# Evolutionary clustering of MMPI-2 psychometric data on candidates for child adoption

Emanuele Musumeci, 1653885

## I. Introduction

In this project I faced the challenge of clustering samples from a high-dimensional dataset obtained through the Minnesota Multiphasic Personality Inventory (MMPI-2) psychological test, admistered to a pool of candidate couples for child adoption. The main difficulty of this task was the lack of understanding of the semantics for the various psychometric scales and their importance in the psychological profiling for the case at study. I addressed this issue by choosing evolutionarily which psychometric scales to use for the clustering, using a genetic algorithm to optimize a fitness function purely based on the geometric properties of the dataset, obtained at each iteration by considering the new subset of psychometric scales and discarding the others.

## II. Dataset

The dataset used is the result of a MMPI-2 psychological interview [1] on couples that were candidates for child adoption. The result of this interview process is a pool of samples each one representing the psychological profile of a single person, including the anagraphic data and the answers returned to the boolean questionnaire used for the evaluation. The clustering process is focused on the psychometric evaluation part and ignores the rest (the anagraphic data and the boolean questionnaire items). A minimal knowledge about the psychometric scales was used to divide them in the following groups:

- *Validity scales*: determine if patients are non-responding or inconsistent responding (CNS, VRIN, TRIN), overreporting or exaggerating symptoms (F, Fb, Fp, FBS) or underreporting/downplaying psychological symptoms (L, K, S)
- *Clinical scales*: diagnoses of various psychiatric conditions (Hypocondriasis ...)
- *Content scales*: Used to measure symptoms of psychiatric conditions (Anxiety, obsessivenes, ...) so they need to be read in addition to clinical scales
- *Supplemental scales*: used in supplement to the content scales to determine if some of the symptoms are caused by different additional possible causes (”controlled hostility”, alcoholism ...)
- *Psy-5 scales*: measure dimensional traits of personality disorders

Given though that without a field expert it is impossible to determine the relative importance of these scales for the current task without a process based on trial-and-error, the subdivision in scale groups is only used to ”deactivate” the whole group instead of deactivating single scales during the evolutionary optimization process. This approach though doesn’t lead to good result so it was discarded in experimentation phase.

## III. Synthetic dataset

In order to prove the efficacy of this method, a first dataset was created synthetically to establish a baseline model, by sampling data points from three normal distributions with close centroids and large variance, in order to make the task more difficult. The image below represents an instance of this kind of dataset:

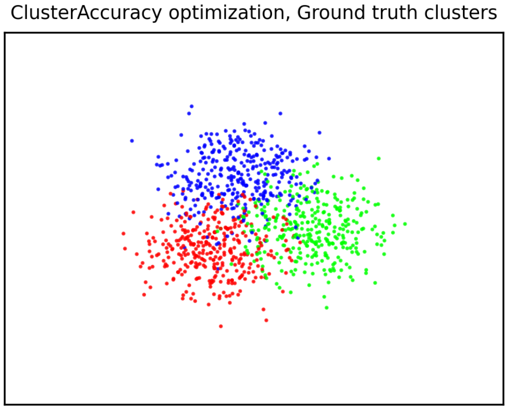


Fig. 1: Dataset created synthetically to establish a baseline performance for this algorithm

Results of this experiment are discussed in the *Experimental Analysis* section.

## IV. Dataset preprocessing

The provided dataset featured data fields that were empty for all samples or other fields that were missing values for some samples. To solve these two issues, a minimal preprocessing was required:

- **Fields** that were **completely empty** were simply removed altogether
- **Fields** that were **missing values for some samples** (which could only be found in the boolean items) could be:
  - Replaced with the most common value for that category
  - Replaced using a probabilistic approach (compute probability for both outcomes (0 or 1) for that question and toss a coin to choose the value for the empty field

  The first criterion was chosen.

## V. Clustering

The basic idea of this project revolves around determining in an unsupervised way which psychometric scales are important for the clustering process and which ones can be instead considered redundant towards the evaluation. To achieve this, the problem is formalized as ”synthesizing a binary string, where each *bit* corresponds to one of the psychometric scales and controls whether that scale is to be used or not in the clustering process, by optimizing a certain objective function”.

The optimization process is performed by using a *genetic algorithm* where the objective function is actually used as a fitness function.

The clustering process is instead performed by trying three different clustering methods (*K-Means*, *Gaussian Mixture Models* and *Spectral clustering*) and evaluating which one leads to the best fitness, according to the chosen fitness method.

In some experiments, a specific prior was assumed: given that the dataset also reports the couple instance to which the interviewed person belongs, instead of clustering based on single scale values, samples are classified based on the difference between their psychometric scales and the ones of the other component of the couple. In this way we're assuming that the attempt to falsify the results of the psychological evaluation would result in anomalies in the distribution of this inter-couple differences. The validity of this prior is verified by some of the experiments discussed in the *Experimental Analysis* section.

### A. Genetic algorithm

The optimization process is performed by using a genetic algorithm.

