Principal Component Analysis and Factor Analysis

Therese Leinonen

university of groningen

Seminar in Statistics and Methodology, 25th February, 2009

Overview

- Introduction
- Basic mathematics behind PCA and FA
- Background: on acoustic measures for vowel quality
- Example: PCA on Bark filtered vowel spectra
- PCA vs. FA

Introduction

- factor analysis (FA) and principal component analysis (PCA) are data reduction methods
- by analyzing the correlation between variables in a data set the variables can be reduced to a smaller amount of **factors** (FA) or **principal components** (PCA)
- both methods give a set of **loadings** an a set of **scores**
- **loadings** are correlations between original variables and extracted factors/components
- **scores** are values each data item gets on the extracted factors/components after data reduction

Basic mathematics behind PCA and FA

- starting point: a correlation matrix or a variance-covariance matrix
- by determining the eigenvalues and eigenvectors of the matrix variables that correlate highly can be clustered together on components(/factors)
- the eigenvectors (=loadings) are ordered by eigenvalue, highest to lowest, which gives the components in order of significance
- by choosing the most significant components the eigenvectors of these can be used to derive the scores (=reduced data set) for each subject
- the interpretability of the solutions can be enhanced by different rotation techniques

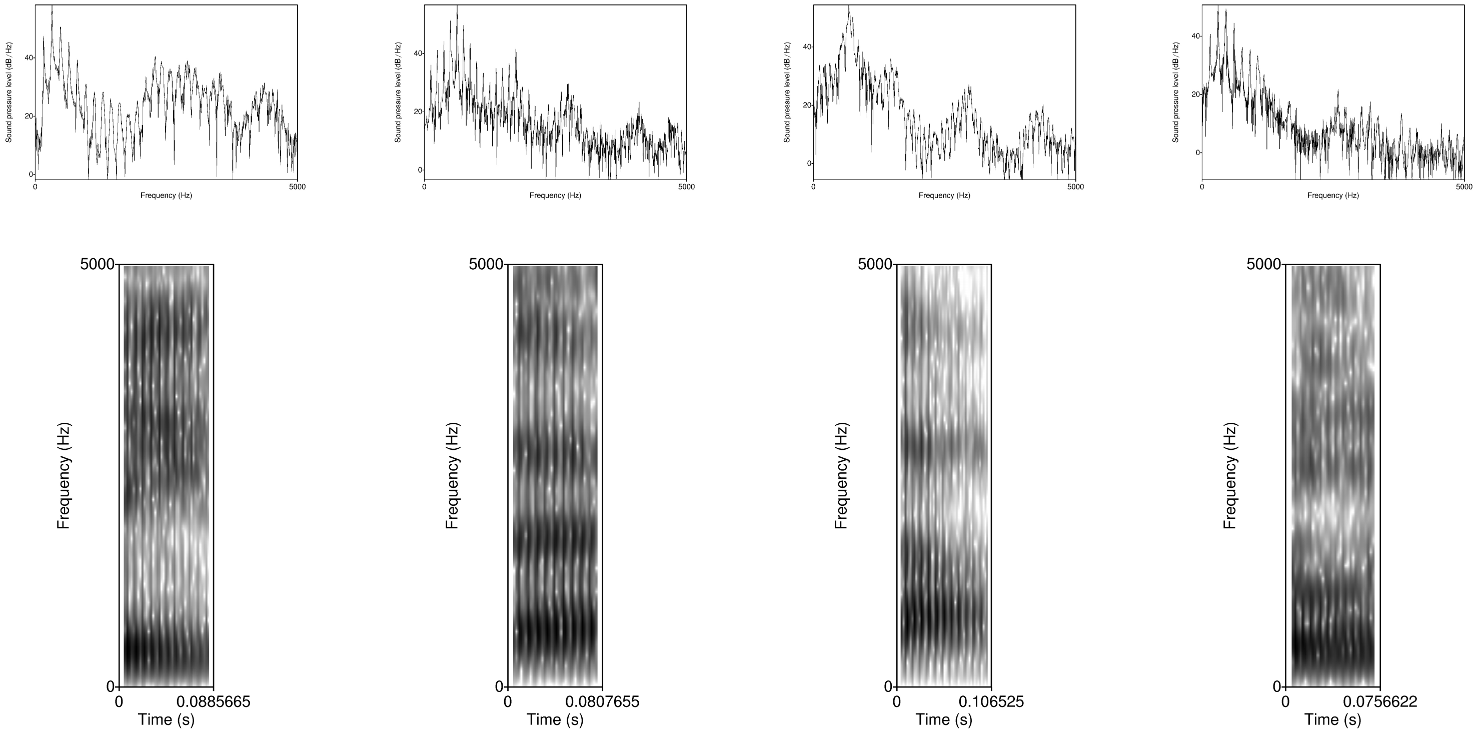
Background: on acoustic measures for vowel quality

