Chapter 4

Principal Component and Factor Analysis

4.1 Introduction

Let $\mathbf{x}$ denote an $n \times p$ data matrix of n p-variate observations. Principal Component Analysis is a technique applied when some of the variables are highly correlated. The aim is to find m linear combinations of the variables, where $m < p$, which describe the sample covariance or correlation structure of the data set.

PCA may be carried out on either S, the sample covariance matrix, or R, the sample correlation matrix. The sample correlation matrix is preferable if the p variables in the data set have widely varying scales.

The aim is

- data reduction (reducing p variables to m linear combinations of the variables)
- interpretation (we examine which variables influence the principal components and, from this, try to determine hidden factors; the principal components are *factors*).

4.2 Principal Component Analysis

Since PCA is concerned with the covariance / correlation structure of the variables, the data matrix is first *centred*, so that the columns are all mean zero. Let

$$H = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^t \tag{4.1}$$

where I_n denotes the $n \times n$ identity matrix and $\mathbf{1}_n$ denotes the n-vector with each entry 1. Let

$$\mathbf{z} = H\mathbf{x},$$

then the entries of the $n \times p$ matrix $\mathbf{z}$ are

$$z_{ij} = x_{ij} - \overline{x}_{.j}.$$

The sample covariance matrix S of $\mathbf{x}$ is given by:

$$S = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \overline{x}_{.i})(x_{kj} - \overline{x}_{.j}) = \frac{1}{n-1}\mathbf{z}^t\mathbf{z}. \tag{4.2}$$

A principal component analysis simply finds the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ of the sample covariance matrix S and the corresponding eigenvectors $P_{.j} : j = 1, \ldots, p$. The principal components are the uncorrelated linear combinations; $\mathbf{y} = \mathbf{z}P$, where $\mathbf{y}_{.1}$ has the largest possible statistical variance among orthonormal transformations of $\mathbf{z}$ and $\mathbf{y}_{.q}$ has the largest statistical variance under the constraint that it is uncorrelated with $(\mathbf{y}_{.1}, \ldots, \mathbf{y}_{.,q-1})$.

Lemma 4.1. *Let S be the sample covariance matrix defined by Equation* (4.2) *and let λ_1 be the* largest *eigenvalue of S and let $\underline{\gamma}$ denote the corresponding normalised eigenvector; namely,*

$$S\underline{\gamma} = \lambda_1\underline{\gamma}, \qquad \sum_{j=1}^{p}\gamma_j^2 = 1.$$

Let $z_{ij} = x_{ij} - \bar{x}_{.j}$. Then, for any p- vector $\underline{a}$, with $\sum_{j=1}^{p} a_j^2 = 1$,

$$\sum_{i=1}^{n}\left(\sum_{j=1}^{p} a_j z_{ij}\right)^2 \leq \sum_{i=1}^{n}\left(\sum_{j=1}^{p}\gamma_j z_{ij}\right)^2.$$

Proof

$$\begin{aligned}\sum_{i=1}^{n}\left(\sum_{j=1}^{p} a_j z_{ij}\right)^2 &= \sum_{i=1}^{n}\sum_{jk=1}^{p} a_j a_k z_{ij} z_{ik} \\ &= (n-1)\sum_{jk} a_j a_k S_{jk} = (n-1)\underline{a}^t S\underline{a} = (n-1)\underline{a}^t P^t D P\underline{a} = (n-1)\underline{b}^t D\underline{b},\end{aligned}$$

where $\underline{b} = P\underline{a}$. Note that $\underline{b}^t\underline{b} = \underline{a}^t P^t P\underline{a} = \underline{a}^t\underline{a} = 1$, so $\underline{b}$ is a unit vector. Since $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ where $\lambda_1 \geq \ldots \geq \lambda_p$, it follows that the expression is maximised if $\underline{b} = (1, 0, \ldots, 0)$, so

$$\sum_{i=1}^{n}\left(\sum_{j=1}^{p} a_j z_{ij}\right)^2 \leq (n-1)\lambda_1.$$

Meanwhile, since $\underline{\gamma}$ is a unit eigenvector of S with eigenvalue λ_1, it follows that

$$\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\gamma_j z_{ij}\right)^2 = (n-1)\underline{\gamma}^t S\underline{\gamma} = (n-1)\lambda_1\underline{\gamma}^t\underline{\gamma} = (n-1)\lambda_1$$

and the result follows. □

Notation λ_k will be used to denote the kth largest eigenvalue.

