MAESTRÍA EN INVESTIGACIÓN OPERATIVA Y ESTADÍSTICA

Universidad Tecnológica de Pereira

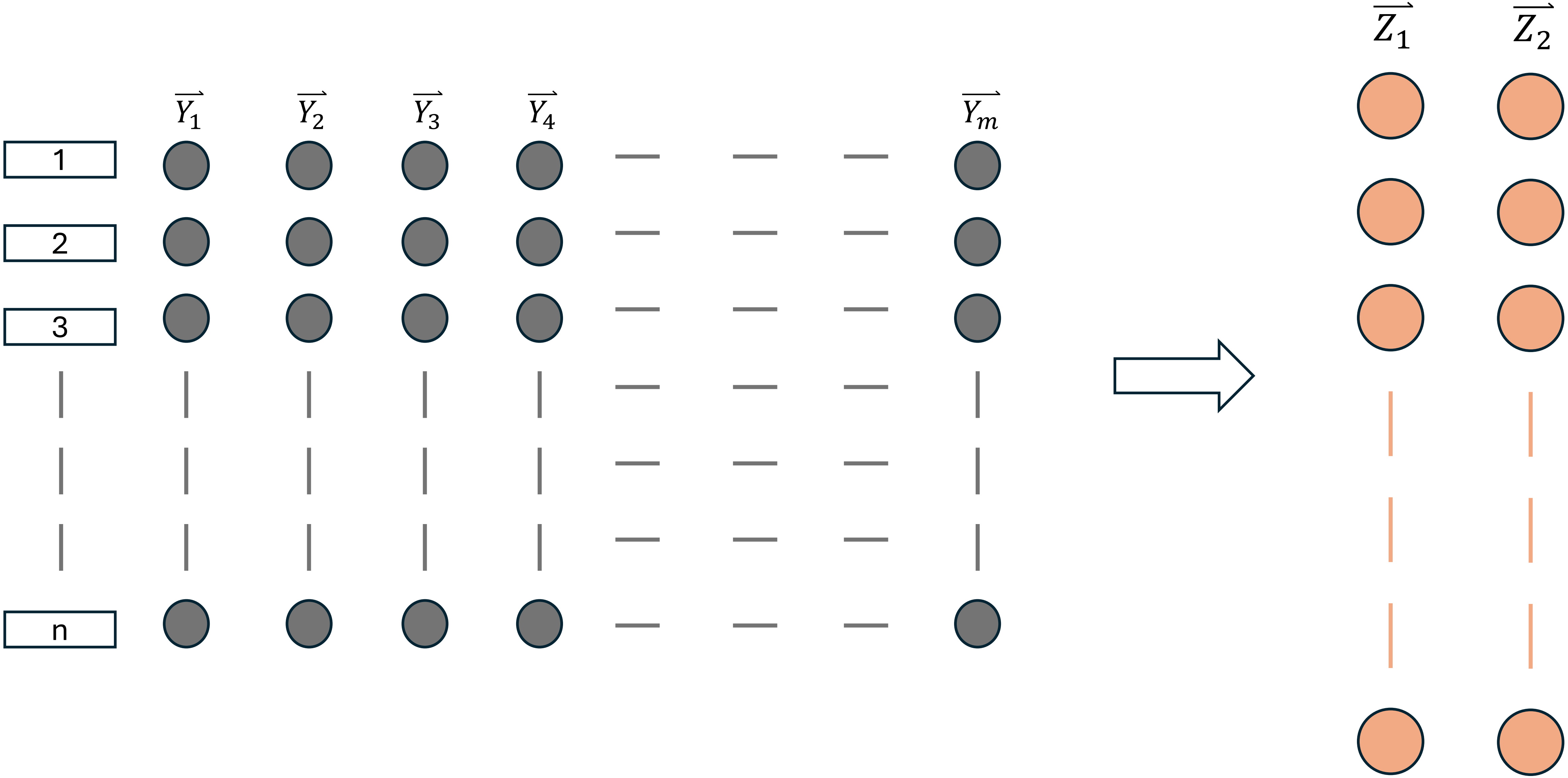
Multivariate statistics

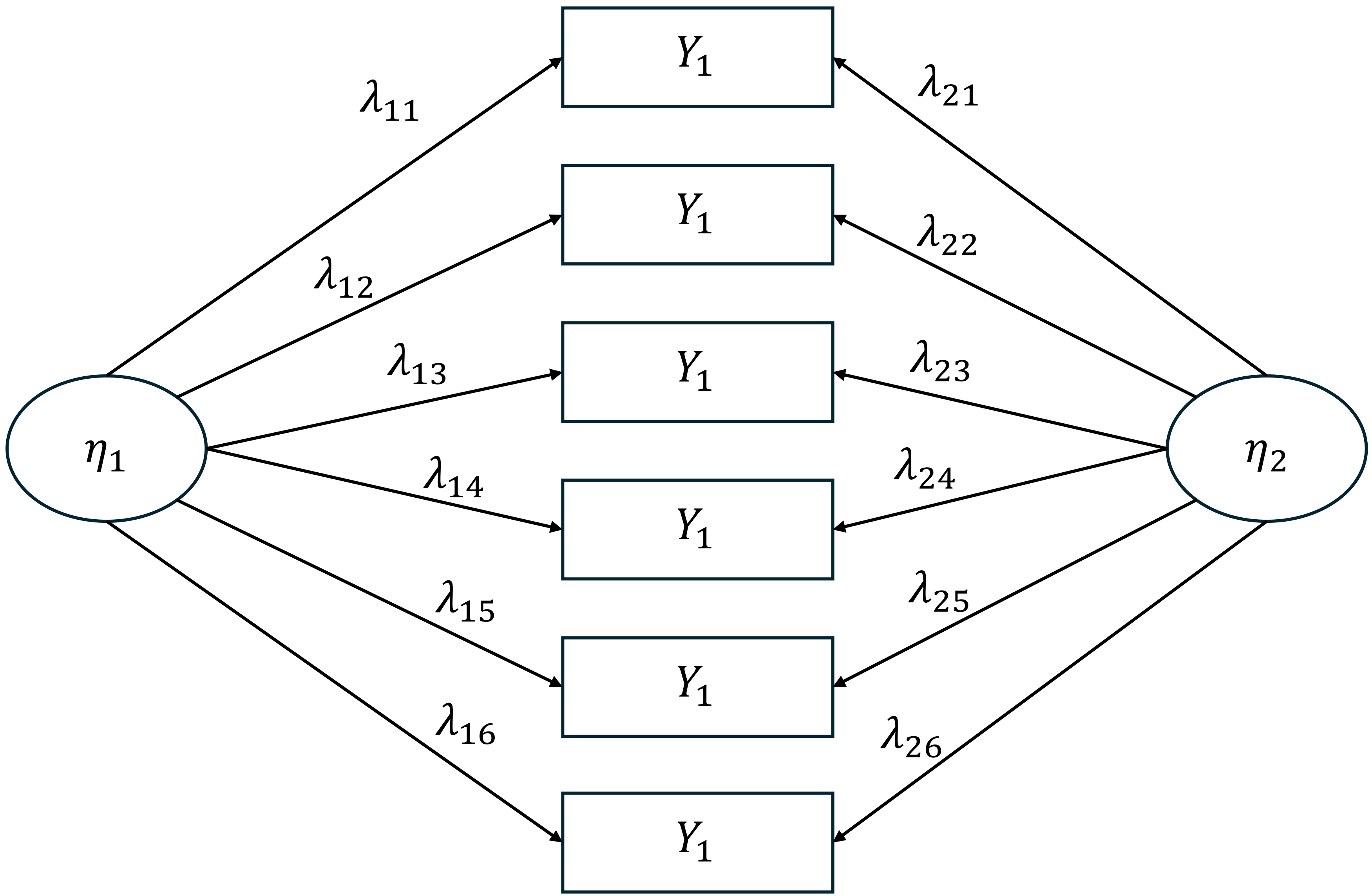
Exploratory Factor Analysis

May 25, 2024

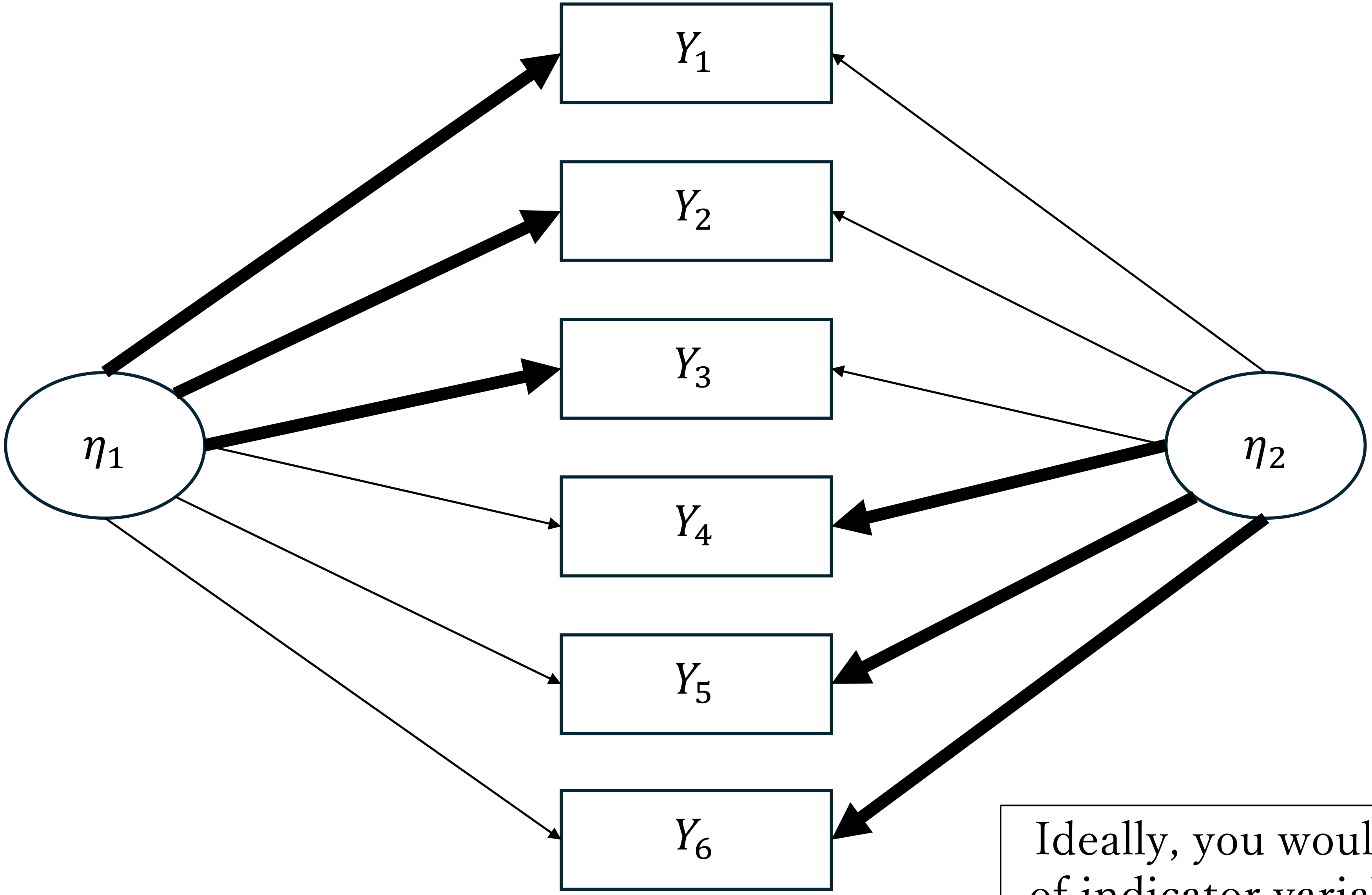
Orlando Sabogal-Cardona
@Antonio Sabogal
orlando.sabogal.20@ucl.ac.uk

General idea





Y_1
Y_1
Y_1
Y_1
Y_1
Y_1
η_1
η_2
λ_{11}
λ_{12}
λ_{13}
λ_{14}
λ_{15}
λ_{16}
λ_{21}
λ_{22}
λ_{23}
λ_{24}
λ_{25}
λ_{26}



Ideally, you would prefer groups of indicator variables with strong loadings for specific factors

