An Introduction to Factor Analysis

Simon Jackman

Spring 2005

The Factor Analysis Latent Variable Model

Consider the case of survey data, where we denote respondent i's answer to survey question j, as x_{ij}, $(i = 1, \ldots, n; j = 1, \ldots, k)$. Factor analysis posits that x_{ij} is a combination of p unobserved factors, each written using the Greek letter ξ: i.e.,

$$x_{ij} = \lambda_{j1}\xi_{i1} + \lambda_{j2}\xi_{i2} + \ldots + \lambda_{jp}\xi_{ip} + \delta_{ij} \tag{1}$$

where the λ terms are *factor loadings* to be estimated, and δ_{ij} is the measurement error in x_{ij}, or that part of x_{ij} that can not be accounted for by the p underlying factors. It is possible to consider non-linear or multiplicative factor models (e.g., Bollen, 1989, 403ff), but the simple linear, additive structure in equation (1) is by far the more widely used factor analysis model.

Factor Loadings

The factor loadings λ are parameters to be estimated that tap how the unobserved factors account for the observed variables: the larger the values of λ, the more a particular variable is said to "load" on the corresponding factor. Note that the factor loadings λ vary across survey items, but not across individuals. Put differently, items vary in the way they are explained by the underlying factors, but the relationships between underlying factors and observed responses is constant across individuals (hence the absence of an i subscript indexing λ). Note also that there are fewer underlying factors than there are variables ($p < k$), consistent with the notion that like any statistical procedure, factor analysis is a device for "data reduction", taking a possibly rich though unwieldy set of survey responses and summarizing them with a simpler underlying structure.

Measurement Errors

The δ terms for measurement errors simply reflect the idea that survey responses are not deterministically generated by the underlying factor structure. Like any statistical model, the factor structure is an approximation or a simplification that only captures so much of the survey responses under study. Another way of thinking about measurement error is to imagine a respondent being asked to generate a response to a survey question on successive days: the observed responses would presumably fluctuate about a mean (tapped by the structural part of the factor analysis model), but the response on any given day will be a little above or below the average response. Conditional on the structure we posit to underlie the survey responses, this random component of the survey response is tapped by δ.

A Model for Multiple Responses

We can write an equation of the form of equation (1) for each item being analyzed. Each equation expresses the corresponding survey response as a combination of "structure" and "noise": the underlying factors and measurement error, respectively. Matrix notation allows the entire system of equations to be written quite compactly: i.e., for each respondent

$$\mathbf{x}_i = \mathbf{\Lambda}\boldsymbol{\xi}_i + \boldsymbol{\delta}_i, \tag{2}$$

where $\mathbf{x}_i$ is a k by 1 vector of observed survey responses, $\mathbf{\Lambda}$ is a k by p matrix of factor loadings to be estimated, $\boldsymbol{\xi}_i$ is p by 1 vector of scores on the p underlying factors, and $\boldsymbol{\delta}$ is a k by 1 vector of measurement errors. In turn, we can lose the i subscript indexing individual respondents by stacking equation (2) over respondents to yield

$$\mathbf{X} = \mathbf{\Xi}\mathbf{\Lambda}' + \mathbf{\Delta}, \tag{3}$$

where $\mathbf{X}$ is a n by k matrix of observed survey responses, $\mathbf{\Xi}$ is a n by p matrix of scores on the underlying factors, $\mathbf{\Lambda}'$ is the transpose of the k by p matrix of factor loadings, and $\mathbf{\Delta}$ is a n by k matrix of measurement errors.

Analysis of Covariances and Correlations

Since factor analysis usually works with the variances and covariances of the observed x variables, it is sometimes referred to as "the analysis of covariance structures". Some hint of this is apparent in equation (1), where the absence of an intercept term suggests that the means of the observed variables are either zero or of no direct interest. Indeed, this is typically the case in factor analysis, where the task is to learn about inter-relationships among variables rather than model the levels of each variable. Moreover, it is generally not possible to estimate *both* the factor loadings and intercept terms (cf Jöreskog and Sörbom, 1989, ch10). See also Bollen (1989, 306--311). Consequently, all the x variables and the unobserved ξ are presumed to have zero means, constraining any intercept term in equation (1) to zero. In addition, for the ordinal variables frequently encountered in surveys, the latent variable approach to generating a correlation matrix posits the variances of the latent variables to be 1, making the all covariances between the latent variables interpretable as correlations.

