Factor analysis

Angela Montanari

1 Introduction

Factor analysis is a statistical model that allows to explain the correlations between a large number of observed correlated variables through a small number of uncorrelated unobservable ones: the factors.

The origin of factor analysis dates back to a work by Spearman in 1904. At that time psychometricians were deeply involved in the attempt to suitably quantify human intelligence, and Spearman's work provided a very clever and useful tool that is still at the bases of the most advanced instruments for measuring intelligence.

Intelligence is the prototype of a vast class of variables that are not directly observed (they are called latent variables) but can be measured in an indirect way through the analysis of observable variables closely linked to the latent ones. Latent variables are common to many research fields besides psychology, from medicine to genetics, form finance to economics and this explain the still vivid interest towards Factor analysis.

Spearman considered the following correlation matrix between children's examination performance in Classics (x_1), French (x_2) and English (x_3):

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

He noticed a high positive correlation between the scores and hypothesized that it was due to the correlation of the three observed variables with a further unobserved variable that he called intelligence or general ability. If his assumption was true, than he expected that the partial correlation coefficients, computed between the observed variables after controlling for the common latent one, would vanish.

Starting from this intuition he formulated the following model, which, as we will see in the following, can perfectly fulfill the goal:

$$x_1 = \lambda_1 f + u_1 \qquad x_2 = \lambda_2 f + u_2 \qquad x_3 = \lambda_3 f + u_3$$

where f is the underlying *common factor*, $\lambda_1, \lambda_2, \lambda_3$ are the *factor loadings* and u_1, u_2, u_3 are the *unique or specific factors*.

The factor loadings indicate how much the common factor contributes to the different observed values of the x variables; the unique factors represent residuals, random noise terms that besides telling that an examination only offers an approximate measure of the subject's ability, also describe, for each individual, how much his result on a given subject, say French, differs from his general ability. Spearman's model can be generalized to include more than one common factor:

$$x_i = \lambda_{i1} f_1 + \lambda_{i2} f_2 + \cdots + \lambda_{ik} f_k + \cdots + \lambda_{im} f_m + u_i$$

2 The factor model

Let $\mathbf{x}$ be a p-dimensional random vector with expected value μ and covariance matrix $\mathbf{\Sigma}$. An m factor model for $\mathbf{x}$ holds if it can be decomposed as:

$$\mathbf{x} = \mathbf{\Lambda f} + \mathbf{u} + \mu$$

If we assume to deal with mean centered $\mathbf{x}$ variables then, with no loss in generality, the model will be

$$\mathbf{x} = \mathbf{\Lambda f} + \mathbf{u} \tag{1}$$

where $\mathbf{\Lambda}$ is the $p \times m$ factor loading matrix; $\mathbf{f}$ is the $m \times 1$ random vector of common factors and $\mathbf{u}$ is the $p \times 1$ random vector of unique factors.

The model looks like a linear regression model, but in this case all the elements in right hand side of the equal sign are unknown. In order to reduce indeterminacy we can impose the following constraints:

- $$E(\mathbf{f}) = \mathbf{0} \qquad E(\mathbf{u}) = \mathbf{0}$$

  this condition is perfectly coherent with the fact that we work with mean centered data

- $$E(\mathbf{f}\mathbf{f}^T) = \mathbf{I}$$

  i.e. the common factors are standardized uncorrelated random variables: their variances are equal to 1 and their covariances are 0. This assumption could also be relaxed, while the following ones are strictly required.

- $$E(\mathbf{u}\mathbf{u}^T) = \mathbf{\Psi}$$

  where

  $$\mathbf{\Psi} = \begin{bmatrix} \psi_{11} & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \psi_{kk} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & \psi_{pp} \end{bmatrix}$$

  is a diagonal matrix. This means that the unique factors are uncorrelated and may be heteroscedastic.

- $$E(\mathbf{f}\mathbf{u}^T) = 0 \qquad E(\mathbf{u}\mathbf{f}^T) = 0$$

  that means that the unique factors are uncorrelated with the common ones.