EFA

The common factor model

$$Y = \Lambda\xi + \varepsilon$$

Used to predict the covariance/correlation matrix

$$\boxed{\Sigma = \Lambda\Psi\Lambda' + \Theta}$$

$Y: Matrix\ of\ observed\ indicator\ variables$

$\xi: Matrix\ of\ factors$

$\Lambda: Factor\ loading\ matrix$

$\varepsilon: Matrix\ of\ unique\ factors\ (source\ in\ variance\ not\ associated\ with\ \varepsilon)$

$n: number\ of\ observations$
$p: number\ of\ observed\ variables$
$Y: n * p$
$\Lambda: p * m$
$\xi: n * m$
$\varepsilon: n * p$

$S: Observed\ covariance\ or\ correlation\ matrix$

$\Sigma: Model\ correlation\ or\ covariance\ matrix$

$\Psi: Correlation\ matrix\ of\ the\ factors$

$\Theta: Diagonal\ matrix\ of\ unique\ error\ variances$

$m: number\ of\ factors$
$\mathrm{S}: p * p$
$\Sigma: p * p$
$\Psi: m * m$
$\Theta: p * p$

Make "Σ" *as similar to* "S" as possible

Four steps

- Number of factors to extract
- Rotation
- Estimation method
- Matrix
- Additional: factor scores

Factor extraction

- How many factors?
- There are several approaches based on “statistical criterion”.
- In practice you regularly use a few
- Usually not interpretable. Therefore, we need rotation.

Factor selection

- Probably the most important decision.
- Underfactoring or overfactoring compromise model validity and estimates. (Better to overfactor than to underfactor)

Factor selection: eigenvalues

- Unreduced correlation matrix: correlation matrix with units (1s) in the diagonal
- Reduced correlation matrix: correlation matrix with communalities estimates in the diagonal.
- The unreduced or reduced version are used when ML estimation is not used.

Kaiser-Guttman rule (Kaiser criterion)

- Compute eigenvalues from the correlation matrix
- Select eigenvalues greater than one.

- Question for you: what is the logic behind the Kaiser-Guttman rule?

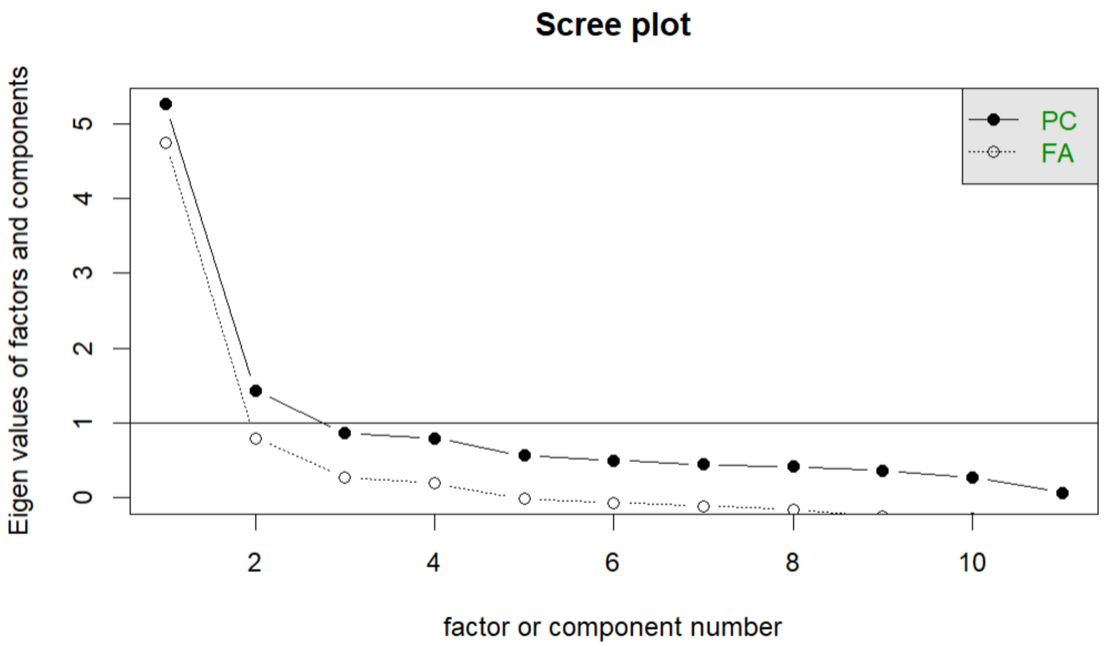
- This approach is often criticized.
- Another name: Latent Root

MAESTRÍA EN INVESTIGACIÓN OPERATIVA Y ESTADÍSTICA

Universidad Tecnológica de Pereira

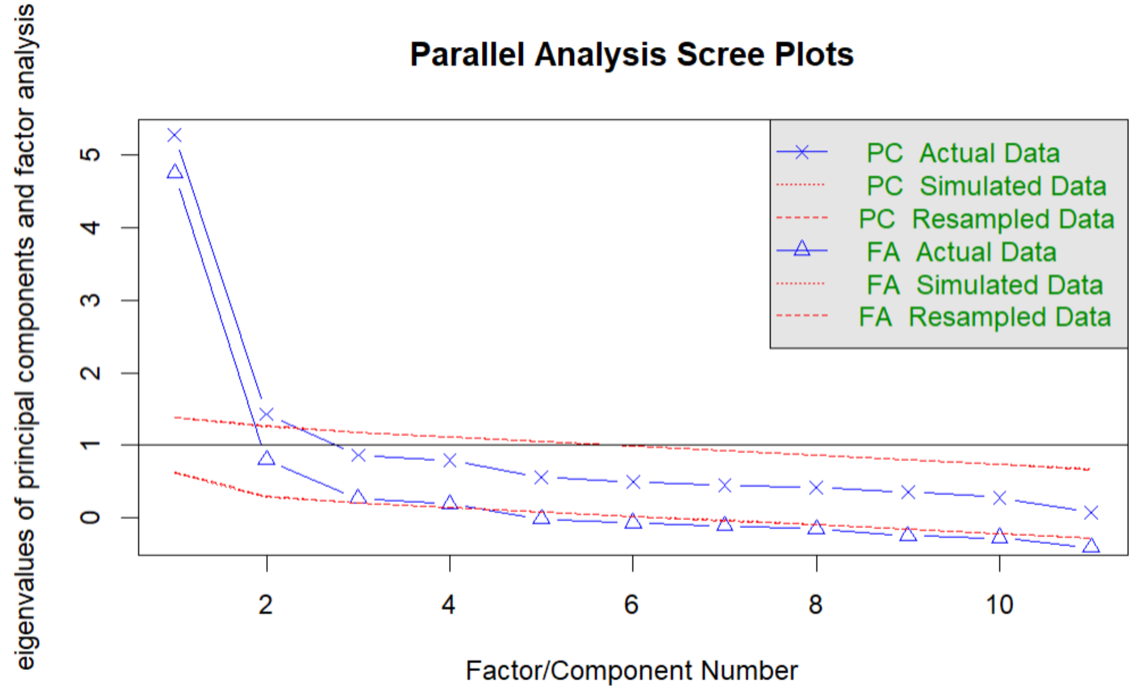
Scree test

- Input or reduced correlation matrix (the latter might be better)
- Graph: eigenvalues on Y and factors on X
- What happens if there is no change in the slope?



