Exploratory Factor Analysis: common factors, principal components, and more

Psychology 588: Covariance structure and factor models

Common factor model

- As learned earlier for CFA, both CFA and EFA are of (almost) the same model form, which defines q indicator variables as linear combinations of n "common" factors plus q mutually orthogonal "unique" factors as follows:

$$\mathbf{x} = \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta}, \quad E\left(\boldsymbol{\xi}\boldsymbol{\delta}'\right) = \mathbf{0}_{n\times q}, \quad E\left(\delta_i \delta_j\right) = 0 \text{ for } i \neq j$$

Given $E\left(\mathbf{x}\right) = \mathbf{0}_{q\times 1}, \quad E\left(\boldsymbol{\xi}\right) = \mathbf{0}_{n\times 1}, \quad E\left(\boldsymbol{\delta}\right) = \mathbf{0}_{q\times 1}$

- Difference between CFA and EFA resides in whether some selective elements of loading matrix $\boldsymbol{\Lambda}$ are constrained at particular constants (e.g., at 0 or by equality) --- consequently, the parameters are estimated by different methods

- Technically, CFA and EFA differ by the degree of constraints
  - EFA --- loading matrix has minimal constraints for a unique solution with fixed orientation of factors

    To fix orientation of factor axes, $n \times n$ "constraints" need be imposed, and that is done by n scaling constraints, $n(n-1)/2$ elements for orthogonal factors ξ and $n(n-1)/2$ elements for the loading matrix Λ (i.e., canonical form)
  - CFA --- further constraints (motivated by theoretical hypotheses) imposed to see if they agree with the data
- Furthermore, CFA may allow correlated measurement errors (so long as all parameters identifiable), but EFA doesn't allow such relaxation by definition

Implied cov matrix of the common factor model

$$\boldsymbol{\Sigma} \equiv E(\mathbf{x}\mathbf{x}') = E(\boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta})(\boldsymbol{\xi}'\boldsymbol{\Lambda}' + \boldsymbol{\delta}'), \quad E(\mathbf{x}) = \mathbf{0}$$

$$= \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}$$

$$\boldsymbol{\Phi} \equiv E(\boldsymbol{\xi}\boldsymbol{\xi}')$$

$$\boldsymbol{\Theta} \equiv E(\boldsymbol{\delta}\boldsymbol{\delta}') \equiv \text{diag}\left(\theta_1, \theta_2, ..., \theta_q\right)$$

- It's a convention in EFA to scale the common factors to have a variance of 1 (instead of setting their metric equal to one of their indicators) --- by this scaling, $\boldsymbol{\Phi}$ becomes a correlation matrix
- When considered for realized data (i.e., "subjects"), $\boldsymbol{\xi}$ is often called "factor score" matrix

Orthogonal factor model

- Common factors $\boldsymbol{\xi}$ are constrained to be mutually orthogonal, and so with the unit-variance scaling, factors becomes mutually orthogonal z-scores:

$$\mathbf{x} = \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta}, \quad E(\boldsymbol{\xi}\boldsymbol{\xi}') = \mathbf{I}$$

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}$$

- Thanks to the rotational indeterminacy, orthogonal common factors are typically estimated in a canonical form for a computationally unique (unrotated) solution without loss of generality

- Principal components might be considered as such orthogonal unrotated factors

Variance and covariance by the model

- From the orthogonal factor model, $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda} + \boldsymbol{\Theta}$:

communality $\left(\equiv h_i^2 \right)$

uniqueness

$$\operatorname{var}(\mathbf{x}_i) = \lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{in}^2 + \theta_i, \quad i = 1,...,q$$

$$\operatorname{cov}(\mathbf{x}_i, \mathbf{x}_j) = \lambda_{i1}\lambda_{j1} + \lambda_{i2}\lambda_{j2} + \cdots + \lambda_{in}\lambda_{jn}, \quad i \neq j = 1,...,q$$

$$\operatorname{var}(\xi_k) = \lambda_{1k}^2 + \lambda_{2k}^2 + \cdots + \lambda_{qk}^2, \quad k = 1,...,n$$

$$\operatorname{cov}(\mathbf{x}, \boldsymbol{\xi}) = E\left((\boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta})\boldsymbol{\xi}' \right) = \boldsymbol{\Lambda}$$

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1n} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{q1} & \lambda_{q2} & \cdots & \lambda_{qn} \end{bmatrix} \begin{matrix} h_1^2 \\ h_2^2 \\ \vdots \\ h_q^2 \end{matrix}$$

$$\operatorname{var}(\xi_k)$$

Principal components as special orthogonal factors

- Principal components are defined as linear combinations of $\mathbf{x}$ that are mutually orthogonal and successively account for maximum variance of the data

