Version: August 6, 2011

Notes on Factor Analysis

The first question we need to address is why go to the trouble of developing a specific factor analysis model when principal components and "Little Jiffy" seem to get at this same problem of defining factors:

(1) In a principal component approach, the emphasis is completely on linear combinations of the observable random variables. There is no underlying (latent) structure of the variables that I try to estimate. Statisticians generally love models and find principal components to be somewhat inelegant and nonstatistical.

(2) The issue of how many components should be extracted is always an open question. With explicit models having differing numbers of "factors", we might be able to see which of the models fits "best" through some formal statistical mechanism.

(3) Depending upon the scale of the variables used (i.e., the variances), principal components may vary and there is no direct way of relating the components obtained on the correlation matrix and the original variance-covariance matrix. With some forms of factor analysis, such as maximum likelihood (ML), it is possible to go between the results obtained from the covariance matrix and the correlations by dividing or multiplying by the standard deviations of the variables. In other words, we can have a certain type of "scale invariance" if we choose, for example, the maximum likelihood approach.

(4) If one wishes to work with a correlation matrix and have a means of testing whether a particular model is adequate or to develop confidence intervals and the like, it is probably preferable to use the ML approach. In PCA on a correlation matrix, the results that are usable for statistical inference are limited and very strained generally (and somewhat suspect).

To develop the factor analysis model, assume the p *observable* random variables, $\mathbf{X}' = [X_1, \ldots, X_p]$, are MVN($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$). Without loss of generality, we can assume that $\boldsymbol{\mu}$ is the zero vector. Also, suppose that each X_i can be represented by a linear combination of some m *unobservable* or latent random variables, $\mathbf{Y}' = [Y_1, \ldots, Y_m]$, plus an error term, e_i:

$$X_i = \lambda_{i1}Y_1 + \cdots + \lambda_{im}Y_m + e_i, \text{ for } 1 \leq i \leq p \ .$$

Here, $Y_1, \ldots, Y_m$ are the common factor variables; $e_1, \ldots, e_p$ are the specific factor variables; λ_{ij} is the *loading* (i.e., the covariance) of the i^{th} response variable, X_i, on the j^{th} common factor variable.

If $\mathbf{e}' = [e_1, \ldots, e_p]$, then $\mathbf{X} = \boldsymbol{\Lambda}\mathbf{Y} + \mathbf{e}$, where

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_{11} & \cdots & \lambda_{1m} \\ \vdots & & \vdots \\ \lambda_{p1} & \cdots & \lambda_{pm} \end{pmatrix} \ .$$

For notation, we let the variance of e_i be ψ_i, $1 \leq i \leq p$, and refer to ψ_i as the *specific variance* of the i^{th} response variable; $e_i \sim$ N(0,ψ) and all the e_is are independent of each other; $Y_i \sim$ N(0,1) and all the Y_is are independent of each other and of the e_is. Also,

we define the diagonal matrix containing the specific variances to be

$$\mathbf{\Psi} = \begin{pmatrix} \psi_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \psi_p \end{pmatrix} .$$

$$\mathrm{Var}(X_i) = \mathrm{Var}(\lambda_{i1}Y_1 + \cdots \lambda_{im}Y_m + e_i) =$$
$$\mathrm{Var}(\lambda_{i1}Y_1) + \cdots + \mathrm{Var}(\lambda_{im}Y_m) + \mathrm{Var}(e_i) =$$
$$\lambda_{i1}^2 + \cdots + \lambda_{im}^2 + \psi_i \ .$$

The expression, $\Sigma_{j=1}^m \lambda_{ij}^2$, is called the communality of the i^{th} variable, X_i.

Because terms involving different unobservable and specific variables are zero because of independence, we have

$$\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(\lambda_{i1}Y_1 + \cdots \lambda_{im}Y_m + e_i, \lambda_{j1}Y_1 + \cdots \lambda_{jm}Y_m + e_j) =$$
$$\lambda_{i1}\lambda_{j1} + \cdots + \lambda_{im}\lambda_{jm} \ .$$

As a way of summarizing the results just given for the variances and covariances of the observable variables in terms of the loadings and specific variances, the factor analytic model is typically written as

$$\mathbf{\Sigma}_{p\times p} = \mathbf{\Lambda}_{p\times m}\mathbf{\Lambda}'_{m\times p} + \mathbf{\Psi}_{p\times p} \ .$$

There is a degree of indeterminacy in how this model is phrased, because for any $m \times m$ orthogonal matrix $\mathbf{T}$, we have the same type of decomposition of $\mathbf{\Sigma}$ as

$$\mathbf{\Sigma}_{p\times p} = (\mathbf{\Lambda T})_{p\times m}(\mathbf{\Lambda T})'_{m\times p} + \mathbf{\Psi}_{p\times p} \ .$$

Thus, we have a rotation done by $\mathbf{T}$ to generate a new loading matrix, $\mathbf{\Lambda T}$.

