00:01:41
说话：

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 $i = 1, \cdots, n$, $k = 1, \cdots, s$,

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

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 β_{0k} is an intercept, $\boldsymbol{\beta}_{1k}$ and $\boldsymbol{\beta}_{2k}$ are the $r \times 1$ and $q \times 1$ vectors of regression coefficients, ϵ_{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.

STAT5030 Linear Models (Qualifying exam 2019-2020)

1. Consider a linear regression model

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots x_{ip}\beta_p + \epsilon_i, \quad i = 1, \ldots, n.$$

The ridge regression is to apply squared penalty on the least squares estimate by minimizing

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2,$$

where $\lambda \ge 0$ is a tuning parameter. By convention, the response is centered and the covariates are standardized. The error term ϵ has zero mean. The resulting estimate is denoted by $\hat{\boldsymbol{\beta}}^{\text{ridge}}$.

   i. Denote the design matrix by $\boldsymbol{X}_{n\times p} = (x_1, \ldots, x_p)$. Derive the explicit expression of $\hat{\boldsymbol{\beta}}^{\text{ridge}}$ in detailed steps.

   ii. Show the details how to compute the ridge solution via the singular value decomposition (SVD).

   iii. Show that there always exists a λ such that the mean squared error (MSE) of $\hat{\boldsymbol{\beta}}^{\text{ridge}}$ is less than the MSE of $\hat{\boldsymbol{\beta}}^{\text{ols}}$, the ordinary least square estimate. (Please provide detailed derivation of each step).

2. In the following, $\boldsymbol{I}_m$ is an $m \times m$ identity matrix, $\mathbf{0}_m$ is an $m \times 1$ vector of zero elements, and $\boldsymbol{J}_m = \mathbf{1}_m\mathbf{1}_m'$, where $\mathbf{1}_m$ is an $m \times 1$ vector of 1's. You may use, without proof, the fact that

$$[\boldsymbol{I}_m + \phi\boldsymbol{J}_m]^{-1} = \left[\boldsymbol{I}_m - \frac{\phi}{1+m\phi}\boldsymbol{J}_m\right].$$

   i. Consider the following linear model:

$$\begin{aligned} Y_{ijt} &= \gamma_i + \tau_j + \epsilon_{ijt}, \\ \epsilon_{ijt} &\sim N(0, \sigma_E^2), \gamma_i \sim N(0, \sigma_\gamma^2), i = 1, 2; j = 1, 2; t = 1, 2; \end{aligned} \tag{1}$$

   where all random variables on the right hand side of the model are mutually independent. Write the model as $\boldsymbol{Y} = \boldsymbol{Z\gamma} + \boldsymbol{X\tau} + \boldsymbol{\epsilon}$, where

$$\boldsymbol{Y} = [Y_{111}, Y_{112}, Y_{121}, Y_{122}, Y_{211}, Y_{212}, Y_{221}, Y_{222}], \boldsymbol{\gamma} = [\gamma_1, \gamma_2], \boldsymbol{\tau} = [\tau_1, \tau_2]$$

   and find $\boldsymbol{Z}, \boldsymbol{X}$. Next, find the variance-covariance matrix of $\boldsymbol{Z\gamma} + \boldsymbol{\epsilon}$.

   ii. State the distribution of $\boldsymbol{Y}$ and find the best linear unbiased estimator of $\boldsymbol{\tau}$ in part (a). Give a condition for $\boldsymbol{C}'\boldsymbol{\tau}$ to be estimable under model (1), where C' is $q \times p$ of rank q (and $q \ge 1$). Justify your answer.

   iii. For given constant vector $\boldsymbol{d}$ and estimable set of functions $\boldsymbol{C}'\boldsymbol{\tau}$, state a test statistic for testing

$$H_0 : \boldsymbol{C}'\boldsymbol{\tau} = \boldsymbol{d} \quad \text{versus} \quad H_1 : \boldsymbol{C}'\boldsymbol{\tau} \neq \boldsymbol{d},$$

   where $\boldsymbol{C}'$ is $q \times p$ of rank q (and $q \ge 1$). Find the expected value of the numerator of the test statistic.

   iv. Let $\phi = \sigma_\gamma^2/\sigma_E^2$ and let $\boldsymbol{C}' = [1, -1]$. Assuming that the distribution of your test statistic in part(c) is non-central F, does the power of this test depend on the value of σ_γ? If so, in which way?