The main steps of the algorithm are performed as follows:

1) **Initial population generation**: an initial population of `POPULATION_SIZE` individuals is generated randomly, each individual being a binary string of `STRING_SIZE` bits.
2) **Fitness evaluation**: each population individual is evaluated according to a certain fitness function and a fitness value is assigned to each one. Fitness functions will be discussed also in the following section.
3) **Tournament selection**: `POPULATION_SIZE` pools are drawn randomly from the population, each one consisting of a certain number of individuals, and the individual with the best fitness value is selected from each pool.
4) **Crossover**: The individuals selected during the *tournament selection* step will be considered the parents of the next population generation. `POPULATION_SIZE` couples will be drawn from the set of parents. Then from each couple of parents, two children are generated in the following way:
   a) A crossover point is chosen randomly (it's the same for both strings): for example, given the two parents `11010010` and `10100101`, the chosen crossover point is after the *3rd* bit (`110|10010` and `101|00101` respectively)
   b) The two strings are split at the crossover point and the second halves are swapped among the two: the first child becomes `110|00101` and the second child becomes `101|11010`
5) **Mutation**: Each bit of the children is flipped with a certain probability
6) **Early stopping**: If there is no improvement in the best fitness value in the population after `PATIENCE` turns, break the optimization loop

### B. Fitness functions

The experiments featured various fitness functions:

- For the baseline experiments, a fitness function wrapping an accuracy metric was used on the ground truth data of synthetically created datasets. See the *Experimental Analysis* section for a discussion of the baseline experiment.

Notice that this approach was not possible on the psychometric clustering task as ground truth data was not available, so heuristic fitness estimators were needed.

- ***Minimum inter-cluster distance***: the fitness value $V$ is the minimum distance between two data points belonging to two different clusters:

$$V = \min_{\substack{s_i \in C_i \\ s_j \in C_i \\ i \neq j}} \{d(\mathbf{s_i}, \mathbf{s_j})\} \tag{1}$$

where
  - $C_i$ amd $C_j$ are different clusters in the dataset
  - $d(.,.)$ is a distance measure (in our case the simple euclidean distance was used)

.

- ***Minimum centroid distance***: the fitness value $V$ is the minimum distance between two cluster centroids:

$$V = \min_{i \neq j} \{d(\frac{\sum_{s_i \in \mathbf{C_i}} s_i}{|C_i|}, \frac{\sum_{s_j \in \mathbf{C_j}} s_j}{|C_j|})\} \tag{2}$$

### *C. Clustering*

Experiments were performed using three different clustering methods:

- **K-Means**: partitions samples into $K$ clusters (where $K$ is a given number) as follows:
  1) Choose randomly $K$ samples as initial clusters (and therefore initial centroids as well)
  2) Assign each sample to the cluster with the nearest centroid
  3) Recompute the centroids
  4) If there are no changes in the centroid positions (or the euclidean distance of the new centroids from the old ones is lower than a certain threshold) return the assignment otherwise go back to step 2.
- **Gaussian mixture model**: tries to fit a given number $K$ of *normal distributions* to subsets of the original dataset, by estimating their mean and variance using the *Expectation-Maximization* algorithm [2].
- **Spectral clustering**:
  1) Compute a *similarity matrix* of a given set of points (an evaluation of similarity for each pair of points in the dataset) or a *neighbors matrix* (built as an adjacency matrix where two data points are considered adjacent if one of them is among the k-nearest neighbors of the others)
  2) Compute eigenvalues and eigenvectors of this matrix
  3) Sort eigenvectors based on eigenvalues (in decreasing order)
  4) The first non-zero eigenvalue is called the *spectral gap* and gives us some notion of the density of the graph (the minimum number of connections all nodes have in the graph).
  5) Now, from the second non-zero eigenvalue onwards, we'll have to look for a ”gap” between subsequent eigenvalues: each eigenvector with eigenvalues smaller than the ”gap” represents a ”cut” in the graph, a subdivision of the dataset samples in two groups. Therefore having $K$ eigenvectors before this gap will mean having (most likely) $K + 1$ clusters.
  6) For each eigenvector (called a *Fiedler vector*) the value for each column of the matrix tells us, for the corresponding dataset point, to which one of the two sides of the cut the node belongs.
  7) We take these column eigenvectors and use them as a matrix to cluster the dataset using *K-Means*: the assignment for the $l$-th row of this matrix will correspond to the $l$-th sample of the dataset.

### *D. Visualization*

Some words should be spent on how the 2D and 3D visualizations are created:

1) The clustering is performed using the selected subset of fields
2) The high-dimensional dataset obtained by projecting the original dataset on the new subset of fields is then projected to a 2D or 3D space through Principal Component Analysis and the obtained distribution of points is plotted as a scatter plot.

## VI. Experimental analysis

### *A. Baseline*

Given the lack of hand-crafted ground truth data, the accuracy of this method can not be tested on the original dataset. Therefore a baseline dataset was created synthetically by sampling data points from three normal distributions, with mean $\mu \sim [-10, 10]$ and variance $\sigma = 10$.

An example of this synthetic dataset is shown in Figure 1. Three experiments were performed on this kind of dataset to establish a baseline, in order to

prove the efficacy of this optimization process. For all three instances, data was normalized and then experiments were conducted with **K-Means**, **GMM** and **Spectral clustering** algorithms. The results are shown in table 4, with *K-Means* returning the best result at an $81\%$ accuracy.
A comparison of the final result of each experiment with ground truth clusters is shown in figures , and .

