INF a 4 - ANALYSE DE DONNEES ET RECONNAISSANCE DE FORMES

-

BE3 - ANALYSES FACTORIELLES ET CLASSIFICATION NON SUPERVISEE

Nicolas MARSHALL

Octobre 2012

Les objectifs de ce bureau d'étude sont de réaliser dans un premier temps une analyse factorielle des correspondances binaires (AFC), ainsi qu'une analyse factorielle des correspondances multiples (ACM). Dans un second temps, une classification non supervisée sera réalisée en utilisant deux méthodes : une classification ascendante hiérarchique, ainsi que l'algorithme des centres mobiles. Enfin, une combinaison des méthodes d'analyse factorielles et de classification non supervisée sera réalisée afin d'enrichir l'étude de tableaux de données.

1. AFC

1.1. Démarche

Les données du fichier « csp.mat » contiennent des effectifs au croisement de chaque paire de modalités (CSP des parents, études des enfants). Une analyse factorielle de type AFC est donc parfaitement adaptée sans traitement préalable des données. Elle se fait en plusieurs étapes :

1. Lecture des données
2. Construction du tableau des fréquences relatives
3. Construction des tableaux des profils-lignes et profils-colonnes
4. Calcul des matrices X, M et D
5. Calcul des axes factoriels et des valeurs propres associées

6. Calcul des facteurs associés aux axes factoriels
7. Calcul des indices d'aide à l'interprétation des résultats de l'analyse factorielle
8. Visualisation des résultats
9. Interprétation des résultats

NB : Dans tous les fichiers « .m », les commentaires sont mis pour expliquer à quoi correspond chaque bloc de commandes.

1.2. Résultats et Interprétation

On étudie d'abord la répartition de l'inertie sur les axes factoriels. On regarde pour cela les valeurs propres associées à ceux-ci, qui indiquent pour chaque axe l'inertie associée :

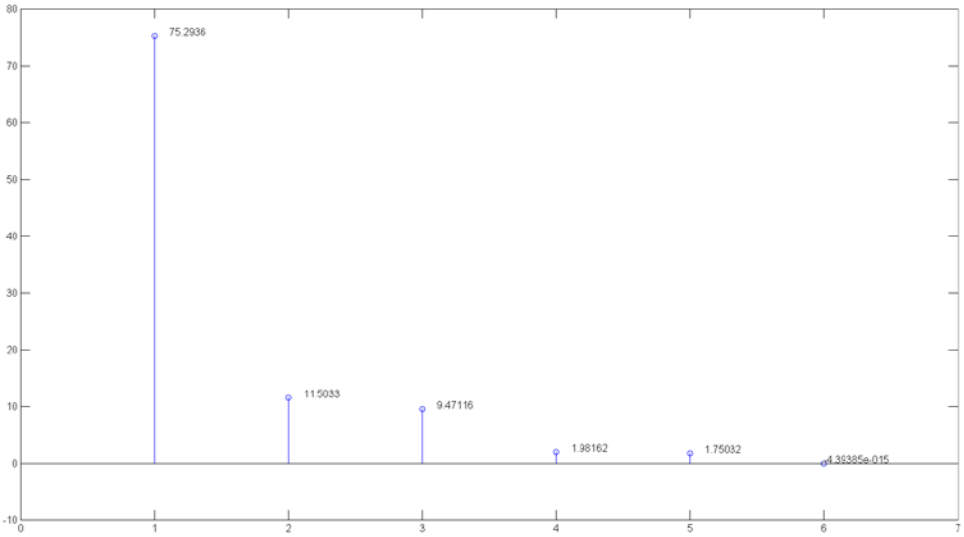

Figure 1.1 : Diagramme des valeurs propres

On s'aperçoit qu'une grande partie de l'inertie est concentrée sur les deux premiers axes factoriels (86.8 %). Les représentations sur le premier plan factoriel des individus et des variables auront donc une bonne qualité globale en termes d'inertie représentée.

