More on Factor Analysis: Estimation

36-350, Data Mining

29 September 2008

1 Recap

Let's recall what the factor analysis model looks like. We have n observations of p factors; X_{ij} is the value of feature j for object i. Our hypothesis is that these values arise from a linear combination of k factors, F_{ir}, plus noise, ϵ_{ij}:

$$X_{ij} = \epsilon_{ij} + \sum_{r=1}^{k} F_{ir} w_{rj} \tag{1}$$

We assume that the features have been centered to have mean zero.[1] We also assume that all the ϵ_{ij} have mean zero and are uncorrelated with each other and with the factors F_{ir}, but the variance depends on j, $\mathrm{Var}[\epsilon_{ij}] = \psi_j$. The parameters of the model are the factor loadings w_{rj}, the specific variances ψ_j, and the actual factor scores F_{ir}. We assume that the factors are uncorrelated with each other and uncorrelated across cases. For convenience (there's no loss of generality) we take the means of the factors to be zero and their variances to be one.

In matrix form, the model is

$$\mathbf{X} = \epsilon + \mathbf{F}\mathbf{w} \tag{2}$$

By taking the product $\mathbf{X}^T\mathbf{X}$, we get n times the sample covariance matrix $\mathbf{V}$, and so the model boils down to

$$\mathbf{V} = \Psi + \mathbf{w}^T\mathbf{w} \tag{3}$$

where Ψ is the diagonal matrix whose entries are the ψ_j. In this form, the factor scores have have disappeared. (Hopefully this will make the estimation problem simpler, and not impossible.)

We left the story just after observing Eq. 3 is really one equation for each element of $\mathbf{V}$, i.e., p^2 equations, but that there are only $p + pk$ unknowns on the right hand side (the diagonal elements of Ψ, plus the elements of $\mathbf{w}$), and that systems with more equations than unknowns generally cannot be solved. This makes it sound like it's actually impossible to estimate the factor analysis model!

Let's see how to dig ourselves out of this hole.

[1] It's common to also standardize them to have variance 1, but not necessary.

2 Estimation by Linear Algebra

The means of escape is linear algebra.

2.1 A Clue from Spearman's One-Factor Model

Remember from last time Spearman's model with a single general factor. The covariance between features a and b in that model is the product of their factor weightings:

$$V_{ab} = w_a w_b$$

The exception is that $V_{aa} = w_a^2 + \psi_a$, rather than w_a^2. However, if we look at $\mathbf{U} = \mathbf{V} - \mathbf{\Psi}$, that's the same as $\mathbf{V}$ off the diagonal, and a little algebra shows that its diagonal entries are, in fact, just w_a^2. (EXERCISE: Do that algebra.) So if we look at any two rows of $\mathbf{U}$, they're proportional to each other:

$$U_{a\cdot} = \frac{w_a}{w_b} U_{b\cdot}$$

This means that, when Spearman's model holds true, there is actually only *one* linearly-independent row in in $\mathbf{U}$. Rather than having p^2 equations, we've only got p independent equations.[2]

Recall from linear algebra that the **rank** of a matrix is how many linearly independent rows it has.[3] Ordinarily, the matrix is of **full rank**, meaning all the rows are linearly independent. What we have just seen is that when Spearman's model holds, the matrix $\mathbf{U}$ is *not* of full rank, but rather of rank 1. More generally, when the factor analysis model holds with k factors, the matrix has rank k.

2.2 Estimating Factor Loadings and Specific Variances

We are now in a position to set up the classic method for estimating the factor model.

As in previous subsection, define $\mathbf{U} = \mathbf{V} - \mathbf{\Psi}$. This is the **reduced** or **adjusted** covariance matrix. The diagonal entries are no longer the variances of the features, but the variances minus the specific variances. These **common variances** or **commonalities** show how much of the variance in each feature is associated with the variances of the latent factors. $\mathbf{U}$ is still, like $\mathbf{V}$, a positive symmetric matrix. We can't actually calculate $\mathbf{U}$ until we know, or have a guess, as to $\mathbf{\Psi}$. A reasonable and common starting-point is to do a linear regression of each feature j on all the other features, and then set ψ_j to the residual sum of squares for that regression.

Because $\mathbf{U}$ is a positive symmetric matrix, we know from linear algebra that it can be written as

$$\mathbf{U} = CDC^T \tag{4}$$

[2] This creates its own problems when we try to estimate the factor scores, as we'll see.