Parallel Analysis

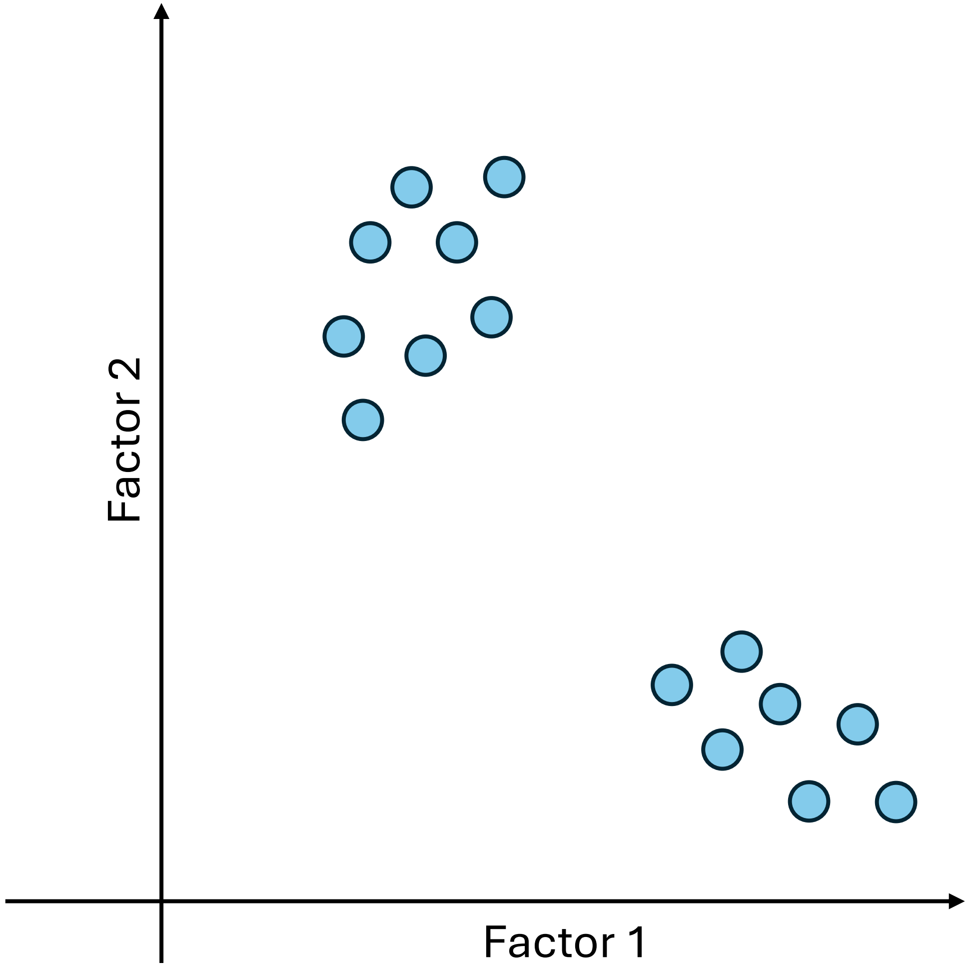
- Eigenvalues of sample data vs eigenvalues estimated from random numbers (mean of multiple batches)
- Look for the point where the two lines cross.
- Question for you: what is the logic behind this?



