Exploratory Factor Analysis and Principal Component Analysis

- Topics:
  - What are EFA and PCA for?
  - Analysis steps:
    - Which extraction method? (and conceptual differences)
    - How many factors?
    - Which rotation → which interpretation?
    - (Don't) generate factor scores

Big Picture of Instrument Development

- Primary concerns about the use of an instrument to measure one or more latent traits have a hierarchical structure:
  - **Validity**: Extent to which an instrument measures what it is supposed to
    - Validity is always a matter of degree and depends critically on how it is used
    - Almost always demonstrated by **external evidence**: relationships to measures of other constructs in expected directions (e.g., discriminant and convergent validity)
  - An essential **precursor** to validity is **reliability**: Extent to which an instrument measures a latent trait with **sufficient consistency** (i.e., extent to which the same result would be obtained repeatedly)
    - "Validity is measuring the right thing; reliability is measuring the thing right"
    - Reliability indices will be provided differently across LTMMs (stay tuned)
  - An important **precursor** to reliability is **dimensionality**: Accuracy of the mapping of the observed indicators to the latent traits they measure
    - Reliability is per trait! Most reliability indices assume **unidimensional traits**
    - So obtaining evidence for the **expected trait dimensionality** of the chosen indicators must come first… how to do so will be a big focus this semester

Stopping by EFA and PCA on the Way

- This course is dedicated to latent trait **measurement models**…
  - Confirmatory factor models (≈ linear factor models), item response models (≈ nonlinear factor models), and others can be used to provide **evidence for expected trait dimensionality** (and detect problems)

- Today we'll visit EFA and PCA to describe how these devices are similar to and different than confirmatory factor models
  - I'm hitting the major points only—it's honestly not worth learning more, because these techniques (EFA especially) are generally pretty useless
    - Supplemental material will be provided after CFA for those who want more
  - The results from exploratory factor analyses can be misleading:
    - If outcomes do not meet assumptions of model or method (e.g., non-normality)
    - If the constraints made in the analysis make no sense (i.e., the definition of EFA)
    - Because there is no empirical basis for saying which solution is more "correct"

- My thesis: it is not your data's job to tell you what it measures!
  - You should at least have a **clue**, even if you don't have the right answer

EFA vs. PCA

- 2 very different schools of thought on **exploratory factor analysis** (**EFA**) vs. **principal components analysis** (**PCA**):
  1. EFA and PCA are TWO ENTIRELY DIFFERENT THINGS… How dare you even put them into the same sentence!
  2. PCA is a special kind (or extraction type) of EFA… although they serve different purposes, the results are often the same anyway, so what's the big deal?

- My world view:
  - I'll describe them more along with school of thought #2. I want you to know what their (severe) limitations are. I want you to know that **they are not testable models**.
  - **It is not your data's job to tell you what constructs you are measuring!! If you don't have any idea at all, game over.**

Primary Purposes of EFA and PCA

- **EFA**: "Determine the nature of and the number of latent variables that account for observed variation and covariation among set of observed indicators (≈ items or variables)"
  - In other words, **what traits cause these observed responses**?
  - Factors (traits) evidenced by patterns of correlation among indicators
  - If the indicators are not correlated in the first place, game over!
  - Getting a "solution" (mapping of indicators to traits) is the point

- **PCA**: "Reduce multiple observed variables into fewer components that summarize their variance"
  - In other words, how can I **abbreviate** this set of observed variables?
  - Correlation among indicators is unnecessary (is not the focus)
  - Solution is a means to an end—get "**components**" to use elsewhere

Planning a Factor Analytic Study (from Tabachnick and Fidell)

- **Hypothesize the number of factors** you are trying to measure (5-6 factors is recommended for a stable solution)

- Get **5-6 good indicators** (items or variables) *per factor*
  - At least some should be 'marker indicators', such that you know a priori which factor each indicator should be related to
  - Avoid multidimensional indicators (that measure >1 factors)
  - Watch out for 'outlier indicators'—if an indicator is not related to the others, it will not be part of a useful factor solution
  - Older programs (e.g., SAS and SPSS) assume multivariate normality of the indicators, although Mplus allows EFA for other responses