The following constraints and identities make the covariance structure representation of the factor analysis model tractable:

$$E(\mathbf{\Delta}'\mathbf{\Xi}) = \mathbf{0} \tag{4a}$$

$$E(\mathbf{\Xi}'\mathbf{\Delta}) = \mathbf{0} \tag{4b}$$

$$E(\mathbf{\Delta}'\mathbf{\Delta}) = \mathbf{\Theta}_{\boldsymbol{\delta}} \tag{4c}$$

$$E(\mathbf{\Xi}'\mathbf{\Xi}) = \mathbf{\Phi} \tag{4d}$$

$$E(\mathbf{X}'\mathbf{X}) = \mathbf{\Sigma} \tag{4e}$$

The first two of these constraints are identical, stating that there is no correlation between the measurement errors and the scores on the underlying factors. The three remaining conditions simply define some variance-covariance matrices of interest: the k by k variance-covariance of the measurement errors is denoted $\boldsymbol{\Theta}_{\boldsymbol{\delta}}$; the p by p variance-covariance matrix of the underlying factors is denoted $\boldsymbol{\Phi}$; and the k by k variance-covariance matrix of the data is denoted $\boldsymbol{\Sigma}$. Using these assumptions and definitions, equation (3) can be manipulated to yield (e.g., Bollen, 1989, 35)

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}_{\boldsymbol{\delta}} \tag{5}$$

That is, the variances and covariances among the observed variables can be decomposed into a component attributable to the underlying factors (and the relationships among those factors), and the measurement error variances and covariances. Quite simply, the statistical problem here is to estimate the elements of the matrices on the right-hand side of equation (5) --- the parameters constituting the factor analysis model --- using the information in the variance-covariance matrix of the observed data.

Identification Constraints

Constraints are typically needed to make all of these parameters estimable. To see why, note that there are only $k(k + 1)/2$ unique elements in $\boldsymbol{\Sigma}$ (variance-covariance matrices are symmetric), but as many as $kp + p(p + 1)/2 + k(k + 1)/2$ elements on the right-hand side of equation (5), in $\boldsymbol{\Lambda}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Theta}_{\boldsymbol{\delta}}$, respectively. That is, there are more parameters to estimate than the available pieces of sample information. Common strategies that help with this problem are[1]

1. make p small relative to k. That is, assume there are a relatively small number of factors underlying the data, which is an attractive assumption in any event, on grounds of parsimony.

2. constrain off-diagonal elements of $\boldsymbol{\Theta}_{\boldsymbol{\delta}}$ to zero. These parameters are the measurement error covariances, and in general are not particularly interesting parameters and for pairs of survey responses reasonably thought to be unrelated, setting the corresponding off-diagonal element of $\boldsymbol{\Theta}_{\boldsymbol{\delta}}$ to zero is not an implausible restriction, and unlikely to be substantively consequential.

3. constrain $\boldsymbol{\Phi}$ to be an identity matrix, or a diagonal matrix. The former restriction assumes that the underlying factors have equal variances (set to one), and are uncorrelated; this is the default assumption in many popular programs for exploratory factor analysis that produce orthogonal factors. Assuming $\boldsymbol{\Phi}$ to be diagonal eases the equal variance restriction, but retains the property that the estimated factors are orthogonal.

[1] It is important to remember that having more sample variances and covariances than parameters to estimate is a necessary but not sufficient condition for making the parameters of the factor analysis model identifiable.

4. constrain elements of $\mathbf{\Lambda}$ to zero; i.e., some variables might reasonably be presumed not to load on a particular factor.

5. alternatively, one element per column of $\mathbf{\Lambda}$ can be set to 1, effectively "defining" the particular factor with reference to the corresponding variable. For instance, in my analysis here, I define the first factor as a "left-right" item by constraining the loading of the the left-right self-placement variable on that factor to be 1.

Exploratory factor analysis typically works with a small number of factors relative to the number of variables, but allows all variables to load on all factors. Accordingly, exploratory factor analysis constrains other parts of the model along the lines suggested above: $\mathbf{\Phi}$ is constrained to be an identity matrix (to yield orthogonal factors each with unit variance), and $\mathbf{\Theta_\delta}$ is constrained to be diagonal.