Definition 4.2 (Principal Component, Loading Vector)**.** *Let* $\mathbf{x}$ *denote the* $n \times p$ *data matrix,* n *multivariate observations on* p *variables. Let* P *denote the orthonormal matrix and* D *the diagonal matrix with elements arranged in decreasing order such that* $S = PDP^t$*. Let* H *be the* $n \times n$ *matrix defined by Equation* (4.1)*. The* columns *of the matrix*

$$\mathbf{y} = H\mathbf{x}P \tag{4.3}$$

are called the sample principal components*. The* i*th element of the* k*th column represents the* score *of the* k*th principal component for the* i*th observation. The* k*th column of the orthonormal matrix* P *is the* loading vector *for the* k*th principal component.*

The following theorem is stated without proof; it is left as an exercise.

Theorem 4.3. *Let* $\mathbf{x}$ *be an* $n \times p$ *data matrix;* n *p-variate observations. Let* P *denote the orthonormal matrix and* D *the diagonal matrix with elements from highest to lowest such that* $S = PDP^t$*, where* S *is the unbiased sample covariance matrix. Let*

$$\mathbf{y} = H\mathbf{x}P$$

where

$$H = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^t.$$

Let

$$S^{(\mathbf{y})} = \frac{1}{n-1}\mathbf{y}^t\mathbf{y}.$$

Then

1. $$S^{(\mathbf{y})} = D$$

2. *Let* $\underline{z}_k = (z_{1,k}, \ldots, z_{n,k})^t$ *where* $z_{jk} = x_{jk} - \frac{1}{n}\sum_{i=1}^n x_{ik}$*. Then the linear combinations*

$$P_{1m}\underline{z}_1 + \ldots + P_{pm}\underline{z}_p, \qquad m = 1, \ldots, q \leq p$$

   span the parallelepiped with the largest volume among all parallelepipeds spanned by standardised linear combinations of $\mathbf{x}$ *in variable space.*

3. *Let* $c_{kl}^{(Y)} = \sum_{i=1}^n y_{ik}y_{il}$*. The largest volume is given by*

$$\sqrt{det\left((c_{kl}^{(Y)})_{(k,l)\in\{1,\ldots,q\}^2}\right)} = (n-1)^{q/2}\sqrt{d_{11}\ldots d_{mm}}.$$

4. $$tr(S^{(\mathbf{x})}) = tr(S^{(\mathbf{y})}) = \lambda_1 + \ldots + \lambda_p.$$

Recall For a symmetric $p \times p$ matrix, the sum of the trace is equal to the sum of the eigenvalues. It follows that $\text{tr}(S) = \sum_{j=1}^{p} \lambda_j = \text{tr}(D)$.

Interpretation

- Plotting $\mathbf{y}_{.,1}, \ldots, \mathbf{y}_{.,q}$ for $q < p$ in object space gives the 'best' q dimensional summary of $\mathbf{x}$ if one is looking for the parallelepiped of largest volume spanned by standardised linear combinations of $\mathbf{x}$ in variable space (recall that the point $\underline{\overline{x}}^t := (\overline{x}_{.,1}, \ldots, \overline{x}_{.,p})$ is the origin in variable space.

- It often turns out that all the loadings for the *first* principal component are positive. If this is the case, then it can be interpreted as a measurement of size. If this is the case, then it necessarily follows that all the other principal components have both positive and negative loadings and are therefore interpreted in terms of shape. Since the general idea is to reduce the data and to only use m principal components where $m < p$, they will not cover all possibilities for shape.

