STAT 5020 : Topics in Multivariate Analysis

Assignment 3 (Due date: 12-Apr-2023)

Academic year 22/23, 2nd term

1. Consider the following linear SEM with dichotomous, continuous, and binary variables $\mathbf{y}_i = (y_{i1}, y_{i2}, y_{i3}, y_{i4}, y_{i5}, y_{i6}, y_{i7}, y_{i8}, y_{i9}, y_{i10})^T$:

$$\begin{aligned} y_{ik}^* &= \mu_k + \boldsymbol{\lambda}_k \boldsymbol{\omega}_i + \epsilon_{ik},\ k = 1, 2, 3 \\ y_{ik} &= \mu_k + \boldsymbol{\lambda}_k \boldsymbol{\omega}_i + \epsilon_{ik},\ k = 4, 5, 6, 7 \\ \vartheta_{ik} &= \mu_k + \boldsymbol{\lambda}_k \boldsymbol{\omega}_i,\ k = 8, 9, 10 \\ \eta_i &= b * d_i + \gamma_1 * \xi_{1i} + \gamma_2 * \xi_{2i} + \delta_i, \\ \boldsymbol{\xi}_i &\sim N(0, \boldsymbol{\Phi}),\ \delta_i \sim N(0, \psi_\delta),\ \boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Psi}_{\boldsymbol{\epsilon}}),\ i = 1, \ldots 500 \end{aligned} \tag{1}$$

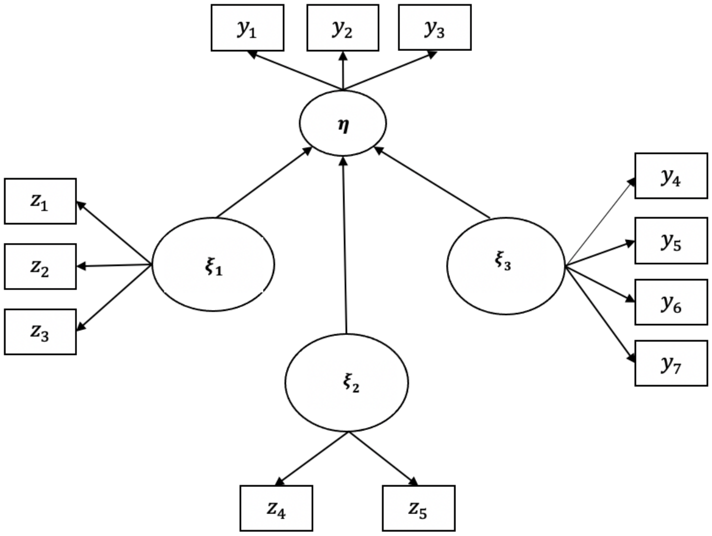
Among the manifest variables $\mathbf{y}_i$, the first three are dichotomous, the next four are continuous, and the last three are binary. $\boldsymbol{\omega}_i = (\eta_i, \xi_1, \xi_2)^T$ is a 3×1 vector of latent variables, y_{ik}^* is the latent continuous measurement for dichotomous y_{ik}, and ϑ_{ik} is the canonical parameter for binary y_{ik}. The fixed covariate d_i is sampled from $Bernoulli(0.7)$. The true values of model parameters are given by

$$\boldsymbol{\Lambda} = \begin{pmatrix} 1 & 0 & 0 \\ 0.8 & 0 & 0 \\ 0.8 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0.7 & 0 \\ 0 & 0.9 & 0 \\ 0 & 0.7 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0.9 \\ 0 & 0 & 0.8 \end{pmatrix},\ \boldsymbol{\mu} = 0,\ \boldsymbol{\Psi}_\epsilon = \text{diag}(1, 1, 1, 0.3, 0.3, 0.25, 0.25),$$

$$b = 0.3, \boldsymbol{\gamma} = (0.4, 0.5)^T, \boldsymbol{\Phi} = \begin{pmatrix} 1 & 0.2 \\ 0.2 & 0.81 \end{pmatrix},\ \psi_\delta = 0.36.$$

Please conduct a simulation study for model (1). Use bias and RMS to summarize the result of Bayesian analysis based on 10 replications.

2. A dataset is taken from 3,074 public and 2,909 private high school seniors to explore the effect of home background (ξ_1), academic orientation (ξ_2), and extra-curricular activity (ξ_3) on students' occupational aspiration (η). Each of the four variables are latent traits measured from a set of manifest variables as follows (loadings/residuals terms omitted):



where z_1-z_5 are ordered categorical variables, y_1-y_3 are continuous, and y_4-y_7 are from the EFDs.

a. Specify a SEM for this multisample problem, write your model in a matrix form, and state the conditions needed for model identification.

b. Describe the major difference in the posterior inference of SEM with multisample data.

c. Briefly describe how to test the invariant constraint on factor loadings across the subpopulations using Bayes factor and DIC. [Hint: the major steps of BF/DIC calculation across iterations]