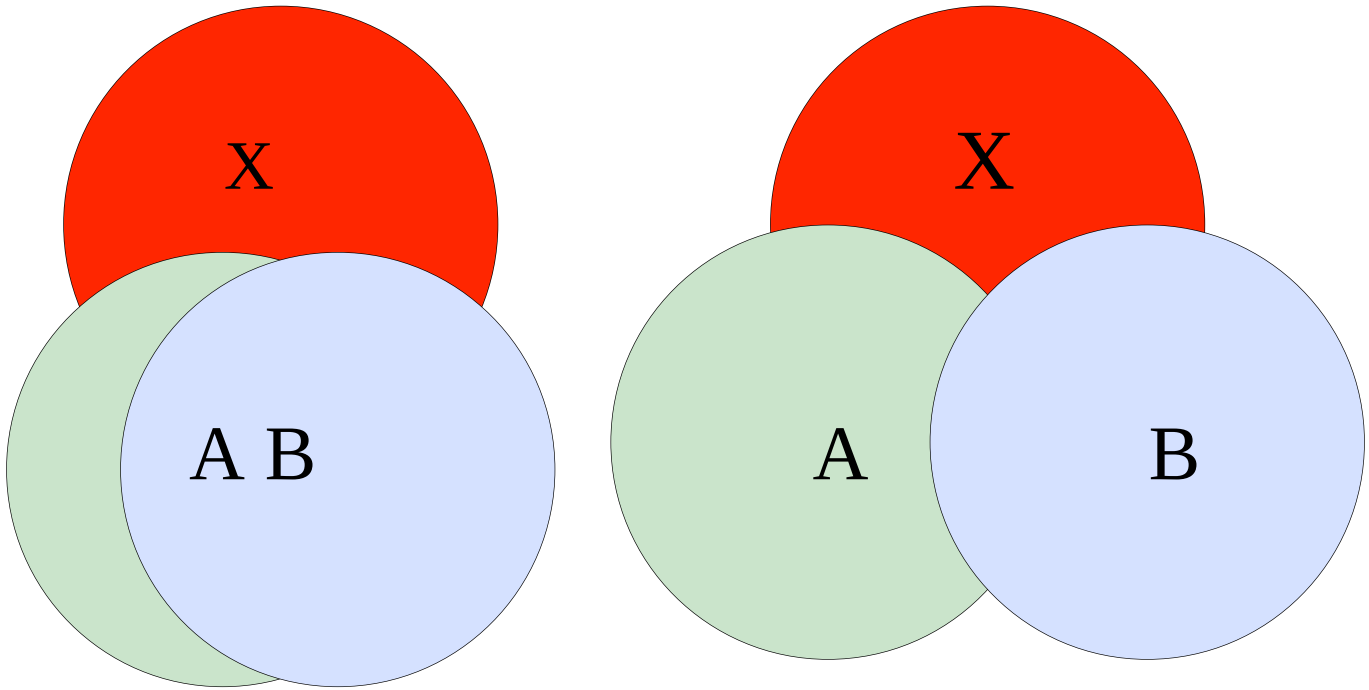
Multicollinearity
Data Reduction
(FA & PCA)

Multicollinearity



- high degree of correlation amongst IVs
  - ex: height and weight, household income and water consumption, mileage and price of a car

Multicollinearity in IVs

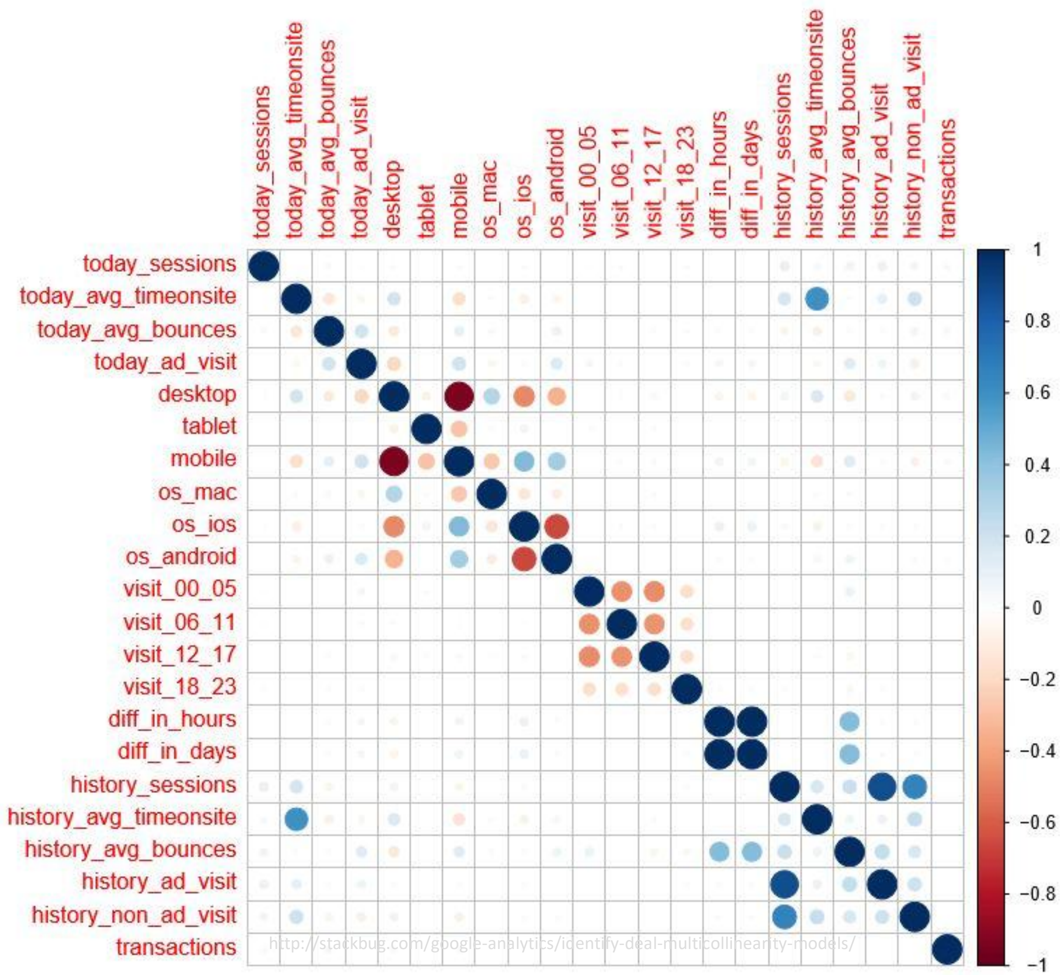
- causes unwanted effects
  - saps statistical power of the analysis
  - can cause switch in signs of the coefficients (in regression), overestimate standard errors, reduced precision in estimating the coefficients' effects, etc…
  - will result in less reliable statistical inferences
- higher number of IV —> increase in sample size required
- what can you do?
  - removing highly correlated IVs / features / items / predictors
  - combine them/uncover latent dimensions **[PCA, FA]**

Multicollinearity

- Some Solutions:
  - **Feature or Variable Selection**
  - **Reduce by Combining Variables**
- choice depends upon
  - research inquiry
  - interpretability

Feature or Variable Selection

- **Correlation:** helps identify collinear variables



where R_j^2 is the R^2-value obtained by regressing the j^{th} predictor on the remaining predictors.

Feature or Variable Selection

- **Variance Inflation Factor (VIF)**
  - The R-square term tells us
    - how predictable one IV is from the set of other IVs
    - 1 = not correlated.
    - Between 1 and 5 = moderately correlated.
    - Greater than 5 = highly correlated.

$$VIF_j = \frac{1}{1 - R_j^2}$$

Where, R_j^2 is the R^2-value obtained by regressing the j^{th} predictor on the remaining predictors.

EXAMPLE

Feature or Variable Selection

	Gender	Age	Years of service	Education level	Salary
0	0.0	27.0	1.7	0.0	39343.0
1	1.0	26.0	1.1	1.0	43205.0
2	1.0	26.0	1.2	0.0	47731.0
3	0.0	27.0	1.6	1.0	46525.0
4	0.0	26.0	1.5	1.0	40891.0

	variables	VIF
0	Gender	2.207155
1	Age	13.706320
2	Years of service	10.299486
3	Education level	2.409263

	variables	VIF
0	Gender	1.863482
1	Years of service	2.478640
2	Education level	2.196539

Dropping Age

	variables	VIF
0	Gender	2.168068
1	Education level	2.407695
2	Age_at_joining	3.326991

(Age - Years of service)

Combining Age & Service

https://www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/

Multicollinearity

- **Solutions:**
  - *Feature or Variable Selection*

  - $\boxed{\textit{Reduce by Combining Variables}}$

- choice depends upon
  - research inquiry
  - interpretability vs model performance

Feature Set Reduction

- Why?
  - increase in dimensions -> complex data -> harder to interpret
  - additional variables = additional processing time and space
  - avoid curse of dimensionality -> amount of data needed to support the result often grows exponentially with the dimensionality
  - reduce overfitting
  - help eliminate irrelevant features
  - easier visualisation

Research Question?

