36-722: Factor Analysis

Brian Junker

October 6, 2005

- The Orthogonal Factor Model
- Interpretation and Exemplars
- Scale Invariance
- Factor Indeterminacy
- Likelihood Ratio Test of Fit
- Factor Rotation
- Two Approaches to Factor Analysis

The Orthogonal Factor Model

When we considered the "scree plot" (plot of eigenvalues in a pca decomposition) as a tool for deciding how much smoothing is appropriate, we considered a model of the form

$$x_i = Az_i + \varepsilon_i$$

A (slight) generalization of this model is the orthogonal factor model:

The *orthogonal factor model* for the $p \times 1$ random vector $X = (X_1, \ldots, X_p)^T$ is

$$X = \mu + QF + U$$

where

- $E[X] = \mu$; $E[F] = 0$; $E[U] = 0$
- $\mathrm{Var}(F) = I_{k \times k}$ *(F = "common factors")*;
- $\mathrm{Var}(U) = \Psi = diag(\psi_{11}, \ldots, \psi_{pp})$ *(U = "unique (specific) factors")*;
- $\mathrm{Cov}(F, U) = 0$.
- $Q = [q_{j\ell}]_{p \times k}$; *($Q$ = "factor loadings").*

The *observable* variables X_j are represented in terms of the *latent variables* F_ℓ and U_j as

$$X_j = \sum_{\ell=1}^{k} q_{j\ell} F_\ell + U_j + \mu_j$$

- Clearly

$$\begin{aligned} \Sigma &= Var(X) = E[(X-\mu)(X-\mu)^T] \\ &= E[(QF+U)(QF+U)^T] = E[QFF^TQ^T] + E[UU^T] \\ &= Q\mathrm{Var}(F)Q^T + \mathrm{Var}(U) = QQ^T + \Psi \end{aligned}$$

- $\sigma_{X_jX_j} = \sum_{\ell=1}^{k} q_{j\ell}^2 + \psi_{jj}$
  - $h_j^2 = \sum_{\ell=1}^{k} q_{j\ell}^2$ is called the "*communality*";
  - ψ_{jj} is called the "*specific variance*"

  Similarly, $\sigma_{X_j,X_k} = \sum_{\ell=1}^{k} q_{j\ell}q_{k\ell}$;
- The loadings $q_{j\ell}$ control the relationship between the X_j's and F_ℓ's. Indeed
  - $\Sigma_{XF} = \mathrm{Cov}(X,F) = E[(QF+U)F^T] = Q$
  - $\mathrm{Corr}(X,F) = D^{-1/2}Q$

Interpretation and Exemplars

- The Wechsler Adult Intelligence Scale (WAIS) includes four subscales (total scores on tests)

X_1 = Information, X_2 = Similarities, X_3 = Arithmetic, X_4 = Picture Completion

  - In many populations we might expect all of these subscales to be positively correlated, suggesting a one-factor FA model ($k = 1$) with positive factor loadings:

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix} = \begin{bmatrix} q_{11} \\ q_{21} \\ q_{31} \\ q_{41} \end{bmatrix} \begin{bmatrix} F_1 \end{bmatrix} + \begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \end{bmatrix} + \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix}$$

    The common factor F_1 might be labelled "general proficiency" or "general intelligence", and performance on each test X_j is a rescaled version $q_{j1}F_1$ of this general factor, plus noise U_j specific to the test or test-taking circumstances.

- If we observe that $\text{Corr}(X_1, X_3)$, $\text{Corr}(X_2, X_4) > 0$, and $\text{Corr}(X_1, X_2) = \text{Corr}(X_3, X_4) = 0$, we might try a more-refined model:

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix} = \begin{bmatrix} q_{11} & 0 \\ 0 & q_{22} \\ q_{31} & 0 \\ 0 & q_{42} \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \end{bmatrix} + \begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \end{bmatrix} + \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix}$$

  where now F_1 is the common factor underlying *Information* and *Arithmetic* performance, and F_2 is the common factor underlying *Similarities* and *Picture Completion*.
- This sort of model-building for mental testing scores is how Factor Analysis was developed by Spearman (one general intelligence factor), Cattell (scree plot) and Thurstone ("Vectors of the Mind") in the early 20^{th} century.
- The second model above exhibits an extreme version of "*simple structure*": Each observable X_j is related to only a few factors F_ℓ. Simple structure generally helps "interpretation" or "labelling" of factors.

- Factor analysis gets used in many other fields as well, where it is suspected that high-dimensional *observable* data is being driven by a low-dimensional *latent* process, plus noise. For example,
  - Marketing, to identify key salient features in consumer preference...
  - Chemometrics, especially mixture modeling and "noisy" spectral decomposition, identifying common components of sensor signals...
  - Stock market modeling, e.g. in identifying common elements of multiple markets...
  - Modeling of multivariate time series as a function of a smaller number of "latent" series plus white noise...
  - Smoothing in which the uniquenesses U_j are dropped and $X_{smooth} = QF + \mu$.