- Get a **big enough sample** with sufficient trait variability
  - But the much-cited "at least 5 people per indicator" has been shown to be inadequate: it depends far more on the commonality of the items

Steps in EFA (and PCA)

1. **Choose an estimator/extraction method**
2. Determine number of factors
3. Select a rotation
4. Interpret solution (may need to repeat steps 2 and 3)
5. (Don't) generate factor scores

Extraction Methods
(school of thought #1, please don't yell at me)

- The Question: How many factors do I need to reproduce the observed correlation matrix among the indicators?
  - **But 'which' correlation matrix are we starting from???**

- Primary difference between PCA and EFA:
  - **PCA**: Analyze **ALL** the variance in the indicators
    - On the diagonal of the analyzed correlation matrix are 1's
  - **EFA**: Analyze **COMMON** variance (covariance) in the indicators
    - On the diagonal of the correlation matrix are essentially the R^2 for each indicator when predicted by all the other indicators
    - These R^2 values are called "**commonalities**" (labeled H^2)
    - So the leftover **non-common variance** (which we'll eventually call "error variance") **gets dropped prior to analysis**

Extraction Methods: PCA

- PCA: Extracts # COMPONENTS = # indicators
  - Will perfectly reproduce original correlation matrix
  - Is unique mathematical solution (via the magic of matrix algebra)
  - Components are uncorrelated (or "orthogonal")
  - Extracted in order of most variance accounted for of the indicators
  - Provides component **loadings** (the L's) that relate the observed **indicators** (the I's) to the extracted **components** (the C's)

- Example with 5 indicators:
  - $C_1 = L_{11}I_1 + L_{12}I_2 + L_{13}I_3 + L_{14}I_4 + L_{15}I_5$
  - $C_2 = L_{21}I_1 + L_{22}I_2 + L_{23}I_3 + L_{24}I_4 + L_{25}I_5$
  - $C_3 = L_{31}I_1 + L_{32}I_2 + L_{33}I_3 + L_{34}I_4 + L_{35}I_5$
  - $C_4 = L_{41}I_1 + L_{42}I_2 + L_{43}I_3 + L_{44}I_4 + L_{45}I_5$
  - $C_5 = L_{51}I_1 + L_{52}I_2 + L_{53}I_3 + L_{54}I_4 + L_{55}I_5$

Keep **all** possible components?
= "Full" Component Solution
(which has no point)

Keep **fewer** components?
= "Truncated" Component Solution
(which is usually the case)

PCA, continued

- Consider this correlation matrix
- There appears to be 2 kinds of information in these 4 indicators
  - I_1 **and** I_2 I_3 **and** I_4

	I_1	I_2	I_3	I_4
I_1	1.0			
I_2	.7	1.0		
I_3	.3	.3	1.0	
I_4	.3	.3	.5	1.0

- Looks like the PCs should be formed as
  - $C_1 = L_{11}I_1 + L_{12}I_2$ → capturing the information in I_1 **and** I_2
  - $C_2 = L_{23}I_3 + L_{24}I_4$ → capturing the information in I_3 **and** I_4

- But PCA doesn't "group indicators"—it "reproduces variance"
  - Note the cross-correlations among these kinds of indicators…

PCA, continued

- So, because of the cross correlations, to **maximize the variance reproduced**, C_1 will be formed as:

$$C_1 = .5I_1 + .5I_2 + .4I_3 + .4I_4$$

  - Each contributes; higher loadings for I_1 **and** I_2

	I_1	I_2	I_3	I_4
I_1	1.0			
I_2	.7	1.0		
I_3	.3	.3	1.0	
I_4	.3	.3	.5	1.0

- Because C_1 didn't focus on the I_1 **and** I_2 indicator group OR on the I_3 **and** I_4 indicator group, there is still variance to account for in both, and so C_2 will be formed as:

$$C_2 = .3I_1 + .3I_2 - .4I_3 - .4I_4$$

  - Each contributes; now higher loadings for I_3 **and** I_4

- PCA maximizes variance accounted for; it does not find groups of indicators that measure the same latent trait
  - How would you interpret these components if used subsequently???
  - The less correlated the indicators, the harder to interpret components

