Mélange de gaussiennes

1 Hypothèses

On suppose que le nombre de clusters K est connu. L'ensemble des observations $(X_1, \dots, X_n)$ est constitué de K sous-populations, chaque sous-population correspondant à un cluster. On note $C(i) \in \{1, \dots, K\}$ l'étiquette (non observée) de l'observation i.

Chaque sous-population suit une distribution connue, caractérisée par un paramètre θ_k, pour $k \in \{1, \dots, K\}$. Les paramètres θ_k sont totalement ou partiellement inconnus et doivent donc être estimés conjointement avec les affectations $C(i)$.

2 Approche par maximum de vraisemblance

Sous ces hypothèses, on peut écrire la fonction de vraisemblance complète (vraisemblance des données complétées par les labels) :

$$L(\theta, C) = \prod_{k=1}^{K} \prod_{\{i : C(i)=k\}} f(X_i \mid \theta_k).$$

où $f(\cdot \mid \theta_k)$ désigne la densité du modèle sous-jacent pour le cluster k.

2.1 Cas gaussien

Si chaque sous-population suit une loi normale multivariée $\mathcal{N}(\mu_k, \Sigma_k)$, la vraisemblance complète s'écrit :

$$L(\theta, C) = \prod_{k=1}^{K} \prod_{i : C(i)=k} (2\pi)^{-d/2} \, |\Sigma_k|^{-1/2} \exp\Big(-\frac{1}{2} (X_i - \mu_k)^\top \Sigma_k^{-1} (X_i - \mu_k) \Big),$$

où d est la dimension des données.

2.2 Optimisation

Comme les étiquettes $C(i)$ sont inconnues, l'optimisation directe de la vraisemblance complète est impossible. La stratégie standard consiste à maximiser la vraisemblance marginale :

$$L(\theta) = \prod_{i=1}^{n} \left(\sum_{k=1}^{K} \pi_k \, f(X_i \mid \theta_k) \right),$$

où $\pi_k = \mathbb{P}(C(i) = k)$ sont les proportions (ou poids de mélange).

La maximisation de cette vraisemblance s'effectue classiquement via l'algorithme EM (Expectation–Maximization), qui alterne :

— E-step : estimation des probabilités d'appartenance τ_{ik} ;
— M-step : mise à jour des paramètres π_k, μ_k, Σ_k.

3 Propriétés et contraintes sur les matrices de covariance

L'estimation est fortement influencée par les contraintes imposées sur les matrices Σ_k. Ces contraintes déterminent la forme géométrique des clusters (sphériques, ellipsoïdes, étirés, orientés, etc.).

Voici les cas usuels :

— Covariances sphériques, même taille : $\Sigma_k = \sigma^2 I$.
Les clusters sont des boules de même rayon ; comportement proche du k-means.
— Covariances sphériques, tailles différentes : $\Sigma_k = \sigma_k^2 I$.
Clusters sphériques mais de dispersion différente.
— Covariances diagonales :
Les dimensions sont supposées indépendantes. Une standardisation préalable peut être effectuée. Les clusters sont des ellipsoïdes alignés avec les axes.
— Covariances ellipsoïdales de même forme mais de tailles différentes : $\Sigma_k = \lambda_k D D^\top$.
Même orientation, dispersion variable.
— Covariances générales (Σ_k non contraintes) :
Cas le plus flexible. Les clusters sont des ellipsoïdes de forme complètement libre, au risque de sur-ajustement lorsque d est grand.

Ces différentes paramétrisations correspondent aux modèles GPCM (Banfield & Raftery), tels qu'implémentés dans `mclust`.

4 Choix du nombre de clusters

Dans une approche par maximum de vraisemblance, le nombre de clusters K n'est pas estimé directement par EM. On ajuste plutôt différents modèles pour plusieurs valeurs possibles de K, puis on sélectionne le meilleur selon un critère pénalisé, typiquement :

— AIC (Akaike Information Criterion) : Favorise des modèles plus complexes.
— BIC (Bayesian Information Criterion) : Critère plus parcimonieux, souvent utilisé pour les mélanges gaussiens.