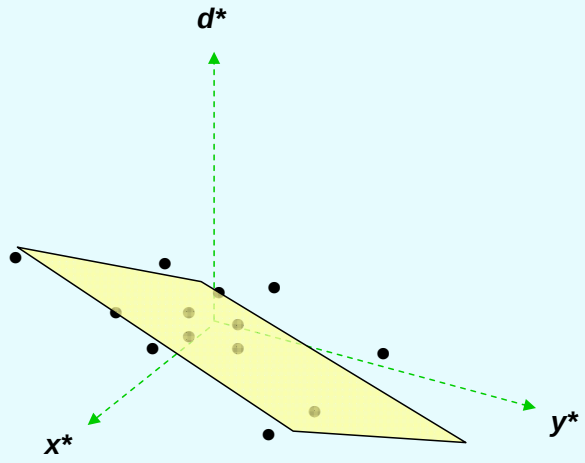


www.multivariatestatistics.org

Multiple linear regression, variables standardized

*d**	*y**	*x**
0.503	0.132	-0.156
0.052	-0.802	0.036
-0.400	0.413	-0.988
-1.152	1.720	-0.668
-0.099	-0.288	-0.860
0.804	-0.895	1.253
0.052	0.039	-1.373
-1.603	0.272	-0.860
0.503	-0.288	-0.412
-0.249	2.561	-0.348
-0.701	0.926	-1.116
0.654	-0.335	0.613
-1.603	2.281	-1.373
-0.249	0.086	0.549
0.202	1.020	1.637
-1.152	-0.802	0.613
-1.603	0.880	1.381
1.406	-0.054	-0.028
0.804	-0.662	0.292
-0.249	0.506	-0.092
-1.001	-0.101	-0.988
-0.249	-1.222	-1.309
0.955	-0.989	1.317
-0.550	-0.101	-0.668
1.858	-1.175	1.445
-0.099	-0.942	0.228
2.159	-1.129	0.677
1.406	-0.522	1.125
-0.400	0.693	-1.501
1.105	-1.222	1.573

$$d^* = -.446y^* + 0.347x^*$$

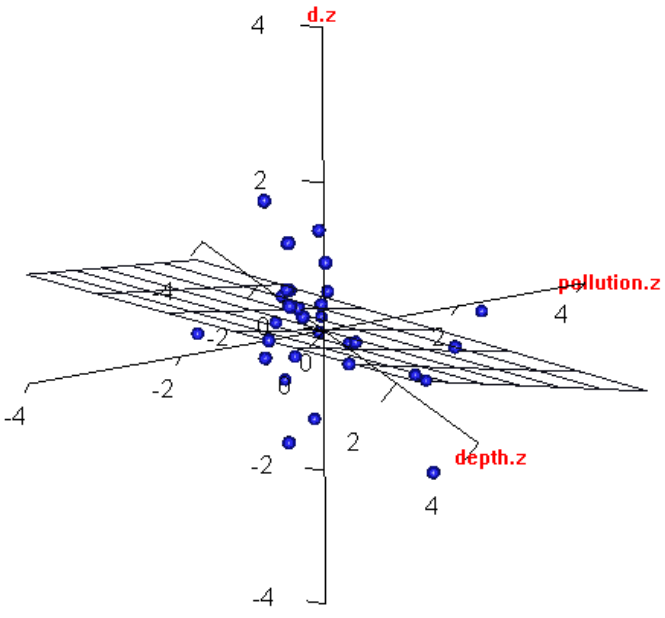


Explanatory variables ***x*** and ***y*** and response variable ***d*** standardized

Response & predictors all standardized

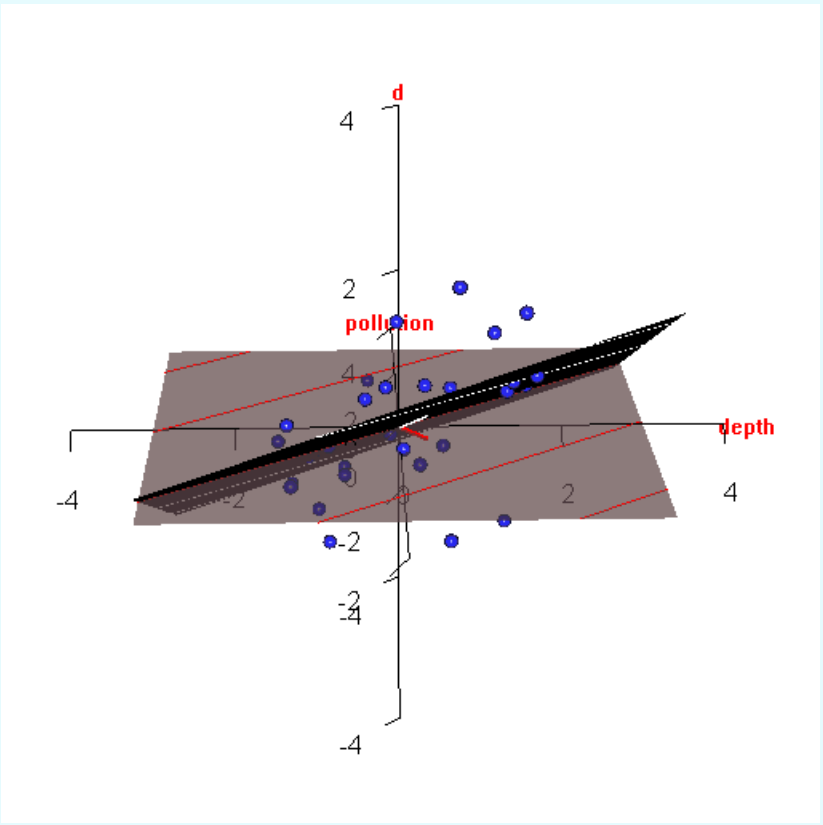
d.z	Depth.z	Pollution.z
0.47	-0.16	0.13
0.02	0.04	-0.80
-0.44	-0.99	0.41
-1.19	-0.67	1.72
-0.14	-0.86	-0.29
0.77	1.25	-0.90
0.02	-1.37	0.04
-1.64	-0.86	0.27
0.47	-0.41	-0.29
-0.29	-0.35	2.56
-0.74	-1.12	0.93
0.62	0.61	-0.33
-1.64	-1.37	2.28
-0.29	0.55	0.09
0.17	1.64	1.02
-1.19	0.61	-0.80
-1.64	1.38	0.88
1.37	-0.03	-0.05
0.77	0.29	-0.66
-0.29	-0.09	0.51
-1.04	-0.99	-0.10
-0.29	-1.31	-1.22
0.92	1.32	-0.99
-0.59	-0.67	-0.10
1.82	1.45	-1.18
-0.14	0.23	-0.94
2.12	0.68	-1.13
1.37	1.12	-0.52
-0.44	-1.50	0.69
1.07	1.57	-1.22

$$d.z = -0.446\ Pollution.z + 0.347\ Depth.z$$



Line of steepest ascent and contours

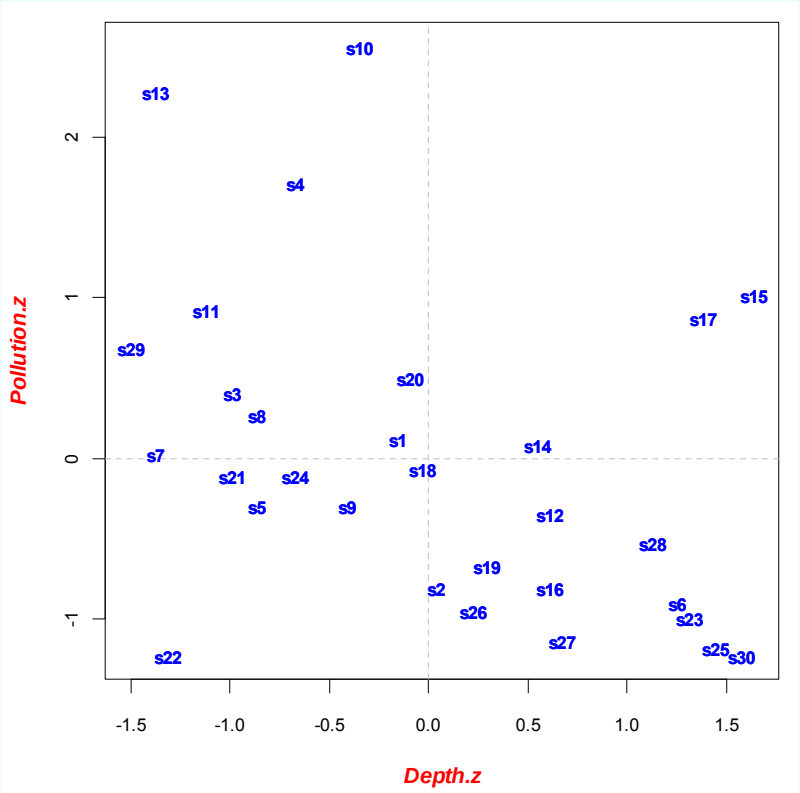
d.z	Depth.z	Pollution.z
0.47	-0.16	0.13
0.02	0.04	-0.80
-0.44	-0.99	0.41
-1.19	-0.67	1.72
-0.14	-0.86	-0.29
0.77	1.25	-0.90
0.02	-1.37	0.04
-1.64	-0.86	0.27
0.47	-0.41	-0.29
-0.29	-0.35	2.56
-0.74	-1.12	0.93
0.62	0.61	-0.33
-1.64	-1.37	2.28
-0.29	0.55	0.09
0.17	1.64	1.02
-1.19	0.61	-0.80
-1.64	1.38	0.88
1.37	-0.03	-0.05
0.77	0.29	-0.66
-0.29	-0.09	0.51
-1.04	-0.99	-0.10
-0.29	-1.31	-1.22
0.92	1.32	-0.99
-0.59	-0.67	-0.10
1.82	1.45	-1.18
-0.14	0.23	-0.94
2.12	0.68	-1.13
1.37	1.12	-0.52
-0.44	-1.50	0.69
1.07	1.57	-1.22



www.multivariatestatistics.org

Scatterplot of standardized values

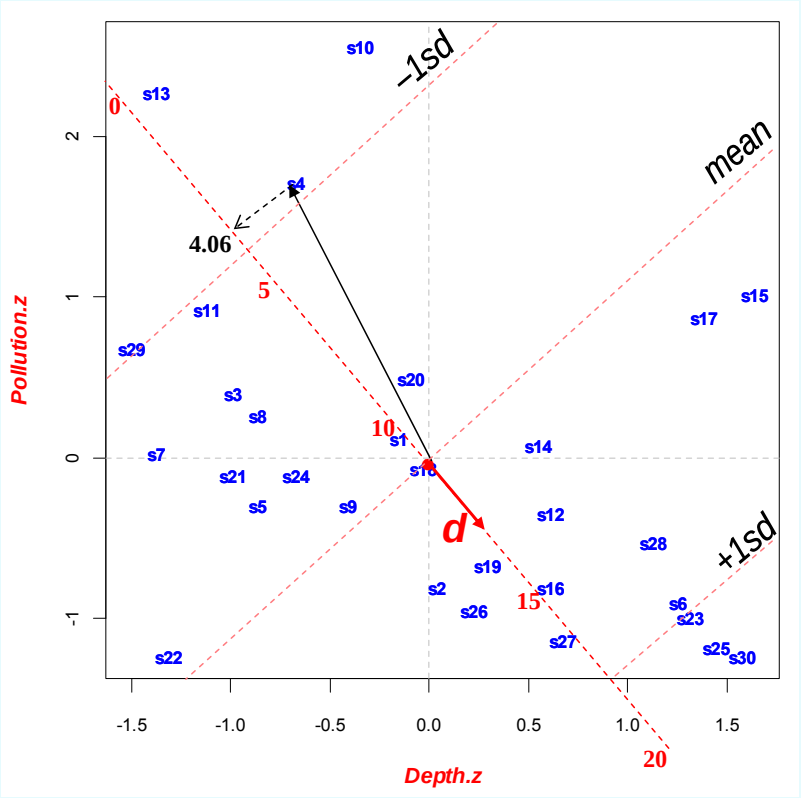
Depth.z	Pollution.z
-0.16	0.13
0.04	-0.80
-0.99	0.41
-0.67	1.72
-0.86	-0.29
1.25	-0.90
-1.37	0.04
-0.86	0.27
-0.41	-0.29
-0.35	2.56
-1.12	0.93
0.61	-0.33
-1.37	2.28
0.55	0.09
1.64	1.02
0.61	-0.80
1.38	0.88
-0.03	-0.05
0.29	-0.66
-0.09	0.51
-0.99	-0.10
-1.31	-1.22
1.32	-0.99
-0.67	-0.10
1.45	-1.18
0.23	-0.94
0.68	-1.13
1.12	-0.52
-1.50	0.69
1.57	-1.22



Everything on the "ground"

d		d.z	Depth.z	Pollution.z
	s1	0.47	-0.16	0.13
	s2	0.02	0.04	-0.80
	s3	-0.44	-0.99	0.41
3	s4	-1.19	-0.67	1.72
	s5	-0.14	-0.86	-0.29
	s6	0.77	1.25	-0.90
	s7	0.02	-1.37	0.04
	s8	-1.64	-0.86	0.27
	s9	0.47	-0.41	-0.29
	s10	-0.29	-0.35	2.56
	s11	-0.74	-1.12	0.93
	s12	0.62	0.61	-0.33
0	s13	-1.64	-1.37	2.28
	s14	-0.29	0.55	0.09
	s15	0.17	1.64	1.02
	s16	-1.19	0.61	-0.80
	s17	-1.64	1.38	0.88
	s18	1.37	-0.03	-0.05
	s19	0.77	0.29	-0.66
	s20	-0.29	-0.09	0.51
	s21	-1.04	-0.99	-0.10
	s22	-0.29	-1.31	-1.22
	s23	0.92	1.32	-0.99
	s24	-0.59	-0.67	-0.10
23	s25	1.82	1.45	-1.18
	s26	-0.14	0.23	-0.94
	s27	2.12	0.68	-1.13
	s28	1.37	1.12	-0.52
	s29	-0.44	-1.50	0.69
18	s30	1.07	1.57	-1.22

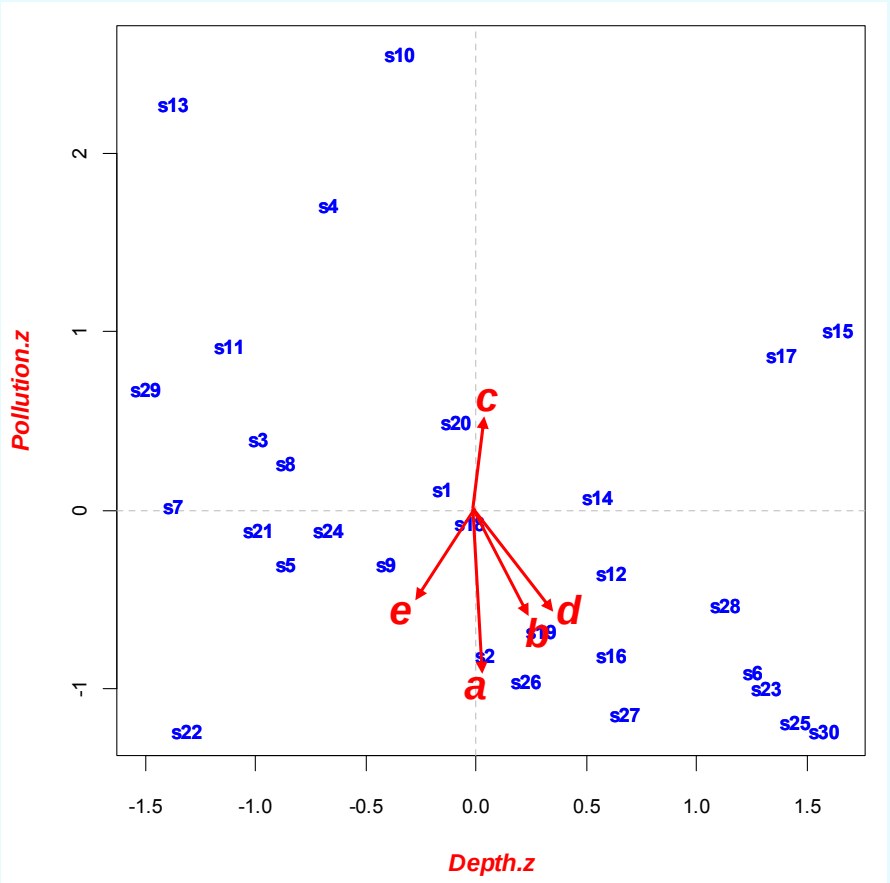
Error is – 1.06



$$d.z = -0.446\,Pollution.z + 0.347\,Depth.z$$

variance explained (R^2): 44.2%

Regressions of all five responses



variance explained:

a: 52.9%

b: 39.1%

c: 21.8%

d: 44.2%

e: 23.5%

overall: 36.3%

significance:

	P	*D*
a:	**	
b:	**	
c:	*	
d:	**	*
e:	*	*

Generalized linear model biplots

- Usual linear regression model:

  Mean μ of species d : $\mu = c + ax + by$

  d normally distributed about this (varying) mean

- Poisson regression model:

  Log of mean μ of species d : $\log(\mu) = c^* + a^* x + b^* y$

  d Poisson distributed …

- Logistic regression model:

  Logit of probability p of species d: $\log\left(\frac{p}{1-p}\right) = c^\circ + a^\circ x + b^\circ y$

  Presence of d binomially distributed …

$$\text{logit}(p_d) = \log\left(\frac{p_d}{1-p_d}\right) = 2.712 - 1.177\,y^* - 0.137\,x^*$$

Logistic regression biplots

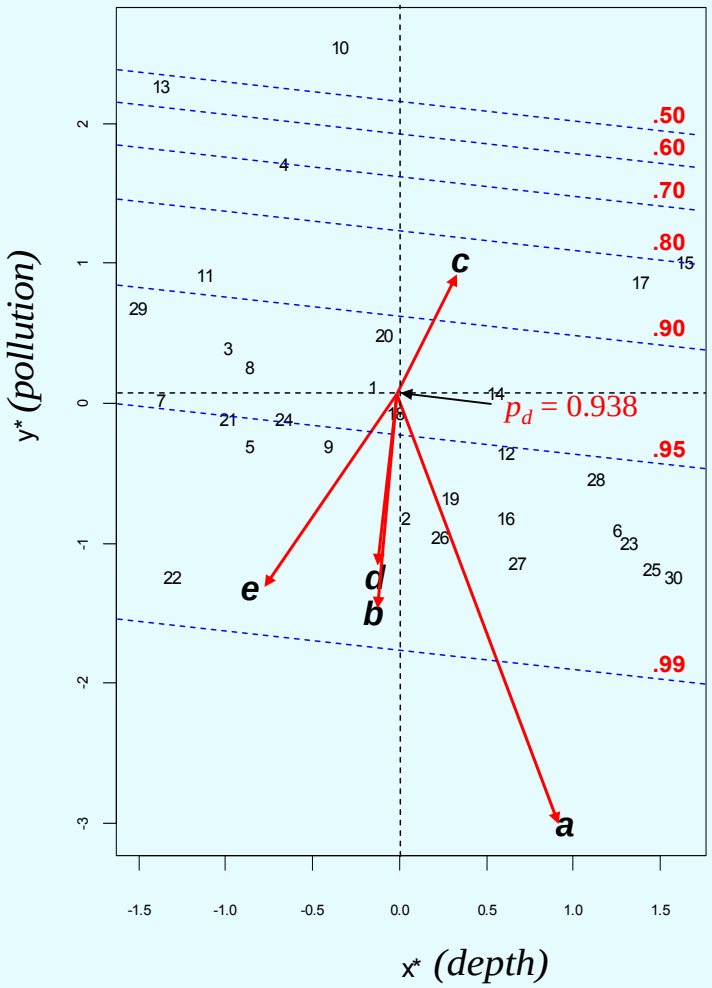
SITE NO.	SPECIES COUNTS a	b	c	d	e
s1	0	2	9	14	2
s2	26	4	13	11	0
s3	0	10	9	8	0
s4	0	0	15	3	0
s5	13	5	3	10	7
s6	31	21	13	16	5
s7	9	6	0	11	2
s8	2	0	0	0	1
s9	17	7	10	14	6
s10	0	5	26	9	0
s11	0	8	8	6	7
s12	14	11	13	15	0
s13	0	0	19	0	6
:	:	:	:	:	:

transform to presence/absence (0/1) data →

0	1	1	1	1
1	1	1	1	0
0	1	1	1	0
0	0	1	1	0
1	1	1	1	1
1	1	1	1	1
1	1	0	1	1
1	0	0	0	1
1	1	1	1	1
0	1	1	1	0
0	1	1	1	1
1	1	1	1	0
0	0	1	0	1
:	:	:	:	:

$$\text{logit}(p_d) = \log\left(\frac{p_d}{1-p_d}\right) = 2.712 - 1.177\,y^* - 0.137\,x^*$$

Logistic regression biplot



error deviance:

a: 0.464

b: 0.756

c: 0.911

d: 0.798

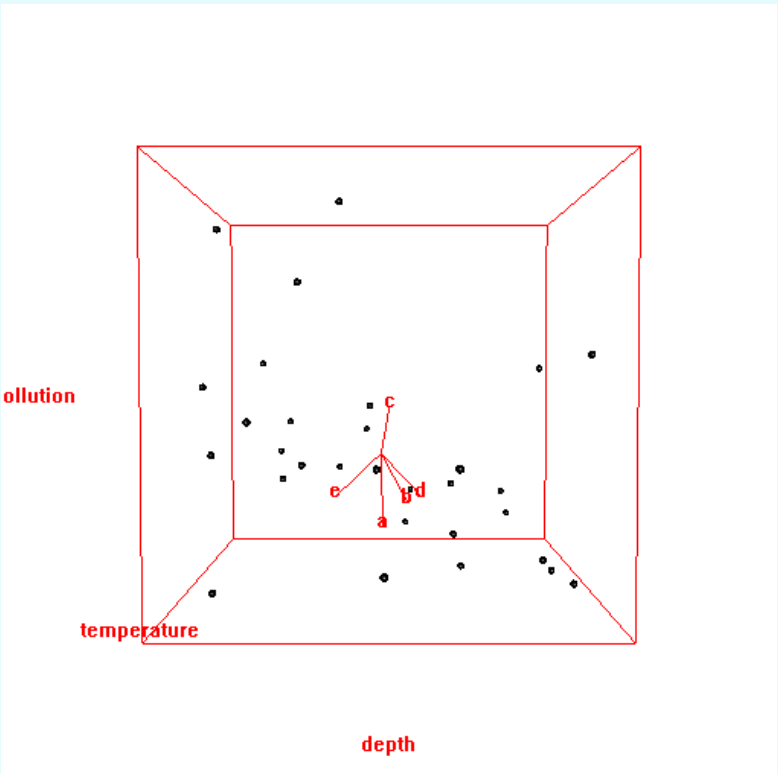
e: 0.832

average:

0.752

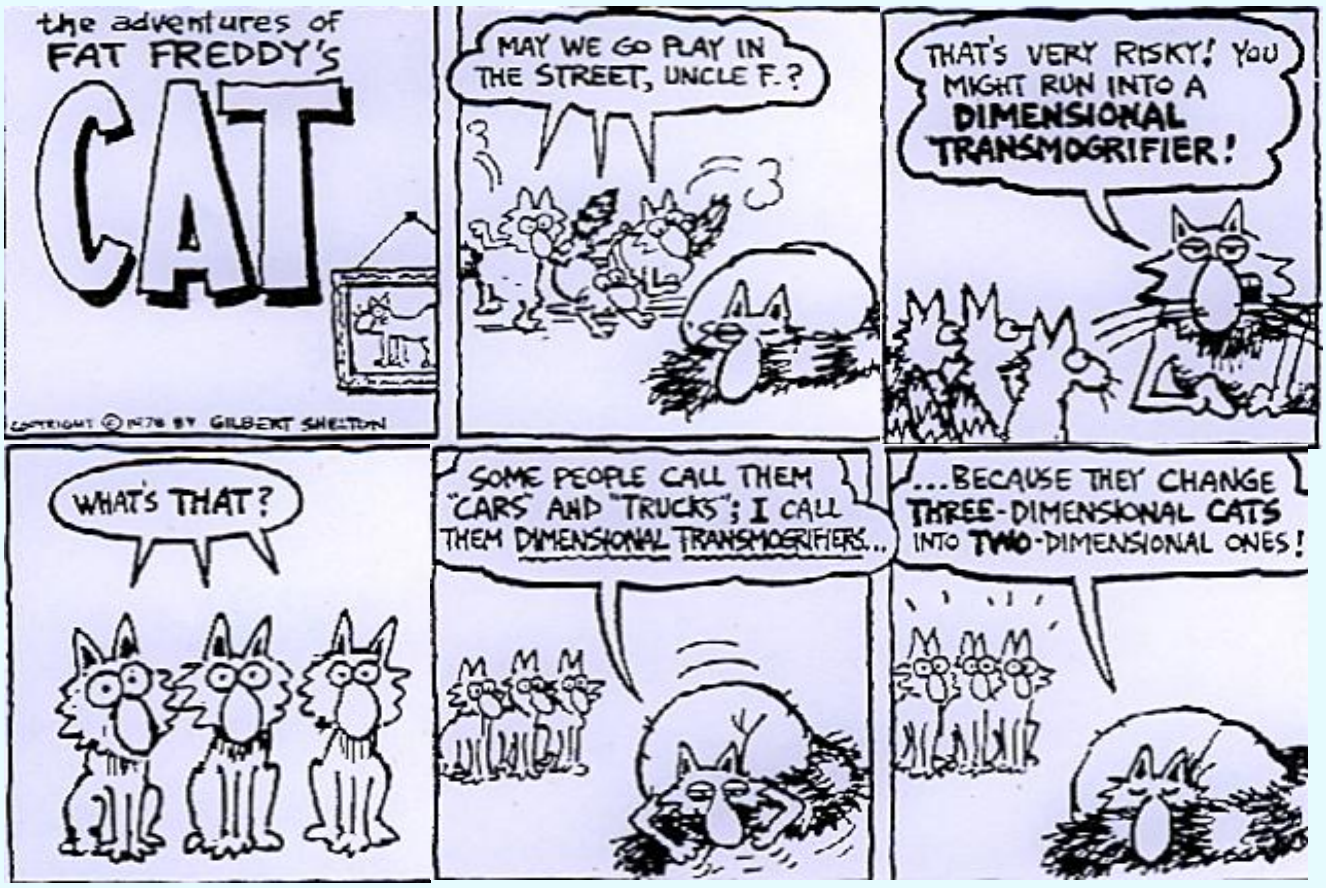
What happens for three predictors?

- Each regression model can be represented as a point vector in three-dimensional space.
- Reconstruct the data from projections of cases onto variable directions, but only as well as measured by R^2 ; in this example the increase in explained variance from two-dimensional to three-dimensional (adding temperature as an explanatory variable) is from 36.3% to 37.1%, hence temperature is explaining very little extra variance.



There will be a particular orientation of the vectors that gives maximum variance explained in the two-dimensional projection

Dimensional Transmogrifier



with thanks to Jörg Blasius

Dimensional Transmogrifier = Singular Value Decomposition (SVD)

- The SVD is a matrix decomposition of the (possibly centred and possibly normalized) $\mathbf{Y}$:

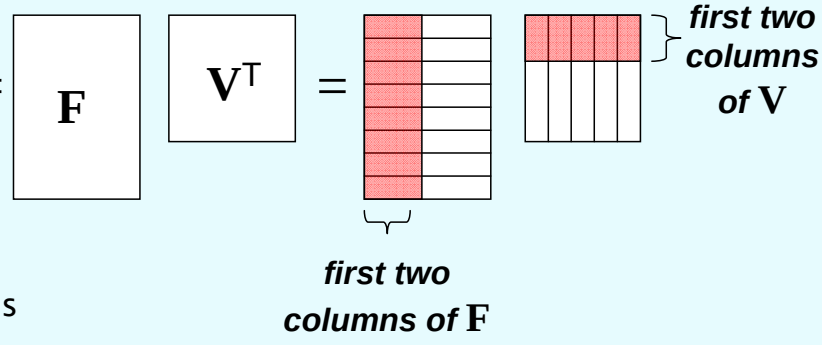
  $\mathbf{Y} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^T$ where $\mathbf{U}$ and $\mathbf{V}$ are matrices with orthonormal columns: $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$ and $\mathbf{D}_\alpha$ is a diagonal matrix of singular values $\alpha_1 \geq \alpha_2 \geq \cdots > 0$

  Standard coordinates of columns: $\mathbf{V}$

  Principal coordinates of rows: $\mathbf{F} = \mathbf{U}\mathbf{D}_\alpha$

  So:

$$\mathbf{Y} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^T = \mathbf{F}\mathbf{V}^T =$$



original 'abcde' data

SITE	a	b	c	d	e	sum
s1	0	2	9	14	2	27
s2	26	4	13	11	0	54
s3	0	10	9	8	0	27
s4	0	0	15	3	0	18
s5	13	5	3	10	7	38
s6	31	21	13	16	5	86
s7	9	6	0	11	2	28
s8	2	0	0	0	1	3
s9	17	7	10	14	6	54
s10	0	5	26	9	0	40
⋮			⋮			⋮
s29	11	0	7	8	0	26
s30	24	37	5	18	1	85
sum	404	262	252	327	89	1334

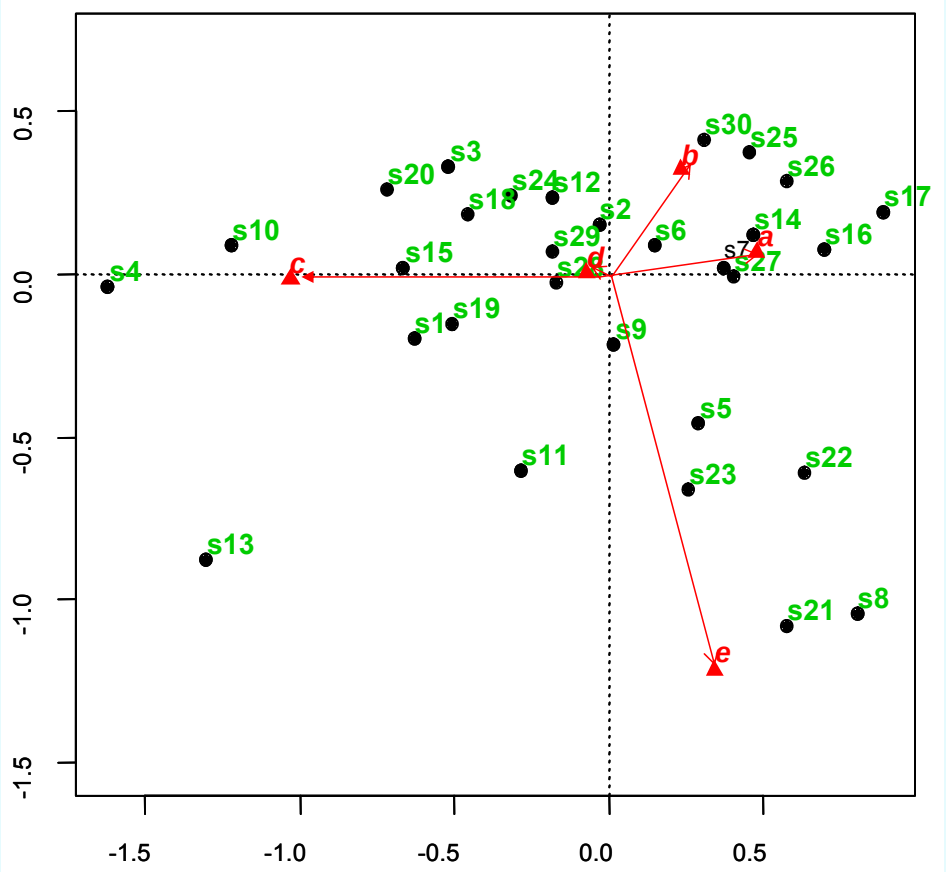
row profiles

SITE	a	b	c	d	e
s1	0.000	0.074	0.333	0.519	0.074
s2	0.481	0.074	0.241	0.204	0.000
s3	0.000	0.370	0.333	0.296	0.000
s4	0.000	0.000	0.833	0.167	0.000
s5	0.342	0.132	0.079	0.263	0.184
s6	0.360	0.244	0.151	0.186	0.058
s7	0.321	0.214	0.000	0.393	0.071
s8	0.667	0.000	0.000	0.000	0.333
s9	0.315	0.130	0.185	0.259	0.111
s10	0.000	0.125	0.650	0.225	0.000
⋮			⋮		
s29	0.423	0.000	0.269	0.308	0.000
s30	0.282	0.435	0.059	0.212	0.012
ave.	0.303	0.196	0.189	0.245	0.067

