Principal Component Analysis (PCA) and Exploratory Factor Analysis (EFA)

Ana Alina Tudoran

Aarhus University, 2024

anat@econ.au.dk

Agenda

General purpose

Principal Component Analysis

Exploratory Factor Analysis

Introduction

- PCA and FA are applied to discover which variables in the set form *coherent subsets* relatively independent of one another
- These variables are combined into *factors* or components
- Factors are thought to reflect underlying processes that have created *the correlations among variables*
- Objectives:
  - for developing objective measures for abstract constructs e.g. personality, intellectual ability
  - for summarizing data and testing a theory about underlying process
  - for data reduction and multicolinearity resolution

Introduction

- Start with a very large number of items (observed variables), collected using a *survey*
- Possibly these items were extracted based on a qualitative research
- In *exploratory* FA, factors are discovered
- In *confirmatory* FA, factors are confirmed

Introduction

- Mathematically, factors/components are *linear combinations* of the items
- They aim to summarize the *observed correlation matrix*
- The number of factors/components extracted is usually far fewer than the number of items

Main steps

- Select and measure a set of variables
- Check the correlation matrix to run PCA or (E)FA
- Extract a set of factors
- Determine the number of factors
- Rotate the factors to increase interpretation - *many rotations*
- Interpret the factors - *scientific utility*
- Establish the construct validity of the factors - *a large step*

Matrices

- *Observed correlation matrix*
- *Reproduced correlation matrix*
- The difference is the *residual matrix*

Matrices in the solution

- *Orthogonal rotation* - factors are uncorrelated
  - loading matrix - *meaning of factors*
- *Oblique rotation* - factors are correlated
  - factor correlation matrix
  - structure matrix
  - pattern matrix - *meaning of factors*
- Both rotations return a factor-score coefficients matrix - to predict factor scores for each individual
- STATA, R, LIREL, AMOS etc, may call these matrices differently

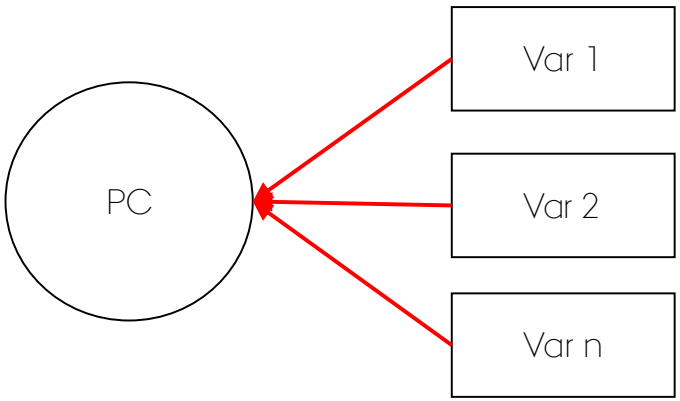
The differences between PCA and FA

- Preparation of the correlation matrix
  - PCA: all the variance in the observed variables is analyzed
  - FA: only the common variance in the observed variables is analyzed
- Underlying theory
  - PCA: dimension reduction and visualization
  - FA: theory development and testing

Theoretically PCA vs. FA

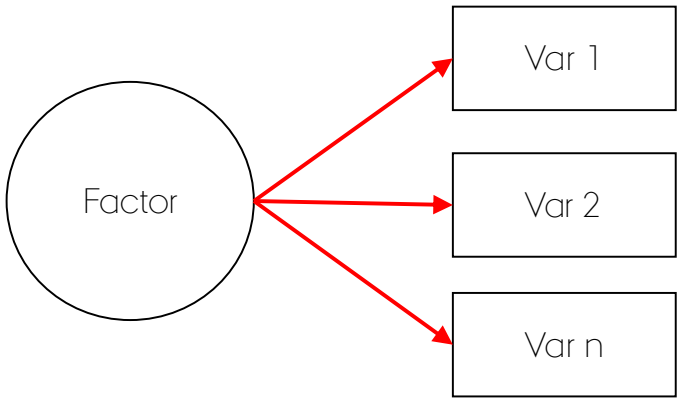
Principal Component Analysis

- The component is a linear combination of observed variables



(Exploratory) Factor Analysis

- Observed variables are a linear combination of the underlying and unique factors



Agenda

General purpose

Principal Component Analysis

Exploratory Factor Analysis

PCA - An example

- Consider a data set of p consumer perceptions about a company, $X_1, X_2, ...X_p$
- To better understand the relationships between these p features, we could look at their correlation matrix
- Bivariate correlation too many, not too informative overall
- PCA aims to find groups of these highly correlated variables

How PCA works?

- Given p observable variables, $X_1, X_2, ...X_p$ in a dataset with n observations
- $Z_1, Z_2, ...Z_M$, principal components represent linear combinations of the original variables

