Exploratory Factor Analysis and Principal Component Analysis

- Today's Topics:
  - What are EFA and PCA for?
  - Planning a factor analytic study
  - Analysis steps:
    - Extraction methods
    - How many factors
    - Rotation and interpretation
    - (Don't) generate factor scores
  - Wrapping Up…

Where we are headed…

- This course is dedicated to latent trait measurement models…
  - Confirmatory factor models (≈ linear factor models), item response models (≈ nonlinear factor models), and others, too!

- Now we'll visit EFA and PCA to illustrate how these devices are similar to and different than confirmatory factor models
  - Hitting the major points only—it's not worth learning more, because these techniques are antiquated and generally pretty terrible
  - The results from exploratory analyses can be misleading:
    - If data do not meet assumptions of model or method selected (non-normal)
    - If constraints made by analysis are implausible (the definition of EFA)
    - Results are certain to be idiosyncratic to the sample analyzed

- 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 answer right

EFA vs. PCA

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

- My world view:
  - I'll describe them via school of thought #2.
    I want you to know what their 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 causes these observed responses?
  - Factors predict the patterns of correlation among indicators
  - If there is no correlation among indicators, game over
  - Solution is an end (i.e., is of interest) in and of itself

- PCA: "Reduce multiple observed variables into fewer components that summarize their variance"
  - In other words, how can I abbreviate this set of variables?
  - Indicators don't have to be correlated
  - Solution is usually a means to an end

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 (measures 2+ 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 variability
  - "At least 5 people per indicator"
  - Somewhere past 200 or so… "300 is comforting"

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 hurt 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 being predicted by all the other indicators
    - These R^2 values are called **commonalities** (H^2)
    - Means that the leftover non-common variance (which we'll eventually call error variance) gets dropped prior to analysis

Extraction Methods: PCA
(school of thought #1, please don't hurt me)

- PCA: Extracts # COMPONENTS = # indicators
  - Will perfectly reproduce original correlation matrix
  - Unique mathematical solution
  - Components are uncorrelated (orthogonal)
  - Extracted in order of most variance accounted for in indicators
  - Provides component loadings (the L's) that relate each observed indicator (the I's) to each extracted component (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 components?
= Full Component Solution

Keep fewer components?
= Truncated Component Solution

PCA, continued

- Consider this correlation matrix
- There appears to be 2 kinds of information in these 4 indicators
  - I_1 & I_2 I_3 & 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 & I_2
  - $C_2 = L_{23}I_3 + L_{24}I_4$ → capturing the information in I_3 & I_4

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

PCA, continued

- So, because of the cross correlations, in order to maximize the variance reproduced, C_1 will be formed more like …

$$\mathbf{C_1 = .5I_1 + .5I_2 + .4I_3 + .4I_4}$$

  - Notice that all the variables contribute to defining C_1
  - Notice the slightly higher loadings for I_1 & I_2

- Because C_1 didn't focus on the I_1 & I_2 indicator group or I_3 & I_4 indicator group, there will still be variance to account for in both, and C_2 will be formed, probably something like…

$$\mathbf{C_2 = .3I_1 + .3I_2 - .4I_3 - .4I_4}$$

  - Notice that all the variables contribute to defining C_2
  - Notice the slightly higher loadings for I_3 & I_4

- PCA maximizes variance accounted for; it does not find groups of indicators that measure the same thing

PCA: 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 C_1 → $.8^2 + .7^2 + .2^2 + .2^2 = 1.21$

- Eigenvalue / # indicators = variance accounted for across indicators by that component

  % C_1 → 1.21 / 4 = .3025 or 30.25%

- If you square and sum across the values in a row, you get the extracted communality for that indicator (started at 1 in PCA):

  R^2 for I_1 → $.8^2 + -.2^2 = .68$ or 68% of its variance

  - Note this won't work unless the solution stays orthogonal…

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

EFA Extraction Methods: PF vs. ML

- PCA-based methods of "extraction" for EFA:
  - No model fit, but no multivariate normality required
  - Iterative procedure focused on finding communalities
    - Starts as R^2 from prediction by other indicators ("Initial")
    - Ends up with R^2 from prediction by all the factors ("Extraction")
    - Watch out for "Heywood cases" → $R^2 > 1$
  - Goal is to maximize variance extracted

