Introduction to multivariate analysis, ordination and PCA

Contents

Learning targets

- Explain the specifics of multivariate analysis

- List ordination methods and ability to select based on research question

- Explain the mathematical basis of and apply PCA

- Interpreting results from a PCA

Learning targets and study questions

- Explain the specifics of multivariate analysis

  - Why should you favour multivariate approaches for multivariate data?

  - Outline the differences to the univariate case when diagnosing multivariate outliers and normality.

- List ordination methods and ability to select based on research question

  - Outline the aim of ordination

  - Distinguish constrained and unconstrained ordination

  - Which criteria should guide the selection of an ordination method?

Learning targets and study questions

- Explain the mathematical basis of and apply PCA

  - Explain the variance-covariance matrix. What are eigenvalues and eigenvectors?

  - How do eigenvalues relate to the variance of a PC?

  - Outline criteria to determine the optimal number of PCs.

  - What is sparse PCA and how does it influence the evaluation of descriptor contribution to PCs?

- Interpreting results from a PCA

  - Explain biplots with respect to (a) correlation between variables and (b) relation between sites/species and variables.

  - How does scaling influence interpretation of a biplot?

  - Which objects from a PCA would be extracted as non-collinear predictors for a multiple regression analysis?

From univariate to multivariate statistics

	univariate	**multivariate**
Variables (vars.)	Single/multiple explanatory vars., single response var. y	Single/multiple explanatory vars., multiple response vars. $y_1, …, y_n$
Distribution of response	One-dimensional	n-dimensional
Data format	y is vector	$y_1, …, y_n$ constitute matrix
Example	Species richness and environmental variables	Taxa and environmental variables

Multivariate data analysis: Introduction

Some advantages of multivariate over univariate methods for analysing multivariate data

- Not all research questions can be answered with univariate statistical methods
  e.g. *What are the most important environmental variables determining community composition?*

- Multivariate methods allow for dimension reduction and visualisation of multidimensional data
  e.g. *Ordination, Cluster dendrogram*

- Joint (multivariate) analysis can reduce noise and increase power of statistical tests

Multivariate approaches in R

Available techniques and developments: CRAN Task View

CRAN Task View: Multivariate Statistics

Maintainer: Paul Hewson

Contact: Paul.Hewson at plymouth.ac.uk

Version: 2014-12-18

Base R contains most of the functionality for classical multivariate analysis, somewhere. There are a large number of packages on CRAN which extend this methodology, a brief overview is given below. Application-specific uses of multivariate statistics are described in relevant task views, for example whilst principal components are listed here, ordination is covered in the Environmetrics task view. Further information on supervised classification can be found in the MachineLearning task view, and unsupervised classification in the Cluster task view.

The packages in this view can be roughly structured into the following topics. If you think that some package is missing from the list, please let me know.

Visualising multivariate data

- *Graphical Procedures:* A range of base graphics (e.g. `pairs()` and `coplot()`) and lattice functions (e.g. `xyplot()` and `splom()`) are useful for visualising pairwise arrays of 2-dimensional scatterplots, clouds and 3-dimensional densities. `scatterplot.matrix` in the car provides usefully enhanced pairwise scatterplots. Beyond this, scatterplot3d provides 3 dimensional scatterplots, aplpack provides bagplots and `spin3R()`, a function for rotating 3d clouds. misc3d, dependent upon rgl, provides animated functions within R useful for visualising densities. YaleToolkit provides a range of useful visualisation techniques for multivariate data. More specialised multivariate plots include the following: `faces()` in aplpack provides Chernoff's faces; `parcoord()` from MASS provides parallel coordinate plots; `stars()` in graphics provides a choice of star, radar and cobweb plots respectively. `mstree()` in ade4 and `spantree()` in vegan provide minimum spanning tree functionality. calibrate supports biplot and scatterplot axis labelling. geometry, which provides an interface to the qhull library, gives indices to the relevant points via `convexhulln()`. ellipse draws ellipses for two parameters, and provides `plotcorr()`, visual display of a correlation matrix. denpro provides level set trees for multivariate visualisation. Mosaic plots are available via `mosaicplot()` in graphics and `mosaic()` in vcd that also contains other visualization techniques for multivariate categorical data. gclus provides a number of cluster specific graphical enhancements for scatterplots and parallel coordinate plots See the links for a reference to GGobi. rggobi interfaces with GGobi. xgobi interfaces to the XGobi and XGvis programs which allow linked, dynamic multivariate plots as well as projection pursuit. Finally, iplots allows particularly powerful dynamic interactive graphics, of which interactive parallel co-ordinate plots and mosaic plots may be of great interest. Seriation methods are provided by seriation which can reorder matrices and dendrograms.
- *Data Preprocessing:* `summarize()` and `summary.formula()` in Hmisc assist with descriptive functions; from the same package `varclus()` offers variable clustering while `dataRep()` and `find.matches()` assist in exploring a given dataset in terms of representativeness and finding matches. Whilst `dist()` in base and `daisy()` in cluster provide a wide range of distance measures, proxy provides a framework for more distance measures, including measures between matrices. simba provides functions for dealing with presence / absence data including similarity matrices and reshaping.

Hypothesis testing

- ICSNP provides Hotellings T2 test as well as a range of non-parametric tests including location tests based on marginal ranks, spatial median and spatial signs computation, estimates of shape. Non-parametric two sample tests are also available from cramer and spatial sign and rank tests to investigate location, sphericity and independence are available in SpatialNP.

Multivariate distributions

- *Descriptive measures:* `cov()` and `cor()` in stats will provide estimates of the covariance and correlation matrices respectively. ICSNP offers several descriptive measures such as `spatial.median()` which provides an estimate of the spatial median and further functions which provide estimates of scatter. Further robust methods are provided such as `cov.rob()` in MASS which provides robust estimates of the variance-covariance matrix by minimum volume ellipsoid, minimum covariance determinant or classical product-moment. covRobust provides robust covariance estimation via nearest neighbor variance estimation. robustbase provides robust covariance estimation via fast minimum covariance determinant with

Multivariate outlier checking

Multivariate and univariate exploration/diagnosis similar for e.g. multicollinearity and transformation, but differ for checking of e.g. outliers and distributional assumptions
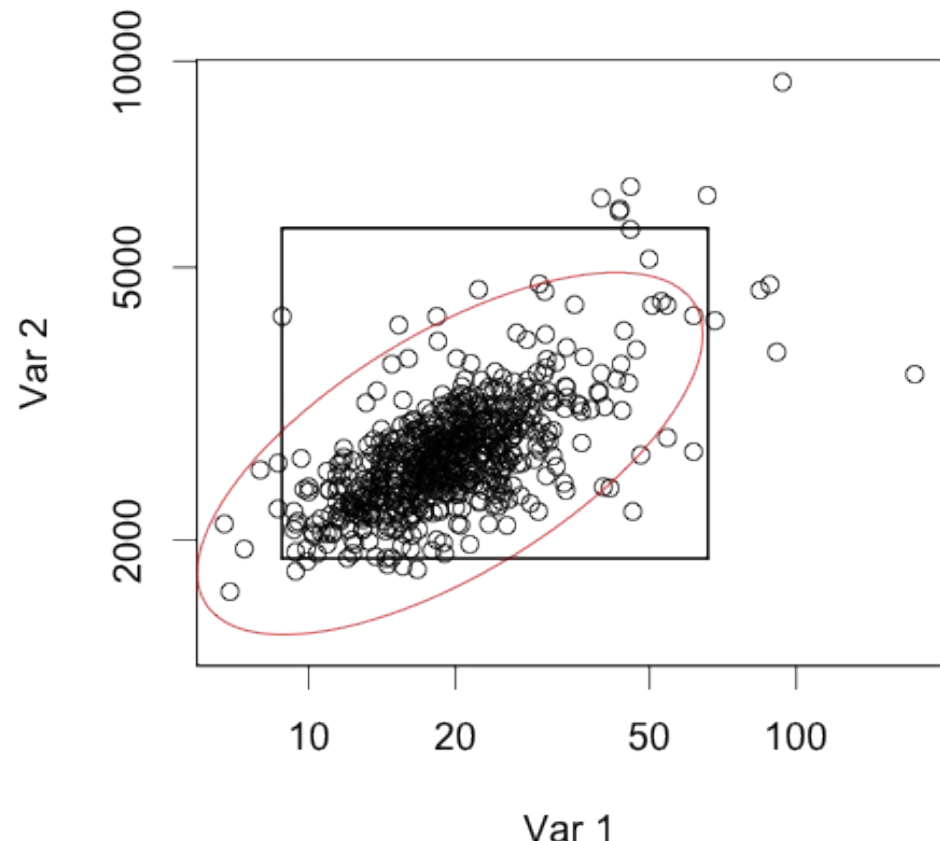
<u>Multivariate outliers</u>
use joint multivariate distribution of variables to find outliers (i.e. extreme points from centre of multivariate sample)

Outside of box:
Outliers, considering both variables individually (99% quantile of each variable)
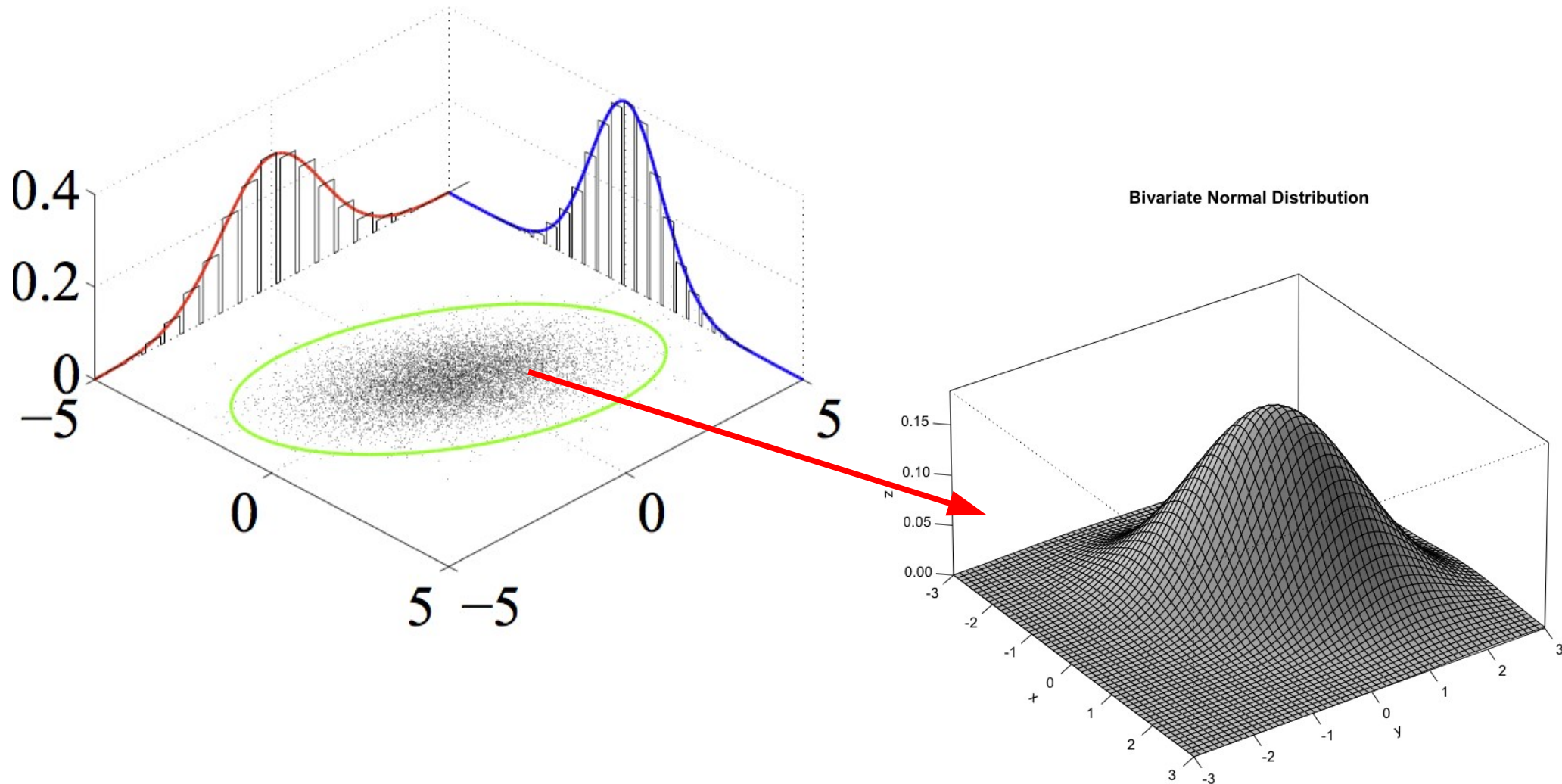
Outside of ellipse:
Outliers, considering both variables jointly (99% quantile using Mahalanobis distance)
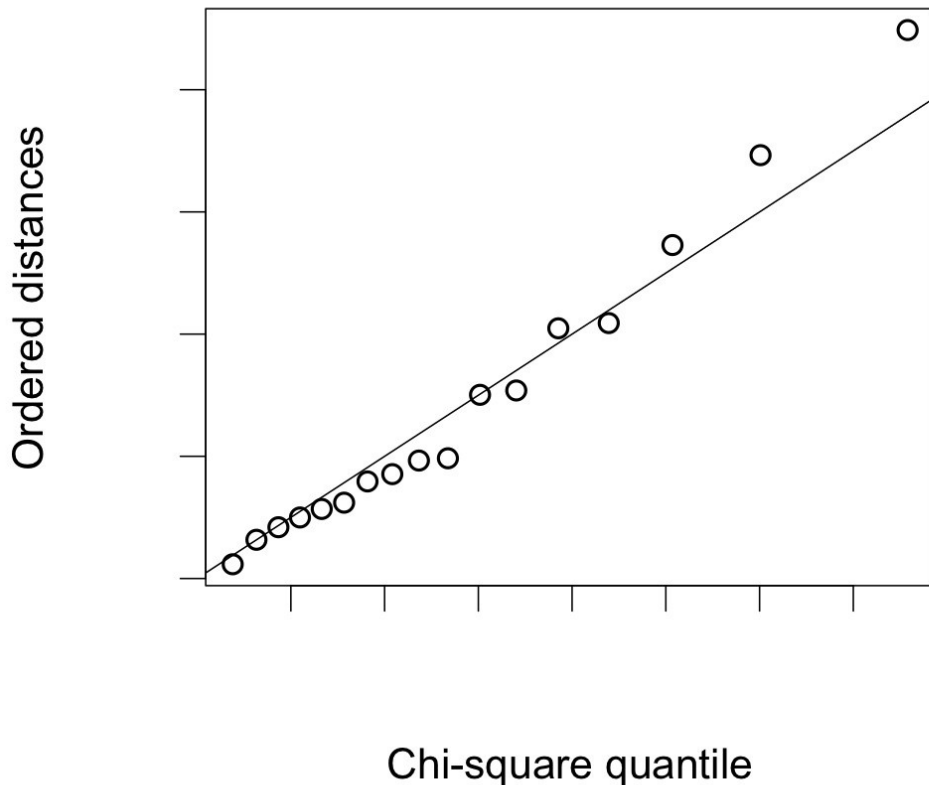
Multivariate normal distribution

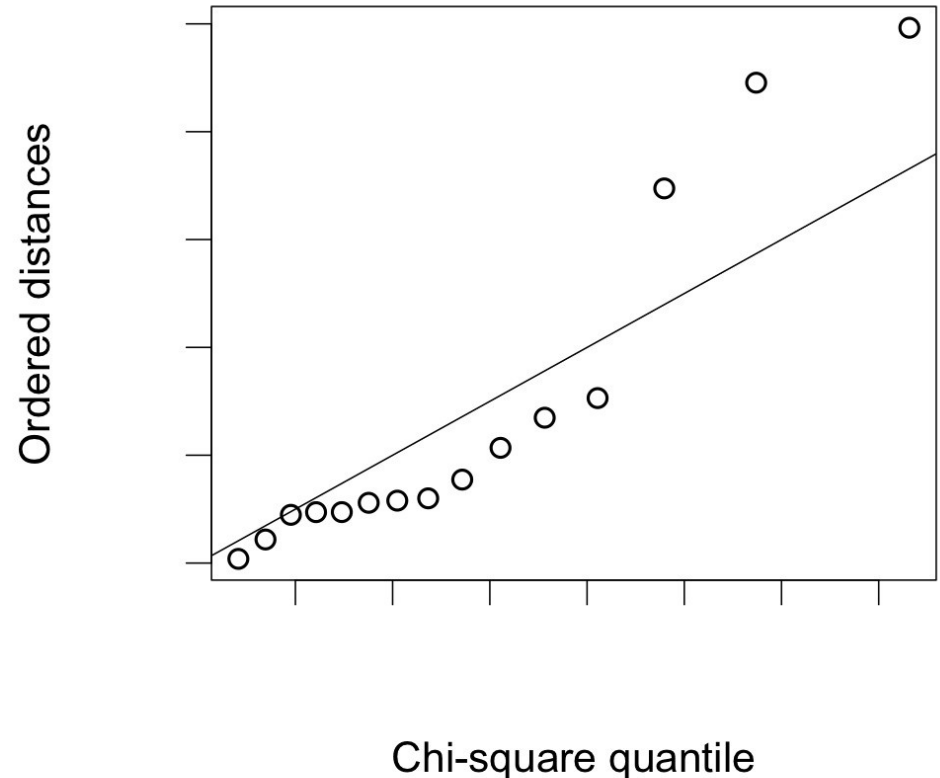Multivariate normal distribution is evaluated instead of univariate normal distribution

Bivariate Normal Distribution

http://commons.wikimedia.org/wiki/File:Multivariate_normal_sample.svg

Multivariate normal distribution

Visual check of multivariate normality with *q-q* plots for sample Mahalanobis distances (to centroid) and theoretical quantiles from the χ^2 distribution

Q-Q plot with data sampled from a multivariate normal distribution

Ordered distances

Chi-square quantile

Q-Q plot with data sampled from a exponential distribution

Ordered distances

Chi-square quantile

Covariance matrix

For a data matrix **Y** containing the variables $y_1,...,y_p$

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \\ \vdots \\ \mathbf{y}_i' \\ \vdots \\ \mathbf{y}_n' \end{pmatrix} = \text{(units)} \begin{matrix} 1 \\ 2 \\ \\ i \\ \\ n \end{matrix} \begin{pmatrix} \overset{1}{y_{11}} & \overset{2}{y_{12}} & \cdots & \overset{j}{y_{1j}} & \cdots & \overset{p}{y_{1p}} \\ y_{21} & y_{22} & \cdots & y_{2j} & \cdots & y_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ y_{i1} & y_{i2} & \cdots & y_{ij} & \cdots & y_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nj} & \cdots & y_{np} \end{pmatrix}$$