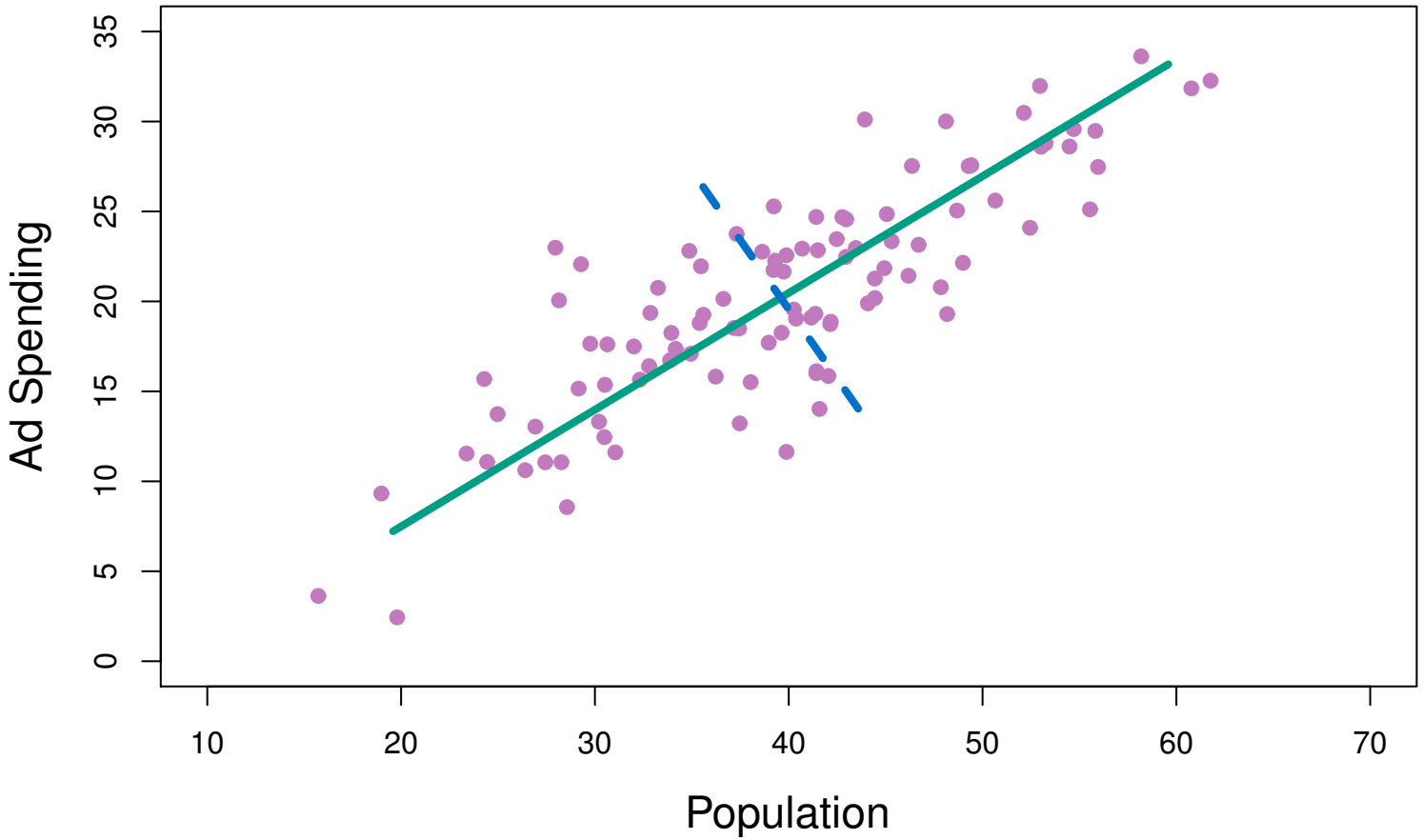
$$Z_m = \sum_{j=1}^{p} \Phi_{jm} X_j$$

- Φ are called **loadings** found in **matrix rotation** in R

Interpretation of components

- The *1st* principal component, Z_1, is that pc along which the observations *vary most*
  - If we project the data on the line represented by Z1 (i.e. PC1), then the resulting projected observations would have the largest possible variance
- The *2nd* principal component, Z_2 is that along which the observations vary second most, subject to the *constraint* that Z_1 uncorrelated with Z_2, if orthogonal rotation.
- so on..

Example with 2 items and 2 PC extracted



PC1

- The 1st PC mathematically is:

$$Z_1 = 0.839 \times (pop - \overline{pop}) + 0.544 \times (ad - \overline{ad})$$

where loadings:

$$\Phi_{11}^2 + \Phi_{21}^2 = 1$$

- This particular linear combination also defines the line that is closest to all n of the observations.

PC2

- The 2st PC mathematically is:

$$Z_2 = 0.544 \times (pop - \overline{pop}) - 0.839 \times (ad - \overline{ad})$$

where loadings:

$$\Phi_{11}^2 + \Phi_{21}^2 = 1$$

- This particular linear combination is that along which the observations vary second most, subject to the *constraint* that Z_1 uncorrelated with Z_2.

What the scores are?

- **Scores** are constructed as weighted averages of the original variables for each individual
- In our example:

$$z_i = 0.839 \times (pop_i - \overline{pop}) + 0.544 \times (ad_i - \overline{ad})$$

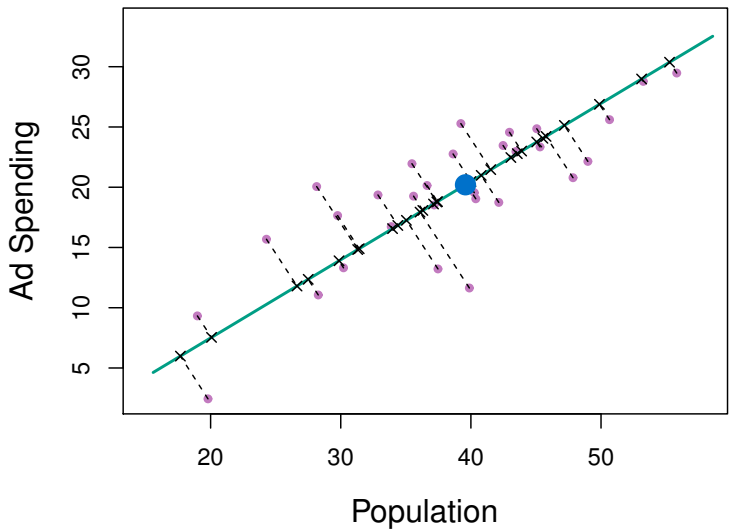
$$z_i = 0.544 \times (pop_i - \overline{pop}) - 0.839 \times (ad_i - \overline{ad})$$

