Analyse des correspondances (AFC)

François Husson & Magalie Houée-Bigot

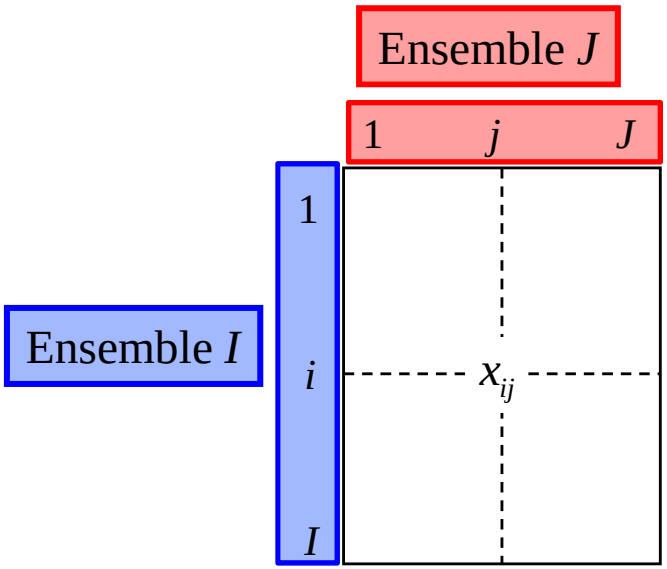
Department of applied mathematics - Agrocampus Rennes

husson@agrocampus-ouest.fr

Analyse des correspondances (AFC)

Analyse Factorielle des Correspondances (AFC)

Tableau de correspondances



x_{ij} : nombre d'individus appartenant
à l'élément i de l'ensemble I
à l'élément j de l'ensemble J

Personnages de Phèdre (Racine)	Mots	Nombre de fois que le personnage i a utilisé le mot j
Parfums	Descripteur	Nombre de fois où le parfum i a été décrit par le mot j
Milieux	Espèces	Abondance de l'espèce j dans le milieu i

$\Longrightarrow$ Exemples où le test d'indépendance du χ^2 peut être appliqué

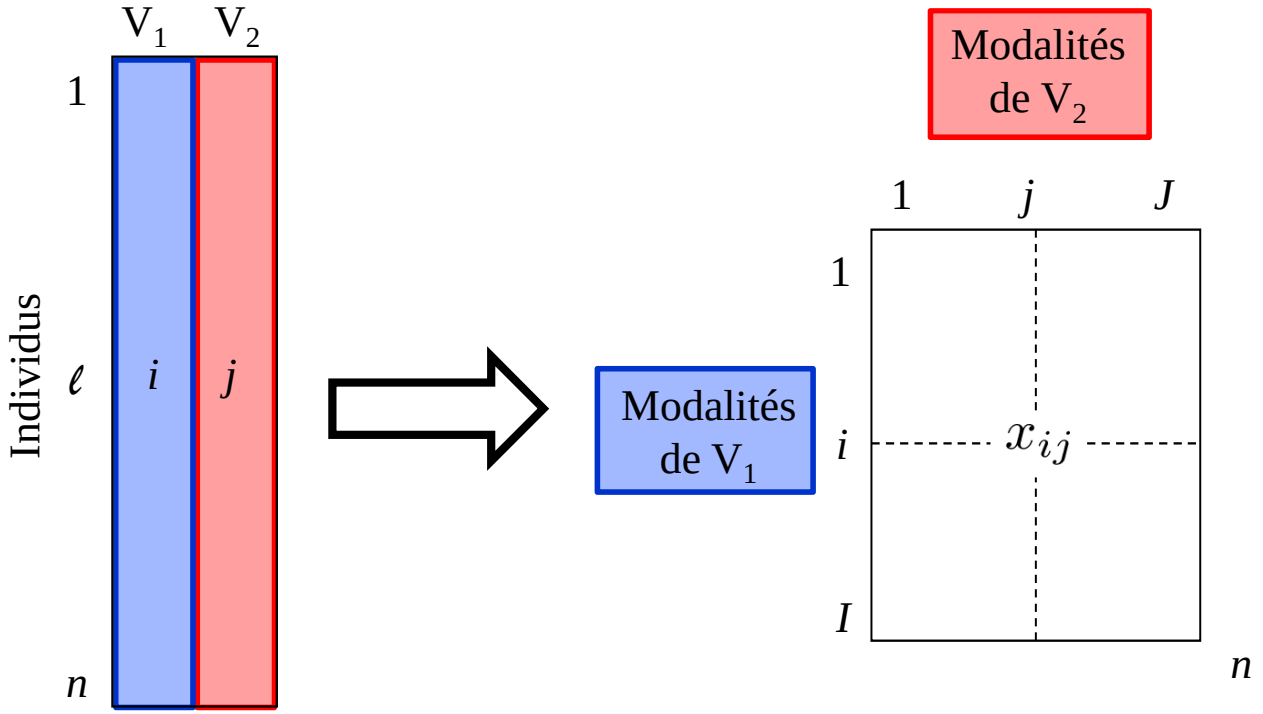
Données sur les prix Nobel

	Chimie	Economie	Littérature	Medecine	Paix	Physique	Somme
Allemagne	24	1	8	18	5	24	80
Canada	4	3	2	4	1	4	18
France	8	3	11	12	10	9	53
GB	23	6	7	26	11	20	93
Italie	1	1	6	5	1	5	19
Japon	6	0	2	3	1	11	23
Russie	4	3	5	2	3	10	27
USA	51	43	8	70	19	66	257
Somme	121	60	49	140	51	149	570

Y a-t'il un lien entre les pays et les catégories de prix ? Certains pays ont-ils des spécificités ?

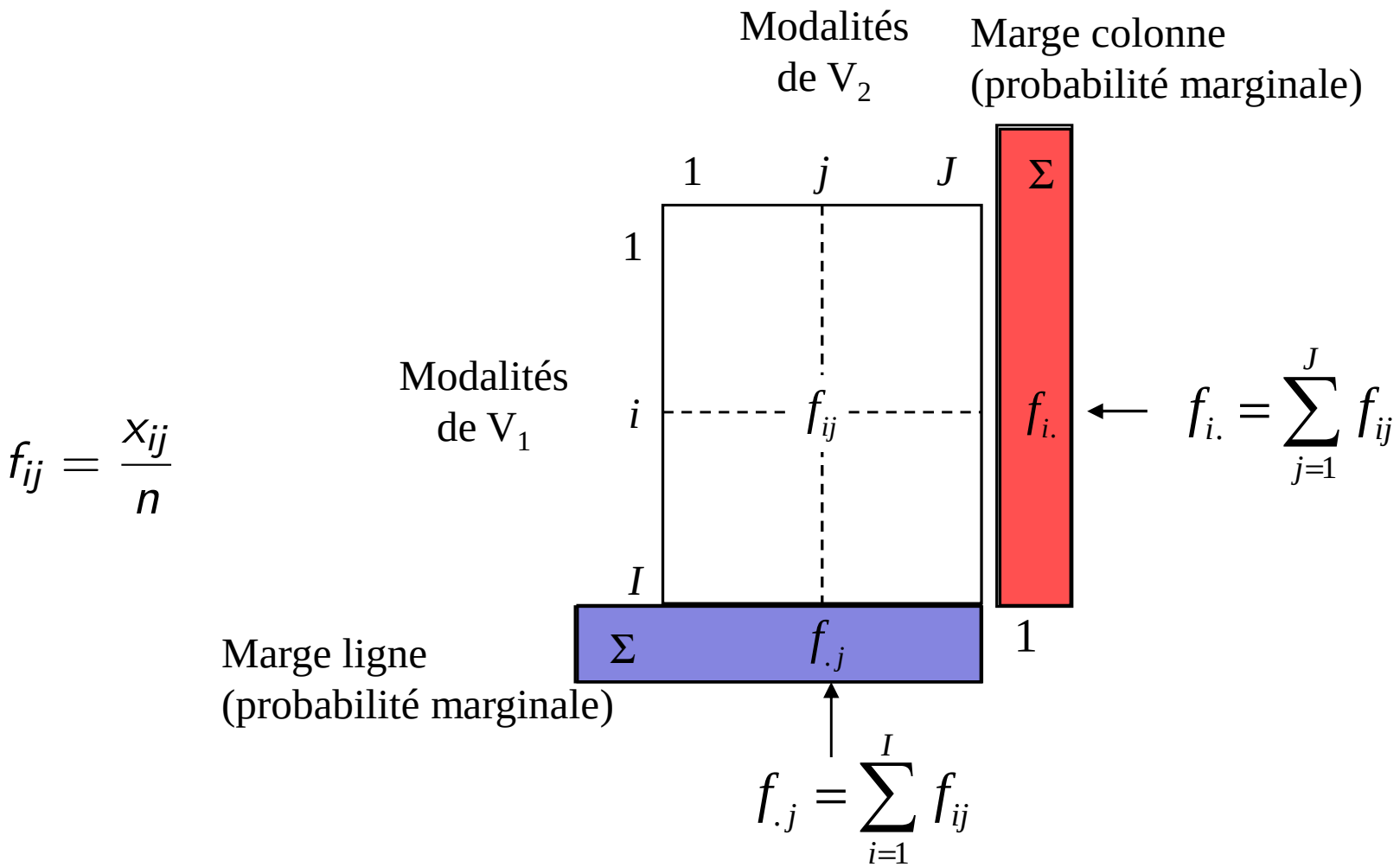
Données

n individus et 2 variables qualitatives



Distribution des *n* individus dans les $I \times J$ cases du tableau

Du tableau de contingences au tableau de probabilités



Liaison entre V1 et V2 : écart entre les données observées et le modèle d'indépendance

Analyse Factorielle des Correspondances (AFC)

Liaisons et indépendance entre deux variables qualitatives

Modèle d'indépendance :

Evènements indépendants : $P(A \text{ et } B) = P(A) \times P(B)$

Variables qualitatives indépendantes : $\forall i,\ \forall j,\ f_{ij} = f_{i.} \times f_{.j}$
$\Rightarrow$ Probabilité conjointe = produit des probabilités marginales

Autres écritures : $\dfrac{f_{ij}}{f_{i.}} = f_{.j}$ $\qquad$ $\dfrac{f_{ij}}{f_{.j}} = f_{i.}$
$\Rightarrow$ Probabilité conditionnelle = probabilité marginale

Liaisons entre deux variables qualitatives

Ecart entre données obs (f_{ij}) et modèle d'indépendance ($f_{i.}\, f_{.j}$)

