MSc II: Further Statistical Analysis (Autumn Term)
Advanced Multivariate Analysis

5 Factor Analysis

Roughly speaking, the aim of Factor Analysis, as with Principal Components Analysis, is to replace the original variables by a smaller number of 'underlying' variables. However, there are important differences between the two techniques. For example,

1. Principal Components Analysis (PCA) can be regarded as a technique for accounting maximally for the <u>variances</u> of the responses, $\mathbf{x}_i$ $(i = 1, ..., p)$. On the other hand, Factor Analysis (FA) is a method that accounts maximally for the <u>covariances</u> of the responses.

2. PCA is just a linear transformation (translation and orthogonal rotation) which, for a particular scaling of the variables, gives the best k-dimensional representation of the data in terms of the squared Euclidean distance (or equivalently in terms of variances). FA is based on a statistical model for the data.

Factor analysis was originally developed by psychologists and the starting point was in the field of education, with a paper by Spearman concerned with data on exam marks. Marks on six exams for a sample of children were obtained. Looking at the corresponding covariance matrix, Spearman noticed that in each row, the values decreased from left to right, roughly proportionally. i.e given

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{16} \\ s_{21} & s_{22} & \cdots & s_{26} \\ \vdots & & & \vdots \\ s_{61} & s_{62} & \cdots & s_{66} \end{pmatrix} = (s_{ij})$$

$\dfrac{s_{ik}}{s_{jk}}$ appeared to be independent of k, for all i, j.

Spearman proposed the following model to explain this.

Suppose $x_i = \lambda_i \mathrm{f} + \mathrm{e}_i$, $i = 1, \ldots, 6$, where

f is a random variable <u>common</u> to all $x_i, i = 1, \ldots, 6$

e_i is a random variable <u>specific</u> to $x_i, i = 1, \ldots, 6$

Assume further that f, $\mathrm{e}_1, \ldots, e_6$ are all uncorrelated, so that the covariances between them are zero. Then,

$$\begin{aligned}\mathrm{Cov}(x_i, x_j) = \sigma_{ij} &= \mathrm{Cov}(\lambda_i \mathrm{f} + \mathrm{e}_i, \lambda_j \mathrm{f} + \mathrm{e}_j) \\ &= \lambda_i \lambda_j \mathrm{Var}(\mathrm{f})\end{aligned}$$

so that, $\dfrac{\sigma_{ik}}{\sigma_{jk}} = \dfrac{\lambda_i}{\lambda_j}$ is independent of k.

This would account for the proportionality. Note, that this model replaces the 15 covariances in $\Sigma = (\sigma_{ij})$ by the 6 parameters $\lambda_i, i = 1, \ldots, 6$.

f is regarded as an underlying (unmeasurable directly) <u>common</u> factor.

$\lambda_1, \ldots, \lambda_6$ are called <u>factor loadings</u>

$\mathrm{e}_1, \ldots, \mathrm{e}_6$ are called the <u>specific</u> or <u>unique</u> factors (e.g. specific to the particular examination subject, say mathematics or classics).

5.1 The general k-factor model

Let $\mathbf{x} = (x_1, \ldots, x_p)^T$ be a random vector with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. Consider the model:

$$\begin{aligned} x_1 &= \lambda_{11}\mathrm{f}_1 + \ldots + \lambda_{1k}\mathrm{f}_k + \mathrm{e}_1 + \mu_1 \\ x_2 &= \lambda_{21}\mathrm{f}_1 + \ldots + \lambda_{2k}\mathrm{f}_k + \mathrm{e}_2 + \mu_2 \\ &\vdots \\ x_p &= \lambda_{p1}\mathrm{f}_1 + \ldots + \lambda_{pk}\mathrm{f}_k + \mathrm{e}_p + \mu_p \end{aligned} \tag{1}$$

where,

f_j is the jth <u>common factor</u> random variable.
λ_{ij} is the parameter representing the importance of the jth factor in the composition of x_i (i.e. the <u>loading</u> of x_i on the jth common factor.
e_i is the ith <u>specific factor</u> random variable.

