Factor Analysis

36-350, Data Mining

23 September 2009

Contents

1 From Principal Components to Factor Analysis

There are two ways to go from principal components analysis to factor analysis — two motivating stories.

1.1 Measurement Error

Suppose that the numbers we write down as our observations aren't altogether accurate — that our numbers are the true variables plus some measurement noise. (Or, if we're not making the measurements ourselves but just taking numbers from some database, that whoever created the database wasn't able to measure things perfectly.) PCA doesn't care about this — it will try to reproduce true-value-plus-noise from a small number of components. But that's kind of weird — why try to reproduce the noise?[1] Can we do something like PCA, where we reduce a large number of features to additive combinations of a smaller number of variables, but which allows for noise?

The simplest model, starting from PCA, would be something like this. Each object or record has p features, so X_{ij} is the value of feature j for object i. As before, we'll center all the observations (subtract off their mean). We now postulate that there are q **factor** variables, and each observation is a linear combination of **factor scores** F_{ir} plus noise:

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

The weights w_{rj} are called the **factor loadings** of the observable features; they say how much feature j changes, on average, in response to a one-unit change in factor score r. Notice that we are allowing each feature to go along with more than one factor (for a given j, w_{rj} can be non-zero for multiple r). This would correspond to our measurements running together what are really distinct variables.

Here ϵ_{ij} is as usual the noise term for feature j on object i. We'll assume this has mean zero and variance ψ_j — i.e., different features has differently-sized noise terms. The ψ_j are known as the **specific variances**, because they're specific to individual features. We'll further assume that $\mathbf{E}\left[\epsilon_{ij}\epsilon_{lm}\right] = 0$, unless $i = l$, $j = m$ — that is, each object and each feature has uncorrelated noise.

We can also re-write the model in vector form,

$$\vec{X_i} = \vec{\epsilon_i} + \vec{F_i}\mathbf{w} \tag{2}$$

with $\mathbf{w}$ being a $q \times p$ matrix. If we stack the vectors into a matrix, we get

$$\mathbf{X} = \epsilon + \mathbf{F}\mathbf{w} \tag{3}$$

This is the factor analysis model. The only (!) tasks are to estimate the factor loadings $\mathbf{w}$, the factor scores $\mathbf{F}$, and the specific variances ψ_j.

A common question at this point is, or should be, where does the model (1) come from? The answer is, we *make it up*. More formally, we *posit* it, and all the stuff about the distribution of the noise, etc., as a *hypothesis*. All the rest of our reasoning is conditional, premised on the assumption that the posited hypothesis is in fact true. It is unfortunately too common to find people who just state the hypothesis in a semi-ritual manner and go on. What we should really do is try to *test* the hypothesis, i.e., to check whether it's actually right. We will come back to this.

1.2 Preserving correlations

PCA aims to preserve variance, or (what comes to the same thing) minimize mean-squared residuals (reconstruction error). But it doesn't preserve correlations. That is, the correlations of the features of the image vectors are not

[1] One reason would be if we're not sure what's noise, or if what seems to be noise for one purpose is signal for something else. But let's press onward.

the same as the correlations among the features of the original vectors (unless $q = p$, and we're not really doing any data reduction). We might value those correlations, however, and want to preserve them, rather than the variance.[2] That is, we might ask for a set of vectors whose image in the feature space will have the same correlation matrix as the original vectors, or as close to the same correlation matrix as possible while still reducing the number of dimensions.

This *also* leads to the factor analysis model, as we'll see, but we need to take a somewhat circuitous root to get there.

1.3 Roots of Factor Analysis in Causal Discovery

The roots of factor analysis go back to work by Charles Spearman just over a century ago (Spearman, 1904); he was trying to discover the hidden structure of human intelligence. His observation was that schoolchildren's grades in different subjects were all correlated with each other. He went beyond this to observe a particular *pattern* of correlations, which he thought he could explain as follows: the reason grades in math, English, history, etc., are all correlated is performance in these subjects is all correlated with *something else*, a general or **common** factor, which he named "general intelligence", for which the natural symbol was of course g or G.

Put in a form like Eq. 1, Spearman's model becomes

$$X_{ij} = \epsilon_{ij} + G_i w_j \tag{4}$$

(Since there's only one common factor, the factor loadings w_j need only one subscript index.) If we assume that the features and common factor are all centered to have mean 0, and that there is no correlation between ϵ_{ij} and G_i for any j, then the correlation between the j^{th} feature, $X_{\cdot j}$, and G is just w_j.