If a factor model satisfying the above mentioned conditions holds, the covariance matrix of the observed variables $\mathbf{x}$ can be decomposed as follows:

$$\begin{aligned} \mathbf{\Sigma} &= E(\mathbf{x}\mathbf{x}^T) = E[(\mathbf{\Lambda}\mathbf{f} + \mathbf{u})(\mathbf{\Lambda}\mathbf{f} + \mathbf{u})^T] \\ &= E(\mathbf{\Lambda}\mathbf{f}\mathbf{f}^T\mathbf{\Lambda}^T + \mathbf{\Lambda}\mathbf{f}\mathbf{u}^T + \mathbf{u}\mathbf{f}^T\mathbf{\Lambda}^T + \mathbf{u}\mathbf{u}^T] \\ &= \mathbf{\Lambda}E(\mathbf{f}\mathbf{f}^T)\mathbf{\Lambda}^T + \mathbf{\Lambda}E(\mathbf{f}\mathbf{u}^T) + E(\mathbf{u}\mathbf{f}^T)\mathbf{\Lambda}^T + E(\mathbf{u}\mathbf{u}^T) \qquad (2) \\ &= \mathbf{\Lambda}\mathbf{I}\mathbf{\Lambda}^T + \mathbf{\Lambda}\mathbf{0} + \mathbf{0}\mathbf{\Lambda}^T + \mathbf{\Psi} \\ &= \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi} \end{aligned}$$

The opposite is also true: if the covariance matrix of the observed variables $\mathbf{x}$ can be decomposed as in equation (2) then the linear factor model (1) holds.

As $\mathbf{\Psi}$ is diagonal, decomposition (2) clearly shows that the common factors account for all the observed covariances and provides theoretical motivation to Spearman's intuition.

It is worth having a closer look at the diagonal elements of the matrices on both sides of the equality sign in decomposition (2).

$$Var(x_i) = \sigma_{ii} = \sum_{k=1}^{m} \lambda_{ik}^2 + \psi_{ii} = h_i^2 + \psi_{ii}$$

The quantity $\sum_{k=1}^{m} \lambda_{ik}^2 = h_i^2$ is called called the *communality*; it represents the part of the variance of x_i that is explained by the common factors or, in other words, it is the variance of x_i that is shared with the other variables via the common factors. ψ_{ii} is called the *unique variance*: it is the variance of the i-th unique factor and represents the part of the variance of x_i not accounted for by the common factors.

From the assumptions on the common and the unique factors that have been previously described a further interesting characterization for $\mathbf{\Lambda}$ derives. Let's consider the covariance between the observed variables $\mathbf{x}$ and the common factors $\mathbf{f}$:

$$Cov(\mathbf{x}, \mathbf{f}) = E(\mathbf{x}\mathbf{f}^T) = E[(\mathbf{\Lambda}\mathbf{f} + \mathbf{u})\mathbf{f}^T] = \mathbf{\Lambda}E(\mathbf{f}\mathbf{f}^T) + E(\mathbf{u}\mathbf{f}^T) = \mathbf{\Lambda}$$

This means that the factor loading matrix $\mathbf{\Lambda}$ is also the covariance matrix between $\mathbf{x}$ and $\mathbf{f}$.

Scale equivariance

The factor model (1) is scale equivariant.

Let's consider the re-scaled $\mathbf{x}$ variables obtained as $\mathbf{y} = \mathbf{C}\mathbf{x}$ where $\mathbf{C}$ is a diagonal matrix. Let's also assume that the linear factor model holds for $\mathbf{x}$ and put $\mathbf{\Lambda} = \mathbf{\Lambda}_x$ and $\mathbf{\Psi} = \mathbf{\Psi}_x$. The factor model for the new $\mathbf{y}$ variables becomes:

$$\mathbf{y} = \mathbf{C}\mathbf{x} = \mathbf{C}(\mathbf{\Lambda}_x\mathbf{f} + \mathbf{u}) = \mathbf{C}\mathbf{\Lambda}_x\mathbf{f} + \mathbf{C}\mathbf{u} = \mathbf{\Lambda}_y\mathbf{f} + \mathbf{C}\mathbf{u}$$

Thus the factor model also holds for $\mathbf{y}$. The new factor loading matrix is but a re-scaled version of the old one $\mathbf{\Lambda}_y = \mathbf{C}\mathbf{\Lambda}_x$ and the unique variance is $\mathbf{\Psi}_y = E(\mathbf{C}\mathbf{u}\mathbf{u}^T\mathbf{C}) = \mathbf{C}\mathbf{\Psi}_x\mathbf{C}$.