[3] We could also talk about the columns; it wouldn't make any difference.

where $\mathbf{C}$ is the matrix whose columns are the eigenvectors of $\mathbf{U}$, and $\mathbf{D}$ is the diagonal matrix whose entries are the eigenvalues. That is, if we use all p eigenvectors, we can reproduce the covariance matrix exactly. Suppose we instead use $\mathbf{C_k}$, the $p \times k$ matrix whose columns are the eigenvectors going with the k largest eigenvalues, and likewise make $\mathbf{D_k}$ the diagonal matrix of those eigenvalues. Then $\mathbf{C_k D_k C_k}^T$ will be a symmetric positive $p \times p$ matrix. It won't *quite* equal $\mathbf{U}$, but it will come closer as we let k grow towards p, and at any given k, this matrix comes closer to being $\mathbf{U}$ than any other we could put together which had rank k.

Now define $\mathbf{D_k}^{1/2}$ as the $k \times k$ diagonal matrix of the square roots of the eigenvalues. Clearly $\mathbf{D_k} = \mathbf{D_k}^{1/2}\mathbf{D_k}^{1/2}$. So

$$\mathbf{C_k D_k C_k}^T = \mathbf{C_k D_k}^{1/2}\mathbf{D_k}^{1/2}\mathbf{C_k}^T = \left(\mathbf{C_k D_k}^{1/2}\right)\left(\mathbf{C_k D_k}^{1/2}\right)^T \tag{5}$$

So we have

$$\mathbf{U} \approx \left(\mathbf{C_k D_k}^{1/2}\right)\left(\mathbf{C_k D_k}^{1/2}\right)^T \tag{6}$$

but at the same time we know that $\mathbf{U} = \mathbf{w}^T\mathbf{w}$. So first we identify $\mathbf{w}$ with $\left(\mathbf{C_k D_k}^{1/2}\right)^T$:

$$\widehat{\mathbf{w}} = \left(\mathbf{C_k D_k}^{1/2}\right)^T \tag{7}$$

Now we use $\mathbf{w}$ to re-set $\mathbf{\Psi}$, so as to fix the diagonal entries of the covariance matrix.

$$\widehat{\mathbf{w}} = \left(\mathbf{C_k D_k}^{1/2}\right)^T \tag{8}$$

$$\widehat{\psi_j} = V_{jj} - \sum_{r=1}^{k} w_{rj}^2 \tag{9}$$

$$\mathbf{V} \approx \widehat{\mathbf{V}} \equiv \widehat{\mathbf{\Psi}} + \widehat{\mathbf{w}}^T\widehat{\mathbf{w}} \tag{10}$$

The "predicted" covariance matrix $\widehat{\mathbf{V}}$ in the last line is exactly right on the diagonal (by construction), and should be closer off-diagonal than anything else we could do with the same number of factors — i.e., the same rank for the $\mathbf{U}$ matrix. However, our estimate of $\mathbf{U}$ itself has in general changed, so we can try iterating this (i.e., re-calculating $\mathbf{C_k}$ and $\mathbf{D_k}$), until nothing changes.

Let's think a bit more about how well we're approximating $\mathbf{V}$. The approximation will always be exact when $k = p$, so that there is one factor for each feature (in which case $\mathbf{\Psi} = 0$ always). Then all factor analysis does for us is to rotate the coordinate axes in feature space, so that the new coordinates are uncorrelated. (This is the same was what PCA does with p components.) The approximation can *also* be exact with fewer factors than features if the reduced covariance matrix is of less than full rank, and we use at least as many factors as the rank.

3 Maximum Likelihood Estimation

It has probably not escaped your notice that the estimation procedure above requires a starting guess as to $\mathbf{\Psi}$. This makes its consistency somewhat shaky. (If we continually put in ridiculous values for $\mathbf{\Psi}$, there's no reason to expect that $\widehat{\mathbf{w}} \to \mathbf{w}$, even with immensely large samples.) On the other hand, we know from our elementary statistics courses that maximum likelihood estimates are generally consistent, unless we choose a spectacularly bad model. Can we use that here?

We can, but at a cost. We have so far got away with just making assumptions about the means and covariances of the factor scores $\mathbf{F}$. To get an actual likelihood, we need to assume something about their distribution as well.