Affichons maintenant ces représentations du premier plan factoriel :

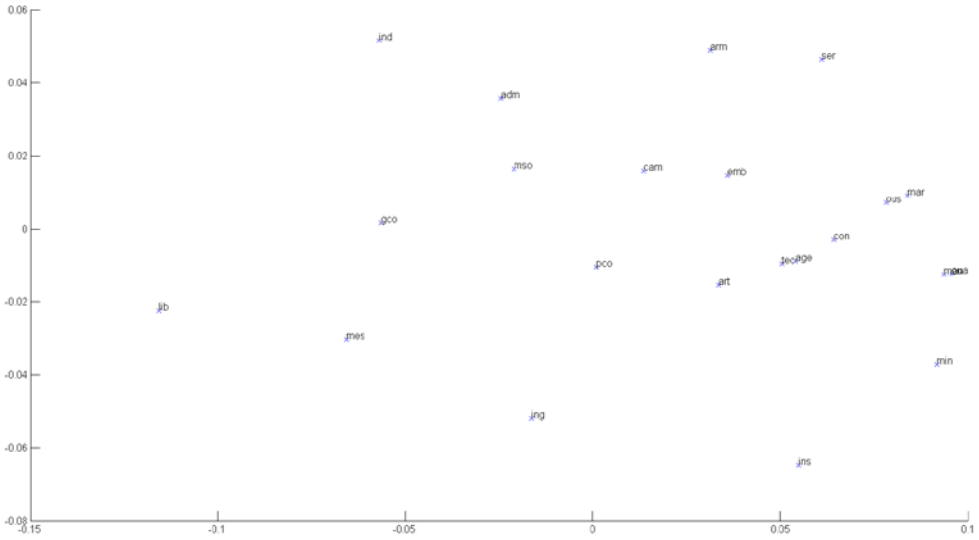

Figure 1.2 : Plan factoriel des individus

Au niveau du plan factoriel des individus, on voit deux facteurs majeurs ressortir dans la distribution des CSP des parents : de gauche à droite, le facteur principal qui joue dans le choix d'études des enfants est la rémunération des parents : à gauche, les métiers les mieux rémunérés, et à droite, les métiers rémunérés plus faiblement. De bas en haut, le second facteur identifiable est la mixité au sein de la profession : en bas, les métiers les plus mixtes et en haut, les métiers encore quasiment exclusivement masculins.

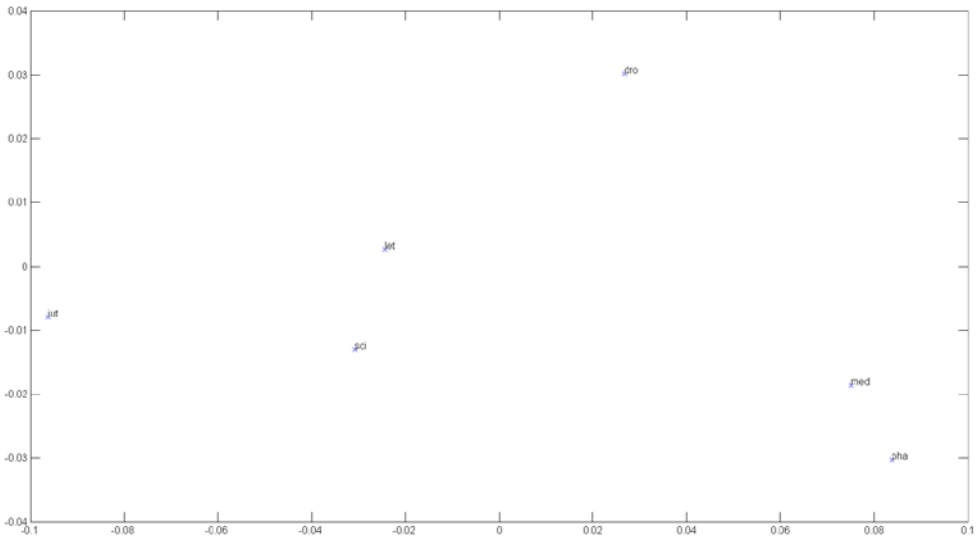

Figure 1.3 : Plan factoriel des variables

Au niveau du plan factoriel des variables, on peut distinguer visuellement trois groupes : le groupe {sci, let, iut }, où les enfants font des études onéreuses (ou relativement) et pour lesquels les parents

ont pour une bonne partie des revenus confortables. Pour les groupes {dro} et {med, pha}, les études des enfants sont en opposition avec les métiers des parents, mais correspondent globalement avec les moyens des parents comme pour le premier groupe, c'est-à-dire qu'on a surtout affaire à des études en université.

2. ACM

2.1. Démarche

Cette fois les données ne sont pas présentes initialement sous une forme adéquate. Pour pouvoir procéder à une ACM dessus, il faut au préalable les mettre en classes. Pour chaque valeur (au croisement d'un individu (CSP des parents) et d'une variable (études des enfants)), on compare la valeur au reste de la ligne (aux autres études faites par les enfants) et on lui attribue une modalité parmi les suivantes : {très faible, faible, moyen, fort, très fort}.

