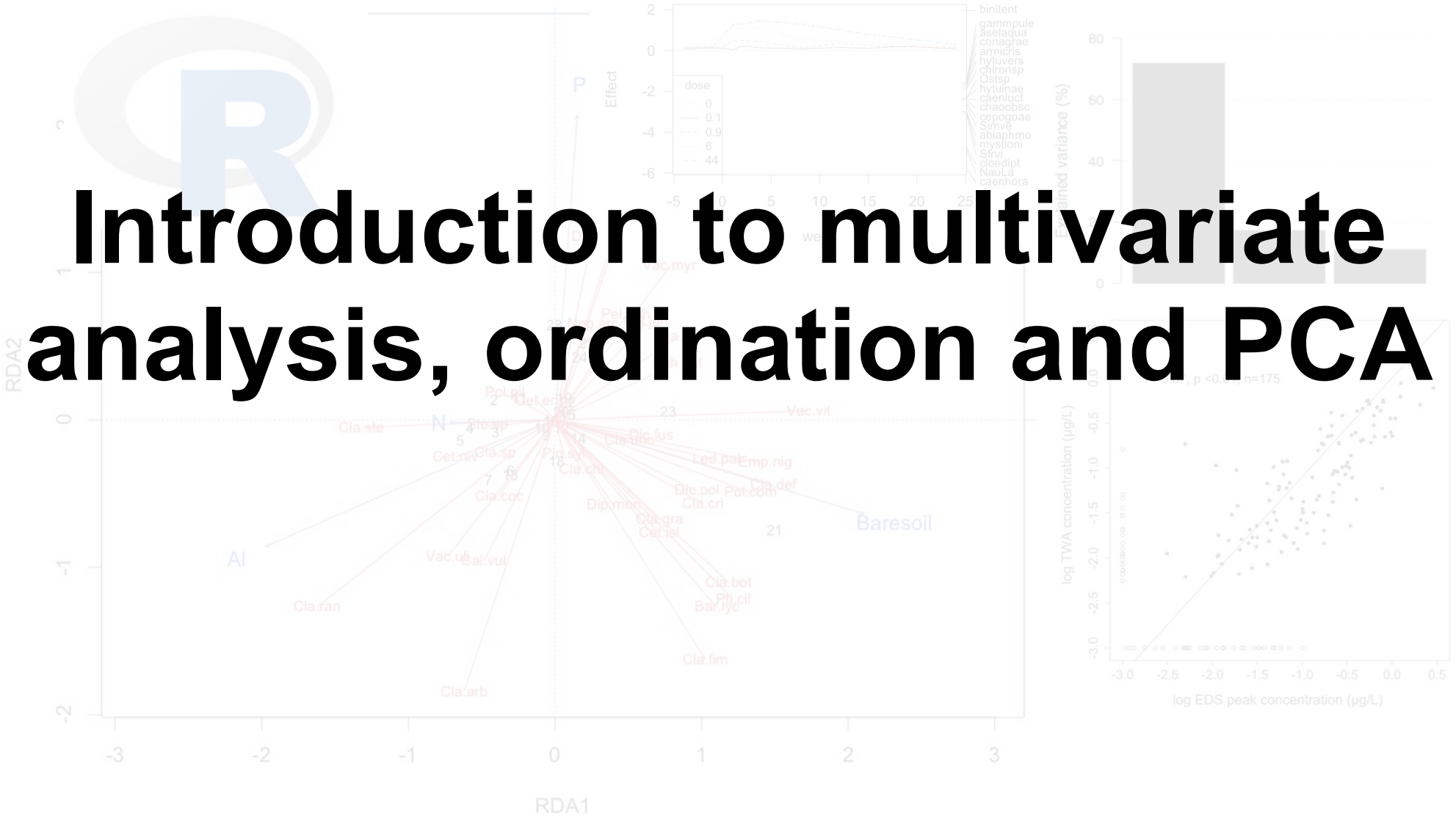
Tools for complex data analysis

University of Koblenz-Landau 2019/20



Introduction to multivariate analysis, ordination and PCA

Ralf B. Schäfer

Learning targets

- Explain the specifics of multivariate analysis
- List and select ordination methods based on research goal
- Explain mathematical basis of 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?
  - What is the difference to the univariate case when diagnosing multivariate normality? What is the covariance matrix?
- List and select ordination methods based on research goal
  - Explain the aims of ordination.
  - Which criteria influence the selection of an ordination method? Discuss the relationship of the criteria to the research goal.

Learning targets and study questions

- Explain mathematical basis of PCA.
  - What are eigenvalues and eigenvectors? Provide a geometrical and algebraic explanation.
  - How do eigenvalues relate to the variance captured by 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 relationship between a) variables, b) sites and c) variables and axes.
  - Outline PCA assumptions and related diagnostic tools.
  - Which objects from a PCA would be extracted as non-collinear predictors for a multiple regression analysis?

Introduction to multivariate analysis, ordination and PCA

Contents

From univariate to multivariate statistics

	univariate	**multivariate**
Variables (vars.)	Single/multiple predictors, single response variable Y	Single/multiple predictors, multiple response vars. $Y_1, \ldots, Y_n$
Distribution of response	One-dimensional	n-dimensional
Data format	Y is vector	$Y_1, \ldots, Y_n$ constitute matrix
Example	Species richness explained by environmental variables	Community explained by environmental variables

WHY DO I HAVE TO LEARN MULTIVARIATE ANALYSES?

www.colourbox.de

Multivariate data analysis: Introduction

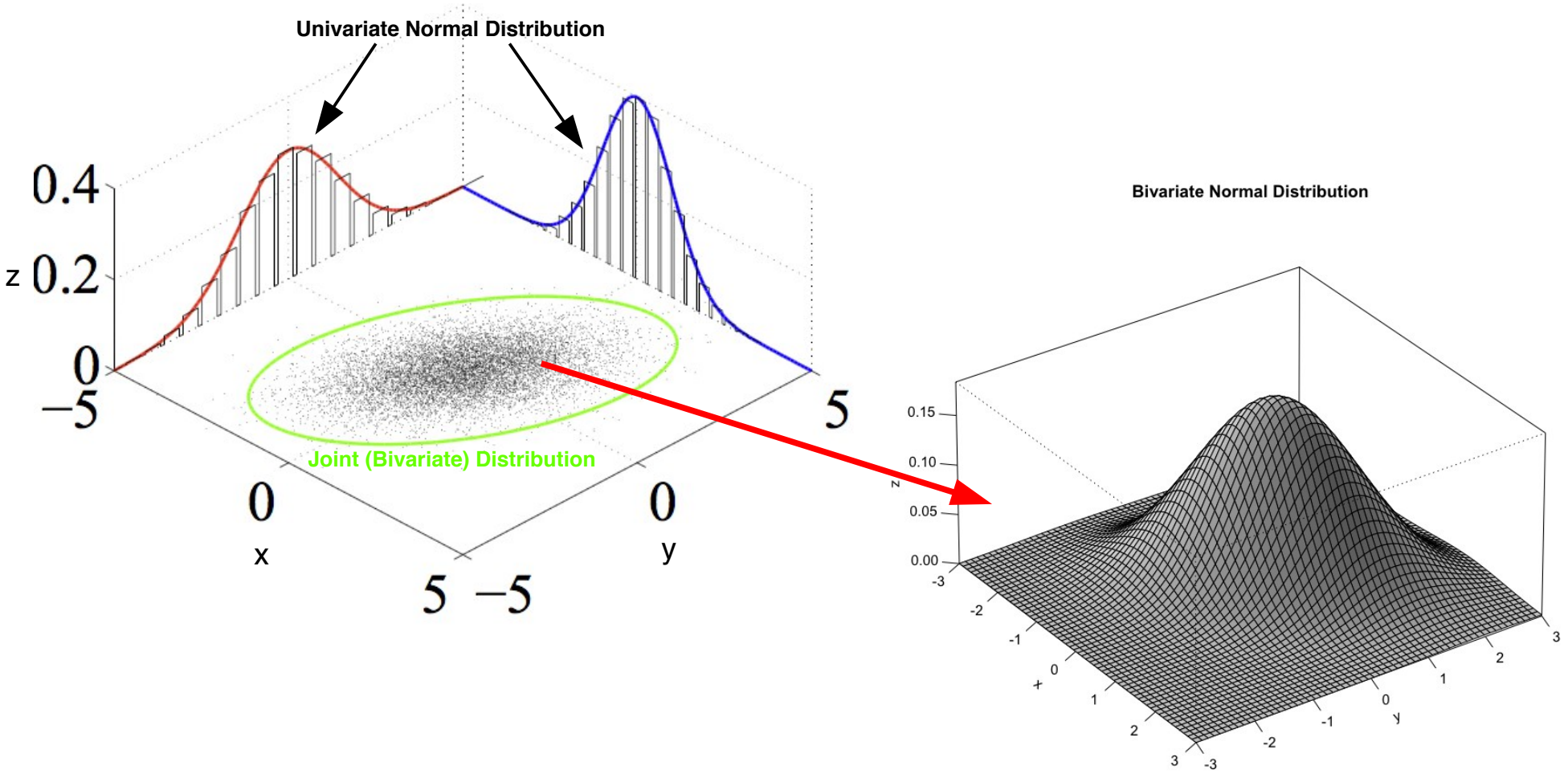
Some advantages of multivariate over univariate methods for analysing multivariate data

- 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 when assessing statistical hypotheses

Specifics of multivariate analysis

Tools and models used in univariate analysis can be used for example for diagnosing multicollinearity, but multivariate response requires adaptation of or new tools and models

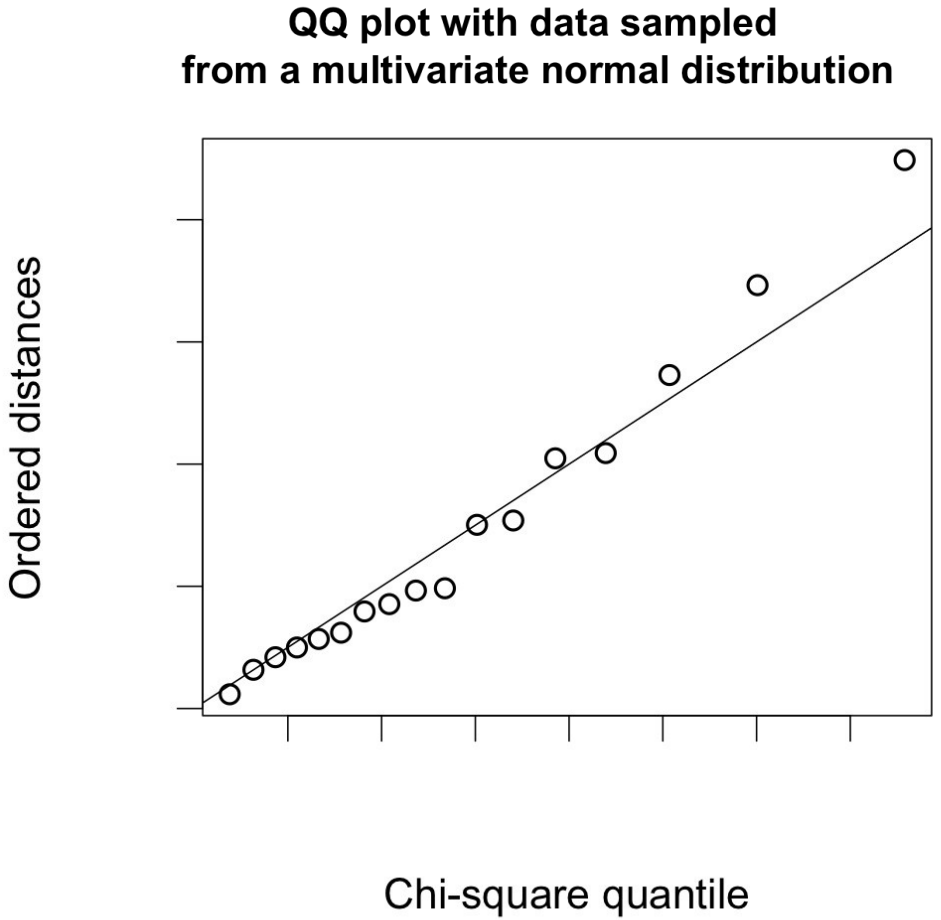
Multivariate normal distribution

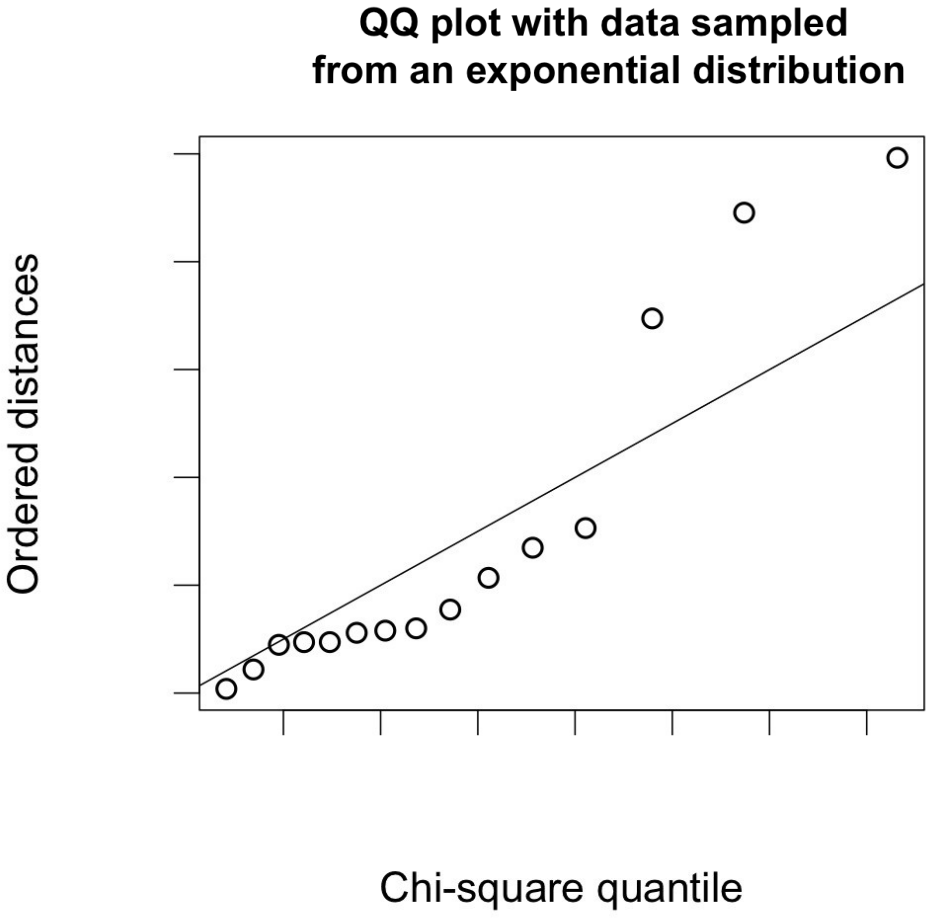


http://commons.wikimedia.org/wiki/File:Multivariate_normal_sample.svg

Specifics of multivariate analysis

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





Covariance matrix

For a data matrix $\boldsymbol{Y}$ containing the variables $Y_1,\ldots,Y_p$

$$\boldsymbol{Y} = \begin{pmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,j} & y_{1,k} & \cdots & y_{1,p} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,j} & y_{2,k} & \cdots & y_{2,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{i,1} & y_{i,2} & \cdots & y_{i,j} & y_{i,k} & \cdots & y_{i,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{n,1} & y_{n,2} & \cdots & y_{n,j} & y_{n,k} & \cdots & y_{n,p} \end{pmatrix}$$

the sample covariance matrix $\boldsymbol{S}$, which is an estimate of $\boldsymbol{\Sigma}$, is:

$$\hat{\Sigma} = \boldsymbol{S} = \begin{pmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,j} & s_{1,k} & \cdots & s_{1,p} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,j} & s_{2,k} & \cdots & s_{2,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{j,1} & s_{j,2} & \cdots & s_{j,j} & s_{j,k} & \cdots & s_{j,p} \\ s_{k,1} & s_{k,2} & \cdots & s_{k,j} & s_{k,k} & \cdots & s_{k,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{p,1} & s_{p,2} & \cdots & s_{p,j} & s_{p,k} & \cdots & s_{p,p} \end{pmatrix}$$

where $s_{j,k} = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_{i,j}-\overline{y_j}\right)\left(y_{i,k}-\overline{y_k}\right)$

Covariance and Mahalanobis distance

In the diagonal, the equation simplifies to:

$$s_{j,j} = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_{i,j}-\overline{y_j}\right)^2$$

$$\hat{\Sigma} = \boldsymbol{S} = \begin{pmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,j} & s_{1,k} & \cdots & s_{1,p} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,j} & s_{2,k} & \cdots & s_{2,p} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ s_{j,1} & s_{j,2} & \cdots & s_{j,j} & s_{j,k} & \cdots & s_{j,p} \\ s_{k,1} & s_{k,2} & \cdots & s_{k,j} & s_{k,k} & \cdots & s_{k,p} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ s_{p,1} & s_{p,2} & \cdots & s_{p,j} & s_{p,k} & \cdots & s_{p,p} \end{pmatrix}$$

Covariances

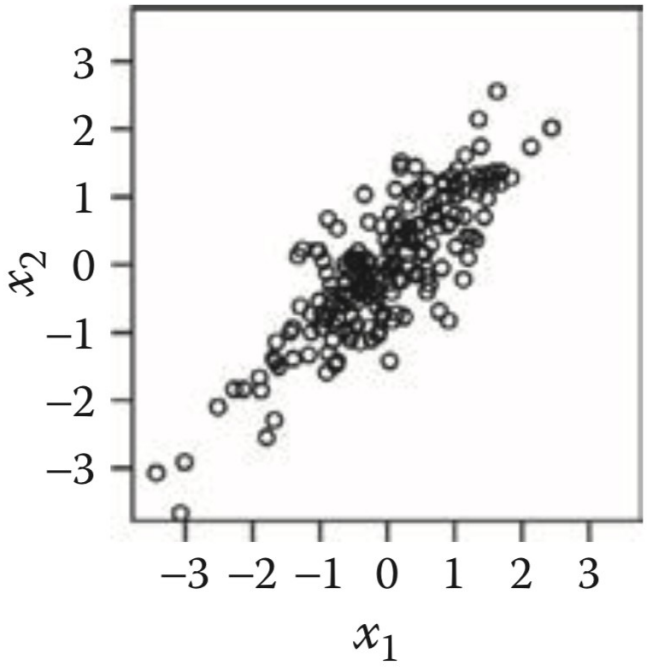
Variances

The Mahalanobis distance d_{M} incorporates $\boldsymbol{\Sigma}$ when measuring the multivariate distance between two vectors. For example, d_{M} for the observation x to the mean vector μ is:

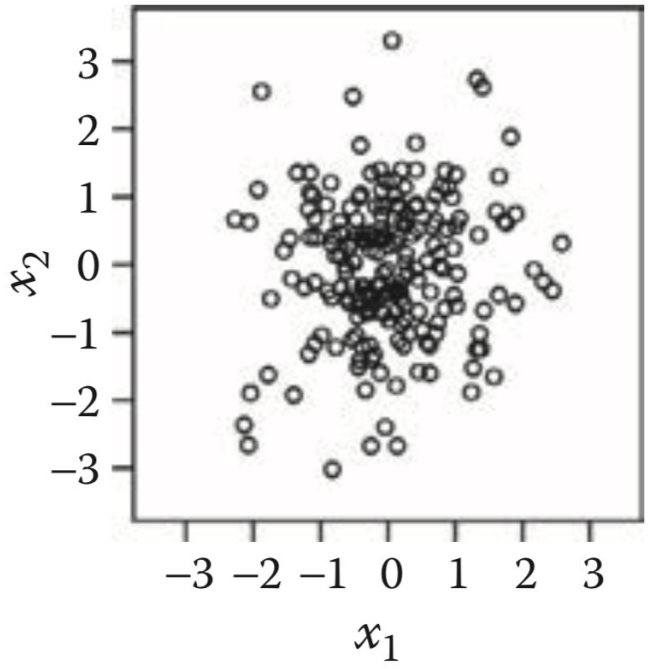
$$d_{\mathrm{M}} = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

$\rightarrow$ can be used to detect outliers

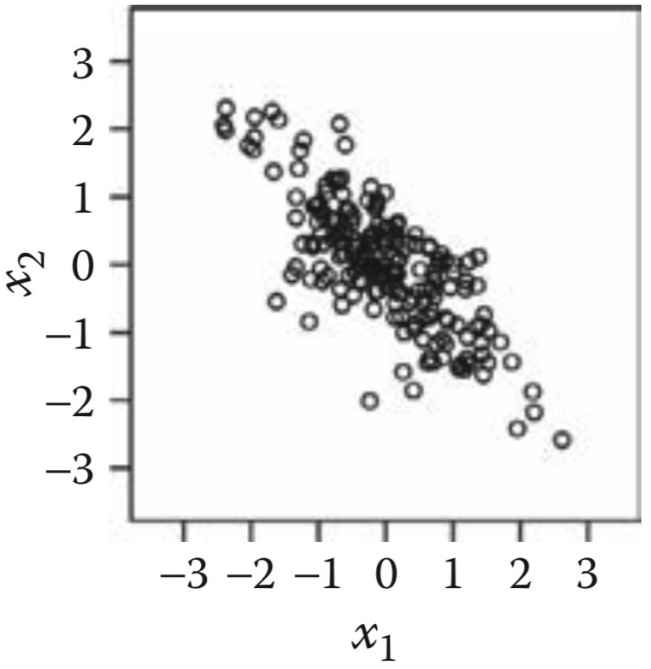
Covariance matrices for two variables



$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$$



$$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$



$$\boldsymbol{\Sigma}_3 = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 1 \end{pmatrix}$$

Varmuza & Filzmoser 2009: 40

Multivariate approaches in R

Available methods and developments: CRAN Task View

CRAN Task View: Multivariate Statistics

Maintainer: Paul Hewson
Contact: Paul.Hewson at plymouth.ac.uk
Version: 2018-07-21
URL: https://CRAN.R-project.org/view=Multivariate

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.

