Principal Components and Factor Analysis

Goals

- What is factor analysis?
- What are factors?
- Representing factors
  - Graphs and equations
- Extracting factors
  - Methods and criteria
- Interpreting factor structures
  - Factor rotation
- Reliability
  - Cronbach’s alpha

When and Why Do We Use Factor Analysis?

- Take many variables and explain them with a few “factors” or “components”
- To see whether different measures are tapping aspects of a common dimension
  - This is similar to canonical correlation in some ways
General Steps in Factor Analysis

- Step 1: Select and measure a set of variables in a given domain
- Step 2: Screen data in order to prepare the correlation matrix
- Step 3: Factor extraction
- Step 4: Factor rotation to increase interpretability
- Step 5: Interpretation of factors
- Further Steps: Validate and determine reliability of the scales

Correlation Matrix

<table>
<thead>
<tr>
<th></th>
<th>Talk 1</th>
<th>Talk 2</th>
<th>Talk 3</th>
<th>Talk 4</th>
<th>Talk 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Talk 1</td>
<td>1.000</td>
<td>0.505</td>
<td>0.402</td>
<td>0.352</td>
<td>0.303</td>
</tr>
<tr>
<td>Talk 2</td>
<td>0.505</td>
<td>1.000</td>
<td>0.450</td>
<td>0.352</td>
<td>0.336</td>
</tr>
<tr>
<td>Talk 3</td>
<td>0.402</td>
<td>0.450</td>
<td>1.000</td>
<td>0.352</td>
<td>0.343</td>
</tr>
<tr>
<td>Talk 4</td>
<td>0.352</td>
<td>0.352</td>
<td>0.352</td>
<td>1.000</td>
<td>0.312</td>
</tr>
<tr>
<td>Talk 5</td>
<td>0.303</td>
<td>0.336</td>
<td>0.343</td>
<td>0.312</td>
<td>1.000</td>
</tr>
</tbody>
</table>

- In factor analysis, we look to reduce the correlation matrix into smaller sets of uncorrelated dimensions

What is a Factor?

- If several variables correlate highly, they might measure aspects of a common underlying dimension
  - These dimensions are called factors
- Factors are classification axes along which the measures can be plotted
  - The greater the loading of variables on a factor, the more that factor explains relationships between those variables
What is a “Good” Factor?

- A good factor should...
  - Make sense
  - Be easy to interpret
  - Have a simple structure
  - Lack complex loadings

Problems with Factor Analysis

- Unlike many of the other analyses we have covered, there is no statistical criterion to serve as a comparison for the linear combination
- It is more art than science
  - There are a number of extraction methods (e.g., principal-components analysis)
  - There are a number of rotation methods (e.g., Orthogonal, Oblique)
  - Choice of the number of factors to extract
  - Communality estimates

Types of Factor Analysis

- Exploratory factor analysis
  - Summarizing data by grouping correlated variables
  - Investigating sets of measured variables related to theoretical constructs
  - Usually done near the onset of research
  - This is the type of factor analysis that we will address
Types of Factor Analysis

• Confirmatory Factor Analysis
  • More advanced technique
  • When factor structure is known...or at least theorized
  • This basically involves testing the generalization of a factor structure to new data
  • This is tested through Structural Equation Modeling which we will discuss later

Basic Terminology

• Orthogonal Rotation
  • Loading Matrix: correlation between each variable and the factor
• Oblique Rotation
• Factor Correlation Matrix: correlations between the factors
  • Structure Matrix: correlation between factors and variables
  • Pattern Matrix: unique relationship between each factor and variable uncontaminated by overlap between the factors
• Factor Coefficient matrix: coefficients used to calculate factor scores (like regression coefficients)

Factor Analysis vs. Principal-Components Analysis

• FA produces factors
  • PCA produces components
• Factors cause variables
  • Components are aggregates of the variables
• FA analyzes only the variance shared among the variables (common variance without error or unique variance)
  • PCA analyses all of the variance
• FA: "What are the underlying processes that could produce these correlations?"
  • PCA: Just summarize empirical associations, very data driven
The goal of principal-components analysis is to identify a new set of a few variables that explain all (or nearly all) of the total variance.