1. Significativité de la liaison (de l'écart) : test du χ^2

$$\chi^2_{obs} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(\text{eff. observé} - \text{eff. théorique})^2}{\text{effectif théorique}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n\, f_{ij} - n\, f_{i.}\, f_{.j})^2}{n\, f_{i.}\, f_{.j}}$$

$$\chi^2_{obs} = \sum_{i=1}^{I}\sum_{j=1}^{J} n\, \frac{(\text{probabilité observée} - \text{probabilité théorique})^2}{\text{probabilité théorique}} = n\, \Phi^2$$

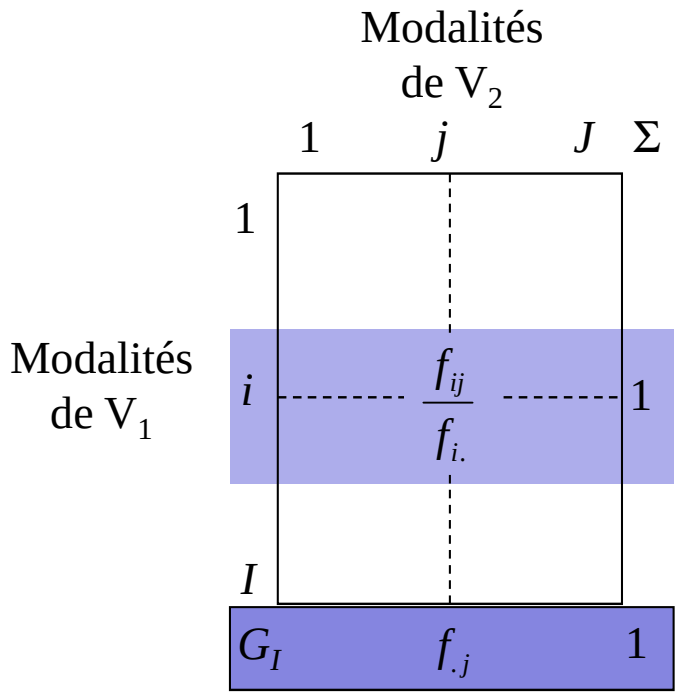
2. Intensite de la liaison $= \Phi^2 =$ écart entre probabilités théoriques et observées

3. Nature de la liaison = association entre modalités

L'AFC travaille sur le tableau des probabilités
ne dit rien sur la significativité
visualise la nature de la liaison entre les deux variables

Comment l'AFC appréhende l'écart à l'indépendance ?

Analyse par lignes : $\frac{f_{ij}}{f_{i.}} = f_{.j}$



Profil ligne *i* = distribution conditionnelle de V2 sachant que l'on possède la modalité *i* de V1

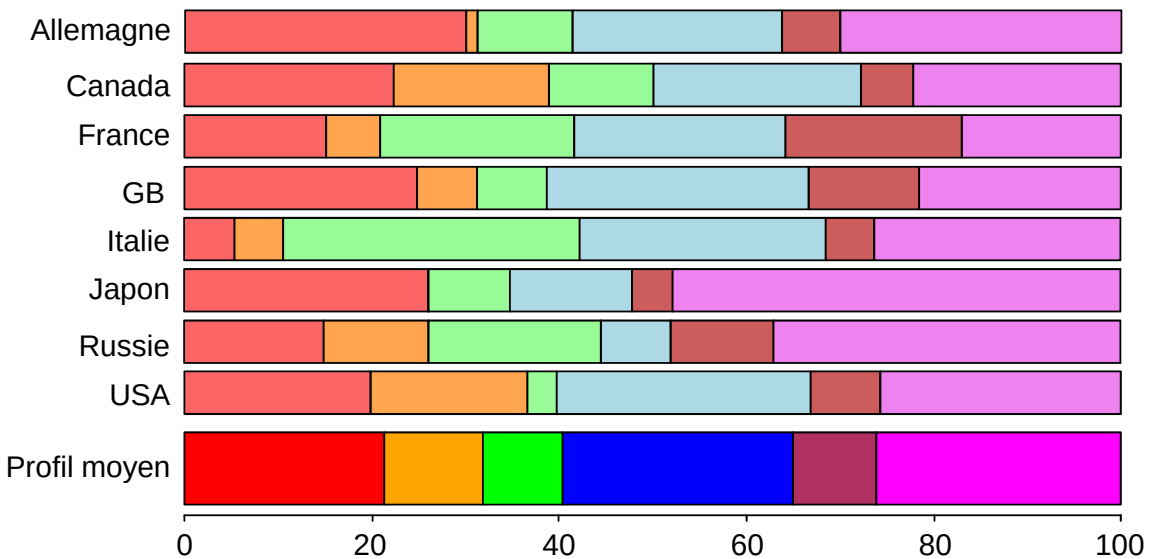
L'AFC compare les profils lignes au profil moyen

Profil ligne moyen = distribution marginale de V2
Profil de l'ensemble des individus étudiés

Approche multidimensionnelle de l'écart à l'indépendance

Comparaison du profil ligne au profil moyen

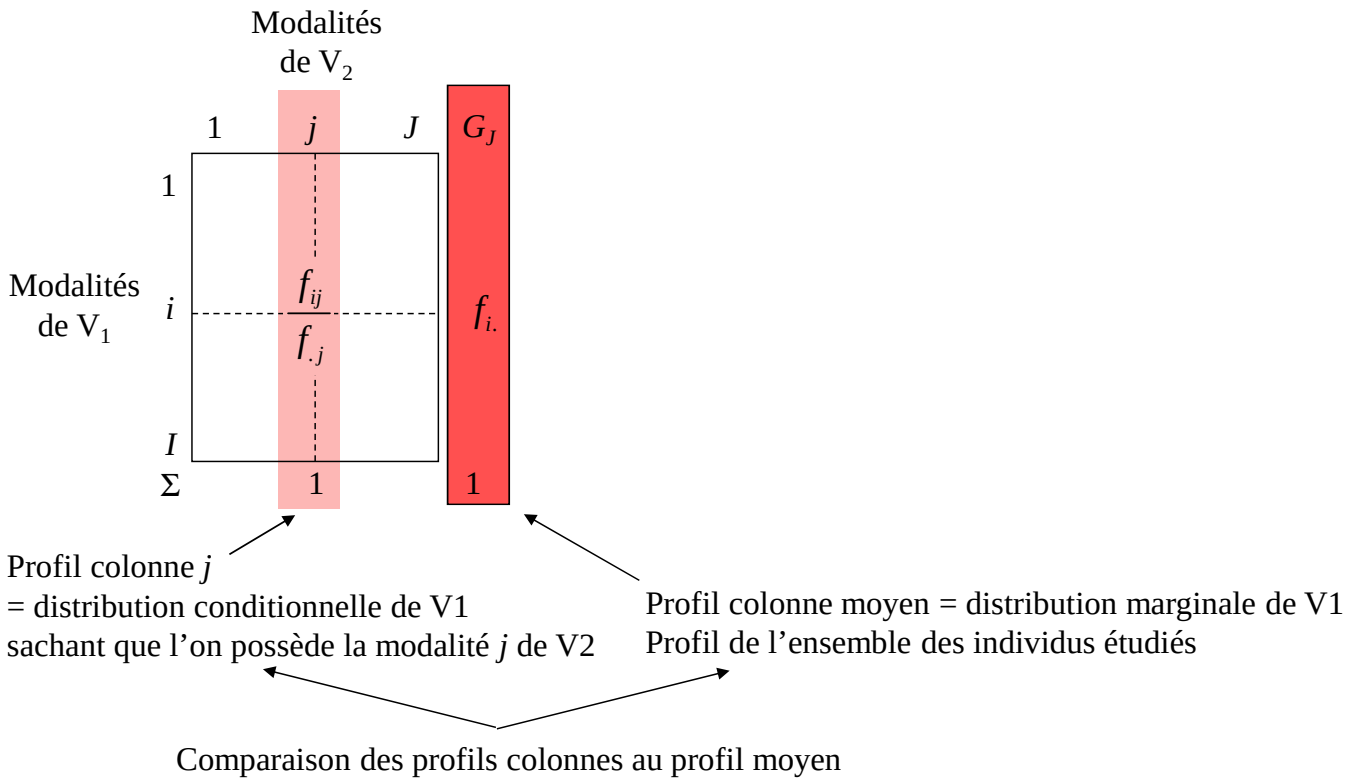
	Chimie	Eco	Lit.	Médecine	Paix	Physique	Somme
Allemagne	30.0	1.2	10.0	22.5	6.2	30.0	100
Canada	22.2	16.7	11.1	22.2	5.6	22.2	100
France	15.1	5.7	20.8	22.6	18.9	17.0	100
GB	24.7	6.5	7.5	28.0	11.8	21.5	100
Italie	5.3	5.3	31.6	26.3	5.3	26.3	100
Japon	26.1	0.0	8.7	13.0	4.3	47.8	100
Russie	14.8	11.1	18.5	7.4	11.1	37.0	100
USA	19.8	16.7	3.1	27.2	7.4	25.7	100
Profil moyen	21.2	10.5	8.6	24.6	8.9	26.1	100



Les Italiens obtiennent-ils des prix Nobel dans des disciplines particulières ?

Comment l'AFC appréhende l'écart à l'indépendance ?

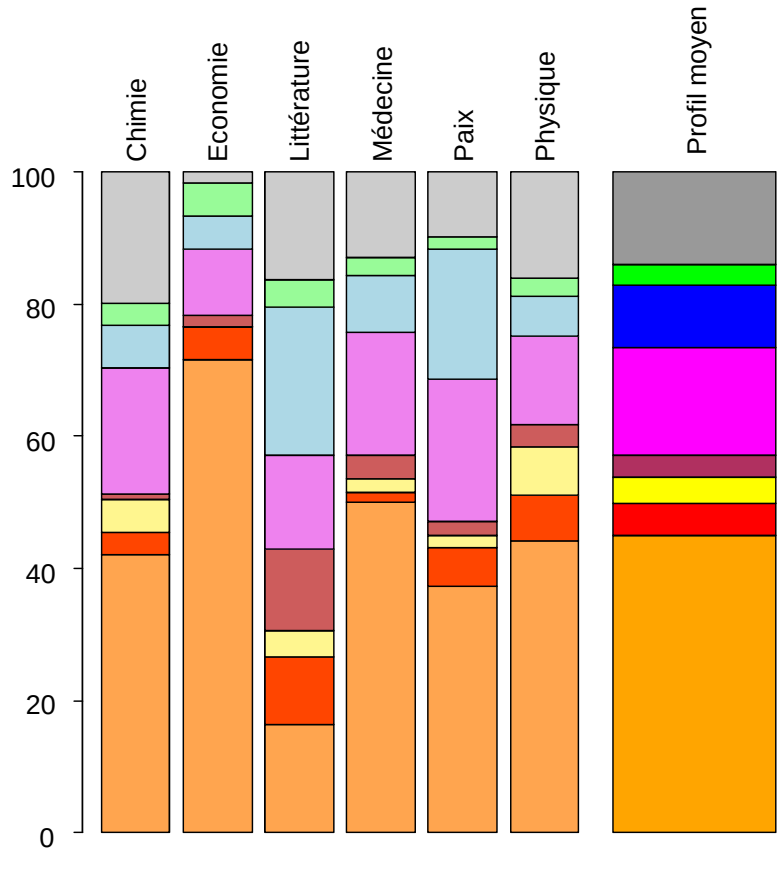
Analyse par colonnes : $\frac{f_{ij}}{f_{.j}} = f_{i.}$