PCA: More on Component Matrix

	C_1	C_2
I_1	.8	−.2
I_2	.7	−.1
I_3	.2	.5
I_4	.2	.4

- Row = indicators, column = component, Value = correlation for indicator with component
- If you square and sum the values in a column, you get the **Eigenvalue** for a component
- Eigenvalue for $\boldsymbol{C_1} = .8^2 + .7^2 + .2^2 + .2^2 = 1.21$
- Variance accounted for across indicators by that component → eigenvalue / # indicators
- For $\boldsymbol{C_1} = 1.21 / 4 = .3025$ or 30.25%

- If you square and sum across the values in a row, you get that indicator's extracted communality (i.e., called "H^2", like R^2 in regression)
- $\boldsymbol{H^2}$ for I_1 → $.8^2 + -.2^2 = .68$ or 68% of its variance
  - This won't work unless the components stay orthogonal (uncorrelated)

- Same exact logic and procedure applies to EFA, but they are called "Factor Matrices" instead (because "factors" instead of "components")

EFA: Extraction Methods

- PCA-based methods of "extraction" for EFA:
  - Known as "principal axes" or "principal factors" (default in SPSS!)
  - No model fit provided, but no multivariate normality required
  - Focused on finding communalities to maximize variance extracted → devices invented in absence of statistical computing resources!
    - Starts as H^2 from prediction by other indicators ("Initial")
    - Ends up with H^2 from prediction by all the factors ("Extraction")
    - Watch out for "Heywood cases" → $H^2 > 1$

- Alternative Extraction Method: Maximum Likelihood (ML)
  - Goal is to find most likely estimates for loadings and error variances
  - Provides model fit because uses same log-likelihood as CFA/SEM
    - Most programs require multivariate normality (there are other options in Mplus)
  - Start here if you must do EFA at all! (which you don't)

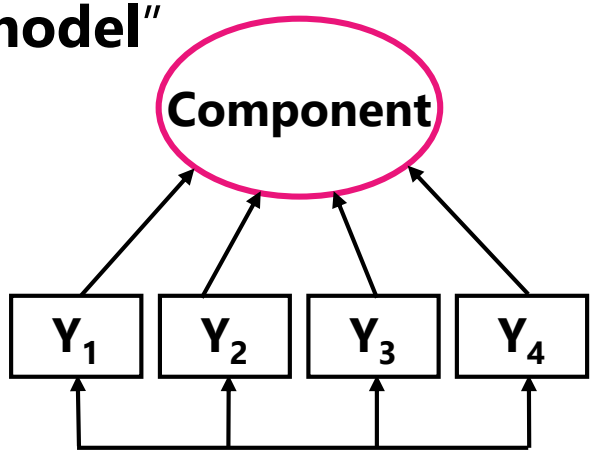
Big Conceptual Difference between PCA and EFA

- In **PCA**, the **component** is just the **weighted sum of the indicators**, and so any correlation between the indicators is essentially irrelevant
  - But if the indicators are not correlated, then how would you interpret the results for the component as a variable in subsequent analyses???
  - Type of trait measured by a component is sometimes called an "**emergent**" construct – i.e., it emerges from the indicators ("**formative**" model)
  - Examples: Socio-Economic Status, Lack of Free Time, Resources Available

- In **EFA**, the **indicator responses are thought to be caused by the factors**, and thus should be uncorrelated once controlling for the factor(s)
  - Type of trait measured by a factor is often called a '**reflective**' construct (i.e., the indicator responses are a reflection of your status on the latent trait)
  - Examples: Pretty much everything else…
  - Remember: Trait = factor = construct = true score → thing you are measuring