- measuring formant frequencies is the traditional way of analyzing vowel quality acoustically

- formants = peaks in vowel spectra resulting from resonance in the vocal tract

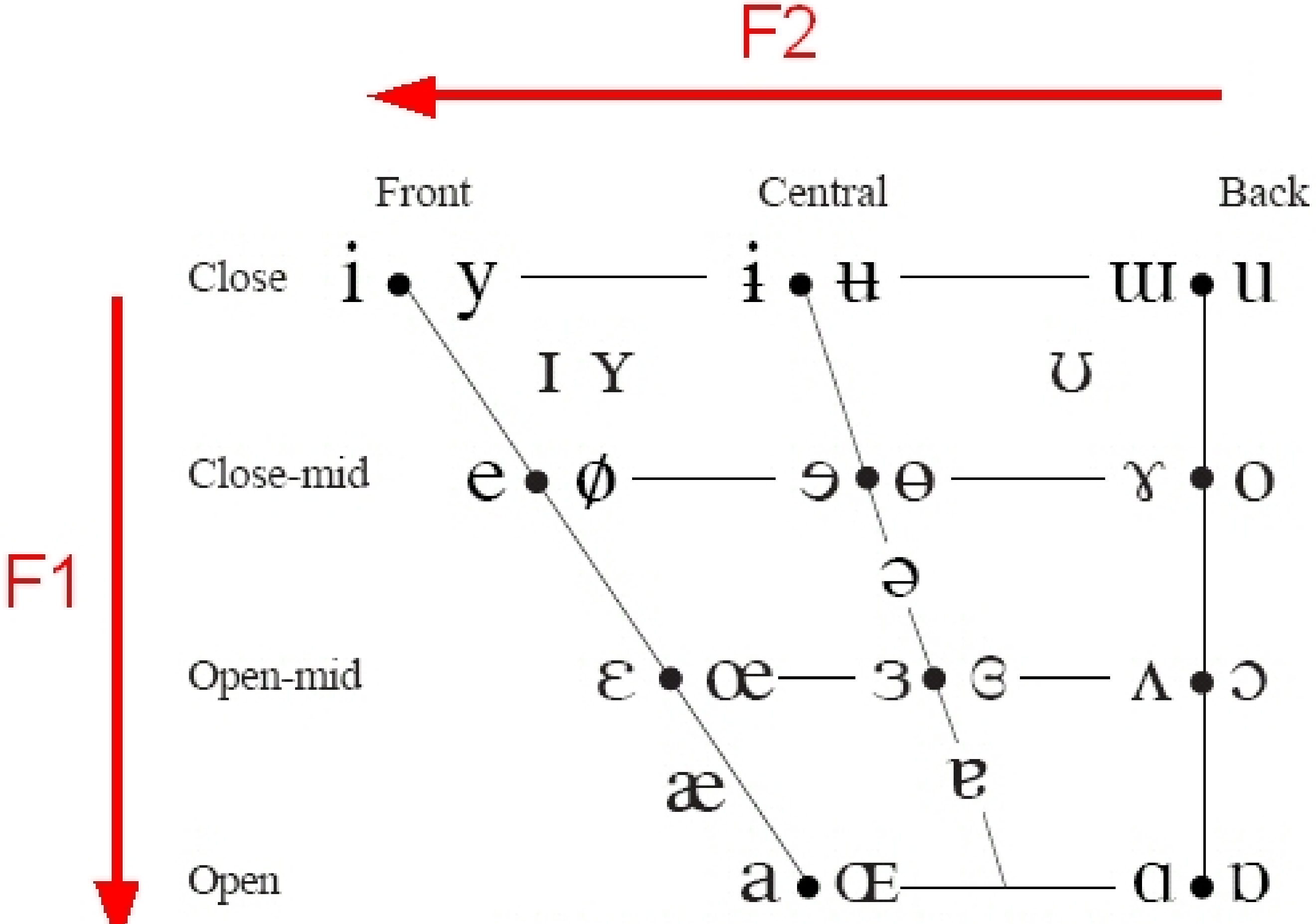
- the first two formants (F1 and F2), corresponding well with vowel height and backness, are usually enough to distinguish vowels from each other

- problem with formant measurements: automatic detection of peaks in the spectrum not reliable spectrum



spectrogram

vowel [i] [æ] [a] [u]



Where symbols appear in pairs, the one
to the right represents a rounded vowel.