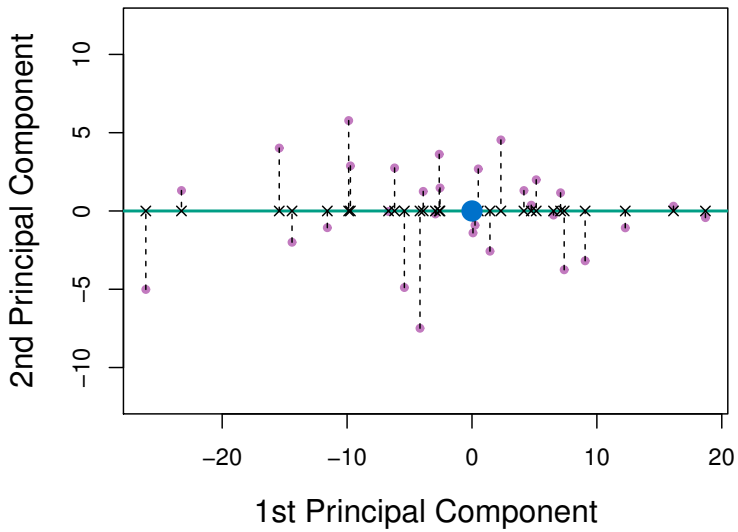
for all i=1:N individuals in the dataset

- In R: **matrix x**

Scores visually

- First pc score for any ith observation is the distance in the x-direction of the ith cross from zero





Negative and Positive score

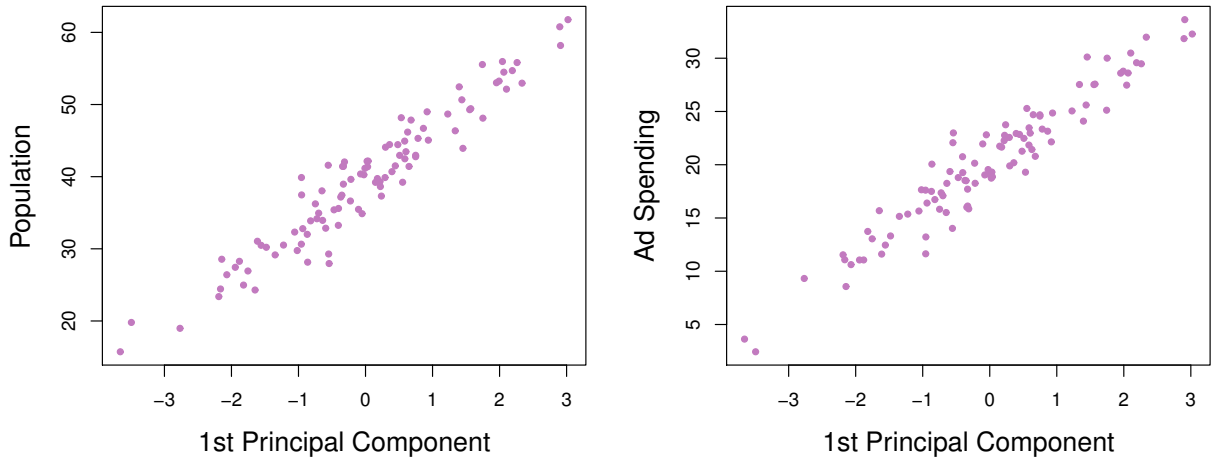
- Given the loadings are all positive (as here), if the z score is negative, that is:

$$z_1 = 0.839 \times (pop_i - \overline{pop}) + 0.544 \times (ad_i - \overline{ad}) < 0$$

- It means the ID i scores below the average on both variables
- A positive score suggests the opposite

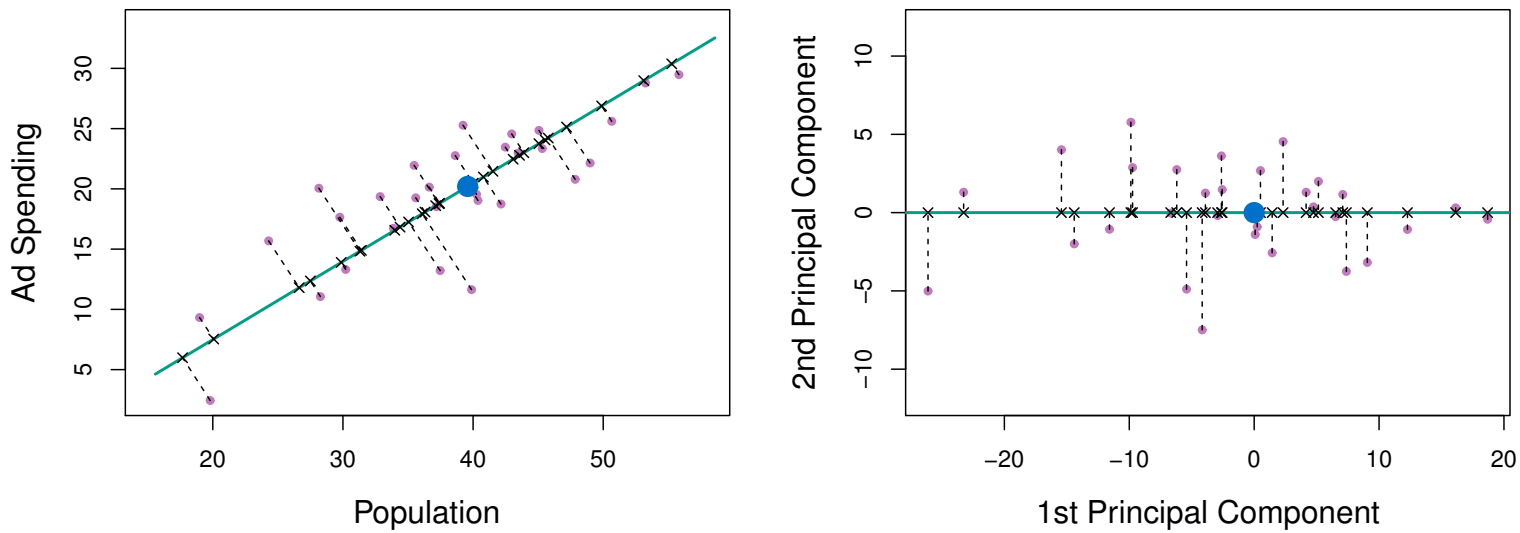
How representative are PC1 scores?