- These principal-components are a linear function of the original variables.
  - The first principal-component maximizes the amount of variance that is explained.
- Eigenvector: the linear function.
  - Their goal is to explain as much variability as possible.
  - The number varies between analyses.
- Eigenvalue: the total amount of variance that is explained by an eigenvector.

There are a number of strategies for deciding when to stop extracting factors or components:
- Percentage of variance that is explained.
- A specific number of factors may be extracted.
- Kaiser's stopping rule: only extracts (and retains) those factors with eigenvalues of at least 1.
- Scree test is a graphical procedure that depicts the eigenvalues.
The values in the equation represent the weights of a variable on a factor. These values are the same as the coordinates on a factor plot. They are called Factor Loadings. These values are stored in a Factor pattern matrix (A).

**Mathematical Representation**

\[ Y = b_1X_1 + b_2X_2 \ldots b_nX_n \]

Factor, \( b \) Variable, \( + b \) Variable, ..., \( b \) Variable.

\[ Y = b_1X_1 + b_2X_2 \ldots b_nX_n \]

Sociability = \( b \) Talk1 + \( b \) Social Skills + \( b \) Interest + \( b \) Talk2 + \( b \) Selfish + \( b \) Liar

Consideration = \( b \) Talk1 + \( b \) Social Skills + \( b \) Interest + \( b \) Talk2 + \( b \) Selfish + \( b \) Liar

**Factor Loadings**

- The \( b \) values in the equation represent the weights of a variable on a factor.
- These values are the same as the coordinates on a factor plot.
- They are called Factor Loadings.
- These values are stored in a Factor pattern matrix (A).

\[
A = \begin{bmatrix}
0.87 & 0.01 \\
0.96 & 0.03 \\
0.92 & 0.04 \\
0.00 & 0.82 \\
-0.10 & 0.75 \\
0.09 & 0.70
\end{bmatrix}
\]
How many participants?

- Subjects-to-variables ratio should be at least 5-to-1 (5 participants for every variable)

The quality of analysis depends upon the quality of the data ("Garbage In" → "Garbage Out")

- Test variables should correlate quite well
  - $r > 0.3$
- Avoid Multicollinearity:
  - several variables highly correlated, $r > 0.80$
- Avoid Singularity:
  - some variables perfectly correlated, $r = 1$
- Screen the correlation matrix and eliminate any variables that obviously cause concern
Further Considerations

- Determinant: Indicator of multicollinearity
  - Should be greater than 0.00001
- Kaiser-Meyer-Olkin (KMO): Measures sampling adequacy
  - Should be greater than 0.5
- Bartlett's Test of Sphericity: Tests whether the R-matrix is an identity matrix
  - Should be significant at $p < .05$
- Anti-Image Matrix: Measures of sampling adequacy on diagonal
  - Off-diagonal elements should be small
- Reproduced: Correlation matrix after rotation
  - Most residuals should be $< |0.05|$

Finding Factors: Communality

- Common Variance:
  - Variance that a variable shares with other variables
- Unique Variance:
  - Variance that is unique to a particular variable
- The proportion of common variance in a variable is called the communality
  - Communality = 1, All variance shared
  - Communality = 0, No variance shared
  - $0 <$ Communality $< 1 = Some$ variance shared
Finding Factors

- We find factors by calculating the amount of common variance
  - Circularity
- Principal Components Analysis:
  - Assume all variance is shared
  - All communalities = 1
- Factor Analysis
  - Estimate communality
  - Use squared multiple correlation (SMC)
- Principal Components and Factor Analysis will identify similar factors when there are a large number of variables (i.e., more than 30) and the communalities are high (i.e., greater than .7)
Factor Extraction