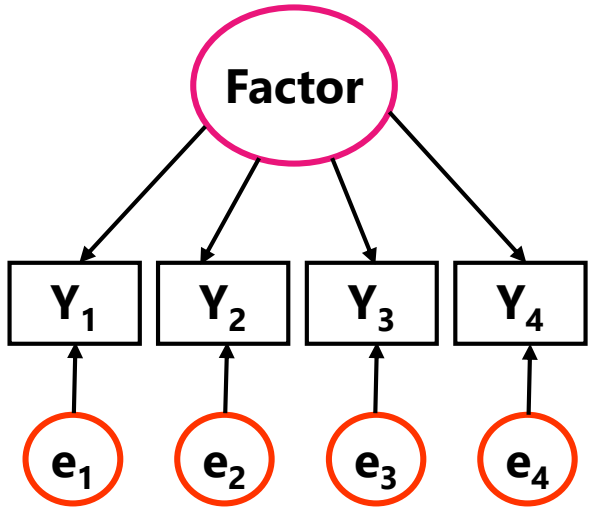
PCA vs. EFA/CFA

"**Formative model**"



This is NOT a testable measurement model, because how do we know if the indicators have been combined "correctly"?

"**Reflective model**"



This IS a testable measurement model, because it predicts the observed covariances between the indicators through the factor loadings (arrows)—**the factor IS the reason for the covariance.**

PCA vs. EFA/CFA, continued

- In **PCA**, **components are** built from **linear combinations** of the indicator responses (e.g., 5 indicators and 2 components):
  - $\boldsymbol{C_1} = L_{11}I_1 + L_{12}I_2 + L_{13}I_3 + L_{14}I_4 + L_{15}I_5$
  - $\boldsymbol{C_2} = L_{21}I_1 + L_{22}I_2 + L_{23}I_3 + L_{24}I_4 + L_{25}I_5$
  - Note that $\boldsymbol{C_1}$ and $\boldsymbol{C_2}$ are the **OUTCOMES → is a FORMATIVE model**
    - **This is NOT a testable measurement model by itself!**

- In **EFA**, **factors are thought to cause the indicator responses** in a **REFLECTIVE** model (e.g., 5 indicators and 2 factors):
  - $I_1 = L_{11}\boldsymbol{F_1} + L_{12}\boldsymbol{F_2} + e_1$
  - $I_2 = L_{21}\boldsymbol{F_1} + L_{22}\boldsymbol{F_2} + e_2$
  - $I_3 = L_{31}\boldsymbol{F_1} + L_{32}\boldsymbol{F_2} + e_3$
  - $I_4 = L_{41}\boldsymbol{F_1} + L_{42}\boldsymbol{F_2} + e_4$
  - $I_5 = L_{51}\boldsymbol{F_1} + L_{52}\boldsymbol{F_2} + e_5$
  - Note that F_1 and F_2 are the **PREDICTORS → testable reflective model**

Why Not to Use the EFA Framework

- In EFA, **all factors predict all indicators**, which is why there is a need to decide on a "good" solution (# factors, rotation)
  - Even if you suspect this is not the case, **you can't change it**!

- But this isn't possible mathematically without some kind of identification constraints, and the ones in EFA make no sense!
  - For example, for two factors: The sum across items of all squared loadings times the item's unique variance must be 0…. Huh?
  - These **constraints** are not testable and **are not interpretable**

- If you need to do "exploratory" factor analyses, do them in an **LTMM framework**, in which **you impose the constraints**
  - Fit of alternative models can be **compared empirically**
  - So you don't need "strong theory" to get started, **you just need a clue**! The results will help diagnose misfit (you can remedy)

Steps in EFA (and PCA)

1. Choose an estimator/extraction method
2. **Determine number of factors**
3. Select a rotation
4. Interpret solution (may need to repeat steps 2 and 3)
5. (Don't) generate factor scores

How many factors/components?

- In other words, "How many constructs am I measuring?"
  - Now do you see why the data shouldn't be telling you this?