Une fois cette mise en classes effectuée, la démarche est celle vue en cours et elle est très similaire à la précédente pour l'AFC.

2.2. Résultats et interprétation

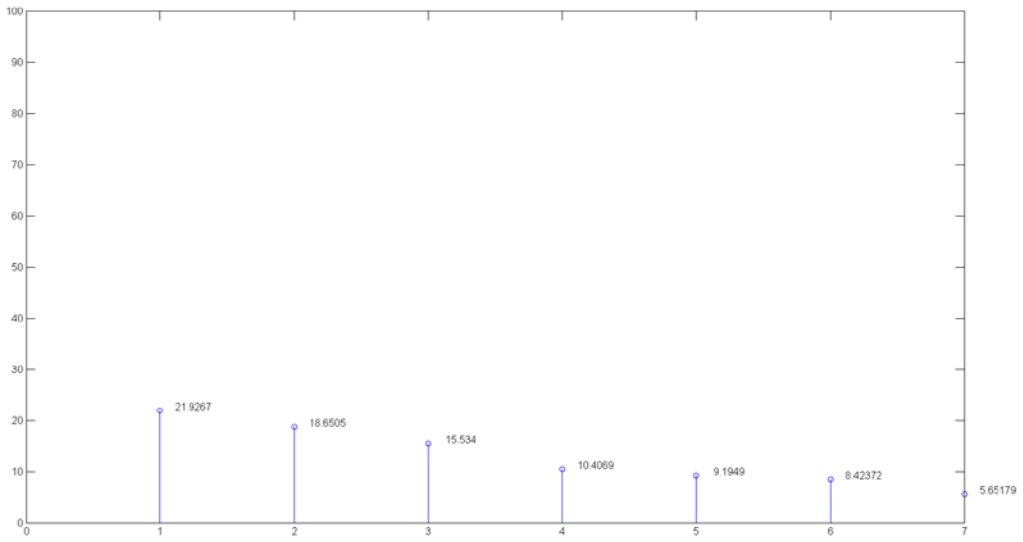

Figure 2.1 : Diagramme des valeurs propres

Cette fois, on voit une qualité de représentation en forte baisse : moins de la moitié de l'information est concentrée sur le premier plan factoriel. Les interprétations possibles des facteurs seront donc logiquement plus erronées et on en tiendra compte dans une moindre mesure que les résultats de l'AFC qui étaient pour leur part tout à fait satisfaisants.

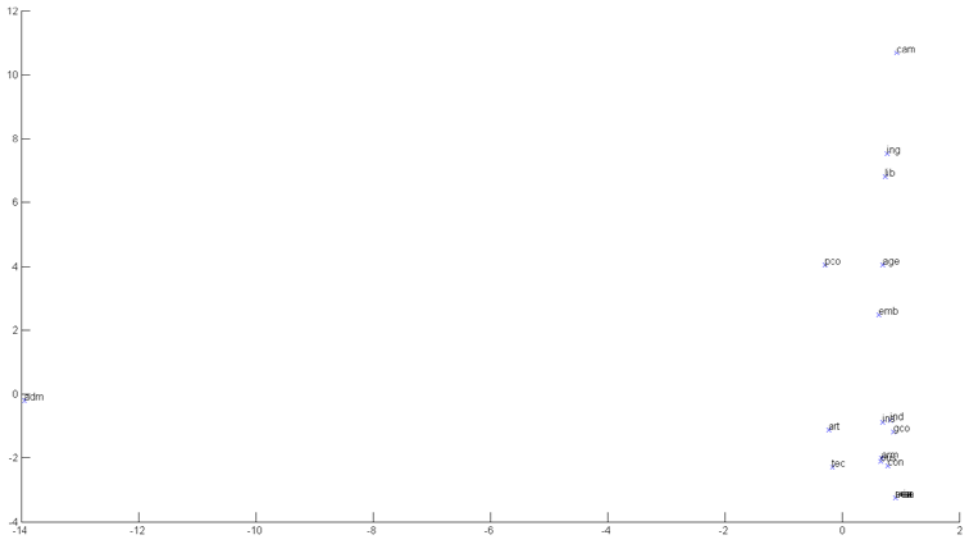

Figure 2.2 : Plan factoriel des individus

Comme prévu la représentation met moins en évidence les différences qu'il existe entre les individus : on ne distingue à peine qu'une variable importante, les individus étant tassés dans une partie de la figure à cause d'un individu rebelle (auquel on pourrait attribuer une importance moindre dans les calculs via la matrice de poids des individus) et ne montrant une variation que dans le sens vertical (second axe factoriel). La différence majeure entre les individus tient surtout de la catégorie de salaire des parents (salaires plus élevés en haut du graphique).

