1 Analyse Factorielle des Correspondances

L'analyse factorielle des correspondances noté A.F.C. d'une manière similaire à L'ACP vise à rassembler en un nombre réduit de dimensions, la plus grande partie de l'information initiale. La différence entre ces deux méthodes est que l'AFC ne dépend pas des valeurs absolues mais dépend des correspondances entre les variables, c'est-à-dire aux valeurs relatives. Cette réduction est d'autant plus utile que le nombre de dimensions initial est élevé. La notion de " réduction " est commune à toutes les techniques factorielles c'est-à-dire où l'on extrait des facteurs.

L'AFC offre la particularité (contrairement à l'ACP) de fournir un espace de représentation commun aux variables et aux individus. Pour cela l'AFC raisonne à partir de tableau de fréquences.

1.1 Présentation des données

Si on disposition n individus sur lesquelles on observe deux variables qualitatives, on peut résumé ces observations dans un tableau dit tableau de contingence ou de **correspondance** et cela on exprimant toutes les modalités des deux variables puis en calculant le n_{ij} qui représente le nombre d'individus possédant à la fois la modalité i et la modalité j. D'où l'appellation d'effectif conjoint. Au final on obtient le tableau à double entrées (voir la figure 1.1). L'application de l'ACP sur ce genre de tableau est impossible du fait que les modalités sont nominales et aussi que les n_{ij} ne représentent pas des mesures des variables étudier. L'analyse Factorielle des Correspondances s'applique a un tableau de ce genre.

Le tableau de contingence noté X (voir la figure 1.1) se présente sous la forme

		V 2							
		Modalité1	**Modalité2**	**Modalité3**	**...**	**Modalité j**	**...**	**Modalité q**	**Effectif marginal**
V1	**Modalité1**	n_{11}	n_{12}	n_{13}		n_{1j}		n_{1q}	$n_{1.}$
	Modalité2	n_{21}	n_{22}	n_{23}		n_{2j}		n_{2q}	$n_{2.}$
	Modalité3	n_{31}	n_{32}	n_{33}		n_{3j}		n_{3q}	$n_{3.}$
	Modalité i	n_{i1}	n_{i2}	n_{i3}		n_{ij}		n_{iq}	$n_{i.}$
	Modalité p	n_{p1}	n_{p2}	n_{p3}		n_{pj}		n_{pq}	$n_{p.}$
	Effectif marginal	$n_{.1}$	$n_{.2}$	$n_{.3}$		$n_{.j}$		$n_{.q}$	n

Figure 1.1 – Tableau de contingence

de matrice à p lignes et q colonnes (voir la matrice 1.1).

$$X = \begin{bmatrix} n_{11} & n_{12} & \dots & n_{1j} & \dots & n_{1q} \\ n_{21} & n_{22} & \dots & n_{2j} & \dots & n_{2q} \\ \vdots & \ddots & & \ddots & \ddots & \\ n_{i1} & n_{i2} & \dots & n_{ij} & \dots & n_{iq} \\ . & . & \dots & . & \dots & . \\ n_{p1} & n_{p2} & \dots & n_{pj} & \dots & n_{pq} \end{bmatrix}. \tag{1.1}$$

On déduit d'un tableau de contingence le tableau des fréquence noté $F = \frac{X}{n}$ d'élément $f_{ij} = \frac{n_{ij}}{n}$ où $n = \sum_{i,j} n_{ij}$ et $\sum_{i,j} f_{ij} = 1$.

Exemple 1 ***Souffrances psychiques et troubles du développement chez l'enfant et l'adolescent.***

Le tableau des données (voir la figure 1.1) est le résultat de dépistage effectuer sur un échantillon de 3100 élèves de la commune de Biskra en 2018 pour bute d'étudier les éventuels relations entre les troubles psychologiques et les niveaux scolaires.