### B. MMPI clustering

In order to find the best models, a full grid search was performed on the following hyperparameters:

| Hyperparameter | Possible values |
|---|---:|
| N. of clusters | 2, 4 |
| Clustering algo. | K-Means, GMM, Spectral |
| Scale groups | True, False |
| Couple differences | True, False |
| Fitness func. | Min. point dist., Centroid dist. |

The *Scale groups* parameter determines if whole scale groups or single scales should be activated/de-activated throughout the optimization. The *Couple differences* parameter instead determines whether to use values of the individual or their differences with respect to the other member of the couple, for the clustering process. A ranking of results will be discussed for each number of clusters, separately for each fitness function, as we don't have a way to compare the accuracy of the different fitness evaluation criteria.

### C. 2 clusters

The clustering into 2 clusters returns the following rankings:

- **Minimum centroid distance**: as we can see from table 12, the only model to actually show an improvement due to the optimization process is the one using K-Means clustering for differences in values of the same couple on single scales (and not scale groups), reaching a centroid distance of $93.67$ at iteration $14$. A graphical comparison of the clusters obtained at iteration 14 with respect to the clusters at the beginning of the optimization process is shown in figure 6 while figure 6 and figure **??** show respectively the dataset in 3D using all data fields per-sample and the "transformed" one obtained by projecting data points on the subset of data fields selected at the end of the optimization process.
- **Minimum inter-cluster distance**: as we can see from table 8, the best performing model with this fitness metric is the one using K-Means clustering for single values (and not couple differences), optimizing the selection of single scales, reaching a minimum inter-cluster distance of $57.49$ at iteration $10$, followed by a similar model using instead Gaussian Mixture Models, with an inter-cluster distance of $55.42$ at iteration $11$. Figure 8 shows a comparison of the clustering at the beginning and at the end of the optimization process and also in this case figure 8 and figure 8 show respectively the dataset in 3D using all data fields per-sample and using only the selected ones.

### D. 4 clusters

The clustering into 4 clusters instead returns more various results:

- **Minimum centroid distance**: as we can see from table 10, there are many models showing an actual optimization process but the best performing one uses K-Means clustering on couple differences for single scales, reaching a centroid distance of $173.85$, right at iteration $1$. The graphical comparison of the initial clusters versus the final ones is shown in figure 10. Also in this case, a visualization of the original dataset versus the transformed one are shown.
- **Minimum inter-cluster distance**: as we can see from the table, here instead the best performing method is using Gaussian Mixture Models on single samples and optimizing the usage of whole scale groups, reaching a minimum inter-cluster distance of $56.56$ at iteration $9$, followed by a similar model using instead Spectral clustering, with an inter-cluster distance of $47.68$ at iteration $1$.

## REFERENCES

[1] Mmpi-2 test. [Online]. Available: https://en.wikipedia.org/wiki/Minnesota_Multiphasic_Personality_Inventory

[2] Expectation maximization algorithm. [Online]. Available: https://en.wikipedia.org/wiki/Expectation\%E2\%80\%93maximization_algorithm