- The sum of the unbiased sample variances, that is $\text{tr}(S)$ is also called the total sample variation of $\mathbf{x}$. If only m principal components are used, then
$$\frac{\lambda_1 + \ldots + \lambda_m}{\lambda_1 + \ldots + \lambda_p}$$
represents the proportion of the variance explained by the first m sample principal components. There are two usual criteria for deciding how many to use:

  1. The m sample prinicipal components explain 90% of the variation.
  2. (Kaiser's criterion) The variances of the sample principal components beyond the mth principal components account for less than the average $\frac{1}{p}\text{tr}(S)$.

  When $p \leq 20$, the second of these tends to include too few components.

- After deciding on the number of principal components m to include, the data is represented only using the first m principal components: using
$$\mathbf{y} = H\mathbf{x}P = \mathbf{z}P,$$
it follows that
$$\mathbf{x} = \mathbf{1}_n \underline{\overline{x}}^t + \mathbf{y}P^t$$
the components $m+1, \ldots, p$ are estimated by 0, giving $\widehat{\mathbf{x}}$, the estimate of $\mathbf{x}$ as:
$$\hat{\mathbf{x}} = \mathbf{1}\overline{\mathbf{x}}^t + (\mathbf{y}_{.,1} | \ldots | \mathbf{y}_{.,m}) \begin{pmatrix} P_{11} & \ldots & P_{p1} \\ \vdots & \ddots & \vdots \\ P_{1m} & \ldots & P_{pm} \end{pmatrix}.$$

Recall that the vector $(P_{1k}, \ldots, P_{p,k})^t$ is the kth loading vector.

- A PCA on the correlation is equivalent to using *standardised* variables, which is preferable if variables on a *smaller* scale give significant information, which can be lost if the raw data is used.

  Note The analysis is not scale invariant; while covariance and correlation give the same eigenvectors, the order of their corresponding eigenvalues may change, the loadings will change and their interpretation may change.

- It may be possible to decide that some of the p variables are redundant, based on the PCA analysis on the *correlation* matrix. This is done by considering the *last* sample principal component and discarding the variable assigned to the loading with largest absolute value. Then continue with the loadings of the *second* last principal component, and so on. Stop discarding if certain criteria are satisfied: for example,

  1. the eigenvalue corresponding to the loadings is greater than 0.7 (this seems to work in practice)
  2. the sample principal components corresponding to the loadings you have not yet considered explain less than 80% of the variation.

  Using either criterion, at least four variables should always be retained.

The columns of the matrix P are often called the *coefficients*.

Prinipal component analysis is only useful as a tool if some of the eigenvalues of the statistical correlation matrix are very small. Absolutely nothing is achieved by a principal component analysis if all the eigenvalues of the correlation matrix are significant.

4.3 How to do a Principal Component Analysis

Throughout this discussion, variance refers to statistical variance, covariance to statistical covariance and correlation to statistical correlation. Firstly, suppose that the PCA is being carried out on the *covariance*. The procedure is as follows: suppose there are n independent observations from $(X_1, \ldots, X_p)$. Firstly, the data is centralised:

$$\mathbf{z} = H\mathbf{x}$$

and the statistical covariance is computed;

$$S = \frac{1}{n-1}\mathbf{z}^t\mathbf{z}.$$

The *first* principal component is

$$\mathbf{y}_{.1} = P_{11}\mathbf{z}_{.1} + \ldots + P_{p1}\mathbf{z}_{.p},$$

where $P_{11}, \ldots, P_{p1}$ are chosen to maximise $\mathrm{Var}(\mathbf{y}_{.1})$ subject to the constraint that $\sum_{k=1}^{p} P_{k1}^2 = 1$. That is, to maximise

$$P_{.1}^t S P_{.1}$$

where $P_{.1}$ is taken as a column vector, subject to the constraint. Once the first component has been established, the second component

$$\mathbf{y}_{.2} = P_{12}\mathbf{z}_{.1} + \ldots + P_{p2}\mathbf{z}_{.p}$$

is established by finding $(P_{12}, \ldots, P_{p2})$ that maximises $\mathrm{Var}(\mathbf{y}_{.2})$ subject to the constraints