column profiles

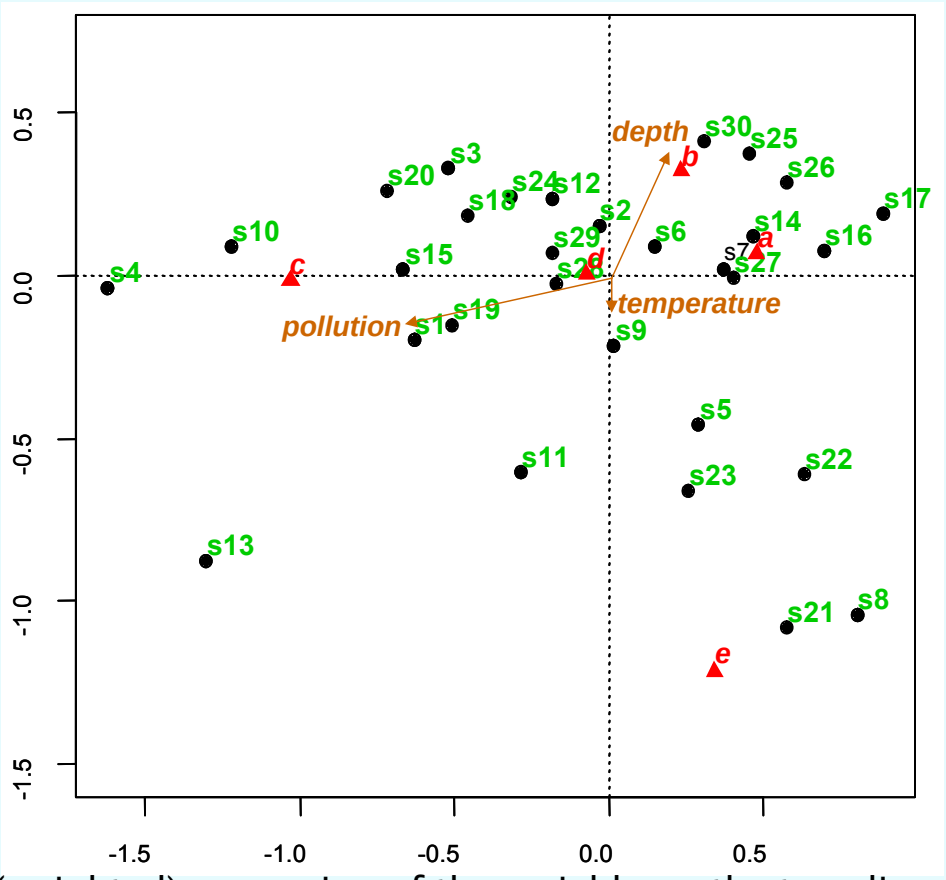
SITE	a	b	c	d	e	ave.
s1	0.000	0.008	0.036	0.043	0.022	0.020
s2	0.064	0.015	0.052	0.034	0.000	0.040
s3	0.000	0.038	0.036	0.024	0.000	0.020
s4	0.000	0.000	0.060	0.009	0.000	0.013
s5	0.032	0.019	0.012	0.031	0.079	0.028
s6	0.077	0.080	0.052	0.049	0.056	0.064
s7	0.022	0.023	0.000	0.034	0.022	0.021
s8	0.005	0.000	0.000	0.000	0.011	0.002
s9	0.042	0.027	0.040	0.043	0.067	0.040
s10	0.000	0.019	0.103	0.028	0.000	0.030
⋮			⋮			⋮
s29	0.027	0.000	0.028	0.024	0.000	0.019
s30	0.059	0.141	0.020	0.055	0.011	0.064

Each species is regressed onto the dimensions



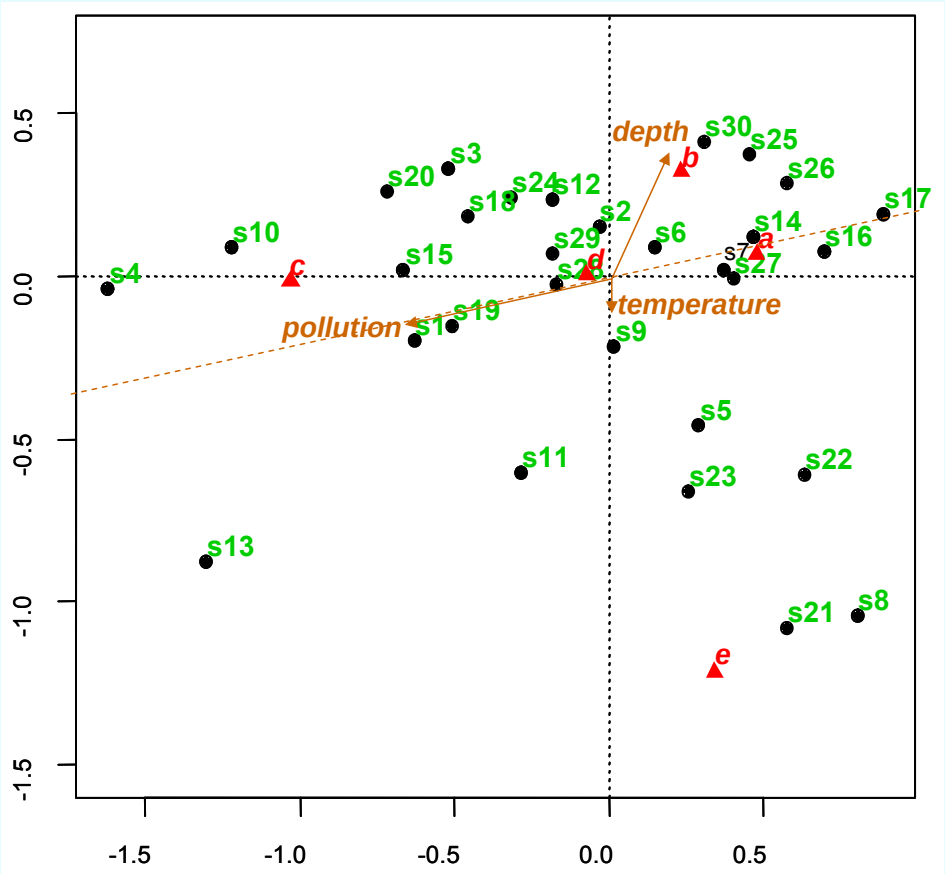
Do a (weighted) regression of the variable on the two dimensions, then use the regression coefficients as coordinates

Adding explanatory variables to the map



Do a (weighted) regression of the variable on the two dimensions, then use the regression coefficients as coordinates

Adding explanatory variables to the map



Variance explained of **supplementary variables**:

Polln: 69.5%
Depth: 30.4%
Temp: 2.1%

Explanatory variables not specifically "optimized" in the display (this is called "**indirect gradient analysis**")

Canonical correspondence analysis (CCA)

ENVIRONMENTAL VARS				SPECIES COUNTS				
Polln	Depth	Temp.	Sedmnt	a	b	c	d	e
4.8	72	3.5	S	0	2	9	14	2
2.8	75	2.5	C	26	4	13	11	0
5.4	59	2.7	C	0	10	9	8	0
8.2	64	2.9	S	0	0	15	3	0
3.9	61	3.1	C	13	5	3	10	7
2.6	94	3.5	G	31	21	13	16	5
4.6	53	2.9	S	9	6	0	11	2
5.1	61	3.3	C	2	0	0	0	1
3.9	68	3.4	C	17	7	10	14	6
10.0	69	3.0	S	0	5	26	9	0
6.5	57	3.3	C	0	8	8	6	7
3.8	84	3.1	S	14	11	13	15	0
9.4	53	3.0	S	0	0	19	0	6
4.7	83	2.5	C	13	0	0	9	0
6.7	100	2.8	C	4	0	10	12	0
2.8	84	3.0	G	42	20	0	3	6
6.4	96	3.1	C	4	0	0	0	0
4.4	74	2.8	G	21	15	33	20	0
3.1	79	3.6	S	2	5	12	16	3
5.6	73	3.0	S	0	10	14	9	0
4.3	59	3.4	C	8	0	0	4	6
1.9	54	2.8	S	35	10	0	9	17
2.4	95	2.9	G	6	7	1	17	10
4.3	64	3.0	C	18	12	20	7	0
2.0	97	3.0	G	32	26	0	23	0
2.5	78	3.4	S	32	21	0	10	2
2.1	85	3.0	G	24	17	0	25	6
3.4	92	3.3	G	16	3	12	20	2
6.0	51	3.0	S	11	0	7	8	0
1.9	99	2.9	G	24	37	5	18	1

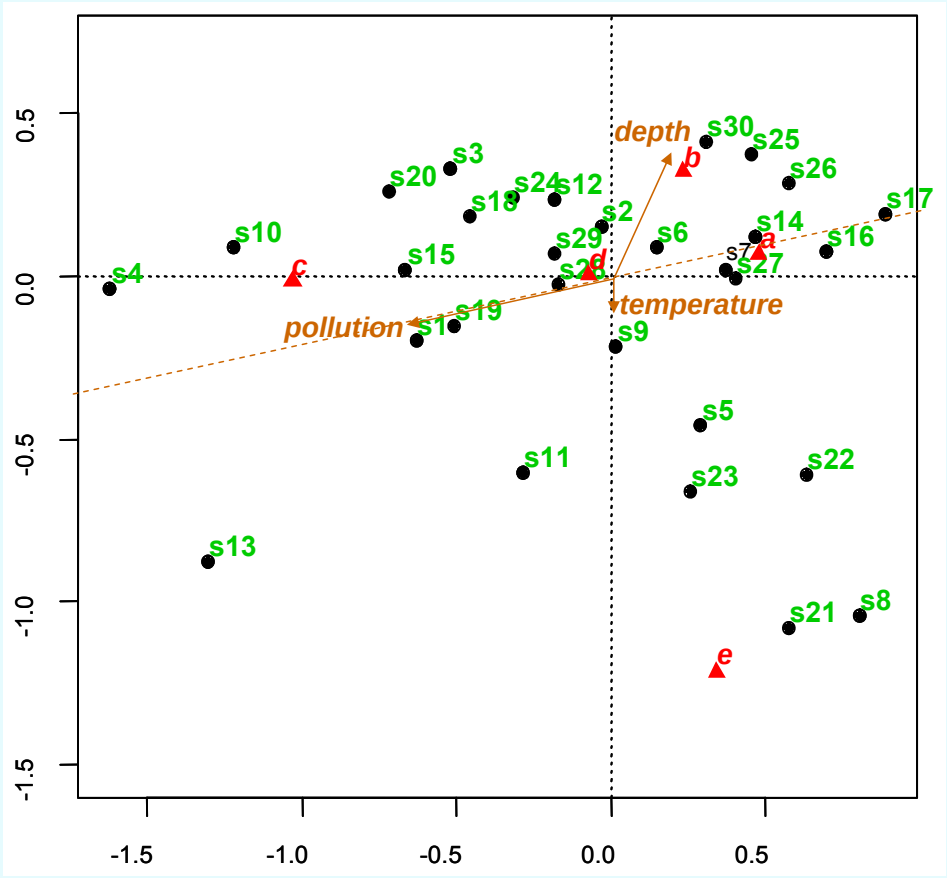
?

CA analyses the counts, finds their main dimensions and then we can relate these dimensions to the environmental variables (e.g., as supplementary points) – they might or might not be related to the environmental variables.

CCA does a more specific analysis: the dimensions are forced to be linear combinations of the environmental variables – we call this a constrained, or restricted, version of CA.

Thus the CCA axes are not the best axes of the species data (these would be found by CA). They are the best axes in the space correlated directly with the environmental predictors.

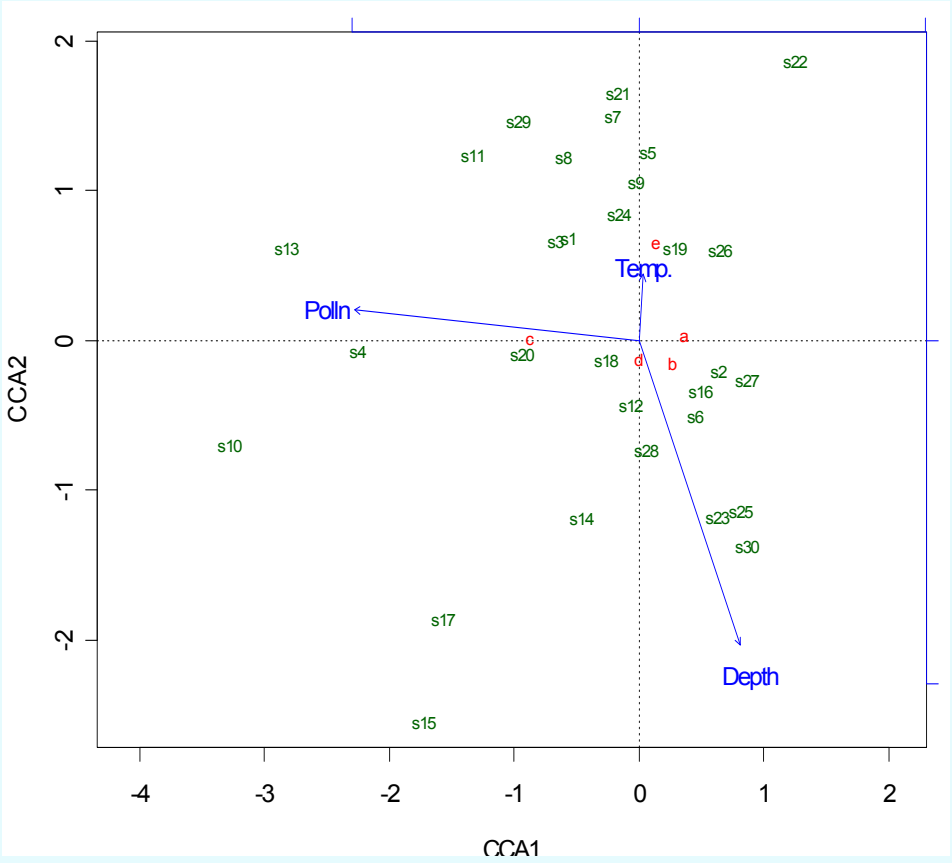
Recall: Indirect Gradient Analysis



Variance explained of **supplementary variables**:

Polln: 69.5%
Depth: 30.4%
Temp: 2.1%

CCA with 3 continuous explanatory variables




```
plot(cca(abcde,xyz), display=c("lc","bp","sp"))
```

Variance explained of constraining variables in 2-d solution:

Polln: 99.4%
Depth: 91.3%
Temp: 3.7%

Using R function cca in **vegan** package

CCA dimensionality & parts of inertia

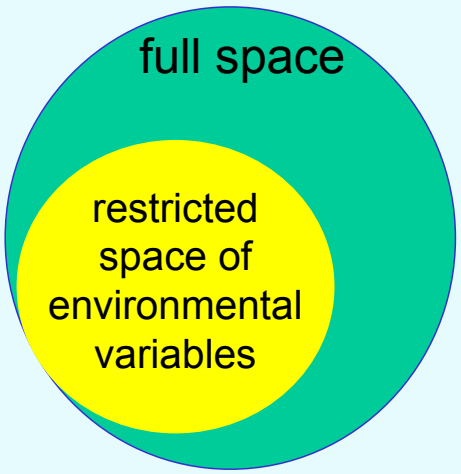
ENVIRONMENTAL VARS

Polln	Depth	Temp.	Sedmnt
4.8	72	3.5	S
2.8	75	2.5	C
5.4	59	2.7	C
8.2	64	2.9	S
3.9	61	3.1	C
2.6	94	3.5	G
4.6	53	2.9	S
5.1	61	3.3	C
3.9	68	3.4	C
10.0	69	3.0	S
6.5	57	3.3	C
3.8	84	3.1	S
9.4	53	3.0	S
4.7	83	2.5	C
6.7	100	2.8	C
2.8	84	3.0	G
6.4	96	3.1	C
4.4	74	2.8	G
3.1	79	3.6	S
5.6	73	3.0	S
4.3	59	3.4	C
1.9	54	2.8	S
2.4	95	2.9	G
4.3	64	3.0	C
2.0	97	3.0	G
2.5	78	3.4	S
2.1	85	3.0	G
3.4	92	3.3	G
6.0	51	3.0	S
1.9	99	2.9	G

?

