# STAT5020 Topics in Multivariate Analysis (Qualifying exam 2019-2020)

1. (50%) A linear structural equation model (SEM), denoted as **Model I**, is defined as

$$\mathbf{y}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\omega}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\eta}_i = \boldsymbol{\Pi}\boldsymbol{\eta}_i + \boldsymbol{\Gamma}\boldsymbol{\xi}_i + \boldsymbol{\delta}_i, \tag{1}$$

where $\mathbf{y}_i$ is a $p \times 1$ vector of observed variables, $\boldsymbol{\mu}$ is a vector of intercepts, $\boldsymbol{\Lambda}$ is a $p \times q$ factor loading matrix, $\boldsymbol{\omega}_i = (\boldsymbol{\eta}_i^T, \boldsymbol{\xi}_i^T)^T$, $\boldsymbol{\eta}_i$ and $\boldsymbol{\xi}_i$ are $q_1 \times 1$ and $q_2 \times 1$ vectors of latent variables and $\boldsymbol{\Pi}$ and $\boldsymbol{\Gamma}$ are $q_1 \times q_1$ and $q_1 \times q_2$ matrices of unknown regression coefficients, respectively, and $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\Psi}_\delta$ are the covariance matrices of $\boldsymbol{\xi}_i$, $\boldsymbol{\epsilon}_i$, and $\boldsymbol{\delta}_i$, respectively.

   (a) (10%) Describe the assumptions and identifiability conditions of Model I.

   (b) (10%) In the classical covariance structural analysis (CSA), the covariance matrix of $\mathbf{y}_i$ under Model I is formulated as a matrix function of the unknown parameter vector $\boldsymbol{\theta}$, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. Derive the specific form of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$.

   (c) (10%) In CSA, the maximum likelihood estimator of $\boldsymbol{\theta}$ is obtained through the following discrepancy function $F(\boldsymbol{\theta}) = \log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| + \mathrm{tr}\mathbf{S}\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} - \log|\mathbf{S}| - p$, where $\mathbf{S}$ is the sample covariance matrix of $\mathbf{y}_i$. Show how to obtain this function.

   (d) (10%) Explain why the classical CSA approach cannot be applied to the analyses of advanced SEMs, such as nonlinear, multilevel, and mixture SEMs.

   (e) (10%) Define a nonlinear SEM and describe its statistical inference.

2. (50%) For $i = 1, \cdots, n$, let $\mathbf{u}_i = (u_{i1}, \cdots, u_{ip})^T$ be a $p \times 1$ vector of observed variable and $\boldsymbol{\omega}_i = (\omega_{i1}, \ldots, \omega_{iq})^T$ be a $q \times 1$ random vector of latent variables. A factor analysis model is defined as follows:

$$\mathbf{u}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\omega}_i + \boldsymbol{\zeta}_i, \tag{2}$$

where $\boldsymbol{\mu}$ is a $p \times 1$ vector of intercepts, $\boldsymbol{\Lambda}$ is a $p \times q$ factor loading matrix, $\boldsymbol{\omega}_i \sim N[\mathbf{0}, \boldsymbol{\Phi}]$, and $\boldsymbol{\zeta}_i$ is a $p \times 1$ vector of random errors independent of $\boldsymbol{\omega}_i$ and distributed as $N[\mathbf{0}, \boldsymbol{\Psi}]$ with a diagonal covariance matrix $\boldsymbol{\Psi}$. Let $\mathbf{z}_i = (z_{i1}, \ldots, z_{is})^T$ be an $s \times 1$ random vector of ordinal variables, where $z_{ik}$ takes integer values in $\{1, 2, \ldots, b_k\}$, and $\mathbf{y}_i = (y_{i1}, \ldots, y_{is})^T$ be the vector of underlying continuous variables. The relationship between $\mathbf{y}_i$ and $\mathbf{z}_i$ is defined as follows: for $i = 1, \cdots, n$, $k = 1, \cdots, s$,

$$z_{ik} = m \quad \text{if} \ \ \alpha_{k,m} \le y_{ik} < \alpha_{k,m+1}, \tag{3}$$

where $\{-\infty = \alpha_{k,1} < \alpha_{k,2} < \cdots < \alpha_{k,b_k} < \alpha_{k,b_k+1} = +\infty\}$ is a set of thresholds. Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ir})^T$ be an $r \times 1$ vector of observable covariates. To assess the effects of $\mathbf{x}_i$ and $\boldsymbol{\omega}_i$ on $z_{ij}$, a regression model is considered as follows:

$$y_{ik} = \beta_{0k} + \boldsymbol{\beta}_{1k}^T\mathbf{x}_i + \boldsymbol{\beta}_{2k}^T\boldsymbol{\omega}_i + \epsilon_{ik}, \tag{4}$$

where $\beta_{0k}$ is an intercept, $\boldsymbol{\beta}_{1k}$ and $\boldsymbol{\beta}_{2k}$ are the $r \times 1$ and $q \times 1$ vectors of regression coefficients, $\epsilon_{ik}$ is a random error distributed as $N[0, \sigma_k^2]$ and independent of $\boldsymbol{\omega}_i$.

   Denote by **Model II** the model defined by (2)–(4). Answer the following questions:

   (a) (10%) Draw a path diagram for Model II.

   (b) (10%) Discuss the identifiability issues of Model II.

   (c) (10%) Specify prior distributions for the parameters.

   (d) (10%) Derive the posterior distributions of the unknowns.

   (e) (10%) Discuss the most challenging part of the posterior inference.