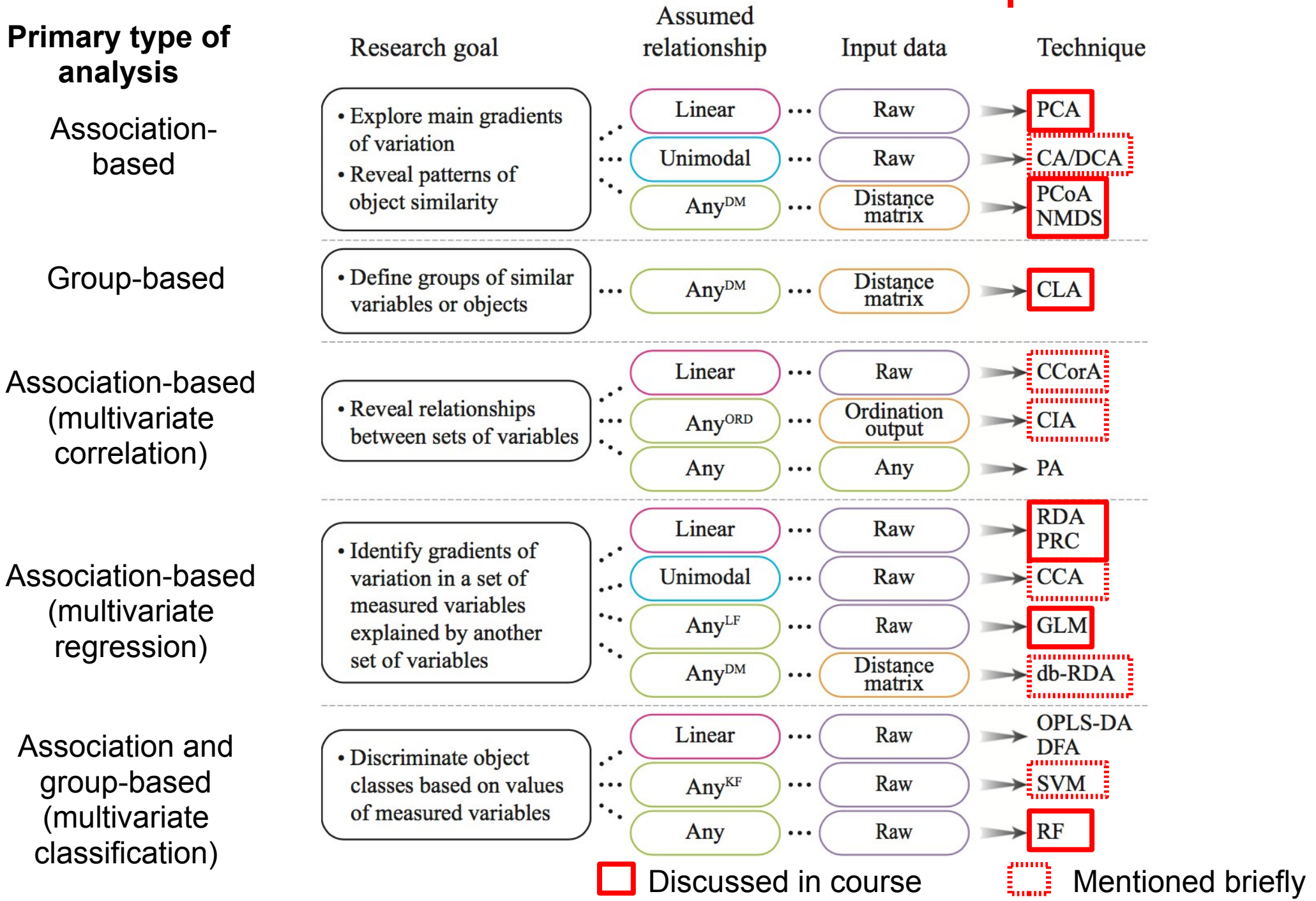
Hypothesis testing

- 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 `covMCD()` and the Orthogonalized pairwise estimate of Gnanadesikan-Kettenring via `covOGK()`. Scalable robust methods are provided within rrcov also using fast minimum covariance determinant with `covMcd()` as well as M-estimators with `covMest()`. corpcor provides shrinkage estimation of large scale covariance and (partial) correlation matrices.
- *Densities (estimation and simulation):* `mvnorm()` in MASS simulates from the multivariate normal distribution. mvtnorm also provides simulation as well as probability and quantile functions for both the multivariate t distribution and multivariate normal distributions as well as density functions for the multivariate normal distribution. mnormt provides multivariate normal and multivariate t density and distribution functions as well as random number simulation. sn provides density, distribution and random number generation for the multivariate skew normal and skew t distribution. delt provides a range of functions for estimating multivariate densities by CART and greedy methods. Comprehensive information on mixtures is given in the Cluster view, some density estimates and random numbers are provided by `rmvnorm.mixt()` and `dmvnorm.mixt()` in ks, mixture fitting is also provided within bayesm. Functions to simulate from the Wishart distribution are provided in a number of places, such as `rwishart()` in bayesm and `rwish()` in MCMCpack (the latter also has a density function `dwish()`). `bkde2D()` from KernSmooth and `kde2d()` from MASS provide binned and non-binned 2-dimensional kernel density estimation, ks also provides multivariate kernel smoothing as does ash and GenKern. prim provides patient rule induction methods to attempt to find regions of high density in high dimensional multivariate data, feature also provides methods for determining feature significance in multivariate data (such as in relation to local modes).
- *Assessing normality:* mvnormtest provides a multivariate extension to the Shapiro-Wilks test, mvoutlier provides multivariate outlier detection based on robust methods. ICS provides tests for multi-normality. `mvnorm.etest()` in energy provides an assessment of normality based on E statistics (energy); in the same package `k.sample()` assesses a number of samples for equal distributions. Tests for Wishart-distributed covariance matrices are given by `mauchly.test()` in stats.
- *Copulas:* copula provides routines for a range of (elliptical and archimedean) copulas including normal, t, Clayton, Frank, Gumbel, fgac provides generalised archimedian copula.

Overview multivariate techniques



Paliy & Shankar 2016 *Mol. Ecol.* 25: 1032

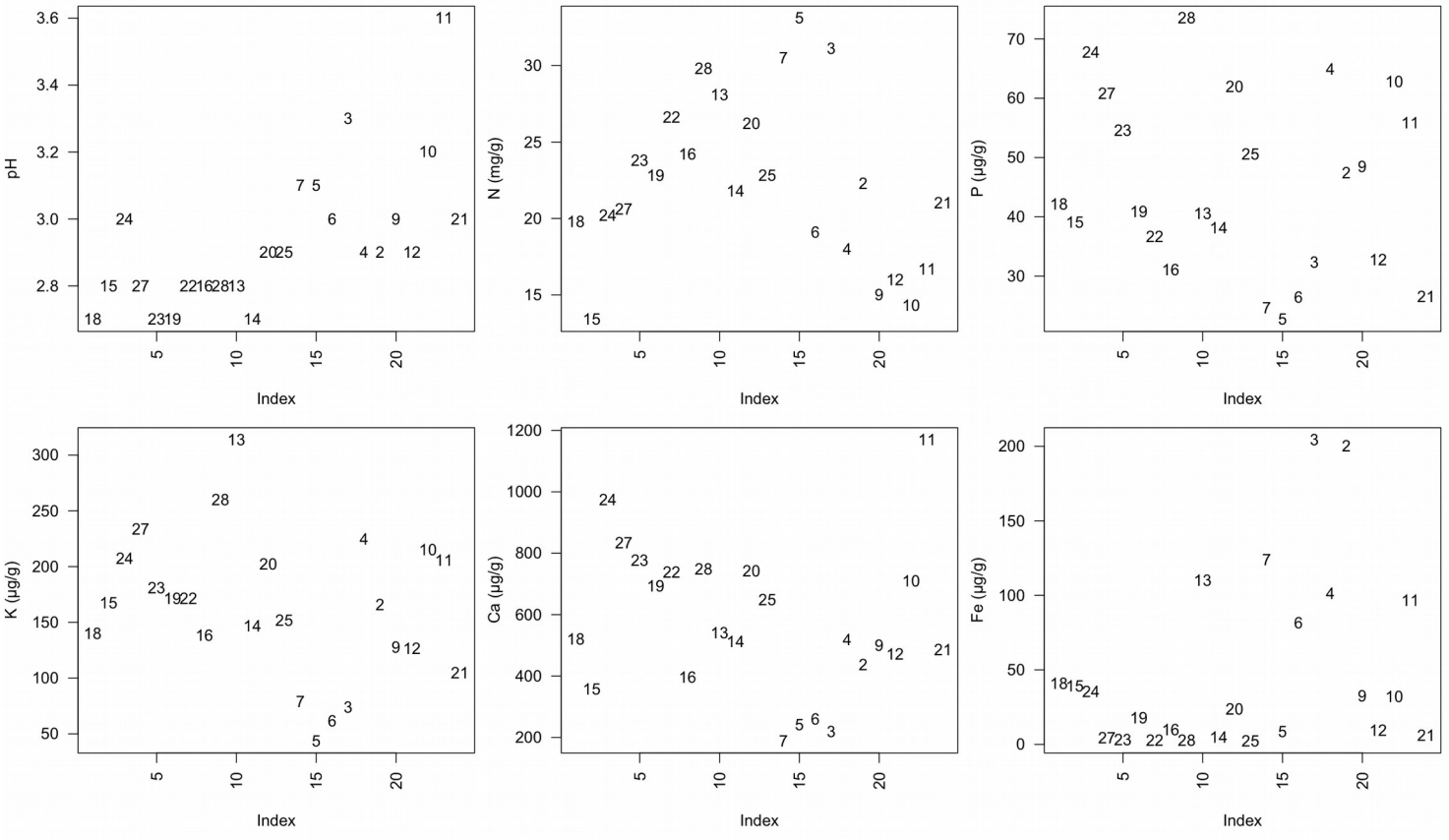
Introduction to multivariate analysis, ordination and PCA

Contents

Ordination: Introduction

- Representation of objects (e.g. sites, samples) along one or multiple axes
- <u>Case study:</u> Measurements of 14 physicochemical variables (e.g. pH, ions) in soil of 24 sites. Research goal: Exploration – what is the main gradient and how are the sites related?

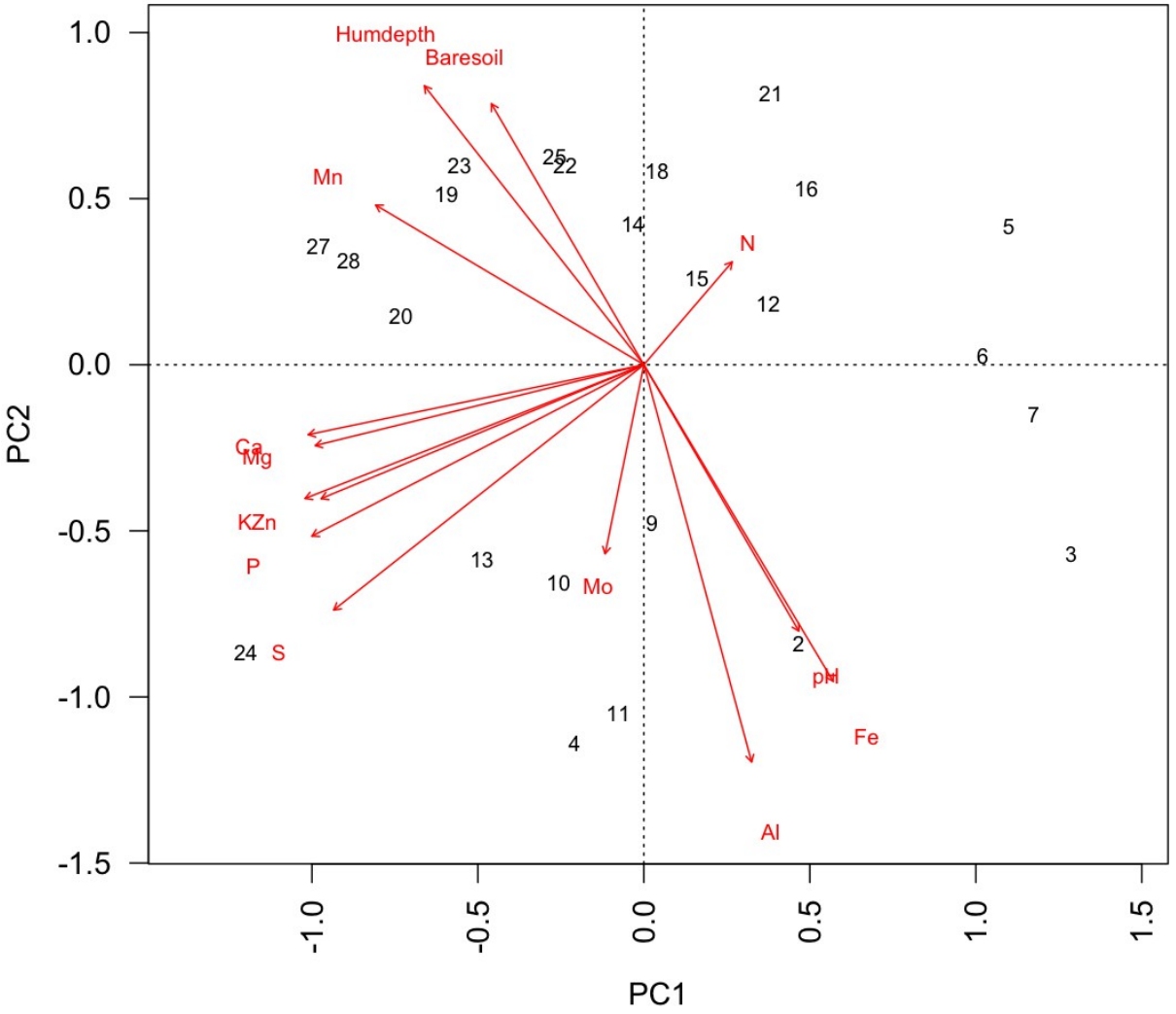
- Univariate approach: inspection of e.g. means and variances, correlations between variables, pattern of sites with respect to variables
- Many different kinds of statistics
  → Difficult to interpret



Ordination: Introduction

- Multivariate approach: Principal Component analysis (PCA)
- Representation of the first two major gradients in the data → Two-dimensional representation of the major variance in 14 dimensions (i.e. variables)

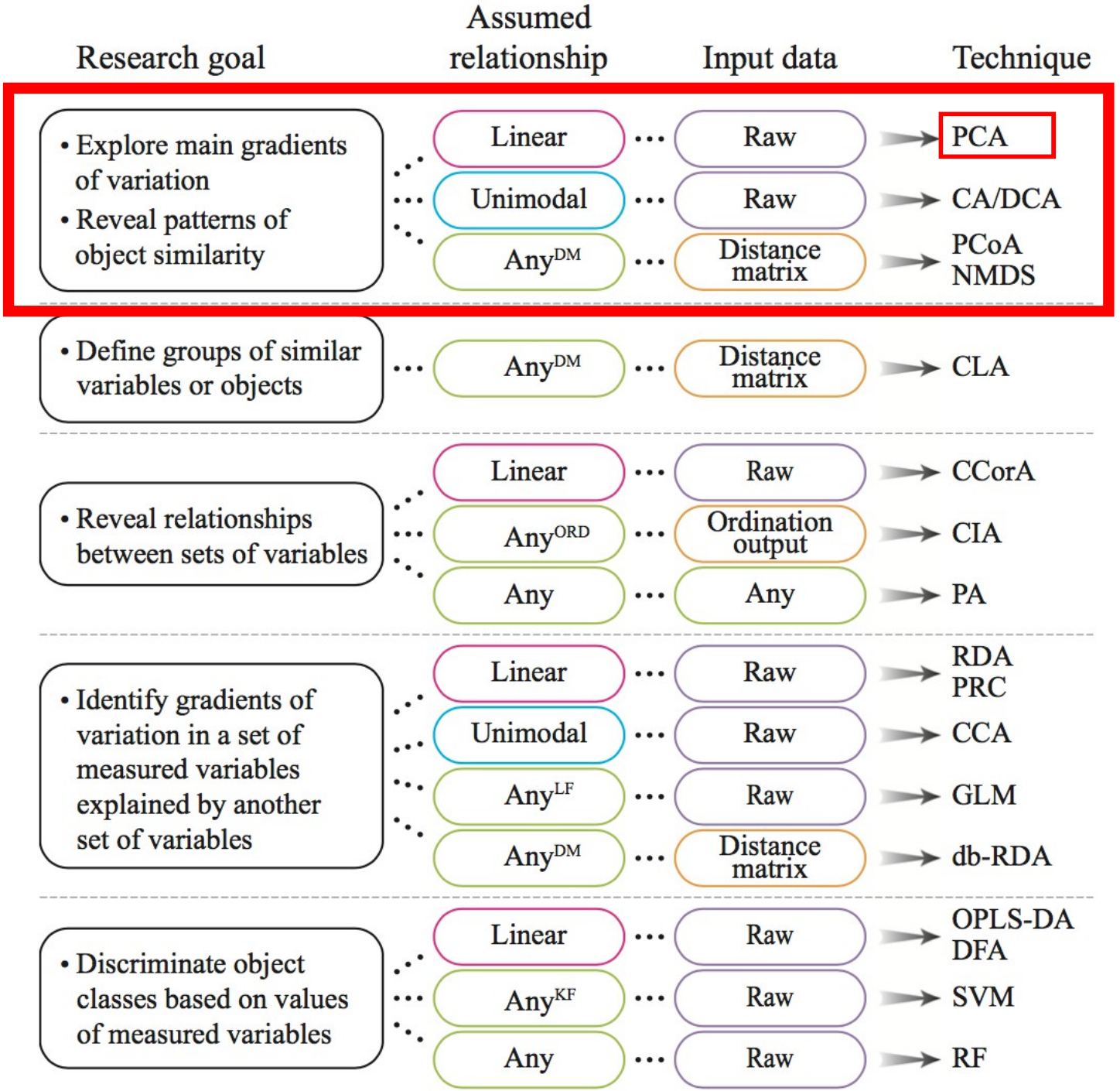
- Ordination plot displays the following information:
  - Contribution of variables to major gradients
  - Relationship between variables
  - Relationship between sites
  - Pattern of sites with respect to variables
- Easier to interpret!



Aims of ordination

- Dimension reduction – (graphical) representation of data in reduced (lower-dimensional) space (and potentially omission of variables/gradients/axes that capture low amount of variance)
- Aggregation of variables into gradients and extraction of main gradients
- Constrained ordination: extraction of gradients that are explained by variables of second data set. Supervised learning.
- Unconstrained ordination: extraction without consideration of variables outside of data set. Unsupervised learning.

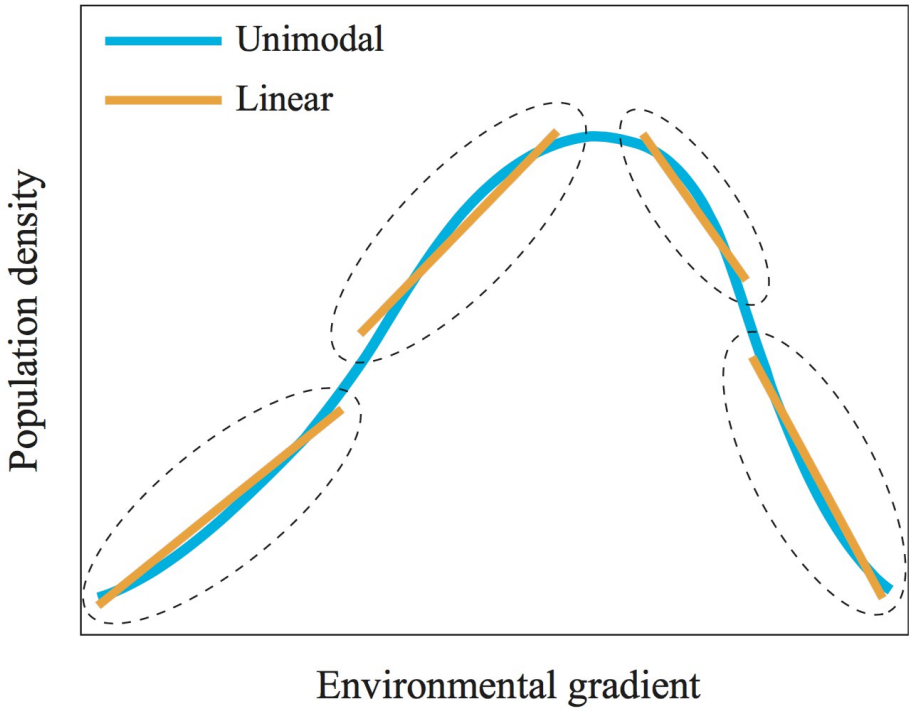
Unconstrained ordination



Paliy & Shankar 2016 *Mol. Ecol.* 25: 1032

Ordination: Overview

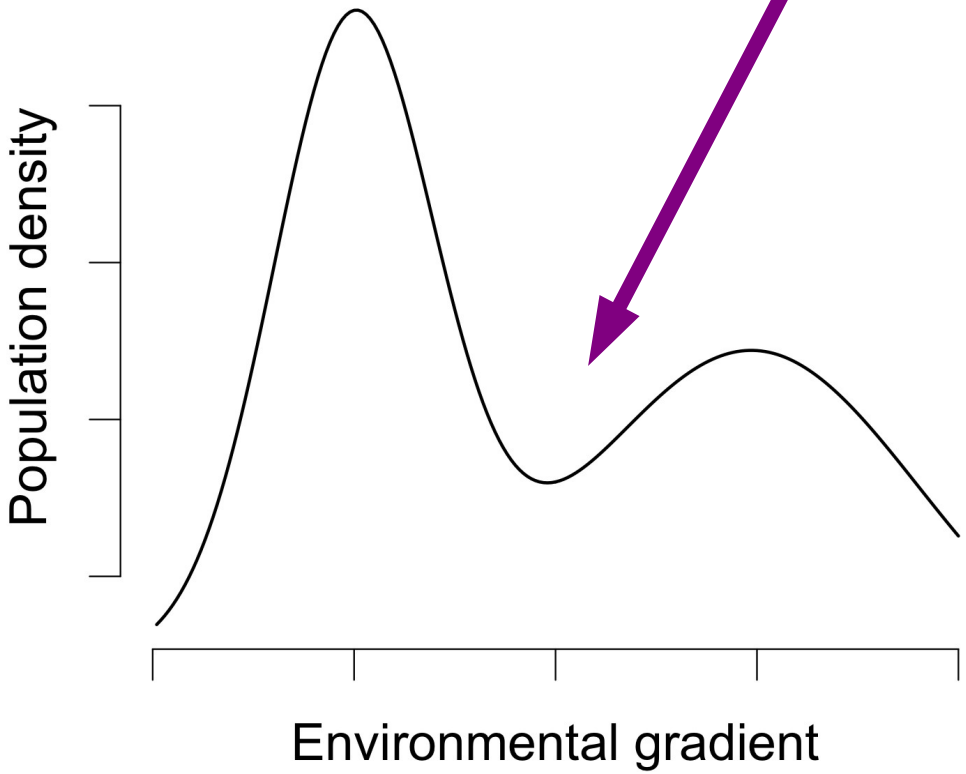
Shape of response	Linear	Unimodal	Any
Unconstrained methods (examples)	PCA	CA	Distance-based: NMDS; GAM-based: UAO (U-VGAM)
Constrained methods (examples)	RDA	CCA	Distance-based: db-RDA; GAM-based: CAO (RR-VGAM)



Paliy & Shankar 2016 *Mol Ecol*:1032

Ordination: Overview

Shape of response	Linear	Unimodal	Any
Unconstrained methods (examples)	PCA	CA	Distance-based: NMDS; GAM-based: UAO (U-VGAM)
Constrained methods (examples)	RDA	CCA	Distance-based: db-RDA; GAM-based: CAO (RR-VGAM)