Approche multidimensionnelle de l'écart à l'indépendance

Comparaison du profil colonne au profil moyen

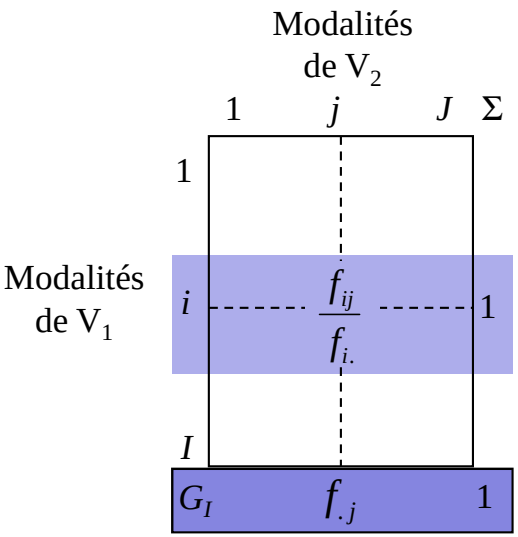
	Chimie	Eco	Lit	Méd	Paix	Phys	Profil moyen
Allemagne	19.8	1.7	16.3	12.9	9.8	16.1	14.0
Canada	3.3	5.0	4.1	2.9	2.0	2.7	3.2
France	6.6	5.0	22.4	8.6	19.6	6.0	9.3
GB	19.0	10.0	14.3	18.6	21.6	13.4	16.3
Italie	0.8	1.7	12.2	3.6	2.0	3.4	3.3
Japon	5.0	0.0	4.1	2.1	2.0	7.4	4.0
Russie	3.3	5.0	10.2	1.4	5.9	6.7	4.7
USA	42.1	71.7	16.3	50.0	37.3	44.3	45.1
Somme	100	100	100	100	100	100	100

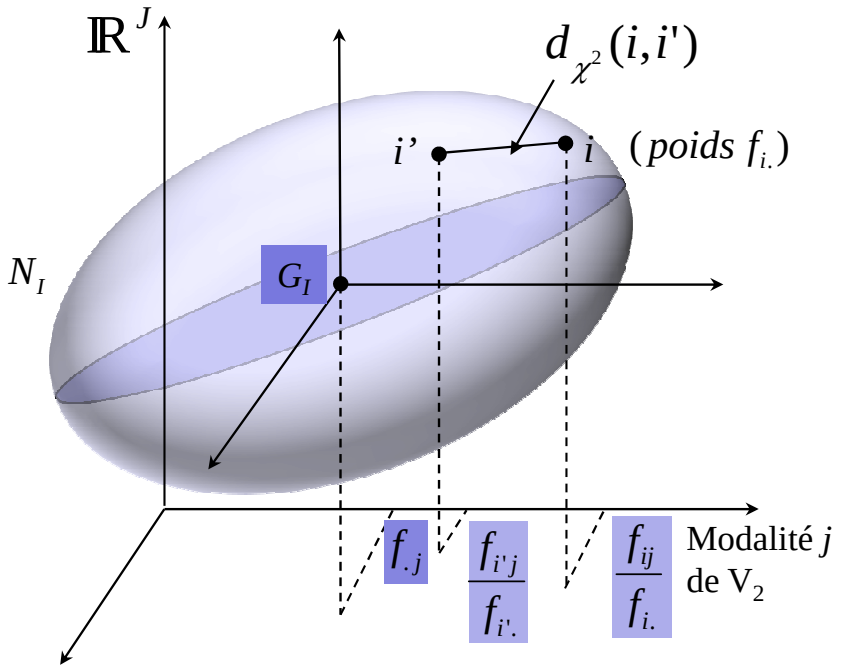


La répartition par pays des prix Nobel en littérature est elle la même que la répartition de l'ensemble des prix Nobel ?

Analyse Factorielle des Correspondances (AFC)

Le nuage des (profils) lignes

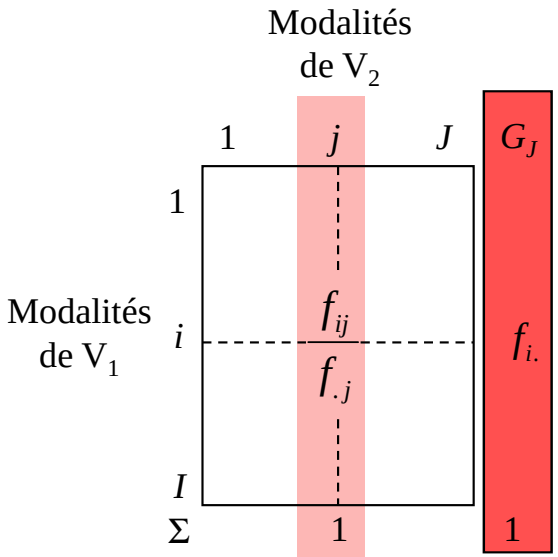


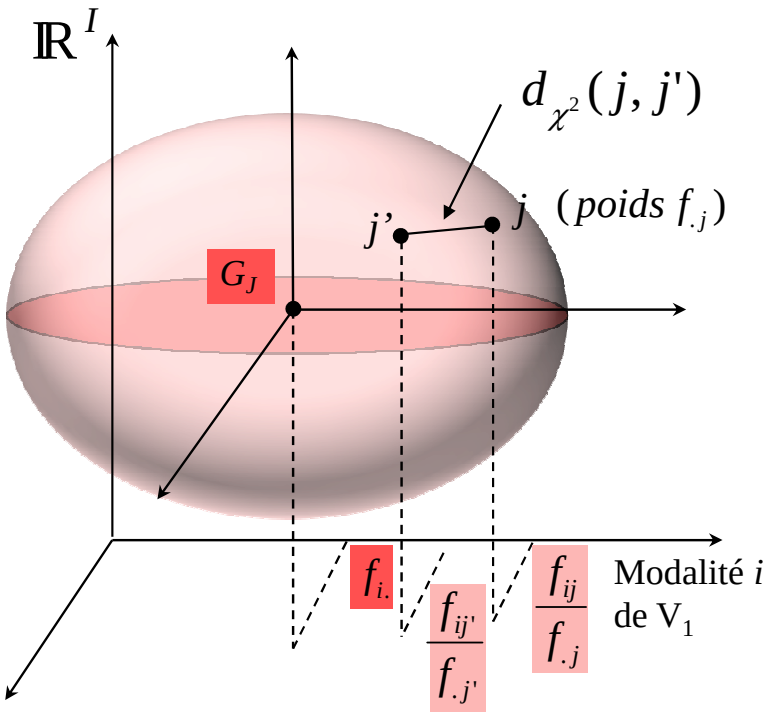


Distance entre deux profils : $d^2_{\chi^2}(i,i') = \sum_{j=1}^{J} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$

Distance au profil moyen G_I : $d^2_{\chi^2}(i,G_I) = \sum_{j=1}^{J} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - f_{.j} \right)^2$

Le nuage des (profils) colonnes





Distance entre deux profils : $d^2_{\chi^2}(j, j') = \sum_{i=1}^{I} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2$

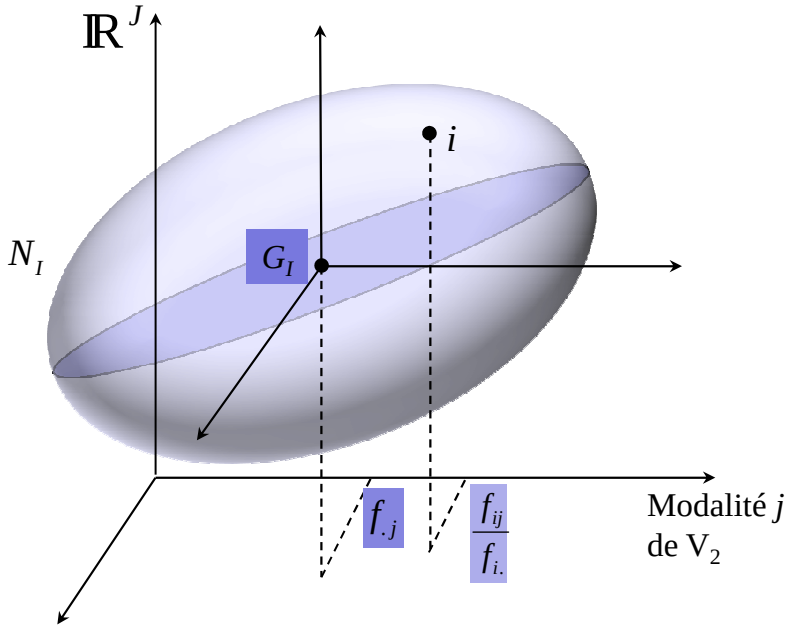
Distance au profil moyen G_J : $d^2_{\chi^2}(j, G_J) = \sum_{i=1}^{I} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - f_{i.} \right)^2$

Que se passe-t-il s'il y a indépendance ?

$$\text{Pour tout } i,\ \frac{f_{ij}}{f_{i.}} = f_{.j}$$

$\Rightarrow$ les profils sont confondus avec le profil moyen $\Rightarrow$ N_I réduit à G_I

$\Rightarrow$ L'inertie du nuage est nulle



Idem pour les colonnes : pour tout j, $\dfrac{f_{ij}}{f_{.j}} = f_{i.}$

Ecart à l'indépendance et inertie

Plus les données s'écartent de l'indépendance et plus les profils s'écartent de l'origine

$$\begin{aligned} Inertie(N_I/G_I) &= \sum_{i=1}^{I} Inertie(i/G_I) = \sum_{i=1}^{I} f_{i.} d^2_{\chi^2}(i, G_I) \\ &= \sum_{i=1}^{I} f_{i.} \left(\sum_{j=1}^{J} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - f_{.j} \right)^2 \right) \\ &= \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(f_{ij} - f_{i.} f_{.j})^2}{f_{i.} f_{.j}} = \frac{\chi^2}{n} = \phi^2 \end{aligned}$$