  - $n = q, \quad \mathbf{y}_{q\times 1} = \mathbf{V}'_{q\times q}\mathbf{x}_{q\times 1}, \quad \mathbf{V}'\mathbf{V} = \mathbf{V}\mathbf{V}' = \mathbf{I}_q$

    $$E\left(\mathbf{y}\mathbf{y}'\right) = \mathrm{diag}\left(\mathrm{var}(y_1),\ldots,\mathrm{var}(y_q)\right)$$

    $$\mathrm{tr}\left(\mathbf{y}\mathbf{y}'\right) = \mathrm{tr}\left(\mathbf{x}\mathbf{x}\right)$$

  - $n < q, \quad \mathbf{y}_{n\times 1} = \mathbf{V}'_{n\times q}\mathbf{x}_{q\times 1}, \quad \mathbf{V}'\mathbf{V} = \mathbf{I}_n, \quad \mathbf{V}\mathbf{V}' \neq \mathbf{I}_q$

    $$\mathrm{tr}\left(\mathbf{y}\mathbf{y}'\right) < \mathrm{tr}\left(\mathbf{x}\mathbf{x}\right)$$

- By spectral decomposition of $\mathbf{\Sigma}$,

$$\mathbf{\Sigma} = \mathbf{VEV}', \quad \mathbf{V}'\mathbf{V} = \mathbf{VV}' = \mathbf{I}_q$$

$$\mathbf{E} = \mathrm{diag}\left(e_1,...,e_q\right), \quad e_1 \geq ... \geq e_q$$

where $e_1, \ldots, e_q$ are eigenvalues in descending order

- If $n < q$, sum of variances of n components is the maximum among all sets of n linear combinations; geometrically speaking, the n components span a subspace of the original q-dimensional data space, on which the projections of the data points have a maximum variance --- Eckart-Young theorem

- By singular value decomposition of $\mathbf{X}$:

$$\mathbf{X}_{q\times N} = \mathbf{V}\mathbf{S}\mathbf{U}', \quad \mathbf{V}'\mathbf{V} = \mathbf{U}'\mathbf{U} = \mathbf{I}_q$$

$$\mathbf{S} = \mathrm{diag}\left(s_1,...,s_q\right) = \mathbf{E}^{0.5}, \quad s_1 \geq ... \geq s_q$$ --- singular values

- Rank-n approximation by SVD:

$$\mathbf{X} = \mathbf{V}\mathbf{S}\mathbf{U}' = \mathbf{V}_1\mathbf{S}_1\mathbf{U}_1' + \mathbf{V}_2\mathbf{S}_2\mathbf{U}_2'$$

$$\mathbf{Y}_{n\times N} = \mathbf{S}_{1(n\times n)}\mathbf{U}'_{1(n\times N)} \quad \text{or} \quad \mathbf{y}_{n\times 1} = \mathbf{S}_{1(n\times n)}\mathbf{u}_{1(n\times 1)}$$

- VAF by n components: $\mathrm{tr}\left(\mathbf{Y}\mathbf{Y}'\right) = \mathrm{tr}\left(\mathbf{S}_1^2\right) = \sum_{k=1}^{n} e_k < \mathrm{tr}\left(\mathbf{X}\mathbf{X}'\right)$

Common factors vs. principal components

- While two conditions of the PC model, dual orthogonality and successive maximization, guarantee a unique solution (identifiability), the orthogonality on the component weights is totally arbitrary under the CF model and so undone by “rotation”

- There are two ways of looking at principal components:
  - Pearson’s view --- components as n linear combinations (functions) of data variables, $\mathbf{y} = \mathbf{V}'\mathbf{x}$
  - Hotelling’s view --- components as n explanatory variables of data variables, $\mathbf{x} \approx \mathbf{V}\mathbf{y}$

- With loss function of least-squares, PCA minimizes it w.r.t. $\mathbf{x}$ (like the OLS for regression) while CFA minimizes it w.r.t. $\boldsymbol{\Sigma}$ (i.e., the ULS)

Principal factor method

- Since eigenvectors (i.e., the “loadings” defining principal components) yield $\hat{\mathbf{x}}$ that is meant to maximally reproduce var($\mathbf{x}$) ($=\sum_i h_i^2 + \theta_i$ by the CF model), they don’t produce the least-squares estimates for covariances

- Alternatively, loadings may be obtained from a modified covariance matrix, $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} - \boldsymbol{\Theta}$

$$\tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} h_1^2 & \sigma_{21} & \cdots & \sigma_{q1} \\ \sigma_{21} & h_2^2 & \cdots & \sigma_{q2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{q1} & \sigma_{q2} & \cdots & h_q^2 \end{bmatrix}$$