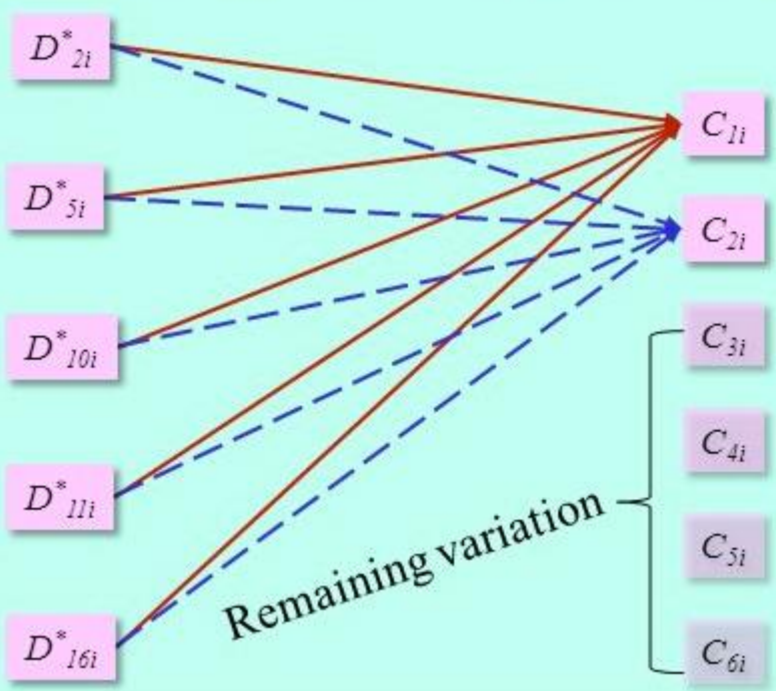
*Rather than asking … **"Can We <u>Forge These Several Indicators</u> Together Into A <u>Smaller Number Of Composites</u> With Defined Statistical Properties?"***

*We could ask … **"Are There A Number Of <u>Unseen (Latent) Factors (Constructs)</u> Acting "<u>Beneath</u>" These Indicators To Forge Their Observed Values?"***

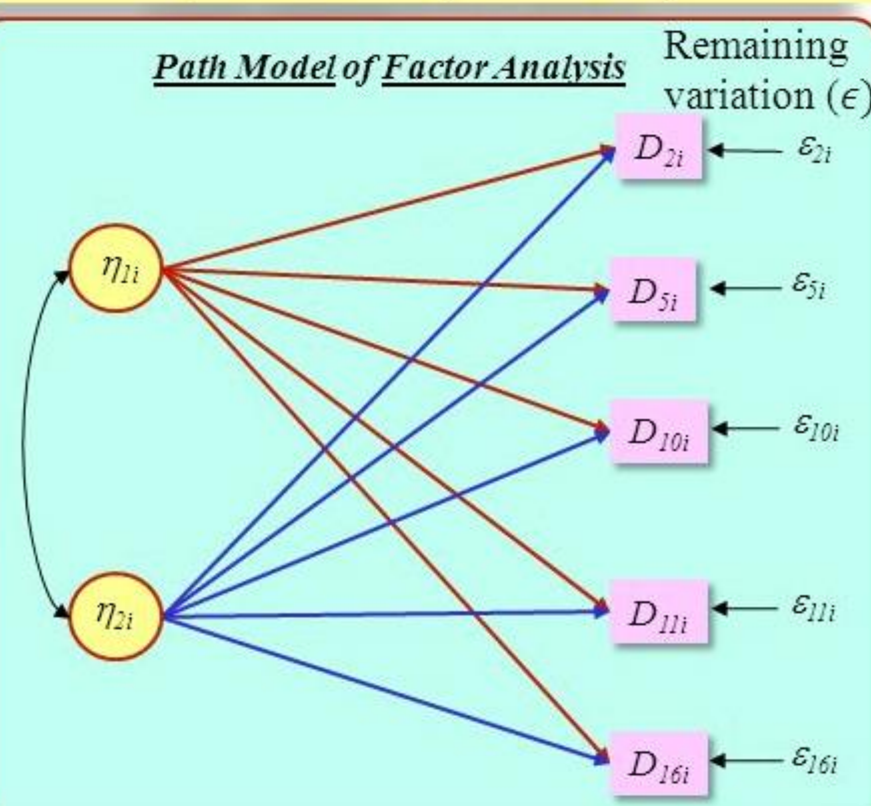
Then, we would need …
***<u>Principal Components Analysis</u>* (PCA)**

Instead, we would need …
***<u>Factor Analysis</u>* (*CFA* or *EFA?*)**

<u>Path Model</u> of <u>Principal Components Analysis</u>

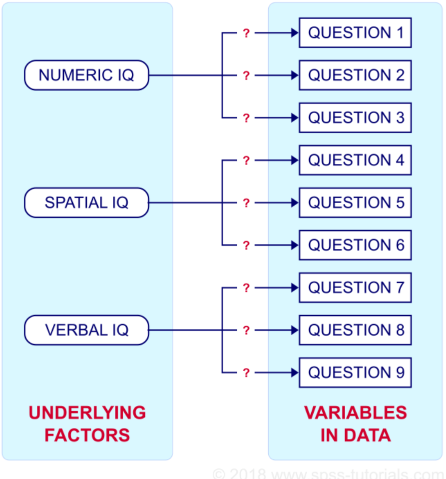


<u>Path Model</u> of <u>Factor Analysis</u>



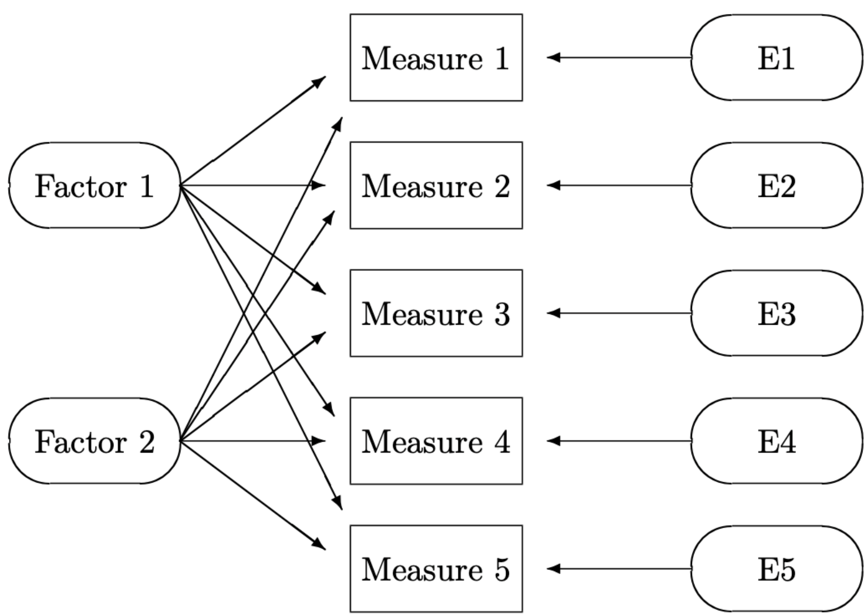
Factor Analysis

- idea—> there are underlying “latent” variables or “factors”, and several variables might be measures of the same factor
- underlying/latent dimensions are not directly observable
- hidden constructs/factors give rise to observed variables



Factor Analysis

- condense information into **factors** with minimum information loss
- predetermined no. of factors (intrinsic dimensionality estimation)



Factor Analysis

$$\mathbf{X} = \mu + \mathbf{L}\mathbf{f} + \epsilon$$

$$\begin{aligned} X_1 &= \mu_1 + l_{11} f_1 + l_{12} f_2 + \cdots + l_{1m} f_m + \epsilon_1 \\ X_2 &= \mu_2 + l_{21} f_1 + l_{22} f_2 + \cdots + l_{2m} f_m + \epsilon_2 \\ &\vdots \\ X_p &= \mu_p + l_{p1} f_1 + l_{p2} f_2 + \cdots + l_{pm} f_m + \epsilon_p \end{aligned}$$

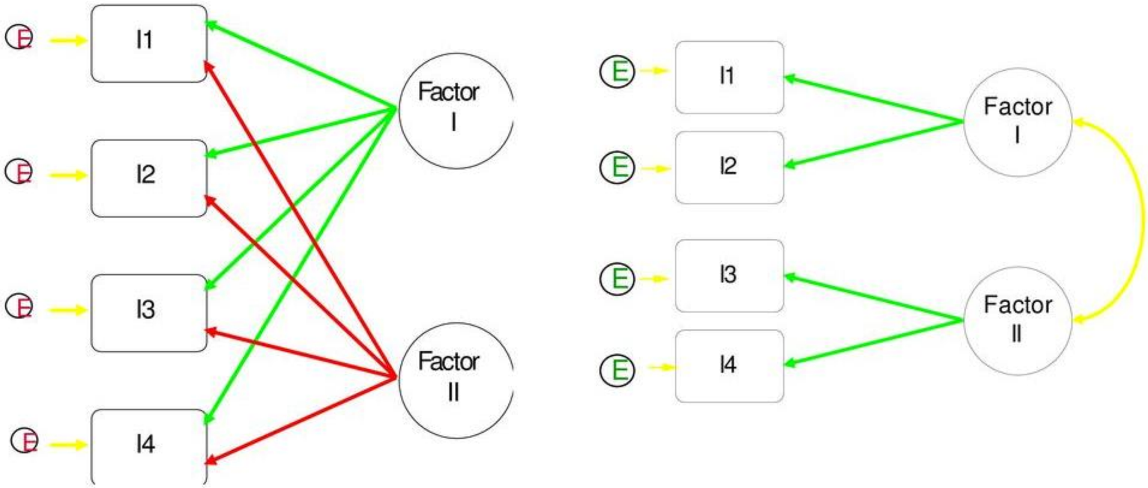
$$\mathbf{L} = \begin{pmatrix} l_{11} & l_{12} & \dots & l_{1m} \\ l_{21} & l_{22} & \dots & l_{2m} \\ \vdots & \vdots & & \vdots \\ l_{p1} & l_{p2} & \dots & l_{pm} \end{pmatrix} = \text{matrix of factor loadings}$$

$$\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_p \end{pmatrix} = \text{vector of specific factors}$$

error terms, what the Factors cannot explain in each variable

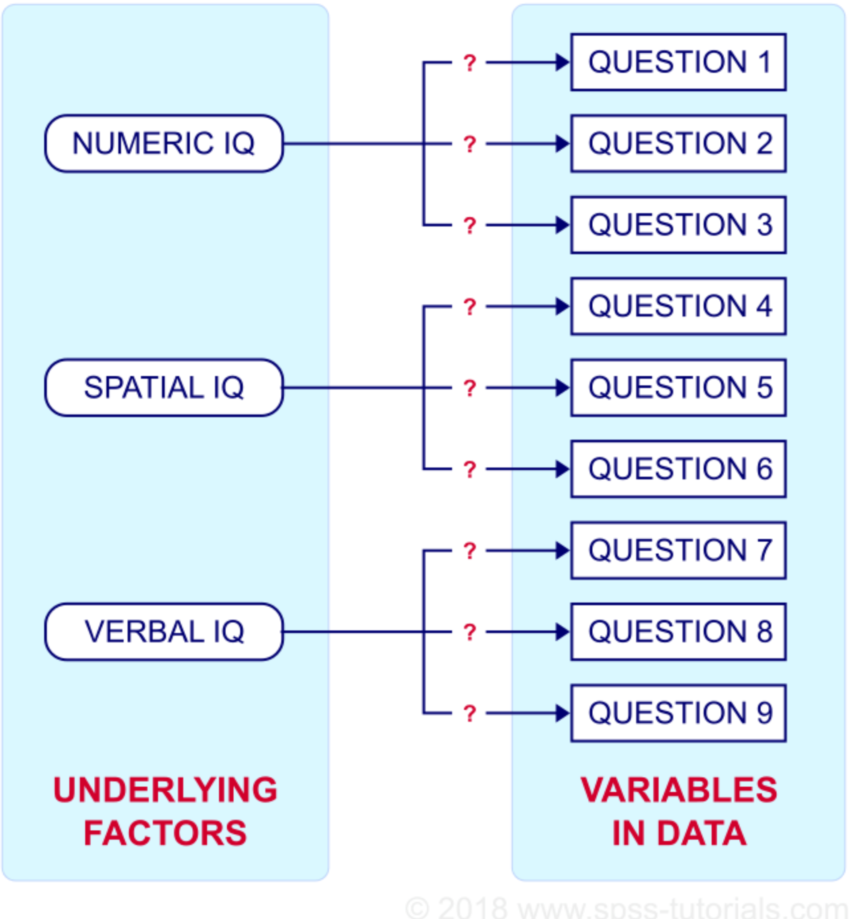
Factor Analysis

- **Exploratory Factor Analysis**: *data-driven*
  - explore underlying structure
- **Confirmatory Factor Analysis**: *theory-driven*
  - confirm or reject pre-established theory

