Factor Analysis

Robert Jacobs
Department of Brain & Cognitive Sciences
University of Rochester
Rochester, NY 14627, USA

August 8, 2008

Reference: Much of the material in this note was taken from Krzanowski, W. J. (1988). *Principles of Multivariate Analysis*. Oxford, UK: Oxford University Press.

We seek a model that 'explains' the observed associations between variables. Suppose that p variables $x = [x_1, \ldots, x_p]^T$ have been observed on each of n sample individuals. Also suppose that the variables are continuous, and that associations between them are therefore measured by correlation coefficients. Let $\mathbf{R}$ denote the matrix of correlation coefficients. Our objective is an adequate 'explanation' of all the entries of $\mathbf{R}$. It is evident that any satisfactory explanation must draw on information from outside this set of correlation values, as otherwise we will become enmeshed in circular arguments. Let us therefore suppose that the relevant extra information resides in a further q variables $z = [z_1, \ldots, z_q]^T$ that could be measured on each sample member (but have not been measured).

To see how extra variables can explain the observed correlations, recall that a correlation r_{ij} between two variables x_i and x_j could arise as a result of their mutual association with an extraneous variable z_k. The variation in z_k values over the sample, and the associations between z_k and each of x_i and x_j, induces an association between x_i and x_j. The partial correlation $r_{ij|k}$ measures the association between x_i and x_j when the value of z_k is held fixed, and is therefore the 'residual' correlation between x_i and x_j after removal of the (linear) effect of z_k on each. If this partial correlation is close to zero (or more precisely if the null hypothesis of zero population partial correlation is tenable), then we can say that z_k has 'explained' the correlation r_{ij} between x_i and x_j. If $r_{ij|k}$ is not close to zero, then there still exists some association between x_i and x_j that is unexplained, and we need to consider a further variable z_l to account for this remaining correlation. If the partial correlation $r_{ij|kl}$ is not significantly different from zero, then z_k and z_l jointly explain the correlation r_{ij}. Otherwise we consider a third variable z_m, and so on. A satisfactory explanation of all the entries in $\mathbf{R}$ will thus be obtained when we find a set of variables $z = [z_1, \ldots, z_q]^T$ such that the partial correlation between any two variables x_i and x_j, on fixing the values of z, is not significantly different from zero. The most parsimonious explanation is achieved when q is as small as possible.

The converse of the above argument is that if the variables $z_1, \ldots, z_q$ are to provide a complete explanation of the entries of $\mathbf{R}$, then the partial correlation between any two elements of x for a fixed value of z must be compatible with a population value of zero. This means that if only those individuals with specified values of z were to be sampled, then no association would be detectable between any two elements of x. In other words, x_i and x_j are conditionally independent given the

values of $z_1, \ldots, z_q$, for all i and j. This property of local independence is a necessary requirement for a set of variables z to provide an explanation for the entries $\mathbf{R}$.

As it stands, the above discussion is an academic one because we don't know anything about the $z_1, \ldots, z_q$ variables (we don't know what they are, we don't know how many of them there are, and we don't know their values). Our data simply consists of n observations on each of the p variables $x_1, \ldots, x_p$. The great insight, however, comes from assuming that $z_1, \ldots, z_q$ are unobserved random variables. Therefore, it may be possible to estimate their moments (e.g., means, variances) through statistical analysis. Since the $z_i \;\; (i = 1, \ldots, q)$ are unobservable, they are termed *latent* variables. Sometimes they are also known as *factors*.

