STA 437/2005: Methods for Multivariate Data

Week 10: Factor Analysis

Piotr Zwiernik

University of Toronto

Low-dimensional structures in multivariate statistics

Discussing covariance matrix estimation we mentioned some special structures that are routinely assumed in multivariate statistics.

We will discuss in detail two such constraints:

- $\Sigma = L + S$ where L is low-rank and S is sparse.
- the inverse of Σ^{-1} is sparse.

In Factor Analysis: $\Sigma = WW^\top + \Psi$ with $W \in \mathbb{R}^{m \times r}$, Ψ diagonal.

We explain here how such structure can occur by discussing motivating examples.

Factor Analysis: Motivating Examples

Example: Capital Asset Pricing Model (CAPM)

Models stock returns based on a common factor; the market return. For each stock:

$$X_i = \mu_i + w_i Z + \varepsilon_i$$

This is one of the most basic models in finance.

Example: Human Intelligence

Cognitive abilities modeled by latent intelligence factor.

This could be further generalized to account for multiple types of intelligence.

Factor Analysis Model

The model assumes the following stochastic representation of $X = (X_1, \dots, X_m)$:

$$X = \mu + WZ + \varepsilon, \qquad Z \sim N_r(0, I_r), \quad \varepsilon \sim N_m(0, \Psi), \quad Z \perp\!\!\!\perp \varepsilon,$$

where Ψ is a diagonal covariance matrix.

The latent factors Z

As the two examples suggest, often in this context Z has a specific interpretation.

In PPCA we have the same representation with $\Psi = \sigma^2 I_m$ (isotropic noise).

In FA more emphasis on interpreting the latent factors.

Parametrization and identifiability

$X = \mu + WZ + \varepsilon$ is Gaussian with the induced covariance structure:

$$\Sigma = WW^T + \Psi.$$

Lack of identifiability

As for PPCA, W is not uniquely identified.

- Replacing W with WU for $U \in O(m)$ does not change the distribution.
- This has important consequences for model intepretability.

Dealing with non-uniqueness of W

Approach 1: Constraint W so that $W^T\Psi^{-1}W$ diagonal.

- Multiply $X = \mu + WZ + \varepsilon$ by $\Psi^{-1/2}$ to get $\tilde{X} = \tilde{\mu} + \tilde{W}Z + \tilde{\varepsilon}$ with $\tilde{\varepsilon} \sim N(0, I_m)$.
- We have $W^T\Psi^{-1}W = \tilde{W}^\top\tilde{W}$ so this corresponds to orthogonality of the columns of $\tilde{W}^\top$.

Approach 2: Apply varimax rotation for interpretability.

- Consider any $\widehat{W} \in \mathbb{R}^{m\times r}$. We find U such that $\widehat{W}U$ more interpretable.
- Define $M \in \mathbb{R}^{m\times r}$ by $M_{ij} = \frac{(WU)_{ij}^2}{\sum_{k=1}^r (WU)_{ij}^2}$ then find the appropriate U maximizing:

$$\|M - \tfrac{1}{m}\mathbf{1}_m\mathbf{1}_m^\top M\|_F^2.$$

- This results with solutions such that each column of M has a bunch of big entries and the remaining ones are negligible.

Fitting the factor analysis model

Data: $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from the model $X = \mu + WZ + \varepsilon$, $\varepsilon \sim N(0, \Psi)$.

The most canonical way to estimate the parameters is via the maximum likelihood.

The MLE is not given in a closed form.

- We could use the EM algorithm.

Alternatively use the fact that MLE has closed form if $\Psi = \sigma^2 I_m$.

Suppose Ψ is known.

Denote $\tilde{X} = \Psi^{-1/2} X$, $\tilde{\mu} = \Psi^{-1/2}\mu$, $\tilde{W} = \Psi^{-1/2} W$ and $\tilde{\varepsilon} = \Psi^{-1/2}\varepsilon \sim N(0, I_m)$

$$\tilde{X} = \tilde{\mu} + \tilde{W}Z + \tilde{\varepsilon}. \quad \text{(PPCA with } \sigma^2 = 1\text{)}$$

Define $\tilde{S}_n = \Psi^{-1/2} S_n \Psi^{-1/2}$ with spectral decomposition $\tilde{S}_n = U \tilde{\Lambda} U^\top$. The MLE:

$$\widehat{W} \ = \ U_r \Theta R,$$

where:

- R is *any* orthogonal matrix, U_r first r columns of U,
- Θ is a diagonal matrix with i-th entry equal to $\sqrt{\max\{0, \tilde{\lambda}_i - 1\}}$.

We can now apply this iteratively, where the update on Ψ

Choosing the Number of Factors

Determining the number of latent factors r is critical in factor analysis.

Overestimating r leads to overfitting, underestimating r leads to loss of structure.

There are several common method. We focus on Horn's Parallel Analysis (PA).

Key ideas of Horn's Parallel Analysis

If no latent signal, the sample **correlation** matrix should resemble the identity matrix.

Depending on (n, m) the actual eigenvalues can still be far from 1.

PA compares eigenvalues of observed data with those obtained from Monte Carlo simulations of purely random noise.

The algorithm

1. Compute eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ of sample correlation matrix R_n, where

$$R_n = D^{-1/2} S_n D^{-1/2}. \tag{1}$$

2. Generate B simulated datasets $X^{(b)} \in \mathbb{R}^{n \times m}$ from $\mathcal{N}(0,1)$.
3. Compute sample correlation matrices $R_n^{(b)}$ for each simulated dataset.
4. Compute average null eigenvalues:

$$\lambda_j^{\text{random}} = \frac{1}{B} \sum_{b=1}^{B} \lambda_j^{(b)}, \quad j = 1, \ldots, m. \tag{2}$$

5. Retain factors where

$$\lambda_j > \lambda_j^{\text{random}}. \tag{3}$$

Application: Factor Analysis on Personality Traits Data

We analyze real survey data from the `bfi` dataset in the `psych` R package.

The dataset consists of 2,800 responses to 25 personality-related questions.

These questions measure the Big Five Personality Traits:

- Neuroticism (N)
- Extraversion (E)
- Conscientiousness (C)
- Agreeableness (A)
- Openness (O)

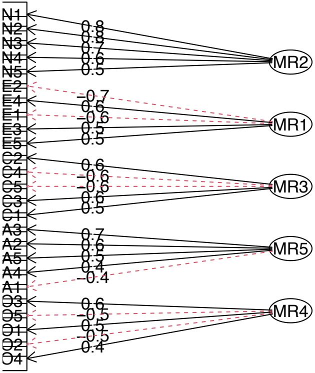
Responses are on the 1-6 scale indicating agreement strength.

We discard demographic variables (gender, age, education).

Factor Analysis Results

PA suggests retaining **five** factors, matching the Big Five personality traits (nice!).

Factor loadings after varimax rotation confirm that the extracted factors correspond to the expected latent traits.



Summary

Factor Analysis is a popular method in multivariate statistics.

It is similar to PPCA and it has clear motivating examples.

The lack of identifiability creates a chalange in the interetation of factor loadings.

Choosing the number of factors (if there is no clear insight) may be also hard.

- Horn's Parallel Analysis is a simple solution that tends to perform well in practice.

The resulting form of the covariance matrix $\Sigma = WW^\top + \Psi$ can be exploited and generalized in many creative ways.