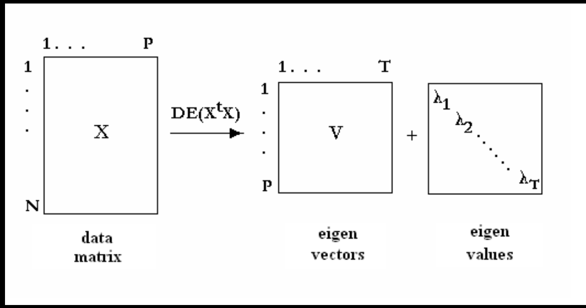
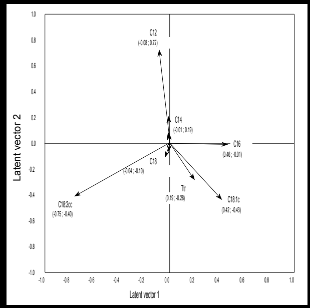
R algorithms
for the calculation of markers to be used
in the construction of predictive and interpolative
biplot axes in routine multivariate analyses

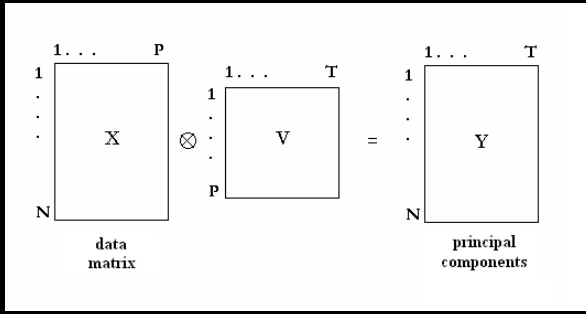
M. Rui Alves [1,2] and **M. Beatriz Oliveira** [2]

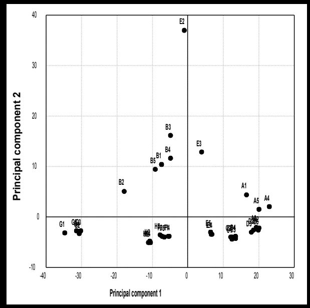
(1) Escola Superior de Tecnologia e Gestão, IPVC, Viana do Castelo, Portugal
(2) REQUI*M*TE, Faculdade de Farmácia, Universidade do Porto, Porto, Portugal

Example: PCA of a matrix of fatty acids in margarines









Journal of Chemometrics (2003), 17, 594-602

Examples of matrices (fatty acids in margarines)

matrix of latent values

Número	valor	Valor em percentagem	acumulado	acumulado em percentagem
1	11628.1	72.32	11628.10	72.32
2	3121.90	19.42	14750.00	91.74
3	1040.60	6.47	15790.60	98.21
4	214.72	1.34	16005.32	99.55
5	59.53	0.37	16064.85	99.92
6	8.26	0.05	16073.11	99.97
7	3.30	0.02	16076.41	99.99
8	1.70	0.01	16078.11	100.00
9	0.00	0.00	16078.11	100.00
10	0.00	0.00	16078.11	100.00
11	0.00	0.00	16078.11	100.00
12	0.00	0.00	16078.11	100.00
13	0.00	0.00	16078.11	100.00
14	0.00	0.00	16078.11	100.00
15	0.00	0.00	16078.11	100.00
16	0.00	0.00	16078.11	100.00
17	0.00	0.00	16078.11	100.00

