Analyse des Correspondances Multiples

ACM

Analyse des correspondances multiples

Objectif : Même objectif que l'ACP ou AFC ! Visualiser, résumer l'information contenue dans des masses volumineuses de données

Les données :

- I individus décrit par J variables nominales ou ordinales

Exemple : réponse à une enquête basée sur des questions fermées à choix multiples (catégories d'âge, de revenus, activités sportives, ...)

L'ACM vise à mettre en évidence :

- **Les relations entre les modalités des différentes variables**
- **Les relations entre les individus**
- **Les relations entre les variables à partir des relations entre modalités**

Plan

- Problématique
- Données et transformation du tableau de données
- Objectifs
- Étude des individus
- Études des modalités
- Aide à l'interprétation

Problématique

- L'analyse factorielle des correspondances, vue dans le chapitre précédent, s'applique à des situations où les individus sont décrits par deux variables nominales.
- Mais il est fréquent que l'on dispose d'individus décrits par plusieurs (deux ou plus) variables nominales ou ordinales.
- C'est notamment le cas lorsque nos données sont les résultats d'une enquête basée sur des questions fermées

Données et transformation du tableau

- ACM

Tableau de Données

- *J* variables qualitatives mesurées sur *I* individus
- Exemple : enquêtes basées sur des questions fermées à choix multiples

  (catégories d'âge, de revenus, activités sportives, …)
  - Quelle est votre catégorie professionnelle ?
    {cadre, employé, ouvrier, etc}
  - Il faut fermer les centrales nucléaires ; êtes vous
    {pas d'accord, sans opinion, d'accord,...}
- Les individus sont les enquêtés
- Les variables sont les questions
- Les modalités des variables sont ordonnées ou non

Tableau protocole : 3 questions, 7 modalités

	Sexe	Revenu	Preference
s1	F	M	A
s2	F	M	A
s3	F	E	B
s4	F	E	C
s5	F	E	C
s6	H	E	C

Tableau disjonctif complet

	Sexe:F	Sexe:H	Rev:M	Rev:E	Pref:A	Pref:B	Pref:C
s1	1	0	1	0	1	0	0
s2	1	0	1	0	1	0	0
s3	1	0	0	1	0	1	0
s4	1	0	0	1	0	0	1
s5	1	0	0	1	0	0	1
s6	0	1	0	1	0	0	1

La disjonction complète

DEPARTEMENTS	BLE	VIN	LAIT
DEP 1	NON	ROUGE	PEU
DEP 2	OUI	ROSE	MOYEN
DEP 3	OUI	BLANC	MOYEN

LA DISJONCTION EST UNE CODIFICATION EN DONNEES BINAIRES

CREATION D'UNE VARIABLE POUR CHAQUE MODALITE

	BLE		VIN			LAIT		
DEPARTEMENTS	OUI	NON	ROUGE	ROSE	BLANC	PEU	MOYEN	BCP
DEP 1	0	1	1	0	0	1	0	0
DEP 2	1	0	0	1	0	0	1	0
DEP 3	1	0	0	0	1	0	1	0

Tableau disjonctif complet

- X = tableau disjonctif complet;
- K = nombre total de modalités
- La somme des éléments d'une même ligne est constante et vaut J (=nombre de variables)
- La somme de tous les éléments du tableau vaut n
- La somme des éléments d'une même colonne n'est pas constante mais égale à l'effectif marginal possédant la modalité k de la variable j considérée

	1…	Variable j	…J	
	1…	Modalités …	…K_j	
1	…		…	
…				
i	…	X_{ik}	…	$X_{i.} \cdot J$
…				
I	…		…	
		$X_{.k} \cdot I$		$X_{..} \cdot IJ$

Le tableau disjonctif complet ou TDC comporte une colonne pour chaque modalité des variables étudiées et une ligne pour chaque individu statistique. Les cellules du tableau contiennent 1 ou 0 selon que l'individu considéré présente la modalité ou non.

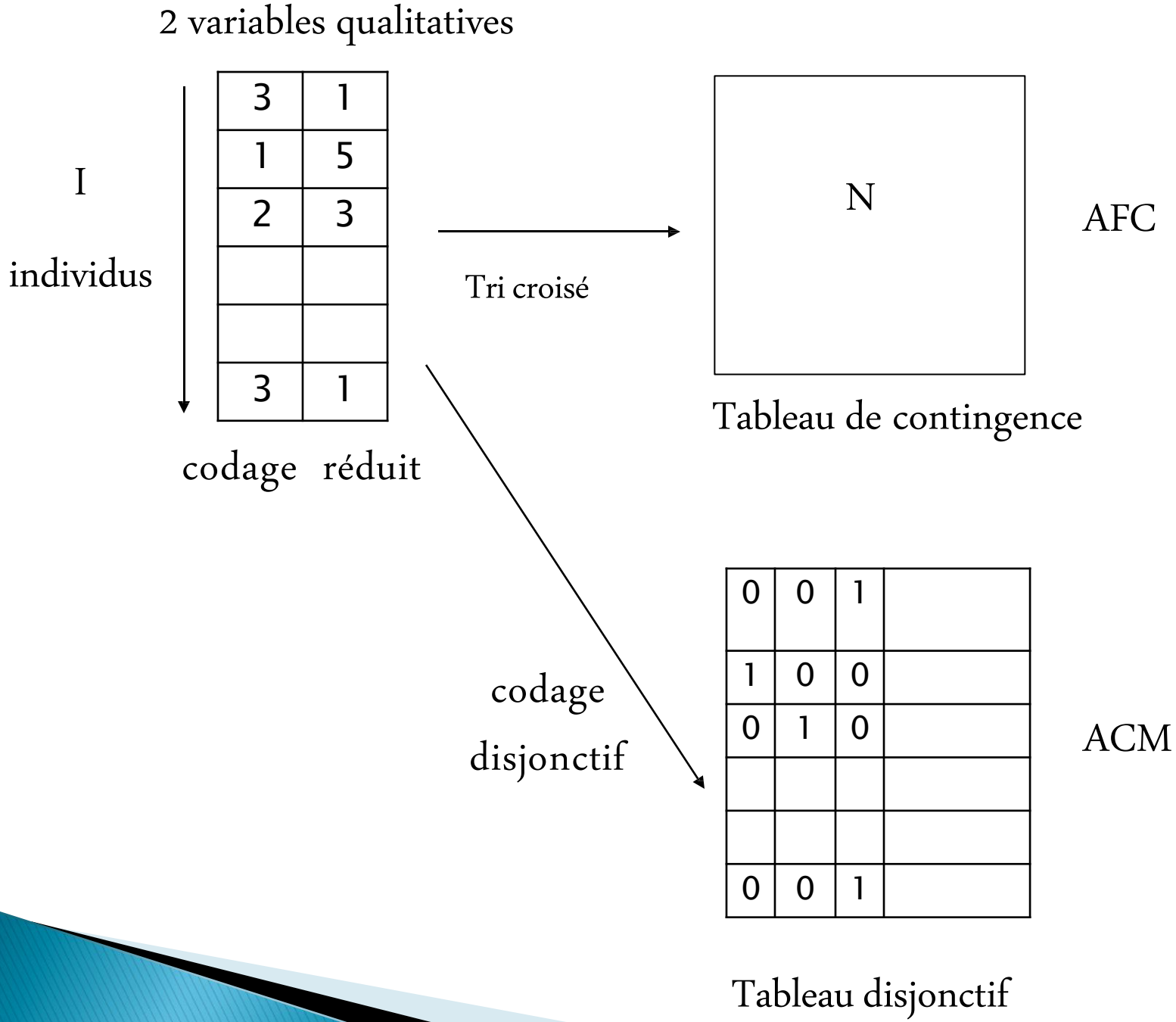
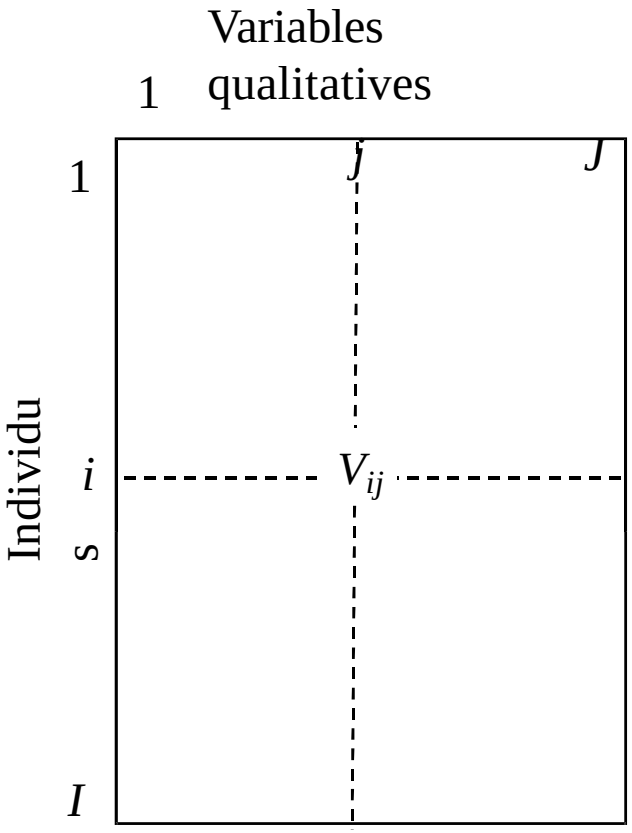