Introduction to multivariate analysis, ordination and PCA

Contents

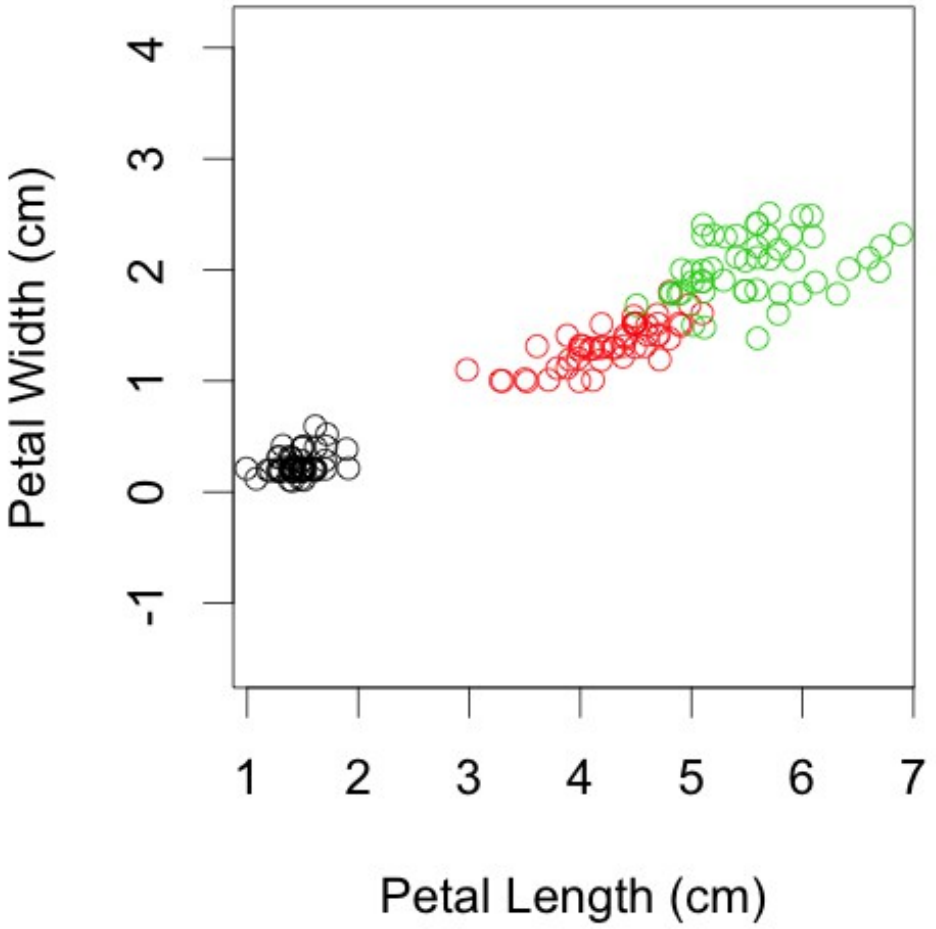
Principal Component Analysis

Example-based introduction

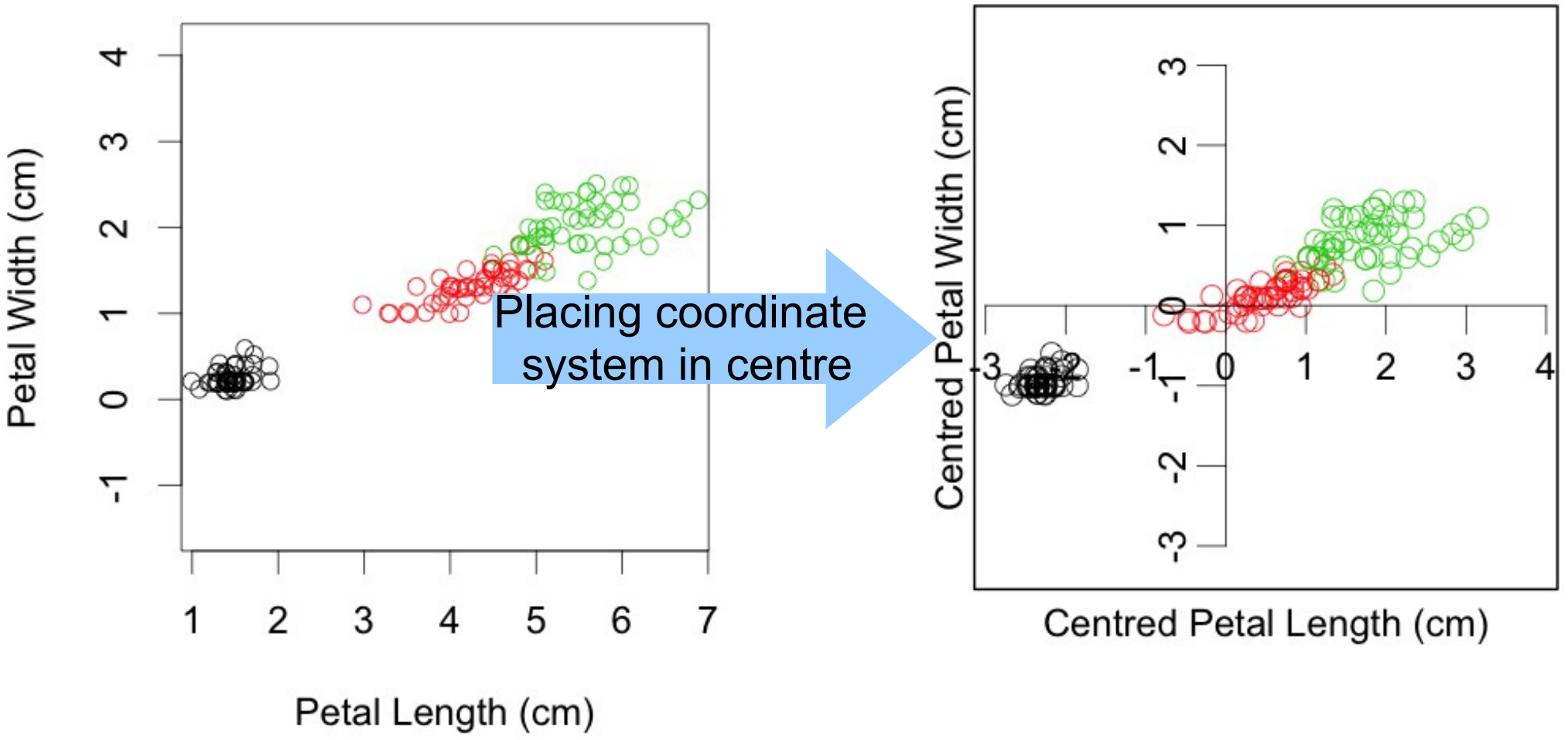
Iris data set: sepal length & width, petal length & width for 50 flowers from 3 species of Iris. Visual demonstration for 2 variables.

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

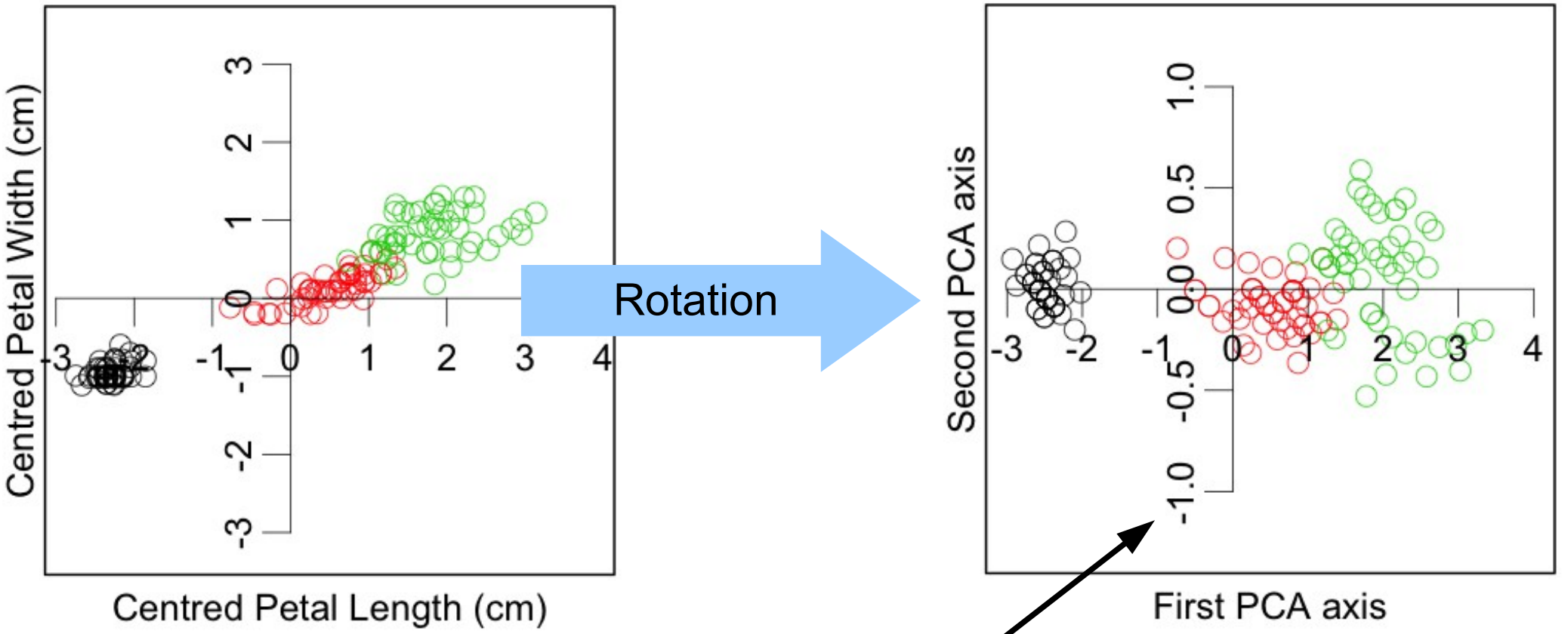
Goal: Dimension reduction. Represent as much variance as possible with first few axes



Introduction to PCA

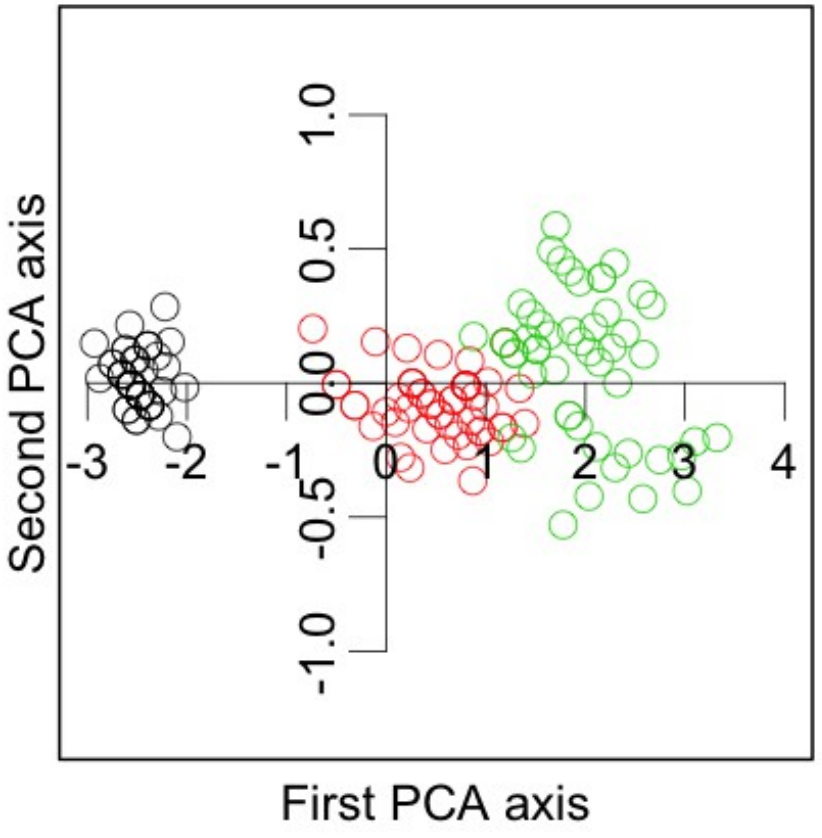


Introduction to PCA



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

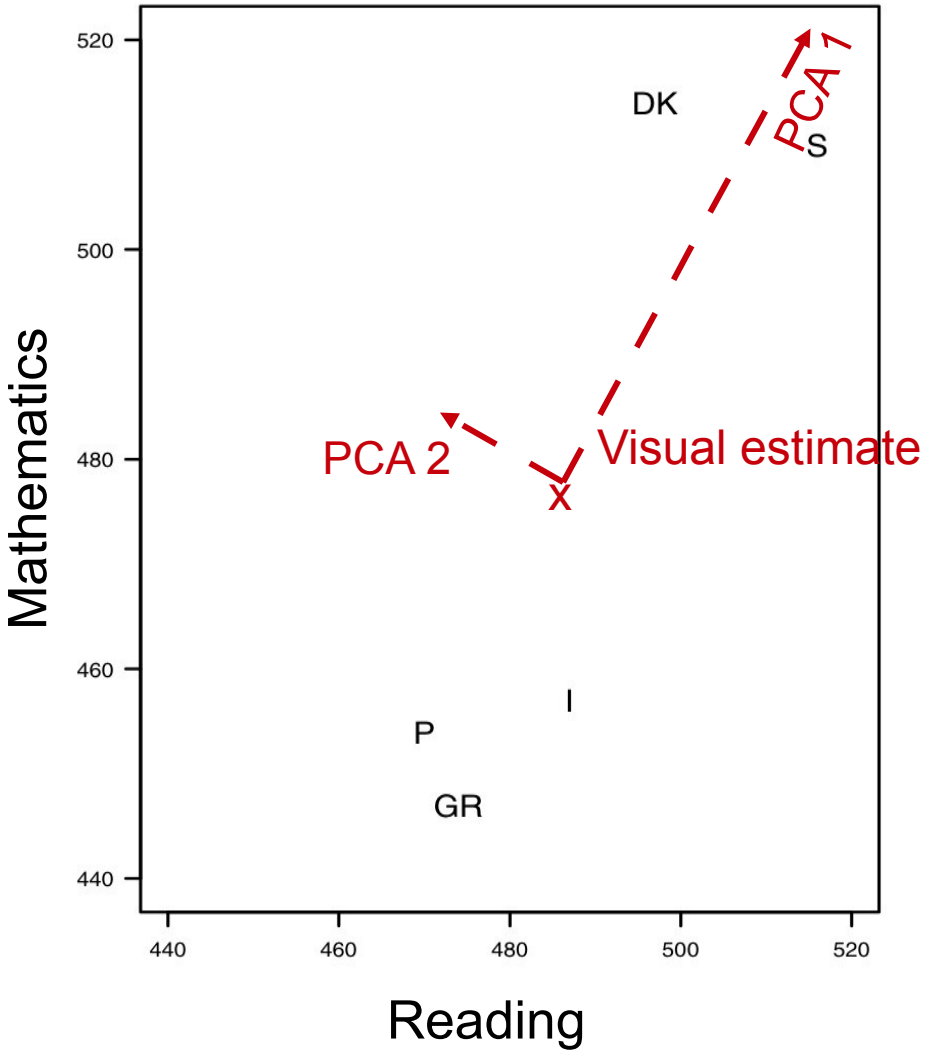
What is an Eigenvalue?

First axis captures 99% of variance!

Mathematical background of PCA

Example: Scores in international education survey (PISA)

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



Centering leads to

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

Search for first axis with maximum variation!

Example taken from Handl (2010)

Mathematical background of PCA

Search for first axis with maximum variation!

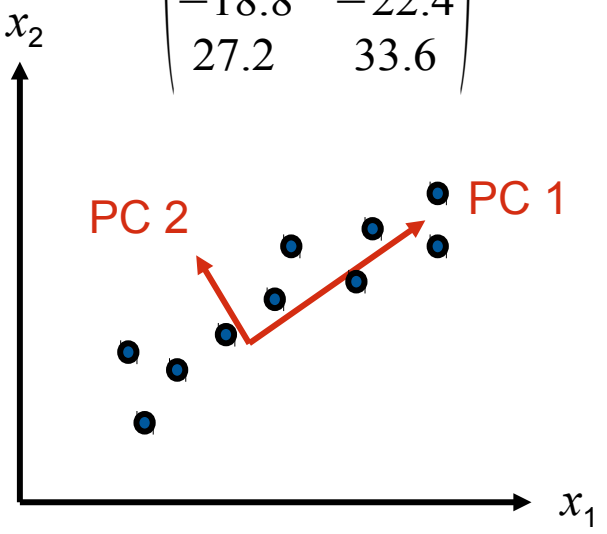
$$\widetilde{\boldsymbol{X}} = \begin{array}{c} \begin{matrix} \widetilde{X}_1 & \widetilde{X}_2 \end{matrix} \\ \begin{pmatrix} 8.2 & 37.6 \\ -14.8 & -29.4 \\ -1.8 & -19.4 \\ -18.8 & -22.4 \\ 27.2 & 33.6 \end{pmatrix} \end{array}$$

Scores on new axis given as

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

and maximum variation means:

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



Generalise problem: define a_1, a_2 as elements of vector a

and likewise $\widetilde{X}_1, \widetilde{X}_2$ of matrix $\boldsymbol{X}$: $\arg\max_{a} \ \mathrm{Var}\left(a \widetilde{\boldsymbol{X}}\right)$

Trivial solution: choose high values for $a_1, a_2, \ldots, a_n$

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

Mathematical background of PCA

Solve:

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

This can be expressed as (see Handl & Kuhlenkasper 2017: 87)

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

Covariance matrix

Using the Lagrange function yields:

$$L(a,\lambda) = a^T \Sigma\, a - \lambda\left(a^T a - 1\right)$$

Eigenvalue problem

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

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

Introduction to multivariate analysis, ordination and PCA

Contents

Mathematical basics: Eigenvalues

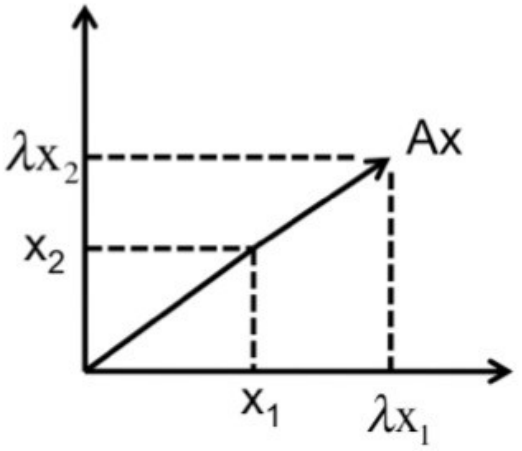
Idea: Conversion of matrix into a matrix with linear independent variables

$$\begin{pmatrix} a_{1,1} & \dots & a_{1,p} \\ \dots & a_{2,2} & \dots \\ a_{n,1} & \dots & a_{n,p} \end{pmatrix} \quad \text{Conversion} \Rightarrow \quad \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_p \end{pmatrix}$$

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: Eigenvalues

$$\boldsymbol{A}x = \lambda x \quad \Leftrightarrow \quad \boldsymbol{A}x - \lambda x = 0$$

$$\Leftrightarrow \quad \left(\boldsymbol{A} - \lambda\, \boldsymbol{I}\right)x = 0$$

$$\Leftrightarrow \begin{pmatrix} a_{1,1}-\lambda_1 & \dots & a_{1,p} \\ \dots & \dots & \dots \\ a_{n,1} & \dots & a_{n,p}-\lambda_p \end{pmatrix} \begin{pmatrix} x_1 \\ \dots \\ x_n \end{pmatrix} = \begin{pmatrix} 0 \\ \dots \\ 0 \end{pmatrix}$$

Homogeneous linear equation system (HLS)

Ignore trivial solution: $x = 0$ $\quad \Rightarrow \quad \boldsymbol{A} - \lambda\, \boldsymbol{I} = 0$

The given HLS has only a non-trivial solution if the columns of $\boldsymbol{A} - \lambda\, \boldsymbol{I}$ are linearly dependent, which is the case if the determinant = 0.

$$\Rightarrow \quad det\left(\boldsymbol{A} - \lambda\, \boldsymbol{I}\right) = 0 \quad \Leftrightarrow \quad \left|\boldsymbol{A} - \lambda\, \boldsymbol{I}\right| = 0$$

Example I: Calculation of Eigenvalues

Sample Variance-Covariance matrix from PISA example:

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

Following $|\boldsymbol{A} - \lambda \boldsymbol{I}| = 0$ we obtain:

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

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

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

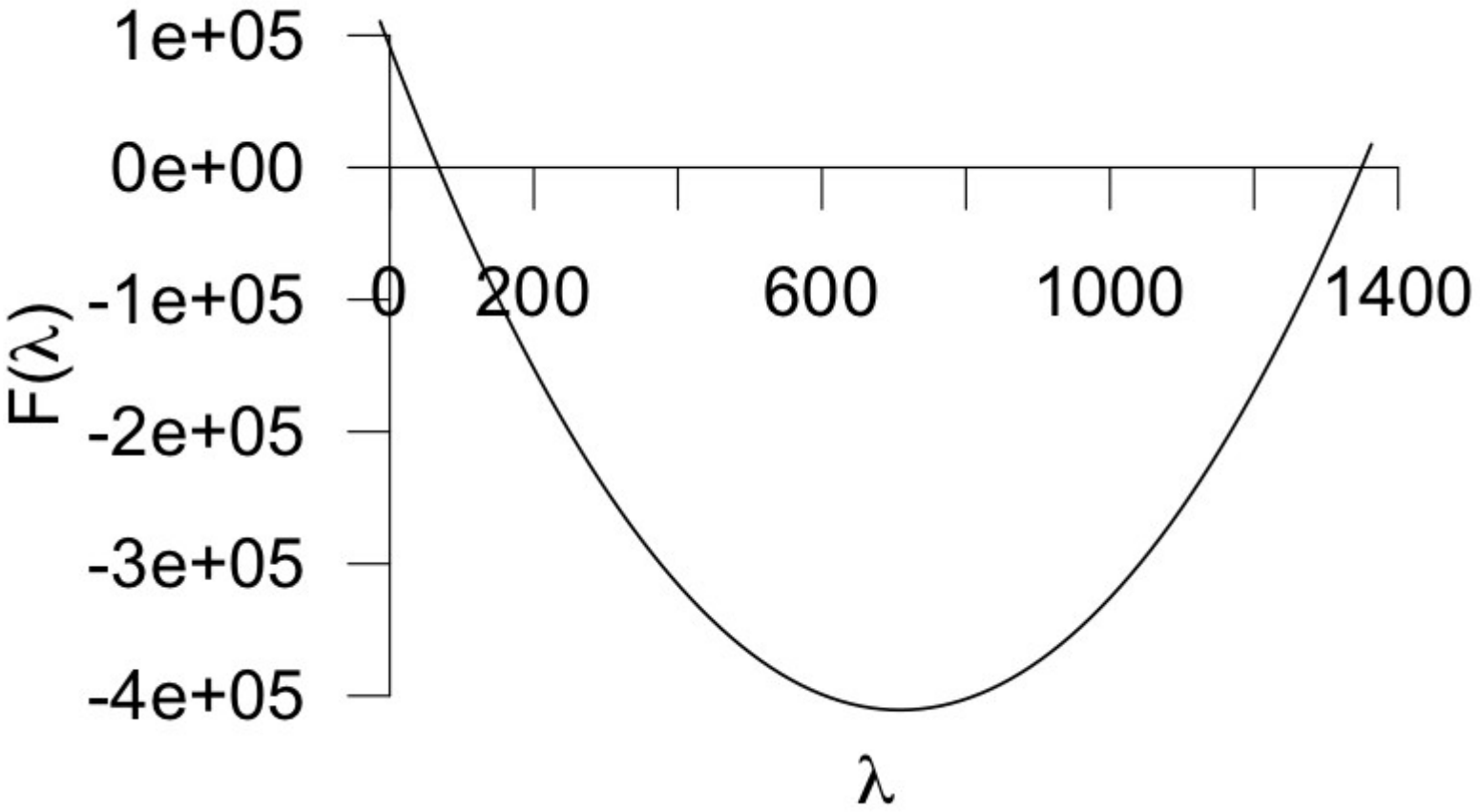
$$(345.7 - \lambda)(1071.30 - \lambda) - 528.35^2 = 0$$