Les différents troubles sont

Niveaux / Troubles psychologiques	Primaire (filles)	Primaire (garçons)	Moyen (filles)	Moyen (garçons)	Secondaire (filles)	Secondaire (garçons)
Bégaiement	7	2	4	5	0	2
Troubles de l'articulation	22	36	2	6	0	0
Dyslexie	38	34	0	1	0	0
Encoprésie	2	1	0	1	0	1
Enurésie	14	22	10	15	6	2
Retard scolaire	40	38	30	32	12	14
Echec scolaire	33	30	22	20	20	18
Refus scolaire	2	6	14	24	3	22
Phobie scolaire	8	1	0	0	0	0
Angoisse	0	0	45	40	80	68
Anxiété	0	0	40	66	83	73
Phobie	3	4	6	1	0	1
Hystérie	0	0	2	0	4	0
Névrotiques obsessionnelles	0	2	8	2	3	6
Agressivité	5	20	2	40	0	8
Hyperactivité	32	66	4	12	0	10
Fugue scolaire	0	0	2	0	1	2
Asthme	2	4	4	2	5	6
Migraines	0	0	0	28	38	2
Diabète	4	3	4	4	2	2
Maltraitance à l'école	6	4	2	1	0	0
Viol	2	2	6	4	0	0
Etat dépressif	0	0	4	1	6	2

FIGURE 1.2 – Tableau des souffrances des élèves de Biskra

Troubles du langage oral et écrit : Bégaiement, Troubles de l'articulation, Dyslexie.

Troubles sphinctériens : Encoprésie, Enurésie.

Difficultés scolaires : Retard scolaire, Échec scolaire, Refus scolaire, Phobie scolaire.

Troubles Névrotiques : Angoisse, Anxiété, Phobie, Hystérie, Névrotique obsessionnelle.

Troubles du comportement : Agressivité/Violence, Hyperactivité, Fugue.

Troubles psychosomatiques : Asthme, Migraines, Diabète.

Maltraitance : maltraitance à l'école, viol/sexuel.

Troubles de l'humeur : État dépressif.

Donc on veut étudier les liaisons possibles entre les modalités des deux variables qualitatives "trouble" et "niveau scolaire".

Les effectifs marginaux

Les effectifs marginaux du tableau X sont données par le vecteur ligne d'éléments $n_{\bullet j} = \sum_{i=1}^{p} n_{ij},\ j = 1, \dots q$ et le vecteur colonne d'éléments $n_{i\bullet} = \sum_{j=1}^{q} n_{ij}\ i = 1, \dots p$. Aussi les effectifs marginaux du tableau F (le tableau des fréquences) sont données par le vecteur ligne d'éléments $f_{\bullet j} = \sum_{i=1}^{p} \frac{n_{\bullet j}}{n}$ et le vecteur colonne d'éléments $f_{i\bullet} = \sum_{j=1}^{q} \frac{n_{i\bullet}}{n}$.

Dans ce qui suit, on défini un point de vue général du tableau dit profil.

1.1.1 Profil ligne et profil colonne

Pour analyser le tableau de contingence, ce ne sont pas les effectifs brutes qui nous intéresse, mais les répartitions en proportion à l'intérieure d'une ligne ou d'une colonne. En effet, pour mesurer les distances entre modalités, il est nécessaire de calculer au préalable la distribution de chaque modalité d'une variable en fonction de l'autre variable. Les profils (lignes ou colonne) sont les distributions des modalités des deux variables. Un profil ligne est la distribution conditionnelle de la deuxième variable sachant que l'on possède la modalité i de la première variable. Il est déterminé en divisant la ligne par la fréquence total de la ligne.
Soit le profil ligne "i" donnée par

$$i = \left(\frac{f_{i1}}{f_{i\bullet}}, \frac{f_{i2}}{f_{i\bullet}}, \cdots, \frac{f_{iq}}{f_{i\bullet}} \right).$$

Un profil colonne est est la distribution conditionnelle de la première variable sachant que l'on possède la modalité j de la deuxième variable. Il est déterminé en divisant la colonne par la fréquence totale de la colonne.
Soit le profil colonne "j"

$$j = \left(\frac{f_{1j}}{f_{\bullet j}}, \frac{f_{2j}}{f_{\bullet j}}, \cdots, \frac{f_{pj}}{f_{\bullet j}} \right).$$