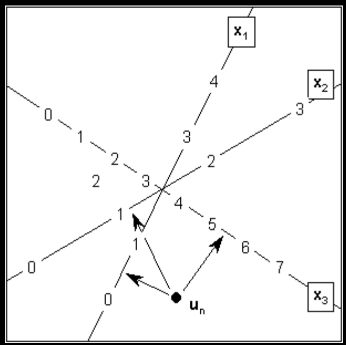
part of matrix of components

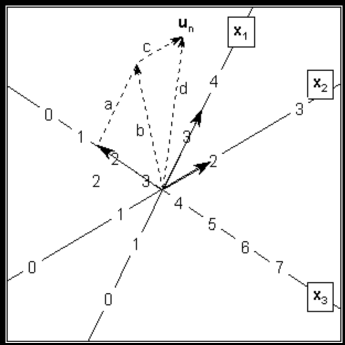
relações entre ácidos gordos e vectores próprios			
Ácido gordo	vector 1	vector 2	Comunalidade
C_8	-0.0065	0.0771	0.0060
C_{10}	-0.0062	0.0664	0.0044
C_{12}	-0.0819	0.7182	0.5226
C_{14}	-0.0145	0.1885	0.0357
C_{15}	0.0003	0.0002	0.0000
C_{16}	0.4625	-0.0107	0.2141
$C_{16:1c}$	0.0008	-0.0003	0.0000
C_{17}	0.0008	-0.0011	0.0000
C_{18}	-0.0362	-0.1015	0.0116
$C_{18:1c}$	0.4164	-0.4333	0.3612
C_{19}	0.0020	-0.0021	0.0000
$C_{18:2cc}$	-0.7521	-0.3975	0.7237
C_{20}	0.0014	-0.0054	0.0000
$C_{18:3c}$	-0.0048	-0.0590	0.0035
$C_{18:2cc}$	0.0003	-0.0003	0.0000
C_{22}	-0.0055	-0.0057	0.0001
C_{24}	0.1966	-0.2802	0.1172

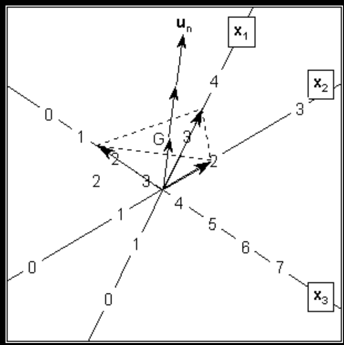
Gower's concepts for biplots

<u>Predictive biplots</u>:
Interpreteing results in terms of initial variables

<u>Interpolative biplots</u>:
Positioning new units in pre-existing graphs, mainly in routine quality control







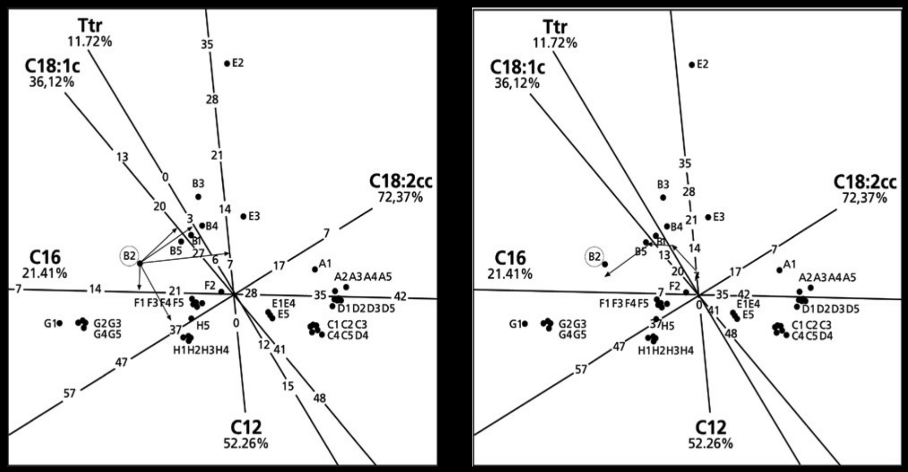
	x_1	x_2	x_3
u =	[0.5	1.2	5.4]

	x_1	x_2	x_3
u =	[3.5	2.0	1.5]

PCA biplots: fatty acids in margarines

Predictive biplots
Interpreting results

Interpolative biplots
Positioning new units



Journal of Chemometrics (2001), 15, 71-84

Journal of Chemometrics (2003), 17, 594-602

Problems on computation

- Gower e Hand, in their book *Biplots*, say:

  "(...) *The main computational problems* [of biplots] *are in integrating different bits of available software and in finding good portable graphic facilities*.

