Factor Analysis

Robert L. Wolpert
Department of Statistical Science
Duke University, Durham, NC, USA

1 Factor Models

The multivariate regression model $Y = XB + U$ expresses each row $Y_{i\cdot} \in \mathbb{R}^p$ as a linear combination $X_{i\cdot}B$ of the q columns of X, plus some mean-zero random error $U_{i\cdot}$. Perhaps fewer than q terms would suffice to predict Y adequately or, more realistically, perhaps some number $m < q$ of linear combinations of the columns of X would suffice. That is the essence of Factor Analysis, a method proposed in 1904 by psychologist Charles Spearman in his effort to show that all measures of mental ability (mathematical skill, vocabulary, other verbal skills, artistic skills, logical reasoning ability, etc.) could be explained by a single quantity he called g for "general intelligence" (also the idea behind IQ). The same goal comes up in many application areas— to explain some rather large number q of measurements in terms of a much smaller number m. An extreme example of recent importance is the search for "genes" within DNA sequence data where q might be many thousands, much larger than the number n of subjects— a challenge for the traditional multivariate tools, leading to the development of "data mining" methodology.

The simplest possibility would be for most of the variability of $X_{i\cdot}$ to be explained by a *linear* relation to a small number m of "factors," an idea we now pursue. For a nice introduction from a social scientist's perspective, see Richard Darlington's Cornell notes at `http://www.psych.cornell.edu/Darlington/factor.htm`. An interesting misuse of FA was used to design Duke's abysmal Course Evaluation Forms.

1.1 A Single Vector

Let $X \in \mathbb{R}^p$ be a random vector with mean $\mu \in \mathbb{R}^p$ and covariance $\Sigma \in \mathcal{S}_+^p$ and let $m \in \mathbb{N}$. X is said to follow a "m-factor model" if we can write

$$X = \Lambda f + \mu + u \tag{1}$$

for some constant $(p \times m)$ matrix Λ (the "loading matrix") and random vectors $f \in \mathbb{R}^m$ (the "factors") and $u \in \mathbb{R}^p$. The elements of f are called "common factors", while those of u are (specific or) "unique factors" or "errors." Of course this will get more interesting below when we have more than one vector X.

Without any loss of generality we may insist that:

$$\begin{aligned} \mathsf{E}[f] &= 0 & \mathsf{Cov}[f] &= \mathsf{E} f\, f' = I \\ \mathsf{E}[u] &= 0 & \mathsf{Cov}[u] &= \mathsf{E} u\, u' = \Psi = \mathrm{diag}(\psi_{11}, \ldots, \psi_{pp}) \\ & & \mathsf{Cov}[f, u] &= \mathsf{E} f\, u' = 0 \end{aligned}$$

It follows that the covariance $\Sigma = \mathsf{E}(X - \mu)(X - \mu)'$ may be expressed

$$\Sigma = \Lambda\Lambda' + \Psi \tag{2}$$

as the sum of a common component and a unique component.

Even if a m-factor model holds for X, it is not unique— for any $(m \times m)$ orthogonal matrix G, we can re-write Equation (1) as

$$X = \Lambda G' G f + u + \mu$$

with common factor Gf (also with mean $\mathsf{E}[Gf] = 0$ and covariance $\mathsf{E}[Gff'G'] = I$). For fixed Ψ, this is the only indeterminacy— *i.e.*, once Ψ is fixed then Λ is determined up to an orthogonal rotation. The indeterminacy can be resolved (if desired) by imposing some arbitrary $\frac{1}{2}m \times (m-1)$-dimensional linear constraint, like insisting that $\Lambda'\Psi^{-1}\Lambda$ be diagonal or filling Λ with as many zeros as possible.

