Notes on the Statistics of Factor Analysis

Simon Jackman

Spring 2002

A Likelihood Model for Factor Analysis

If $\mathbf{x}_i$ is assumed to be multivariate normal $\forall\ i$, then the elements of the sample covariance matrix

$$\mathbf{S} = \frac{1}{n - 1}\mathbf{X}'\mathbf{X}$$

follow a Wishart distribution with $(n-1)$ degrees of freedom. The log-likelihood function is (neglecting terms constant with respect to $\mathbf{\Sigma}$)

$$\ln\mathcal{L} = -\frac{1}{2}(n - 1)\left[\ln|\mathbf{\Sigma}| + \text{tr}(\mathbf{S}\mathbf{\Sigma}^{-1})\right] \tag{1}$$

where the (orthogonal) factor analysis model postulates $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}$. In this way the log-likelihood is a function of the factor loadings $\mathbf{\Lambda}$ and the measurement error variances $\mathbf{\Psi}$.

Effective Number of Free Parameters

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

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

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

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

which follows from transforming the original model

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

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

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

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

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

which is true since

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

which is required by the model. In addition,

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

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

That is, while there are kp free parameters in $\mathbf{\Lambda}$, and k free parameters in $\mathbf{\Psi}$, the requirement that $\mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{\Lambda}$ 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 $\mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{\Lambda}$ 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 $\mathbf{\Lambda}$
- k measurement error variances to estimate (no error covariances) in $\mathbf{\Psi}$
- *less* $\frac{1}{2}p(p-1)$ parameters imposed by the constraint that $\mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{\Lambda}$ 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 estimated $\mathbf{\Lambda}$ are such that the observed covariance matrix $\mathbf{S}$ is perfectly recovered by $\hat{\mathbf{\Sigma}}$. In this case the log-likelihood reduces to

$$\ln\mathcal{L}_0 = -\frac{1}{2}(n - 1)\left[\ln|\mathbf{S}| + k\right], \tag{2}$$

since in this case $\mathbf{S} = \mathbf{\Sigma}$ and the trace of $\mathbf{I}_p$ is p.

Comparing the two log-likelihoods in equations (1) and (2) 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 $v =$ number of parameters 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.

Twice the difference between the two log-likelihoods is

$$(n - 1)F(\mathbf{S}, \mathbf{\Sigma}(\mathbf{\Lambda}, \mathbf{\Psi})) = (n - 1)\left[\ln|\mathbf{\Sigma}| + \mathrm{tr}(\mathbf{S}\mathbf{\Sigma}^{-1}) - \ln|\mathbf{S}| - k\right]. \tag{3}$$

Minimizing this function is equivalent to maximizing the log-likelihood (i.e., the optimal estimate of $\mathbf{\Sigma}$ is $\mathbf{S}$), and this is the function that is used in practice for testing goodness-of-fit.

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 in equation (3).

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. We could then repeat the factor analysis, with $p = 2$, 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).

Also, the test is heavily "statistical" and in my experience leads to an overly high number of factors being retained. 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 $\mathbf{X}$ that is distinguishable from sampling variability, so many factors added will have little explanatory power in substantive sense, though will be statistically significant.

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 $\mathbf{x}$ 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 $\mathbf{X}$ 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.

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

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

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

References

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

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

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