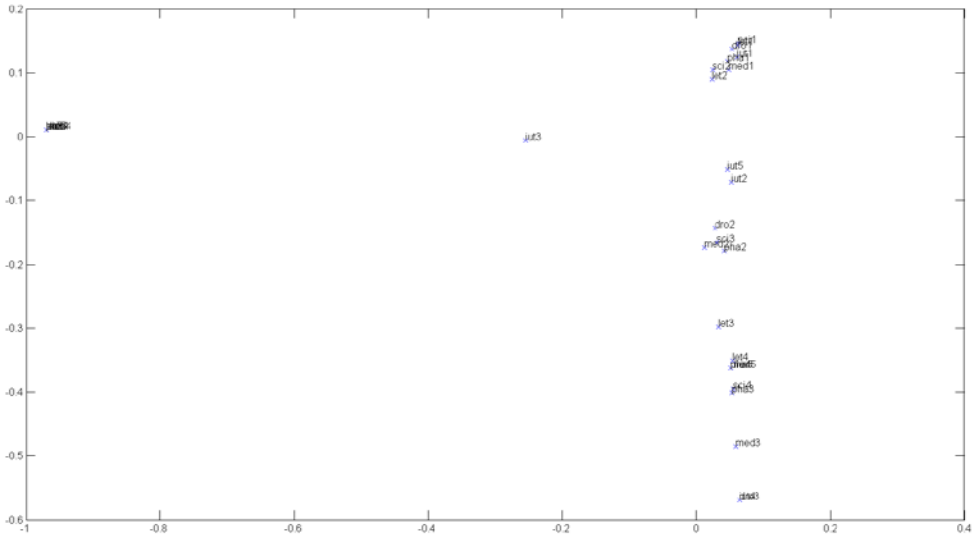

Figure 2.3 : Plan factoriel des variables

La qualité n'est pas vraiment meilleure sur ce graphe. On distingue cependant les plus grands taux d'études scientifiques (médecine et pharma inclus) vers le bas du graphique, tandis que le haut à gauche correspond plus à des études de type littéraire.

3. Classification non supervisée

3.1. CAH sur les résultats de l'AFC

Pour une classification ascendante hiérarchique on commence par réaliser un dendrogramme (fonction « linkage » sous matlab).