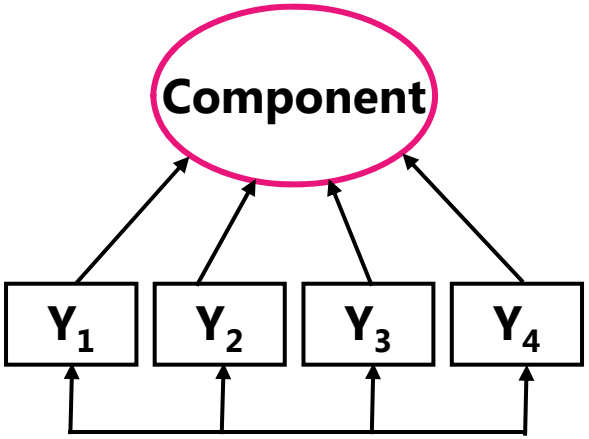
- ML = Maximum Likelihood
  - Focuses on coming up with 'best guesses' for loadings and error variances, not directly for communalities
  - Assessment of model fit because uses same log-likelihood as CFA/SEM
    - Most programs require multivariate normality (there are other options in Mplus)

Big Conceptual Difference between PCA and EFA

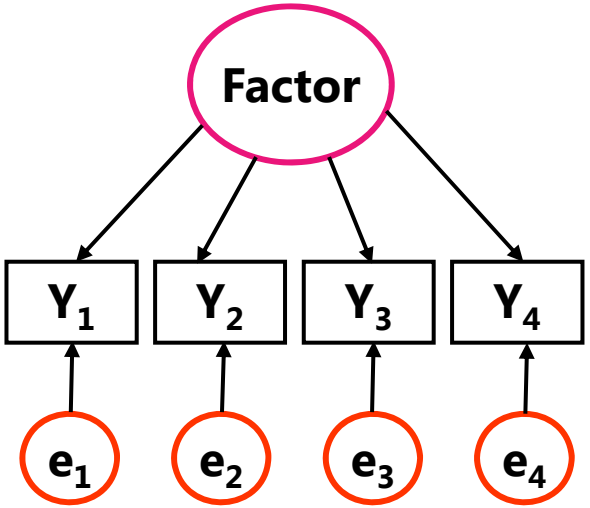
- In PCA, we get components that are **outcomes** built from linear combinations of the indicators:
  - $\mathbf{C_1} = L_{11}I_1 + L_{12}I_2 + L_{13}I_3 + L_{14}I_4 + L_{15}I_5$
  - $\mathbf{C_2} = L_{21}I_1 + L_{22}I_2 + L_{23}I_3 + L_{24}I_4 + L_{25}I_5$
  - … and so forth – note that C is the **OUTCOME**
    - **This is not a testable measurement model by itself.**

- In EFA, we get factors that are thought to be the **cause** of the observed indicators (here, 5 indicators, 2 factors):
  - $I_1 = L_{11}\mathbf{F_1} + L_{12}\mathbf{F_2} + e_1$
  - $I_2 = L_{21}\mathbf{F_1} + L_{22}\mathbf{F_2} + e_1$
  - $I_3 = L_{31}\mathbf{F_1} + L_{32}\mathbf{F_2} + e_1$
  - … and so forth… but note that F is the **PREDICTOR → testable**

PCA vs. EFA/CFA



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



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.**

Big Conceptual Difference between PCA and EFA

- In **PCA**, the component is just the sum of the parts, and there is no inherent reason why the parts should be correlated (they just are)
  - But they should be (otherwise, there's no point in trying to build components to summarize the variables → "component" = "variable")
  - The type of construct measured by a component is often called an "**emergent**" construct – i.e., it emerges from the indicators ("**formative**").
  - Examples: "Lack of Free time", "SES", "Support/Resources"

- In **EFA**, the indicator responses are caused by the factors, and thus should be uncorrelated once controlling for the factor(s)
  - Type of construct that is measured by a factor is often called a '**reflective**' construct – i.e., the indicators are a reflection of your status on the latent variable
  - Examples: Pretty much everything else…

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 computer 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 out of total possible variance
    - In EFA, 'total' is just out of 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)

