Chapter 15. Principal Component Analysis (PCA)

1. Ecological data are often multivariate – a site is described by many biotic or abiotic variables. Sometimes, biotic and abiotic measures are used together to analyze the association between sites, but more often they are used separately, such as the examples shown below.

> dune

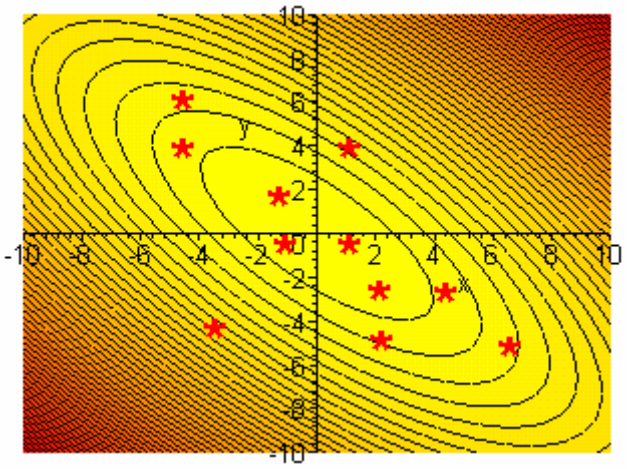
	Belper	Empnig	Junbuf	Junart	Airpra	Elepal	Rumace	Viclat	Brarut	Ranfla	Cirarv
X2	3	0	0	0	0	0	0	0	0	0	0
X13	0	0	3	0	0	0	0	0	0	2	0
X4	2	0	0	0	0	0	0	0	2	0	2
X16	0	0	0	3	0	8	0	0	4	2	0
X6	0	0	0	0	0	0	6	0	6	0	0
X1	0	0	0	0	0	0	0	0	0	0	0
X8	0	0	0	4	0	4	0	0	2	2	0
X5	2	0	0	0	0	0	5	0	2	0	0
X17	0	0	0	0	2	0	0	0	0	0	0
X15	0	0	0	3	0	5	0	0	4	2	0
X10	2	0	0	0	0	0	0	1	2	0	0
X11	0	0	0	0	0	0	0	2	4	0	0
X9	0	0	4	4	0	0	2	0	2	0	0
X18	2	0	0	0	0	0	0	1	6	0	0
X3	2	0	0	0	0	0	0	0	2	0	0
X20	0	0	0	4	0	4	0	0	4	4	0
X14	0	0	0	0	0	4	0	0	0	2	0
X19	0	2	0	0	3	0	0	0	3	0	0
X12	0	0	4	0	0	0	2	0	4	0	0
X7	0	0	2	0	0	0	3	0	2	0	0

> dune.env

	A1	Moisture	Management	Use Manure
1	3.5	1	BF Haypastu	2
2	6.0	5	SF Haypastu	3
3	4.2	2	SF Haypastu	4
4	5.7	5	SF Pasture	3
5	4.3	1	HF Haypastu	2
6	2.8	1	SF Haypastu	4
7	4.2	5	HF Pasture	3
8	6.3	1	HF Hayfield	2
9	4.0	2	NM Hayfield	0
10	11.5	5	NM Haypastu	0
11	3.3	2	BF Hayfield	1

12	3.5	1	BF	Pasture	1
13	3.7	4	HF	Hayfield	1
14	4.6	1	NM	Hayfield	0
15	4.3	2	SF	Haypastu	4
16	3.5	5	NM	Hayfield	0
17	9.3	5	NM	Pasture	0
18	3.7	5	NM	Hayfield	0
19	5.8	4	SF	Haypastu	2
20	2.8	1	HF	Pasture	3

2. The distribution of sites in multi-dimensional spaces, showing two abiotic variables. The location of each site is determined by the values of the two variables.



$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}, \qquad \boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \begin{bmatrix} \sigma_{11} & -\sigma_{12} \\ -\sigma_{21} & \sigma_{22} \end{bmatrix}$$

$$(\boldsymbol{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = \frac{1}{1-\rho_{12}^2} \left[\left(\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}} \right)^2 + \left(\frac{x_2 - \mu_2}{\sqrt{\sigma_{22}}} \right)^2 - 2\rho_{12} \left(\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}} \right) \left(\frac{x_2 - \mu_2}{\sqrt{\sigma_{22}}} \right) \right]$$

This is also called Mahalanobis distance

3. In order to study the relationship between sites in a multi-dimensional space, we need certain quantities to measure the distance between sites. Numerous indices have been developed to measure distance/similarity/dissimilarity/association/resemblance

between sites. See Chapter 7 of Legendre brothers' Numerical Ecology (1998, Elsevier).