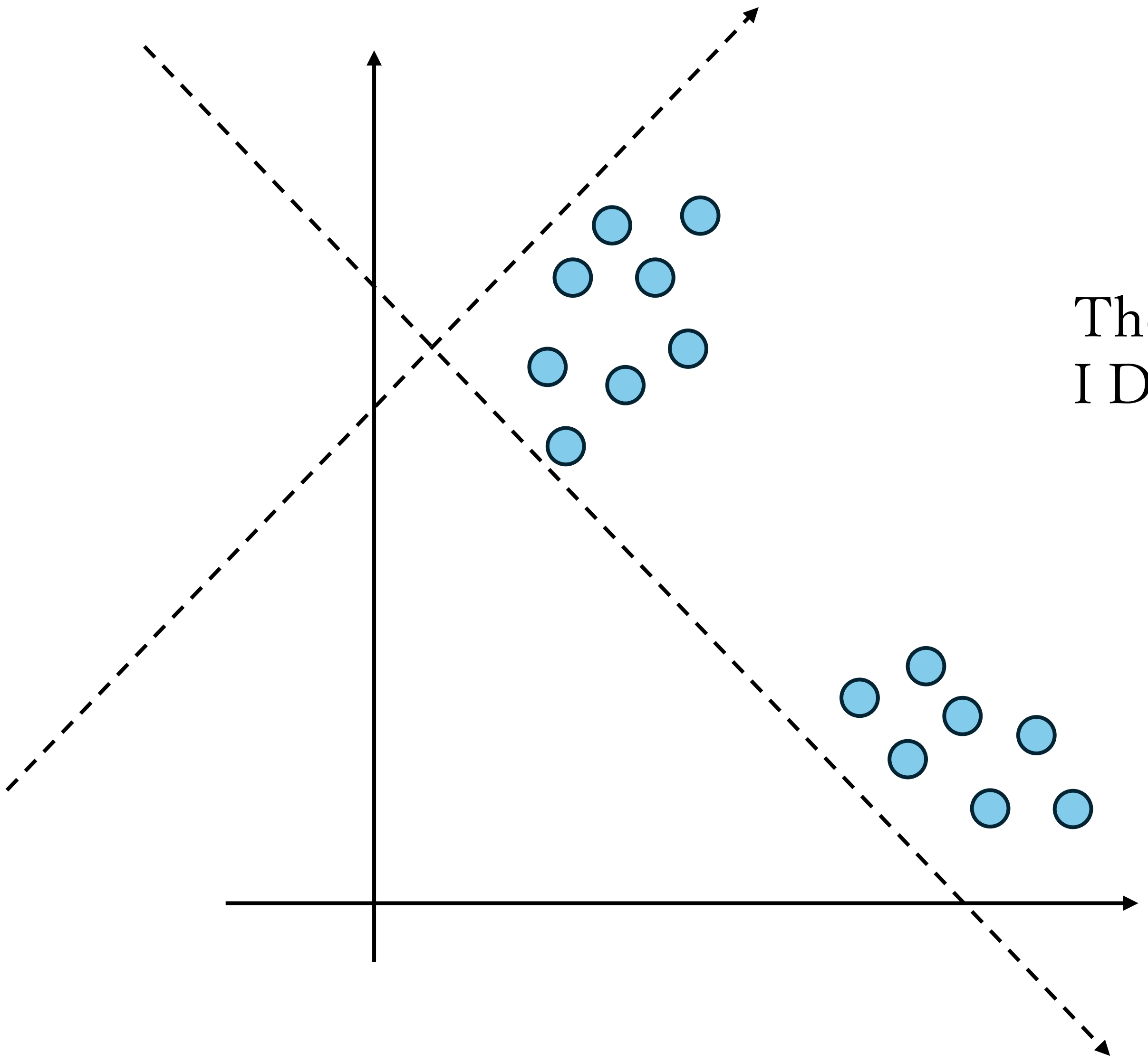
Factor selection

- Final decision should be guided by substance, theory, and interpretability. Do not limit to mathematical/statistical criteria.

Factor rotation

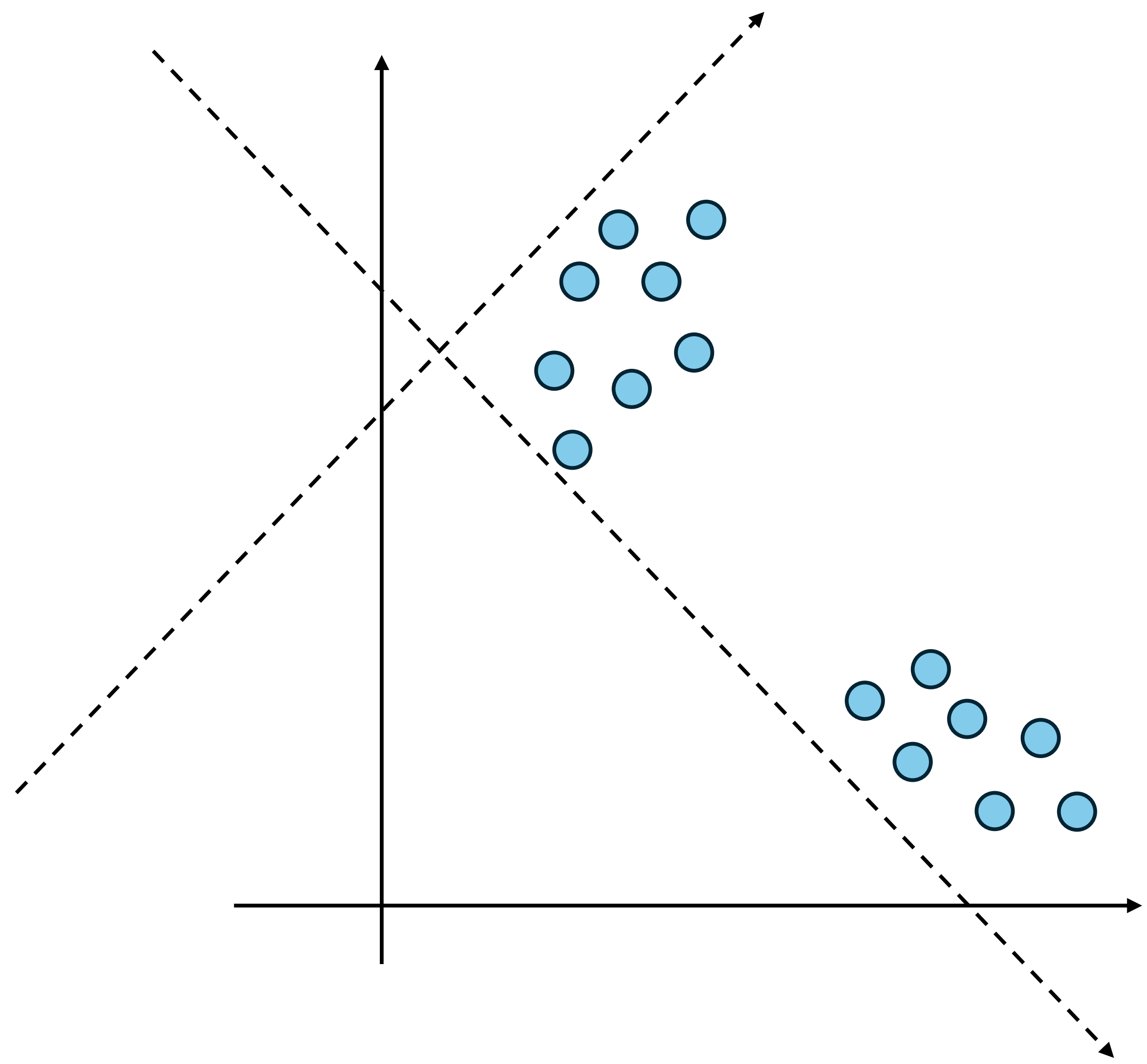


Factor rotation (orthogonally)