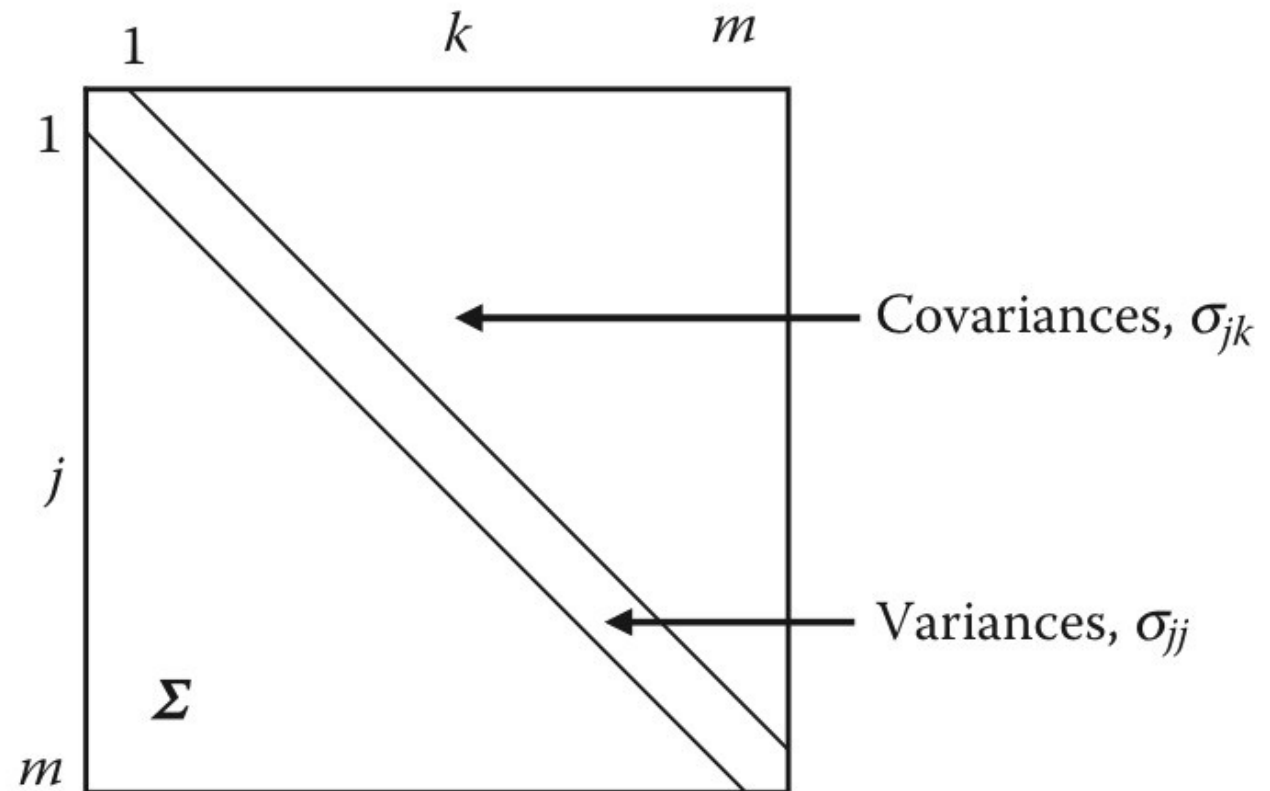the sample covariance matrix for these variables is **S** (Sigma)

$$\mathbf{S} = (s_{jk}) = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix}$$

is obtained by: $\quad s_{jk} = \dfrac{1}{n-1} \sum_{i=1}^{n} (y_{ij} - \overline{y}_j)(y_{ik} - \overline{y}_k)$

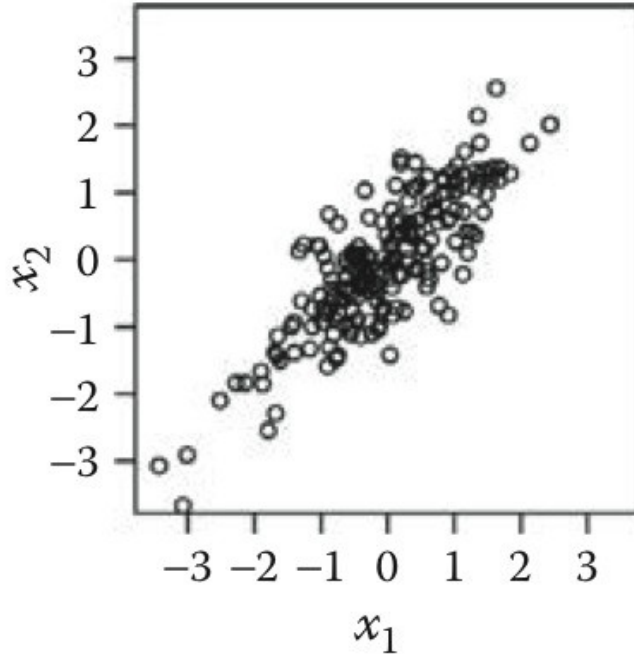taken from Rencher (2012): *Methods of multivariate Analysis*. 55f

Covariance matrix

In the diagonal, the equation simplifies to:

$$s_{jj} = \frac{1}{n-1}\sum_{i=1}^{n}(y_{ij}-\overline{y_j})^2$$

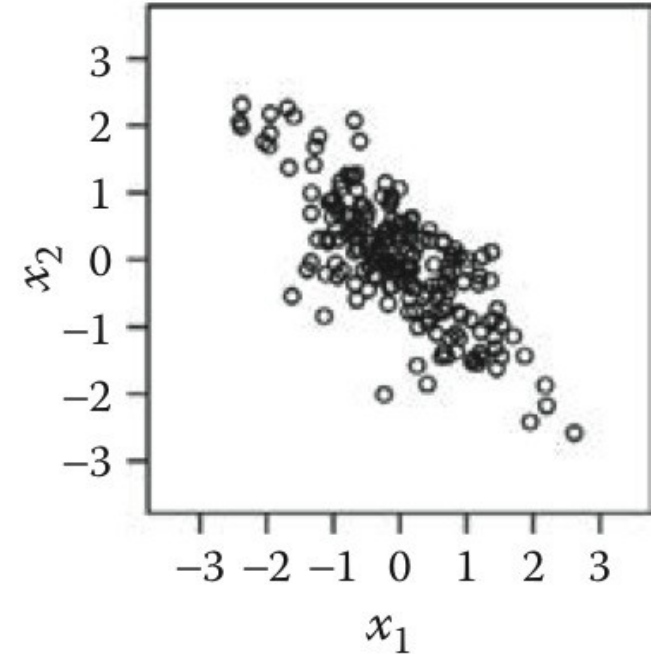# Covariance matrix for two variables

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix} \qquad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \boldsymbol{\Sigma}_3 = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 1 \end{pmatrix}$$

Multivariate distance from mean to observation is measured with Mahalanobis distance D_M:

$$D_M(x) \; = \; \sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^t \, \boldsymbol{S}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}$$

$\rightarrow D_M$ is distance of vector $\boldsymbol{x}$ from the mean vector $\boldsymbol{\mu}$ weighed by their covariance (given in sample covariance matrix $\boldsymbol{S}$)

13

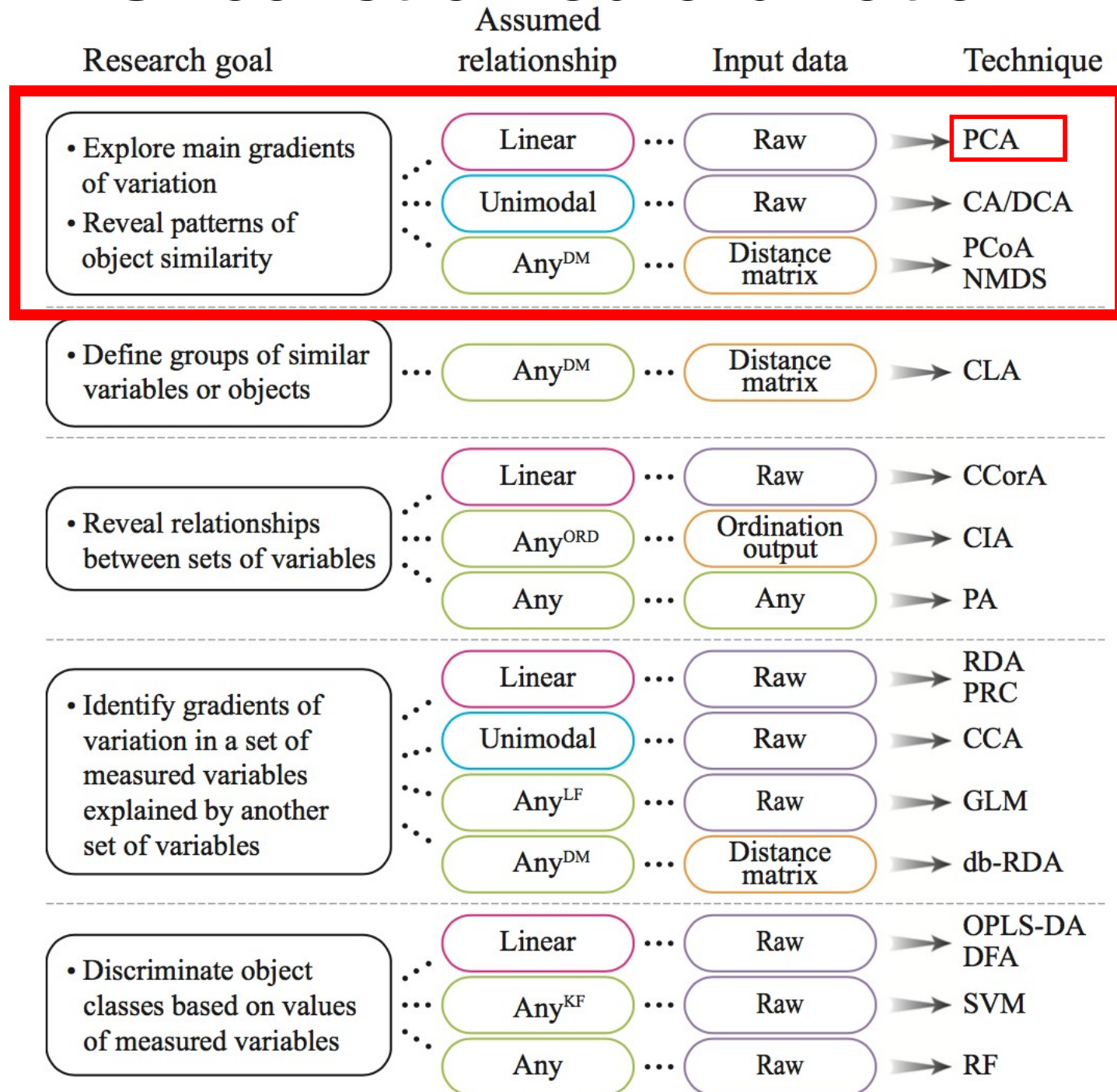# Introduction to multivariate analysis, ordination and PCA

Contents

1. Learning targets and introduction to multivariate analysis

2. **Overview ordination**

3. Introduction to PCA and mathematical background

4. PCA results and interpretation

5. PCA diagnosis and extensions, brief tutorial

Ordination: Introduction

- Extraction of new axes from high dimensional data that sequentially maximise the variance
  - → Dimension reduction (e.g. omission of axes that capture low amount of variance)
  - → Aggregation of variables into gradients
  - → Graphical representation in lower dimensionality

- Unconstrained ordination: extraction without consideration of variables outside of data set

- Constrained ordination: extraction of information, which can be explained by variables of a second data set
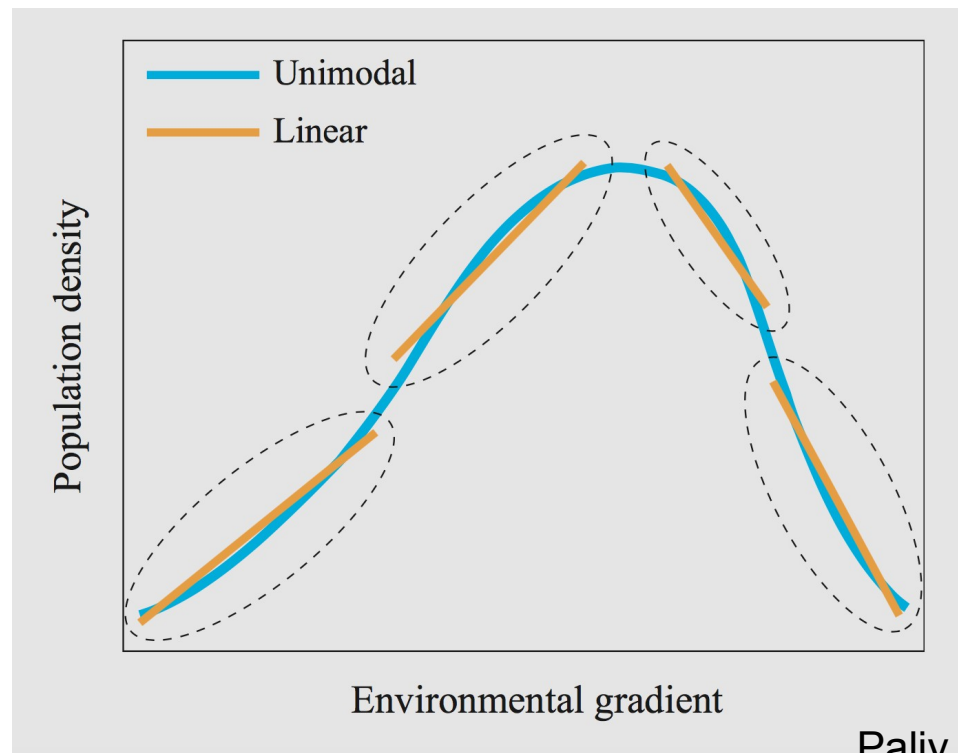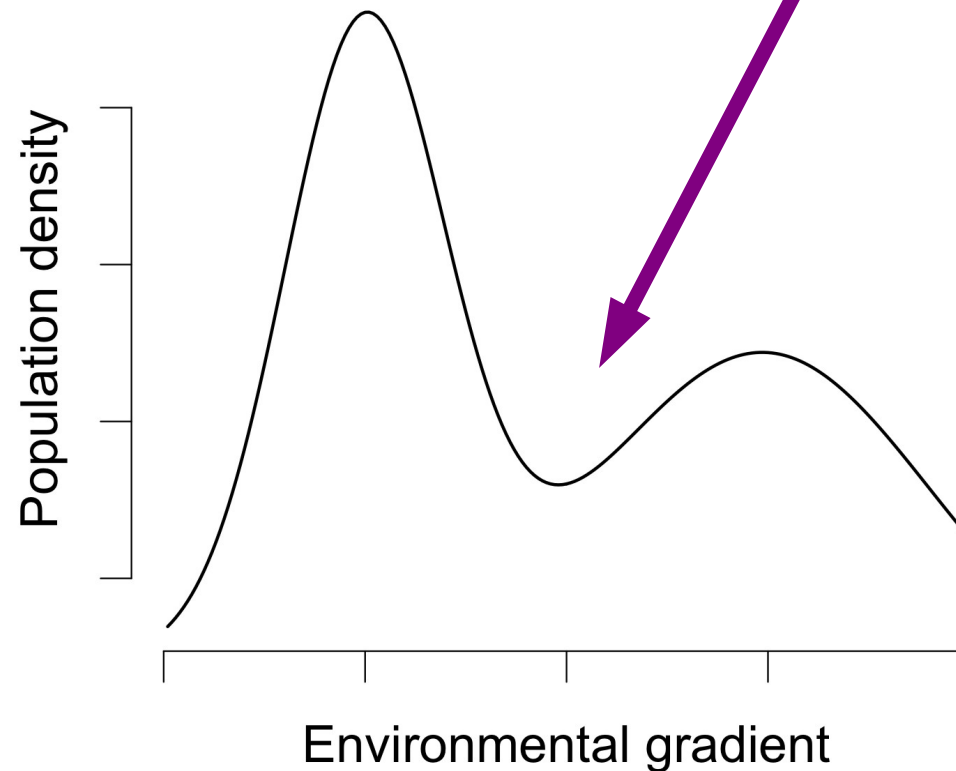
Unconstrained ordination

Research goal		Assumed relationship		Input data		Technique
• Explore main gradients of variation	⋮	Linear	⋯	Raw	→	PCA
• Reveal patterns of object similarity	⋮	Unimodal	⋯	Raw	→	CA/DCA
	⋮	Any[DM]	⋯	Distance matrix	→	PCoA NMDS
• Define groups of similar variables or objects	⋯	Any[DM]	⋯	Distance matrix	→	CLA
• Reveal relationships between sets of variables	⋮	Linear	⋯	Raw	→	CCorA
	⋮	Any[ORD]	⋯	Ordination output	→	CIA
	⋮	Any	⋯	Any	→	PA
• Identify gradients of variation in a set of measured variables explained by another set of variables	⋮	Linear	⋯	Raw	→	RDA PRC
	⋮	Unimodal	⋯	Raw	→	CCA
	⋮	Any[LF]	⋯	Raw	→	GLM
	⋮	Any[DM]	⋯	Distance matrix	→	db-RDA
• Discriminate object classes based on values of measured variables	⋮	Linear	⋯	Raw	→	OPLS-DA DFA
	⋮	Any[KF]	⋯	Raw	→	SVM
	⋮	Any	⋯	Raw	→	RF

Ordination: Overview

Shape of response	Linear	Unimodal	Any
Example for unconstrained method	PCA	CA	Distance-based: NMDS; GAM-based: UAO (U-VGAM)
Example for constrained method	RDA	CCA	Distance-based: db-RDA; GAM-based: CAO (RR-VGAM)

Paliy & Shankar 2016 *Mol Ecol*:1032

Ordination: Overview

Shape of response	Linear	Unimodal	Any
Example for unconstrained method	PCA	CA	Distance-based: NMDS; GAM-based: UAO (U-VGAM)
Example for constrained method	RDA	CCA	Distance-based: db-RDA; GAM-based: CAO (RR-VGAM)