SPECIES COUNTS

a	b	c	d	e
0	2	9	14	2
26	4	13	11	0
0	10	9	8	0
0	0	15	3	0
13	5	3	10	7
31	21	13	16	5
9	6	0	11	2
2	0	0	0	1
17	7	10	14	6
0	5	26	9	0
0	8	8	6	7
14	11	13	15	0
0	0	19	0	6
13	0	0	9	0
4	0	10	12	0
42	20	0	3	6
4	0	0	0	0
21	15	33	20	0
2	5	12	16	3
0	10	14	9	0
8	0	0	4	6
35	10	0	9	17
6	7	1	17	10
18	12	20	7	0
32	26	0	23	0
32	21	0	10	2
24	17	0	25	6
16	3	12	20	2
11	0	7	8	0
24	37	5	18	1



inertia in full space

inertia in restricted space

inertia outside restricted space

(we could also decompose this inertia NOT explained by environmental variables, along principal axes)

inertia explained by (e.g., 2-d) solution

inertia not explained by (e.g., 2-d) solution

CCA inertia decomposition for *abcde/pdt* data

Recall CA results

Explained inertia:
Axis 1: 0.288 (53.0%)
Axis 2: 0.120 (22.2%)
In 2-d: 0.408 (75.2%)

Total inertia:
0.5436

inertia in full space
0.5436 (100%)

inertia in restricted space
0.2399 (44.1%)

inertia outside restricted space
0.3038 (55.9%)

(we could also decompose this inertia NOT explained by environmental variables, along principal axes)

inertia explained by 2-d solution
0.2352 (98.0%)

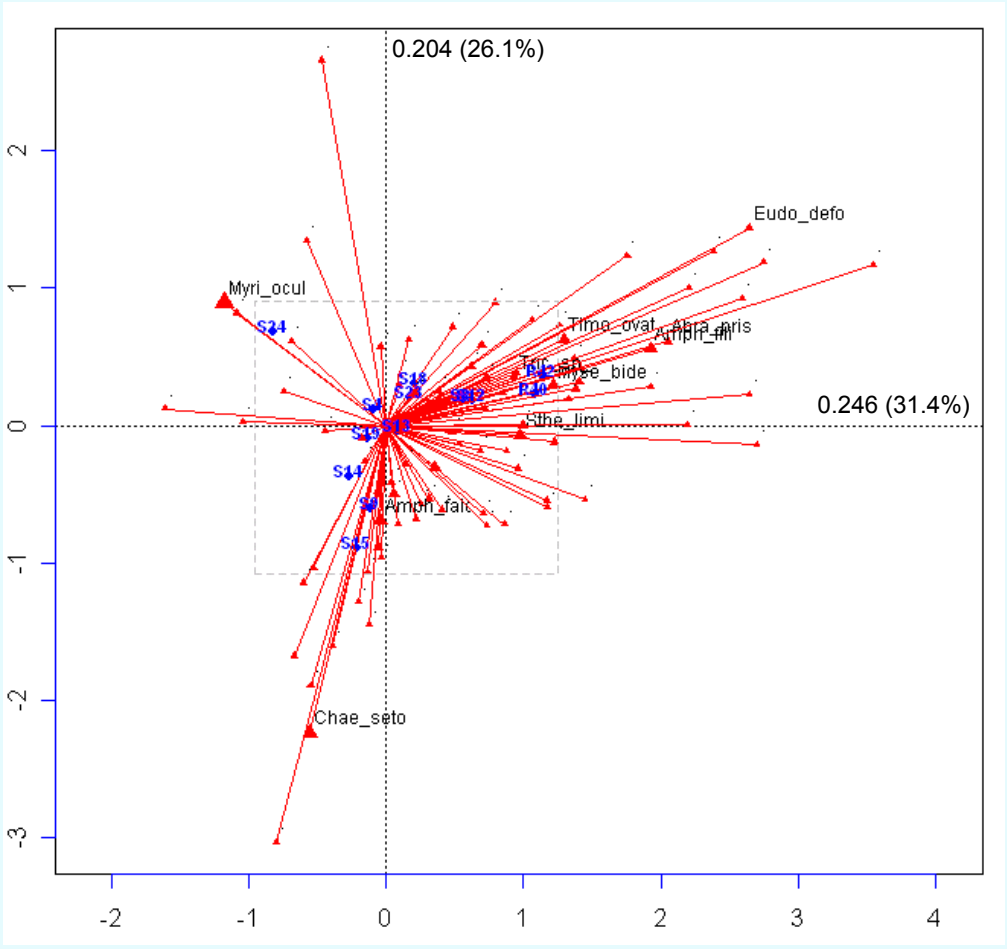
inertia not explained by 2-d solution
0.0047 (2.0%)

Data set "benthos"

A marine ecological data set with 92 species counted at 13 locations: 11 polluted locations (S..) and 2 unpolluted reference locations (R..)

	S4	S8	S9	S12	S13	S14	S15	S18	S19	S23	S24	R40	R42
Myri_ocul	193	79	150	72	141	302	114	136	267	271	992	5	12
Chae_seto	34	4	247	19	52	250	331	12	125	37	12	8	3
Amph_falc	49	58	66	47	78	92	113	38	96	76	37	0	5
Myse_bide	30	11	36	65	35	37	21	3	20	156	12	58	43
Goni_macu	35	39	41	37	32	45	41	41	31	29	64	32	23
Amph_fili	19	39	11	38	18	20	11	22	30	40	3	55	65
Timo_ovat	1	78	5	95	20	0	7	55	50	44	3	0	2
⋮	⋮												⋮
Falc_cros	1	0	2	0	0	0	0	0	0	1	0	0	0
Epit_trev	0	0	1	0	2	0	0	0	0	1	0	0	0
Ophi_flex	0	1	1	0	0	0	0	1	0	0	0	0	0
*Eucl_sp_*	0	0	0	0	1	0	0	1	1	0	0	0	0
Scal_infl	0	1	0	0	0	1	0	0	0	0	0	0	1
Eumi_ocke	0	0	1	0	0	1	1	0	0	0	0	0	0
Modi_modi	0	0	0	1	1	0	0	1	0	0	0	0	0

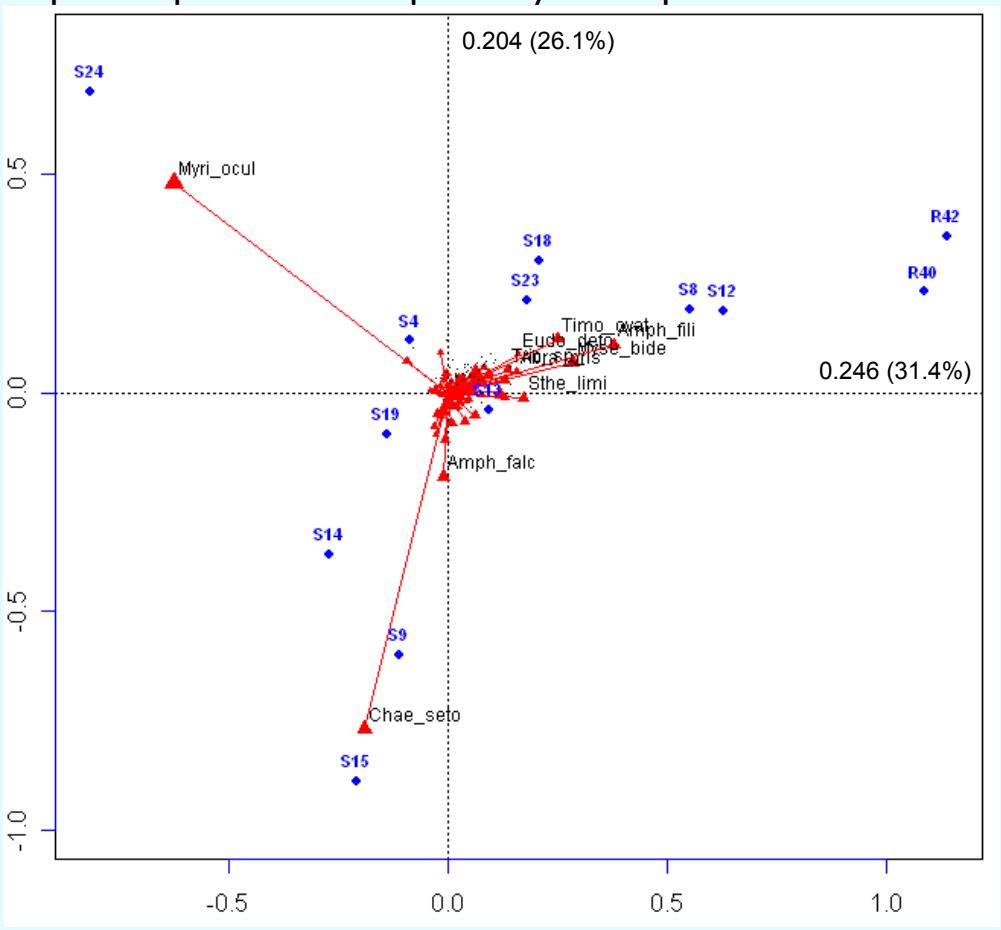
Asymmetric CA biplot of 'benthos'



explained inertia = 57.5%

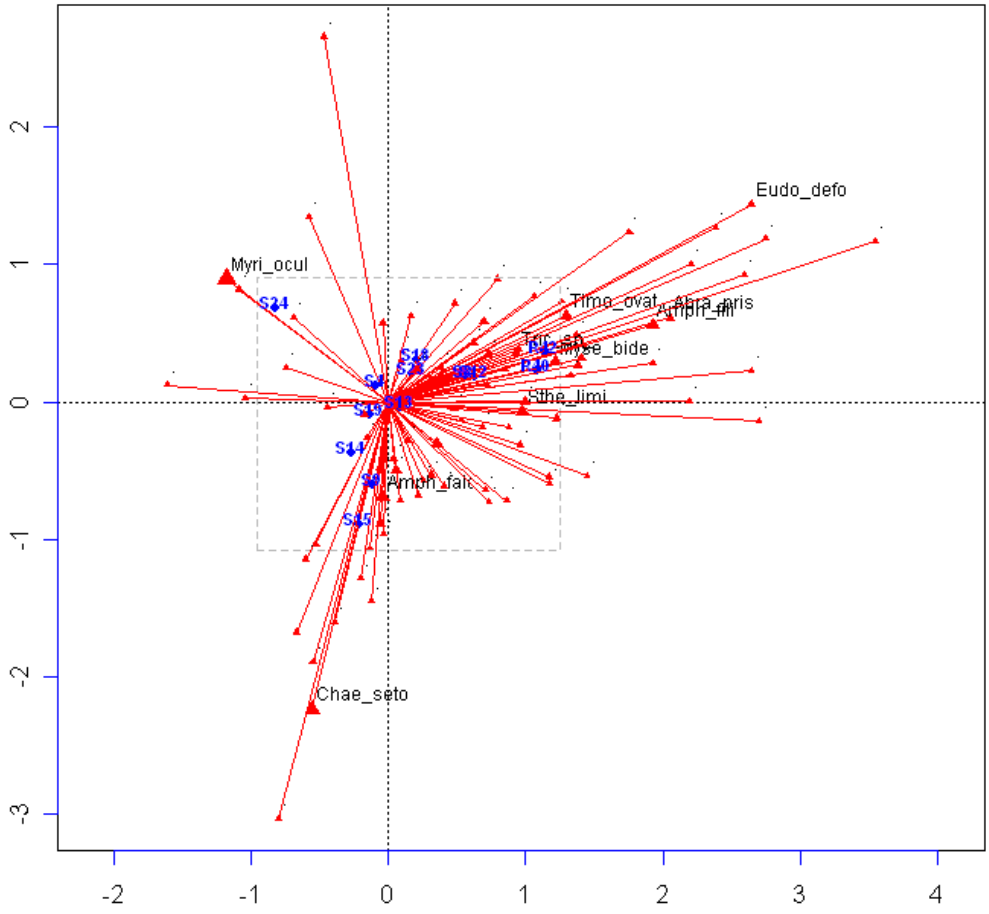
CA contribution biplot of 'benthos'

(each species point is multiplied by the square root of its mass)



explained inertia = 57.5%

Transition to contribution biplot



Data set “benthos” again

(Greenacre, *Correspondence Analysis in Practice*, 2007: chaps10&24)

13 sampling stations (last 2 references)

92 species

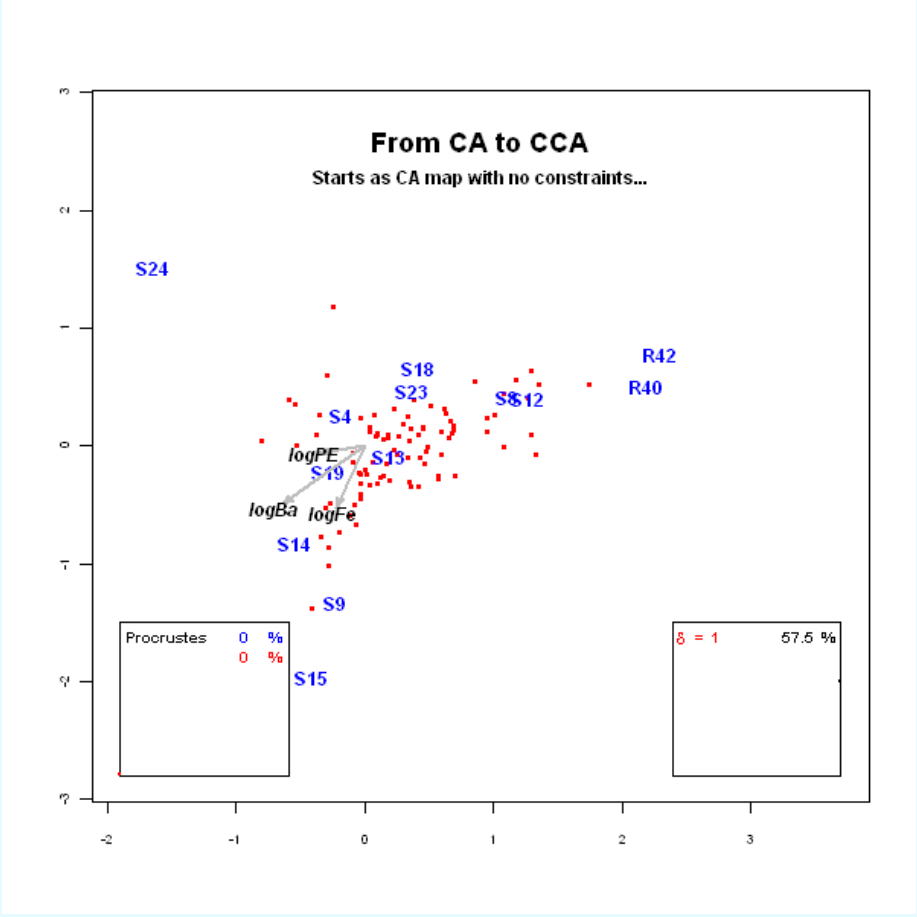
	S4	S8	S9	S12	S13	S14	S15	S18	S19	S23	S24	R40	R42
Myri_ocul	193	79	150	72	141	302	114	136	267	271	992	5	12
Chae_seto	34	4	247	19	52	250	331	12	125	37	12	8	3
Amph_falc	49	58	66	47	78	92	113	38	96	76	37	0	5
Myse_bide	30	11	36	65	35	37	21	3	20	156	12	58	43
Goni_macu	35	39	41	37	32	45	41	41	31	29	64	32	23
Amph_fili	19	39	11	38	18	20	11	22	30	40	3	55	65
Timo_ovat	1	78	5	95	20	0	7	55	50	44	3	0	2
Sthe_limi	22	33	29	26	16	19	27	21	22	22	12	32	19
Chae_niti	25	21	35	14	20	30	17	9	41	36	11	14	10
*Tric_sp_*	9	26	5	30	35	2	11	13	5	63	1	0	1
⋮	⋮												⋮
Ophi_flex	0	1	1	0	0	0	0	1	0	0	0	0	0
*Eucl_sp_*	0	0	0	0	1	0	0	1	1	0	0	0	0
Scal_infl	0	1	0	0	0	1	0	0	0	0	0	0	1
Eumi_ocke	0	0	1	0	0	1	1	0	0	0	0	0	0
Modi_modi	0	0	0	1	1	0	0	1	0	0	0	0	0

expl.vars

Ba	1656	1373	3680	2094	2813	4493	6466	1661	3580	2247	2034	40	85
Fe	2022	2398	2985	2535	2612	2515	3421	2381	3452	3457	2311	1804	1815
PELITE	2.9	14.9	3.8	5.3	4.2	9.1	5.3	4.1	7.4	3.1	6.5	2.5	2
⋮	⋮												⋮

(transformed to logarithms for CCA)

Regular CA: explanatory variables added



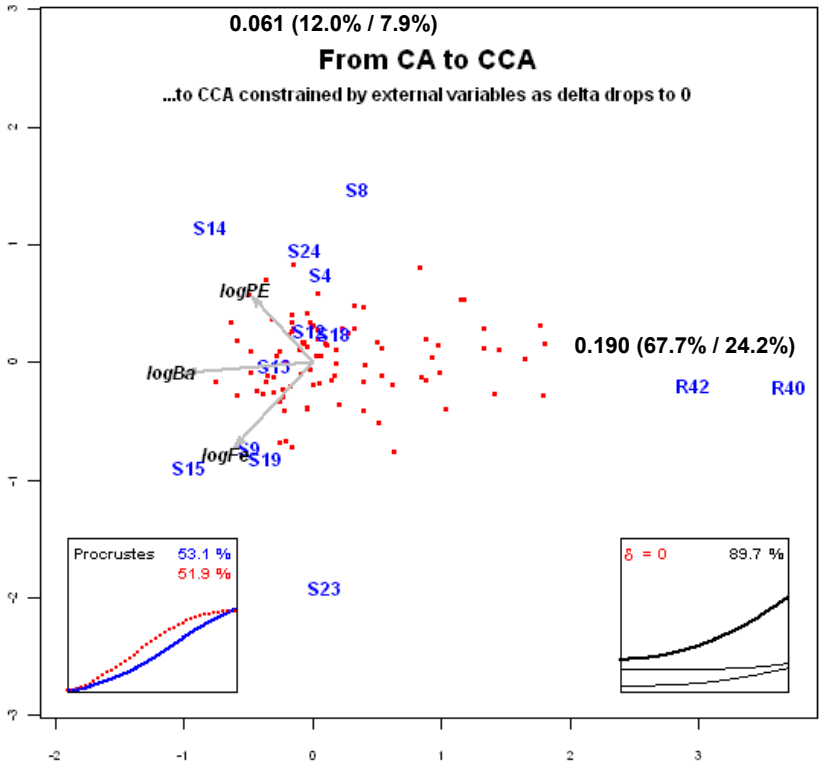
Total inertia=0.783 Dimensionality=12 57.5% in 2-d solution

Top (>1%) contributing species shown (account for 75% of the solution)

Axes 1 and 2 have regressions on *log(Ba)*, *log(Fe)* and *log(PE)* with R^2 of 0.494 and 0.319.

