Problem 1: EM for Exponential Models

Assume that the data is generated by an exponential model \( P(y|\theta) = \theta e^{-\theta y} \) for \( y \geq 0 \). Use maximum likelihood to estimate \( \theta \) from the full dataset.

Now sample from an NMAR missing data mechanism \( P(m=1|y, \phi) = e^{-y\phi} \) to remove data. Set \( \phi = 0.0001 \). What is the ML estimate of \( \theta \) from the remaining data?

Formulate the task of estimating \( \theta \) and \( \phi \) in terms of maximum likelihood using the missing data mechanism \( P(M|Y_{obs}, Y_{mis}, \phi) \) and \( P(Y_{obs}, Y_{mis}|\theta) \). Explain how the EM algorithm can be used to estimate \( \theta, \phi \).

Implement EM and use it to: (a) estimate \( \theta \) assuming that \( \phi \) is known, (b) to estimate \( \theta, \phi \). Compare the estimates of \( \theta \) to those obtained for the full dataset, and for the remaining data.

Problem 2: EM for Gaussian Models

Repeat the previous question, but assume that the data is generated by a Gaussian model \( P(y|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/(2\sigma^2)} \).

The missing data mechanism is \( P(m=1|y, \phi) = e^{-(y-\mu_\phi)^2/(2\sigma^2_\phi)} \). Set \( \mu_\phi = 5000, \sigma^2_\phi = 1000 \).

Estimate \( \mu_\theta, \sigma^2_\theta \) for the full dataset, and for the remaining data after sampling to remove missing data.

Apply EM: (a) to estimate \( \mu_\theta, \sigma^2_\theta \), assuming that \( \mu_\phi, \sigma^2_\phi \) are known, (b) to estimate \( \mu_\theta, \sigma^2_\theta, \mu_\phi \), assuming that \( \sigma^2_\phi \) is known.