Analyse Factorielle des Correspondances

Denis Laloë

Analyse Factorielle des Correspondances

Février 2023

Denis Laloë
GABI - GiBBS

Février 2023

Introduction
Les analyses factorielles

Ensemble de méthodes partageant +/- une approche et une formalisation commune. Remonte au début du 20ème siècle

Quelques jalons

- ACP : Pearson K. 1901. On lines and planes of closest fit to systems of points in space. Philos Mag A. 6: 559-572
- ACP : Hotelling H. 1933. Analysis of a complex of statistical variables into principal components. J Educ Psychol. 25: 417-441
- SVD : Eckart C and Young G. 1936. The approximation of a matrix by another of a lower rank. Psychometrika. 1: 211-218.1
- Correspondances : Fisher, R.A. 1940. The precision of discriminant functions. Annals of Eugenics, 10, 422- 429.
- Correspondances : Benzécri J.P., 1965-66, Leçons sur l'analyse factorielle et la reconnaissance des formes, Rennes

Introduction

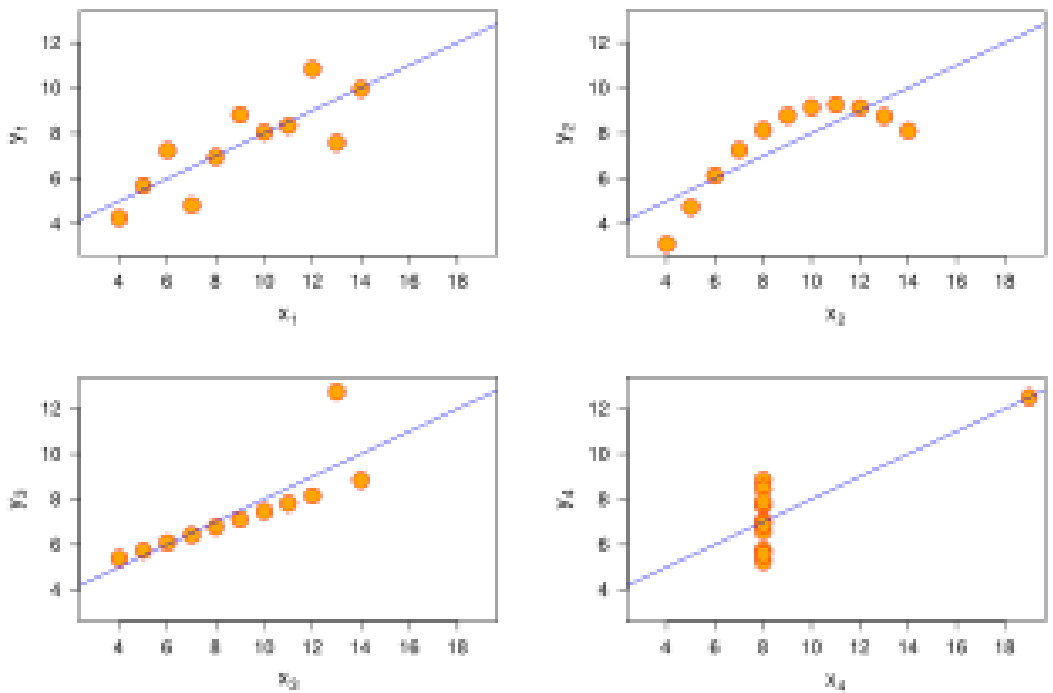
Données

- *Le modèle doit suivre les données, non l'inverse, J P Benzécri*
- Observation *vs* Expérimentation
  - Donnée préexistante (Sciences sociales / Ecologie)
  - Pas de structure a priori : induction
  - Synthèse (vision holistique / corrélations partielles / causalité) :
    - Approche de Durkheim : Pour dégager des relations causales, une relation binaire ne suffit pas, il faut intégrer plusieurs variables dans l'analyse et considérer leurs relations
    - Approche de Benzécri : C'est de la synthèse, sans a priori, que les causes émergent.

Quelques références : Armatte, Benzecri, Bressoux, Rouanet et Leroux

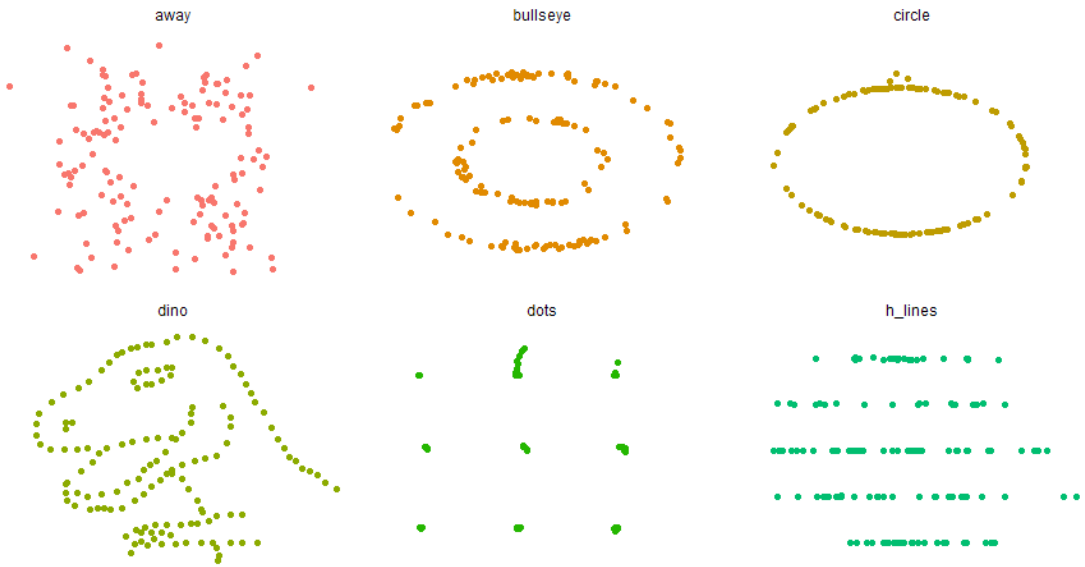
Représentation des données

Approche géométrique : représentation de données sous forme de nuages de points (plutôt que des résumés quantitatifs) F J Anscombe, 1973



Représentation des données

Approche géométrique : représentation de données sous forme de nuages de points (plutôt que des résumés quantitatifs)



Représentation des données

Approche géométrique : représentation de données sous forme de nuages de points (plutôt que des résumés quantitatifs)

Cleveland and McGill, 1984

The real power of a Cartesian graph does not derive only from one's ability to perceive the x and y values separately but from one's ability to understand the relationship of x and y.

