

Computational Statistics

Chapter 3: EM algorithm

1. Let $\mathbf{Y} = (Y_1, \dots, Y_n)$ be an i.i.d. sample from a mixture of a normal distribution $\mathcal{N}(\mu, \sigma)$ and a uniform distribution $\mathcal{U}([-a, a])$, with pdf

$$g(y; \theta) = \pi \phi(y; \mu, \sigma) + (1 - \pi)c, \quad (1)$$

where $\phi(\cdot; \mu, \sigma)$ is the normal pdf, $c = (2a)^{-1}$, π is the proportion of the normal distribution in the mixture and $\theta = (\mu, \sigma, \pi)^T$ is the vector of parameters. Typically, the uniform distribution corresponds to outliers in the data. The proportion of outliers in the population is then $1 - \pi$. We want to estimate parameter θ using the EM algorithm

- (a) Using the functions `sample`, `rnorm` and `runif`, generate a sample of size $n = 100$. Draw a box plot of the data.
 - (b) Write an EM algorithm for this problem.
 - (c) Apply the EM algorithm to the data, with different initializations. Draw the estimated probabilities $1 - z_i^{(t)}$ of being an outlier, as a function of y_i . Does it make sense?
 - (d) Compare the estimates with those computed using the `optim` function.
2. We will now apply the same idea to linear regression. We assume that we have an independent sample $\mathbf{Y} = (Y_1, \dots, Y_n)$, where the distribution of each observation Y_i is a mixture of a normal distribution $\mathcal{N}(v_i^T \beta, \sigma^2)$ and a uniform distribution $\mathcal{U}([-a, a])$, v_i being a vector of covariates and β a vector of coefficients. The pdf of Y_i is

$$g(y; \theta) = \pi \phi(y; v_i^T \beta, \sigma) + (1 - \pi)c, \quad (2)$$

with $\theta = (\beta^T, \sigma, \pi)^T$ and $c = (2a)^{-1}$.

- (a) Using the functions `rnorm` and `runif`, generate a sample of size $n = 100$, with $v_i \sim \mathcal{U}([-6, 6])$, $\beta = (1, 2)^T$, $\sigma = 2$, $a = 20$ and $\pi = 0.5$. Draw a scatter plot of the data.
- (b) Compute the ordinary least squares estimates (OLS) of the coefficients, and draw the corresponding line.

- (c) Write an EM algorithm for this problem. (In the M-step, you will have to solve a weighted least-squares problem. You can use the `lm` functions with input parameter `weights`).
- (d) Apply the EM algorithm to the data, taking the OLS estimates as initial values. Draw the line with coefficients equal to the MLEs. Plot the points (v_i, y_i) such that $z_i^{(t)} < 0.5$ as filled circles. Does it make sense?