We shall require that:

(i) The f_i be uncorrelated, each with mean 0 and variance 1.
(ii) The e_i be uncorrelated, each with mean 0 and with $\mathrm{Var}(\mathrm{e}_i) = \psi_i$.
(iii) $\mathrm{Cov}(\mathrm{f}_i, \mathrm{e}_j) = 0$, for all $i = 1, \ldots, k$ and $j = 1, \ldots, p$.

From this model, the variance and covariances of the $x_1, \ldots, x_p$ can be expressed in terms of the model parameters λ_i and ψ_i.

Clearly,

$$\begin{aligned}\mathrm{Var}(x_i) = \sigma_{ii} &= \lambda_{i1}^2 + \lambda_{i2}^2 + \ldots + \lambda_{ik}^2 + \psi_i \\ &= h_i^2 + \psi_i\end{aligned} \tag{2}$$

where,

h_i^2 is called the <u>communality</u> of x_i, i.e. the variance of x_i accounted for by its dependence on the k common factors, and
ψ_i is called the <u>specific</u> or <u>unique</u> variance of x_i.

Also,

$$\begin{aligned}\mathrm{Cov}(x_i, x_j) &= \mathrm{Cov}(\lambda_{i1}\mathrm{f}_1 + \ldots + \lambda_{ik}\mathrm{f}_k + \mathrm{e}_i, \lambda_{j1}\mathrm{f}_1 + \ldots + \lambda_{jk}\mathrm{f}_k + \mathrm{e}_j) \\ &= \lambda_{i1}\lambda_{j1} + \lambda_{i2}\lambda_{j2} + \ldots + \lambda_{ik}\lambda_{jk}\end{aligned} \tag{3}$$

Thus, the covariance is entirely due to the common dependence on the factors $\mathrm{f}_1, \ldots, \mathrm{f}_k$.

In matrix form,

$$\Sigma = \Lambda\Lambda^T + \Psi \tag{4}$$

where,

$$\Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \ldots & \lambda_{1k} \\ \lambda_{21} & \lambda_{22} & \ldots & \lambda_{2k} \\ \vdots & & & \\ \lambda_{p1} & \lambda_{p2} & \ldots & \lambda_{pk} \end{pmatrix} \text{ and } \Psi = \mathrm{diag}(\psi_i) = \begin{pmatrix} \psi_1 & 0 & \ldots & 0 \\ 0 & \psi_2 & \ldots & 0 \\ \vdots & & & \\ 0 & 0 & \ldots & \psi_p \end{pmatrix}$$

Thus, model (1), subject to conditions (i), (ii) and (iii) can be written as

$$\mathbf{x} = \Lambda\mathbf{f} + \mathbf{e} + \boldsymbol{\mu} \tag{5}$$

subject to

$$\mathrm{E}(\mathbf{f}) = \mathbf{0}, \mathrm{Cov}(\mathbf{f}, \mathbf{f}) = \mathrm{I} \tag{6}$$

$$\mathrm{E}(\mathbf{e}) = \mathbf{0}, \mathrm{Cov}(\mathbf{e}, \mathbf{e}) = \Psi \tag{7}$$

$$\mathrm{Cov}(\mathbf{f}, \mathbf{e}) = 0 \tag{8}$$

<u>Definition</u>: The k-factor model holds for the random vector $\mathbf{x}$ if $\mathbf{x}$ can be written in the form (5) and (6)-(8) hold.

Theorem 1
The k-factor model holds for $\mathbf{x}$ if and only if the covariance matrix, Σ, of $\mathbf{x}$ can be written in the form

$$\Sigma = \Lambda\Lambda^T + \Psi$$

Proof: (See Mardia, Kent and Bibby).

Theorem 2
Factor analysis is unaffected by a rescaling of the variables.

Proof: As Exercise.