Lewandowski et Spence, 1989

- Conservative judges of correlation, tending to estimate the squares of the correlation
- If outliers are present, they exhibit less bias in their estimates of correlation than do some robust numerical estimates

Analyse géométrique des données

- Approche géométrique : représentation de données sous forme de nuages de points (plutôt que des résumés quantitatifs)
- Efficience : synthèse par optimisation d'un critère (Inertie,...)

Une référence

Schéma de dualité

$$(\mathbf{X},\mathbf{Q},\mathbf{D}) \Leftrightarrow \begin{array}{ccc} \langle p \rangle & \xrightarrow{\mathbf{Q}} & \langle p \rangle \\ \mathbf{X}' \uparrow & & \downarrow \mathbf{X} \\ \langle n \rangle & \xleftarrow[\mathbf{D}]{} & \langle n \rangle \end{array}$$

- $\mathbf{X}_{(}\mathbf{n},\mathbf{p})$: matrice de données (transformées)
- $\mathbf{Q}_{(}\mathbf{p},\mathbf{p})$: matrice d'un produit scalaire, carrée symétrique
- $\mathbf{D}_{(}\mathbf{n},\mathbf{n})$: matrice d'un produit scalaire, carrée symétrique

Schéma de dualité - suite

Décomposition en valeurs singulières d'une matrice $\mathbf{M}_{(}\mathbf{n}, \mathbf{p})$:

- $\mathbf{M}_{(}\mathbf{n}, \mathbf{p}) = \mathbf{USV^t}$
- $\mathbf{U}_{(}\mathbf{n}, \mathbf{n})$: matrice de vecteurs singuliers (orthonormés)
- $\mathbf{V}_{(}\mathbf{p}, \mathbf{p})$: matrice de vecteurs singuliers (orthonormés)
- $\mathbf{\Lambda}_{(}\mathbf{n}, \mathbf{p})$: matrice diagonale des valeurs singulières (racines carrées des valeurs propres λ de $\mathbf{M^tM}$)
- Théorème de Eckart-Young - Meilleure approximation (norme de Frobenius) de rang k:
  $\mathbf{M^{[k]}} = \sum_{i=1}^{k} \sqrt{\lambda_i} \mathbf{u_i v_i^t}$

Schéma de dualité - suite

Schéma de dualité $< \mathbf{X}_{(}\mathbf{n}, \mathbf{p}), \mathbf{Q}_{(}\mathbf{p}, \mathbf{p}), \mathbf{D}_{(}\mathbf{n}, \mathbf{n}) >$

Analyse factorielle : Décomposition en valeurs singulières de la matrice $\mathbf{D^{1/2}XQ^{1/2}} = \mathbf{U\Lambda^{1/2}V^t}$

- $\mathbf{U}_{(}\mathbf{n}, \mathbf{n})$: base de l'espace des colonnes
- $\mathbf{V}_{(}\mathbf{p}, \mathbf{p})$: base de l'espace des lignes
- $\mathbf{\Lambda}_{(}\mathbf{n}, \mathbf{p}) = diag(\lambda_1, ..., \lambda_i, ...)$

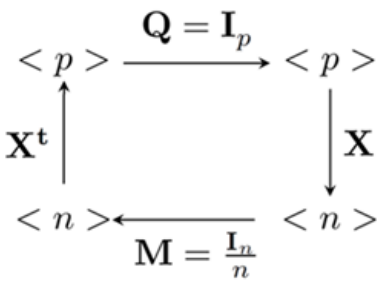
Schéma de dualité
ACP normée

$$(\mathbf{X},\mathbf{Q},\mathbf{D}) \Leftrightarrow \begin{array}{ccc} \langle p \rangle & \xrightarrow{\mathbf{Q}} & \langle p \rangle \\ \mathbf{X}' \uparrow & & \downarrow \mathbf{X} \\ \langle n \rangle & \xleftarrow[\mathbf{D}]{} & \langle n \rangle \end{array}$$

- $\mathbf{X}_{(}\mathbf{n},\mathbf{p})$: matrice de données (variables centrées réduites)
- $\mathbf{Q}_{(}\mathbf{p},\mathbf{p})$: matrice diagonale (poids des colonnes) $diag(1,...,1)$
- $\mathbf{D}_{(}\mathbf{n},\mathbf{n})$: : matrice diagonale (poids des lignes) $diag(1/n,...,1/n)$

Le schéma de dualité - ACP

Maximisation of the correlation between variables and components



Maximisation of the dispersion of individuals

Observations

V=X'X/*n*

W=XX' /*n*

Diagonalisation

X'X **XX'**

mêmes valeurs propres non nulles

$\lambda 1 > \lambda 2 > ... > \lambda r > 0$

$\Lambda = diag(\lambda_1, ..., \lambda_r)$

VA=AΔ
A'A=I
Principal axes

WB=BΔ
B'B=I
Principal components

Variable scores
C=X'B

Observation scores
L=XA

Transition formulae

$$XA\Delta^{-0,5} = B$$
$$X'B\Delta^{-0,5} = A$$

Singular value decomposition

$$X = B\Lambda^{0.5}A'$$

Best approximation (rank *l*)

