Biplots

The biplot extends the idea of a simple scatterplot of two variables to the case of many variables, with the objective of visualizing a maximum amount of information in the data as possible.

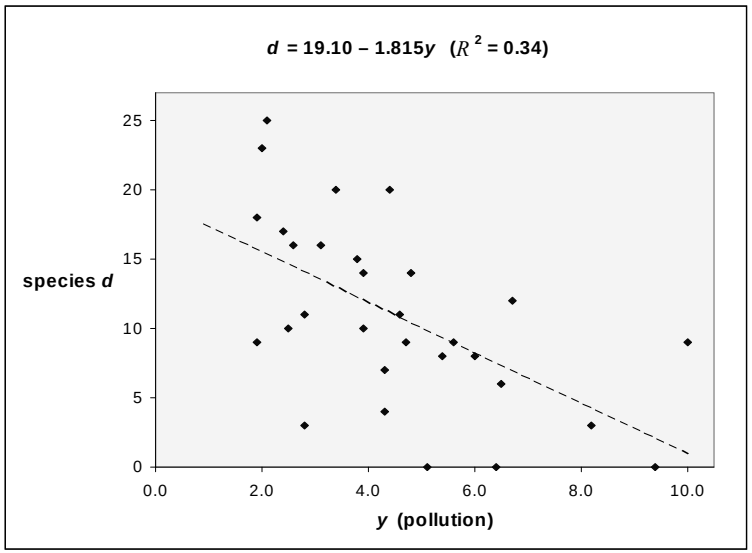
We first look at how a regression model with two explanatory variables can be depicted as a single point, and then extend this idea to principal component analysis and correspondence analysis

Some ecological data to illustrate regression biplots

SITE	SPECIES COUNTS					ENVIRONMENTAL VARS			
NO.	*a*	*b*	*c*	*d*	*e*	*Polln*	*Depth*	*Temp.*	*Sedmnt*
s1	0	2	9	14	2	4.8	72	3.5	S
s2	26	4	13	11	0	2.8	75	2.5	C
s3	0	10	9	8	0	5.4	59	2.7	C
s4	0	0	15	3	0	8.2	64	2.9	S
s5	13	5	3	10	7	3.9	61	3.1	C
s6	31	21	13	16	5	2.6	94	3.5	G
s7	9	6	0	11	2	4.6	53	2.9	S
s8	2	0	0	0	1	5.1	61	3.3	C
s9	17	7	10	14	6	3.9	68	3.4	C
s10	0	5	26	9	0	10.0	69	3.0	S
s11	0	8	8	6	7	6.5	57	3.3	C
s12	14	11	13	15	0	3.8	84	3.1	S
s13	0	0	19	0	6	9.4	53	3.0	S
s14	13	0	0	9	0	4.7	83	2.5	C
s15	4	0	10	12	0	6.7	100	2.8	C
s16	42	20	0	3	6	2.8	84	3.0	G
s17	4	0	0	0	0	6.4	96	3.1	C
s18	21	15	33	20	0	4.4	74	2.8	G
s19	2	5	12	16	3	3.1	79	3.6	S
s20	0	10	14	9	0	5.6	73	3.0	S
s21	8	0	0	4	6	4.3	59	3.4	C
s22	35	10	0	9	17	1.9	54	2.8	S
s23	6	7	1	17	10	2.4	95	2.9	G
s24	18	12	20	7	0	4.3	64	3.0	C
s25	32	26	0	23	0	2.0	97	3.0	G
s26	32	21	0	10	2	2.5	78	3.4	S
s27	24	17	0	25	6	2.1	85	3.0	G
s28	16	3	12	20	2	3.4	92	3.3	G
s29	11	0	7	8	0	6.0	51	3.0	S
s30	24	37	5	18	1	1.9	99	2.9	G