Now we can begin to see how factor analysis reproduces correlations. Under these assumptions, it follows that the correlation between the j^{th} feature and the l^{th} feature, call that ρ_{jl}, is just the product of the factor loadings:

$$\rho_{jl} = w_j w_l \tag{5}$$

Up to this point, this is all so much positing and assertion and hypothesis. What Spearman did next, though, was to observe that this hypothesis carried a very strong implication about the *ratios* of correlation coefficients. Pick any

[2] Why? Well, originally the answer was that the correlation coefficient had just been invented, and was about the only way people had of measuring relationships between variables. Since then it's been propagated by statistics courses where it is the only way people are *taught* to measure relationships. The great statistician John Tukey once wrote "Does anyone know when the correlation coefficient is useful? If so, why don't they tell us?"

four features, j, l, r, s. Then, if the model (4) is true,

$$\begin{aligned}
\frac{\rho_{jr}/\rho_{lr}}{\rho_{js}/\rho_{ls}} &= \frac{w_j w_r / w_l w_r}{w_j w_s / w_l w_s} && (6)\\
&= \frac{w_j/w_l}{w_j/w_l} && (7)\\
&= 1 && (8)
\end{aligned}$$

The relationship

$$\rho_{jr}\rho_{ls} = \rho_{js}\rho_{lr} \tag{9}$$

is called the “tetrad equation”, and we will meet it again later when we consider methods for causal discovery.

Spearman found that the tetrad equation held in his data on school grades (to a good approximation), and concluded that a single general factor of intelligence must exist. This was, of course, logically fallacious.

Later work, using large batteries of different kinds of intelligence tests, showed that the tetrad equation does not hold in general, or more exactly that departures from it are too big to explain away as sampling noise. (Recall that the equations are about the true correlations between the variables, but we only get to see sample correlations, which are always a little off.) The response, done in an *ad hoc* way by Spearman and his followers, and then more systematically by Thurstone, was to introduce *multiple* factors. This breaks the tetrad equation, but still accounts for the correlations among features by saying that features are really directly correlated with factors, and uncorrelated conditional on the factor scores.[3] Thurstone's form of factor analysis is basically the one people still use — there have been refinements, of course, but it's mostly still his method.

2 Preliminaries to Factor Estimation

Assume all the factor scores are uncorrelated with each other and have variance 1; also that they are uncorrelated with the noise terms. We'll solve the estimation problem for factor analysis by reducing it to an eigenvalue problem again.

Start from the matrix form of the model, Eq. 3, which you'll recall was

$$\mathbf{X} = \epsilon + \mathbf{F}\mathbf{w} \tag{10}$$

We know that $\mathbf{X}^T\mathbf{X}$ is a $p \times p$ matrix, in fact it's n times the sample covariance

[3] You can (and should!) read the classic “The Vectors of Mind” paper (Thurstone, 1934) online.

matrix **V**. So

$$\begin{aligned}
n\mathbf{V} &= \mathbf{X}^T\mathbf{X} && (11)\\
&= (\epsilon + \mathbf{F}\mathbf{w})^T (\epsilon + \mathbf{F}\mathbf{w}) && (12)\\
&= \left(\epsilon^T + \mathbf{w}^T\mathbf{F}^T\right)(\epsilon + \mathbf{F}\mathbf{w}) && (13)\\
&= \epsilon^T\epsilon + \epsilon^T\mathbf{F}\mathbf{w} + \mathbf{w}^T\mathbf{F}^T\epsilon + \mathbf{w}^T\mathbf{F}^T\mathbf{F}\mathbf{w} && (14)\\
&= n\Psi + 0 + 0 + n\mathbf{w}^T I\mathbf{w} && (15)\\
&= n\Psi + n\mathbf{w}^T\mathbf{w} && (16)\\
\mathbf{V} &= \Psi + \mathbf{w}^T\mathbf{w} && (17)
\end{aligned}$$

where Ψ is the diagonal matrix whose entries are the ψ_j. The cross-terms cancel because the factor scores are uncorrelated with the noise, and the $\mathbf{F}^T\mathbf{F}$ term is just n times the covariance matrix of the factor scores, which by assumption is the identity matrix.

At this point, the actual factor scores have dropped out of the problem, and all we are left with are the more "structural" parameters, namely the factor loadings **w** and the specific variances ψ_j. We know, or rather can easily estimate, the covariance matrix **V**, so we want to solve Eq. 17 for these unknown parameters.

The problem is that we want $q < p$, but on its face (17) gives us p^2 equations, one for each entry of **V**, and only $p + pq$ unknowns (the diagonal elements of Ψ, plus the elements of **w**). 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![4]

3 Estimation by Linear Algebra

The means of escape is linear algebra.

3.1 A Clue from Spearman's One-Factor Model

Remember that in 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:

$$V_{ab} = w_a w_b \tag{18}$$

[4] Actually, the book-keeping for the number of degrees of freedom is a little more complicated, though the point is sound. First of all, there are not p^2 independent equations but only $p(p+1)/2$ of them, because **V** is a symmetric matrix. (Since Ψ is diagonal, $\Psi + \mathbf{w}^T\mathbf{w}$ is automatically symmetric.) On the other hand, each of the q rows of **w** must be orthogonal to all the others, which gives $q(q-1)/2$ constraints on the unknowns. So the number of degrees of freedom for **V** is $p(p+1)/2$, and the number of degrees of freedom for the unknown parameters is $p + pq - q(q-1)/2$. If the former exceeds the later, there are degrees of freedom left over to estimate the parameters — but there may be no exact solution. If on the other hand the parameters have more degrees of freedom than **V** does, then there cannot possibly be a unique solution, and the model is hopelessly unidentifiable no matter how much data we have. Most software, including R's default factor analysis function, will simply refuse to work with such a model.

