

Algebra of Principal Component Analysis

Data: $\mathbf{Y} = \begin{bmatrix} 2 & 1 \\ 3 & 4 \\ 5 & 0 \\ 7 & 6 \\ 9 & 2 \end{bmatrix}$ Centre each column on its mean: $\mathbf{Y}_c = [y - \bar{y}] = \begin{bmatrix} -3.2 & -1.6 \\ -2.2 & 1.4 \\ -0.2 & -2.6 \\ 1.8 & 3.4 \\ 3.8 & -0.6 \end{bmatrix}$

Covariance matrix (2 variables): $\mathbf{S} = \frac{1}{n-1} \mathbf{Y}_c' \mathbf{Y}_c = \begin{bmatrix} 8.2 & 1.6 \\ 1.6 & 5.8 \end{bmatrix}$

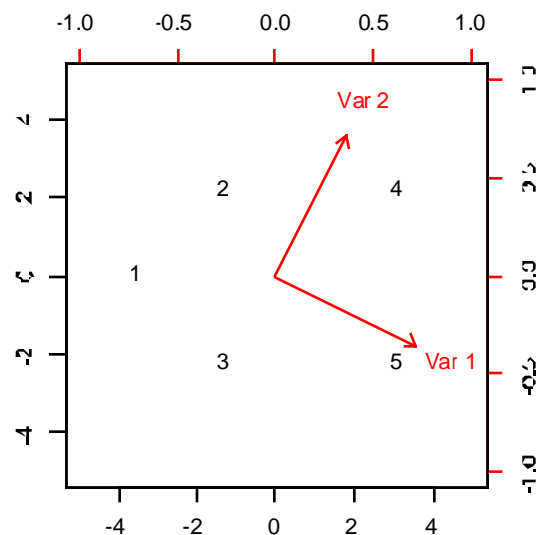
Equation for eigenvalues and eigenvectors of \mathbf{S} : $(\mathbf{S} - \lambda_k \mathbf{I}) \mathbf{u}_k = \mathbf{0}$

Eigenvalues: $\lambda_1 = 9, \lambda_2 = 5$ Matrix of eigenvalues: $\Lambda = \begin{bmatrix} 9 & 0 \\ 0 & 5 \end{bmatrix}$

Matrix of eigenvectors: $\mathbf{U} = \begin{bmatrix} 0.8944 & -0.4472 \\ 0.4472 & 0.8944 \end{bmatrix}$

Positions of the 5 objects in ordination space: $\mathbf{F} = [y - \bar{y}] \mathbf{U}$

$$\mathbf{F} = \begin{bmatrix} -3.2 & -1.6 \\ -2.2 & 1.4 \\ -0.2 & -2.6 \\ 1.8 & 3.4 \\ 3.8 & -0.6 \end{bmatrix} \begin{bmatrix} 0.8944 & -0.4472 \\ 0.4472 & 0.8944 \end{bmatrix} = \begin{bmatrix} -3.578 & 0 \\ -1.342 & 2.236 \\ -1.342 & -2.236 \\ 3.130 & 2.236 \\ 3.130 & -2.236 \end{bmatrix}$$



