Factors, what factors?

A Whirlwind Tour of Exploratory Factor Analysis

Matthew Sigal

March 9th, 2018

A Cornerstone of Psychology...

Exploratory Factor Analysis

- Basis in psychometric research on intelligence and cognitive abilities
- Also used in personality, psychopathology, other areas
- Used to assess constructs that can't be directly measured:
  - e.g., intelligence, attitudes, personality traits, preferences
- Also can be used to test "factorial invariance" across groups

The Goal:

Develop a model which represents the pattern of associations among a potentially large number of empirically observed variables in terms of a small number of unobserved, or latent, variables (or "factors").

Exploratory Factor Analysis

Driving force is: **Parsimony!**

How many different *underlying constructs* (common factors or latent variables) are needed to account for or explain the *correlations* among a set of observed variables?

EFA assumes that there exists a small number of factors within a given topic domain, which influence the observed variables to varying extents and is responsible for the correlations among them.

Holzinger and Swineford (1939)

- Mental ability test scores from 301 7th and 8th grade children
- 9 test scores – *36 bivariate correlations*

Table 1: Correlation Matrix

	x1	x2	x3	x4	x5	x6	x7	x8	x9
x1	1	0.30	0.44	0.37	0.29	0.36	0.07	0.22	0.39
x2	0.30	1	0.34	0.15	0.14	0.19	-0.08	0.09	0.21
x3	0.44	0.34	1	0.16	0.08	0.20	0.07	0.19	0.33
x4	0.37	0.15	0.16	1	0.73	0.70	0.17	0.11	0.21
x5	0.29	0.14	0.08	0.73	1	0.72	0.10	0.14	0.23
x6	0.36	0.19	0.20	0.70	0.72	1	0.12	0.15	0.21
x7	0.07	-0.08	0.07	0.17	0.10	0.12	1	0.49	0.34
x8	0.22	0.09	0.19	0.11	0.14	0.15	0.49	1	0.45
x9	0.39	0.21	0.33	0.21	0.23	0.21	0.34	0.45	1

Exploratory Factor Analysis

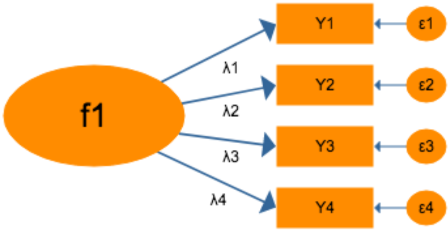
FA is used to establish whether and to what extent certain observed, operational variables can be used to represent *hypothetical* latent variables or constructs. Can be used with:

- a battery of test scores (continuous data; using **R**); or,
- to model individual items within a test (categorical data; using polychoric correlations)

The Common Factor Model

Observed variables depend on two different types of latent variables:

1. **Common factors** influence *more than one* observed variable and account for the correlations among all observed variables and a portion of the variance of each observed variable.
2. **Unique factors** influence only *one* observed variable and represent the part of the observed variable not explained by common factors.
    - Specific - systematic variation affecting a single observation
    - Error - random variation



The Common Factor Model

$$Y_{pi} = \left(\sum_{m=1}^{M} \lambda_{pm} f_{mi} \right) + \epsilon_{pi}$$

- Y_{pi} is the observed score on the pth observed variable for individual i
- λ_{pm} is the factor loading of the pth observed variable on the mth factor
- f_{mi} is a factor score on the mth common factor for individual i
- ϵ_{pi} is the value on the pth unique factor for individual i.

This relates a given observed variable (P) to the *set of common factors* (M) and a *unique factor* (ϵ).

The Common Factor Model

If $M = 1 \rightarrow Y_{pi} = \lambda_{p1} f_{1i} + \epsilon_{pi}$

If $M = 2 \rightarrow Y_{pi} = \lambda_{p1} f_{1i} + \lambda_{p2} f_{2i} + \epsilon_{pi}$

Factor Analysis as Multiple Regression

This is essentially a linear multiple regression model in which the given observed variable (Y) is the outcome and the common factors ($F_1 \ldots F_M$) are the predictor variables!