- **Kaiser's Extraction**
  - Kaiser (1960): retain factors with Eigenvalues > 1

- **Scree Plot**
  - Cattell (1966): use 'point of inflexion' of the scree plot

- **Which Rule?**
  - Use Kaiser’s Extraction when
    - less than 30 variables, communalities after extraction > 0.7
    - sample size > 250 and mean communality ≥ 0.6
    - Scree plot is good if sample size is > 200

---

<table>
<thead>
<tr>
<th>Component</th>
<th>Total Variance</th>
<th>Cumulative Variance</th>
<th>Total Variance</th>
<th>Cumulative Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.159</td>
<td>1.159</td>
<td>1.159</td>
<td>1.159</td>
</tr>
<tr>
<td>2</td>
<td>1.159</td>
<td>2.317</td>
<td>1.159</td>
<td>2.317</td>
</tr>
<tr>
<td>3</td>
<td>0.860</td>
<td>3.177</td>
<td>1.159</td>
<td>3.177</td>
</tr>
<tr>
<td>4</td>
<td>0.860</td>
<td>3.177</td>
<td>1.159</td>
<td>3.177</td>
</tr>
<tr>
<td>5</td>
<td>0.860</td>
<td>3.177</td>
<td>1.159</td>
<td>3.177</td>
</tr>
<tr>
<td>6</td>
<td>0.860</td>
<td>3.177</td>
<td>1.159</td>
<td>3.177</td>
</tr>
<tr>
<td>7</td>
<td>0.860</td>
<td>3.177</td>
<td>1.159</td>
<td>3.177</td>
</tr>
<tr>
<td>8</td>
<td>0.860</td>
<td>3.177</td>
<td>1.159</td>
<td>3.177</td>
</tr>
<tr>
<td>9</td>
<td>0.860</td>
<td>3.177</td>
<td>1.159</td>
<td>3.177</td>
</tr>
<tr>
<td>10</td>
<td>0.860</td>
<td>3.177</td>
<td>1.159</td>
<td>3.177</td>
</tr>
<tr>
<td>11</td>
<td>0.860</td>
<td>3.177</td>
<td>1.159</td>
<td>3.177</td>
</tr>
<tr>
<td>12</td>
<td>0.860</td>
<td>3.177</td>
<td>1.159</td>
<td>3.177</td>
</tr>
<tr>
<td>13</td>
<td>0.860</td>
<td>3.177</td>
<td>1.159</td>
<td>3.177</td>
</tr>
<tr>
<td>14</td>
<td>0.860</td>
<td>3.177</td>
<td>1.159</td>
<td>3.177</td>
</tr>
<tr>
<td>15</td>
<td>0.860</td>
<td>3.177</td>
<td>1.159</td>
<td>3.177</td>
</tr>
<tr>
<td>16</td>
<td>0.860</td>
<td>3.177</td>
<td>1.159</td>
<td>3.177</td>
</tr>
<tr>
<td>17</td>
<td>0.860</td>
<td>3.177</td>
<td>1.159</td>
<td>3.177</td>
</tr>
<tr>
<td>18</td>
<td>0.860</td>
<td>3.177</td>
<td>1.159</td>
<td>3.177</td>
</tr>
<tr>
<td>19</td>
<td>0.860</td>
<td>3.177</td>
<td>1.159</td>
<td>3.177</td>
</tr>
<tr>
<td>20</td>
<td>0.860</td>
<td>3.177</td>
<td>1.159</td>
<td>3.177</td>
</tr>
<tr>
<td>21</td>
<td>0.860</td>
<td>3.177</td>
<td>1.159</td>
<td>3.177</td>
</tr>
<tr>
<td>22</td>
<td>0.860</td>
<td>3.177</td>
<td>1.159</td>
<td>3.177</td>
</tr>
</tbody>
</table>

Scree Plots

- Point of Inflection

---

- **Kaiser's Extraction**
  - Kaiser (1960): retain factors with Eigenvalues > 1

- **Scree Plot**
  - Cattell (1966): use 'point of inflexion' of the scree plot

- **Which Rule?**
  - Use Kaiser’s Extraction when
    - less than 30 variables, communalities after extraction > 0.7
    - sample size > 250 and mean communality ≥ 0.6
    - Scree plot is good if sample size is > 200
To aid interpretation it is possible to maximize the loading of a variable on one factor while minimizing its loading on all other factors. This is known as Factor Rotation. There are two types:

- Orthogonal (factors are uncorrelated)
- Oblique (factors intercorrelate)
What Do the Factors Represent?