Confirmatory factor analysis eases these constraints to allow more realistic models to be estimated, but at the cost of constraints on the factor loadings. For instance, we can "free" the off-diagonal element of $\mathbf{\Phi}$ (so as to estimate the correlation between the factors), and also estimate off-diagonal elements of $\mathbf{\Theta_\delta}$ (measurement error covariances). The numerous restrictions on the factor loadings allow these relaxations of the other parts of the model.

Estimation via Principal Components/Eigen-Decomposition

Consider the case where $\mathbf{\Phi}$ is constrained to be an identity matrix, and so the model in equation (5) reduces to

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi} \tag{6}$$

A popular and computationally cheap method for estimating the model parameters is via principal components, exploiting the fact that a covariance matrix (i.e., a positive definite, square, symmetric matrix) can be decomposed as follows

$$\mathbf{\Sigma} = \mathbf{Z}'\mathbf{\Gamma}\mathbf{Z} \tag{7}$$

where $\mathbf{\Gamma}$ is a diagonal matrix containing the eigenvalues of $\mathbf{\Sigma}$ in decreasing order ($\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_k \geq 0$), and $\mathbf{Z}$ is a k by k matrix of orthogonal eigenvectors. Each eigenvector can be usefully considered as a vector of coefficients that could be used forming uncorrelated linear combinations of the $\mathbf{X}$ variables. For instance, using the jth eigenvector in this way produces a new variable $\mathbf{y}_j = \mathbf{X}\mathbf{z}_j$, which is the jth *principal component* of $\mathbf{X}$ ($\mathbf{y}_j$ is a n by 1 vector, $\mathbf{X}$ is a n by k matrix, and $\mathbf{z}_j$ is a k by 1 vector).

Principal components have properties that make them especially useful for factor analysis. The first principal component has the largest variance among all linear combinations of $\mathbf{X}$.[2] The second principal component has the largest variance

[2] The $\mathbf{X}$ variables are typically standardized so as to ensure that a variable with a large variance does not unduly dominate the analysis. This standardization is implicit when performing factor analysis on correlation matrices.

among linear combinations of $\mathbf{X}$ subject to the constraint that it is uncorrelated with the first principal component, and so on for subsequent principal components. Accordingly, each eigenvector of the correlation matrix is also a vector of principal components factor loadings.

While there are as many principal components as there are $\mathbf{X}$ variables, the idea behind factor analysis to come up with a parsimonious representation of the structure underlying the $\mathbf{X}$ variables. In practice, then, only the first few principal components are retained, corresponding to a few factors. For any p factor model (with $p > k$), only the first p eigenvectors in $\mathbf{Z}$ are retained, and so the "full" k dimensional decomposition in equation (7) is not used; i.e., some of the variation in $\mathbf{X}$ is considered random, and is relegated to the $\mathbf{\Psi}$ matrix in equation (6). The factor analysis model estimated by principal components is

$$\mathbf{\Sigma} = \mathbf{Z}_{(p)}\mathbf{Z}'_{(p)} + \mathbf{\Psi}, \tag{8}$$

where $\mathbf{Z}_{(p)}$ is the k by p matrix containing the first p eigenvectors of $\mathbf{\Sigma}$.

Another important property of factor analysis via principal components is that the model in equation (8) is not unique. Any rotation of $\mathbf{Z}$ that preserves its orthogonal structure fits the data just as well as the unrotated solution in equation (8). That is, the principal components factor loadings $\mathbf{Z}_{(p)}$ can be multiplied by a p by p orthogonal matrix[3] $\mathbf{G}$ to yield $\mathbf{Z}^*_{(p)} = \mathbf{Z}_{(p)}\mathbf{G}$ and so

$$\begin{aligned}
\mathbf{\Sigma} &= \mathbf{Z}^*_{(p)}\mathbf{Z}^{*\prime}_{(p)} + \mathbf{\Psi} \\
&= (\mathbf{Z}_{(p)}\mathbf{G})(\mathbf{G}'\mathbf{Z}'_{(p)}) + \mathbf{\Psi} \\
&= \mathbf{Z}_{(p)}\mathbf{Z}'_{(p)} + \mathbf{\Psi},
\end{aligned}$$

i.e., the factor loadings are identified only up to orthogonal rotations. The problem then becomes one of choosing among rotations that are optimal on other criteria. One popular choice is the *varimax* rotation (Kaiser, 1958), which produces factors loadings that have maximal variance, taking on values close to 1 and 0 in absolute value. This helps ensure that the factors are reasonably distinct, with variables tending to load either quite strongly or quite weakly on any given factor.

