Universiteit Utrecht

Theory Construction and Statistical Modeling

Factor analysis

Welcome!

Factor Analysis

- Exploratory Factor Analysis (EFA) and Principal Components Analysis (PCA)
- Two related techniques
- Both often described as types of *factor analysis*
  - In R: use the package "psych"
    - `install.packages("psych"); library(psych)`
    - Functions: `principal()` and `fa()`
  - Controversy discussed in Preacher & McCallum
  - Confirmatory Factor Analysis (CFA) next week

EFA and PCA

- Statistical techniques in which researchers want to know, very generally:

Given a set of observed variables, how can I transform them to make a smaller set, while still retaining as much **information** as possible

- As much as possible, similar variables in my original set should relate to the same variable in my new set
- E.g. If I have 10, 50 or 100 variables, how can I make 2, 3 or 4 variables that capture as much as possible
- **Data-driven** approaches!

When is it useful?

1. Develop **measurement** tools or tests for latent variables
   – Personality, Intelligence, Depression
2. Investigate the dimensions of test items
3. Data reduction
   – Also called “dimension reduction”
   – E.g., solves multicollinearity in linear regression

When is it useful?

1. **Develop measurement tools or tests for latent variables**
   - Personality, Intelligence, Depression
2. Investigate the dimensions of test items
3. Data reduction
   - Also called "dimension reduction"
   - E.g., solves multicollinearity in linear regression

In Practice: Developing a measurement scale

1. Create a questionnaire with a very large number of items about a topic of interest
   – Student aptitude: school history, family history, health, personality, previous grades
2. Give questionnaire to random sample
3. Derive factors
   – E.g. Intelligence, Work ethic, Independence
4. Delete or add items depending on factor loadings
5. Repeat steps 2 to 4
6. Test validity of factors
   – E.g. predict future grades

Difference between PCA and EFA

- Goal:
  - **PCA**: reduce correlated observed variables to a smaller set of independent composite variables.
    - Data reduction!
    - Components describe the total **variance** in the dataset
  - **(E)FA**: assume or wish to test a theoretical model of latent factors causing observed variables.
    - Model says that observed variables covary **<u>because</u>** all variables are caused by an unobserved factor
    - Don’t know exactly how many factors or which factors cause which variables – ***Exploratory Factor Analysis (EFA)***
    - Strong theory on latent structure that you want to confirm/disconfirm – *Confirmatory Factor analysis*

  - PCA rotates axes to explain as much **variance** as possible,
    EFA **models the covariance matrix**.

Variance and Covariance

Sample Covariances (Girls)

	wordmean	sentence	paragrap	lozenges	cubes	visperc
wordmean	68,260					
sentence	28,845	25,197				
paragrap	21,718	12,864	12,516			
lozenges	23,947	13,228	9,056	61,726		
cubes	6,840	4,036	3,356	17,416	20,265	
visperc	13,037	12,645	8,335	26,531	14,931	47,175

PCA analyzes variance
EFA analyzes covariance

$$Cov_{xy} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{N-1}$$

or

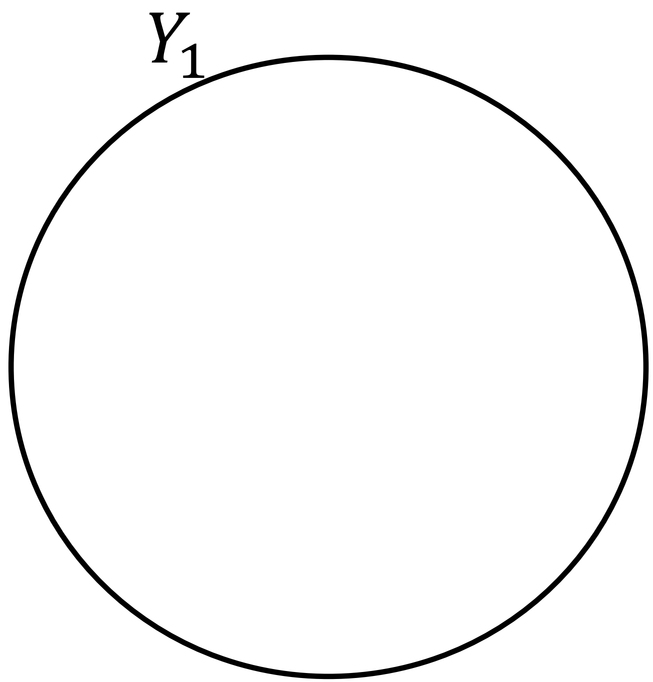
PCA EFA

Sample Correlations (Girls)

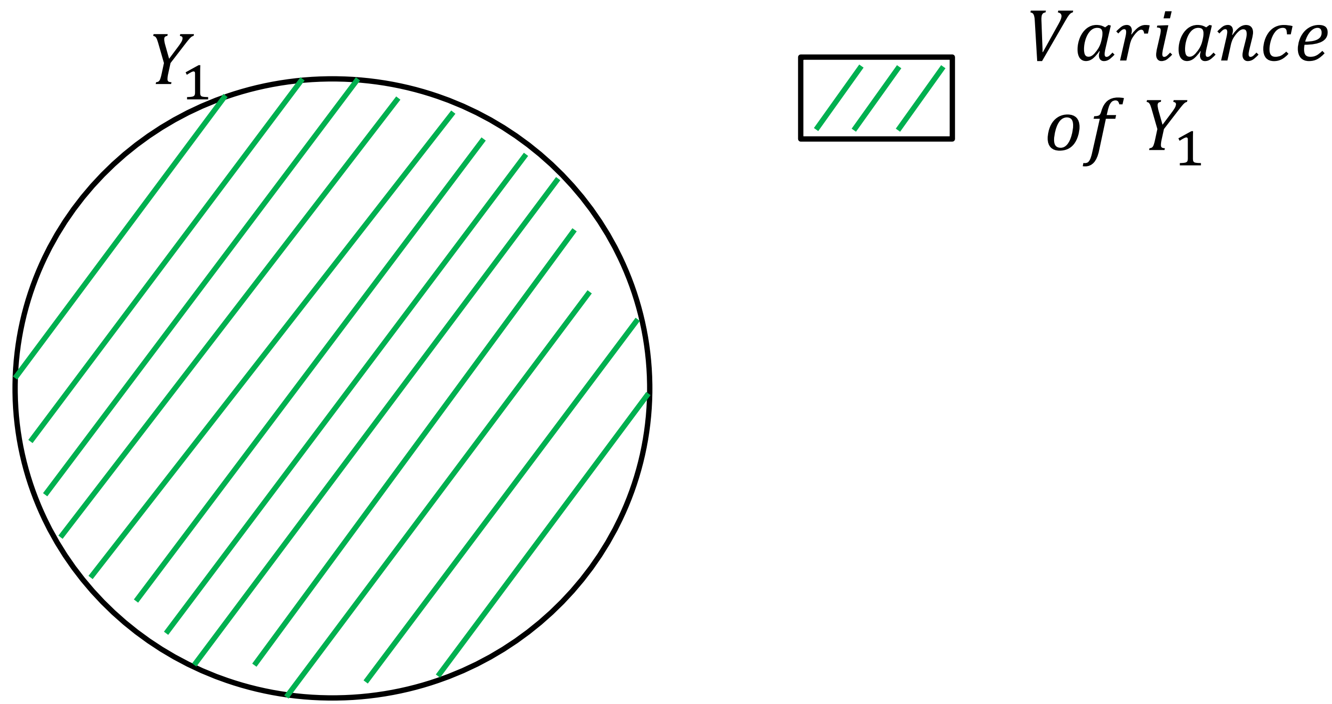
	wordmean	sentence	paragrap	lozenges	cubes	visperc
wordmean	1,000					
sentence	,696	1,000				
paragrap	,743	,724	1,000			
lozenges	,369	,335	,326	1,000		
cubes	,184	,179	,211	,492	1,000	
visperc	,230	,367	,343	,492	,483	1,000

