**Exercice 4.a.**

Pour tester les performances de l’algorithme, j’ai construit des données synthétiques labellisées :

|  |  |
| --- | --- |
| /var/folders/ty/1lsvbm5d23jfrmrt1cqkzy080000gn/T/com.microsoft.Word/Content.MSO/DB53A904.tmp  *Données labellisées générées avec des gaussiennes* | /var/folders/ty/1lsvbm5d23jfrmrt1cqkzy080000gn/T/com.microsoft.Word/Content.MSO/2FF169B2.tmp  *Résultat du k-means* |

La convergence dépend cependant de l’initialisation. J’ai ensuite pu comparer la distorsion des résultats avec l’algorithme classique du K-Means et celui de K-Means++.

J’ai obtenu le tableau suivant en simulant l’algorithme 1000 fois :

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithme** |  | **Distorsion** |  |
|  | **Moyenne** | **Écart-type** | **Max** |
| Kmeans | 1665.29 | 1796.92 | 23645.19 |
| Kmeans++ | 1162.58 | 590.24 | 3203.85 |

L’algorithme Kmeans++ fonctionne non seulement mieux en moyenne, mais est nettement plus robuste (la distorsion maximale est nettement plus faible).

**Exercice 4.b**

Sous l’hypothèse d’une variance proportionnelle à l’identité, la vraisemblance d’une distribution du mélange de gaussienne peut s’écrire :

Avec , étant le vecteur indicateur aléatoire qui attribue un élément à un cluster.

*E-Step* : l’espérance de la vraisemblance en conditionnellement à est :

avec . On calcule donc chaque

*M-Step :*

Le problème de maximisation du logarithme de la vraisemblance étant concave sur un domaine convexe, le maximiseur annule la dérivant. La vraisemblance en deux termes. D’un côté, on dérive par rapport à

Soit :

En notant :

De la même manière :

L’autre terme dépend de et en le dérivant par rapport à :

**Exercice 4.d**

|  |  |  |
| --- | --- | --- |
| **Set** | **Isotropic** | **General** |
| Train | -19740.56 | -69820.00 |
| Test | -19518.09 | -66032.70 |

*Complete log-likelihood*

Les données ne sont pas réparties selon des cercles dans le plan. C’est pourquoi, elles s’expliquent difficilement avec des covariances diagonales. Cela se voit visuellement, car l’ellipse de covariance du cluster en haut à gauche est très large. Ainsi, des points situés loin de lui sont attribués par erreur. Cela se voit également à travers ce tableau