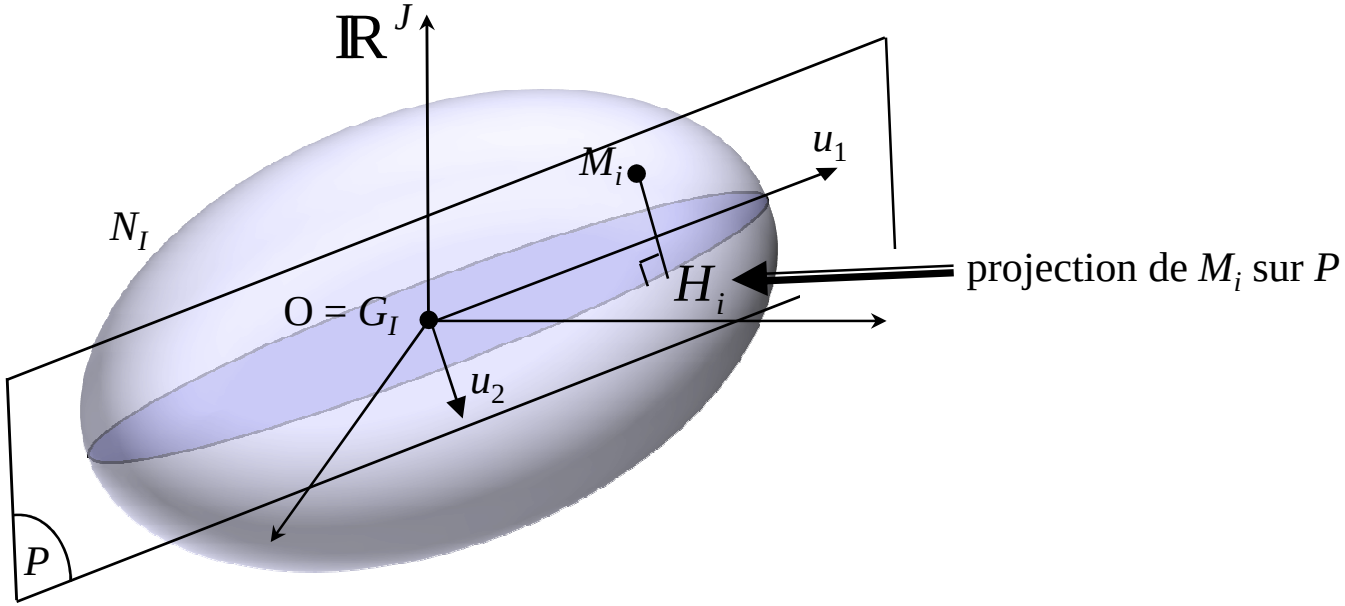
ϕ^2 mesure l'intensité de la liaison

Etudier l'inertie de N_I revient à étudier l'écart à l'indépendance

Idem pour N_J : $Inertie(N_J/G_J) = Inertie(N_I/G_I)$ (dualité)

Représentation du nuage des lignes (ou des colonnes)

Décomposition de l'inertie de N_I par analyse factorielle
Projection de N_I sur une suite d'axes orthogonaux d'inertie maximum

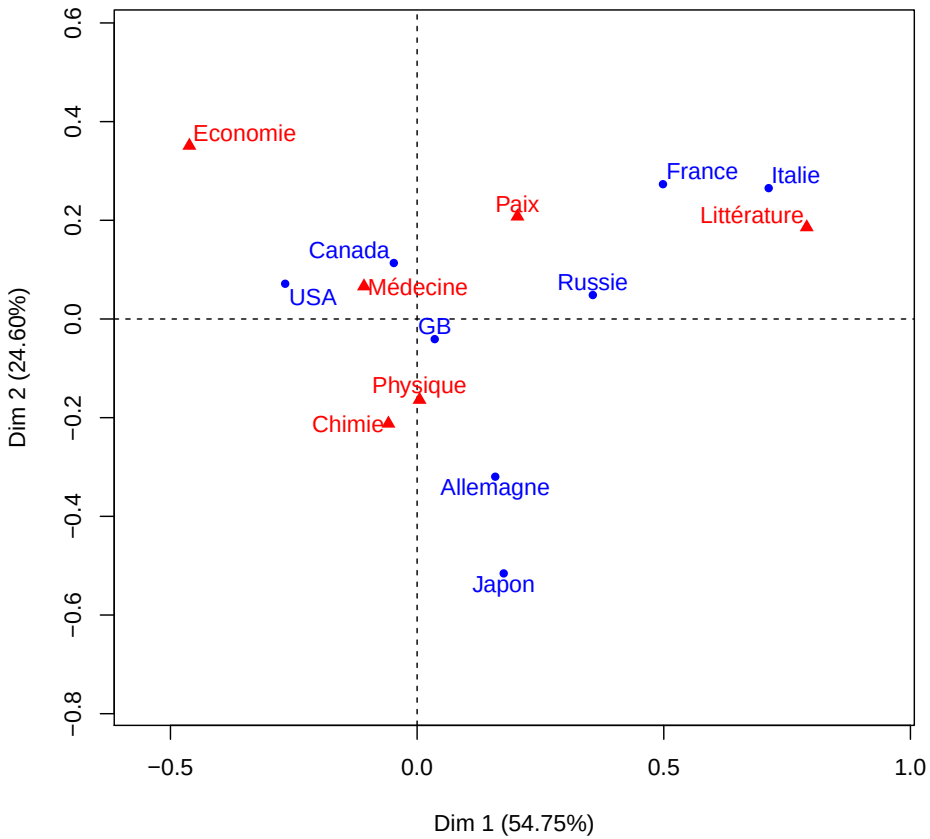


Trouver P tel que $\sum_{i=1}^{I} f_{i.} \, (OH_i)^2$ est maximum

u_1 axe d'inertie maximum
u_2 axe d'inertie maximum avec $u_2 \perp u_1$

Inertie associée à l'axe s : $\sum_{i=1}^{I} f_{i.} \, (OH_i^s)^2 = \lambda_s$

Règles d'interprétation sur l'exemple



1er axe : opposition sciences - autre catégorie

2ème axe : opposition physique/chimie - science éco

	Chimie	Economie	Littérature	Médecine	Paix	Physique	Somme
Italie	5.26	5.26	31.58	26.32	5.26	26.32	100
GB	24.73	6.45	7.53	27.96	11.83	21.51	100
Profil moyen	**21.23**	**10.53**	**8.60**	**24.56**	**8.95**	**26.14**	**100**

Analyse Factorielle des Correspondances (AFC)

Pourcentages d'inertie

1. Qualité de représentation de N_I par l'axe de rang s

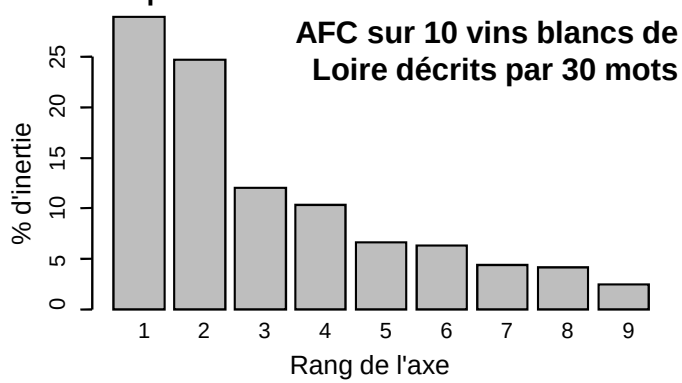
$$\frac{\text{inertie projetée de } N_I \text{ sur } u_s}{\text{inertie totale de } N_I} = \frac{\sum_{i=1}^{I} f_{i.} \left(OH_i^s\right)^2}{\sum_{i=1}^{I} f_{i.} \left(OM_i\right)^2} = \frac{\lambda_s}{\sum_{k=1}^{K} \lambda_k}$$

	Inertie	Inertie (%)
F1	0.0833	54.75
F2	0.0374	24.60
F3	0.0217	14.23
F4	0.0079	5.18
F5	0.0019	1.25
Sum	0.1522	100

$\Rightarrow$ Ecart à l'indépendance bien résumé par les deux premiers axes (79 %)

2. Inerties projetées s'additionnent d'un axe à l'autre (axes orthogonaux)
$\sum_{k=1}^{K} \lambda_k = \text{Inertie } (N_I) = \Phi^2$
Ici $n\Phi^2 = 570 \times 0.1522 = \chi^2 = 86.75 \quad \Rightarrow \quad$ Proba. critique $= 2.77\ 10^{-6}$

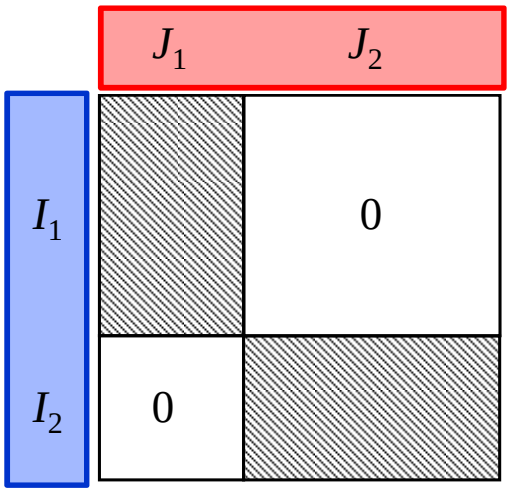
3. La décroissance des inerties suggère le nombre d'axes à conserver

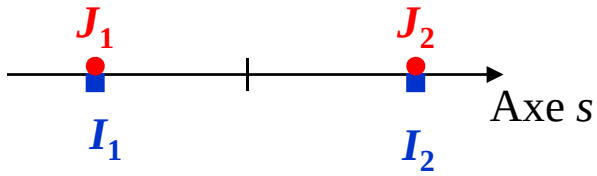


Inerties (= valeurs propres)

En AFC : $0 \leq \lambda_s \leq 1$ En ACP (normée) : $1 \leq \lambda_1$

A quelle structure correspond une valeur propre de 1 ?