- To work with bipots we used:
  - Genstat 5.3.1 package to develop the algorithms and carry out the analyses
  - Statistica for Windows to draw all the graphs based on the converted ASCII outputs produced by Genstat

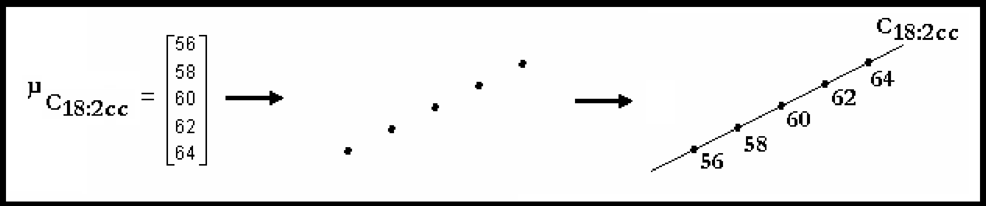
- We started to work with R in an attempt to provide a final, complete and more covenient solution

Projection of markers

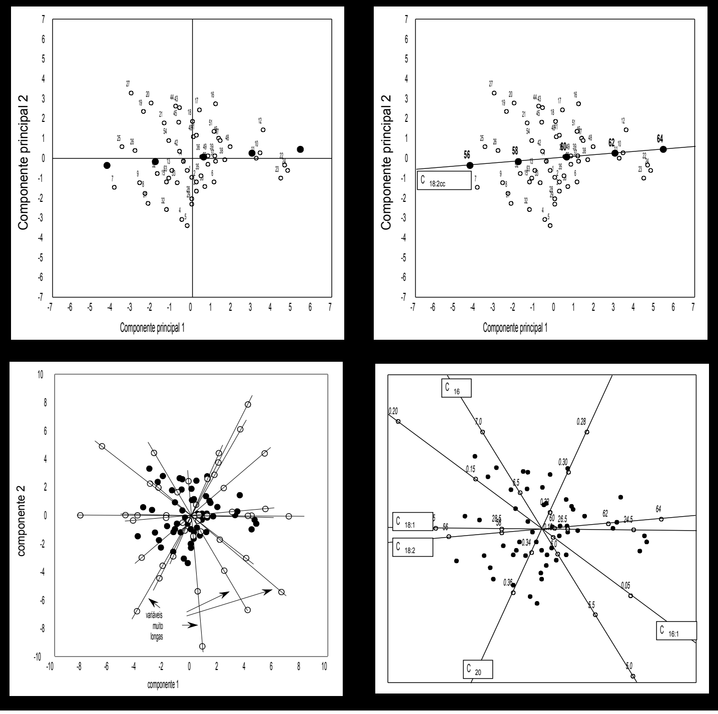
Projection of Markers

$$\left[\left(\mu_p - 1\bar{x}_p\right)s_p^{-1}\right]\left[e_p^t\left(V_\rho^{-1}\right)^t\right]\left[e_p^t\left(V_\rho^{-1}\right)^t\left(V_\rho^{-1}\right)e_p\right]^{-1}$$

$$\left[\left(\mu_p - 1\bar{x}_p\right)s_p^{-1}\right]\left[e_p^t V_\rho\right]$$



Building biplots (PCA of fatty acids in sunflower oils)



Strategies for the automation of the process

Two strategies were devised:

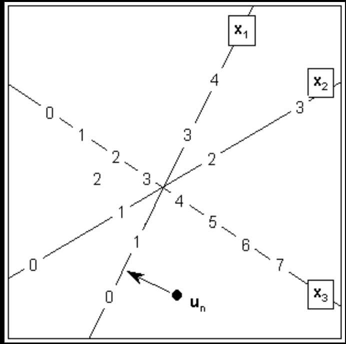
The first, more obvious, is to create an object containing the markers, the axis, the scale values and variable's name.

- Project the object.
- Live it in the graph if it fits well
- Delete it otherwise (too long or too short vectors)
- This procedure would require interactivity facilities

The second, more mathematical:

- Find a way for the evaluation of variables' predictive powers
- Leave in the graph variables displaying high predictive power
- Delete them otherwise
- Draw the graphs (only with automatically selected variables)

Algorithm for the selection based on predictive powers



$$u_{n(pred)} = \text{read in the graph} = 0{,}5$$

$$u_{n(inicial)} = \text{initial value} = 0{,}4$$

$$erro_{(u_n)} = 0{,}5 - 0{,}4 = 0{,}1$$

$$erro_{(x_p)} = \frac{1}{N}\sum_{n=1}^{N} \frac{u_{n(pred)} - u_{n(incial)}}{s_p}$$