Estimation - Maximum Likelihood

The more modern approach is to estimate the parameters of the factor analysis model via maximum likelihood. If the $\mathbf{x}_i$ are assumed to be iid multivariate normal $\forall\, i$, i.e., $\mathbf{x}_i \overset{\text{iid}}{\sim} N(\boldsymbol{\mu}, \mathbf{\Sigma})$, then the joint density of the data is

$$f(\mathbf{X}; \boldsymbol{\mu}, \mathbf{\Sigma}) = (2\pi)^{-\frac{kn}{2}} |\mathbf{\Sigma}|^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right].$$

[3] If $\mathbf{G}$ is an orthogonal matrix then $\mathbf{G}\mathbf{G}' = \mathbf{I}$.

This provides the basis for maximum likelihood estimation: we can treat $\boldsymbol{\mu}$ as known (i.e., using the sample mean $\bar{\mathbf{x}}$ is the MLE of $\boldsymbol{\mu}$) and then embed the factor analysis model for $\boldsymbol{\Sigma}$ in the likelihood function. We work with the concentrated log likelihood (i.e., the log-likelihood that results from treating $\boldsymbol{\mu}$ as fixed at its sample estimate):

$$-\frac{n}{2}\ln|2\pi\boldsymbol{\Sigma}| - \frac{n}{2}\text{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1}) \tag{9}$$

where $\mathbf{S} = \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$. Anderson (2003, 14.3) considers properties of the maximum likelihood estimator for the orthogonal factor model $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\Psi}}$; see also Mardia, Kent and Bibby (1979) and Lawley and Maxwell (1971).

As a practical matter, and at least for the orthogonal factor model, it is easier to *minimize* the following function with respect to $\boldsymbol{\Sigma}$

$$F(\boldsymbol{\Lambda}, \boldsymbol{\Psi}; \mathbf{S}) = \text{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S}) - \ln|\boldsymbol{\Sigma}^{-1}\mathbf{S}| - k \tag{10}$$

i.e., minimizing equation (10) wrt $\boldsymbol{\Sigma}$ yields the same result as maximizing the log-likelihood in equation (9. Computational strategies for carrying out this optimization are discussed in Mardia, Kent and Bibby (1979, 264-266), summarizing the pioneering work by Jöreskog (1967).

Effective Number of Free Parameters

Following Lawley and Maxwell (1971, 7ff), note that the orthogonal factor analysis model implies

$$\boldsymbol{\Sigma} - \boldsymbol{\Psi} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}'.$$

Suppose now that each variable in $\mathbf{X}$ is rescaled so that the residual measurement error variances (the diagonal elements of $\boldsymbol{\Psi}$) are all one. This means that

$$\boldsymbol{\Sigma}^* = \boldsymbol{\Lambda}^*\boldsymbol{\Lambda}^{*\prime} + \mathbf{I}$$

which follows from transforming the original model

$$\boldsymbol{\Psi}^{-\frac{1}{2}}\boldsymbol{\Sigma}\boldsymbol{\Psi}^{-\frac{1}{2}} = \boldsymbol{\Psi}^{-\frac{1}{2}}\boldsymbol{\Lambda}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-\frac{1}{2}} + \boldsymbol{\Psi}^{-\frac{1}{2}}\boldsymbol{\Psi}\boldsymbol{\Psi}^{-\frac{1}{2}}$$

and hence $\boldsymbol{\Sigma}^* = \boldsymbol{\Psi}^{-\frac{1}{2}}\boldsymbol{\Sigma}\boldsymbol{\Psi}^{-\frac{1}{2}}$. Furthermore, this transformation of the model implies that $\boldsymbol{\Sigma} - \boldsymbol{\Psi}$ is transformed to become

$$\begin{aligned}\boldsymbol{\Psi}^{-\frac{1}{2}}(\boldsymbol{\Sigma} - \boldsymbol{\Psi})\boldsymbol{\Psi}^{-\frac{1}{2}} &= \boldsymbol{\Psi}^{-\frac{1}{2}}\boldsymbol{\Sigma}\boldsymbol{\Psi}^{-\frac{1}{2}} - \boldsymbol{\Psi}^{-\frac{1}{2}}\boldsymbol{\Psi}\boldsymbol{\Psi}^{-\frac{1}{2}} \\ &= \boldsymbol{\Sigma}^* - \mathbf{I}_k\end{aligned}$$