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

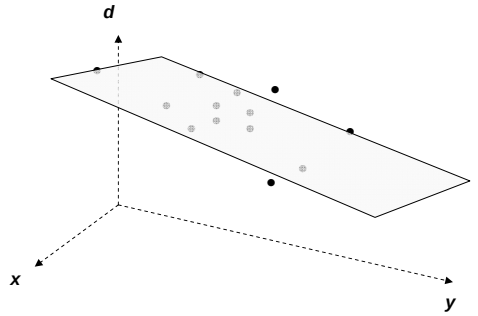
d	*y* (pollution)
14	4.8
11	2.8
8	5.4
3	8.2
10	3.9
16	2.6
11	4.6
0	5.1
14	3.9
9	10.0
6	6.5
15	3.8
0	9.4
9	4.7
12	6.7
3	2.8
0	6.4
20	4.4
16	3.1
9	5.6
4	4.3
9	1.9
17	2.4
7	4.3
23	2.0
10	2.5
25	2.1
20	3.4
8	6.0
18	1.9



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

$$\boldsymbol{d} = 6.135 - 1.388\boldsymbol{y} + 0.148\boldsymbol{x}$$

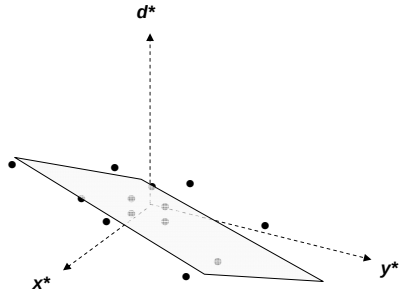


Regression model is a (hyper)plane

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

Small detour: R code to do regression

```
> summary(lm(d~y+x))

Call:
lm(formula = d ~ y + x)

Residuals:
     Min       1Q   Median       3Q      Max
-11.7001  -2.4684   0.1749   3.0563   9.1803

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.13518    6.25721   0.980  0.33554
y           -1.38766    0.48745  -2.847  0.00834 **
x            0.14822    0.06684   2.217  0.03520 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.162 on 27 degrees of freedom
Multiple R-squared: 0.4416,     Adjusted R-squared: 0.4003
F-statistic: 10.68 on 2 and 27 DF,  p-value: 0.0003831
```

From usual regression coefficients to standardized ones

$$d = a + bx + cy$$

$$\bar{d} = a + b\bar{x} + c\bar{y}$$

$$d - \bar{d} = b(x - \bar{x}) + c(y - \bar{y}) = bs_x \frac{(x - \bar{x})}{s_x} + cs_y \frac{(y - \bar{y})}{s_y}$$

$$\frac{d - \bar{d}}{s_d} = \frac{bs_x}{s_d} \frac{(x - \bar{x})}{s_x} + \frac{cs_y}{s_D} \frac{(y - \bar{y})}{s_y}$$

$$\boxed{d^* = b\frac{s_x}{s_d}x^* + c\frac{s_y}{s_d}y^*}$$

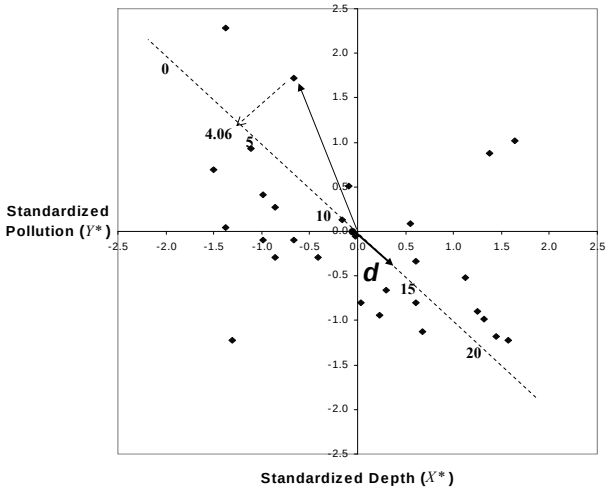
R code to calculate standardized regression coefficients

```
> lm(d~y+x)$coefficients
(Intercept)           y           x
  6.1351841  -1.3876604   0.1482187

> lm(d~y+x)$coefficients[2]*sd(y)/sd(d)
         y
-0.4457286

> lm(d~y+x)$coefficients[3]*sd(x)/sd(d)
        x
0.3471993
```