$$r = \frac{Cov_{XY}}{S_X S_Y}$$

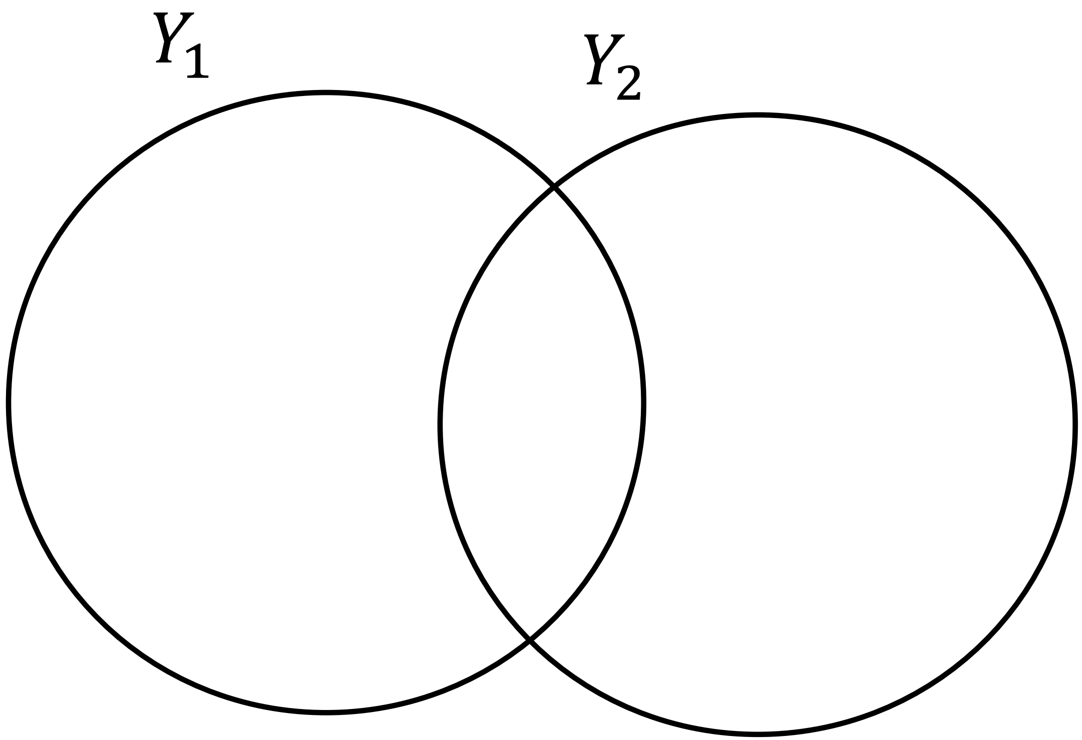
Variance and Covariance



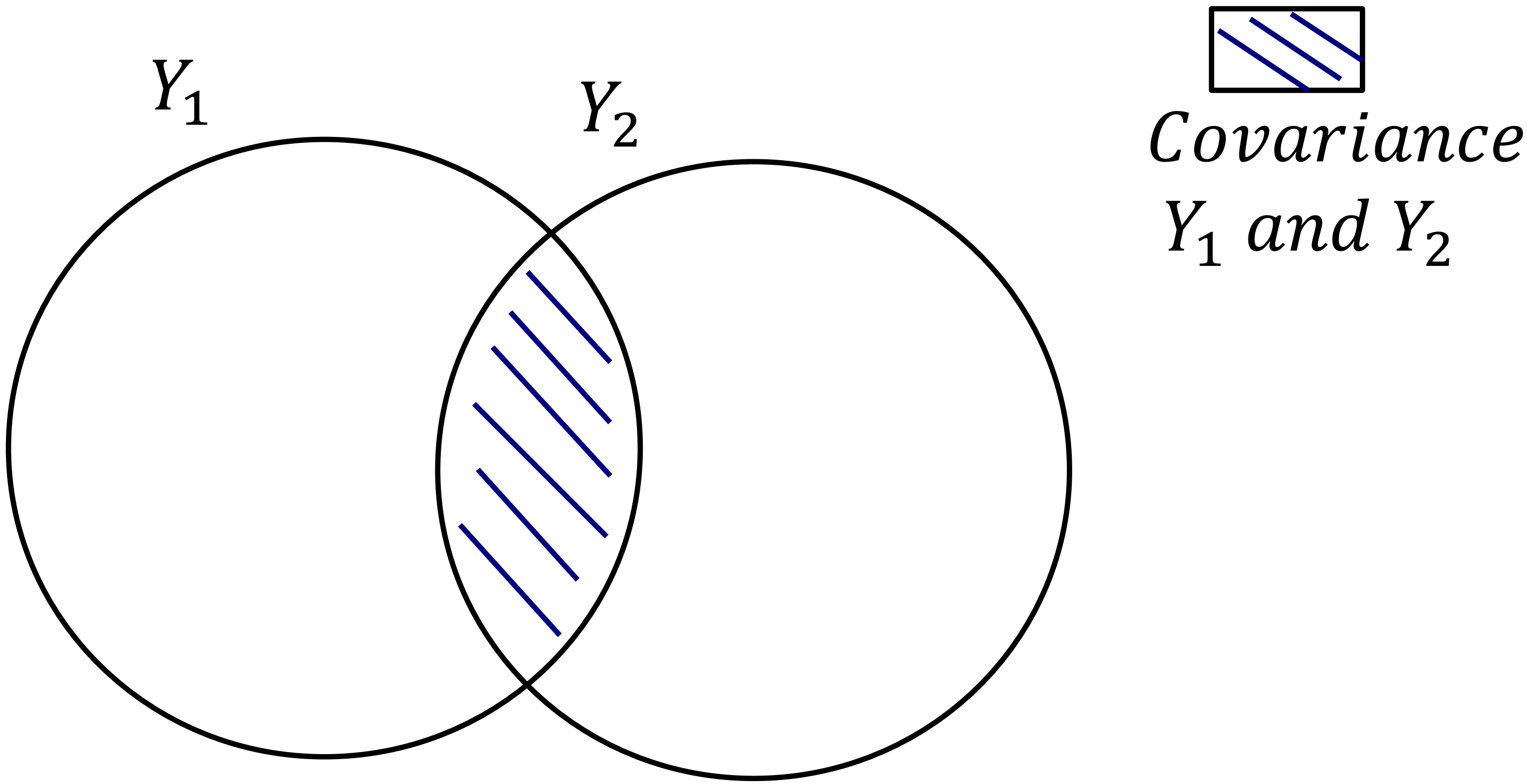
Variance and Covariance



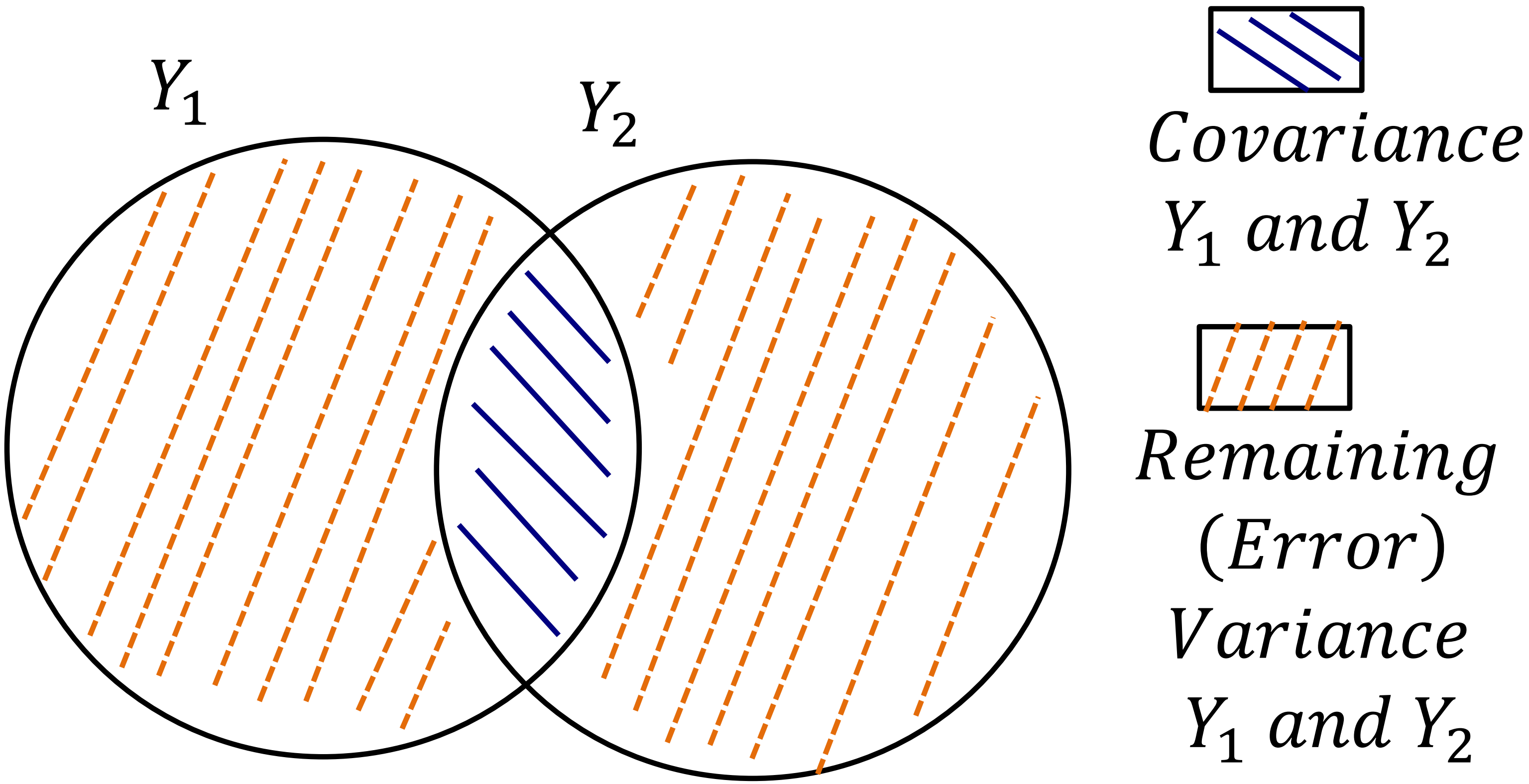
Variance and Covariance



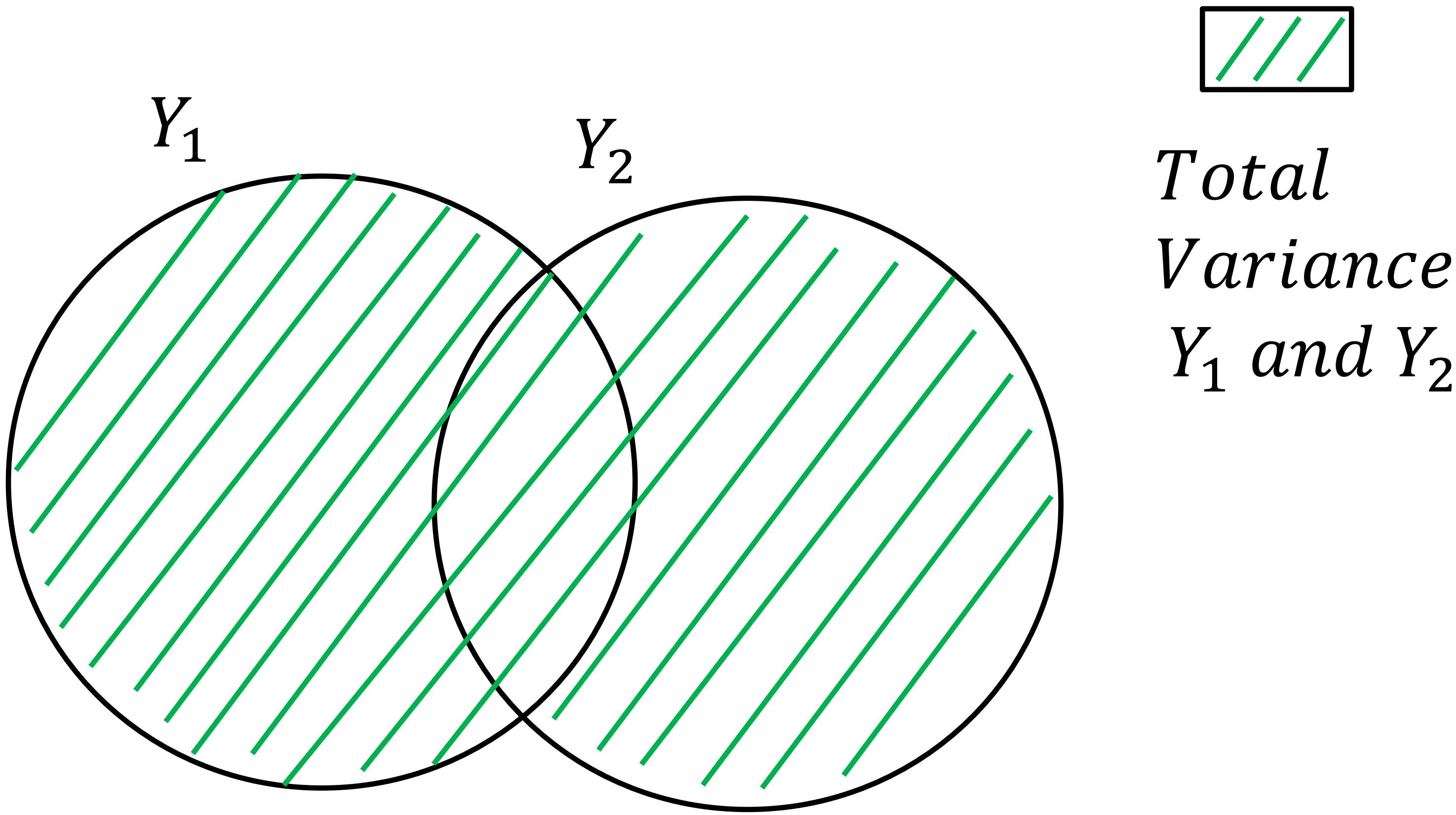
Variance and Covariance



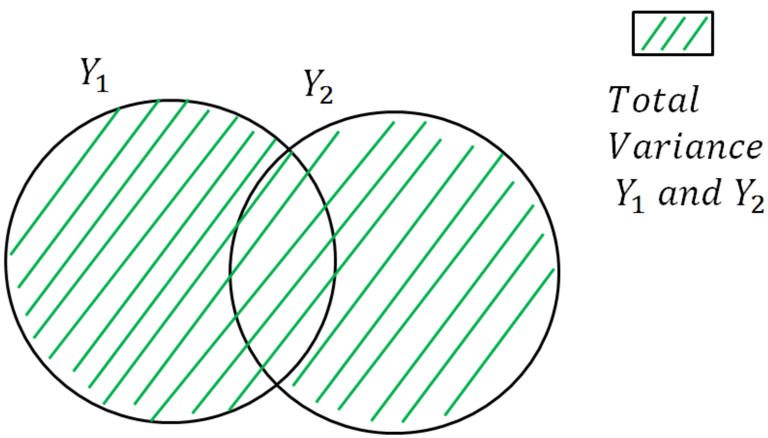
Variance and Covariance



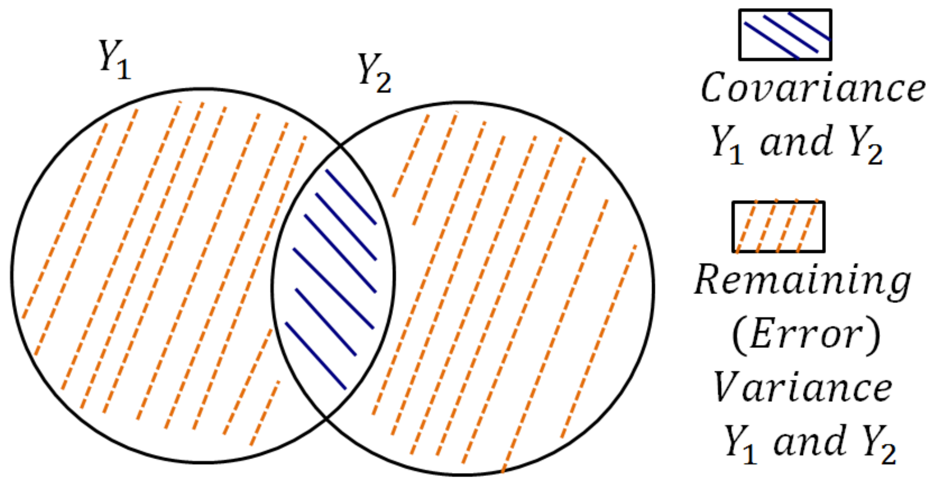
Variance and Covariance



Principal Components Analysis

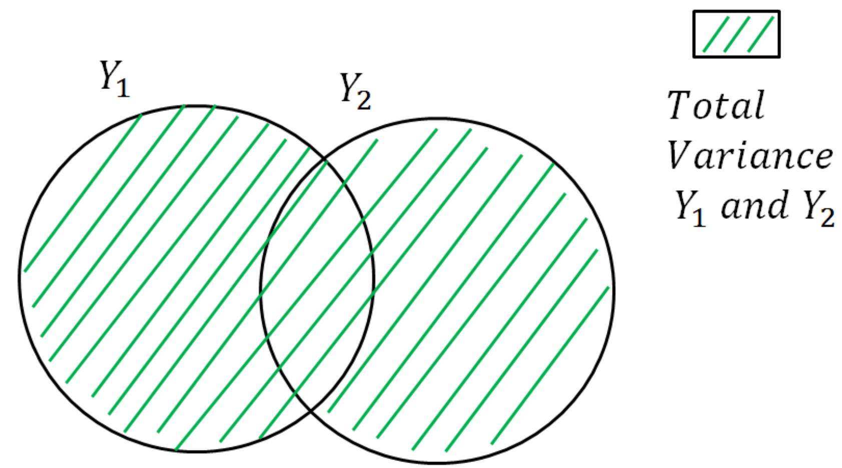


Exploratory Factor Analysis



PCA: Summarize variance

- For *n* variables, you obtain *n* components
- The first component explains most variance, second explains second-most, etc.
- Each component is uncorrelated with all others
(but see Rotation)

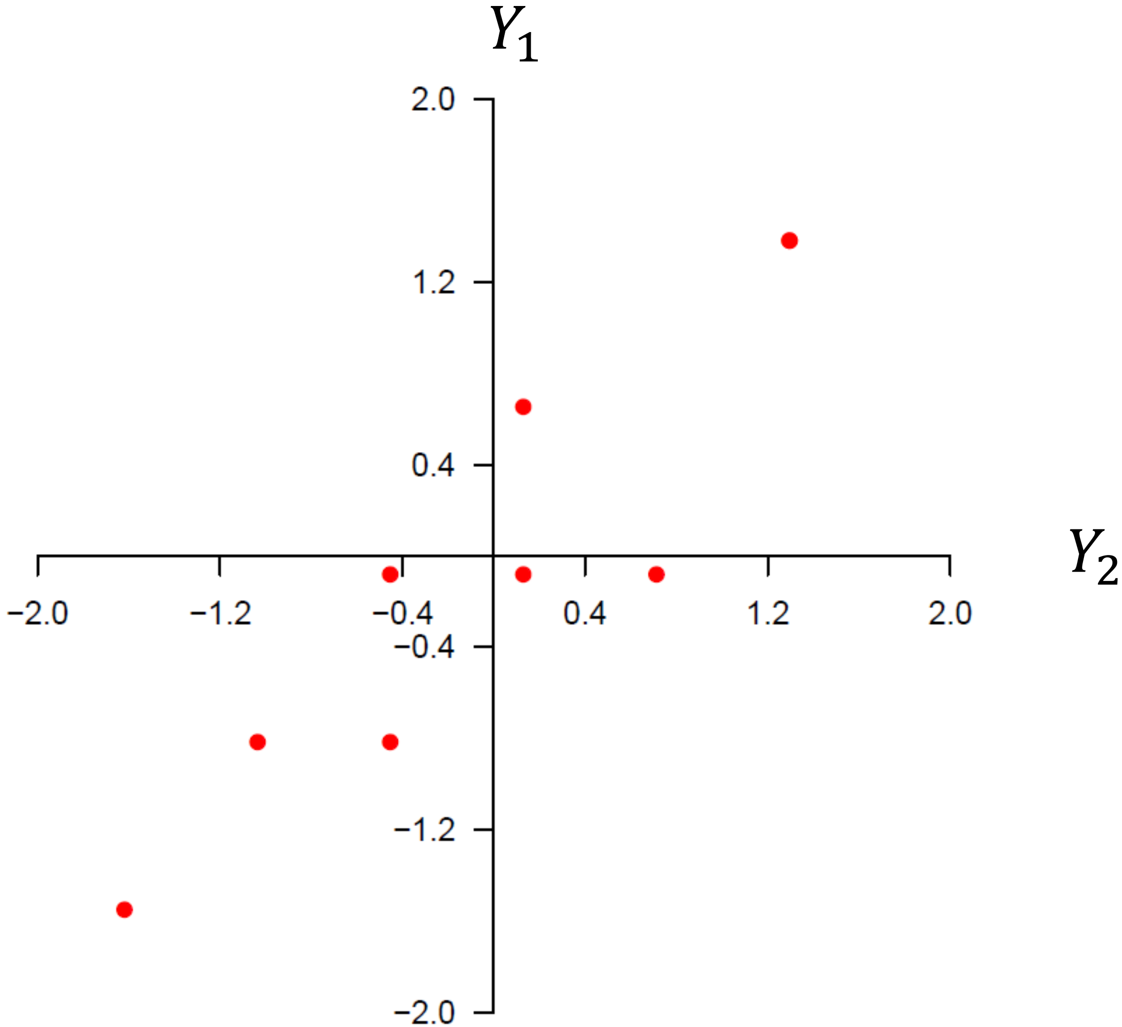


- Usually we retain the first few components that eplain **most** variance: Data reduction