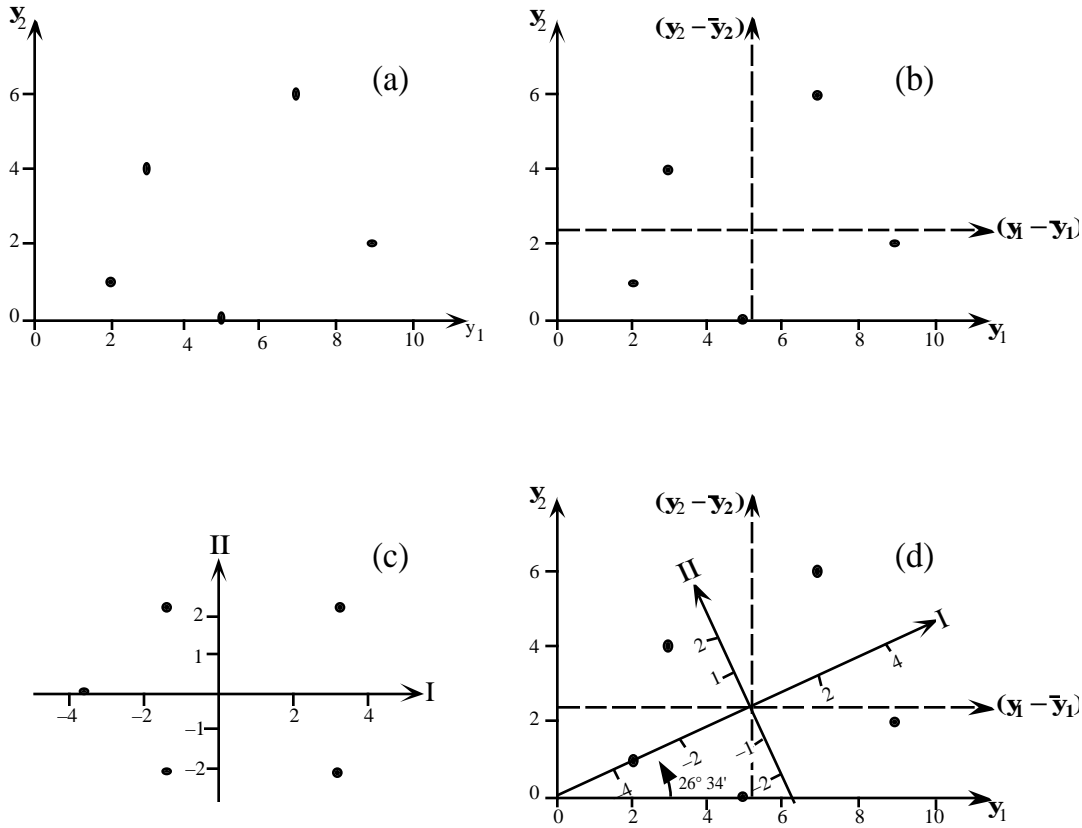


Figure 9.2 Numerical example of principal component analysis. (a) Five objects are plotted with respect to descriptors y_1 and y_2 . (b) After centring the data, the objects are now plotted with respect to $(y_1 - \bar{y}_1)$ and $(y_2 - \bar{y}_2)$, represented by dashed axes. (c) The objects are plotted with reference to principal axes I and II, which are centred with respect to the scatter of points. (d) The two systems of axes (b and c) can be superimposed after a rotation of $26^\circ 34'$.