2 variables qualitatives
I
individus
3 1
1 5
2 3
3 1
codage réduit
Tri croisé
N
AFC
Tableau de contingence
codage
disjonctif
0 0 1
1 0 0
0 1 0
0 0 1
ACM
Tableau disjonctif

Tableau données

- *J* variables qualitatives mesurées sur I individus



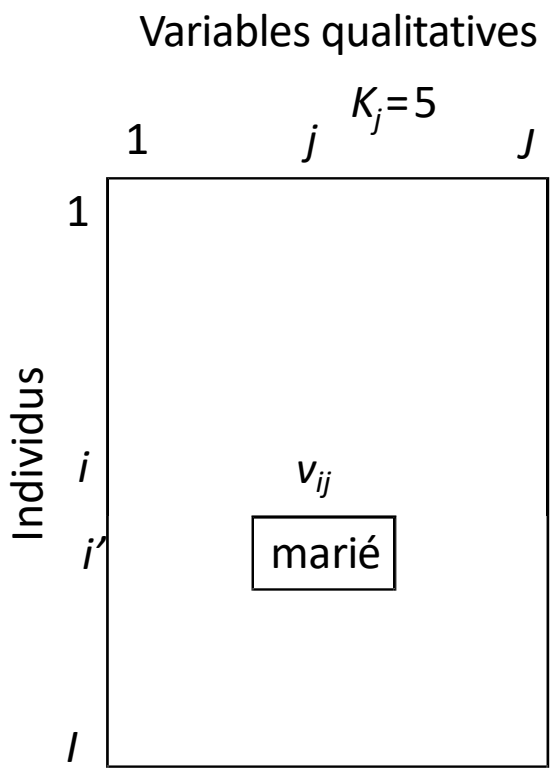
I individus
J variables qualitatives

v_{ij} : modalité de la variable *j* possédée par l'individu *i*

Exemple : enquête où *I* personnes sont interrogées sur *J* questions à choix multiples

- Chaque colonne est constituée d'observations des modalités

Les données



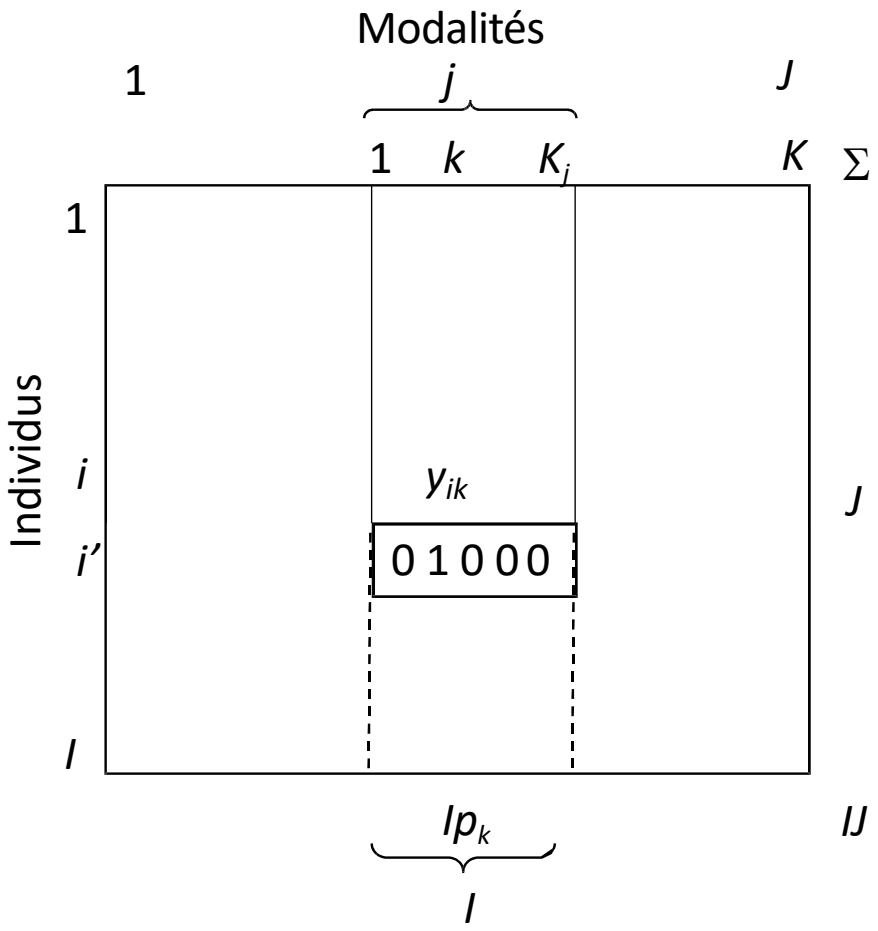


Tableau disjonctif complet (TDC)

Transformation du tableau disjonctif complet

Le poids d'un individu est $\frac{1}{I}$

p_k : proportion d'individus possédant la modalité K

y_{ik} = 1 si i possède la modalité k de la variable j (quel que soit p_k)
= 0 sinon

- $x_{ik} = y_{ik}/p_k$

$$\frac{\Sigma_{i=1}^{I} x_{ik}}{I} = \frac{1}{I}\frac{\Sigma_{i=1}^{I} y_{ik}}{p_k} = \frac{1}{I}\frac{I \times p_k}{p_k} = 1$$

Centrage : $x_{ik} = y_{ik}/p_k - 1$

Jeu de données : Histoire de vie 2003

- Description du jeu données
- Questions, individus, variables…

- Échantillon de 2000 personnes et 20 variables issues de l'enquête Histoire de Vie, réalisée en France en 2003 par l'INSEE.
- Un data frame avec 2000 lignes et 20 variables inclus dans la librarie **questionr** de **R**

Objectifs

- ACM

Objectifs

- Cette méthode est particulièrement adaptée à l'exploration d'enquêtes.
- Nous nous plaçons donc dans la situation où nous disposons de I individus, décrits par J variables
- L'ACM vise à mettre en évidence :
  - les relations entre les modalités des différentes variables ;
    - les relations entre individus
    - Un individu = une ligne du TDC = ensemble de ses modalités
    - Ressemblance des individus/Variabilité des individus
  - les relations entre les variables à partir des relations entre modalités.
  - Dualité : Quelles variables expliquent le plus la variabilité entre individus ?