Another view of regression & prediction

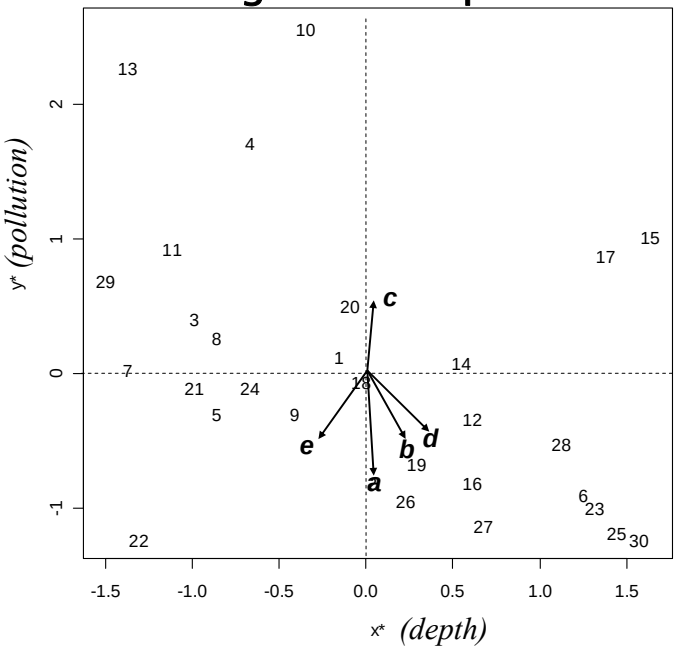
d	y	x	y*	x*
14	4.8	72	0.132	-0.156
11	2.8	75	-0.802	0.036
8	5.4	59	0.413	-0.988
3	8.2	64	1.720	-0.668
10	3.9	61	-0.288	-0.860
16	2.6	94	-0.895	1.253
11	4.6	53	0.039	-1.373
0	5.1	61	0.272	-0.860
14	3.9	68	-0.288	-0.412
9	10.0	69	2.561	-0.348
6	6.5	57	0.926	-1.116
15	3.8	84	-0.335	0.613
0	9.4	53	2.281	-1.373
9	4.7	83	0.086	0.549
12	6.7	100	1.020	1.637
3	2.8	84	-0.802	0.613
0	6.4	96	0.880	1.381
20	4.4	74	-0.054	-0.028
16	3.1	79	-0.662	0.292
9	5.6	73	0.506	-0.092
4	4.3	59	-0.101	-0.988
9	1.9	54	-1.222	-1.309
17	2.4	95	-0.989	1.317
7	4.3	64	-0.101	-0.668
23	2.0	97	-1.175	1.445
10	2.5	78	-0.942	0.228
25	2.1	85	-1.129	0.677
20	3.4	92	-0.522	1.125
8	6.0	51	0.693	-1.501
18	1.9	99	-1.222	1.573

$$d^* = -0.446y^* + 0.347x^*$$



variance explained (R^2): 44.2%

Regression biplot



variance explained:

a: 52.9%

b: 39.1%

c: 21.8%

d: 44.2%

e: 23.5%

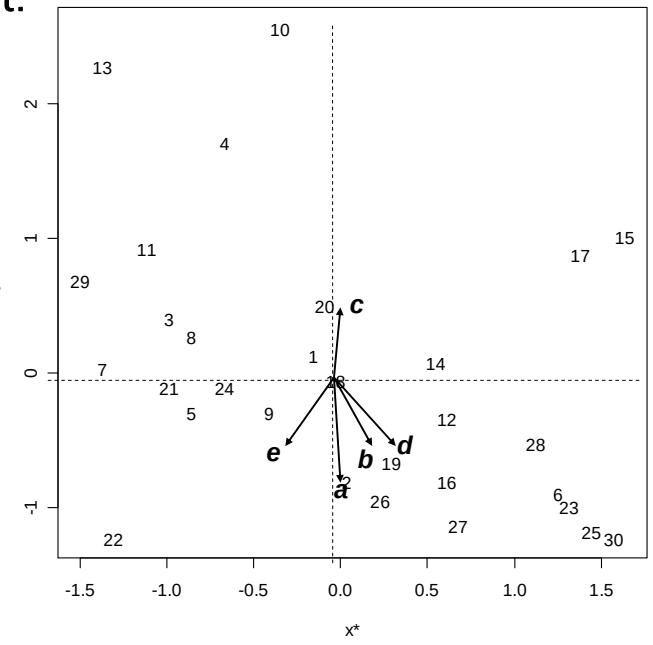
overall: 36.3%

significance:

	y	x
a:	**	
b:	**	
c:	*	
d:	**	*
e:	*	*

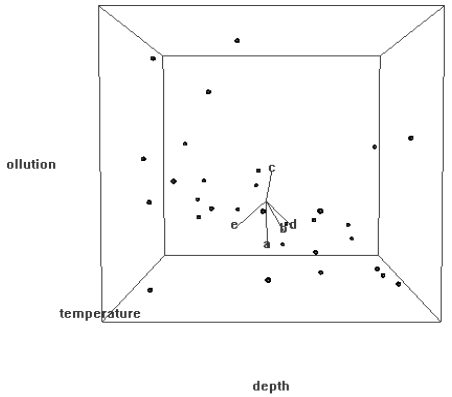
Regression biplot: summary

- A regression model can be represented as a point vector in space (in this case two-dimensional space because two explanatory variables)
- Reconstruct the data from projections of cases onto variable directions, but only as well as measured by R^2
- If x and y are uncorrelated, then the regression coefficients are just correlation coefficients (in this case, of the species with the explanatory variables)



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.



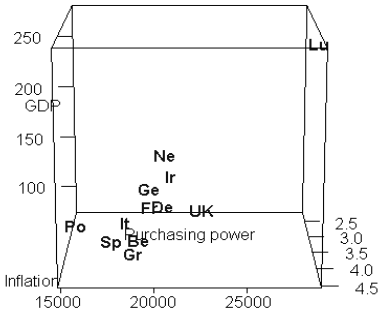
There will be a particular orientation of the vectors that gives maximum variance explained in the two-dimensional projection

Visualizing trivariate continuous data (repeat)

Continuous variables

X1 – Purchasing power/capita (euros)
X2 – GDP/capita (index)
X3 – inflation rate (%)

	Country	*X1*	*X2*	*X3*
Be	Belgium	19200	115.2	4.5
De	Denmark	20400	120.1	3.6
Ge	Germany	19500	115.6	2.8
Gr	Greece	18800	94.3	4.2
Sp	Spain	17600	102.6	4.1
Fr	France	19600	108.0	3.2
Ir	Ireland	20800	135.4	3.1
It	Italy	18200	101.8	3.5
Lu	Luxembourg	28800	276.4	4.1
Ne	Netherlands	20400	134.0	2.2
Po	Portugal	15000	76.0	2.7
UK	United Kingdom	22600	116.2	3.6

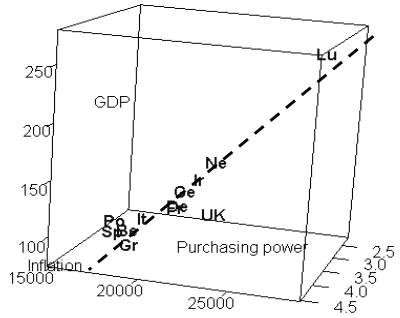


Visualizing trivariate continuous data (repeat)

Continuous variables

X1 – Purchasing power/capita (euros
X2 – GDP/capita (index)
X3 – inflation rate (%)

	Country	*X1*	*X2*	*X3*
Be	Belgium	19200	115.2	4.5
De	Denmark	20400	120.1	3.6
Ge	Germany	19500	115.6	2.8
Gr	Greece	18800	94.3	4.2
Sp	Spain	17600	102.6	4.1
Fr	France	19600	108.0	3.2
Ir	Ireland	20800	135.4	3.1
It	Italy	18200	101.8	3.5
Lu	Luxembourg	28800	276.4	4.1
Ne	Netherlands	20400	134.0	2.2
Po	Portugal	15000	76.0	2.7
UK	United Kingdom	22600	116.2	3.6