Eckart and Young

$$\hat{\mathbf{X}}_l = \sum_{i=1,l} \sqrt{\lambda_i}\, \mathbf{b_i a_i'}$$

Table de contingence

Deux variables qualitatives A et B, avec respectivement I et J modalités

	B_1	...	B_j	...	B_J	Total
A_1	n_{11}	...	n_{ij}	...	n_{iJ}	$n_{1.}$
.	...	...	...	...	...	...
A_i	n_{i1}	...	n_{ij}	...	n_{iJ}	$n_{i.}$
.	...	...	...	...	...	...
A_I	n_{I1}	...	n_{Ij}	...	...	$n_{I.}$
Total	$n_{.1}$	...	$n_{.j}$	...	$n_{.J}$	$n_{..}$

Table de fréquences; distributions conditionnelles

	B_1	...	B_j	...	B_J	Marge colonnes
A_1	$f_{11} = \frac{n_{11}}{n_{..}}$	...	$f_{1j} = \frac{n_{1j}}{n_{..}}$	...	$f_{1J} = \frac{n_{1J}}{n_{..}}$	$f_{1.} = \frac{n_{1.}}{n_{..}}$
.	...	...	...	...	...	...
A_i	$f_{i1} = \frac{n_{i1}}{n_{..}}$	...	$f_{ij} = \frac{n_{ij}}{n_{..}}$	...	$f_{iJ} = \frac{n_{iJ}}{n_{..}}$	$f_{i.} = \frac{n_{i.}}{n_{..}}$
.	...	...	...	...	...	...
A_I	$f_{I1} = \frac{n_{I1}}{n_{..}}$	...	$f_{Ij} = \frac{n_{Ij}}{n_{..}}$	...	$f_{iJ} = \frac{n_{IJ}}{n_{..}}$	$f_{I.} = \frac{n_{I.}}{n_{..}}$
Marge lignes	$f_{.1} = \frac{n_{.1}}{n_{..}}$	...	$f_{.j} = \frac{n_{.j}}{n_{..}}$	...	$f_{.J} = \frac{n_{.J}}{n_{..}}$	1

Distributions conditionnelles

- $p(A = i|B = j) = \frac{p(A = i \cap B = j)}{p(B = j)} = \frac{nij}{n.j}$
- $p(B = j|A = i) = \frac{p(A = i \cap B = j)}{p(A = i)} = \frac{nij}{ni.}$

Distances entre points

distances (χ^2) entre profils

distance du χ^2 : distance euclidienne pondérée par $\frac{1}{f_{.j}}$

- Points ligne : $d^2(i,i') = \sum_{j=1}^{J} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$
- Points colonne : $d^2(j,j') = \sum_{i=1}^{I} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2$

Equivalence distributionnelle

On peut regrouper deux modalités ayant même profil, sans modification de l'analyse: on ne perd pas d'information en agrégeant des classes à profil identique; on n'en gagne pas en les subdivisant

Une table de contingence
Couleur des yeux * Couleur des cheveux

	C_Blond	C_Marron	C_Noir	C_Roux
Y_bleu	94	84	20	17
Y_Marron	7	119	68	26
Y_Noisette	10	54	15	14
Y_vert	16	29	5	14

De la table de contingence à la table des fréquences

Fréquences des cellules $f_{ij} = \frac{n_{ij}}{n..}$

	C_Blond	C_Marron	C_Noir	C_Roux
Y_bleu	0.1587838	0.1418919	0.0337838	0.0287162
Y_Marron	0.0118243	0.2010135	0.1148649	0.0439189
Y_Noisette	0.0168919	0.0912162	0.0253378	0.0236486
Y_vert	0.0270270	0.0489865	0.0084459	0.0236486

Fréquences marginales -lignes- $f_{i.} = \frac{n_{i.}}{n..}$

Y_bleu	Y_Marron	Y_Noisette	Y_vert
0.3631757	0.3716216	0.1570946	0.1081081

Fréquences marginales -colonnes- $f_{.j} = \frac{n_{.j}}{n..}$

C_Blond	C_Marron	C_Noir	C_Roux
0.214527	0.4831081	0.1824324	0.1199324

Distributions conditionnelles

Distributions conditionnelles par ligne $f_{i/j} = \frac{n_{ij}}{n.j}$

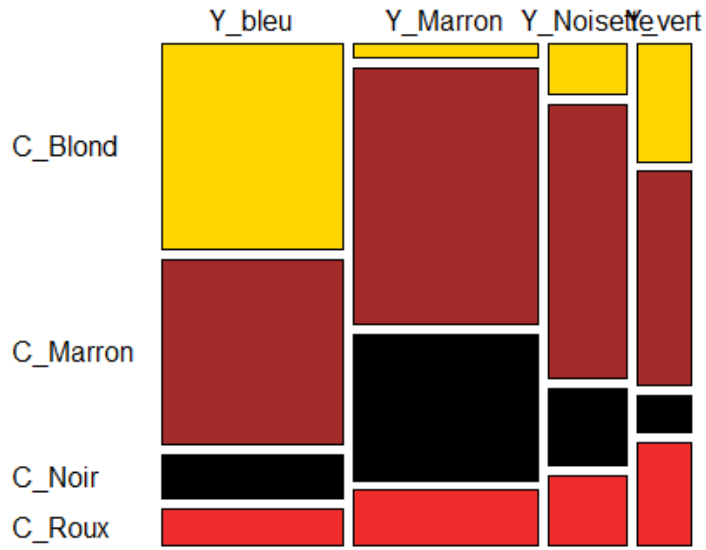
	C_Blond	C_Marron	C_Noir	C_Roux	Total
Y_bleu	0.4372093	0.3906977	0.0930233	0.0790698	1
Y_Marron	0.0318182	0.5409091	0.3090909	0.1181818	1
Y_Noisette	0.1075269	0.5806452	0.1612903	0.1505376	1
Y_vert	0.2500000	0.4531250	0.0781250	0.2187500	1