0.1 Iterated Principal (Axis) Factor Analysis

Suppose I assume the factor analytic model to hold for the population correlation matrix, $\mathbf{P} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}$, and am given the sample correlation matrix, $\mathbf{R}$. The Guttman lower bound to the communality of a variable is the squared multiple correlation of that variable with the others, and can be used to give an initial estimate, $\hat{\mathbf{\Psi}}$, of the matrix of specific variances by subtracting these lower bounds from 1.0 (the main diagonal entries in $\mathbf{R}$). A component analysis (with m components) is carried out on $\mathbf{R} - \hat{\mathbf{\Psi}}$ and then normalized to produce a factoring, say, $\mathbf{BB}'$. We estimate $\mathbf{\Psi}$ by using the diagonal of $\mathbf{R} - \mathbf{BB}'$, and iterate the process until convergence. (Little Jiffy (the principal component solution to the factor analysis model) could be viewed as a "one shot" process, with specific variances set at 0.0.)

0.2 Maximum Likelihood Factor Analysis (MLFA)

The method of MLFA holds out the hope of being a scale-invariant method, implying that the results from a correlation or the covariance matrix can be transformed into each other though simple multiplications by the variable standard deviations. So if λ_{ij} is a loading from a (population) correlation matrix, then $\lambda_{ij}\sigma_i$ is the corresponding loading from the (population) covariance matrix.

MLFA begins with the assumption that $\mathbf{X}_{p\times 1} \sim \text{MVN}(\mathbf{0}, \mathbf{\Sigma}_{p\times p} = \mathbf{\Lambda}_{p\times m}\mathbf{\Lambda}'_{m\times p} + \mathbf{\Psi}_{p\times p})$. If there is a unique diagonal matrix, $\mathbf{\Psi}$, with positive elements such that the m largest roots (eigenvalues) of $\mathbf{\Sigma}^* = \mathbf{\Psi}^{-1/2}\mathbf{\Sigma}\mathbf{\Psi}^{-1/2}$ are distinct and greater than unity, and the $p - m$

remaining roots are each unity (this is true if the model holds), then $\mathbf{\Lambda} = \mathbf{\Psi}^{1/2}\mathbf{\Omega}\mathbf{\Delta}^{1/2}$, where $\mathbf{\Sigma}^* - \mathbf{I} = \mathbf{\Omega}_{p\times m}\mathbf{\Delta}_{m\times m}\mathbf{\Omega}'_{m\times p}$. In other words, once you get $\mathbf{\Psi}$, you are "home free" because $\mathbf{\Lambda}$ comes along by a formula.

So, we start with some $\mathbf{\Psi}$ (and generating $\mathbf{\Lambda}$ automatically), and improve upon this initial value by maximizing the log-likelihood

$$\ell(\mathbf{\Lambda}, \mathbf{\Psi}) = -\frac{n}{2}(\ln|\mathbf{\Sigma}| + \mathrm{Tr}(\mathbf{S}\mathbf{\Sigma}^{-1})) + \text{constant} \ .$$

Equivalently, we can minimize

$$\mathrm{F}(\mathbf{\Lambda}, \mathbf{\Psi}) = \ln|\mathbf{\Sigma}| + \mathrm{Tr}(\mathbf{S}\mathbf{\Sigma}^{-1})) - \ln|\mathbf{S}| - p \ .$$

The particular iterative optimization procedure used to obtain better and better values for $\mathbf{\Psi}$ is typically the Davidon-Fletcher-Powell method.

In practice, one has a large sample likelihood ratio test available of

$$H_0 : \mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi} \ ,$$

using a test statistic of $(n - (2p+5)/6 - 2m/3)\mathrm{F}(\hat{\mathbf{\Lambda}}, \hat{\mathbf{\Psi}})$, compared to a chi-squared random variable with $\frac{1}{2}[(p-m)^2 - (p+m)]$ degrees of freedom. Generally, the residuals one gets from an MLFA tend to be smaller than from a PCA, even though the cumulative variance explained in a PCA is usually larger; these are somewhat different criteria of fit.

In MLFA, one typically needs a rotation (oblique or orthogonal) to make the originally generated factors intelligible. Also, we now have various forms of confirmatory factor analysis (CFA) where some of

the loadings might be fixed and others free to vary. CFA seems to be all the rage in scale development, but I would still like to see what a PCA tells you in an exploratory and optimized context. Finally, and although we talked about using and plotting component scores on our subjects in PCA, the comparable factor scores here should *not* be used. There has been an enormous controversy about their indeterminacy; among people who are thinking straight (e.g., SYSTAT and Leland Wilkinson), factor scores are just not given.

When one allows correlated factors (e.g., using an oblique rotation), the factor analytic model is generalized to

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}' + \mathbf{\Psi}$$

where $\mathbf{\Phi}$ is the $m \times m$ covariance matrix among the m factors. In terms of terminology, the matrix, $\mathbf{\Lambda}$, is called the factor *pattern* matrix; $\mathbf{\Lambda}\mathbf{\Phi}$ is called the factor *structure* matrix and contains the covariances between the observed variables and the m common factors.

There is one property of MLFA that sometimes (in fact, often) rears its ugly head, involving what are called Heywood cases (or improper solutions) in which the optimization procedure wants to make some of the ψ_is go negative. When this appears to be happening, the standard strategy is to remove the set of variables for which the ψ_is want to go negative, set them equal to zero exactly; the removed set is then subjected to a principal component analysis, and a "kluge" made of the principal components and the results from an MLFA on a covariance matrix residualized from the removed set. Obviously, the nice scale invariance of a true MLFA approach disappears when

these improper solutions are encountered. You can tell immediately that you have this kind of hybrid solution when some of the specific variances are exactly zero.