- If the items have approx. a linear relationship, a single-number summary will work well in representing the original items
- Example below: PC 1 captures most of the information in the pop and ad predictors



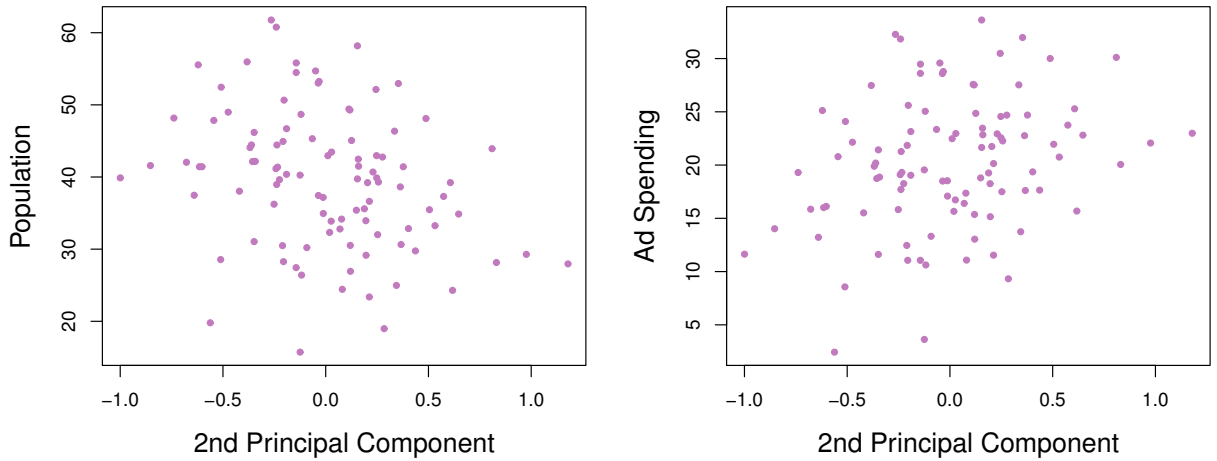
PC2 captures much less information

- We note in Figure 16.5 that the PC scores in the second dimension display a lower variability (between -5 and 5)



How representative are PC2 scores?

- Example below: there is little relationship between PC2 and the two variables



- Concl.: here one only needs the first PC to accurately represent pop and ad.

How many PC to select?

Criteria:

1. Cumulative proportion of variance explained >0.60
2. Eigenvalues >1
3. Scree plot and elbow rule for 1.and 2.
4. Based on supervised techniques using cross-validation
5. A-priori based on previous studies

Variance explained

Select the number of PC such that the cumulative proportion of variance explained is at least .60

```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5
Standard deviation     1.8721 1.7399 1.3178 1.1319 0.78636
Proportion of Variance 0.2921 0.2523 0.1447 0.1068 0.05153
Cumulative Proportion  0.2921 0.5443 0.6891 0.7958 0.84734
                          PC6     PC7     PC8     PC9    PC10
Standard deviation     0.7454 0.67782 0.53969 0.46246 0.41499
Proportion of Variance 0.0463 0.03829 0.02427 0.01782 0.01435
Cumulative Proportion  0.8936 0.93193 0.95620 0.97402 0.98837
                          PC11    PC12
Standard deviation     0.36269 0.08930
Proportion of Variance 0.01096 0.00066
Cumulative Proportion  0.99934 1.00000
```

Variance explained

- Total Variance present in the data set (assuming the variables were standardized):

$$\sum_{j=1}^{p} Var(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2$$

- Variance Explained (VE) by each component, m:

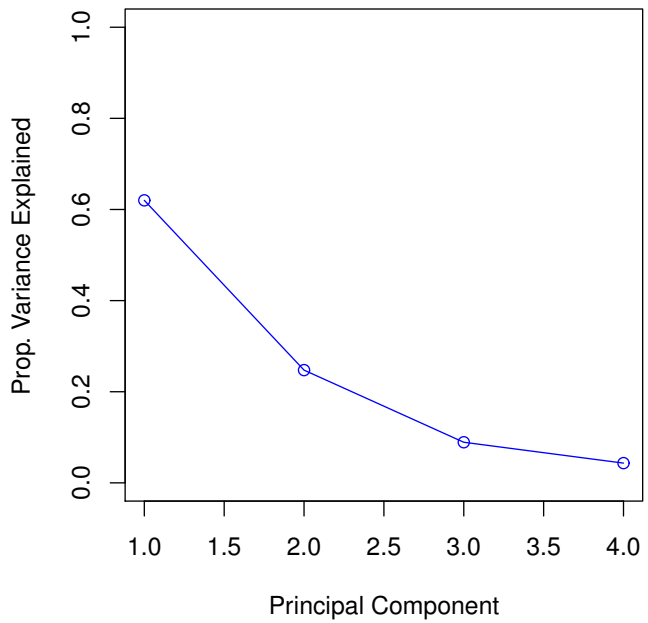
$$\frac{1}{n} \sum_{i=1}^{n} z_{im}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \Phi_{jm} x_{ij} \right)^2$$