Profil ligne moyen est la distribution marginale de la deuxième variable ; autrement dit ; c'est le vecteur ligne des effectifs marginaux du tableau des fréquences conditionnelles

$$G_p = (f_{\bullet 1}, f_{\bullet 2}, \cdots, f_{\bullet q}).$$

Profil colonne moyen est est la distribution marginale de la première variable ; autrement dit ; c'est le vecteur colonne des effectifs marginaux du tableau des fréquences conditionnelles

$$G_q = (f_{1\bullet}, f_{2\bullet}, \cdots, f_{p\bullet}).$$

Lorsqu'on parle de liaison, on pense toujours à l'utilisation du terme indépendance. C'est-à-dire qu'avant de traité l'existence de liaison, on doit vérifier si les deux variables sont indépendantes ou non. Et du fait qu'on travail avec deux variables qualitatives, on utilise le test du χ^2 donné si dessous.

1.1.2 Test d'indépendance du χ^2

Ce test permet de contrôler l'indépendance de deux caractères (variable qualitatives) d'une population. En supposant que l'hypothèse nulle est définie par H_0 : "Les deux variables sont indépendantes".
En général, on prouve l'indépendance entre deux variables $(V_1,\ V_2)$ en utilisant la formule suivante

$$f_{ij} = f_{i\bullet} f_{\bullet j}.$$

C'est-à-dire que la fréquence conjointe est égale au produit des fréquences marginales. Mais, la définition d'indépendance qui nous intéresse le plus est celle qui utilise les fréquence conditionnelle donné comme suit

$$f_{i|j} = \frac{f_{ij}}{f_{\bullet j}} = f_{i\bullet}.$$

C'est-à-dire que la fréquence conditionnelle est égale à la fréquence marginale.
Dans le cas théorique, si on considère que H_0 est vraie, cela implique que le tableau sera constituer des cases obtenues à laide de la formule

$$n_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n} \Leftrightarrow f_{ij} = f_{i\bullet} f_{\bullet j}.$$

L'idée est de comparer les fréquences théoriques $(f_{i\bullet} f_{\bullet j})$ aux fréquences observées (f_{ij}). D'où l'écriture de la statistique du test sous la forme suivante

$$\chi^2_{obs} = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n})^2}{\frac{n_{i\bullet} n_{\bullet j}}{n}}$$

$$= \sum_{i=1}^{p} \sum_{j=1}^{q} n \frac{(f_{ij} - f_{i\bullet} f_{\bullet j})^2}{f_{i\bullet} f_{\bullet j}}$$

$$= n\phi^2.$$

Sous H_0, cette variable suit la loi de χ^2 à $(p-1)(q-1)$ degré de liberté. ϕ^2 représente l'intensité de l'indépendance. Sur une table de la loi du khi deux, on extrait $\chi^2_{\alpha,k}$ où α représente le risque du premier genre et k représente le degré de liberté.

— Si $\chi^2_{obs} \leq \chi^2_{\alpha,k}$ on accepte H_0 et on dit qu'il y a indépendance entre les variables avec une probabilité α.
— Si $\chi^2_{obs} > \chi^2_{\alpha,k}$ on rejette H_0 et on dit qu'il n'y a pas d'indépendance entre les deux variables. Dans ce cas on peut continuer l'étude.