which is symmetric and has rank p. Accordingly, we can decompose $\boldsymbol{\Sigma}^* - \mathbf{I}$ into $\boldsymbol{\Omega}\mathbf{Y}\boldsymbol{\Omega}'$, where $\mathbf{Y}$ is a diagonal matrix of order p and $\boldsymbol{\Omega}$ is a k by p matrix such that $\boldsymbol{\Omega}'\boldsymbol{\Omega} = \mathbf{I}_p$. The elements of $\mathbf{Y}$ contain the p non-zero eigenvalues of $\boldsymbol{\Sigma}^* - \mathbf{I}_k$ and the columns of $\boldsymbol{\Omega}$ are the corresponding eigenvectors. This decomposition implies a unique solution for $\boldsymbol{\Lambda}$:

$$\boldsymbol{\Lambda} = \boldsymbol{\Psi}^{\frac{1}{2}}\boldsymbol{\Omega}\mathbf{Y}^{\frac{1}{2}},$$

which is true since

$$\begin{aligned}\boldsymbol{\Lambda}\boldsymbol{\Lambda}' &= \boldsymbol{\Psi}^{\frac{1}{2}}\boldsymbol{\Omega}\boldsymbol{\Upsilon}^{\frac{1}{2}}\boldsymbol{\Upsilon}^{\frac{1}{2}}\boldsymbol{\Omega}'\boldsymbol{\Psi}^{\frac{1}{2}} \\ &= \boldsymbol{\Psi}^{\frac{1}{2}}\boldsymbol{\Omega}\boldsymbol{\Upsilon}\boldsymbol{\Omega}'\boldsymbol{\Psi}^{\frac{1}{2}} \\ &= \boldsymbol{\Psi}^{\frac{1}{2}}(\boldsymbol{\Sigma}^{*} - \mathbf{I}_k)\boldsymbol{\Psi}^{\frac{1}{2}} \\ &= \boldsymbol{\Sigma} - \boldsymbol{\Psi},\end{aligned}$$

which is required by the model. In addition,

$$\begin{aligned}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} &= \boldsymbol{\Upsilon}^{\frac{1}{2}}\boldsymbol{\Omega}'\boldsymbol{\Psi}^{\frac{1}{2}}\boldsymbol{\Psi}^{-1}\boldsymbol{\Psi}^{\frac{1}{2}}\boldsymbol{\Omega}\boldsymbol{\Upsilon}^{\frac{1}{2}} \\ &= \boldsymbol{\Upsilon}^{\frac{1}{2}}\boldsymbol{\Omega}'\boldsymbol{\Omega}\boldsymbol{\Upsilon}^{\frac{1}{2}} \\ &= \boldsymbol{\Upsilon},\end{aligned}$$

since $\boldsymbol{\Omega}'\boldsymbol{\Omega} = \mathbf{I}_p$. But $\boldsymbol{\Upsilon}$ is a diagonal matrix of order p, which effectively imposes constraints on $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$.

That is, while there are kp free parameters in $\boldsymbol{\Lambda}$, and k free parameters in $\boldsymbol{\Psi}$, the requirement that $\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}$ be diagonal imposes $\frac{1}{2}p(p - 1)$ constraints on the model parameters. To see this, note that if there were no constraints on the p by p symmetric matrix $\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}$ we would have $\frac{1}{2}p(p + 1)$ free parameters; the constraint implies just p free parameters, for a difference of $\frac{1}{2}p(p - 1)$ parameters. In total then, the orthogonal, unit variance p factor analysis model has

- kp factor loadings to estimate (each of k variables loading onto all p factors) in $\boldsymbol{\Lambda}$
- k measurement error variances to estimate (no error covariances) in $\boldsymbol{\Psi}$
- *less* $\frac{1}{2}p(p-1)$ parameters imposed by the constraint that $\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}$ be diagonal.

for a total of $k + kp - \frac{1}{2}p(p - 1)$ free parameters.