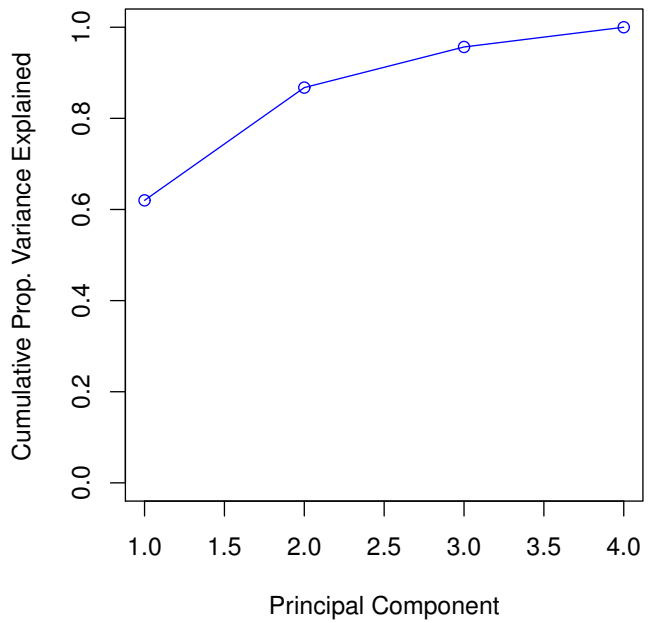
- The proportion of VE of any component is:

$$\frac{\sum_{i=1}^{n} \left(\sum_{j=1}^{p} \Phi_{jm} x_{ij} \right)^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2}$$

Plot of variance explained

Display the proportion of variance explained by each factor and identify the elbow. Select the number of PC above (or incl.) the elbow.





Eigenvalue

Eigenvalue represents the variance in all of the observed variables which is accounted for by that principal component

Communality of an item

- **Communality** represents the variance in a given observed variable explained by all the pc or by the pc which are extracted. Communality = 1-Uniqueness

Communalities

	Initial	Extraction
X6 - Product Quality	1,000	,798
X7 - E-Commerce Activities	1,000	,780
X8 - Technical Support	1,000	,894
X9 - Complaint Resolution	1,000	,890
X10 - Advertising	1,000	,585
X12 - Salesforce Image	1,000	,860
X13 - Competitive Pricing	1,000	,661
X14 - Warranty & Claims	1,000	,891
X16 - Order & Billing	1,000	,806
X18 - Delivery Speed	1,000	,894

Example

Component Matrix[a]

	Component			
	1	2	3	4
X6 - Product Quality	,020	,488	-,115	,739
X7 - E-Commerce Activities	,544	-,541	,380	,216
X8 - Technical Support	,182	,579	,700	-,187
X9 - Complaint Resolution	,819	,267	-,362	-,129
X10 - Advertising	,529	-,413	,161	,330
X12 - Salesforce Image	,631	-,536	,377	,180
X13 - Competitive Pricing	-,011	-,695	,009	-,422
X14 - Warranty & Claims	,294	,544	,688	-,185
X16 - Order & Billing	,783	,270	-,319	-,135
X18 - Delivery Speed	,842	,164	-,370	-,148

a. 4 components extracted.