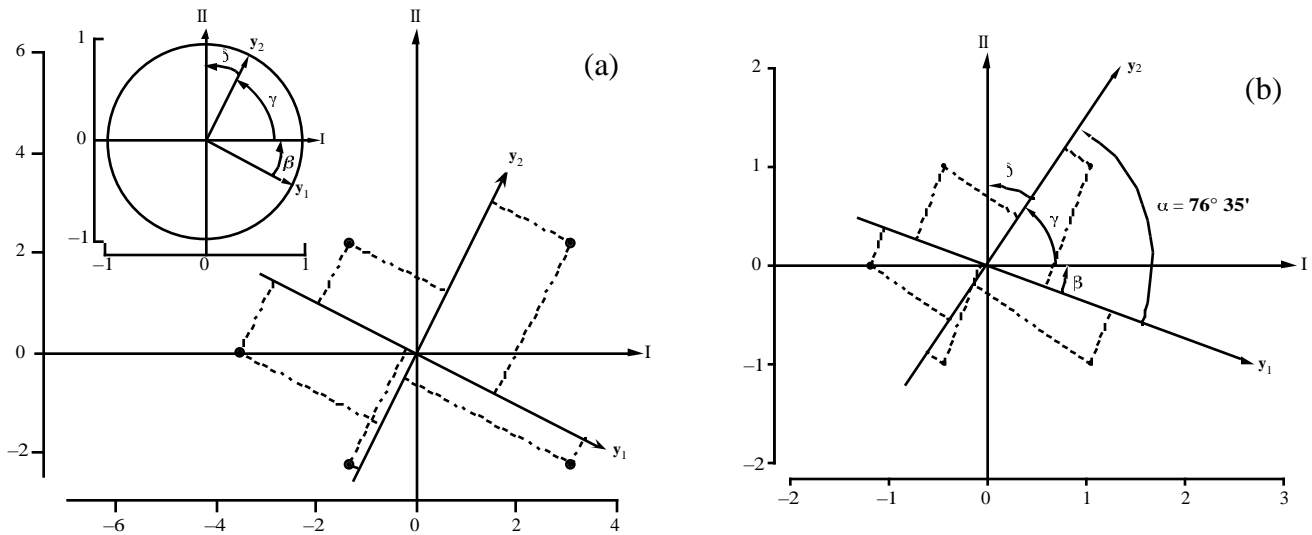


Fig. 9.3 Numerical example from Fig. 9.2. Distance and correlation biplots are discussed in Subsection 9.1.4. **(a) Distance biplot.** The eigenvectors are scaled to lengths 1. Inset: descriptors (matrix \mathbf{U}). Main graph: descriptors (matrix \mathbf{U} ; arrows) and objects (matrix \mathbf{F} ; dots). The interpretation of the object-descriptor relationships is not based on their proximity, but on orthogonal projections (dashed lines) of the objects on the descriptor-axes or their extensions. **(b) Correlation biplot.** Descriptors (matrix $\mathbf{U}\Lambda^{1/2}$; arrows) with a covariance angle of $76^\circ 35'$. Objects (matrix \mathbf{G} ; dots). Projecting the objects orthogonally on a descriptor (dashed lines) reconstructs the values of the objects along that descriptors, to within a multiplicative constant.

Use the following matrices to draw biplots

Distance biplot (scaling 1): objects = \mathbf{F} , variables = \mathbf{U}

Correlation biplot (scaling 2): objects = $\mathbf{G} = \mathbf{F}\Lambda^{-1/2}$, variables = $\mathbf{U}_{sc2} = \mathbf{U}\Lambda^{1/2}$

These two projections respect the biplot rule, that the product of the two projected matrices reconstruct the data \mathbf{Y} :

$$\text{Distance biplot: } \mathbf{F}\mathbf{U}' = \mathbf{Y}$$

$$\text{Correlation biplot: } \mathbf{G}(\mathbf{U}\Lambda^{1/2})' = \mathbf{Y}$$