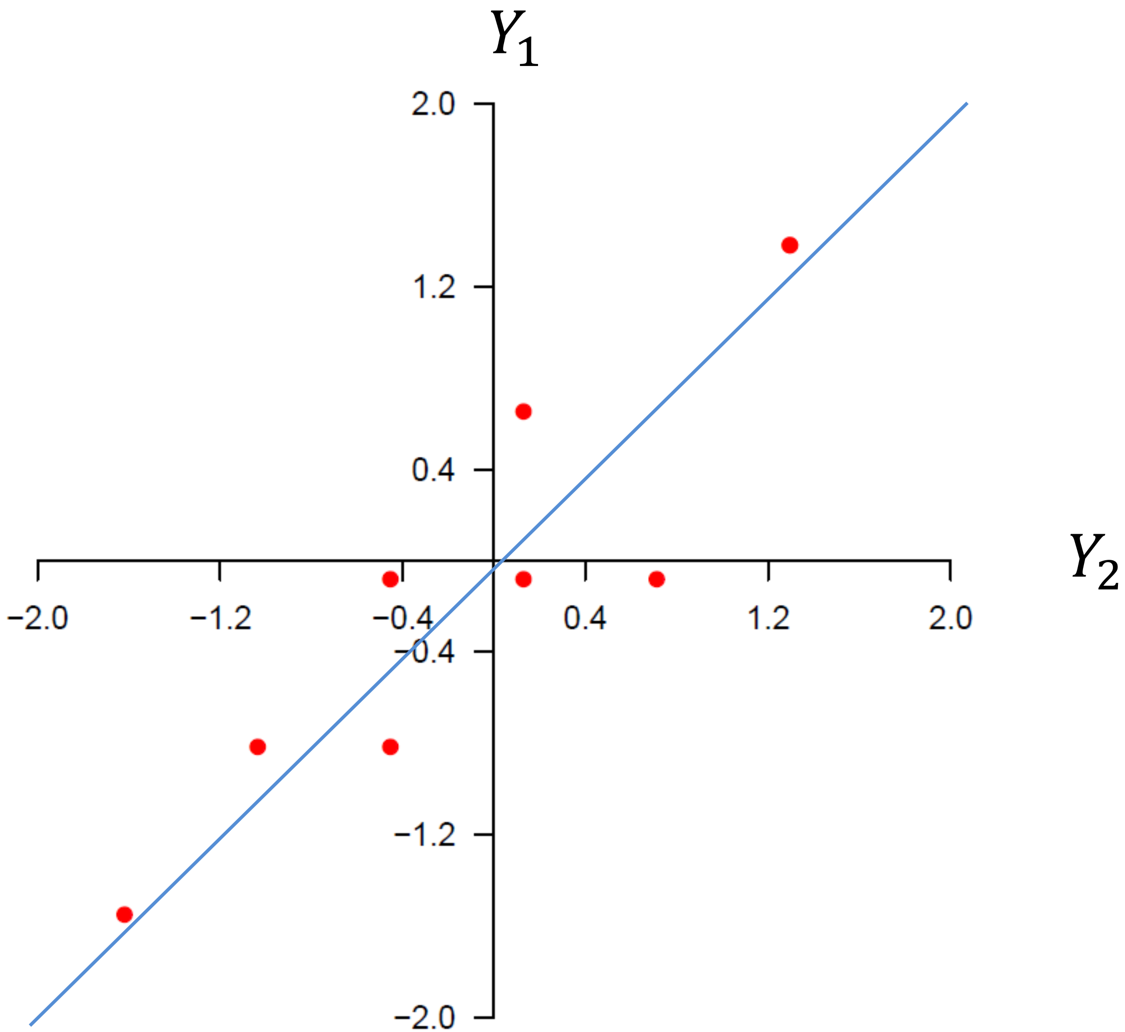
PCA - Visual example

- http://setosa.io/ev/principal-component-analysis/

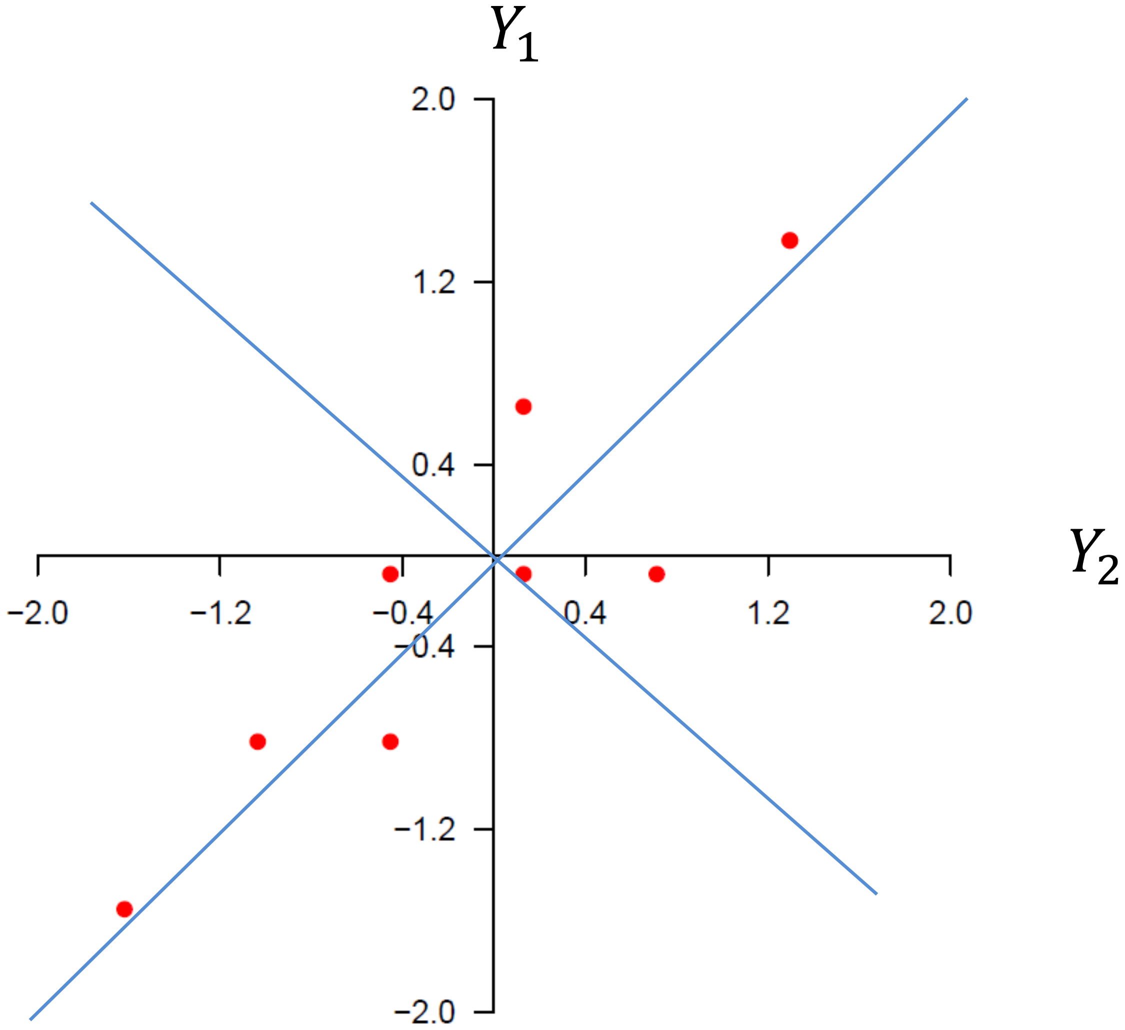
PCA



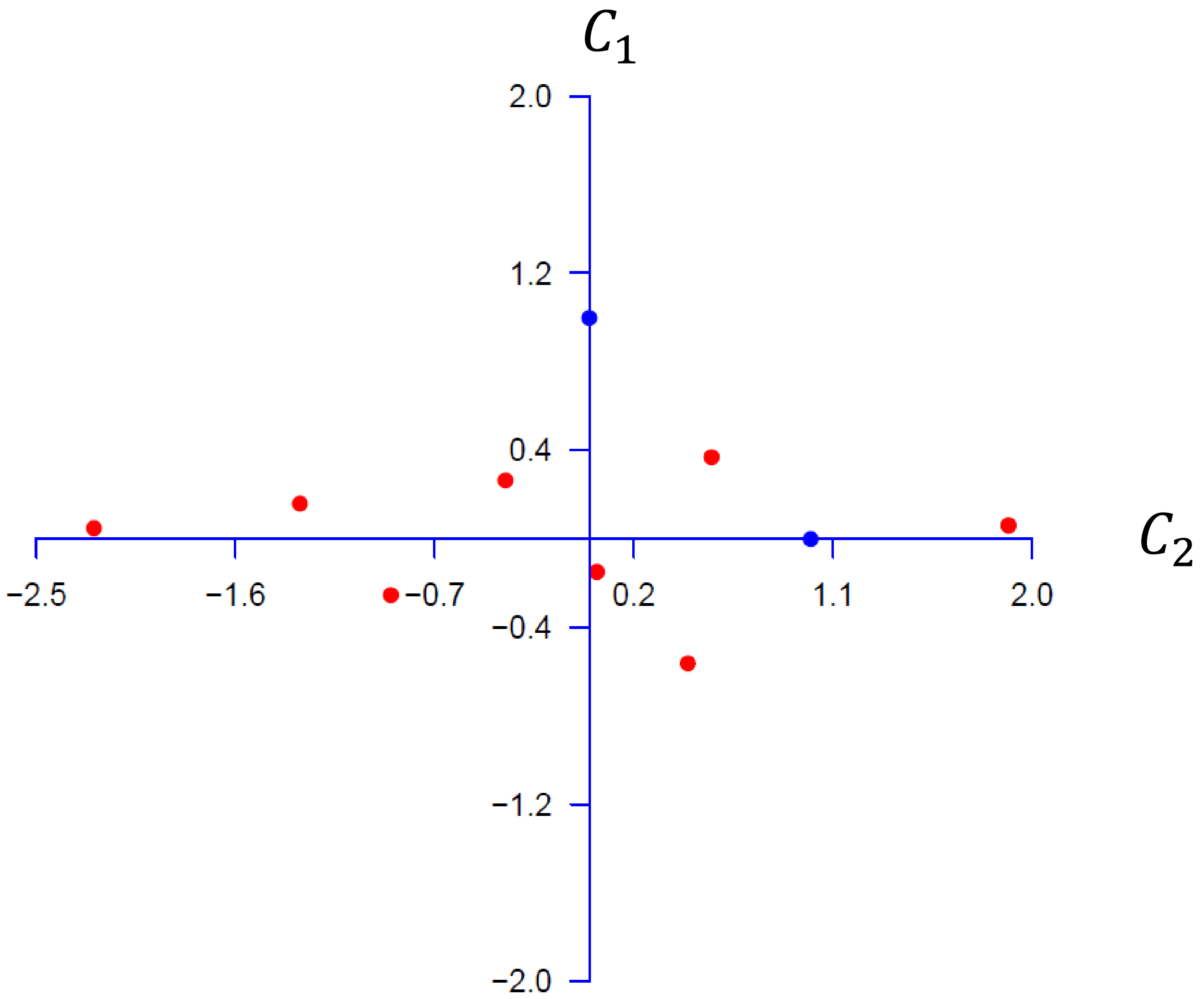
PCA



PCA

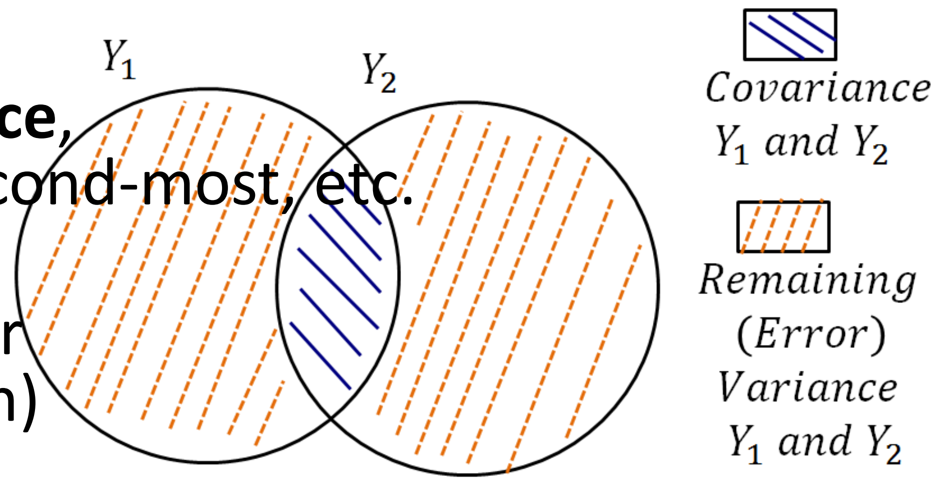


PCA

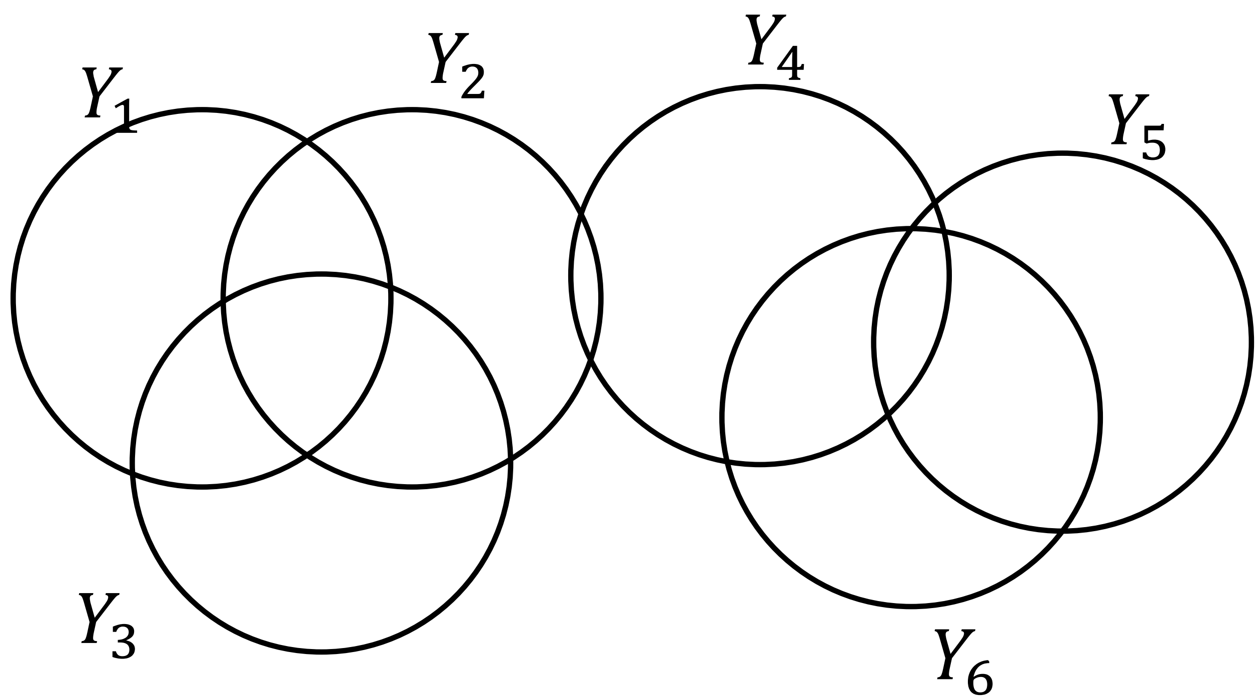


EFA : Explaining covariance

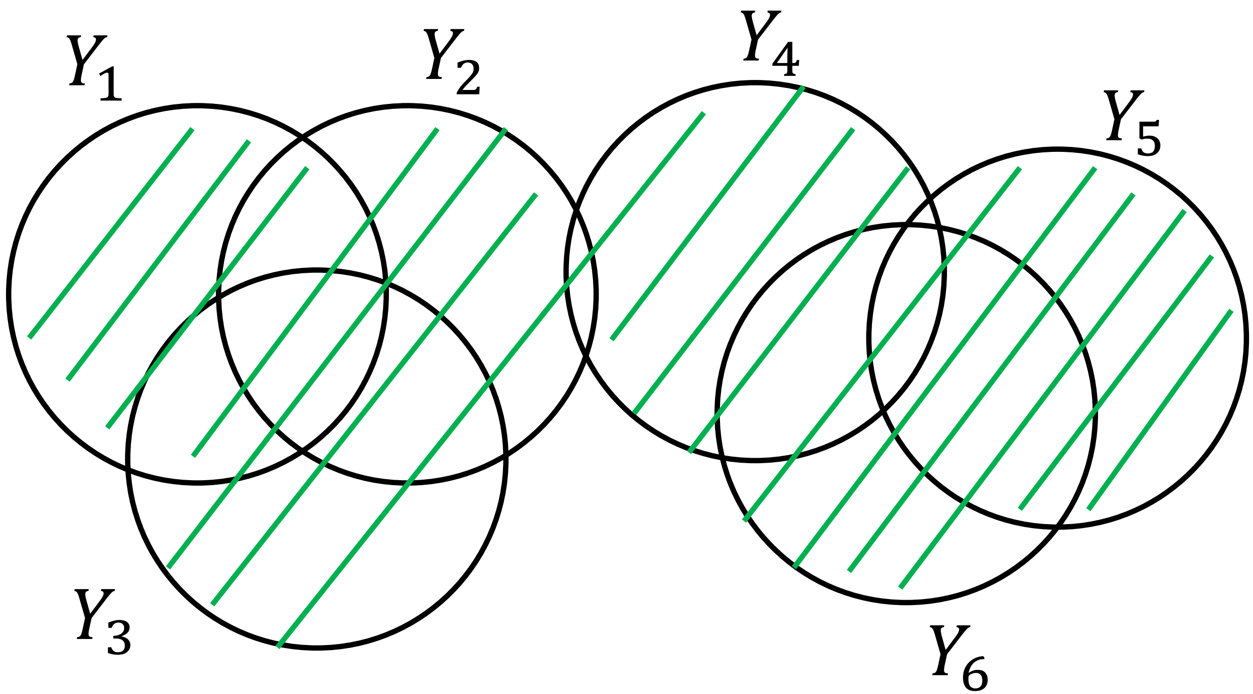
- For *n* variables, estimate **max** *n* new **factors**
  - Usually less than *n*
- I create these so that:
  - The first **factor** explains most **covariance**, the second explains second-most, etc.