[*Hint*: Set $\mathbf{y} = \mathrm{C}\mathbf{x}$ where $\mathrm{C} = \mathrm{diag}(c_i)$. Find $\mathrm{Cov}(\mathbf{y}, \mathbf{y})$ and apply Theorem 1. Give the factor loading matrix for $\mathbf{y}$]

Thus, unlike PCA, FA is scale-free. This gets rid of one form of non-uniqueness. However, there is now another form of non-uniqueness.

Non-uniqueness of Factor Loadings

Suppose the k-factor model holds for $\mathbf{x}$. Then,

$$\mathbf{x} = \Lambda\mathbf{f} + \mathbf{e} + \boldsymbol{\mu}, \qquad \text{and (6)-(8) hold}$$

Let G be any $k \times k$ orthogonal matrix. ($\mathrm{GG}^T = \mathrm{G}^T\mathrm{G} = \mathrm{I}$). Then,

$$\begin{aligned}\mathbf{x} &= \Lambda(\mathrm{GG}^T)\mathbf{f} + \mathbf{e} + \boldsymbol{\mu} \\ &= (\Lambda\mathrm{G})(\mathrm{G}^T\mathbf{f}) + \mathbf{e} + \boldsymbol{\mu}\end{aligned}$$

and $\mathrm{G}^T\mathbf{f}$ and $\mathbf{e}$ also satisfy (6)-(8).

So the k-factor model holds for new factors $\mathrm{G}^T\mathbf{f}$, new loadings $\Lambda\mathrm{G}$ and the original unique random factor $\mathbf{e}$.

As required,

$$\Sigma = (\Lambda\mathrm{G})(\Lambda\mathrm{G})^T + \Psi$$

Thus, for fixed Ψ (i.e. fixed $\mathbf{e}$), the k-dimensional factor space is unique - but there are infinitely many sets of orthogonal factor axes that span it. In this case we can resolve the non-uniqueness by adding another constraint. We require that the loadings (i.e. the

common factor axes) are rotated so as to satisfy, say

$$\Lambda^T \Psi^{-1} \Lambda = \text{diag}(\theta_i) \tag{9}$$

where, $\theta_1 > \theta_2 > \ \ldots > \theta_k > 0$.

Theorem 3
If the k-factor model holds for $\mathbf{x}$, then, given Ψ (i.e. $\mathbf{e}$), the constraint (9) uniquely determines the loading matrix Λ.

Proof: As Exercise.

[*Outline*: If $\Lambda\Lambda^T = \Delta\Delta^T$, and $\Lambda^T\Psi^{-1}\Lambda$ and $\Delta^T\Psi^{-1}\Delta$ are both diagonal matrices with distinct non-zero elements written in decreasing order, then it is required to prove that the k columns of Λ must be identical to the k columns of Δ, except possibly for a difference in sign. Prove by considering the eigenvalues and eigenvectors of the matrix

$$\Psi^{-\frac{1}{2}}\Lambda\Lambda^T\Psi^{-\frac{1}{2}} = \Psi^{-\frac{1}{2}}\Delta\Delta^T\Psi^{-\frac{1}{2}}$$

and hence showing that the k columns of $\Psi^{-\frac{1}{2}}\Lambda$ must be equal to the k corresponding columns of $\Psi^{-\frac{1}{2}}\Delta$, except possibly for a difference in sign].

Note that constraint (9) is arbitrarily chosen and others are possible. It is easy to see (from Theorem 2) that (9) is scale invariant.

5.2 Estimation of Λ and Ψ

Maximum Likelihood Factor Analysis

Suppose that we have a random sample $\mathbf{x}_1, \mathbf{x}_2, \ \ldots, \mathbf{x}_n$ from the population with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. Recall that, under the assumption of multivariate normality, the log-likelihood function is given by

$$\ell(\boldsymbol{\mu}, \Sigma; \mathrm{X}) = -\frac{np}{2}\ln 2\pi - \frac{n}{2}\ln|\Sigma| - \frac{n}{2}\text{tr}(\Sigma^{-1}\mathrm{S}) - \frac{n}{2}(\overline{\mathbf{x}} - \boldsymbol{\mu})^T\Sigma^{-1}(\overline{\mathbf{x}} - \boldsymbol{\mu})$$