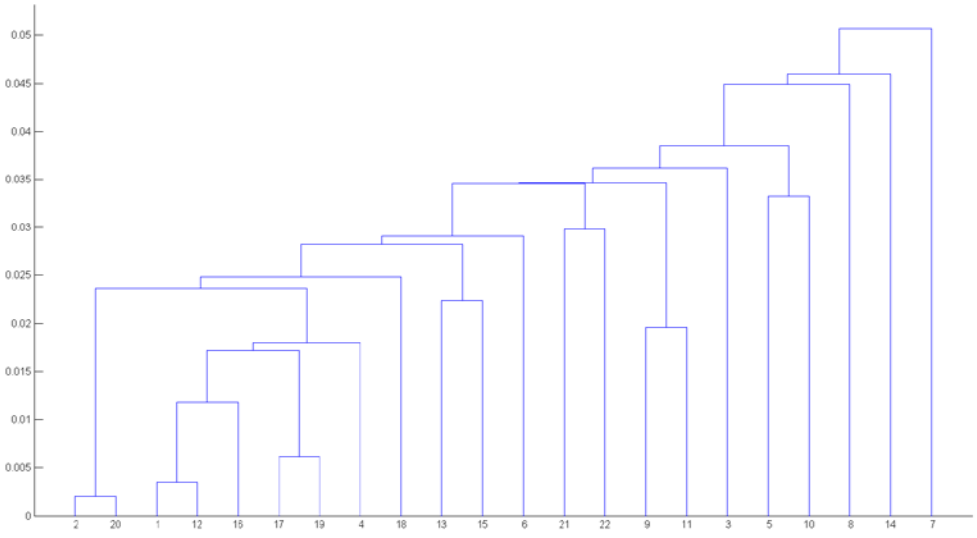

Figure 3.1 : Dendrogramme de la classification ascendante hiérarchique

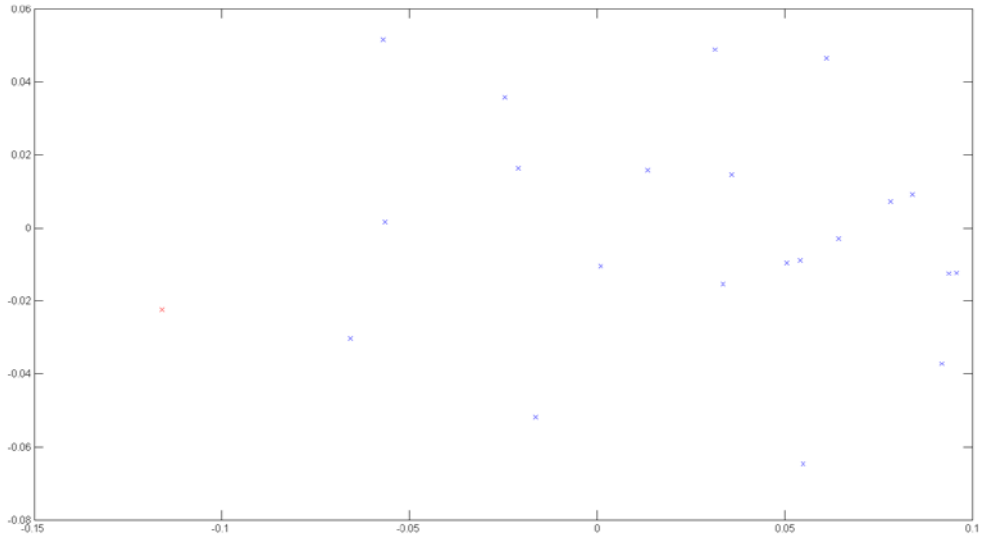

Figure 3.2 : Classification ascendante hiérarchique avec deux classes

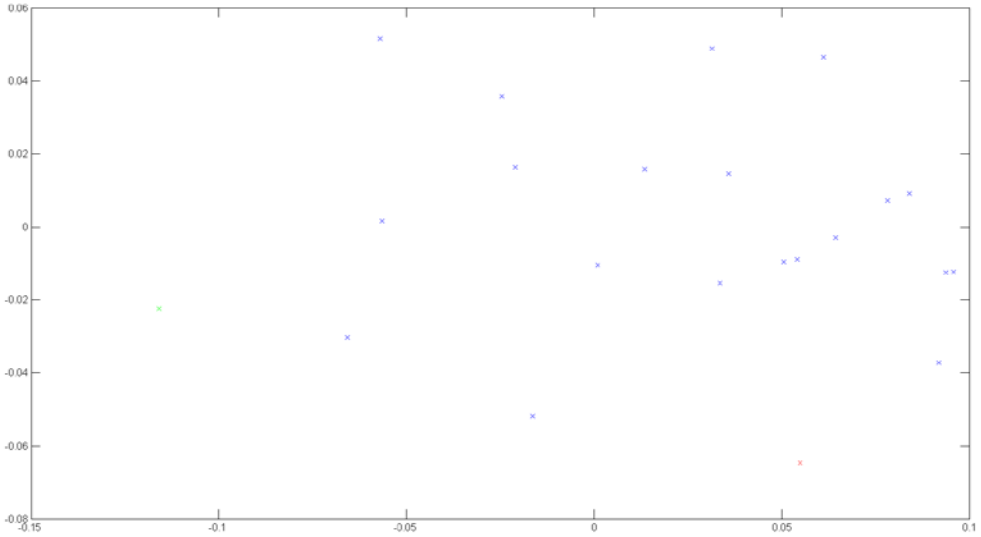

Figure 3.3 : Classification ascendante hiérarchique avec trois classes

Après avoir séparé les points en classes dans le diagramme de sortie de l'AFC, on se rend compte au premier coup d'œil de la particularité de la classification ascendante hiérarchique : les effectifs des classes sont très disparates dans le cas général, puisque cette classification repose sur les distances entre les éléments. On peut donc penser que le cas dans lequel on se retrouve ici (éléments les plus loin du peloton formant une classe à eux tout seul) doit être assez probable si ce genre de classification est utilisé seul. On comprend donc l'intérêt de combiner les différents types de classifications.