Étude des individus

- ACM

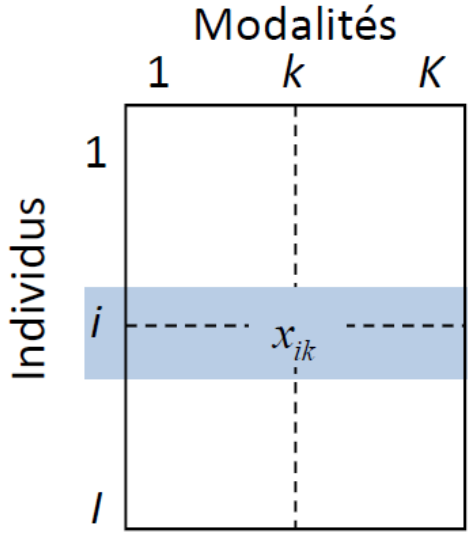
Individus et distance

- 2 individus prennent les mêmes modalités : distance = 0
- 2 individus ont en commun beaucoup de modalités : distance petite
- 2 individus sont d'autant plus éloignés que leurs réponses diffèrent pour un plus grand nombre de questions et pour des modalités rares.
- 2 individus ont en commun une modalité rare : distance petite pour prendre en compte leur spécificité commune
- Un individu est d'autant plus loin de l'origine qu'il comporte des modalités rares

$$d^2_{i,i'} = \sum_{k=1}^{K} \frac{p_k}{J} (x_{ik} - x_{i'k})^2$$

$$= \sum_{k=1}^{K} \frac{p_k}{J} \left(\frac{y_{ik}}{p_k} - \frac{y_{i'k}}{p_k} \right)^2$$

$$= \frac{1}{J} \sum_{k=1}^{K} \frac{1}{p_k} (y_{ik} - y_{i'k})^2$$



Distance / Inertie

$$d^2_{i,i'} = \frac{1}{J}\sum_{k=1}^{K}\frac{1}{p_k}(y_{ik}-y_{i'k})^2$$

$$d(i,G_I)^2 = \frac{1}{J}\sum_{k=1}^{K}\frac{y_{ij}}{p_k} - 1$$

$$\text{Inertie}(N_I) = \sum_{i=1}^{I}\underbrace{\frac{1}{I}d^2(i,O)}_{\text{inertie de } i} = \frac{K}{J} - 1$$

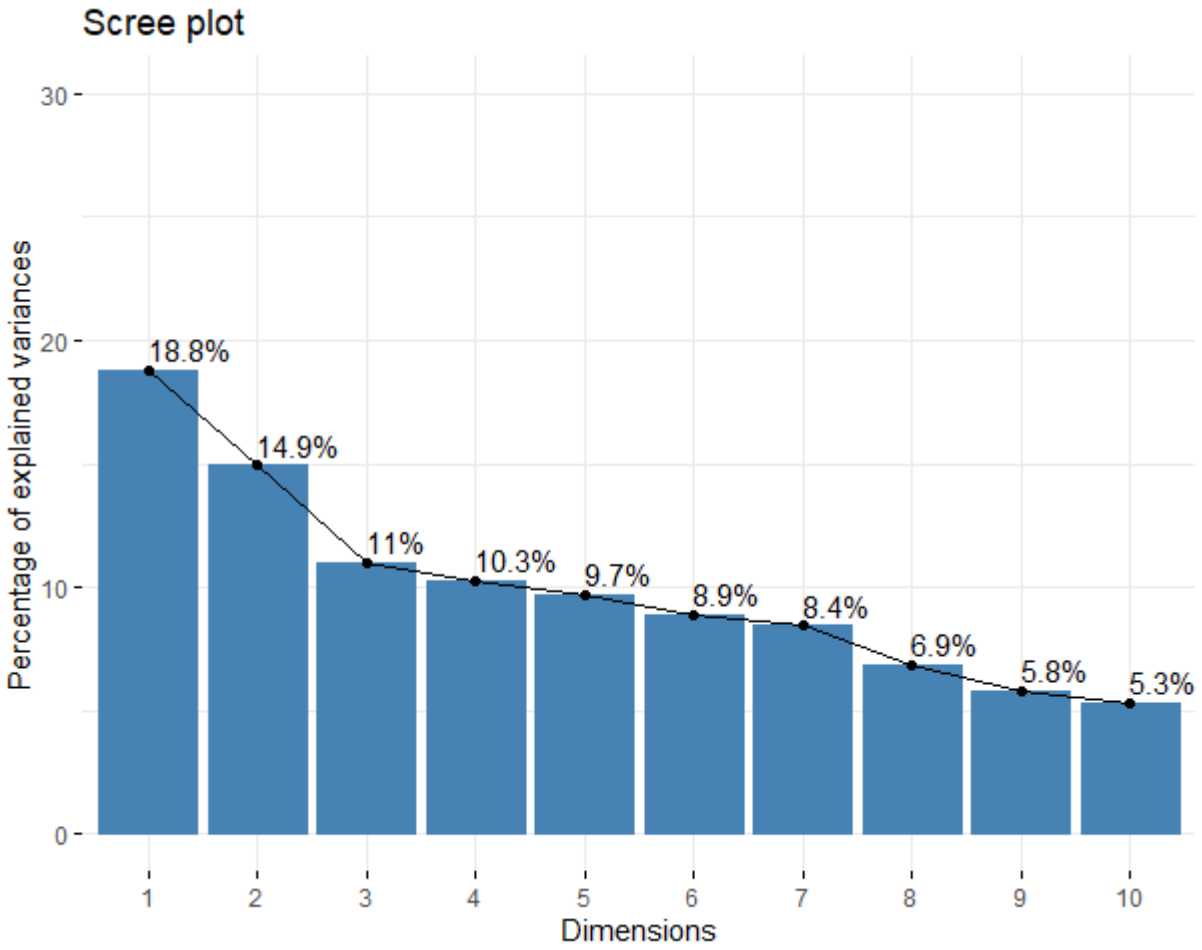
- Les valeurs propres qui étaient très séparées dans l'AFC de N, ne le sont plus dans l'ACM du tableau disjonctif X
- En AFC l'inertie est égale au Khi-deux associé au tableau de contingence divisé par le nombre d'individus
- En ACM l'inertie est égale au nombre moyen de modalités diminué de 1
- En général, vu la nature des données, les inerties portées par les premiers axes sont faibles.

Ajustement du nuage des individus

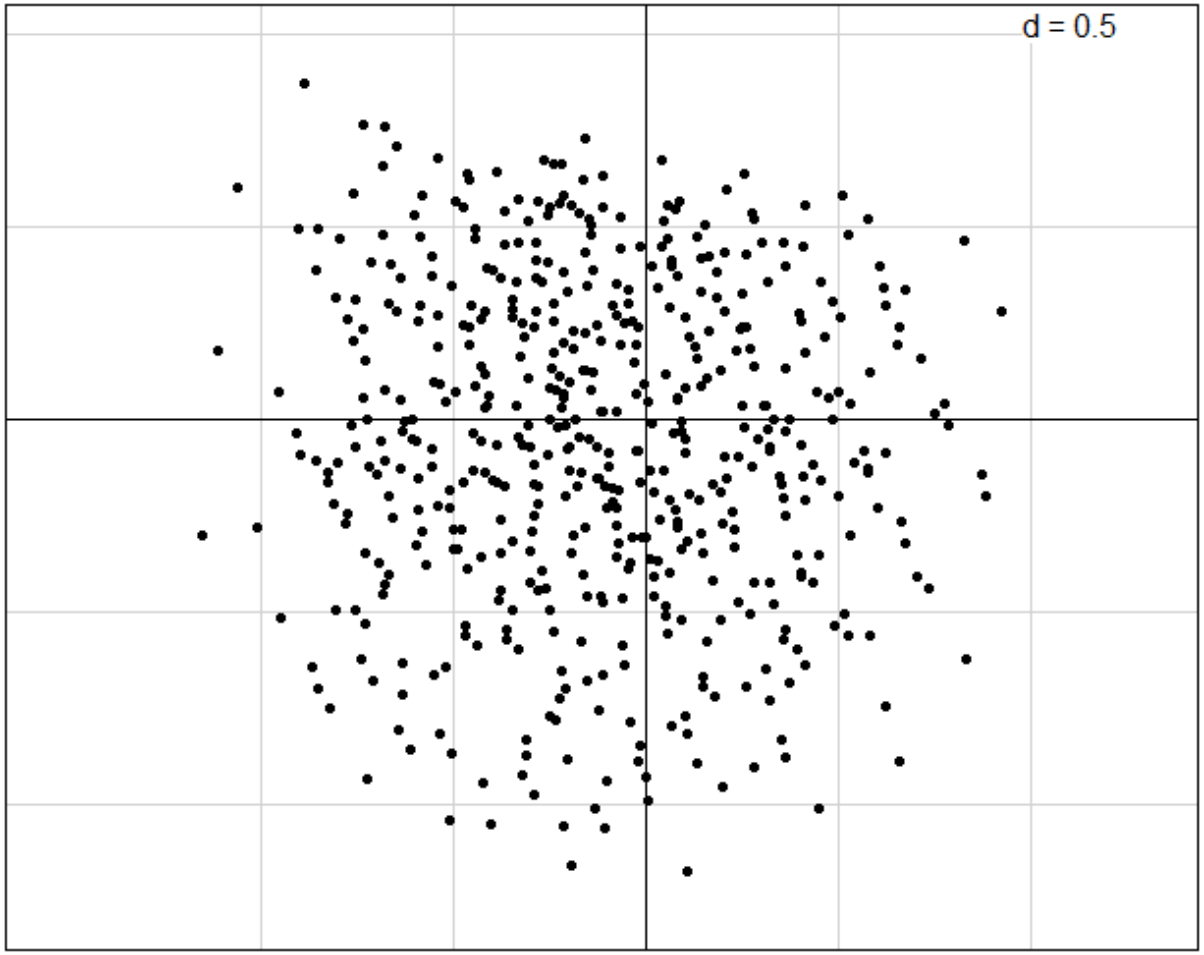
Recherche des dimensions factorielles commme pour toute méthode d'analyse factorielle