So, factor loadings (the λ_{pm}s) are partial regression slope coefficients that give the strength of the relationship between the *m*th common factor and the *p*th observed variable.

In Matrix Form...

Since this is just multivariate multiple regression, we can condense the previous expression using matrix notation:

$$\mathbf{Y} = \Lambda \mathbf{f} + \epsilon$$

- $\mathbf{Y}$ is the $P \times 1$ vector of observed variables
- Λ is a $P \times M$ factor loading matrix
- $\mathbf{f}$ is the $M \times 1$ vector of common factor scores
- ϵ is the $P \times 1$ vector of unique factors

. . . so, what's the problem?

. . . so, what's the problem?

Because the M common factors are latent. . . the individual factor scores ($\mathbf{f}_{mi}$) are unknown and indeterminate.

The goal of EFA is to estimate Λ in spite of this!

Behind the Scenes

Communality

- Akin to R^2 in multiple regression
- Can calculate h^2 for each observed variable (p)
- Is the proportion of that variable's variance explained by the model
- Is a ratio of the variance resulting from the common factors and from the unique factors:

$$h_p^2 = \frac{1 - \mathrm{VAR}(\epsilon_p)}{\mathrm{VAR}(Y_p)}$$

Uniqueness

- The amount of variance *not* account for or explained by the factors:

$$u_p^2 = 1 - h_p^2$$

Estimation

The correlation structure for the P observed variables implied by the factor model is:

$$\hat{\mathbf{P}} = \Lambda \Psi \Lambda' + \Theta$$

- $\hat{\mathbf{P}}$ is the $P \times P$ model-implied correlation matrix for the population
  - If the model is correct in the population, $\hat{\mathbf{P}}$ will equal $\mathbf{P}$
- Λ is the same $P \times M$ matrix of factor loadings
- Ψ is the $M \times M$ matrix of correlations among the common factors
- Θ is a diagonal matrix with diagonal values equal to the uniqueness of the individual observed variables

Estimation

The correlation structure for the P observed variables implied by the factor model is:

$$\hat{\mathbf{P}} = \Lambda\Psi\Lambda' + \Theta$$

- $\hat{\mathbf{P}}$ is the $P \times P$ model-implied correlation matrix for the population
  - If the model is correct in the population, $\hat{\mathbf{P}}$ will equal $\mathbf{P}$
- Λ is the same $P \times M$ matrix of factor loadings
- Ψ is the $M \times M$ matrix of correlations among the common factors
- Θ is a diagonal matrix with diagonal values equal to the uniqueness of the individual observed variables

The factor scores themselves do not appear in this formulation!

Correlational Structure

This is the trick:

- We don't really care about factor scores, so we put them aside
- Want to find a set of parameter values for Λ, Ψ, and Θ that produces a model-implied correlation matrix, $\hat{\mathbf{P}}$ that matches our sample correlation matrix, $\mathbf{R}$, given that it is itself an estimate of the population correlation matrix $\mathbf{P}$.
- Don't actually need raw data values – just the correlation matrix!

Estimation

...Unfortunately, estimating $\hat{\mathbf{P}}$ is pretty difficult.

This has led to many different "factor extraction" techniques (and many simulation studies):

- Principal axis extraction
- Unweighted least-squares estimation
- Generalized least-squares
- Maximum likelihood estimation

Estimation

Begins with the researcher choosing M (the number of common factors) and an extraction method, which generates a starting value for the communality estimates (usually the squared multiple correlations).

Iteration...

1. Estimate factor loadings (given communality estimates)
2. Estimate the communalities (given the factor loadings)
3. Repeat until communalities stop fluctuating

Estimation

Maximum Likelihood fitting function:

$$F_{ML} = \log|\hat{\mathbf{P}}| + \mathrm{tr}(\mathbf{R}\hat{\mathbf{P}}^{-1}) - \log|\mathbf{R}| - P$$

Uses our guess at $\hat{\Lambda}$ and $\hat{\Theta}$ to minimize F_{ML}.
This boils down to a comparison of $\mathbf{R}$ with $\hat{\mathbf{P}}$ and
if $\hat{\mathbf{P}} = \mathbf{R}$, $F_{ML} = 0$.

