STAT 730 Chapter 9: Factor analysis

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 730: Multivariate Data Analysis

Basic idea

Factor analysis attempts to explain the correlation among a large number of variables using a small number of latent *factors*.

Example: Spearman (1904) considered children's exam performance in classics x_1, French x_2, and English x_3. The estimated correlation matrix is $\mathbf{R} = \begin{bmatrix} 1 & 0.83 & 0.78 \\ 0.83 & 1 & 0.67 \\ 0.78 & 0.67 & 1 \end{bmatrix}$. One can postulate a model with a common factor

$$\begin{aligned} x_1 &= \mu_1 + \lambda_1 f + u_1 \\ x_2 &= \mu_2 + \lambda_2 f + u_2 \\ x_3 &= \mu_3 + \lambda_3 f + u_3. \end{aligned}$$

In matrix terms

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda} f + \mathbf{u}.$$

Here, f is the common underlying factor (overall ability) that explains all three scores.

The factor model

Let $\mathbf{x} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The factor model is

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{f} + \mathbf{u},$$

where $\mathbf{x} \in \mathbb{R}^p$ is the response, $\boldsymbol{\Lambda} \in \mathbb{R}^{p \times k}$ are the factor loadings, $\mathbf{f} \in \mathbb{R}^k$ are the common factors (hopefully k is small, e.g. $k = 1$ or $k = 2$ if we're lucky), and $\mathbf{u}$ is error, also called 'specific' or 'unique' factors. It is assumed

$$\mathbf{f} \sim (\mathbf{0}, \mathcal{I}_k), \ \mathbf{u} \sim (\mathbf{0}, \boldsymbol{\Psi}), \boldsymbol{\Psi} = \text{diag}(\psi_{11}, \ldots, \psi_{pp}), \ C(\mathbf{f}, \mathbf{u}) = \mathbf{0}.$$

Note then that

$$V(\mathbf{x}) = \boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}.$$

This must hold if the k-factor model is true.

Motivation from PCA

$$
\begin{aligned}
\mathbf{x} &= \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\Gamma}'(\mathbf{x}-\boldsymbol{\mu}) \\
&= \boldsymbol{\mu} + [\boldsymbol{\Gamma}_k \boldsymbol{\Gamma}_{p-k}] \begin{bmatrix} \boldsymbol{\Gamma}_k' \\ \boldsymbol{\Gamma}_{p-k}' \end{bmatrix} (\mathbf{x}-\boldsymbol{\mu}) \\
&= \boldsymbol{\mu} + [\boldsymbol{\Gamma}_k \boldsymbol{\Gamma}_{p-k}] \begin{bmatrix} \boldsymbol{\Gamma}_k'(\mathbf{x}-\boldsymbol{\mu}) \\ \boldsymbol{\Gamma}_{p-k}'(\mathbf{x}-\boldsymbol{\mu}) \end{bmatrix} \\
&= \boldsymbol{\mu} + \underbrace{\boldsymbol{\Gamma}_k}_{\boldsymbol{\Lambda}} \underbrace{\boldsymbol{\Gamma}_k'(\mathbf{x}-\boldsymbol{\mu})}_{\mathbf{f}} + \underbrace{\boldsymbol{\Gamma}_{p-k}\boldsymbol{\Gamma}_{p-k}'(\mathbf{x}-\boldsymbol{\mu})}_{\mathbf{u}}.
\end{aligned}
$$

This in fact approximately holds when $V(\mathbf{u})$ is small, which happens when the first k principal components explain most of the variability. See MKB Section 9.8 (pp.275–276).

Note that under the general factor model, $\mathbf{u}$ is random scatter in all directions, not necessarily orthogonal to $\boldsymbol{\Lambda}\mathbf{f}$, so this is motivation but the models are different. However, this motivation for factor analysis leads to choosing k based on a PCA scree plot.

Communalities

Recall $V(\mathbf{x}) = \mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}$. Thus

$$V(x_i) = \sigma_{ii} = \underbrace{\sum_{j=1}^{k} \lambda_{ij}^2}_{h_i^2} + \psi_{ii}.$$

h_i^2 is the ith communality, the amount of variance of x_i shared with other variables through the common factors; what is left over is the specific (or unique) variance ψ_{ii}.

Scale invariance

Let $\mathbf{y} = \mathbf{C}\mathbf{x}$ where $\mathbf{C}$ is a diagonal matrix and $\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{f} + \mathbf{u}$ is a k-factor model as on the last slide. Then

$$\mathbf{y} = \mathbf{C}\boldsymbol{\mu} + \mathbf{C}\boldsymbol{\Lambda}\mathbf{f} + \mathbf{C}\mathbf{u},$$

where $V(\mathbf{y}) = [\mathbf{C}\boldsymbol{\Lambda}][\mathbf{C}\boldsymbol{\Lambda}]' + \mathbf{C}\boldsymbol{\Psi}\mathbf{C}$. The factor model thus holds for $\mathbf{y}$.