Population density

Environmental gradient

Introduction to multivariate analysis, ordination and PCA

Contents

1. Learning targets and introduction to multivariate analysis

2. Overview ordination

3. **Introduction to PCA and mathematical background**

4. PCA results and interpretation

5. PCA diagnosis and extensions, brief tutorial
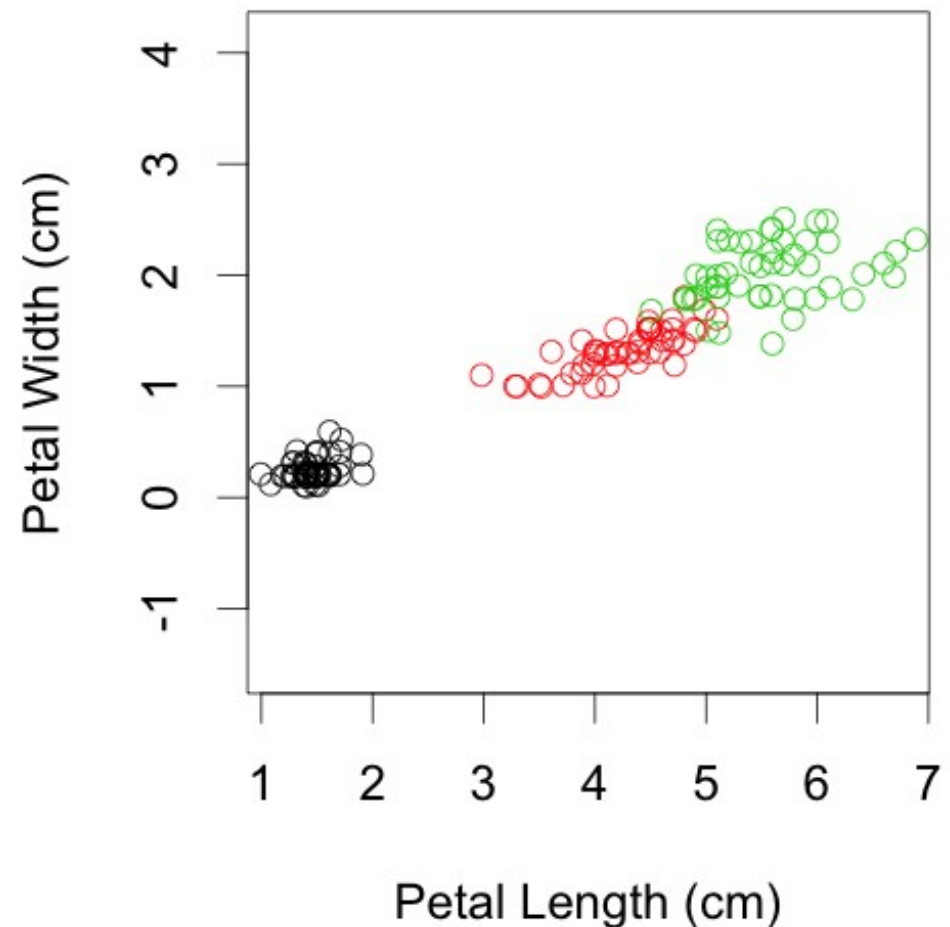
Principal Component Analysis

Example-based introduction

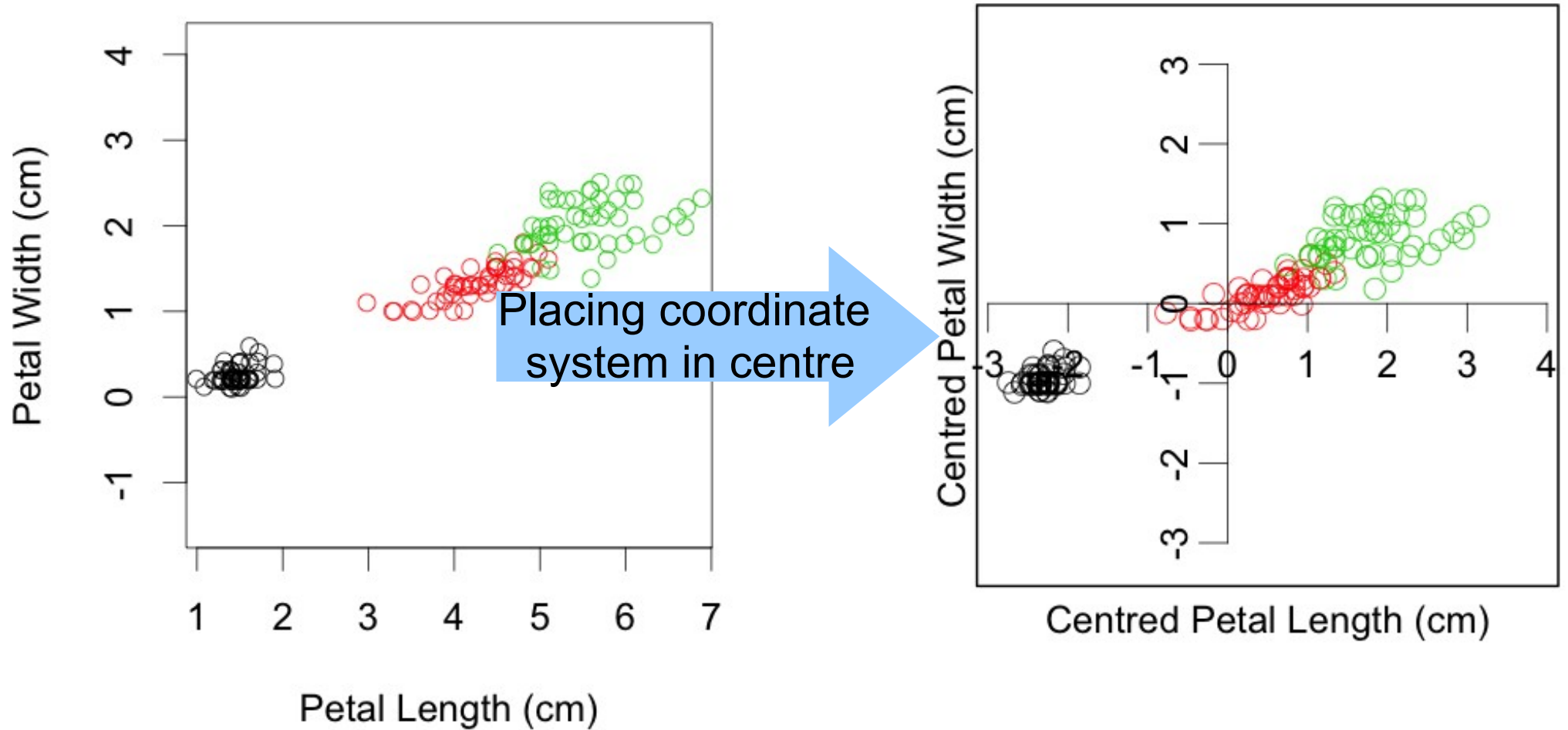Iris data set: sepal length & width, petal length & width for 50 flowers from 3 species of Iris
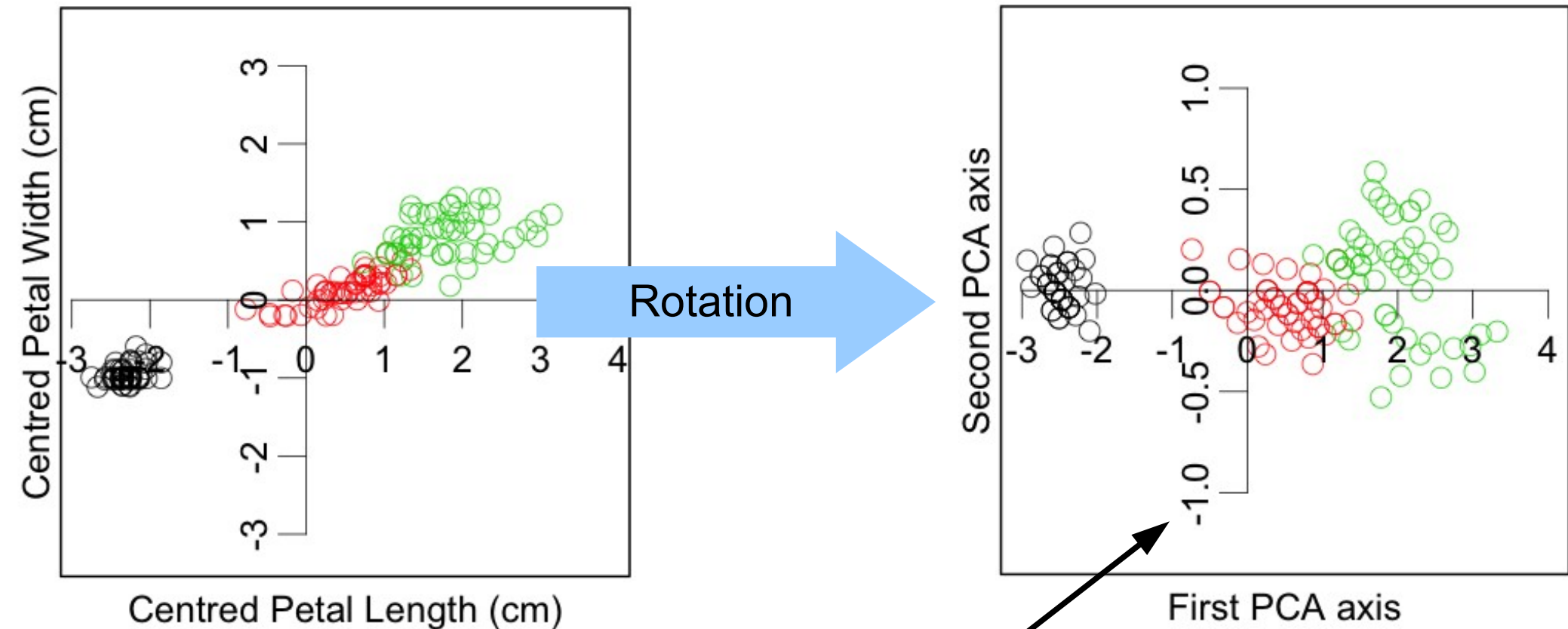

http://de.wikipedia.org/wiki/Schwertlilien

Aim: Represent as much variance as possible on first few axes

Introduction to PCA

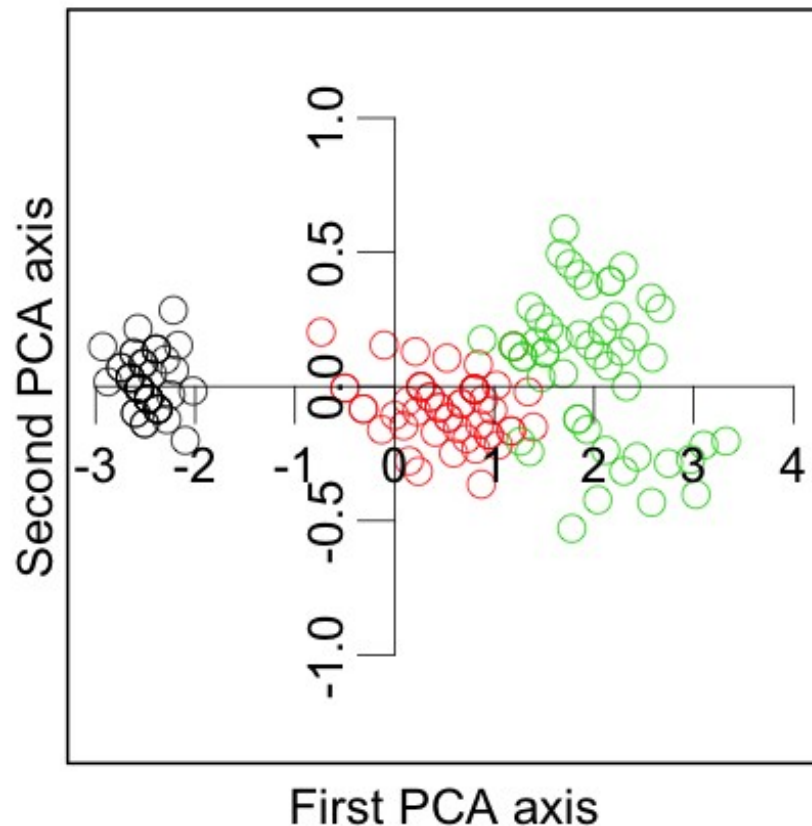

Placing coordinate system in centre

Introduction to PCA

Rescaling of axis based on variation along axis

Results of PCA

Variances
Petal Width 0.58
Petal Length 3.12

Results

Total Variation: 3.70

Importance of components:

	PC1	PC2
Eigenvalue	3.66	0.04
Proportion Explained	0.99	0.01
Cumulative Proportion	0.99	1.00

What is an Eigenvalue?

First axis explains 99% of variation!

Mathematical background of PCA

Country	Reading	Mathematics
DK	497	514
GR	474	447
I	487	457
P	470	454
S	516	510

Centering leads to

$$\widetilde{X} = \begin{pmatrix} 8.2 & 37.6 \\ -14.8 & -29.4 \\ -1.8 & -19.4 \\ -18.8 & -22.4 \\ 27.2 & 33.6 \end{pmatrix}$$

Search for first axis with maximum variation!

Example taken from Handl (2010)
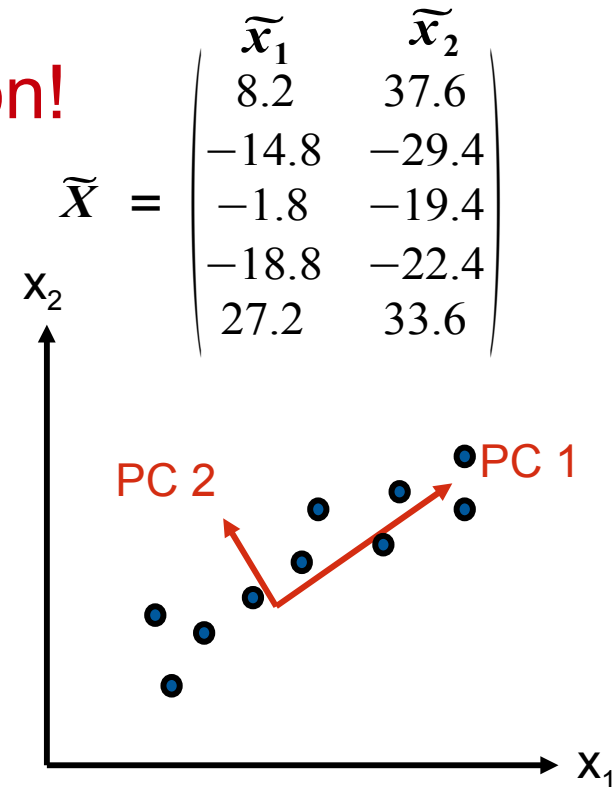
Mathematical background of PCA

Search for first axis with maximum variation!

Scores on new axis given as

$$\mathrm{PC}_1 = a_1 \widetilde{x}_1 + a_2 \widetilde{x}_2$$

and maximum variation means:

$$\max_{a_1, a_2} Var(a_1 \widetilde{x}_1 + a_2 \widetilde{x}_2)$$

$$\widetilde{X} = \begin{pmatrix} \widetilde{x}_1 & \widetilde{x}_2 \\ 8.2 & 37.6 \\ -14.8 & -29.4 \\ -1.8 & -19.4 \\ -18.8 & -22.4 \\ 27.2 & 33.6 \end{pmatrix}$$

Generalise problem: define a_1, a_2 as elements of vector **a** and likewise $\widetilde{x}_1, \widetilde{x}_2$ of matrix **X**: $\max_{a} Var(\boldsymbol{a}\,\widetilde{X})$

Trivial solution: choose high values for a_1, a_2, ... ,a_n

$\rightarrow$ introduce condition: $a_1^2 + a_2^2 + ... + a_n^2 = 1$

Mathematical background of PCA

Solve:

$$\max_{a} \; Var\left(\boldsymbol{a}\,\widetilde{X}\right) \; \text{with} \; \boldsymbol{a}^{t}\,\boldsymbol{a} \; = \; 1$$

This can be expressed as (see Handl 2010: 79)

$$\max_{a} \; \left(\boldsymbol{a}^{t}\,\Sigma\,\boldsymbol{a}\right) \; \text{with} \; \boldsymbol{a}^{t}\,\boldsymbol{a} \; = \; 1$$

Covariance matrix

Using the Lagrange function yields:

$$L(\boldsymbol{a},\lambda) \; = \; \boldsymbol{a}^{t}\,\Sigma\,\boldsymbol{a} - \lambda(\boldsymbol{a}^{t}\,\boldsymbol{a} - 1)$$

Eigenvalue problem

$$\frac{\partial\,L(\boldsymbol{a},\lambda)}{\partial\,\boldsymbol{a}} \; = \; 2\,\Sigma\,\boldsymbol{a} - 2\,\lambda\,\boldsymbol{a} \longrightarrow \boxed{\Sigma\,\boldsymbol{a} \; = \; \lambda\,\boldsymbol{a}}$$

$$\frac{\partial\,L(\boldsymbol{a},\lambda)}{\partial\,\lambda} \; = \; 1 - \boldsymbol{a}^{t}\,\boldsymbol{a}$$
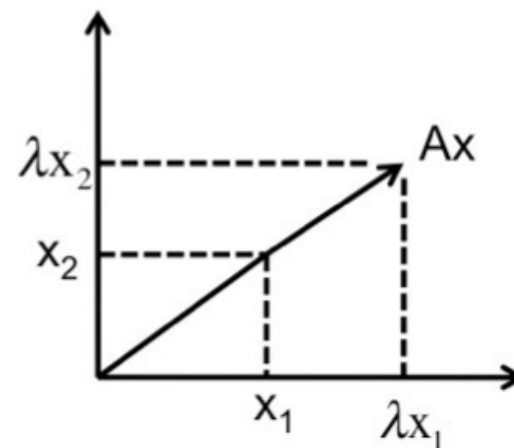
Mathematical basics II: Eigenvalues

Idea: Conversion of a matrix into a matrix with linear independent variables

$$\begin{pmatrix} a_{11} & ... & a_{1n} \\ ... & a_{22} & ... \\ a_{m1} & ... & a_{mn} \end{pmatrix} \xrightarrow{\text{Conversion}} \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_n \end{pmatrix}$$

Eigenvalue problem: $A\,x = \lambda\,x$ Eigenvector

Eigenvalue

Eigenvectors form canonical basis and are only stretched or shrunk by λ when multiplied with A.

Mathematical basics II: Eigenvalues

$$A\,x \;=\; \lambda\,x \quad \Leftrightarrow \quad A\,x - \lambda\,x \;=\; 0$$

$$\Leftrightarrow \quad (A - \lambda\,E)\,x \;=\; 0$$

Ignore trivial solution: $x = 0 \qquad \Rightarrow \quad A - \lambda\,E \;=\; 0$

$$\Leftrightarrow \quad \begin{pmatrix} a_{11} - \lambda_1 & ... & a_{1n} \\ ... & a_{22} - \lambda_2 & ... \\ a_{m1} & ... & a_{mn} - \lambda_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \quad$$ Homogeneous linear equation system (HLS)

HLS have only non-trivial solution if determinant = 0.

$$\Rightarrow \quad det(A - \lambda\,E) \;=\; 0 \quad \Leftrightarrow \quad |A - \lambda\,E| \;=\; 0$$

Example I: Calculation of Eigenvalues

Sample Variance-Covariance matrix from Pisa example:

$$S = \begin{pmatrix} 345.7 & 528.35 \\ 528.35 & 1071.30 \end{pmatrix}$$

Following $|A - \lambda E| = 0$ we obtain:

$$\left| \begin{pmatrix} 345.7 & 528.35 \\ 528.35 & 1071.30 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right| = 0 \Leftrightarrow$$
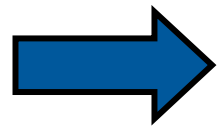
$$\left| \begin{pmatrix} 345.7 & 528.35 \\ 528.35 & 1071.30 \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} \right| = 0 \Leftrightarrow$$

$$\begin{vmatrix} 345.7 - \lambda & 528.35 \\ 528.35 & 1071.30 - \lambda \end{vmatrix} = 0 \Leftrightarrow$$