Two distance measures are most important in multivariate analysis.

(1) The Euclidean distance (metric distance)

$$D(1,2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

This distance is usually applied to environmental measures. It is the distance on which PCA is based on.

In multi-dimensional space, the Euclidean distance is:

$$D(1,2) = \sqrt{\sum_{j=1}^{p} (x_{1j} - x_{2j})^2}$$

(2) χ^2 **distance** – is used to calculate distance among sites using species abundance or other frequency data. No negative values are allowed in the data.

$$D(1,2) = \sqrt{\sum_{j=1}^{p} \frac{1}{x_{+j}/x_{++}} \left(\frac{x_{1j}}{x_{1+}} - \frac{x_{2j}}{x_{2+}} \right)^2} = x_{++} \sqrt{\sum_{j=1}^{p} \frac{1}{x_{+j}} \left(\frac{x_{1j}}{x_{1+}} - \frac{x_{2j}}{x_{2+}} \right)^2}$$

where x_{1j} is cell frequency in a frequency table for site 1, while x_{2j} is cell frequency in a frequency table for site 2. x_{1+} and x_{2+} are row totals. x_{+j} is column total. x_{++} is the total sum of the frequency.

$$[x_{i+}]$$

$$\boldsymbol{x} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 4 & 4 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ 8 \end{bmatrix} \longrightarrow [\boldsymbol{x_{ij}/x_{i+}}] = \begin{bmatrix} 0 & 0.5 & 0.5 \\ 1 & 0 & 0 \\ 0 & 0.5 & 0.5 \end{bmatrix}$$

$$[x_{+j}] = [1 \quad 5 \quad 5][11]$$

The χ^2 distance is

$$D(1,2) = \left[\frac{(0-1)^2}{1/11} + \frac{(0.5-0)^2}{5/11} + \frac{(0.5-0)^2}{5/11} \right]^{1/2} = 3.477$$

$$D(1,3) = \left[\frac{(0-0)^2}{1/11} + \frac{(0.5-0.5)^2}{5/11} + \frac{(0.5-0.5)^2}{5/11} \right]^{1/2} = 0$$

This distance is used in correspondence analysis when computing the association between species or between sites. More generally, it is used for computing the association between the rows or columns of a contingency table. This measure has no upper limit.

(3) Other important association measures include:

Jaccard's coefficient: $S = \dfrac{a}{a+b+c}$

Sørensen's coefficient: $S = \dfrac{2a}{2a+b+c}$

where *a* is the number of species present in both sites, b is the # of species present in *B* but absent from *A*, etc.

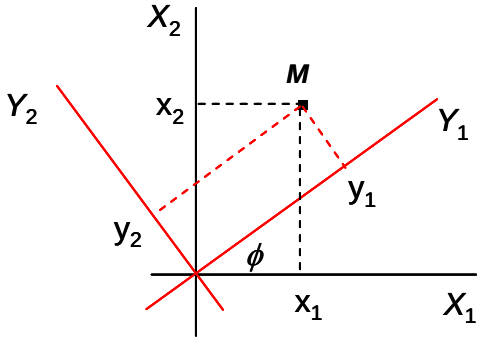
Site *B* \ Site *A*	Presence	Absence
Presence	*a*	*b*
Absence	*c*	*d*

4. PCA is a way of coordinate system rotation

The point *M* has coordinates (x_1, x_2) at the original X_1-X_2 coordinate system. We rotate the *X* system into *Y* position. What are the new coordinates for point *M*?

$$y_1 = x_1 \cos\phi + x_2 \sin\phi$$

$$y_2 = -x_1 \sin\phi + x_2 \cos\phi$$



We re-express this coordinate transformation in matrix notation.

$$\begin{bmatrix} y_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

The principle component analysis is to find a transformed coordinate system (i.e., an appropriate angle ϕ) in which the first axis (Y_1) captures the largest variation in the data and the second axis (Y_2) captures the second largest variation. Furthermore, Y_1 and Y_2 are orthogonal.

5. This can be quite easily achieved using eigenvalue analysis on dispersion (or correlation) matrix of the variables of interest. For the purpose of illustration, let's look at a simple example (see Legendre's Numerical Ecology, page 392).

$$\boldsymbol{X} = \begin{bmatrix} 2 & 1 \\ 3 & 4 \\ 5 & 0 \\ 7 & 6 \\ 9 & 2 \end{bmatrix}.$$