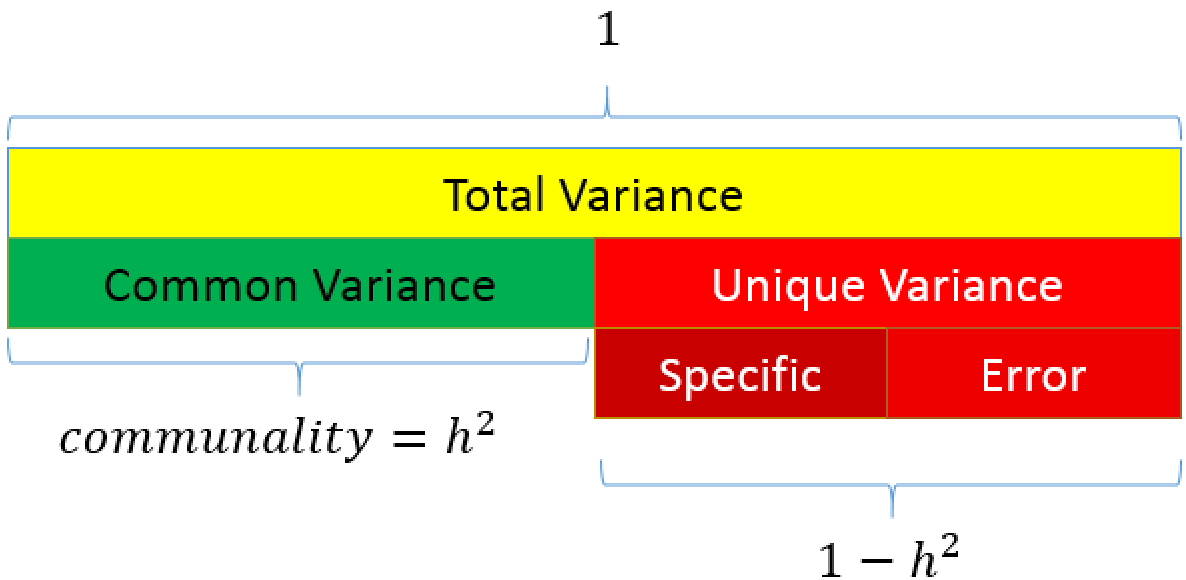


EFA: Factor Analysis Types

- **R-Type** (commonly used)
    - covariation or correlation between variables

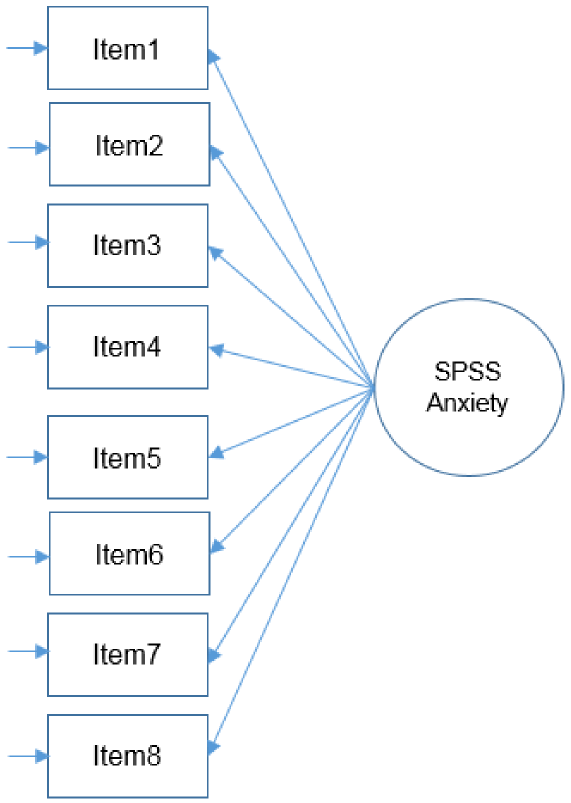


Factor Analysis



The total variance is made up to common variance and unique variance, and unique variance is composed of specific and error variance. If the total variance is 1, then the communality is h^2 and the unique variance is $1-h^2$.

1. Statistics makes me cry
2. My friends will think I’m stupid for not being able to cope with SPSS
3. Standard deviations excite me
4. I dream that Pearson is attacking me with correlation coefficients
5. I don’t understand statistics
6. I have little experience with computers
7. All computers hate me
8. I have never been good at mathematics



Do all these items actually measure what we call “SPSS Anxiety”?

https://stats.idre.ucla.edu/spss/seminars/efa-spss/

EXAMPLE

	Statistics makes me cry	My friends will think I'm stupid for not being able to cope with SPSS	Standard deviations excite me	I dream that Pearson is attacking me with correlation coefficients	I don't understand statistics	I have little experience with computers	All computers hate me	I have never been good at mathematics
Statistics makes me cry	1							
My friends will think I'm stupid for not being able to cope with SPSS	-.099	1						
Standard deviations excite me	-.337	.318	1					
I dream that Pearson is attacking me with correlation coefficients	.436	-.112	-.380	1				
I don't understand statistics	.402	-.119	-.310	.401	1			
I have little experience with computers	.217	-.074	-.227	.278	.257	1		
All computers hate me	.305	-.159	**-.382**	.409	.339	**.514**	1	
I have never been good at mathematics	.331	-.050	-.259	.349	.269	.223	.297	1

Inter-scale/item correlation

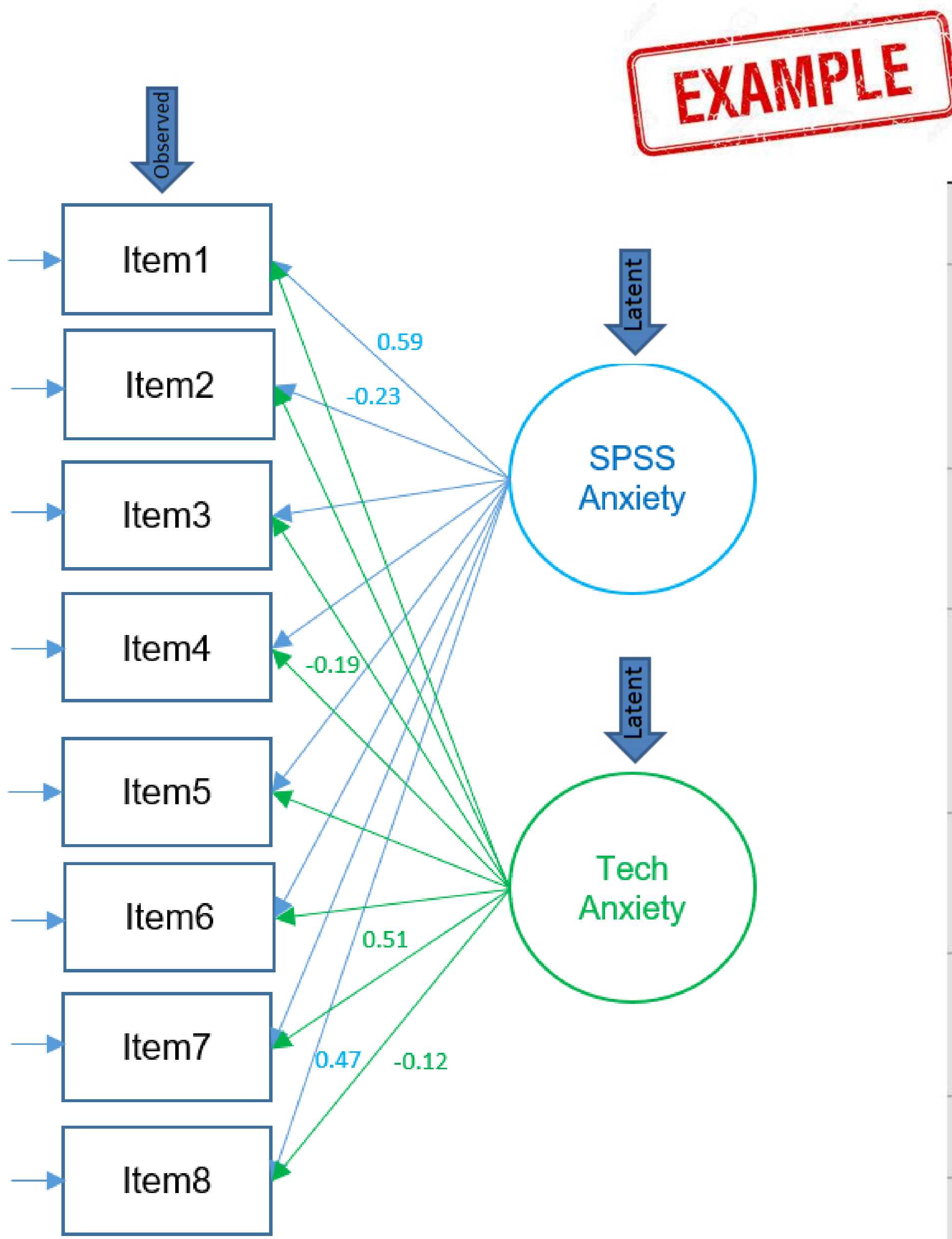
Factor Matrix[a]