Scale Invariance

Suppose we rescale X, letting $Y = CX$, where C is a diagonal matrix (e.g. we change the units of measurement from \$'s to yen, from feet to meters, etc.). We know this can influence PCA. What does it do to factor analysis?

If $X = QF + U + \mu$ then we know $\mathrm{Var}(X) = QQ^T + \Psi$. For Y we have

$$\begin{aligned}
Y &= CX = CQF + U + \mu = Q'F + U' + \mu' \\
\mu' &= C\mu; \; Q' = CQ; \; U' = CU \\
\mathrm{Var}(F) &= I \\
\mathrm{Var}(U') &= C\mathrm{Var}(U)C^T = C\Psi C^T \\
\mathrm{Var}(Y) &= C\mathrm{Var}(X)C^T = C(QQ^T)C^T + C\Psi C^T = (Q')(Q')^T + \Psi'
\end{aligned}$$

So, *the factor loadings and the uniquenesses get rescaled but we get the same factors and factor structure as before*.

In this sense the FA model is like PCA for the correlation matrix (but we shall see that formal inference is easier for FA).

Factor Indeterminacy

Let us suppose

$$X = QF + U + \mu$$

Then for *any orthogonal (rotation) matrix* G,

$$\begin{aligned} X &= Q(GG^T)F + U + \mu = (QG)(G^T F) + U + \mu \\ &= Q'F' + U + \mu \qquad \text{where} \end{aligned}$$

- $Q' = QG$ may not look anything like Q!
- $F' = G^T F$ is still has $E[F'] = GE[F] = 0 = E[F]$ and $\text{Var}(F')G\text{Var}(F)G^T = GIG^T = I = \text{Var}(F)$ (since G^T is orthogonal)!

The *parameters* of the FA model are

- The means $\mu = (\mu_1, \ldots, \mu_p)^T$;
- The uniquenesses $diag(\Psi) = (\Psi_{11}, \ldots, \Psi_{pp})^T$;
- The factor loadings $q_{j\ell}$ in Q.
- And perhaps the dimensionality k of F.

So there is a *big nonidentifiability* problem in estimating the FA model!

Identifying the model; degrees of freedom

We wish to impose relatively simple constraints on the parameters that will force a unique solution (like linear constraints in ANOVA). One computationally convenient set of constraints is to require that

$$Q^T \Psi^{-1} Q \equiv \Delta = diag(\Delta_{11}, \ldots, \Delta_{kk})$$

another constraint sometimes used is to make $Q^T D^{-1} Q$ diagonal, where $D = diag(\Sigma)$.

- The unconstrained model $X = QF + U + \mu$ has $p \cdot k + p$ degrees of freedom: $p \cdot k$ for Q, and p for Ψ (diagonal).
- The constraint above sets $\frac{1}{2}k(k-1)$ elements of Δ equal to zero.

Therefore the d.f. for a LR test of $H_0 : X \sim N_p(\mu, QQ^T + \Psi)$, a k factor FA model, vs. $H_A : X \sim N_p(\mu, \Sigma)$, a general multivariate normal model, is

$$\begin{aligned} d &= (\#\text{ params in unconstrained } \Sigma) - (\#\text{ params in FA model for } \Sigma) \\ &= \frac{1}{2}p(p+1) - (pk + p - \frac{1}{2}k(k-1)) \\ &= \frac{1}{2}(p-k)^2 - \frac{1}{2}(p+k) \end{aligned}$$

Note that if $d \leq 0$ then the asymptotic χ^2 approximation to the LR test will fail; in fact: Σ can be fitted without error either uniquely ($d = 0$) or infinitely many ways ($d < 0$).

The case $d > 0$ is the more statistically interesting since it says that H_0 has less parameters (a more parsimonious model) than H_A.

Likelihood Ratio Test of Fit

Now we consider a data matrix X with rows $x_i \overset{iid}{\sim} N_p(\mu, \Sigma)$. As we have seen before,

$$\begin{aligned} \ell(X;\mu,\Sigma) &\equiv \ell(\mu,\Sigma) = \log L(X;\mu,\Sigma) \\ &= -\frac{n}{2}\log|2\pi\Sigma| - \frac{n}{2}tr\{\Sigma^{-1}S\} - \frac{n}{2}(\overline{x}-\mu)^T\Sigma^{-1}(\overline{x}-\mu) \end{aligned}$$