Construction séquentielle : recherche d'un axe qui maximise l'inertie et qui est orthogonal aux axes précédemment trouvés

Diagramme des inerties

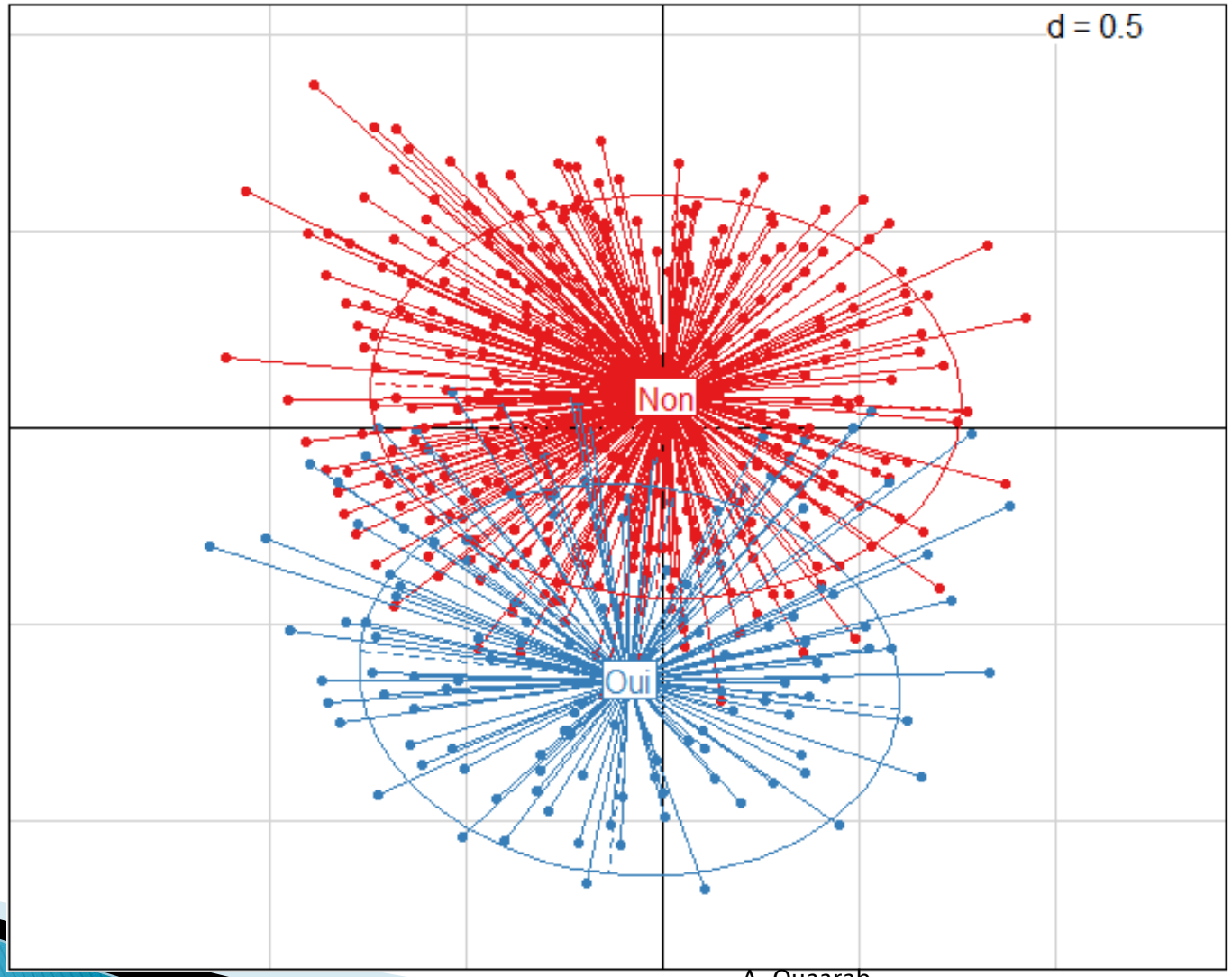


Représentation du nuage des individus



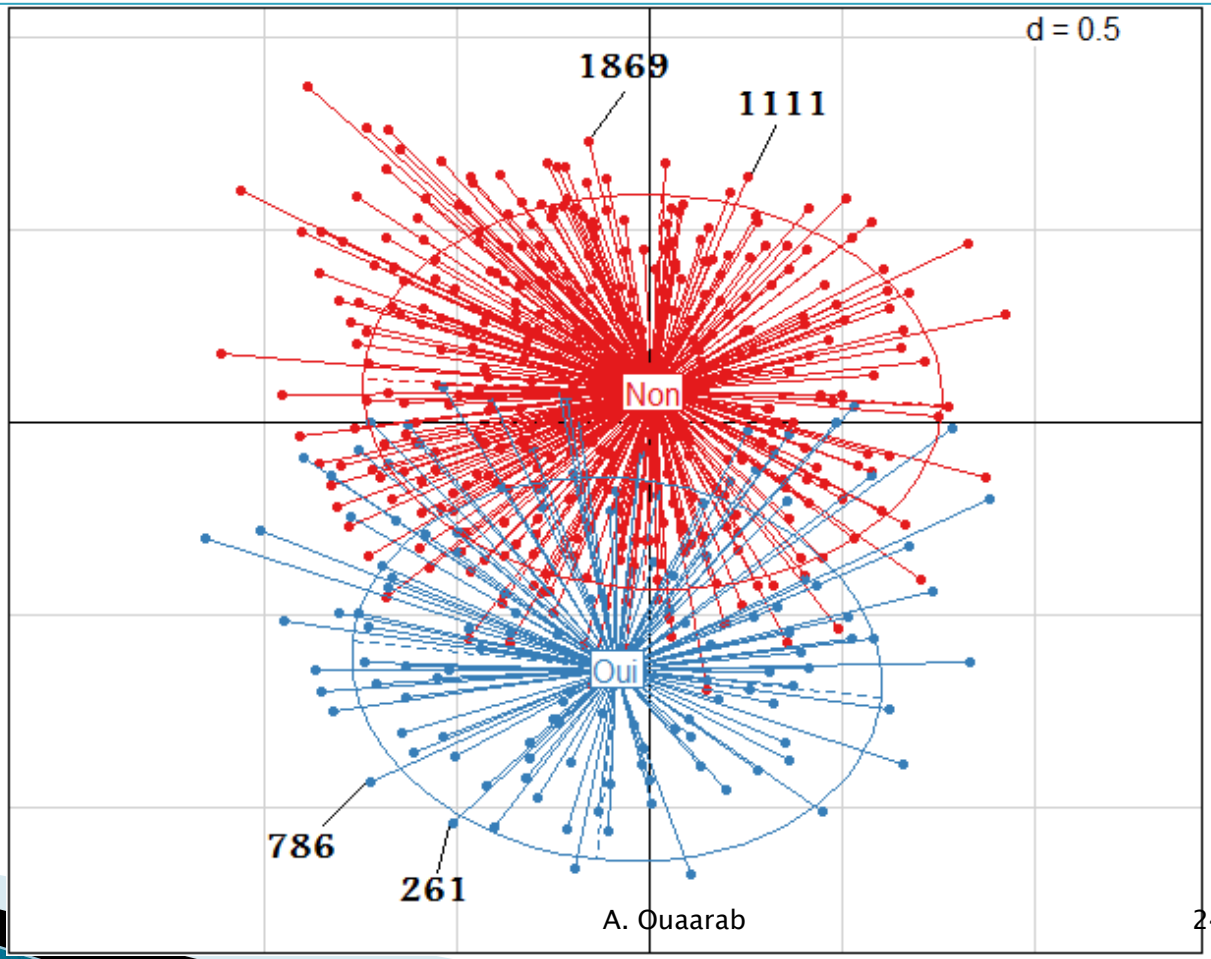
Représentation des individus en fonction de pèche/chasse