Three explan-atory vari-ables have R^2 with two axes of 0.648 (*Ba*), 0.326 (*Fe*), 0.126 (*PE*)

Canonical CA: dimensions restricted to be linear combinations of explanatory variables



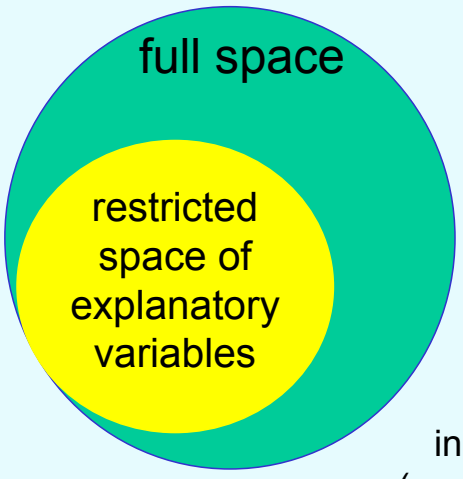
Inertia in restricted space=0.280 Dimensionality=3 89.7% in 2-d solution (with respect to original full space: 32.1%)

Top (>1%) contributing species shown (account for 83% of the solution)

Axes equal to linear combi-nations of *log(Ba)*, *log(Fe)* and *log(PE)* - R^2 of 1 and 1.

Three explanatory variables have R^2 with two axes of 0.998 (*Ba*), 0.886 (*Fe*), 0.529 (*PE*).

CCA inertia decomposition of “benthos”



inertia in full space
0.783 (100%)

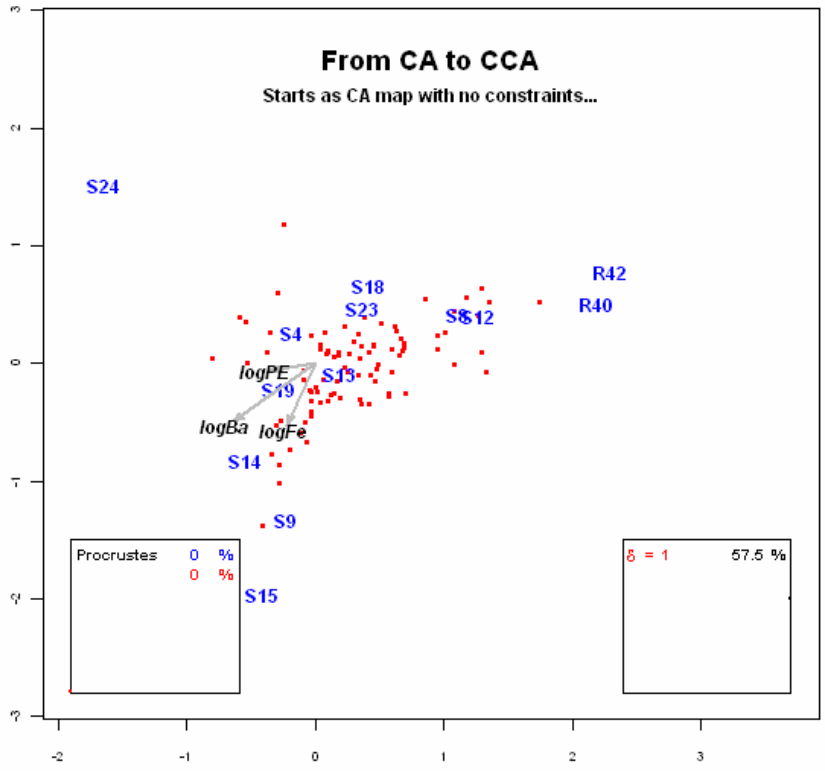
inertia in restricted space
0.280 (35.8%)

inertia outside restricted space
0.503 (64.2%)
(we could also decompose along principal axes this inertia which is ***uncorrelated*** with the explanatory variables)

inertia explained by (e.g., 2-d) solution **0.251**
(89.7% in restricted space)
(32.1% in full space)

inertia not explained by (e.g., 2-d) solution
0.029 (10.3%)

Dynamic transition from CA to CCA



CA is based on the SVD of:

$$\mathbf{S} = \mathbf{D}_r^{-½}(\mathbf{P} - \mathbf{r}\mathbf{c}^{\mathsf{T}})\mathbf{D}_c^{-½} = \mathbf{U}\Gamma\mathbf{V}^{\mathsf{T}}$$

CCA involves (weighted) projection onto the explanatory variables in $\mathbf{X}$:

$$\mathbf{Q} = \mathbf{D}_r^{½}\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{D}_r\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{D}_r^{½}$$

So in CCA the SVD is computed of:

$$\mathbf{S}^{\perp} = \mathbf{Q}\mathbf{S}$$

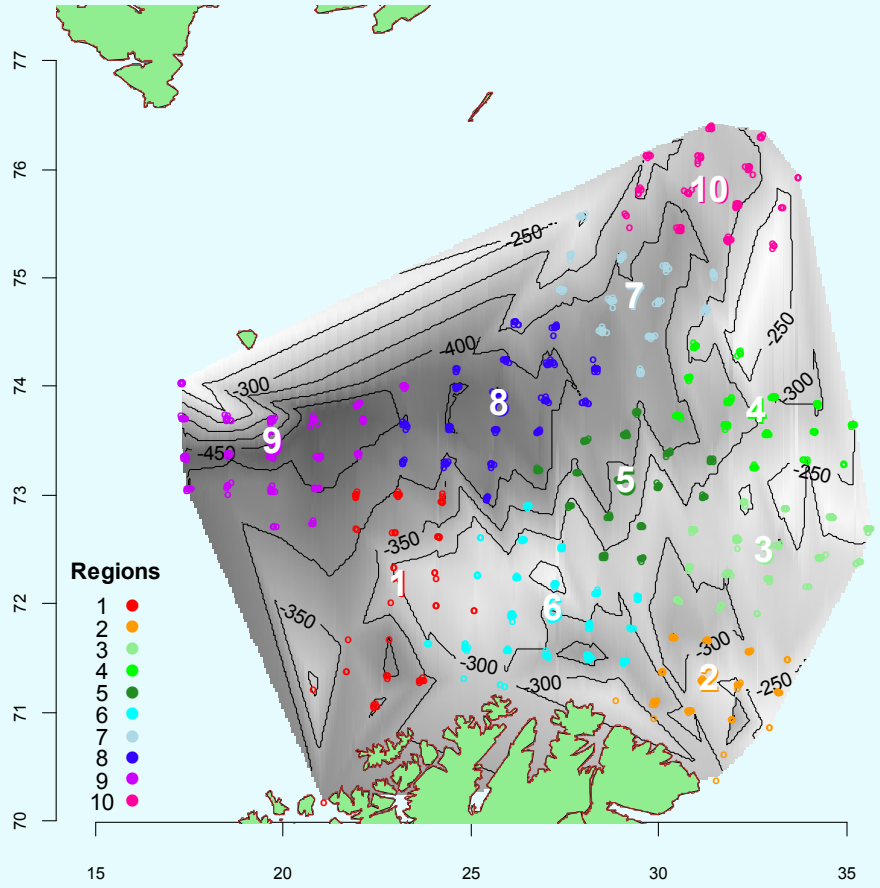
Parametrize the “projection” matrix:

$$\mathbf{Q}^{*} = \mathbf{Q}^{*}(\delta) = \delta\mathbf{I} + (1-\delta)\mathbf{Q}$$

Generate a sequence of analyses by letting δ vary from 1 to 0, each time “projecting” using $\mathbf{Q}$ (function of δ):

$$\mathbf{S}^{*} = \mathbf{Q}^{*}\mathbf{S}$$

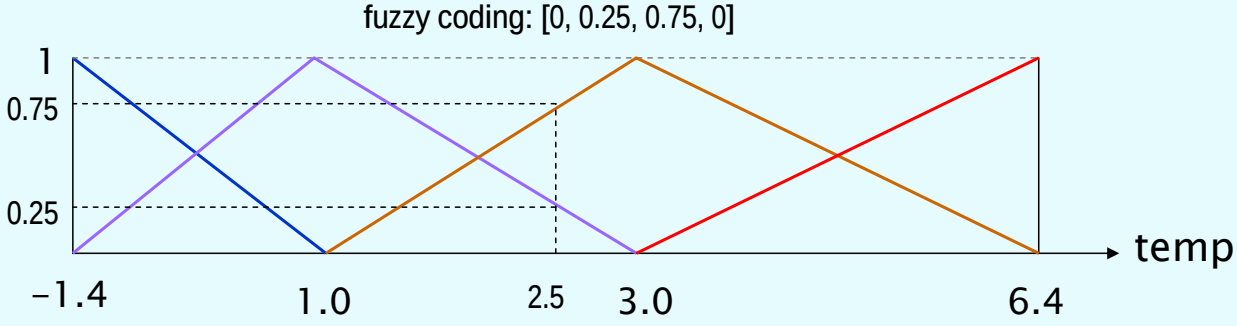
Sampling area in Barents Sea: 1992-2004



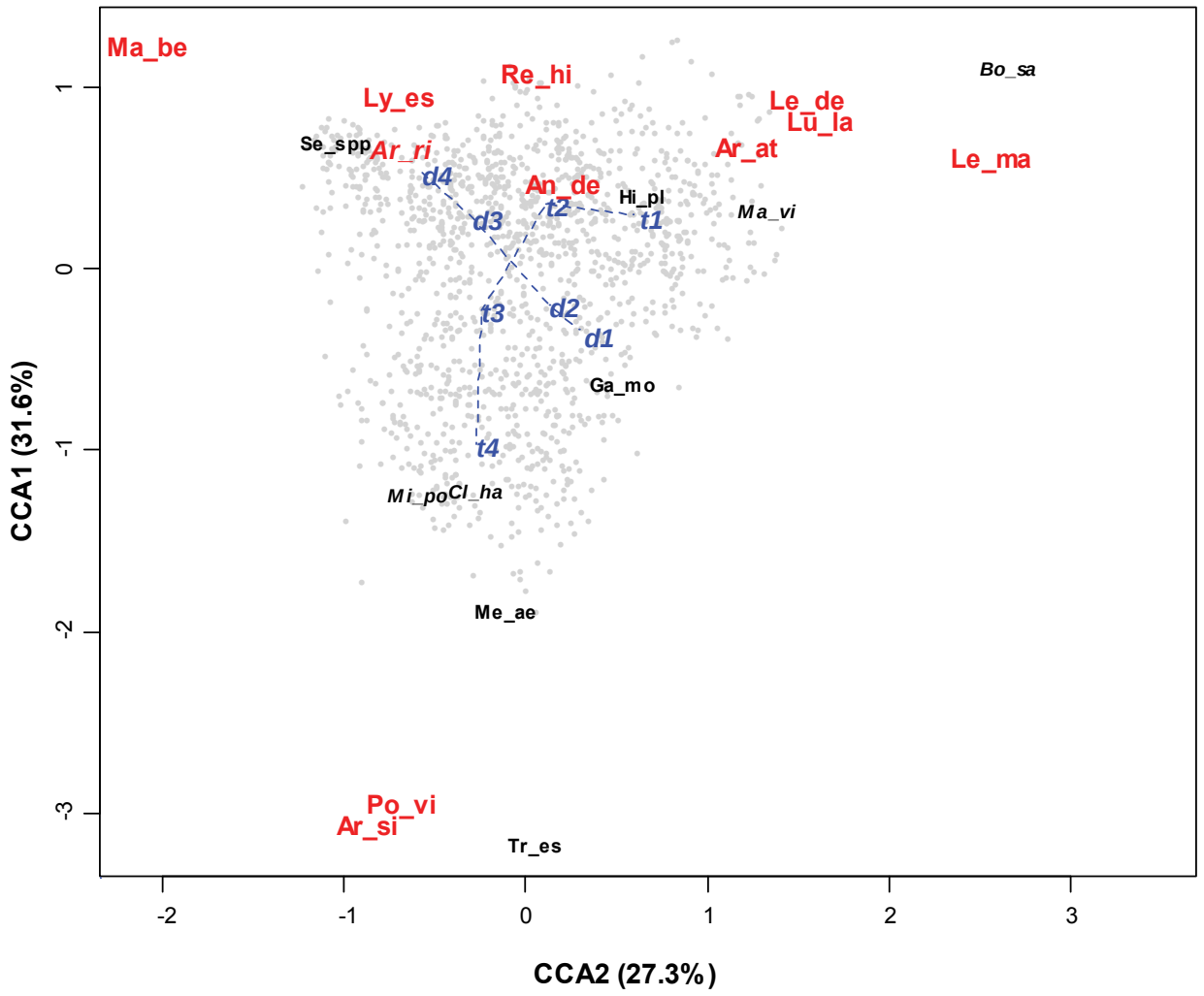
Fish community structure in Barents Sea

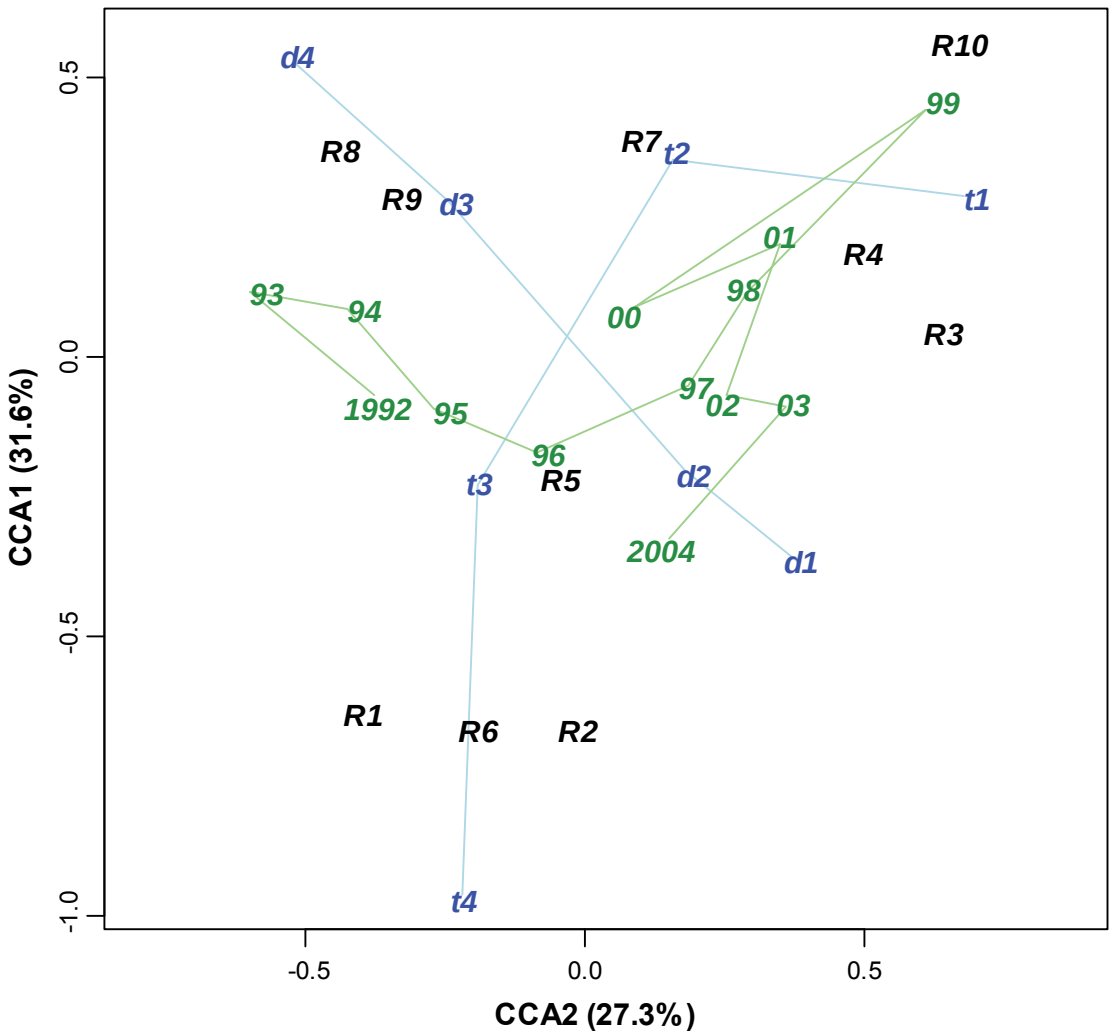
Aschan M., Fossheim M, Greenacre M, Primicerio R (2011) Change in fish community structure in Barents Sea. Under review.