Replacing μ with its MLE $\overline{x}$ we have

$$\ell(\overline{x},\Sigma) = -\frac{n}{2}\left\{\log|2\pi\Sigma| - tr(\Sigma^{-1}S)\right\}$$

and then substituting $\Sigma = QQ^T + \Psi$ we get

$$\ell(\overline{x},Q,\Psi) = -\frac{n}{2}\left\{\log|2\pi(QQ^T+\Psi)| - tr[(QQ^T+\Psi)^{-1}S]\right\}$$

Assuming that $\hat{Q}$ and $\hat{\Psi}$ are the maximum likelihood estimates we get the LR statistic

$$-2\log\left(\frac{\ell(\overline{x}, Q, \Psi)}{\ell(\overline{x}, \Sigma)}\right) = n\log\left(\frac{|\hat{Q}\hat{Q}^T + \hat{\Psi}|}{|S|}\right)$$

which is asymptotically χ^2 under H_0 with d.f. $= \frac{1}{2}(p-k)^2 - \frac{1}{2}(p+k)$.

Bartlett (1954) has shown that the χ^2 approximation is better if we replace n with $n - 1 - (2p + 4k + 5)/6$.

- This can be used as a test of fit for the FA model, vs. the general multivariate normal;
- It is not wise to use this, without further adjustments, for comparing k vs. $k + 1$ factors, e.g. (why?? – see next page)
- A better assessment of the number of factors can be based on prediction error, Bayes factors, an information criterion, or some other measure that does not depend on the LR test being asymptotically χ^2 under H_0.

Naively we would think that a test of H_0 : k factors, vs. H_A : $k+1$ factors, could be based on the likelihood ratio test

$$-2\log\left(\frac{\ell(\overline{x}, Q_k, \Psi)}{\ell(\overline{x}, Q_{k+1}, \Psi)}\right) = n\log\left(\frac{|\hat{Q}_k\hat{Q}_k^T + \hat{\Psi}|}{|\hat{Q}_{k+1}\hat{Q}_{k+1}^T + \hat{\Psi}|}\right)$$

which we would take to be χ^2 under H_0 with d.f. $= \left[\frac{1}{2}(p-k)^2 - \frac{1}{2}(p+k)\right] - \left[\frac{1}{2}(p-(k+1))^2 - \frac{1}{2}(p+(k+1))\right] = 2(p-k)$.

However, the model for H_0 is *at the edge of the parameter space* for the model for H_A in this case (the asymptotic χ^2 theory depends on a Taylor expansion in a neigborhood of H_A around H_0; this is not posssible when H_0 is at the edge of H_A):

Under H_0: $\quad \mathrm{Var}(X) - \mathrm{Var}(U) = Q_kQ_k^T = \Gamma\Lambda^{(k)}\Gamma^T$

where $\Lambda^{(k)} = diag(\lambda_1, \ldots, \lambda_k, 0, \ldots, 0)$, since $Q_kQ_k^T$ has rank k.

Under H_A: $\quad \mathrm{Var}(X) - \mathrm{Var}(U) = Q_{k+1}Q_{k+1}^T = \Gamma^*\Lambda^{(k+1)}(\Gamma^*)^T$

where $\Lambda^{(k+1)} = diag(\lambda_1, \ldots, \lambda_k, \lambda_{k+1}, 0, \ldots, 0)$, since $Q_{k+1}Q_{k+1}^T$ has rank $k+1$.

Although H_0 can be obtained from H_A by the linear constraint $\lambda_{k+1} = 0$, in this case this constraint is at the edge of the parameter space, since $\lambda_{k+1} < 0$ is outside the space of eigenvalues for the positive semi-definite matrix $\mathrm{Var}(X) - \mathrm{Var}(U)$.

Factor Rotation

As we observed above, given the FA model

$$X = QF + U + \mu$$

then for *any orthogonal (rotation) matrix* G,

$$\begin{aligned} X &= Q(GG^T)F + U + \mu = (QG)(G^T F) + U + \mu \\ &= Q'F' + U + \mu \qquad \text{where} \end{aligned}$$

1. $Q' = QG$ may not look anything like Q!
2. $F' = G^T F$ is still has $E[F'] = GE[F] = 0 = E[F]$ and $\text{Var}(F')G\text{Var}(F)G^T = GIG^T = I = \text{Var}(F)$ (since G^T is orthogonal)!

This is analogous to the lack of identifiability in an ANOVA model:

- You can't estimate the grand mean and all the cell means at the same time without constraints.

  - To estimate the ANOVA model we impose constraints to identify the model for estimation (e.g. set intercept = 0, and estimate the "cell means" model).
  - After estimating the model we often re-parametrize for a particular interpretation (e.g. intercept = grand mean, cell effects sum to zero; or intercept = baseline cell, other cell effects are "offsets", etc.)