Distributions conditionnelles par colonne $f_{j/i} = \frac{n_{ij}}{ni.}$

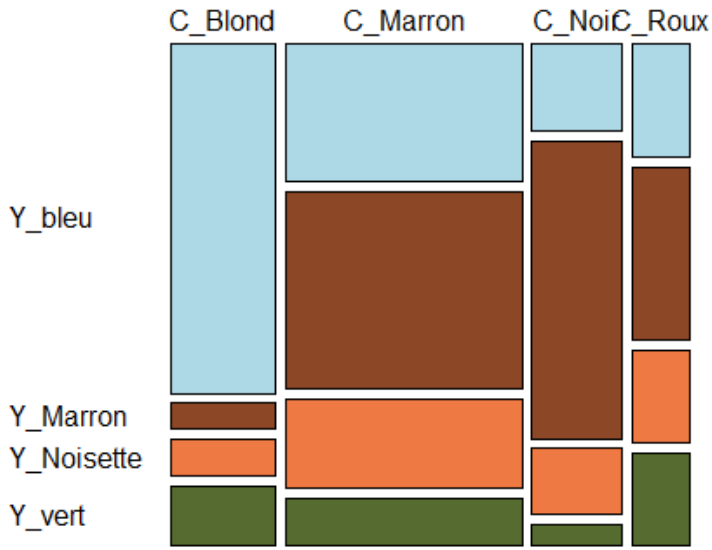
	C_Blond	C_Marron	C_Noir	C_Roux
Y_bleu	0.7401575	0.2937063	0.1851852	0.2394366
Y_Marron	0.0551181	0.4160839	0.6296296	0.3661972
Y_Noisette	0.0787402	0.1888112	0.1388889	0.1971831
Y_vert	0.1259843	0.1013986	0.0462963	0.1971831
Total	1	1	1	1

Une première représentation : les mosaicplots

Répartition Couleur des cheveux / Couleur des Yeux



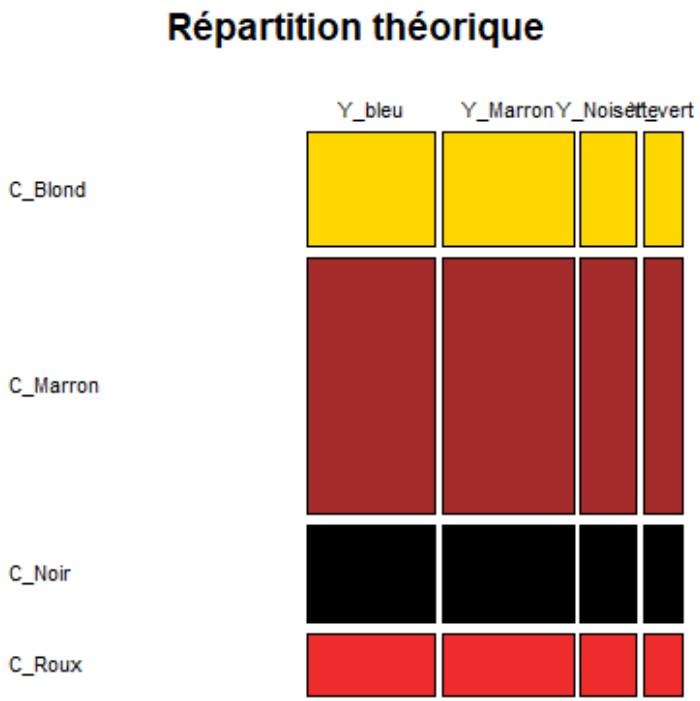
Répartition Couleur des yeux / Couleur des cheveux

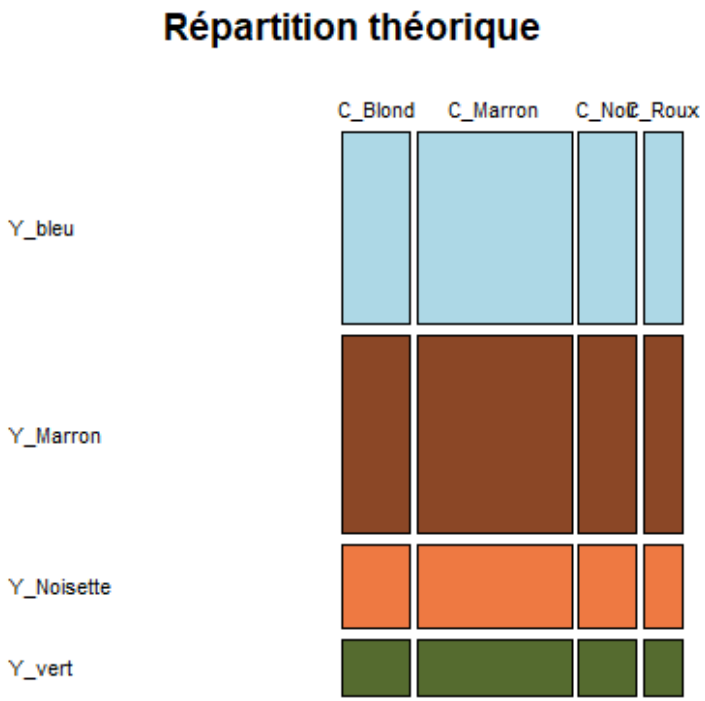


Indépendance des facteurs

Indépendance : $P(A = i \cap B = j) = P(A = i)P(B = j)$

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





Test de χ^2

Indépendance : $P(A = i \cap B = j) = P(A = i)P(B = j)$

$n_{ij} = \frac{n_{i.} n_{.j}}{n_{..}}$

Test de χ^2

$$\chi^2_{obs} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \frac{n_{i.} n_{.j}}{n_{..}})^2}{\frac{n_{i.} n_{.j}}{n_{..}}}$$