EXAMPLE

	Factor	
	1	2
Statistics makes me cry	.588	-.303
My friends will think I'm stupid for not being able to cope with SPSS	-.227	.020
Standard deviations excite me	-.557	.094
I dream that Pearson is attacking me with correlation coefficients	.652	-.189
I don't understand statistics	.560	-.174
I have little experience of computers	.498	.247
All computers hate me	.771	.506
I have never been good at mathematics	.470	-.124

1. Statistics makes me cry
2. My friends will think I'm stupid for not being able to cope with SPSS
3. Standard deviations excite me
4. I dream that Pearson is attacking me with correlation coefficients
5. I don't understand statistics
6. I have little experience with computers
7. All computers hate me
8. I have never been good at mathematics

Factor Loadings: the weight of the factor in predicting the variable/correlations between variables and factors



Note: only selected loadings shown

Factor Matrix[a]

	Factor 1	Factor 2
Statistics makes me cry	.588	-.303
My friends will think I'm stupid for not being able to cope with SPSS	-.227	.020
Standard deviations excite me	-.557	.094
I dream that Pearson is attacking me with correlation coefficients	.652	-.189
I don't understand statistics	.560	-.174
I have little experience of computers	.498	.247
All computers hate me	.771	.506
I have never been good at mathematics	.470	-.124

EXAMPLE

Factor Interpretation

F1: customer experience post boarding

F2: airline booking experience and related perks

F3: flight competitive advantage of the airline compared to its competition

	Factor 1	Factor 2	Factor 3
Great hospitality	0.98	-0.04	0.02
Flight is on time	0.95	-0.01	0.18
Great Food	0.92	0.04	-0.05
Friendly atmosphere	0.62	0.17	-0.33
Frequent flyer program	-0.03	0.97	-0.01
Flights are economic	-0.02	0.96	0.09
No hassles in boarding	-0.07	0.95	0.09
Good flight times	-0.09	0.19	0.96
Seats are comfortable	0.03	0.09	0.95
Loyalty or attachment	-0.19	-0.42	-0.09

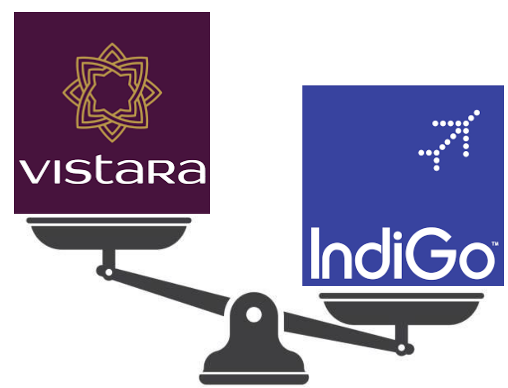
ex: factor loadings for an airlines survey

Factor Scores

- composite scores represented by the latent variable which can be used in subsequent statistical analyses (ex: multiple regression, t-tests, etc.)

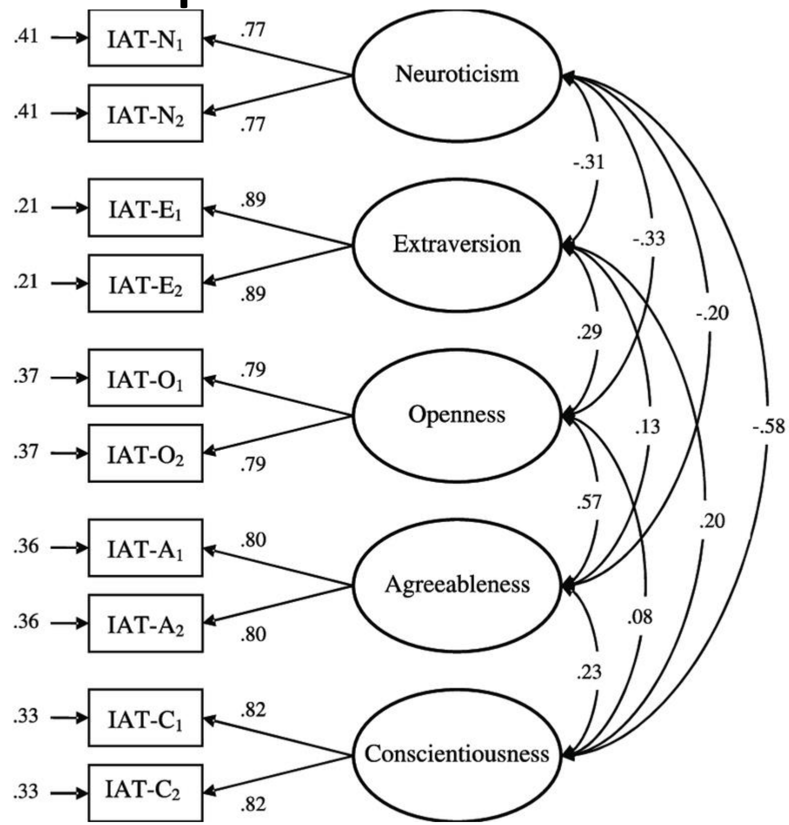
F1: customer experience post boarding

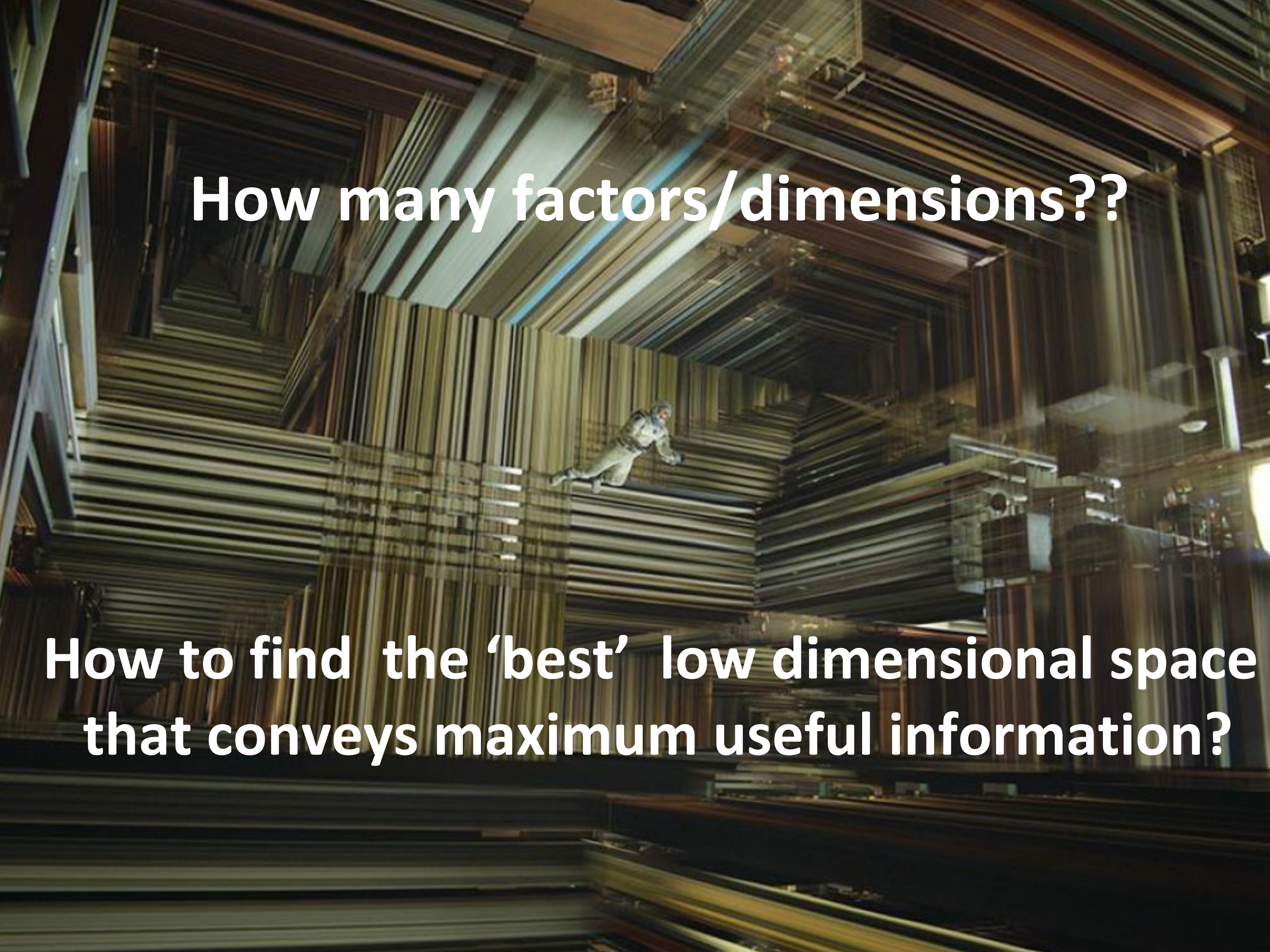
F2: airline booking experience and related perks



Factor Analysis Types

- **Q-Type**
  - similar to clustering of people
  - allows identification of groups
  - ex: participant X's responses are similar to Y's





How many factors/dimensions??

How to find the ‘best’ low dimensional space that conveys maximum useful information?