Exemple 2 *Rappelons que l'objectif dans cet exemple est d'étudier la liaison entre les troubles psychologiques et les niveaux scolaires. D'abord il faut que cette liaison soit significative, c'est pour quoi on applique le test du* χ^2 *sur notre tableau. Le logiciel R permet de calculer la valeur du* χ^2 *qui est égale à 1107.9.Le logiciel donne la P-valeur* $< 2.2e-16$. *Cela signifie que la probabilité que les variables soient indépendantes est inférieure à 2.2e-16. Ce qui nous permet de rejeter l'hypothèse d'indépendance entre les troubles psychologiques et les niveaux scolaires. En résumé, du fait que les variables ne sont pas indépendante avec une probabilité* 99%, *on peut exécuter une AFC pour obtenir notre dépendance.*

1.2 Procédé de l'A.F.C.

Nous avons vus plus haut que pour étudier l'intensité de l'indépendance, nous avons us recours à la définition de l'indépendance donnée par les marges du tableau. Donc Pour dire que nos variables sont liée avec une grande intensité ou avec une petite intensité, il faut qu'on compare la distance qui sépare toutes les lignes (profil ligne) avec la marge ligne (profil ligne moyen), et toutes les colonnes (profil colonne) avec la marge colonne (profil colonne moyen). C'est une approche multidimensionnelle qui utilise la distance du χ^2 pour étudier l'écart à l'indépendance.

1.2.1 Définition de la distance khi deux

Il faut toujours définir la distance utiliser. Pour G_p le profil ligne moyen, soit le profil ligne i. La distance entre le profil i et le profil G_p est donnée par

$$d\chi^2(i, G_p) = \sum_{j=1}^{q} \frac{1}{f_{ij}} \left(\frac{f_{ij}}{f_{i\bullet}} - f_{\bullet j} \right)^2 .$$

De même si on prend deux profils ligne i et $\grave{i}$, la distance entre ces deux profils est donnée par

$$d\chi^2(i,\grave{i}) = \sum_{j=1}^{q} \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - \frac{f_{\grave{i}j}}{f_{\grave{i}\bullet}} \right)^2 .$$

Ici $\frac{1}{f_{\bullet j}}$ est le pois du profil i.
En ce qui concerne G_q le profil colonne moyen, soit le profil colonne j. La distance entre le profil j et le profil G_q est donnée par

$$d\chi^2(i,G_p) = \sum_{i=1}^{p} \frac{1}{f_{i\bullet}} \left(\frac{f_{ij}}{f_{\bullet j}} - f_{i\bullet} \right)^2 .$$

La distance entre deux profils colonne est donnée par

$$d\chi^2(j,\grave{j}) = \sum_{i=1}^{p} \frac{1}{f_{i\bullet}} \left(\frac{f_{ij}}{f_{\bullet j}} - \frac{f_{i\grave{j}}}{f_{\bullet\grave{j}}} \right)^2 .$$

Après avoir défini une distance, on commence notre étude par l'exécution d'une ACP deux fois.

1.2.2 Double A.C.P.

D'un seul tableau de contingence, on ait sortie avec deux tableaux , le premier, celui des profils lignes et le deuxième résume les profils colonnes. Ici on voit bien que nous allons utilisé l'A.C.P. sur les deux tableaux pour aboutir à des représentations graphiques.

Comme dans le premier chapitre, on définit le centre de gravité à l'aide de la distance du χ^2.

Centre de gravité

En A.F.C. nous avons deux centres de gravité, pour

1. Pour le profil ligne, le centre de gravité de nuage N_p est le profil ligne moyen G_p.
2. Pour le profil colonne, le centre de gravité de nuage N_q est le profil ligne moyen G_q.

En A.C.P. pour choisir les axes, il faut calculer l'inertie du nuage. Ici aussi nous avons l'inertie du nuage mais pour deux tableaux.

Inertie des nuages N_p, N_q

1. Pour le profil ligne, soit I_p l'inertie du nuage N_p calculer par rapport au centre de gravité G_p

$$\begin{aligned} I_p &= \text{Inertie}(N_p|G_p) \\ &= \sum_{i=1}^{p} \text{Inertie}(i|G_p) \\ &= \sum_{i=1}^{p} f_{i\bullet} d^2(i, G_p) \\ &= \sum_{i=1}^{p} f_{i\bullet} \sum_{j=1}^{q} \frac{1}{f_{i\bullet}} \left(\frac{f_{ij}}{f_{i\bullet}} - f_{\bullet j} \right)^2 \\ &= \sum_{i=1}^{p} \sum_{j=1}^{q} \left(\frac{f_{ij} - f_{i\bullet} f_{\bullet j}}{f_{i\bullet} f_{\bullet j}} \right) \\ &= \phi^2. \end{aligned}$$