  - Each **factor** is uncorrelated with other **factors** ** (see Rotation)
  - As much as possible each observed variable only relates to one **factor**



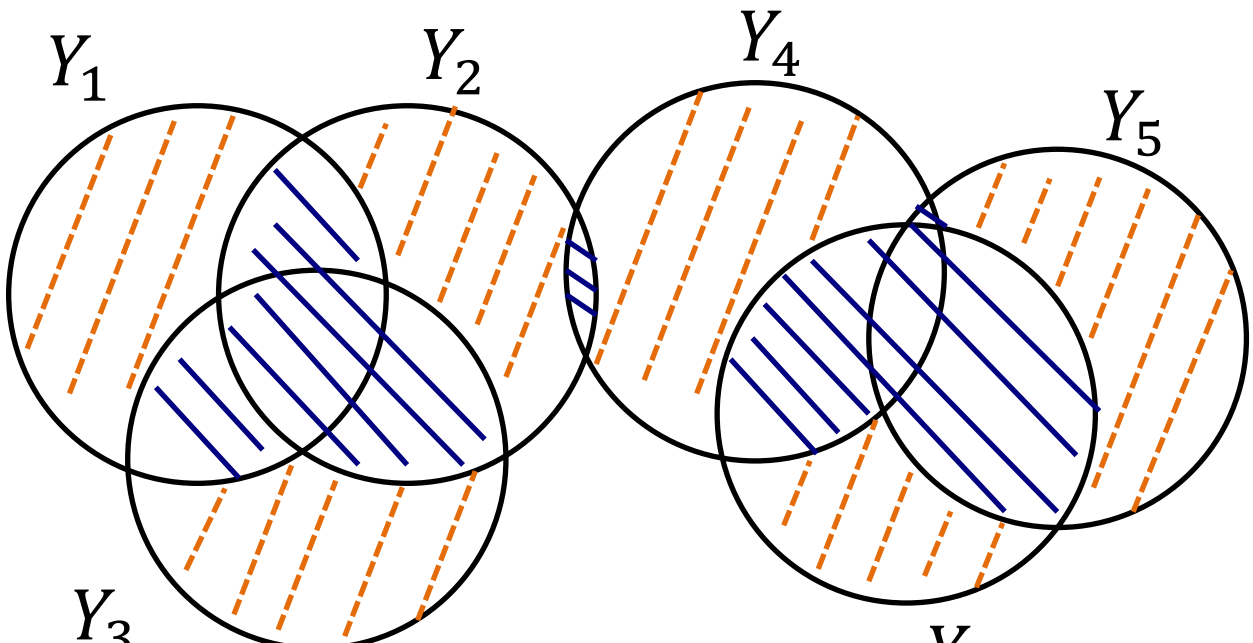
PCA and EFA



PCA: Analyse Variance



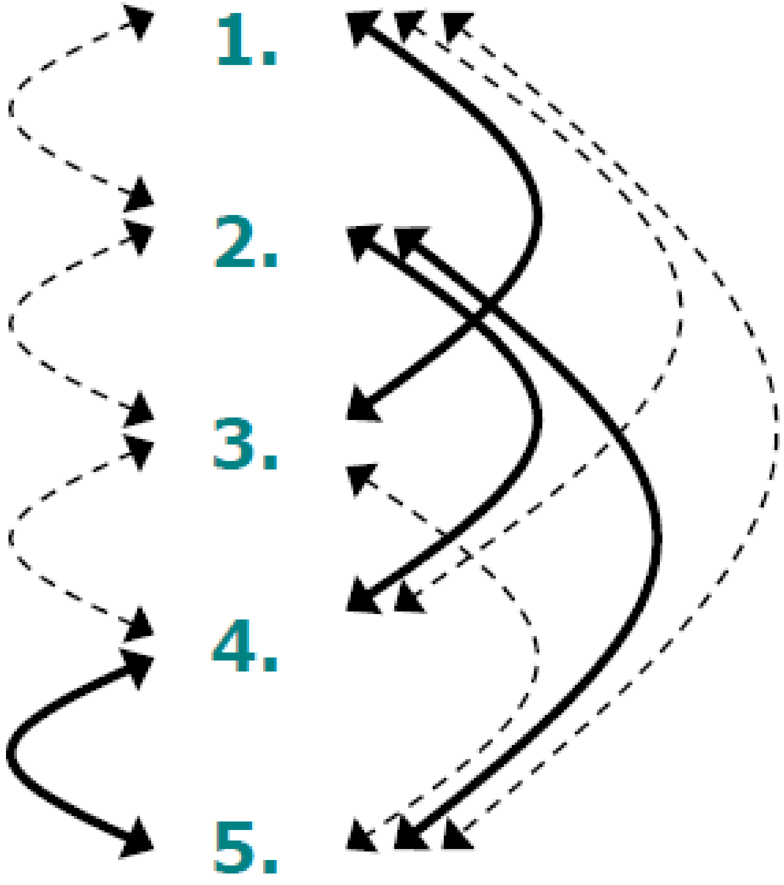
EFA: Analyse Co-variance



PCA –Example 2

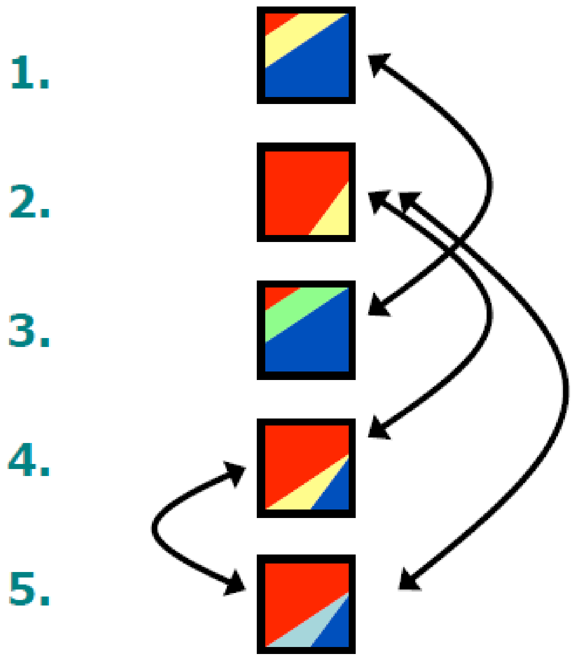
1. I always wear a seatbelt
2. I do not think before I act
3. I would never make a long journey in a sailing boat
4. I am an impulsive person
5. I would like to jump out of an airplane with a parachute