Example I: Calculation of Eigenvalues

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

Characteristic polynomial

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



Example II: Calculation of Eigenvalues and -vectors

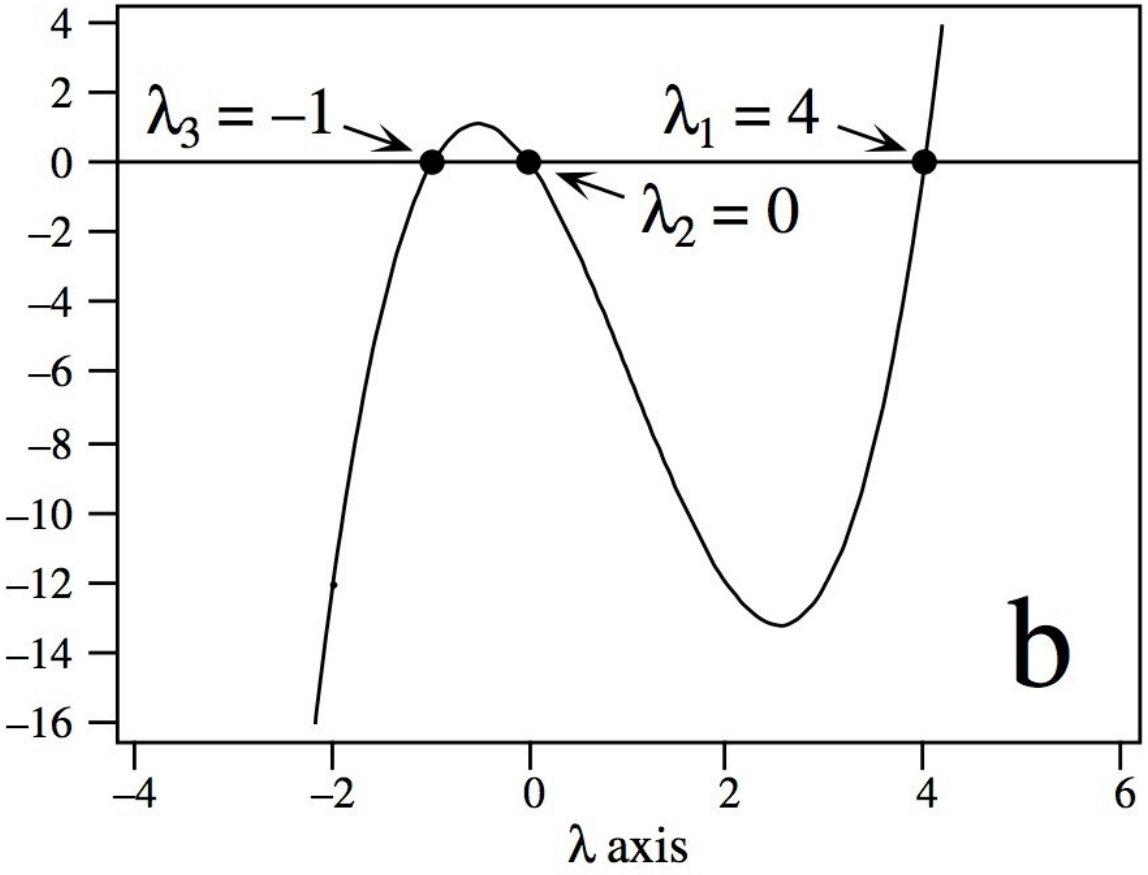
$$\begin{pmatrix} 1 & 3 & -1 \\ 0 & 1 & 2 \\ 1 & 4 & 1 \end{pmatrix}$$

EV calculation following Sarrus

Characteristic polynomial

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

Eigenvalues λ: 4, 0 and -1



Legendre & Legendre 2012: 93

Example II: Calculation of Eigenvalues and -vectors

Calculation of eigenvector for $\lambda = 4$

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

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

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

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

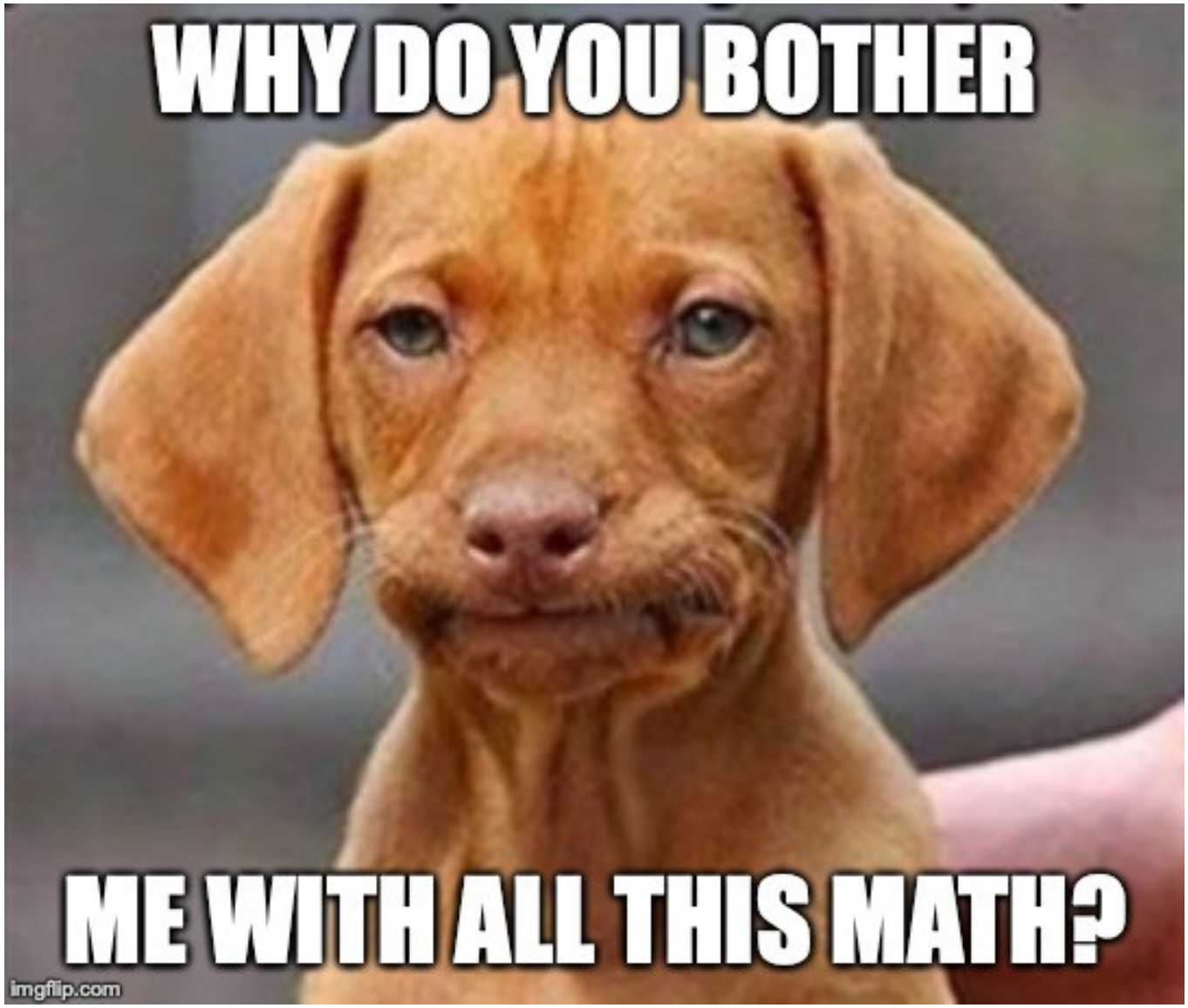
Example II: Calculation of Eigenvalues and -vectors

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

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


WHY DO YOU BOTHER
ME WITH ALL THIS MATH?
imgflip.com

Introduction to multivariate analysis, ordination and PCA

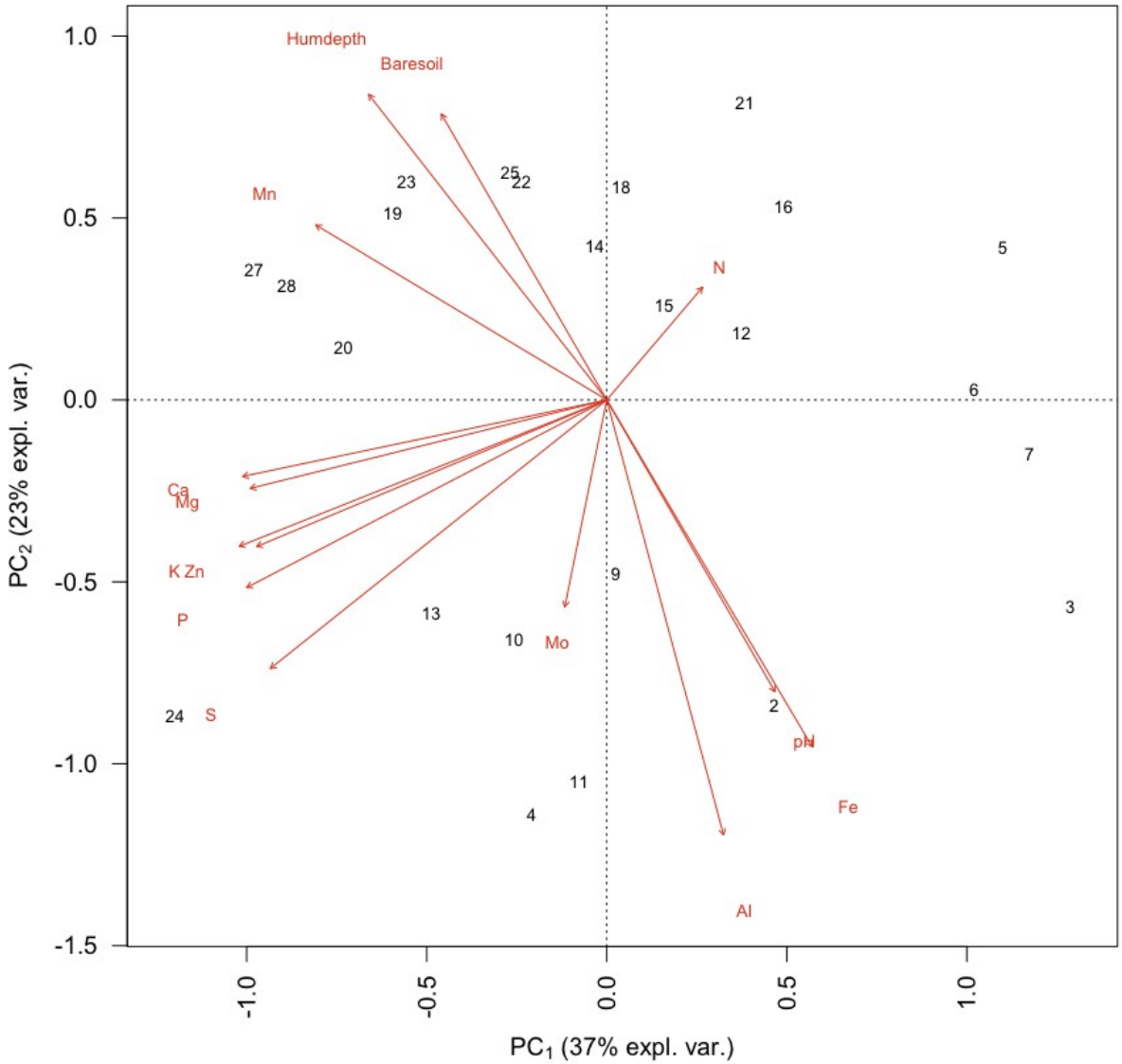
Contents

1. Introduction and specifics of multivariate analysis
2. Overview ordination
3. Introduction to PCA
4. Mathematical background
5. **PCA results and interpretation**
6. PCA diagnosis, tutorial and extensions

Biplot for PCA on 24 soil samples

- First two axes (gradients) capture 60% of variance
- Plot displays following information:

- Position of (site) labels given by scores on (new) axes
- Units on axes = standard deviations
- Contribution of variables to major gradients (the longer the arrow along an axis, the higher the correlation (and representation by axis))
- Relationship between variables (angle between arrows:
  0° or 180° $\rightarrow \rho$ = 1 or -1
  90° $\rightarrow \rho$ = 0
  60° $\rightarrow \rho$ = 0.5)
- Relationship between sites (the closer, the higher similarity (consider variance explained by axes!))
- Pattern of sites with respect to variables (perpendicular projection of site on arrow)



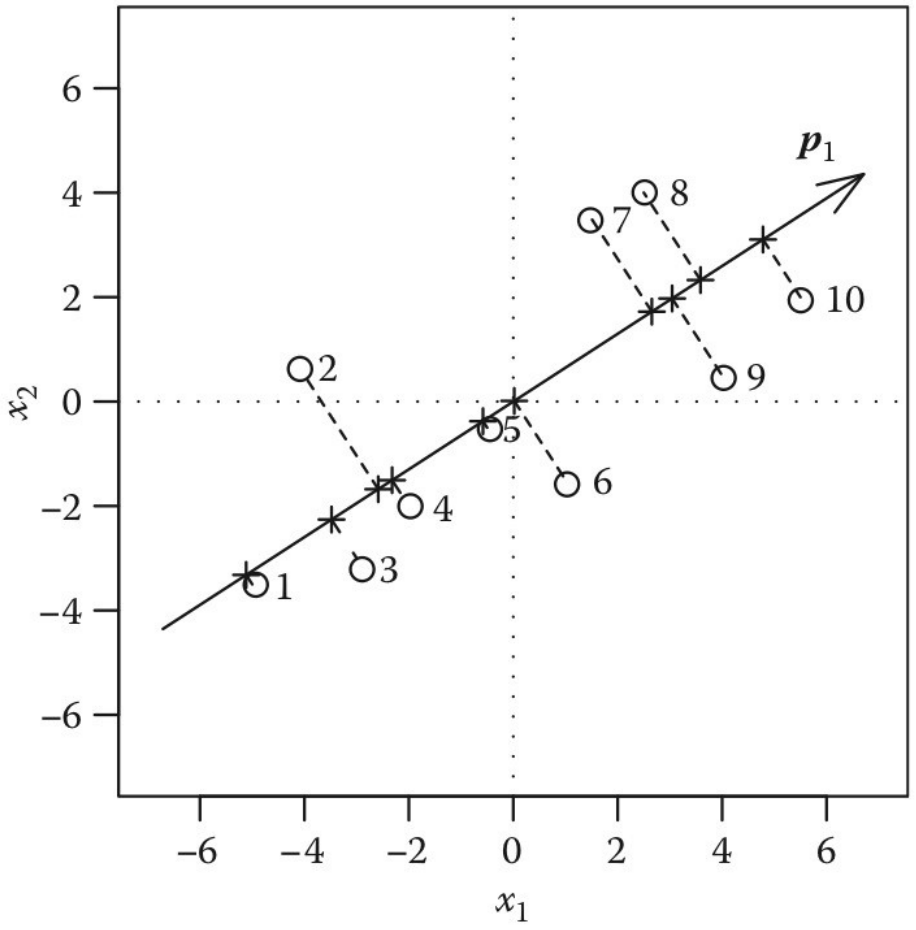
PC scores

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

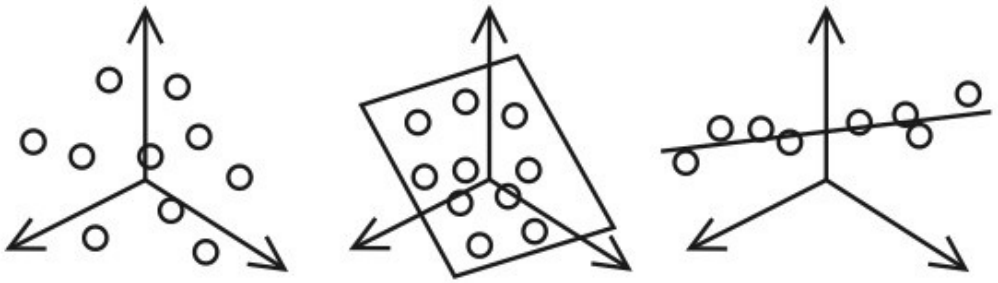
Note: i, Object number; t_1 and t_2 are the PCA scores of PC1 and PC2, respectively; $\bar{x}$, mean; v, variance; v%, variance in percent of total variance.

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



Number of principal components

- Number of descriptors/variables determines number of eigenvalues and thus principal components
- Principal component that relates to the largest eigenvalue captures 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

Varmuza & Filzmoser 2009: Chapter 3

How many principal components needed?

Some criteria to evaluate the optimal number of axes:

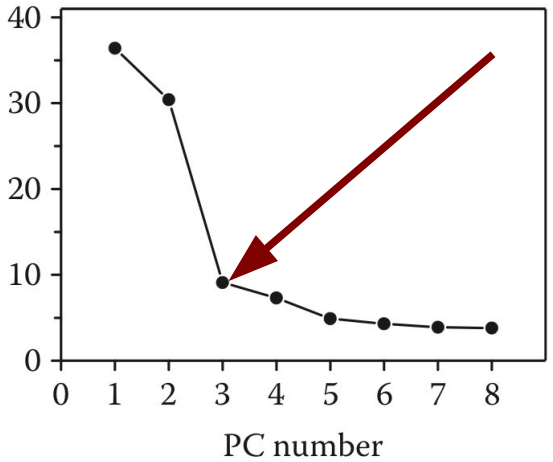
1. Sum criterion

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

2. Broken-Stick criterion

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

3. Scree plot



4. Cross-validation

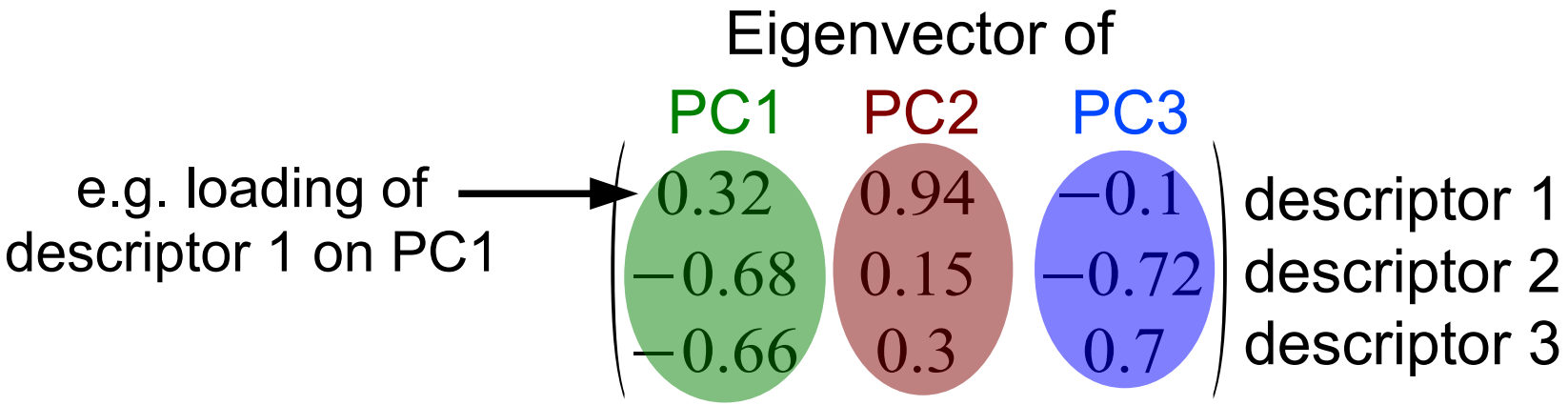
$$\arg\min_{S} \ \mathrm{MSPE}(S) = \frac{1}{I\,K} \sum_{i=1}^{I} \sum_{k=1}^{K} \left(x_{i,k} - \left(\hat{x}_{i,k}\right)^{(S)} \right)^2$$

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

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

Importance of descriptor for PC axes

- Elements of eigenvector matrix = 'loadings', give weight of original descriptor/variable on PC axes



- Easier to interpret: Correlation loadings $r_i = a_i \sqrt{\lambda_i}$
- Interpretation of descriptor/variable importance complicated if many variables load on a PC axis → Sparse PCA – introduces penalty term (cf. LASSO):

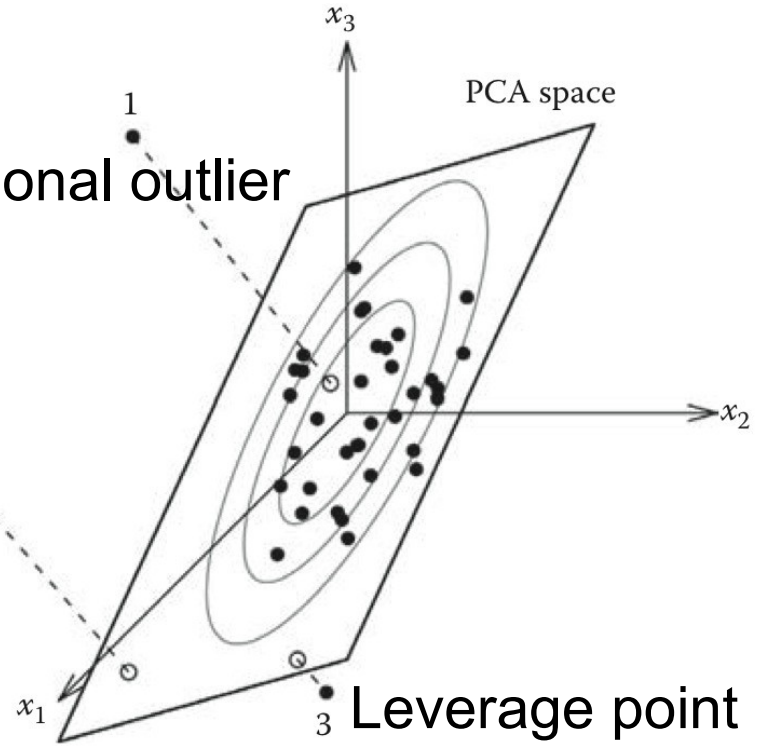
$$\arg\max_{a} \left(a^T \Sigma \; a\right) - \lambda_1 \|a\| \quad \text{with} \quad a^T a = 1$$