Eigenvalue

$$0{,}020^2 + \ldots + 0{,}842^2 = \ldots$$

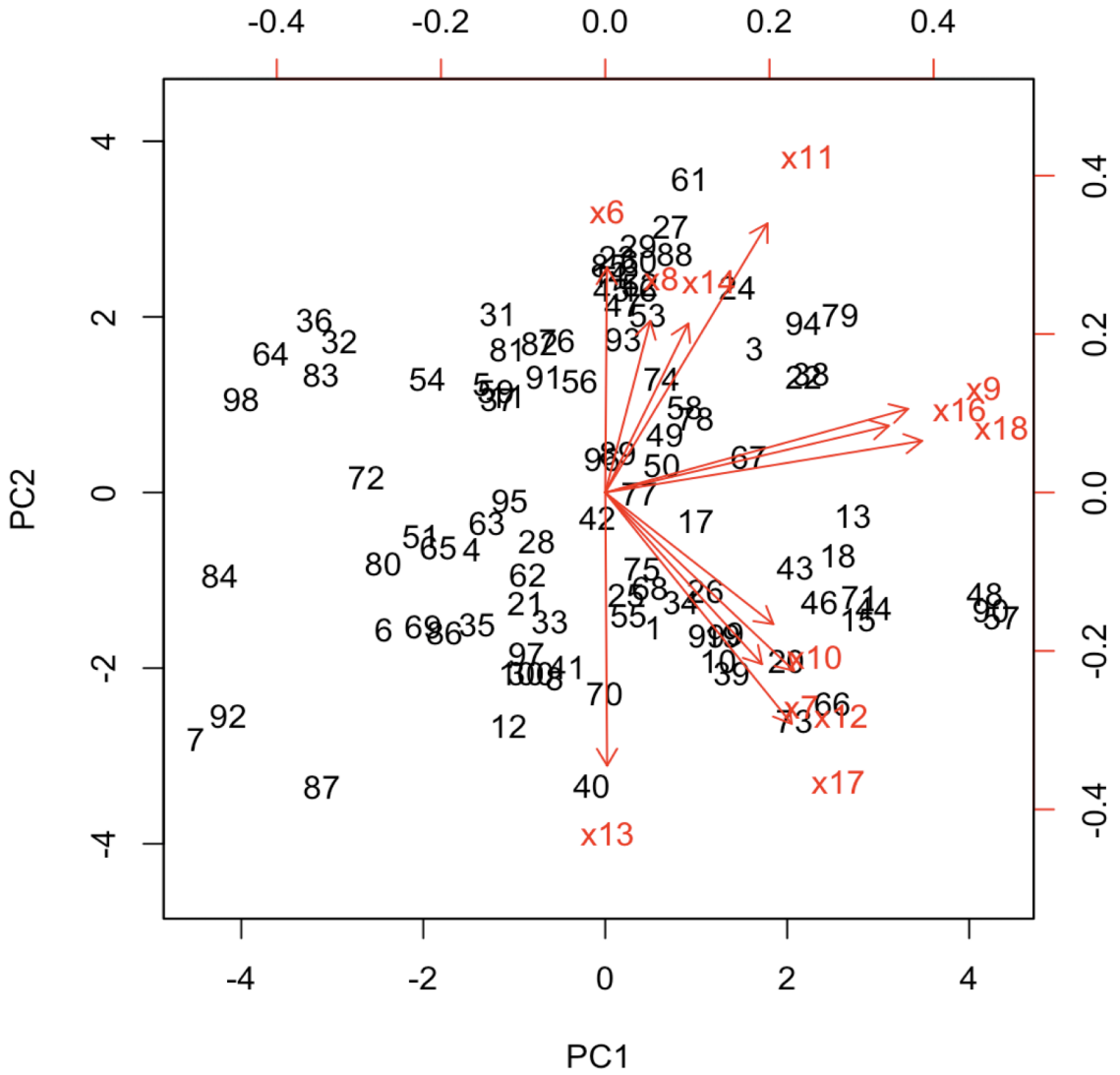
Communality

$$0{,}020^2 \ldots + 0{,}739^2 = 0{,}798$$

Interpret PC results

- **Biplot**: simultaneous display of factor loadings and factor scores in regard to the principal components discovered

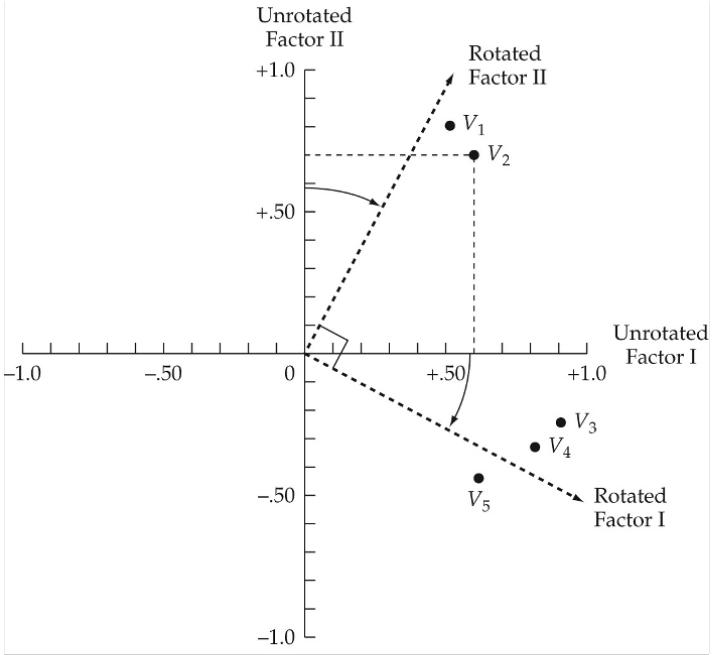
Example

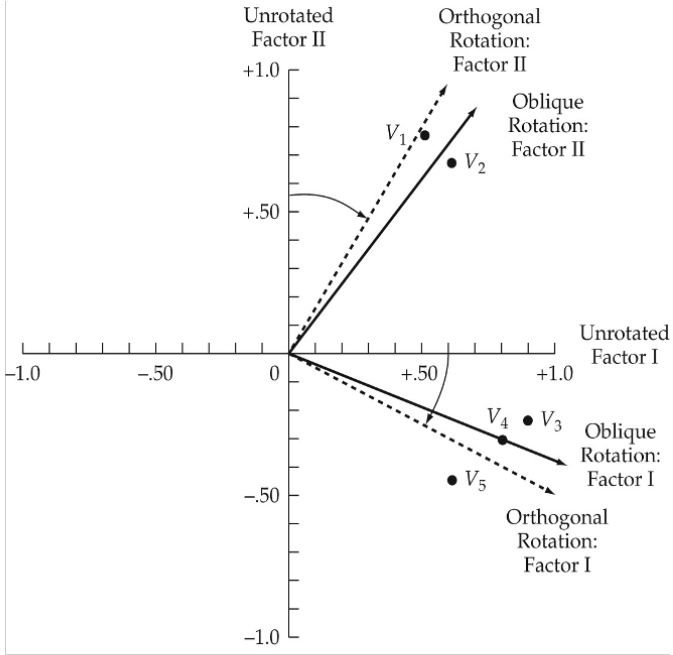

-0.4
-0.2
0.0
0.2
0.4
4
2
0
-2
-4
0.4
0.2
0.0
-0.2
-0.4
PC2
PC1
x11
x6
x14
x9
x16
x18
x10
x12
x17
x13