- Sous l'hypothèse d'indépendance, la statistique de test suit une loi de χ^2 à $(I-1)(J-1)$ degrés de liberté
- $\chi^2_{obs} = 138.29$
- ddl $= 9$
- $p(\chi^2 > 138.29) = 2.3 * 10^{-25}$

Distances entre points

distances (χ^2) entre profils

distance du χ^2 : distance euclidienne pondérée par $\frac{1}{f_{.j}}$

- Points ligne : $d^2(i,i') = \sum_{j=1}^{J} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$
- Points colonne : $d^2(j,j') = \sum_{i=1}^{I} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2$

Equivalence distributionnelle

On peut regrouper deux modalités ayant même profil, sans modification de l'analyse: on ne perd pas d'information en agrégeant des classes à profil identique; on n'en gagne pas en les subdivisant

Prime à la rareté

La distance du χ^2 pondère plus les modalités rares que la distance

Schéma de dualité de l'analyse des correspondances

Table de données **F**

Ecart à l'effectif théorique $\mathbf{F} = \dfrac{n_{ij} - \dfrac{n_{i.}n_{.j}}{n_{..}}}{\dfrac{n_{i.}n_{.j}}{n_{..}}}$

Poids des lignes

Fréquences marginales des lignes $f_{i.} = \dfrac{n_{i.}}{n..}$

Poids des colonnes

Fréquences marginales des lignes $f_{.j} = \dfrac{n_{.j}}{n..}$

Table de l'analyse

Centrage des lignes et des colonnes

$\mathbf{Ff.j} = 0$

$\mathbf{F^t fi.} = 0$

Rang de F

$rang(F) \leqslant min(I-1, J-1)$

Lien de l'inertie de la table avec la valeur du χ^2

$$\chi^2_{obs} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \frac{n_{i.} n_{.j}}{n_{..}})^2}{\frac{n_{i.} n_{.j}}{n_{..}}} = Inertie * n_{..}$$

Déroulé de l'analyse - 1

1. A partir de la table de contingence $T = n_{ij}$, création de
   $$\mathbf{F} = \frac{n_{ij} - \frac{n_{i.} n_{.j}}{n_{..}}}{\frac{n_{i.} n_{.j}}{n_{..}}}$$
2. calcul des fréquences marginales ligne et colonne:
   $\mathbf{D_l} = diag(f_{i.})$; $\mathbf{D_c} = diag(f_{.j})$
3. SVD de $\mathbf{X} = \mathbf{D_l^{1/2} F D_c^{1/2}} = \mathbf{U \Lambda^{1/2} V^t}$
   avec $\mathbf{V^t V} = \mathbf{U^t U} = \mathbf{I}$
   Décomposition de $\mathbf{X^t X} = \mathbf{V \Lambda V^t}$
   Décomposition de $\mathbf{X X^t} = \mathbf{U \Lambda U^t}$
4. Cooordonnées
   - Lignes: $\mathbf{L} = \mathbf{D_l^{-1/2} U \Lambda^{1/2}}$
   - Colonnes: $\mathbf{C} = \mathbf{D_c^{-1/2} V \Lambda^{1/2}}$

Déroulé de l'analyse - 2

- Relations de passage entre **U** et **V**
  - $\mathbf{V} = \mathbf{X}^{\mathbf{t}}\mathbf{U}\mathbf{\Lambda}^{-1/2}$
  - $\mathbf{U} = \mathbf{X}\mathbf{V}\mathbf{\Lambda}^{-1/2}$
- Relations de transition entre colonnes et lignes. On passe, à un coefficient de dilatation près, des coordonnées des lignes aux coordonnées des colonnes:
  - $C(i,k) = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{J} f_{j/i} L(j,k)$
  - $L(j,k) = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{I} f_{i/j} C(i,k)$

Aides à l'interprétation

Contribution absolue

Comment une ligne (colonne) explique un axe

- Contribution d'une ligne i à l'inertie de l'axe k:

$$contribligne(i,k) = \frac{l^2(i,k) * D_l(i)}{\lambda_k}$$

- Contribution d'une colonne j à l'inertie de l'axe k:

$$contribcol(i,k) = \frac{c^2(j,k) * D_c(j)}{\lambda_k}$$

Contribution relative

Comment un axe explique une ligne ou une colonne: Cosinus carré de la ligne (ou colonne) avec l'axe.

Packages R

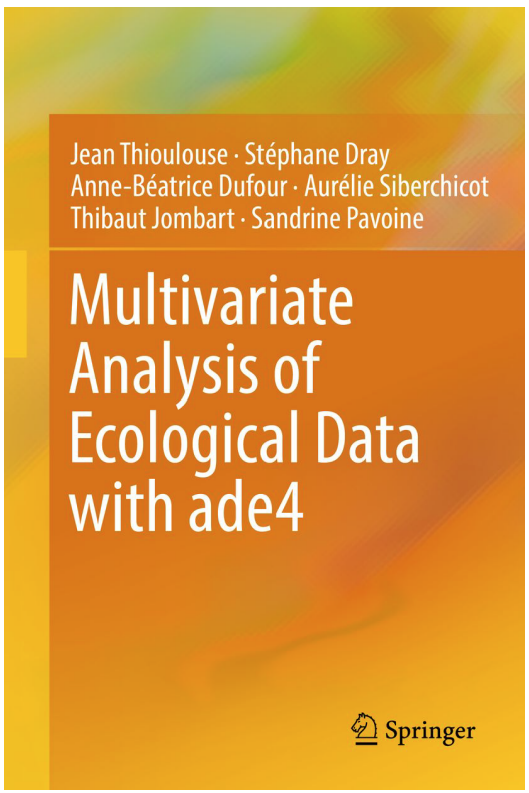
De nombreux packages R font de l'analyse des correspondances. Deux en particulier:

- ade4. https://pbil.univ-lyon1.fr/ADE-4/
- FactoMineR http://factominer.free.fr/

Livres, MOOC, matériel pédagogique...