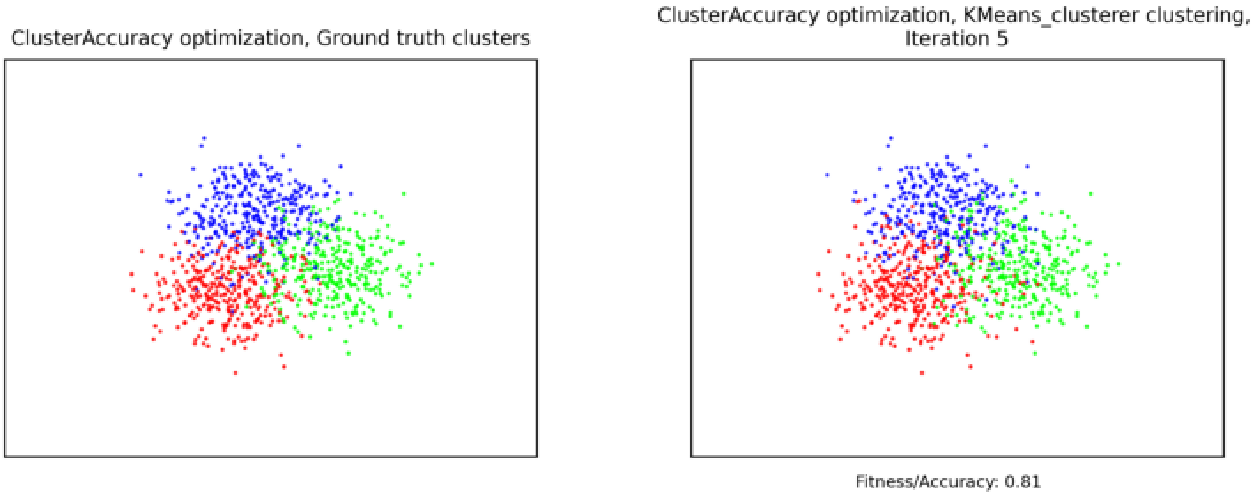


Fig. 2: K-Means baseline experiment results comparison

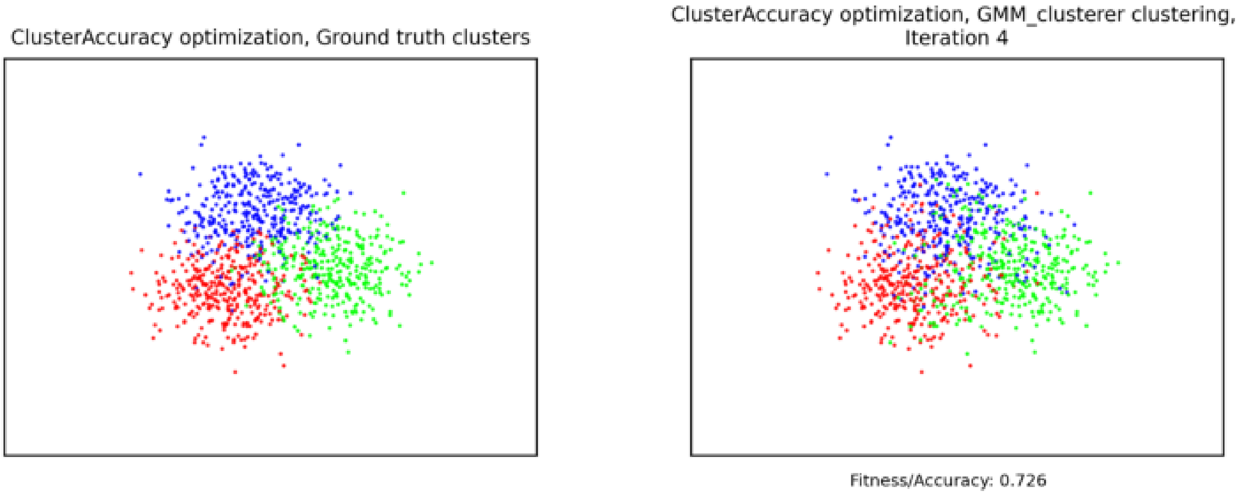


Fig. 3: GMM baseline experiment results comparison

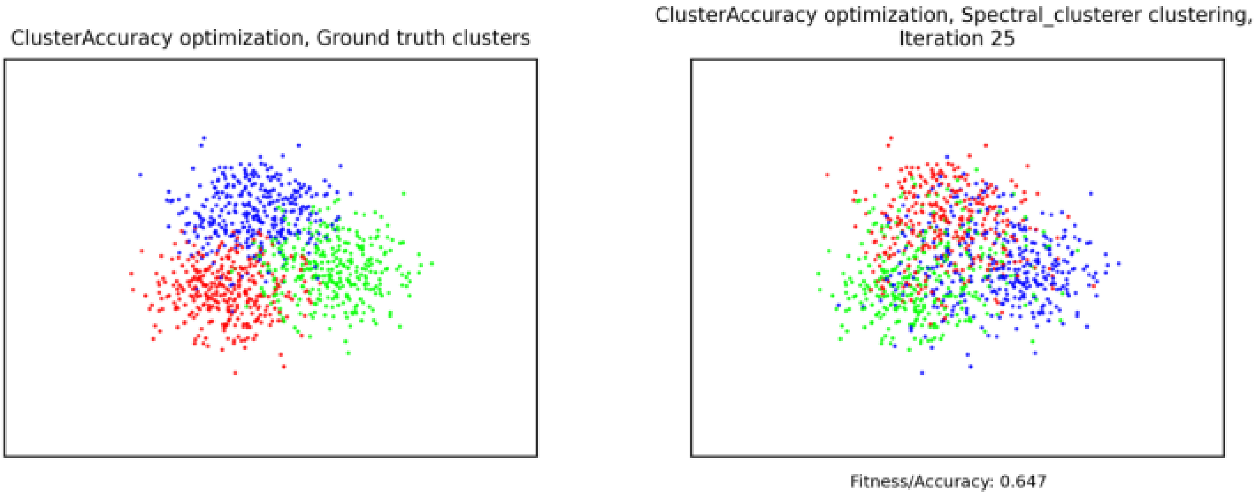


Fig. 4: Spectral clustering baseline experiment results comparison

| **Model** | **Accuracy** | **Iteration** |
|---|---:|---|
| K-Means | 81% | 5 |
| GMM | $72,6\%$ | 4 |
| Spectral | $64,7\%$ | 25 |

Fig. 5: Comparison of the results for the three baseline experiments.