Unlike PCA, FA is unaffected by rescaling the outcomes.

Non-uniqueness of $\mathbf{\Lambda}$ and $\mathbf{f}$

Let $\mathbf{G} \in \mathbb{R}^{k \times k}$ s.t. $\mathbf{G}\mathbf{G}' = \mathbf{G}'\mathbf{G} = \mathcal{I}_k$ and $\mathbf{x} = \boldsymbol{\mu} + \mathbf{\Lambda}\mathbf{f} + \mathbf{u}$ is a k-factor model. Then

$$\mathbf{x} = \boldsymbol{\mu} + [\mathbf{\Lambda}\mathbf{G}][\mathbf{G}'\mathbf{f}] + \mathbf{u}.$$

To achieve identifiability, constraints such as $\mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{\Lambda} = \mathbf{D}$ where $\mathbf{D}$ is diagonal are used. After the factors are estimated $\hat{\mathbf{\Lambda}}$, they are further rotated by $\mathbf{G}$ to $\mathbf{\Delta} = \hat{\mathbf{\Lambda}}\mathbf{G}$ in order to maximize

$$\phi = \sum_{i=1}^{k}\sum_{j=1}^{p}(d_{ij}^2 - \bar{d}_i)^2, \; d_{ij} = \frac{\delta_{ij}}{\sum_{s=1}^{k}\delta_{is}^2}, \bar{d}_j = \frac{1}{p}\sum_{i=1}^{p} d_{ij}^2.$$

This is called *varimax* rotation, developed by Kaiser (1958), and produces loadings $\mathbf{\Delta}$ with a few large loadings and many near-zero loadings. Such rotations are more interpretable, as discussed in class for PCA. An alternative rotation is the *promax*, which produces *correlated* factors; it is varimax followed by an oblique (not orthogonal) Procrustes rotation.

Exploratory vs. confirmatory FA

Near-zero loadings lead to a natural question: *are they actually zero?* A researcher might have in mind one or more factors (math ability, reading ability) and hypothesize that these factors are related to only a subset of observed variables (division, comprehension). Furthermore, the factors may or may not be correlated.

Exploratory FA seeks to find out how many factors are needed and get a rough idea of which factors highly load (are related) to variables. There are no particular hypotheses in mind at this stage. Confirmatory FA seeks to validate (or form) a theoretical hypothesis. Specific (causal) regressions are stipulated and an overall model is fit, essentially setting some loadings to zero, forcing correlations to be zero, etc. The number of factors are given in advance, as well as their relationship to the observed variables. The hypothetical model is then checked for overall fit, thereby "confirming" or "denying" a given hypothesis.

Fitting

When $\mathbf{\Sigma}$ is unconstrained there are $\frac{1}{2}p(p+1)$ free parameters; $\mathbf{\Lambda}$ has kp parameters and $\mathbf{\Psi}$ has p. The constraint $\mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{\Lambda}$ is diagonal eats up $\frac{1}{2}k(k+1)$ parameters, so the total number of "free" parameters ends up being

$$s = \tfrac{1}{2}(p-k)^2 - \tfrac{1}{2}(p+k).$$

One wants $s > 0$, which will happen when $k << p$; $s > 0$ implies that the factor model offers a simpler interpretation than allowing $\mathbf{\Sigma}$ to be completely arbitrary.

When $s > 0$ the factor model can be estimated, yielding $\hat{\mathbf{\Lambda}}$ and $\hat{\mathbf{\Psi}}$ from data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$. There are two main methods:

- Principal factor analysis (pp. 261–263).
- Maximum likelihood (pp. 263–267) assuming $\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{iid}{\sim} N_p(\boldsymbol{\mu}, \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi})$. The MLE approach also allows for a the test $H_0 : k$ is adequate vs. $H_a : \mathbf{\Sigma}$ is unconstrained. This test is described on pp. 265–268 and automatically provided by `factanal` in R.