2. Pour le profil colonne. Soit I_p l'inertie du nuage N_q calculer par rapport au centre de gravité G_p

$$\begin{aligned} I_q =& \text{Inertie}(N_q|G_q) \\ &= \sum_{j=1}^{q} \text{Inertie}(j|G_q) \\ &= \sum_{j=1}^{q} f_{\bullet j} d^2(j, G_q) \\ &= \sum_{j=1}^{q} f_{\bullet j} \sum_{i=1}^{p} \frac{1}{f_{\bullet j}} (\frac{f_{ij}}{f_{\bullet j}} - f_{i\bullet})^2 \\ &= \sum_{j=1}^{q} \sum_{i=1}^{p} (\frac{f_{ij} - f_{\bullet j} f_{i\bullet}}{f_{\bullet j} f_{i\bullet}}) \\ &= \phi^2. \end{aligned}$$

Remarquer que les deux inerties sont égales $I_p = I_q$ et sont déjà obtenus par le test. Ce résultat est la conséquence de la dualité. Donc en AFC, nous avons deux centres de gravités et deux inerties, mais égales à l'aide de la distance du khi-deux. Reste à trouvé les axes orthogonaux qui passe par le centre de gravité G_p (respectivement G_q) et qui maximise I_p (respectivement I_q). Ceci revient à faire le même travail que dans L'ACP.

1.2.3 Recherche des axes

Soit (Δ_1) l'axe qui passe par G_p de vecteur directeur unitaire $\vec{a_1}$ et qui maximise

$$\sum_{i=1}^{p} f_{i\bullet} d_{\chi^2}(\vec{a_1}, G_p).$$

Soit (Δ_2) l'axe qui passe par G_p de vecteur directeur unitaire $\vec{a_2}$ tel que $(\Delta_1) \perp (\Delta_2)$ et maximise

$$\sum_{i=1}^{p} f_{i\bullet} d_{\chi^2}(\vec{a_2}, G_p).$$

Et ainsi de suite, on obtient des p axes (respectivement q) de valeurs propres $(\lambda_i)_{i=1,\dots,p}$ (respectivement $(\lambda_j)_{j=1,\dots,q}$), ordonnés d'une manière décroissante pour les deux profils. Chaque valeur propre correspondant à la part d'inertie projeté sur un axe donné. Sachant que la somme des valeurs propres est toujours égale à l'inertie totale du nuage. On caractérise ainsi chaque axes par le pourcentage d'inertie qu'il permet d'expliquer. C'est pourquoi les valeurs propres en A.F.C. sont toutes inférieurs à 1. En pratique, on retiendra donc que les axes avec les plus forte valeurs propres.

Le choix des axes retenus est principalement basé sur des règles (règle du coude, règle de l'inertie minimale, règle du bon sens) comme en A.C.P.

Remarque 1 *Pour choisir le nombre d'axe s, il faut garder le minimum entre les deux dimensions $(p-1)$ et $(q-1)$. Donc $s \leq min(p-1, q-1)$.*

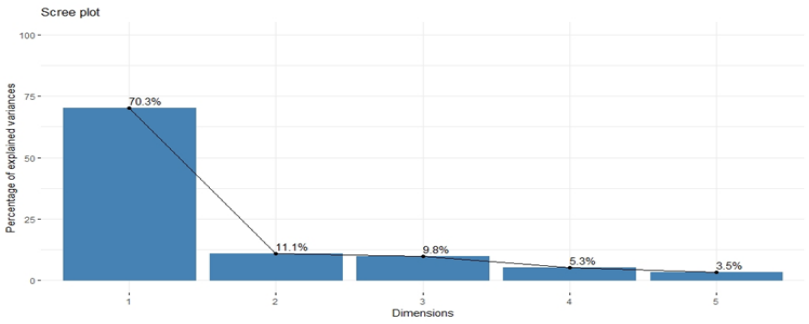

FIGURE 1.3 – Eboullis des valeurs propres

Exemple 3 *Dans notre analyse les deux premiers axes expliquent* 81.4% *de la variance totale (voir la figure 1.2.3). On va interpréter selon ces* 2 *axes.*