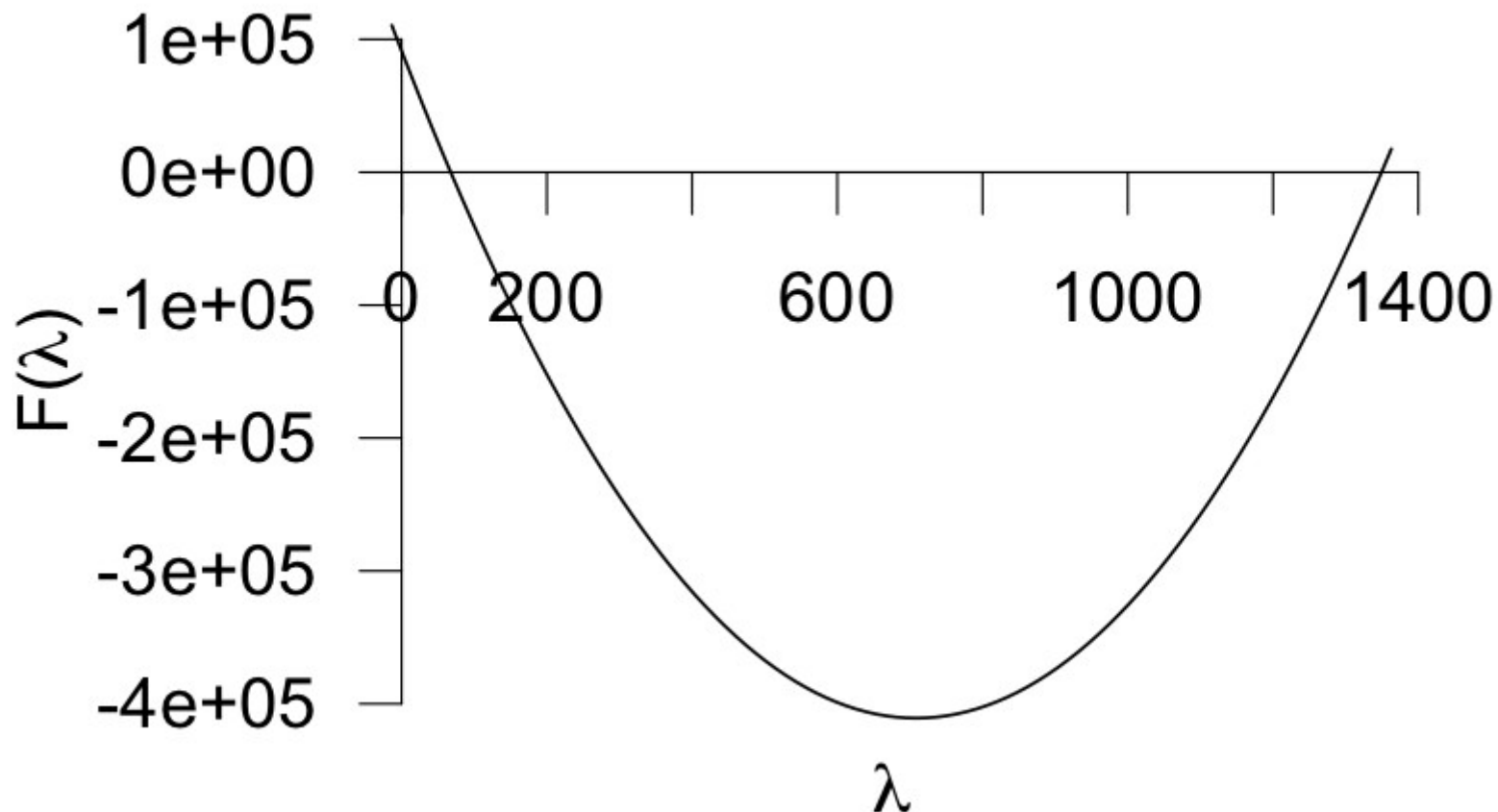$$(345.7 - \lambda)(1071.30 - \lambda) - 528.35^2 = 0$$

Example I: Calculation of Eigenvalues

$$(345.7 - \lambda)(1071.30 - \lambda) - 528.35^2 = 0$$

➡ Characteristical polynom

$$\lambda^2 - 1417\lambda + 91194.69 = 0$$

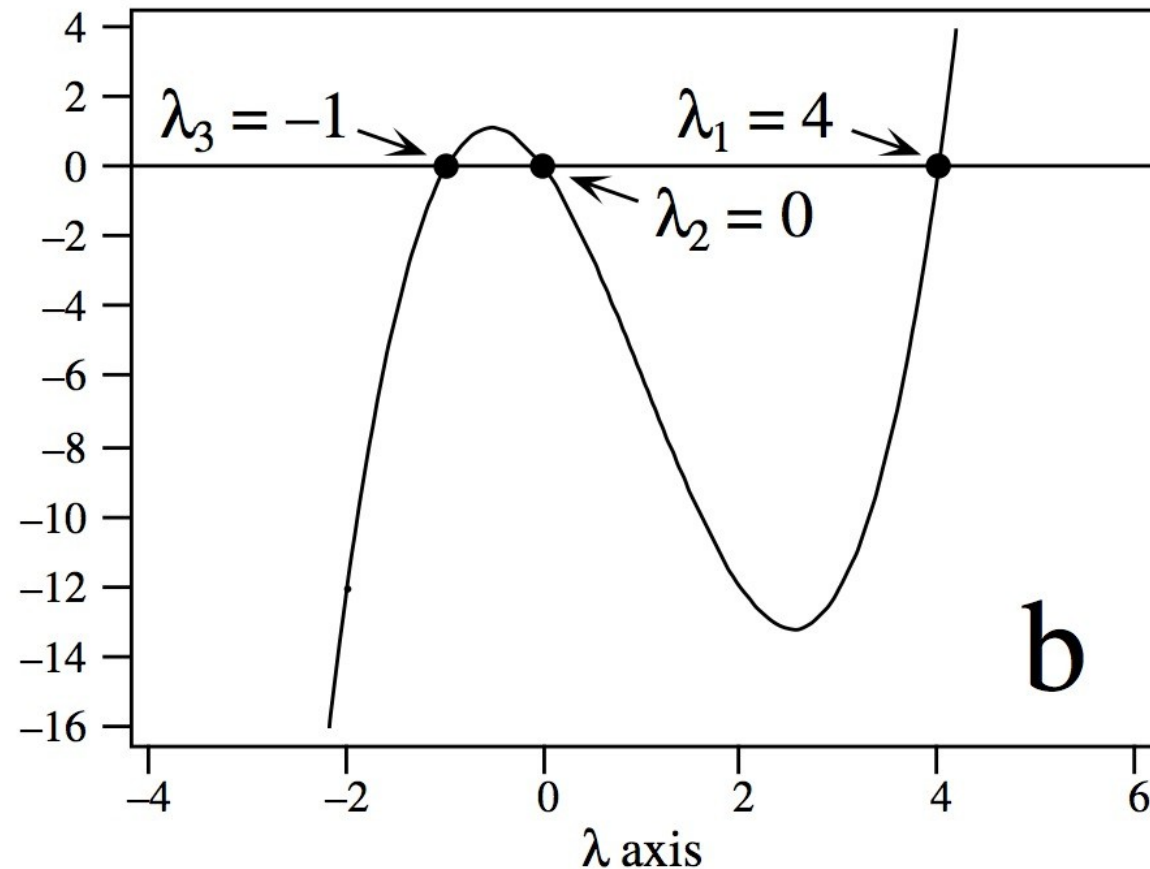# Example II: Calculation of Eigenvalues and -vectors

$$\begin{pmatrix} 1 & 3 & -1 \\ 0 & 1 & 2 \\ 1 & 4 & 1 \end{pmatrix}$$

EV calculation
following Sarrus

Characteristical polynom

$$\lambda^3 - 3\lambda^2 - 4\lambda = 0$$

Eigenvalues λ:
4, 0 and -1

Example II: Calculation of Eigenvalues and -vectors

Calculation of eigenvector for λ = 4

$$(A - \lambda E)x = 0$$

$$\left(\begin{pmatrix} 1 & 3 & -1 \\ 0 & 1 & 2 \\ 1 & 4 & 1 \end{pmatrix} - \lambda_1 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right) \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = 0 \Leftrightarrow \begin{pmatrix} 1-4 & 3 & -1 \\ 0 & 1-4 & 2 \\ 1 & 4 & 1-4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = 0$$

$$\begin{matrix} -3x_1 & 3x_2 & -x_3 & 0 \\ 0 & -3x_2 & 2x_3 & = 0 \\ x_1 & 4x_2 & -3x_3 & 0 \end{matrix} \Leftrightarrow \begin{matrix} -3x_1 & 0 & x_3 & 0 \\ 0 & 1.5x_2 & 0 & = x_3 \\ x_1 & 4x_2 & -3x_3 & 0 \end{matrix}$$

Matrix is singular (no unique solution)
→ fix value of one variable e.g. x_1 = 1.

Example II: Calculation of Eigenvalues and -vectors

$$x_1 = 1 \quad \Rightarrow \quad \begin{matrix} -3 & 0 & 0 & -x_3 \\ 0 & 1.5\, x_2 & 0 & = & x_3 \\ 1 & 4\, x_2 & -3\, x_3 & 0 \end{matrix} \quad \Rightarrow \quad x_1 = 1\, ; \ x_2 = 2\, ; \ x_3 = 3$$

Calculation of eigenvectors for all eigenvalues yields the following matrix of eigenvectors (or multiples of columns):

$$\begin{pmatrix} 1 & 7 & 2 \\ 2 & -2 & -1 \\ 3 & 1 & 1 \end{pmatrix}$$

Eigenvalues: 4; 0 and -1

Introduction to multivariate analysis, ordination and PCA

Contents

1. Learning targets and introduction to multivariate analysis

2. Overview ordination

3. Introduction to PCA and mathematical background

4. **PCA results and interpretation**

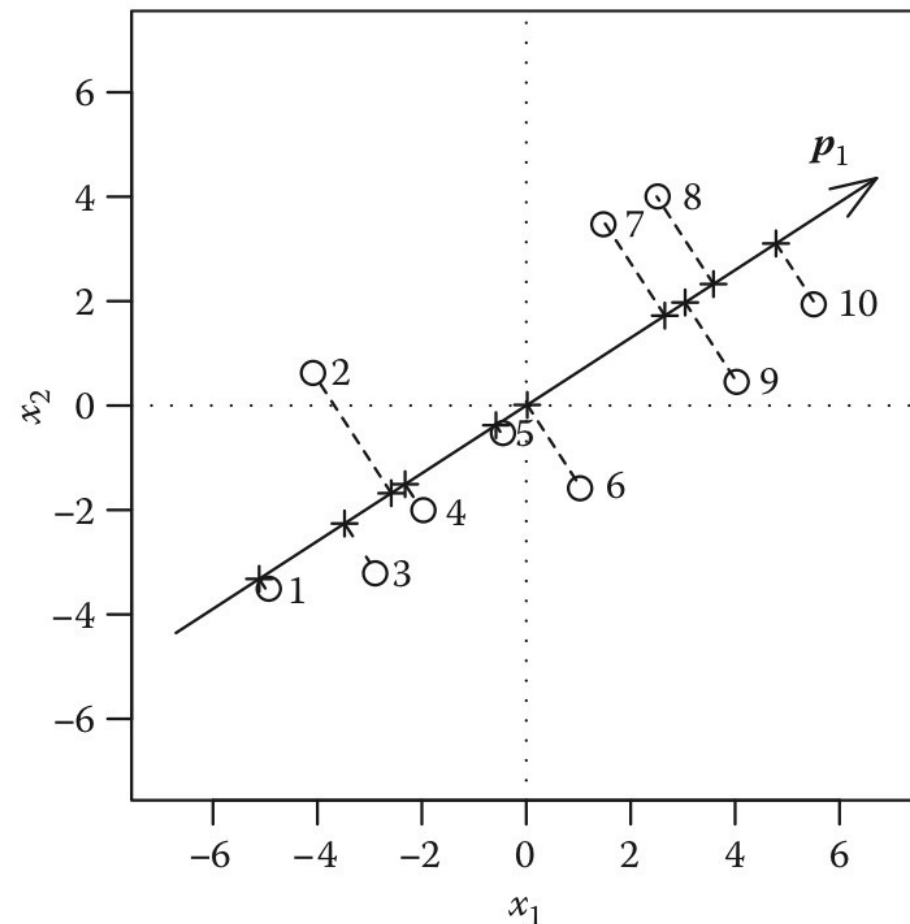5. PCA diagnosis and extensions, brief tutorial

Results of PCA II

PC Scores

Demo Example for PCA with 10 Objects and Two Mean-Centered Variables x_1 and x_2

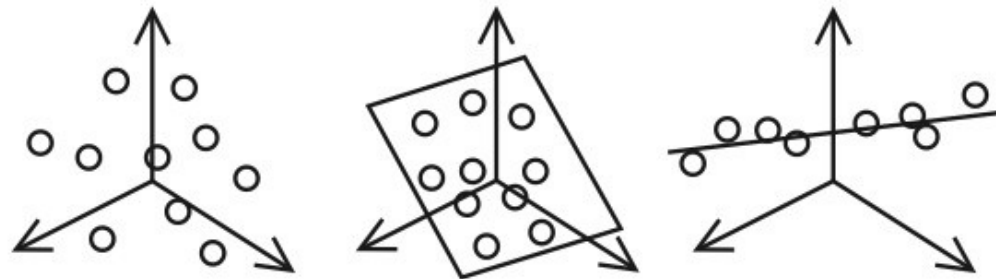| i	x_1	x_2	t_1	t_2
1	−5.0	−3.5	−6.10	−0.21
2	−4.0	0.5	−3.08	2.60
3	−3.0	−3.0	−4.15	−0.88
4	−2.0	−2.0	−2.77	−0.59
5	−0.5	−0.5	−0.69	−0.15
6	1.0	−1.5	0.02	−1.80
7	1.5	3.5	3.16	2.12
8	2.5	4.0	4.27	1.99
9	4.0	0.5	3.63	−1.76
10	5.5	2.0	5.70	−1.32
$\bar{x}$	0.00	0.00	0.00	0.00
v	12.22	6.72	16.22	2.72
$v\%$	64.52	35.48	85.64	14.36

Note: i, Object number; t_1 and t_2 are the PCA scores of PC1 and PC2, respectively; $\bar{x}$, mean; v, variance; $v\%$, variance in percent of total variance.

PC scores result from multiplication of scores from initial axes with eigenvectors

Varmuza & Filzmoser 2009: Chapter 3

Number of principal components

- Number of descriptors/explanatory variables determines number of eigenvalues and thus principal components

- Largest eigenvalue (and the corresponding eigenvector) explains highest share of total variance

- Aim is to represent the major variation with a few principal components → How many components are needed?

| Number of variables | 3 | 3 | 3 |
| Number of relevant components = intrinsic dimensionality | 3 | 2 | 1 |

How many principal components needed?

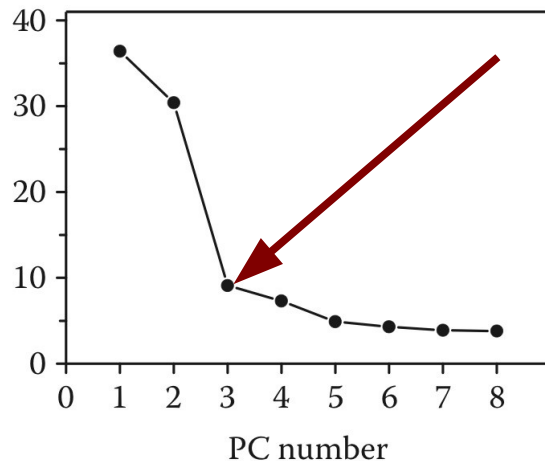Some criteria to evaluate the optimal number of axes:

1. Sum criterion

$$\frac{\sum\limits_{i=1}^{r} \lambda_i}{\sum\limits_{j=1}^{p} \lambda_j} \geq \alpha$$

2. Broken-Stick criterion

$$\lambda_i > \frac{1}{p} \sum\limits_{i=1}^{p} \frac{1}{i}$$

3. Scree plot

PC number

4. Cross-validation

$$\min_{S} \; \mathrm{MSEP}(S) \; = \; \frac{1}{I\,K} \sum\limits_{i=1}^{I} \sum\limits_{k=1}^{K} \left(x_{i,k} - (\hat{x}_{i,k})^{(S)} \right)^2$$

For the matrix $X_{I \times K}$, we search the number of PC S minimizing the mean square error of prediction (MSEP).

 Varmuza & Filzmoser 2009: Chapter 3, Jolliffe 2002: Chapter 6, Josse & Husson 2012

Importance of descriptor for PC axes

- Elements of eigenvector matrix = "loadings", indicate weight of original descriptor on PCs

Eigenvector of

	PC1	PC2	PC3	
e.g. loading of descriptor 1 on PC1 →	0.32	0.94	−0.1	Descriptor 1
	−0.68	0.15	−0.72	Descriptor 2
	−0.66	0.3	0.7	Descriptor 3

- Easier to interpret: Correlation loadings $\quad r_i = \boldsymbol{a_i}\sqrt{\lambda_i}$

- Interpretation of descriptor importance complicated if many variables load on PC

- Sparse PCA – introduces penalty term (cf. LASSO):

$$\max_{a} \; Var(\boldsymbol{a}\,\widetilde{\boldsymbol{X}}) - \lambda\|\boldsymbol{a}\| \; \text{ with } \; \boldsymbol{a}^t\boldsymbol{a} = 1$$

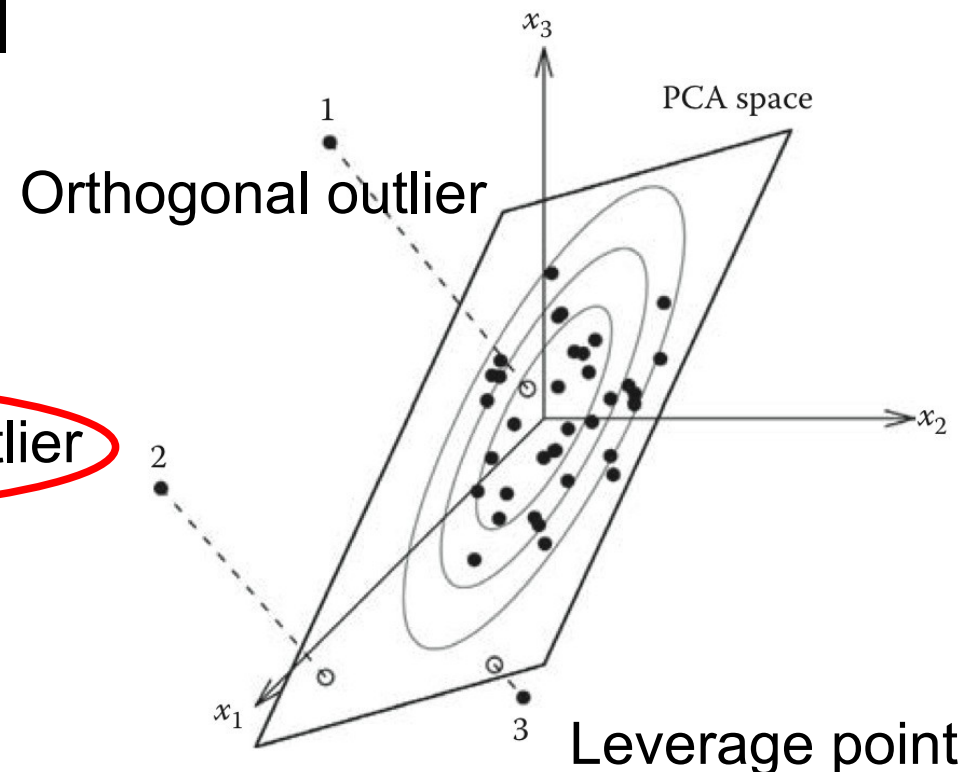# Introduction to multivariate analysis, ordination and PCA

Contents

1. Learning targets and introduction to multivariate analysis

2. Overview ordination

3. Introduction to PCA and mathematical background

4. PCA results and interpretation

5. **PCA diagnosis and extensions, brief tutorial**

PCA assumptions and diagnosis

- Independence of observations (temporal and spatial independence)

- multivariate normality (depending on research question)

- no serious outliers, otherwise use robust PCA (see Varmuza & Filzmoser 2009: chapter 3.5)

- Computation of orthogonal distance (OD) and score distance (SD)

Orthogonal outlier

Leverage point and orthogonal outlier

Problematic

Leverage point

Varmuza & Filzmoser 2009, ch.3

PCA assumptions and limitations

- Linear gradient of descriptors (rarely the case for species data, but often for environmental data)

- Euclidean distances inappropriate for species data with many double zeros (joint absences)

- Adding noise variables to data increases fraction of variance on first axis, but has no meaning

- Useful technique in exploratory analysis and for analysing environmental data (or other data with linear gradients)

- Best results for large n and high $n:p$ (p = descriptors)

Brief tutorial for PCA

1. Check if conditions for descriptors are met (quantitative, multivariate normality, linear)

2. Conduct PCA (or sparse PCA) on scaled descriptors unless they exhibit a similar variation and have been measured on similar scale

3. Check for outliers

4. Determine number of principal components

5. How informative are first two PCs?

6. Which descriptors contribute most to PCs?

7. Visualise and interpret

Principal component regression

- Extract (unscaled) PC scores to use PCs as descriptors in multiple regression analysis

- PCs are orthogonal $\rightarrow$ fix for multicollinearity in regression analysis

- In low n:p situations, the few last PCs are often removed to reduce number of predictors in regression $\rightarrow$ Can be problematic because low variance of PC does not imply low explanatory power for response variable
  $\rightarrow$ not necessarily a fix for low n:p ratios

Introduction to multivariate analysis, ordination and PCA

Contents

Learning targets

- Explain the specifics of multivariate analysis
- List ordination methods and ability to select based on research question
- Explain the mathematical basis of and apply PCA
- Interpreting results from a PCA

Learning targets and study questions

- Explain the specifics of multivariate analysis

  - Why should you favour multivariate approaches for multivariate data?