$$\sum_{k=1}^{p} P_{k2}^2 = 1,$$

$$P_{.2}^t S P_{.2} = 0.$$

That is, $P_{.2}$ is chosen to ensure that the *statistical* correlation is zero. Inductively, once $P_{.j}$ have been established for $j = 1, \ldots, k-1$, $P_{.k}$ are established by maximising the estimate of $\mathrm{Var}(\mathbf{y}_k)$,

$$P_{.k}^t S P_{.k}$$

subject to the constraints that

$$\sum_{l=1}^{p} P_{lk}^2 = 1$$

and the statistical covariances $\mathrm{Cov}(\mathbf{y}_{.,j}, \mathbf{y}_{.,k})$ are zero for $j = 1, \ldots, k-1$. That is

$$P_{.j}^t S P_{.k} = 0, \qquad j = 1, \ldots, k-1.$$

Note that the statistical variances of the principal components are the eigenvalues of the sample covariance matrix and that the columns $P_{.k}$ are the eigenvectors.

Recall that, for a symmetric $m \times m$ matrix C, with eigenvalue $\lambda_1, \ldots, \lambda_m$,

$$\mathrm{tr}(C) = \sum_{j=1}^{m} \lambda_j.$$

Let λ_j denote the estimates of $\mathrm{Var}(Z_j)$. It follows that

$$\sum_{j=1}^{p} S_{jj} = \sum_{j=1}^{p} \lambda_j.$$

Since principal component analysis considers dependence and independence, it is usual to code the variables $\mathbf{x}_{.1}, \ldots, \mathbf{x}_{.p}$ so that they each have mean 0 and variance 1 at the beginning of the analysis. The procedure with this modification is therefore as follows:

1. Compute $\bar{x}_{.k}$ for $k = 1, \ldots, p$ and $S_{kl} = \frac{1}{n-1} \sum_{j=1}^{n} (x_{jk} - \bar{x}_{.k})(x_{jl} - \bar{x}_{.l})$, the sample means and sample covariance matrix.

2. Compute the coded variables, $y_{jk} = \frac{x_{jk} - \bar{x}_{.k}}{\sqrt{S_{kk}}}$.

3. Compute the correlation matrix
$$R_{kl} = \frac{S_{kl}}{\sqrt{S_{kk} S_{ll}}}.$$
This is the covariance matrix for the coded variables.

4. Find the eigenvalues $\lambda_1, \ldots, \lambda_p$ and the corresponding eigenvectors $P_{.1}, \ldots, P_{.p}$ in the way described above.

5. Discard any principal components that do not account for a significant variation in the data. This means that $\mathbf{y}_{.k}, \ldots, \mathbf{y}_{.p}$ are discarded for k such that $\sum_{j=k+1}^{p} \lambda_j \leq \alpha \leq \sum_{j=k}^{p} \lambda_j$ where α is the level of the variation that is to be ignored. Usually, this is roughly 20% or, when the data is standardised, components corresponding to eigenvalues less than 1 are ignored.

4.4 Confidence Intervals for PCA Eigenvalues and Eigenvectors

There exists some results in the literature. The proofs of these are long and technical. Much more seriously, they all rely on the assumption that the data comes from i.i.d. p-variate Gaussian variables and that n is large.

Theorem 4.4 (Lawley (1956))**.** *If λ_i is a distinct eigenvalue of the covariance (correlation) matrix, then*
$$\mathbb{E}[\widehat{\lambda}_i] = \lambda_i + \frac{\lambda_i}{n} \sum_{j \neq i} \frac{\lambda_j}{\lambda_i - \lambda_j} + O(n^{-2})$$
so that the estimate is asymptotically unbiased and:
$$\mathbf{V}(\widehat{\lambda}_i) = \frac{2\lambda_i^2}{n} \left(1 + \frac{1}{n} \sum_{j \neq i} \left(\frac{\lambda_i}{\lambda_i - \lambda_j} \right)^2 \right) + O(n^{-3}).$$

Also, let $\widehat{h}_i$ denote the estimate of the ith eigenvector h_i with λ_i the ith eigenvalue, then