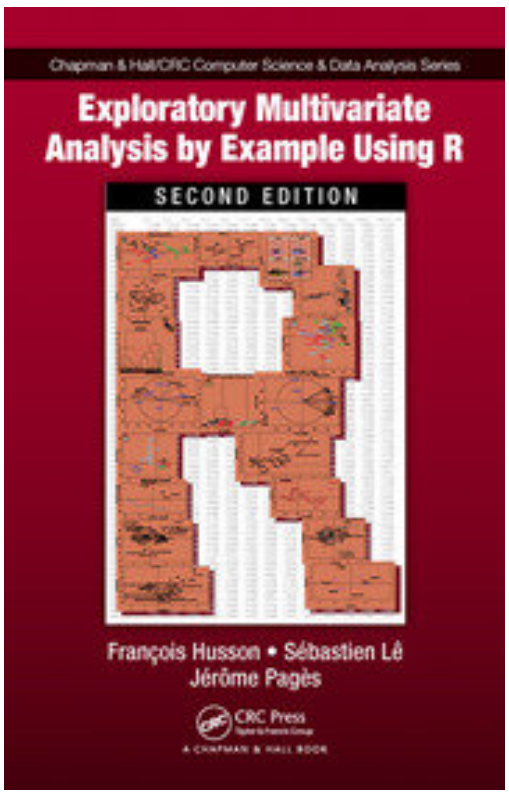
ade4

ade4 est un logiciel développé au laboratoire de Biométrie et Biologie Évolutive (UMR 5558) de l'Université Lyon 1. Il contient des fonctions d'Analyse de Données destinée d'abord à la manipulation des données Écologiques et Environnementales avec des procédures Exploratoires d'essence Euclidienne, d'où la dénomination ade4.



FactoMineR

FactoMineR est un package R dédié à l'analyse exploratoire multidimensionnelle de données (à la Française). Il a été développé et il est maintenu par François Husson, Julie Josse, Sébastien Lê, d'Agrocampus Rennes, et J. Mazet.



CA - FactoMineR

fonction CA de FactoMineR
args(CA)
function (X, ncp = 5, row.sup = NULL, col.sup = NULL, quanti.sup = NULL, quali.sup = NULL, graph = TRUE, axes = c(1, 2), row.w = NULL, excl = NULL)

CA - FactoMineR

Correspondence Analysis (CA) Description Performs Correspondence Analysis (CA) including supplementary row and/or column points.
Usage CA(X, ncp = 5, row.sup = NULL, col.sup = NULL, quanti.sup=NULL, quali.sup = NULL, graph = TRUE, axes = c(1,2), row.w = NULL, excl=NULL)
Arguments X a data frame or a table with n rows and p columns, i.e. a contingency table
ncp number of dimensions kept in the results (by default 5)
graph boolean, if TRUE a graph is displayed axes a length 2 vector specifying the components to plot

?CA

Value Returns a list including:
eig a matrix containing all the eigenvalues, the percentage of variance and the cumulative percentage of variance
col a list of matrices with all the results for the column variable (coordinates, square cosine, contributions, inertia)
row a list of matrices with all the results for the row variable (coordinates, square cosine, contributions, inertia)
Returns the row and column points factor map. The plot may be improved using the argument autolab, modifying the size of the labels or selecting some elements thanks to the plot.CA function