Rotation

- Orthogonal: axes are maintained at 90 degrees
- Oblique: axes are not maintained at 90 degrees

Example





	Unrotated Factor Loadings		Rotated Factor Loadings	
Variables	I	II	I	II
V1	.50	.80	.03	.94
V2	.60	.70	.16	.90
V3	.90	-.25	.95	.24
V4	.80	-.30	.84	.15
V5	.60	-.50	.76	-.13

PCA practical issues

- Measurement scales: continuous
- Sample size: at least 5 observations (ideally 10-20) per observed variable and at least 100 observations overall
- Linear relationship between observed variables
- Normal distribution for each observed variable

Summary

- If you suspect that some variables are redundant in your dataset, PCA can tell you the number of sources of variance in your data
- PCA does not discard any characteristics. It aims to reduce the overwhelming number of variables by constructing PC
- PC aim to account for most of the variance of the original variables

Application PCA

1. Check if PCA applies to your data
2. Run PCA model - $\mathrm{prcomp}(\mathrm{dataset}, \mathrm{scale{=}TRUE})$
3. Evaluate the rotation matrix (loadings)
4. Evaluate the variance explained by each PCs - $\mathrm{summary}()$
5. Decide how many PCs to use - $\mathrm{screeplot}()$
6. Visualize and interpret - $\mathrm{biplot}()$

EFA design

1. **Generate hypotheses about factors**
   - Statistical/practical: 5-6 hypothesized factors
   - Logical: all relevant factors of a process are required

2. **Select observable variables** (items, marker variables)
   - 5-6 items per factor
   - Factors with 1-2 items are not stable!
   - Items should be highly correlated with only one factor

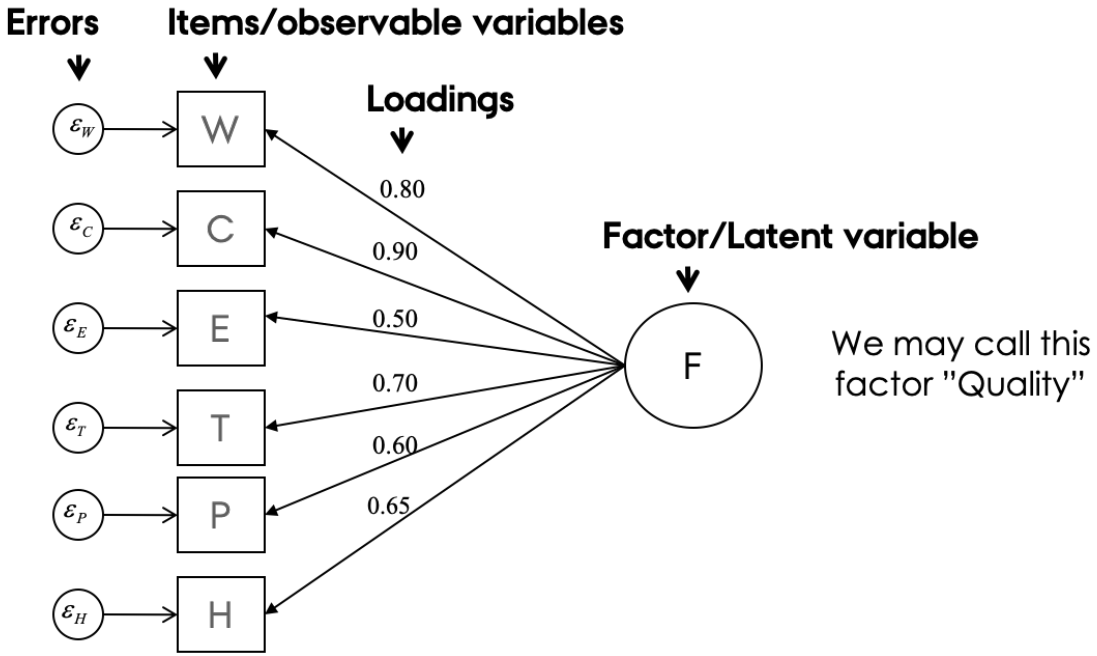
EFA practical issues

1. Sample: min 300 cases, but it depends on the no. of factors
2. Multivariate normality is assumed when statistical inference is used to determine the no. of factors
   - Univariate normality is typically assessed by skewness and kurtosis
   - Linear relationships among pair of variables are assessed by scatterplot
3. Some researchers standardize the variables

Is EFA applicable? Criteria:

1. Matrix of correlations should be >+-0.30
2. Partial correlations (anti-image matrix) should be small
3. Bartlett test - should be not significant; but this test is too dependent on N
4. Kaiser MSA test - should be >0.60