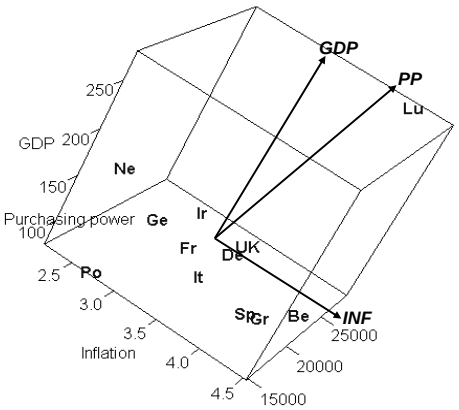
Visualizing trivariate continuous data (repeat)

Continuous variables

X1 – Purchasing power/capita (euro
X2 – GDP/capita (index)
X3 – inflation rate (%)

	Country	*X1*	*X2*	*X3*
Be	Belgium	19200	115.2	4.5
De	Denmark	20400	120.1	3.6
Ge	Germany	19500	115.6	2.8
Gr	Greece	18800	94.3	4.2
Sp	Spain	17600	102.6	4.1
Fr	France	19600	108.0	3.2
Ir	Ireland	20800	135.4	3.1
It	Italy	18200	101.8	3.5
Lu	Luxembourg	28800	276.4	4.1
Ne	Netherlands	20400	134.0	2.2
Po	Portugal	15000	76.0	2.7
UK	United Kingdom	22600	116.2	3.6

cor	*X1*	*X2*	*X3*
X1	1.000	0.929	0.243
X2	0.929	1.000	0.207
X3	0.243	0.207	1.000



General principle for adding biplot axes

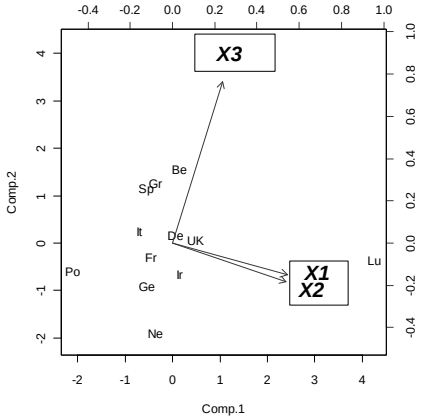
- For a samples by variables matrix situate the samples in some space of representation. This will usually be the low-dimensional subspace onto which the sample points have been projected. In the specific case of CA it could be the space of the row profiles, for example, with respect to principal axes.
- Perform the regression of each centred variable on the (principal) axes, and draw the variable as a vector using the regression coefficients as coordinates. In CA where the rows are weighted, the regression has to be performed with weights.
- The vector indicates the direction of maximum slope of the regression plane (i.e., the gradient), and its contours are perpendicular to this vector. This means the vector can be calibrated in units of the (predicted) variable.
- Special case: samples are standardized on uncorrelated principal axes and variable is standardized, then the regression coefficients are just the variable – axis correlations.

Visualizing trivariate continuous data (repeat)

Continuous variables

X1 – Purchasing power/capita (euro
X2 – GDP/capita (index)
X3 – inflation rate (%)

	Country	*X1*	*X2*	*X3*
Be	Belgium	19200	115.2	4.5
De	Denmark	20400	120.1	3.6
Ge	Germany	19500	115.6	2.8
Gr	Greece	18800	94.3	4.2
Sp	Spain	17600	102.6	4.1
Fr	France	19600	108.0	3.2
Ir	Ireland	20800	135.4	3.1
It	Italy	18200	101.8	3.5
Lu	Luxembourg	28800	276.4	4.1
Ne	Netherlands	20400	134.0	2.2
Po	Portugal	15000	76.0	2.7
UK	United Kingdom	22600	116.2	3.6



R: `biplot(princomp(EU, cor=T), scale=0)`

Asymmetric maps in CA are biplots

- The CA "model" based on the SVD of $\mathbf{S} = \mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}^{\mathsf{T}})\mathbf{D} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^{\mathsf{T}}$