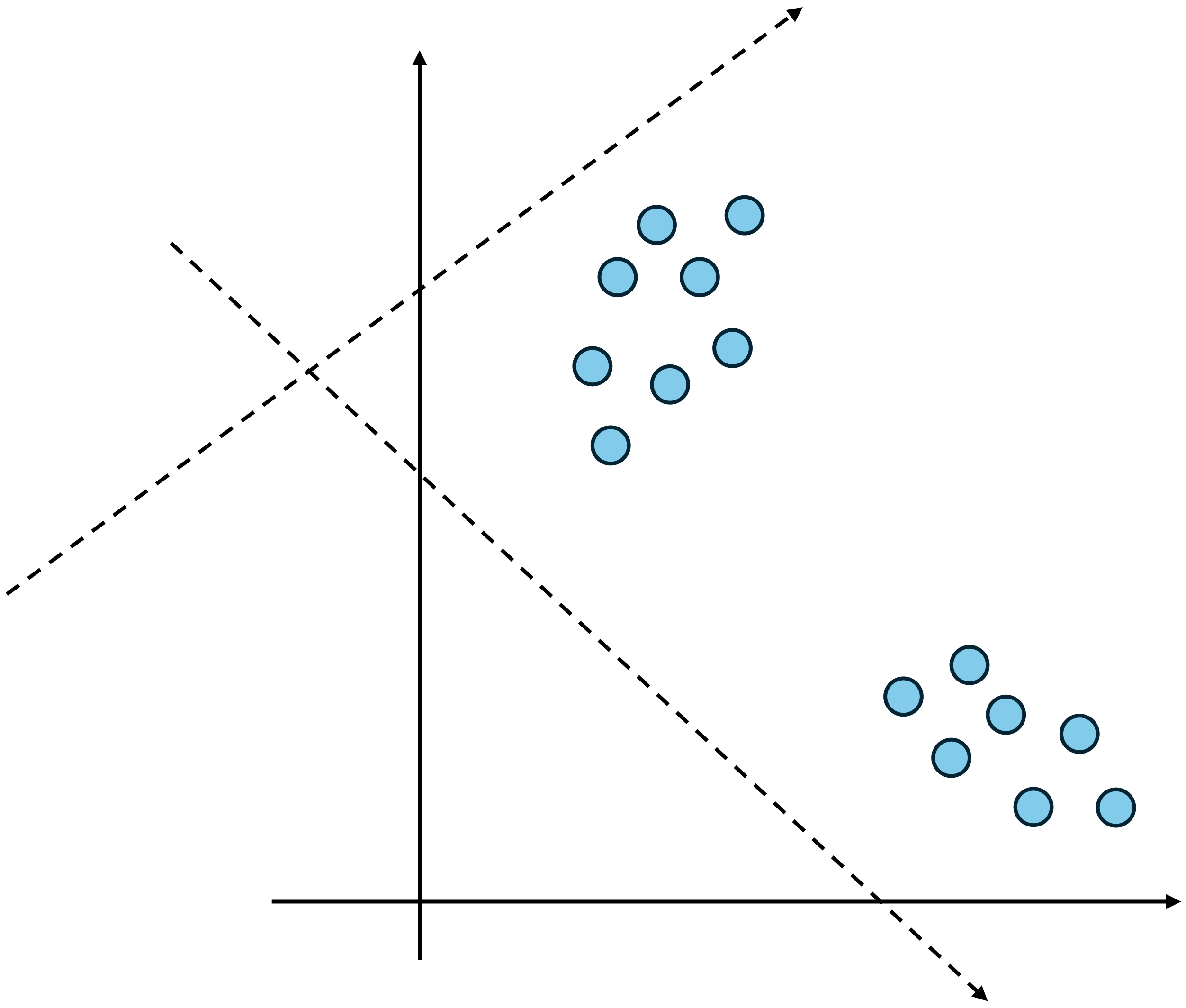


There is a debate.
I DO NOT RECOMMEND IT.

Factor rotation (orthogonally)



Factor rotation (oblique)



Factor rotation

You have options

`rotate` "none", "varimax", "quartimax", "bentlerT", "equamax", "varimin", "geominT" and "bifactor" are orthogonal rotations. "Promax", "promax", "oblimin", "simplimax", "bentlerQ, "geominQ" and "biquartimin" and "cluster" are possible oblique transformations of the solution. The default is to do a oblimin transformation, although versions prior to 2009 defaulted to varimax. SPSS seems to do a Kaiser normalization before doing Promax, this is done here by the call to "promax" which does the normalization before calling Promax in GPArotation.

Source: fa() function from the psych library

Estimation method

Maximum likelihood estimation:

- For continuous indicators
- Multivariate normality of indicators
- Sensitive to sample size (it might not converge)
- Allows for GoF
- With large sample, GoF measures might be sensitive to misspecifications
- Might produce improper solutions

Principal Axis Factoring PAF

- Does not require multivariate normality
- Based on OLS

Others: WLS, unweighted LS, GLS, etc.

Matrix

Correlation matrix:

- Standardized
- We do not worry about issues of scale
- It is often preferable

Covariance matrix:

- Not standardized
- Reflect variability in the data
- It also reflects differences in measurement scales

Matrix (categorical items)

cor	How to find the correlations: "cor" is Pearson", "cov" is covariance, "tet" is tetrachoric, "poly" is polychoric, "mixed" uses mixed cor for a mixture of tetrachorics, polychorics, Pearsons, biserials, and polyserials, Yuleb is Yulebonett, Yuleq and YuleY are the obvious Yule coefficients as appropriate

