MUSÉUM NATIONAL D'HISTOIRE NATURELLE

SEMIN-R

Analyses factorielles avec R

Elisabeth MORAND
INED

Analyses Factorielles avec R

E. Morand

10 Décembre 2009

INED

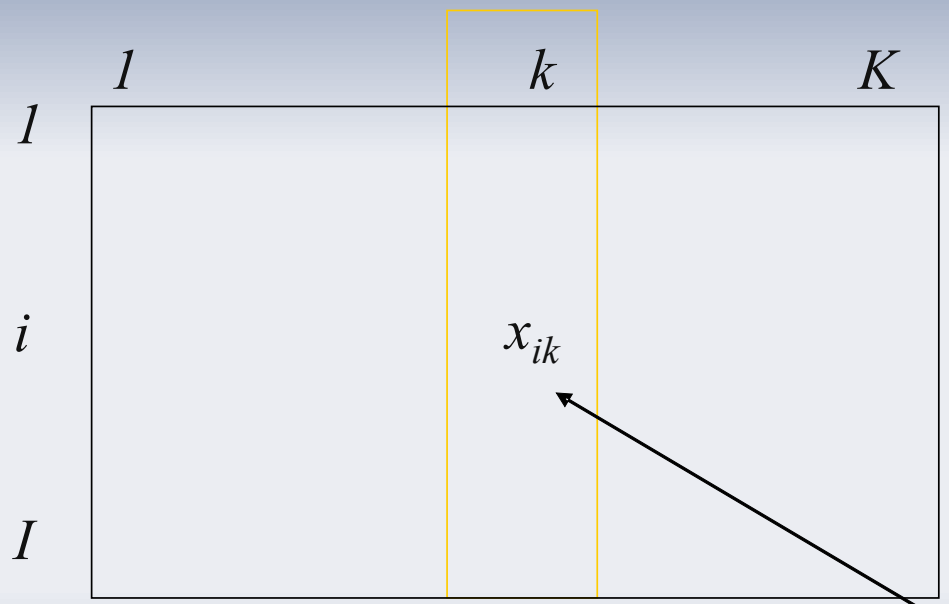
Part I

Analyse en Composantes Principales : ACP

Sommaire

plan

Données



Variables quantitatives

Individus (statistiques)

K le nombre de variables

Valeur de la variable k pour l'individu i

$$\bar{x}_k \qquad s_k \qquad s_k^2 = \frac{1}{I}\sum_{i=1}^{I}\left(x_{ik} - \bar{x}_k\right)^2$$

Un exemple de données

Les températures maximales moyennes mensuelles des villes de France en 2008

villes	latitude	longitude	janv.	février	mars	avril	mai	juin	juil.	août	sept.	oct.	nov.	déc.	amplitude	géographie
angers	47,28	-0,33	10,5	12,2	11,8	14,6	21,5	22,7	24,7	23,5	19,7	16	11	7,5	17,2	ouest
Besançon	47,15	6,02	8,6	11,1	9,6	13,7	21,9	22,3	25	22,7	16,9	13,3	9,3	4	21	Est
Biarritz	43,29	-1,34	13,7	16,3	13,9	17	21	21,3	23,4	24	22,1	18,6	12,7	10,6	13,4	Ouest
Bordeaux	44,5	-0,34	12,1	15,4	13,2	17	22	23,8	26,1	25,8	25,8	18,4	12,4	9	17,1	Ouest

Problématique

- Etude des lignes (individus)
- Etude des colonnes (variables)

Etude des lignes

- Notion de ressemblance
  - 2 jus de fruits sont-ils proches sensoriellement (ou pas) ?
  - 2 personnes sont-elles proches dans leurs habitudes de consommation?
- Proche du point de vue de l'ensemble des variables Mais si on interroge 1000 personnes les regarder 2 à 2 impossible !
- il faut donc une synthèse
- Typologie au sens strict : partition, les classes sont définies et un individus appartient à une et une seule classe
  - Typologie a posteriori
  - Typologie a priori
- Typologie au sens large : le marché se sépare de façon continue

Etude des colonnes

- pas de notion de ressemblance (e.g poids et taille)

Etude des colonnes

- pas de notion de ressemblance (e.g poids et taille)
- on utilise liaison linéaire entre deux variables
    - effet de seuil?

Etude des colonnes

- pas de notion de ressemblance (e.g poids et taille)
- on utilise liaison linéaire entre deux variables
    - effet de seuil?