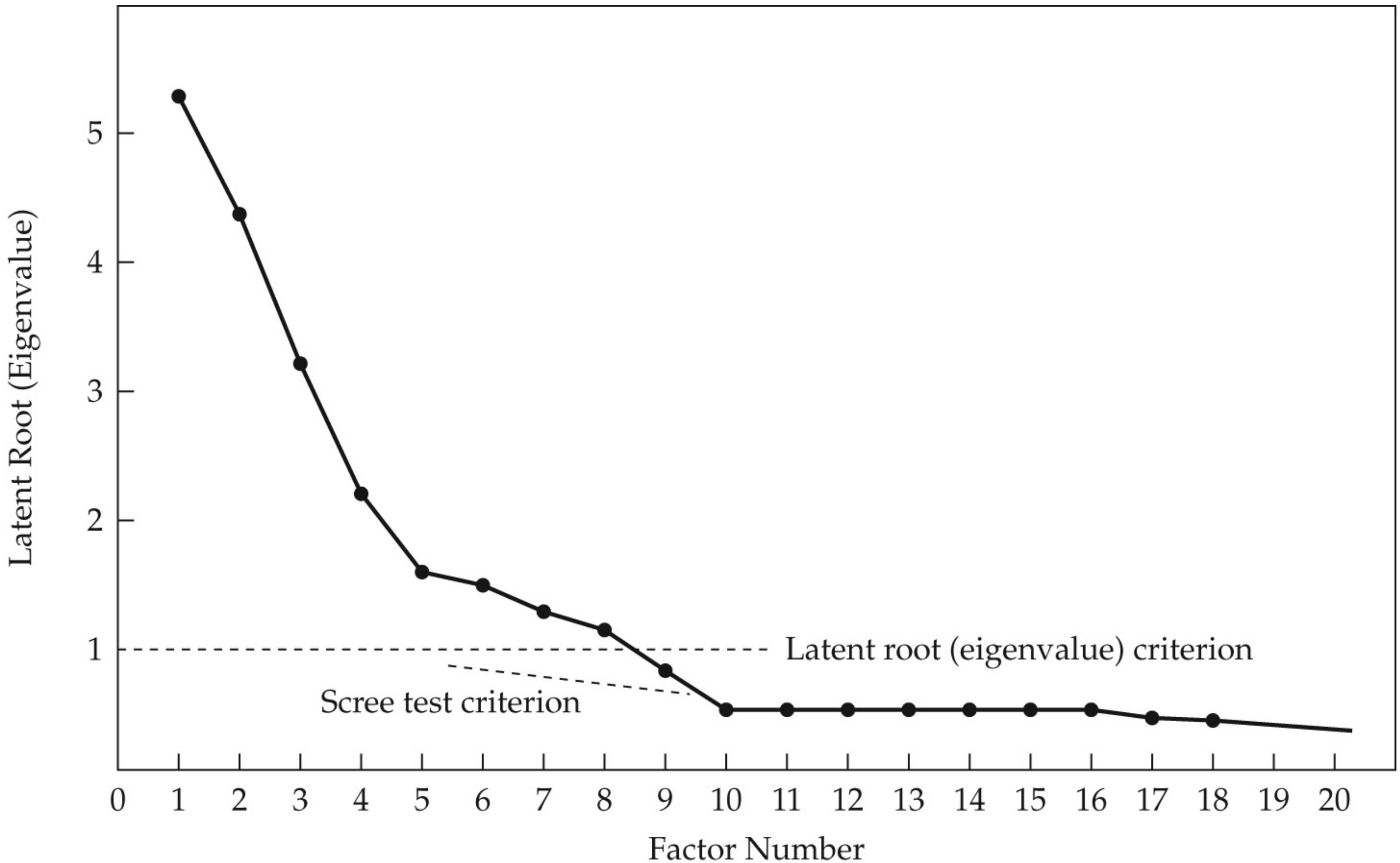
Dimensionality Estimation

- **a priori criterion**
  - define a priori the number of factors to be extracted (testing a hypothesis about the number of factors)
  - trade off - representativeness vs parsimony
- **latent Root criterion**
  - any individual factor should account for the variance of at least one single variable – latent root or eigenvalue >1
- **scree plot/test**
  - point of inflexion in latent root plot

Terminology

- **Scree Plot**
  - plots eigenvalue against component number
  - components with eigenvalues greater than 1 are retained (they are the ’principal’ components)
  - components with eigenvalues less than 1 are of little use because they account for less of the variance than the original variable

Scree Plot


Latent Root (Eigenvalue)
5
4
3
2
1
Latent root (eigenvalue) criterion
Scree test criterion
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Factor Number

Dimensionality Estimation

- **parallel Analysis** (widely used)
  - based on the Monte Carlo simulation
  - creating a random dataset with the same numbers of observations and variables as the original data
  - compare eigenvalues from the random data with original data

Dimensionality Estimation Example

Healthy-Unhealthy Music Scale (HUMS)

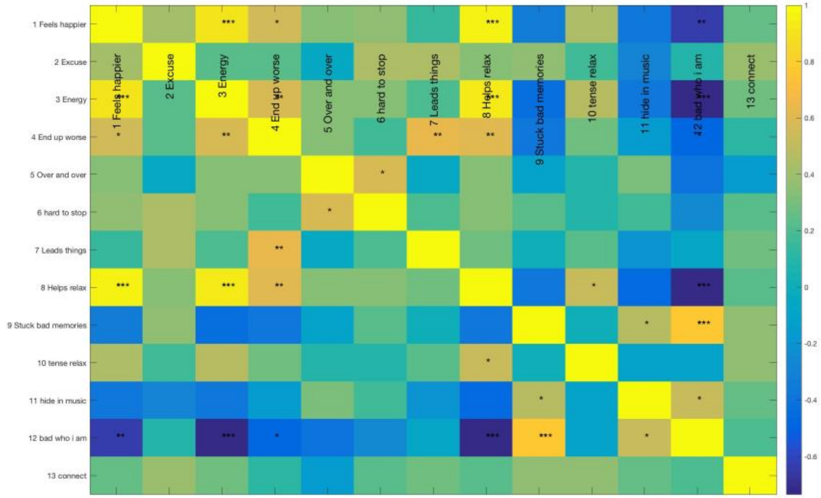
Most people believe that music is a helpful part of their lives, but sometimes it's not. When you answer the questions below, please try to recall actual moments when music has been helpful and when it has not.

Please read each statement and mark how much it applies to you. Mark only one answer for each question.

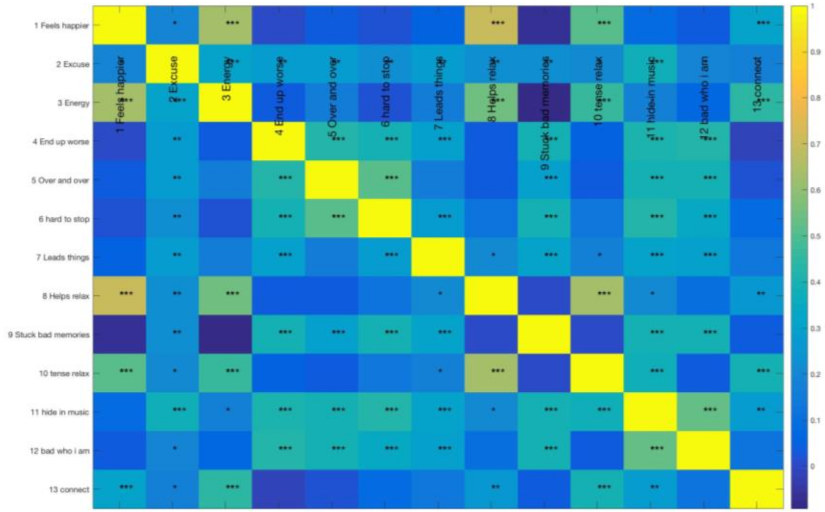
		Never	Rarely	Some-times	Often	Always
1.	When I listen to music I get stuck in bad memories	☐	☐	☐	☐	☐
2.	I hide in my music because nobody understands me, and it blocks people out	☐	☐	☐	☐	☐
3.	Music helps me to relax	☐	☐	☐	☐	☐
4.	When I try to use music to feel better I actually end up feeling worse	☐	☐	☐	☐	☐
5.	I feel happier after playing or listening to music	☐	☐	☐	☐	☐
6.	Music gives me the energy to get going	☐	☐	☐	☐	☐
7.	I like to listen to songs over and over even though it makes me feel worse	☐	☐	☐	☐	☐
8.	Music makes me feel bad about who I am	☐	☐	☐	☐	☐
9.	Music helps me to connect with other people who are like me	☐	☐	☐	☐	☐
10.	Music gives me an excuse not to face up to the real world	☐	☐	☐	☐	☐
11.	It can be hard to stop listening to music that connects me to bad memories	☐	☐	☐	☐	☐
12.	Music leads me to do things I shouldn't do	☐	☐	☐	☐	☐
13.	When I'm feeling tense or tired in my body music helps me to relax	☐	☐	☐	☐	☐