A common re-scaling transformation is standardization, that amounts to choose $\mathbf{C} = \mathbf{\Delta}^{-1/2}$ where, as usual, $\mathbf{\Delta}$ is the diagonal matrix with the variances of the $\mathbf{x}$ variables. The scale equivariance property implies that we can define the model either on the raw or on the standardized data and it is always possible to go from one solution to the other by applying the same

scale change to the model parameters. It is worth reminding that, on the contrary, PCs don't share this property.

Identifiability

A statistical model is said to be identifiable if different values of the parameters generate different predicted values. In other words a statistical model is identifiable if a solution exists to the estimation problem and it is unique. Our linear factor model is not identifiable as there is an infinite number of different matrices $\mathbf{\Lambda}$ that can generate the same $\mathbf{x}$ values.
In fact, if the m-factor model holds, then it also holds if the factors are rotated. Denote by $\mathbf{G}$ an $m \times m$ orthogonal matrix such that $\mathbf{G}\mathbf{G}^T = \mathbf{G}^T\mathbf{G} = \mathbf{I}$. The factor model can equivalently be re-written as

$$\mathbf{x} = \mathbf{\Lambda}\mathbf{f} + \mathbf{u} = \mathbf{\Lambda}\mathbf{I}\mathbf{f} + \mathbf{u} = \mathbf{\Lambda}\mathbf{G}\mathbf{G}^T\mathbf{f} + \mathbf{u}$$

If we set $\mathbf{\Lambda}^* = \mathbf{\Lambda}\mathbf{G}$ and $\mathbf{f}^* = \mathbf{G}^T\mathbf{f}$, the factor model becomes $\mathbf{x} = \mathbf{\Lambda}^*\mathbf{f}^* + \mathbf{u}$. The two factor models are completely equivalent and indistinguishable. They have the same properties:

$$E(\mathbf{f}^*) = E(\mathbf{G}^T\mathbf{f}) = \mathbf{G}^T E(\mathbf{f}) = 0$$
$$E(\mathbf{f}^*\mathbf{f}^{*T}) = E(\mathbf{G}^T\mathbf{f}\mathbf{f}^T\mathbf{G}) = \mathbf{G}^T E(\mathbf{f}\mathbf{f}^T)\mathbf{G} = \mathbf{G}^T\mathbf{I}\mathbf{G} = \mathbf{G}^T\mathbf{G} = \mathbf{I}$$
$$E(\mathbf{f}^*\mathbf{u}^T) = E(\mathbf{G}^T\mathbf{f}\mathbf{u}^T) = \mathbf{G}^T E(\mathbf{f}\mathbf{u}^T) = \mathbf{0}$$

In order to obtain a unique solution constraints on $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ are usually imposed. The most common ones are that either $\mathbf{\Lambda}^T\mathbf{\Psi}^{-1}\mathbf{\Lambda}$ or $\mathbf{\Lambda}^T\mathbf{\Delta}^{-1}\mathbf{\Lambda}$ is diagonal.
The property of rotation invariance will turn out to be really useful for interpretation purposes.

The issue of the existence of the decomposition in equation (2) needs now to be addressed.
The number of available informations is given by the distinct elements of $\mathbf{\Sigma}$, i.e. the variances and the covariances whose number is equal to the sum of the first p natural numbers $p(p+1)/2$. The number of unknown is given by the pm elements of $\mathbf{\Lambda}$ plus the p diagonal elements of $\mathbf{\Psi}$. The above mentioned constraint reduces the number of unknowns of the quantity $m(m-1)/2$ i.e. the number of off-diagonal elements that are constrained to be equal to zero. The number of equations is therefore $p(p+1)/2$ and the number of unknowns is $pm + p - m(m-1)/2$. If the first number is lower than the second we

have an infinite number of solutions. If equality holds then we have a unique exact solution that however is useless from a statistical point of view since it doesn't allow to explain the observed correlations through a small number of latent variables. If $p(p+1)/2 > pm+p-m(m-1)/2$ then a solution, however approximate, exists. The previous inequality, known as Lederman condition, imposes an upper bound on the maximum number of common factors.