Likelihood Ratio Test Statistic

Consider the case where the factor analysis model fits the data perfectly: that is, the observed covariance matrix $\mathbf{S}$ is perfectly recovered by $\hat{\boldsymbol{\Sigma}}$. In this case the log-likelihood reduces to

$$\ln\mathcal{L}_0 = -\frac{n}{2}\ln|2\pi\mathbf{S}| - \frac{nk}{2}, \tag{11}$$

since in this case $\mathbf{S} = \hat{\boldsymbol{\Sigma}}$ and the trace of $\mathbf{S}\hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{I}_k$ is k.

Comparing the two log-likelihoods in equations (9) and (11) allow a likelihood ratio test statistic to be constructed: i.e.,

$$\begin{aligned}q &= \frac{\mathcal{L}}{\mathcal{L}_0} \\ \ln q &= \ln\mathcal{L} - \ln\mathcal{L}_0 \\ -2\ln q &\sim \chi^2_v\end{aligned}$$

where the degrees of freedom parameters v is the number of unique elements in the covariance matrix $\mathbf{S}$ minus the number of parameters estimated, or

$$\begin{aligned} v &= \left[\frac{1}{2}(k+1)k\right] - \left[k + kp - \frac{1}{2}p(p-1)\right] \\ &= \frac{1}{2}\left[k^2 + k - 2k - 2pk + p^2 - p\right] \\ &= \frac{1}{2}\left[(k-p)^2 - (k+p)\right] \end{aligned}$$

where k is the number of $\mathbf{x}$ variables, and p is the number of factors being estimated. If the difference in the unconstrained and constrained likelihoods is large, then the test statistic will be large, and, if sufficiently large, we will reject the constrained model in favor of a less constrained model.

One of the key features of the representation of the optimization problem underlying maximum likelihood factor analysis given in equation (10) is that it is leads directly to the test statistic given above. That is, the test statistic

$$-2\ln q = nF(\hat{\mathbf{\Lambda}}, \hat{\mathbf{\Psi}})$$

where F is defined in equation (10), has an asymptotic χ^2_v distribution, where the degrees of freedom v is defined above.

Testing the Number of Factors

This statistic can be used in testing the number of factors. For instance, we might start with $p = 1$ factors, and compute the test statistic given above. The test statistic gets larger as $\hat{\mathbf{\Sigma}}$ diverges from $\mathbf{S}$ and so if the test statistic exceeds some critical value, we can reject the hypothesis that the p factor solution is an appropriate fit to the data, in favor of a model with more factors. We could then repeat the factor analysis, with $p = 2$, and recompute the test statistic. This procedure could be repeated until the hypothesis of a p factor fit is not rejected, or until the degrees of freedom parameter $v \leq 0$.

Note that this sequential testing procedure is

> ...open to the criticism because the critical values of the test criterion have not been adjusted to allow for the fact that a set of hypotheses is being tested in sequence, with each one dependent on the rejection of all predecessors. Lawley and Maxwell (1971) suggest that this problem is unlikely to cause serious problems in practice (Everitt, 1984, 22).

Relying solely on statistical criteria in an exploratory analyses is not well-advised; typically researchers will have ideas about what the underlying factors look like, and how many factors ought to define a parsimonious model for the data. The testing procedure described above will lead to the additional of factors that

pick up additional variation in **X** that is distinguishable from sampling variability, so many factors added will have little explanatory power in a substantive sense, though will be statistically significant.

As a special case, we can test the adequacy of a zero factor model; i.e., the observed **X** variables are mutually independent. In this special case, the MLE of $\boldsymbol{\Sigma}$ is $\hat{\boldsymbol{\Sigma}}$ = diag**S** (i.e., all off-diagonal elements are zero) and Mardia, Kent and Bibby (1979, 267) show that the test statistic reduces to $-n\ln|\mathbf{R}|$, where **R** is the correlation matrix of the **X** variables.

Caveats

Simulation work also suggests some caution be taken in interpreting χ^2 type ML tests for factor analysis models (e.g., Bollen (1989, 266ff)):

- the tests assume that the residual matrix (see below) is distributed Wishart, which in turn requires that the **x** variables exhibit no kurtosis (such that the data can be validly summarized with the second moments in $\mathbf{X}'\mathbf{X}$, which is true if **X** is multivariate normal).
- the sample is large; samples less than 50 or even 100 tend to lead to too frequent rejections of null hypotheses; some authors suggest rules-of-thumb such as 5 observations for every free parameter.
- the null model is an unrealistic "perfect fit" model; perhaps all we want is a model that gives us a reasonable approximation, rather than a comparison against a perfect fit. A high value of the χ^2 test statistic which leads us to reject the null might lead us to estimate more parameters when we already have a reasonable approximation.