Au final on obtient les représentation graphiques de deux nuages N_p et N_q. Ce qui est incroyable est que, ces deux représentations graphiques donne exactement la même dispersion. Ceci est expliquer par les formules de transition dite propriété barycentriques qui donne

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{j=1}^{q} \frac{f_{ij}}{f_{i\bullet}} G_s(i).$$

Et

$$G_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^{p} \frac{f_{ij}}{f_{i\bullet}} F_s(j).$$

Avec $F_s(i)$ est le vecteur coordonnés de la ligne i sur l'axe du rang s noté (Δ_s). Aussi $\frac{f_{ij}}{f_{i\bullet}}$ (respectivement $\frac{f_{ij}}{f_{\bullet j}}$) est le $i^{\text{ième}}$ élément du profil i (respectivement le $j^{\text{ième}}$ élément du profil j. Et $G_s(j)$ est le vecteur coordonnés de la colonne j sur l'axe du rang s. En fin, λ_s est l'inertie associée à l'axe (Δ_s).

1.3 Interprétation d'une A.F.C.

1.3.1 Qualité de représentation et contribution

Il est nécessaire de intéresser à la qualité de représentation des lignes et des colonnes sur les plans factoriels ainsi qu'à leur contribution.

Qualité de représentation

la qualité de représentation d'un point c'est un indicateur qui nous informe que l'écart d'un profil au profil moyen est bien représenté par l'axe.

$$\cos^2(\overrightarrow{GM_i}, \Delta_s) = \frac{\text{inertie projeté de } M_i \text{ sur } \Delta_s}{\text{inertie totale de } M_i} = \frac{f_{i\bullet}(Gh_i^{\Delta})^2}{f_{i\bullet}(GM_i)^2}.$$

Remarque 2 *Dans un plan donné, on définit également la qualité de représentation globale comme le pourcentage d'inertie qu'explique le plan. C'est par rapport à cette qualité globale que l'on évalue la qualité de représentation d'un profil.*

Contribution

lors de la construction d'un axe factoriel, certains profils ont des rôles plus importants.On calcule un paramètre appelé contribution , qui permet qui calculer cette influence.

Cette contribution sur un axe donné, est définie comme la proportion de l'inertie de cet axe expliquée par le profil ; ce paramètre permet donc de mesurer la contribution des lignes (colonne) à la formation des axes :
contribution du profil ligne i à la l'inertie de l'axe S :

$$Cr(i) = \frac{n_{i\bullet}}{n}(F_s(i))^2 \backslash \lambda_k$$

contribution du profil colonne j à la l'inertie de l'axe S :

$$Cr(j) = \frac{n_{j\bullet}}{n}(F_s(j))^2 \backslash \lambda_k$$

Remarque 3 *Sur une figure ou pense généralement que le point le plus éloigné est celui qui participe le plus à la confection des axes , mais en vérité étant vérifier les contributions , on obtient le contraire.*

Une règle élémentaire pour voir quelle ligne apporte une contribution importante à l'axe S consiste à regarder pour la ligne i si $Cr(i) > \frac{n_{i\bullet}}{n}$, et pour la colonne j si $Cr(j) = \frac{n_{j\bullet}}{n}$

Représentation graphique

L'analyse des plans factoriels permet d'observer les profils proches entre eux ou au contraire éloignés, il est ainsi possible de construire des groupes, d'observer des tendances.

La construction des composantes principales conduit à rendre minimale la déformation des distances du χ^2 entre profils lorsque l'on projette les profils dans le plan factoriel.Ainsi les distances que l'on observe entre les profils dans le plan factoriel sont globalement les plus proches possible des distances réelles entre ces profils.