Estimation Problems

1. Communalities greater than 1 (Heywood case)
2. Non-convergence - iteration fails to settle on a solution

Most often appear when there is:

- linear dependence among the observed variables;
- too many common factors; or,
- sample size is too small.

But is it good?

Often researchers will test $M = 1, 2, 3 \ldots$ This choice should be based upon a variety of criteria.

- Most tests involve looking at the eigenvalues of the correlation matrix (which characterizes the amount of information contained in a factor relative to the overall covariation among the observed variables)
- Interpretational quality often regarded as most important criterion

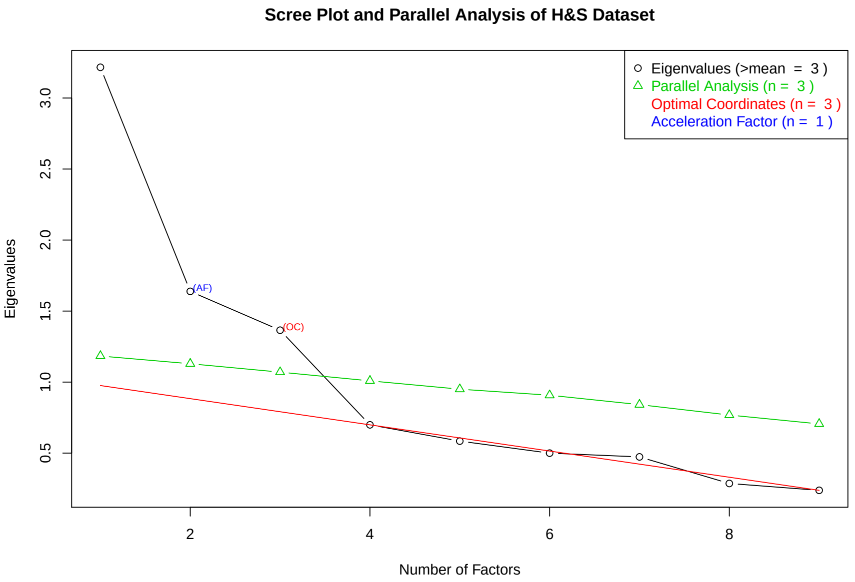
Kaiser Criterion

- Eigenvalues in $\mathbf{R} > 1$
- Default in SPSS and SAS
- Not recommended

Visual Tests

- Scree plots:
  - line chart of eigenvalues of **R** against their ranks in terms of magnitude
  - Look for the "bend" where not much more information is gained
- Parallel analysis:
  - Addition to scree plot, provides a less ambiguous guideline
  - Eigenvalues from **R** are compared vs random simulated data
  - Limit M to when the original line does not give more information than random data

Visual Tests



Statistical Tests and Fit Indices

- Standardized Root-Mean-Square residual (SRMR)
- χ^2 test of exact-fit (almost always significant. . .)
- Root-mean-square error of approximation (RMSEA; smaller)
- Akaike Information Criterion (AIC; smaller)
- Bayesian Information Criterion (BIC; smaller)
- Tucker-Lewis Index (TLI; larger)

But is it good?

A good factor...

- Makes sense
- Will be easy to interpret
- Possesses "simple structure"
- Items have low cross-loadings

How are factor loading interpreted?

Rotation

When $M \geq 2$, there are an infinite number of factor loading matrices that could explain the relations $\rightarrow$ **rotational indeterminacy**.
Initial $\hat{\Lambda}$ estimates are almost always difficult to interpret and needs to be rotated to enhance conceptual understanding.

	ML1	ML2	ML3
x1	0.49	0.31	0.39
x2	0.24	0.17	0.40
x3	0.27	0.41	0.47
x4	0.83	-0.15	-0.03
x5	0.84	-0.21	-0.10
x6	0.82	-0.13	0.02
x7	0.23	0.48	-0.46
x8	0.27	0.62	-0.27
x9	0.38	0.56	0.02

Rotation

The goal of rotation is to fit a geometric projection of the loadings where some are strong and others are near zero for each factor.