 # This data is named example16.dat in R.

(1) Standardize this matrix by centralization:

$$\boldsymbol{X} - \overline{\boldsymbol{X}} = \begin{bmatrix} 2-5.2 & 1-2.6 \\ 3-5.2 & 4-2.6 \\ 5-5.2 & 0-2.6 \\ 7-5.2 & 6-2.6 \\ 9-5.2 & 2-2.6 \end{bmatrix} = \begin{bmatrix} -3.2 & -1.6 \\ -2.2 & 1.4 \\ -0.2 & -2.6 \\ 1.8 & 3.4 \\ 3.8 & -0.6 \end{bmatrix}$$

(2) The dispersion matrix of X is the variance-covariance matrix, calculated as:

$$\boldsymbol{S} = \frac{1}{n-1}\left[\boldsymbol{X} - \overline{\boldsymbol{X}}\right]'\left[\boldsymbol{X} - \overline{\boldsymbol{X}}\right] = \begin{bmatrix} 8.2 & 1.6 \\ 1.6 & 5.8 \end{bmatrix}$$

R-Code: t(XX)%*%XX/4 **# XX is the centralized matrix *X*-*X*′.**

(3) Compute eigenvalues and eigenvectors for the dispersion matrix $\boldsymbol{S}$.

$\boldsymbol{Su} = \lambda\boldsymbol{u} \longrightarrow (\boldsymbol{S} - \lambda\boldsymbol{I})\boldsymbol{u} = \boldsymbol{0}$. A nontrivial solution $\boldsymbol{u} \neq \boldsymbol{0}$ requires that $|\boldsymbol{S} - \lambda\boldsymbol{I}| = 0$.

Substitute $\boldsymbol{S}$ into this determinant:

$$|\boldsymbol{S} - \lambda\boldsymbol{I}| = \left\| \begin{bmatrix} 8.2 & 1.6 \\ 1.6 & 5.8 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right\| = 0.$$

Solve this determinant, we have $\lambda_1 = 9$ and $\lambda_2 = 5$. Eigenvalues capture the variation in the data. The importance of axes I and II is entirely measured by the proportion of the eigenvalues for each axis. For example, the first principal axis describes $\frac{9}{14} = 64.29\%$ of variation in the data, and the second axis describes $\frac{5}{14} = 35.71\%$ of the variation.

(4) For each eigenvalue, there are many eigenvectors (not unique) that meet condition: $(\boldsymbol{S} - \lambda_1 \boldsymbol{I})\boldsymbol{u} = \boldsymbol{0}$. For example, for $\lambda_1 = 9$, we can have

$$\left(\begin{bmatrix} 8.2 & 1.6 \\ 1.6 & 5.8 \end{bmatrix} - \begin{bmatrix} 9 & 0 \\ 0 & 9 \end{bmatrix}\right)\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} -0.8 & 1.6 \\ 1.6 & -3.2 \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} -0.8u_1 + 1.6u_2 \\ 1.6u_1 - 3.2u_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

This leads to one equation: $u_1 = 2u_2$. Given any value of u_2, there is one u_1, therefore, there is no unique solution. Because u_1 and u_2 are proportional, it does not matter which (u_1, u_2) pair we use. For simplicity, we can scale (u_1, u_2) so that $u_1^2 + u_2^2 = 1$ (i.e., $\boldsymbol{u}'\boldsymbol{u} = 1$). Therefore, under this condition we have eigenvector corresponding to eigenvalue of $\lambda_1 = 9$:

$$\boldsymbol{u}_1 = \begin{cases} u_1 = 0.8944 \\ u_2 = 0.4472 \end{cases}$$

Similarly, we have eigenvector for eigenvalue $\lambda_2 = 5$:

$$\boldsymbol{u}_2 = \begin{cases} u_1 = -0.4472 \\ u_2 = 0.8944 \end{cases}$$ (note: it is not a problem if u_1 and u_2 switch the sign.)

The computation of eigenvalues and eigenvectors of $\boldsymbol{S}$ can be easily implemented using R-Code:

> **eigen(s)**

(5) Eigenvectors are called loadings in PCA, which are the coefficients in the linear principle axes (they are transformed axes):

Axis I: $y_{1i} = 0.8944(x_{1i} - \bar{x}_{1i}) + 0.4472(x_{2i} - \bar{x}_{2i})$

Axis II: $y_{2i} = -0.4472(x_{1i} - \bar{x}_{1i}) + 0.8944(x_{2i} - \bar{x}_{2i})$

(6) The new coordinates calculated from y_{1i} and y_{2i} are called principal components. For example the principle components for the first data point are:

$y_{11} = 0.8944 \times (-3.2) + 0.4472 \times (-1.6) = -3.576$

$y_{21} = -0.4472 \times (-3.2) + 0.8944 \times (-1.6) = 0$

6. R implementation

There are a number of ways PCA analysis can be implemented using R. There are two major programs. One is **princomp** built in R, the other is a contributed package **vegan** specifically designed for analyzing ecological data (vegetation analysis). Let's look at **princomp** first.