MLE factor analysis

```
# Spearman's data
R=matrix(c(1,0.83,0.78,0.83,1,0.67,0.78,0.67,1),3,3)
f=factanal(covmat=R,factors=1,rotation="varimax") # varimax, promax, or none
print(f,digits=2,cutoff=.3,sort=TRUE) # compare to bottom p. 260

# open/closed book exams
library(bootstrap)
data(scor) # note that rotation is unnecessary when k=1
plot(prcomp(scor),type="l") # k=1?
f=factanal(scor,factors=1,rotation="varimax")
print(f,digits=2,cutoff=.3,sort=TRUE) # compare to bottom p. 266
```

The option `scores="regression"` or `scores="Bartlett"` produce factor scores (see MKB p. 274). Use the `covmat=` option to enter a correlation or covariance matrix directly. If entering a covariance matrix, include the option `n.obs=` for tests.

Structural equation modeling

Structural equation modeling (SEM) generalizes (confirmatory) factor analysis. The `sem` package was the first package to allow the routine fitting of SEM in R; recently the `lavaan` (latent variable analysis) package was introduced with much more functionality, rivaling the commercial packages `Mplus` and `LISREL`. Path diagrams can be created using `semPlot` using lavaan modeling syntax.

SEM

Structural equations are simply a series of related regression models involving observed and latent variables. The distinction between predictor and response gets blurred as variables can appear on either side of the regression equations, although they may appear on the "left" as a dependent variable only once. General Gaussian covariance structures are allowed for the errors. Maximum likelihood (based on normality) factor analysis is a special case.

SEM are constructed using a hypothesized "causal pathway" but it is important to note that one cannot infer causation from observational data. SEMs provide a structured means to explore possible simple causal paths, but do not confirm their existence.

LISREL model: measurement submodel

Jöreskog and Sörbom came up with the LISREL (linear structural relations) model, as well as the software which is currently over $700; the cost of Mplus is similar. Note that `lavaan` is free. Have two sets of latent factors now (endogenous and exogenous), associated with two sets of observed variables through linear regressions. We observe $(\mathbf{y}_i, \mathbf{x}_i)$ on subject i. The measurement equations are

$$\mathbf{y}_i = \mathbf{\Lambda}^y \mathbf{f}_i^y + \mathbf{u}_i^y,$$
$$\mathbf{x}_i = \mathbf{\Lambda}^x \mathbf{f}_i^x + \mathbf{u}_i^x.$$

The latent "cause-and-effect" factors $\mathbf{f}_i^y$ and $\mathbf{f}_i^x$ are endogenous and exogenous latent variables. In general, endogenous (independent) variables are explained by other variables through linear regression; these are on the left-hand side and include $\mathbf{y}_i$ and $\mathbf{x}_i$. Exogenous variables are on the right-hand side and are (conditionally) fixed explanatory variables. In general, variables can be either, but should appear on the left-side only once.

LISREL model: structural submodel

The structural submodel is

$$\mathbf{f}_i^y = \mathbf{B}^y \mathbf{f}_i^y + \mathbf{B}^x \mathbf{f}_i^x + \mathbf{u}_i^m.$$

Here, $\mathbf{B}^y$ has zeros along the diagonal, $\mathbf{u}_i^x, \mathbf{u}_i^y, \mathbf{u}_i^m$ are mutually uncorrelated and mean-zero, and $C(\mathbf{x}_i, \mathbf{u}_i^m) = \mathbf{C}(\mathbf{y}_i, \mathbf{u}_i^y) = \mathbf{C}(\mathbf{u}_i^x, \mathbf{f}_i^x) = \mathbf{0}$. Also, $\mathbf{u}_i^m \overset{iid}{\sim} (\mathbf{0}, \boldsymbol{\Psi})$ indep. $\mathbf{f}_i^x \overset{iid}{\sim} (\mathbf{0}, \boldsymbol{\Phi})$. The off-diagonals of $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$ can be non-zero reflecting correlations among the latent endogenous and exogenous factors.

Confirmatory FA places only uses $\mathbf{x}_i$ and places structure on $\boldsymbol{\Lambda}^x$ as well as $\boldsymbol{\Phi}$.

`lavaan` can fit much more general models. For example, note that $E(\mathbf{x}_i) = E(\mathbf{y}_i) = \mathbf{0}$ above. We may also want to regress some of the observed variables onto other observed variables, as well as latent variables.

Measures of model fit

- χ^2 simply looks at the observed and estimated covariance matrices, yielding a p-value. However, the test may inappropriately reject in large samples and inappropriately not reject in small samples, so other measures of fit are often used.
- The root mean square error of approximation (RMSEA) also looks at the difference between the fitted model and observed covariance. RMSEA ranges from 0 to 1; smaller is better. `lavaan` provides a p-value for H_0 : RMSEA ≤ 0.05 (want to accept).
- The standardized root mean square residual (SRMR) is yet another discrepancy between the observed and fitted model covariance matrices, and ranges from 0 to 1. Researchers look for SRMR ≤ 0.08.

All of these measures are available in `lavaan`.