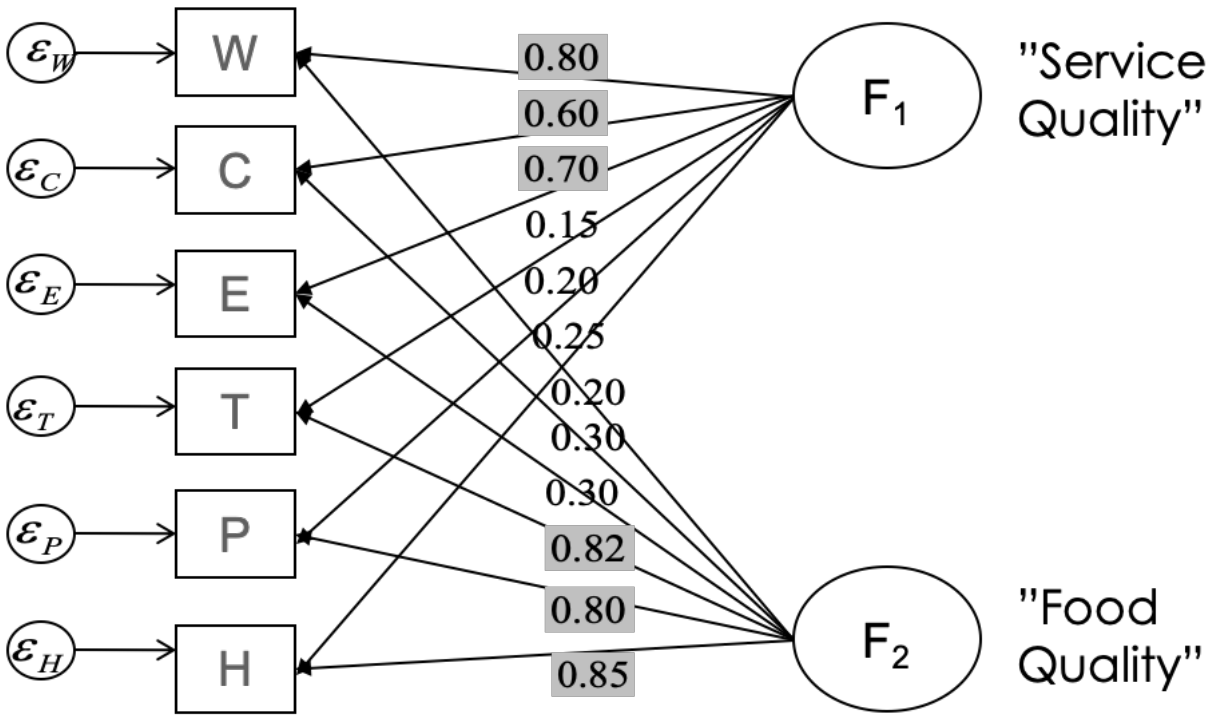
One factor - example



Relationships are regression equations

$$W = 0{,}80 \cdot F + \varepsilon_W$$
$$C = 0{,}90 \cdot F + \varepsilon_C$$
$$E = 0{,}50 \cdot F + \varepsilon_E$$
$$T = 0{,}70 \cdot F + \varepsilon_T$$
$$P = 0{,}60 \cdot F + \varepsilon_P$$
$$H = 0{,}65 \cdot F + \varepsilon_H$$

Two factors - example



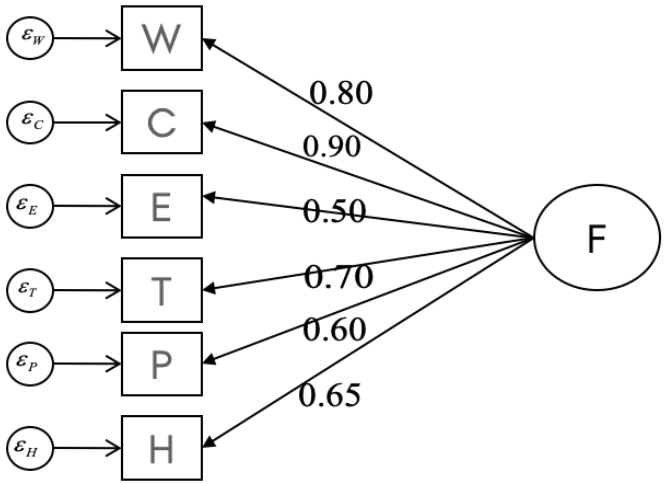
How many Factors to select?

Criteria are the **same as in PCA**.

1. A-priori based on theory
2. Cumulative proportion of variance explained >0.60
3. Eigenvalues >1
4. Scree plot and elbow rule
5. In combination with supervised techniques using cross-validation

Recall eigenvalue = factor importance

Eigenvalue represents the variance in all of the observed variables which is accounted for by that factor.



$$0{,}80^2 + \ldots + 0{,}65^2 = \\ = eigenvalue\ (of\ factor\ F)$$

Guidelines for loadings, eigenvalues and communalities

1. Std. loadings $\mp$ 0.30 to $\mp$ 0.40 are minimally acceptable
2. Std. loadings $>\mp$ 0.50 - necessary for practical significance
3. Communalities $>$0.50 - to retain the item in the analysis
4. Eigenvalues $>$1 - to retain the factor in the analysis

Factor internal consistency: Cronbach´s Alpha

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_i V_i}{V_t}\right) \quad \text{(Cronbach, 1951, p. 299)}$$

n= number of obs

V_t = total variance

V_i= variance of individual items

For each factor α should be $>$0.6

Checklist for EFA

1. Limitations
   - Outliers among cases
   - Sample size and missing data
   - Factorability of data
   - Normality and linearity of variables
   - Multicollinearity and singularity
   - Outliers among variables
2. Major analysis
   - Number of factors
   - Nature of factors
   - Type of rotation
   - Importance of factors
3. Additional analysis
   - Factor scores
   - Distinguish-ability and simplicity of factors
   - Complexity of variables
   - Internal consistency of factors

Application EFA

1. Check if FA applies to your data (same in PCA)
2. Run FA model
3. Evaluate the variance explained by each factor
4. Decide how many factors to extract and retain the factors
5. Evaluate the rotation matrix and rotate it
6. Evaluate the loadings and cross-loadings
7. Evaluate the internal consistency of each factor
8. Interpret the factors and give them names

References

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013)
An Introduction to Statistical Learning, Ch. 6.3.1., Ch.10.1, 10.2
Springer Texts in Statistics.

Tabachnich, B.G. and Fidell, L.(2007)
Using Multivariate Statistics, Ch. 13.
Pearson International Edition.