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

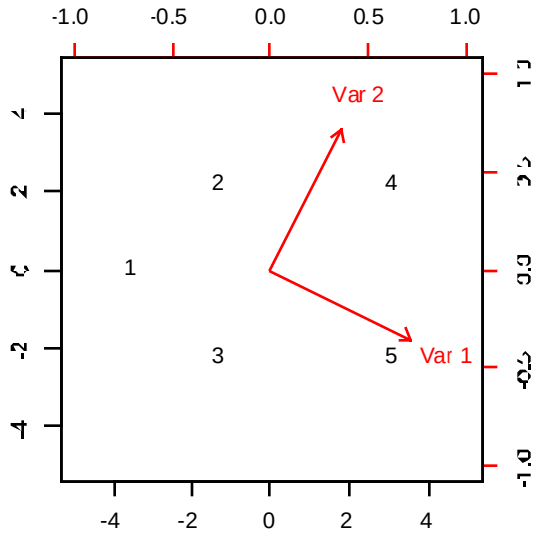
Equation for eigenvalues and eigenvectors of $\mathbf{S}$: $(\mathbf{S} - \lambda_k \mathbf{I})\, \mathbf{u}_k = \mathbf{0}$

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: $\mathbf{F} = [y - \bar{y}]\,\mathbf{U}$