3.2. Centre mobiles sur les résultats de l'AFC

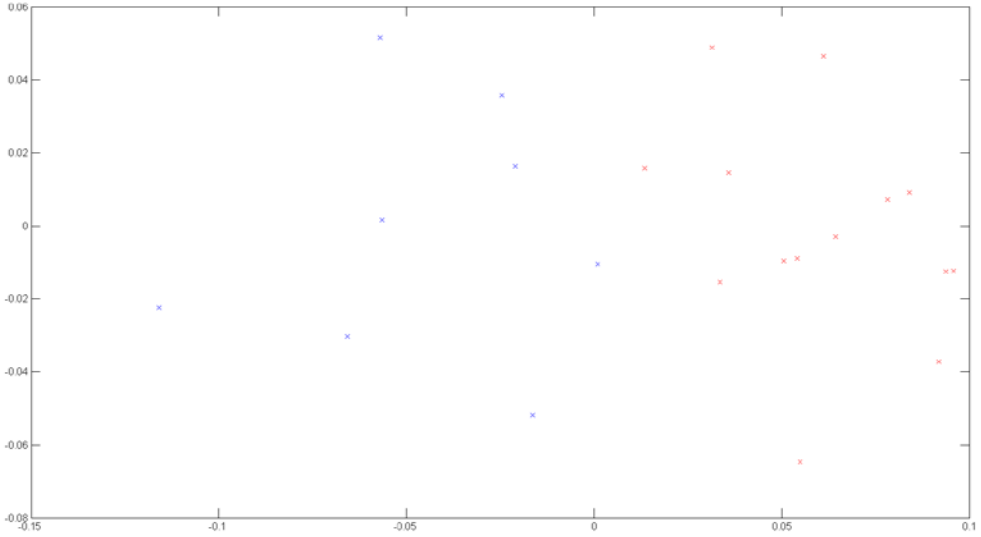

Figure 3.4 : Centres mobiles avec deux classes

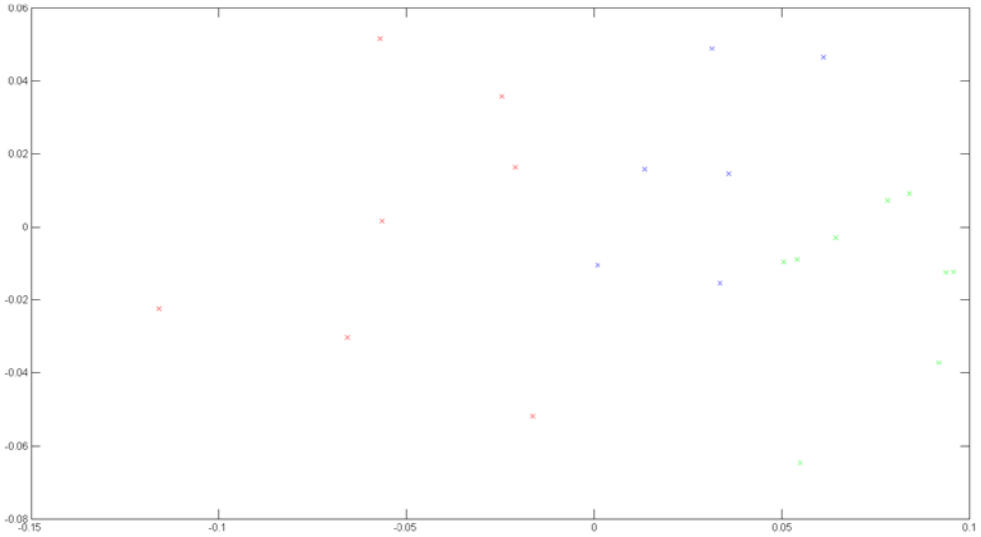

Figure 3.5 : Centres mobiles avec trois classes

On réalise cette fois une classification par la méthode des centres mobiles, et on voit que la répartition des effectifs est plus équilibrée. Cette classification est plus proche de celle qu'aurait pu faire un humain, pourtant la variation intra-classe est parfois assez grande, presque comparable à l'écart interclasse. On peut toujours combiner cette méthode avec une autre pour obtenir de meilleurs résultats, mais comme ceci dépend grandement des échantillons et des analyses faites précédemment dessus, il n'existe pas de « meilleure méthode » de classification.

3.3. CAH sur les résultats de l'ACM

Ici aussi on commence par réaliser un dendrogramme des résultats obtenus par l'ACM.