3 Estimation issues

In practice we do not know $\mathbf{\Sigma}$, but can estimate it by the sample covariance matrix $\mathbf{S}$ (or by the sample correlation matrix $\mathbf{R}$ if we want to deal with standardized data). We seek therefore the estimates $\hat{\mathbf{\Lambda}}$ and $\hat{\mathbf{\Psi}}$ such that $\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T + \hat{\mathbf{\Psi}}$ is as close to $\mathbf{S}$ as possible.

Various estimation methods have been proposed over the years. Most of them lacked in theoretical bases, and that's one reason why factor analysis didn't have much success among the statical community and remained confined to the psychological literature. Only around 1960 Joreskog succeeded in deriving a sound and reliable algorithm for maximum likelihood estimation. This greatly helped the diffusion of factor analysis that, to day, is one of the most widely used dimension reduction models.
We will briefly sketch one of the older heuristic methods, the Principal factor method, that is still often used and describe the best known Maximum likelihood method.

Principal factor analysis
This method is completely distribution free. It is derived from the knowledge of $\mathbf{S}$ or $\mathbf{R}$ only.
As a first step, preliminary estimates $\widetilde{h}_i^2$ of the communalities h_i^2, $i = 1, \ldots, p$ are obtained.

When $\mathbf{S}$ is the starting point, most softwares use $s_{ii}R_{i0}^2$ or $s_{ii}\max_{i'}|r_{ii'}|$; when working on $\mathbf{R}$ the corresponding expressions are simply R_{i0}^2 or $\max_{i'}|r_{ii'}|$. The squared multiple correlation coefficient R_{i0}^2 is referred to a multiple regression model where the variable x_i is considered as the dependent variable and the other x variables are the covariates. It measures the portion of the variance of x_i that is explained by its linear relationship with the other vari-

ables. As the communality is the part of the variance of x_i that is shared with the other variables via the common factors, R^2_{i0} can provide a reasonable approximation.
It can be easily proved that R^2_{i0} can be computed from the correlation matrix $\mathbf{R}$ as $1-1/r^*_{ii}$ where r^*_{ii} is the i-th diagonal element of $\mathbf{R}^{-1}$. This implies that $\mathbf{R}$ is non singular. In case this condition is not satisfied the method based on $\max_{i'} |r_{ii'}|$ can be used.

Let's then define the *reduced covariance matrix* as $\mathbf{S} - \hat{\mathbf{\Psi}}$ (the *reduced correlation matrix* would be $\mathbf{R}-\hat{\mathbf{\Psi}}$) ; it is simply $\mathbf{S}$ (or $\mathbf{R}$) where the diagonal elements have been replaced by $\widetilde{h}^2_i$. By the spectral decomposition theorem, $\mathbf{S} - \hat{\mathbf{\Psi}}$ can be decomposed as

$$\mathbf{S} - \hat{\mathbf{\Psi}} = \mathbf{\Gamma L \Gamma}^T$$

where $\mathbf{\Gamma}$ is the orthonormal matrix whose columns are the eigenvectors of $\mathbf{S}$ and $\mathbf{L}$ is the diagonal matrix of the eigenvalues.

$\mathbf{S} - \hat{\mathbf{\Psi}}$ is no longer positive semi definite. Suppose that the first m eigenvalues are positive. A rank m approximation of $\mathbf{S} - \hat{\mathbf{\Psi}}$ can thus be obtained as

$$\mathbf{S} - \hat{\mathbf{\Psi}} = \mathbf{\Gamma}_m \mathbf{L}_m \mathbf{\Gamma}^T_m \tag{3}$$

where the columns of $\mathbf{\Gamma}_m$ are the m eigenvectors of the reduced covariance matrix corresponding to the m positive eigenvalues, that appear on the diagonal of $\mathbf{L}_m$. Equation (3) can be rewritten as

$$\mathbf{S} - \hat{\mathbf{\Psi}} = \mathbf{\Gamma}_m \mathbf{L}_m \mathbf{\Gamma}^T_m = \mathbf{\Gamma}_m \mathbf{L}^{1/2}_m \mathbf{L}^{1/2}_m \mathbf{\Gamma}^T_m = \hat{\mathbf{\Lambda}} \hat{\mathbf{\Lambda}}^T$$

where $\hat{\mathbf{\Lambda}} = \mathbf{\Gamma}_m \mathbf{L}^{1/2}_m$.