Lorsque le nombre de variables est important (K grand) le nombre les liaisons deux à deux à étudier devient très important. Il faut donc synthétiser l'information

Etude des colonnes

- pas de notion de ressemblance (e.g poids et taille)
- on utilise liaison linéaire entre deux variables
  - effet de seuil?

Lorsque le nombre de variables est important (K grand) le nombre les liaisons deux à deux à étudier devient très important. Il faut donc synthétiser l'information

- Typologie:

  groupe de variable très corréléés entre elles parce que pouvant être considérées comme différentes mesures d'un même facteur sous-jacent
- Variable synthétique a priori par exemple la moyenne

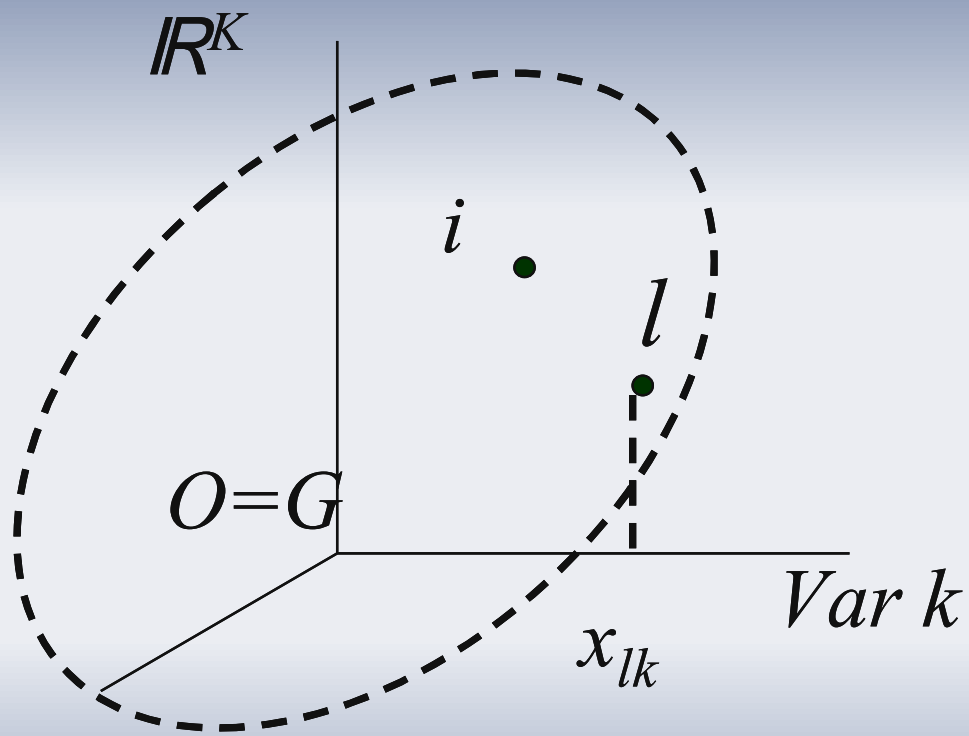
Centrage Réduction

Dans la suite on travaille avec le tableau de données centrées réduites

plan

Nuage des individus

- Un individu = une ligne du tableau soit K valeurs numériques
- un individu = un vecteur dans un espace à K dimensions



Distance entre deux individus :