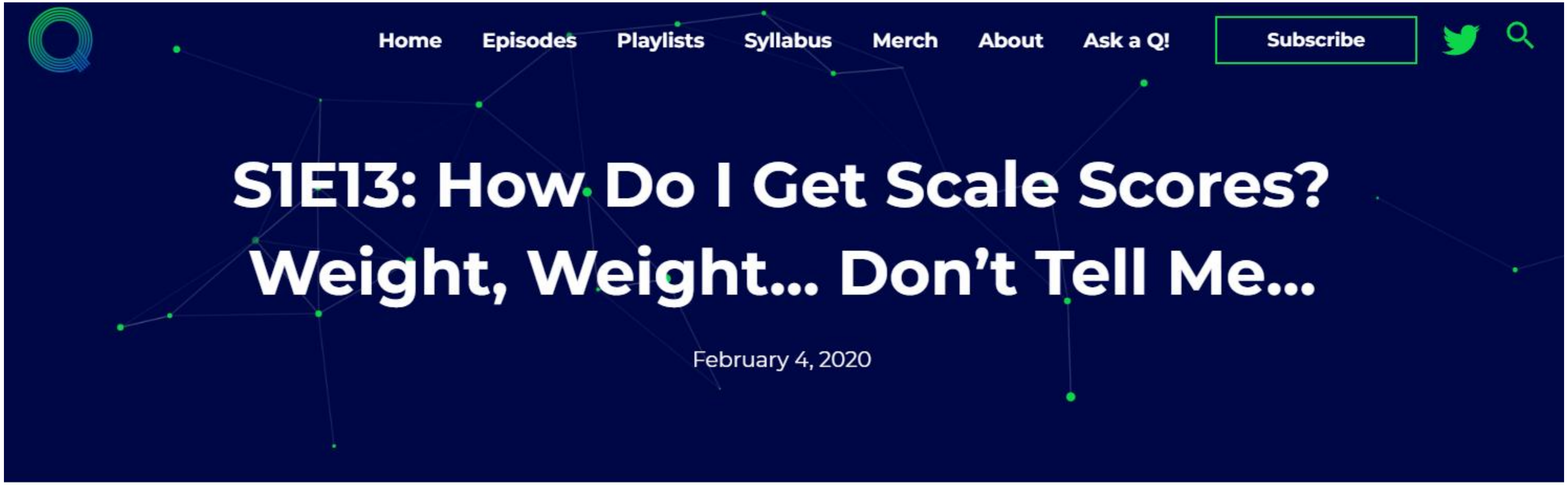
Source: fa() function from the psych library

Factor scores

Regression scores (Thurstone)
Bartlett scores
Anderson-Rubin scores
Others…

Recommended reading:

DiStefano, C., Zhu, M., & Mindrila, D. (2019). Understanding and using factor scores: Considerations for the applied researcher. *Practical assessment, research, and evaluation, 14*(1), 20.



https://quantitudepod.org/episode-13-how-do-i-get-scale-scores-weight-weight-dont-tell-me/

Recommended readings

- McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological methods*, *23*(3), 412.
- Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2019). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical assessment, research, and evaluation*, *17*(1), 3.
- Lee, C. T., Zhang, G., & Edwards, M. C. (2012). Ordinary least squares estimation of parameters in exploratory factor analysis with ordinal data. *Multivariate Behavioral Research*, *47*(2), 314-339.

- Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding statistics: Statistical issues in psychology, education, and the social sciences*, *2*(1), 13-43.
- Howard, M. C. (2016). A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve?. *International journal of human-computer interaction*, *32*(1), 51-62.

Go to the tutorial!

Latent variables: are you really measuring what you are supposed to measure? Are you using the right indicators?

Scale development:

It requires theoretical and empirical foundations.

- Content (face) validity: subjective (expert?) evaluation of correspondence between items and factors.
- Construct validity: degree to which a measure accurately represents what it is supposed to.
- Construct reliability: degree to which the observed variable measures the "true" value. Consistency between multiple measures.
- Internal consistency: Several items should measure the same latent variable.

Construct validity:

- Convergent validity: Correlation between measures of the same concept (expected to be high)
- Discriminant validity: Differences between two similar concepts
- Nomological validity: Predictions of other concepts

"In summary, convergent validity confirms that the scale is correlated with other known measures of the concept; discriminant validity ensures that the scale is sufficiently different from other similar concepts to be distinct; and nomological validity determines whether the scale demonstrates the relationships shown to exist based on theory or prior research." Hair

Items factorability?

Are items intercorrelated enough to "produce" the factors?
Measures of intercorrelation:

- Visual inspection
- Bartlett test of sphericity: applied to the correlation matrix. Indicates significant correlation among some of the variables. Sensitive to (large) sample size.
- Measure of sampling adequacy MSA: Index from 0 to 1 (minimum required: 0.5?). Sensitive to sample size and number of variables.

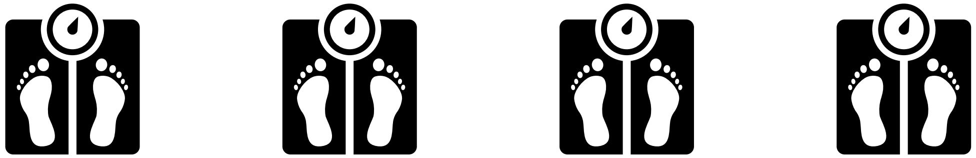
Some notes on reliability

A score has two components: true and error. $X = T + e$

Reliable: when the true component accounts for more variance than the error component.

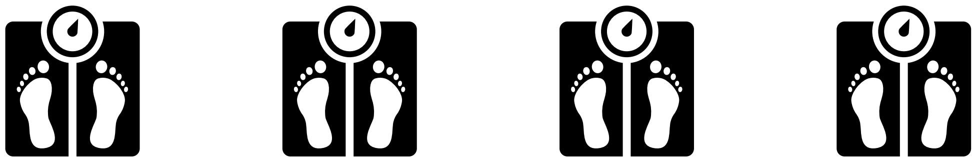
Think of it as a ratio.

Reliability can be tested by administering consecutive surveys and measuring the correlation. (probably unfeasible)



73 Kg | 72.8 Kg | 73 Kg | 73.1 Kg

Reliable →



73 Kg | 72.8 Kg | 73 Kg | 73.1 Kg

Reliable →

73 Kg | 72.8 Kg | 73 Kg | 73.1 Kg

73 Kg | 60.5 Kg | 80.1 Kg | 77.4 Kg

← Not Reliable

MAESTRÍA EN INVESTIGACIÓN OPERATIVA Y ESTADÍSTICA

Universidad Tecnológica de Pereira

X_1	X_2	X_3	X_5
73 Kg	72.8 Kg	73 Kg	73.1 Kg
73 Kg	60.5 Kg	80.1 Kg	77.4 Kg
LV score	LV score	LV score	LV score

Internal consistency reliability explores the relation of each item to the others.
Does respondents provide similar answer across questions?

Do you see the Tau-equivalence? Do you see why it is so relevant?
Do you see why errors are assumed to be random?
Do you see unidimensionality?

