# $k$-means vs. GMM

Comparison of $k$-means and GMM iterations for the clustering of a set of data points, $\{\mathbf{x_n}\}_{\mathbf{n=1}}^{\mathbf{N}}$.

| | $k$**-means** | **GMMs** |
|---|---|---|
| Init: | Select $K$ cluster centers $(\mathbf{m_1^{(1)}}, \dots, \mathbf{m_K^{(1)}})$ | $K$ components with means, $\mu_k$ and covariance $\Sigma_k$ and mixing coefficients, $P_k$ |
| E-step: | Allocate datapoints to clusters<br><br>$S_i^{(t)} = \{\mathbf{x_p} : \lvert\lvert\mathbf{x_p} - \mathbf{m_i^{(t)}}\rvert\rvert^{\mathbf{2}} \leq \lvert\lvert\mathbf{x_p} - \mathbf{m_j^{(t)}}\rvert\rvert^{\mathbf{2}} \forall \mathbf{j}\}$ | Update probability that component, $k$, generating the data point $\mathbf{x_n}$:<br>$\gamma_{nk} = \frac{P_k \mathcal{N}(\mathbf{x_n} \lvert \mu_\mathbf{k}, \mathbf{\Sigma_k})}{\sum_{j=1}^{K} P_j \mathcal{N}(\mathbf{x_n} \lvert \mu_\mathbf{j}, \mathbf{\Sigma_j})}$ |
| M-step: | Re-estimate cluster centers:<br><br>$\mathbf{m_i^{(t+1)}} = \frac{\mathbf{1}}{\lvert\mathbf{S_i^{(t)}}\rvert} \sum_{\mathbf{x_j} \in \mathbf{S_i^{(t)}}} \mathbf{x_j}$ | Calculating the estimated number of cluster members, $N_k$, means, $\mu_k$ and covariance $\Sigma_k$ and mixing coefficients, $P_k$.<br>$N_k = \sum_{n=1}^{N} \gamma_{nk}$,<br>$\mu_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \mathbf{x_n}$,<br>$\Sigma_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} (\mathbf{x_n} - \mu_\mathbf{k}^\mathbf{new})(\mathbf{x_n} - \mu_\mathbf{k}^\mathbf{new})^\mathbf{T}$,<br>$P_k^{new} = \frac{N_k}{N}$, |
| Stop: | Based on no differences in set allocation. | When the likelihood not increasing fast enough:<br>$\ln \Pr(\mathbf{x} \lvert \mu, \mathbf{\Sigma}, \mathbf{P}) = \sum_{\mathbf{n=1}}^{\mathbf{N}} \ln\{\sum_{\mathbf{k=1}}^{\mathbf{K}} \mathbf{P_k} \mathcal{N}(\mathbf{x_n} \lvert \mu_\mathbf{k}, \mathbf{\Sigma_k})\}$ |