If we replace $\boldsymbol{\mu}$ by its m.l.e. $\overline{\mathbf{x}}$ we have

$$\ell(\overline{\mathbf{x}}, \Sigma; \mathrm{X}) = -\frac{n}{2}[p\ln 2\pi + \ln|\Sigma| + \text{tr}(\Sigma^{-1}\mathrm{S})]$$

If we assume that $\Sigma = \Lambda\Lambda^T + \Psi$, then ℓ is a function of Λ and Ψ, and m.l.e.'s of Λ and Ψ are obtained by maximizing ℓ with respect to Λ and Ψ. Equivalently, we shall minimize the following function.

$$F(\Lambda, \Psi) = \ln|\Sigma| + \mathrm{tr}(\Sigma^{-1}\mathrm{S}) - \ln|\mathrm{S}| - p \qquad (*)$$

where $\Sigma = \Lambda\Lambda^T + \Psi$.

Procedures to obtain parameter estimates Λ and Ψ for the k-factor model by minimizing $(*)$ (subject to the constraint (9)) form the basis of algorithms for maximum likelihood factor analysis which are available in many statistical packages (eg SAS and S-PLUS).

5.3 Goodness-of-fit

Factor Analysis proposes a model for the covariance structure of a set of variables. Assuming multivariate normality, we can obtain from data maximum likelihood estimates of the parameters in the model. It is natural then to use the generalised likelihood ratio to test the adequacy of the k-factor model for explaining the observed covariances.

Suppose we have a random sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from MVN$(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then, the null hypothesis is

$$H_0 : \Sigma = \Lambda\Lambda^T + \Psi$$

where Λ is a $p \times k$ matrix and Ψ is a diagonal matrix with non-negative elements. We want to test H_0 against the alternative

$$H_1 : \Sigma \text{ has no constraints}$$

<u>Under H_0</u> the m.l.e.s are $\hat{\boldsymbol{\mu}} = \overline{\mathbf{x}}$ and $\hat{\Sigma} = \hat{\Lambda}\hat{\Lambda}^T + \hat{\Psi}$ where $\hat{\Lambda}$ and $\hat{\Psi}$ are the mle's of Λ and Ψ which minimize $(*)$ above, subject to the constraint (9). Thus,

$$\begin{aligned} \ell_0^* &= \ell(\overline{\mathbf{x}}, \hat{\Sigma}, \mathrm{X}) \\ &= -\frac{1}{2}n[p\ \ln 2\pi + \ln|\hat{\Sigma}| + \mathrm{tr}(\hat{\Sigma}^{-1}\mathrm{S})] \end{aligned}$$

<u>Under H_1</u> the m.l.e.s are $\hat{\boldsymbol{\mu}} = \overline{\mathbf{x}}$ and $\hat{\Sigma} = \mathrm{S}$, and

$$\begin{aligned} \ell_1^* &= \ell(\overline{\mathbf{x}}, \mathrm{S}, \mathrm{X}) \\ &= -\frac{1}{2}n[p\ \ln 2\pi + \ln|\mathrm{S}| + p] \end{aligned}$$

The likelihood ratio statistic λ is defined as

$$\lambda = \frac{L_0^*}{L_1^*}$$

Clearly $0 \leq \lambda \leq 1$ and assymptotically $-2\ln\lambda$ has a χ^2 distribution if H_0 holds. That is

$$\begin{aligned} -2\ln\lambda &= 2(\ell_1^* - \ell_0^*) \\ &= n\mathrm{tr}(\hat{\Sigma}^{-1}\mathrm{S}) + n\ln|\hat{\Sigma}| - n\ln|\mathrm{S}| - np \\ &= nF(\hat{\Lambda}, \hat{\Psi}) \end{aligned} \tag{10}$$

Under H_0 the statistic $-2\ln\lambda$ has an assymptotic χ_r^2 distribution where r is the difference between the number of free parameters in H_1 (n_1) and the number of free parameters in H_0 (n_0).