- Rules about the number of factors or components needed are based on **Eigenvalues**:
  - Eigenvalues = how much of "total" variance in observed indicators is accounted for by each factor or component
    - In PCA, "total" is really the total possible variance
    - In EFA, "total" is just the total possible *common* variance

- 3 proposed methods
  - Kaiser-Guttman Rules (eigenvalues over 1)
  - Scree test (ok, "scree plot", really)
  - Parallel analysis (ok, "parallel plot", really)

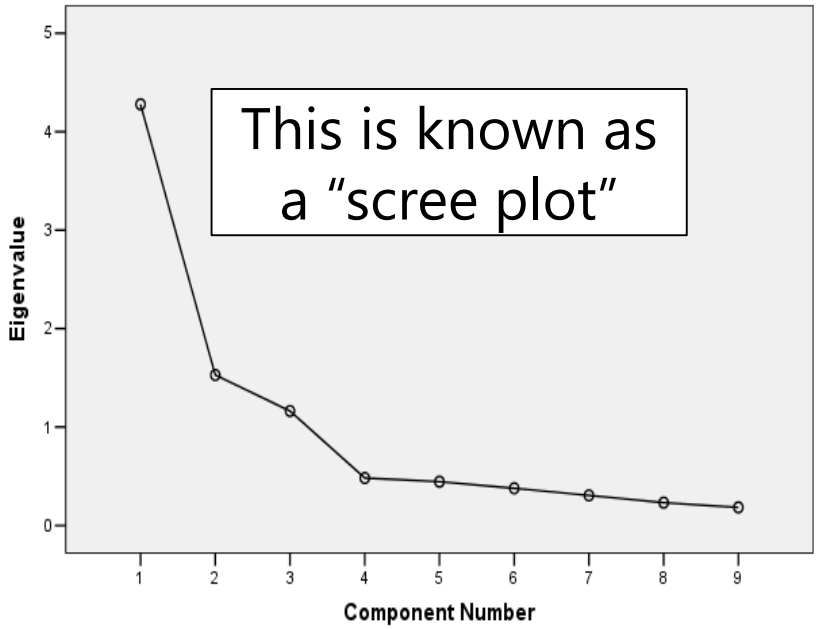
How many factors/components?

- Kaiser-Guttman Rule:
  - Keep any factors with Eigenvalues over 1
    - Supposed to be on non-reduced correlation matrix (i.e., the one with the 1's in the diagonal for all the variance, not just the common variance), but people use it for the reduced EFA corr matrices, too
  - Logic: Eigenvalues are amount of variance accounted for by the factor (where total variance = total # indicators)
    - At the bare minimum, the factor should account for as much variance as one of the original indicators did (i.e., its own variance)
    - Again, this logic only makes sense if you're talking about the non-reduced, total variance matrix… but this appears ambiguous
  - But whatever: Research suggests this rule doesn't work well, anyway… (and of course it is the default in many programs)

How many factors? Stare at a picture…

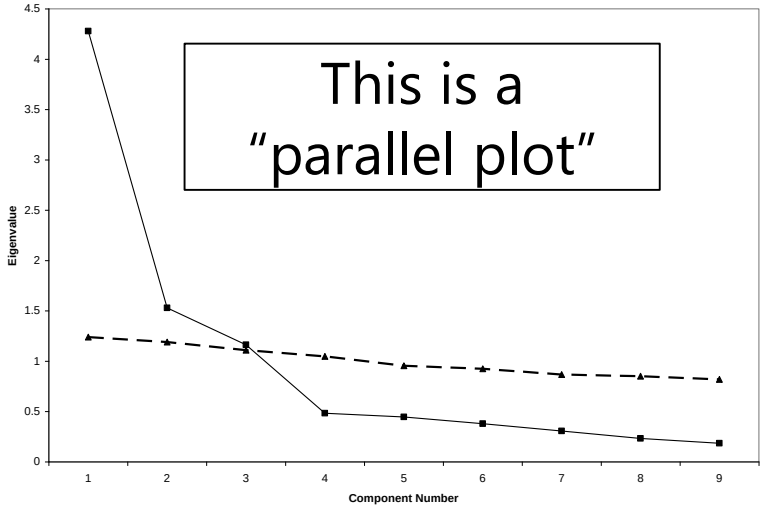
Scree "Test" → Scree plot

- Plot factor number on x-axis, and its Eigenvalue on y-axis
- Look for "break" where the slope changes, and retain the factors before that break
- Available in most programs