The diagonal elements of $\hat{\mathbf{\Lambda}} \hat{\mathbf{\Lambda}}^T$ provide new estimates of the communalities. The estimation procedure can be stopped here or iterated, by replacing the new communality estimates on the diagonal of the reduced covariance matrix, until convergence.
The sum of squares of the i-th row of $\hat{\mathbf{\Lambda}}$ is the communality of the i-th observed variable; the sum of squares of the k-th column of $\hat{\mathbf{\Lambda}}$ is the variance explained by the k-th common factor. If we devide it by the total variance we obtain an indication of the proportion of the total variance explained by the k-th factor. It is worth noticing that, being based on the spectral decomposition, the principal factor solution is not scale equivariant (we have already seen the same holds for PCA). This means that principal factoring

produces different results if we work on $\mathbf{S}$ or on $\mathbf{R}$.
Most statistical softwares perform Factor analysis on $\mathbf{R}$ as a default option. Their output also provides the *residual correlation matrix*, a heuristic tool aimed at assessing the goodness of fit of the m factor solution. It is computed as $\mathbf{R} - \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T$. It's diagonal elements are the estimates of the unique variances while the off- diagonal entries are the residual correlations i.e. the correlation that the common factors have not been able to account for. Values close to 0 indicate a good fit, values far from 0 indicate lack of fit. This might mean that more factors are need in order to better explain the observed correlation, but it also might mean that the relationship between observed variables and latent ones is not linear, as the model assumes.

When working on $\mathbf{R}$ many softwares allow a third option for preliminary estimation of the communalities; it consists in setting all the communalities equal 1. This means that the unique variances are equal to 0. In this case the residual correlation matrix becomes $\mathbf{R} - \hat{\mathbf{\Psi}} = \mathbf{R}$ and the principal factors coincide with the Principal Components. This algebraic equivalence has led to a dangerous confusion between factor analysis and PCA. It must be remembered that PCA is a data transformation while Factor analysis assumes a model. Further more PCA aims at explaining variances while the main interest of Factor Analysis lies on covariances. We will deal with this aspect in more detail in the end of the chapter.

Maximum likelihood factor analysis

The approach based on maximum likelihood for the estimation of the model parameters assumes that the vector $\mathbf{x}$ is distributed according to a multivariate normal distribution, that is:

$$f(\mathbf{x}) = |2\pi\mathbf{\Sigma}|^{-1/2} \exp\{-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\}$$

The likelihood for a sample of n units, is then

$$L(\mathbf{x}, \mu, \mathbf{\Sigma}) = |2\pi\mathbf{\Sigma}|^{-n/2} \exp\{-\frac{1}{2}\sum_{j=1}^{n}(\mathbf{x}_j-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}_j-\mu)\}$$

and the log-likelihood

$$l(\mathbf{x}, \mu, \mathbf{\Sigma}) = \ln L(\mathbf{x}, \mu, \mathbf{\Sigma}) = -\frac{n}{2}\ln|2\pi\mathbf{\Sigma}| - \frac{1}{2}\sum_{j=1}^{n}(\mathbf{x}_j-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}_j-\mu)$$

After estimating μ by $\bar{\bar{\mathbf{x}}}$ and after some algebra it becomes

$$l(\mathbf{x}, \mathbf{\Sigma}) = -\frac{n}{2} \ln |2\pi\mathbf{\Sigma}| - \frac{n}{2} tr\mathbf{\Sigma}^{-1}\mathbf{S}$$

If the factor model holds then $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}$ and the log-likelihood becomes:

$$l(\mathbf{x}, \mathbf{\Sigma}) = l(\mathbf{x}, \mathbf{\Lambda}, \mathbf{\Psi}) = -\frac{n}{2} \ln |2\pi(\mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi})| - \frac{n}{2} tr(\mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi})^{-1}\mathbf{S}$$

This function must be maximized with respect to $\mathbf{\Lambda}$ and $\mathbf{\Psi}$. This problem is not trivial and requires the recourse to numerical methods. In 1967 Joreskog developed an efficient two-stage algorithm. Convergence is quite fast, but it might happen that one or more elements of the estimate of $\mathbf{\Psi}$ become negative. This situation, known as Heywood case, can be solved by constraining the elements of $\mathbf{\Psi}$ to be non negative.
As the maximum likelihood estimates are scale equivariant the estimates referred to standardized data can be obtained from the ones obtained on the raw data by a simple change in the scale. The same holds for the opposite transformation.