Inter-Scale/Item Correlation

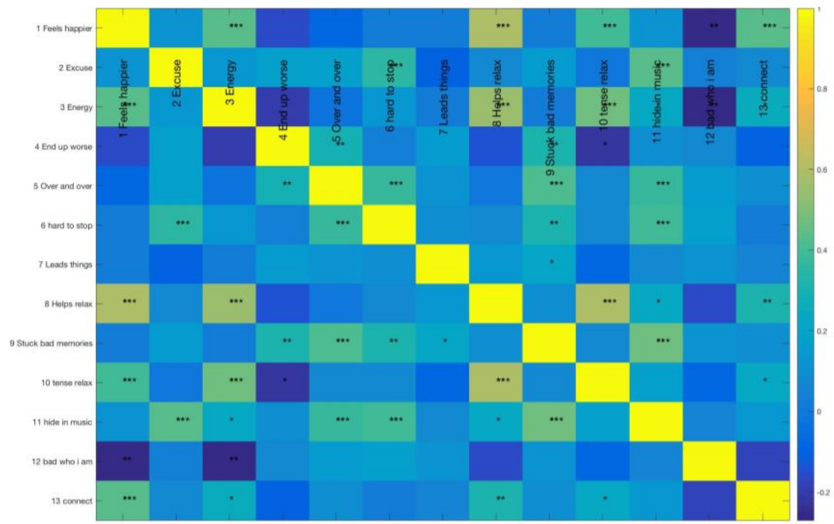
RM class 2018
25 students



141 Indians

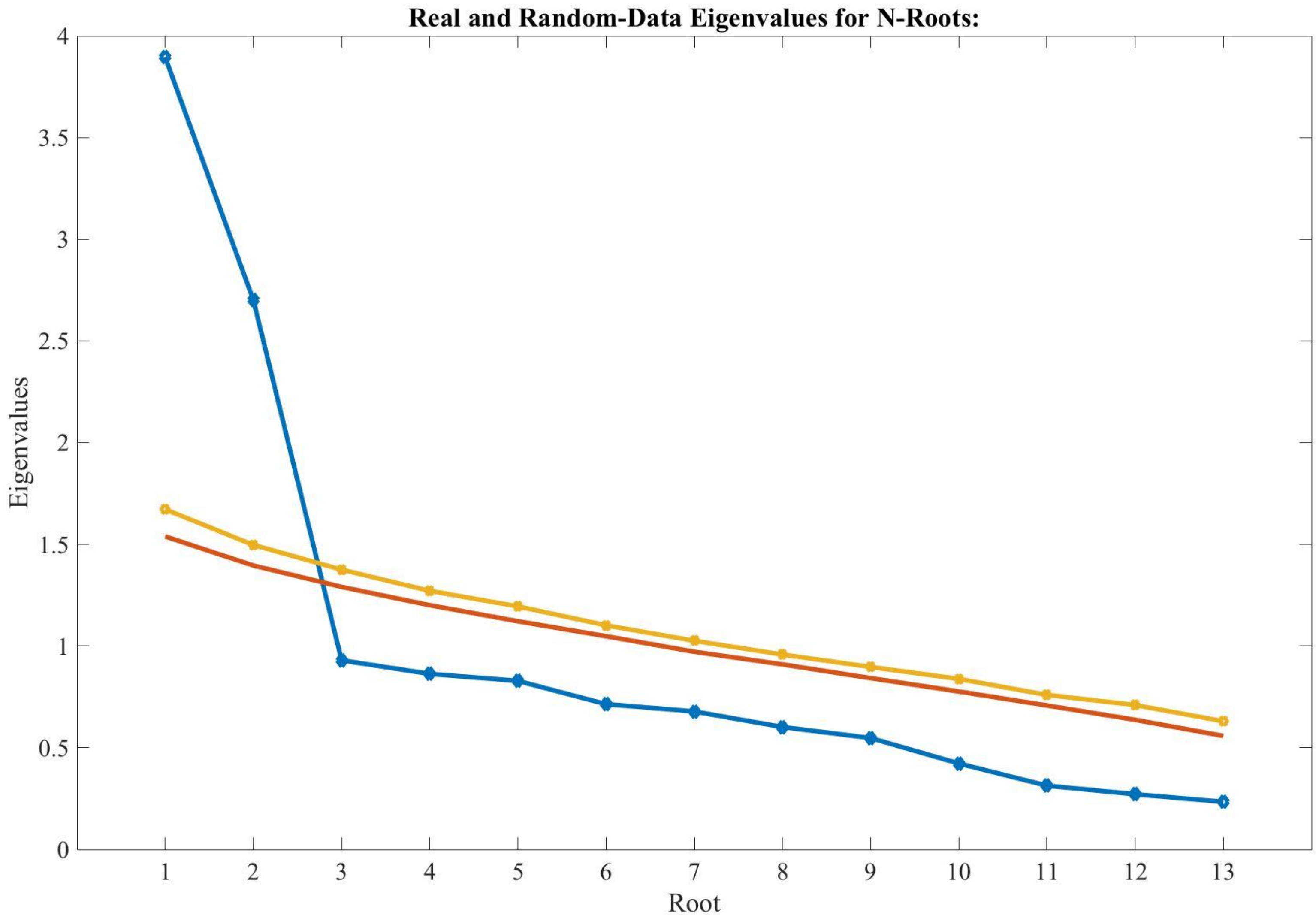


102 British



Parallel Analysis

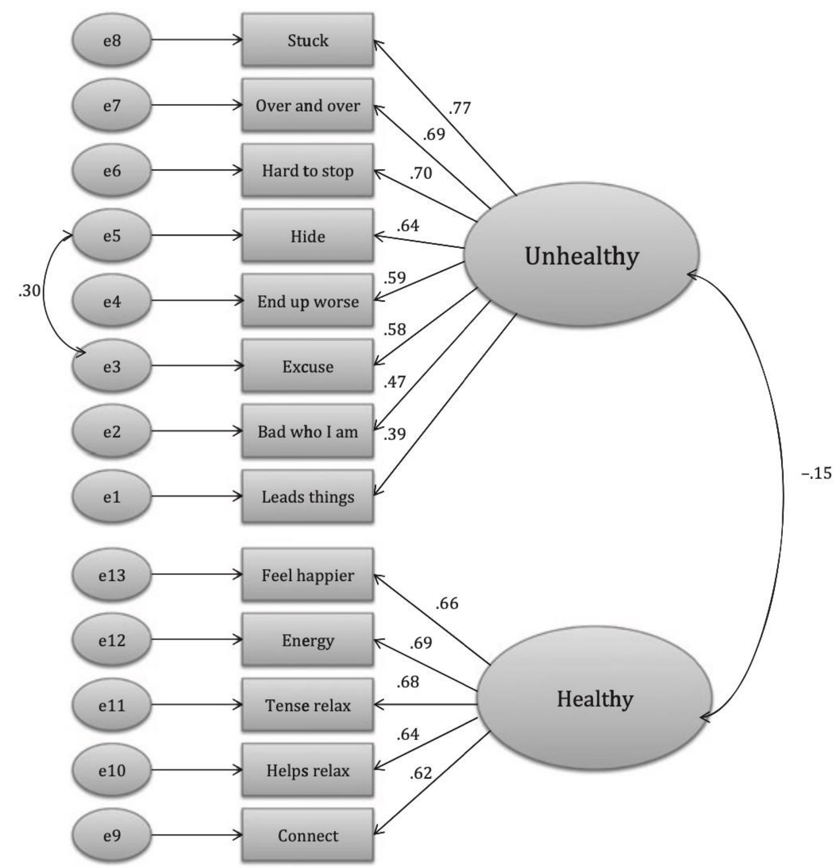
141 Indians



Factor Interpretation

Table 2. The factor loadings (pattern matrix) of the final version of Healthy-Unhealthy Music Scale

Items	F1	F2
When I listen to music I get stuck in bad memories	**.760**	−.033
I like to listen to songs over and over even though it makes me feel worse	**.714**	−.092
It can be hard to stop listening to music that connects me to bad memories	**.658**	.187
I hide in my music because nobody understands me, and it blocks people out	**.639**	.156
When I try to use music to feel better I actually end up feeling worse	**.627**	−.163
Music gives me an excuse not to face up to the real world	**.571**	.249
Music makes me feel bad about who I am	**.521**	−.186
Music leads me to do things I shouldn't do	**.428**	−.103
I feel happier after playing or listening to music	−.157	**.708**
Music gives me the energy to get going	−.005	**.692**
When I'm feeling tense or tired in my body music helps me to relax	−.028	**.667**
Music helps me to relax	.040	**.621**
Music helps me to connect with other people who are like me	−.061	**.608**



Principal Component Analysis

Research Question?

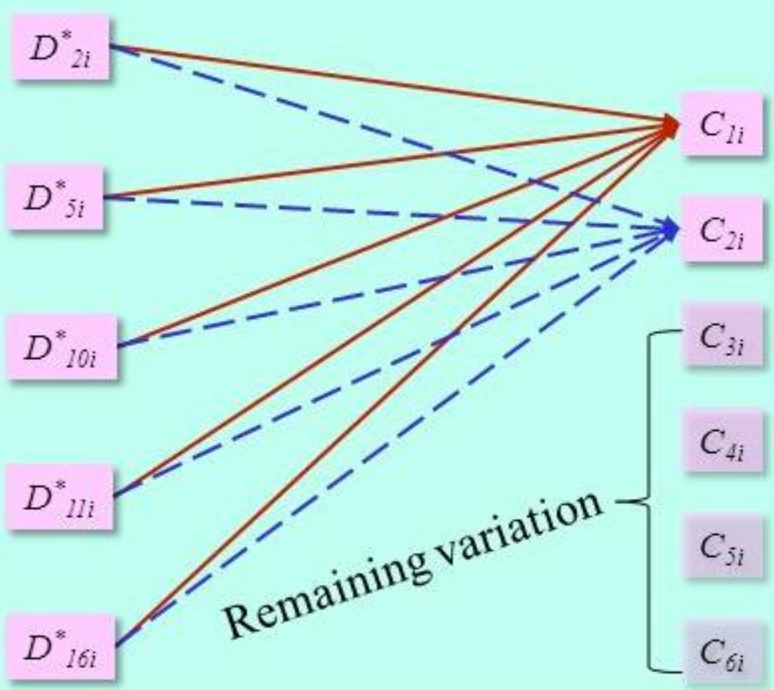
Rather than asking ... ***"Can We <u>Forge These Several Indicators</u> Together Into A <u>Smaller Number Of Composites</u> With Defined Statistical Properties?"***

Then, we would need ...
***<u>Principal Components Analysis</u>* (*PCA*)**

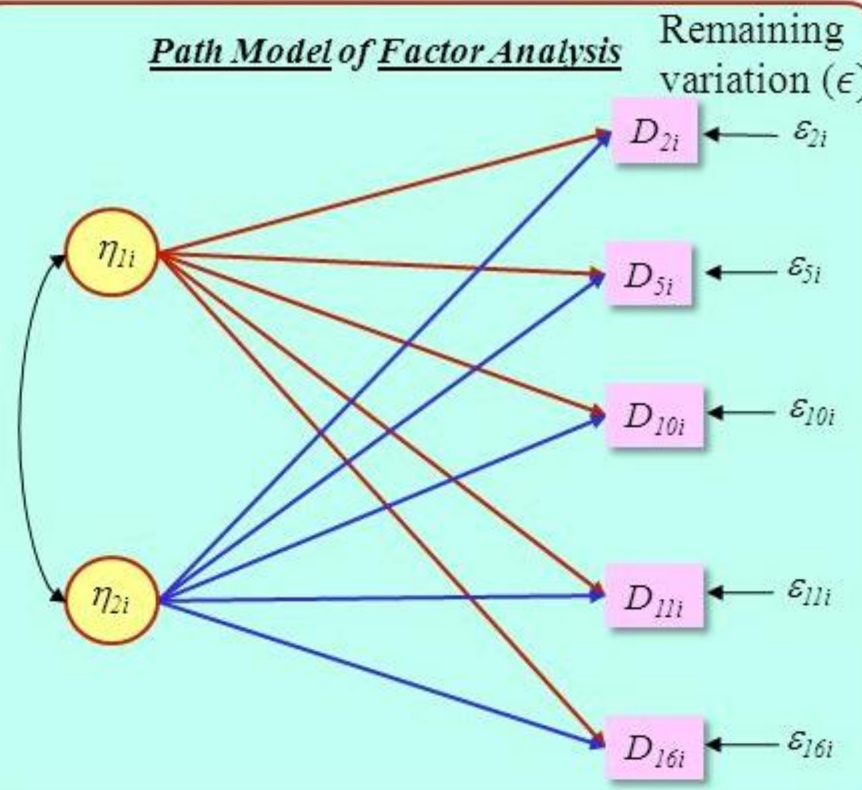
We could ask ... ***"Are There A Number Of <u>Unseen (Latent) Factors (Constructs)</u> Acting "<u>Beneath</u>" These Indicators To Forge Their Observed Values?"***