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. So if we look at any two rows of $\mathbf{U}$, they're proportional to each other:

$$U_{a\cdot} = \frac{w_a}{w_b} U_{b\cdot} \tag{19}$$

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.[5]

Recall from linear algebra that the **rank** of a matrix is how many linearly independent rows it has.[6] 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 q factors, the matrix has rank q.

3.2 Estimating Factor Loadings and Specific Variances

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

As above, 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 mean squared error for that regression.

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

$$\mathbf{U} = CDC^T \tag{20}$$

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 $\mathbf{C_q}$, the $p \times q$ matrix whose columns are the eigenvectors going with the q largest eigenvalues, and likewise make $\mathbf{D_q}$ the diagonal matrix of those eigenvalues. Then $\mathbf{C_q}\mathbf{D_q}\mathbf{C_q}^T$ will be a symmetric positive $p \times p$ matrix. It won't *quite* equal $\mathbf{U}$, but it will come closer as we let q grow towards p, and at any given q, this matrix comes closer to being $\mathbf{U}$ than any other we could put together which had rank q.

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

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

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

$$\mathbf{C_q}\mathbf{D_q}{\mathbf{C_q}}^T = \mathbf{C_q}{\mathbf{D_q}}^{1/2}{\mathbf{D_q}}^{1/2}{\mathbf{C_q}}^T = \left(\mathbf{C_q}{\mathbf{D_q}}^{1/2}\right)\left(\mathbf{C_q}{\mathbf{D_q}}^{1/2}\right)^T \tag{21}$$

So we have

$$\mathbf{U} \approx \left(\mathbf{C_q}{\mathbf{D_q}}^{1/2}\right)\left(\mathbf{C_q}{\mathbf{D_q}}^{1/2}\right)^T \tag{22}$$

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

$$\widehat{\mathbf{w}} = \left(\mathbf{C_q}{\mathbf{D_q}}^{1/2}\right)^T \tag{23}$$

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

$$\begin{aligned}
\widehat{\mathbf{w}} &= \left(\mathbf{C_q}{\mathbf{D_q}}^{1/2}\right)^T && (24)\\
\widehat{\psi_j} &= V_{jj} - \sum_{r=1}^{k} w_{rj}^2 && (25)\\
\mathbf{V} &\approx \widehat{\mathbf{V}} \equiv \widehat{\mathbf{\Psi}} + \widehat{\mathbf{w}}^T\widehat{\mathbf{w}} && (26)
\end{aligned}$$

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 $\mathbf{C_q}$ and $\mathbf{D_q}$), until nothing changes.

Let's think a bit more about how well we're approximating $\mathbf{V}$. The approximation will always be exact when $q = p$, 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.

4 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 $r = 1, \ldots q$ 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

$$L = -\frac{np}{2}\log 2\pi - \frac{n}{2}\log|\mathbf{\Psi} + \mathbf{w}^T\mathbf{w}| - \frac{n}{2}\operatorname{tr}\left((\mathbf{\Psi} + \mathbf{w}^T\mathbf{w})^{-1}\mathbf{V}\right) \tag{27}$$

where $\operatorname{tr}\mathbf{A}$ 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 q 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.[7]

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.

4.1 Estimating Factor Scores

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

$$\widehat{F_{ir}} = \sum_j X_{ij} b_{ij} \tag{28}$$

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.

5 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

$$\mathbf{R}_\alpha = \left[\begin{array}{cc} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{array}\right] \tag{29}$$

[7] 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.)

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 $q \times q$ matrix and its transpose:

$$\begin{aligned} \mathbf{X} &= \epsilon + \mathbf{F}\mathbf{w} & (30) \\ &= \epsilon + \mathbf{F}\mathbf{O}\mathbf{O}^{T}\mathbf{w} & (31) \\ &= \epsilon + \mathbf{G}\mathbf{u} & (32) \end{aligned}$$

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.[8]

Statistically, the fact that different parameter settings give us the same observational consequences means that the parameters of the factor model are **unidentifiable**. The rotation problem is, as it were, the revenant of having an ill-posed problem: we thought we'd slain it through heroic feats of linear algebra, but it's still around and determined to have its revenge.

Mathematically, this should not be surprising at all. The factor live in a q-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 coordinates relate 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?

Exercises

1. Prove Eq. 5.

2. Why is it fallacious to go from "the data have the kind of correlations predicted by a one-factor model" to "the data were generated by a one-factor model"?

3. Show that the correlation between the j^{th} feature and G, in the one-factor model, is w_j.

4. Show that the diagonal entries of $\mathbf{U} = \mathbf{V} - \mathbf{\Psi}$ are given by w_a^2.

References

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

[8] 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.

Spearman, Charles (1904). ““General Intelligence,” Objectively Determined and Measured.” *American Journal of Psychology*, **15**: 201–293. URL `http://psychclassics.yorku.ca/Spearman/`.

Thurstone, L. L. (1934). “The Vectors of Mind.” *Psychological Review*, **41**: 1–32. URL `http://psychclassics.yorku.ca/Thurstone/`.