Introduction to multivariate analysis, ordination and PCA

Contents

PCA assumptions and diagnosis

- Independence of observations (temporal and spatial independence)
- Multivariate normality (depending on research goal)
- No serious outliers, diagnosis via orthogonal distance (OD) and score distance (SD)
- Alternative: robust PCA



Problematic

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 distance used in PCA inappropriate for species data
- Alternatives: Transformation of data (Hellinger or Chord) or using different ordination method (e.g. NMDS)
- Adding noise variables to data increases fraction of variance on first axis, but has no meaning
- Best results for large n and high n:p (p = descriptors)

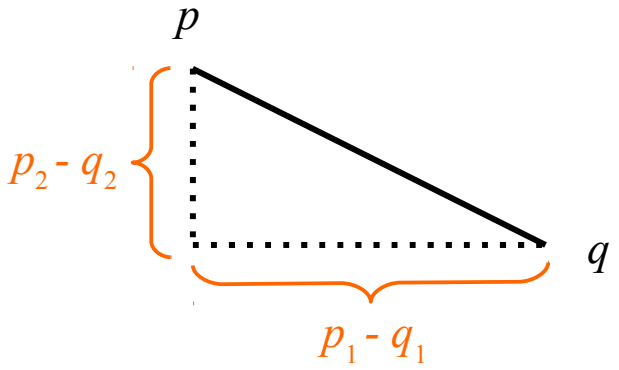
Euclidean distance and species data

$$d_{\text{Euclidean}}(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

Widely used distance measure

Two dimensional case:

$$d_{\text{Euclidean}}(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}$$



Species x Site matrix

Sites	Species		
	y_1	y_2	y_3
x_1	0	1	1
x_2	1	0	0
x_3	0	4	4

Euclidean distance →

Distance matrix

Sites	Sites		
	x_1	x_2	x_3
x_1	0	1.732	4.243
x_2	1.732	0	5.745
x_3	4.243	5.745	0

Sites x_1 and x_2 share any species, but have a smaller distance than sites sharing species (x_1 and x_3) → "Species abundance paradox"

→ Euclidean distance problematic for ecological data

Legendre & Legendre 2012: 300

Brief tutorial for PCA

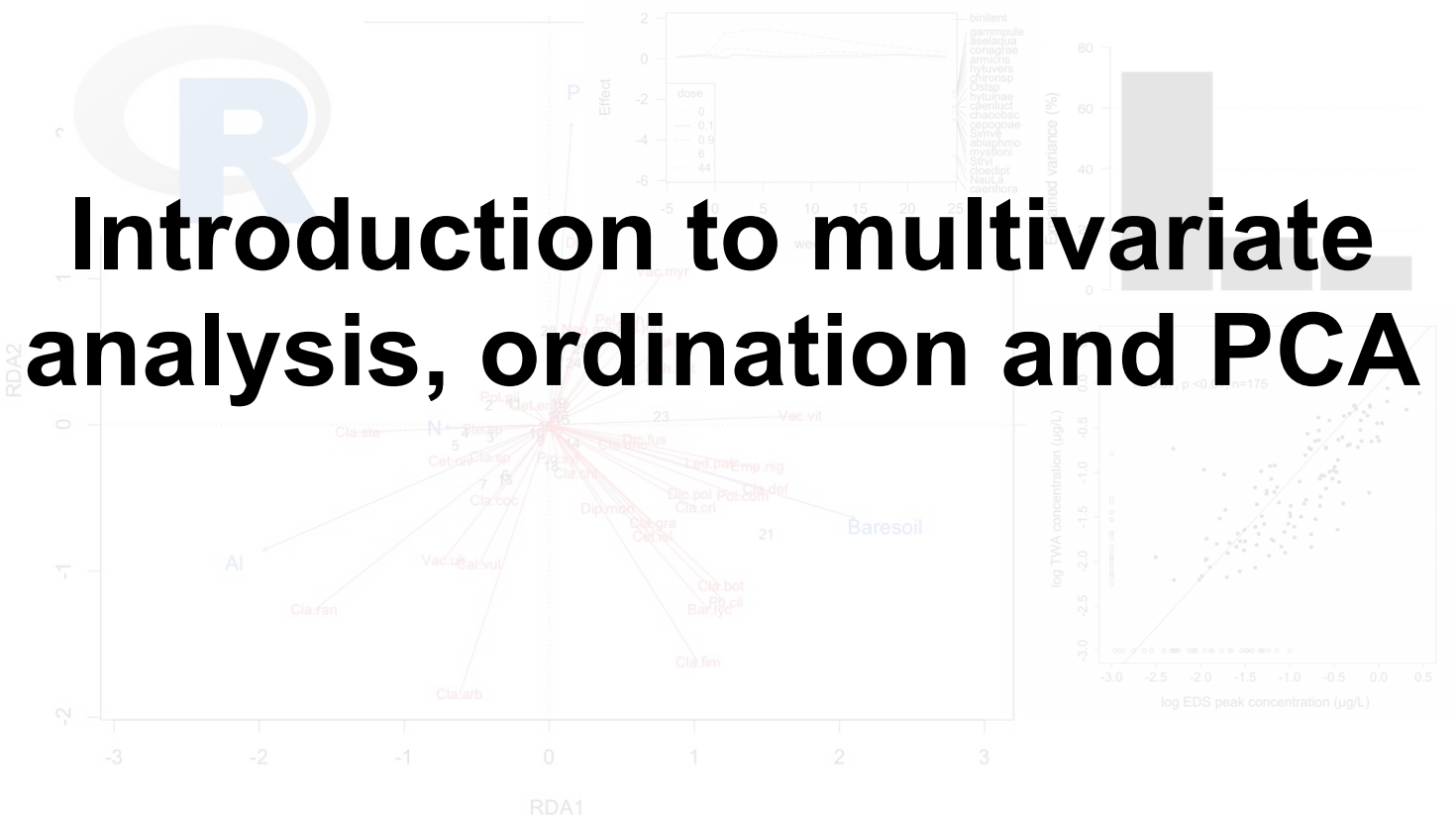
1. Check if conditions for descriptors are met (quantitative, multivariate normality, linear)
2. Conduct PCA (or sparse PCA) on scaled descriptors unless they exhibit a similar variation and have been measured on similar scale
3. Check for outliers
4. Select the optimal number of principal components
5. How informative are the first two PCs?
6. Which descriptors contribute most to PCs?
7. Visualise and interpret

Extension: 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 necessarily imply low explanatory power → not necessarily a fix for low n:p ratios
- Alternative: Sparse principal component-guided regression

Tools for complex data analysis

University of Koblenz-Landau 2019/20



Introduction to multivariate analysis, ordination and PCA

Ralf B. Schäfer

These slides and notes complement the lecture with exercises “Tools for complex data analysis” for ecotoxicologists and environmental scientists. Do not hesitate to contact me if you have any comments or you find any errors (slides, slide notes, or code): schaefer-ralf@uni-landau.de

While I made notes below the slides, some aspects are only mentioned in the R demonstration associated with the lecture.

Learning targets

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

2

Learning targets and study questions

- Explain the specifics of multivariate analysis
  - Why should you favour multivariate approaches for multivariate data?
  - What is the difference to the univariate case when diagnosing multivariate normality? What is the covariance matrix?
- List and select ordination methods based on research goal
  - Explain the aims of ordination.
  - Which criteria influence the selection of an ordination method? Discuss the relationship of the criteria to the research goal.

3

Learning targets and study questions

- Explain mathematical basis of PCA.
  - What are eigenvalues and eigenvectors? Provide a geometrical and algebraic explanation.
  - How do eigenvalues relate to the variance captured by 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 relationship between a) variables, b) sites and c) variables and axes.
  - Outline PCA assumptions and related diagnostic tools.
  - Which objects from a PCA would be extracted as non-collinear predictors for a multiple regression analysis?

4

Introduction to multivariate analysis, ordination and PCA

Contents

5

From univariate to multivariate statistics

	univariate	**multivariate**
Variables (vars.)	Single/multiple predictors, single response variable Y	Single/multiple predictors, multiple response vars. $Y_1, \ldots, Y_n$
Distribution of response	One-dimensional	n-dimensional
Data format	Y is vector	$Y_1, \ldots, Y_n$ constitute matrix
Example	Species richness explained by environmental variables	Community explained by environmental variables

6

The table describes the difference between univariate and multivariate statistics with a focus on the response (i.e. dependent) variable(s). Sometimes models with multiple predictors (e.g. GLMs, multiple regression model) are also considered as multivariate (e.g. Lloyd 2010), though they rather represent multivariable models (see Hidalgo & Goodman 2013).

Cited reference:

Hidalgo B. & Goodman M. (2013) Multivariate or Multivariable Regression? *American Journal of Public Health* 103, 39–40. Freely available under: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3518362/

Lloyd C. 2010: Spatial data analysis. Oxford University Press: Oxford.

WHY DO I HAVE TO LEARN MULTIVARIATE ANALYSES?

www.colourbox.de

7

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 when assessing statistical hypotheses

8

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

Furthermore, response variables typically incorporate some random variation and therefore can display an association with noise variables for example in multiple regression analysis (Flack & Chang 1987). Moreover, the larger the number of noise variables, the higher the probability that one of these is correlated with meaningful predictors, which also leads to associations between the response and noise variables. The simultaneous consideration of several response variables as in multivariate analysis reduces the influence of noise variables and can increase the statistical power when assessing hypotheses. For example, multivariate methods are used in climatology for the reconstruction of the global temperature trend, where single time lines would be insufficient to discover and establish the global warming trend (Ammann & Wahl 2007).

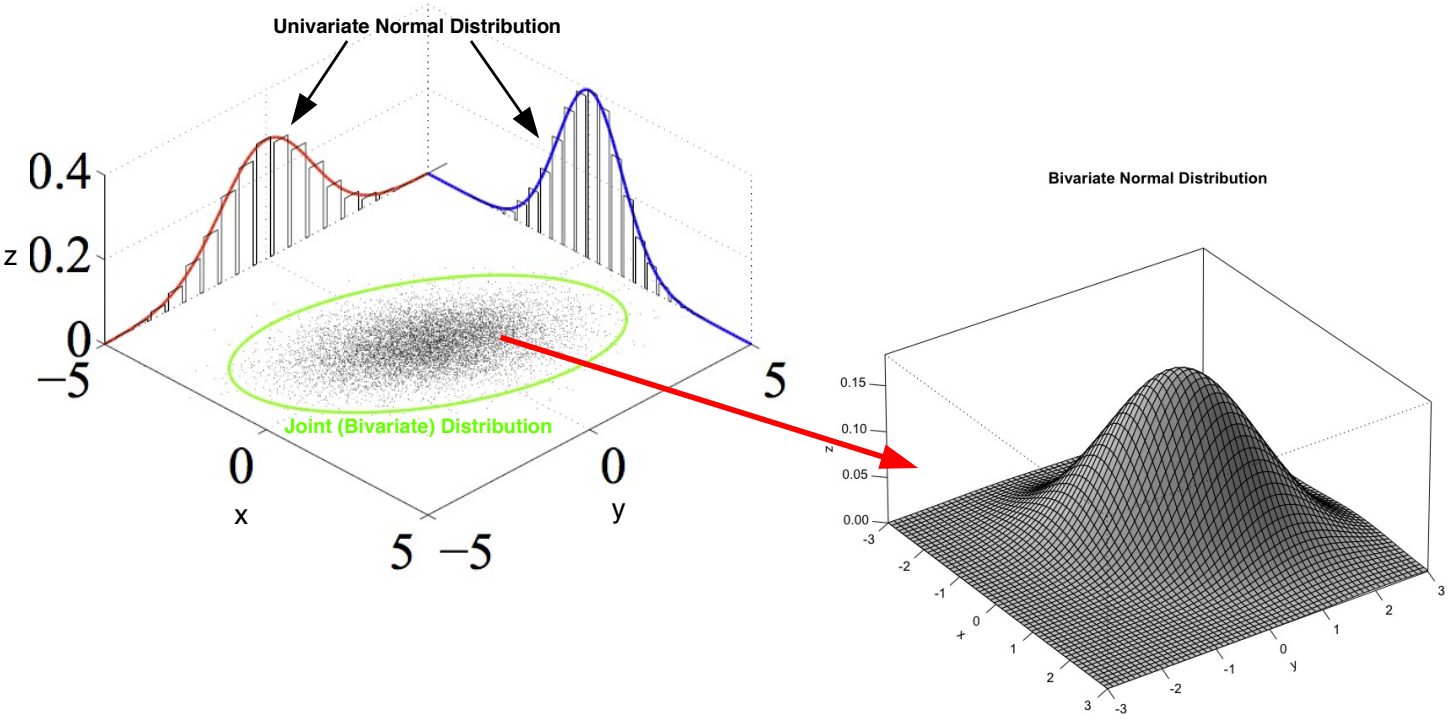
Ammann, E. R. & Wahl C. M. 2007: The importance of the geophysical context in statistical evaluations of climate reconstruction procedures. *Climate Change* 85 (1-2): 71-88

Flack, V. F. & Chang, P. C., Frequency of Selecting Noise Variables in Subset Regression-Analysis - a Simulation Study. *American Statistician* 1987: 84-86

Specifics of multivariate analysis

Tools and models used in univariate analysis can be used for example for diagnosing multicollinearity, but multivariate response requires adaptation of or new tools and models

Multivariate normal distribution

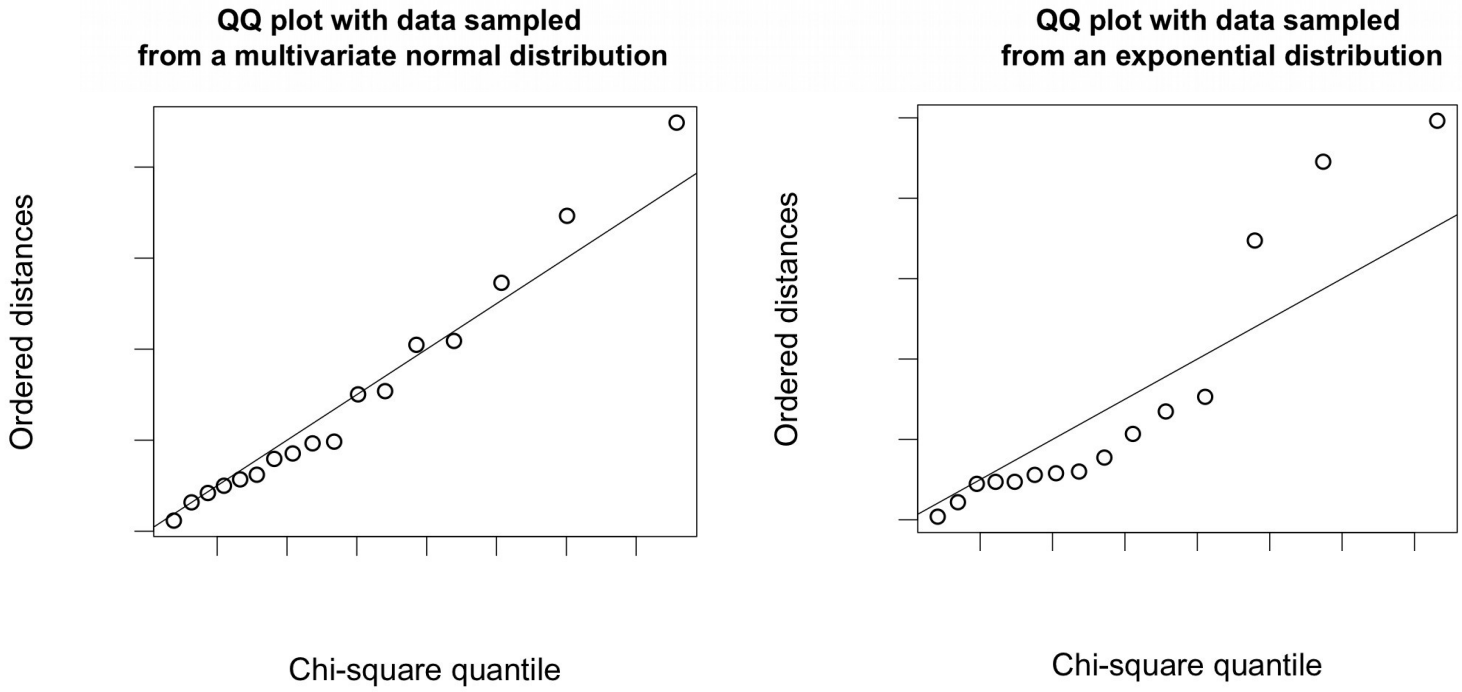


9 http://commons.wikimedia.org/wiki/File:Multivariate_normal_sample.svg

The general approach in multivariate analysis is often very similar to the one used in univariate analysis. Hence, our data analysis cycle also applies to multivariate analysis. However, we often need to adapt the tools and models or use new tools and models to account for the multivariate response. We discuss a few technical aspects that are relevant in the context of multiple methods.

Specifics of multivariate analysis

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



Regarding the calculation of the *QQ*-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 later). The d_Ms are then ordered and compared to the quantiles of a beta or χ^2 distribution. The χ^2 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, briefly discussed later). The criticism on hypothesis testing for normality outlined for the univariate case largely applies to these.

Small, N. J. H. 1978: Plotting squared radii. *Biometrika* 65 (3): 657-658

Covariance matrix

For a data matrix $\boldsymbol{Y}$ containing the variables $Y_1,\ldots,Y_p$

$$\boldsymbol{Y} = \begin{pmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,j} & y_{1,k} & \cdots & y_{1,p} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,j} & y_{2,k} & \cdots & y_{2,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{i,1} & y_{i,2} & \cdots & y_{i,j} & y_{i,k} & \cdots & y_{i,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{n,1} & y_{n,2} & \cdots & y_{n,j} & y_{n,k} & \cdots & y_{n,p} \end{pmatrix}$$

the sample covariance matrix $\boldsymbol{S}$, which is an estimate of $\boldsymbol{\Sigma}$, is:

$$\hat{\Sigma} = \boldsymbol{S} = \begin{pmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,j} & s_{1,k} & \cdots & s_{1,p} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,j} & s_{2,k} & \cdots & s_{2,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{j,1} & s_{j,2} & \cdots & s_{j,j} & s_{j,k} & \cdots & s_{j,p} \\ s_{k,1} & s_{k,2} & \cdots & s_{k,j} & s_{k,k} & \cdots & s_{k,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{p,1} & s_{p,2} & \cdots & s_{p,j} & s_{p,k} & \cdots & s_{p,p} \end{pmatrix}$$

where $s_{j,k} = \frac{1}{n-1}\sum_{i=1}^{n}(y_{i,j}-\overline{y_j})(y_{i,k}-\overline{y_k})$

The formula for the covariance of two variables Y_j and Y_k (i.e. $s_{j,k}$) should be familiar to you from univariate statistics: the covariance is a component of the calculation of the of bivariate *sample Pearson correlation coefficient* r (for details see the document *Key terms and concepts*). For two variables X and Y, $r_{X,Y}$ is the (sample) covariance divided by the product of the two estimates of the standard deviations:

$$r_{X,Y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}}$$

Covariance and Mahalanobis distance

In the diagonal, the equation simplifies to:

$$s_{j,j} = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_{i,j}-\overline{y_j}\right)^2$$

$$\hat{\Sigma} = \boldsymbol{S} = \begin{pmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,j} & s_{1,k} & \cdots & s_{1,p} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,j} & s_{2,k} & \cdots & s_{2,p} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ s_{j,1} & s_{j,2} & \cdots & s_{j,j} & s_{j,k} & \cdots & s_{j,p} \\ s_{k,1} & s_{k,2} & \cdots & s_{k,j} & s_{k,k} & \cdots & s_{k,p} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ s_{p,1} & s_{p,2} & \cdots & s_{p,j} & s_{p,k} & \cdots & s_{p,p} \end{pmatrix}$$

Covariances

Variances

The Mahalanobis distance d_M incorporates $\boldsymbol{\Sigma}$ when measuring the multivariate distance between two vectors. For example, d_M for the observation x to the mean vector μ is:

$$d_M = \sqrt{(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

→ can be used to detect outliers

12

The covariance matrix is symmetrical with respect to the diagonal, i.e. $s_{j,k} = s_{k,j}$.

d_M can be calculated for any two vectors of the same length if either $\boldsymbol{\Sigma}$ is known or $\boldsymbol{S}$ can be estimated. It is the distance between the two vectors weighed by the (sample) covariance and represents the multivariate counter part to the *standard score*, i.e. z-score. The z-score indicates the distance of an observation to the mean measured in fractions of standard deviations (SD). It is calculated by subtracting the mean of X (μ_X) from the value of an observation x_i, and dividing the result by the SD of X:

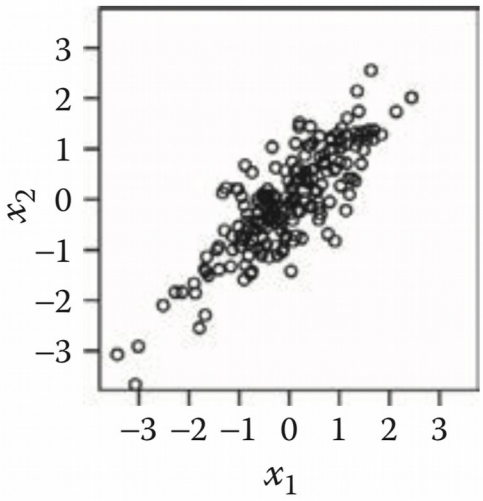
$$z_i = \frac{x_i-\mu_X}{\sqrt{\mathrm{Var}(X)}}$$

The division by SD (also denoted σ_X) and the multiplication with the inverse of $\boldsymbol{\Sigma}$ have the same purpose: to standardise the difference to the sample mean/centroid by the variation in the respective variable(s).

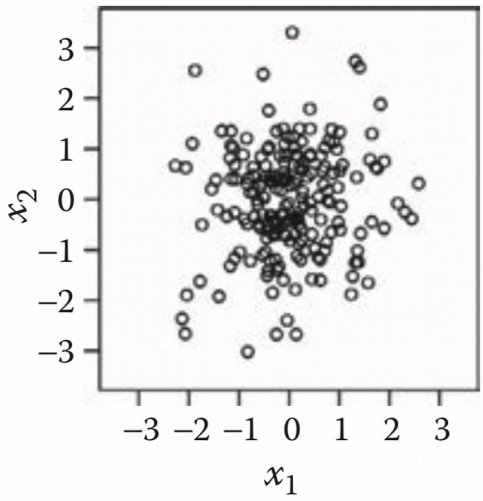
For outlier detection, a robust version of the Mahalanobis distance has been recommended, where the d_M is calculated from a subsample of the data that is free from outliers (for details see Leys et al. 2018).

