**Clustering EEG Signals**

A diagram of a clustering graph

Description automatically generated with medium confidence

\*I wanted clusters but received 3 with best initialization with minimum inertia.

Inertia is a metric used to evaluate the quality of clustering in the K-means algorithm. Specifically, inertia measures the sum of squared distances between each data point in a cluster and the centroid of that cluster. Mathematically, it can be expressed as:

Inertia indicates how internally coherent the clusters are:

* **Lower inertia** means that the data points are closer to their respective cluster centers, suggesting more compact and well-defined clusters.
* **Higher inertia** indicates that the data points are more spread out from their cluster centers, suggesting less compact clusters.

The goal in K-means clustering is to minimize inertia, as lower inertia values generally indicate better clustering results where the points within each cluster are more similar to each other.

So tried **PAM (K Medoid clustering)**

A diagram of a number of dots

Description automatically generated with medium confidence

Here's a comparison of the two:

**1. Cluster Center Selection**

* **K-Means**:
  + **Centroids**: K-Means uses the arithmetic mean of the points in a cluster as the center, known as the *centroid*. The centroid is not necessarily one of the original data points.
  + **Continuous Update**: Centroids are updated iteratively by averaging the coordinates of all points in the cluster.
* **PAM (K-Medoids)**:
  + **Medoids**: PAM selects actual data points as the center of the cluster, called *medoids*. Medoids are the most centrally located points in a cluster, minimizing the sum of dissimilarities (distances) to all other points in the cluster.
  + **Discrete Update**: Medoids are updated by swapping data points between clusters to reduce the overall cost (sum of dissimilarities).

**2. Robustness to Outliers**

* **K-Means**:
  + **Sensitive to Outliers**: Because K-Means uses the mean to compute centroids, it is sensitive to outliers. An outlier can significantly shift the position of the centroid, leading to less meaningful clustering.
* **PAM (K-Medoids)**:
  + **Robust to Outliers**: PAM is more robust to outliers since it uses actual data points (medoids) as cluster centers. Outliers are less likely to be chosen as medoids, so their impact on the clustering is minimized.

**3. Computational Complexity**

* **K-Means**:
  + **Less Complex**: K-Means is computationally efficient with a time complexity of O(n×k×t)O(n \times k \times t)O(n×k×t), where nnn is the number of data points, kkk is the number of clusters, and ttt is the number of iterations. This makes it suitable for large datasets.
* **PAM (K-Medoids)**:
  + **More Complex**: PAM has a higher computational cost with a time complexity of O(k×(n−k)2)O(k \times (n - k)^2)O(k×(n−k)2). This is because PAM involves pairwise distance calculations between all data points and their potential medoids. As a result, PAM is generally slower and less scalable than K-Means, especially on large datasets.

**4. Algorithm Initialization**

* **K-Means**:
  + **Random Initialization**: K-Means typically starts with a random selection of initial centroids. This can lead to different results on different runs unless the random state is fixed. Multiple initializations (e.g., n\_init parameter) are often used to find the best clustering.
* **PAM (K-Medoids)**:
  + **Deterministic Initialization**: PAM usually starts with a deterministic or heuristic selection of initial medoids. It is less dependent on random initialization, leading to more stable and consistent results across runs.

**5. Suitability for Different Data Types**

* **K-Means**:
  + **Euclidean Distance**: K-Means relies on minimizing Euclidean distance, making it best suited for continuous, numerical data. It may not perform well on categorical or mixed-type data without modifications.
* **PAM (K-Medoids)**:
  + **Any Distance Metric**: PAM can work with any distance metric (e.g., Manhattan distance, Euclidean distance, etc.), making it more versatile for different types of data, including categorical, ordinal, and mixed-type data.

**6. Interpretability**

* **K-Means**:
  + **Centroids**: The centroids in K-Means do not correspond to actual data points, which might make the interpretation of clusters less intuitive, especially in non-linear datasets.
* **PAM (K-Medoids)**:
  + **Medoids**: Since medoids are actual data points, they can provide a more interpretable and meaningful representative of each cluster, especially in applications where it is important to identify a specific example from the dataset as a cluster center.

**7. Use Cases**

* **K-Means**:
  + Commonly used in large-scale clustering problems, where speed and efficiency are crucial, such as in customer segmentation, document clustering, image compression, etc.
* **PAM (K-Medoids)**:
  + Preferred in scenarios where robustness to outliers and interpretability are more important, such as in market research, healthcare data analysis, and scenarios with mixed data types.