- We did something similar with the FA model
  - Take $Q^T \Psi^{-1} Q = \Delta$, a diagonal matrix, to identify the model for (ML) estimation.
  - The "rotation matrix" G above suggests how to reparametrize for a particular interpretation.

Factor Rotation and Simple Structure

In an over-parametrized ANOVA model there are a few traditional re-parametrizations that most analyses concentrate on.

In FA, there is no "natural" re-parametrization, but we often try to find "simple structure". In

$$X = QF + U + \mu$$

simple structure basically means:

- Each X_i depends on as few F_ℓ's as possible; or equivalently
- Each row of Q contains as many zero or near-zero entries as possible.

Essentially, we wish to find an orthogonal matrix R such that $Q' = QR$ has lots of zero or near-zero entries. The reparametrization will then be

$$X = (QR)(R^T F) + U + \mu = Q'F' + U + \mu$$

for this R.

Varimax

There are many heuristics for finding R. A common set of heuristics considers a derived matrix

$$Q^* = [|Q'_{j\ell}|]_{p\times k}, \quad Q^* = [Q'^2_{j\ell}]_{p\times k}, \quad \text{or } Q^* = [Q'^4_{j\ell}]_{p\times k}, \quad \text{etc.}$$

and tries to find R to *maximize the spread within each column of* Q^* (this puts lots of zero's or near-zero's in each column of $Q' = QR$).

The VARIMAX method finds R such that the *sum of the column variances of* $Q^* = [Q'^2_{j\ell}]$ *is maximized*:

$$V_R(Q') = \sum_\ell \left(\frac{1}{p} \sum_j Q'^4_{j\ell} - \left(\frac{1}{p} \sum_j Q'^2_{j\ell} \right)^2 \right)$$

where $Q' = QR$.

Other "rotations"

- Reparametrizations that rotate by an orthogonal matrix R are called *orthogonal rotations*. They preserve the property that $\mathrm{Var}(F') = I_{k\times k}$ (where $F' = R^T F$).
  - This used to be a cottage industry, generating names like QUARTIMAX, EQUIMAX, etc.; but VARIMAX is most popular
  - ICA is also a rotation method but with a different criterion (independence and non-normality, vs. simple structure).
- If we relax the criterion $\mathrm{Var}(F') = I_{k\times k}$ then we can take R to be non-orthohonal. Such reparametrizations are called *oblique rotations*. This also used to be a cottage industry, but the most popular is PROMAX:
  - Find the VARIMAX rotation $Q' = QR$; let $C = [Q'^4_{j\ell}]$.
  - Find a further invertible M to minimize $\sum\sum_{j,\ell}(Q''_{j\ell} - C_{j\ell})^2$, where $Q'' = QRM$.

  The final model $X = (QRM)(M^{-1}R^T F) + U + \mu = Q''F'' + U + \mu$ no longer has $\mathrm{Var}(F'') = I$, but may have "nice" simple structure.

Two Approaches to Factor Analysis

The approach we have described so far is often called *Exploratory Factor Analysis (EFA)*:

- Fit the unrestricted FA model (with $Q^T\Psi^{-1}Q = \Delta$ diagonal) with different dimensions (numbers of factors) k until you get "good fit".
- Perform one or more rotations to get a good "interpretation" of the factors.

An alternative approach is often called *Confirmatory Factor Analysis (CFA)*:

- Instead of the constraint "$Q^T\Psi^{-1}Q = \Delta$ diagonal", directly impose linear conditions on Q (most often, set many of the $Q_{j\ell} = 0$ to reflect prior theory about which factors go with which observable variables);
- Test overall fit, test whether some (more) $Q_{j\ell} = 0$, etc.

In EFA, it is hard to develop distribution theory and tests for factor loadings (because of the rotations). CFA is much more like choosing a specific parametrization as in ANOVA, and distribution theory, hypothesis tests, etc. are more available.

CFA as a statistical model

The CFA model has the same form as we have seen already

$$X = QF + U + \mu$$

but there is no ambiguity about the status of the factors; in fact the CFA model can be viewed as a kind of hierarchical Bayes model, e.g.:

$$\begin{aligned}
X \mid \mu, Q, F &\sim N_p(\mu + QF, \Psi) \\
\mu &\sim N_p(0, \Sigma_\mu) \\
vec(Q) &\sim N_{pk}(\mu_Q, \Sigma_Q) \text{ (where some entries of } \Sigma_Q \text{ set to zero)} \\
F &\sim N_k(0, I) \\
diag(\Psi) &\sim \text{independent inverse-}\chi^2\text{'s, etc.}
\end{aligned}$$

In this setting, posterior inference on all the parameters is of interest.

In particular, posterior inference on F is known as "estimating" or "predicting" factor scores; a variety of methods have been developed.