Dimensionality Reduction

Ziping Zhao

School of Information Science and Technology
ShanghaiTech University, Shanghai, China

CS182: Introduction to Machine Learning (Fall 2022)
http://cs182.sist.shanghaitech.edu.cn

Ch. 6 of I2ML (Secs. 6.4, 6.6, and 6.12 – 6.13 excluded)

Outline

Factor Analysis

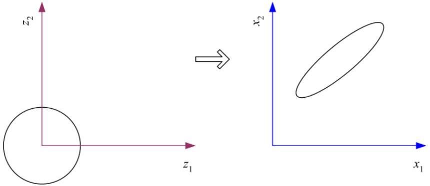
- Factor analysis (FA) or exploratory factor analysis (EFA) assumes that there is a set of latent factors z_j, $j = 1, \ldots, k$, which when acting in combination generate the observed variables $\mathbf{x}$.
- The goal of FA is to characterize the dependency among the observed variables by means of a smaller number of factors, i.e., in a smaller dimensional space without loss of information measured as the correlation between variables.
- Problem settings:
  - Sample $\mathcal{X} = \{\mathbf{x}^t\}$: drawn from some unknown probability density with $\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$ and $\mathrm{Cov}(\mathbf{x}) = \boldsymbol{\Sigma}$.
  - Factors z_j are unit normals and uncorrelated: $\mathbb{E}[z_j] = 0$, $\mathrm{Var}(z_j) = 1$, $\mathrm{Cov}(z_i, z_j) = 0$, $i \neq j$.
  - Noise sources ϵ_i to explain what is not explained by the factors: $\mathbb{E}[\epsilon_i] = 0$, $\mathrm{Var}(\epsilon_i) = \psi_{ii} = \psi_i^2$, $\mathrm{Cov}(\epsilon_i, \epsilon_j) = \psi_{ij} = 0$, $i \neq j$, $\mathrm{Cov}(\epsilon_i, z_j) = 0$, $\forall i, j$

Relationships Between Factors and Input Dimensions

- Each of the d input dimensions x_i, $i = 1, \ldots, d$, can be expressed as a weighted sum of the k $(< d)$ factors z_j, $j = 1, \ldots, k$, plus some residual error term:

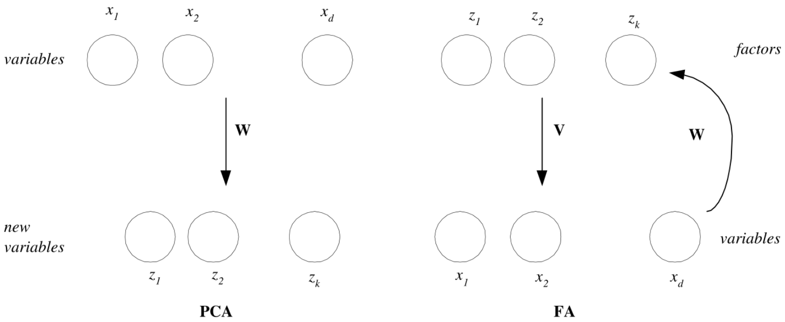
$$x_i - \mu_i = \sum_{j=1}^{k} v_{ij} z_j + \epsilon_i \quad \text{or} \quad \mathbf{x} - \boldsymbol{\mu} = \mathbf{V}\mathbf{z} + \boldsymbol{\epsilon}$$

  where $\mathbf{V} \in \mathbb{R}^{d \times k}$ is a matrix of weights, called factor loadings.
  - Without loss of generality, we can assume that $\boldsymbol{\mu} = \mathbf{0}$.
- The factors z_j are independent unit normals that are stretched, rotated, and translated to generate the inputs $\mathbf{x}$.