Example

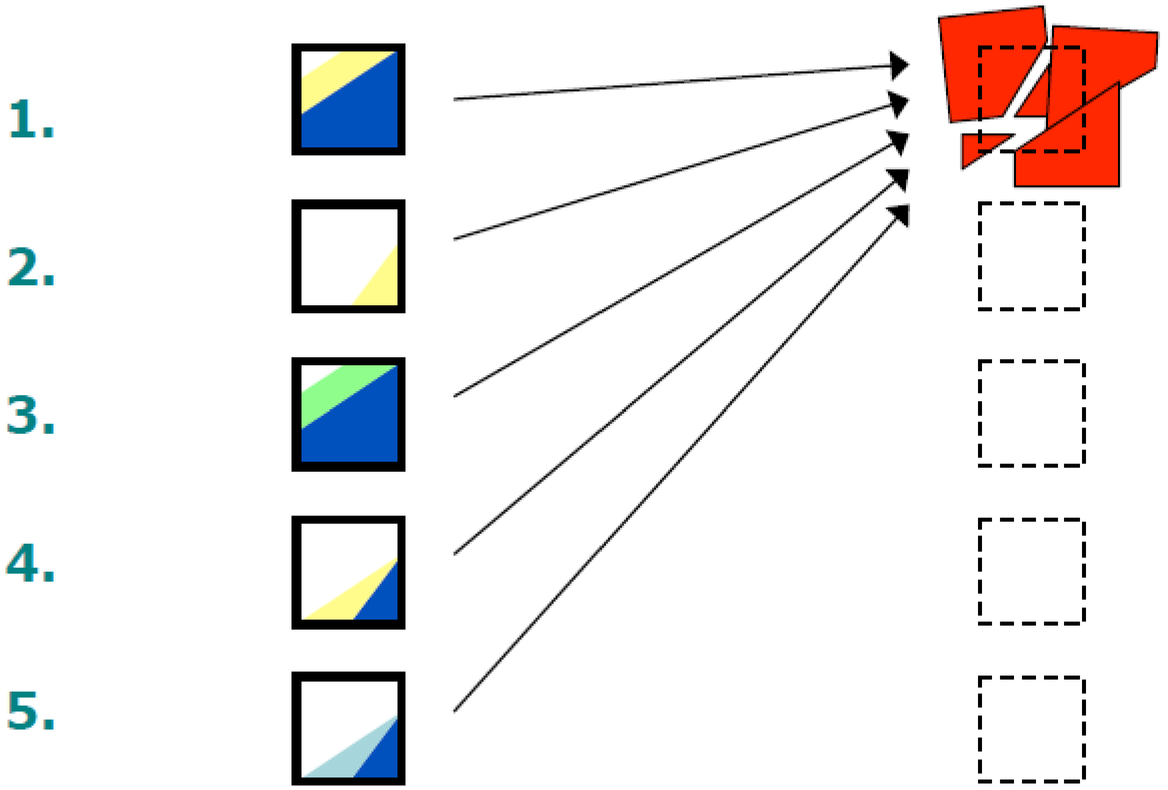


- Five questions
- We observe these correlations

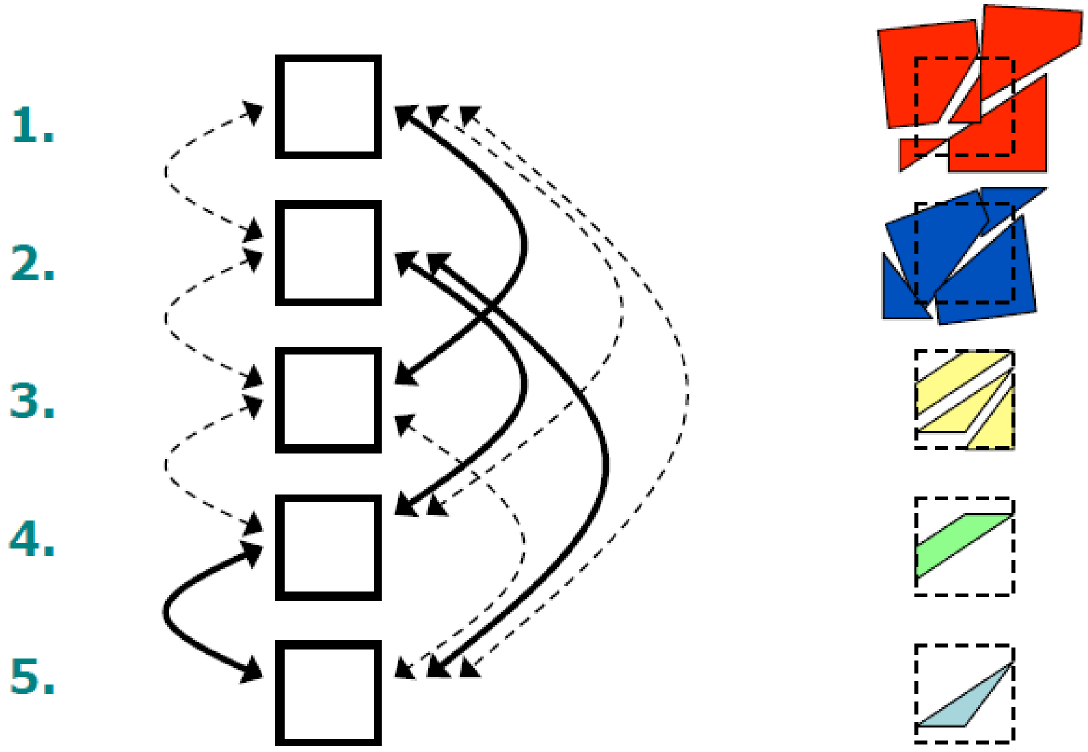
Example



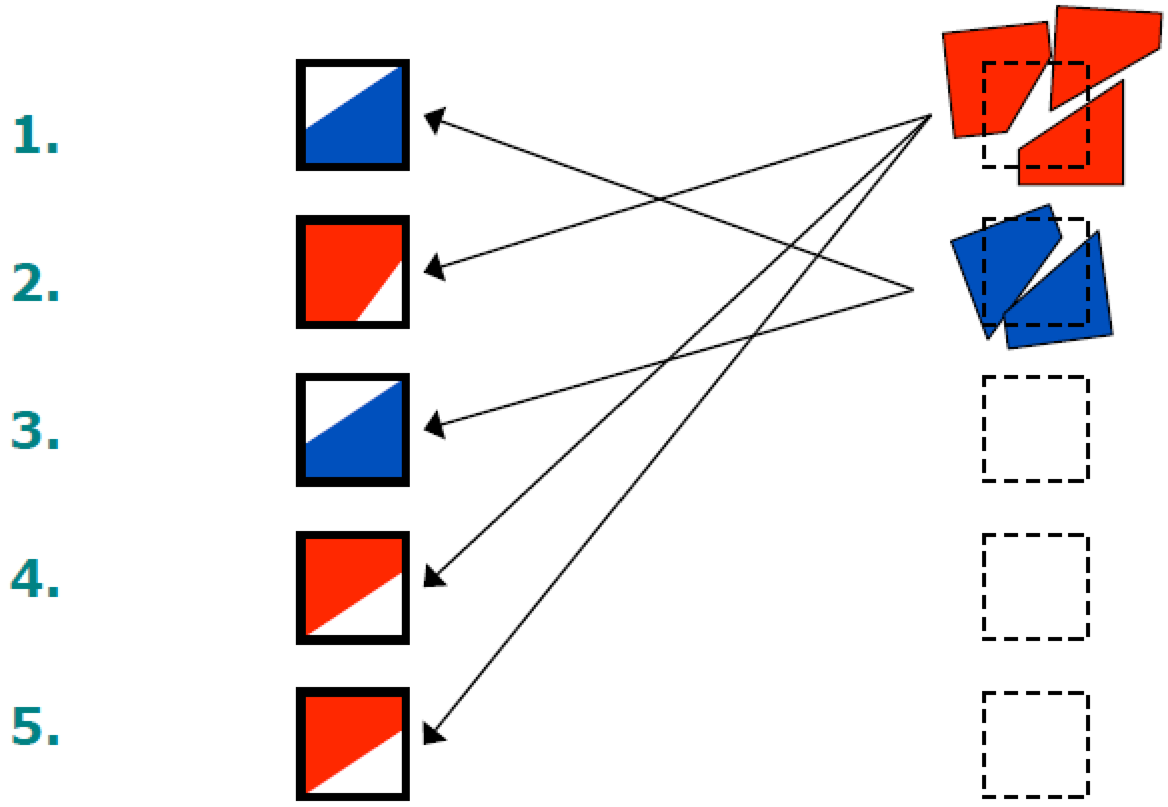
Example



Example



Example



Example

1. I always wear a seatbelt
2. I do not think before I act
3. I would never make a long journey in a sailing boat
4. I am an impulsive person
5. I would like to jump out of an airplane with a parachute

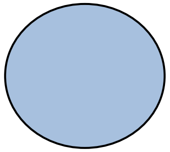
➢ Impulsive

➢ Safety-Conscious

BOX DIAGRAMS OF PCA AND FA

Quick Revision: Path Diagrams

Observed variable
(or **Indicator**)



Latent (unmeasured) variable
(or **Factor**)

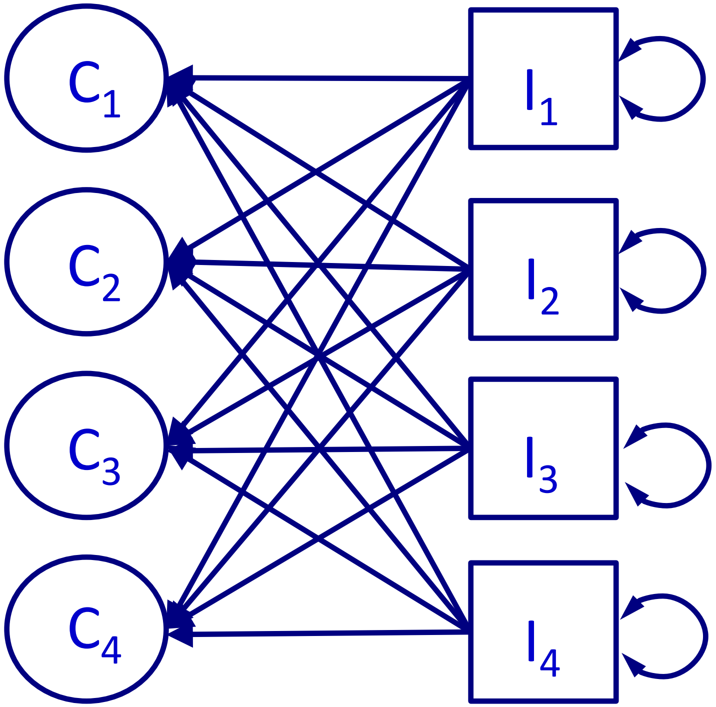
Regression
- (Theoretical) Causal effect *
- Direct Effect *

Covariance
- (no causal hypothesis)

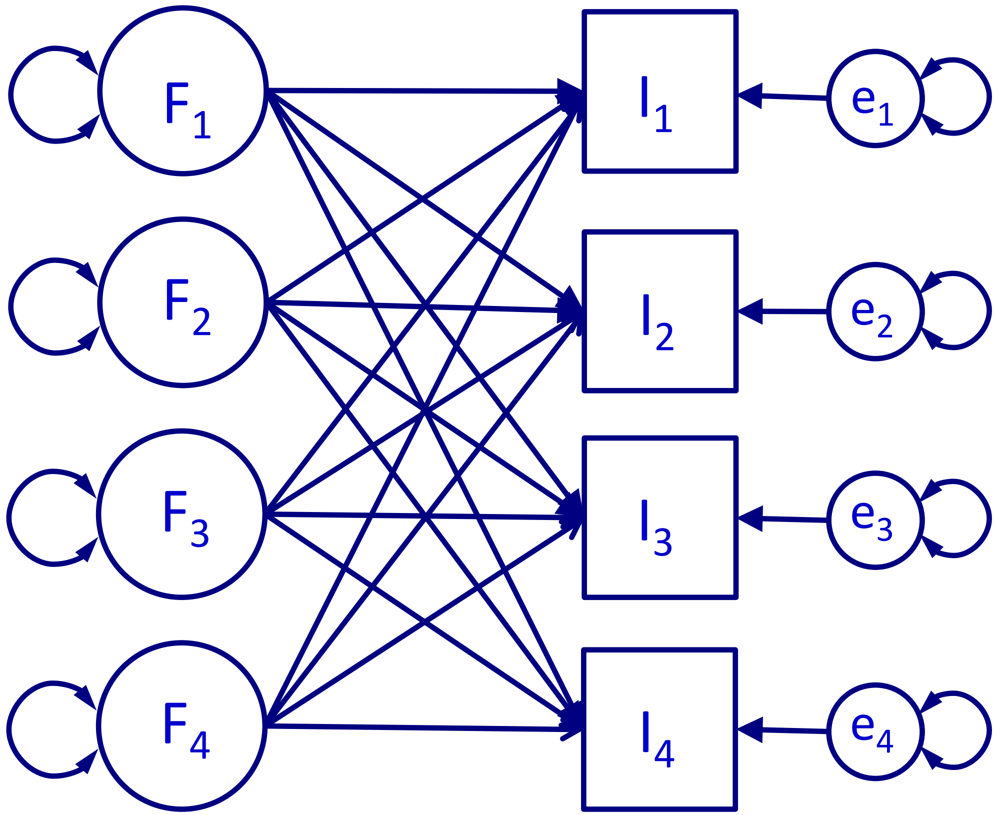
Quick Revision: Interpretation of parameters

- Direct effects, *b*, (X→ Y) as regression coefficients
  - If X goes up with 1 point, y is expected to go up with *b* points (controlling for other predictors).
  - If X goes up with 1 SD, y is expected to go up with *b* SD (controlling for other predictors).
- Factor loadings are direct effects from a factor to an indicator
- Covariances (unstandardized) and correlations (standardized)
- Variances and residual variances

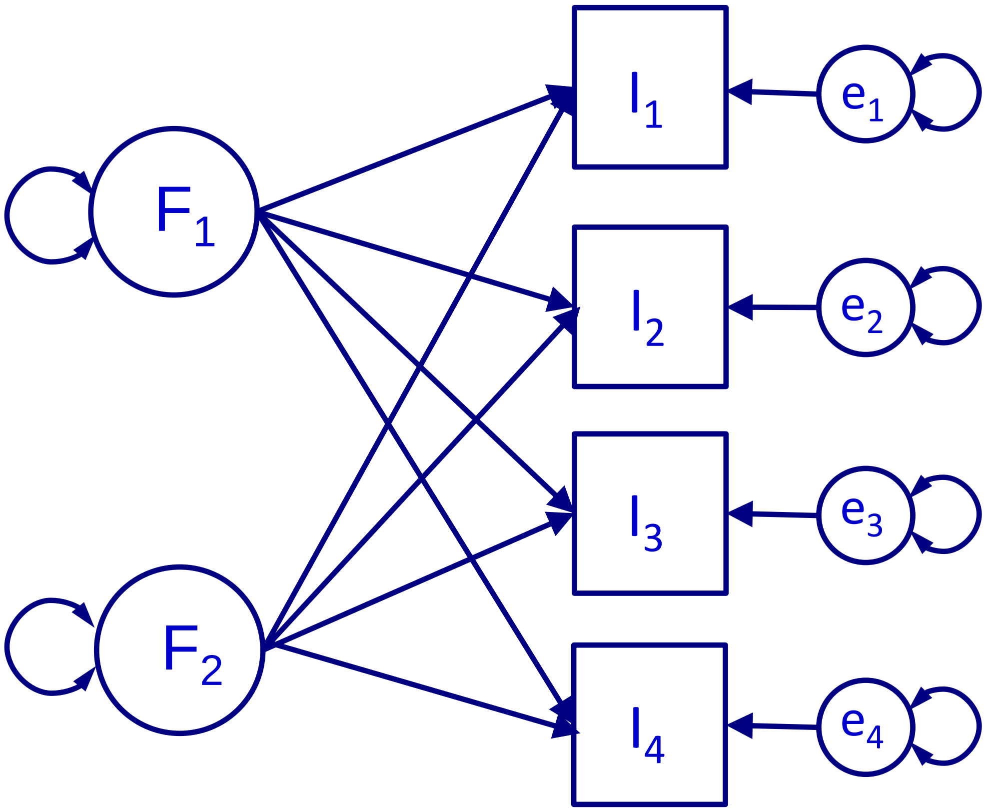
PCA



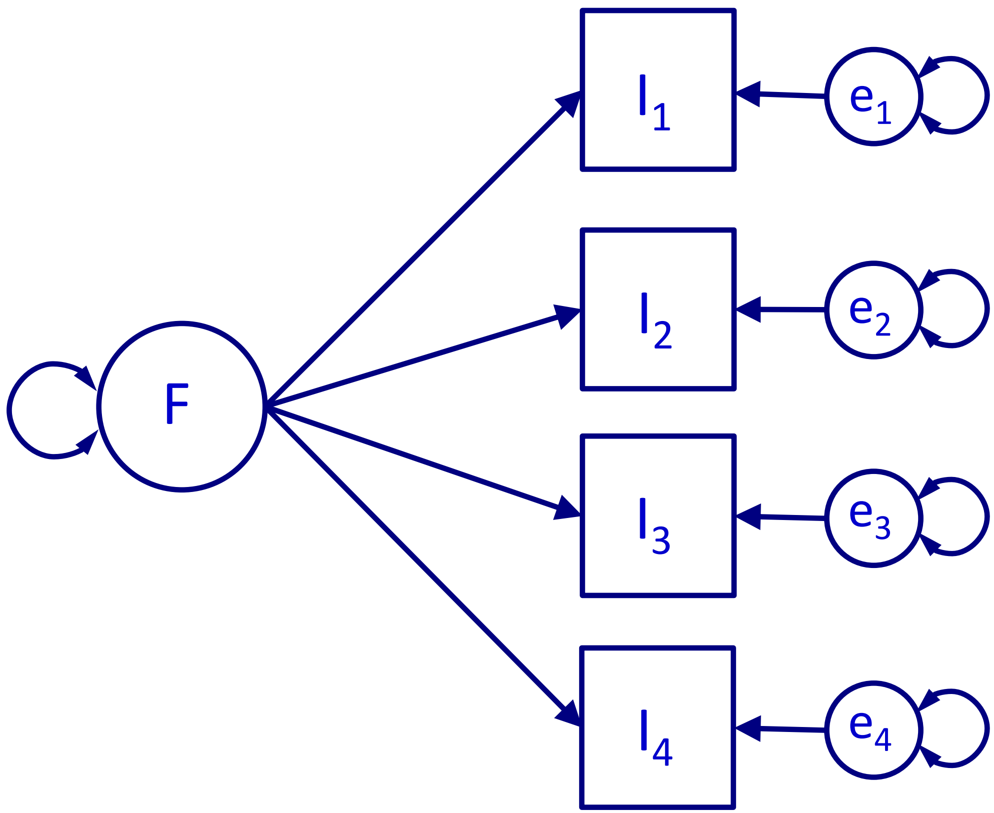
EFA



EFA



EFA

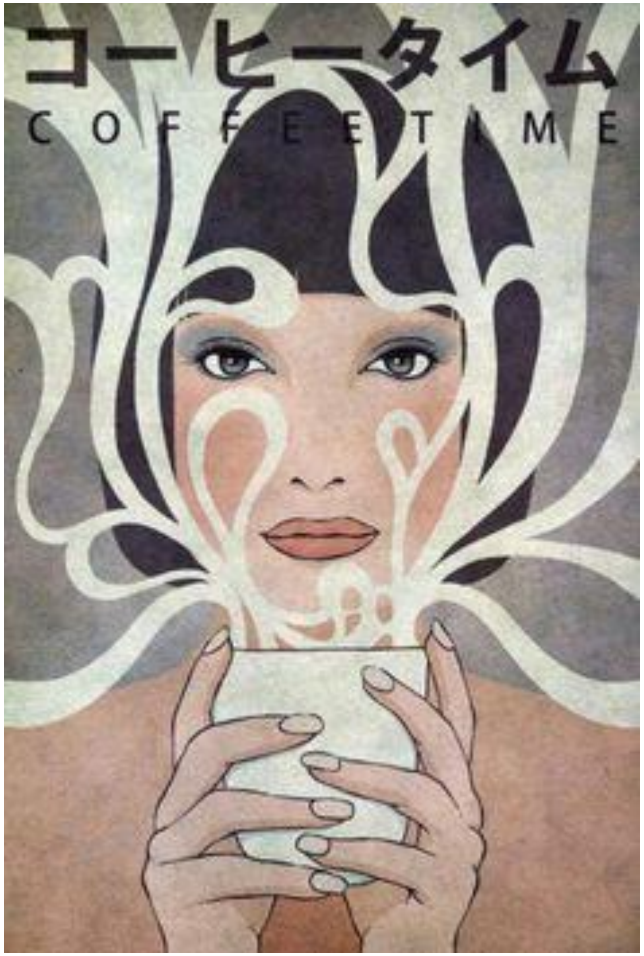


Summary PCA vs. EFA

Principal Components Analysis (PCA)	**Exploratory Factor Analysis (EFA)**
Components Summarize Variance	Factors explain Covariance
Not really a model: • Transformation of the data • No Model Fit	Model: • Some variance is interesting (covariance), some is error • Fit indices possible
Dimension Reduction	Scale construction
library(psych) principal(data, nfactors = *n*)	library(psych) fa(data, nfactors = *n*)
Extraction method: Principal Components	Extraction method: OLS, can also do "ml" (which SPSS uses)