1.
$$\sqrt{n}(\underline{\widehat{\lambda}} - \underline{\lambda}) \longrightarrow_{(d)} N_p(\underline{0}, 2\Lambda^2)$$
where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$

2.
$$\sqrt{n}(\widehat{\underline{h}}_i - \underline{h}_i) \longrightarrow_{(d)} N_p(\underline{0}, E_i)$$

where
$$E_i = \lambda_i \sum_{k \neq i} \frac{\lambda_k}{(\lambda_i - \lambda_k)^2} \underline{h}_k \underline{h}_k^t.$$

3. For each $i = 1, \ldots, p$, $\widehat{\lambda}_i \perp \widehat{\underline{h}}_i$.

Since the 'normality' assumption for these results is not usually satisfied and size of the data set is often insufficient for a 'central limit theorem effect', these results are of limited value. To find confidence intervals for eigenvalues, bootstrap methods may be used.

If we have an $n \times p$ data matrix, a bootstrap method takes randomly chosen subsets of size m, were $m \leq n$ and performs the PCA on the subset of size m. By taking M such randomly chosen subsets, an empirical distribution for the estimate of the eigenvalue λ_i may be constructed.

4.4.1 Using the Principal Components

In the first example in the Lab (the 'Sparrow' data set), we see that by far the most of the variation is accounted for by the first two principal components, so these two components should be useful for most analysis of the data; the other three components should not add much.

In the lab, we consider the question of whether all 5 quantitative variables are required to show differences between the two groups of birds.

4.5 Weighted Projection Methods

Let $\mathbf{x}$ be the $n \times p$ data matrix, corresponding to n p-variate observations $\underline{x}_1, \ldots, \underline{x}_n$. Let $\underline{y}_1, \ldots, \underline{y}_n$ denote the corresponding n points obtained by projecting onto a q dimensional subspace of the object space. The following properties characterise the q dimensional subspace found by PCA.

1. The points $\underline{x}_1, \ldots, \underline{x}_n$ are projected perpendicularly onto $\underline{y}_1, \ldots, \underline{y}_n$.

2. The data points $\underline{y}_1, \ldots, \underline{y}_n$ have the greatest variance among standardised q dimensional subspace projections.

The points are those in the q dimensional space that minimise:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} (d_{ij} - \hat{d}_{ij})^2$$

where

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}, \qquad \hat{d}_{ij} = \sqrt{\sum_{k=1}^{p}(y_{ik} - y_{jk})^2}.$$

A disadvantage of the constraint that $\underline{y}_1, \ldots, \underline{y}_n$ have the largest possible variation is that observations close to the centre in the projected space may be far from the centre in the higher dimensional space. A better quantity to minimise is

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\omega_{ij}(d_{ij} - \hat{d}_{ij})^2$$

where ω_{ij} controls the accuracy of the comparisons. For example, take $\omega_{ij} = d_{ij}$ if accurate representation of large distances is required and $\omega_{ij} = \frac{1}{d_{ij}}$ if accurate representation of small distances is required.

4.6 Factor Analysis

As usual with descriptive statistics, 'Var' represents a *statistical* variance and 'Cov' represents *statistical* covariance; the terms refer to the statistics computed from the data and not to any features of the population distribution.

Factor analysis may be seen as an extension of Principal Component Analysis. Given p variables $X_1, \ldots, X_p$, it is hoped that they can be expressed, or mostly expressed, by a reduced number of *factors*, which are linear combinations of the variables. Based on the original variables, it is hoped that these factors may have an interpretation.

Suppose there are n observations, $(x_{j1}, \ldots, x_{j_p})_{j=1}^{n}$ of the p variables. One starts by applying a *principal component analysis* on the *correlation* (that is, on the standardised data). A suitable value of m is chosen and principal components after level m are neglected. Let $Y_1, \ldots, Y_p$ denote the principal components, with corresponding eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p$. Suppose that the data has been standardised. Then $Y = P^t X$. In co-ordinates,

$$Y_k = P_{1k}X_1 + \ldots + P_{pk}X_p, \qquad k = 1, \ldots, p.$$

These linear combinations of the variables, as discussed earlier, are statistically uncorrelated. Now, choose m, the number of factors in the model. As discussed earlier, there are two usual methods; either let m equal the number of eigenvalues greater than or equal to 1 (Kaiser's method) or else let m denote the lowest number of eigenvalues that account for more than 80% of the variation.

