Mini Project 2

Jackson Passey, Gavin Hatch, Ty Hawkes

2025-10-17

Initialization: Start with initial estimates $\tilde{\boldsymbol{\mu}}$ and $\tilde{\Sigma}$.

For each observation i, partition the vector

$$\boldsymbol{x}_i = \begin{bmatrix} \boldsymbol{x}_i^{(1)} \\ \boldsymbol{x}_i^{(2)} \end{bmatrix},$$

where $\boldsymbol{x}_i^{(1)}$ are the *missing* components and $\boldsymbol{x}_i^{(2)}$ are the *observed* components.

Similarly, partition the mean vector and covariance matrix as

$$\tilde{\boldsymbol{\mu}} = \begin{bmatrix} \tilde{\boldsymbol{\mu}}^{(1)} \\ \tilde{\boldsymbol{\mu}}^{(2)} \end{bmatrix}, \quad \tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} \tilde{\boldsymbol{\Sigma}}_{11} & \tilde{\boldsymbol{\Sigma}}_{12} \\ \tilde{\boldsymbol{\Sigma}}_{21} & \tilde{\boldsymbol{\Sigma}}_{22} \end{bmatrix}.$$

E-step (Expectation): Compute the conditional expectation of the missing components:

$$\tilde{\boldsymbol{x}}_i^{(1)} = \tilde{\boldsymbol{\mu}}^{(1)} + \mathbf{B}_i \left(\boldsymbol{x}_i^{(2)} - \tilde{\boldsymbol{\mu}}^{(2)} \right),$$

where

$$\mathbf{B}_i = \tilde{\boldsymbol{\Sigma}}_{12} \tilde{\boldsymbol{\Sigma}}_{22}^{-1}.$$

M-step (Maximization): Update the mean and covariance estimates using the completed data $\tilde{\boldsymbol{x}}_i = [\, \tilde{\boldsymbol{x}}_i^{(1)}, \boldsymbol{x}_i^{(2)} \,]$:

$$\tilde{\boldsymbol{\mu}}^{(new)} = \frac{1}{n} \sum_{i=1}^{n} \tilde{\boldsymbol{x}}_i, \quad \tilde{\boldsymbol{\Sigma}}^{(new)} = \frac{1}{n} \sum_{i=1}^{n} (\tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{\mu}}^{(new)})(\tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{\mu}}^{(new)})'.$$

Repeat the E- and M-steps until convergence.

Imputation (after convergence): Once convergence is reached, with final estimates $\boldsymbol{\mu}^*$ and $\boldsymbol{\Sigma}^*$, generate multiple imputations as:

$$\boldsymbol{x}_{i,[m]}^{(1)} = \boldsymbol{\mu}^{*(1)} + \mathbf{B}_i^* (\boldsymbol{x}_i^{(2)} - \boldsymbol{\mu}^{*(2)}) + \boldsymbol{e}_{i,[m]}^{(1)},$$

where

$$\mathbf{B}_i^* = \boldsymbol{\Sigma}_{12}^* (\boldsymbol{\Sigma}_{22}^*)^{-1},$$

and

$$\boldsymbol{e}_{i,[m]}^{(1)} \sim \mathcal{N}_q \left(\mathbf{0},\ \boldsymbol{\Sigma}_{11}^* - \boldsymbol{\Sigma}_{12}^* (\boldsymbol{\Sigma}_{22}^*)^{-1} \boldsymbol{\Sigma}_{21}^* \right),$$

with q = number of missing components in observation i.

Conditional Wilk's Λ Distribution

$$\Lambda_{z|y} = \frac{\Lambda_{yz}}{\Lambda_y} \sim \Lambda_{q,\nu_H,\nu_E - p}$$

where $q = 6$ for length of y acids, and $p = 2$ for length of z acids. The Wilk's Λ approximates to the following F distribution:

$$F = \frac{1 - \Lambda_{z|y}^{1/t}}{\Lambda_{z|y}^{1/t}} \frac{df_1}{df_2} \sim F_{df_1, df_2}$$

Where $t = 1$, $df_1 = 2$, and $df_2 = 83$.