Leys C., Klein O., Dominicy Y. & Ley C. (2018). Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. *Journal of Experimental Social Psychology* 74, 150–156.

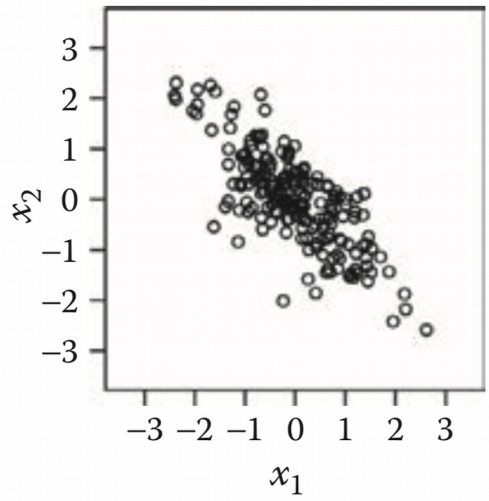
Covariance matrices for two variables



$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$$



$$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$



$$\boldsymbol{\Sigma}_3 = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 1 \end{pmatrix}$$

13

Varmuza & Filzmoser 2009: 40

If the variances are 1 as in this example, the covariance matrices are identical to the correlation matrices.

Multivariate approaches in R

Available methods and developments: CRAN Task View

CRAN Task View: Multivariate Statistics

Maintainer: Paul Hewson
Contact: Paul.Hewson at plymouth.ac.uk
Version: 2018-07-21
URL: https://CRAN.R-project.org/view=Multivariate

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

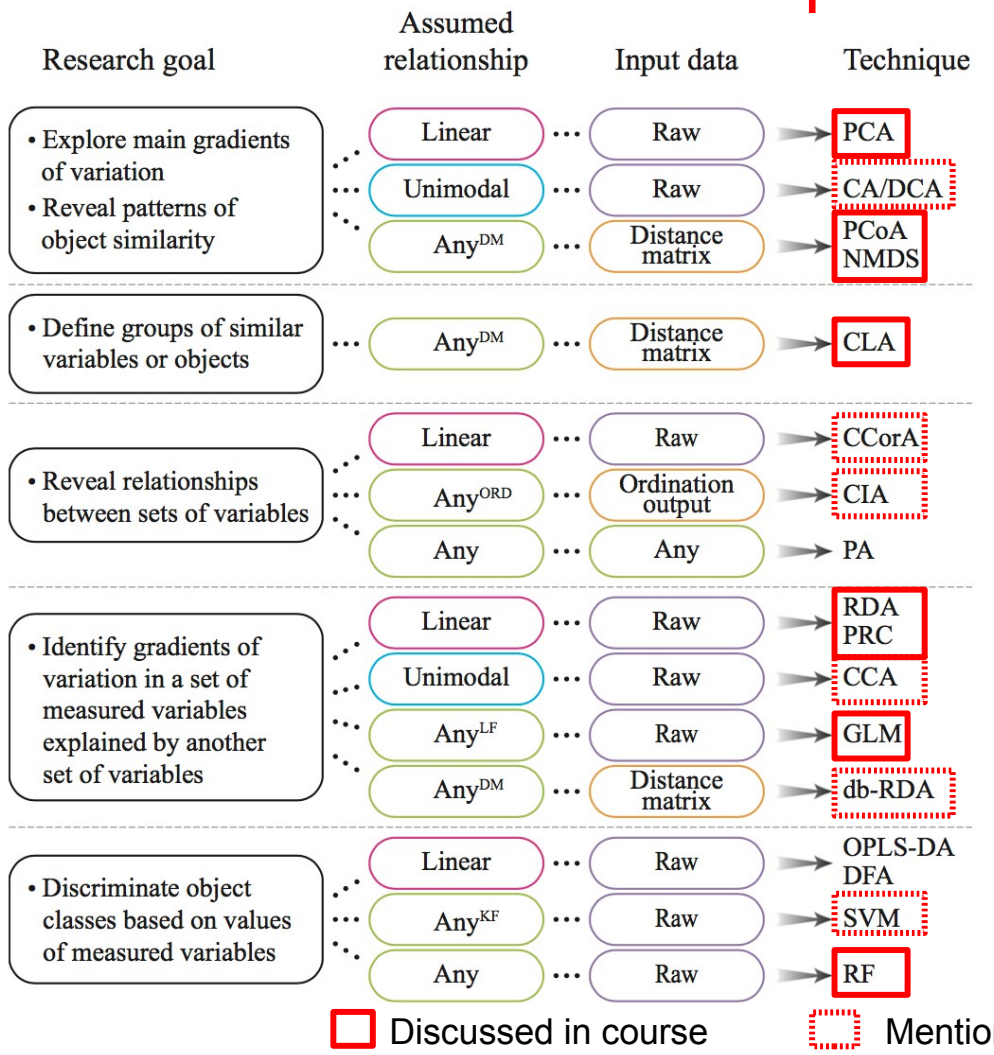
- ICSNP provides Hotellings T2 test as well as a range of non-parametric tests including location tests based on marginal ranks, spatial median and spatial signs computation, estimates of shape. Non-parametric two sample tests are also available from cramer and spatial sign and rank tests to investigate 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 `covMCD()` and the Orthogonalized pairwise estimate of Gnanadesikan-Kettenring via `covOGK()`. Scalable robust methods are provided within rrcov also using fast minimum covariance determinant with `covMcd()` as well as M-estimators with `covMest()`. corpcor provides shrinkage estimation of large scale covariance and (partial) correlation matrices.
- *Densities (estimation and simulation):* `mvnorm()` in MASS simulates from the multivariate normal distribution. mvtnorm also provides simulation as well as probability and quantile functions for both the multivariate t distribution and multivariate normal distributions as well as density functions for the multivariate normal distribution. mnormt provides multivariate normal and multivariate t density and distribution functions as well as random number simulation. sn provides density, distribution and random number generation for the multivariate skew normal and skew t distribution. delt provides a range of functions for estimating multivariate densities by CART and greedy methods. Comprehensive information on mixtures is given in the Cluster view, some density estimates and random numbers are provided by `rmvnorm.mixt()` and `dmvnorm.mixt()` in ks, mixture fitting is also provided within bayesm. Functions to simulate from the Wishart distribution are provided in a number of places, such as `rwishart()` in bayesm and `rwish()` in MCMCpack (the latter also has a density function `dwish()`). `bkde2D()` from KernSmooth and `kde2d()` from MASS provide binned and non-binned 2-dimensional kernel density estimation, ks also provides multivariate kernel smoothing as does ash and GenKern. prim provides patient rule induction methods to attempt to find regions of high density in high dimensional multivariate data, feature also provides methods for determining feature significance in multivariate data (such as in relation to local modes).
- *Assessing normality:* mvnormtest provides a multivariate extension to the Shapiro-Wilks test, mvoutlier provides multivariate outlier detection based on robust methods. ICS provides tests for multi-normality. `mvnorm.etest()` in energy provides an assessment of normality based on E statistics (energy); in the same package `k.sample()` assesses a number of samples for equal distributions. Tests for Wishart-distributed covariance matrices are given by `mauchly.test()` in stats.
- *Copulas:* copula provides routines for a range of (elliptical and archimedean) copulas including normal, t, Clayton, Frank, Gumbel, fgac provides generalised archimedian copula.

14

Overview multivariate techniques



15

Paliy & Shankar 2016 *Mol. Ecol.* 25: 1032

Abbreviations: PCA – Principal Component Analysis, DCA – Detrended Correspondence Analysis, PCoA – Principal Coordinate Analysis, NMDS – Non-Metric Multidimensional Scaling, CLA – Cluster Analyis, CCorA – Canonical Correlation Analysis, CIA – Co-Inertia Analysis, PA – Procrustes Analysis, RDA – Redundancy Analysis, PRC – Principal Response Curves, Canonical Correspondence Analysis, GLM - Generalized Linear Model, db-RDA – Distance-based Redundancy Analysis, OPLS-DA – Orthogonal Partial Least Squares Discriminant Analysis, DFA – Discriminant Factor Analysis, SVM – Supporting Vector Machines, RF – Random Forest.

The overview presents a wide range of techniques of which only some will be discussed in detail in the course (in red). The research goals will be discussed in more detail in the context of these techniques.

Compared to the univariate overview, the overview seems to lack the group-based analyses with the research goal *To identify gradients of variation in a set of measured variables explained by factor(s)*. Related techniques include multivariate analysis of variance (MANOVA) and permutational analysis of variance (PERMANOVA). However, RDA and multivariate GLMs can be used as well for this goal and this may explain why the authors omitted these techniques in the figure (but mention them in the paper).

Very readable introductions into multivariate methods are provided by Ramette (2007) and Paliy & Shankar (2016). For a focus on ecotoxicology, though a bit outdated, see van den Brink et al. (2003). For a focus on multivariate techniques in the context of diversity research see Anderson et al. (2011). For a focus on novel developments, see Warton et al. (2015 a,b).

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. Freely accessible within our university at: http://onlinelibrary.wiley.com/doi/10.1111/j.1461-0248.2010.01552.x/full

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

Ramette, A. (2007) Multivariate analyses in microbial ecology. *FEMS Microbiology Ecology*, 62, 142-160. Freely accessible at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2121141/pdf/fem0062-0142.pdf

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 accessible at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.77.403&rep=rep1&type=pdf

Warton D.I., Foster S.D., De'ath G., Stoklosa J. & Dunstan P.K. (2015a) Model-based thinking for community ecology. *Plant Ecology* 216, 669–682. Freely accessible at: 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. (2015b) So Many Variables: Joint Modeling in Community Ecology. *Trends in Ecology & Evolution*. 30, 766-779.

Introduction to multivariate analysis, ordination and PCA

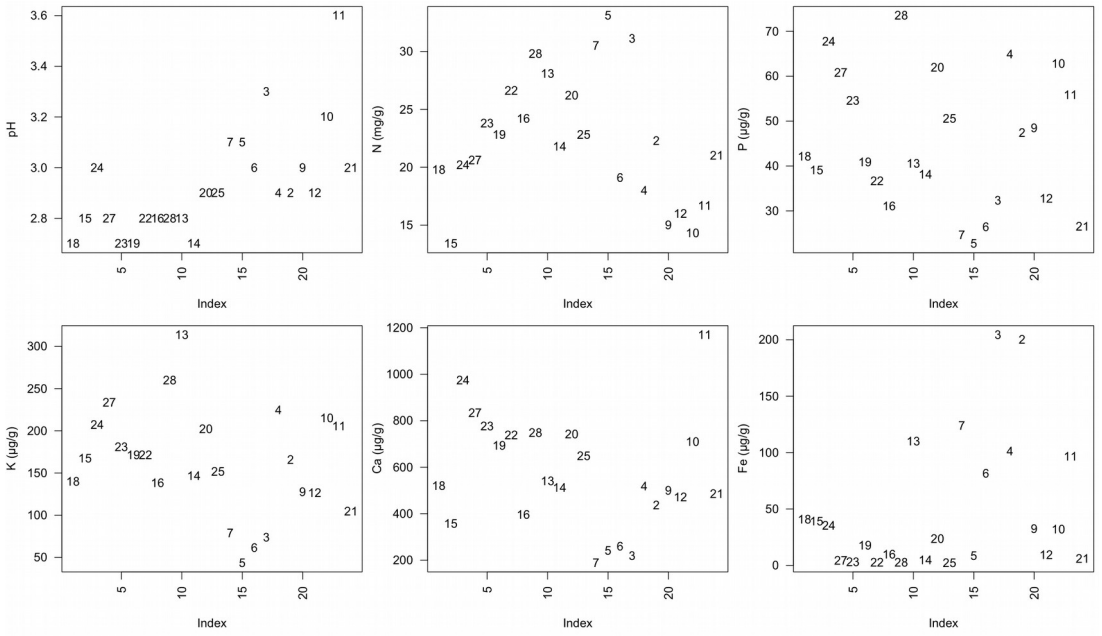
Contents

16

Ordination: Introduction

- Representation of objects (e.g. sites, samples) along one or multiple axes
- Case study: Measurements of 14 physicochemical variables (e.g. pH, ions) in soil of 24 sites. Research goal: Exploration – what is the main gradient and how are the sites related?

- Univariate approach: inspection of e.g. means and variances, correlations between variables, pattern of sites with respect to variables
- Many different kinds of statistics
  → Difficult to interpret



Ordination comes from ecology and is used when referring to methods that order objects along axes (that may represent ecological gradients or time). It is typically used in the context of multivariate analysis, where the aim is to project the data into a reduced (i.e. lower dimensional) space, for example, to allow for graphical representation.

Regarding terminology, by main gradient we refer to the main direction of variation in the data, typically multiple variables contribute to this gradient.

To tackle the research goal using an univariate approach, although not recommended, you could first compare the relative standard deviations of the different variables to identify the longest gradients. Based on the relationship between the variables, examined for example via paired scatterplots, you may then be able to identify the major gradient, but rather qualitatively. However, evaluation of the relationship of the sites with the main gradient would be rather difficult. You would need to simultaneously inspect the plots of the most relevant individual variables.

We exemplify this using a data set with soil information, see Väre et al. (1995) for details. Note that even for six (of 14) variables it is complicated to identify clear patterns for the sites (here, we ignore the question whether these 6 would be the variables constituting the major gradient). However, you could note that site 28 has high concentrations of P and N, but has among the lowest Fe concentrations.

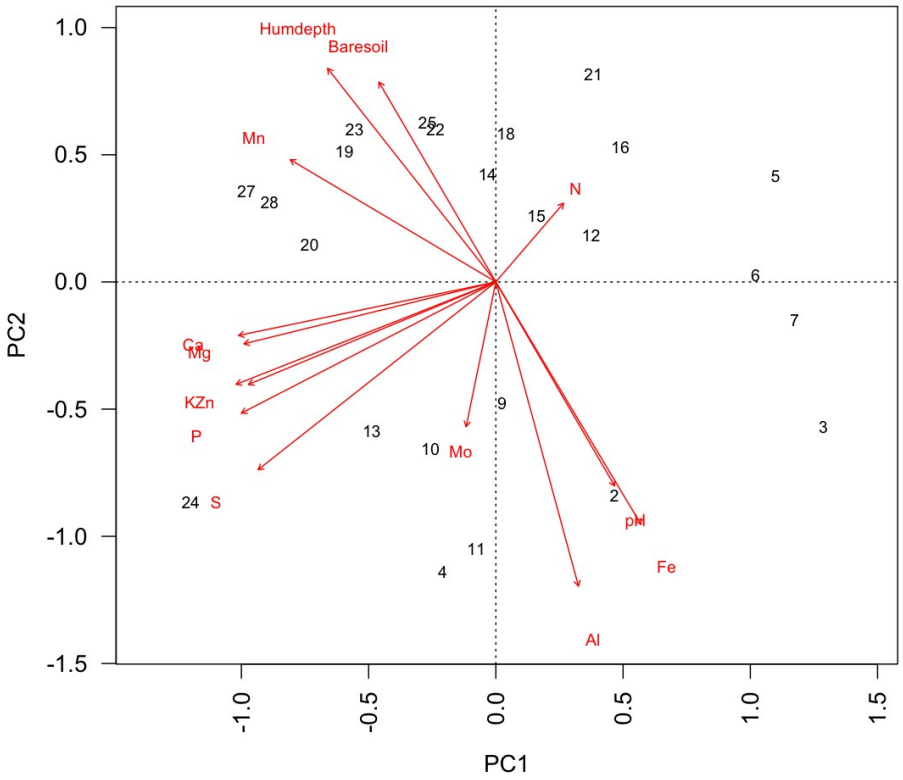
Väre H., Ohtonen R. & Oksanen J. (1995). Effects of reindeer grazing on understorey vegetation in dry Pinus sylvestris forests. *Journal of Vegetation Science* 6, 523–530.

17

Ordination: Introduction

- Multivariate approach: Principal Component analysis (PCA)
- Representation of the first two major gradients in the data → Two-dimensional representation of the major variance in 14 dimensions (i.e. variables)

- Ordination plot displays the following information:
  - Contribution of variables to major gradients
  - Relationship between variables
  - Relationship between sites
  - Pattern of sites with respect to variables
- Easier to interpret!



18

Aims of ordination

- Dimension reduction – (graphical) representation of data in reduced (lower-dimensional) space (and potentially omission of variables/gradients/axes that capture low amount of variance)
- Aggregation of variables into gradients and extraction of main gradients
- Constrained ordination: extraction of gradients that are explained by variables of second data set. Supervised learning.
- Unconstrained ordination: extraction without consideration of variables outside of data set. Unsupervised learning.

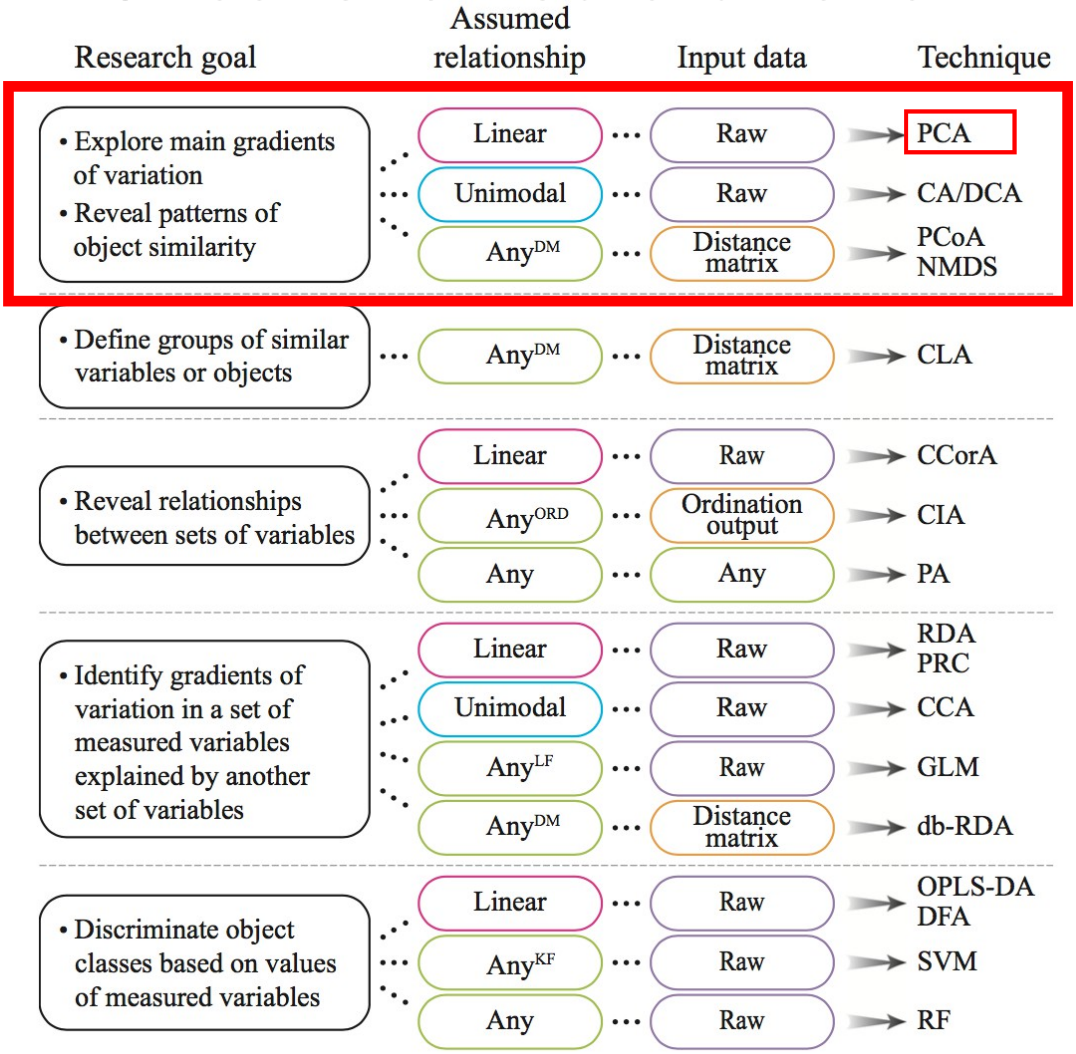
The main research goals for constrained ordination are:
- Explanation
- Determination of probabilities and assessing hypotheses

By contrast, for unconstrained ordination the main research goals are:
- Exploration
- Explanation

Parameter estimation and prediction are rarely research goals related to ordination. Model-based approaches that will be introduced later are more appropriate for these research goals.

Unsupervised learning methods lack a clear measure of accuracy that is comparable across methods (as opposed to supervised learning where, for example, the RMSE can be used to compare the performance of different methods). Therefore, it is also difficult to draw inferences based on these methods. See Hastie, Tibshirani and Friedman (2017: 487) for further discussion of this issue.