$$n_1 = \frac{1}{2}p(p+1)$$

$$n_0 = pk + p - \frac{1}{2}k(k-1)$$

Then, it is easily verified that

$$r = \frac{1}{2}(p-k)^2 - \frac{1}{2}(p+k) \tag{11}$$

In practice the n in (10) is replaced by $n' = n - 1 - \frac{1}{6}(2p+5) - \frac{2}{3}k$ to improve convergence to χ_r^2. This is Bartlett's improved χ_r^2 approximation. So to test H_0, test

$$U = n'F(\hat{\Lambda}, \hat{\Psi}) \sim \chi_r^2$$

This asymptotic approximation is good for $n \geq p + 50$.

5.4 Interpretation of Factor Analysis Output

Since Factor Analysis is scale-free, and maximum likelihood is a scale free estimation procedure, it is the sample correlation matrix R that is generally analysed. From this it follows that:

(a) Since diag(R)=I, the m.l. estimates $\hat{\Lambda}$ and $\hat{\Psi}$ satisfy

$$\sum_{j=1}^{k} \hat{\lambda}_{ij}^2 + \hat{\psi}_i = 1$$

or $\text{diag}(\hat{\Lambda}\hat{\Lambda}^T + \hat{\Psi}) = \text{diag(R)}$, which implies the following relationship between the specific variance $\hat{\psi}_i$ and the communality $\hat{h}_i$.

$$\hat{\psi}_i = 1 - \sum_{j=1}^{k} \hat{\lambda}_{ij}^2 = 1 - \hat{h}_i^2 \tag{12}$$

(b) $\hat{\lambda}_{ij}$ is an estimate of the <u>correlation</u> between x_i and f_j. That is,

$$\mathbf{x} = \Lambda\mathbf{f} + \mathbf{e} + \boldsymbol{\mu}$$
$$\Rightarrow \text{Cov}(\mathbf{x}, \mathbf{f}) = \text{Cov}(\Lambda\mathbf{f} + \mathbf{e} + \boldsymbol{\mu}, \mathbf{f}) = \Lambda$$

and since, in the sample, $\mathbf{x}$ is scaled so that $s_{ii} = 1$, then $\hat{\Lambda}$ estimates $\text{Corr}(\mathbf{x}, \mathbf{f})$.

<u>Given $\hat{\Lambda}$ (and hence $\hat{\Psi}$ from (12)), what can we say about the model fitted?</u>

We have:

$$\hat{\Lambda} = \begin{bmatrix} \hat{\lambda}_{11} & \hat{\lambda}_{12} & \dots & \hat{\lambda}_{1k} \\ \hat{\lambda}_{21} & \hat{\lambda}_{22} & \dots & \hat{\lambda}_{2k} \\ \vdots & & & \\ \hat{\lambda}_{p1} & \hat{\lambda}_{p2} & \dots & \hat{\lambda}_{pk} \end{bmatrix} \qquad \begin{matrix} \sum_{j=1}^{k} \hat{\lambda}_{1j}^2 = \hat{h}_1^2 \\ \sum_{j=1}^{k} \hat{\lambda}_{2j}^2 = \hat{h}_2^2 \\ \vdots \\ \sum_{j=1}^{k} \hat{\lambda}_{pj}^2 = \hat{h}_p^2 \end{matrix}$$

(i) The <u>interpretation</u> of a factor depends on the sizes of the $\hat{\lambda}$'s in the corresponding column of $\hat{\Lambda}$. (All are comparable, estimated correlations). Interpretation is clearly subjective, but crucial since it makes no sense to set up such an elaborate model to explain the correlations if the factors are uninterpretable.

(ii) $\hat{h}_i^2 = \sum_{j=1}^{k} \hat{\lambda}_{ij}^2$, the sum of squared elements in row i of $\hat{\Lambda}$, estimates the <u>communality</u> of x_i. This is the amount of the variance of x_i accounted for by the dependence on the k factors. The relative sizes of $\hat{h}_i^2$ and $\hat{\psi}_i = 1 - \hat{h}_i^2$ are of interest.