  - Outline the differences to the univariate case when diagnosing multivariate outliers and normality.

- List ordination methods and ability to select based on research question

  - Outline the aim of ordination

  - Distinguish constrained and unconstrained ordination

  - Which criteria should guide the selection of an ordination method?

3

Learning targets and study questions

- Explain the mathematical basis of and apply PCA

  - Explain the variance-covariance matrix. What are eigenvalues and eigenvectors?

  - How do eigenvalues relate to the variance of a PC?

  - Outline criteria to determine the optimal number of PCs.

  - What is sparse PCA and how does it influence the evaluation of descriptor contribution to PCs?

- Interpreting results from a PCA

  - Explain biplots with respect to (a) correlation between variables and (b) relation between sites/species and variables.

  - How does scaling influence interpretation of a biplot?

  - Which objects from a PCA would be extracted as non-collinear predictors for a multiple regression analysis?

4

From univariate to multivariate statistics

	univariate	**multivariate**
Variables (vars.)	Single/multiple explanatory vars., single response var. y	Single/multiple explanatory vars., multiple response vars. $y_1, …, y_n$
Distribution of response	One-dimensional	n-dimensional
Data format	y is vector	$y_1, …, y_n$ constitute matrix
Example	Species richness and environmental variables	Taxa and environmental variables

5

The table describes the distinction between univariate and multivariate statistics with a focus on the response (i.e. dependent) variable(s). Often multiple regression analysis is also considered as multivariate (e.g. Lloyd 2010: Spatial data analysis. Oxford Univ. Press: Oxford), though it is rather a multivariable technique (see Hidalgo & Goodman 2013).

Cited reference:
Hidalgo B. & Goodman M. (2013) Multivariate or Multivariable Regression? American journal of public health 103, 10.2105/AJPH.2012.300897.

Multivariate data analysis: Introduction

Some advantages of multivariate over univariate methods for analysing multivariate data

- Not all research questions can be answered with univariate statistical methods
  e.g. *What are the most important environmental variables determining community composition?*

- Multivariate methods allow for dimension reduction and visualisation of multidimensional data
  e.g. *Ordination, Cluster dendrogram*

- Joint (multivariate) analysis can reduce noise and increase power of statistical tests

6

In univariate analyses, only the relationship between single species and environmental variables can be examined, whereas multivariate analyses allow for the simultaneous analysis of how environmental variables act on a community of organisms.

Furthermore, since response variables incorporate some random variation, they can respond to noise variables for example in multiple regression analysis (Flack, V. F.; Chang, P. C., Frequency of Selecting Noise Variables in Subset Regression-Analysis - a Simulation Study. *American Statistician* 1987: 84-86). The simultaneous consideration of several response variables as in multivariate analysis reduces the influence of noise variables and can increase the power of statistical tests. For example, multivariate techniques are used for the reconstruction of the global temperature trend, where single time lines would not suffice to discover the global warming trend (see Ammann & Wahl 2007: The importance of the geophysical context in statistical evaluations of climate reconstruction procedures. *Clim. Change* 85 (1-2): 71-88.)

Multivariate approaches in R

Available techniques and developments: CRAN Task View

CRAN Task View: Multivariate Statistics

Maintainer: Paul Hewson
Contact: Paul.Hewson at plymouth.ac.uk
Version: 2014-12-18

Base R contains most of the functionality for classical multivariate analysis, somewhere. There are a large number of packages on CRAN which extend this methodology, a brief overview is given below. Application-specific uses of multivariate statistics are described in relevant task views, for example whilst principal components are listed here, ordination is covered in the Environmetrics task view. Further information on supervised classification can be found in the MachineLearning task view, and unsupervised classification in the Cluster task view.

The packages in this view can be roughly structured into the following topics. If you think that some package is missing from the list, please let me know.

Visualising multivariate data

- *Graphical Procedures:* A range of base graphics (e.g. `pairs()` and `coplot()`) and lattice functions (e.g. `xyplot()` and `splom()`) are useful for visualising pairwise arrays of 2-dimensional scatterplots, clouds and 3-dimensional densities. `scatterplot.matrix` in the car provides usefully enhanced pairwise scatterplots. Beyond this, scatterplot3d provides 3 dimensional scatterplots, aplpack provides bagplots and `spin3R()`, a function for rotating 3d clouds. misc3d, dependent upon rgl, provides animated functions within R useful for visualising densities. YaleToolkit provides a range of useful visualisation techniques for multivariate data. More specialised multivariate plots include the following: `faces()` in aplpack provides Chernoff's faces; `parcoord()` from MASS provides parallel coordinate plots; `stars()` in graphics provides a choice of star, radar and cobweb plots respectively. `mstree()` in ade4 and `spantree()` in vegan provide minimum spanning tree functionality. calibrate supports biplot and scatterplot axis labelling. geometry, which provides an interface to the qhull library, gives indices to the relevant points via `convexhulln()`. ellipse draws ellipses for two parameters, and provides `plotcorr()`, visual display of a correlation matrix. denpro provides level set trees for multivariate visualisation. Mosaic plots are available via `mosaicplot()` in graphics and `mosaic()` in vcd that also contains other visualization techniques for multivariate categorical data. gclus provides a number of cluster specific graphical enhancements for scatterplots and parallel coordinate plots See the links for a reference to GGobi. rggobi interfaces with GGobi. xgobi interfaces to the XGobi and XGvis programs which allow linked, dynamic multivariate plots as well as projection pursuit. Finally, iplots allows particularly powerful dynamic interactive graphics, of which interactive parallel co-ordinate plots and mosaic plots may be of great interest. Seriation methods are provided by seriation which can reorder matrices and dendrograms.
- *Data Preprocessing:* `summarize()` and `summary.formula()` in Hmisc assist with descriptive functions; from the same package `varclus()` offers variable clustering while `dataRep()` and `find.matches()` assist in exploring a given dataset in terms of representativeness and finding matches. Whilst `dist()` in base and `daisy()` in cluster provide a wide range of distance measures, proxy provides a framework for more distance measures, including measures between matrices. simba provides functions for dealing with presence / absence data including similarity matrices and reshaping.

Hypothesis testing

- ICSNP provides Hotellings T2 test as well as a range of non-parametric tests including location tests based on marginal ranks, spatial median and spatial signs computation, estimates of shape. Non-parametric two sample tests are also available from cramer and spatial sign and rank tests to investigate location, sphericity and independence are available in SpatialNP.

Multivariate distributions

- *Descriptive measures:* `cov()` and `cor()` in stats will provide estimates of the covariance and correlation matrices respectively. ICSNP offers several descriptive measures such as `spatial.median()` which provides an estimate of the spatial median and further functions which provide estimates of scatter. Further robust methods are provided such as `cov.rob()` in MASS which provides robust estimates of the variance-covariance matrix by minimum volume ellipsoid, minimum covariance determinant or classical product-moment. covRobust provides robust covariance estimation via nearest neighbor variance estimation. robustbase provides robust covariance estimation via fast minimum covariance determinant with

Additional methods that are still under development can be found on R Forge:
https://r-forge.r-project.org/softwaremap/tag_cloud.php?tag=multivariate

Multivariate outlier checking

Multivariate and univariate exploration/diagnosis similar for e.g. multicollinearity and transformation, but differ for checking of e.g. outliers and distributional assumptions
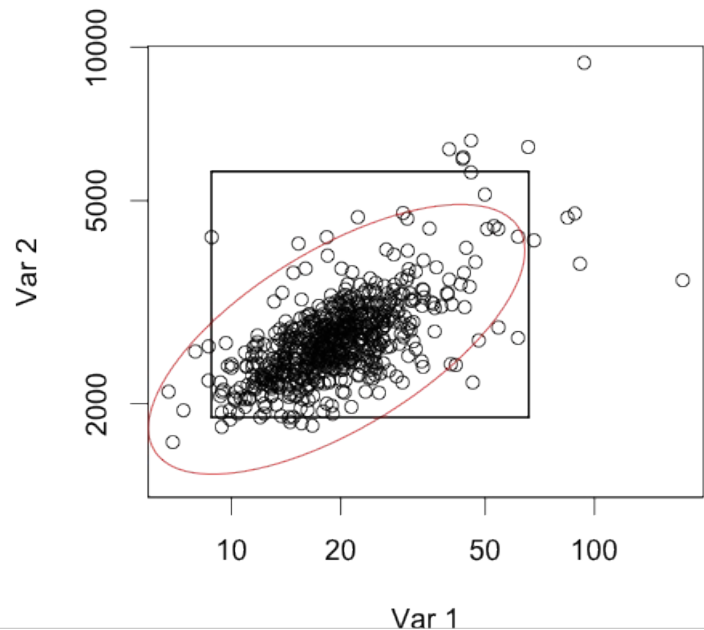
Multivariate outliers
use joint multivariate distribution of variables to find outliers (i.e. extreme points from centre of multivariate sample)

Outside of box:
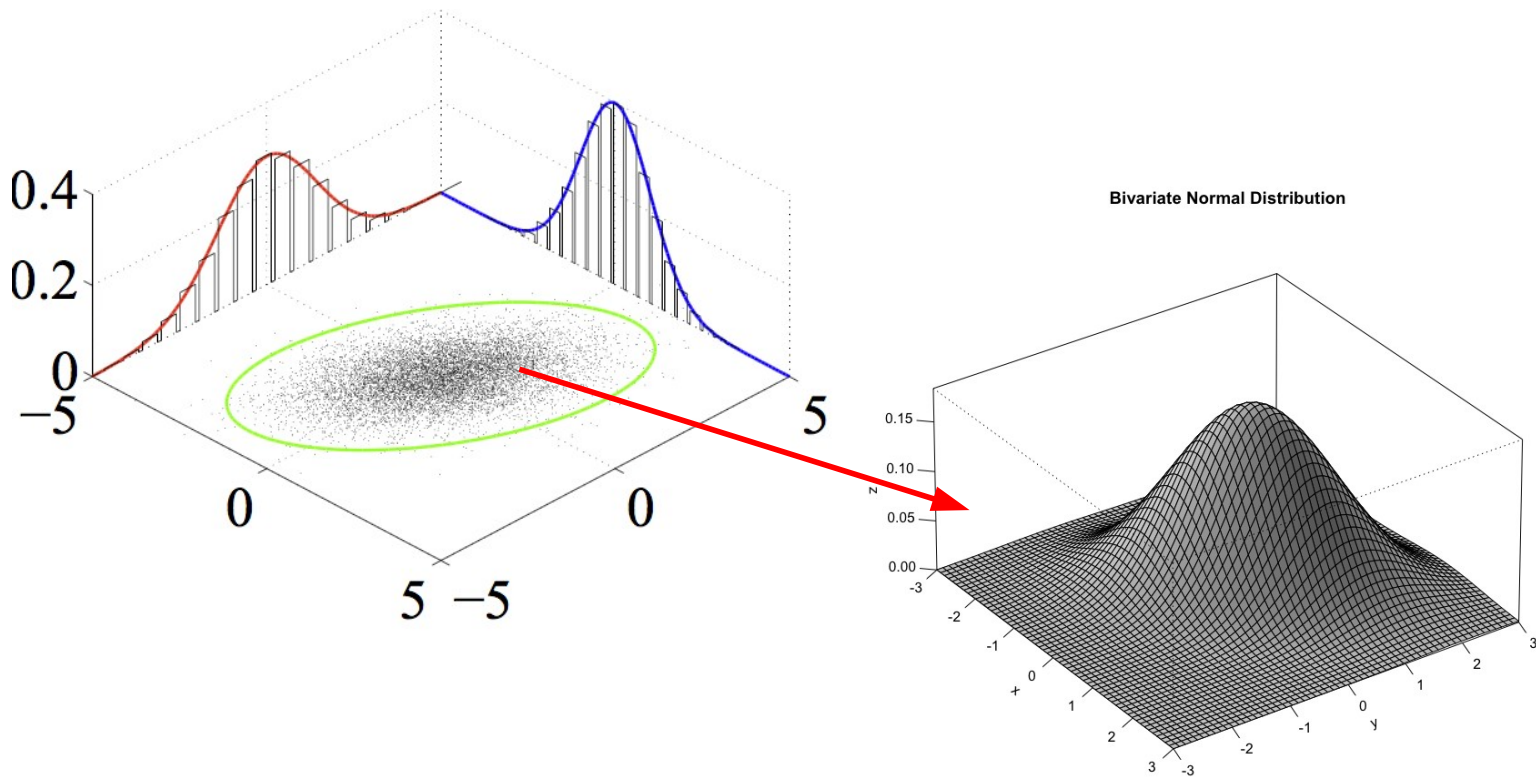Outliers, considering both variables individually (99% quantile of each variable)

Outside of ellipse:
Outliers, considering both variables jointly (99% quantile using Mahalanobis distance)
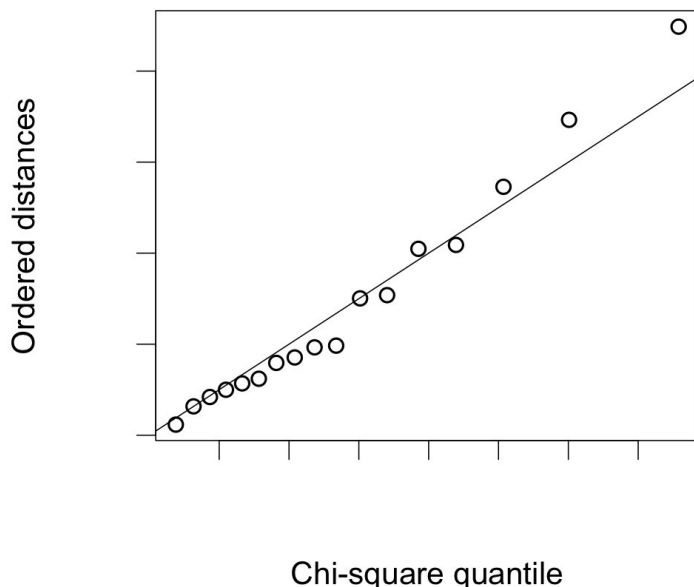

8

Multivariate normal distribution

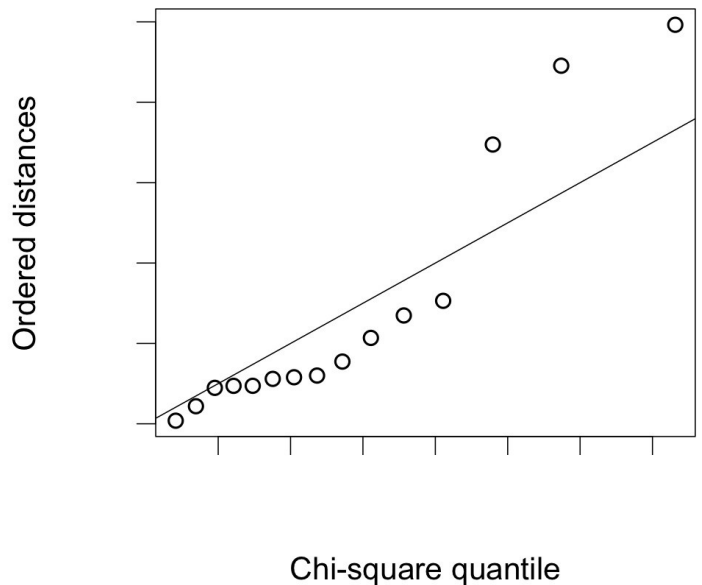Multivariate normal distribution is evaluated instead of univariate normal distribution

Bivariate Normal Distribution

http://commons.wikimedia.org/wiki/File:Multivariate_normal_sample.svg

Multivariate normal distribution

Visual check of multivariate normality with *q-q* plots for sample Mahalanobis distances (to centroid) and theoretical quantiles from the χ^2 distribution

Q-Q plot with data sampled from a multivariate normal distribution

Q-Q plot with data sampled from a exponential distribution

Regarding the calculation of the *q-q* plot: The distance to the multivariate centroid of the sample is calculated for each empirical multivariate observation x_i and weighed by the inverse of the sample covariance matrix (Mahalanobis distance (D_M); further details on the D_M are provided in a few slides). The D_Ms are then ordered and compared to the quantiles of a beta or chi-square distribution. The chi-square distribution can be misleading for a low ratio of observations to variables (<25), see Small (1978) for details. In this case the beta distribution may yield more reliable results.

Several hypothesis tests are available to check for multivariate normality (see CRAN Task View). The criticism on hypothesis testing for normality outlined for the univariate case largely applies to them as well.