- Une modalité au barycentre des individus qui possèdent cette modalité



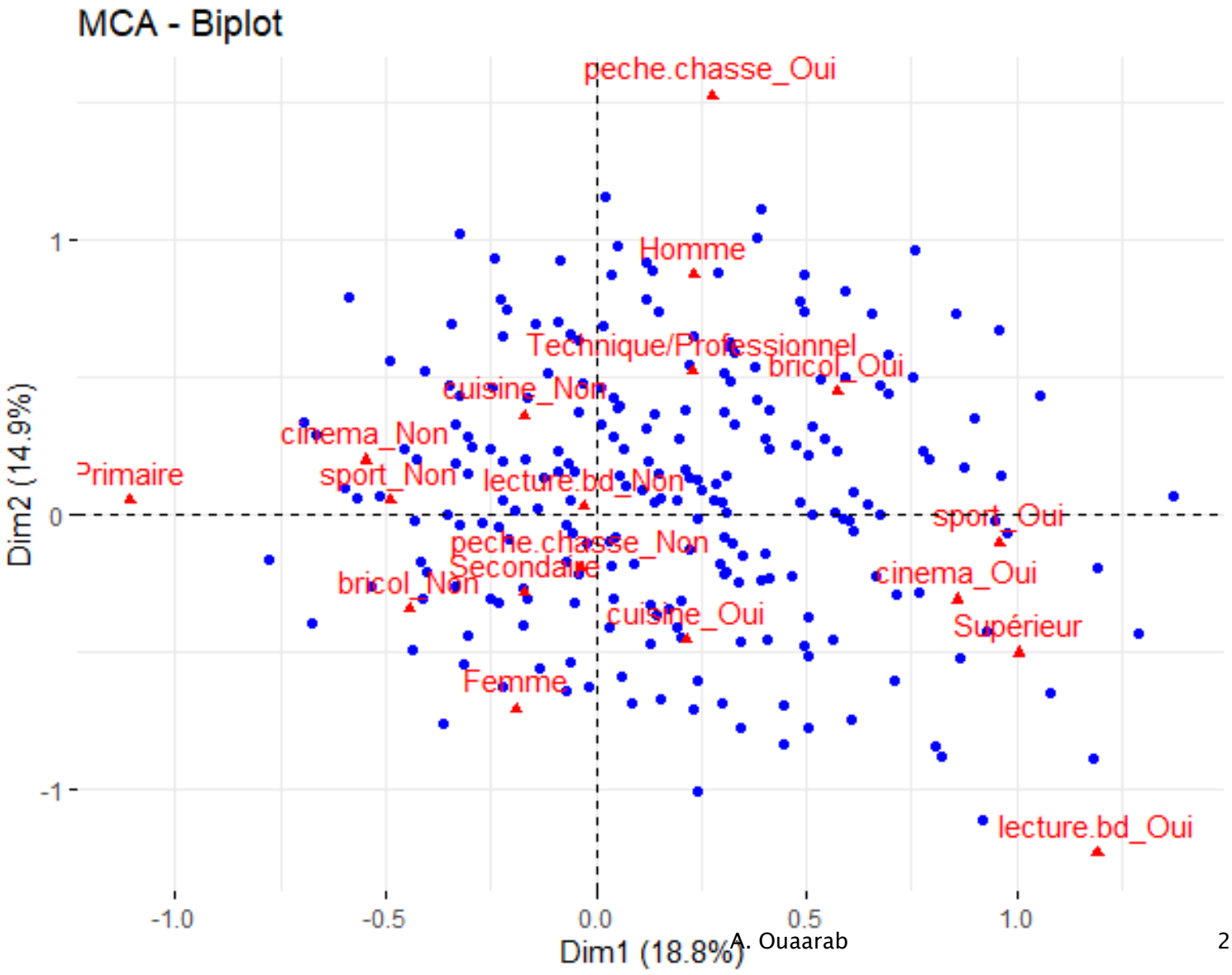
Modalités dans le nuage des individus

ind	grpage	sexe	etud	pech.chas	cinema	cuisine	bricol	sport	lecture.bd
1869	**[45,65)**	F	Scnd	Non	Oui	Oui	Non	Non	**Oui**
1111	**[65,93]**	F	Scnd	Non	Oui	Oui	Non	Non	**Non**
786	**[25,45)**	H	Tech/Prof	Oui	Oui	Non	Oui	Oui	Non
261	**[45,65)**	H	Tech/Prof	Oui	Oui	Non	Oui	Oui	Non



Représentation des individus/modalités

- Chaque modalité est au barycentre des individus qui la prennent
- Chaque individu au barycentre des modalités qu'il possède

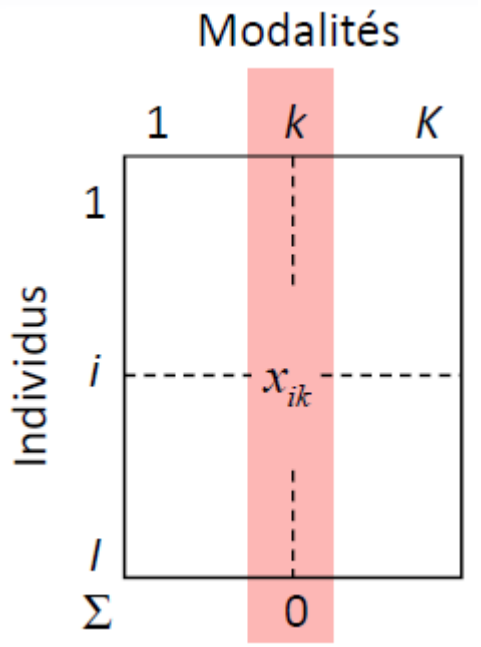


Étude des modalités

- ACM

Distance modalités

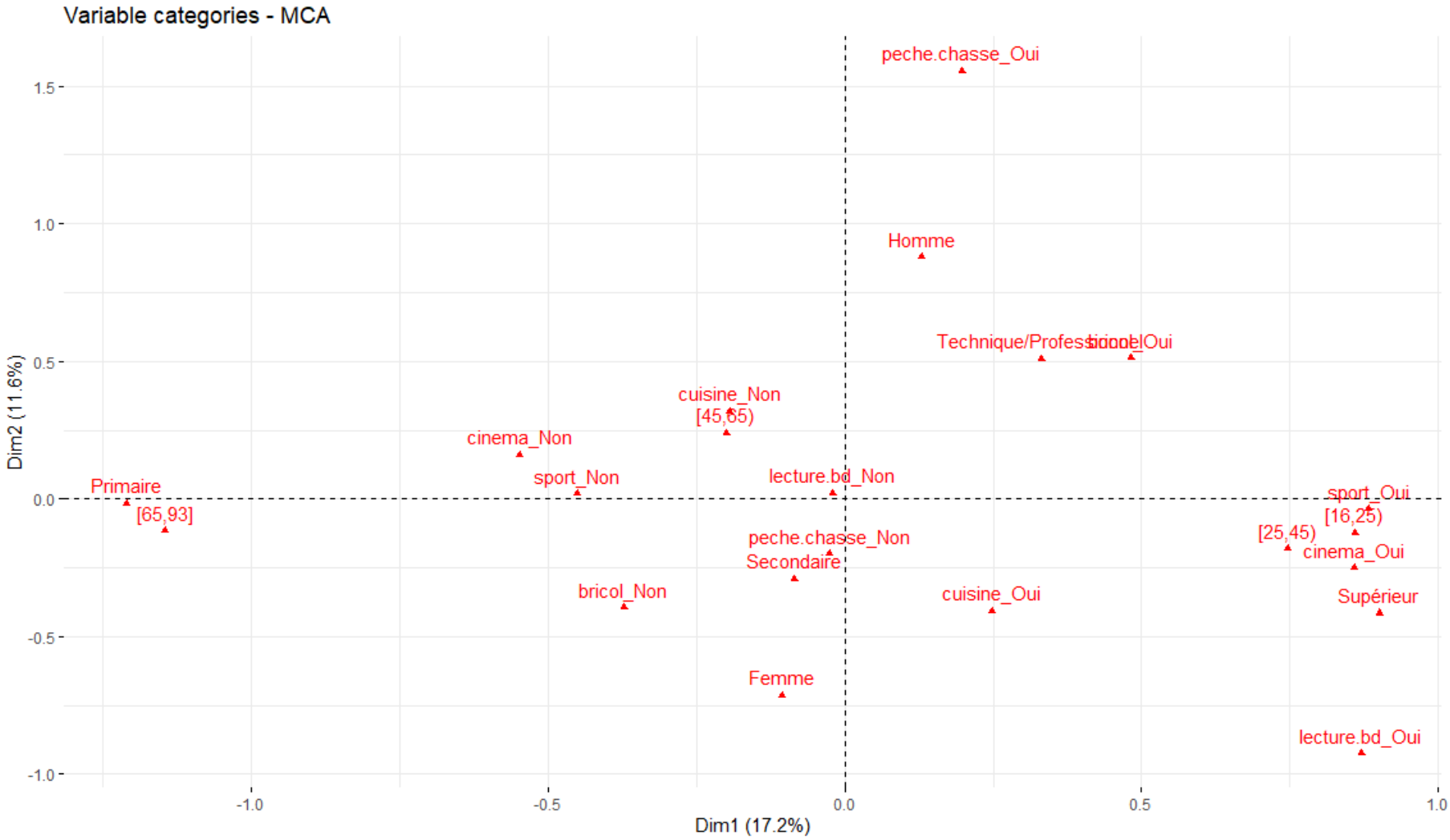
- Une modalité est d'autant plus loin de O que sa fréquence est faible
- Deux modalités sont d'autant plus éloignées qu'elles sont de fréquences faibles et rarement rencontrées simultanément
  - Ce faisant, on réduit la distance lorsque les associations sont nombreuses
  - ⇒ cette distance (issue du produit scalaire) permet bien de distinguer les 2 cas