7. **> example15.pca=princomp(example15.dat)**

> summary(example15.pca)

Importance of components:

Comp.1 Comp.2

Standard deviation 2.6832816 2.0000000

Proportion of Variance 0.6428571 0.3571429 # eigenvalue proportions

Cumulative Proportion 0.6428571 1.0000000

> example15.pca$loadings

Loadings:

Comp.1 Comp.2

x1 0.894 0.447 # eigenvectors

x2 0.447 -0.894

Comp.1 Comp.2

SS loadings 1.0 1.0

Proportion Var 0.5 0.5

Cumulative Var 0.5 1.0

> example15.pca$scores # principal components

	Comp.1	Comp.2
site1	-3.577709	-4.090695e-16
site2	-1.341641	-2.236068e+00
site3	-1.341641	2.236068e+00
site4	3.130495	-2.236068e+00
site5	3.130495	2.236068e+00

> biplot(example15.pca) # plot outputs

8. **bioplot** – The biplot allows one to represent both the original variables and the transformed observations on the principal components axes so that you can graphically view the relationship between those original variables and the principal components.
   Interpretation of the biplot: The *x*-axis represents the scores for the first principal component, the y-axis the scores for the second principal component. The original variables are represented by arrows (**princomp**, no arrows in **vegan**) which graphically indicate the proportion of the original variance explained by the first two principle components.

   Note: Sites-descriptor relationships are not interpreted based on their proximity, but on orthogonal projections of the sites on the descriptor axes.

9. Assessing correlation between principal axes and original descriptors
   There are several ways to assess the contribution of each descriptor to the principal components. Two common practices are:

(1) Examine the sign of each loading. Positive loadings and negative loadings may represent two opposite environmental variables (e.g., wet versus dry, nutrient rich versus nutrient poor).
(2) Correlate the principal components with each original descriptor.

10. Let's try a real data: tiantong.dat

11. Let's try **vegan**.

>example15.rda=rda(example15.dat)

> summary(example15.rda)

Call:

Partitioning of variance:

Total 14
Unconstrained 14

Eigenvalues, and their contribution to the variance

	PC1	PC2
lambda	9.0000	5
accounted for	0.6429	1

Scaling 2 for species and site scores
-- Species are scaled proportional to eigenvalues
-- Sites are unscaled: weighted dispersion equal on all dimensions

Species scores

	PC1	PC2

x1 1.9618 0.7311

x2 0.9809 -1.4622

[Eigenvectors are not the same as the outputs of **princomp** because they are not scaled. $1.9618^2 + 0.9809^2 = 4.8108$. Scaled 1.9618 by sqrt(4.8108) leads to the same eigenvectors as calculated manually in the above: $1.9618/\sqrt{4.8108} = 0.8944$, $0.9809/\sqrt{4.8108} = 0.4472$.]

Site scores (weighted sums of species scores)

	PC1	PC2
site1	-1.6312	1.215e-16
site2	-0.6117	-1.368e+00
site3	-0.6117	1.368e+00
site4	1.4273	-1.368e+00
site5	1.4273	1.368e+00

[The PCA scores are the same as those produced by **princomp** when the PC1 are multiplied by $\sqrt{4.8108} = 2.1933$, and PC2 are multiplied by $\sqrt{2.6726} = 1.6348$.

12. PCA using correlation matrix

What we have learned so far is PCA using covariance (dispersion) matrix (i.e., eigenvalue analysis is all based on this matrix). The alternative is PCA using correlation matrix. Both **princomp** and **vegan** have options for that.

> princomp(example15.dat, cor=T)

> rda(example15.dat, scale=T)

Important note: Correlation matrix makes each descriptor variable more homogeneous. It actually is the *z*-score standardized data.

The outputs of PCA-correlation matrix are different from those of PCA-covariance matrix because the distance between sites is differently defined. Here are two general rules whether covariance matrix or correlation matrix should be used:

(1) Try both, but choose the one which makes the interpretation of your results easier and more interesting.

(2) If one wants to cluster the sites in the reduced space, there are two questions you may want to ask yourself before deciding which matrices to use for PCA. Should the clustering be done with respect to the original descriptors, thus preserving their differences in magnitude? Or, should all descriptors contribute equally to the clustering of sites, independently of the variance exhibited by each one? In the second case, you should proceed from the correlation matrix.

13. Let's try data: **tiantong.dat**.

(1) Produce outputs

(2) Plot the sites in the first two principal axes. Plot *D* sites and *T* sites in different colors. (*D* – degenerated secondary forests, *T* – old growth forests with some degree of disturbances).