Interprétation

Pour interpréter les représentation graphique en A.F.C. on basant sur les produits scalaires entre les vecteurs des des modalités des deux variables.
Si le produit scalaire est positif,on dit qu'il y a une conjonction entre les modalités.
Si le produit scalaire est négative,on dit qu'il y a une opposition entre les modalités.
Si le produit scalaire est nulle,c'est une quadrature.

Exemple 4 *Voici les représentations graphiques de notre étude.*

Représentation graphique des profils lignes.

Représentation graphique des profils colonne.

Représentation graphique des lignes et des colonnes dans le même plan.

Interprétation :

D'après la représentation graphique,on remarque que les éléves du primaire souffrent de troubles d'articulation,hyperactivité,enurésite,encoprésie,retard scolaire (surtout chez les garçons) ;et aussi le bégaiement,maltraitance à l'école,phobie scolaire(surtout chez les filles), dyslexie (chez les deux sexes)par contre agressivité/violences est présente chez les garçons. Une autre remarque très importante est que anxiété et l'angoisse chez les élèves de ce niveau est absente.

De plus,on observe chez les filles au moyen les troubles névrotiques obsessionnelle ,et fugue. Les garçons dans ce niveau souffrent de la migraine et ils sont agres-

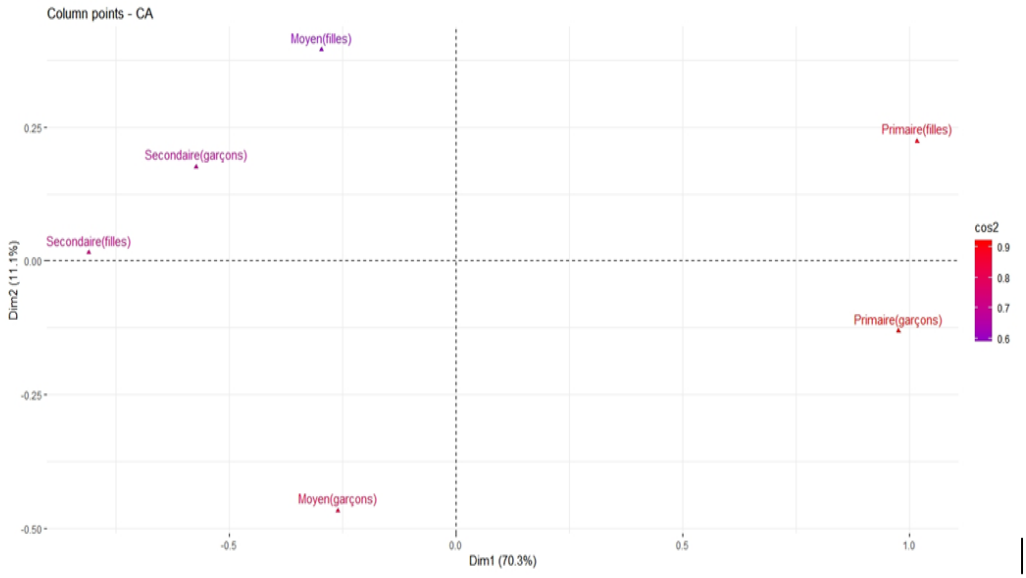

FIGURE 1.4 – Représentation graphique des colonnes

sives ; contrairement au filles En ce qui concerne les élèves au secondaire ,les deux sexes souffrent d'angoisse, anxiété,migraine,hystérie, fugue ,état dépressif, et les troubles d'articulation,dyslexie sont absents chez les élevés de ce niveau.

Donc on peut distinguées deux groupes,le premier groupe : les élevés du primaire qui souffrent des troubles généralement physique par contre les élèves du moyen et du secondaire (qui est le deuxième groupe) les troubles chez eux sont totalement psychique.