$$Var(k) = d^2(k, O) = \frac{1}{p_k} - 1$$

$$d^2(k, k') = \frac{p_k + p_{k'} - 2p_{kk'}}{p_k p_{k'}}$$

Etude des modalités



Inerties – Définitions

L'inertie totale du nuage des modalités est déterminée uniquement par le nombre total de modalités K et le nombre de variables J

- Inertie totale du nuage des modalités : $I = \sum_{j=1}^{J} Inertie(j) = \frac{K-J}{J}$

- Inertie d'une variable $Inertie(\mathrm{j}) = \frac{K_j - 1}{\mathrm{J}}$

k_j est le nombre de modalités de la variable j

- Inertie d'une modalité $Inertie(\mathrm{k}) = \frac{1 - p_k}{\mathrm{J}}$

p_k étant la proportion d'individus possèdant la modalité k de la variable j

➢ A noter que plus une modalité est rare et plus son influence globale est élevée

1) Indépendance des modalités M_k et $M_{k'}$:

$$d^2(k,k') = d^2(O,k) + d^2(O,k')$$

Autrement dit, dans l'espace multidimensionnel, le triangle *Okk'* est alors un triangle rectangle en *O*.

2) Si les modalités *k* et k' s'attirent, l'angle (Ok, Ok') est un angle aigu.

3) Si les modalités k et k' se repoussent, l'angle (Ok, Ok') est un angle obtus.

4) Si l'effectif conjoint $n_{kk'}$ des modalités k et k' est nul (en particulier si k et k' sont deux modalités d'une même question) :

$$d^2(k,k') = d^2(O,k) + d^2(O,k') + 2$$

Proximités entre les modalités

- Si deux modalités d'une même variable sont proches, cela signifie que les individus qui possèdent l'une des modalités et ceux qui possèdent l'autre sont globalement similaires du point de vue des autres variables

➢ Si deux modalités de deux variables différentes sont proches, cela peut signifier que ce sont globalement les mêmes individus qui possèdent l'une et l'autre

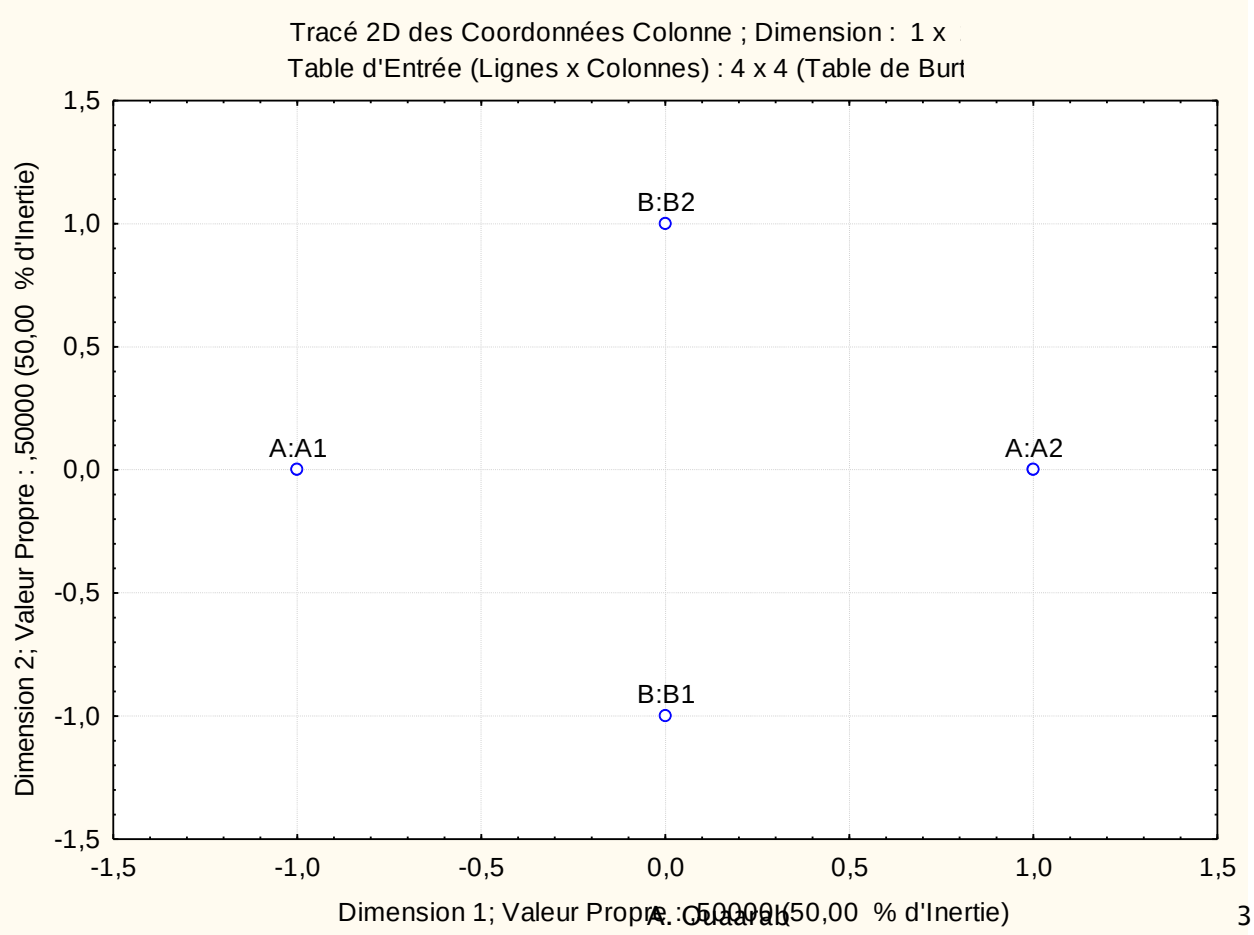
Deux questions à deux modalités chacune.