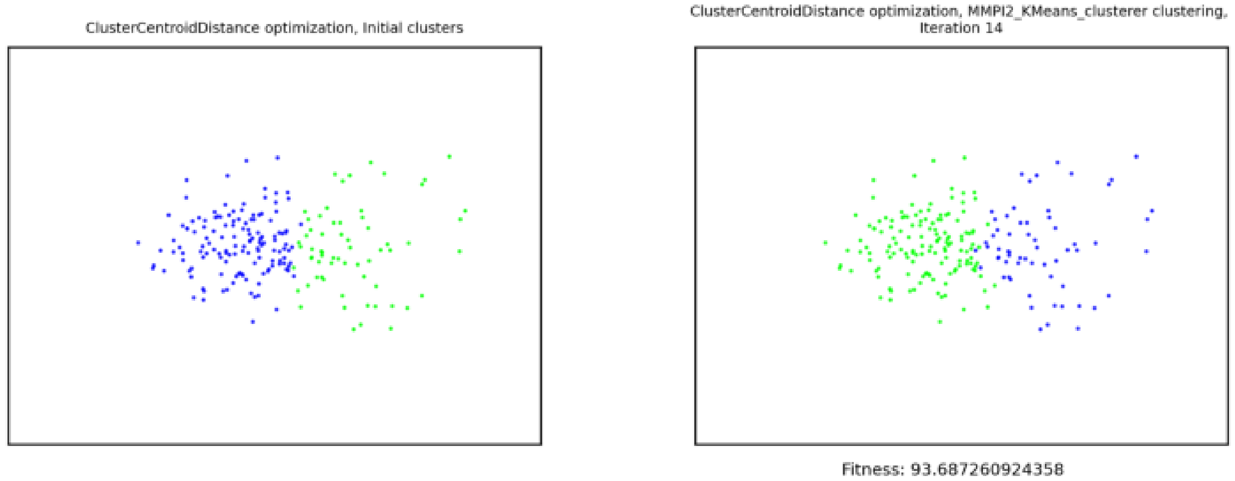
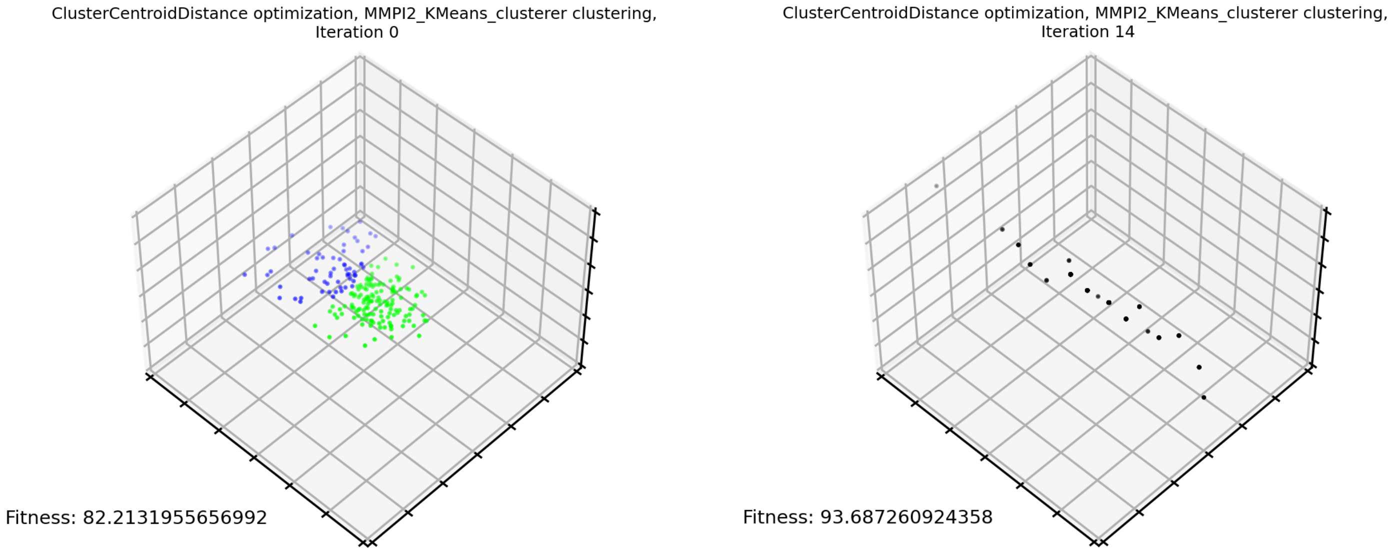


Fig. 6: Comparison of the results for the best model for the 2 clusters problem using the Centroid distance fitness function (see table below)



??

| N. clusters | Clust. algo. | Group scales | Couple difference | Fitness function | Fitness | Iteration |
|---|---|---|---|---|---|---|
| 2 | K-Means | No | Yes | Centroid distance | 93.69 | 14 |

Fig. 7: Comparison of the results for the experiments featuring 2 clusters with the Centroid distance fitness function (the other models showed no improvement).