Now let's consider the factor analysis model. The first step is an appropriate model, which in this case is any model that satisfies the requirement of local independence outlined above. The factor analysis model is the simplest model to satisfy this requirement. We assume that x is a linear function of z:

$$\begin{aligned} x_1 &= \mu_1 + \gamma_{11} z_1 + \ldots + \gamma_{1q} z_q + e_1 \\ x_2 &= \mu_2 + \gamma_{21} z_1 + \ldots + \gamma_{2q} z_q + e_2 \\ &\vdots \\ x_p &= \mu_p + \gamma_{p1} z_1 + \ldots + \gamma_{pq} z_q + e_p \end{aligned} \tag{1}$$

where the μ_i and γ_{ij} are constants, while the z_i and e_i are random variables. The minimal set of assumptions about these random variables (to ensure that the local independence property is satisfied) is that the e_i are uncorrelated with each other and with the z_i. It is then evident that if the values of the z_i are specified we can write $x_i = v_i + e_i \;\; (i = 1, \ldots, p)$ where the v_i are constants, so that $\text{corr}(x_i, x_j) = 0$ for all i, j.

It'll be useful if we use vector notation:

$$x = \mu + \mathbf{\Gamma} z + e. \tag{2}$$

It is worth re-writing this as

$$x - \mu = \mathbf{\Gamma} z + e. \tag{3}$$

Note that the entire right-hand side of this equation consists of conceptual entities (i.e. unobservable variables). In order to make any progress with this equation, we need to make some assumptions. In the factor analysis model, we assume that z and e are random variables with Normal distributions. In regard to z, we assume that its mean is the zero vector and that its covariance matrix is the identity matrix $\mathbf{I}$. In regard to e, we assume that its mean is the zero vector and that its covariance matrix, denoted $\mathbf{\Psi}$, is a diagonal matrix whose diagonal entries are $\psi_1^2, \ldots, \psi_p^2$. Given these assumptions, it is easy to show that x is Normally distributed with mean μ and covariance matrix $\mathbf{\Gamma}\mathbf{\Gamma}^T + \mathbf{\Psi}$.

As an aside, let's show this:

$$E[(x - \mu)(x - \mu)^T] \;\; = \;\; E[(\mu + \mathbf{\Gamma} z + e - \mu)(\mu + \mathbf{\Gamma} z + e - \mu)^T]$$

$$
\begin{aligned}
&= E[(\mathbf{\Gamma}z + e)(\mathbf{\Gamma}z + e)^T] \\
&= E[\mathbf{\Gamma}zz^T\mathbf{\Gamma}^T] + E[\mathbf{\Gamma}ze^T] + E[ez^T\mathbf{\Gamma}^T] + E[ee^T] \\
&= \mathbf{\Gamma}E[zz^T]\mathbf{\Gamma} + \mathbf{\Gamma}E[ze^T] + E[ez^T]\mathbf{\Gamma} + E[ee^T] \\
&= \mathbf{\Gamma}\mathbf{I}\mathbf{\Gamma} + \mathbf{\Gamma}0 + 0\mathbf{\Gamma} + \Psi \\
&= \mathbf{\Gamma}\mathbf{\Gamma}^T + \Psi.
\end{aligned}
$$

As a second aside, let's also examine the covariance between x and z:

$$
\begin{aligned}
E[(x-\mu)z^T] &= E[(\mu + \mathbf{\Gamma}z + e - \mu)z^T] \\
&= E[(\mathbf{\Gamma}z + e)z^T] \\
&= \mathbf{\Gamma}E[zz^T] + E[ez^T] \\
&= \mathbf{\Gamma}.
\end{aligned}
$$

Using these asides, the joint distribution of data x and factors z is

$$
P\left(\begin{bmatrix} x \\ z \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} \mu \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{\Gamma}\mathbf{\Gamma}^T + \Psi & \mathbf{\Gamma} \\ \mathbf{\Gamma}^T & \mathbf{I} \end{bmatrix}\right). \tag{4}
$$

Consequently, we can compute $P(x|z)$ or $P(z|x)$ (see the appendix at the end of this note for how to compute conditional distributions from a joint distribution that is Normal). For example, the probability of $x^{(t)}$ given $z^{(t)}$ (i.e. x and z for data item t) is the Normal distribution:

$$
(2\pi)^{p/2}|\Psi|^{-1/2}\exp\{-\frac{1}{2}(x^{(t)} - \mu - \mathbf{\Gamma}z^{(t)})^T\Psi^{-1}(x^{(t)} - \mu - \mathbf{\Gamma}z^{(t)})\}. \tag{5}
$$