PCs: $\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{D}_\alpha$, $\mathbf{G} = \mathbf{D}_c^{-1/2}\mathbf{V}\mathbf{D}_\alpha$ SCs: $\mathbf{\Phi} = \mathbf{D}_r^{-1/2}\mathbf{U}$, $\mathbf{\Gamma} = \mathbf{D}_c^{-1/2}\mathbf{V}$

is: $p_{ij} = r_i c_j (1 + \sum_k \alpha_k u_{ik} v_{jk})$

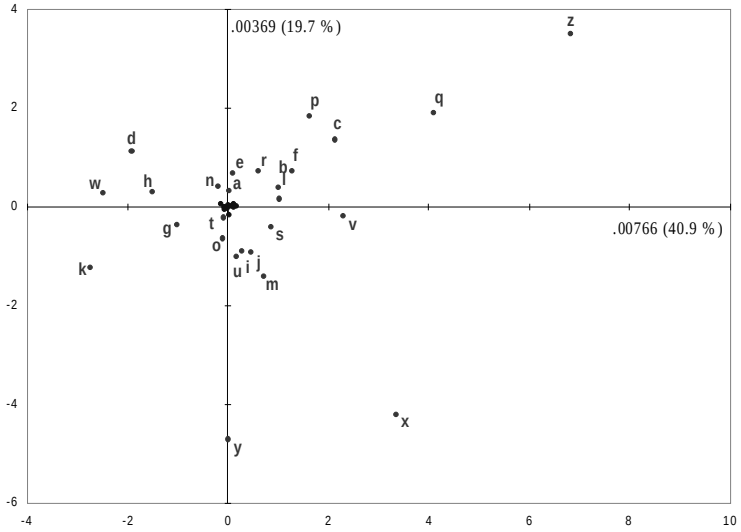
which, from the row profile point of view, can be written as either of these:

$$\frac{\frac{p_{ij}}{r_i} - c_j}{c_j} = \sum_k f_{ik}\gamma_{jk} \qquad \frac{p_{ij}}{r_i} - c_j = \sum_k f_{ik}(c_j\gamma_{jk}) \qquad \frac{\frac{p_{ij}}{r_i} - c_j}{c_j^{1/2}} = \sum_k f_{ik}(c_j^{1/2}\gamma_{jk})$$

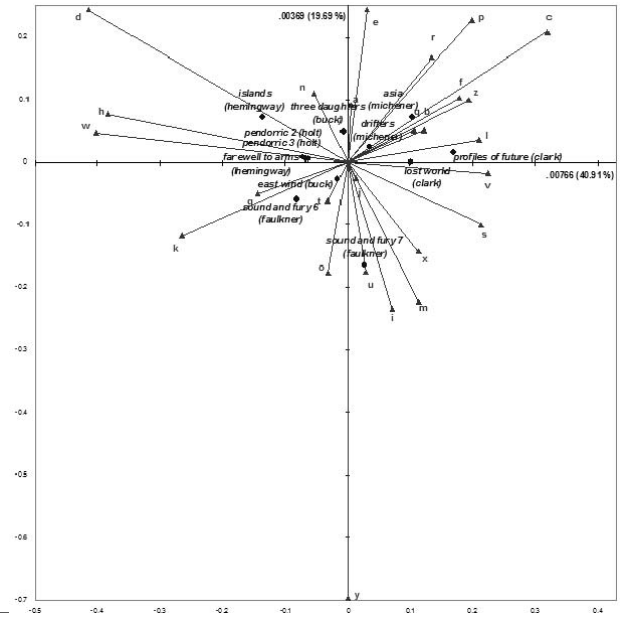
"classic" asymmetric map — Gabriel's biplot — Greenacre's standard (or contribution) biplot

