Choix du nombre de clusters :

La méthode K-means est une méthode d’apprentissage non-supervisée. Lors de son application les données sont séparées en plusieurs classes prédéterminés de façon que les individus ayant le plus de similarité. C’est ainsi qu’une des tâches clefs est de trouver le nombre approprié de classes, k. Il existe plusieurs techniques pour déterminer le nombre de classes. Nous discuterons que des cas les plus connus :

1. **Méthode du pouce [[1]](#footnote-1)**

Cette méthode est une méthode approximative où le nombre de classes, k est déterminé par :

Une image contenant dessin, table

Description générée automatiquement

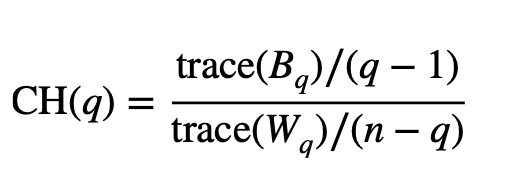
1. **L’indice de qualité**

Afin d’évaluer la qualité de la classification, les indices inertiels[[2]](#footnote-2), soient l’inertie intra-classes et l’inertie l’inter-classes sont souvent utilisés. L’inertie intra-classes ‘mesure le degré d’homogénéité entre les objets appartenant à la même classe‘ tandis que l’inertie inter-classes ‘mesure le degré d’hétérogénéité entre les classes.’

Il existe plusieurs indices de qualité, par exemple l’indice de Dunn, l’indice de Calinski et Harabasz (CH) ou encore l’indice de Silhouette.

Le premier calcule la distance minimale inter-classes et ainsi plus cette distance est grande, meilleur est classification.

Nous notons l’indice de CH comme étant :



, où Bq est la somme d’erreurs au carré inter-classes, et

Wq est la somme d’erreurs au carré intra-classes.

Cette méthode maximise l’indice de CH(q) [[3]](#footnote-3) par rapport à q (q>1) afin d’estimer .

Introduit par Kauffman et Rousseew, l’indice de Silhouette nous donne une représentation visuelle de la distance entre un point d’une classe avec les points des classes voisines. Plus le coefficient ainsi calculé est proche de 1, plus la distance avec les classes voisines (inertie inter) est grande. Ceci représente le nombre de classes optimale. À l’inverse un coefficient proche de -1 nous indique une mauvaise classification de l’observation.

1. **Méthode du coude**

La méthode du coude est une technique visuelle très connue. L’idée derrière cette technique est d’implémenter la méthode K-means en parcourant k valeurs. À chacune des k valeurs, la somme des erreurs au carré est calculée et est affiché sur un graphique, nous permettant à mieux visualiser les résultats. L’objectif est de choisir la valeur k (qui sera le nombre de classes) créant un effet de ‘coude’, c’est-à-dire provoquant une baisse une conséquente, plus soudaine de la somme des erreurs au carré. Nous disons ceci en gardant en tête que la somme des erreurs aura toujours tendance à baisser, plus la valeur de k est grande.

1. **La validation croisée**

La validation croisée regarde la stabilité des classes. Les données sont séparées en au moins deux parties. La première est utilisée pour former les classes tandis que la deuxième sert de validation. Lorsque nous parlons de stabilité, nous parlons de la fréquence à laquelle des classes similaires sont formées lorsque plusieurs itérations sont effectuées. Ainsi une plus grosse fréquence de l’apparition de mêmes classes équivaut à une plus grosse stabilité de ces classes[[4]](#footnote-4).

Dans le cadre de notre étude, nous choisissons de travailler avec la méthode du coude.

1. <https://www.researchgate.net/publication/313554124_Review_on_Determining_of_Cluster_in_K-means_Clustering> [↑](#footnote-ref-1)
2. <https://hal.archives-ouvertes.fr/hal-00614071/document> [↑](#footnote-ref-2)
3. <https://link.springer.com/content/pdf/10.1007/s41019-019-0091-y.pdf> [↑](#footnote-ref-3)
4. <https://www.researchgate.net/publication/313554124_Review_on_Determining_of_Cluster_in_K-means_Clustering> [↑](#footnote-ref-4)