ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Factor Analysis

Soel Micheletti

June 2022

Department of Computer Science, ETH Zürich

In a nutshell, the basic idea of factor analysis is the following: we have p covariates, and we assume that these are driven by m common unobserved factors [1]. The goal is recovering these unobserved factors. Factor analysis is particularly useful (but not limited) to the case where $m << p$. There are many reasons why factor analysis is interesting. A non-exhaustive list is the following:

- We are in a very high-dimensional setting (i.e. p is very large for the given problem).
- Factor analysis is a possible candidate algorithm if we want to perform dimensionality reduction.
- We assume that the unobserved latent factors are more interesting than the observed quantitative measurements.

This seems pretty similar as PCA. Despite the fact that the two methods are connected, they have different goals. PCA finds the m-dimensional representation of the p-dimensional data that minimizes the reconstruction error. Factor analysis groups the p variables in m groups, where all the variables belonging to the same group are considered driven by the corresponding latent factor.

The setting

The p covariates of each samples are collected in a vector $\mathbf{X} \in \mathbb{R}^p$. The random vector $\mathbf{X}$ has mean $\mu \in \mathbb{R}^p$, i.e. we assume that $\mathbb{E}[X_i] = \mu_i$. We want to find meaningful factors $\mathbf{f} \in \mathbb{R}^m$ such that

$$\mathbf{X} = \mu + \mathbf{L}\mathbf{f} + \varepsilon$$

where $\mathbf{L} \in \mathbb{R}^{p \times m}$ is called *matrix of factor loadings* and the error vector $\varepsilon \in \mathbb{R}^p$ is called *specific factor*. Note that, if we get a new p-dimensional sample, we consider $\mathbf{L}$ fixed and we solve a system of equations to recover the corresponding low-dimensional representation $\mathbf{f}_i$.

Similarly as in the well-studied linear regression framework, we need to impose some assumptions. Common assumptions which can (partially) be relaxed include:

1. The noise has zero mean, i.e. $\mathbb{E}[\varepsilon_i] = 0$ for $i \in [p]$.
2. The common factors have mean zero, i.e. $\mathbb{E}[\mathbf{f}_i] = 0$ for $i \in [m]$.

[1] More formally, we assume that each of the p observed variables is a function (up to random noise) of (one of) the latent factors

3. The common factors have homoscedastic variance (which can be normalized to one).
4. $Var[\varepsilon_i] = \psi_i$.
5. The common factors are uncorrelated, i.e. $Cov[\mathbf{f}_i, \mathbf{f}_j] = 0$ for all $i \neq j$.
6. $Cov[\varepsilon_i, \varepsilon_j] = 0$ for all $i \neq j$.
7. $Cov[\varepsilon_i, \mathbf{f}_j] = 0$ for all $i \in [p]$, $j \in [m]$.

Given the assumptions above, one can easily derive the following properties:

- $Var[\mathbf{X}_i] := \sigma_i^2 = \sum_{j=1}^{m} l_{ij}^2 + \psi_i$
- $Cov[\mathbf{X}_i, \mathbf{X}_j] := \sigma_{ij} = \sum_{k=1}^{m} l_{ik} l_{jk}$
- $Cov[\mathbf{X}_i, \mathbf{f}_j] = l_{ij}$
- We can write the variance-covariance of the model as

$$\Sigma = \mathbf{L}\mathbf{L}^T + \psi$$

where $\psi \in \mathbb{R}^{p \times p} = diag(\psi_i)$.

Interpretation of the results

We define the communalities h_i^2 for $i \in [p]$ as

$$h_i^2 = \sum_{j=1}^{p} l_{ij}^2$$

We can interpret the communalities as the proportion of variance explained by the m factors for the corresponding variable. Note that this is different than the proportion of total variation explained by the m factors, which can be computed as

$$\frac{\sum_{i=1}^{p} h_i^2}{p}$$

Recall that the main goal of factor analysis is discerning some underlying factors describing the data. By running some factor analysis algorithm (such as MLE, we omit the details) it could happen that some variables have large loading factors for multiple factors. This is undesirable, as we wish to group different variables together and consider them driven by the same factor. Rotation are a very useful tool towards this goal.

First, note that the equation $\mathbf{X} = \mu + \mathbf{Lf} + \varepsilon$ has multiple solutions. More specifically, given a solution $\mathbf{L}$, $\mathbf{f}$, we can derive other solutions. Concretely, we compute $\mathbf{X} = \mu + \mathbf{L}^*\mathbf{f}^* + \varepsilon$ with $\mathbf{L}^* = \mathbf{LT}$, $\mathbf{f}^* = \mathbf{T}^T\mathbf{f}$, and $\mathbf{T}$ an appropriate orthogonal matrix. We omit the details, but a common choice is the Varimax rotation. Varimax ensures that per factor certain variables load as high as possible and the other variables load as low as possible. In other words, the factor loading is high for exactly one factor for each variable. This is obtained when the variance of the factor charges per factor is constrained to be as high as possible.