Example: PCA on Bark filtered vowel spectra

- method introduced by Pols, Tromp and Plomp (1973), recently applied for analyzing sub-phonemic variation in Dutch vowels by Jacobi (2008)

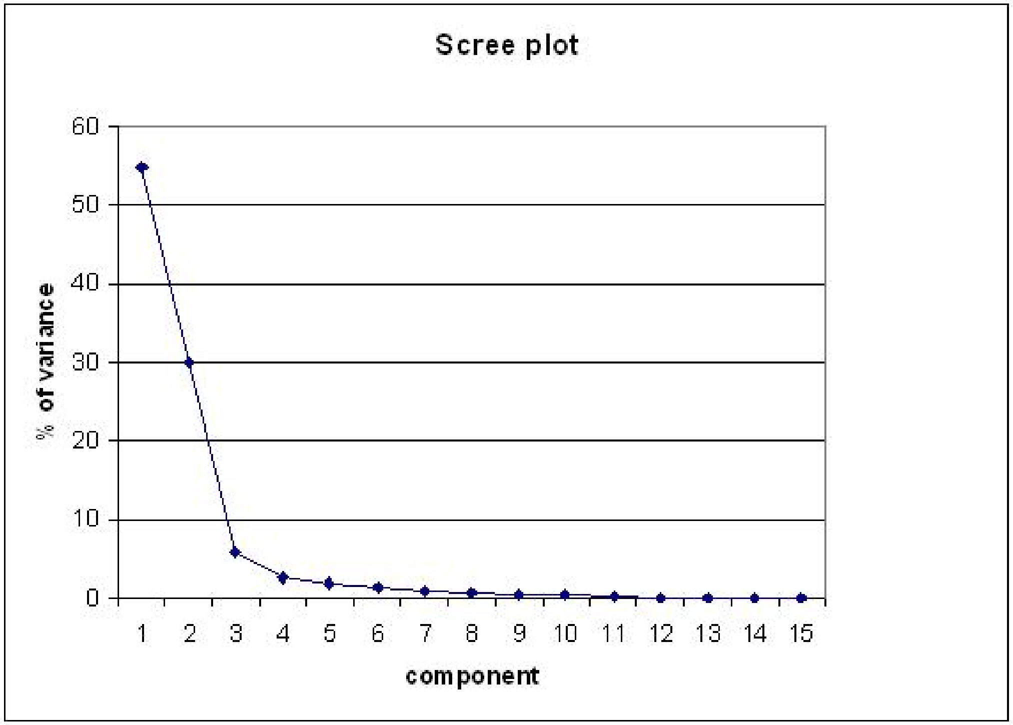
- alternative to formant measurements, can be fully automated

- vowel spectra filtered up to 17 Bark with every filter covering 1 Bark, mean intensity (dB) per filter band measured for each vowel pronunciation => every pronunciation is described by 17 variables

- since we know that vowel quality can generally be described with much fewer variables (2?) we use PCA for data reduction

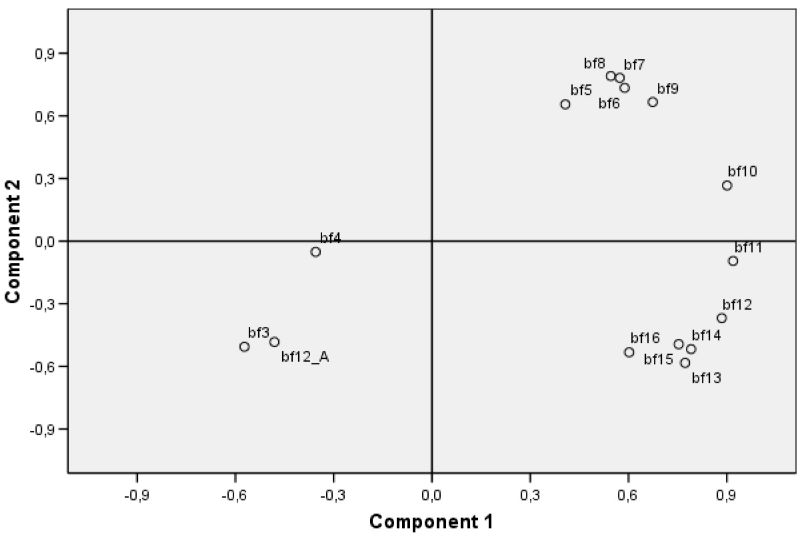
How many components should be extracted?