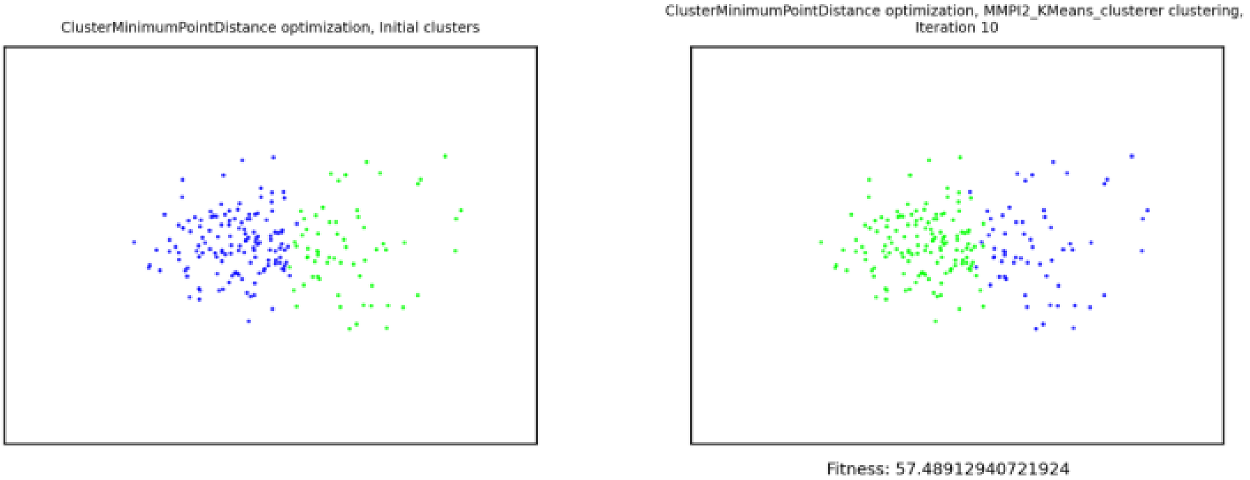
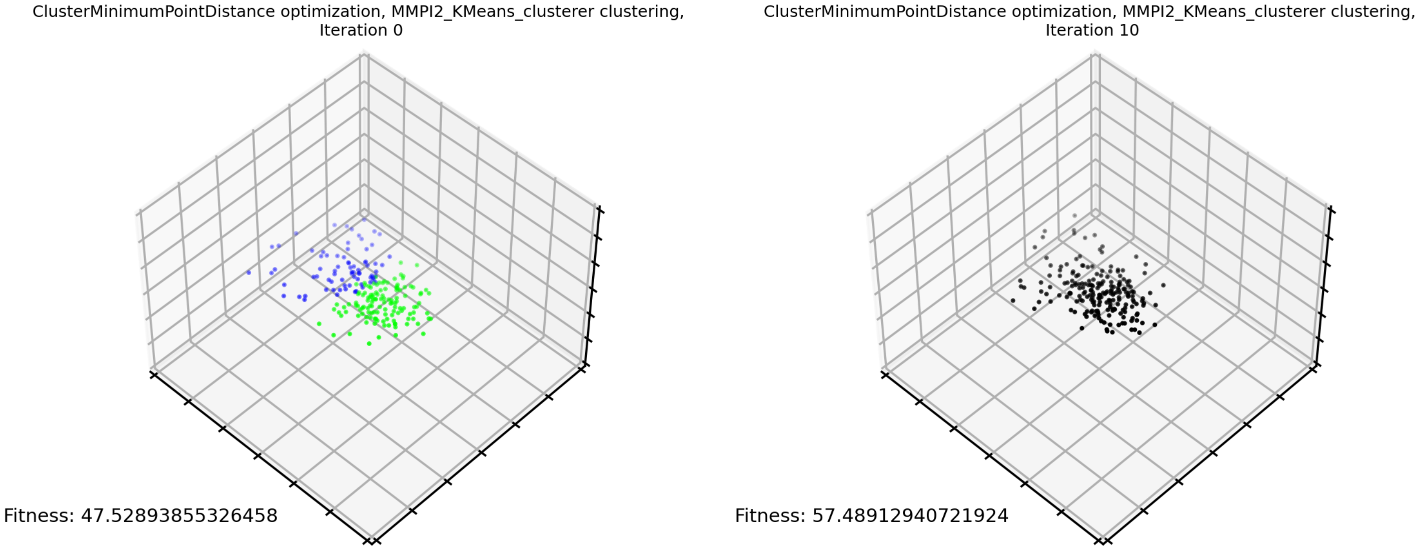


Fig. 8: Comparison of the results for the best model for the 2 clusters problem using the Minimum inter-cluster distance fitness function (see table below)



| N. clusters | Clust. algo. | Group scales | Couple diff. | Fitness function | Fitness | Iteration |
|---|---|---|---|---|---|---|
| 2 | K-Means | No | No | Min. inter-cluster dist. | 57.49 | 10 |
| 2 | GMM | No | No | Min. inter-cluster dist. | 54.41 | 11 |
| 2 | Spectral | No | Yes | Min. inter-cluster dist. | 54.35 | 8 |
| 2 | K-Means | No | Yes | Min. inter-cluster dist. | 13.42 | 13 |

Fig. 9: Comparison of the results for the experiments featuring 2 clusters with the Minimum inter-cluster distance fitness function (the other models showed no improvement).