- In that the unique variances (hence communalities) are not known, the adjustment should be made iteratively, typically with initial communality estimates of variances of x_i predicted by all other variables (by the OLS regression)

- Given such communality estimates, the loading matrix can be obtained by the spectral decomposition of $\tilde{\boldsymbol{\Sigma}}$, and communalities are updated with the new estimates of loadings --- these alternating updates iterate until the loading estimates converge

- One complication due to the adjustment is, so called, Heywood case of negative error variance --- common to most extraction methods of common factors, not only to the principal factors method; and more likely to occur with small q/n ratio (i.e., consequence of overfitting)

- Under the common factor model, communalities are a function of factor loadings (i.e., $h_i^2 = \boldsymbol{\lambda}_i' \boldsymbol{\lambda}_i$), and so covariances (i.e., off-diagonal elements of $\boldsymbol{\Sigma}$) contain all information on "common" factors

- Accordingly, MINRES minimizes residuals of only the off-diagonal entries of $\boldsymbol{\Sigma}$ as:

$$\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} - \boldsymbol{\Lambda}\boldsymbol{\Lambda}' - \boldsymbol{\Theta} = \boldsymbol{\Sigma}^* - \left(\boldsymbol{\Lambda}\boldsymbol{\Lambda}' - \mathbf{H}^2 \right)$$

$$\boldsymbol{\Sigma}^* = \begin{bmatrix} 0 & \sigma_{21} & \cdots & \sigma_{q1} \\ \sigma_{21} & 0 & \cdots & \sigma_{q2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{q1} & \sigma_{q2} & \cdots & 0 \end{bmatrix}, \quad \mathbf{H}^2 = \operatorname{diag}\left(h_1^2, ..., h_q^2 \right)$$

- Note that the common factor model is a recursive model and so the diagonal entries of Σ have no residuals --- in this regard, MINRES optimizes the least-squares function w.r.t. only the relevant quantities

- MINRES is the best analytic (non-parametric) factoring method for the common factor model

- For those familiar with MATLAB, a function file called “minresfac.m” available in my netfiles under “\data” (with syntax, “lambda_hat = minresfac(data_cov,n)”); since the solution is only locally optimal, multiple runs would be needed

Maximum likelihood method

- With normally distributed $\boldsymbol{\xi}$ and $\boldsymbol{\delta}$, their linear combinations $\mathbf{x}$ is also normally distributed (or put differently, normally distributed $\mathbf{x}$ indicates normal $\boldsymbol{\xi}$ and $\boldsymbol{\delta}$), and so its likelihood function is known (essentially the same as for CFA)

$$\log L = -\frac{N-1}{2}\left(\log\left|\hat{\boldsymbol{\Sigma}}\right| + \mathrm{tr}\left(\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{S}\right)\right) + c$$

$$c = \text{ function independent of } \hat{\boldsymbol{\Sigma}}$$

- Unlike analytic factoring methods (e.g., principal factoring and MINRES), a χ^2 test is available for the model fit with the ML estimator since the sampling distribution of its χ^2 estimate is known based on the assumed data distribution

- Note that the likelihood function is scale free, and so it would not make any difference whether covariance or correlation data are used

- One technical difference --- the ML estimator uses a side condition of $\mathbf{\Lambda' S^{-1} \Lambda}$ being diagonal instead of the usual canonical form for a unique solution

- Residuals of the covariance matrix are generally better minimized by those implementing the common factor model properly (such as MINRES and ML); while PCA (and principal factors) tend to maximize total variance of $\mathbf{x}$

- The following are results from 4 different factoring methods of men's track and field records of 55 countries on 8 Olympic games (data given in the course website as "trackm.xls")

Communalities for the track data for men (taken from Johnson & Wichern, Applied multivariate statistical analysis 6th, 2007) by 4 factoring methods; Correlation matrix is analyzed for an equal contribution by the games

	PCA	PF	ML	MINRES
100m	0.950	0.942	0.919	0.925
200m	0.940	0.930	0.924	0.913
400m	0.892	0.871	0.849	0.855
800m	0.900	0.895	0.865	0.878
1500m	0.938	0.942	0.918	0.925
5000m	0.965	0.980	0.966	0.962
10000m	0.974	0.997	0.982	0.979
Marathon	0.943	0.921	0.914	0.904
sum	7.502	7.478	7.337	7.341
MAR	0.013	0.014	0.007	0.006

- MAR: mean absolute residuals of correlations
- What's the df of the ML fitting ($n = 2$)?