

## Chapter 2

# Dimension reduction techniques - 3

### 2.1

### 2.2

### 2.3 Factor Analysis

#### 2.3.1

#### 2.3.2 Estimation of the factors

As before, we have a data matrix  $\mathbf{X}$ , with sample mean  $\bar{\mathbf{X}}$  and variance matrix  $S$ . We wish to find estimates  $\hat{\Lambda}$  and  $\hat{\Psi}$  such that  $S = \hat{\Lambda}\hat{\Lambda}^t + \hat{\Psi}$  is approximately satisfied. Using scale invariance, we proceed by standardizing the data so that we only deal with the correlation matrix  $R$ . Since we are interested in relationships between the variables, the scaling is often ignored and we estimate the factor loadings only from  $R$ .

The matrix  $S$  gives us an idea whether factor analysis should be useful or not. If the off-diagonal elements of  $S$  are very small, i.e. the correlation matrix is close to identity, the factor analysis will not prove useful. In that case, the specific factors will be dominant, and the common factors will be masked. In presence of large off-diagonal entries, a factor model can be entertained, and we try to estimate the factor loadings  $\Lambda$  and the specific variances  $\Psi$ . To solve

the equation approximately, we can employ either of two methods, the principal component method, and the maximum likelihood method.

### **Principal factor (component) method :**

We start with the principal component decomposition of the sample covariance matrix  $S$ . Writing  $S = GLG^t = \sum_{i=1}^p \lambda_i g_{(i)} g_{(i)}^t$  is by itself a factor decomposition with  $\hat{\Lambda} = GL^{1/2}$  and  $\hat{\Psi} = 0$ . However, there is essentially no useful reduction and interpretation available from this, since here  $k = p$ , which means the common factors do not provide any summary, and the specific factors are all identical to 0.

To rectify this, suppose we write  $\tilde{\Sigma} = \sum_{i=1}^k \lambda_i g_{(i)} g_{(i)}^t$  where  $k$  is a small specified number (the number of factors we decide to have). We write  $\hat{\Lambda} = [\sqrt{l_1}g_{(1)}, \dots, \sqrt{l_k}g_{(k)}]$ , a loading of only  $k$  factors. Writing  $\hat{\psi}_{ii} = s_{ii} - \sum_{j=1}^k \hat{\lambda}_{ij}^2$ , we have an approximation of the factor model  $S = \hat{\Lambda}\hat{\Lambda}^t + \hat{\Psi}$ . The approximation is exact at the diagonal and it can be shown that

$$\|S - (\hat{\Lambda}\hat{\Lambda}^t + \hat{\Psi})\| \leq \sum_{j=k+1}^p l_j^2$$

the right hand term is small since the PCA ensures the last eigenvalues to be tiny and can be thrown off without too much harm. The proportion of total variance explained by  $k$  factors is  $\frac{l_1 + \dots + l_k}{tr(S)}$  and gives us a good idea of how many factors to be included. In practice, it is advisable to keep only a few factors rather than many of them, since the whole idea of factor analysis is to provide some useful factors that can be interpreted and held responsible for the group of variables.

### **Maximum likelihood method :**

In the days of manual computing, the problem of maximizing the normal likelihood ( $\hat{\mu} = \bar{\mathbf{X}}$  is already replaced)

$$L(\Sigma) = C|\Sigma|^{n/2} e^{-\frac{1}{2}\Sigma^{-1}S}$$

over all  $\Sigma = \Lambda\Lambda^t + \Psi$  was impossible to solve. However, with the advent of modern computing, this can be done numerically. To ensure identifiability, we

need to impose some more conditions, e.g.  $\Lambda^t \Psi^{-1} \Lambda = \Delta$ , a diagonal matrix. This can be performed using an intensive computer program, and we leave it there. In particular, the communalities and the specific variances are estimated numerically, and can be used to deduce the number of underlying factors that need to be retained.

### 2.3.3 Factor rotation

Since the factor models are transformable under orthogonal transformations (remember  $\mathbf{x} = \tilde{\Lambda}(\Gamma^t f) + u + \mu$  where  $\tilde{\Lambda} = \Lambda \Gamma$  for some orthogonal  $\Gamma$ ), a rotation can be applied to the factor loadings, and still lead us to the original covariance structure. In essence, the factors can be viewed from a rotated position without disturbing the structure of the model. After obtaining the estimates  $\hat{\Lambda}$  and  $\hat{\Psi}$ , one can perform a rotation of the loadings as well, without altering the covariance, residuals, communalities etc. This comes in handy in case the estimated factor loadings are not readily interpretable, and we need to rotate them until we get a simple structure. By that, we strive to get a pattern where each variable loads highly on single factor, and relatively less loading on other factors. It may not be always possible to get such a simple structure. When we have 2 or 3 factors, this can be frequently determined graphically. A simple plot of the factor loadings show the necessary pattern and we can rotate the factors accordingly.

For higher dimensions, an analytical tool has been provided by Kaiser. This is known as the varimax rotation and tries to maximize a quantity which reflects the sum of the variances of squared factor loadings. In essence, it tries to maximize the variances of the factor loadings among all possible rotations of them. We also need to scale the factor loadings by their communalities to obtain simple structures. All modern softwares employ this varimax rotations while computing the factor loadings. Effectively, maximizing the variances in the factor loadings correspond to spread out the loadings in each factor, which in turn will give rise to one large factor and others close to zero. (To get the

basic idea, see this. Suppose we want to maximize  $\sum x_i^2$  subject to  $\sum x_i = c$  and  $x_i \geq 0$ . Then, the solution is  $x_1 = c$  and the rest are 0). Let,

$$V = \sum_{j=1}^m [\sum_{i=1}^p \lambda_{ij}^4 - (\sum_{i=1}^p \lambda_{ij}^2)^2/p]$$

where  $\lambda_{ij}$  is the factor loading corresponding to  $j$ th factor and  $i$ th variable. A scaling is also advisable such as  $\lambda_{ij}^* = \lambda_{ij}/h_i$  where  $h_i$  is the square root of the communality.