Eigenvalues larger than one

A rule of thumb often encountered in applied exploratory factor analysis is to retain as many factors as there are eigenvalues greater than one. This a fairly arbitrary criterion, but for orthogonal solutions, it has the virtue that the eigenvalues are directly tied to the proportion of variance in **X** explained by successive factors.

For $\mathbf{X}'\mathbf{X}$ with rank k, the sum of the eigenvalues is k and the eigenvalues associated with each successive (principal component) factor decline towards zero. Large eigenvalues indicate principal components picking up a relatively large proportion of the variation in **X**, while small eigenvalues are associated with principal components picking up relatively small proportions of the variation in **X**. Eigenvalues less than 1 indicate that the corresponding principal component is picking up less variation than is in each variable: that is, the principal component accounts for less variation in **X** than does the average principal component.

Scree plots are a useful diagnostic tool for examining the eigenvalues of $\mathbf{X}'\mathbf{X}$. An example is in my paper, in Figure 2. As the size of the eigenvalues diminishes,

the proportion of explained variation diminishes, such that very little variation is picked up by additional factors after 3 or 4 factors enter the solution.

Estimation - Weighted Least Squares

A weighted least squares estimator finds estimates $\hat{\mathbf{\Sigma}}$ that minimizes the criterion

$$w = [vec(\mathbf{S}) - vec(\hat{\mathbf{\Sigma}})]' \mathbf{W}^{-1} [vec(\mathbf{S}) - vec(\hat{\mathbf{\Sigma}})], \quad (12)$$

where **W** is a matrix of weights, and the *vec* operator turns the lower triangle of its matrix argument into a vector (Bollen, 1989, 425). If **S** is a k by k matrix then $vec(\mathbf{S})$ is a vector of length $k(k + 1)/2$, containing the unique elements of **S** (k diagonal elements, plus $k(k - 1)/2$ unique off-diagonal elements). Accordingly, **W** is a $k(k + 1)/2$ by $k(k + 1)/2$ matrix. If the elements of **W** contain consistent estimates of the variances and covariances of the $vec(\mathbf{S})$, then the Browne (1984) asymptotically-best distribution-free WLS estimator results from choosing $\hat{\mathbf{\Sigma}}$ to minimize the criterion in equation (12).

This ADF-WLS estimator is especially useful when working with ordinal survey responses, where non-normality in the "normal scores" is almost guaranteed. **W** is usually estimated in the course of generating **S** from the normal score representation of the ordinal responses. Each diagonal element of **W** is an estimate of the variance of the corresponding sample correlation in $vec(\mathbf{S})$, while each off-diagonal elements is an estimate of the covariance in pairs of sample correlations. These off-diagonal quantities turn out to be functions of the cross-kurtoses in the raw data; if the raw data were distributed iid multivariate normal, then a consistent estimate of **W** would be an identity matrix. The ADF-WLS estimator is distribution-free in the sense that the raw data can be of almost any distribution, but with an appropriate **W** matrix we can obtain estimates with properties such as consistency and asymptotic normality. This makes asymptotically-valid hypothesis testing and inference-making possible even when the raw data are non-normal.

Goodness of Fit

I have already noted that when dealing with orthogonal factors, the relative magnitudes of the eigenvalues determine the proportion of the variance explained. This hold for any orthogonal rotation of a principal components solution such as the common varimax rotation.

There is also a way to do residual analysis with the factor analysis model. Note the model we fit is

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

with a perfect fit being when $\hat{\mathbf{\Sigma}} = \mathbf{S}$ (the sample covariance or correlation matrix). The simplest summary measure of goodness-of-fit involves simply comparing $\hat{\mathbf{\Sigma}}$ with **S**. One should always inspect this "residual matrix" ($\mathbf{S} - \hat{\mathbf{\Sigma}}$) for large elements which suggest model inadequacy; note that is matrix will be symmetric and thus

have only $k(k - 1)/2$ unique elements. Various summary measures have been proposed: one popular candidate is *root mean-square residual* (RMR):

$$\text{RMR} = \left[2 \sum_{i=1}^{k} \sum_{j=1}^{i} \frac{\left(s_{ij} - \hat{\sigma}_{ij}\right)^2}{k(k+1)} \right]^{\frac{1}{2}}$$

i.e., the square-root of mean of the squared elements of the residual matrix.