$$d^2(i,l) = \sum_{k=1}^{K} (x_{ik} - x_{lk})^2$$

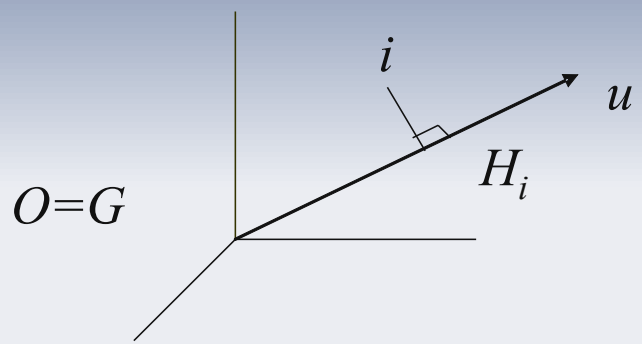
Projection

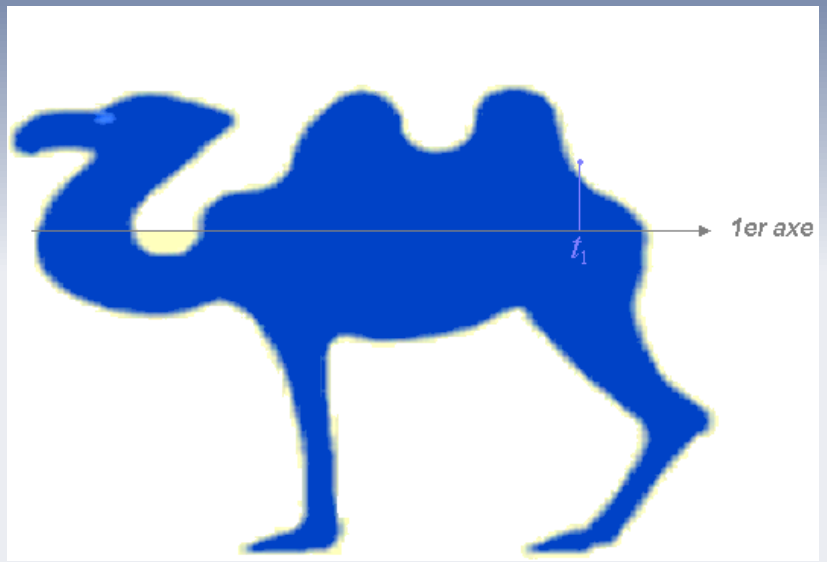
- Le problème : en 2 dimensions on voit bien. Au-delà de 3 dimensions pas accessible à nos sens
- Réduire l'espace de représentation: projection

projection

On projette le nuage des individus (N_I) sur une suite d'axes orthogonaux d'inertie maximum

Construction de la projection





- i est bien représenté si (iH_i) petit et si OH_i est grand pour tous les i
- On maximise

$$\sum_i (OH_i)^2$$

Choix des axes

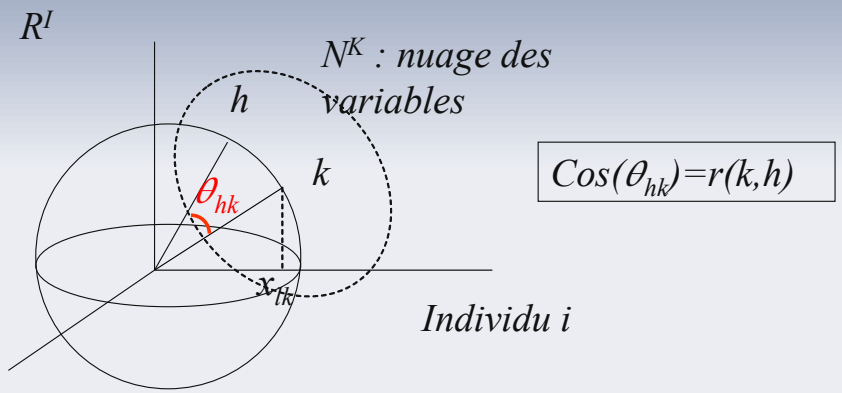
On cherche les principales dimensions de variabilité

plan

Cas des villes de France

- La proximité des villes ici est la proximité climatique
- des villes proches et éloignées mais suivant quels critères?
  - Toutes les températures?
  - Les températures en été?
- Retour aux variables

Nuage des variables



Les données sont centrées réduites donc

$$d^2(O,h) = \sum_{i=1}^{I} x_{ih}^2 = var(h) = 1$$

projection des variables

Projeté le nuage des variables (N^K) sur une suite d'axes d'inerties maximum.
Ceci revient à chercher l'axe qui ressemble le plus (très corrélé) aux variables de départ. On retrouve la notion de variable synthétique.

Dualité

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{k=1}^{K} x_{ik} G_s(k)$$

$$G_s(k) = \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^{I} (1/I) x_{ik} F_s(i)$$

- $F_s(i)$ coordonnée de i sur l'axe de rang s
- $G_s(k)$ coordonnée de la variable k sur l'axe de rang s
- λ_s variance de la composante F_s

plan

Exemple

Un individu (une ligne) est une ville. Chaque colonne est une variable décrivant la ville :

- températures maximales moyennes chaque mois
- longitude
- latitude
- amplitude

On dispose aussi d'une variable qualitative : la région (géographie)

problématique

- Y a-t-il des variables corrélées, i.e. apportant à peu près la même info ?
- Quelles sont les grandes caractéristiques mesurées ?
- Quelles sont les positions relatives des villes selon ces caractéristiques ?

Réalisation de l'Analyse en composantes principales avec FactoMineR

On commence par étudier les températures. Les 12 colonnes correspondants aux 12 moyennes mensuelles sont prises dans l'analyse.