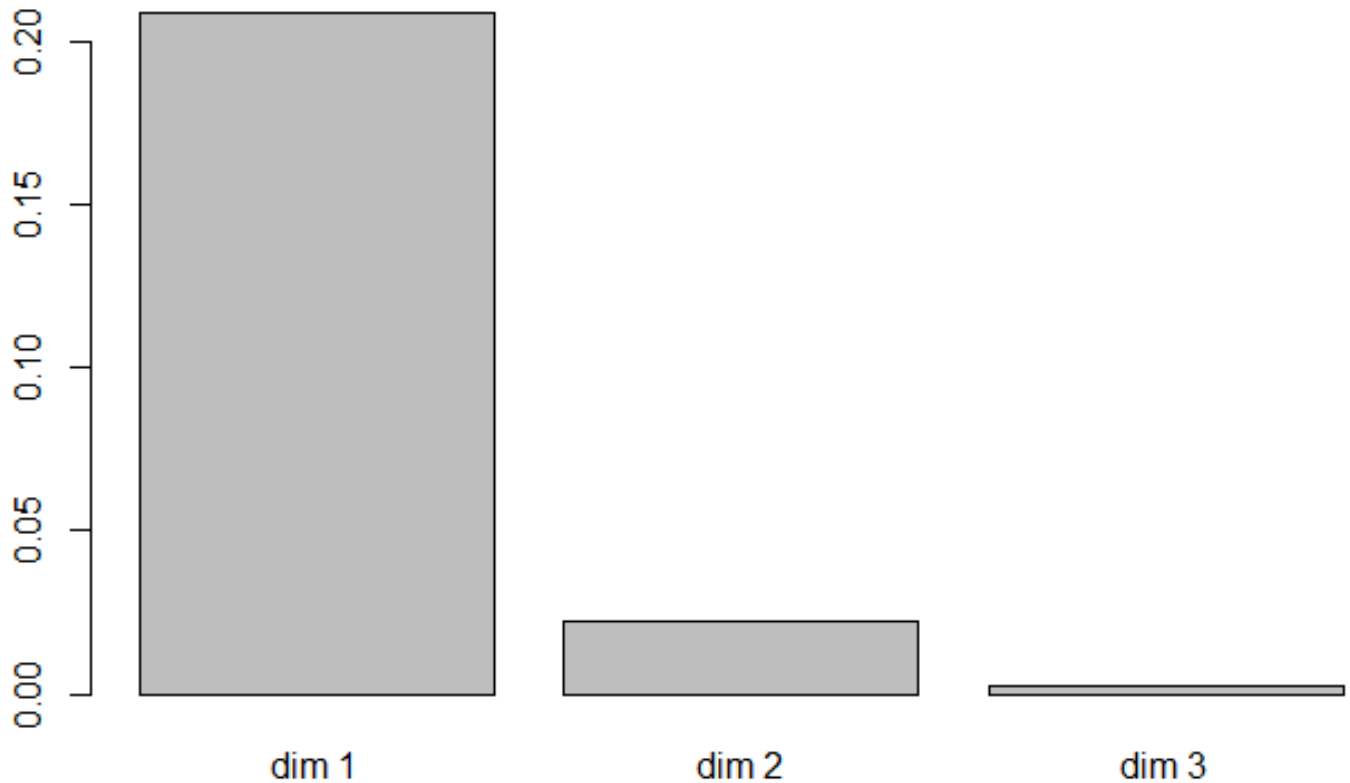
Analyse du tableau Cheveux*Yeux

	C_Blond	C_Marron	C_Noir	C_Roux
Y_bleu	94	84	20	17
Y_Marron	7	119	68	26
Y_Noisette	10	54	15	14
Y_vert	16	29	5	14

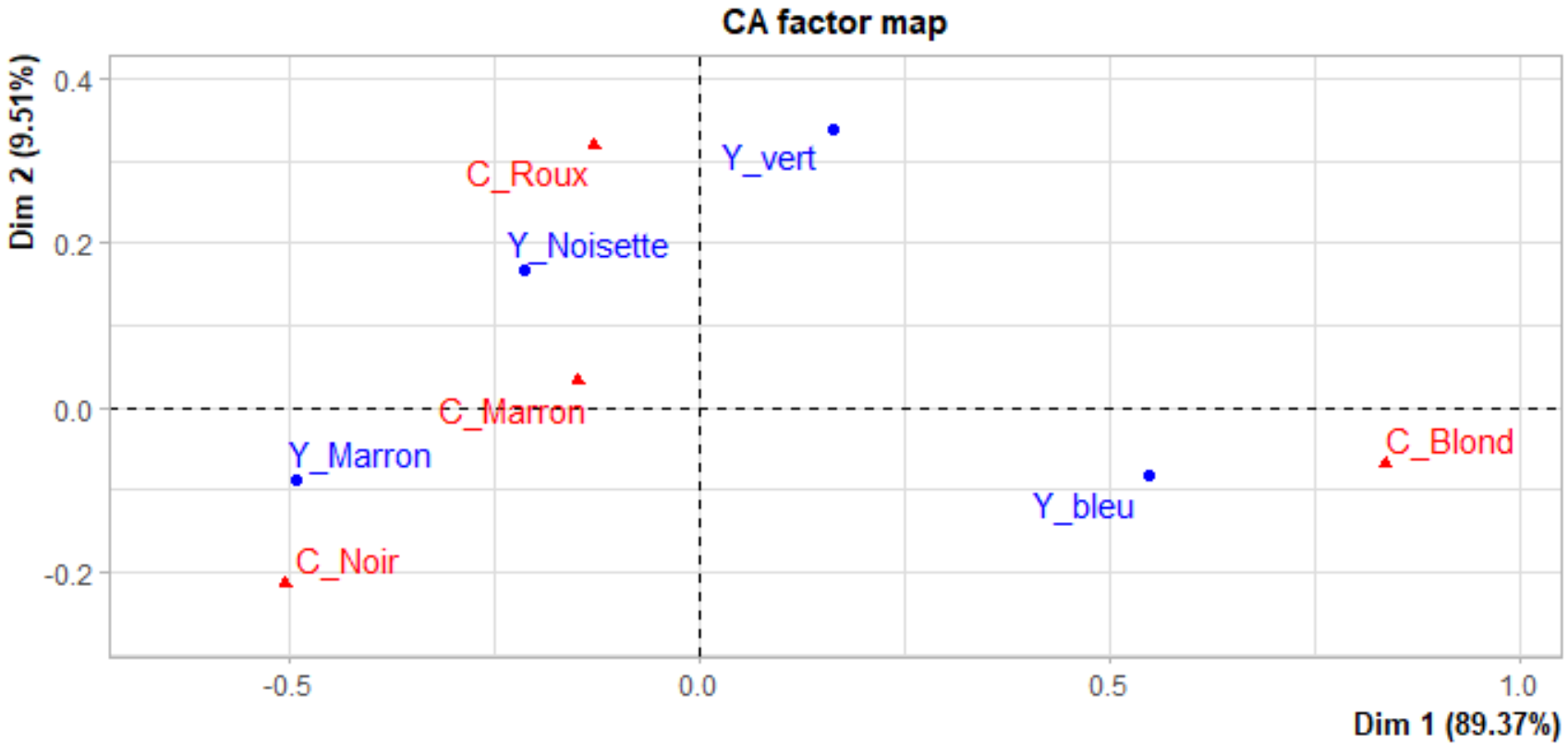
chev.CA¡-CA(cheveux)