- We assume that algebraic factors represent psychological constructs
  - Factor 1: Fear of statistics (e.g., "I can’t sleep for thoughts of eigenvectors")
  - Factor 2: Fear of peer evaluation (e.g., "My friends are better at statistics than me")
  - Factor 3: Fear of computers (e.g., "All computers hate me")
  - Factor 4: Fear of mathematics (e.g., "I have never been good at mathematics")
- The nature of these psychological dimensions is ‘guessed at’ by looking at the loadings for a factor
- There is no way to "know" what the factor represents…rather we have to decide ourselves
  - The same set of items may be referred to as “social dominance” by one researcher but “aggression” by another

Reliability

- Test-Retest Method
  - Complete the same measure on two occasions and calculates the correlation
- Alternate Form Method
  - Complete two slightly different forms of the same measure and calculates the correlation
- Split-Half Method
  - Splits the questionnaire into two random halves and calculates the correlation
- Cronbach’s Alpha
  - Basically splits the questionnaire into all possible halves, calculates the scores, correlates them, and averages the correlation for all splits
  - Ranges from 0 (no reliability) to 1 (complete reliability)

Interpreting Cronbach’s Alpha

- Kline (1999)
  - Reliable if $\alpha > .7$
- Depends on the number of items
  - More questions = bigger $\alpha$
- Treat Subscales separately
- Remember to reverse score reverse phrased items!
  - If not, $\alpha$ is reduced and can even be negative
## Reliability for Fear of Statistics Subscale

<table>
<thead>
<tr>
<th>Cronbach’s Alpha</th>
<th>Cronbach’s Alpha Based on Standardized Items</th>
<th>N of Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.821</td>
<td>0.823</td>
<td>8</td>
</tr>
</tbody>
</table>

## Reliability for the Fear of Peer Evaluation Subscale

<table>
<thead>
<tr>
<th>Cronbach’s Alpha</th>
<th>Cronbach’s Alpha Based on Standardized Items</th>
<th>N of Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.578</td>
<td>0.572</td>
<td>5</td>
</tr>
</tbody>
</table>

## Reliability for Fear of Computers Subscale

<table>
<thead>
<tr>
<th>Cronbach’s Alpha</th>
<th>Cronbach’s Alpha Based on Standardized Items</th>
<th>N of Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.822</td>
<td>0.821</td>
<td>7</td>
</tr>
</tbody>
</table>
# Reliability for Fear of Mathematics Subscale

## Item Total Statistics

<table>
<thead>
<tr>
<th>Item Description</th>
<th>Scale Mean</th>
<th>Scale Standard Deviation</th>
<th>Internal Consistency</th>
<th>Reliability Coefficient</th>
<th>Cronbach's Alpha Before Truncation</th>
<th>Cronbach's Alpha After Truncation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Have never been good at mathematics</td>
<td>4.31</td>
<td>1.677</td>
<td>0.884</td>
<td>0.494</td>
<td>0.400</td>
<td>0.400</td>
</tr>
<tr>
<td>Need to think in numbers to understand</td>
<td>4.16</td>
<td>1.677</td>
<td>0.884</td>
<td>0.494</td>
<td>0.400</td>
<td>0.400</td>
</tr>
<tr>
<td>Look at a complex when see an equation</td>
<td>4.05</td>
<td>1.677</td>
<td>0.884</td>
<td>0.494</td>
<td>0.400</td>
<td>0.400</td>
</tr>
</tbody>
</table>

## Reliability Statistics

<table>
<thead>
<tr>
<th>Cronbach's Alpha</th>
<th>Cronbach's Alpha Based on Standardized Items</th>
<th>N of Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.19</td>
<td>0.19</td>
<td>3</td>
</tr>
</tbody>
</table>