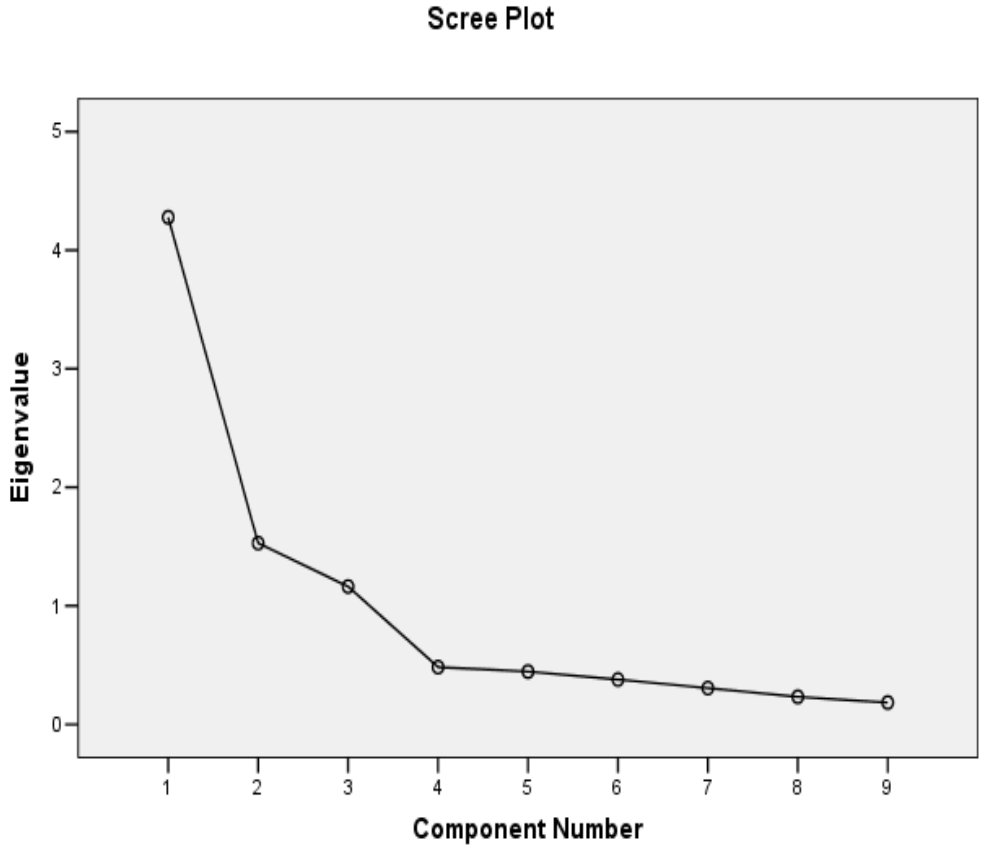
How many factors?

- 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 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 total, non-reduced 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?

Scree "Test" → Scree plot

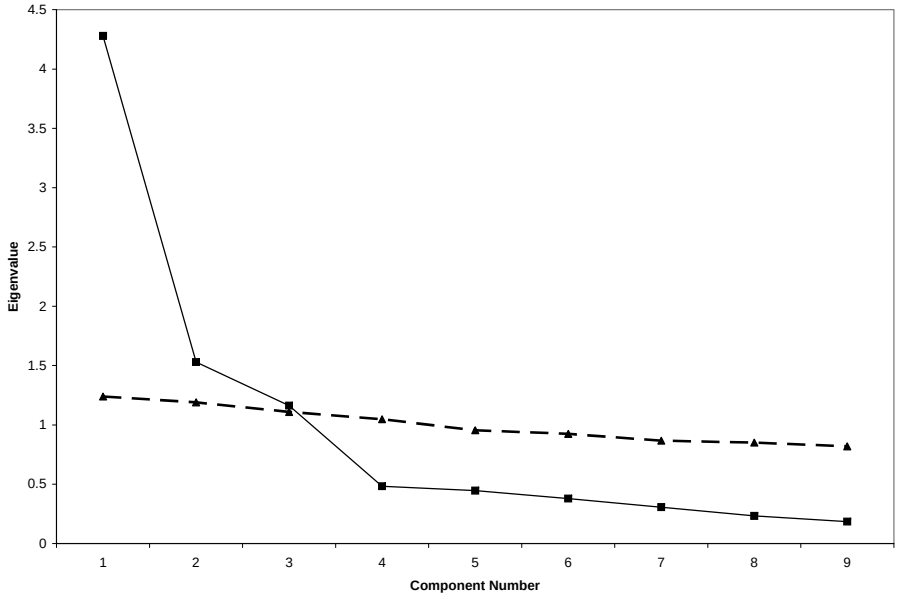
- Plot factor number on x-axis, its Eigenvalue on y-axis
- Look for 'break' in the curve where the slope changes, and retain the number of factors before that break
- Available in most programs
- Research suggests it works 'most of the time'



How many factors?