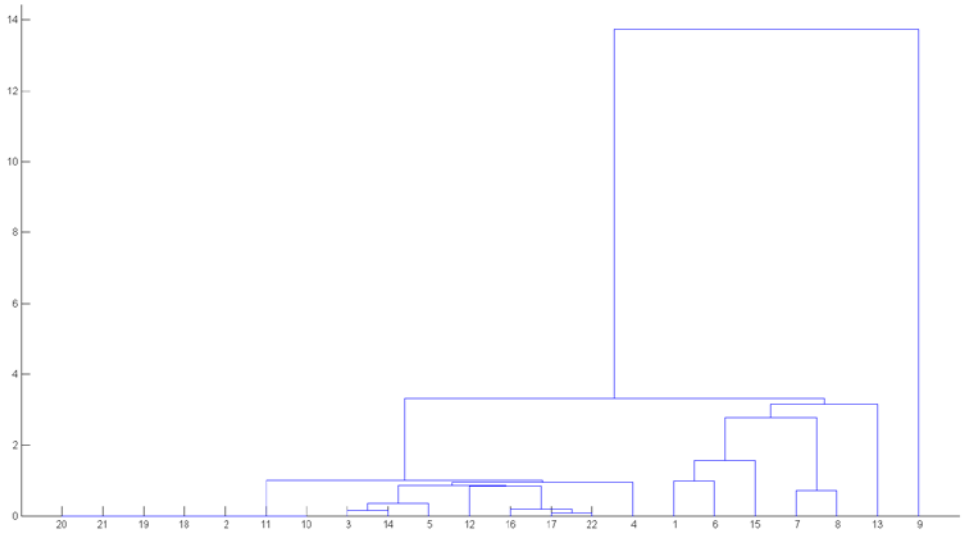

Figure 3.6 : Dendrogramme de la classification ascendante hiérarchique

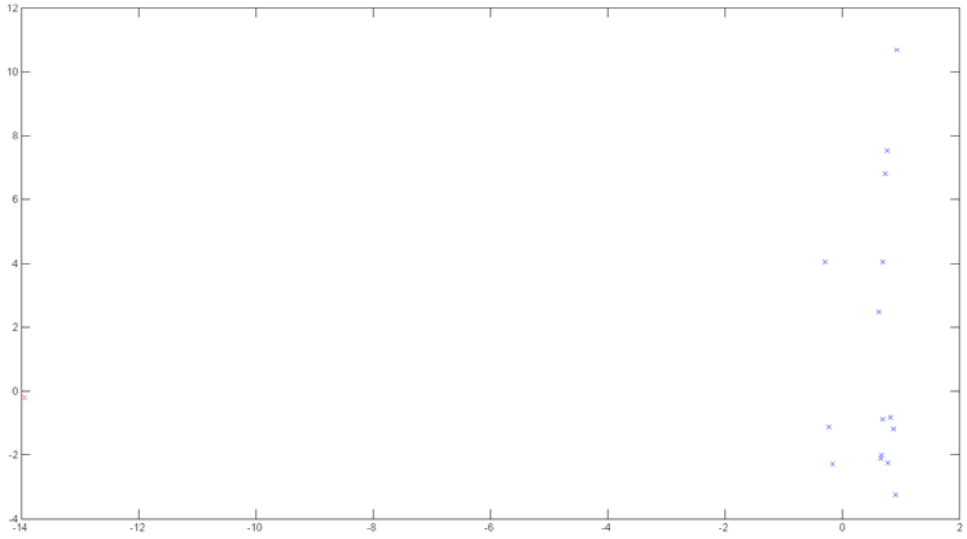

Figure 3.7 : Classification ascendante hiérarchique avec deux classes

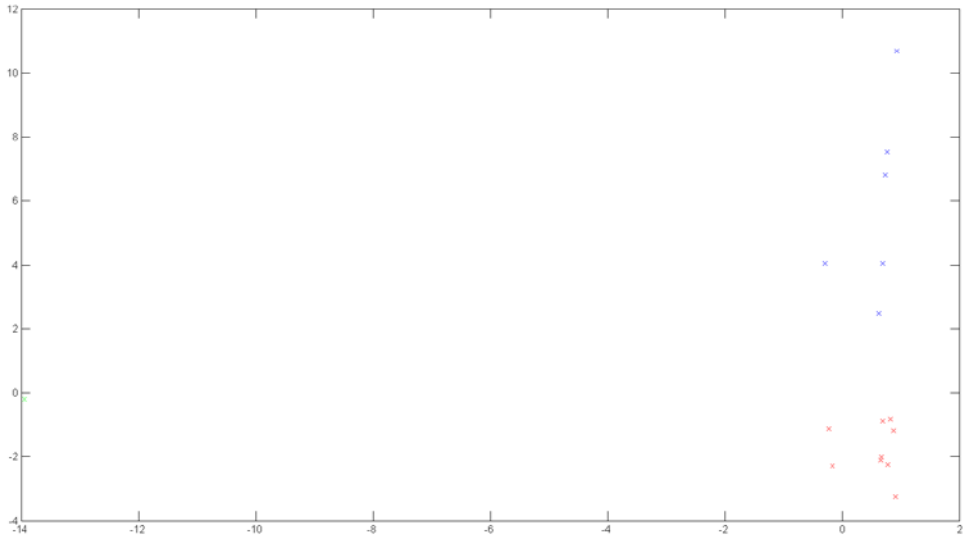

Figure 3.8 : Classification ascendante hiérarchique avec trois classes

Ici, avec deux classes on obtient des résultats comparables à la CAH précédente (c'est-à-dire que l'un des points forme une classe à lui tout seul), mais on observe une meilleure répartition