Parallel "Test" → Parallel plot

- Plot your Eigenvalues against those obtained from random simulated data using same study design
- Find point where real data crosses fake data; retain # factors above
- Not available in most programs
  - Available SAS code reference given in Brown (2015) chapter 3



Steps in EFA (and PCA)

1. Choose an estimator/extraction method
2. Determine number of factors/components
3. **Select a rotation**
4. Interpret solution (may need to repeat steps 2 and 3)
5. (Don't) generate factor scores

What is Rotation For?

- Although the component or factor matrix has the loadings of each indicator for each component or factor, those original loadings hardly ever get used directly to interpret the factors

- Instead, we often '**rotate**' the factor solution (as shown next)
- But alternative rotations result in equivalently-fitting, differently interpreted model solutions

- What this means is that factor loadings are NOT unique: for every solution there is an infinite number of possible sets of factor loadings, **each as 'right' as the next** (which is really dissatisfying in practice!)
  - e.g., competing EFA solutions across publications

Goal of Rotation: Simple Structure

- The idea of rotation is to re-arrange the factor loadings to obtain **simple factor structure**
  - Each factor should have indicators with strong loadings
    - Obvious which indicators measure it (+/-) and which don't
  - Each indicator should load strongly on only one factor
    - Know what each item is 'for'
    - Construct measured is readily identifiable
    - Indicators should have large communalities

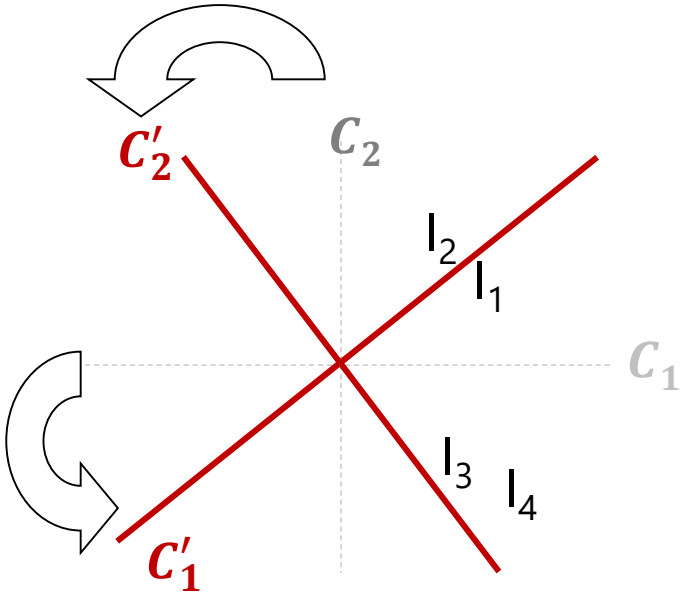
- Two kinds of rotations:
  - Orthogonal (uncorrelated factors—seriously??)
  - Oblique (correlation among factors in another matrix)

"Simple Structure" via Rotation

- Factor Rotations—changing the "viewing angle" of the factor space—are the major approach to providing simple structure
  - Goal is to get "simple structure" by getting the factor or component vectors to "spear" the indicator clusters