Instead, we would need ...
***<u>Factor Analysis</u>* (*CFA* or *EFA?*)**

<u>Path Model</u> of <u>Principal Components Analysis</u>

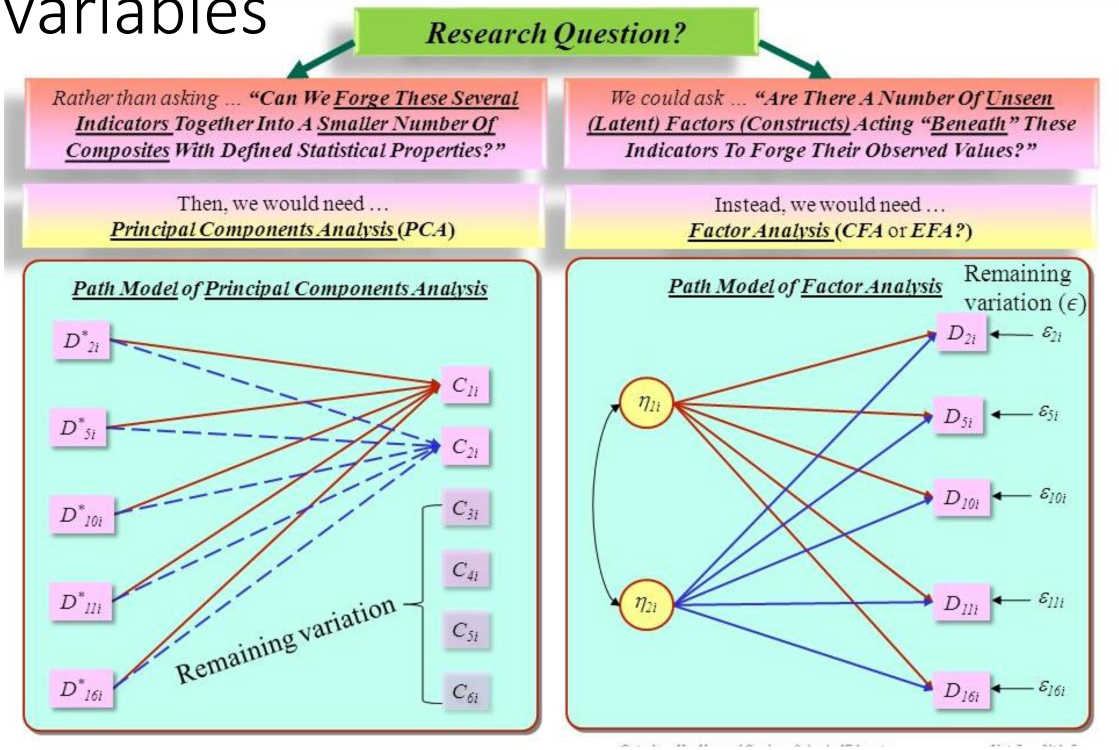


<u>Path Model</u> of <u>Factor Analysis</u>

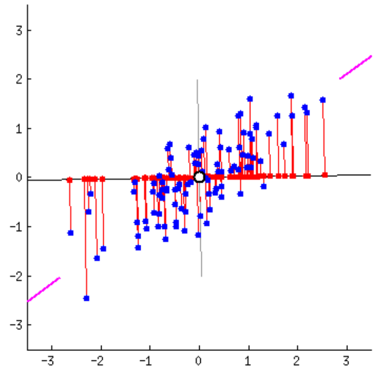


PCA

- idea —> reduce the number of variables of a data set while preserving as much information as possible.
- dimensionality reduction by creating linear combinations of variables



PCA



- Example: Combining two variables into a single component
  - Fit a regression line that represents the ’best’ summary of the linear relationship between the variables
  - This line, representing a new component, would capture most of the ’essence’ of the two variables

PCA

- If there are more than two variables...
  - this process is repeated until all variables have been assigned to a component
  - gives as many components as variables in decreasing order of variance explained
  - however, only the first few components are likely to be useful..

PCA

- Assumptions:
  - at least interval level data
  - a linear relationship between all variables
  - sampling adequacy (KMO, ~15 cases/variable), Bartlett’s test of sphericity
  - normally distributed (no outliers)

- Subtract mean from data (center X)
- (Typically) scale each dimension by its variance
  - Helps to pay less attention to magnitude of dimensions
- Compute covariance matrix S $\mathbf{S} = \frac{1}{N}\mathbf{X}^\intercal\mathbf{X}$
- Compute k largest eigenvectors of S
- These eigenvectors are the k principal components

https://www.youtube.com/watch?v=g-Hb26agBFg

https://www.youtube.com/watch?v=PFDu9oVAE-g

Principal Components

- principal components: linear combinations of original variables that result in an axis or a set of axes that explain most of the variability in the dataset
- variables that correlate highly with each other are grouped together into underlying variables, or components
- In mathematical terms, we can say that the first Principal Component is the eigenvector of the covariance matrix corresponding to the maximum eigenvalue

Component Scores & Loadings

- each original variable is assigned a component score and a component loading
- **Component scores** = score/projection on a given component (can be used in subsequent statistical analyses, e.g., regression)
- **Component loadings** = correlation of the original variable with a given component - can be used to determine the importance of a particular variable to a component (Higher loadings = more important)

Dimensionality Estimation

- **Percentage of Variance** criterion
  - achieving a specified cumulative percentage of total variance.
    - typical values – natural sciences ~95%;
    - typical values – social sciences > ~60%
- **Parallel Analysis** (widely used)
  - based on the Monte Carlo simulation
  - creating a random dataset with the same numbers of observations and variables as the original data
  - compare eigenvalues from the random data with original datas

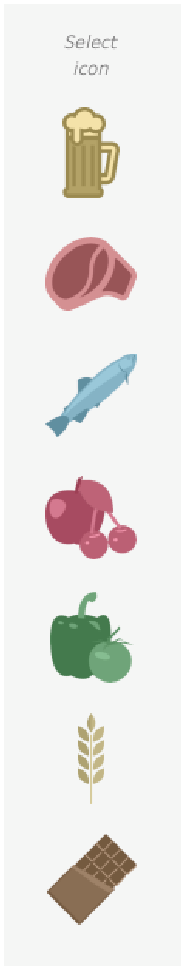
Dimensionality Estimation

- **latent Root criterion**
  - any individual factor should account for the variance of at least one single variable – latent root or eigenvalue >1
- **scree plot/test**
  - point of inflexion in latent root plot

Rotation (similar to FA)

- the reference axes of the factors are turned about the origin until some other position has been reached
- the ultimate effect of rotating the factor matrix is to redistribute the variance from earlier factors to later ones to achieve a simpler, theoretically more meaningful factor pattern

EXAMPLE

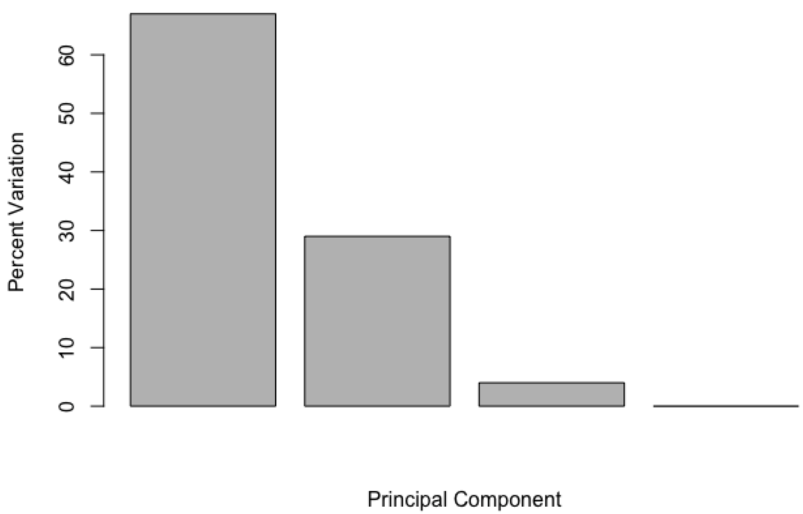


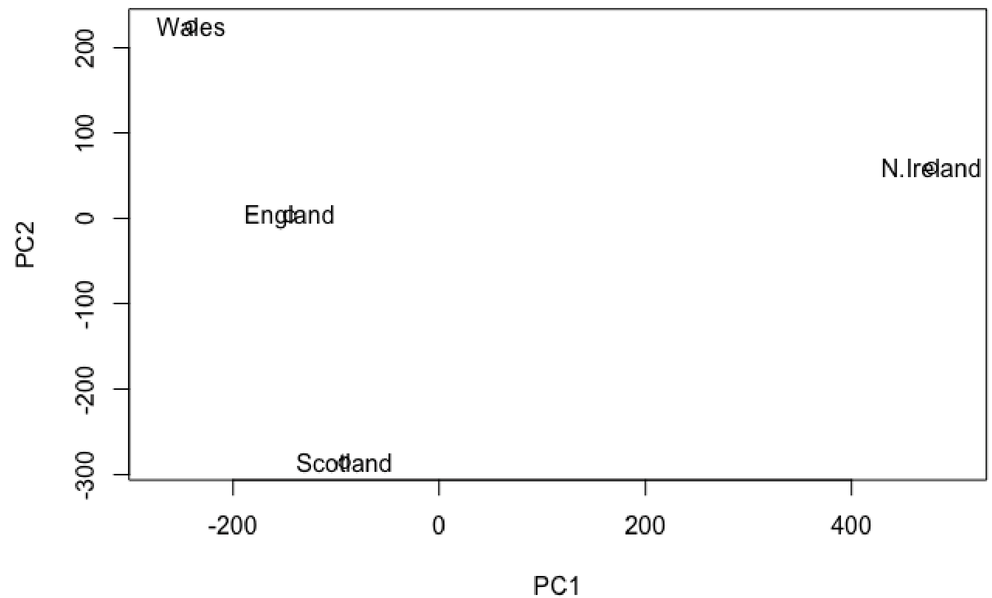
	England	Wales	Scotland	N.Ireland
Cheese	105	103	103	66
Carcass_meat	245	227	242	267
Other_meat	685	803	750	586
Fish	147	160	122	93
Fats_and_oils	193	235	184	209
Sugars	156	175	147	139
Fresh_potatoes	720	874	566	1033
Fresh_Veg	253	265	171	143
Other_Veg	488	570	418	355
Processed_potatoes	198	203	220	187
Processed_Veg	360	365	337	334
Fresh_fruit	1102	1137	957	674
Cereals	1472	1582	1462	1494
Beverages	57	73	53	47
Soft_drinks	1374	1256	1572	1506
Alcoholic_drinks	375	475	458	135
Confectionery	54	64	62	41

data set of foods commonly consumed (in grams per person, per week) in different parts of UK