In large samples, with large number of correlated variables, practical differences are often small

Break

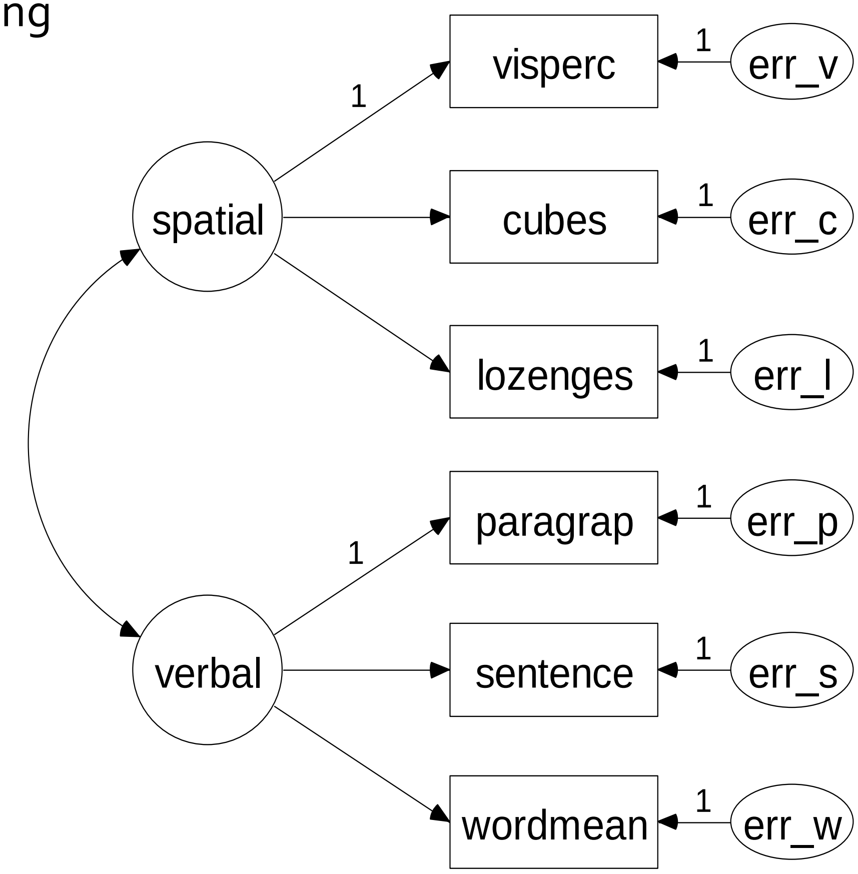


Steps to take

- Analysis requires decisions
  - 1.Extraction method
    - PCA = “Principal Components”
    - EFA = “OLS/Maximum Likelihood”
  - 2.Number of factors
  - 3.Rotation method
  - 4.(Factor scores)

Example

- Six observed variables (intelligence tests)
  - visual perception, cubes, lozenges,
  - paragraph, sentence, word meaning
- 2 factors
- Simulated data



1. Components or Factors?

```
principal(df, nfactors = 2)
```

	RC1	RC2	h2	u2	com
visperc	0.81	0.08	0.66	0.34	1.0
cubes	0.77	0.07	0.59	0.41	1.0
lozenges	0.78	0.13	0.62	0.38	1.1
paragrap	0.17	0.79	0.64	0.36	1.1
sentence	0.11	0.78	0.62	0.38	1.0
wordmean	0.07	0.74	0.56	0.44	1.0

	RC1	RC2
SS loadings	1.90	1.80
Proportion Var	0.32	0.30
Cumulative Var	0.32	**0.62**
Proportion Explained	0.51	0.49
Cumulative Proportion	0.51	1.00

1. Components or Factors?

```
fa(df, nfactors = 2)
              MR1   MR2  h2   u2     com
visperc       0.74 -0.03 0.53 0.47   1
cubes         0.60  0.01 0.36 0.64   1
lozenges      0.65  0.04 0.44 0.56   1
paragrap      0.01  0.72 0.52 0.48   1
sentence      0.01  0.65 0.42 0.58   1
wordmean      0.00  0.55 0.31 0.69   1

                          MR1  MR2
SS loadings               1.33 1.25
Proportion Var            0.22 0.21
Cumulative Var            0.22 0.43
Proportion Explained      0.52 0.48
Cumulative Proportion     0.52 1.00

With factor correlations of
    MR1  MR2
MR1 1.00 0.38
MR2 0.38 1.00
```

2. EFA: How Many Factors?

- If (proto) *theory* predicts *k* factors, try *k* factors
- Parallel analysis
- Guttman-Kaiser criterion (Eigenvalue ≥ 1) best with small number of reliable variables
- Scree plot best with large number of unreliable variables
- Pick the solution that makes most interpretative sense

2. EFA: How Many Factors?

- Guttman-Kaiser criterion (Eigenvalue ≥1) best with small number of reliable variables
- Eigenvalues relate to how much of the total variance each component/factor accounts for
  - First explains most, second explains second-most, etc.
- $\frac{Eigenvalue}{Total\ Number\ observed\ items} = Variance\ explained\ by\ factor$

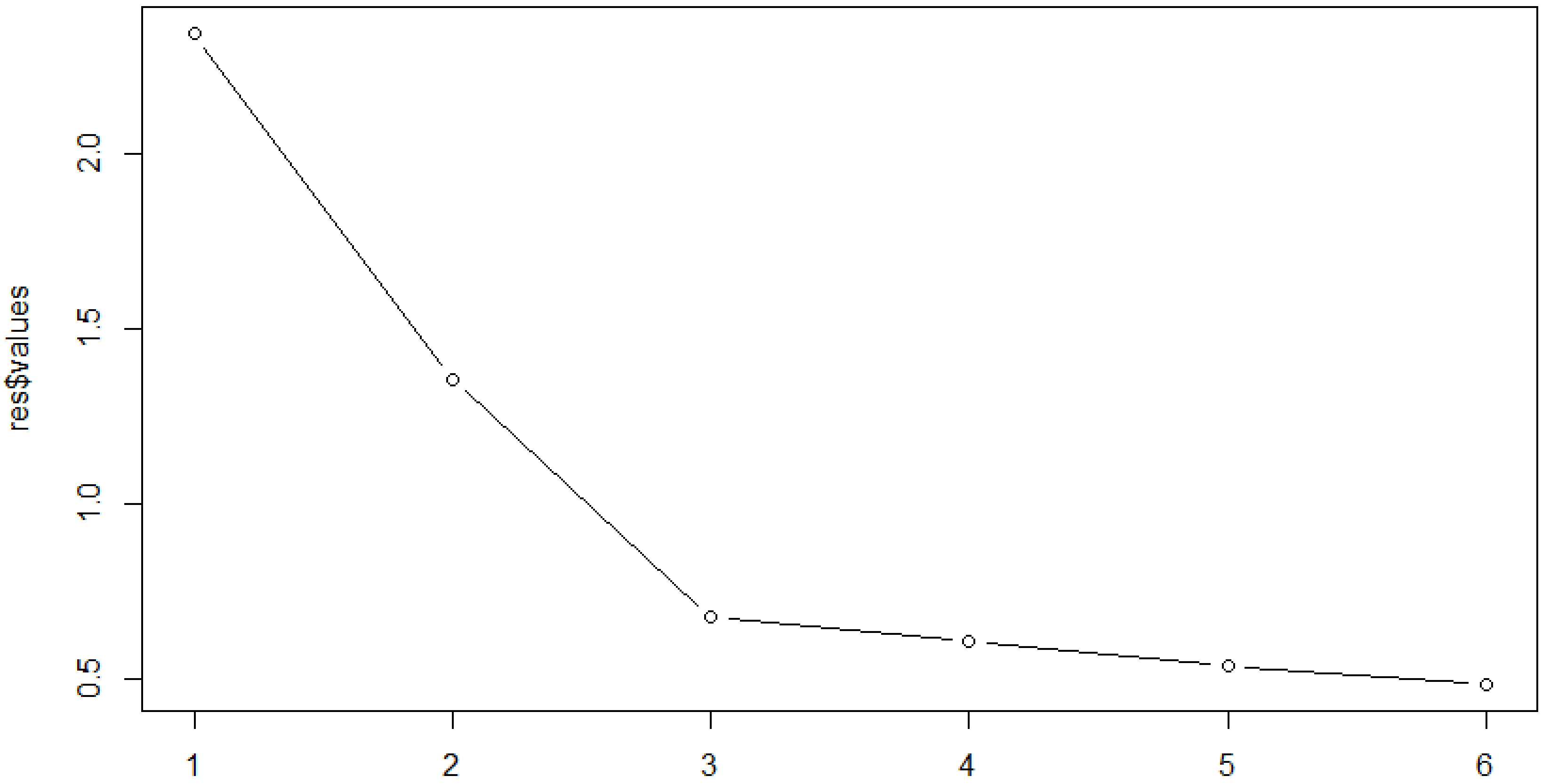
```
> res <- principal(df, nfactors = 5)
> res$values
[1] 2.34 1.35 0.67 0.60 0.53 0.48
> res$values > 1
[1] TRUE TRUE FALSE FALSE FALSE FALSE
```

```
>
```

2. EFA: How Many Factors?

- Scree plot best with large number of unreliable variables
- Pick the number of factors “above the elbow”

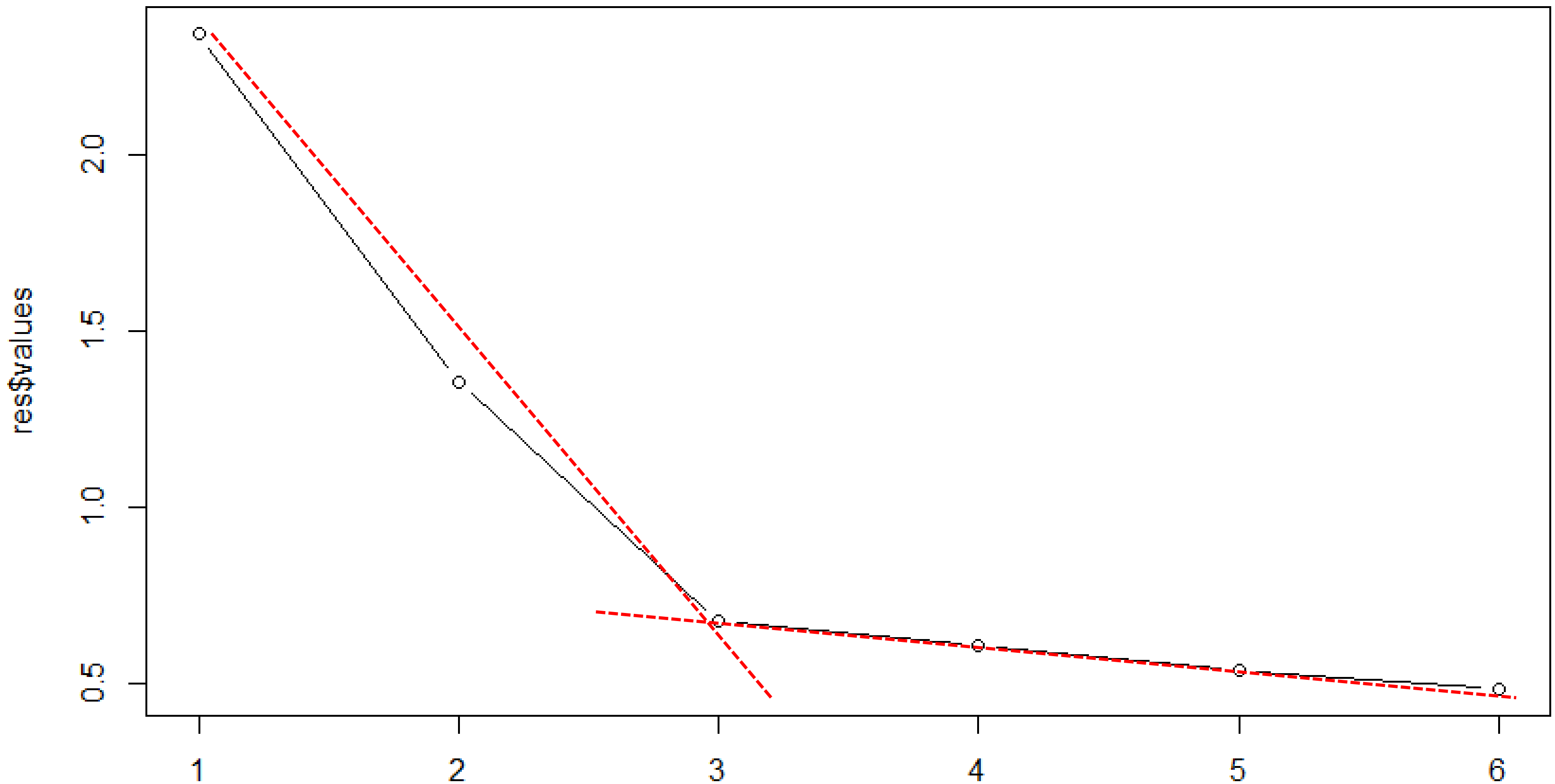
```
> plot(1:6, res$values, type = "b")
```



2. EFA: How Many Factors?

- Scree plot best with large number of unreliable variables
- Pick the number of factors “above the elbow”