Factor Scores

Given estimates of the various component of the factor-analysis model, it is possible to generate scores on the underlying dimensions for each respondent. A useful way to think about factor scores is consider predicting the scores on the underlying factors from a regression on the the raw data $\mathbf{X}$. That is, consider the "regression" equation

$$\boldsymbol{\Xi} = \mathbf{X}\mathbf{A} + \boldsymbol{\Omega}, \tag{13}$$

where $\mathbf{A}$ is a k by p matrix of unknown coefficients relating the n by k matrix of data $\mathbf{X}$ to the n by p matrix of factor scores, and $\boldsymbol{\Omega}$ is a matrix of disturbances.

An estimate of $\mathbf{A}$ can be obtained via analogy with the regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{14}$$

for which a popular estimator of $\boldsymbol{\beta}$ is given by

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= [\text{var}(\mathbf{X})]^{-1}\text{cov}(\mathbf{X}, \mathbf{y}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \end{aligned}$$

This estimator in turn yields estimates of $\mathbf{y}$ conditional on $\mathbf{X}$ of the form $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.

Applying this approach to equation (13) yields $\hat{\mathbf{A}} = [\text{var}(\mathbf{X})]^{-1}\text{cov}(\mathbf{X}, \boldsymbol{\Xi})$. From equation (4e) we have $[\text{var}(\mathbf{X})]^{-1} = \boldsymbol{\Sigma}^{-1}$,. To obtain an expression for $\text{cov}(\mathbf{X}, \boldsymbol{\Xi})$ I substitute equation (3) into the definition of covariance:

$$\begin{aligned} E(\mathbf{X}'\boldsymbol{\Xi}) &= E\left[(\boldsymbol{\Xi}\boldsymbol{\Lambda}' + \boldsymbol{\Delta})'\boldsymbol{\Xi}\right], \\ &= E\left[(\boldsymbol{\Delta}' + \boldsymbol{\Lambda}\boldsymbol{\Xi}')\boldsymbol{\Xi}\right], \\ &= E\left[\boldsymbol{\Delta}'\boldsymbol{\Xi} + \boldsymbol{\Lambda}\boldsymbol{\Xi}'\boldsymbol{\Xi}\right], \\ &= \boldsymbol{\Lambda}\boldsymbol{\Phi}, \end{aligned}$$

since by assumption $E(\boldsymbol{\Delta}'\boldsymbol{\Xi}) = \mathbf{0}$ (equation 4b) and $E(\boldsymbol{\Xi}'\boldsymbol{\Xi}) = \boldsymbol{\Phi}$ (equation 4d). Replacing these quantities with their corresponding sample estimates yields $\hat{\mathbf{A}} = \hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Phi}}$. Finally, substituting into equation (13) yields $\hat{\boldsymbol{\Xi}} = \mathbf{X}\hat{\mathbf{A}} = \mathbf{X}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Phi}}$.

References

Anderson, T. W. 2003. *An Introduction to Multivariate Statistical Analysis*. Third ed. Hoboken, New Jersey: Wiley.

Bollen, Kenneth A. 1989. *Structural Equations With Latent Variables*. New York: Wiley.

Browne, M. W. 1984. ''Asymptotically distribution free methods for the analysis of covariance structures.'' *British Journal of Mathematical and Statistical Psychology* 37:62--83.

Everitt, B. S. 1984. *An Introduction to Latent Variable Models*. London: Chapman and Hall.

Jöreskog, Karl G. 1967. ''Some contributions to maximum likelihood factor analysis.'' *Psychometrika* 32:443--482.

Jöreskog, Karl G. and Dag Sörbom. 1989. *LISREL 7 User's Reference Guide*. Chicago: Scientific Software International Inc.

Kaiser, H. F. 1958. ''The varimax criterion for analytic rotation in factor analysis.'' *Psychometrika* 23:187--200.

Lawley, D. N. and A. E. Maxwell. 1971. *Factor Analysis as a Statistical Method*. Second ed. London: Butterworths.

Mardia, K.V., J.T. Kent and J.M. Bibby. 1979. *Multivariate Analysis*. San Diego: Academic Press.