- The absolute distance between any two points stays the same.
- Rotation does not affect communality estimates or the predicted/residual correlation matrices.

$$\hat{\Lambda}_r = \hat{\Lambda}\mathbf{T}$$

Types of Rotation

Orthogonal

The transformation matrix $\mathbf{T}$ is a square, orthogonal matrix ($\mathbf{T}\mathbf{T}' = \mathbf{I}$).

- Varimax is most popular (default in SPSS and SAS)
- Ensures that factors remain uncorrelated ($\hat{\Psi} = \mathbf{I}$)
- Not encouraged!

Oblique

More realistic that factors are correlated to some extent. Oblique rotations define ($\hat{\Psi} = \mathbf{T}^{-1}\mathbf{T}'^{-1}$).

- Promax and oblimin rotations are most commonly used
- Oblimin weight can be modified to balance between row and column parsimony.

Types of Rotation

	ML1	ML2	ML3
x1	0.49	0.31	0.39
x2	0.24	0.17	0.40
x3	0.27	0.41	0.47
x4	0.83	-0.15	-0.03
x5	0.84	-0.21	-0.10
x6	0.82	-0.13	0.02
x7	0.23	0.48	-0.46
x8	0.27	0.62	-0.27
x9	0.38	0.56	0.02

(a) None

	ML1	ML3	ML2
x1	0.28	0.62	0.15
x2	0.10	0.49	-0.03
x3	0.03	0.66	0.13
x4	0.83	0.16	0.10
x5	0.86	0.09	0.09
x6	0.80	0.21	0.09
x7	0.09	-0.07	0.70
x8	0.05	0.16	0.71
x9	0.13	0.41	0.52

(b) Varimax

	ML1	ML2	ML3
x1	0.15	0.04	0.61
x2	0.01	-0.12	0.52
x3	-0.11	0.03	0.70
x4	0.84	0.00	0.01
x5	0.90	0.01	-0.08
x6	0.81	-0.01	0.07
x7	0.04	0.74	-0.21
x8	-0.05	0.72	0.05
x9	0.01	0.48	0.34

(c) Promax

	ML1	ML3	ML2
x1	0.19	0.60	0.03
x2	0.04	0.51	-0.12
x3	-0.07	0.69	0.02
x4	0.84	0.02	0.01
x5	0.89	-0.07	0.01
x6	0.81	0.08	-0.01
x7	0.04	-0.15	0.72
x8	-0.03	0.10	0.70
x9	0.03	0.37	0.46

(d) Oblimin

Table 2: Rotation and Factor Loadings

How are factor loadings interpreted?

Tableplots:

	x1	x2	x3	x4	x5	x6	x7	x8	x9
ML1	19	4	−7	84	89	81	4	−3	3
ML3	60	51	69	2	−7	8	−15	10	37
ML2	3	−12	2	1	1	−1	72	70	46

Where does this fit in?

Structural Equation Modelling

A general framework encompassing a wide variety of methods and models represented via path diagrams.

- EFA: I don't know what is going on
- CFA: Let's test what is going on
- Path Analysis: I think these things are related in a particular way but only the things I see are real
- Latent Variable Modelling: Fully generalizable framework that incorporates both latent and manifest variables

Where does this fit in?

Structural Equation Modelling

A general framework encompassing a wide variety of methods and models represented via path diagrams.

- EFA: I don't know what is going on
- CFA: Let's test what is going on
- Path Analysis: I think these things are related in a particular way but only the things I see are real
- Latent Variable Modelling: Fully generalizable framework that incorporates both latent and manifest variables

PCA is not on this list for a good reason!

- Principal Components Analysis only does data reduction
- Assumes that variables are measured without error

Finally. . .

EFA is a process

- Solutions should replicate with new samples
- Over a series of studies:
  - Develop a good idea of how variables relate to underlying factors
  - Formulate specific hypotheses about the values of the coefficients
  - Can conduct CFA to test structure (constrain λ values to 0)