PCA vs. FA

- The target of FA is opposite to that of PCA:
  - PCA (from $\mathbf{x}$ to $\mathbf{z}$):
$$\mathbf{z} = \mathbf{W}^T(\mathbf{x} - \mu)$$
  - FA (from $\mathbf{z}$ to $\mathbf{x}$ – generative model):
$$\mathbf{x} - \mu = \mathbf{V}\mathbf{z} + \epsilon$$



Covariance Matrix

- Given that $\text{Var}(z_j) = 1$ and $\text{Var}(\epsilon_j) = \psi_i^2$,

$$\text{Var}(x_i) = \sum_{j=1}^{k} v_{ij}^2 \text{Var}(z_j) + \text{Var}(\epsilon_i) = \sum_{j=1}^{k} v_{ij}^2 + \psi_i^2$$

where the first part $\sum_{j=1}^{k} v_{ij}^2$ is the variance explained by the common factors and the second part ψ_i^2 is the variance specific to x_i. Similarly, for $i \neq i'$, we have

$$\text{Cov}(x_i, x_{i'}) = \sum_{j=1}^{k} v_{ij} v_{i'j}$$

- Then, the covariance matrix:

$$\boldsymbol{\Sigma} = \text{Cov}(\mathbf{x}) = \text{Cov}(\mathbf{Vz} + \boldsymbol{\epsilon}) = \text{Cov}(\mathbf{Vz}) + \text{Cov}(\boldsymbol{\epsilon}) = \mathbf{V}\text{Cov}(\mathbf{z})\mathbf{V}^T + \boldsymbol{\Psi} = \mathbf{V}\mathbf{V}^T + \boldsymbol{\Psi}$$

where $\boldsymbol{\Psi} = \text{diag}(\boldsymbol{\psi})$ with $\boldsymbol{\psi} = [\psi_1^2, \ldots, \psi_d^2]$.

2-Factor Example for Illustration – I

- Let

$$\mathbf{x} = [x_1, x_2, x_3]^T \qquad \mathbf{V} = \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \\ v_{31} & v_{32} \end{bmatrix} \qquad \mathbf{z} = [z_1, z_2]^T$$

- Since

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} = \mathbf{V}\mathbf{V}^T + \mathbf{\Psi} = \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \\ v_{31} & v_{32} \end{bmatrix} \begin{bmatrix} v_{11} & v_{21} & v_{31} \\ v_{12} & v_{22} & v_{32} \end{bmatrix} + \begin{bmatrix} \psi_1^2 & 0 & 0 \\ 0 & \psi_2^2 & 0 \\ 0 & 0 & \psi_3^2 \end{bmatrix}$$

  we have

$$\sigma_{12} = \mathrm{Cov}(x_1, x_2) = v_{11}v_{21} + v_{12}v_{22}$$

- If x_1 and x_2 have high covariance, then they are related through a factor:
  - If it is the first factor, then v_{11} and v_{21} will both be high.
  - If it is the second factor, then v_{12} and v_{22} will both be high.
- If x_1 and x_2 have low covariance, then they depend on different factors:
  - In each of the products $v_{11}v_{21}$ and $v_{12}v_{22}$, one term will be high and the other low.

2-Factor Example for Illustration – II

- Because

$$\begin{aligned}\mathrm{Cov}(x_1, z_1) =& \mathrm{Cov}(v_{11}z_1 + v_{12}z_2 + \epsilon_1, z_1) \\ =& \mathrm{Cov}(v_{11}z_1, z_1) = v_{11}\mathrm{Var}(z_1) = v_{11}\end{aligned}$$

  and similarly,

$$\begin{aligned}&\mathrm{Cov}(x_1, z_2) = v_{12} \\ &\mathrm{Cov}(x_2, z_1) = v_{21} \quad \mathrm{Cov}(x_2, z_2) = v_{22} \\ &\mathrm{Cov}(x_3, z_1) = v_{31} \quad \mathrm{Cov}(x_3, z_2) = v_{32}\end{aligned}$$

  so we have

$$\mathrm{Cov}(\mathbf{x}, \mathbf{z}) = \mathbf{V}$$

  i.e., the factor loadings $\mathbf{V}$ represent the covariances (or correlations) between the variables and the factors.