Un-rotated

	C_1	C_2
I_1	.7	.5
I_2	.6	.6
I_3	.6	-.5
I_4	.7	-.6



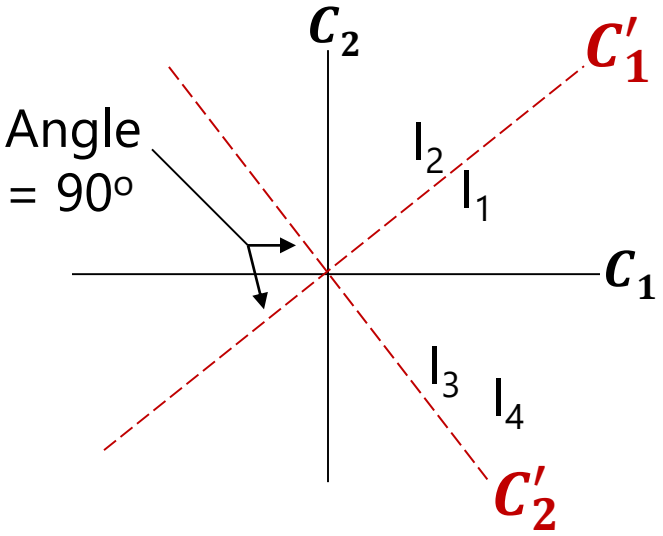
Rotated

	C_1'	C_2'
I_1	.7	-.1
I_2	.7	.1
I_3	.1	-.5
I_4	.2	-.6

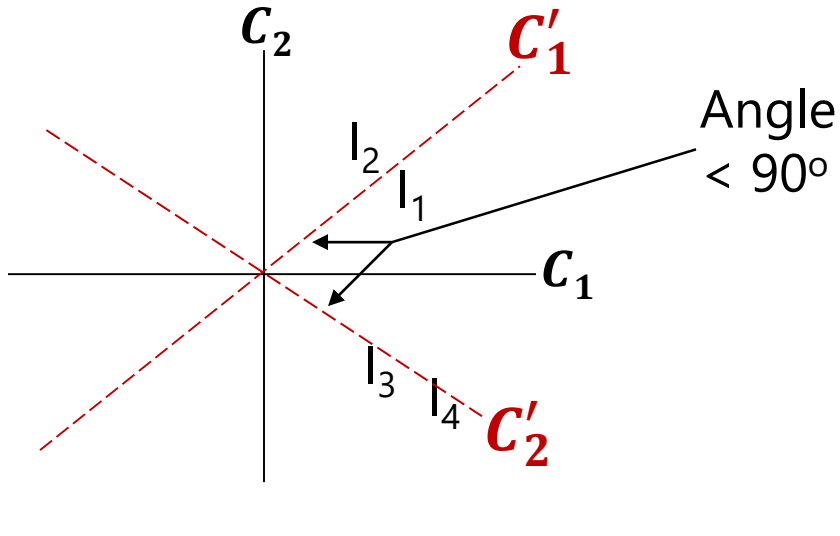
Major Types of Rotation

- Orthogonal Rotation—resulting factors are uncorrelated
  - More parsimonious and efficient, but less "natural"
- Oblique Rotation—resulting factors are correlated
  - More "natural" and better "spearing", but more complicated

Orthogonal Rotation



Oblique Rotation



Types of **Orthogonal** Rotation

- **Varimax**—most commonly used and often program default
  - "Simplifies factors" by maximizing variance of loadings within factors (high loadings → higher, low loadings → lower)
  - Tends to produce group factors (factors are more equitable)
- **Quartimax**
  - "Simplifies indicators" by maximizing variance of loadings within indicators (minimizes #factors each indicator loads on)
  - Tends to "move" indicators from extraction less than varimax
  - Tends to produce a general and small group factors
- **Equimax**
  - Designed to "balance" varimax and quartimax tendencies
  - Didn't work very well (particularly if you don't know how many factors you should have)—can't do simultaneously—whichever is done first dominates the final structure

Types of **Oblique** (Correlated) Rotation

- **Direct Oblimin**
  - Spearing indicator clusters as well as possible to produce lowest occurrence of cross-loading indicators
  - Depends on value of "allowed correlation" (δ in SPSS, Γ also):
    - $\delta = -4$ solution is orthogonal
    - $\delta < 0$ solutions are increasingly orthogonal
    - $\delta = 0$ factors are fairly highly correlated (Direct Quartimin)
    - $\delta = 1$ factors are very highly correlated
    - This parameter matters, so try a few versions…
- **Promax**
  - Computes best orthogonal solution and then "relaxes" orthogonality constraints to better "spear" indicator clusters with factor vectors (give simpler structure)
- **Geomin** (default in Mplus)
  - Uses iterative algorithm that attempts to provide a good fit to the non-rotated factor loadings while minimizing a penalty function