Parallel "Test" → Parallel plot

- Plot Eigenvalues from your solution against those obtained from simulated data using randomly generated numbers
  - Use mean across simulations (same sample size, same # indicators, same # factors)
- Find point where real data crosses fake data – retain # factors above that point
- Not available in SPSS
  - Available SAS code reference given in Brown chapter 3



Intermediate Summary…

- PCA and EFA are both exploratory techniques geared loosely towards examining the structure underneath a series of continuous indicators (items or subscales):
  - PCA: How do indicators linearly combine to produce a set of uncorrelated linear composite outcomes?
  - EFA: What is the structure of the latent factors that produced the covariances among the observed indicators (factor = predictor)?

- Involves sequence of sometimes ambiguous decisions:
  - Extraction method
  - Number of factors
  - Next up: rotation, interpretation, and factor scores…

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

- Different rotations result in equivalently-fitting, but 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

Goal of Rotation: Simple Structure

- The idea of rotation is to redefine the factor loadings to obtain simple 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

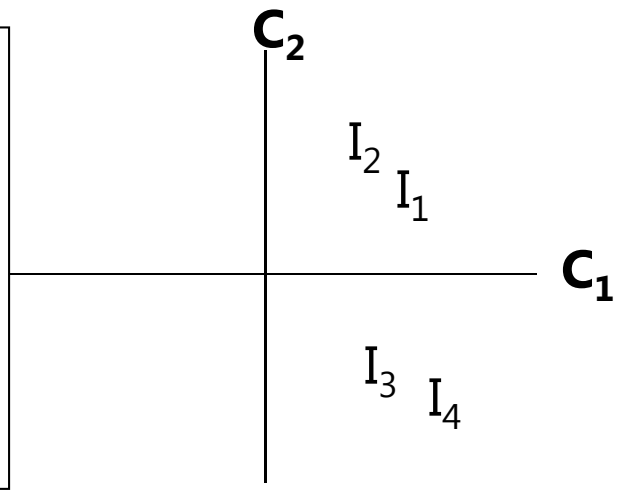
- Two kinds of rotations:
  - Orthogonal (uncorrelated factors—seriously??)
  - Oblique (correlation among factors in another matrix)

"Simple Structure" via Rotation

- We're usually factoring to find "groups of indicators", but the extraction process is trying to "reproduce variance"
- Factor Rotations—changing the "viewing angle" of the factor space—are the major approach to providing simple structure
- **Simple Structure**: factor vectors spear the indicator clusters, such that each indicator loads only on one factor

Un-rotated

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

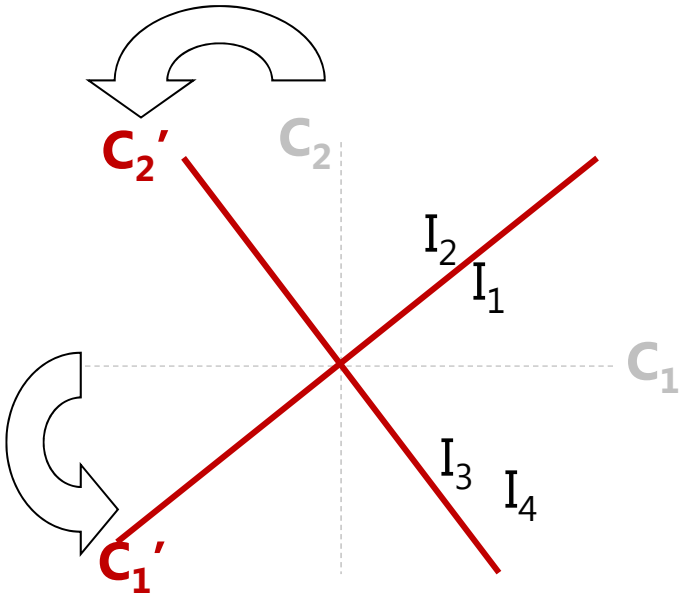


"Simple Structure" via Rotation

- 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 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