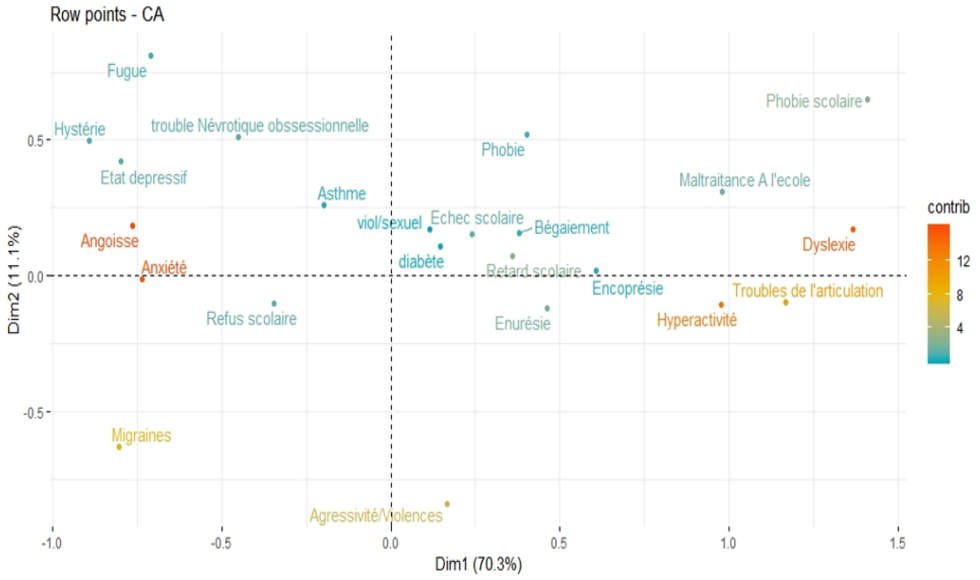

Figure 1.5 – Représentation graphique des lignes

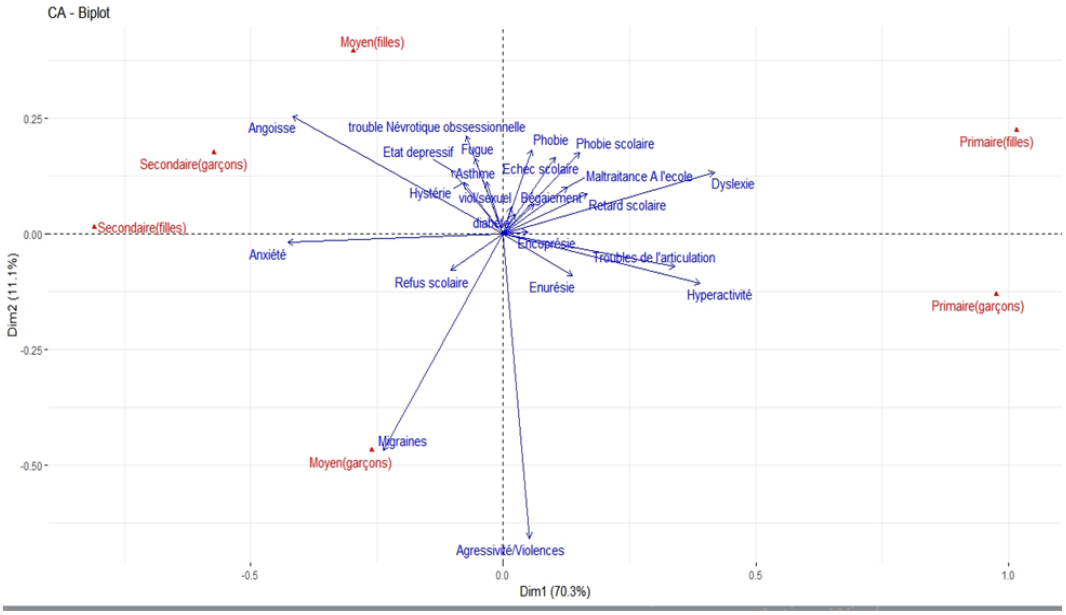

Figure 1.6 – Représentation graphique des deux variables dans le même plan