PCA example, three variables

```
# Create the data matrix
data.3 <- matrix(c(2,3,5,7,9,10,1,4,0,6,2,5,0,1,-1,-1,1,0),6,3)
data.3
      [,1] [,2] [,3]
[1,]    2    1    0
[2,]    3    4    1
[3,]    5    0   -1
[4,]    7    6   -1
[5,]    9    2    1
[6,]   10    5    0

# Centre the variables
data.cent <- scale(data.3, center=TRUE, scale=FALSE)
data.cent
      [,1] [,2] [,3]
[1,]   -4   -2    0
[2,]   -3    1    1
[3,]   -1   -3   -1
[4,]    1    3   -1
[5,]    3   -1    1
[6,]    4    2    0

# Compute the covariance matrix
data.cov <- cov(data.cent) # or, because the data are not standardized: cov(data.3)
      [,1] [,2] [,3]
[1,] 10.4  3.2  0.0
[2,]  3.2  5.6  0.0
[3,]  0.0  0.0  0.8

# Compute the eigenvalues and eigenvectors
data.eig <- eigen(data.cov)
data.eig

$values
[1] 12.0  4.0  0.8

$vectors
      [,1]      [,2] [,3]
[1,] 0.8944272  0.4472136  0
[2,] 0.4472136 -0.8944272  0
[3,] 0.0000000  0.0000000  1

# Compute the output matrices for scaling 1
U <- data.eig$vectors
F.mat <- data.cent %*% U

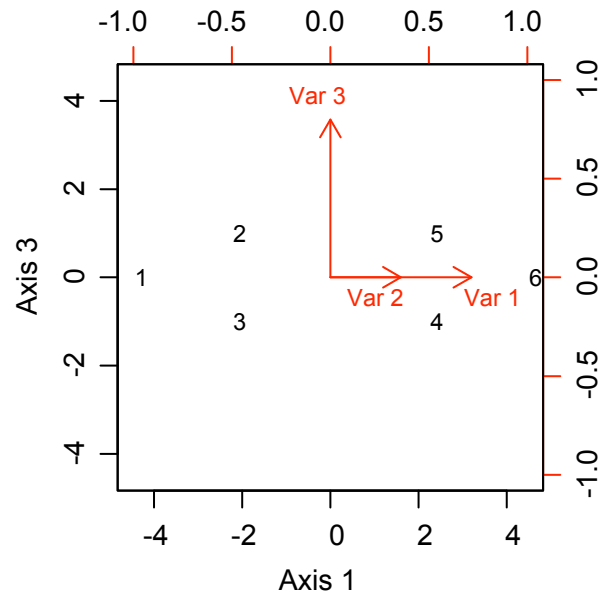
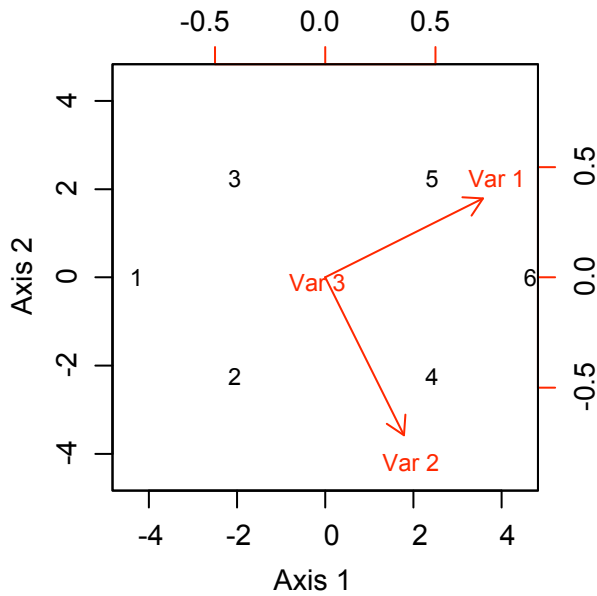
# Compute the output matrices for scaling 2
U.sc2 <- U %*% diag(data.eig$values^(0.5))
G.mat <- F.mat %*% diag(data.eig$values^(-0.5))
```

```

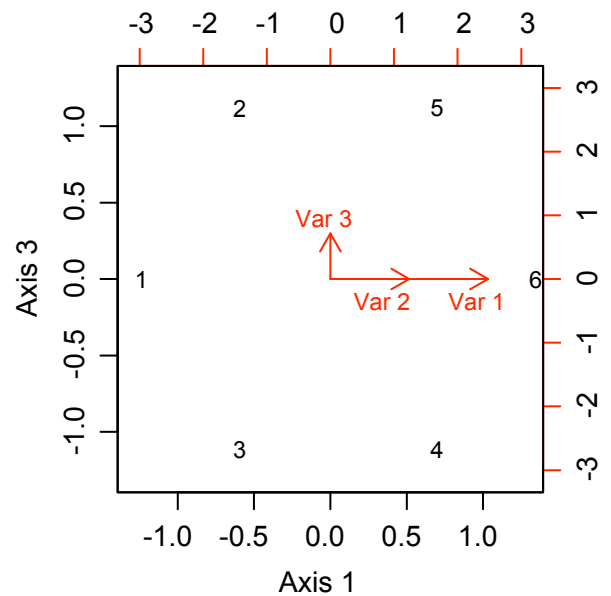
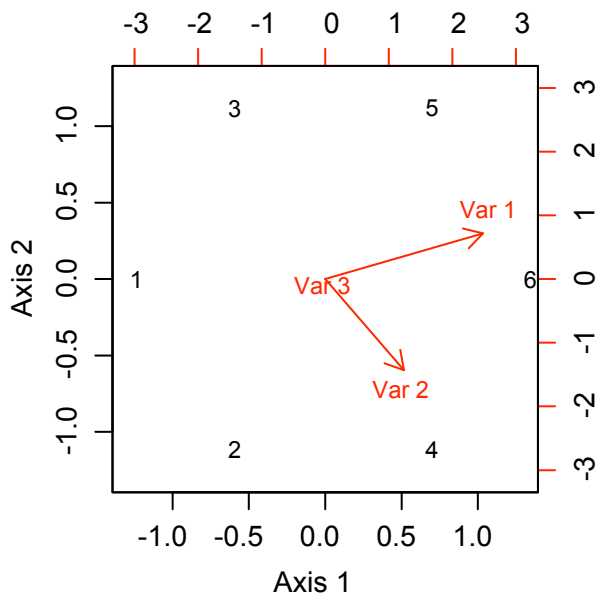
# Draw the scaling 1 and 2 biplots
par(mfrow=c(2,2))
biplot(F.mat[,c(1,2)], U[,c(1,2)])
biplot(F.mat[,c(1,3)], U[,c(1,3)])
biplot(G.mat[,c(1,2)], U.sc2[,c(1,2)])
biplot(G.mat[,c(1,3)], U.sc2[,c(1,3)])