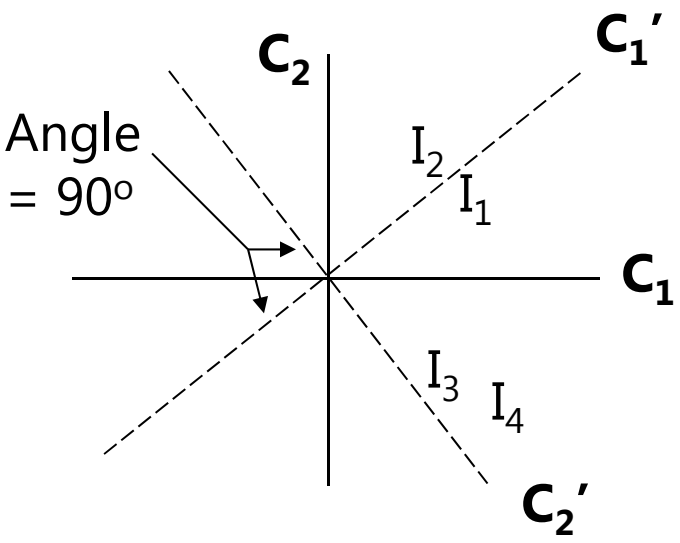
Rotated

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

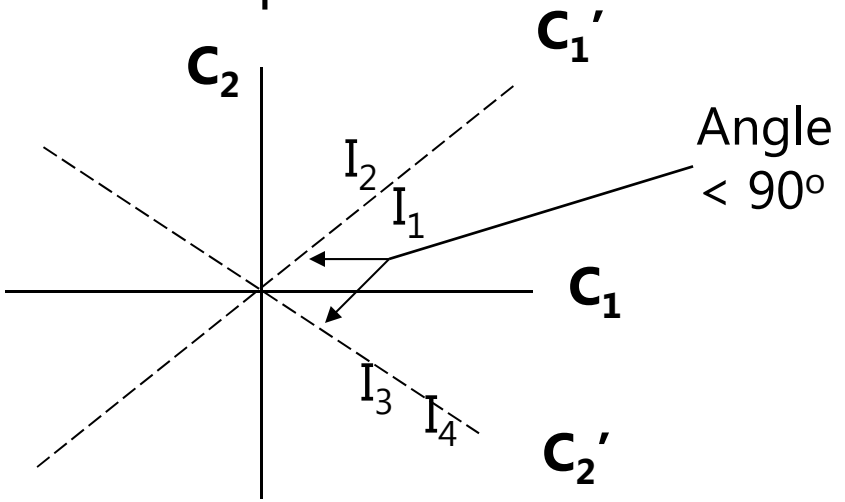
Major Types of Rotation

- 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 common 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 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 is the process of "naming factors" based on the indicators that "load on" them
- Which indicators "load" is decided based on a "cutoff"
  - Cutoffs usually range from .3 to .4 (+/-)
  - Note that significance tests of loadings are not usually given!!
    - Although can be obtained separately though other procedures or in Mplus
- Higher cutoffs decrease # loading indicators
  - Factors may be ill-defined, some indicators may not load
- Lower cutoffs increase # loading indicators
  - 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 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

- Items that load on >1 factor = "multivocal"
  - Does the indicator just happen to measure two things? (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 (see Grice, 2001):
  - There is an infinite number of possible factor scores that all have the same mathematical characteristics
  - Different approaches can yield very different results
- A simple, yet effective solution is simply sum the items that load highly on a factor…"Unit-weighting"
  - Research has suggested that this 'simple' solution is more effective when applying the results of a factor analysis to different samples – factor loadings don't replicate all that well
  - Just make sure to standardize the indicators first if they are on different numerical scales
- Or just use SEM. You don't need the factor scores anyway….
  - Stay tuned for a reasonable way to use them when you can't do SEM…

Wrapping Up: "Exploratory" Factor Analysis

- Exploring means trying alternatives
  - # factors, rotations, cutoffs for loadings, factor scores…
- Best-case scenario: we get about the same answer regardless of solution choices
  - More realistic scenario: we have to pick one and defend it
  - Report all factor loadings 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 at least some idea of what you are measuring in the first place (even if you aren't quite right)!