Cas 1 : les effectifs des modalités sont donnés par :

	A1	A2	Total
B1	50	50	100
B2	50	50	100
Total	100	100	200

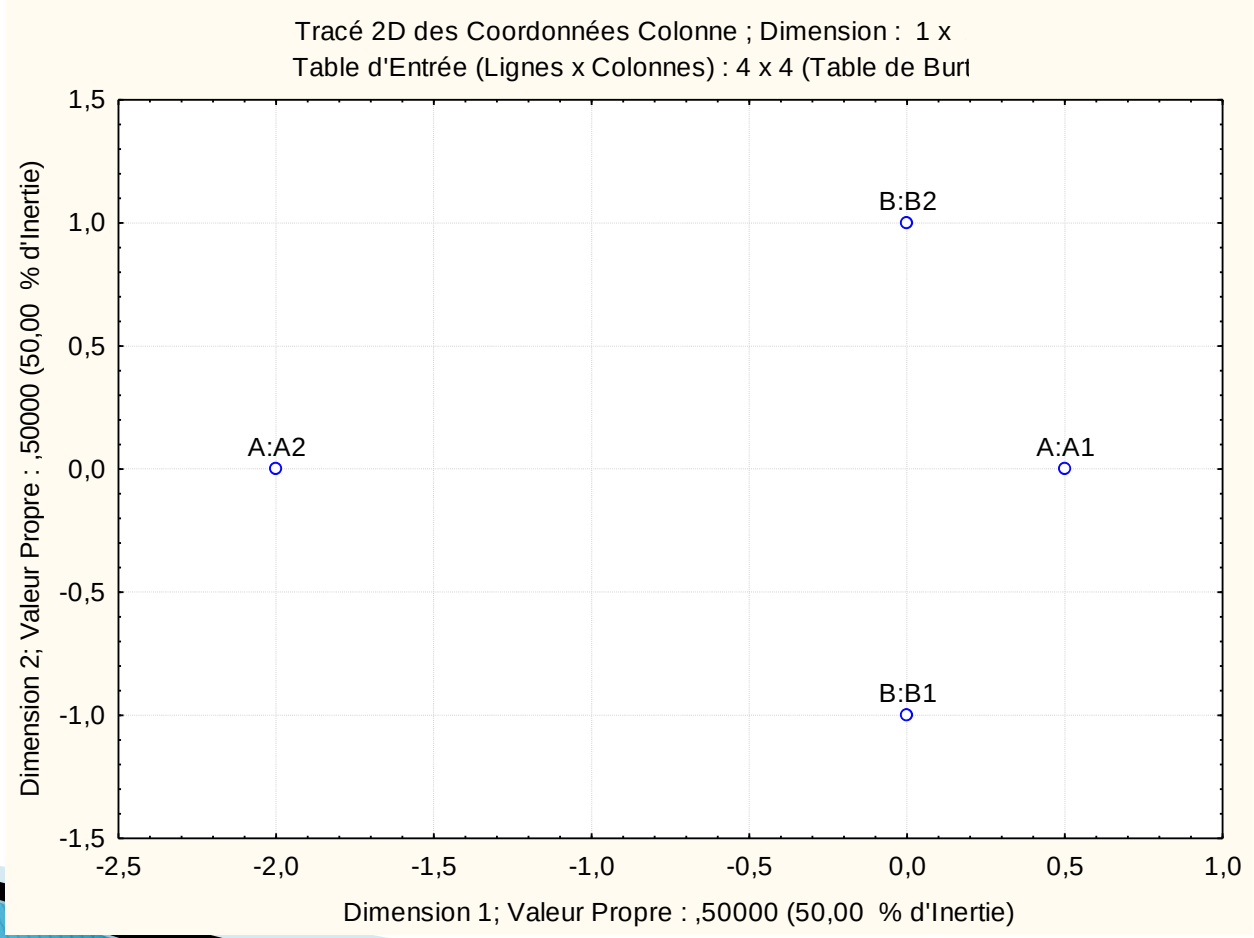
Prévoir la forme de la représentation par rapport au premier plan factoriel.

Réponse :



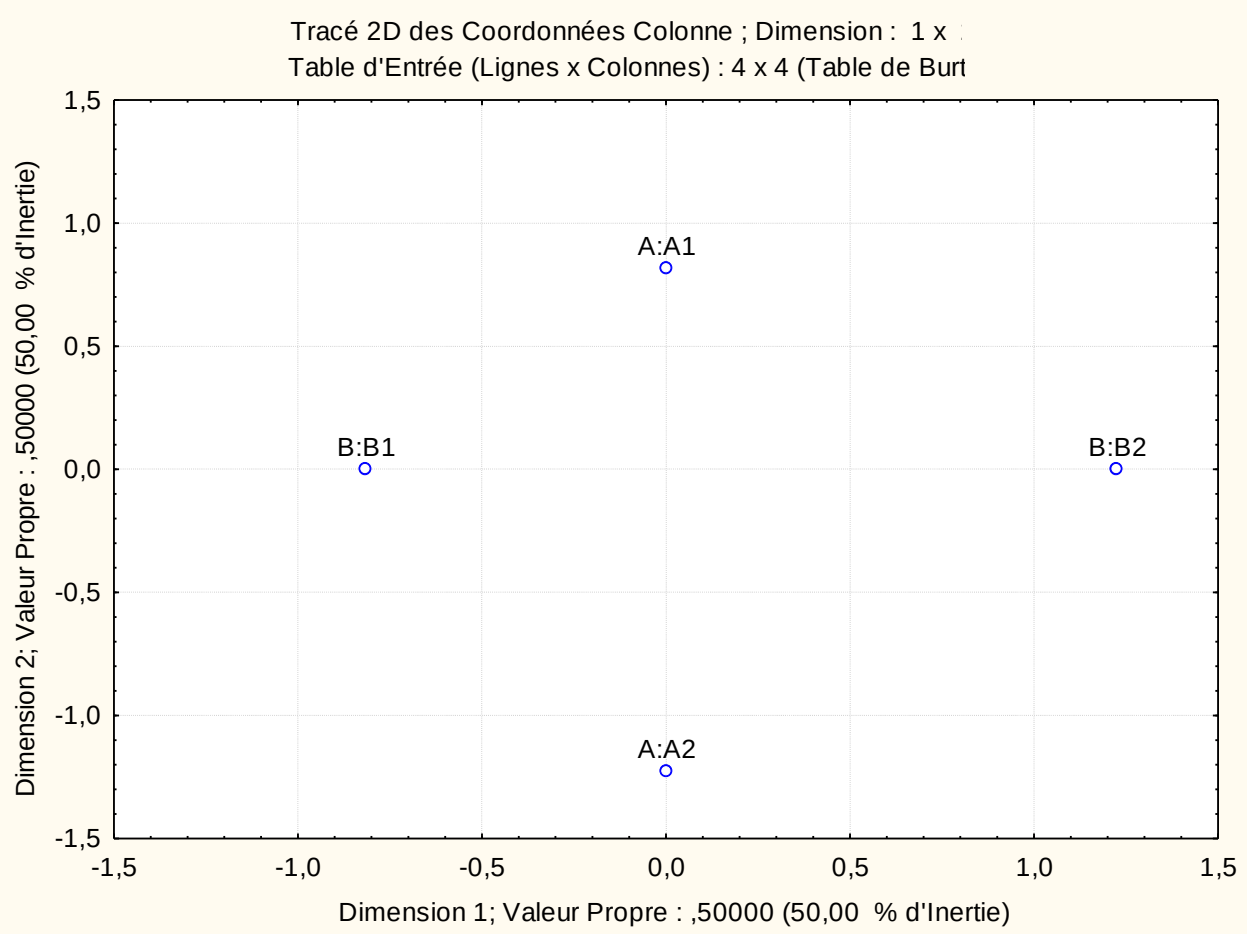
Cas 2 : les effectifs des modalités sont donnés par :