- scree plot
- eigenvalues higher than 1
- research design

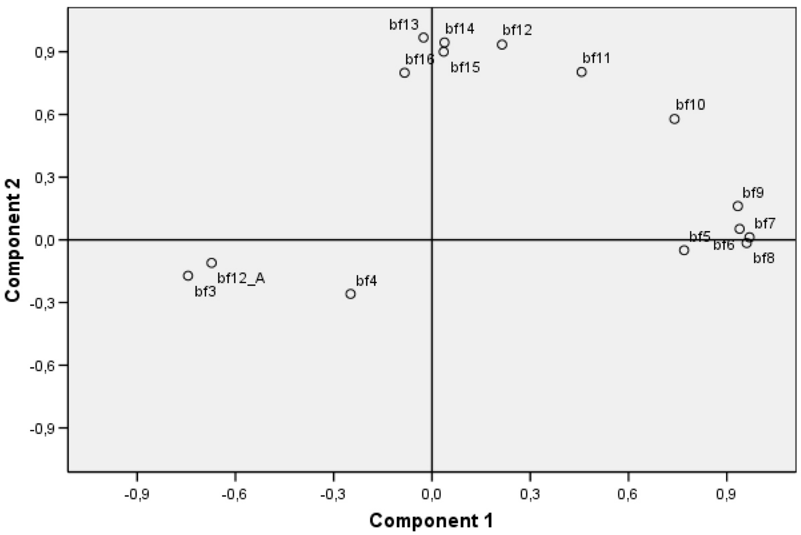


Loadings

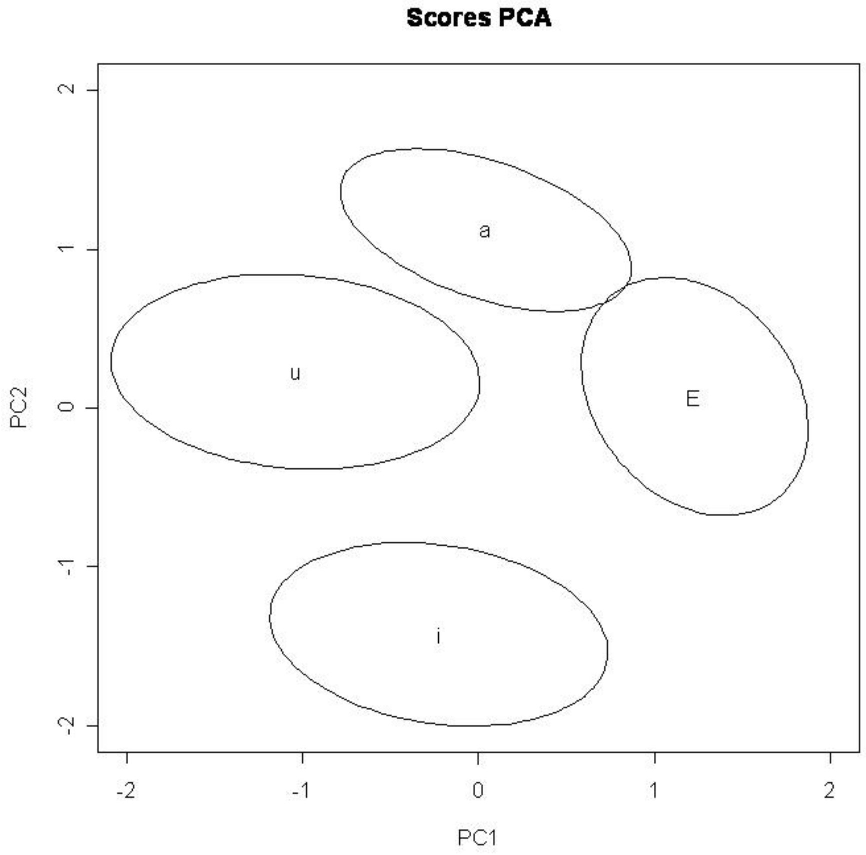
Component Plot

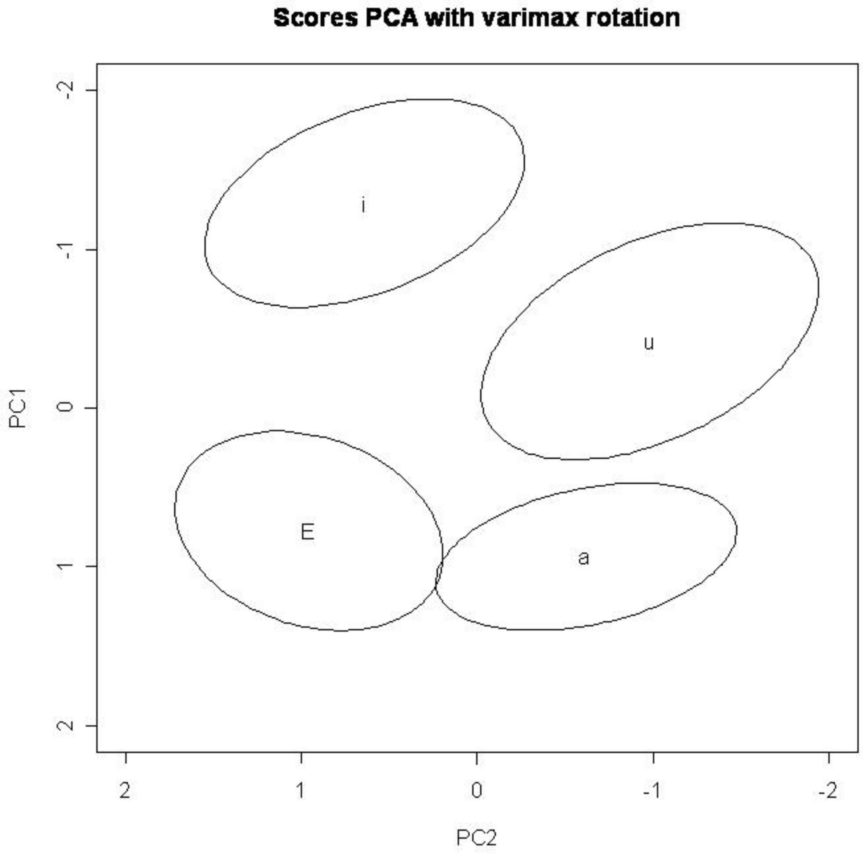


Component Plot in Rotated Space



Scores

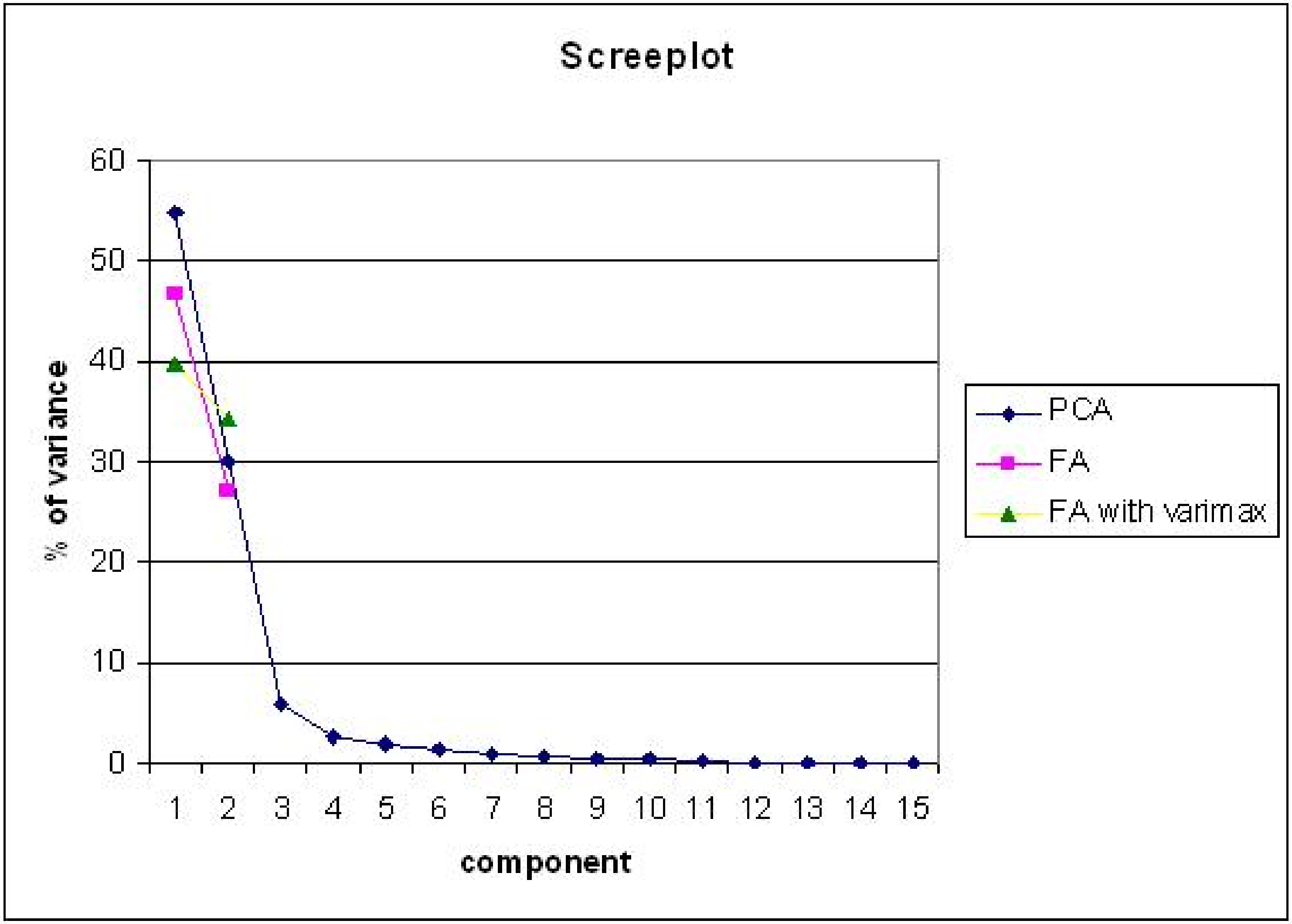

Scores PCA


Scores PCA with varimax rotation

PCA vs. FA

- PCA analyzes all variance present in the data set, while FA analysis only common variances (=uncontaminated by unique and error variability) => FA less sensitive to noise in the data
- in PCA the first component explains as much as possible of the total variance, the second component as much as possible of the still remaining variance etc. => when interpreting components one should bare in mind that most of the variance has been explained by previous components