```

Scaling 1 biplots, axes (1, 2) and (1, 3)



Scaling 2 biplots, axes (1, 2) and (1, 3)



Data transformations

Transform physical variables (*Ecology*) or characters (*Taxonomy*)

- Univariate distributions are not symmetrical
⇒ Apply skewness-reduction transformation
- Variables are not in the same physical units

⇒ Apply standardization $z_i = \frac{y_i - \bar{y}}{s_y}$ or ranging $y'_i = \frac{y_i - y_{min}}{y_{max} - y_{min}}$

Transform community composition data (*Ecology*)

(species presence-absence or abundance)

- Reduce asymmetry of distributions
⇒ Apply $\log(y + c)$ transformation
- Make community composition data suitable for Euclidean-based ordination methods (PCA, RDA)
⇒ Use the chord, chi-square, profile, or Hellinger transformations (Legendre & Gallagher 2001)

Some uses of principal component analysis (PCA)

- Two-dimensional ordination of the *objects*:

- Sampling sites in ecology
- Individuals or taxa in taxonomy

A 2-dimensional ordination diagram is an interesting graphical support for representing other properties of multivariate data, e.g., clusters.

- Detect outliers or erroneous data in data tables
- Find groups of *variables* that behave in the same way:
 - Species in ecology
 - Morphological/behavioural/molecular variables in taxonomy
- Simplify (collinear) data; remove noise
- Remove an identifiable component of variation
e.g., size factor in log-transformed morphological data

Algebra of Correspondence Analysis

$$\text{Frequency data table } \mathbf{Y} = \begin{bmatrix} \mathbf{f}_{ij} \end{bmatrix} = \begin{bmatrix} 10 & 10 & 20 \\ 10 & 15 & 10 \\ 15 & 5 & 5 \end{bmatrix} \begin{bmatrix} \mathbf{f}_{i+} \\ 40 \\ 35 \\ 25 \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{f}_{+j} \end{bmatrix} = \begin{bmatrix} 35 & 30 & 35 \end{bmatrix} \quad 100 = \mathbf{f}_{++}$$

$$\begin{aligned} p_{ij} &= \mathbf{f}_{ij} / \mathbf{f}_{++} \\ p_{i+} &= \mathbf{f}_{i+} / \mathbf{f}_{++} \\ p_{+j} &= \mathbf{f}_{+j} / \mathbf{f}_{++} \end{aligned}$$

$$\text{Matrix } \mathbf{Q} = [\bar{q}_{ij}] = \left[\frac{p_{ij} - p_{i+}p_{+j}}{\sqrt{p_{i+}p_{+j}}} \right] = \frac{(O_{ij} - E_{ij}) / \sqrt{E_{ij}}}{\sqrt{\mathbf{f}_{++}}}$$

$$\text{Matrix } \mathbf{Q} = \begin{bmatrix} -0.10690 & -0.05774 & 0.16036 \\ -0.06429 & 0.13887 & -0.06429 \\ 0.21129 & -0.09129 & -0.12667 \end{bmatrix}$$

$$\text{Cross-product matrix: } \mathbf{Q}'\mathbf{Q} = \begin{bmatrix} 0.06020 & -0.02204 & -0.03980 \\ -0.02204 & 0.03095 & -0.00661 \\ -0.03980 & -0.00661 & 0.04592 \end{bmatrix}$$

$$\text{Compute eigenvalues and eigenvectors of } \mathbf{Q}'\mathbf{Q} : \quad (\mathbf{Q}'\mathbf{Q} - \lambda_k \mathbf{I}) \mathbf{u}_k = \mathbf{0}$$