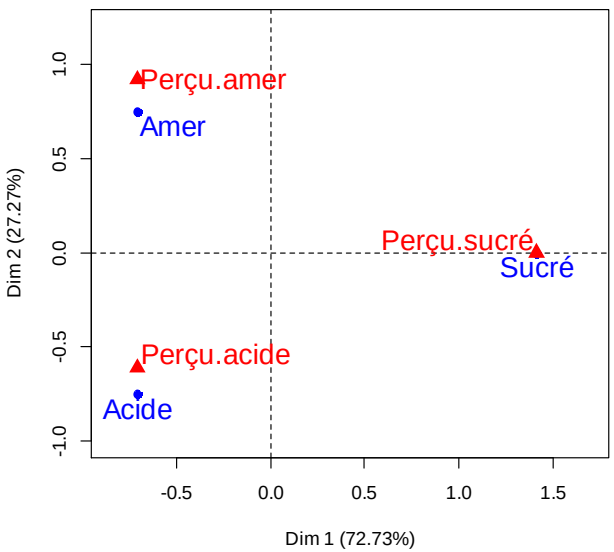
$\Rightarrow$ Partition en deux classes des lignes et des colonnes
Association exclusive des classes

Inerties (= valeurs propres)

Données : reconnaissance de trois saveurs (sucré, acide, amer)
Pour chaque saveur, on a demandé à dix personnes de reconnaître la saveur d'une solution qui leur était présentée

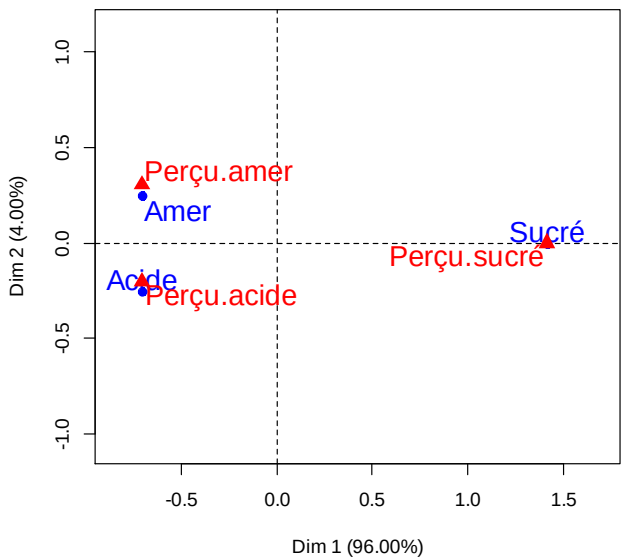
	Perçu sucré	Perçu acide	Perçu amer
Sucré	10	0	0
Acide	0	9	1
Amer	0	3	7

AFC	V. Propre	%
Axe 1	1	72,727
Axe 2	0,375	27,273
Somme	1,375	100



	Perçu sucré	Perçu acide	Perçu amer
Sucré	10	0	0
Acide	0	7	3
Amer	0	5	5

AFC	V. Propre	%
Axe 1	1	96
Axe 2	0,042	4
Somme	1,042	100



Inerties (= valeurs propres)

	Chimie	Economie	Littérature	Médecine	Paix	Physique
Allemagne	24	1	8	18	5	24
Canada	4	3	2	4	1	4
France	8	3	11	12	10	9
GB	23	6	7	26	11	20
Italie	1	1	6	5	1	5
Japon	6	0	2	3	1	11
Russie	4	3	5	2	3	10
USA	51	43	8	70	19	66

	Inertie	Inertie (%)
F1	0.0833	54.75
F2	0.0374	24.60
F3	0.0217	14.23
F4	0.0079	5.18
F5	0.0019	1.25
Sum	0.1522	100

$\lambda_1 = 0.0833 \ll 1 \Rightarrow$ on est loin d'une association exclusive entre une ligne et une colonne

$\Phi^2 = 0.1522 \ll 5 \Rightarrow$ on est loin d'une liaison parfaite, i.e. d'une association exclusive entre les modalités des deux variables

Analyse Factorielle des Correspondances (AFC)

Représentation simultanée des lignes et colonnes

Relation de transition = propriétés barycentriques

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

$F_s(i)$: coord. de la ligne i sur l'axe de rang s
$\frac{f_{ij}}{f_{i.}}$: jème élément du profil i
$G_s(j)$: coord. de la colonne j sur l'axe de rang s
λ_s : inertie associée à l'axe s (en AFC $\lambda_s \leq 1$)

Le long de l'axe de rang s, on calcule le barycentre de toutes les colonnes, chaque colonne j étant affectée du poids $f_{ij}/f_{i.}$

Le barycentre est ensuite d'autant plus écarté de l'origine que λ_s est petit : $1/\sqrt{\lambda_s} \geq 1$

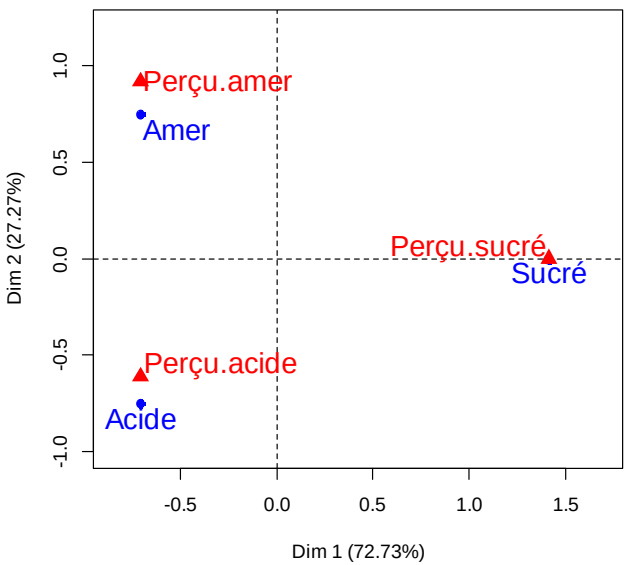
$$G_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^{I} \frac{f_{ij}}{f_{.j}} F_s(i)$$

Représentation simultanée et inertie

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

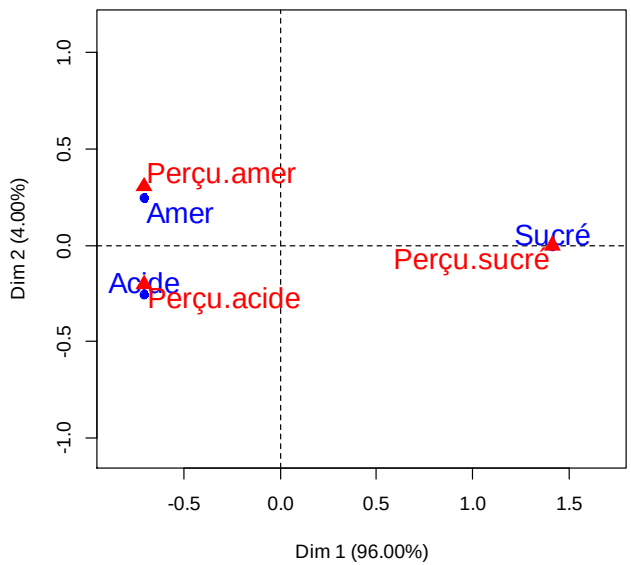
	Perçu sucré	Perçu acide	Perçu amer
Sucré	10	0	0
Acide	0	9	1
Amer	0	3	7

AFC	V. Propre	%
Axe 1	1	72,727
Axe 2	0,375	27,273
Somme	1,375	100



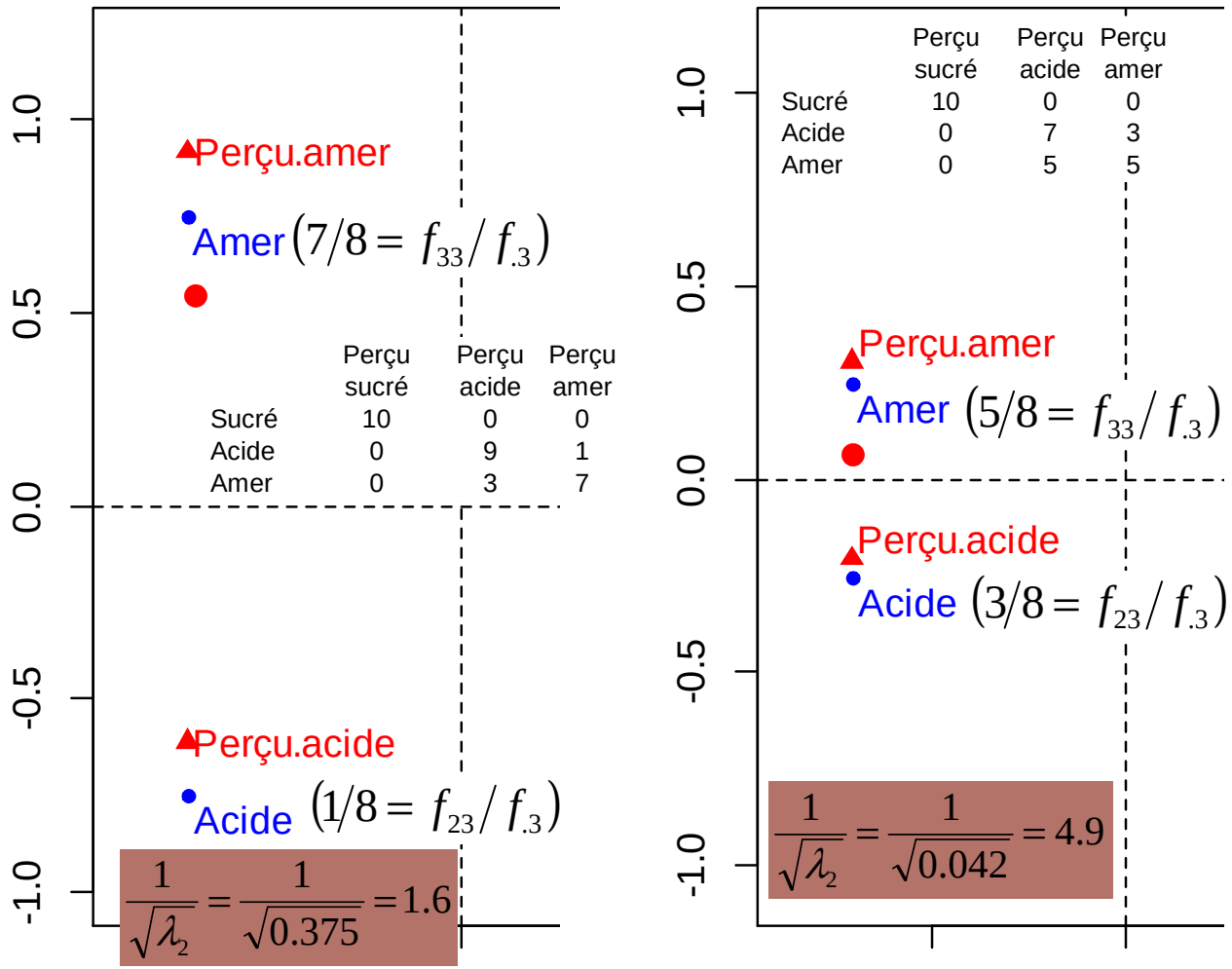
	Perçu sucré	Perçu acide	Perçu amer
Sucré	10	0	0
Acide	0	7	3
Amer	0	5	5