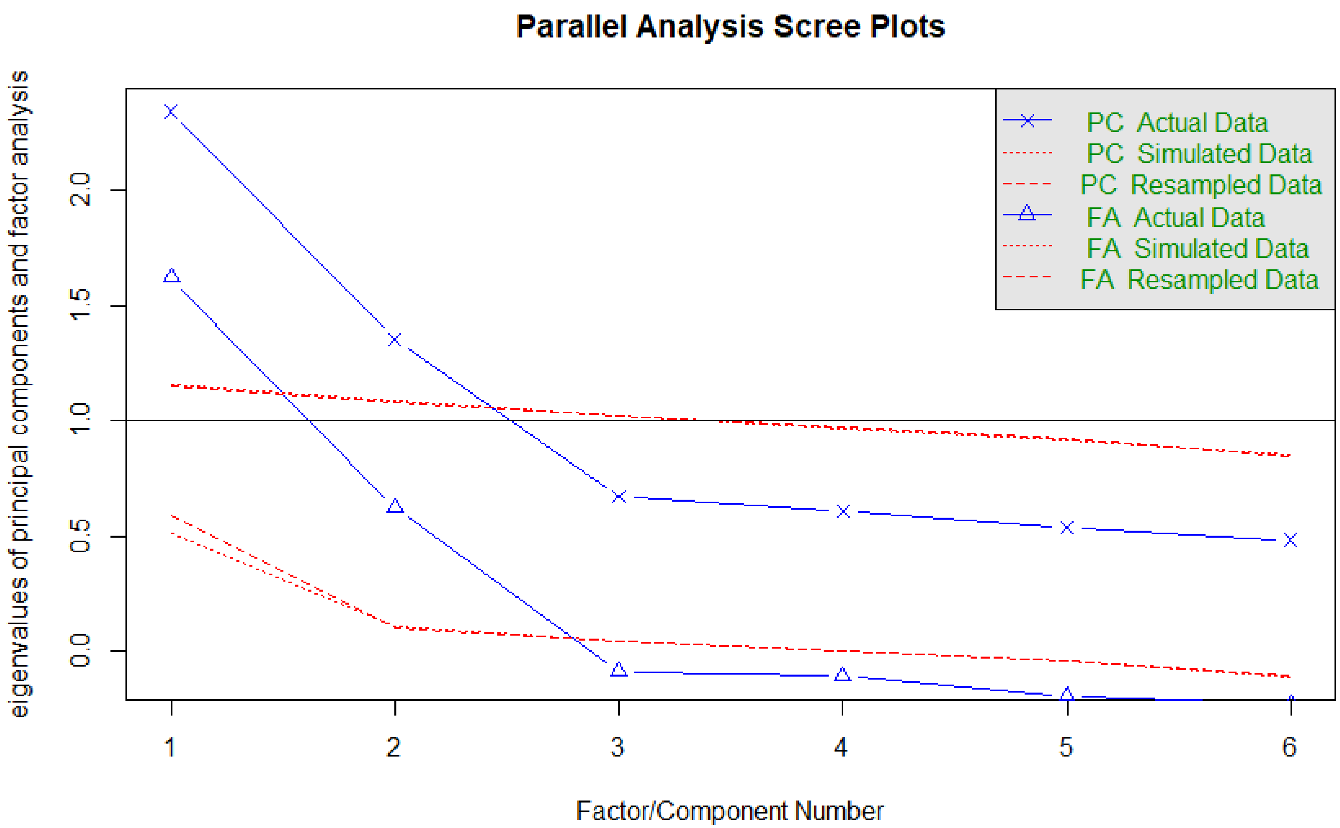
```
> plot(1:6, res$values, type = "b")
```



2. EFA: How Many Factors?

- Parallel Analysis (Horn, 1965)

```
> fa.parallel(df)
Parallel analysis suggests that
the number of factors = 2 and
the number of components = 2
```



2. EFA: How Many Factors?

- Pick the solution that makes most sense wrt interpretation
- If *theory* predicts *k* factors, try *k* factors
- Try out different numbers of factor solutions
  - Sometimes different rules-of-thumb give different solutions
- Look at the factor loadings
- Pick the solution which gives you meaningful factors/components

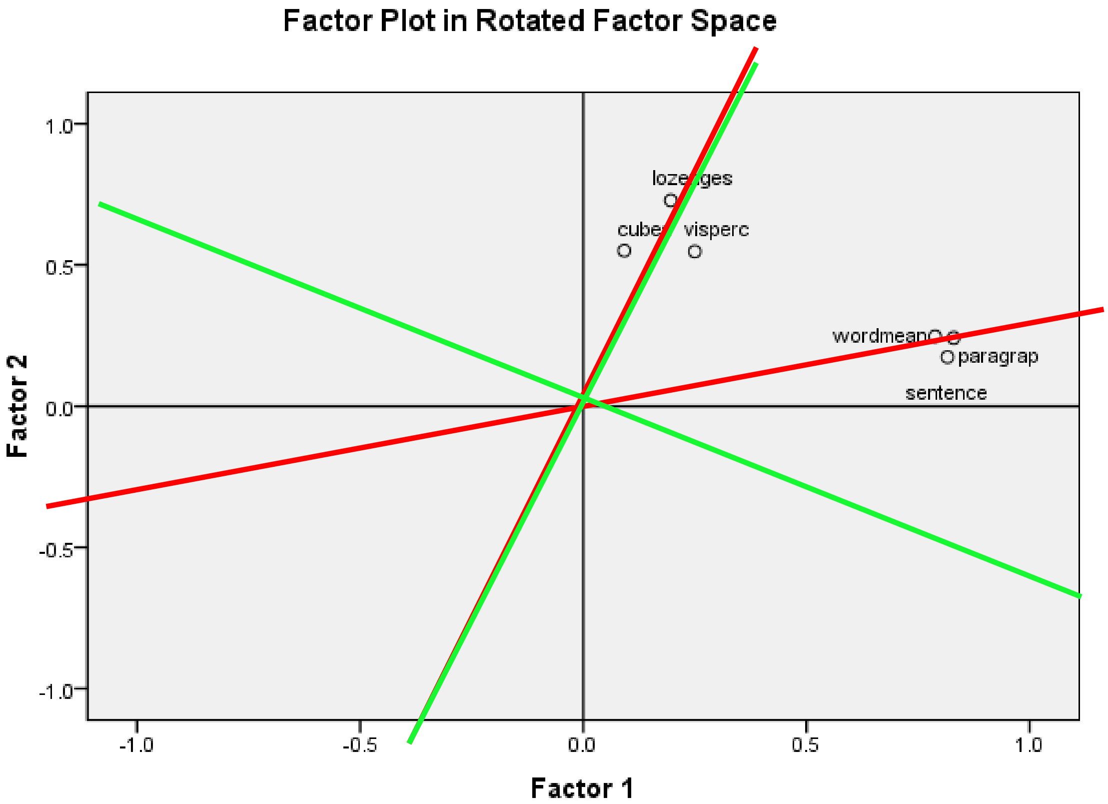
3. Factor Rotation

Orthogonal rotation:

Factors rotate, but ‘angle’ is always 90 degrees. Factors are not correlated!

Oblique rotation: factors rotate to minimize distance between items and factor (oblique)

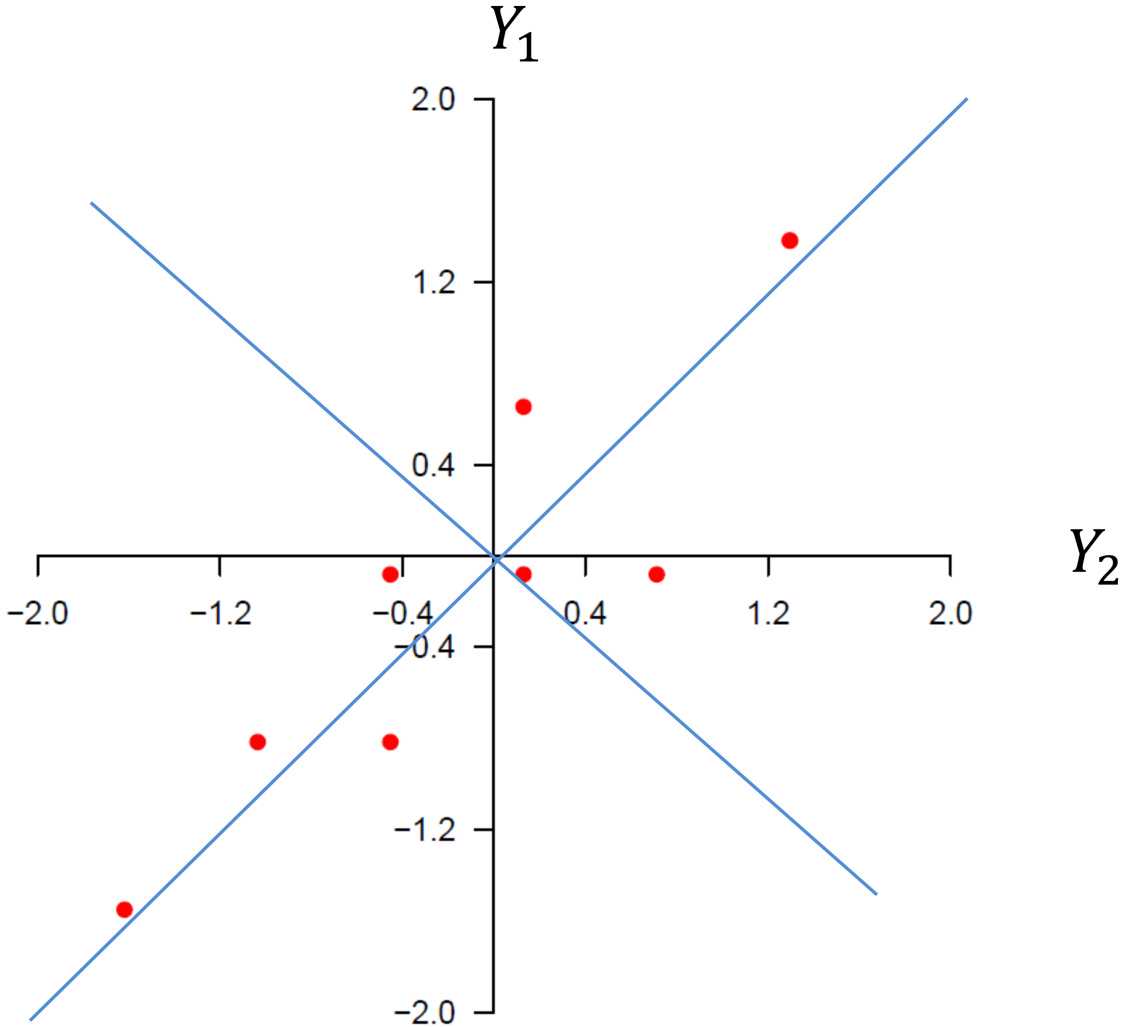
Factors are correlated!



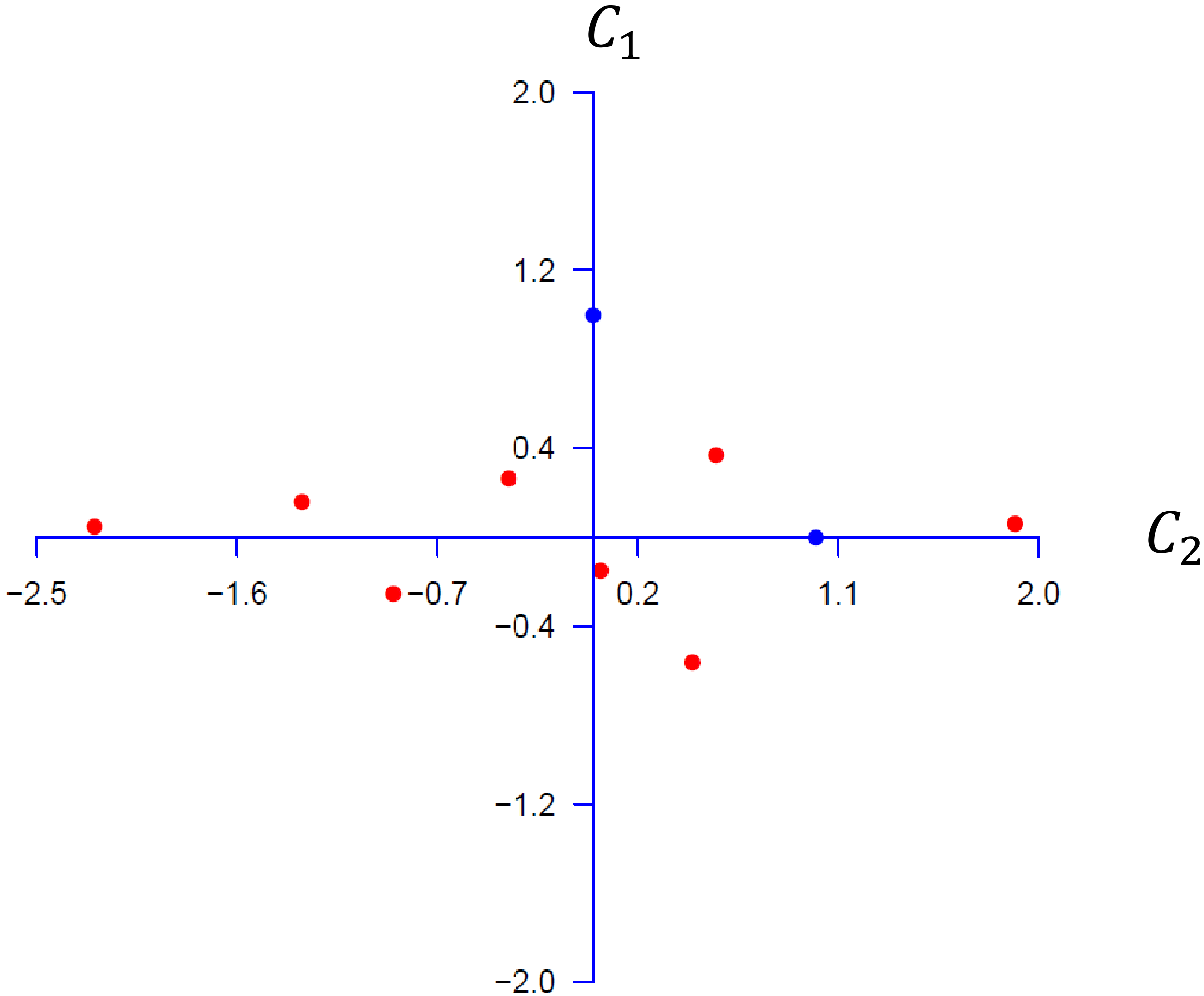
Reading question 5: what is the purpose of factor rotation?

The procedure of rotating the factor axes makes sure items load as much on only one factor as possible. There are two methods: Orthogonal rotation, in which two latent factors are not allowed to correlate (i.e. the axes describe a 90 degree angle), and oblique (oblimin or promax) rotation, in which the factors are allowed to correlate.