dans le cas de trois classes, ce qui montre que la méthode de classification peut s'avérer pertinente pour certains sets de données en entrée.

3.4. Centre mobiles sur les résultats de l'ACM

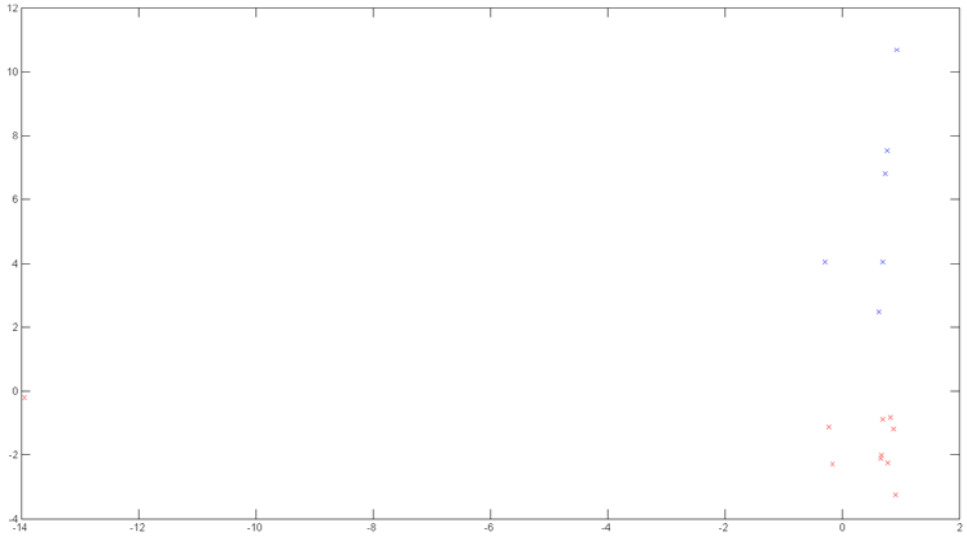

Figure 3.9 : Centres mobiles avec deux classes

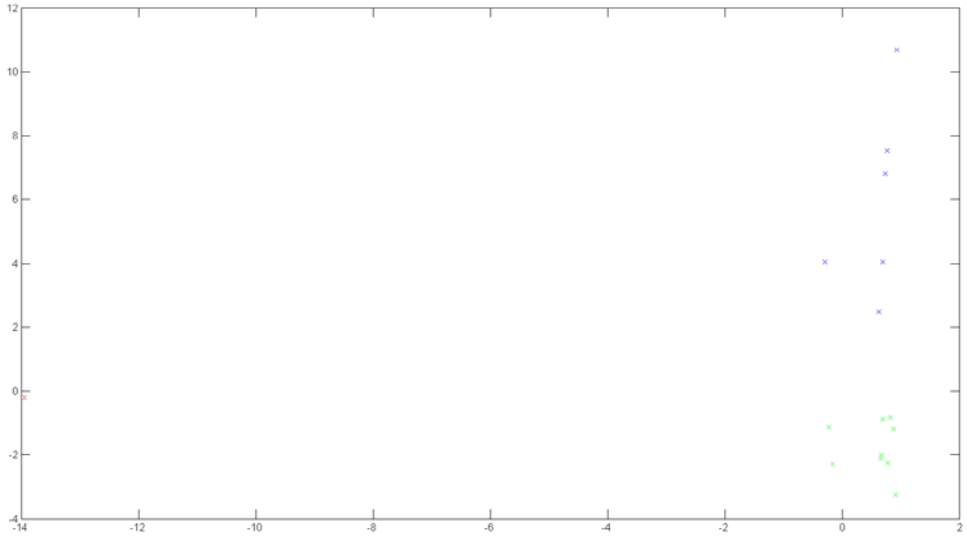

Figure 3.10 : Centres mobiles avec trois classes

La répartition est encore proche ici de ce qu'aurait fait un humain. De manière générale, les commentaires faits pour le cas de l'AFC restent valables ici.