Small 1978: Plotting squared radii. Biometrika 65 (3): 657-658

Covariance matrix

For a data matrix **Y** containing the variables **y**$_1$,...,**y**$_p$

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \\ \vdots \\ \mathbf{y}_i' \\ \vdots \\ \mathbf{y}_n' \end{pmatrix} = \text{(units)} \begin{matrix} 1 \\ 2 \\ \\ i \\ \\ n \end{matrix} \overset{\overset{\text{(variables)}}{\begin{matrix} 1 & 2 & & j & & p \end{matrix}}}{\begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1j} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2j} & \cdots & y_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ y_{i1} & y_{i2} & \cdots & y_{ij} & \cdots & y_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nj} & \cdots & y_{np} \end{pmatrix}}$$

the sample covariance matrix for these variables is **S** (Sigma)

$$\mathbf{S} = (s_{jk}) = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix}$$

is obtained by: $s_{jk} = \dfrac{1}{n-1} \sum_{i=1}^{n} (y_{ij} - \overline{y}_j)(y_{ik} - \overline{y}_k)$

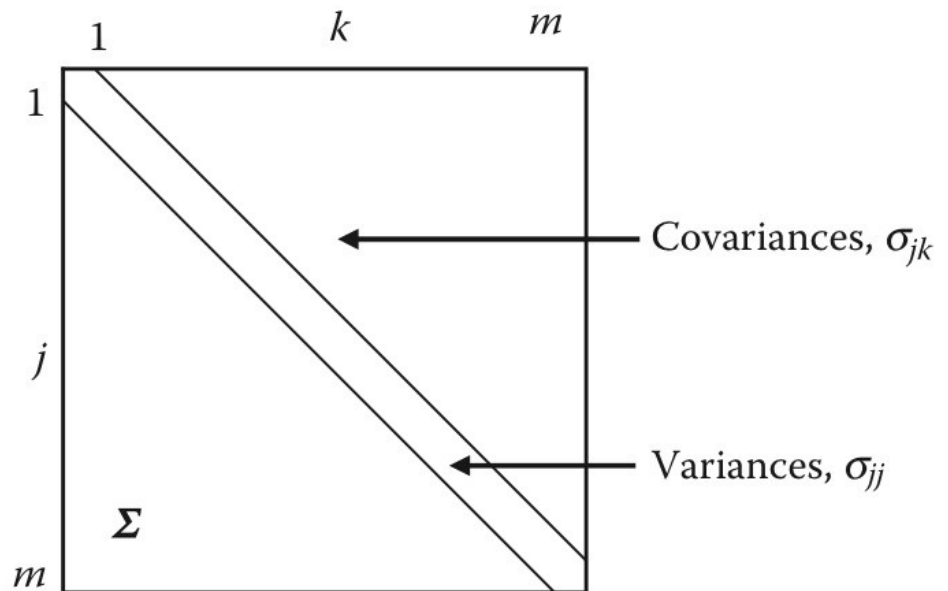taken from Rencher (2012): *Methods of multivariate Analysis*. 55f

---

The formula for the covariance should be very familiar to you from univariate statistics: we encountered the covariance already in the section on bivariate correlation (the bivariate correlation for two variables **x**, **y** is the covariance divided by the product of the two standard deviations of the variables):

$$r_{x,y} = \frac{\dfrac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\dfrac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1} \dfrac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n-1}}}$$
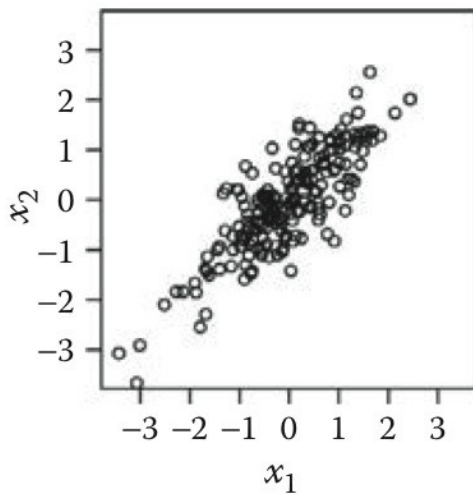
Covariance matrix
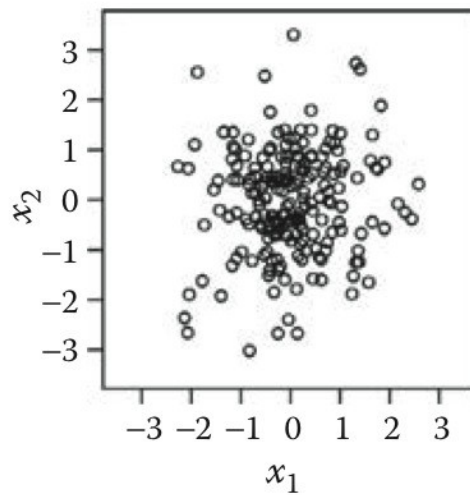
In the diagonal, the equation simplifies to:

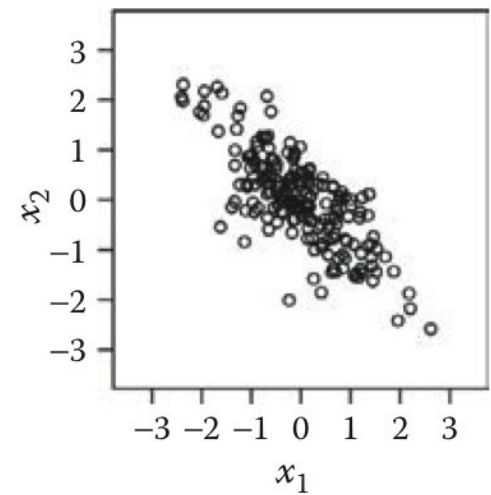$$s_{jj} \;=\; \frac{1}{n-1} \sum_{i=1}^{n} (y_{ij} - \overline{y}_j)^2$$

Covariance matrix for two variables

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix} \qquad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \boldsymbol{\Sigma}_3 = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 1 \end{pmatrix}$$

Multivariate distance from mean to observation is measured with Mahalanobis distance D_M:

$$D_M(x) \; = \; \sqrt{(\boldsymbol{x} - \boldsymbol{\mu})^t \, \boldsymbol{S}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})}$$

$\rightarrow D_M$ is distance of vector $\boldsymbol{x}$ from the mean vector $\boldsymbol{\mu}$ weighed by their covariance (given in sample covariance matrix $\boldsymbol{S}$)

Varmuza 2009

The D_M can also be calculated for other vectors than μ, for example to quantify the multivariate distance between any two vectors of the same length.

D_M represents the multivariate extension of the z-score, which is calculated by subtracting the mean of x (given as μ) from each observation i from x, and dividing the result by the standard deviation of x:

$$z \; = \; \frac{\boldsymbol{x} - \mu}{\sqrt{\dfrac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu)^2}}$$

The division by the standard deviation and the multiplication with the inverse of the sample covariance matrix have the same purpose: to standardise the difference to the sample mean/centroid by the variation in the respective variable(s).

Introduction to multivariate analysis, ordination and PCA

Contents

1. Learning targets and introduction to multivariate analysis

2. **Overview ordination**

3. Introduction to PCA and mathematical background

4. PCA results and interpretation

5. PCA diagnosis and extensions, brief tutorial

Ordination: Introduction

- Extraction of new axes from high dimensional data that sequentially maximise the variance
  - → Dimension reduction (e.g. omission of axes that capture low amount of variance)
  - → Aggregation of variables into gradients
  - → Graphical representation in lower dimensionality

- Unconstrained ordination: extraction without consideration of variables outside of data set

- Constrained ordination: extraction of information, which can be explained by variables of a second data set

Very readable introductions into multivariate methods are:
Ramette, A. (2007) Multivariate analyses in microbial ecology. FEMS Microbiology Ecology, 62, 142-160. Freely available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2121141/pdf/fem0062-0142.pdf

Paliy O. & Shankar V. (2016) Application of multivariate statistical techniques in microbial ecology. Molecular Ecology 25, 1032–1057. http://onlinelibrary.wiley.com/doi/10.1111/mec.13536/abstract

with a strong focus on ecotoxicology:
van den Brink, P. J.; van den Brink, N. W.; Ter Braak, C. J.F. (2003): Multivariate analysis of ecotoxicological data using ordination: demonstrations of utility on the basis of various examples. Australasian Journal of Ecotoxicology 9, 141–156. Freely available at:
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.77.403&rep=rep1&type=pdf

and finally a paper discussing different multivariate techniques in the context of diversity research:
Anderson, M. J.; Crist, T. O.; Chase, J. M.; Vellend, M.; Inouye, B. D.; Freestone, A. L. et al. (2011): Navigating the multiple meanings of diversity: a roadmap for the practicing ecologist. Ecology Letters 14, 19–28.
http://onlinelibrary.wiley.com/doi/10.1111/j.1461-0248.2010.01552.x/full
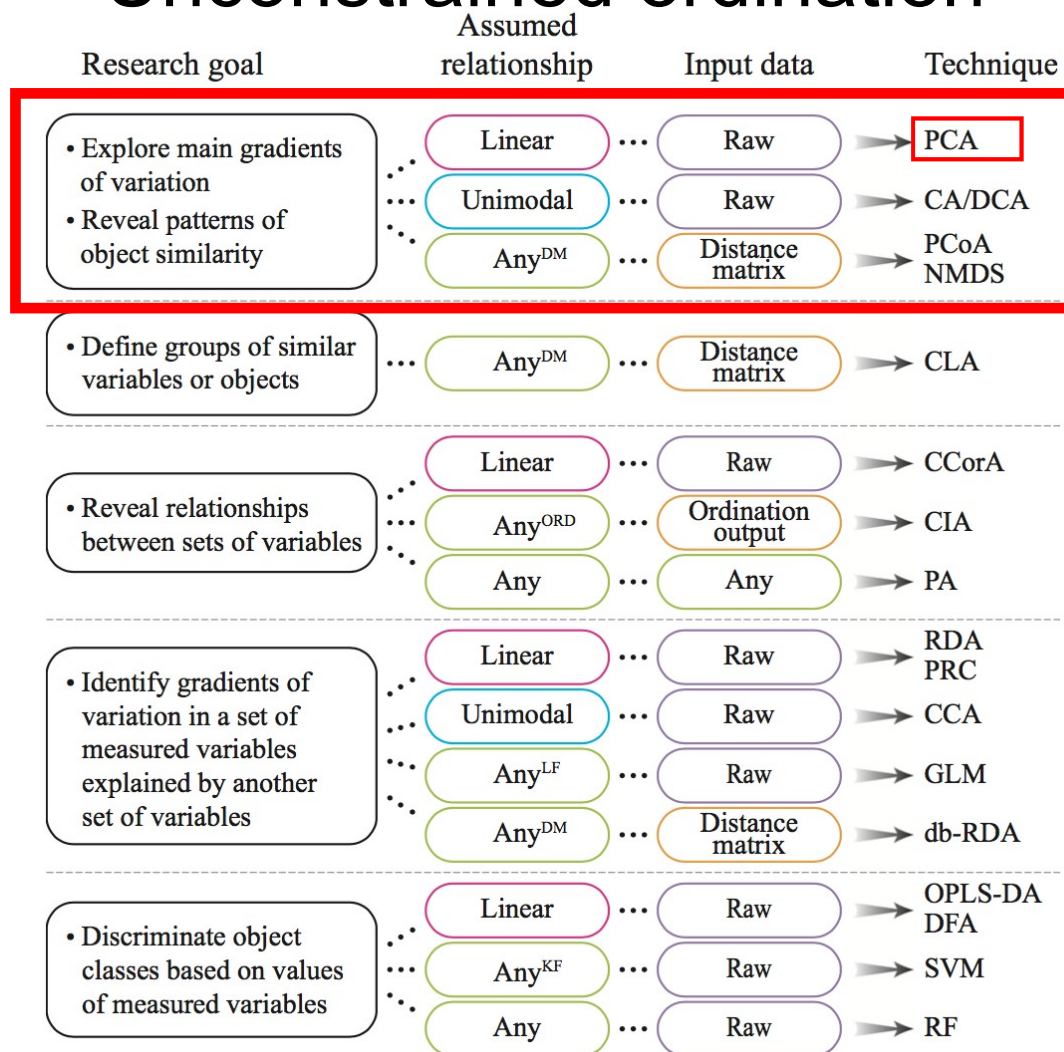
For novel developments, see:
Warton D.I., Foster S.D., De'ath G., Stoklosa J. & Dunstan P.K. (2015) Model-based thinking for community ecology. Plant Ecology 216, 669–682.
https://www.researchgate.net/publication/276481409_Model-based_thinking_for_community_ecology
Warton D.I., Blanchet F.G., O'Hara R.B., Ovaskainen O., Taskinen S., Walker S.C., et al. (2015) So Many Variables: Joint Modeling in Community Ecology. Trends in Ecology & Evolution. 30, 766-779
https://www.researchgate.net/publication/283447639_So_Many_Variables_Joint_Modeling_in_Community_Eco

Unconstrained ordination

| Research goal	Assumed relationship	Input data	Technique
• Explore main gradients of variation • Reveal patterns of object similarity	Linear	Raw	PCA
	Unimodal	Raw	CA/DCA
	Any[DM]	Distance matrix	PCoA NMDS
• Define groups of similar variables or objects	Any[DM]	Distance matrix	CLA
• Reveal relationships between sets of variables	Linear	Raw	CCorA
	Any[ORD]	Ordination output	CIA
	Any	Any	PA
• Identify gradients of variation in a set of measured variables explained by another set of variables	Linear	Raw	RDA PRC
	Unimodal	Raw	CCA
	Any[LF]	Raw	GLM
	Any[DM]	Distance matrix	db-RDA
• Discriminate object classes based on values of measured variables	Linear	Raw	OPLS-DA DFA
	Any[KF]	Raw	SVM
	Any	Raw	RF

Paliy O. & Shankar V. (2016) Application of multivariate statistical techniques in microbial ecology. *Molecular Ecology* 25, 1032–1057.

Ordination: Overview

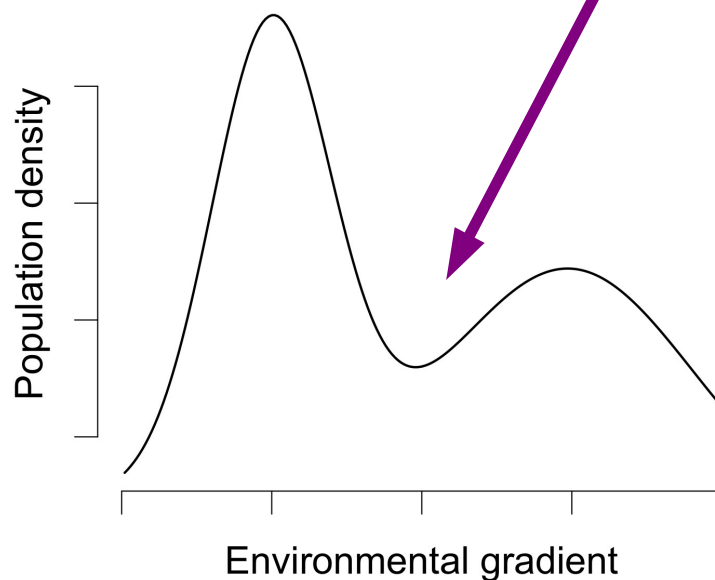| Shape of response	Linear	Unimodal	Any
Example for unconstrained method	PCA	CA	Distance-based: NMDS; GAM-based: UAO (U-VGAM)
Example for constrained method	RDA	CCA	Distance-based: db-RDA; GAM-based: CAO (RR-VGAM)
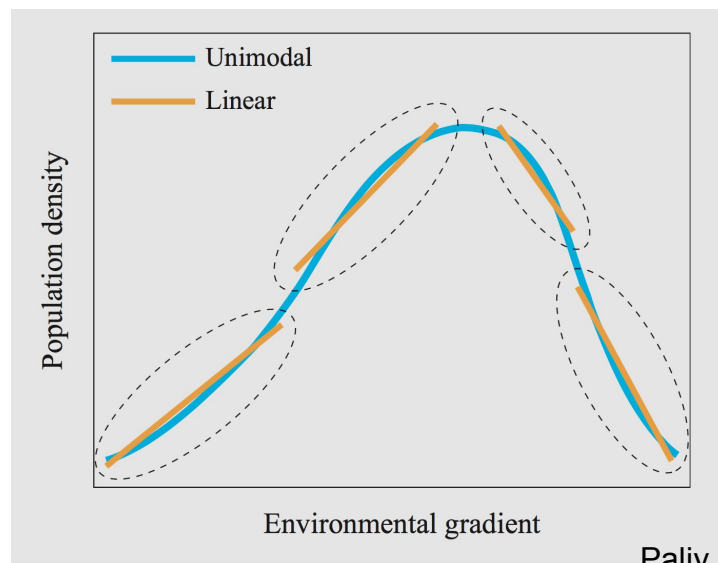
Paliy & Shankar 2016 *Mol Ecol*:1032

PCA: Principal Component Analysis
RDA: Redundancy Discriminant Analysis
CA: Correspondence Analysis
CCA: Canonical Correspondence Analysis
NMDS: Non-Metric Multidimensional Scaling
db-RDA: distance-based RDA
UAO: Unconstrained Additive Ordination
CAO: Constrained Additive Ordination
GAM: Generalised Additive Model
VGAM: Vectorised Generalised Additive Model
U-VGAM: Unconstrained-VGAM
RR-VGAM: Reduced Rank-VGAM

Ordination: Overview

Shape of response	Linear	Unimodal	Any
Example for unconstrained method	PCA	CA	Distance-based: NMDS; GAM-based: UAO (U-VGAM)
Example for constrained method	RDA	CCA	Distance-based: db-RDA; GAM-based: CAO (RR-VGAM)

PCA: Principal Component Analysis
RDA: Redundancy Discriminant Analysis
CA: Correspondence Analysis
CCA: Canonical Correspondence Analysis
NMDS: Non-Metric Multidimensional Scaling
db-RDA: distance-based RDA
UAO: Unconstrained Additive Ordination
CAO: Constrained Additive Ordination
GAM: Generalised Additive Model
VGAM: Vectorised Generalised Additive Model
U-VGAM: Unconstrained-VGAM
RR-VGAM: Reduced Rank-VGAM

Introduction to multivariate analysis, ordination and PCA

Contents

1. Learning targets and introduction to multivariate analysis

2. Overview ordination

3. **Introduction to PCA and mathematical background**

4. PCA results and interpretation

5. PCA diagnosis and extensions, brief tutorial
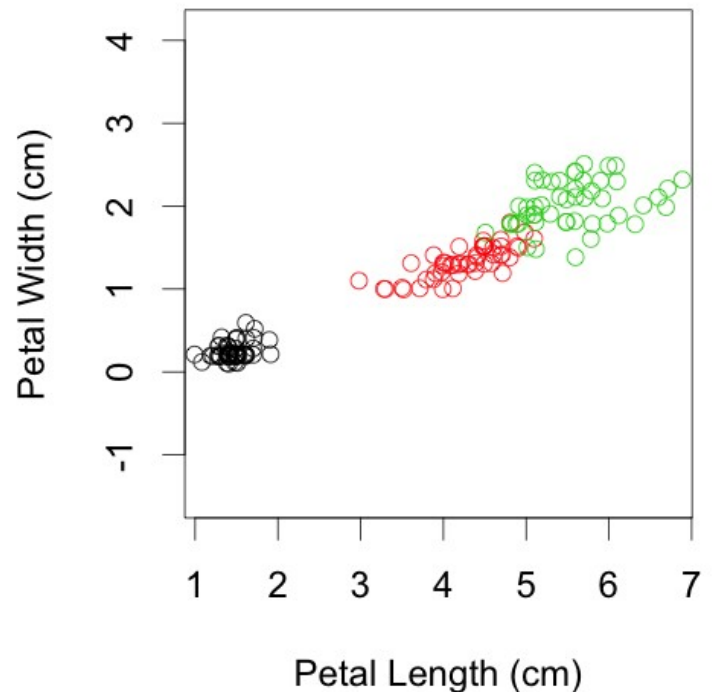
Principal Component Analysis

Example-based introduction

Iris data set: sepal length & width, petal length & width for 50 flowers from 3 species of Iris

Aim: Represent as much variance as possible on first few axes

http://de.wikipedia.org/wiki/Schwertlilien

The example has been taken from: http://blogs.nature.com/boboh/2012/01/17/pca-and-pcoa-explain (links seems not to work anymore)

PCA can be regarded as a coordinate system with linearly independent axes that is placed in the centre of the data points. The axes are then rotated until the first axis explains the maximum variance, the second axis the highest remaining variance and so on. The number of axes always equals the number of variables. PCA is motivated by the expectation that the first few axes explain the major part of the total variation. The data are centred during analysis (the mean is subtracted) to facilitate interpretation.