Analyse du tableau Cheveux*Yeux

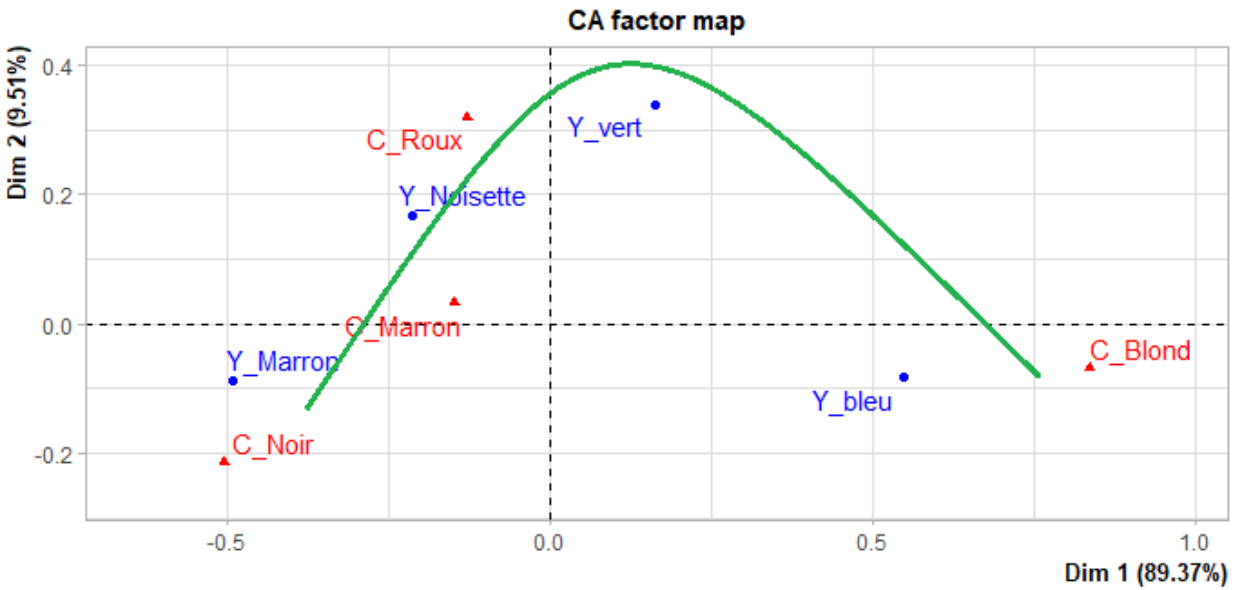
Eboulis des valeurs propres (Screeplot)



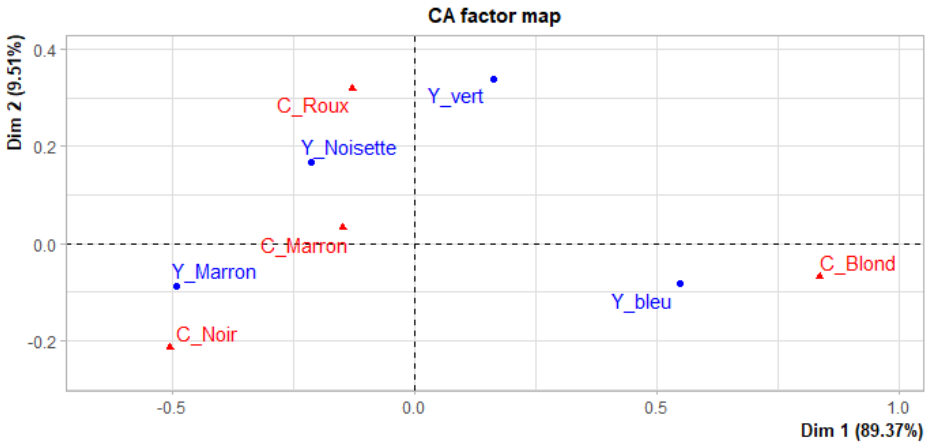
Analyse du tableau Cheveux*Yeux

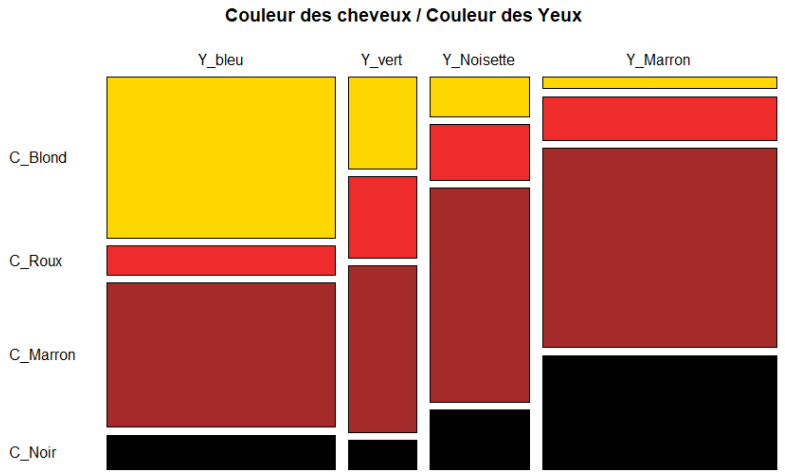


Effet "fer à cheval" (Guttman)



Effet "fer à cheval" (Guttman)





Déclinaison du schéma de dualité

Analyse	Nature des données	Tableau de données	Pondération lignes	Pondérations colonnes
ACP	Quantitatif	$\frac{x_{ij} - x_{.j}}{s_{.j}}$	$\frac{\mathbf{I}}{n}$	$\mathbf{I}$
AFC	Tableau de contingence	$\frac{f_{ij} - f_{i.} f_{.j}}{f_{i.} f_{.j}}$	$f_{i.}$	$f_{.j}$
ACM	Qualitatif (tableau disjonctif)	$\frac{\delta_{ikj} - f_{kj}}{f_{kj}}$	$\frac{\mathbf{I}}{n}$	$diag(f_i)$/nfacteurs