The M-step in EM Algorithm

Via MLE we get the following estimates:

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} x_n \qquad \Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (x_n - \mu_k)(x_n - \mu_k)^T \qquad w_k = \sum_{n=1}^{N} r_{nk}$$

the higher the responsibility of a data point for a cluster is, the more influence it has on what the mean and variance is

Note: $N_k = \sum_{n=1}^{N} r_{nk}$ is now based on soft assignments now

if data points are unlikely to belong to cluster k, the N_k small,
if data points are likely to belong to cluster k, then N_k large

Take Aways

- GMM does **soft assignment**, every data point belongs to every cluster with some probability
- Data points that are more likely to be in a cluster have **more influence** over its parameters
- GMM uses the EM algorithm to iteratively update the cluster distributions:
  - first assign a responsibility to each data point (**E-step**)
  - then using them to calculate weighted means and variances for each cluster (**M-step**)
- Responsibilities measure **the probability of a data point being in each cluster** (technically the posterior probability).
- Responsibilities contain information about how common a cluster is as well as **the likelihood of a data point belonging to that cluster**