Introduction to PCA

Placing coordinate system in centre

Introduction to PCA
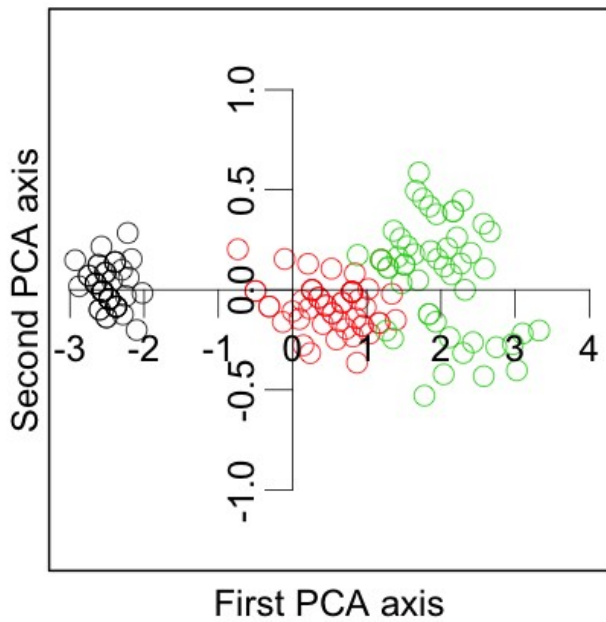

Rescaling of axis based on variation along axis
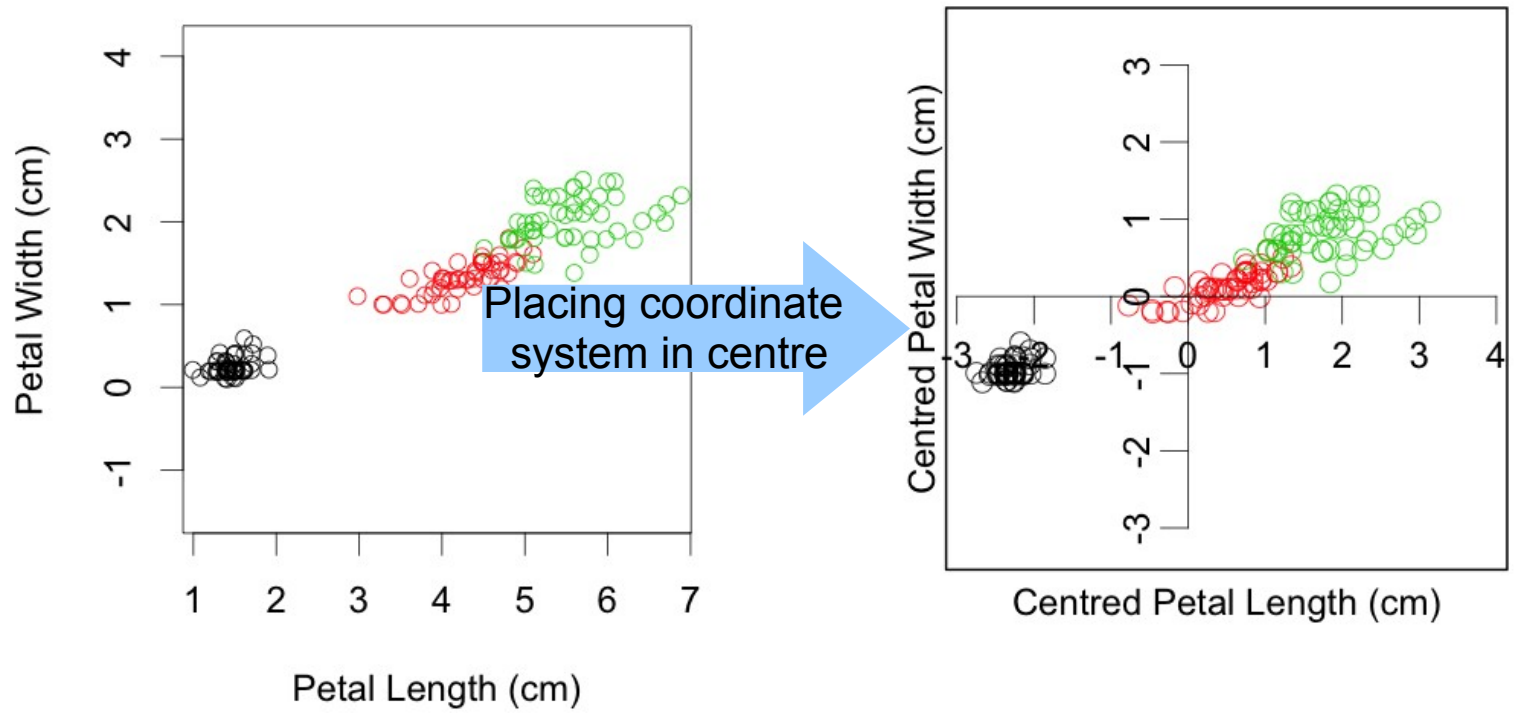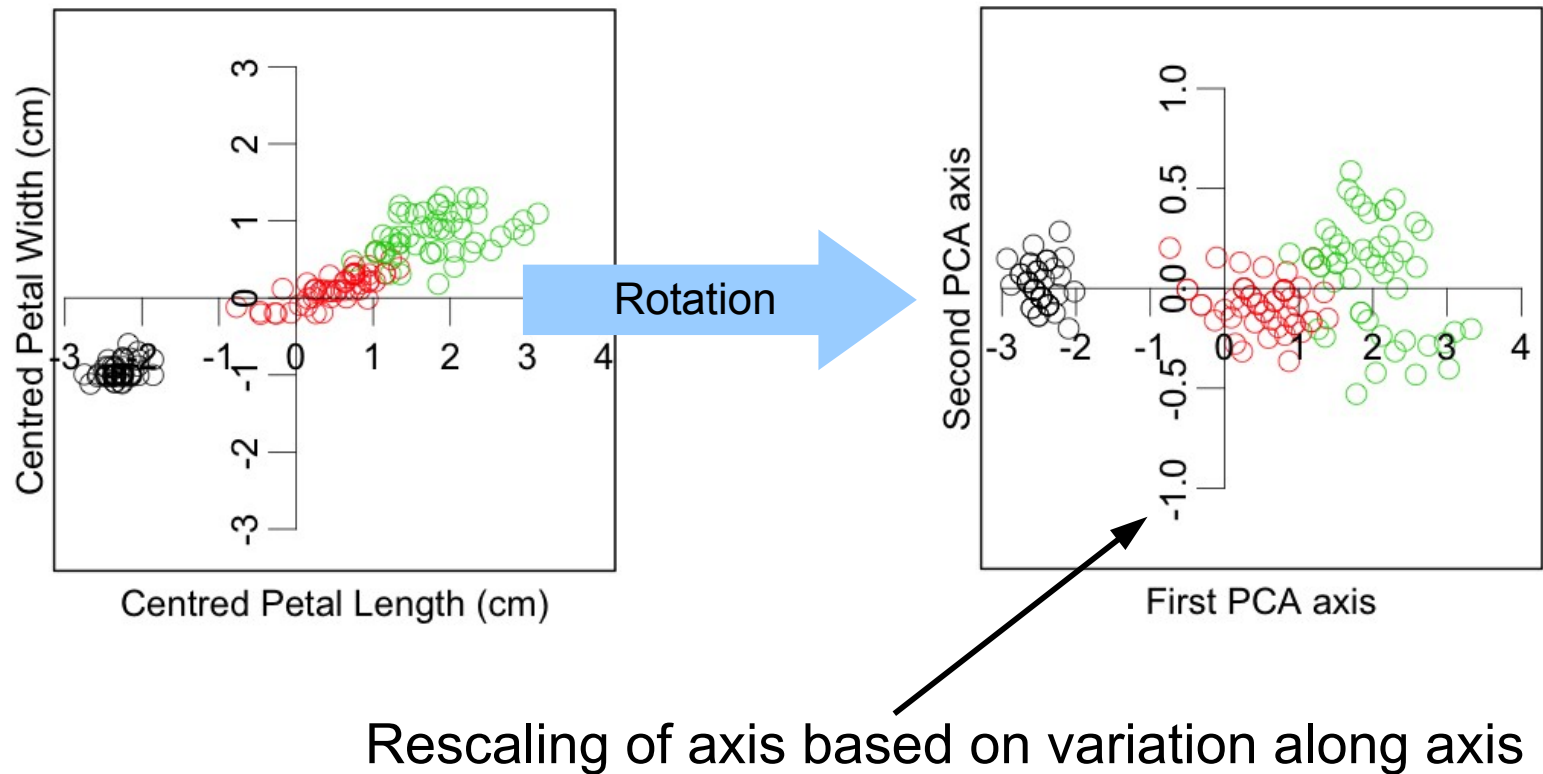
The term scaling can be used with two different meanings in the context of PCA. First, it refers to the question whether predictors are standardised to unit variance before analysis. Second, the PCA results can be scaled either to display the distances between objects or to display the relationship between descriptors (see Borcard 2011:121 and Jolliffe (2002): 90ff for details). Here, we refer to the second meaning.

Jolliffe I.T. (2002) Principal component analysis, 2nd edn. Springer, New York.

Results of PCA

Variances
Petal Width 0.58
Petal Length 3.12

<u>Results</u>

Total Variation: 3.70

Importance of components:

	PC1	PC2
Eigenvalue	3.66	0.04
Proportion Explained	0.99	0.01
Cumulative Proportion	0.99	1.00

What is an Eigenvalue?

First axis explains 99% of variation!

Mathematical background of PCA

Country	Reading	Mathematics
DK	497	514
GR	474	447
I	487	457
P	470	454
S	516	510

Centering leads to

$$\widetilde{X} = \begin{pmatrix} 8.2 & 37.6 \\ -14.8 & -29.4 \\ -1.8 & -19.4 \\ -18.8 & -22.4 \\ 27.2 & 33.6 \end{pmatrix}$$

Search for first axis with maximum variation!

Handl 2010: Multivariate Analysemethoden: Theorie und Praxis multivariater Verfahren unter besonderer Berücksichtigung von S-PLUS. Springer, Berlin.

The '~' on top of X indicates that it is an estimate for a transformed variable (centering represents a transformation).
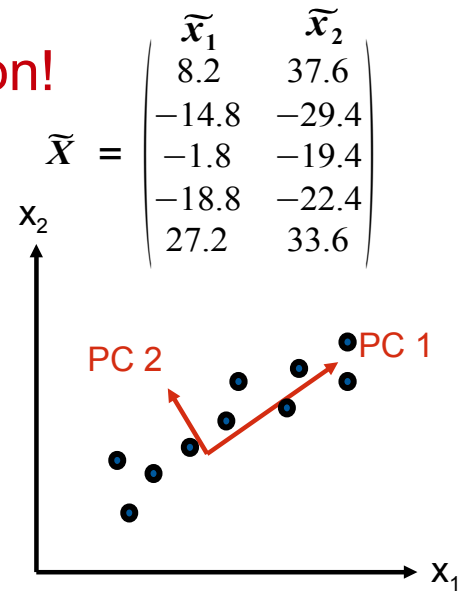
Mathematical background of PCA

Search for first axis with maximum variation!

$$\widetilde{X} = \begin{pmatrix} \widetilde{x}_1 & \widetilde{x}_2 \\ 8.2 & 37.6 \\ -14.8 & -29.4 \\ -1.8 & -19.4 \\ -18.8 & -22.4 \\ 27.2 & 33.6 \end{pmatrix}$$

Scores on new axis given as

$$PC_1 = a_1 \widetilde{x}_1 + a_2 \widetilde{x}_2$$

and maximum variation means:

$$\max_{a_1,\ a_2}\ Var\left(a_1 \widetilde{x}_1 + a_2 \widetilde{x}_2\right)$$

Generalise problem: define a_1, a_2 as elements of vector $\boldsymbol{a}$ and likewise $\widetilde{x}_1, \widetilde{x}_2$ of matrix $\boldsymbol{X}$: $\max_{a}\ Var\left(\boldsymbol{a}\,\widetilde{X}\right)$

Trivial solution: choose high values for a_1, a_2, … ,a_n
→ introduce condition: $a_1^2 + a_2^2 + ... + a_n^2 = 1$

The score of objects on the first PC axis are obtained through a linear combination of the scores on the initial axes (after centering). The general equation for *q* dimensions for the first PC axis is:

$$PC_1 = a_1 \boldsymbol{x}_1 + a_2 \boldsymbol{x}_2 + ... + a_q \boldsymbol{x}_q$$

As we could artificially inflate the variance, we introduce a condition for the coefficients of the linear combination.

Regarding the Pisa data, for $a_1 = 1$ und $a_2 = 0$ the variance is given by the variance for $\widetilde{x}_1$. If you set $a_1 = 0.6$ and $a_2 = 0.8$ (meeting the condition: $0.6^2 + 0.8^2 = 1$), this results in a variance of 1317, which is higher than the individual variances of 1071 and 346. We need to find the values for a_1 and a_2 (under condition outlined above) that maximise variation.

Mathematical background of PCA

Solve:

$$\max_{a} \; Var(a\,\widetilde{X}) \quad \text{with} \quad a^t a = 1$$

This can be expressed as (see Handl 2010: 79)

$$\max_{a} \; (a^t \Sigma \; a) \quad \text{with} \quad a^t a = 1$$

           Covariance matrix

Using the Lagrange function yields:

$$L(a,\lambda) = a^t \Sigma \; a - \lambda(a^t a - 1)$$

Eigenvalue problem

$$\frac{\partial L(a,\lambda)}{\partial a} = 2\Sigma a - 2\lambda a \longrightarrow \boxed{\Sigma a = \lambda a}$$

$$\frac{\partial L(a,\lambda)}{\partial \lambda} = 1 - a^t a$$

The description here uses the known variance-covariance matrix Σ. In practice, we usually have to estimate this matrix from the sample and use this sample variance-covariance matrix S. If the data are standardized before analysis (divided by the variance) then this matrix equals the correlation matrix R.
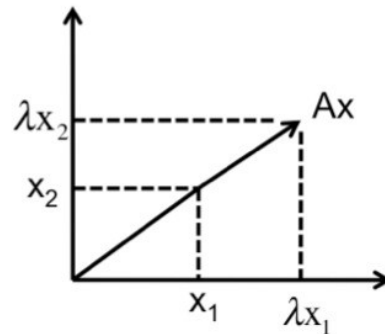
Mathematical basics II: Eigenvalues

Idea: Conversion of a matrix into a matrix with linear independent variables

$$\begin{pmatrix} a_{11} & ... & a_{1n} \\ ... & a_{22} & ... \\ a_{m1} & ... & a_{mn} \end{pmatrix} \quad \text{Conversion} \quad \Longrightarrow \quad \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_n \end{pmatrix}$$

Eigenvalue problem: $\quad A\,x = \lambda\,x \quad$ Eigenvector

Eigenvalue

Eigenvectors form canonical basis and are only stretched or shrunk by λ when multiplied with A.

Eigenvalues and eigenvectors play an important role in the mathematics behind most ordination methods. The eigenvectors and eigenvalues of matrix A can be obtained by solving the (special) eigenvalue problem. It is important to understand that the eigenvalue problem can also be rewritten as linear function $f(x) = \lambda * x$ and that this vector x is then only stretched by the factor λ for all $f(x)$. The matrix with the eigenvalues is also termed "canonical" form, a label encountered in many methods (e.g. canonical correspondence analysis, canonical correlation etc.).

Geometrically, the multiplication of a matrix with a vector, where the vector defines a point in a source coordinate system, rotates and stretches the vector to a new position. Consider the matrix $\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ that defines the rotation/stretching of a point $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

Examples for related matrices:

$\begin{pmatrix} 5 & 0 \\ 0 & 1 \end{pmatrix}$ Stretching x_1 $\begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$ 180 degree rotation

Mathematical basics II: Eigenvalues

$$A x = \lambda x \quad \Leftrightarrow \quad A x - \lambda x = 0$$

$$\Leftrightarrow \quad (A - \lambda E) x = 0$$

Ignore trivial solution: $x = 0$ $\quad \Rightarrow \quad A - \lambda E = 0$

$$\Leftrightarrow \quad \begin{pmatrix} a_{11} - \lambda_1 & \ldots & a_{1n} \\ \ldots & a_{22} - \lambda_2 & \ldots \\ a_{m1} & \ldots & a_{mn} - \lambda_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \quad \begin{array}{l} \text{Homogeneous} \\ \text{linear equation} \\ \text{system (HLS)} \end{array}$$

HLS have only non-trivial solution if determinant = 0.

$$\Rightarrow \quad det(A - \lambda E) = 0 \quad \Leftrightarrow \quad |A - \lambda E| = 0$$

E is the identity matrix.

A homogenous linear equation system $Ax = 0$ has as the unique solution $x_1 = x_2 = x_3 = \ldots = x_n = 0$ for a regular matrix (quadratic matrix for which the inverse exists). If the matrix is singular, at least one row is a linear combination of the other rows and consequently at least one $x \neq 0$. However, there is no unique solution in this case. The determinant (a special function, which assigns a unique number to every $n \times n$ matrix) for singular matrices is 0. This means that for the non-trivial case, the determinant is 0, otherwise all rows would be linearly independent (which also means that they do not share any variance and a PCA would not produce meaningful results).

Of course, for most data sets the matrices are rarely quadratic. Fortunately, the eigenvalues of a matrix can also be computed by singular value decomposition (special case of the Schur decomposition), which is not restricted to quadratic matrices. This is implemented in R with svd().

Example I: Calculation of Eigenvalues

Sample Variance-Covariance matrix from Pisa example:

$$S = \begin{pmatrix} 345.7 & 528.35 \\ 528.35 & 1071.30 \end{pmatrix}$$

Following $|A - \lambda E| = 0$ we obtain:

$$\left| \begin{pmatrix} 345.7 & 528.35 \\ 528.35 & 1071.30 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right| = 0 \Leftrightarrow$$

$$\left| \begin{pmatrix} 345.7 & 528.35 \\ 528.35 & 1071.30 \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} \right| = 0 \Leftrightarrow$$

$$\left| \begin{matrix} 345.7 - \lambda & 528.35 \\ 528.35 & 1071.30 - \lambda \end{matrix} \right| = 0 \Leftrightarrow$$
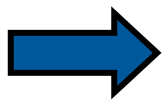
$$(345.7 - \lambda)(1071.30 - \lambda) - 528.35^2 = 0$$

For small, quadratic matrices, the determinant can be calculated by hand. For example The determinant of a matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is given as a*d – b*c.
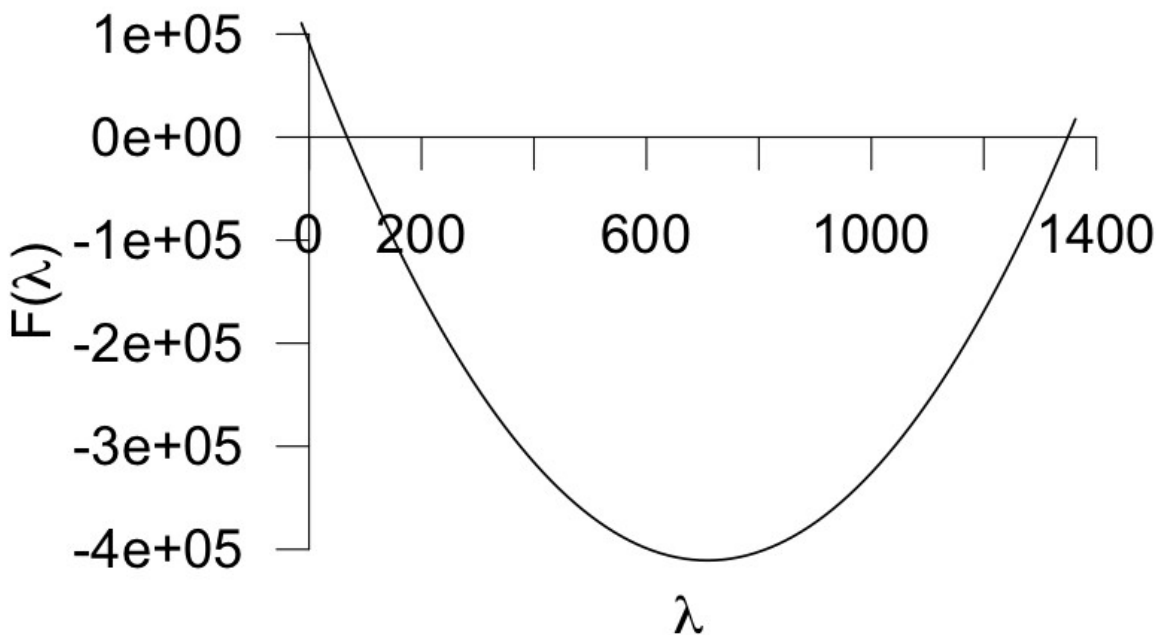
Example I: Calculation of Eigenvalues

$$(345.7-\lambda)(1071.30-\lambda)-528.35^2 \;=\; 0$$

Characteristical polynom

$$\lambda^2-1417\,\lambda+91194.69 \;=\; 0$$

In the quadratic case, the eigenvalues (λ) can easily be found using the pq formula:

$$\lambda_{1,2} \;=\; -\frac{p}{2} \;\pm\; \sqrt{\left(\frac{p}{2}\right)^2-q}$$

Evaluation the equation for our numbers yields:

$$\lambda_{1,2} \;=\; -\frac{-1417}{2} \;\pm\; \sqrt{\left(\frac{-1417}{2}\right)^2-91194.69}$$

Thus, λ_1 yields to 1349.42 and λ_2 yields to 67.6.

