Review for Midterm

zhiyanggeezhou.github.io

Zhiyang Zhou (zhiyang.zhou@umanitoba.ca)

Mar 16, 2022

Statistical modeling

- To figure out the joint distribution of random variables of interest

Random vector

- Mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
- Model: $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{\text{iid}}{\sim} (\mu, \boldsymbol{\Sigma})$
  - Unbiased estimators for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are sample mean vector $\bar{\mathbf{X}}$ and sample covariance matrix $\mathbf{S}$, respectively.

Multivariate normal (MVN)

- Standard normal random vector: entries are iid of $N(0, 1)$
- (General) MVN
  - Defined upon a standard normal random vector via an affine transformation
  - Characterized by $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$
- Properties on MVN
  - Affine-transformed MVN random vector is still of MVN
  - If $[\mathbf{X}_1^\top, \mathbf{X}_2^\top]^\top \sim MVN$, then $\mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2 \Leftrightarrow \operatorname{cov}(\mathbf{X}_1, \mathbf{X}_2) = 0$
  - ...
- Model: $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{\text{iid}}{\sim} MVN_p(\mu, \boldsymbol{\Sigma})$, $n > p$
  - Checking and improving the normal assumption
  - $\hat{\boldsymbol{\mu}}_{\text{ML}} = \bar{\mathbf{X}}$ and $\hat{\boldsymbol{\Sigma}}_{\text{ML}} = n^{-1}(n-1)\mathbf{S}$
  - Sampling distribution of $\hat{\boldsymbol{\mu}}_{\text{ML}}$ and $\hat{\boldsymbol{\Sigma}}_{\text{ML}}$
  - Inference on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$
    - Likelihood ratio test (LRT)
      - Hypotheses
      - Name of approach
      - Value of test statistic
      - Rejection region/p-value
      - Conclusion: e.g., at the α level, we reject/do not reject H0, i.e., we believe. . .
    - Confidence region for an unknown vector: a dual problem of hypothesis testing
    - Simultaneous confidence intervals:
      - Construct a CI for each random scalar of interest, e.g., entries of $\boldsymbol{\mu}$, simultaneously
      - To make sure the coverage probability of the intersection of multiple CIs is at least $1 - \alpha$

- Model: $\mathbf{X}_{11}, \ldots, \mathbf{X}_{1n_1} \overset{\text{iid}}{\sim} MVN_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, $n_1 > p$, and $\mathbf{X}_{21}, \ldots, \mathbf{X}_{2n_2} \overset{\text{iid}}{\sim} MVN_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, $n_2 > p$
  - Inference on $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ via LRT

Multivariate analysis of variance (MANOVA)

- One-way MANOVA
  - Model: $\mathbf{X}_{ij} = \boldsymbol{\mu} + \boldsymbol{\tau}_i + \mathbf{E}_{ij}$ with $\mathbf{E}_{ij} \overset{\text{iid}}{\sim} MVN_p(\mathbf{0}, \boldsymbol{\Sigma})$ and $\sum_i \boldsymbol{\tau}_i = 0$
  - Testing $H_0 : \boldsymbol{\tau}_1 = \cdots = \boldsymbol{\tau}_m = 0$ v.s. H_1 : otherwise
    * Sums of squares and cross products matrices (SSP)
    * Wilk's lambda test (a modification of LRT)
- Two-way MANOVA
  - Model: $\mathbf{X}_{ijk} = \boldsymbol{\mu} + \boldsymbol{\tau}_i + \boldsymbol{\beta}_j + \boldsymbol{\gamma}_{ij} + \mathbf{E}_{ijk}$ with $\mathbf{E}_{ijk} \overset{\text{iid}}{\sim} MVN_p(\mathbf{0}, \boldsymbol{\Sigma})$, $i = 1, \ldots, m$, $j = 1, \ldots, b$, $k = 1, \ldots, n$
    * $\boldsymbol{\tau}_i$: the main effect of factor 1 at level i
    * $\boldsymbol{\beta}_j$: the main effect of factor 2 at level j
    * $\boldsymbol{\gamma}_{ij}$: the interaction of factors 1 and 2 whose levels are i and j, respectively
    * Identifiability: $\sum_i \boldsymbol{\tau}_i = \sum_j \boldsymbol{\beta}_j = \sum_i \boldsymbol{\gamma}_{ij} = \sum_j \boldsymbol{\gamma}_{ij} = \mathbf{0}$
  - Testing first the interaction (via the Wilk's lambda test) and then main effects if the interaction is insignificant
- Testing the equality of covariance matrices
  - Model: m independent samples, where
    * $\mathbf{X}_{11}, \ldots, \mathbf{X}_{1n_1} \overset{\text{iid}}{\sim} MVN_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$
    * $\vdots$
    * $\mathbf{X}_{m1}, \ldots, \mathbf{X}_{mn_m} \overset{\text{iid}}{\sim} MVN_p(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$
  - Testing $H_0 : \boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_m$ v.s. H_1 : otherwise via the Box's M test statistic (a modification of LRT)