Unconstrained ordination

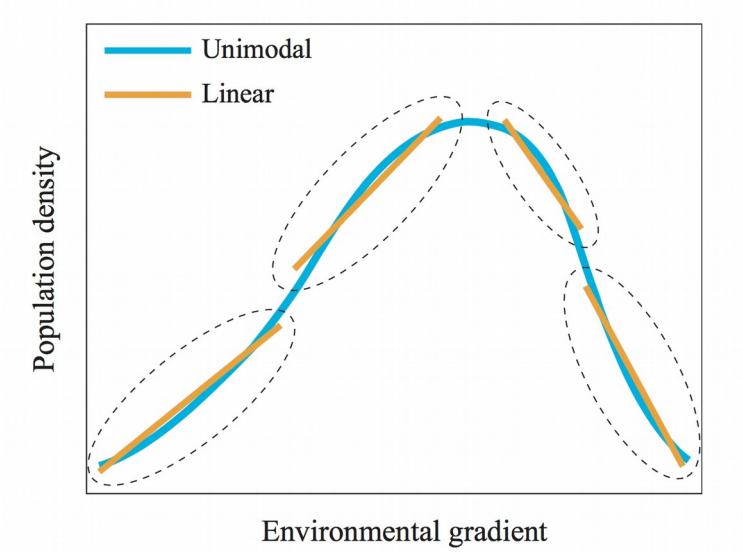


Paliy & Shankar 2016 *Mol. Ecol.* 25: 1032

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

Ordination: Overview

Shape of response	Linear	Unimodal	Any
Unconstrained methods (examples)	PCA	CA	Distance-based: NMDS; GAM-based: UAO (U-VGAM)
Constrained methods (examples)	RDA	CCA	Distance-based: db-RDA; GAM-based: CAO (RR-VGAM)



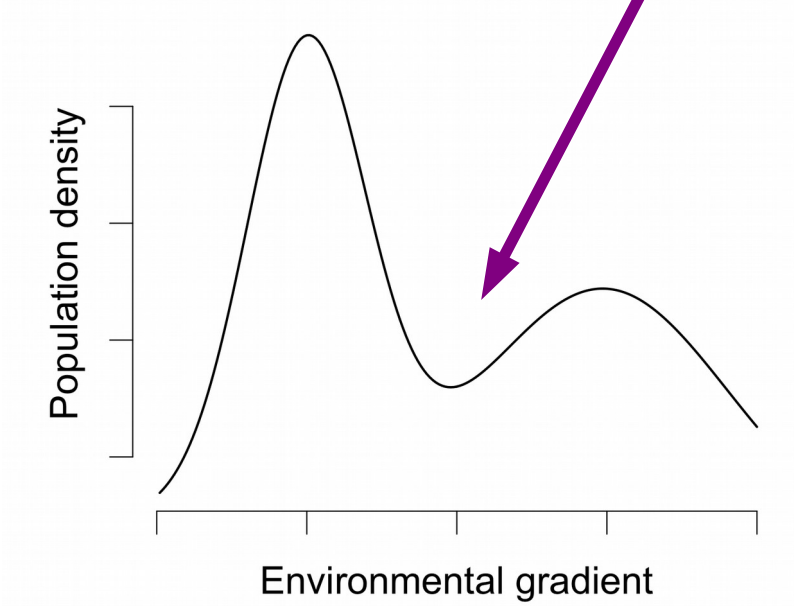
21

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
Unconstrained methods (examples)	PCA	CA	Distance-based: NMDS; GAM-based: UAO (U-VGAM)
Constrained methods (examples)	RDA	CCA	Distance-based: db-RDA; GAM-based: CAO (RR-VGAM)



22

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

23

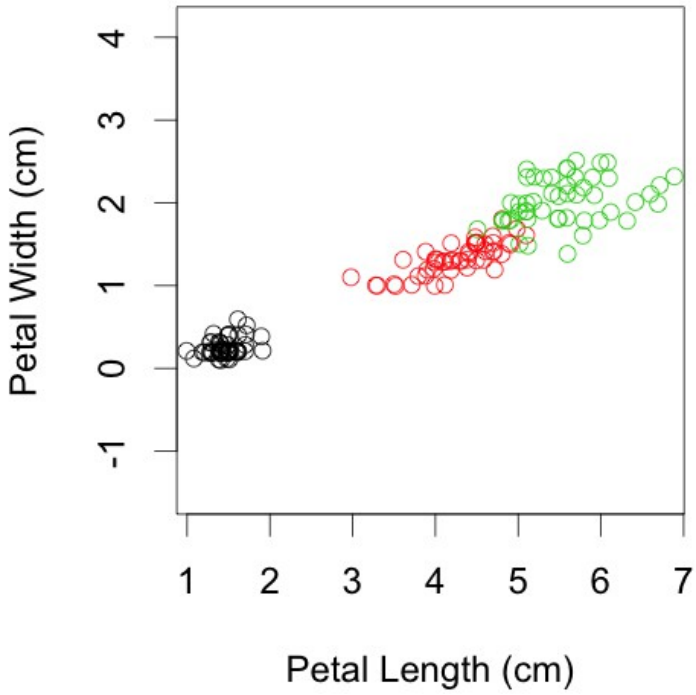
Principal Component Analysis

Example-based introduction

Iris data set: sepal length & width, petal length & width for 50 flowers from 3 species of Iris. Visual demonstration for 2 variables.

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

Goal: Dimension reduction. Represent as much variance as possible with first few axes



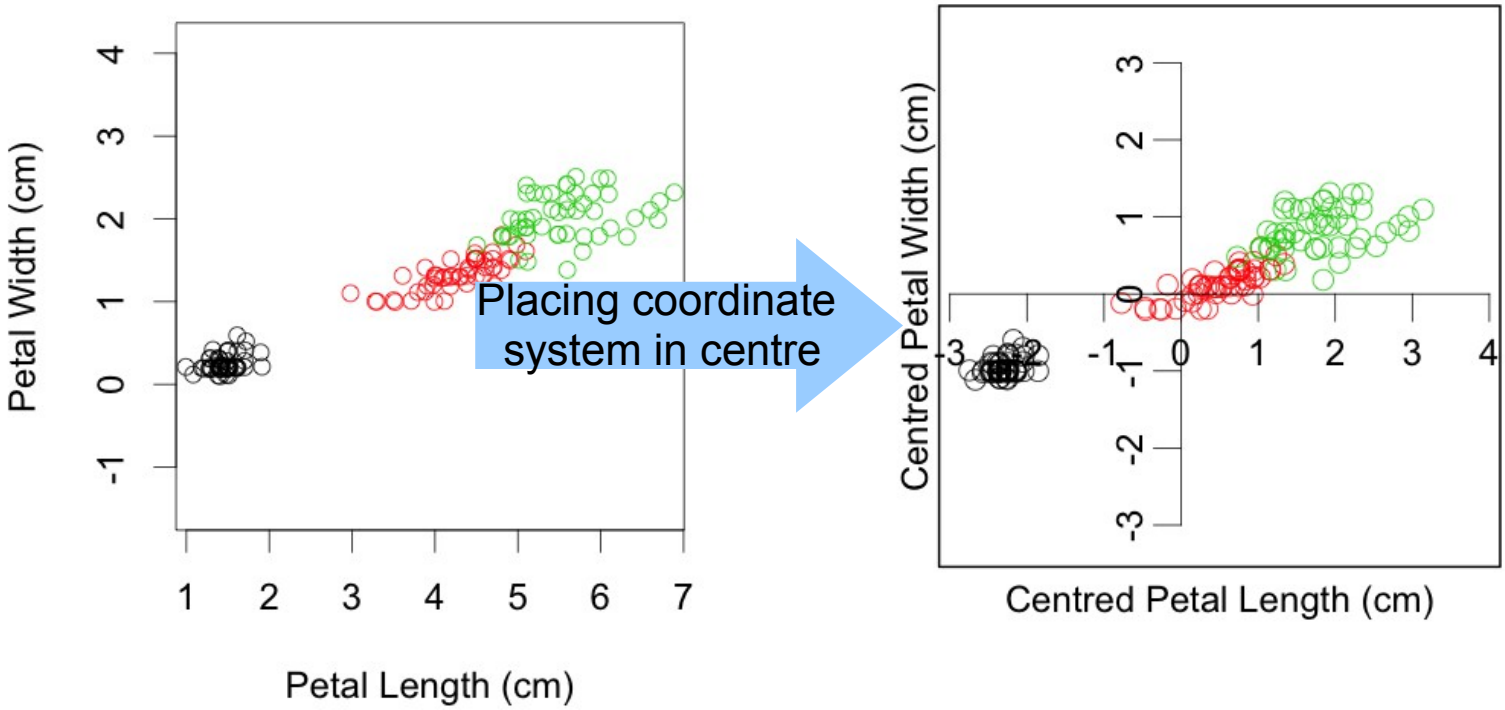
24

The example has been adapted from a blog contribution, but the content is not accessible anymore: http://blogs.nature.com/boboh/2012/01/17/pca-and-pcoa-explained

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.

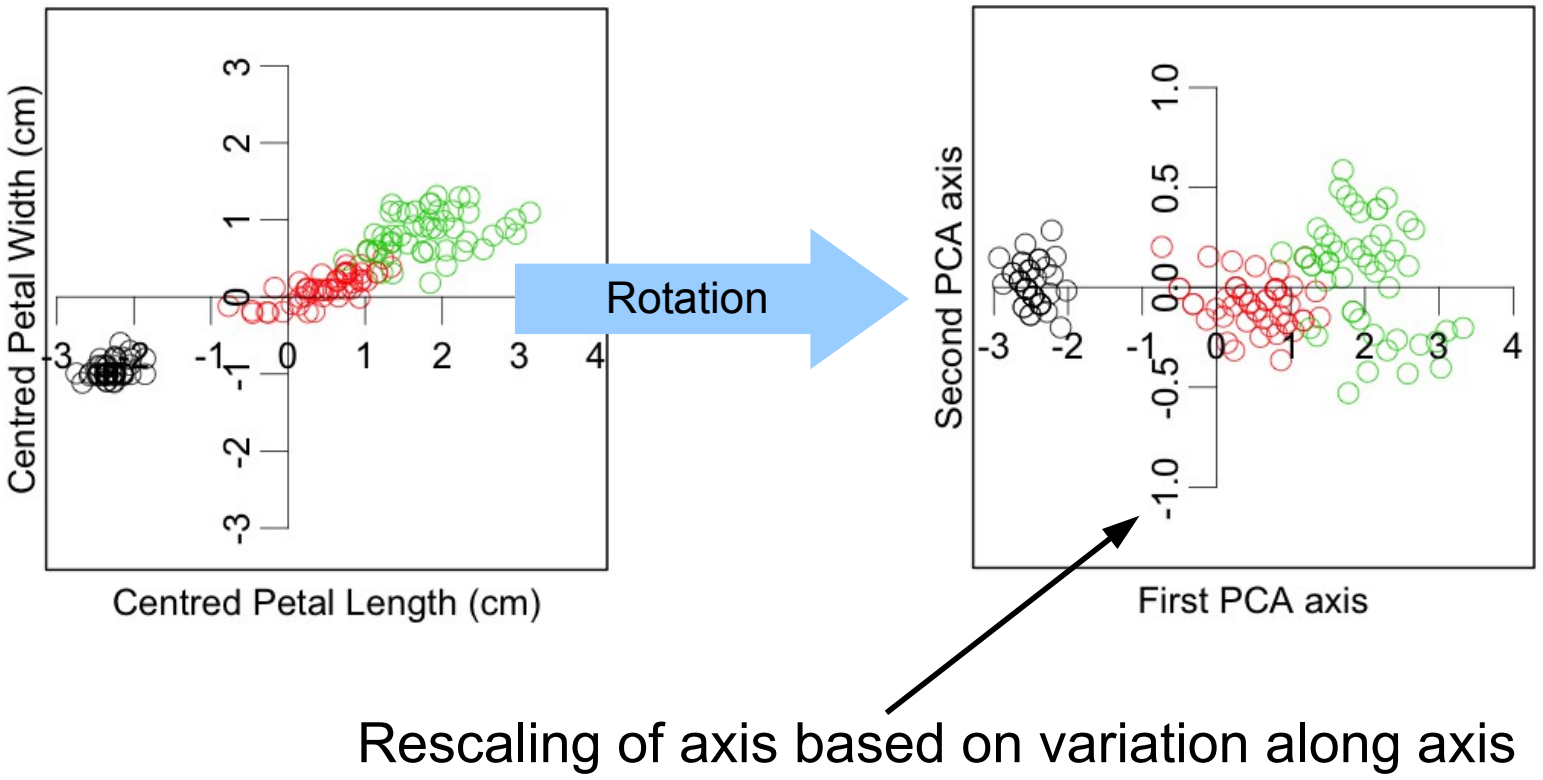
If you struggle with the topic, I suggest to start with a layman introduction into PCA provided in a blog, which is based on a teapot (no joke): http://blog.ephorie.de/intuition-for-principal-component-analysis-pca

Introduction to PCA



25

Introduction to PCA



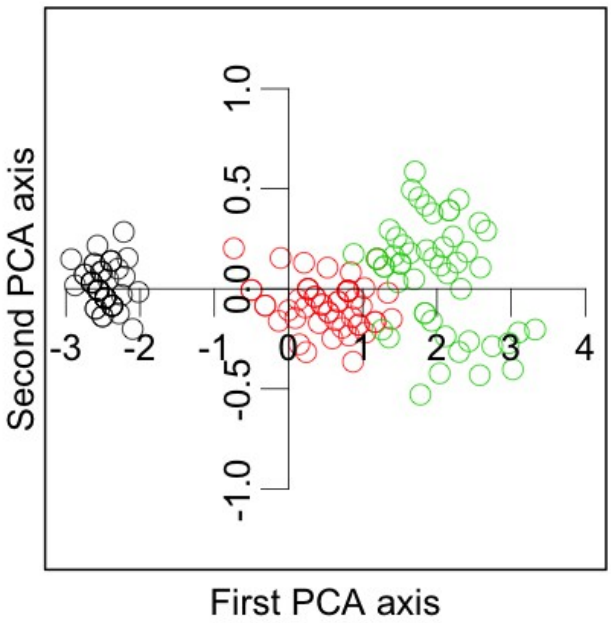
26

The term scaling is used with two different meanings in the context of PCA and ordination in general. First, it refers to the question whether predictors are standardised to unit variance before analysis (i.e. divided by their standard deviation). Second, the PCA results can be scaled either to display the distances between objects or to display the relationship between descriptors (see Borcard, Gillet & Legendre (2018: 156-160) and Jolliffe (2002: 90ff) for details). Here, we refer to the second meaning.

Borcard D., Gillet F. & Legendre P. (2011) Numerical ecology with R, Springer, New York.

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

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 captures 99% of variance!

27

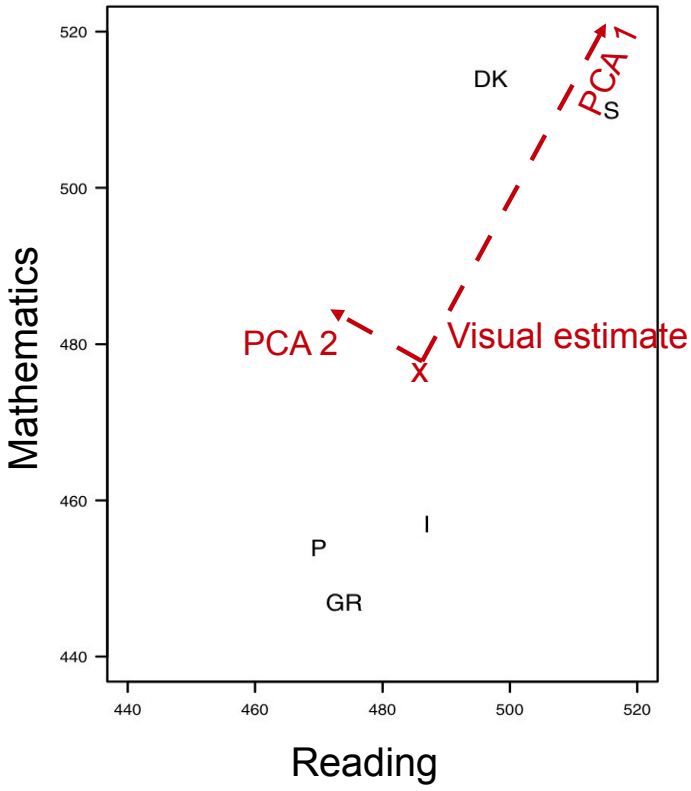
Regarding terminology (see Legendre & Legendre (2012: 429)), the principal component describes the positions of the objects in the new coordinate system, whereas principal component axis refers only to the axis (and not to the objects).

The eigenvalues indicate how much of the total variance is captured by an axis in the new coordinate system. The higher the eigenvalue, the more variance is captured.

Mathematical background of PCA

Example: Scores in international education survey (PISA)

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



Centering leads to

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

Search for first axis with maximum variation!

28

Example taken from Handl (2010)

The ~ on top of X indicates that the matrix has resulted from the transformation of the matrix X (centering represents a transformation).

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

Mathematical background of PCA

Search for first axis with maximum variation!

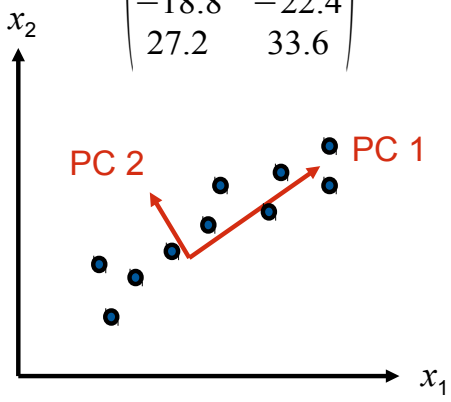
$$\tilde{X} = \begin{pmatrix} \tilde{X}_1 & \tilde{X}_2 \\ 8.2 & 37.6 \\ -14.8 & -29.4 \\ -1.8 & -19.4 \\ -18.8 & -22.4 \\ 27.2 & 33.6 \end{pmatrix}$$

Scores on new axis given as

$$PC_1 = a_1 \tilde{X}_1 + a_2 \tilde{X}_2$$

and maximum variation means:

$$\arg\max_{a_1, a_2} \; \mathrm{Var}\left(a_1 \tilde{X}_1 + a_2 \tilde{X}_2\right)$$



Generalise problem: define a_1, a_2 as elements of vector a and likewise $\tilde{X}_1, \tilde{X}_2$ of matrix $\boldsymbol{X}$: $\arg\max_a \; \mathrm{Var}\left(a \tilde{\boldsymbol{X}}\right)$

Trivial solution: choose high values for $a_1, a_2, \ldots, a_n$

→ introduce condition: $a_1^2 + a_2^2 + \ldots + a_n^2 = 1$

29

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

$$PC_1 = a_1 \tilde{X}_1 + a_2 \tilde{X}_2 + \ldots + a_q \tilde{X}_q$$

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 equals the variance for $\tilde{X}_1$. 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 the condition outlined above) that maximise the variation.

Mathematical background of PCA

Solve:

$$\arg\max_{a} \; \mathrm{Var}(a\,\tilde{X}) \;\text{ with }\; a^T a = 1$$

This can be expressed as (see Handl & Kuhlenkasper 2017: 87)

$$\arg\max_{a} \; (a^T \Sigma\, a) \;\text{ with }\; a^T a = 1$$

Covariance matrix

Using the Lagrange function yields:

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

Eigenvalue problem

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

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

30

The description here uses the known variance-covariance matrix $\boldsymbol{\Sigma}$, in other words the variance-covariance matrix of the statistical population. In practice, we usually have to estimate this matrix from the sample data and use the sample variance-covariance matrix $\boldsymbol{S}$. If the data are standardized before analysis (divided by the variance) then this matrix equals the correlation matrix $\boldsymbol{R}$.

Handl A. & Kuhlenkasper T. (2017). Multivariate Analysemethoden: Theorie und Praxis mit R, 3. revised ed. Springer Spektrum, Berlin. Freely accessible within our university: https://www.uni-koblenz-landau.de/de/bibliothek

Introduction to multivariate analysis, ordination and PCA

Contents

31

Mathematical basics: Eigenvalues

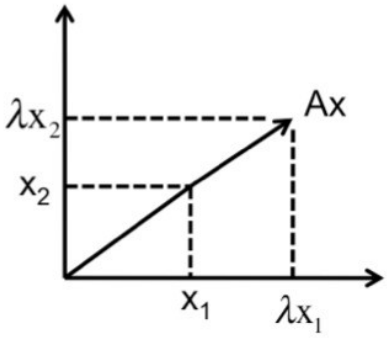
Idea: Conversion of matrix into a matrix with linear independent variables

$$\begin{pmatrix} a_{1,1} & \dots & a_{1,p} \\ \dots & a_{2,2} & \dots \\ a_{n,1} & \dots & a_{n,p} \end{pmatrix} \quad \text{Conversion} \quad \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_p \end{pmatrix}$$

Eigenvalue problem: $\boldsymbol{A}\,x = \lambda x$ Eigenvector

Eigenvalue

Eigenvectors form canonical basis and are only stretched or shrunk by λ when multiplied with $\boldsymbol{A}$.



32

Eigenvalues and eigenvectors play an important role in the mathematics behind most ordination methods. The eigenvectors and eigenvalues of matrix $\boldsymbol{A}$ 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).

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 and stretching of an observation given by the vector: $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

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: Eigenvalues

$$\boldsymbol{A}x = \lambda x \quad \Leftrightarrow \quad \boldsymbol{A}x - \lambda x = 0$$

$$\Leftrightarrow \quad (\boldsymbol{A} - \lambda \boldsymbol{I})x = 0$$

$$\Leftrightarrow \begin{pmatrix} a_{1,1}-\lambda_1 & \dots & a_{1,p} \\ \dots & \dots & \dots \\ a_{n,1} & \dots & a_{n,p}-\lambda_p \end{pmatrix} \begin{pmatrix} x_1 \\ \dots \\ x_n \end{pmatrix} = \begin{pmatrix} 0 \\ \dots \\ 0 \end{pmatrix}$$

Homogeneous linear equation system (HLS)

Ignore trivial solution: $x = 0$ $\quad \Rightarrow \quad \boldsymbol{A} - \lambda \boldsymbol{I} = 0$

The given HLS has only a non-trivial solution if the columns of $\boldsymbol{A} - \lambda \boldsymbol{I}$ are linearly dependent, which is the case if the determinant = 0.

$$\Rightarrow \quad det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0 \quad \Leftrightarrow \quad |\boldsymbol{A} - \lambda \boldsymbol{I}| = 0$$

33

Remember that we defined $\boldsymbol{I}$ as the identity matrix.

A homogenous linear equation system $\boldsymbol{A}x = 0$ has the unique solution $x_1 = x_2 = x_3 = \dots = x_n = 0$ for a regular matrix (symmetric 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 means that they are non-correlated and a PCA would in most such cases be pointless).

In PCA, the variance-covariance or correlation matrix is used, which both are symmetric matrices. However, the eigenvalues of a non-symmetric matrix can also be computed by singular value decomposition (special case of the Schur decomposition). This is implemented in R with svd().

Example I: Calculation of Eigenvalues

Sample Variance-Covariance matrix from PISA example: $\boldsymbol{S} = \begin{pmatrix} 345.7 & 528.35 \\ 528.35 & 1071.30 \end{pmatrix}$

Following $|\boldsymbol{A} - \lambda \boldsymbol{I}| = 0$ we obtain:

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

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

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

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

34

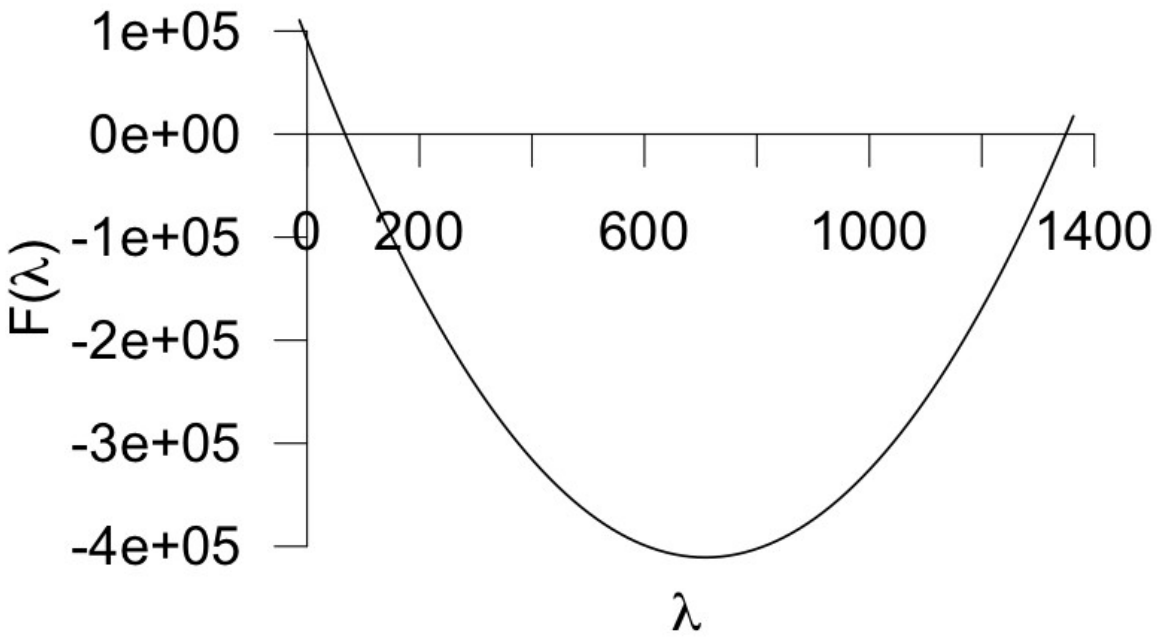
For small, symmetric matrices, the determinant can be calculated by hand. For example The determinant of a matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is given as $a\,d - b\,c$.