(iii) $\sum_{i=1}^{p} \hat{\lambda}_{ij}^2$ is the sum of squared elements in column j of $\hat{\Lambda}$. This is the amount of the total variance (p) accounted for by factor j.

(iv) $\sum_{i=1}^{p}\sum_{j=1}^{k}\hat{\lambda}_{ij}^2 = \sum_{i=1}^{p}\hat{h}_i^2$ is the amount of the total variance (p) accounted for by all the factors.

(v) $\sum_{l=1}^{k}\hat{\lambda}_{il}\hat{\lambda}_{jl}$ estimates $\mathrm{Corr}(x_i, x_j)$. These correlations are what the model is attempting to reproduce. It is of interest to see how close these estimates are to the observed r_{ij}, the elements of R. It is usual to examine the elements of $(\mathrm{R} - \hat{\Lambda}\hat{\Lambda}^T)$. The off-diagonal elements are the correlation residuals. The diagonal elements represent the estimated unique variances.

Interpretation and Rotation of Factors

The constraint that $\Delta = \hat{\Lambda}^T\hat{\Psi}^{-1}\hat{\Lambda}$ be diagonal is a mathematical convenience to make the factor loadings unique. It results in an ordering of factors such that the first factor has maximal contribution to the common variance of the observed variables, the second has maximal contribution to this variance subject to being uncorrelated with the first, and so on. Nevertheless it may not lead to the most easily interpretable factors.

Since the factors may be transformed using

$$\mathbf{x} = (\hat{\Lambda}\mathrm{G})(\hat{\Lambda}\mathbf{f}) + \mathbf{e} + \boldsymbol{\mu}$$

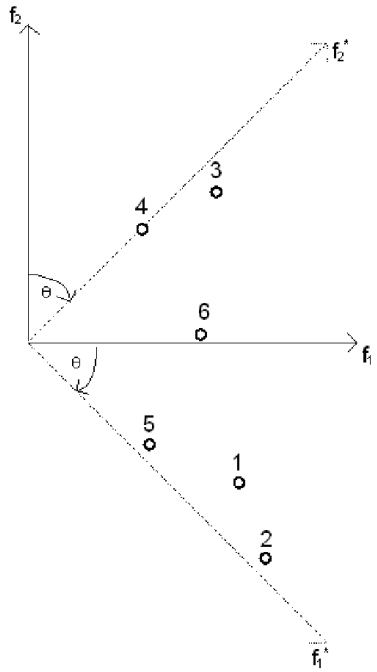
i.e.

$$\hat{\Sigma} = (\hat{\Lambda}\mathrm{G})(\hat{\Lambda}\mathrm{G})^T + \hat{\Psi}$$

where $\mathrm{G}^T\mathrm{G} = \mathrm{G}\mathrm{G}^T = \mathrm{I}$, without affecting the validity of the model - we may choose an orthogonal transformation matrix G so as to make the factors as 'meaningful' as possible. Such transformations amount to an orthogonal rotation of the axes of the k-dimensional factor space and are commonly called <u>rotations</u> of the <u>loadings</u>.

What criteria should we use to find a 'good' rotation?

Consider the following example.



$$\underset{6\times 2}{\hat{\Lambda}} = \begin{bmatrix} \hat{\lambda}_{11} & \hat{\lambda}_{12} \\ \vdots & \\ \hat{\lambda}_{61} & \hat{\lambda}_{62} \end{bmatrix}$$

Initially;
f_1 - all x_i have high loadings,
f_2 - 2 large +ve and 3 large −ve.

After rotation;
f_1^* - associated with 1, 2 and 5,
f_2^* - associated with 3 and 4,
and variable 6 has medium loadings on both f_1^* and f_2^*.

That is the rotation has made interpretation easier.

Possible criteria for a 'good' structure

Interpretation is most straightforward if the variables are split into disjoint sets, each associated with a single factor. This aim is behind Thurstone's 'simple structure', which has the following properties. (For more detail see Everitt and Dunn, p278).

