Lab 14

Mixture Models

Ex. 1. (2.5p) A health research institute wants to analyze the relationship between the number of hours of physical exercise per week and cholesterol levels in individuals from several distinct demographic groups. These individuals come from 3 to 5 different subpopulations (e.g., age, lifestyle, genetics), and the proportions of each subpopulation in the sample are unknown.

The model assumes that each individual belongs to one of the subpopulations, each having its own regression model (polynomial). The choice of subgroup is determined by the following mixture:

$$\text{Cholesterol}_i = \sum_{i=1}^{K} w_k \cdot \mathcal{N}(\mu_{k,i}, \sigma_k^2),$$

where K is the number of subpopulations, w_k are the weights of the subpopulations, $\mu_{k,i} = \alpha_k + \beta_k t_i + \gamma_k t_i^2$ (polynomial model), and σ_k is the standard deviation of each model. Observations regarding the number of hours of physical exercise per week and cholesterol levels are collected in the file date_colesterol.csv.

1. Estimate the weights and regression coefficients for each subpopulation, for each $K \in \{3, 4, 5\}$. (1.5p)
2. How many subpopulations best represent the observed data? Justify this using Bayesian criteria such as WAIC and/or LOO. (1p)