Recall that P is orthonormal and hence $P^{-1} = P^t$. It follows that $X = PY$. In co-ordinates,

$$X_j = P_{j1}Y_1 + \ldots P_{jp}Y_p, \qquad j = 1, \ldots, p.$$

Set $F_j = \frac{Y_j}{\sqrt{\lambda_j}}$ for $j = 1, \ldots, m$, then the $(F_j)_{j=1}^m$ are uncorrelated and $\text{Var}(F_j) = 1$ for each $j = 1, \ldots, m$.

Let

$$A_{jk} = \sqrt{\lambda_k} P_{jk} \qquad j = 1, \ldots, p, \qquad k = 1, \ldots m.$$

Let $\epsilon_a = \sum_{k=m+1}^p A_{ak} Y_k$. Then, for $j = 1, \ldots, p$,

$$X_j = A_{j1} F_1 + \ldots + A_{jm} F_m + \epsilon_j, \qquad j = 1, \ldots, p.$$

The $F_1, \ldots, F_m$ are uncorrelated factors with $\text{Var}(F_j) = 1$ for all $j = 1, \ldots, m$.

Definition 4.5 (Specificity)**.** *The quantity $Var(\epsilon_a)$ is known as the* specificity *of X_a, the part of the variance that is unrelated to the common factors.*

The elements $A_{a1}, \ldots, A_{am}$ are known as the provisional *factor loadings* for variable a.

Definition 4.6 (Factor Loadings)**.** *Once the errors $\epsilon_1, \ldots, \epsilon_p$ have been determined, along with the factors $F_1, \ldots, F_m$ to be used, the* factor loadings *for the factor a are the coefficients $A_{a1}, \ldots, A_{am}$ such that*

$$X_a = \sum_{j=1}^m A_{aj} F_j + \epsilon_a.$$

Definition 4.7 (Communality)**.** *The* communality *of a variable X_j in a factor analysis is defined as $\sum_{k=1}^m A_{jk}^2$, where m is the number of factors. It gives the correlation between X_j and the part of X_j explained by the factors.*

An orthonormal transformation of uncorrelated variables yields uncorrelated variables. Therefore, any orthonormal transformation D yielding factors F^* given by

$$\mathbf{F}^* = D\mathbf{F}$$

will produce a suitable decomposition of X into uncorrelated factors. The second stage of the analysis is to find a rotation matrix D that produces rotated factors that are most convenient.

The last stage is to calculate the factor scores $(F_{j1}^*, \ldots, F_{jm}^*)$ for each observation $j = 1, \ldots, n$.

Note that the factors produced by a principal component analysis are orthogonal (i.e. uncorrelated). In the second stage, an orthonormal transformation will preserve this feature. If other transformations are used, the factors will not be independent.

The Varimax Rotation This is the transformation taken from the *orthogonal* transforms that maximises the variance of the squared loadings; that is, choose D to maximise

$$\mathcal{V} := \frac{1}{p}\sum_{l=1}^{k}\left(\sum_{j=1}^{p} A_{jl}^4 - \left(\frac{1}{p}\sum_{j=1}^{p} A_{jl}^2\right)^2\right).$$

The logic behind this is that if this is large, then each values of A_{jk} is close to either 0 or 1, so that the variable is explained as much as possible by a single factor.