Factor Analysis

- Since
$$\mathbf{\Sigma} = \mathbf{V}\mathbf{V}^T + \mathbf{\Psi}$$
if there are only a few factors (i.e., $k \ll d$), we can get a simplified structure for $\mathbf{\Sigma}$.
- The number of parameters is reduced from $d(d+1)/2$ (for $\mathbf{S}$) to $dk + d$ (for $\mathbf{V}\mathbf{V}^T + \mathbf{\Psi}$).
- Special cases:
  - Probabilistic PCA (PPCA): $\mathbf{\Psi} = \psi^2\mathbf{I}$ (i.e., all ψ_i^2 are equal)
  - Conventional PCA: $\mathbf{\Psi} = \mathbf{0}$, (i.e., $\psi_i^2 = 0$)
- The solution of factor loadings are not unique
$$\mathbf{V}\mathbf{V}^T = \mathbf{V}\mathbf{T}\mathbf{T}^T\mathbf{V}^T = (\mathbf{V}\mathbf{T})(\mathbf{V}\mathbf{T})^T = \tilde{\mathbf{V}}\tilde{\mathbf{V}}^T$$
for any orthogonal matrix $\mathbf{T} \in \mathbb{R}^{k \times k}$.
- The factors can be rotated to give maximum loading on as few factors as possible for each variable, to make the factors interpretable, for knowledge extraction.

Estimation of FA – I

- Given $\mathbf{S}$ as the estimator of $\boldsymbol{\Sigma}$, we want to find $\mathbf{V}$ and $\boldsymbol{\Psi}$ such that

$$\mathbf{S} = \mathbf{V}\mathbf{V}^T + \boldsymbol{\Psi} \qquad (\text{or}: \ \mathbf{S} \approx \mathbf{V}\mathbf{V}^T + \boldsymbol{\Psi})$$

- A naive method is to obtain $\mathbf{V}$ firstly via PCA and then $\boldsymbol{\Psi}$ by taking directly the residual's sample variance.
- A joint estimation method over $\mathbf{V}$ and $\boldsymbol{\Psi}$ can also be chosen to minimize

$$\begin{aligned} \underset{\mathbf{V},\boldsymbol{\Psi}}{\text{minimize}} \quad & \|\mathbf{S} - (\mathbf{V}\mathbf{V}^T + \boldsymbol{\Psi})\|_F^2 \\ \text{subject to} \quad & \boldsymbol{\Psi} \succ \mathbf{0} \end{aligned}$$

Estimation of FA – II

- The MLE for FA directly learns the parameters from raw data $\mathbf{x}^t$.
- It assumes that the data are generated from a certain statistical model, typically the multivariate Gaussian distribution.
- Then the parameters are estimated by maximizing the likelihood function

$$
\begin{aligned}
\underset{\boldsymbol{\mu},\mathbf{V},\boldsymbol{\Psi}}{\text{minimize}} \quad & \frac{N}{2}\log\det(\boldsymbol{\Sigma}) + \frac{1}{2}\sum_t (\mathbf{x}^t - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}^t - \boldsymbol{\mu}) \\
\text{subject to} \quad & \boldsymbol{\Sigma} = \mathbf{V}\mathbf{V}^T + \boldsymbol{\Psi} \\
& \boldsymbol{\Psi} \succ \mathbf{0}
\end{aligned}
$$

- Computationally this process is complex.
- In general, there is no closed-form solution to this optimization problem so iterative methods are applied.

Dimensionality Reduction

- FA can be used for dimensionality reduction when $k < d$.
- For dimensionality reduction, FA offers no advantage over PCA except the interpretability of factors allowing the identification of common causes, a simple explanation, and knowledge extraction.