Factor Analysis

Psy 524

Ainsworth

What is Factor Analysis (FA)?

- FA and PCA (principal components analysis) are methods of data reduction
  - Take many variables and explain them with a few “factors” or “components”
  - Correlated variables are grouped together and separated from other variables with low or no correlation

What is FA?

- Patterns of correlations are identified and either used as descriptives (PCA) or as indicative of underlying theory (FA)
- Process of providing an operational definition for latent construct (through regression equation)

What is FA?

- FA and PCA are not much different than canonical correlation in terms of generating canonical variates from linear combinations of variables
  - Although there are now no “sides” of the equation
  - And your not necessarily correlating the “factors”, “components”, “variates”, etc.

General Steps to FA

- Step 1: Selecting and Measuring a set of variables in a given domain
- Step 2: Data screening in order to prepare the correlation matrix
- Step 3: Factor Extraction
- Step 4: Factor Rotation to increase interpretability
- Step 5: Interpretation
- Further Steps: Validation and Reliability of the measures

“Good Factor”

- A good factor:
  - Makes sense
  - will be easy to interpret
  - simple structure
  - Lacks complex loadings

Problems w/ FA

- Unlike many of the analyses so far there is no statistical criterion to compare the linear combination to
  - In MANOVA we create linear combinations that maximally differentiate groups
  - In Canonical correlation one linear combination is used to correlate with another

Problems w/ FA

- It is more art than science
  - There are a number of extraction methods (PCA, FA, etc.)
  - There are a number of rotation methods (Orthogonal, Oblique)
  - Number of factors to extract
  - Communality estimates
  - ETC…
- This is what makes it great…

Problems w/ FA

- Life (researcher) saver
  - Often when nothing else can be salvaged from research a FA or PCA will be conducted

Types of FA

- Exploratory FA
  - Summarizing data by grouping correlated variables
  - Investigating sets of measured variables related to theoretical constructs
  - Usually done near the onset of research
  - The type of FA and PCA we are talking about in this chapter

Types of FA

- Confirmatory FA
  - More advanced technique
  - When factor structure is known or at least theorized
  - Testing generalization of factor structure to new data, etc.
  - This is tested through SEM methods discussed in the next chapter

Terminology

- Observed Correlation Matrix
- Reproduced Correlation Matrix
- Residual Correlation Matrix

Terminology

- Orthogonal Rotation
  - Loading Matrix – correlation between each variable and the factor
- Oblique Rotation
  - Factor Correlation Matrix – correlation 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