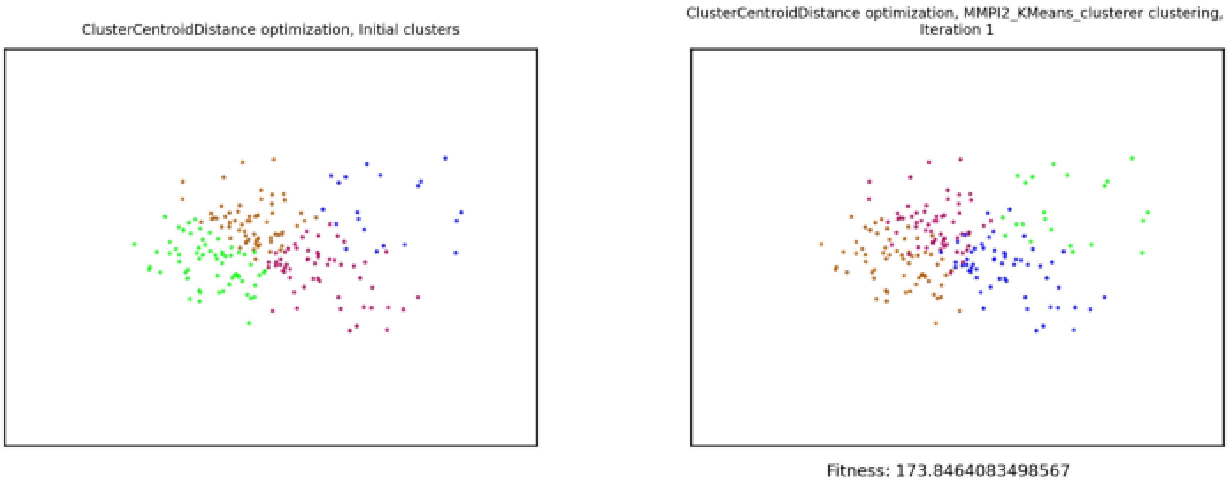
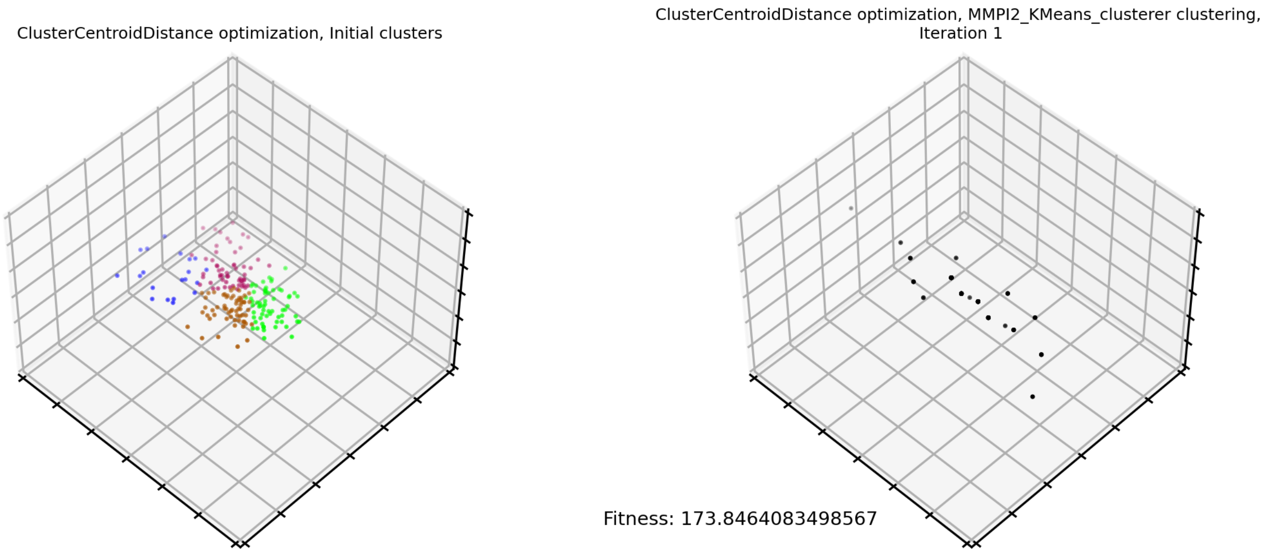


Fig. 10: Comparison of the results for the best model for the 4 clusters problem using the centroid distance fitness function (see table below)



| N. clusters | Clust. algo. | Group scales | Couple difference | Fitness function | Fitness | Iteration |
|---|---|---|---|---|---|---|
| 4 | K-Means | No | Yes | Centroid distance | 173.85 | 1 |
| 4 | GMM | No | No | Centroid distance | 100.88 | 14 |
| 4 | GMM | No | No | Centroid distance | 94.4 | 13 |
| 4 | K-Means | Yes | Yes | Centroid distance | 56.91 | 1 |
| 4 | K-Means | Yes | No | Centroid distance | 41.02 | 1 |

Fig. 11: Comparison of the results for the experiments featuring 4 clusters with the Centroid distance fitness function (the other models showed no improvement).