Example II: Calculation of Eigenvalues and -vectors
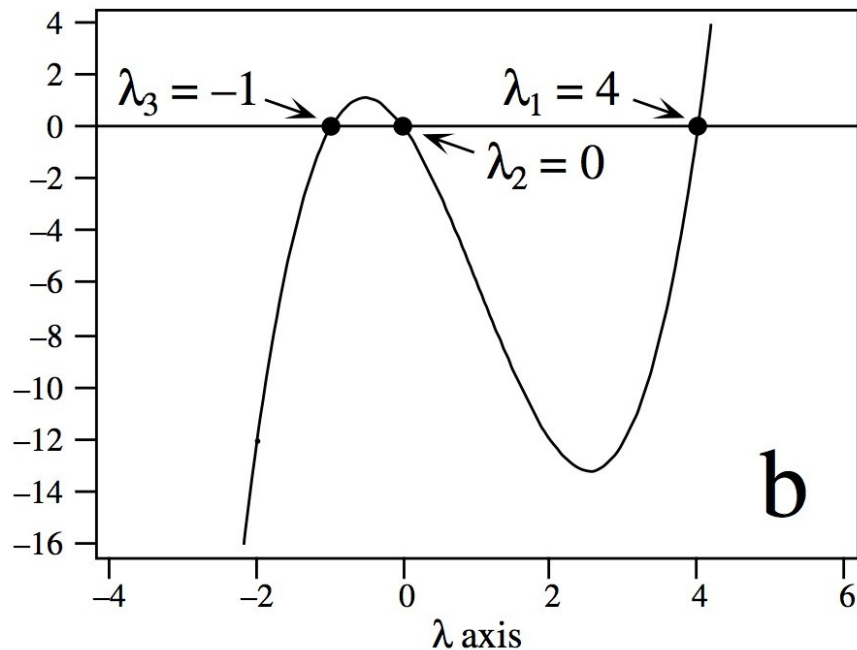
$$\begin{pmatrix} 1 & 3 & -1 \\ 0 & 1 & 2 \\ 1 & 4 & 1 \end{pmatrix}$$

EV calculation following Sarrus

Characteristical polynom

$$\lambda^3 - 3\lambda^2 - 4\lambda = 0$$

Eigenvalues λ:
4, 0 and -1

Note that this example is only to demonstrate the calculation of eigenvalues and vectors but is not based on real data. The eigenvalues of a PCA would usually all be positive.

Example II: Calculation of Eigenvalues and -vectors

Calculation of eigenvector for $\lambda = 4$

$$(A - \lambda E)x = 0$$

$$\left(\begin{pmatrix} 1 & 3 & -1 \\ 0 & 1 & 2 \\ 1 & 4 & 1 \end{pmatrix} - \lambda_1 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right)\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = 0 \Leftrightarrow \begin{pmatrix} 1-4 & 3 & -1 \\ 0 & 1-4 & 2 \\ 1 & 4 & 1-4 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = 0$$

$$\begin{matrix} -3x_1 & 3x_2 & -x_3 & 0 \\ 0 & -3x_2 & 2x_3 & = 0 \\ x_1 & 4x_2 & -3x_3 & 0 \end{matrix} \Leftrightarrow \begin{matrix} -3x_1 & 0 & x_3 & 0 \\ 0 & 1.5x_2 & 0 & = x_3 \\ x_1 & 4x_2 & -3x_3 & 0 \end{matrix}$$

Matrix is singular (no unique solution)
$\rightarrow$ fix value of one variable e.g. $x_1 = 1$.

Example II: Calculation of Eigenvalues and -vectors

$$x_1 = 1 \quad \Rightarrow \quad \begin{matrix} -3 & 0 & 0 \\ 0 & 1.5\,x_2 & 0 \\ 1 & 4\,x_2 & -3\,x_3 \end{matrix} = \begin{matrix} -x_3 \\ x_3 \\ 0 \end{matrix} \quad \Rightarrow \quad x_1 = 1\,;\ \ x_2 = 2\,;\ \ x_3 = 3$$

Calculation of eigenvectors for all eigenvalues yields the following matrix of eigenvectors (or multiples of columns):

$$\begin{pmatrix} 1 & 7 & 2 \\ 2 & -2 & -1 \\ 3 & 1 & 1 \end{pmatrix}$$

Eigenvalues: 4; 0 and -1

As an exercise, compute the eigenvectors for the eigenvalues 0 and -1.

Note that the condition $\boldsymbol{a}^t\,\boldsymbol{a} = 1$ will typically lead to values in the matrix of eigenvectors < 1, contrary to the given example. Moreover, the eigenvalues are typically positive.

Introduction to multivariate analysis, ordination and PCA

Contents

1. Learning targets and introduction to multivariate analysis

2. Overview ordination

3. Introduction to PCA and mathematical background

4. **PCA results and interpretation**

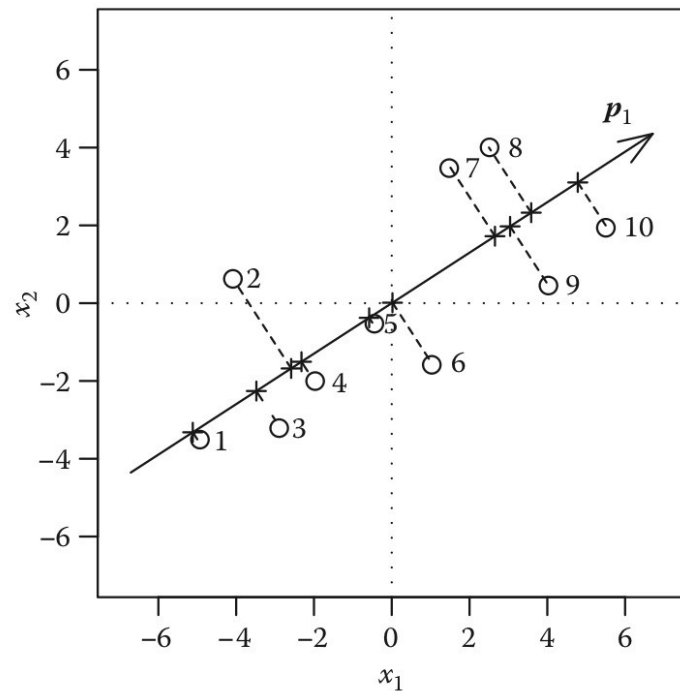5. PCA diagnosis and extensions, brief tutorial

34

PC Scores

Demo Example for PCA with 10 Objects and Two Mean-Centered Variables x_1 and x_2

i	x_1	x_2	t_1	t_2
1	−5.0	−3.5	−6.10	−0.21
2	−4.0	0.5	−3.08	2.60
3	−3.0	−3.0	−4.15	−0.88
4	−2.0	−2.0	−2.77	−0.59
5	−0.5	−0.5	−0.69	−0.15
6	1.0	−1.5	0.02	−1.80
7	1.5	3.5	3.16	2.12
8	2.5	4.0	4.27	1.99
9	4.0	0.5	3.63	−1.76
10	5.5	2.0	5.70	−1.32
$\bar{x}$	0.00	0.00	0.00	0.00
v	12.22	6.72	16.22	2.72
$v\%$	64.52	35.48	85.64	14.36

Note: i, Object number; t_1 and t_2 are the PCA scores of PC1 and PC2, respectively; $\bar{x}$, mean; v, variance; $v\%$, variance in percent of total variance.

PC scores result from multiplication of scores from initial axes with eigenvectors

The figure displays the scores t_1 for the points i on the first PC.

Number of principal components

- Number of descriptors/explanatory variables determines number of eigenvalues and thus principal components

- Largest eigenvalue (and the corresponding eigenvector) explains highest share of total variance

- Aim is to represent the major variation with a few principal components → How many components are needed?
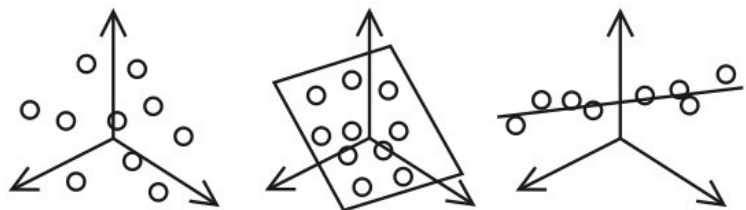
Number of variables	3	3	3
Number of relevant components = intrinsic dimensionality	3	2	1

How many principal components needed?

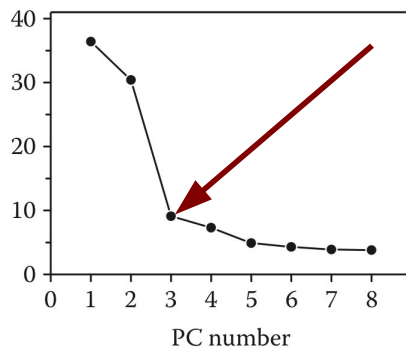Some criteria to evaluate the optimal number of axes:

1. Sum criterion

$$\frac{\sum\limits_{i=1}^{r} \lambda_i}{\sum\limits_{j=1}^{p} \lambda_j} \geq \alpha$$

2. Broken-Stick criterion

$$\lambda_i > \frac{1}{p} \sum_{i=1}^{p} \frac{1}{i}$$

3. Scree plot

4. Cross-validation

$$\min_{S} \ \text{MSEP}(S) \ = \ \frac{1}{IK} \sum_{i=1}^{I} \sum_{k=1}^{K} \left(x_{i,k} - (\hat{x}_{i,k})^{(S)} \right)^2$$

For the matrix $X_{I \times K}$, we search the number of PC S minimizing the mean square error of prediction (MSEP).

Varmuza & Filzmoser 2009: Chapter 3, Jolliffe 2002: Chapter 6, Josse & Husson 2012

The sum criterion and the scree plot are relatively simple rules without a thorough statistical foundation. In case of the sum criterion, the first $i = 1,..,r$ eigenvalues that are larger than α (proportion of cumulative variance) are selected. In other words, the minimal number of eigenvalues that capture at least % of total variance. Typically, a value between 0.7 and 0.9 is chosen for α. Generally, the optimal α will decrease with an increase in the number of variables p in the data set.

The Broken-Stick model is based on the idea that we break a stick of unit length into p pieces, where p is given by the number of eigenvalues/principal components. If the resulting pieces are sorted in decreasing order with respect to their length, the i-th length of the stick is given by the formula. The i-th eigenvalue should be larger than the i-th length from the broken-stick model, because otherwise it would not capture more variance than a random process. A modelling study (Jackson 1993) found the broken stick model among the most reliable methods for the determination of the number of principal components, though novel methods have been developed in the last two decades. Example for broken stick model: Imagine we have 2 variables, then the largest eigenvalue should be higher than 1/2 * (1+1/2) = 0.75. The second eigenvalue should then be higher than 1/2 * 1/2 = 0.25.

Cross-validation is computationally costly and may not be applicable for large data sets (for example see Saccenti & Camacho (2015)). Hence, approximations of cross-validation have been developed (Josse & Husson 2012), though they may be less reliable for some data sets. The performance of methods varies with the properties of the data sets, for an overview see Camacho & Ferrer (2014) and Saccenti & Camacho (2015). Note that statistical tests are in many cases less reliable than cross-validation and are not discussed here, but within the given references.

Camacho J. & Ferrer A. (2014) Cross-validation in PCA models with the element-wise k-fold (ekf) algorithm: Practical aspects. Chemometrics and Intelligent Laboratory Systems 131, 37–50.

Jackson, D.A. (1993) Stopping rules in principal components-analysis - a comparison of heuristic and statistical approaches. Ecology, 74, 2204-2214.

Jolliffe I.T. (2002) Principal component analysis, 2nd edn. Springer, New York.

Josse J. & Husson F. (2012) Selecting the number of components in principal component analysis using cross-validation approximations. Computational Statistics & Data Analysis 56, 1869–1879.

Saccenti E. & Camacho J. (2015) Determining the number of components in principal components analysis: A comparison of statistical, crossvalidation and approximated methods. Chemometrics and Intelligent Laboratory Systems 149, Part A, 99–116.

Importance of descriptor for PC axes

- Elements of eigenvector matrix = "loadings", indicate weight of original descriptor on PCs
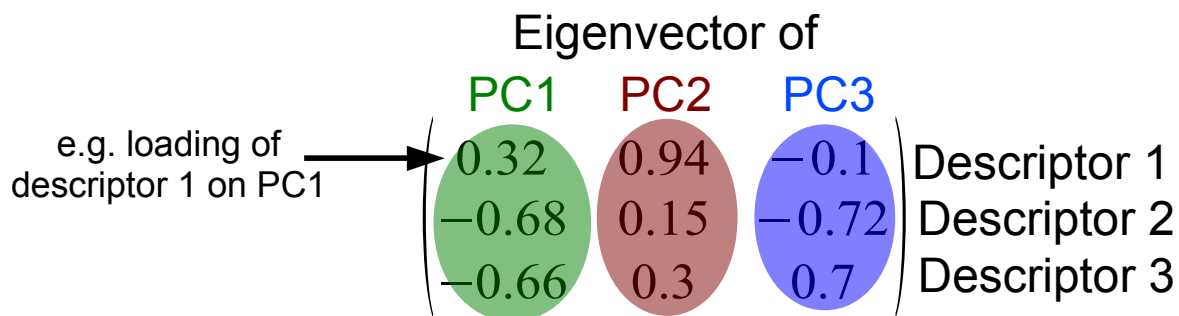
- Easier to interpret: Correlation loadings $r_i = \boldsymbol{a}_i \sqrt{\lambda_i}$

- Interpretation of descriptor importance complicated if many variables load on PC

- Sparse PCA – introduces penalty term (cf. LASSO):

$$\max_{a} \ Var(\boldsymbol{a}\, \widetilde{\boldsymbol{X}}) - \lambda \|\boldsymbol{a}\| \ \text{ with } \ \boldsymbol{a}^t \boldsymbol{a} = 1$$

The correlation loadings are equivalent to correlating the original descriptors with the PC scores.

Note that the λ used in the context of sparse PCA relates to the penalty term and not to eigenvalues. This must not be confused (i.e. the λ in the context of correlation loadings are the eigenvalues).

$\|\boldsymbol{a}\|$ is the norm of $\boldsymbol{a}$ (function that assigns a value to a vector)

For further details refer to the course material on the LASSO and the following articles:

Croux C., Filzmoser P. & Fritz H. (2011) Robust sparse principal component analysis. TU Vienna, Vienna. http://www.statistik.tuwien.ac.at/forschung/SM/SM-2011-2complete.pdf

Zou H., Hastie T. & Tibshirani R. (2006) Sparse Principal Component Analysis. Journal of Computational and Graphical Statistics 15, 265–286.

# Introduction to multivariate analysis, ordination and PCA

Contents

1. Learning targets and introduction to multivariate analysis

2. Overview ordination

3. Introduction to PCA and mathematical background

4. PCA results and interpretation

5. **PCA diagnosis and extensions, brief tutorial**

39

PCA assumptions and diagnosis

- Independence of observations (temporal and spatial independence)

- multivariate normality (depending on research question)

- no serious outliers, otherwise use robust PCA (see Varmuza & Filzmoser 2009: chapter 3.5)

- Computation of orthogonal distance (OD) and score distance (SD)

Leverage point and orthogonal outlier

Problematic

Varmuza & Filzmoser 2009, ch.3

Orthogonal outlier

x_3

PCA space

x_2

x_1
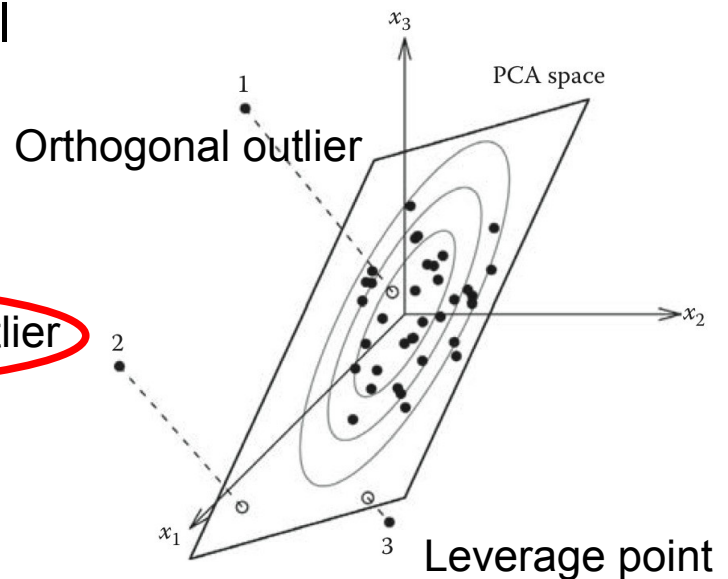
Leverage point

In the case of multivariate normal distribution of the data, the score distances can be approximated with a chi-square distribution to detect leverage points. Orthogonal distances can be approximated with a normal distribution. See Vermuza & Filzmoser 2009, chapter 3.7.3 for details.

Multivariate normality is not required if the main aim of the PCA is visual exploration. In this case, also discrete variables can be reliably displayed, as long as the data values are numerical and the distances between them meaningful. However, if the PCs are used in subsequent regression analysis (Principal component regression), multivariate normality is required. Similarly, hypothesis tests for the number of PCs require multivariate normality.

PCA assumptions and limitations

- Linear gradient of descriptors (rarely the case for species data, but often for environmental data)

- Euclidean distances inappropriate for species data with many double zeros (joint absences)

- Adding noise variables to data increases fraction of variance on first axis, but has no meaning

- Useful technique in exploratory analysis and for analysing environmental data (or other data with linear gradients)

- Best results for large n and high $n{:}p$ (p = descriptors)

The most important assumption is that of a linear gradient of the descriptors over the object space. If this assumption is not met, as for example in the case of species data, this can lead to serious problems (see the example in the related R course script. It nicely displays how unimodal gradients influence the PCA). Non-linearity can be checked using scatter plots of the descriptors before analysis. Strongly skewed data and non-linear relationships may cause problems. For such data, different ordination methods are available or the data can be transformed. See Legendre & Legendre 2012: 450ff and Borcard et al. 2011: 130f. for details.

Brief tutorial for PCA

1. Check if conditions for descriptors are met (quantitative, multivariate normality, linear)

2. Conduct PCA (or sparse PCA) on scaled descriptors unless they exhibit a similar variation and have been measured on similar scale

3. Check for outliers

4. Determine number of principal components

5. How informative are first two PCs?

6. Which descriptors contribute most to PCs?

7. Visualise and interpret

A detailed key for PCA is given in Legendre & Legendre 2012: 453

Principal component regression

- Extract (unscaled) PC scores to use PCs as descriptors in multiple regression analysis

- PCs are orthogonal → fix for multicollinearity in regression analysis

- In low n:p situations, the few last PCs are often removed to reduce number of predictors in regression
  → Can be problematic because low variance of PC does not imply low explanatory power for response variable
  → not necessarily a fix for low n:p ratios

See Jolliffe (2002): 173 for examples where PCs that capture minor amounts of variance exhibit high explanatory power. This is explained when considering that the last few PCs represent the constant elements (low variance) across the descriptors.