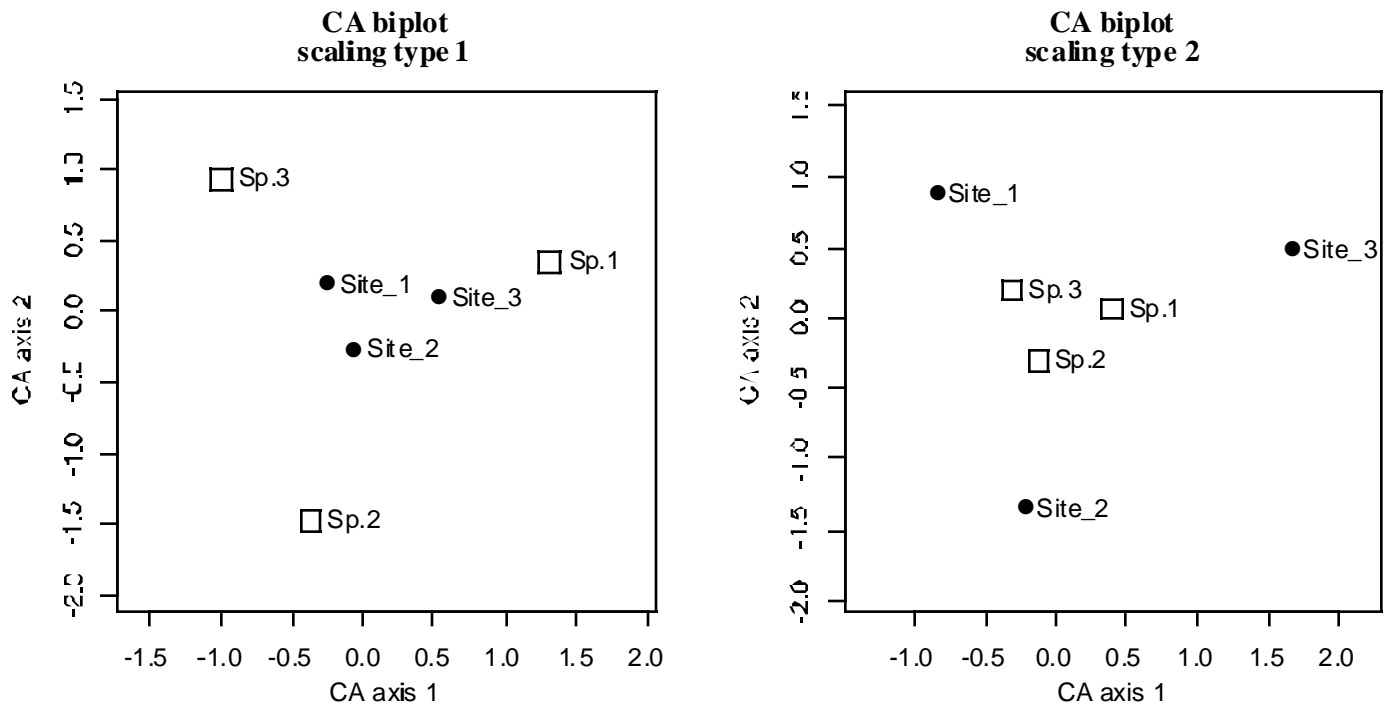
$$\text{Eigenvalues: } \lambda_1 = 0.096, \lambda_2 = 0.041 \quad \text{Matrix of eigenvalues: } \Lambda = \begin{bmatrix} 0.096 & 0 \\ 0 & 0.041 \end{bmatrix}$$

There are never more than $k = \min(r - 1, c - 1)$ eigenvalues > 0 in CA

$$\text{Matrix of eigenvectors of } \mathbf{Q}'\mathbf{Q}_{(c \times c)} : \quad \mathbf{U}_{(c \times k)} = \begin{bmatrix} 0.78016 & 0.20336 \\ -0.20383 & -0.81145 \\ -0.59144 & 0.54790 \end{bmatrix}$$

$$\text{Matrix of eigenvectors of } \mathbf{Q}\mathbf{Q}'_{(r \times r)} : \quad \hat{\mathbf{U}}_{(r \times k)} = \mathbf{Q}\mathbf{U}\Lambda^{-1/2} = \begin{bmatrix} -0.53693 & 0.55831 \\ -0.13043 & -0.79561 \\ 0.83349 & 0.23516 \end{bmatrix}$$