Fitting the common factor model by maximum likelihood also allows to select, through a suitable likelihood ratio test, the number of common factors to be included in the model. The null hypothesis is $H_0 : \mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}$ with $\mathbf{\Lambda}$ a $p \times m$ matrix against the alternative $H_1 : \mathbf{\Sigma}$ *is unstructured*.

The test statistic, after applying Bartlett's correction in order to improve its convergence to a χ^2 distribution, is:

$$W = \{n - \frac{2p + 11}{6} - \frac{2m}{3}\}\{\ln |\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T + \hat{\mathbf{\Psi}}| - \ln |\mathbf{S}|\}$$

where $\hat{\mathbf{\Lambda}}$ and $\hat{\mathbf{\Psi}}$ are the maximum likelihood estimates of the corresponding model parameters.

Under H_0, W is distributed according to a χ^2 with $\frac{1}{2}[(p-m)^2 - (p+m)]$ degrees of freedom.

Examples on the use of FA and on the interpretation of the results will be provided in the notes for the lab session.

4 Rotation of factors

In order to improve interpretability of the factor loadings we can relay on the invariance to orthogonal rotation property of the factor model. In 1947

Thurston gave a definition of how an interpretable (simple) factor structure should be. The variables should be divisible into groups such that the loadings within each group are high on a single factor, perhaps moderate to low on a few factors and negligible on the remaining factors. One way to obtain a factor loading matrix satisfying such a condition is given by the so-called Varimax rotation. It looks for an orthogonal rotation of the factor loading matrix, such that the following criterion is maximized

$$V = \sum_{k=1}^{m} \left\{ \frac{\sum_{i=1}^{p} \beta_{ik}^4}{p} - \left(\frac{\sum_{i=1}^{p} \beta_{ik}^2}{p} \right)^2 \right\}$$

where

$$\beta_{ik} = \frac{\lambda_{ik}}{(\sum_{k=1}^{m} \lambda_{ik}^2)^{1/2}} = \frac{\lambda_{ik}}{h_i}$$

It should be noted that V is the sum of the variances of the squared normalized (within each row) factor scores for each factor. Maximizing it causes the large coefficients to become larger and the small coefficients to approach 0. Other rotation methods are implemented in most statistical softwares. Quartimax rotation maximizes the overall variance of the factor loading matrix, thus usually producing, as the first factor, a general factor with positive and almost equal loadings on all the variables. Rotations leading to correlated common factors are also possible.

5 Estimation of factor scores

After the factor loadings and the unique variances have been estimated, we might be interested in estimating, for each statistical unit whose observed vector is $\mathbf{x}_j$, the corresponding vector of factor scores $\mathbf{f}_j$. If for instance the first factor is intelligence, this could also allow us to rank the individuals according to their scores on this factor, from the most to the least intelligent ones. Two methods exist in popular use for factor score estimation.

The method proposed by Thompson defines the factor scores as linear combinations of the observed variables chosen so as to minimize the squared expected prediction error. For the k-th factor f_k, the corresponding estimate is given by $f_k^* = \mathbf{a}_k^T \mathbf{x} = \mathbf{x}^T \mathbf{a}_k$ where $\mathbf{a}_k$ is a $p \times 1$ vector. According to Thompson's approach $\mathbf{a}_k$ should be chosen so that $E(f_k^* - f_k)^2 = E(\mathbf{x}^T \mathbf{a}_k -$

$f_k)^2$ is minimized. After differentiating with respect to $\mathbf{a}_k$ and setting the derivatives equal to 0 we obtain

$$E[2\mathbf{x}(\mathbf{x}^T\mathbf{a}_k - f_k)] = 2[E(\mathbf{x}\mathbf{x}^T)\mathbf{a}_k - E(\mathbf{x}f_k)] = 2(\mathbf{\Sigma}\mathbf{a}_k - \mathbf{\Lambda}_k) = 0$$

where $\mathbf{\Lambda}_k$ is the k-th column of $\mathbf{\Lambda}$.