- 1360 stations on a grid in the period 1992-2004
- 82 fish species
- environmental variables depth, bottom temperature, geographical region, year
- all environmental variables coded as categorical: depth, temperature and region all coded as **fuzzy variables**



Aşan Z.,Greenacre M, (2010) Biplots of fuzzy coded data. *Fuzzy Sets and Systems* .


Ma_be
Re_hi
Bo_sa
Ly_es
Le_de
Lu_la
Se_spp
Ar_ri
Ar_at
Le_ma
d4
An_de
Hi_pl
t2
Ma_vi
d3
t1
t3
d2
d1
Ga_mo
t4
Mi_po
Cl_ha
Me_ae
Po_vi
Ar_si
Tr_es
CCA1 (31.6%)
CCA2 (27.3%)
-3
-2
-1
0
1
-2
-1
0
1
2
3


d4
R10
99
R8
R7
t2
t1
R9
d3
01
R4
93
94
98
00
R3
97
1992
95
02
03
96
t3
R5
d2
2004
d1
R1
R6
R2
t4
CCA1 (31.6%)
CCA2 (27.3%)
0.5
0.0
-0.5
-1.0
-0.5
0.0
0.5

Gaceta Sanitaria, 2002

ESPECIAL

Correspondence analysis of the Spanish National Health Survey

M. Greenacre
Department of Economics and Business. Centre for Research in Health Economics (CRES).
Universitat Pompeu Fabra. Barcelona.

Correspondencia: Michael Greenacre. Universitat Pompeu Fabra. C/ Ramon Trias Fargas, 25-27. 08005 Barcelona.
Correo electrónico. michael@upf.es

Recibido: 14 de junio de 2001.
Aceptado: 19 de octubre de 2001.

(El análisis de correspondencias en la explotación de la Encuesta Nacional de Salud)

Summary

This report gives a comprehensive explanation of the multivariate technique called correspondence analysis, applied in the context of a large survey of a nation's state of health, in this case the Spanish National Health Survey. It is first shown how correspondence analysis can be used to interpret a simple cross-tabulation by visualizing the table in the form of a map of points representing the rows and columns of the table. Combinations of variables can also be interpreted by coding the data in the appropriate way. The technique can also be used to deduce optimal scale values for the levels of a categorical variable, thus giving quantitative meaning to the categories. Multiple correspondence analysis can analyze several categorical variables simultaneously, and is analogous to factor analysis of continuous variables. Other uses of correspondence analysis are illustrated using different variables of the same Spanish database: for example, exploring patterns of missing data and visualizing trends across surveys from consecutive years.

Key words: Correspondence analysis. Health survey. Principal component analysis. Statistical graphics.

Resumen

Este artículo desarrolla una amplia explicación de una técnica de análisis multivariada denominada análisis de correspondencias, aplicándola a datos de una encuesta nacional de salud, en este caso la Encuesta Nacional de Salud española (ENS). Primero se indica cómo puede utilizarse el análisis de correspondencias para interpretar una tabla de contingencia visualizándola en forma de un gráfico de puntos que representan las filas y columnas de la tabla. También pueden ser interpretadas diferentes combinaciones de las variables codificando los datos de la manera apropiada. Esta técnica puede emplearse también para obtener valores óptimos de escala para los niveles de una variable categórica, dándole de este modo un sentido cuantitativo a este tipo de variables. El análisis de correspondencias múltiple puede analizar varias variables categóricas simultáneamente, y es análogo al análisis de factores de las variables continuas. Otras aplicaciones del análisis de correspondencias se ilustran usando diferentes variables de la ENS; por ejemplo, para analizar pautas en los datos perdidos y visualizando tendencias entre encuestas de años consecutivos.

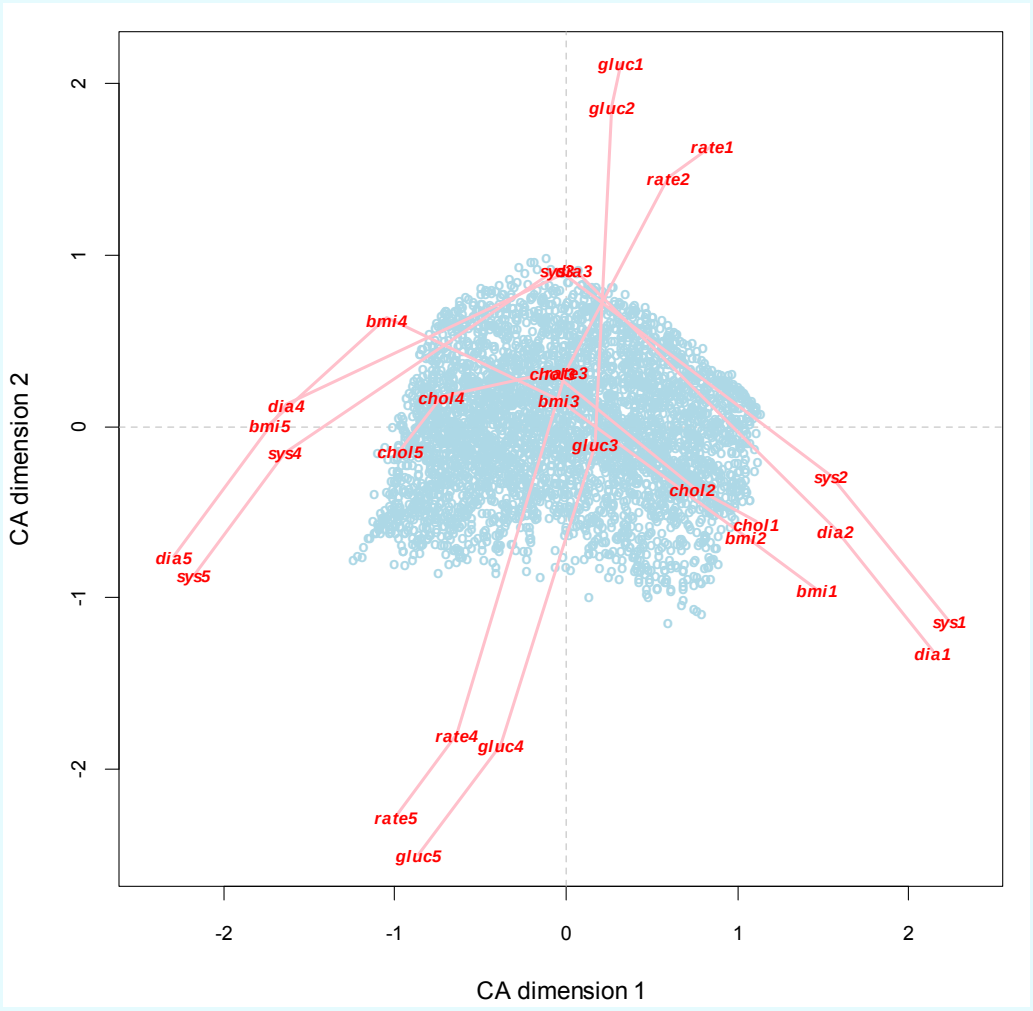
Palabras clave: Análisis de correspondencias. Encuesta de salud. Análisis de componentes principales. Gráficos estadísticos.

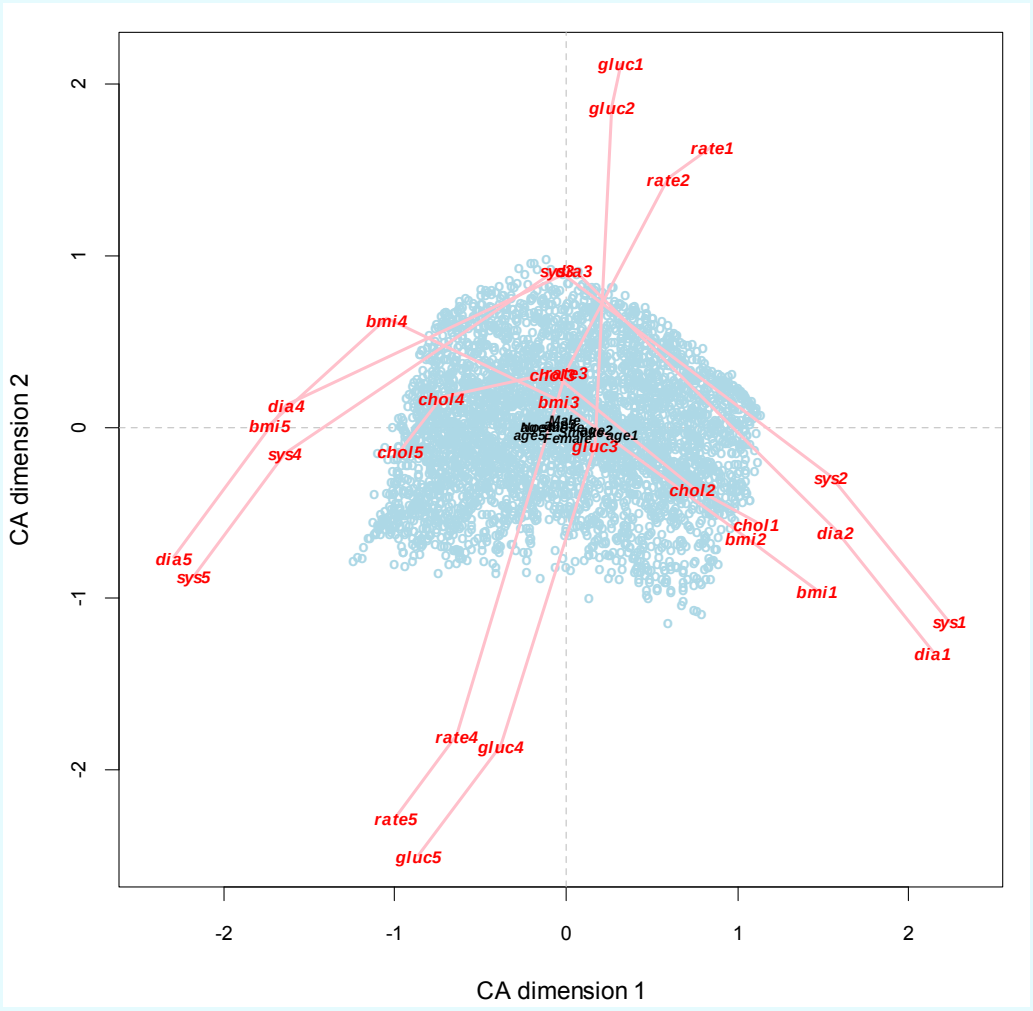
A medical example
The Framingham data set

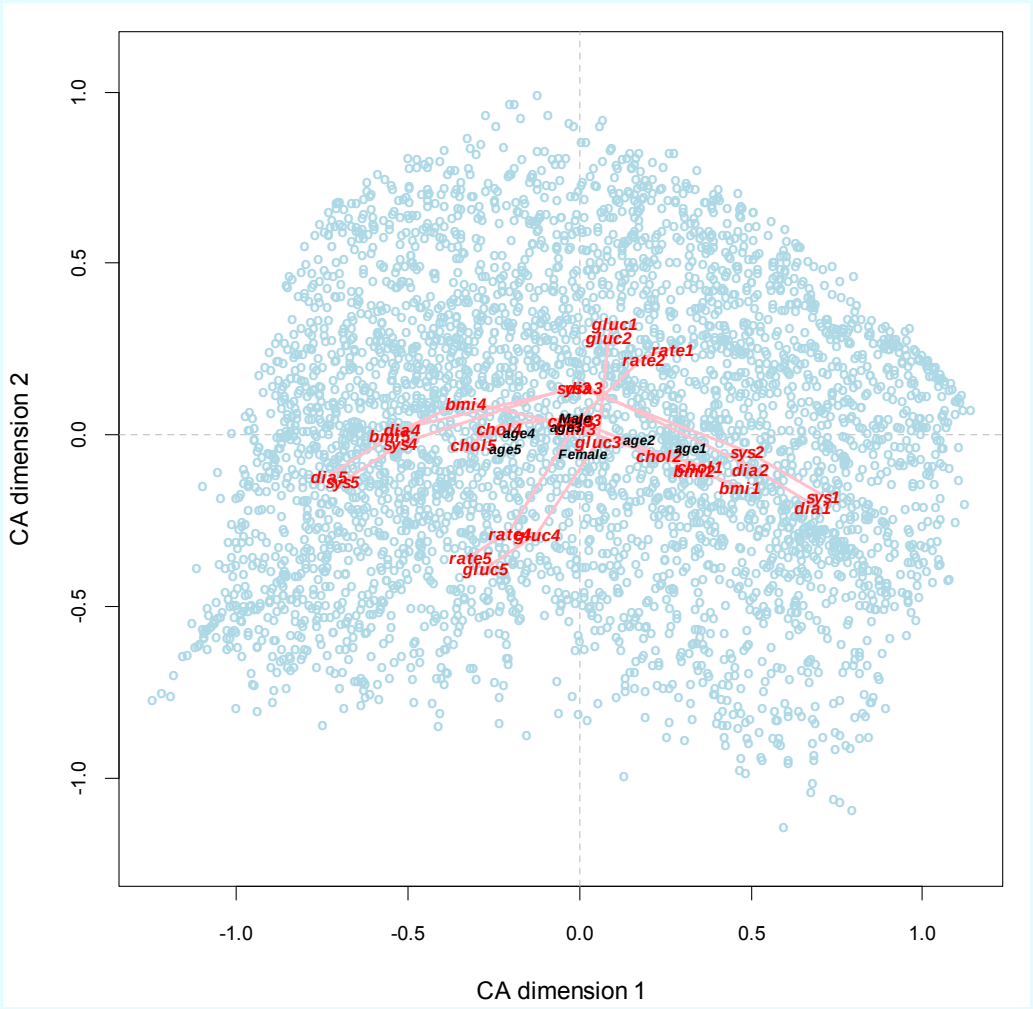
Large cohort study – we use here a subset of the Framingham data provided with the online edX course from Harvard School of Public Health: "Health in Numbers: Quantitative Methods in Clinical Research" (www.edx.org). N=4012

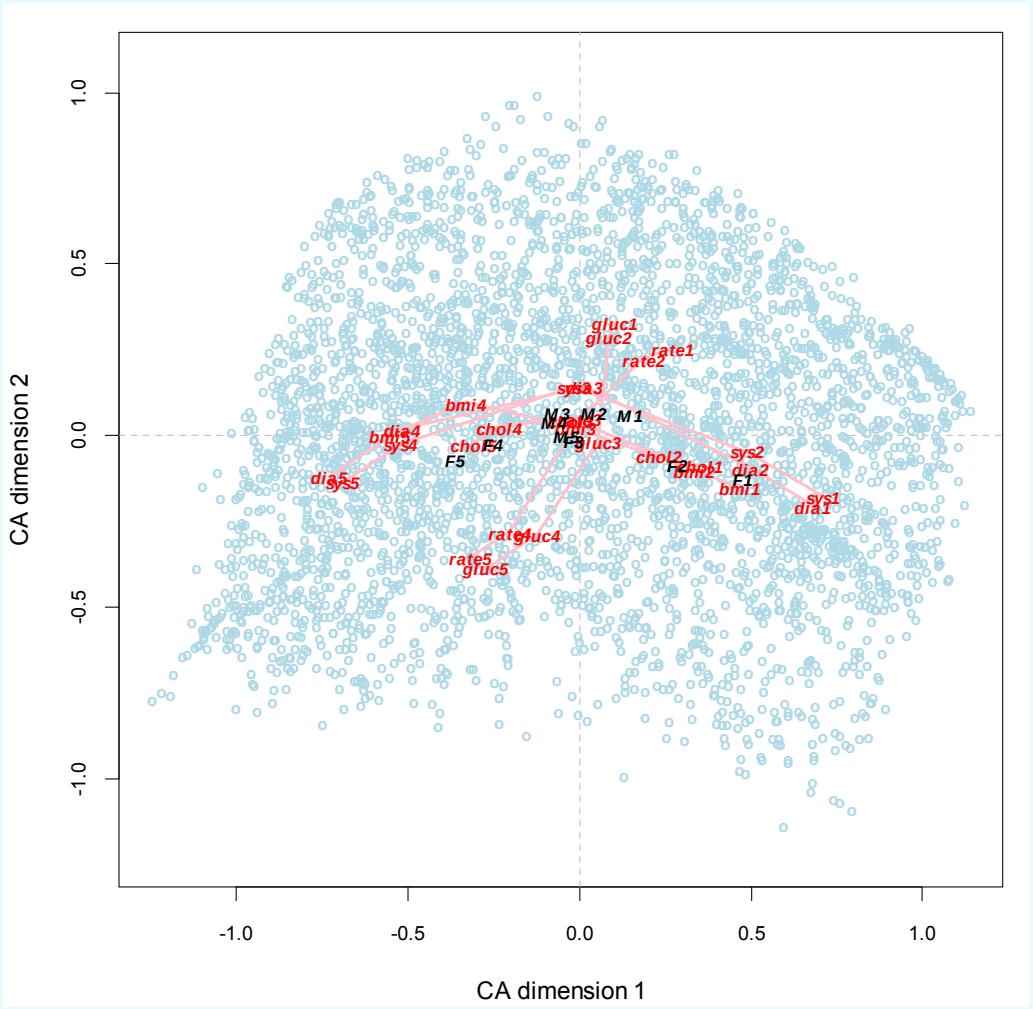
Look at interrelationahips of some of the principal quantitative measures: systolic and diastolic blood pressure, heart rate, glucose and cholesterol level,…, and the association with gender and age.