R ca package: **`map="rowprincipal"` `map="rowgab"` `map="rowgreen"`**

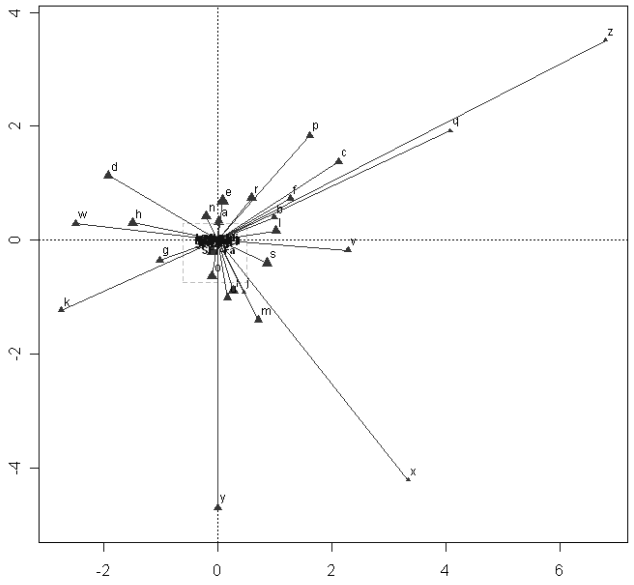
CA asymmetric map of "author" data



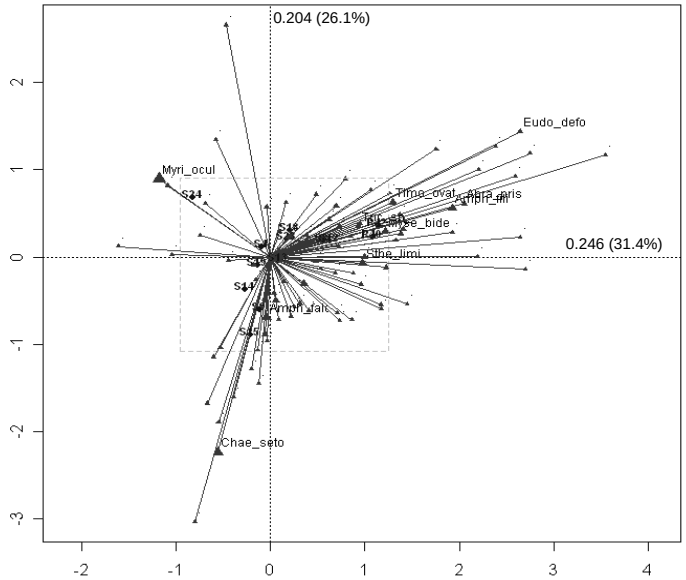
CA contribution biplot of "author" data (each letter vertex point is multiplied by the square root



Dynamic transition to CA contribution biplot



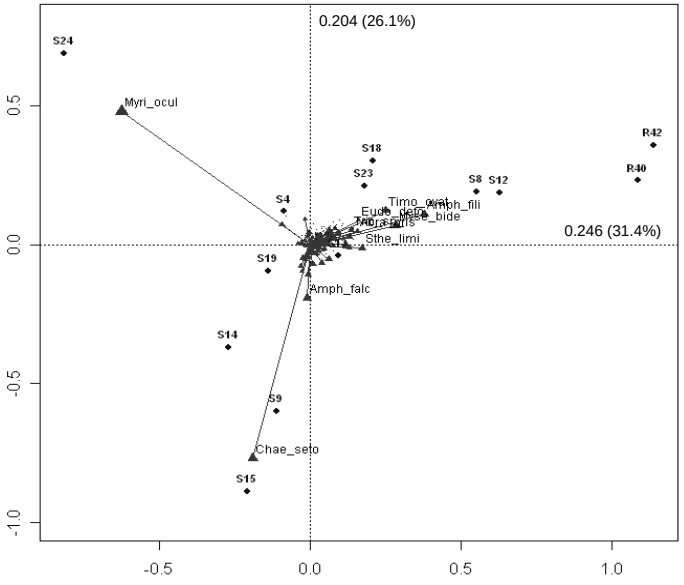
Asymmetric CA biplot of 'benthos'



explained inertia = 57.5%

Standard CA biplot of 'benthos'

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



explained inertia = 57.5%

Transition to standard biplot