Does *every* covariance satisfy Equation (2), or is there anything special about factor models? In general, the dimension of the set of possible covariance matrices $\Sigma \in \mathcal{S}_+^p$ is $p(p+1)/2$ (for example, that's the number of non-zero entries in the Cholesky decomposition of a positive definite matrix), while Σ satisfying Equation (2) is determined by the pm elements of Λ and the p diagonal elements of Ψ. Insisting that the $(m \times m)$ symmetric matrix

$\Lambda'\Psi^{-1}\Lambda$ be diagonal introduces an additional $m(m-1)/2$ constraints, so Σ lies in a set whose dimension is

$$s = \left\{\frac{p(p+1)}{2}\right\} - \left\{pm + p - \frac{m(m-1)}{2}\right\} = \frac{(p-m)^2 - (p+m)}{2} \tag{3}$$

smaller than that of $\mathcal{S}_+^p$ (if, as usual, $s > 0$), making "m-Factor Model" a genuine distinction. Note $s >$ whenever $m < p + \frac{1}{2} - \sqrt{2p + 1/4}$.

1.2 Fitting the Model

If $s < 0$ in Equation (3) above then there are infinitely-many solutions Λ, Ψ to Equation (2), and "the factor model" isn't well-defined. For $s = 0$ there's (typically) a unique solution; for $s > 0$ there will typically be no *exact* solutions, but we can try for an approximate fit.

Let's now suppose we have n iid random vectors $X_{i\cdot}'$ each satisfying Equation (1), assembled as the rows of an $(n \times p)$ matrix X.

Since μ is of little interest here we estimate it by $\hat{\mu} = \bar{x} = \frac{1}{n}X'\mathbf{1}$, and as usual estimate Σ by its MLE (for a normal model)

$$\hat{\Sigma} = \frac{1}{n}(X - \mathbf{1}\bar{x}')'(X - \mathbf{1}\bar{x}'); \tag{4}$$

now the goal is to find m, a $(p \times m)$ matrix $\hat{\Lambda}$, and a diagonal matrix $\hat{\Psi} \in \mathcal{S}_+^m$ that satisfy

$$\hat{\Sigma} \approx \hat{\Lambda}\hat{\Lambda}' + \hat{\Psi}. \tag{5}$$

For any proposed $\hat{\Lambda}$ we can satisfy Equation (5) on the diagonals exactly by setting each

$$\hat{\psi}_{jj} := \hat{\sigma}_{jj} - (\hat{\Lambda}\hat{\Lambda}')_{jj} = \hat{\sigma}_{jj} - \sum_{\ell} \hat{\lambda}_{j\ell}^2 \tag{6}$$

(assuming that's non-negative), so we now turn to finding m and $\hat{\Lambda} \in \mathbb{R}^{p \times m}$. The usual practice is to find the smallest value of m that seems tenable for a given data set, and then to estimate $\hat{\Lambda}$ (and hence $\hat{\Psi}$) for that m.

Since the Factor model of Equation (1) is invariant under changing the location or scale of $X_{i\cdot}$, we now simplify the presentation by assuming that $\bar{x} = 0$ and $\hat{\sigma}_{jj} = \frac{1}{n}\sum_i x_{ij}^2 = 1$ (just replace X with $\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right) X \left(\operatorname{diag}\hat{\Sigma}\right)^{-1/2}$) so now $R = \hat{\Sigma}$ is the sample *correlation* matrix with diagonal entries of $r_{jj} = 1$ and Equation (6) becomes $\hat{\psi}_{jj} = 1 - \sum_\ell \hat{\lambda}_{j\ell}^2$.

Here's one way to estimate Ψ and Λ, called "Principal Factor Analysis." Standardize as above so the rows of X have sample mean zero and sample variance one, and let R be the sample correlation matrix. Pick any point $\left\{\psi_{jj}^0\right\} \in \mathbb{R}_+^m$ as an initial estimate (a common choice is $\psi_{jj}^0 = 1 - \max_{k \neq j} |r_{jk}|$) and construct $\Psi^0 := \text{diag}(\left\{\psi_{jj}^0\right\})$. The matrix $R - \Psi^0$ may not be positive-definite but it's at least symmetric, so it has real eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$, some of which will be positive if $\left\{\psi_{jj}^0\right\} \in \mathbb{R}_+^p$ aren't too big. Let m be the number of positive eigenvalues and let $\hat{\Lambda}$ be the $(p \times m)$ matrix whose m columns are the associated eigenvectors, each scaled by the square root of its eigenvalue. Then $\hat{\Sigma} = R$ will satisfy Equation (5). Now set $\hat{\psi}_{jj} = [R - \hat{\Lambda}\hat{\Lambda}']_{jj}$ to get the diagonal elements exactly. Very efficient algorithms implementing this approach are built into standard statistical software packages— for example, `factanal()` in `R`.