EXAMPLE

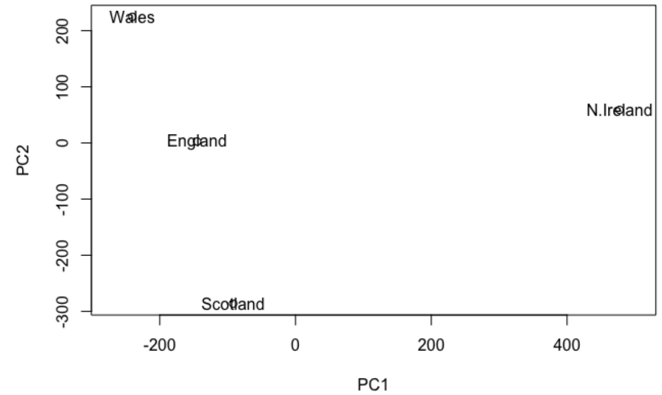
PCA

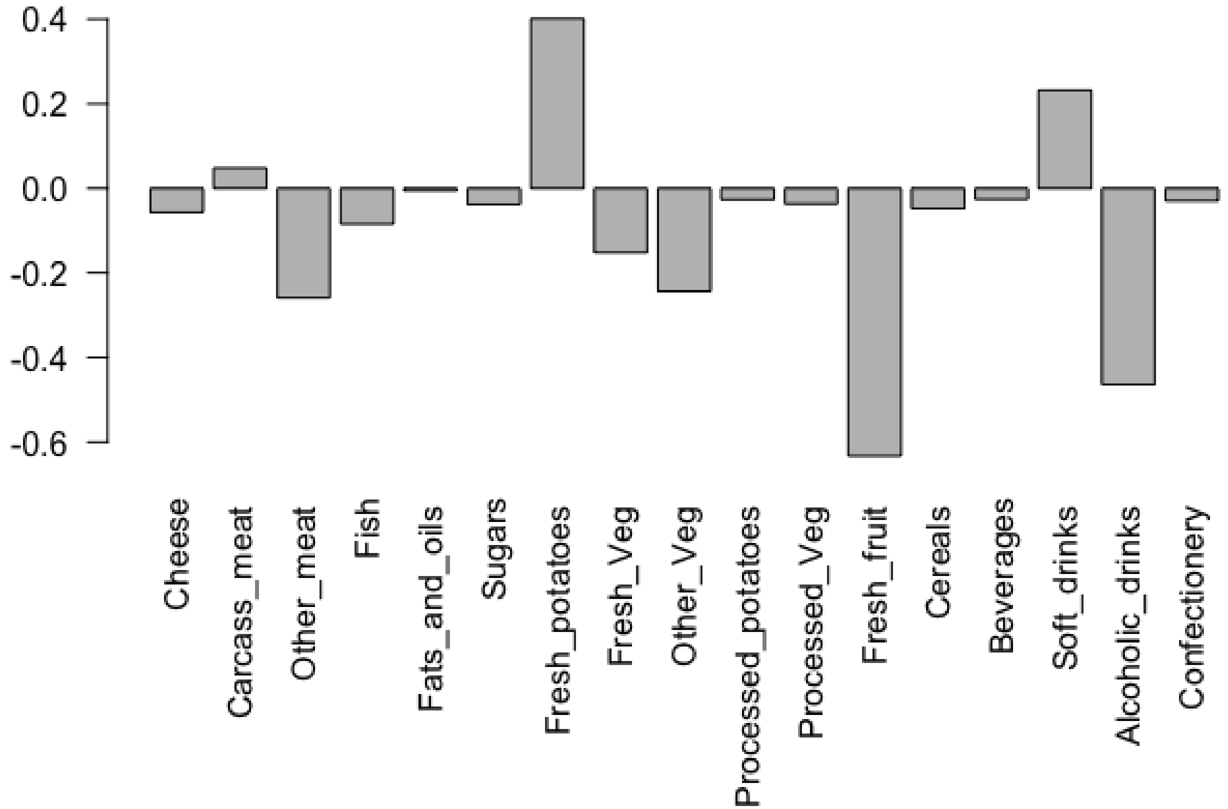




data set of foods commonly consumed (in grams per person, per week) in different parts of UK

PCA

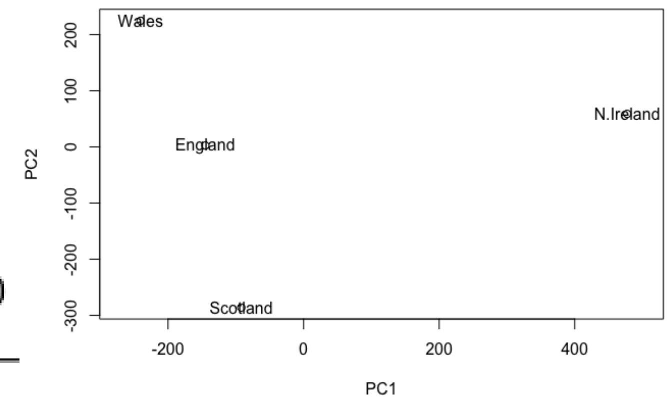


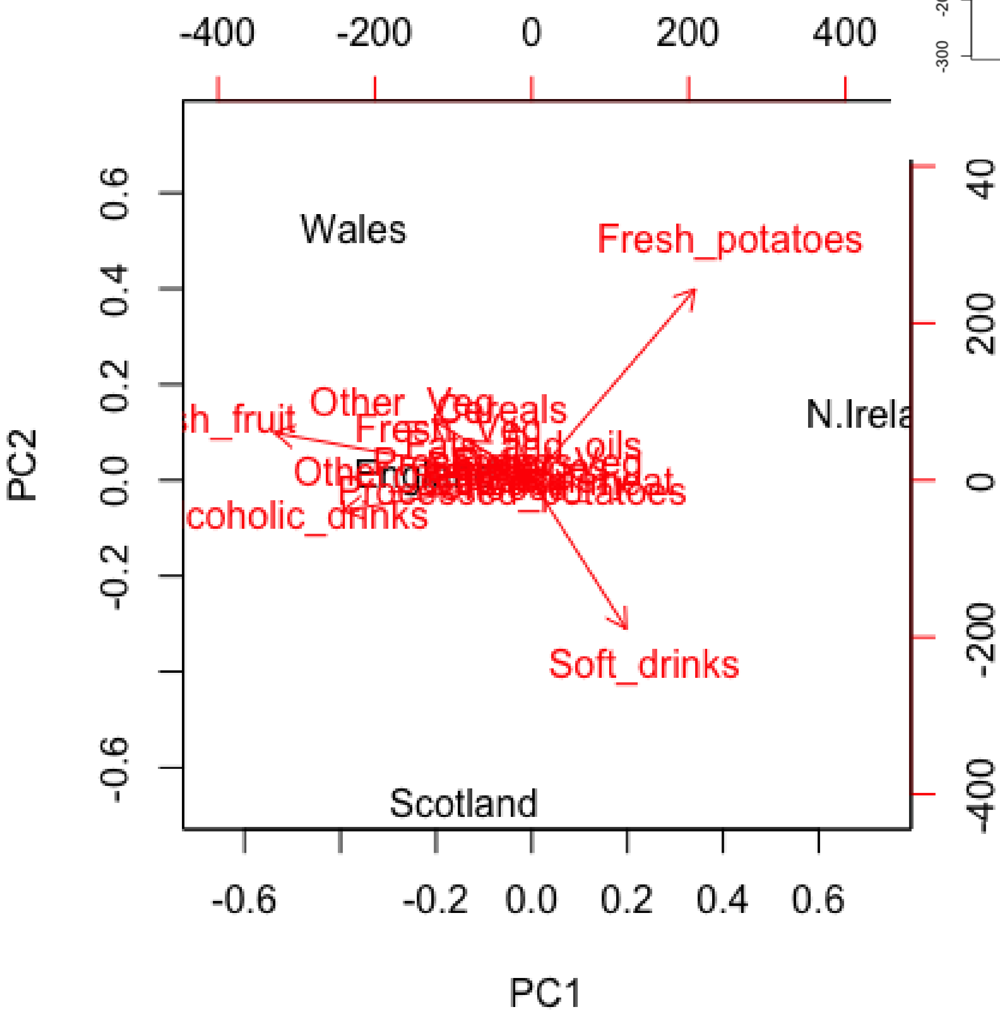


data set of foods commonly consumed (in grams per person, per week) in different parts of UK


EXAMPLE

Biplot


Wales
England
N.Ireland
Scotland
PC1
PC2
-200
0
200
400
-300
-200
-100
0
100
200


-400
-200
0
200
400
Wales
Fresh_potatoes
Other_Veg
Cereals
Fresh_Veg
Soft_drinks
Scotland
N.Irela
coholic_drinks
PC1
PC2
-0.6
-0.2
0.0
0.2
0.4
0.6
-0.2
0.0
0.2
0.4
0.6
-400
-200
0
200
40

Factor analysis	Principal component analysis
Number of factors pre-determined	Number of components evaluated ex post
Many potential solutions	Unique mathematical solution
Factor matrix is estimated	Component matrix is computed
Factor scores are estimated	Component scores are computed
More appropriate when searching for an underlying structure	More appropriate for data reduction (no prior underlying structure assumed)
Factors are not necessarily sorted	Factors are sorted according to the amount of explained variability
Only common variability is taken into account	Total variability is taken into account
Estimated factor scores may be correlated	Component scores are always uncorrelated
A distinction is made between common and specific variance	No distinction between specific and common variability
Preferred when there is substantial measurement error in variables	Preferred as a preliminary method to cluster analysis or to avoid multicollinearity in regression
Rotation is often desirable as there are many equivalent solutions	Rotation is less desirable, unless components are difficult to be interpreted and explained variance is spread evenly across components