STAT 3690 Lecture 24

zhiyanggeezhou.github.io

Zhiyang Zhou (zhiyang.zhou@umanitoba.ca)

Mar 28, 2022

Factor analysis

- A special kind of latent variable model
  - Latent variable model: latent/unobserved variables give rise to observed data through a specified model, i.e., a regression model with unobserved covariates
- Model (population version)
$$\mathbf{Y} - \boldsymbol{\mu} = \mathbf{L}\mathbf{F} + \mathbf{E}$$
  - $\mathbf{Y} = [Y_1, \ldots, Y_p]^\top$: random & observable, $\mathbf{Y} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - $\mathbf{L} = [\ell_{ij}]_{p\times q}$: fixed & unknown, a matrix of factor loadings
  - $\mathbf{F}$: random & unobservable, q-vector of latent/common factors
  - $\mathbf{E}$: random & unobservable, p-vector of error/specific factors
  - Restrictions for identifiability
    * Common factors are of zero mean, mutually uncorrelated and standardized: $\mathbf{F} \sim (\mathbf{0}, \mathbf{I})$
    * Specific factors are centered and mutually uncorrelated and each of them affects only one entry of $\mathbf{Y}$: $\mathbf{E} \sim (\mathbf{0}, \boldsymbol{\Psi})$ with $\boldsymbol{\Psi} = \operatorname{diag}(\psi_1, \ldots, \psi_p)$
    * Common and specific factors are uncorrelated: $\operatorname{cov}(\mathbf{F}, \mathbf{E}) = \mathbf{0}$
- Covariance structure
  - $\operatorname{var}(\mathbf{Y}) = \boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\top + \boldsymbol{\Psi}$
    * $\operatorname{var}(Y_i) = \sum_{j=1}^q \ell_{ij}^2 + \psi_i$
  - $\operatorname{cov}(\mathbf{Y}, \mathbf{F}) = \mathbf{L}$
  - $\sum_{i=1}^p \ell_{ij}^2$: the variance contributed by the jth common factor

Estimating $\mathbf{L}$ and $\boldsymbol{\Psi}$

- Selection of q, i.e., the number of common factors
  - PCA stopping rule
  - Take q such that $\sum_{i,j} \ell_{ij}^2/\operatorname{tr}(\mathbf{S})$ is over a preset percentage
    * $\sum_{i=1}^q \ell_{ij}^2/\operatorname{tr}(\mathbf{S})$: the proportion of variation explained by the jth common factor
  - Take q as the number of positive eigenvalues of $\mathbf{S}$
  - Take q as the number of eigenvalues of $\mathbf{S}$ that are above average
  - Take q as the number of eigenvalues greater than one for the correlation matrix
  - According to domain-knowledge expertise
- PC method
  1. Perform eigendecomposition on $\mathbf{S} = \sum_{j=1}^p \lambda_j w_j w_j^\top$
  2. Pick up q largest eigenvalues $\lambda_1, \ldots, \lambda_q$ and corresponding eigenvectors $w_1, \ldots, w_q$
  3. $\widehat{\mathbf{L}} = [\sqrt{\lambda_1} w_1, \ldots, \sqrt{\lambda_q} w_q]$ and $\widehat{\boldsymbol{\Psi}} = \operatorname{diag}(\mathbf{S} - \widehat{\mathbf{L}}\widehat{\mathbf{L}}^\top)$

- Exercise: `psych::bfi` involves 2800 subjects. For each of them, 25 personality assessments, as well as gender, education and age, are included.

```r
install.packages(c('psych'))
library(psych)
library(tidyverse)
head(psych::bfi)
data = bfi %>%
  select(-gender, -education, -age) %>%
  filter(complete.cases(.)) # Remove demographic variable and keep complete data
S = cov(data)
decomp = prcomp(data) # decompose the covariance matrix

# PCA stopping rule
(q = which(cumsum(decomp$sdev^2)/sum(decomp$sdev^2)>.9)[1])
# the overall proportion of variation explained by common factors
(q = which(
  cumsum(sort(colSums((decomp$rotation %*% diag(decomp$sdev))^2), decreasing = T))/sum(diag(S))>.9
)[1])
# the number of eigenvalues above the average
(q = sum(eigen(S, only.values = T)$values > mean(eigen(S, only.values = T)$values)))
# the number of eigenvalues greater than one for the correlation matrix
(q = sum(eigen(cor(data))$values > 1))

L_pc = decomp$rotation[,1:q] %*% diag(decomp$sdev[1:q])
Psi_pc = diag(diag(S - tcrossprod(L_pc)))

S_pc = tcrossprod(L_pc) + Psi_pc
lattice::levelplot(S - S_pc) # fitting error
lattice::levelplot((S - S_pc)/S) # difference in percentage
```