Define a tolerance value

$$if \quad error_{(x_p)} \text{¿} \; tolerance \Rightarrow accept \; x_p$$

$$if \quad error_{(x_p)} \text{¿} \; tolerance \Rightarrow reject \; x_p$$

Evaluation of predictive power and decision

$$\Pr ediction = X \otimes V_{2\dim} \otimes V[k,]_{2\dim}^{t}$$

$$UnitsStdE = abs\left(X[,k] - \Pr ediction\right)$$

$$MeanStdE = N^{-1} \times 1^{t} \otimes UnitSdtE$$

$$if \left(MeanStdE < Tolerance\right)$$

$$project \quad variable \quad \left[\left(\mu_p - x_p\right) s_p^{-1}\right]\left[e_p^t \left(V_p^{-1}\right)\right]\left[e_p^t \left(V_p^{-1}\right)^t \left(V_p^{-1}\right) e_p\right]^{-1}$$

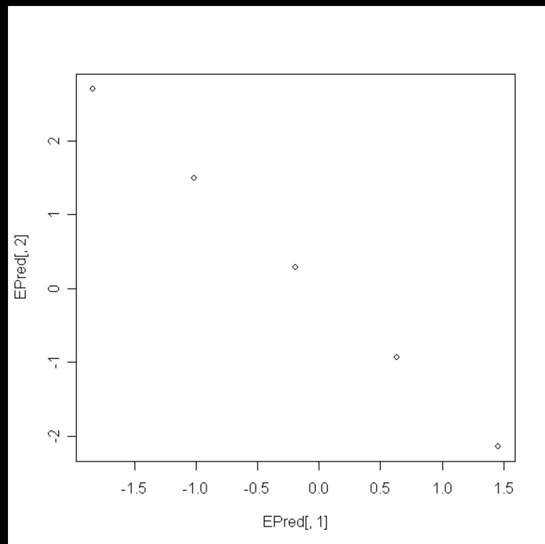
$$else$$

$$pr\text{int}\left(\text{ ital } deleted \right)$$


```
for (i in 1:(Q-1)) {
  for (j in (i+1):Q) {
    print("component"); print(i); print("component"); print(j)
    # latent variables for a pair of components only
    V2Dim[,1] <- RedV[,i]
    V2Dim[,2] <- RedV[,j]
    MStdE <- list()
    for (k in 1:P) {
      # evaluation of variable' s predictive power
      print("variavel") ; print(k)
      VarDir <- matrix(data=(V2Dim[k,]),nrow=1,ncol=2)
      Pred <- XStd %*% V2Dim %*% t(VarDir)
      UnitStdE <- abs(XStd[,k] - Pred)
      VarE <- (t(ColN1s) %*% UnitStdE)/N
      MStdE[[k]] <- VarE
      print(MStdE[[k]])
        if (MStdE[[k]] < Tolerance) {
          Zeros <- matrix(0,c(P,1))
          Zeros[k,] <- 1
          Adj1 <- t(Zeros) %*% V2Dim %*% t(V2Dim) %*% Zeros
          Adj2 <- (ColM1s %*% (1/Adj1) %*% t(Col2_1s))
          EPred <- Adj2 * (ScMat[[k]] %*% V2Dim)
          print(EPred)
          plot(EPred[,1],EPred[,2])
          }
        else
          print("deleted")
      }   }  }
```

Example of results provided by the algorithm

```
[1] "component"
[1] 1
[1] "component"
[1] 2
[1] "variavel"
[1] 1
        [,1]
[1,] 0.6583496
[1] "deleted"
[1] "variavel"
[1] 2
        [,1]
[1,] 0.5805472
[1] "deleted"
[1] "variavel"
[1] 3
        [,1]
[1,] 0.2413512
          [,1]       [,2]
[1,]  2.6066496  4.1854190
[2,]  1.4223328  2.2837971
[3,]  0.2380160  0.3821751
[4,] -0.9463008 -1.5194468
[5,] -2.1306176 -3.4210688
```



What needs to be done and acknowledgments

- Reminding Gower and Hand: “(...) *The main computational problems* [dos biplots] *are in integrating different bits of available software and in finding good portable graphic facilities*.
- The algorithms are done, although probably not in the best or more efficient way, but they provide correct results
- Produce biplots by making the graphs of components and merging the graphs of individual variables as objects
- Multivariate analyses can therefore be made fully automatic, including the interpretation processes
- Paul Murrell and Robert Gittins provided valuable information on graphs and strategies to finalize the work, but unfortunately we did not find the time necessary to do it

Thank you