Multivariate linear model

- Linear model: responses are linear functions with respect to unknown parameters
- Model
  - Population version
    * $\mathrm{E}([Y_1, \ldots, Y_p]^\top \mid X_1, \ldots, X_q) = \mathbf{B}^\top [1, X_1, \ldots, X_q]^\top$
    * $\mathrm{cov}([Y_1, \ldots, Y_p]^\top \mid X_1, \ldots, X_q) = \boldsymbol{\Sigma} > 0$
  - Sample version
$$\underset{n \times p}{\mathbf{Y}} = \underset{n \times (q+1)}{\mathbf{X}} \; \underset{(q+1) \times p}{\mathbf{B}} + \underset{n \times p}{\mathbf{E}}$$
    * $\mathbf{Y} = [Y_{ij}]_{n \times p}$
    * Design matrix
$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{q1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{nq} \end{bmatrix}_{n \times (q+1)}$$
      · $\mathrm{rk}(\mathbf{X}) = q + 1 < p + q + 1 \le n$
    * $\mathbf{E} = [\mathbf{E}_{1\cdot}, \ldots, \mathbf{E}_{n\cdot}]^\top$, where $\mathbf{E}_{i\cdot}^\top$ is the ith row of $\mathbf{E}$
    * Assume the independence across i, i.e.,
      · $[Y_{i1}, \ldots, Y_{ip}, X_{i1}, \ldots, X_{iq}]^\top \overset{\text{iid}}{\sim} [Y_1, \ldots Y_p, X_1, \ldots, X_q]^\top$
      · $\mathbf{E}_{1\cdot}, \ldots, \mathbf{E}_{n\cdot} \overset{\text{iid}}{\sim} (\mathbf{0}_p, \boldsymbol{\Sigma})$
  - Relationship with univariate linear model and MANOVA
  - Least squares (LS) estimation (without normality assumption) and maximum likelihood (ML) estimation (assuming the conditional distribution of $[Y_1, \ldots, Y_p]^\top$ is of MVN)

- Inference on $\mathbf{B}^\top \boldsymbol{a}$ and $\mathbf{Y}_0$: confidence region and prediction region

- Model comparison/selection
  - Testing for nested models
    - $H_0 : \mathrm{E}(\mathbf{Y} \mid \mathbf{X}) = \mathbf{X}_{(0)}\mathbf{B}_{(0)}$ (nested model) vs. $H_1 : \mathrm{E}(\mathbf{Y} \mid \mathbf{X}) = \mathbf{X}_{(0)}\mathbf{B}_{(0)} + \mathbf{X}_{(1)}\mathbf{B}_{(1)}$ (full model)
  - Comparing non-nested models
    - Information criteria

- Model checking
  - Multivariate influence measures
    - Outliers identification via the (externally) Studentized residuals and Cook's distance
  - Normality of residuals