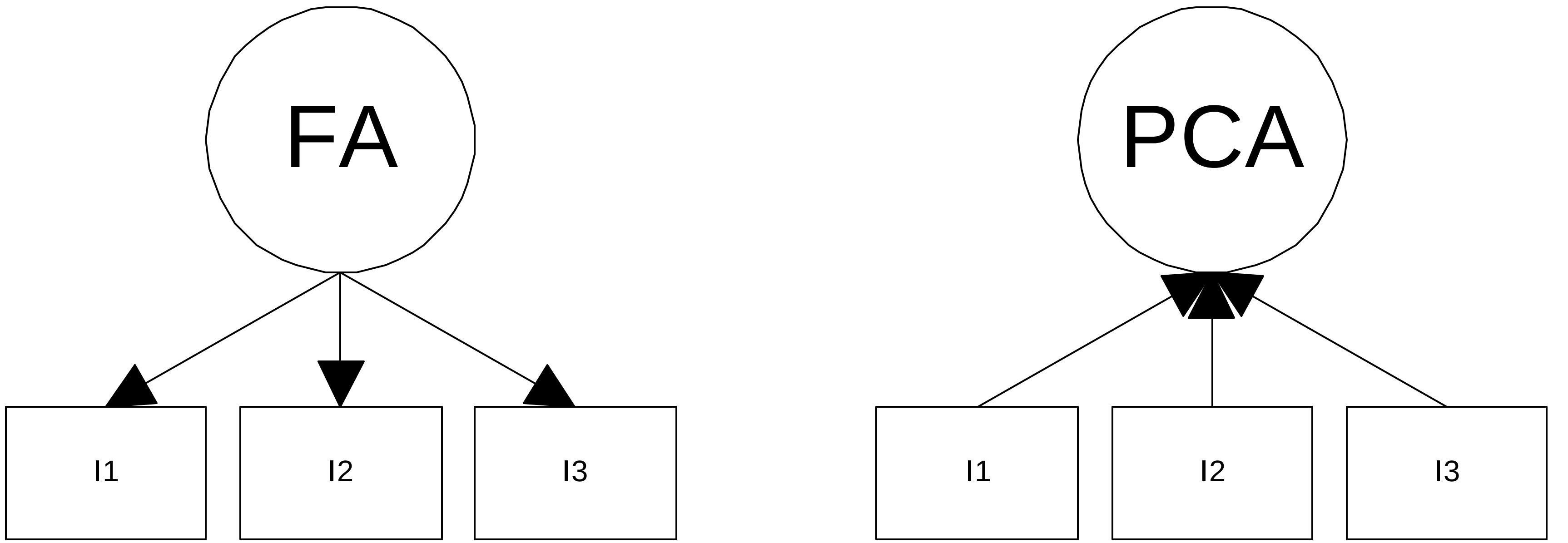
Terminology

- Factor Coefficient matrix – coefficients used to calculate factor scores (like regression coefficients)

FA vs. PCA conceptually

- FA produces factors; PCA produces components
- Factors cause variables; components are aggregates of the variables

Conceptual FA and PCA



FA vs. PCA conceptually

- FA analyzes only the variance shared among the variables (common variance without error or unique variance); PCA analyzes all of the variance
- FA: “What are the underlying processes that could produce these correlations?”; PCA: Just summarize empirical associations, very data driven

Questions

- Three general goals: data reduction, describe relationships and test theories about relationships (next chapter)
- How many interpretable factors exist in the data? or How many factors are needed to summarize the pattern of correlations?

Questions

- What does each factor mean? Interpretation?
- What is the percentage of variance in the data accounted for by the factors?

Questions

- Which factors account for the most variance?
- How well does the factor structure fit a given theory?
- What would each subject’s score be if they could be measured directly on the factors?

Considerations (from Comrey and Lee, 1992)

- Hypotheses about factors believed to underlie a domain
  - Should have 6 or more for stable solution
- Include marker variables
  - Pure variables – correlated with only one factor
  - They define the factor clearly
  - Complex variables load on more than on factor and muddy the water

Considerations (from Comrey and Lee, 1992)

- Make sure the sample chosen is spread out on possible scores on the variables and the factors being measured
- Factors are known to change across samples and time points, so samples should be tested before being pooled together

Assumptions

- Assumes reliable correlations
  - Highly affected by missing data, outlying cases and truncated data
  - Data screening methods (e.g. transformations, etc.) can greatly improve poor factor analytic results

Assumptions

- Sample Size and Missing Data
  - True missing data (MCAR) are handled in the usual ways (ch. 4) but regression methods may overfit
  - Factor analysis needs large samples and it is one of the only draw backs
    - The more reliable the correlations are the smaller the number of subjects needed
    - Need enough subjects for stable estimates

Assumptions

- Comrey and Lee
  - 50 very poor, 100 poor, 200 fair, 300 good, 500 very good and 1000+ excellent
  - Shoot for minimum of 300 usually
  - More highly correlated markers less subjects

Assumptions

- Normality
  - Univariate - normally distributed variables make the solution stronger but not necessary
  - Multivariate – is assumed when assessing number of factors; usually tested univariately

Assumptions

- No outliers – obvious influence on correlations would bias results
- Multicollinearity/Singularity
  - In PCA it is not problem; no inversions
  - In FA, if det(R) or any eigenvalue approaches 0 -> multicollinearity is likely
  - Also investigate inter-item SMCs approaching 1