AFC	V. Propre	%
Axe 1	1	96
Axe 2	0,042	4
Somme	1,042	100

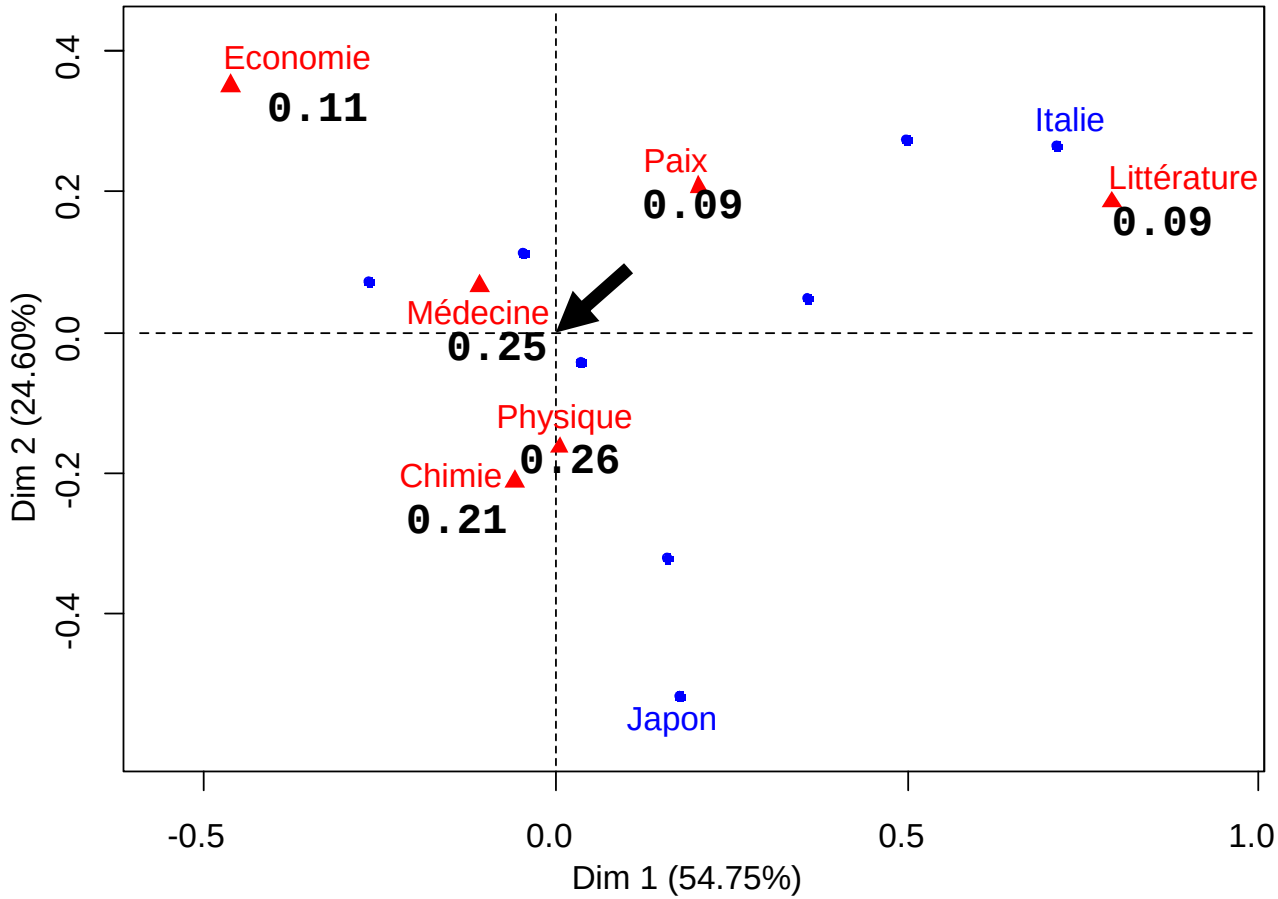


Représentation simultanée et inertie

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

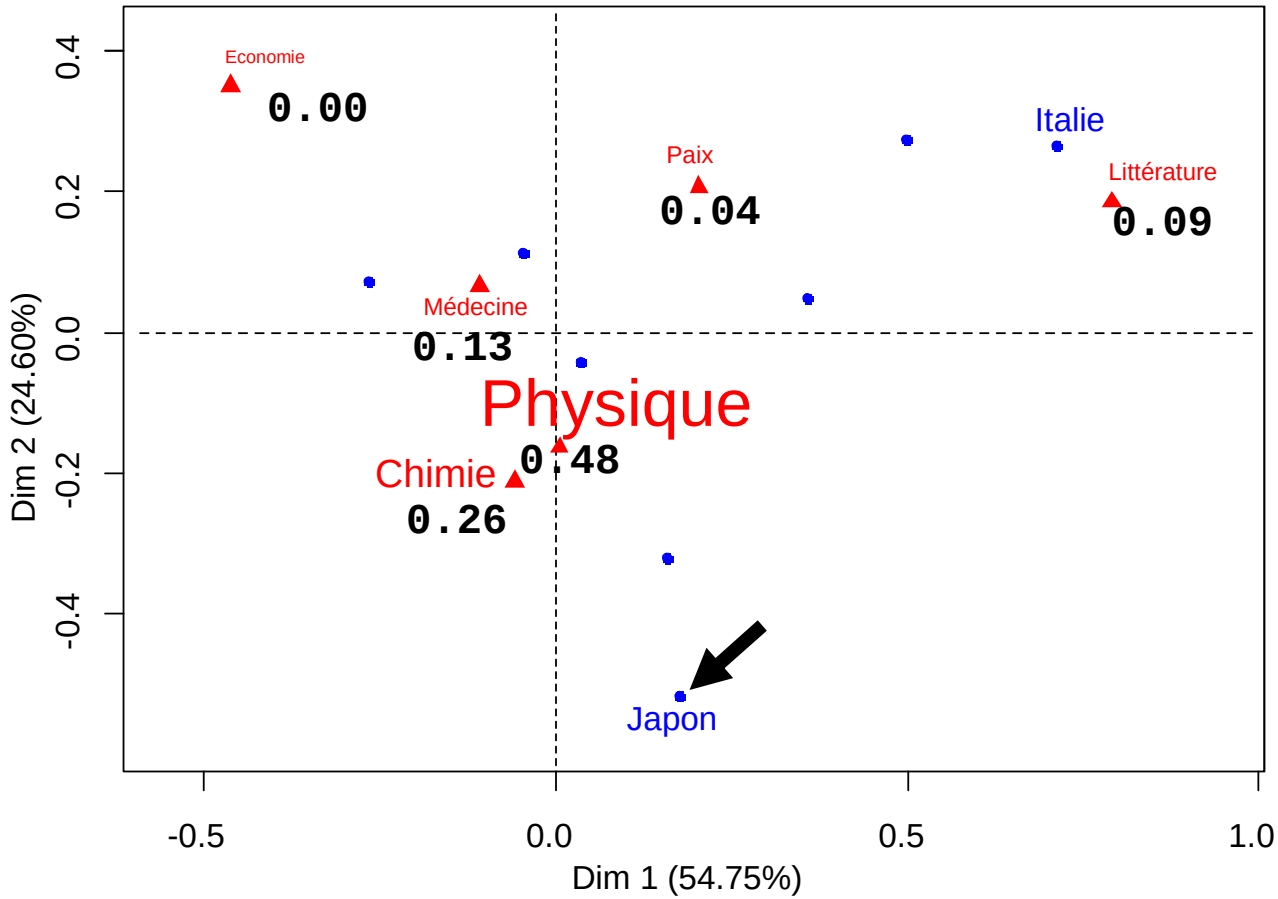


Propriété barycentrique



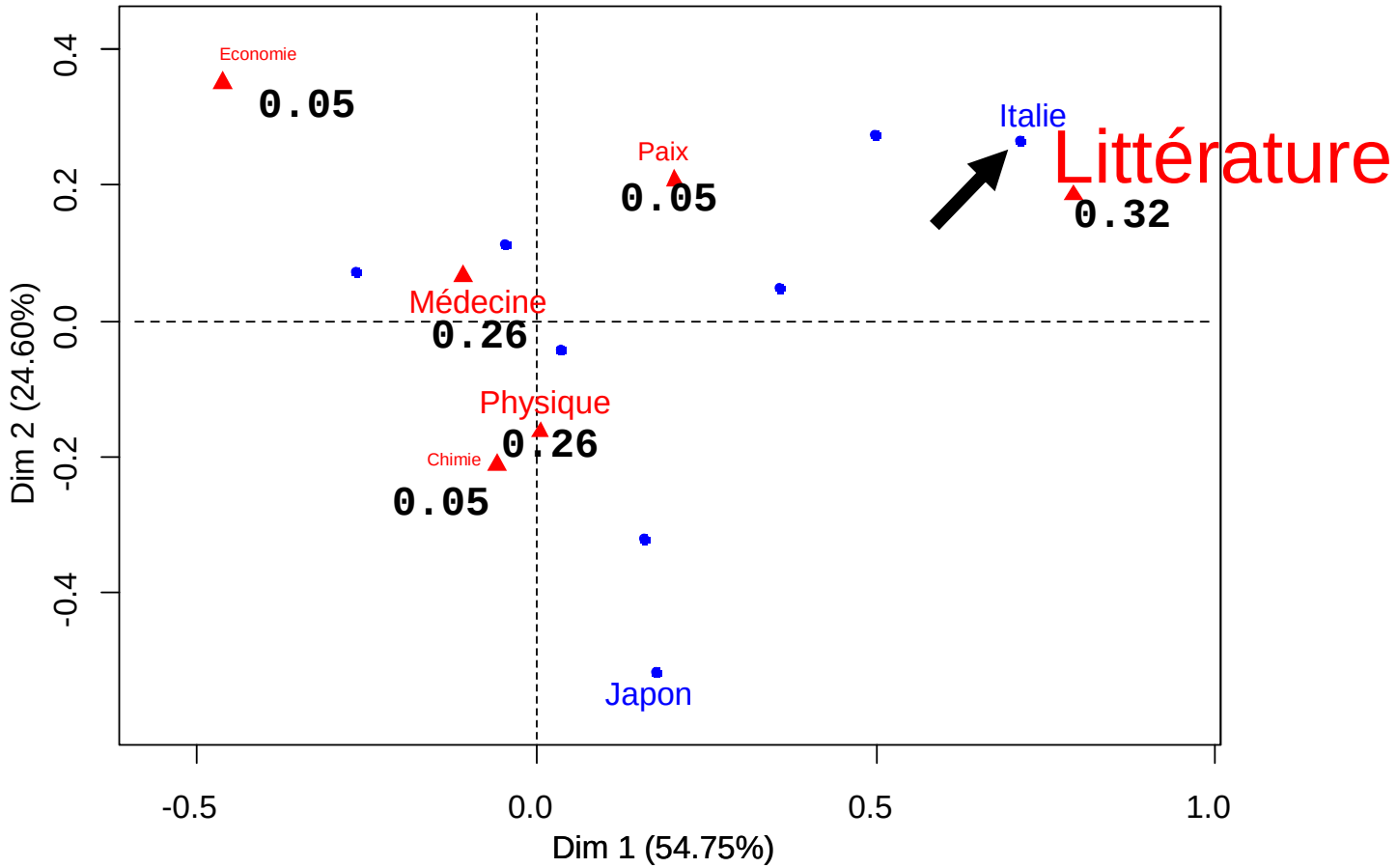
	Chimie	Economie	Littérature	Médecine	Paix	Physique
Italie	5.26	5.26	31.58	26.32	5.26	26.32
Japon	26.09	0.00	8.70	13.04	4.35	47.83
Profil moyen	21.23	10.53	8.60	24.56	8.95	26.14