- PCA should be used if you want an empirical summary of the data, FA if the study is based on assumed underlying factors (Tabachnik and Fidell 2007)
- however, if the study includes a large number of variables (>30) and communalities are high (>0.7) different solutions with PCA and FA are unlikely (Field 2005)

PCA vs. FA


Screeplot
60
50
40
30
20
10
0
% of variance
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
component
PCA
FA
FA with varimax

PCA vs. FA: correlating scores with formants

	F1	F2
F1	1	-0.115
F2	-0.115	1

	PC1	PC2
F1	0.606	0.589
F2	0.420	-0.727

	PC1 rotated	PC2 rotated
F1	0.866	0.122
F2	-0.385	0.741

	factor1 rotated	factor2 rotated
F1	0.848	0.123
F2	-0.394	0.752

How can we use the results of a PCA or FA?

1. interpret the loadings and scores of the analysis as such

2. use the results in subsequent analyses (MANOVA, multiple regression etc.)

References

Field, A.(2005), *Discovering Statistics Using SPSS*, 2nd edn, SAGE, London.

Jacobi, I.(2008), *On Variation and Change in Diphthongs and Long Vowels of Spoken Dutch*, PhD thesis, Universiteit van Amsterdam.

Pols, L. C. W., Tromp, H. R. C. and Plomp, R.(1973), Frequency analysis of Dutch vowels from 50 male speakers, *Journal of the Acoustical Society of America* **53**, 1093–1101.

Tabachnik, B. G. and Fidell, L. S.(2007), *Using Mulitvariate Statistics*, 5th edn, Pearson.