(Factor) Analyze This: PCA or EFA

SAM WOOLFORD PHD, PSTAT®, CQE

SWOOLFORD@BENTLEY.EDU

FOR NIH

JULY 31, 2015

Agenda

Overview

Principal Components Analysis

Exploratory Factor Analysis

Example

Similarities and Differences

Confirmatory Factor Analysis

Factor Analysis: Overview

Typical Applications: PCA or EFA?

- Reduce a large number of variables to a smaller number of factors for further analysis
- Reallocate the variation in a large number of variables
- Create orthogonal representations of the original variables
  - Solve problems of multi-collinearity
- Identify underlying dimensions in the data (constructs)
- Regression with many correlated variables
- Create a hypothesis for a CFA analysis

Factor analysis is an exploratory technique for summarizing the information in observed variables into a smaller set of factors

Factor Analysis: Overview

Start with p correlated “quantitative” variables x_i, i=1,. . ., p

- Each variable should be correlated with at least one other (>.3)

Conceptual foundation for factors (desirable)

- Assumes some underlying structure for the variables exists
  - Should be a homogeneous structure (e.g. gender)

Sample size

- 5-10 observations per variable
- More than 50 observations (200?)
- 2-5 variables per factor (3 or more for CFA)

Factor Analysis: Overview

Level of standardization

- Use correlation matrix
  - Standardized variables give equal importance to all variables as variances are all the same
- Use covariance matrix (data is mean centered)
  - Weight assigned to a variable in a factor is effected by the relative variance of the variable
  - Mean corrected data gives more importance to variables with greater variation

Factor Analysis: Overview

Measures of Adequacy

- Bartlettes test of sphericity
  - Tests that correlation matrix is not diagonal (signif. correlations exist)
- Measure of sampling adequacy (KMO statistic)
  - Predicts whether the data will factor well
  - Overall
    - >.6 is adequate; >.8 is good
  - Individual variables
    - Drop variables with values <.5
    - Delete the lowest first and then continue one at a time until all remaining variables have values >.5

Factor Analysis: Overview

Factors for Factor Extraction

- Partitioning of variable variance
  - Correlation represents shared variance among variables
    - Factor analysis groups variables with high correlation (i.e. shared variance)
  - Components of variable variance
    - Common variance – variance shared with other variables (communality)
    - Unique variance – variance associated with a specific variable and not explained by correlation with other variables
    - Error variance – variance also unexplained by correlation with other variables due to factors such as measurement error or unreliability in data gathering
- Objective of analysis
  - Modeling total variance or common variance only

Factor Analysis: Overview

Principal components

- Assumes that all variation is common variation
  - Accounts for total variation
    - Assumes that the factors can represent the variation in the variables exactly
      - Unique and error variance is a small proportion of the total
  - Diagonal of correlation matrix is taken to be 1
- Objective is data reduction
  - Minimum number of factors that capture the maximum amount of total variation

Factor Analysis: Overview

Exploratory factor analysis

(Aka Common factor analysis or Principal factor analysis)

- Assumes that factors explain only the common shared variance among the variables
  - Unique and error variance is eliminated from the analysis
  - Accounts for the covariance among the variables
    - Diagonal of correlation matrix is taken to be the communalities
- Objective is to identify latent constructs represented in the original variables
  - Often used as a hypothesis generator for confirmatory factor analysis

Factor Analysis: Overview

Results from PCA and EFA may be similar

- When number of variables exceeds 30 or
- Communalities exceed .6 for most variables
- Can run both models and evaluate any differences

Confirmatory factor analysis

- Not an exploratory technique
- Requires hypotheses regarding factors and variables

Principal Components

Model

- Find a set of coefficients a_{ij} such that
  - $y_i = a_{i1}x_1 + \ldots + a_{ip}x_p$ for $i=1,..,p$
  - $y_1,\ldots,y_p$ are uncorrelated
  - Choose direction for y_1 to capture as much of the variation in the x's as possible
  - Choose direction for y_2 to capture as much of the remaining variability as possible while being orthogonal to y_1
  - All variance in x's is transferred to the factors
  - Note that if all the x's are independent then y_1 will be equal to the x with the largest variance and there will be one component for each variable
  - No error in model
    - High correlation (>.9) may cause computational issues

Principal Components

Analysis of factors

- Factor Retention
  - Variance of factor (eigenvalue or latent root) > 1
  - Percentage of variance > 60%
  - Scree test
- Factor Rotation
  - Given the communalities, there are multiple solutions for loadings that result in the same communality and correlation
    - “Correct” answer is dependent upon interpretability
  - Reallocates the variable variance among the factors
    - Shifts variance from earlier factors to later ones
  - Each rotation uses a different criteria to determine the model parameters
    - Orthogonal and non-orthogonal methods are available