Propriété barycentrique



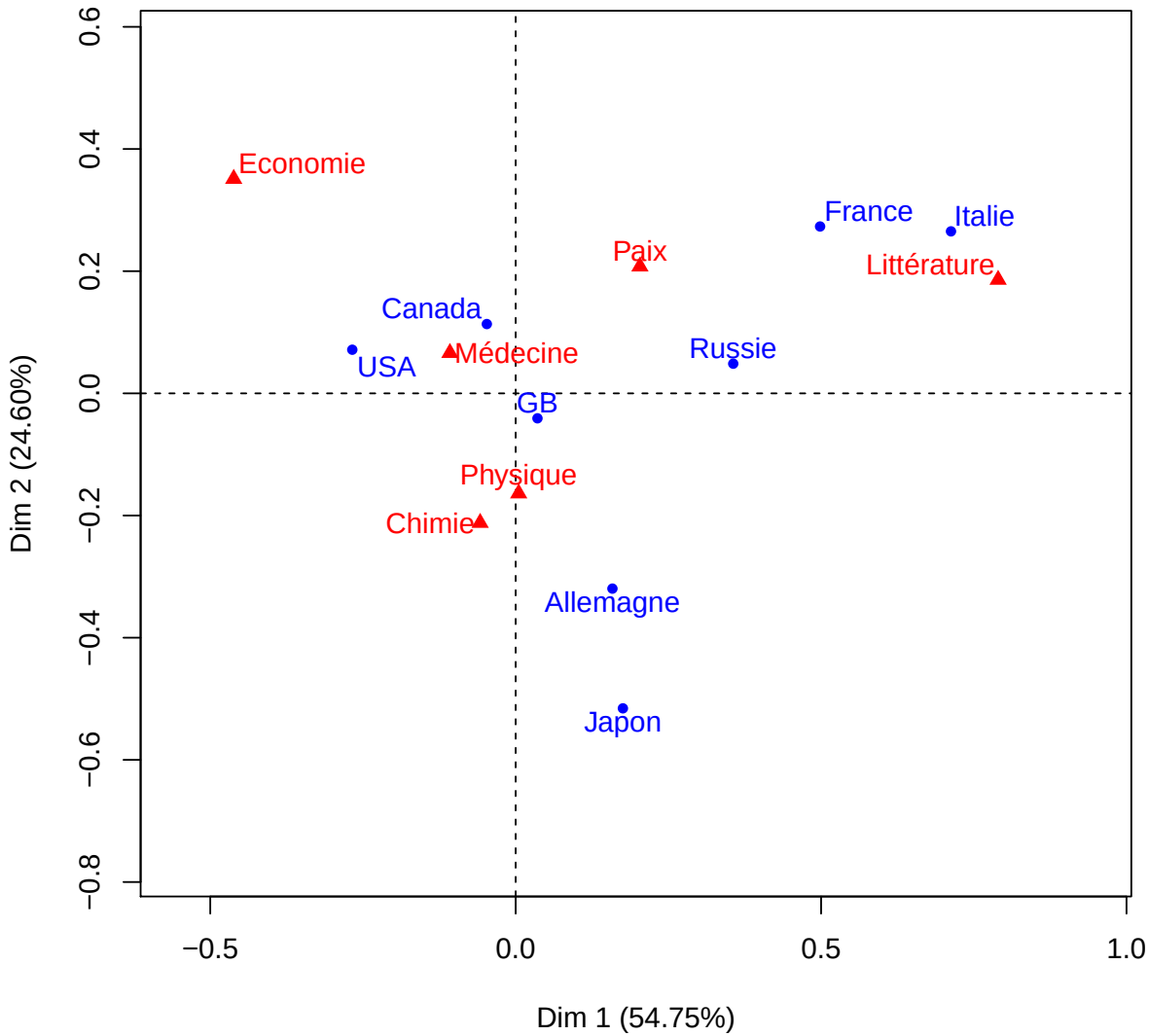
	Chimie	Economie	Littérature	Médecine	Paix	Physique
Italie	5.26	5.26	31.58	26.32	5.26	26.32
Japon	26.09	0.00	8.70	13.04	4.35	47.83
Profil moyen	21.23	10.53	8.60	24.56	8.95	26.14

Propriété barycentrique



	Chimie	Economie	Littérature	Médecine	Paix	Physique
Italie	5.26	5.26	31.58	26.32	5.26	26.32
Japon	26.09	0.00	8.70	13.04	4.35	47.83
Profil moyen	21.23	10.53	8.60	24.56	8.95	26.14

Propriété barycentrique

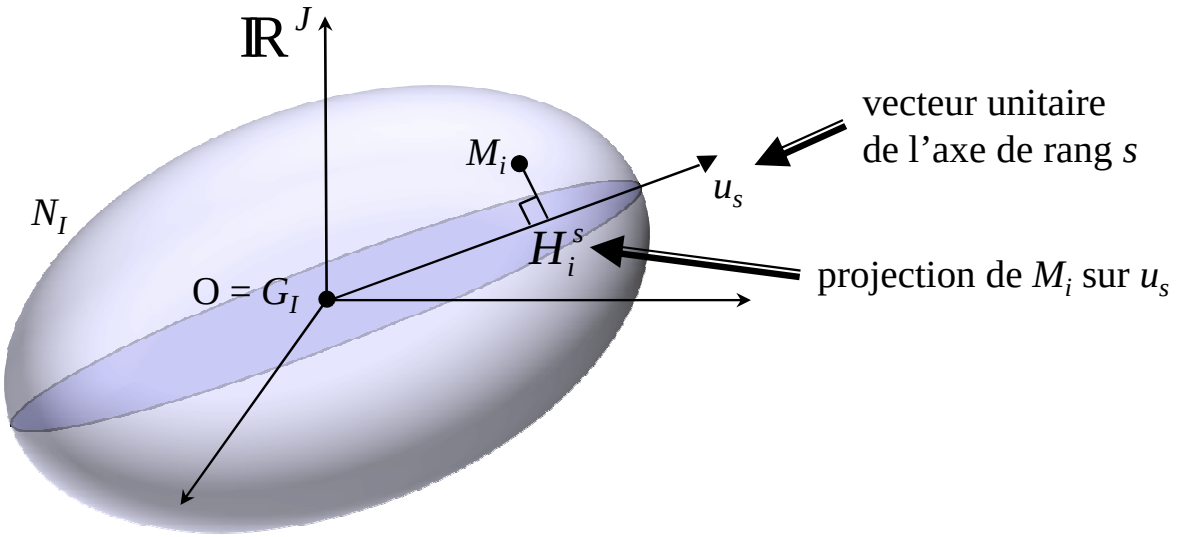


Analyse Factorielle des Correspondances (AFC)

Aides à l'interprétation : qualité de représentation

Indicateur de qualité de représentation d'un point (idem nuage) :

$$\frac{\text{inertie projetée de } M_i \text{ sur } u_s}{\text{inertie totale de } M_i} = \frac{f_{i.}(OH_i^s)^2}{f_{i.}(OM_i)^2} = \cos^2(\overrightarrow{OM_i}, u_s)$$

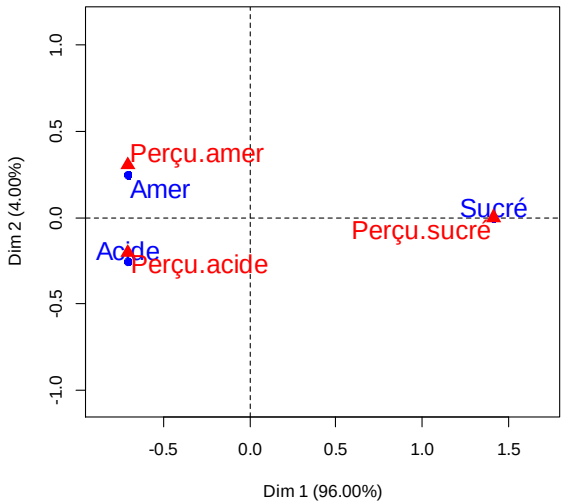


Indicateur montre dans quelle mesure l'écart d'un profil au profil moyen est complètement représenté par l'axe (ou par un plan)

Qualité de représentation : exemple

	Perçu sucré	Perçu acide	Perçu amer
Sucré	10	0	0
Acide	0	7	3
Amer	0	5	5

AFC	V. Propre	%
Axe 1	1	96
Axe 2	0,042	4
Somme	1,042	100



Qualité de représentation (cos^2)

	Axe1	Axe2
Sucré	1.000	0.000
Acide	0.889	0.111
Amer	0.889	0.111
Perçu.sucré	1.000	0.000
Perçu.acide	0.923	0.077
Perçu.amer	0.842	0.152

$\Rightarrow$ Interprétation des graphes basée sur points remarquables ayant une bonne qualité de représentation

Aides à l'interprétation : contribution

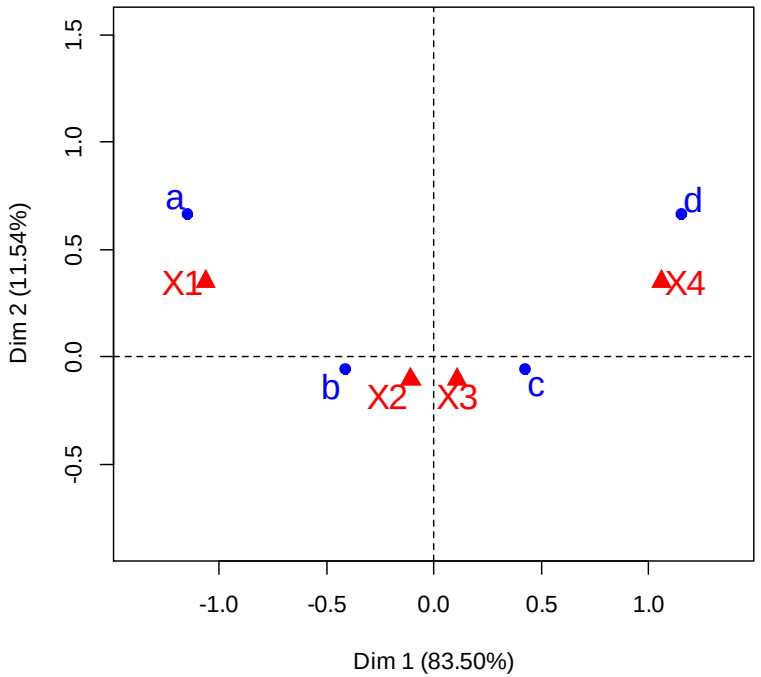
Indicateur brut : inertie projetée de M_i sur $u_s = f_{i.}(OH_i^s)^2$