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



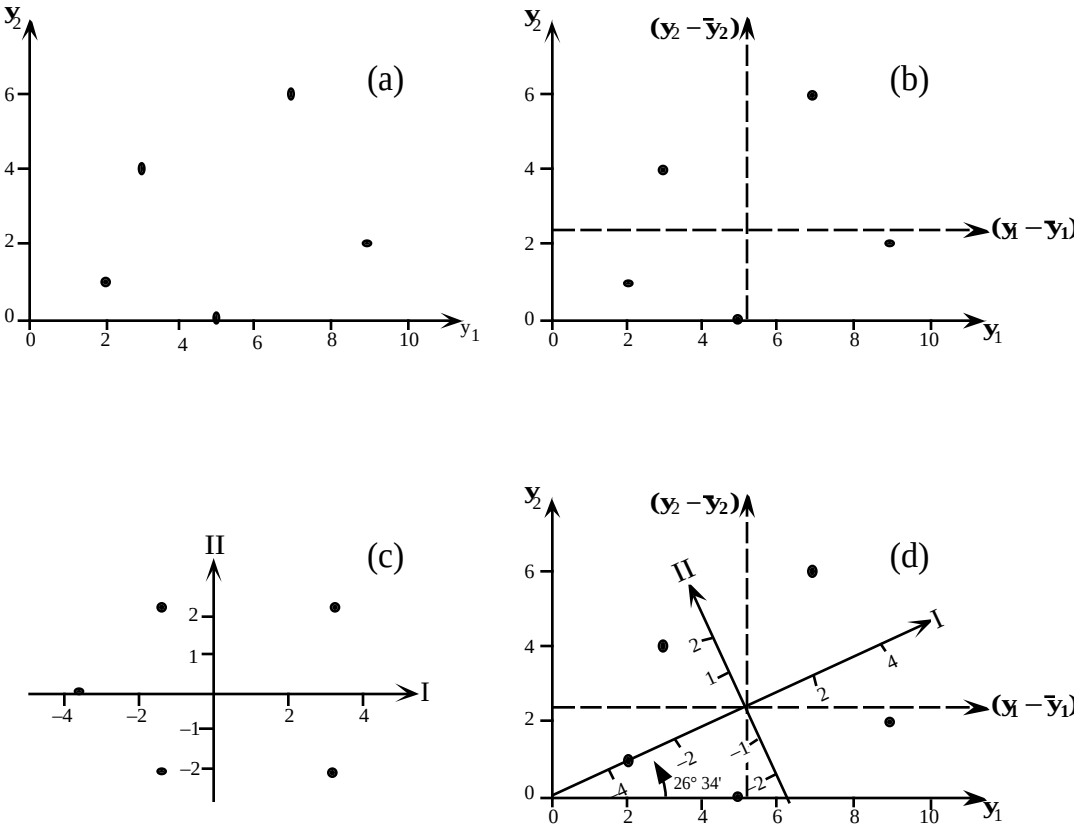


Figure 9.2 Numerical example of principal component analysis. (a) Five objects are plotted with respect to descriptors $\mathbf{y}_1$ and $\mathbf{y}_2$. (b) After centring the data, the objects are now plotted with respect to $(\mathbf{y}_1 - \bar{\mathbf{y}}_1)$ and $(\mathbf{y}_2 - \bar{\mathbf{y}}_2)$, 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°34'.

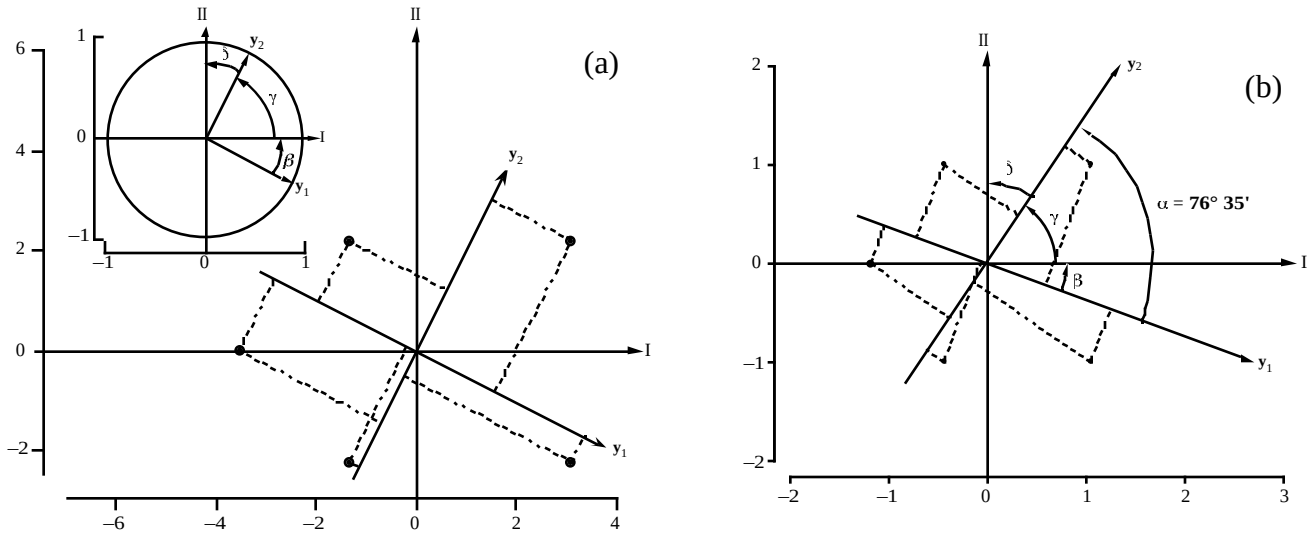


Fig. 9.3 Numerical example from Fig. 9.2. Distance and correlation biplots are discussed in Subsection 9.1.4. **(a) Distance biplot.** The eigenvectors are scaled to lengths 1. Inset: descriptors (matrix **U**). Main graph: descriptors (matrix **U**; arrows) and objects (matrix **F**; dots). The interpretation 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°35'. Objects (matrix **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 = **F**, variables = **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 **Y**:

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

Data transformation

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 $y'_i = \frac{y_i - y_{min}}{y_{max} - y_{min}}$