	A1	A2	Total
B1	80	20	100
B2	80	20	100
Total	160	40	200



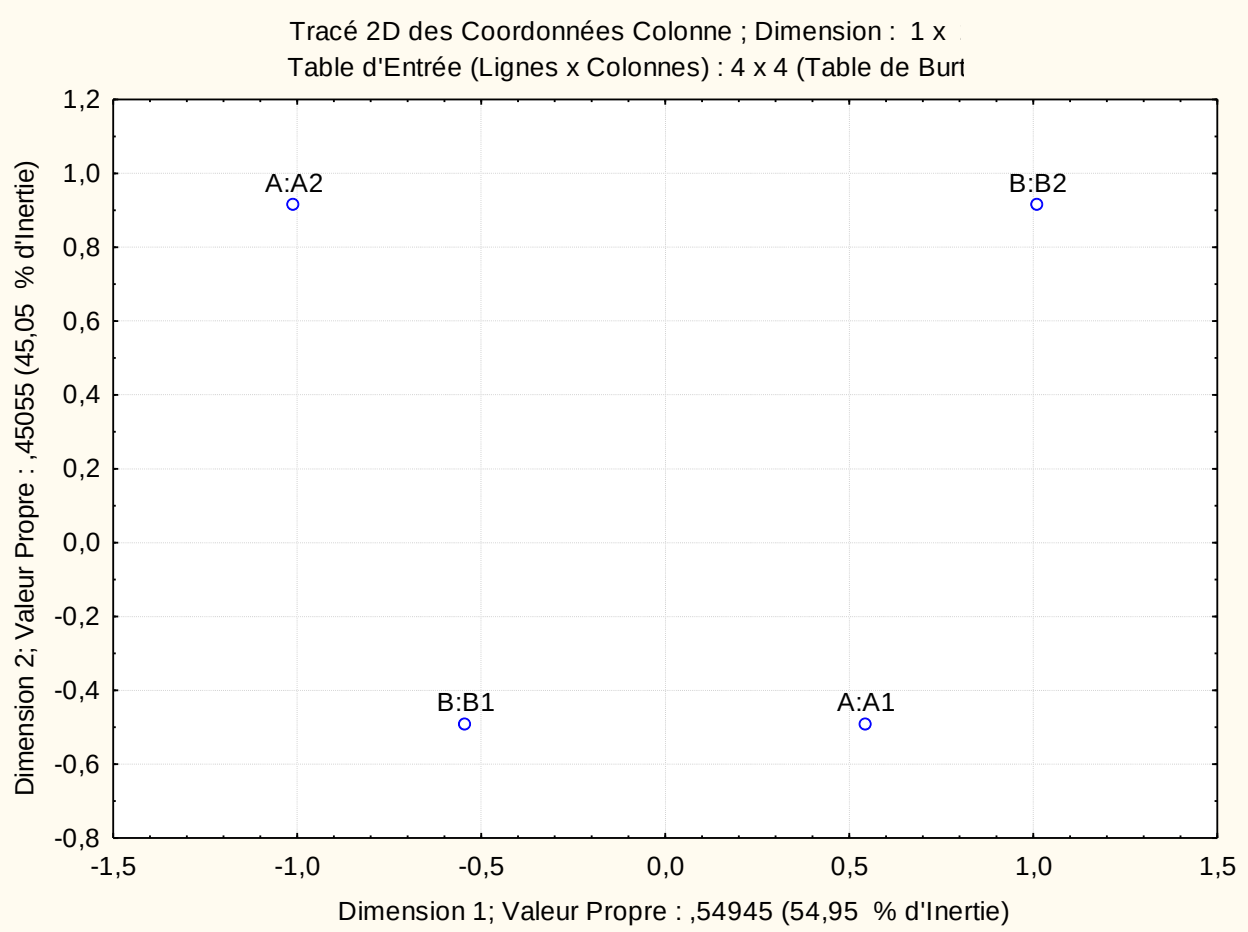
Cas 3 : les effectifs des modalités sont donnés par :

	A1	A2	Total
B1	72	48	120
B2	48	32	80
Total	120	80	200



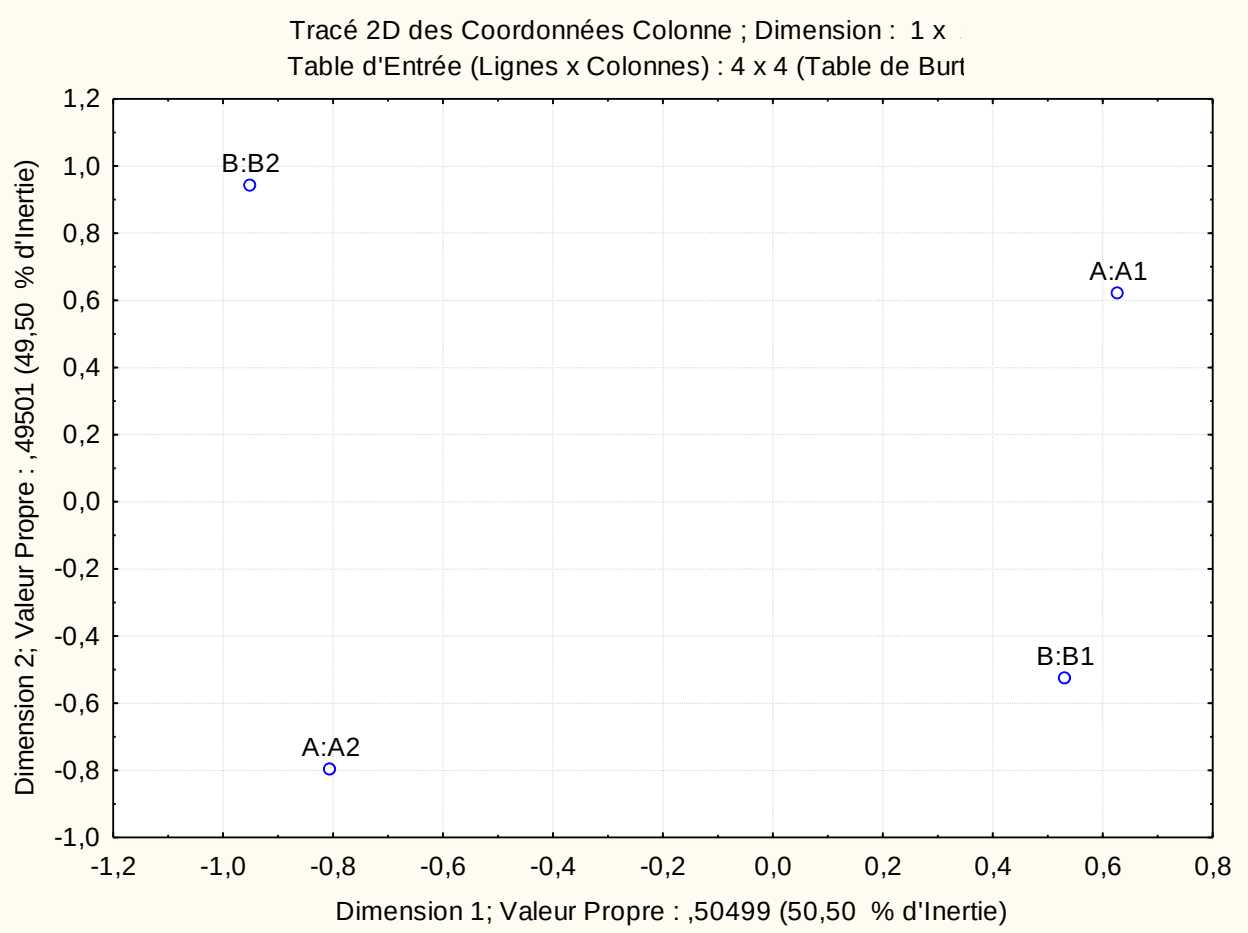
Cas 4 : les effectifs des modalités sont donnés par :

	A1	A2	Total
B1	80	50	130
B2	50	20	70
Total	130	70	200



Cas 5 : les effectifs des modalités sont donnés par :

	A1	A2	Total
B1	73	56	129
B2	40	32	72
Total	113	88	201



Aide à l'interprétation

- ACM

Contribution et qualité de représentation

Modalités extrêmes ne contribuent pas nécessairement beaucoup (cela dépend de leur fréquence)
⇒ $\cos^2$ petits … ce qui est attendu car bcp de dimensions

- Contribution absolue d'une variable :

$$CTR(j) = \sum_{k=1}^{K_j} CTR(k) = \frac{\eta^2(F_s, v_{.j})}{J}$$

- Contribution relative : $CTR(j) = \frac{\eta^2(F_s, v_{.j})}{J\lambda_s}$

Contribution d'autant plus forte que la modalité est plus rare

Avec :
$\eta^2(F_s, v_{.j})$: Rapport de corrélation entre la variable j et la composante s
λ_s *:* moyenne des carré des rapports de corrélation

Contributions à l'inertie totale

- Une modalité est d'autant plus éloignée du centre que son effectif est faible.

- L'inertie totale apportée par cette modalité décroît en fonction de l'effectif.

Il convient donc d'éviter de travailler avec des catégories d'effectif trop faible qui risquent de perturber les résultats de l'analyse (absence de robustesse).

La contribution d'une variable à l'inertie totale est d'autant plus importante que son nombre de modalités est élevé.