Assumptions

- Factorable R matrix
  - Need inter-item correlations > .30 or FA is unlikely
  - Large inter-item correlations does not guarantee solution either
    - Duos
    - Multidimensionality
  - Matrix of partials adjusted for other variables
  - Other tests

Assumptions

- Variables as outliers
  - Some variables don’t work
  - Explain very little variance
  - Relates poorly with factor
  - Low SMCs with other items
  - Low loadings

Extraction Methods

- There are many (dozens at least)
- All extract orthoganal sets of factors (components) that reproduce the R matrix
- Different techniques – some maximize variance, others minimize the residual matrix (R – reproduced R)
- With large stable sample they all should be relatively the same

Extraction Methods

- Usually un-interpretable without rotation (next)
- Differ in output depending on combinations of
  - Extraction method
  - Communality estimates
  - Number of factors extracted
  - Rotational Method

Extraction Methods

- PCA vs. FA (family)
  - PCA begins with 1s in the diagonal of the correlation matrix; all variance extracted; each variable giving equal weight; outputs inflated communality estimate
  - FA begins with a communality estimates (e.g. SMC) in the diagonal; analyzes only common variance; outputs a more realistic communality estimate

Extraction Methods

- PCA analyzes variance
- FA analyzes covariance (communality)
- PCA reproduces the R matrix (near) perfectly
- FA is a close approximation to the R matrix

Extraction Methods

- PCA – the goal is to extract as much variance with the least amount of factors
- FA – the goal is to explain as much of the correlations with a minimum number of factors
- PCA gives a unique solution
- FA can give multiple solutions depending on the method and the estimates of communality

Extraction Methods

- PCA
  - Extracts maximum variance with each component
  - First component is a linear combination of variables that maximizes component score variance for the cases
  - The second (etc.) extracts the max. variance from the residual matrix left over after extracting the first component (therefore orthogonal to the first)
  - If all components retained, all variance explained

Extraction Methods

- Principal (Axis) Factors
  - Estimates of communalities (SMC) are in the diagonal; used as starting values for the communality estimation (iterative)
  - Removes unique and error variance
  - Solution depends on quality of the initial communality estimates

Extraction Methods

- Maximum Likelihood
  - Computationally intensive method for estimating loadings that maximize the likelihood (probability) of the correlation matrix.
- Unweighted least squares – ignores diagonal and tries to minimize off diagonal residuals
  - Communalites are derived from the solution
  - Originally called Minimum Residual method (Comrey)