- Multistate qualitative variables

  In some cases, transform them to dummy (binary) variables

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, 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

$$\text{Frequency data table } \mathbf{Y} = \left[f_{ij} \right] = \begin{bmatrix} 10 & 10 & 20 \\ 10 & 15 & 10 \\ 15 & 5 & 5 \end{bmatrix} \quad \begin{bmatrix} f_{i+} \\ 40 \\ 35 \\ 25 \end{bmatrix}$$

$$\left[f_{+j} \right] = \begin{bmatrix} 35 & 30 & 35 \end{bmatrix} \quad 100 = f_{++}$$

$$p_{ij} = f_{ij} / f_{++}$$

$$p_{i+} = f_{i+} / f_{++}$$

$$p_{+j} = f_{+j} / f_{++}$$

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

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

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

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

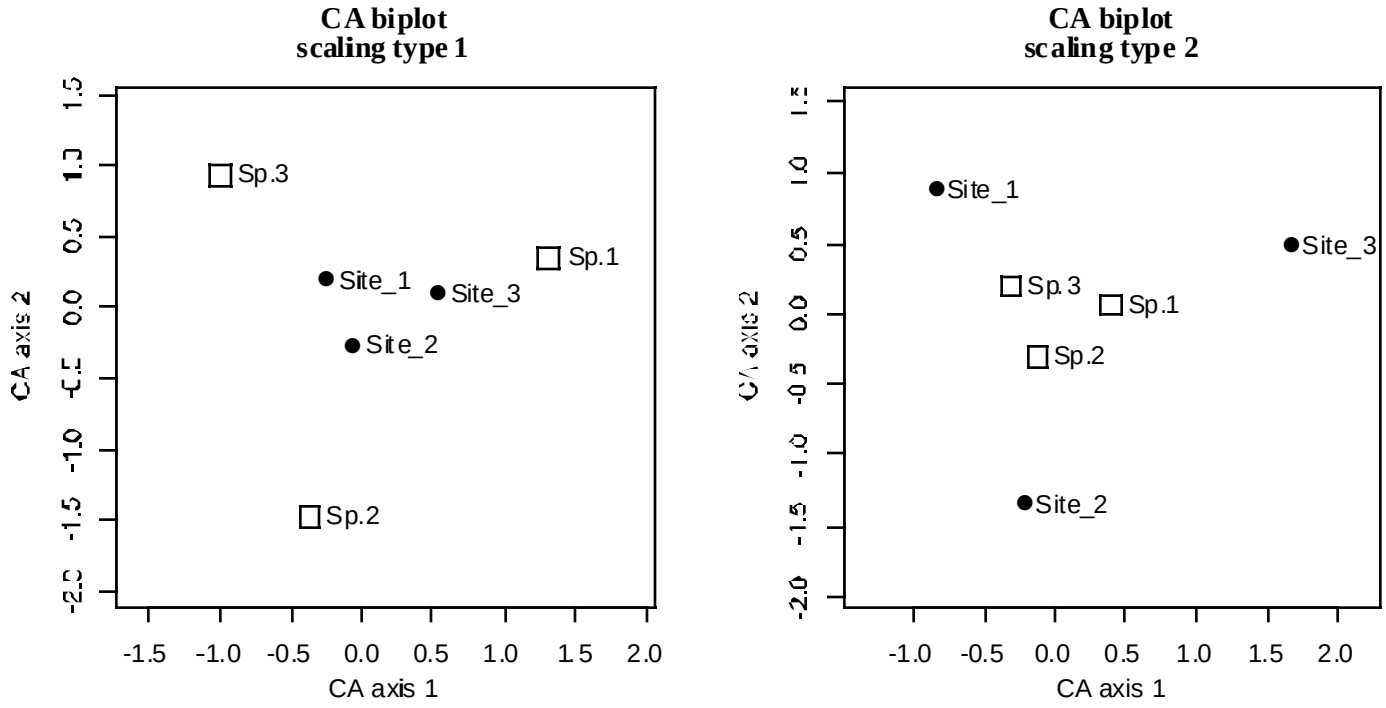
Eigenvalues: $\lambda_1 = 0.096$, $\lambda_2 = 0.041$ Matrix of eigenvalues: $\Lambda = \begin{bmatrix} 0.096 & 0 \\ 0 & 0.041 \end{bmatrix}$

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

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

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

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



Calculation details

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

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

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

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

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

Biplot, scaling type 1: plot **F** for sites, **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.