Code all quantitative variables fuzzily, 5 categories in this case.

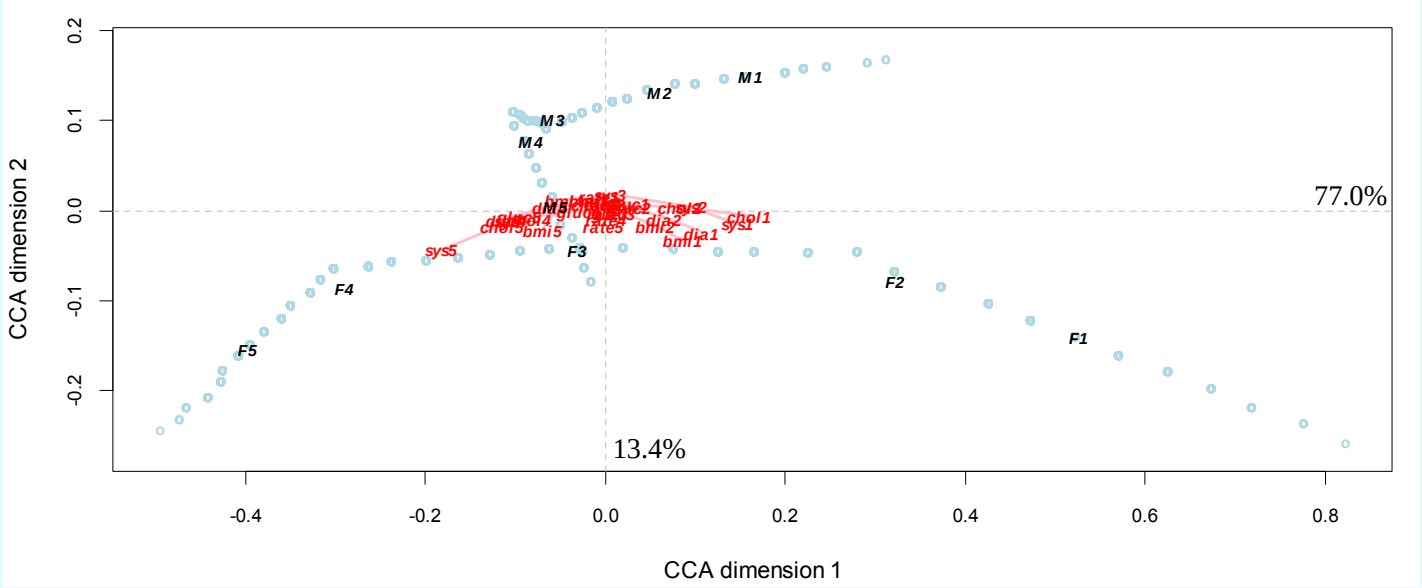

gluc1
gluc2
rate1
rate2
bmi4
dia4
bmi5
sys4
chol5
chol4
bmi3
gluc3
chol2
chol1
bmi2
sys2
dia2
bmi1
sys1
dia1
dia5
sys5
rate4
gluc4
rate5
gluc5
CA dimension 2
CA dimension 1
-2
-1
0
1
2


gluc1
gluc2
rate1
rate2
bmi4
dia4
bmi5
sys4
chol5
chol4
bmi3
gluc3
chol2
chol1
bmi2
sys2
dia2
bmi1
sys1
dia1
dia5
sys5
rate4
gluc4
rate5
gluc5
CA dimension 2
CA dimension 1
-2
-1
0
1
2


CA dimension 2
CA dimension 1
1.0
0.5
0.0
-0.5
-1.0
gluc1
gluc2
rate1
rate2
sys3
bmi4
chol4
dia4
bmi5
sys4
chol5
age4
age5
Male
Female
gluc3
age2
age1
chol2
sys2
chol1
bmi2
dia2
bmi1
sys1
dia1
dia5
sys5
rate4
gluc4
rate5
gluc5


CA dimension 2
CA dimension 1
1.0
0.5
0.0
-0.5
-1.0
gluc1
gluc2
rate1
rate2
sys3
bmi4
chol4
dia4
bmi5
sys4
chol5
F4
F5
M3
M2
M1
M4
M5
F3
gluc3
chol2
F2
chol1
bmi2
sys2
dia2
F1
bmi1
sys1
dia1
dia5
sys5
rate4
gluc4
rate5
gluc5

CCA map of constrained part of the data

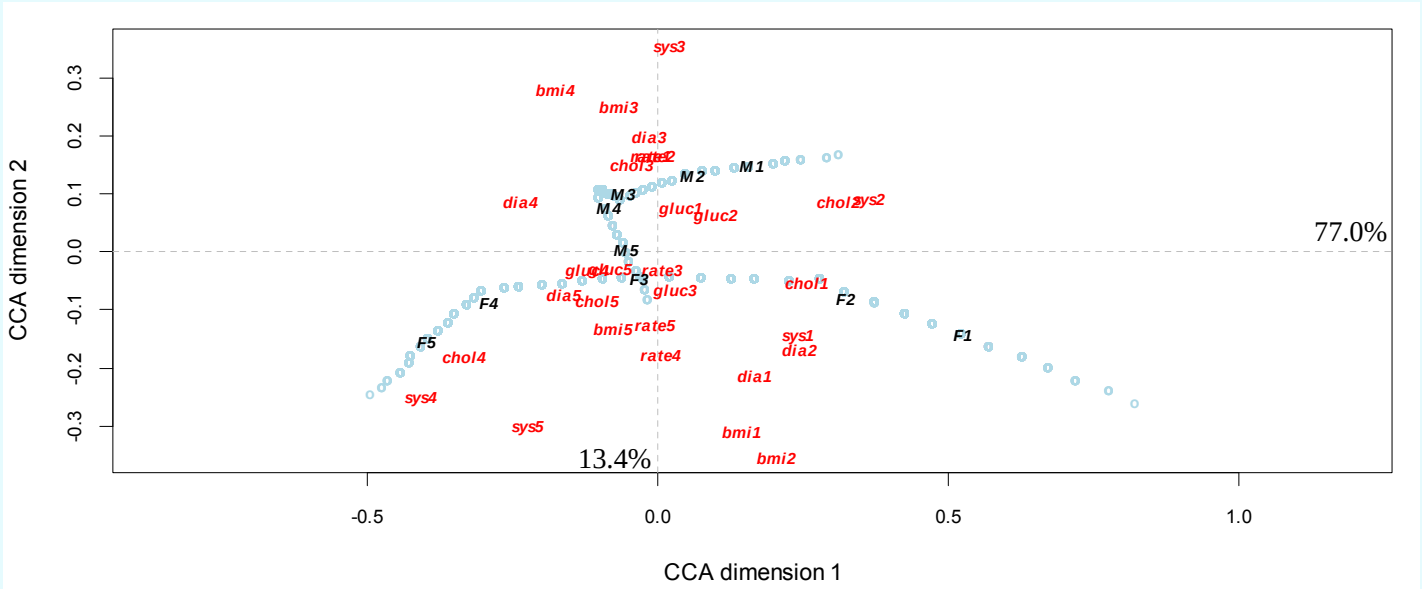


Total variance (inertia) = 1.884

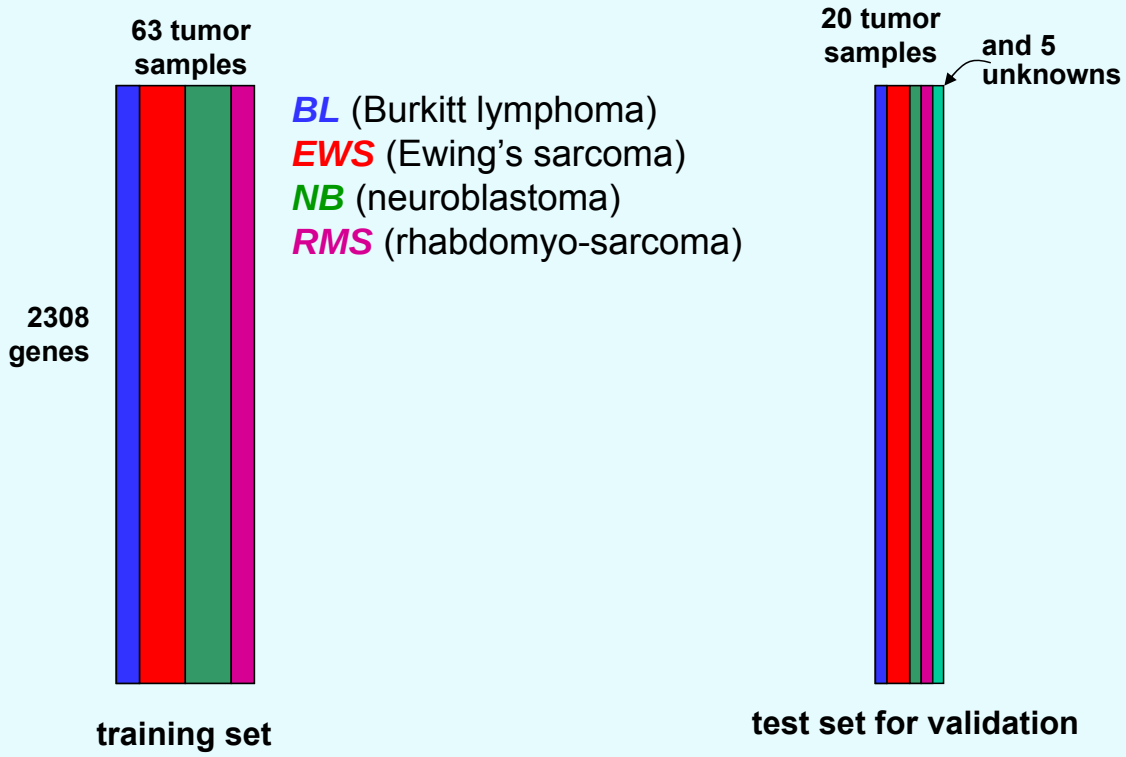
Part that is explained by sex-age interaction = 0.080 (4.3%)

Most of this 4.3% explained in above CCA map

CCA map of constrained part of the data Contribution Biplot

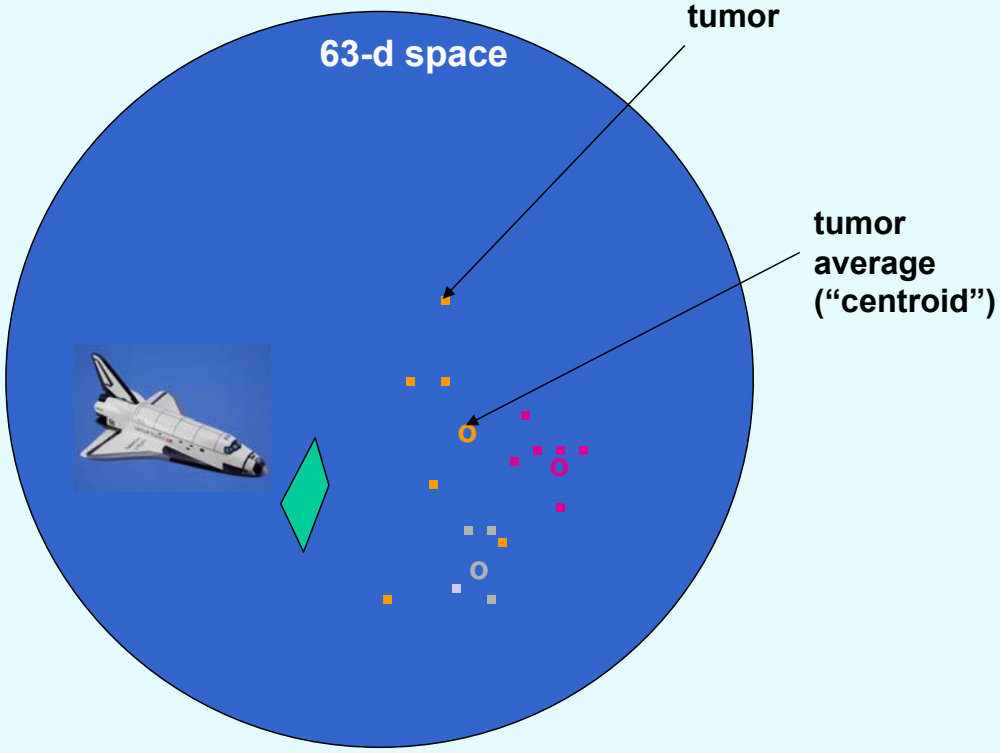


An example from Hastie et al. (2009)
The Elements of Statistical Learning. 2nd Edition.

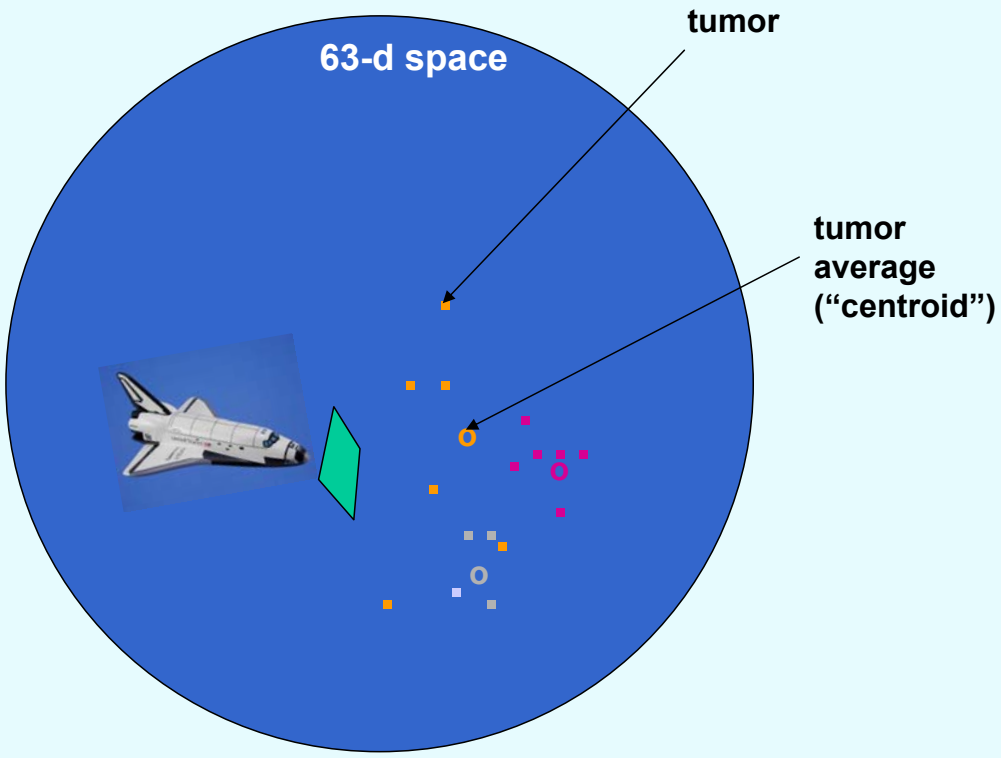


Objective: to identify a small subset of genes that is determinant in classifying the type of tumour