(1) Loadings should be either large and positive or near zero as far as possible. (A large number of intermediate values complicates the interpretation).

(2) Each variable should be highly loaded on at most one factor.

(3) On each factor there should be a reasonable number of negligable loadings.

Such an ideal rotation rarely exists in practice.

Rotation can be done graphically but it is not considered wise to allow the rotation to be chosen subjectively. There are a number of analytical methods of rotation. We consider just one.

Varimax Rotation

It is required to choose $\Gamma = (\gamma_{ij}) = \Lambda \mathrm{G}$ according to some kind of mathematical criterion.

The varimax criterion maximizes the function ϕ, where ϕ is the sum of the variances of the squared elements within each column of the loading matrix $\Lambda = (\lambda_{ij})$, after each row i has been normalized by h_i, the square root of its communality. i.e.

$$\phi = \sum_{j=1}^{k} \sum_{i=1}^{p} (d_{ij}^2 - \overline{d}_j)^2 = \sum_{j=1}^{k} \sum_{i=1}^{p} d_{ij}^4 - p \sum_{j=1}^{k} \overline{d}_j^2$$

where $d_{ij} = \gamma_{ij}/h_i$ and $\overline{d}_j = \frac{1}{p} \sum_{i=1}^{p} d_{ij}^2$.

ϕ is a function of G and can be maximized numerically with respect to G.

5.5 Factor Scores

So far we have been concerned with the way the observed variables are functions of the unknown factors.

e.g. In Spearman's exam data the factor model is

$$x_i = \lambda_{i1} \mathrm{f}_1 + \mathrm{e}_i + \mu_i$$

and describes how a child's test scores depend on his/her overall 'intelligence'.

Conversely, we could ask: Given a child's set of test scores, what can we say about his or her 'score' on the overall intelligence factor?

In general it may be required to find the values or scores of each individual on the derived factors.

In PCA each component, by definition, is a linear combination of the original variables. For the factor model things are not so simple. There is no obvious way of inverting the model to obtain the f_j in terms of the x_i, and the problem of finding the factor scores (from a fully specified factor model) has to be regarded as a problem of estimation.

There are a number of commonly used methods of estimation, but these will not be considered here. [See, for example, Mardia, Kent and Bibby].

5.6 PCA and FA

There are obviously many similarities between principal components analysis and factor analysis, but there are some important differences also. We summarise a few points.

1. Both are ways of factorising a covariance matrix.

   PCA: $\mathbf{S} = \mathbf{\Gamma}\mathbf{\Gamma}^T$ FA: $\mathbf{S} \approx \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T + \hat{\mathbf{\Psi}}$

2. PCA: method for accounting for variances.
   FA: method for explaining covariances.

   Both are pointless if all the observed variables are approximately uncorrelated.

3. PCA: factorisation is unique for a particular scaling of variables.
   FA: factorisation is only unique up to a rotation of the factor axes.

   Note: some people rotate Principal components - but this leads to a loss properties of maximum variance etc.

4. PCA: not scale invariant.
   FA: model scale invariant. MLFA estimates scale invariant.

5. PCA: just an orthogonal rotation of variable axes - no assumptions required and no tests possible if none are made.
   FA: based on a 'proper' statistical model. This implies that many more assumptions are required, but for MLFA test of goodness-of-fit is possible. However in most practical applications k, the number of factors is unknown and the researcher resorts to trying $k = 1, k = 2, k = 3$, etc. until the model 'fits'. It should be noted that the χ^2 statistics obtained in this fashion are not independent and the true significance level of the test may be very different from the nominal value used at each stage.

6. PCA: The number of components worth retaining may be varied without the components retained altering in any way.
   FA: Increasing the number of factors from k to $k + 1$ gives a completely new matrix of loadings which may look quite different from the k-factor matrix of loadings.

Finally, see the extract from Chatfield and Collins in the Examples handout discussing the drawbacks of FA. Theirs is an extreme view but should serve as a warning against the unthinking use of FA - or indeed PCA.