The usual assumption is that $F_{ir} \sim \mathcal{N}(0,1)$, and that the factor scores are independent across factors $r = 1, \ldots k$ and individuals $i = 1, \ldots n$. With this assumption, the features have a multivariate normal distribution $\vec{X}_i \sim \mathcal{N}(0, \mathbf{\Psi} + \mathbf{w}^T\mathbf{w})$. This means that the log-likelihood is

$$L = -\frac{np}{2}\log 2\pi - \frac{n}{2}\log|\mathbf{\Psi} + \mathbf{w}^T\mathbf{w}| - \frac{n}{2}\,\mathrm{tr}\left((\mathbf{\Psi} + \mathbf{w}^T\mathbf{w})^{-1}\mathbf{V}\right) \tag{11}$$

where $\mathrm{tr}\,\mathbf{A}$ is the **trace** of the matrix $\mathbf{A}$, the sum of its diagonal elements.

One can either try direct numerical maximization, or use a two-stage procedure. Starting, once again, with a guess as to $\mathbf{\Psi}$, one finds that the optimal choice of $\mathbf{\Psi}^{1/2}\mathbf{w}^T$ is given by the matrix whose columns are the k leading eigenvectors of $\mathbf{\Psi}^{1/2}\mathbf{V}\mathbf{\Psi}^{1/2}$. Starting from a guess as to $\mathbf{w}$, the optimal choice of $\mathbf{\Psi}$ is given by the diagonal entries of $\mathbf{V} - \mathbf{w}^T\mathbf{w}$. So again one starts with a guess about the unique variances (e.g., the residuals of the regressions) and iterates to convergence.[4]

The differences between the maximum likelihood estimates and the "principal factors" approach can be substantial. If the data appear to be normally distributed (as shown by the usual tests), then the additional efficiency of maximum likelihood estimation is highly worthwhile. Also, as we'll see next time, it is a lot easier to test the model assumptions is one uses the MLE.

3.1 Estimating Factor Scores

The probably the best method for estimating factor scores is the "regression" or "Thomson" method, which says

$$\widehat{F}_{ir} = \sum_j X_{ij} b_{ij} \tag{12}$$

and seeks the weights b_{ij} which will minimize the mean squared error, $\mathbf{E}[(\widehat{F}_{ir} - F_{ir})^2]$. You will see how this works in a homework problem.

[4] The algebra is tedious. See section 3.2 in Bartholomew (1987) if you really want it. (Note that Bartholomew has a sign error in his equation 3.16.)

4 The Rotation Problem

Recall from linear algebra that a matrix $\mathbf{O}$ is **orthogonal** if its inverse is the same as its transpose, $\mathbf{O}^T\mathbf{O} = \mathbf{I}$. The classic examples are rotation matrices. For instance, to rotate a two-dimensional vector through an angle α, we multiply it by

$$\mathbf{R}_\alpha = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix}$$

The inverse to this matrix must be the one which rotates through the angle $-\alpha$, $\mathbf{R}_\alpha^{-1} = \mathbf{R}_{-\alpha}$, but trigonometry tells us that $\mathbf{R}_{-\alpha} = \mathbf{R}_\alpha^T$.

To see why this matters to us, go back to the matrix form of the factor model, and insert an orthogonal matrix $k \times k$ and its transpose:

$$\begin{aligned} \mathbf{X} &= \epsilon + \mathbf{F}\mathbf{w} && (13) \\ &= \epsilon + \mathbf{F}\mathbf{O}\mathbf{O}^T\mathbf{w} && (14) \\ &= \epsilon + \mathbf{G}\mathbf{u} && (15) \end{aligned}$$

We've changed the factor scores to $\mathbf{G} = \mathbf{F}\mathbf{O}$, and we've changed the factor loadings to $\mathbf{u} = \mathbf{O}^T\mathbf{w}$, but nothing about the features has changed *at all*. We can do as many orthogonal transformations of the factors as we like, with no observable consequences whatsoever.[5]

Mathematically, this should not be all that surprising. The factor live in a k-dimensional vector space of their own. We should be free to set up any coordinate system we feel like on that space. Changing coordinates in factor space will, however, require a compensating change in how factor space relates to feature space (the factor loadings matrix $\mathbf{w}$). That's all we've done here with our orthogonal transformation.

Substantively, this should be rather troubling. If we can rotate the factors as much as we like without consequences, how on Earth can we interpret them?

References

Bartholomew, David J. (1987). *Latent Variable Models and Factor Analysis*. New York: Oxford University Press.

[5] Notice that the log-likelihood only involves $\mathbf{w}^T\mathbf{w}$, which is equal to $\mathbf{w}^T\mathbf{O}\mathbf{O}^T\mathbf{w} = \mathbf{u}^T\mathbf{u}$, so even assuming Gaussian distributions doesn't change things.