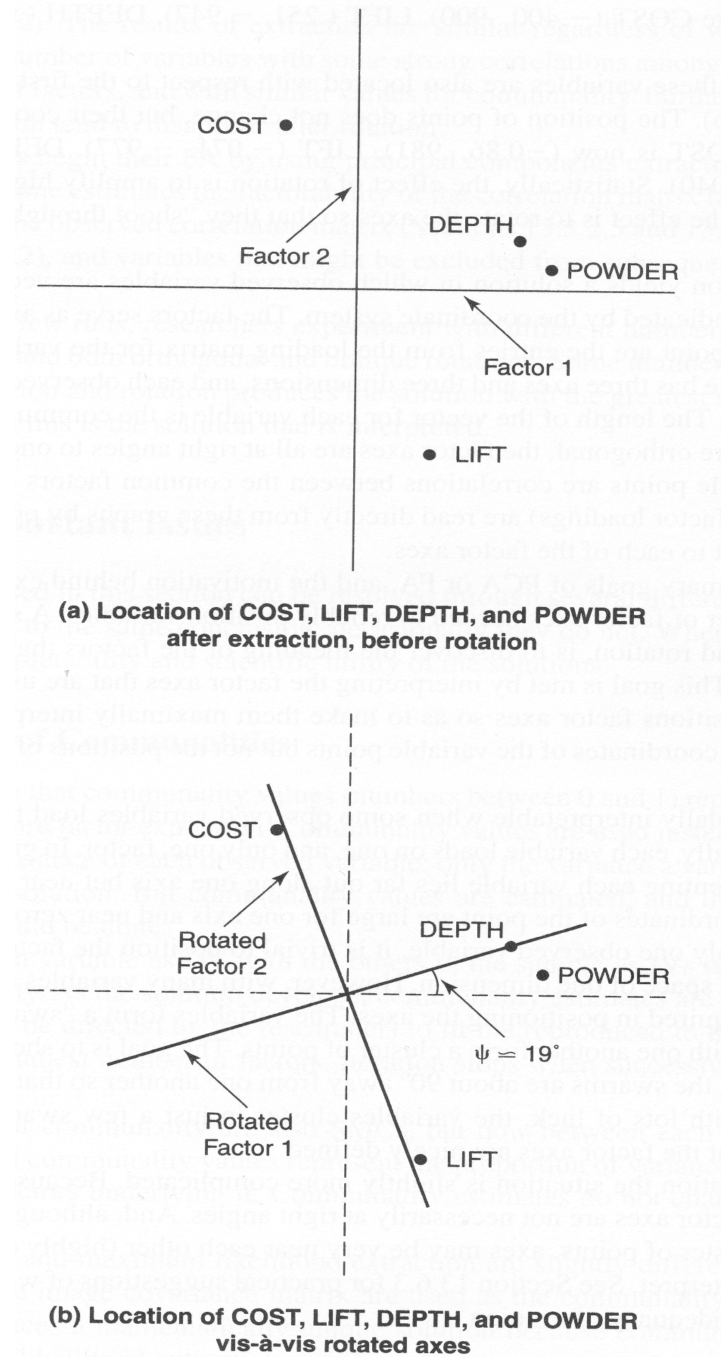
Extraction Methods

- Generalized (weighted) least squares
  - Also minimizes the off diagonal residuals
  - Variables with larger communalities are given more weight in the analysis
- Many Other methods

Rotation Methods

- After extraction (regardless of method) good luck interpreting result
- Rotation is used to improve interpretability and utility
- A orthogonally rotated solution is mathematically equivalent to un-rotated and other orthogonal solutions
- Stable and large N -> same result

Geometric Rotation



Geometric Rotation

- Factor extraction equivalent to coordinate planes
- Factors are the axes
- Length of the line from the origin to the variable coordinates is equal to the communality for that variable
- Orthogonal Factors are at right angles

Geometric Rotation

- Factor loadings are found by dropping a line from the variable coordinates to the factor at a right angle
- Repositioning the axes changes the loadings on the factor but keeps the relative positioning of the points the same

Rotation Methods

- Orthogonal vs. Oblique
  - Orthogonal rotation keeps factors un-correlated while increasing the meaning of the factors
  - Oblique rotation allows the factors to correlate leading to a conceptually clearer picture but a nightmare for explanation

Rotation Methods

- Orthogonal Rotation Methods
  - Varimax – most popular
    - Simple structure by maximizing variance of loadings within factors across variables
    - Makes large loading larger and small loadings smaller
    - Spreads the variance from first (largest) factor to other smaller factors

Rotation Methods

- Orthogonal Rotation Methods
  - Quartimax
    - Opposite of Varimax
    - Simplifies variables by maximizing variance with variables across factors
    - Varimax works on the columns of the loading matrix; Quartimax works on the rows
    - Not used as often; simplifying variables is not usually a goal

Rotation Methods

- Orthogonal Rotation Methods
  - Equamax is a hybrid of the earlier two that tries to simultaneously simplify factors and variables
    - Not that popular either

Rotation Methods

- Oblique Rotation Techniques
  - Direct Oblimin
    - Begins with an unrotated solution
    - Has a parameter (gamma in SPSS) that allows the user to define the amount of correlation acceptable; gamma values near -4 -> orthogonal, 0 leads to mild correlations (also direct quartimin) and 1 highly correlated

Rotation Methods

- Oblique Rotation Techniques
  - Promax – most recommended
    - Solution is rotated maximally with an orthogonal rotation
    - This is followed by oblique rotation
    - Orthogonal loadings are raised to powers in order to drive down small loadings
    - Simple structure is reached
    - Easy and quick method