As an alternative approach to estimating $\hat{\Lambda}$ and $\hat{\Psi}$, we may assume a multivariate Gaussian model and maximize the log likelihood over all choices of Λ, Ψ, or can construct joint prior distribution on them and evaluate the posterior mean using MCMC.

2 Example

Psychologists often use Factor Analysis in an effort to "explain" the variation in a relatively high-dimensional (q) data-set by a latent relatively low-dimensional (m) trait, like "intelligence" or "verbal and quantitative abilities". Here (taken from Mardia et al. (1979, Ch. 9), who excerpted it from Spearman (1904, *p.* 275), the paper in which Factor Analysis was introduced) is the sample correlation matrix for test scores of children on $p = 3$ verbal topics (Classics, English, and French):

$$R = \begin{bmatrix} 1 & 0.83 & 0.78 \\ 0.83 & 1 & 0.67 \\ 0.78 & 0.67 & 1 \end{bmatrix}$$

Note $m = p + \frac{1}{2} - \sqrt{2p + 1/4} = 1$ factor will lead to $s = 0$ for $p = 3$ and hence a unique FA model; upon solving for Λ and Ψ we find

$$R = \begin{bmatrix} 0.983 \\ 0.844 \\ 0.794 \end{bmatrix} \begin{bmatrix} 0.983 & 0.844 & 0.794 \end{bmatrix} + \begin{bmatrix} 0.034 & 0 & 0 \\ 0 & 0.287 & 0 \\ 0 & 0 & 0.370 \end{bmatrix}$$

exactly. One (perhaps Spearman) might then argue that abilities $X' = [X_1, X_2, X_3]$ in Classics, English, and French all reflect a single "verbal skill" factor $f \sim \mathsf{No}(0,1)$ through the relation

$$X = \begin{bmatrix} 0.983 \\ 0.844 \\ 0.794 \end{bmatrix} f + \begin{bmatrix} U_1 \sim \mathsf{No}(0, 0.034) \\ U_2 \sim \mathsf{No}(0, 0.287) \\ U_3 \sim \mathsf{No}(0, 0.370) \end{bmatrix}$$

3 Interpretation and Dangers

The underlying idea of FA is that the information in a data set of high dimension p may be summarized adequately by a much smaller number m of quantities linearly-related to the original data. For example, subjects' success in answering a large number p of questions might reflect their differences in a quite small number of variables. Spearman's original hypothesis was that human mental acuity could be explained by *a single* variable, "general intelligence" g; current SAT tests report a three-dimensional score (it used to be just two-dimensional). In a comic misuse of FA the questions on early versions of Duke's Course Evaluation forms were designed by eliminating questions with low loading on the first factor, on the assumption that course "quality" was a one-dimensional quantity (with more reflection one might expect some students to prefer more quantitative courses, others more verbal ones; some might prefer a more intense experience, others a broader and lighter one; some might prefer an entertaining experience while others might prefer a penetrating one). Instead of seeking questions that would help describe these various features of classes to help students make informed choices, all questions were removed from draft versions of the Course Evaluation Forms except those with a high load on the first factor, which the administrator managing the process assumed would be overall quality. Sigh.

Like Principal Components, Factor Analysis is often applied as a preparatory step to reduce the dimension of a problem before some other analysis (usually multivariate regression) is applied. If (as usual) the same data set is used to determine the factors and for the subsequent regression, then this approach introduces a bias that will lead to confidence intervals that are a bit too short, and that will cover true parameter values with less than the nominal probability (say, 95%). Probably the best alternative is Bayesian Model Averaging (BMA), which has its own advantages and disadvantages.

References

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979), *Multivariate Analysis*, New York, NY: Academic Press.

Spearman, C. (1904), "'General Intelligence,' objectively determined and measured," *American Journal of Psychology*, 15, 201–292.