Example I: Calculation of Eigenvalues

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

Characteristic polynomial

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



In the case of a polynomial with 2 as the highest degree, the eigenvalues (λ) can easily be found using the pq formula:

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

Evaluating the equation for our values yields to:

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

Thus, $\lambda_1 = 1349.42$ and $\lambda_2 = 67.6$.

For details on the calculations see Handl (2010:123).

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

Example II: Calculation of Eigenvalues and -vectors

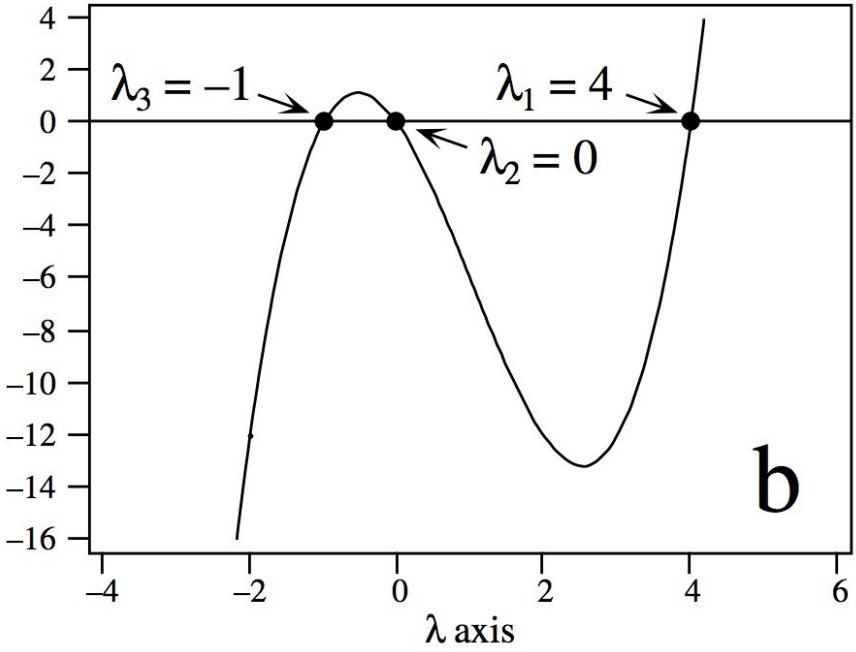
$$\begin{pmatrix} 1 & 3 & -1 \\ 0 & 1 & 2 \\ 1 & 4 & 1 \end{pmatrix}$$

EV calculation following Sarrus

Characteristic polynomial

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

Eigenvalues λ: 4, 0 and -1



36

Legendre & Legendre 2012: 93

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.

For more details on the mathematical basis of eigenvalues and eigenvectors refer to Legendre & Legendre (2012: 89).

Legendre P. & Legendre L. (2012) Numerical ecology, 3rd English ed. Elsevier, Amsterdam; Boston.

Example II: Calculation of Eigenvalues and -vectors

Calculation of eigenvector for $\lambda = 4$

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

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

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

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

37

Example II: Calculation of Eigenvalues and -vectors

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

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

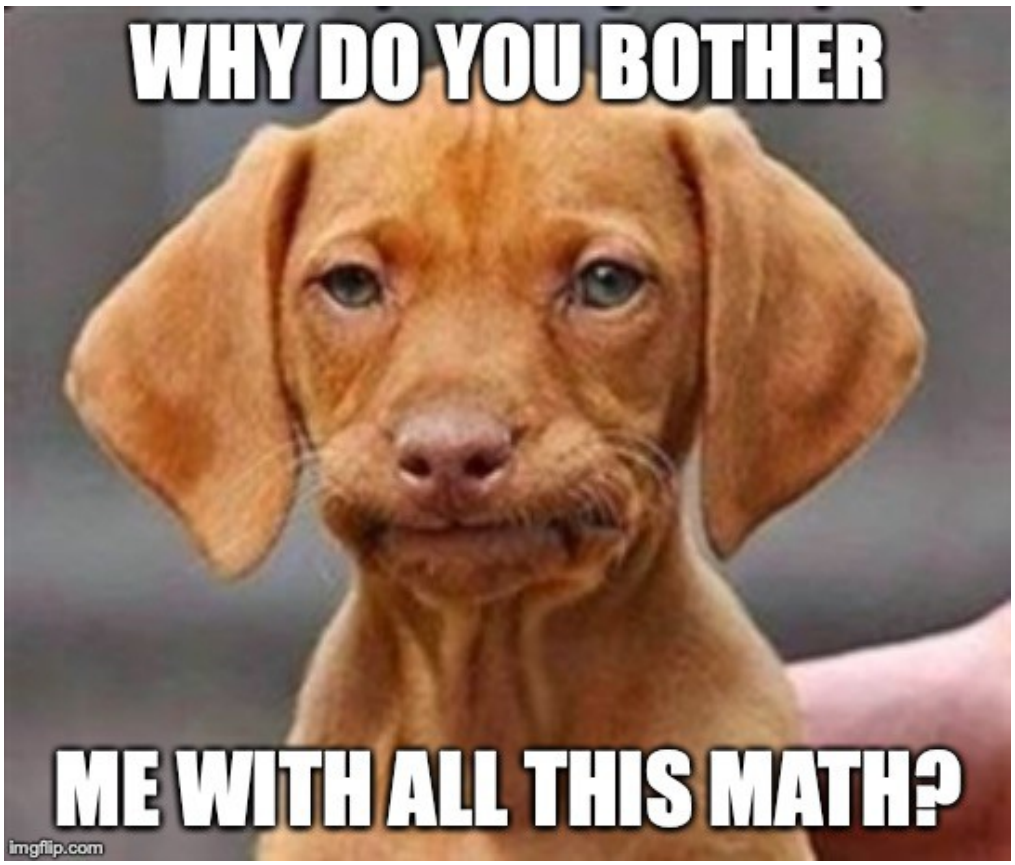
$$\begin{pmatrix} 1 & 7 & 2 \\ 2 & -2 & -1 \\ 3 & 1 & 1 \end{pmatrix}$$

Eigenvalues: 4; 0 and -1

38

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

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



39

Eigenvalues and eigenvectors are essential elements of several multivariate methods. Although the topic may be challenging, once you understand the general idea, the conceptual background of other methods relying on eigenvalues and eigenvectors will be more accessible to you.

Introduction to multivariate analysis, ordination and PCA

Contents

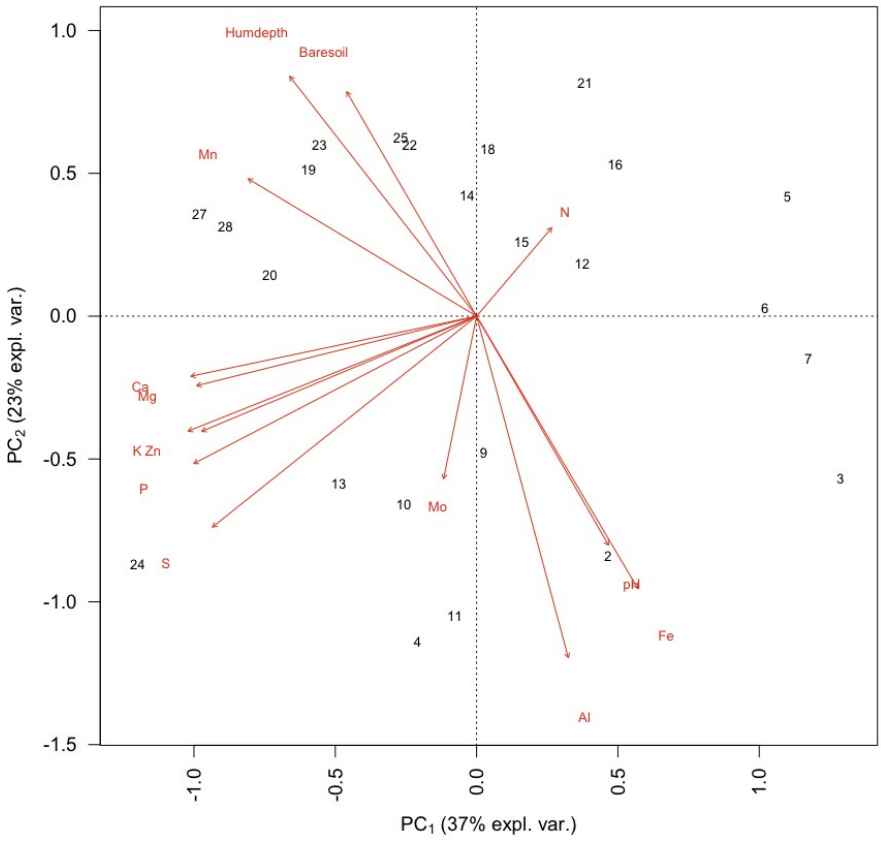
1. Introduction and specifics of multivariate analysis
2. Overview ordination
3. Introduction to PCA
4. Mathematical background
5. **PCA results and interpretation**
6. PCA diagnosis, tutorial and extensions

40

Biplot for PCA on 24 soil samples

- First two axes (gradients) capture 60% of variance
- Plot displays following information:

- Position of (site) labels given by scores on (new) axes
- Units on axes = standard deviations
- Contribution of variables to major gradients (the longer the arrow along an axis, the higher the correlation (and representation by axis))
- Relationship between variables (angle between arrows:
  0° or 180° $\rightarrow \rho = 1$ or -1
  90° $\rightarrow \rho = 0$
  60° $\rightarrow \rho = 0.5$)
- Relationship between sites (the closer, the higher similarity (consider variance explained by axes!))
- Pattern of sites with respect to variables (perpendicular projection of site on arrow)



The interpretation is also influenced by the scaling of the plot (mentioned earlier), we will discuss this topic in more detail in the R demonstration. In the given example, the issue of scaling is of lower relevance.

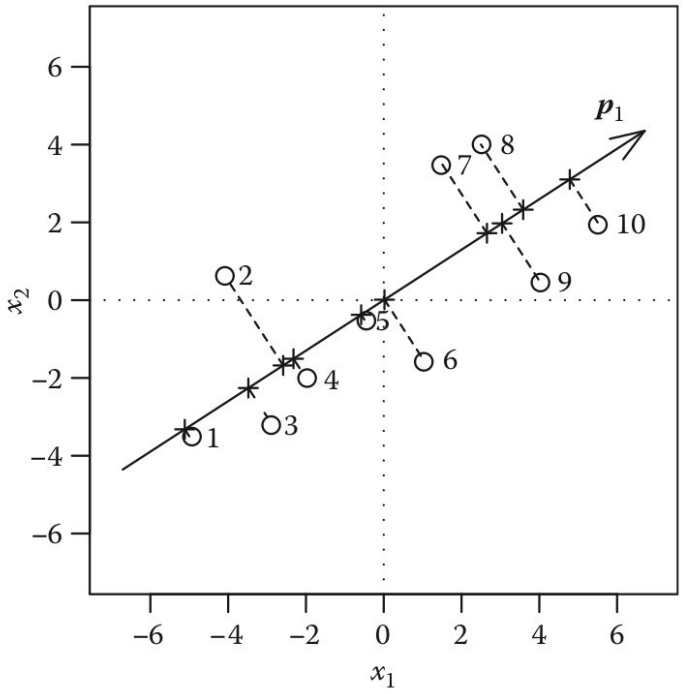
PC scores

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

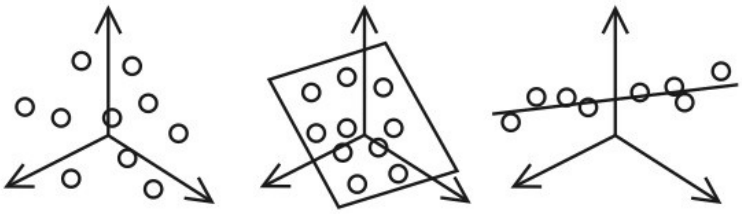


 Varmuza & Filzmoser 2009: Chapter 3

The figure displays the scores t_1 for the object number *i* on the first PC (designated with p_1).

Number of principal components

- Number of descriptors/variables determines number of eigenvalues and thus principal components
- Principal component that relates to the largest eigenvalue captures 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

Varmuza & Filzmoser 2009: Chapter 3

43

How many principal components needed?

Some criteria to evaluate the optimal number of axes:

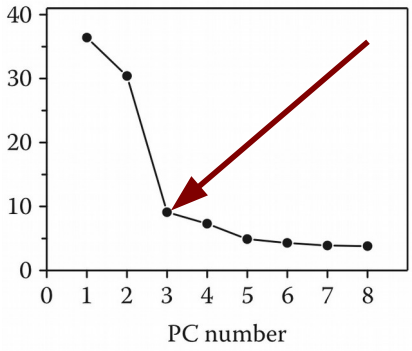
1. Sum criterion

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

2. Broken-Stick criterion

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

3. Scree plot



4. Cross-validation

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

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

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 on the slide. 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.

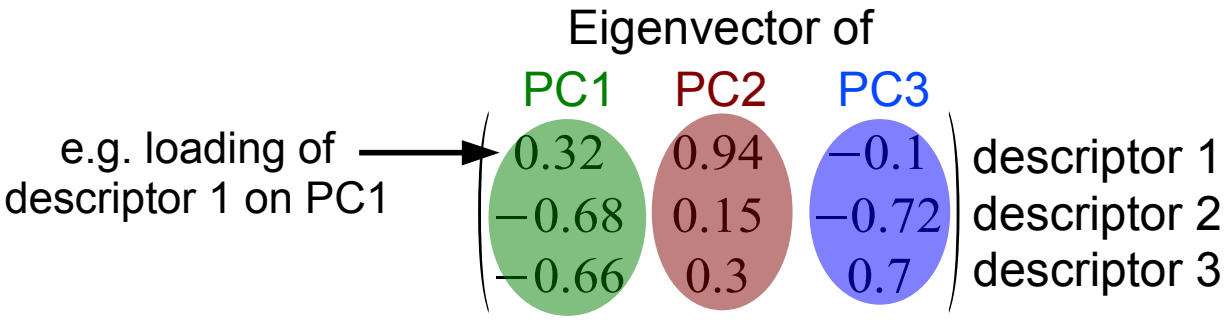
Jolliffe I.T. (2002) Principal component analysis, 2nd edn. Springer, New York.

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', give weight of original descriptor/variable on PC axes



- Easier to interpret: Correlation loadings $r_i = a_i \sqrt{\lambda_i}$
- Interpretation of descriptor/variable importance complicated if many variables load on a PC axis → Sparse PCA – introduces penalty term (cf. LASSO):

$$\arg\max_{a} \left(a^T \Sigma \; a\right) - \lambda_1 \|a\| \quad \text{with} \quad a^T a = 1$$

45

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

Note that the λ_1 used in the context of sparse PCA relates to the ℓ_1 penalty term and not to eigenvalues. This must not be confused (i.e. the λ in the context of correlation loadings are the eigenvalues).
$\|a\|$ is the norm of a (function that assigns a value to a vector)
For further details on sparse PCA refer Croux et al. (2011), Zou et al. (2006) and Hastie et al. (2015). Note that sparse PCA can also provide a solution for low n:p situations (and even $n < p$ situations) where the ordinary PCA fails (see Hastie et al. (2015:204 ff)).

Croux C., Filzmoser P. & Fritz H. (2011) Robust sparse principal component analysis. TU Vienna, Vienna. Freely accessible at: http://www.statistik.tuwien.ac.at/forschung/SM/SM-2011-2complete.pdf (more or less identical to: Croux C., Filzmoser P. & Fritz H. (2013) Robust Sparse Principal Component Analysis. *Technometrics* 55, 202–214.)

Hastie T., Tibshirani R. & Wainwright M. (2015). Statistical learning with sparsity: the lasso and generalizations. CRC Press, Taylor & Francis Group, Boca Raton. Freely accessible at: https://web.stanford.edu/~hastie/StatLearnSparsity_files/SLS_corrected_1.4.16.pdf

Zou H., Hastie T. & Tibshirani R. (2006) Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics* 15, 265–286. Freely accessible at: https://web.stanford.edu/~hastie/Papers/spc_jcgs.pdf

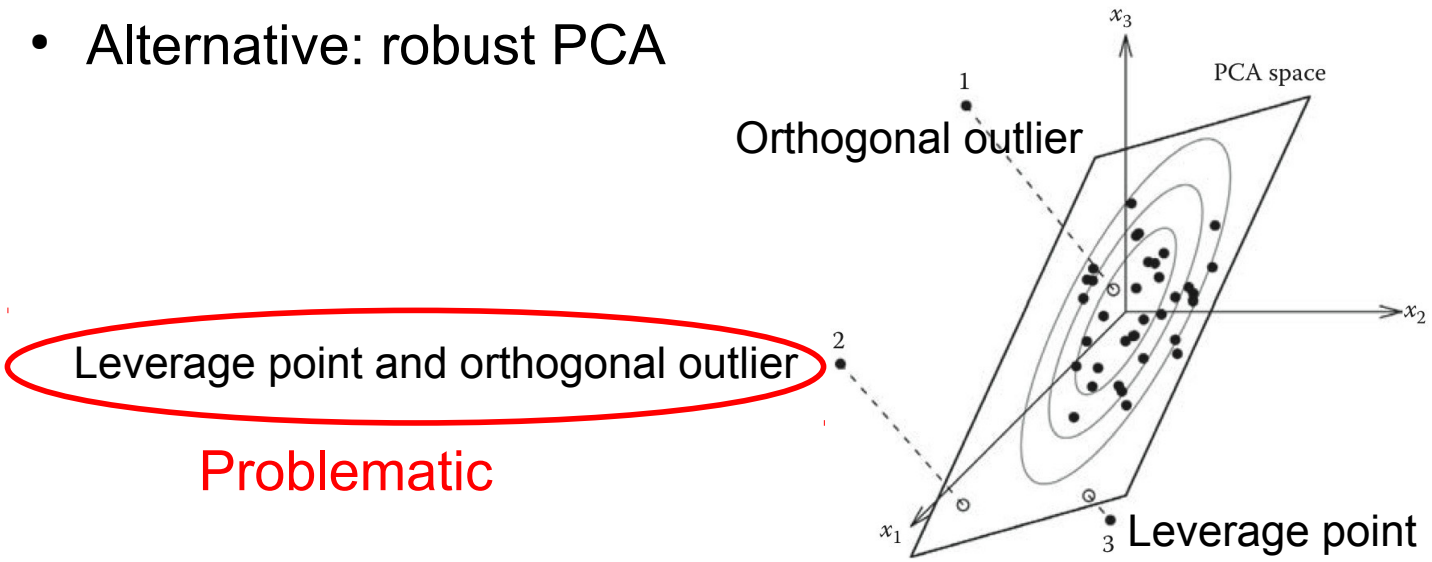
Introduction to multivariate analysis, ordination and PCA

Contents

46

PCA assumptions and diagnosis

- Independence of observations (temporal and spatial independence)
- Multivariate normality (depending on research goal)
- No serious outliers, diagnosis via orthogonal distance (OD) and score distance (SD)
- Alternative: robust PCA



Varmuza & Filzmoser 2009, ch.3

In the case of multivariate normal distribution of the data, the score distances can be approximated with a χ^2 distribution to detect leverage points. Orthogonal distances can be approximated with a normal distribution. See Varmuza & Filzmoser 2009, chapter 3.7.3 for details.

In case of serious outliers, use the robust PCA (see Varmuza & Filzmoser 2009, chapter 3.5).

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 are 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 (not recommended) require multivariate normality.

PCA assumptions and limitations

- Linear gradient of descriptors (rarely the case for species data, but often for environmental data)
- Euclidean distance used in PCA inappropriate for species data
- Alternatives: Transformation of data (Hellinger or Chord) or using different ordination method (e.g. NMDS)
- Adding noise variables to data increases fraction of variance on first axis, but has no meaning
- Best results for large n and high n:p (p = descriptors)

48

The most important assumption is that of a linear gradient of the descriptors over the object space. Linearity can be checked using scatter plots of the descriptors before analysis. Strongly skewed data and non-linear relationships may cause problems. If the assumption of linearity is violated, as for example can be the case for ecological (species) data, this can lead to serious problems (see the example in the R demonstration. It nicely displays how unimodal gradients distort the visualisation in a PCA). For such data, different ordination methods (discussed later) are available or the data can be transformed. Even if the gradients exhibit a linear relationship, the PCA relies on the Euclidean distance and we will in the following discuss the general problem of using the Euclidean distance for species data. See Legendre & Legendre (2012: 450ff) and Borcard et al. (2018: 130) for details.

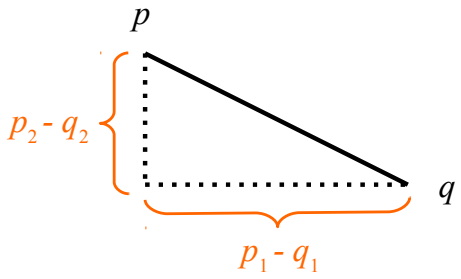
Euclidean distance and species data

$$d_{\text{Euclidean}}(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

Widely used distance measure

Two dimensional case:

$$d_{\text{Euclidean}}(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}$$



Species x Site matrix

Sites	Species		
	y_1	y_2	y_3
x_1	0	1	1
x_2	1	0	0
x_3	0	4	4

Euclidean distance →

Distance matrix

Sites	Sites		
	x_1	x_2	x_3
x_1	0	1.732	4.243
x_2	1.732	0	5.745
x_3	4.243	5.745	0

Sites x_1 and x_2 share any species, but have a smaller distance than sites sharing species (x_1 and x_3) → "Species abundance paradox"

→ Euclidean distance problematic for ecological data

49

Legendre & Legendre 2012: 300

Ecologically, two sites that share the same species pool are more similar than two sites that have any species in common. In the example, the Euclidean distance is smaller for two sites that share any species, demonstrating that it can be misleading. As a fix for species data, the Hellinger-transformation and Chord-transformation can be used before methods relying on the Euclidean distance.

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. Select the optimal number of principal components
5. How informative are the 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

Legendre P. & Legendre L. (2012) Numerical ecology, 3rd English ed. Elsevier, Amsterdam; Boston.

Extension: 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 necessarily imply low explanatory power → not necessarily a fix for low n:p ratios
- Alternative: Sparse principal component-guided regression

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.

For details on sparse principal component -guided regression see Tay et al. (2018).

Tay J.K., Friedman J. & Tibshirani R. (2018). Principal component-guided sparse regression. eprint arXiv:1810.04651, Freely accessible at: https://ui.adsabs.harvard.edu/abs/2018arXiv181004651T