Principal Components

Analysis of factors

- Factor Interpretation
  - Based on factor loadings of ±.5 or greater
    - Loadings represent correlations between observed variables and factor
    - Cut off depends on sample size
    - ±.7 or greater represents that the variable shares half or more of its variance with the factor
  - Communalities represent variance shared between observed variables and factor
    - May delete variables that have low communalities (<.5)
  - Need to evaluate interpretation of any large cross-loading

Exploratory Factor Analysis

Model

- Total variance of variables is decomposed into two components
  - One component results from underlying unobserved construct(s)
    - The communality is the proportion of variance of a variable due to common underlying factors
  - One component is unique to the observed variable
- Find a set of latent factors $\xi_1,\ldots, \xi_m$ ($m<p$) such that
  - $x_i=\lambda_{i1} \xi_1 + \ldots + \lambda_{im} \xi_m + \varepsilon_i$ for $i=1,..,p$
    - x_i and ξ_i are standardized (mean 0, variance 1)
    - ε_i has mean 0 and are uncorrelated with ξ_j (for all i and j) and with ε_j (for $i \neq j$)
    - Consider similarity to a regression equation where the independent variables are unknown
    - Correlation between the x's are due to correlation between x's and ξ's

Exploratory Factor Analysis

Objective

- Identify a small number of common factors that linearly explain the correlation between the original variables
  - Goal is to predict the correlation matrix of the original variables with communalities on the diagonal
    - Residual correlations should be small
    - Equivalent to analyzing the covariance matrix
    - Principal axis factoring in SPSS
  - Different from principal components where the goal is to explain as much variance as possible in a few linear combinations of the original variables
- Exploratory if you don't have a hypothetical model
  - All variables load on all factors

Exploratory Factor Analysis

Analysis of factors

- Factor Retention
  - Variance of factor (eigenvalue or latent root) > 1
  - Percentage of variance > 60%
  - Scree test
- Factor Rotation
  - Given the communalities, there are multiple solutions for loadings that result in the same communality and correlation
    - “Correct” answer is dependent upon interpretability
  - Reallocates the variable variance among the factors
    - Shifts variance from earlier factors to later ones
  - Each rotation uses a different criteria to determine the model parameters
    - Orthogonal and non-orthogonal methods are available

Exploratory Factor Analysis

Analysis of factors

- Factor Interpretation
  - Based on factor loadings of ±.5 or greater
    - Loadings represent correlations between observed variables and factor
    - Cut off depends on sample size
    - ±.7 or greater represents that the variable shares half or more of its variance with the factor
  - Communalities represent variance shared between observed variables and factor
    - May delete variables that have low communalities (<.5)
  - Need to evaluate interpretation of any large cross-loading

Example

Data

- Prices for commodities in 23 cities in the US
  - Bread
  - Burger
  - Milk
  - Oranges
  - Tomatoes
- Goal is to represent the cost of living in different cities in the US

Similarities and Differences

	PCA	EFA
Similarities	• Correlated observed variables • Sample size requirements • Measures of adequacy • Analysis of factors requires decisions regarding factor retention, rotation and interpretation	
Differences	• Models total variance • Models factors in terms of observed variables • Objective is typically data reduction	• Models common variance only • Models observed variables in terms of factors • Objective is typically to identify latent factors

Confirmatory Factor Analysis

Model

- Assumes that the factor structure is known (as opposed to exploratory)
  - Number of factors
  - Orthogonality of factors (correlated or not)
  - Indicators for each factor
- Goal is to empirically estimate and confirm the model
  - Often use exploratory factor analysis to help identify a potential structure
  - Use confirmatory factor analysis to verify it and estimate it
  - Like hypothesis testing (exploratory is hypothesis generating)

Confirmatory Factor Analysis

Estimation

- Equates the model covariance with the sample covariance to solve for model parameters
  - Can create requirements on number of observed variables per construct
  - Utilizes the covariance matrix (as opposed to the usual use of the correlation matrix in exploratory)
  - MLE requires observed data is multivariate normal
- Reference metrics
  - Need to fix the metric of the factor in order to estimate (standardized in effect does that-value is scale free)
  - Alternative is to set one of the loadings for each factor to 1
    - This has effect of setting the scale of the factor to be that of that indicator

Confirmatory Factor Analysis

Diagnostics

- Fit statistics
  - Some statistical some rules of thumb
- Parameter statistics
- Construct validity
  - Convergent and divergent validity
- Modification indices