Compute matrices \mathbf{F} and \mathbf{V} for scaling 1 biplot, and $\hat{\mathbf{V}}$ and $\hat{\mathbf{F}}$ for scaling 2 biplot:



Calculation details

Compute matrices \mathbf{V} , $\hat{\mathbf{V}}$, \mathbf{F} , and $\hat{\mathbf{F}}$ used in the ordination biplots:

$$\mathbf{V}_{(c \times k)} = \mathbf{D}(p_{+j})^{-1/2} \mathbf{U} \quad \text{where } p_{+j} = f_{+j}/f_{++}$$

$$\hat{\mathbf{V}}_{(r \times k)} = \mathbf{D}(p_{i+})^{-1/2} \hat{\mathbf{U}} \quad \text{where } p_{i+} = f_{i+}/f_{++}$$

$$\mathbf{F}_{(r \times k)} = \hat{\mathbf{V}} \mathbf{\Lambda}^{1/2}$$

$$\hat{\mathbf{F}}_{(c \times k)} = \mathbf{V} \mathbf{\Lambda}^{1/2}$$

Biplot, scaling type 1: plot \mathbf{F} for sites, \mathbf{V} for species:

- This projection preserves the chi-square distance among the sites.
- The sites are at the centroids (barycentres) of the species.

Biplot, scaling type 2: plot $\hat{\mathbf{V}}$ for sites, $\hat{\mathbf{F}}$ for species:

- This projection preserves the chi-square distance among the species.
- The species are at the centroids (barycentres) of the sites.

Principal coordinate analysis (PCoA)

Example: a Euclidean distance matrix computed from the data of the PCA example.

$$\mathbf{D} = \begin{bmatrix} 0.00000 & 3.16228 & 3.16228 & 7.07107 & 7.07107 \\ 3.16228 & 0.00000 & 4.47214 & 4.47214 & 6.32456 \\ 3.16228 & 4.47214 & 0.00000 & 6.32456 & 4.47214 \\ 7.07107 & 4.47214 & 6.32456 & 0.00000 & 4.47214 \\ 7.07107 & 6.32456 & 4.47214 & 4.47214 & 0.00000 \end{bmatrix}$$

Transform \mathbf{D} to a new matrix $\mathbf{A} = [a_{hi}]$: $a_{hi} = -0.5D_{hi}^2$

Centre matrix \mathbf{A} to produce matrix $\mathbf{\Delta}$ with sums of the rows and columns equal to 0:

$$\mathbf{\Delta} = [\delta_{hi}] = [a_{hi} - \bar{a}_h - \bar{a}_i + \bar{a}]$$

$$\mathbf{\Delta} = \begin{bmatrix} 12.8 & 4.8 & 4.8 & -11.2 & -11.2 \\ 4.8 & 6.8 & -3.2 & 0.8 & -9.2 \\ 4.8 & -3.2 & 6.8 & -9.2 & 0.8 \\ -11.2 & 0.8 & -9.2 & 14.8 & 4.8 \\ -11.2 & -9.2 & 0.8 & 4.8 & 14.8 \end{bmatrix}$$

Compute the eigenvalues and eigenvectors of matrix $\mathbf{\Delta}$. Scale the eigenvectors to lengths equal to the square roots of their respective eigenvalues, $\sqrt{\mathbf{u}'_k \mathbf{u}_k} = \sqrt{\lambda_k}$.

Eigenvalues:	$\lambda_1 = 36$	$\lambda_2 = 20$
Objects	Eigenvectors	
\mathbf{x}_1	-3.578	0.000
\mathbf{x}_1	-1.342	-2.236
\mathbf{x}_1	-1.342	2.236
\mathbf{x}_1	3.130	-2.236
\mathbf{x}_1	3.130	2.236
Eigenvector length	$\sqrt{36} = 6.000$	$\sqrt{20} = 4.472$
PCoA eigenvalues / $(n-1)$	9	5

Compare the PCoA eigenvalues and eigenvectors to the PCA eigenvalues and matrix \mathbf{F} .

PCoA is used for ordination of \mathbf{D} matrices produced by functions other than the Euclidean.