

Chapter 2

Dimension reduction techniques - 5

2.1

2.2

2.3 Factor Analysis

2.3.1

2.3.2

2.3.3

2.3.4 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 Γ), 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,

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

where λ_{ij} is the factor loading corresponding to j .

2.3.5 Factor scores

Sometimes, the estimated values of the common factors, namely \hat{f} are quantities of interest. Usually, since f is an unobserved variable, rather than a parameter, the estimates are similar to the error estimates in usual regression. They are used for diagnostic purposes. However, the estimation is not easy because they, along with u , outnumber the observed x -s. Usually, we treat the factor loadings Λ and specific variances Ψ as known and replace them by their estimates. The sum of squares of the errors, weighted by their variances is then,

$$\sum_{i=1}^p \frac{u_i^2}{\psi_{ii}} = (\mathbf{x} - \boldsymbol{\mu} - \Lambda f)^t \Psi^{-1} (\mathbf{x} - \boldsymbol{\mu} - \Lambda f)$$

We seek to minimize this quantity at an observed value \mathbf{x}_j and call the corresponding factor score as \hat{f}_j . Replacing μ by $\bar{\mathbf{X}}$, Λ and Ψ by their estimates, we obtain the estimated j th factor score as

$$\hat{f}_j = (\Lambda^t \hat{\Psi} \Lambda)^{-1} \Lambda^t \hat{\Psi} (\mathbf{x}_j - \bar{\mathbf{x}})$$

We need to also use the uniqueness condition $\Lambda^t \Psi \Lambda = \Delta$ to use their estimates.

Another method is to use a regression of f on x -s. Observe that, $\text{var} \begin{pmatrix} \mathbf{x} \\ f \end{pmatrix} = \begin{pmatrix} \Sigma & \Lambda \\ \Lambda^t & I \end{pmatrix}$, yielding $\mathbb{E}(f|\mathbf{x}) = \Lambda^t \Sigma^{-1} (\mathbf{x} - \mu)$. Replacing the parameters by their estimates, we get a regression estimate of f , namely

$$\hat{f}_j = \hat{\Lambda}^t S^{-1} (\mathbf{x}_j - \bar{\mathbf{x}})$$

Remarks : Though we have learned some methods in factor analysis, some basic notions should be kept in mind.

- One should perform both principal factor analysis and maximum likelihood factor analysis, and see whether they give rise to similar factor loadings (include varimax rotation in each case).
- The choice of m should be guided by the principal of Occam's razor, and used to keep as few factors as possible for useful interpretation.
- Sometimes, large data sets can be split in a few parts and the factor analysis done on them separately. the final analysis should be a combination of those partial analysis.

2.4 Canonical correlation

As we mentioned briefly in Chapter ??, the CCA seeks to identify and quantify the associations between two sets of variables. Examples include relating Government policy variables with economic goal variables and associating college performances of a group of students with their pre-college performances etc. Here, we focus on the correlation between a linear combination of one set

of variables to a linear combination in another set. We first seek to maximize this correlation. Then, we seek to find a pair of linear combinations uncorrelated with the initial pair and having the largest correlation among those, and so on. This is analogous to what we do in PCA in spirit, but the main focus is on a correlation. They will be called the canonical variables and the successive correlations as canonical correlations. The technique attempts to concentrate on a high-dimensional relationship between two sets of variables into a few pairs of canonical variables.

As before, we have $\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}$ a $(p+q)$ dimensional random vector with two sets of p and q variables. We assume $p \leq q$ throughout. We assume $\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}$ has mean zero and its variance is partitioned as $\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. Clearly, Σ_{12} , the $p \times q$ dimensional matrix measures the association between the two sets. The main task of CCA is to summarize the association between \mathbf{x} and \mathbf{y} using a few linear combinations and their correlations.

Let $U = a^t \mathbf{x}$ and $V = b^t \mathbf{y}$ be a typical pair of linear combinations of \mathbf{x} and \mathbf{y} . Then,

$$\text{Corr}(U, V) = \frac{a^t \Sigma_{12} b}{\sqrt{a^t \Sigma_{11} a} \sqrt{b^t \Sigma_{22} b}}$$

The first pair of canonical variables (U_1, V_1) will be the pair with unit variance and maximum correlation. The second pair will be the pair (U_2, V_2) with unit variance and maximum correlation among all choices uncorrelated with (U_1, V_1) and so on. The k th pair will be uncorrelated with the first $k-1$ pairs.

Result : Define $\Psi_1 = \Sigma_{11}^{-1/2} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1/2}$ and $\Psi_2 = \Sigma_{22}^{-1/2} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1/2}$.

Then, they have the same set of positive eigenvalues. We'll call them $\tilde{\rho}_1^2 \geq \dots \geq \tilde{\rho}_p^2 \geq 0$. Let (e_1, \dots, e_p) and (f_1, \dots, f_p) be the corresponding eigenvectors of Ψ_1 and Ψ_2 . Namely, $\Psi_1 = \sum_{i=1}^p \tilde{\rho}_i^2 e_i e_i^t$ and $\Psi_2 = \sum_{i=1}^p \tilde{\rho}_i^2 f_i f_i^t$. Then, $U_i = e_i^t \Sigma_{11}^{-1/2} \mathbf{x}$ and $V_i = f_i^t \Sigma_{22}^{-1/2} \mathbf{y}$ are the canonical variables and $\text{Corr}(U_i, V_i) = \tilde{\rho}_i$. e_i -s and f_i -s are related by the fact that each f_i is pro-

portional to $\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1/2}e_i$.

If we standardize the variables \mathbf{x} and \mathbf{y} beforehand, they will also give the same canonical correlations and the same canonical variables. In fact, the variables $\Sigma_{11}^{-1/2}\mathbf{x}$ and $\Sigma_{22}^{-1/2}\mathbf{y}$ will also yield the same set. So, it is clear that the correlations within the two groups do not effect the canonical correlation. Any non-singular linear transformations of \mathbf{x} and \mathbf{y} will share the same property. That is, the CCA is invariant under any rotation, reflection or amplification of the original variables within the two sets.

Interpretation : The interpretations of canonical variables are primarily in terms of the standardized variables. This variables can be identified in terms of the original variables. The correlation between the canonical variables and the original variables can be used to determine which variable contributes most to the relationship between the two sets. Moreover, CCA generalizes other correlation methods such as the two-variable product-moment correlation and multiple correlation. The quantity $\tilde{\rho}_k^2$ gives us the proportion of variance of the canonical variable U_k explained by the set \mathbf{y} and called the k th shared variance between the two sets. The largest value $\tilde{\rho}_1^2$ is a measure of the set overlap.

In contrary to PCA, there is no certainty that the first few canonical variables may capture the variability of the individual sets of variables. It may happen as well, that by leaving out the smaller eigenvalues and associated canonical variables, we actually lose a lot of features of the data. Unfortunately, that is an artifact of the CCA.

Sample canonical correlations : As before, we have a sample of n realizations on $p + q$ variables $\begin{pmatrix} \mathbf{X}_i \\ \mathbf{Y}_i \end{pmatrix}$ and construct the sample covariance matrix $\begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$. The linear combinations with maximum correlations from those two sets of realizations can be computed using the similar eigenvalue technique. However, there is no certainty that the canonical variables will

represent the individual variables in terms of variability and special features and the inference should be restricted to the measure of association between the two sets of variables only.