Hence, $\mathbf{a}_k = \mathbf{\Sigma}^{-1}\mathbf{\Lambda}_k$ and $f_k^* = \mathbf{\Lambda}_k^T\mathbf{\Sigma}^{-1}\mathbf{x}$. It will then be $\mathbf{f}^* = \mathbf{\Lambda}^T\mathbf{\Sigma}^{-1}\mathbf{x}$. After some algebra a different expression for $\mathbf{f}^*$ can be obtained. It is $\mathbf{f}^* = (\mathbf{I} + \mathbf{\Lambda}^T\mathbf{\Psi}^{-1}\mathbf{\Lambda})^{-1}\mathbf{\Lambda}^T\mathbf{\Psi}^{-1}\mathbf{x}$. The two formulations give the same result, but some text books present the second one only. Thompson's estimator is biased. In fact

$$E(\mathbf{f}^*|\mathbf{f}) = E(\mathbf{\Lambda}^T\mathbf{\Sigma}^{-1}\mathbf{x}|\mathbf{f}) = \mathbf{\Lambda}^T\mathbf{\Sigma}^{-1}E(\mathbf{x}|\mathbf{f}) = \mathbf{\Lambda}^T\mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{f} \neq \mathbf{f}$$

It has however been obtained by minimizing $E(f_k^* - f_k)^2$ and so it is the estimator having the least mean squared prediction error.
It is worth mentioning that Thompson's estimator is also known in the literature as "regression estimator". If we assume that the observed variables $\mathbf{x}$, the common factors $\mathbf{f}$ and the unique factors $\mathbf{u}$ are normally distributed and consider the vector $\mathbf{y}^T = (\mathbf{f}^T, \mathbf{x}^T)$ formed by compounding the common factors and the observed variables, under the factor model assumptions, we obtain that $\mathbf{y}$ is also distributed according to a 0 mean multivariate normal distribution with covariance matrix given by

$$\begin{bmatrix} \mathbf{I} & \mathbf{\Lambda}^T \\ \mathbf{\Lambda} & \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi} \end{bmatrix}$$

From the properties of the multivariate normal distribution, the expected value of $\mathbf{f}$ given $\mathbf{x}$ is

$$E(\mathbf{f}|\mathbf{x}) = \mathbf{\Lambda}^T(\mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi})^{-1}\mathbf{x} = \mathbf{\Lambda}^T\mathbf{\Sigma}^{-1}\mathbf{x}$$

that coincides with the previously obtained $\mathbf{f}^*$.

An alternative estimator is due to Bartlett.
After the factor loadings and the unique variances have been estimated, the factor model can be seen as a linear multivariate regression model where $\mathbf{f}$ is the unknown vector parameter and the residuals (i.e. the unique factors) are uncorrelated but heteroscedastic. Estimation can be addressed by wighted least squares.

We want to find an estimate of $\mathbf{f}$, say $\hat{\mathbf{f}}$, such that

$$\mathbf{u}^T\boldsymbol{\Psi}^{-1}\mathbf{u} = (\mathbf{x}-\boldsymbol{\Lambda}\mathbf{f})^T\boldsymbol{\Psi}^{-1}(\mathbf{x}-\boldsymbol{\Lambda}\mathbf{f})$$

is minimum. After differentiating with respect to $\mathbf{f}$ and setting the derivatives equal to 0 we obtain

$$-2\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}(\mathbf{x}-\boldsymbol{\Lambda}\mathbf{f}) = 2(\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\mathbf{f} - \boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}\mathbf{x}) = 0$$

and hence

$$\hat{\mathbf{f}} = (\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}\mathbf{x}$$

Bartlett's estimator $\hat{\mathbf{f}}$ is unbiased. In fact

$$\begin{aligned} E(\hat{\mathbf{f}}|\mathbf{f}) &= E\{(\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}\mathbf{x}|\mathbf{f}\} = \\ &= (\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}E(\mathbf{x}|\mathbf{f}) = \\ &= (\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\mathbf{f} = \mathbf{f} \end{aligned}$$

Of course it has larger mean squared prediction error than Thompson's estimator. The choice between the two is dependent on the research goals.

6 Factor analysis and PCA

It is worth concluding this chapter by stressing the connections and the differences between Factor Analysis and PCA.
Both methods have the aim of reducing the dimensionality of a vector of random variables. But while Factor Analysis assumes a model (that may fit the data or not), PCA is just a data transformation and for this reason it always exists.
Furthermore while Factor Analysis aims at explaining (covariances) or correlations, PCA concentrates on variances.