Recommended reading:

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. Psychological Methods, 23(3), 412–433. https://doi.org/10.1037/met0000144

Cronbach's alpha

Reliability coefficient: internal consistency.
It tells you how well a set of items work together as a group.
1: Perfect internal consistency; 0: No internal consistency

For binary data, Kuder-Richardson (K-R) coefficient of equivalence

Cronbach's alpha

$$\alpha = \frac{N * \bar{c}}{\bar{v} + (N - 1) * \bar{c}}$$

$N: Number\ of\ variables$

$\bar{c}: Average\ covariance\ between\ variable\ pairs$

$\bar{v}: Average\ variance\ of\ variables$

Cronbach's alpha

$$\alpha = \frac{N * \bar{c}}{\bar{v} + (N - 1) * \bar{c}}$$

$N: Number\ of\ variables$

$\bar{c}: Average\ covariance\ between\ variable\ pairs$

$\bar{v}: Average\ variance\ of\ variables$

X_1	X_2	X_3	X_4
3			
2	4		
3	3	4	
1	2	3	5

X_1	X_2	X_3	X_4
3			
2	4		
3	3	4	
1	2	3	5

Cronbach's alpha (standardized)

$$\alpha_{std} = \frac{N * \bar{r}}{\bar{1} + (N - 1) * \bar{r}}$$

N: $Number\ of\ variables$

$\bar{r}$: $Average\ intervariable\ correlation$

$$\bar{r} = \frac{\sum_i^N \sum_j^N r_{ij}}{N(N - 1)/2} \quad \forall\, i \neq j$$

- When items are measured on different scales.
- Enables comparisons among studies.
- Similar interpretation.

Cronbach's alpha

Do not forget it: it is a reliability coefficient

It is based on the correlation matrix. Therefore, continuity in the variables is assumed.

You can get a polychoric alpha or a thetrachoric alpa.

McDonald's Omega

- Generally, more robust and more accurate.
- Does not require Tau-Equivalence assumption (items have the same true score variance).

$$\omega_t = \frac{\left(\sum_{i=1}^{k} \lambda_i\right)^2}{\left(\sum_{i=1}^{k} \lambda_i\right)^2 + \sum_{i=1}^{k} \psi_i + \sum_{i=1}^{k} \theta_{ij}}$$

$\lambda_i : Factor\ loadings$

$\psi_i : Unique\ Variances$

$\theta_{ij} : Error\ Variances$

Kaiser-Meyer-Olkin KMO

- Test for adequacy: suitability of the data for factor analysis.
- It tells you how well items explain each other.
- Estimates the proportion of variance that might be common variance.
- Ranges from 0 to 1.

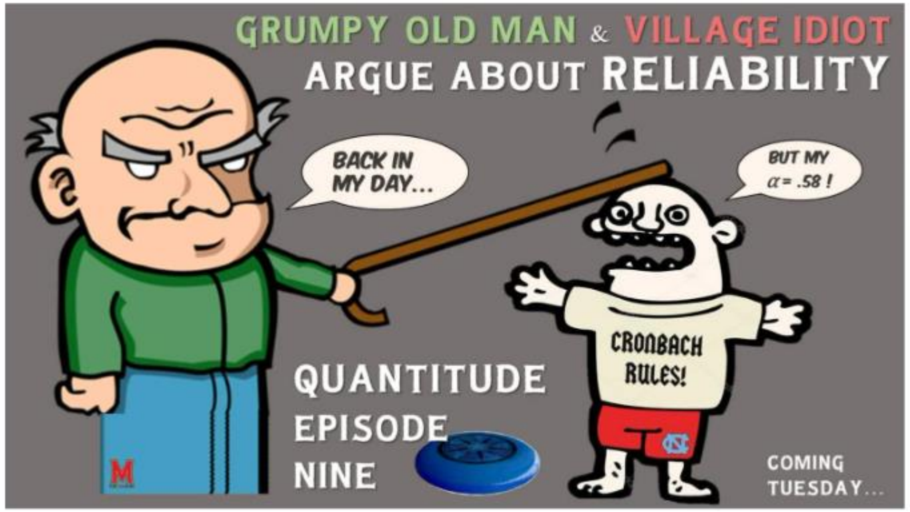
$$KMO = \frac{\sum_{i \neq j} r^2{}_{ij}}{\sum_{i \neq j} r^2{}_{ij} + \sum_{i \neq j} p^2{}_{ij}}$$

r : *correlation*

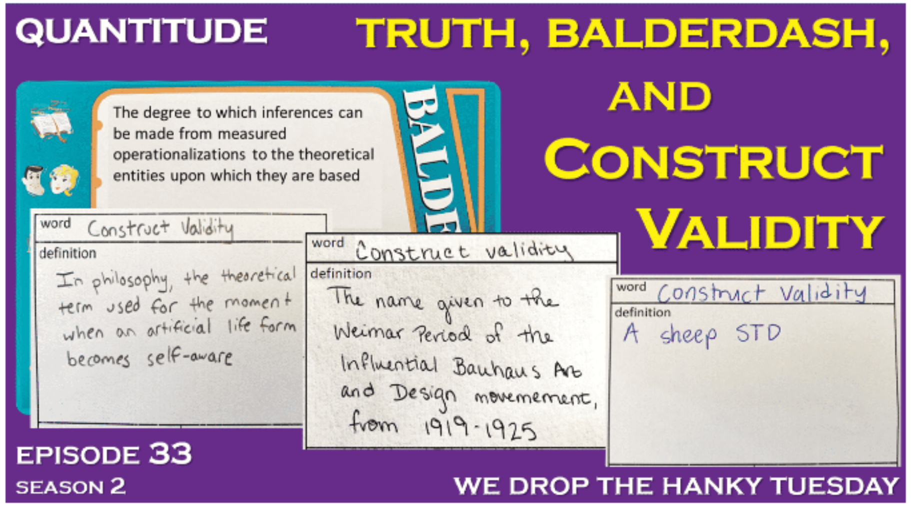
p : *partial correlation*

Bartlett's test of sphericity

Test the presence of "statistically significant" correlations. Null hypothesis: correlation matrix does not have correlations. You want p value < 0.05.



https://quantitudepod.org/episode-9-grumpy-old-man-village-idiot-argue-about-reliability/



https://quantitudepod.org/s2e33-truth-balderdash-and-construct-validity/

MAESTRÍA EN INVESTIGACIÓN OPERATIVA Y ESTADÍSTICA

Universidad Tecnológica de Pereira

Go to the tutorial!

MAESTRÍA EN INVESTIGACIÓN OPERATIVA Y ESTADÍSTICA

Universidad Tecnológica de Pereira

Thank you!

Orlando Sabogal-Cardona

@Antonio Sabogal

orlando.sabogal.20@ucl.ac.uk