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)
$$\mathbf{Y} - \boldsymbol{\mu} = \mathbf{L}\mathbf{F} + \mathbf{E}$$
  - $\mathbf{Y} = [Y_1, \dots, Y_p]^\top$: 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, latent vector/common factors
  - $\mathbf{E}$: random & unobservable, error vector/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 $\boldsymbol{\Psi} = \mathrm{diag}(\psi_1, \dots, \psi_p)$
    * Common and specific factors are uncorrelated: $\mathrm{cov}(\mathbf{F}, \mathbf{E}) = \mathbf{0}$
  - To estimate $\mathbf{L}$ and $\boldsymbol{\Psi}$
- Covariance structure
  - $\mathrm{var}(\mathbf{Y}) = \boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\top + \boldsymbol{\Psi}$
    * $\mathrm{var}(Y_i) = \sum_{j=1}^q \ell_{ij}^2 + \psi_i$
  - $\mathrm{cov}(\mathbf{Y}, \mathbf{F}) = \mathbf{L}$

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

- PC method
  1. Perform eigendecomposition on $\mathbf{S} = \mathbf{W}\Lambda\mathbf{W}^\top = \sum_{j=1}^p \lambda_j w_j w_j^\top$
  2. Select q 1) according to PCA stopping rule, 2) as the number of positive eigenvalues of $\mathbf{S}$, OR 3) as the number of eigenvalues greater than one for the correlation matrix
  3. $\widehat{\mathbf{L}} = [\sqrt{\lambda_1} w_1, \dots, \sqrt{\lambda_q} w_q]_{p\times q}$ and $\widehat{\boldsymbol{\Psi}} = \mathrm{diag}(\mathbf{S} - \widehat{\mathbf{L}}\widehat{\mathbf{L}}^\top)$

- ML method
  - Further assumptions
    * $\mathbf{F} \sim MVN_q(\mathbf{0}, \mathbf{I})$
    * $\mathbf{E} \sim MVN_p(\mathbf{0}, \boldsymbol{\Psi})$
    * $\mathbf{L}\boldsymbol{\Psi}^{-1}\mathbf{L}^\top$ is diagonal
  - `factanal` & `psych::fa`

- Comments on estimation
  - Other methods
  - Different statistical softwares may apply different methods
    - Have to look into help manuals to figure out what is going on for different
  - Compare the outputs of multiple estimation methods
    - For a good fit, similar answers would be reached regardless of the method