Note that, by standardisation, $\mathrm{Var}(F_j) = 1$ for all j and $\mathrm{Cov}(F_j, F_k) = 0$ for $j \neq k$. If ϵ is small (as it should be if the variables are properly explained by m factors), then the correlation structure of $\mathbf{X}$ (where the variables have been standardised) is given by

$$\mathrm{Cov}(X_j, X_k) = \mathrm{Cov}\left(\sum_{a=1}^{m} A_{ja}F_a, \sum_{b=1}^{m} A_{kb}F_b\right) = \sum_{a=1}^{m} A_{ja}A_{ka}.$$

The Value of Factor Analysis

Factor analysis is often useful for gaining *qualitative* insight into the structure of multivariate data, but it should be regarded purely as a piece of *descriptive statistics*; it has no value whatsoever for formal inferential statistics. It is not appropriate if it is carried out on a single small sample that cannot be replicated and then assuming that the factors obtained must represent underlying variables. Simulations have shown that even if a postulated factor model is correct, the chance of recovering it using the available methods is not very high.

4.7 Example: Country Employment Profiles

The second exercise in the Lab is analysis of the country employment profile data. Firstly a PCA is carried out and the anlysis is continued to give a *factor analysis*.

In that example, for the standardised variables, there are four eigenvalues greater than 1 in the principal component analysis, so the 'rule of thumb' suggests that *four* factors are appropriate using (initially) $F_j = \frac{Z_j}{\sqrt{\lambda_j}}$, giving

$$X_j = A_{j1}F_1 + A_{j2}F_2 + A_{j3}F_3 + A_{j4}F_4 + \epsilon_j, \ldots j = 1, \ldots, 9.$$

It is useful if each variable can be expressed in terms of as few factors as possible. The next step is therefore to try a rotation, which keeps the factors uncorrelated. That is, $\mathbf{F}^* = \Theta\mathbf{F}$, where Θ is a rotation matrix, which tries to ensure that for each variable the *loading* is weighted as much as possible towards *one predominant* factor. The varimax seems to work quite well.

For the employment data, this yields the model (where the communality is indicated on the right)

$$X_1 = -\mathbf{0.85}F_1^* - 0.10F_2^* - 0.27F_3^* - 0.36F_4^* + \epsilon_1 \qquad 0.93$$

$$X_2 = -0.11F_1^* - 0.30F_2^* - \mathbf{0.86}F_3^* - 0.10F_4^* + \epsilon_2 \qquad 0.85$$

$$X_3 = 0.03F_1^* - 0.32F_2^* + \mathbf{0.89}F_3^* - 0.09F_4^* + \epsilon_3 \qquad 0.91$$

$$X_4 = 0.19F_1^* + 0.04F_2^*\mathbf{0.64}F_3^* + 0.14F_4^* + \epsilon_4 \qquad 0.46$$

$$X_5 = 0.02F_1^* - 0.08F_2^* + 0.04F_3^* + \mathbf{0.95}F_4^* + \epsilon_5 \qquad 0.92$$

$$X_6 = 0.35F_1^* + 0.48F_2^* + 0.15F_3^* + \mathbf{0.65}F_4^* + \epsilon_6 \qquad 0.79$$

$$X_7 = 0.08F_1^* + \mathbf{0.93}F_2^* + 0.00F_3^* - 0.01F_4^* + \epsilon_7 \qquad 0.87$$

$$X_8 = \mathbf{0.91}F_1^* + 0.18F_2^* + 0.12F_3^* + 0.04F_4^* + \epsilon_8 \qquad 0.88$$

$$X_9 = \mathbf{0.73}F_1^* - \mathbf{0.57}F_2^* + 0.03F_3^* - 0.14F_4^* + \epsilon_9 \qquad 0.87.$$

The 'varimax' rotation has conveniently expressed each variable in terms of a *predominant* factor plus other less important factors for that variable. The only variable that seems to have *two* predominant factors is X_9.

Following the varimax rotation, the results are interpreted by considering the four factors in terms of the variables. From that, it may be possible to give useful labels to each factor.

Here, it is clear that F_1^* has high positive loadings for X_1 (agriculture, forestry and fishing) and high negative loadings for X_8 (social and personal services) and X_9 (transport and communications). Therefore, F_1^* measures the extent to which people are employed in agriculture rather than services and communications. It could be labelled 'rural industries rather than social service and communication'.

Factor F_2^* turns out to have a high negative loading for X_7 (finance). The loading for X_9 (transport and communication) seems to be higher than the others. A possible labelling could be 'lack of finance industries'. □