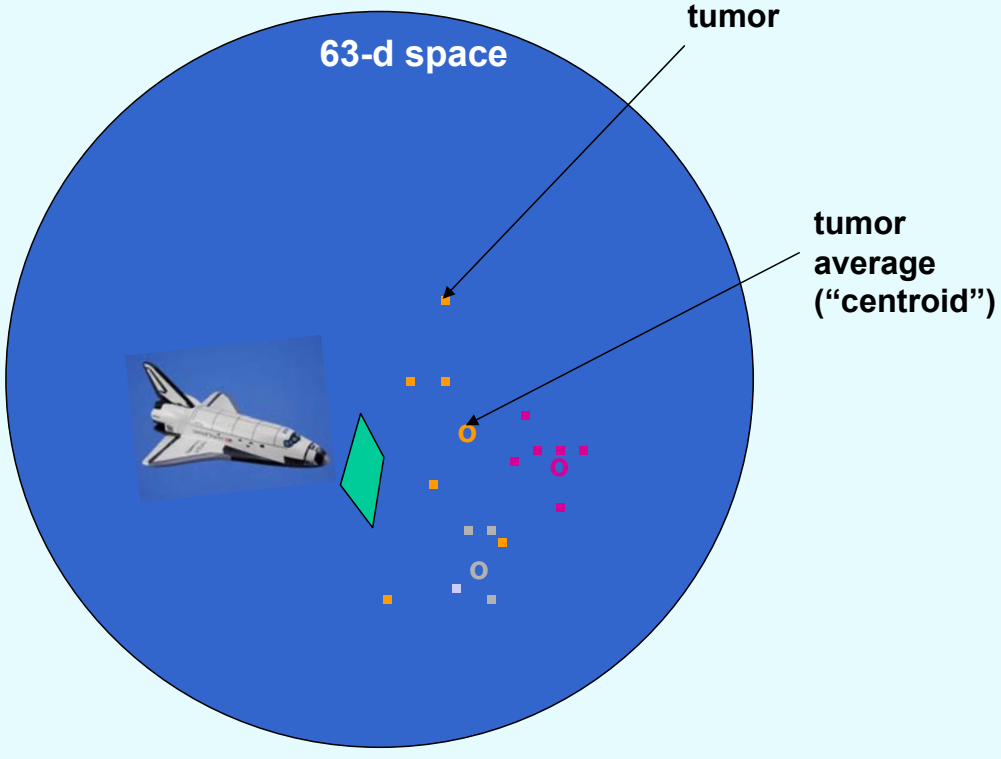
Guided tour through n-dimensional space



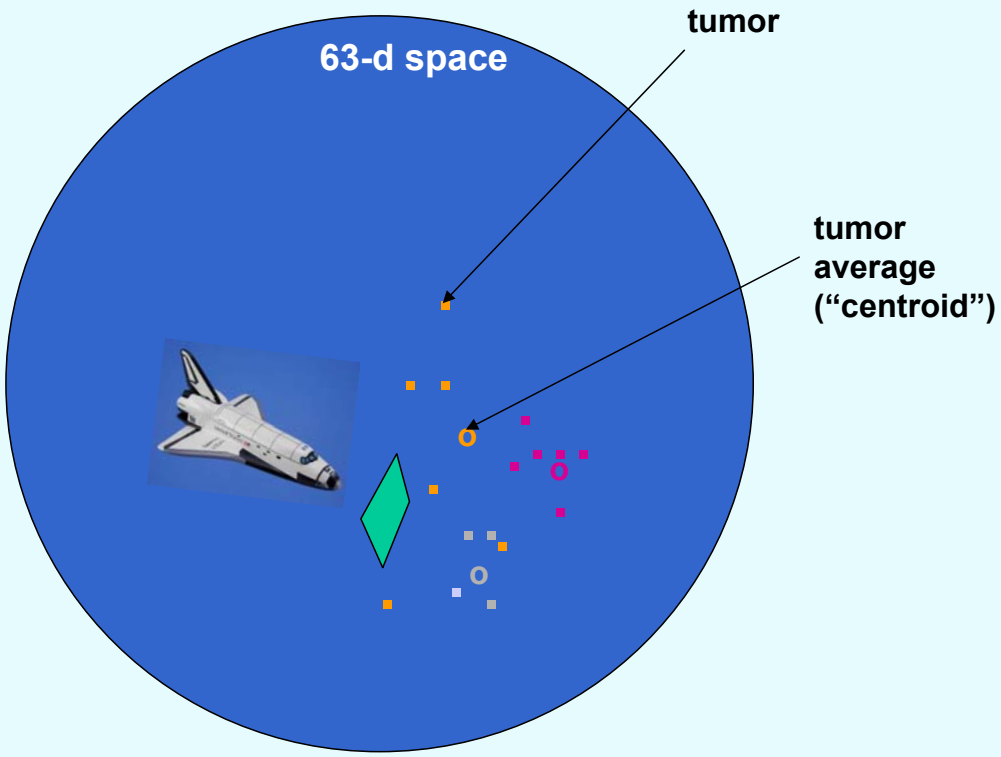
Guided tour through n-dimensional space



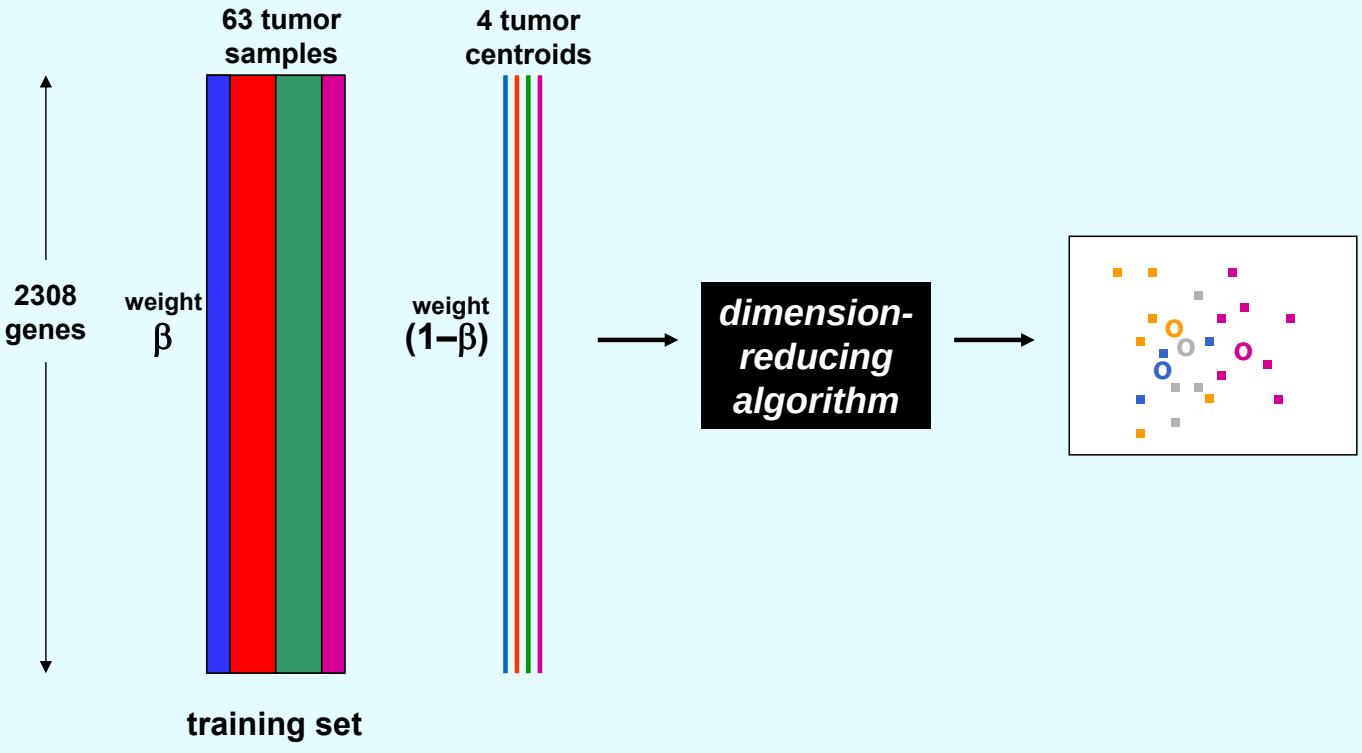
Guided tour through n-dimensional space



Guided tour through n-dimensional space



How do we do the tour?

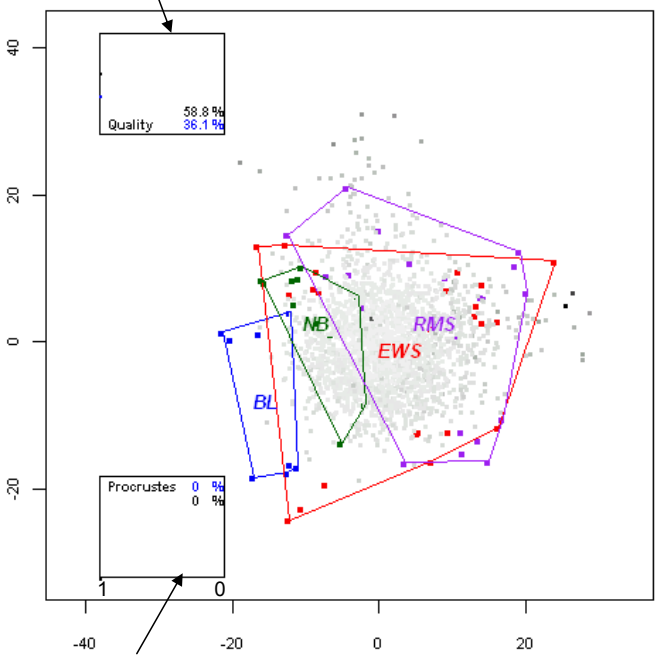


Let weight β vary smoothly from 1 to 0 (typically 1, 0.99, 0.98,…, 0.01, 0) – each time we get a different view as we move into the multidimensional space. β is a tuning parameter.

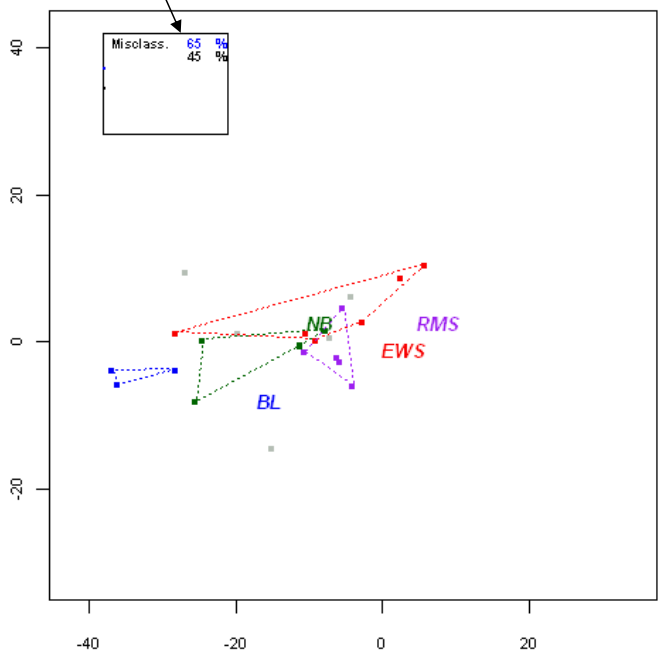
Shows how well the training set centroids are being depicted (blue:2-d,black:3-d)

Shows the misclassification rate for the test data, i.e. the error of our model (blue:2-d,black:3-d)

Training data



Test data

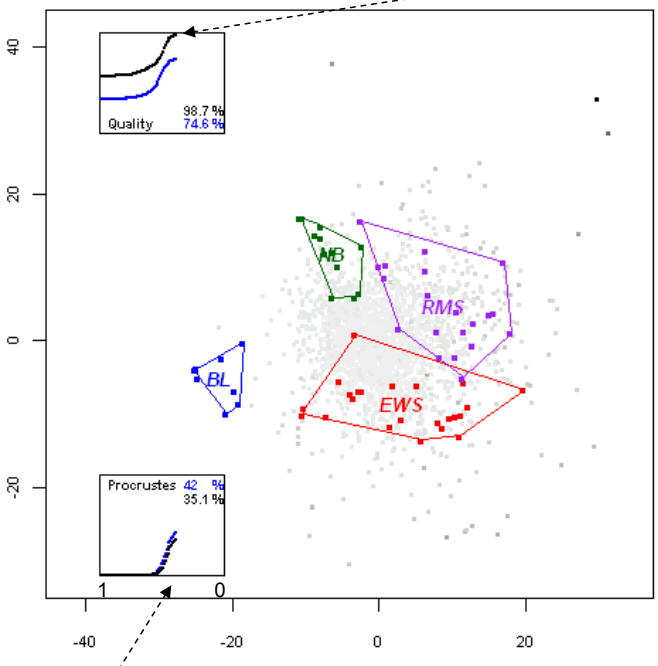


Shows how far we have moved into "deep space" (blue:2-d,black:3-d)

Voyage through 63-dimensional space, looking for best predictor of tumors in test data

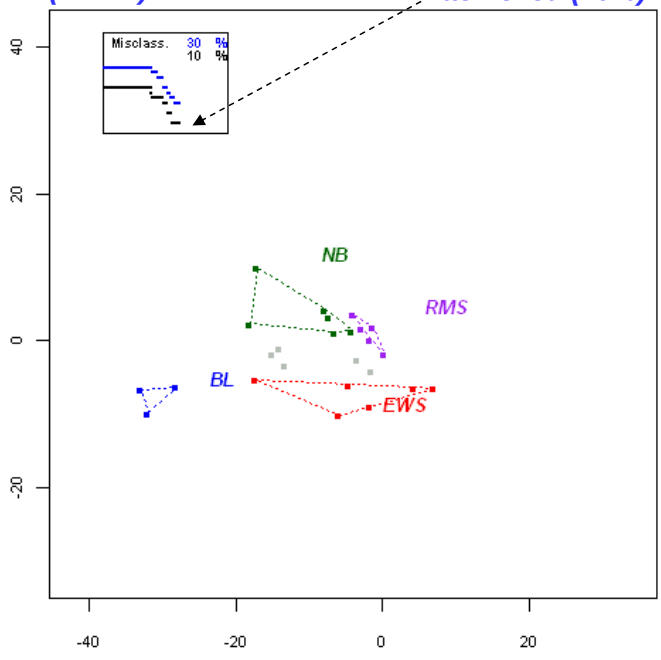
Training data

Quality of representation of tumor centroids almost perfect (98.7%)



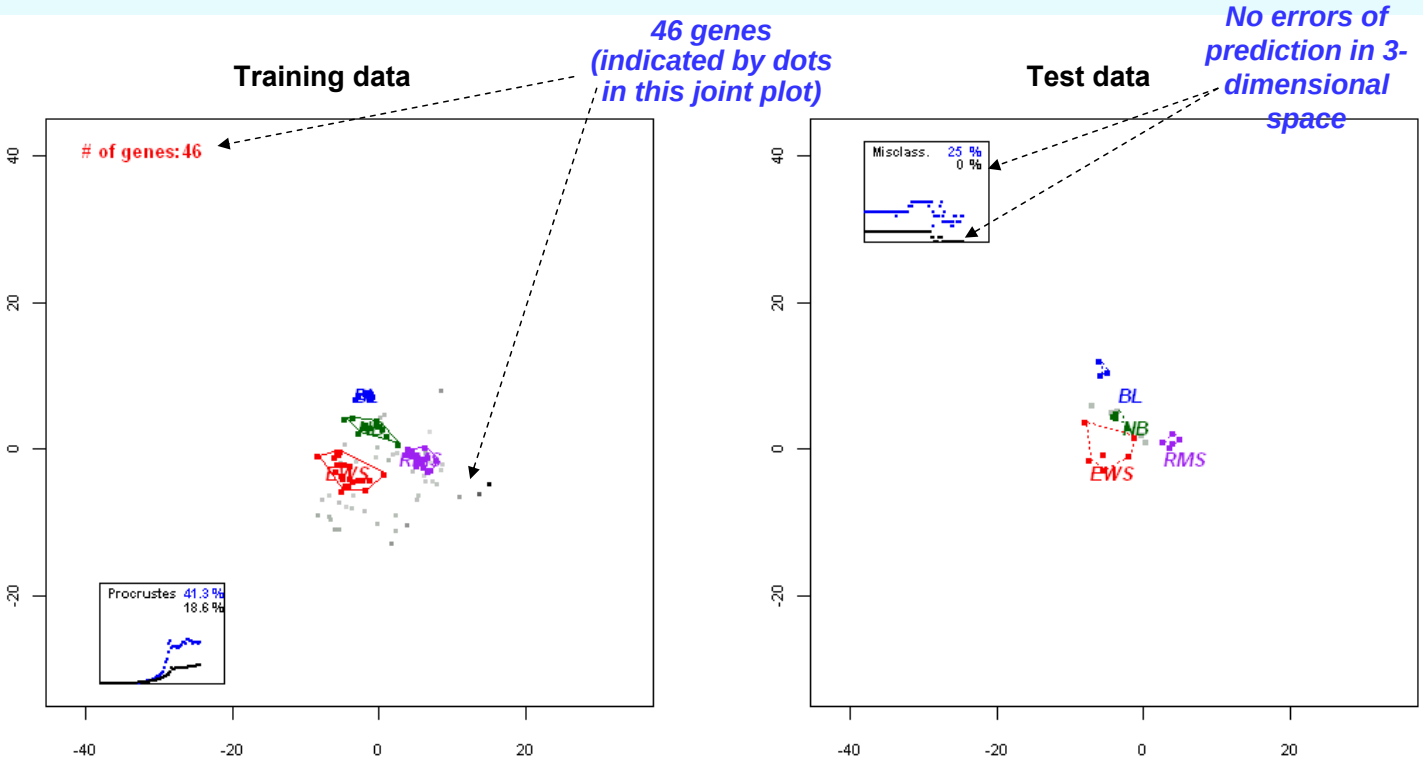
Test data

Lowest misclassification achieved (10%)



This frame is at approximate **0.4 : 0.6** *distribution of weight between tumors and tumor centroids* **(β = 0.6).**

Peeling away genes with least predictive power, seeing effect on prediction of test data.



Rotation of 3-dimensional solution space where prediction takes place; test samples are classified to the closest centroid of the training set.