The joint probability over all data items $x^{(t)}, t = 1, \ldots, T$, is the likelihood function which is maximized when one estimates the values of the free parameters of the model (these parameters are the matrix of linear coefficients $\mathbf{\Gamma}$ and the covariance matrix Ψ). For reasons not described here, it can be tricky to find parameter values that maximize this function and so specialized techniques are often used. One may also be interested in the probability of $z^{(t)}$ given $x^{(t)}$. This is given by a Normal distribution whose mean is

$$
\mathbf{\Gamma}^T(\mathbf{\Gamma}\mathbf{\Gamma}^T + \Psi)^{-1}(x^{(t)} - \mu). \tag{6}
$$

To make this discussion concrete, let's consider an example, namely intelligence testing. Let $x_1, \ldots, x_p$ be the scores obtainable in a battery of tests (e.g., arithmetic, algebra, history, reading, and comprehension), so that the observations for the ith individual in a sample would be denoted by $x_i = [x_{i1}, \ldots, x_{ip}]^T$ where $i = 1, \ldots, n$. The different tests exhibit associations as given by the sample correlation matrix $\mathbf{R}$. To explain the associations, the factor analysis model postulates that each test score is made up of contributions from a number q of common factors (the z_i), together with a 'residual' specific to that test (the e_i). In the example under consideration, common factors

relevant to the given tests might be 'intelligence' (z_1), 'numerical ability' (z_2), 'verbal ability' (z_3), and 'memory' (z_4). Each test requires a combination of these skills, but clearly each skill will be more important for some tests than for others. For example, we would expect arithmetic to have heavier contributions from z_1 and z_2 than from z_3 and z_4, while history would not require z_2 at all but would rely about equally on the other three qualities. The constant γ_{ij} above expresses the importance of factor z_j in test x_i, and is usually known as the loading of factor j on test i. Each individual in the sample is assumed to possess a value for each of these factors (the set of factor scores for that individual), but these values are unobservable.

While the reasonableness of such arguments may not be in doubt, it nevertheless remains true that the factor analysis model seems a very flimsy basis for which to attempt an explanation of a system, because of the lack of observable quantities in it. From the available data (the n vectors x_i), we need to estimate:

- The number q of common factors (clearly we are interested in the smallest value of q that yields an adequate fit to the data).
- The factor loadings γ_{ij}. Since the factors are unobservable, the factor loadings provide the only means of 'labeling' each factor. By identifying which factors have high loading for each variable x_i, we can perhaps attach meanings to the factors. Thus high loadings on factor 1 for all tests in the example above would identify z_1 as 'general intelligence'; high loadings on factor 2 for x_1 and x_2 (arithmetic and algebra) but low loadings for the other x_i would suggest that z_2 was 'numerical ability'.
- The specific variances $\psi_1^2, \ldots, \psi_p^2$. These quantities determine how much of the variability of each variable is not attributable to the common factors.
- The factor scores $z_i = [z_{i1}, \ldots, z_{iq}]^T$ which provide a ranking or scaling of the sample individuals with respect to each identified factor.

Finally, it is worth comparing and contrasting factor analysis models with principal component models. Recall that in PCA we seek a new set of variables $y_1, \ldots, y_p$ as linear combinations of the observed (mean-centered) variables $x_1, \ldots, x_p$ in such a way as to maximize successively the variance of the y_i. If λ_i is the ith largest eigenvalue of the covariance matrix of $x = [x_1, \ldots, x_p]^T$, and $\alpha_i = [\alpha_{i1}, \ldots, \alpha_{ip}]^T$ is its corresponding eigenvector, then the principal components are given by

$$\begin{aligned} y_1 &= \alpha_{11}x_1 + \cdots + \alpha_{1p}x_p \\ &\vdots \\ y_p &= \alpha_{p1}x_1 + \cdots + \alpha_{pp}x_p \end{aligned} \tag{7}$$

and $\text{var}(y_i) = \lambda_i \;\; (i = 1, \ldots, p)$. Since the matrix (α_{ij}) is orthogonal, we can invert this transformation to give

$$\begin{aligned} x_1 &= \alpha_{11}y_1 + \cdots + \alpha_{p1}y_p \\ &\vdots \\ x_p &= \alpha_{1p}y_1 + \cdots + \alpha_{pp}y_p \end{aligned} \tag{8}$$