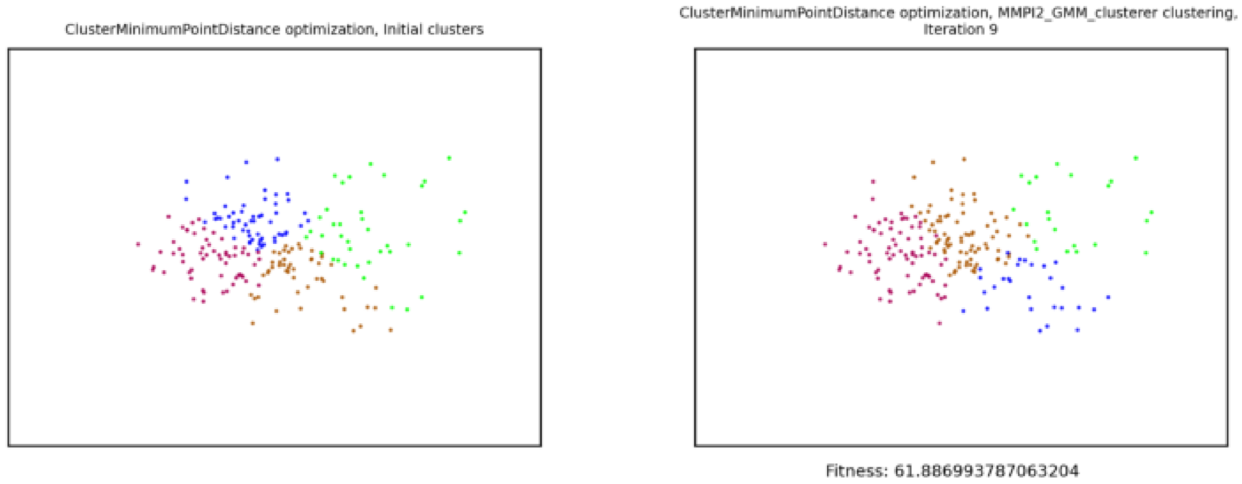
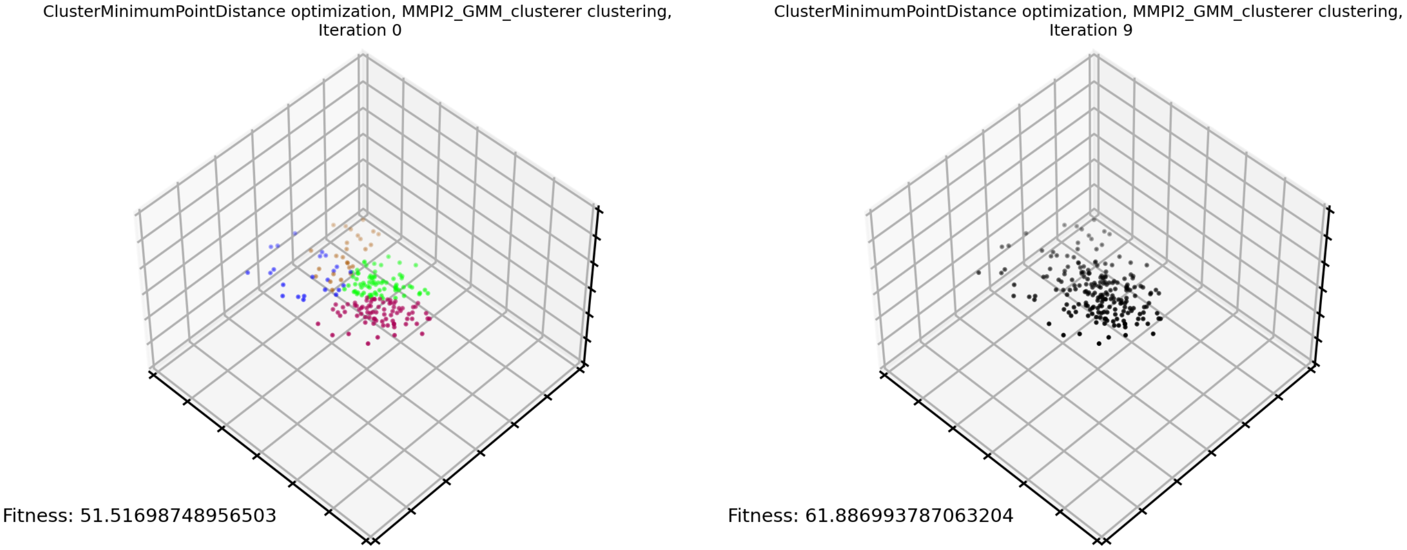


Fig. 12: Comparison of the results for the best model for the 4 clusters problem using the Minimum inter-cluster distance fitness function (see table below)



| N. clusters | Clust. algo. | Group scales | Couple diff. | Fitness function | Accuracy | Iteration |
|---|---|---|---|---|---|---|
| 4 | GMM | Yes | No | Min. inter-cluster dist. | 56.56 | 9 |
| 4 | Spectral | Yes | No | Min. inter-cluster dist. | 47.67 | 1 |
| 4 | Spectral | Yes | Yes | Min. inter-cluster dist. | 47.67 | 1 |
| 4 | GMM | No | No | Min. inter-cluster dist. | 47.41 | 10 |
| 4 | Spectral | No | Yes | Min. inter-cluster dist. | 44.61 | 9 |

Fig. 13: Comparison of the results for the experiments featuring 4 clusters with the Minimum inter-cluster distance fitness function (the other models showed no improvement).