Steps in EFA (and PCA)

1. Choose an estimator/extraction method
2. Determine number of factors/components
3. Select a rotation
4. **Interpret solution** (may need to repeat steps 2 and 3)
5. **(Don't) generate factor scores**

Interpreting Factors

- **Interpretation**: process of "naming factors" based on which indicators "load on" them (or load most highly) within a given rotation
- Which indicators "load" is decided based on a "cutoff"
  - **Loading cutoffs** usually range from **.3 to .4** (absolute magnitude)
  - Note that standard errors or significance tests of loadings are not usually given!!
    - Although can be obtained separately though other procedures or in Mplus
- Higher cutoffs decrease number of indicators that "load"
  - Factors may be ill-defined, some indicators may not load
- Lower cutoffs increase number of indicators that "load"
  - Indicators more likely to be load on more than one factor
- General and "larger" factors include more indicators, account for more variance → more parsimonious (but may lump stuff together)
- Unique and "smaller" factors include fewer indicators and may be more focused → often more specific (but too many is not helpful)

Which Set of Loadings?

- Orthogonal Rotation:
  - "**Rotated Factor** (or Component) **Matrix**"
  - Correlation of indicator with the factor… the end.

- Oblique Rotations: 3 different matrices are relevant
  - Loadings in "**Pattern Matrix**": Partial correlation of indicator with the factor, controlling for the other factors
    - Most often used to interpret the solution
  - Loadings in "**Structure Matrix**": Bivariate correlation of indicator with the factor
    - Loadings will pry be higher than in the pattern matrix
  - "Factor Correlation Matrix": Correlations among factors
  - Pattern Matrix * Factor Correlation Matrix = Structure Matrix

"Bad" Kinds of Factors and Items

- EFA starts with correlations, so **any item properties besides the construct** that influence correlations can unknowingly influence factor solutions:
  - Differential skewness → lower correlation
  - Difficulty factors → indicators with higher means group together
  - Wording direction → reverse-coded indicators may group together
  - Common method → indicators from same source of observation or about the same object may group together regardless of construct
  - **EFA offers no way to disentangle** these possibilities via constraints!

- Items that load on >1 factor = "multivocal"
  - Does the indicator just happen to measure two things? (pry not good)
  - Or do you have a 'third construct' that is different than, but related to, the factors it is currently loading on? (perhaps better?)
  - Multivocal items can be theoretically informative—they could be explored further, even though this may mean more research adding additional indicators that help resolve some of these issues

Factor Scores in EFA: Just Say No

- **Factor Indeterminacy** precludes meaningful EFA factor scores:
  - There is an infinite number of possible factor scores from exploratory solutions that all have the same mathematical characteristics
  - Different approaches can yield very different results!

- LTMMs can be used to provide predicted factor scores that do not have the same indeterminacy problems
  - But they may not adequately capture factor score unreliability
  - Btw, sum scores actually do imply a specific type of LTMM—stay tuned

- Or **just use SEM** for a simultaneous approach. Then you don't need the factor scores anyway….
  - Stay tuned for a reasonable way to use them when you can't do SEM…

Summary: Exploratory Factor Analysis

- **EFA requires lots of arbitrarily made decisions**
  - # factors, type of rotation, cutoffs for loadings… interpretation!

- Best-case scenario: we get about the same answer regardless of these arbitrary solution choices
  - More realistic scenario: we have to pick something and defend it
  - Report all factor loadings and factor correlations so that readers have same information you did to make their own decisions…

- Then comes replication with another similar sample…
  - THEN it's time for LTMM so we can actually test alternative models, not just describe a correlation matrix…
  - **Or just use a LTMM if you have some idea of what you are measuring in the first place (even if you aren't quite right)!**