K means algorithm

Let $n \geqslant 1$ and $X_1, \ldots, X_n$ in $\mathbb{R}^p$. The K-means algorithm aims at minimizing over all partitions $G = (G_1, \ldots, G_K)$ of $\{1, \ldots, p\}$ the criterion

$$\mathcal{L}(G) = \sum_{k=1}^{K} \sum_{a \in G_k} \|X_a - \bar{X}_{G_k}\|^2 \quad \text{with} \quad \bar{X}_{G_k} = \frac{1}{|G_k|} \sum_{b \in G_k} X_b \, .$$

1. Prove that

$$\mathcal{L}(G) = \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{a,b \in G_k} \langle X_a, X_a - X_b \rangle = \frac{1}{2} \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{a,b \in G_k} \|X_a - X_b\|^2 \, .$$

2. Assume now that the observations are independent. Write $\mathbb{E}[X_a] = \mu_a \in \mathbb{R}^p$ so that $X_a = \mu_a + \varepsilon_a$ with $\varepsilon_1, \ldots, \varepsilon_n$ centered and independent. Define $v_a = \mathrm{trace}(\mathbb{V}[X_a])$. Prove that

$$\mathbb{E}[\mathcal{L}(G)] = \frac{1}{2} \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{a,b \in G_k} \left(\|\mu_a - \mu_b\|^2 + v_a + v_b \right) \mathbf{1}_{a \neq b} \, .$$

   What is the value of $\mathbb{E}[\mathcal{L}(G)]$ when all the within-group variables have the same mean?

3. We assume now that there exists a partition $G^* = (G_1^*, \ldots, G_K^*)$ such that within-group variables have the same mean and the same volume. More precisely, we assume that there exists $m_1, \ldots, m_K \in \mathbb{R}^p$ and $\gamma_1, \ldots, \gamma_K > 0$ such that $\mu_a = m_k$ and $v_a = \gamma_k$ for all $a \in G_k^*$ and $k = 1, \ldots, K$. Compute $\mathbb{E}[\mathcal{L}(G^*)]$.

4. In the special case where $\gamma_1 = \ldots = \gamma_K = \gamma$, which partition $G = (G_1, \ldots, G_K)$ minimizes $\mathbb{E}[\mathcal{L}(G)]$?