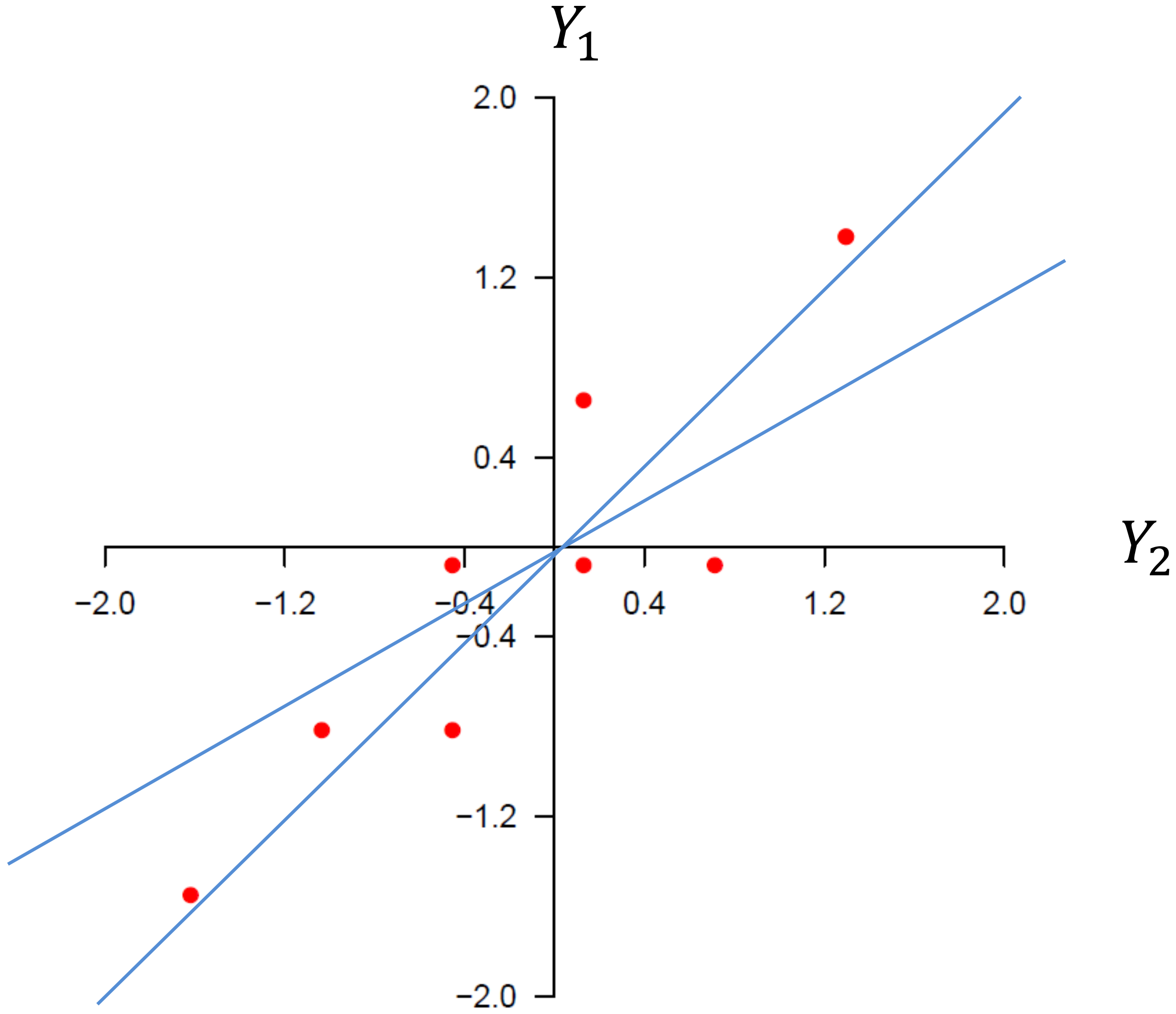
Orthogonal Rotation 1



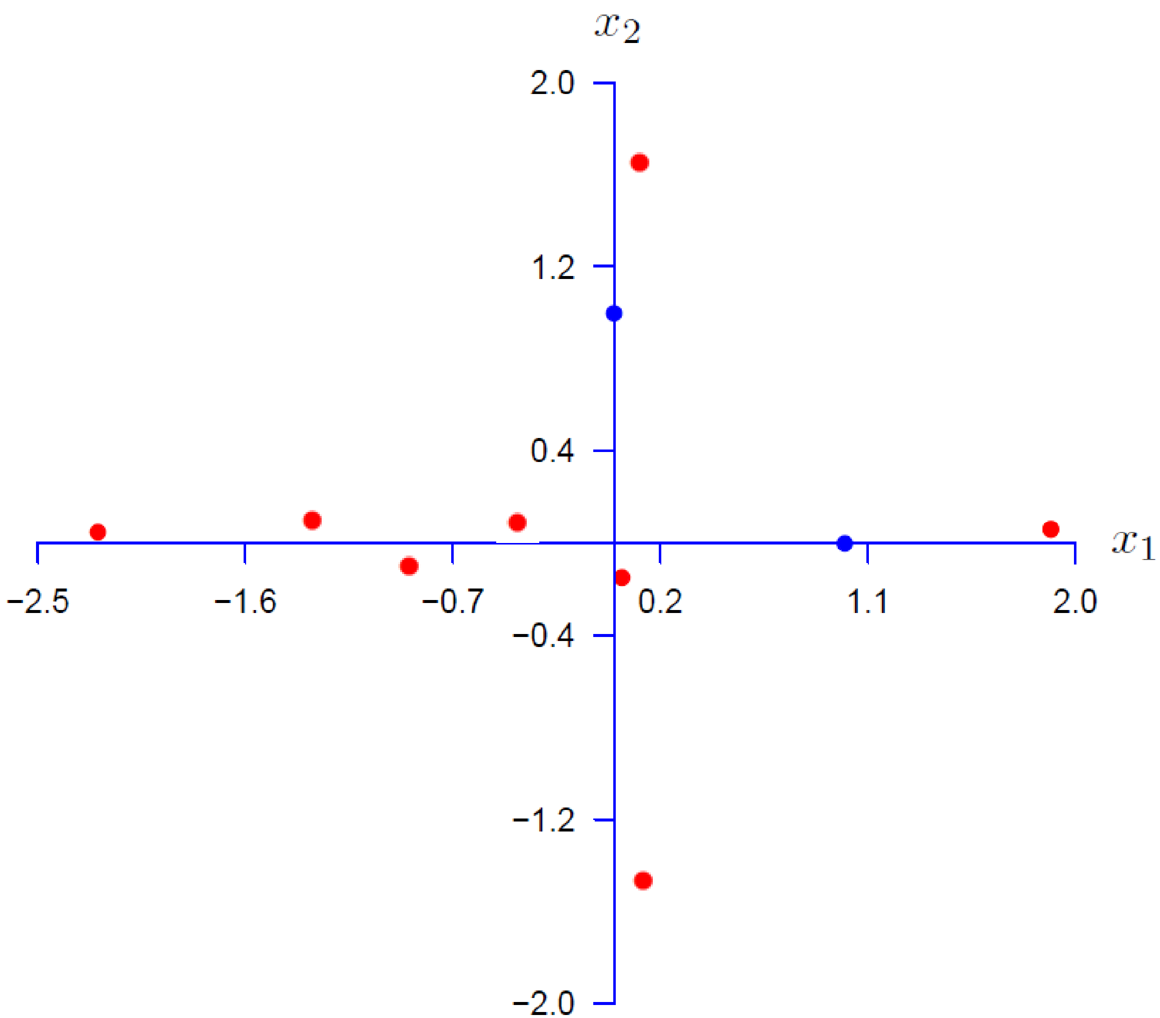
Orthogonal Rotation 2



Oblique Rotation 1



Oblique rotation 2



3. Factor rotation

- Orthogonal: uncorrelated factors
  - Varimax
  - Simple
  - Interpretation may be easier
  - Factor loadings show up in the **Factor Matrix**
- Oblique: correlated factors
  - Promax, Oblimin
  - More realistic
  - Easier to get items to load on only one factor
  - Factor loadings show up in the **Pattern Matrix**

Varimax vs promax

Rotated Factor Matrix[a]

	Factor	
	1	2
item1	.097	.700
item2	.097	.700
item3	.097	.700
item4	.700	.097
item5	.700	.097
item6	.700	.097

Extraction Method: Maximum Likelihood.
Rotation Method: Varimax with Kaiser Nor

a. Rotation converged in 3 iterations.

Pattern Matrix[a]

	Factor	
	1	2
item1	.000	.707
item2	.000	.707
item3	.000	.707
item4	.707	.000
item5	.707	.000
item6	.707	.000

Extraction Method: Maximum Likelihood.
Rotation Method: Promax with Kaiser Nc

a. Rotation converged in 3 iterations.

4. Factor scores

- Useful to save the factor scores:

```
fa(df, nfactors = 2,
          scores = “regression”)
```

  – Multiplication of item scores:

    sum(individual itemscore * factor loading)

    Three ways: “regression”, “Anderson” or “Bartlett”

    - Small difference

  - Use these factors as observed variables in your analysis

    - Ignores measurement error

  - Not needed if you continue with SEM!

4. Factor scores

```
> res <- fa(df, nfactors = 2, scores = "Bartlett")
> head(res$scores)

MR1 MR2
[1,] -0.5609899 -0.03047855
[2,] 0.5132644 1.29435355
[3,] 0.2444246 -1.19983489
[4,] -0.8724184 1.30067344
[5,] -0.1687548 1.02015701
[6,] 1.1181263 -0.51572749
```

EFA: Optimal decisions and defaults

Decision about	Optimal	Default
Extraction	Theory: • Factors Data reduction: • Components	fa(): • Factors principal(): • Components
# Factors	Theory Parallel analysis	
Rotation	Oblique	fa(): • oblimin principal(): • Varimax
Factor scores	Bartlett	fa(): • scores = “regression” principal(): • method = “regression”

Example EFA

- Allen & Mayers (1996) three part model of commitment

- Affective commitment
  - 5 items

I would be very happy to spend the rest of my career with this organization.
I really feel as if this organization's problems are my own.

- Continuance commitment
  - 5 items

Too much of my life would be disrupted if I decided I wanted to leave my organization now.
I feel that I have too few options to consider leaving this organization.

- Normative commitment
  - 4 items

I would feel guilty if I left my organization now.
This organization deserves my loyalty.

Example

Think about (and report)

– Extraction method
– Number of factors
– Rotation method

Number of factors

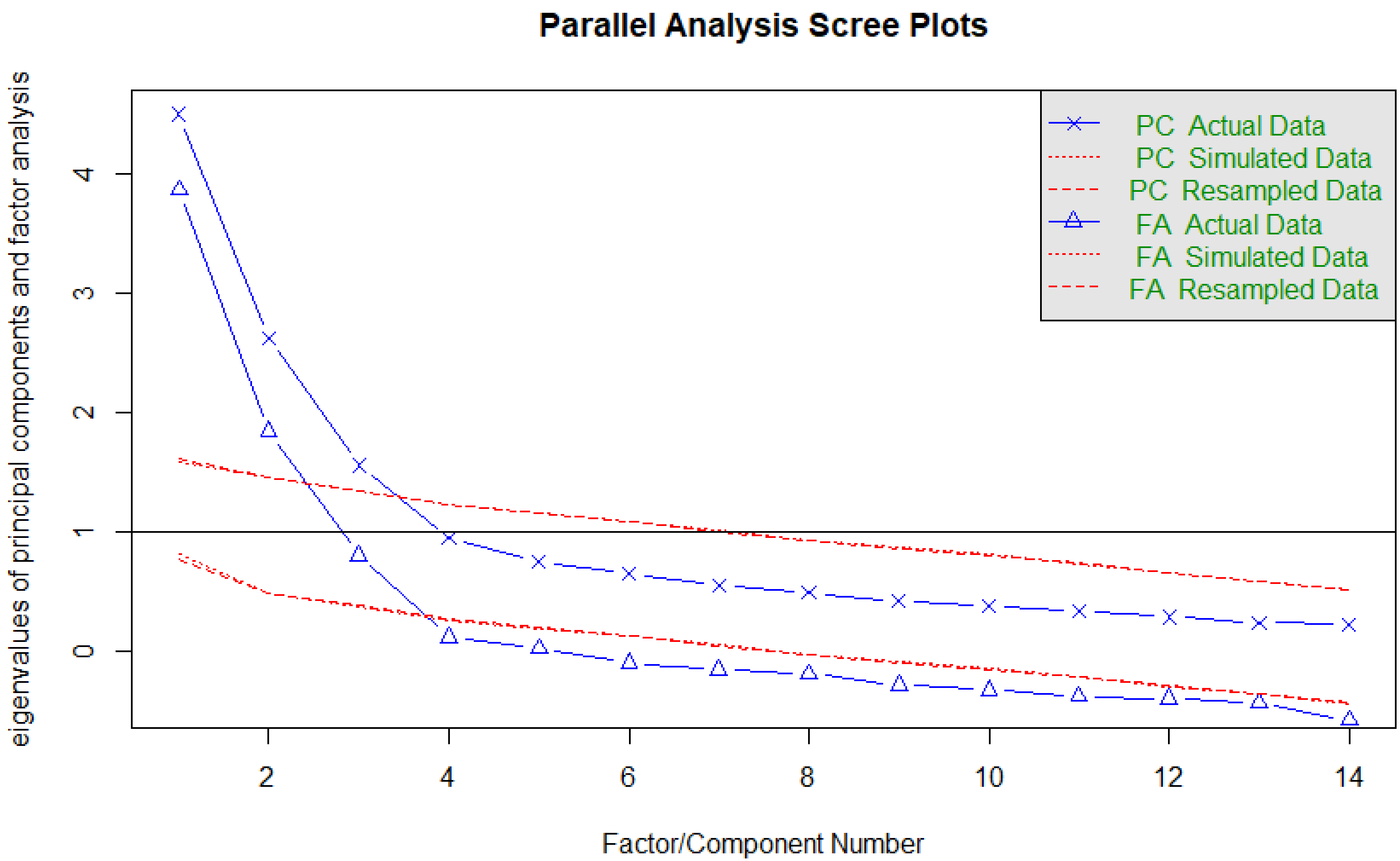
```
> res <- fa(df, nfactors = 6)
> res
Factor Analysis using method =  minres
Call: fa(r = df, nfactors = 6)
Standardized loadings (pattern matrix) based upon correlation matrix
…

                       MR1  MR2  MR5  MR4  MR3  MR6
SS loadings           2.87 1.64 1.37 1.26 1.28 0.65
Proportion Var        0.20 0.12 0.10 0.09 0.09 0.05
Cumulative Var        0.20 0.32 0.42 0.51 0.60 0.65
Proportion Explained  0.32 0.18 0.15 0.14 0.14 0.07
Cumulative Proportion 0.32 0.50 0.65 0.79 0.93 1.00

> res$values
 [1]  4.21663707  2.23703890  1.23475959  0.49065841  0.44348134
      0.44266071  0.11342196  0.05770514  0.03241050
[10]  0.01736304 -0.01664042 -0.03138805 -0.06975776 -0.10589982
```

Number of factors

```
> fa.parallel(df)
Parallel analysis suggests that the number of factors = 3 and the number of components = 3
```



Rotated factor loadings

```
> res <- fa(df, nfactors = 3)
> res
Factor Analysis using method =  minres
Call: fa(r = df, nfactors = 3)
Standardized loadings (pattern matrix) based upon correlation matrix
     MR1   MR3   MR2   h2   u2 com
A1  0.51  0.19  0.03 0.36 0.64 1.3
A2  0.77  0.06 -0.21 0.67 0.33 1.2
A3  0.86 -0.01  0.01 0.73 0.27 1.0
A4  0.72  0.11 -0.08 0.59 0.41 1.1
A5  0.85 -0.09  0.17 0.69 0.31 1.1
C1  0.06  0.31  0.60 0.56 0.44 1.5
C2  0.08  0.12  0.52 0.32 0.68 1.2
C3 -0.17 -0.06  0.72 0.54 0.46 1.1
C4  0.19 -0.02  0.32 0.13 0.87 1.7
C5  0.08 -0.04  0.65 0.42 0.58 1.0
N1  0.16  0.65  0.05 0.55 0.45 1.1
N2  0.09  0.67  0.00 0.50 0.50 1.0
N3 -0.12  0.90  0.00 0.75 0.25 1.0
N4  0.08  0.71  0.04 0.57 0.43 1.0

With factor correlations of
      MR1  MR3   MR2
MR1  1.00 0.34 -0.03
MR3  0.34 1.00  0.25
MR2 -0.03 0.25  1.00
```

Typical step-by-step procedure for assessing quality of measurement?

- 1. check data -> outliers, missing data etc.
- 2. check correlations
- 3. More than 1 factor/component?
- 4. include only those items that form a scale
- 5. compute reliability (Cronbach's alpha) of indicators for every factor using `psych::alpha()`

Additional Reading

- Andy Field is a useful reference
- DO NOT use **Edition 3** or earlier
  - Mixes up PCA and EFA
- See instead **Edition 4** onwards

Universiteit Utrecht

See you Thursday!