Despite these conceptual differences, there have been attempts, mainly in the past, to use PCA in order to estimate the factor model. In the following we will show that indeed PCA may be inadequate when the goal of the research is fitting a factor model.

Let $\mathbf{x}$ be the usual p dimensional random vector and $\mathbf{y}$ the p dimensional vector of the corresponding principal components $\mathbf{y} = \mathbf{A}^T\mathbf{x}$ with $\mathbf{A}$ the

orthonormal matrix whose columns are the eigenvectors of the covariance matrix of the $\mathbf{x}$ variables.

Because of the properties of $\mathbf{A}$ it will also be $\mathbf{x} = \mathbf{A}\mathbf{y}$. $\mathbf{A}$ can be decomposed into two sub matrices $\mathbf{A}_m$ containing the eigenvectors corresponding to the first m eigenvalues and $\mathbf{A}_{p-m}$ containing the remaining ones $\mathbf{A} = (\mathbf{A}_m|\mathbf{A}_{p-m})$. A similar partition is considered for the vector $\mathbf{y}$, $\mathbf{y} = \begin{pmatrix} \mathbf{y}_m \\ - \\ \mathbf{y}_{p-m} \end{pmatrix}$ hence

$$\begin{aligned} \mathbf{x} &= (\mathbf{A}_m|\mathbf{A}_{p-m}) \begin{pmatrix} \mathbf{y}_m \\ - \\ \mathbf{y}_{p-m} \end{pmatrix} = \mathbf{A}_m\mathbf{y}_m + \mathbf{A}_{p-m}\mathbf{y}_{p-m} \\ &= \mathbf{A}_m\mathbf{L}_m^{1/2}\mathbf{L}_m^{-1/2}\mathbf{y}_m + \mathbf{A}_{p-m}\mathbf{y}_{p-m} \end{aligned} \tag{4}$$

where $\mathbf{L}_m$ is the diagonal matrix of the first m eigenvalues.
Setting $\mathbf{A}_m\mathbf{L}_m^{1/2} = \mathbf{\Lambda}$, $\mathbf{L}_m^{-1/2}\mathbf{y}_m = \mathbf{f}$ and $\mathbf{A}_{p-m}\mathbf{y}_{p-m} = \eta$ equation (4) can be rewritten as $\mathbf{x} = \mathbf{\Lambda}\mathbf{f} + \eta$.
It can be easily checked that the new common factors $\mathbf{f}$ have the properties required by the linear factor model

$$E(\mathbf{f}\mathbf{f}^T) = E(\mathbf{L}_m^{-1/2}\mathbf{y}_m\mathbf{y}_m^T\mathbf{L}_m^{-1/2}) = \mathbf{L}_m^{-1/2}E(\mathbf{y}_m\mathbf{y}_m^T)\mathbf{L}_m^{-1/2} = \mathbf{L}_m^{-1/2}\mathbf{L}_m\mathbf{L}_m^{-1/2} = \mathbf{I}$$

because the covariance matrix of the first m PCS is $\mathbf{L}_m$ and

$$E(\mathbf{f}\eta^T) = E(\mathbf{L}_m^{-1/2}\mathbf{y}_m\mathbf{y}_{p-m}^T\mathbf{A}_{p-m}^T) = \mathbf{L}_m^{-1/2}E(\mathbf{y}_m\mathbf{y}_{p-m}^T)\mathbf{A}_{p-m}^T = 0$$

because the first m and the last $p-m$ PCs are uncorrelated.
The new unique factors η, however, are not uncorrelated and this contradicts the linear factor model assumption according to which the common factors completely explain the observed covariances:

$$E(\eta\eta^T) = E(\mathbf{A}_{p-m}\mathbf{y}_{p-m}\mathbf{y}_{p-m}^T\mathbf{A}_{p-m}^T) = \mathbf{A}_{p-m}E(\mathbf{y}_{p-m}\mathbf{y}_{p-m}^T)\mathbf{A}_{p-m}^T = \mathbf{A}_{p-m}\mathbf{L}_{p-m}\mathbf{A}_{p-m}^T$$

$\mathbf{L}_{p-m}$ is the covariance matrix of the last $p-m$ PCs and therefore it is diagonal; but, in general, its diagonal elements are different and therefore $\mathbf{A}_{p-m}\mathbf{L}_{p-m}\mathbf{A}_{p-m}^T$ is not diagonal.