Indicateur relatif : $\dfrac{\text{inertie proj. de } M_i \text{ sur } u_s}{\text{inertie de l'axe } s} = \dfrac{f_{i.}(OH_i^s)^2}{\lambda_s}$

- On peut additionner les contributions de plusieurs éléments
- Elles indiquent dans quelle mesure on peut considérer qu'un axe est dû à un élément ou à quelques éléments
- Compromis opérationnel entre distance à l'origine et poids
- Utiles pour les grands tableaux pour sélectionner un sous-ensemble d'éléments au début de l'interprétation (conjointement à la qualité de représentation)

Contribution : exemple

	X1	X2	X3	X4
a	1	1	0	0
b	5	10	10	0
c	0	10	10	5
d	0	0	1	1



	Inertie	%
Axe 1	0.258	83.501
Axe 2	0.036	11.538
Axe 3	0.015	4.96

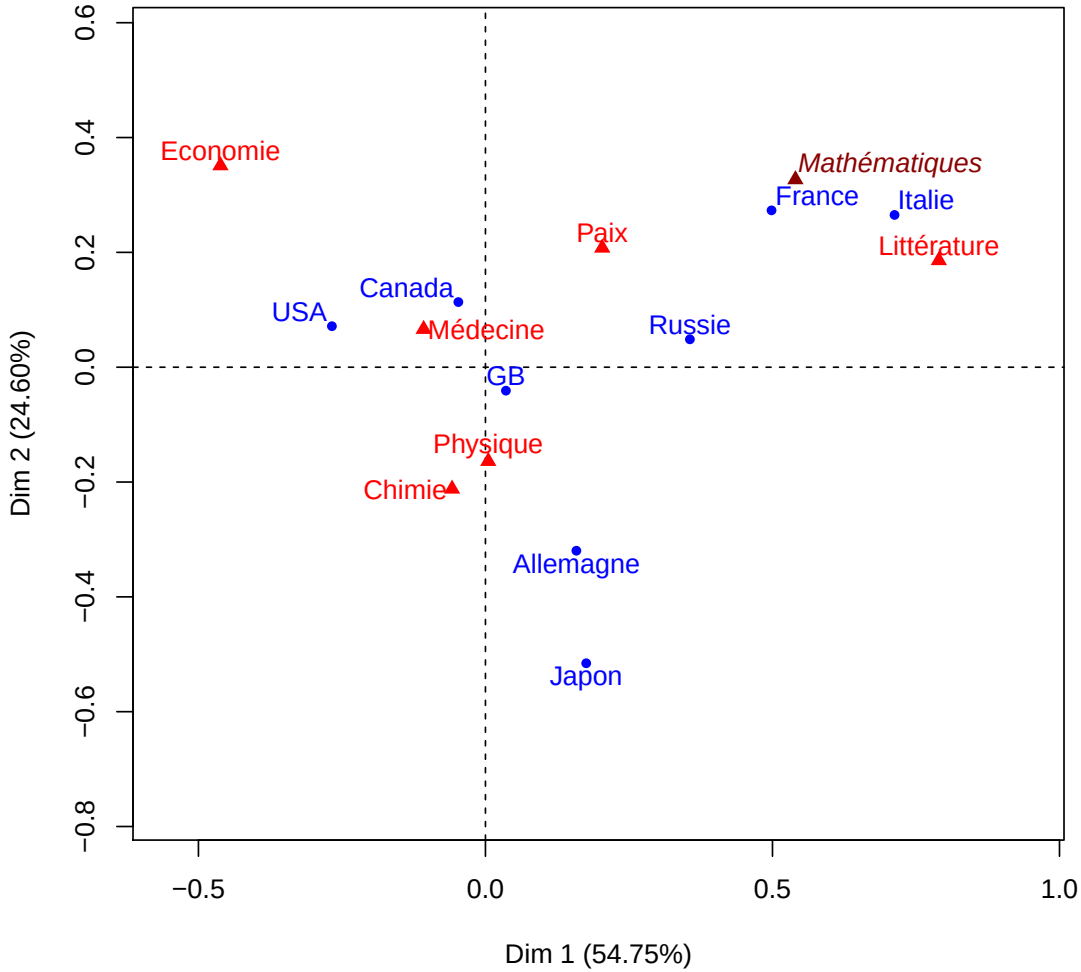
	Axe1	Axe2
a	18.879	46.296
b	31.121	3.704
c	31.121	3.704
d	18.879	46.296
Σ	100	100

$\Rightarrow$ Les points extrêmes ne sont pas nécessairement ceux qui contribuent le plus à la construction des axes

Eléments supplémentaires

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

Les mathématiques sont du côté de la France et de la Russie, et du côté de la littérature et de la paix, à l'opposé des sciences



Equivalence distributionnelle

Equivalence distributionnelle : si plusieurs lignes ayant le même profil sont regroupées en une seule, les résultats de l'AFC sont strictement équivalents (idem pour le regroupement de colonnes)

Application en analyse textuelle :
Grâce à l'équivalence distributionnelle, si 2 mots (ou plus) sont employés dans les mêmes circonstances, leurs coordonnées sont proches et faire l'analyse avec les deux termes ou avec un terme unique qui regroupe ces deux notions est strictement équivalent
$\Rightarrow$ notion très utile (regroupement des singuliers et pluriels, des conjugaisons des verbes, etc.)

Nombre maximum d'axes et V de Cramer

Nuage des lignes : I points dans un espace à J dimensions

$$\left.\begin{array}{r} J \text{ dim. mais 1 contrainte (profils)} \Rightarrow S \leq J-1 \\ I \text{ points évoluent dans au plus } I-1 \text{ dim.} \Rightarrow S \leq I-1 \end{array}\right\} S \leq \min(I-1, J-1)$$

$$\Longrightarrow \Phi^2 = \sum_{k=1}^{\min(I-1,J-1)} \lambda_k \leq \min(I-1, J-1)$$

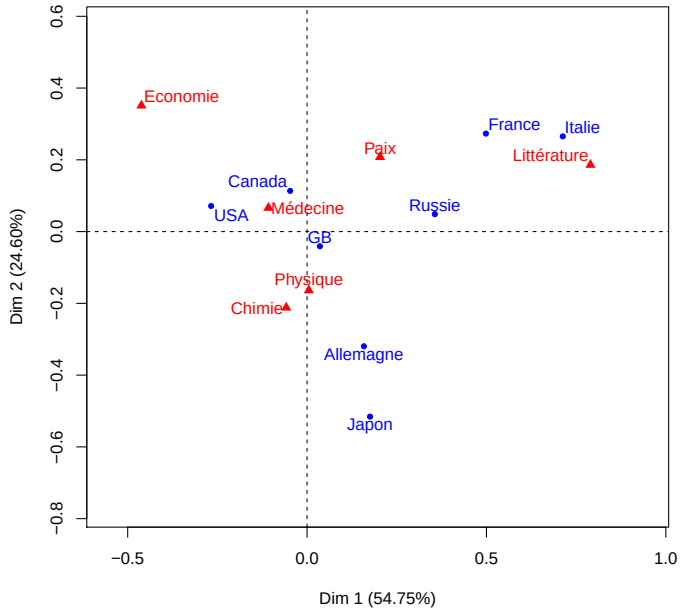
d'où l'idée d'un indicateur borné de la liaison entre 2 variables :

$$\text{V de Cramer} = \frac{\Phi^2}{\min(I-1, J-1)} \in [0;1]$$

	Prix Nobel	Trois saveurs	Trois saveurs
V de Cramer	$0.1522/5 = 0.03044$	$1.375/2 = 0.6875$	$1.042/2 = 0.521$

Bilan sur l'exemple

	Chimie	Economie	Littérature	Médecine	Paix	Physique
Allemagne	24	1	8	18	5	24
Canada	4	3	2	4	1	4
France	8	3	11	12	10	9
GB	23	6	7	26	11	20
Italie	1	1	6	5	1	5
Japon	6	0	2	3	1	11
Russie	4	3	5	2	3	10
USA	51	43	8	70	19	66



L'AFC apporte une visualisation synthétique de l'écart à l'indépendance qui aide la compréhension du tableau (a fortiori avec de grands tableaux)

Sur ces données

- L'essentiel de l'écart à l'indépendance est structuré par une opposition sciences - autres et dans une moindre mesure une opposition physique/chimie - science économique
- La position des pays illustre leur spécificité dans l'obtention des prix Nobel

Conclusion

Pour étudier la liaison entre deux variables qualitatives, on construit un tableau de contingence

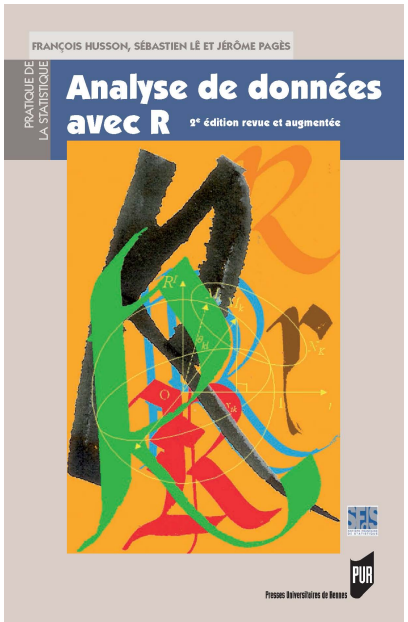
Cette liaison réside dans l'écart entre le tableau de contingence et le modèle d'indépendance

L'analyse des correspondances :

- construit un nuage des lignes (et un nuage des colonnes) dont l'inertie totale mesure l'intensité de l'écart à l'indépendance
- décompose cette inertie totale sur une suite d'axes d'importance décroisante représentant chacun un aspect synthétique de la liaison entre les deux variables
- fournit une représentation des lignes et des colonnes dans laquelle la position d'un point reflète sa participation à l'écart à l'indépendance

Bibliographie

Pour approfondir l'analyse des correspondances dans le même esprit que cette vidéo :



Husson F., Lê S. & Pagès J. (2017)
Exploratory Multivariate Analysis by Example Using R
2nd edition, 230 p., CRC/Press.