Consequently, if the $p - q$ components with smallest variance are treated as 'noise' and set equal to a 'residual' η_i, then we obtain

$$\begin{aligned} x_1 &= \alpha_{11}y_1 + \cdots + \alpha_{1q}y_q + \eta_1 \\ &\vdots \\ x_p &= \alpha_{1p}y_1 + \cdots + \alpha_{qp}y_q + \eta_p \end{aligned} \tag{9}$$

or equivalently

$$\begin{aligned} x_1 &= (\alpha_{11}\sqrt{\lambda_1})\frac{y_1}{\sqrt{\lambda_1}} + \cdots + (\alpha_{q1}\sqrt{\lambda_q})\frac{y_q}{\sqrt{\lambda_q}} + \eta_1 \\ &\vdots \\ x_p &= (\alpha_{1p}\sqrt{\lambda_1})\frac{y_1}{\sqrt{\lambda_1}} + \cdots + (\alpha_{qp}\sqrt{\lambda_q})\frac{y_q}{\sqrt{\lambda_q}} + \eta_p. \end{aligned} \tag{10}$$

If we write $\gamma_{ij} = (\alpha_{ji}\sqrt{\lambda_j})$ and $z_i = \frac{y_i}{\sqrt{\lambda_i}}$, then we have

$$\begin{aligned} x_1 &= \gamma_{11}z_1 + \cdots + \gamma_{1q}z_q + \eta_1 \\ &\vdots \\ x_p &= \gamma_{p1}z_1 + \cdots + \gamma_{pq}z_q + \eta_p. \end{aligned} \tag{11}$$

Putting back the means μ_i of the x_i into the right-hand side of these equations, we recover a set of equations exactly of the form of the factor analysis model. Moreover, since the y_i are uncorrelated and have variances λ_i, the z_i are also uncorrelated and each has variance 1. Hence the standardized components obey the same assumptions as do the factors in a factor analysis model. For this reason, PCA and FA have been inextricably linked and much confused as techniques over the years. It is therefore important to appreciate precisely what the aims of the two techniques are.

The vital distinction between them is that principal components are the optimal entities for describing or explaining the *variances* in a multivariate system, while factors are appropriate when trying to explain the *covariances* in the system. Note that one important difference between PCA and FA has not yet been highlighted. This is that the e_i of FA are not the same as the η_i of PCA. The e_i are assumed to be uncorrelated with the z_i and with each other. Now consider the η_i of PCA. From their definition we can write

$$\begin{aligned} \eta_1 &= \alpha_{q+1,1}y_{q+1} + \cdots + \alpha_{p1}y_p \\ &\vdots \\ \eta_p &= \alpha_{q+1,p}y_{q+1} + \cdots + \alpha_{pp}y_p. \end{aligned} \tag{12}$$

Since the y_i are all mutually uncorrelated, then the η_i are indeed uncorrelated with the z_i. However, since the same y_j occur in different η_i, the η_i are *not* mutually uncorrelated. Thus the z_i as derived from PCA do not explain all the correlation structure in x. Hence PCA and FA are not the same (!!!).

Appendix

In this appendix, we present a fact found in many multivariate statistics books. Given a joint distribution over several variables that is Normal, we are interested in computing a conditional distribution.

If we partition the set of variables into two subsets, labeled x_1 and x_2, then we can partition the mean vector and covariance matrix in the obvious way:

$$p\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right). \tag{13}$$

Let $x_2 = a$. The conditional distribution of x_1 given that $x_2 = a$ is as follows:

$$p(x_1|x_2 = a) \sim N[\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(a - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}]. \tag{14}$$

In the main body of this note, we used this equation to compute the conditional distributions $p(x|z)$ and $p(z|x)$.