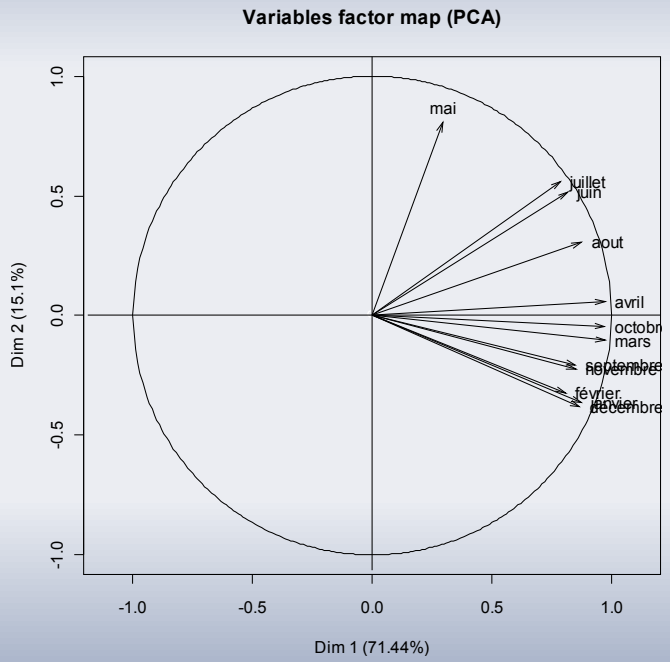
Les valeurs propres

- Valeur propre λ_i = quantité d'information apportée par l'axe i
- Rapport $\lambda_i / \sum \lambda_i = \%$ d'information apportée par l'axe i
- Propriété en ACP normée : $\sum \lambda_i$ = nombre de variables

exemple

	eigen value	percentage of variance	cumulative percentage of variance
comp1	8.57	71.44	71.44
comp2	1.81	15.10	86.54
comp3	0.55	4.55	91.10
comp4	0.47	3.90	95.00
comp5	0.24	2.00	96.99
comp6	0.10	0.84	97.83
comp7	0.08	0.64	98.47
comp8	0.07	0.56	99.03
comp9	0.04	0.37	99.39
comp10	0.04	0.30	99.69
comp11	0.02	0.17	99.86
comp12	0.02	0.14	100.00

Cercle des corrélations



- Représentation graphique des variables en projection sur un plan
- Les projections des variables corrélées sont proches
- Coordonnée sur un axe = corrélation avec le facteur
- Déformation possible due à la projection
- Attention aux variables mal représentées
- Variable proche du cercle = variable bien représentée en projection

Qualité de représentation d'une variable, contribution

Qualité de représentation d'une variable

$$qlt_k(s) = \frac{\text{inertie projetée de } k \text{ sur } v_s}{\text{inertie totale de } k} = cos^2(\theta_k^s)$$

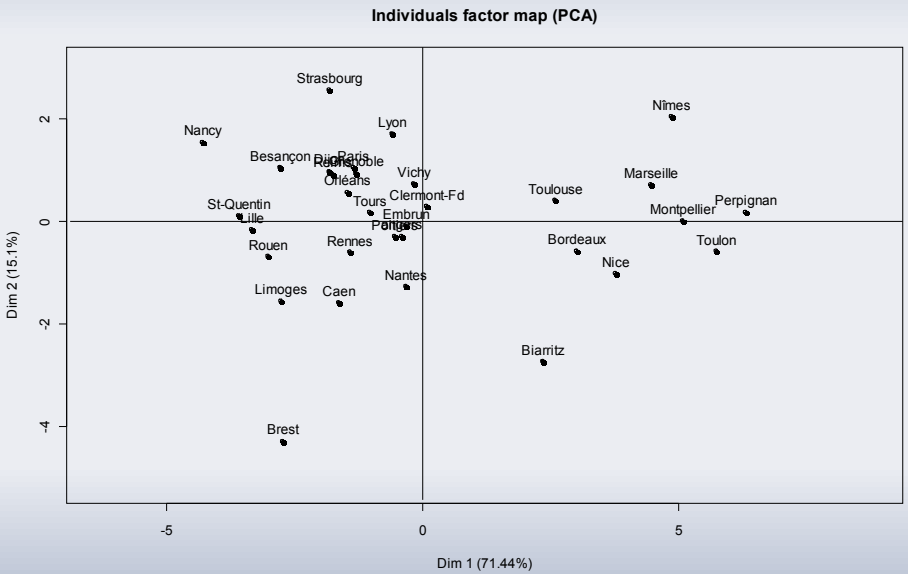
Or $cos^2(\theta_k^s) = r^2(k, v_s)$

contibution d'une variable à la construction d'un axe

Décomposition de l'inertie variable par variable

$$\frac{(OH_k^s)^2}{\lambda_s}$$

représentation des individus

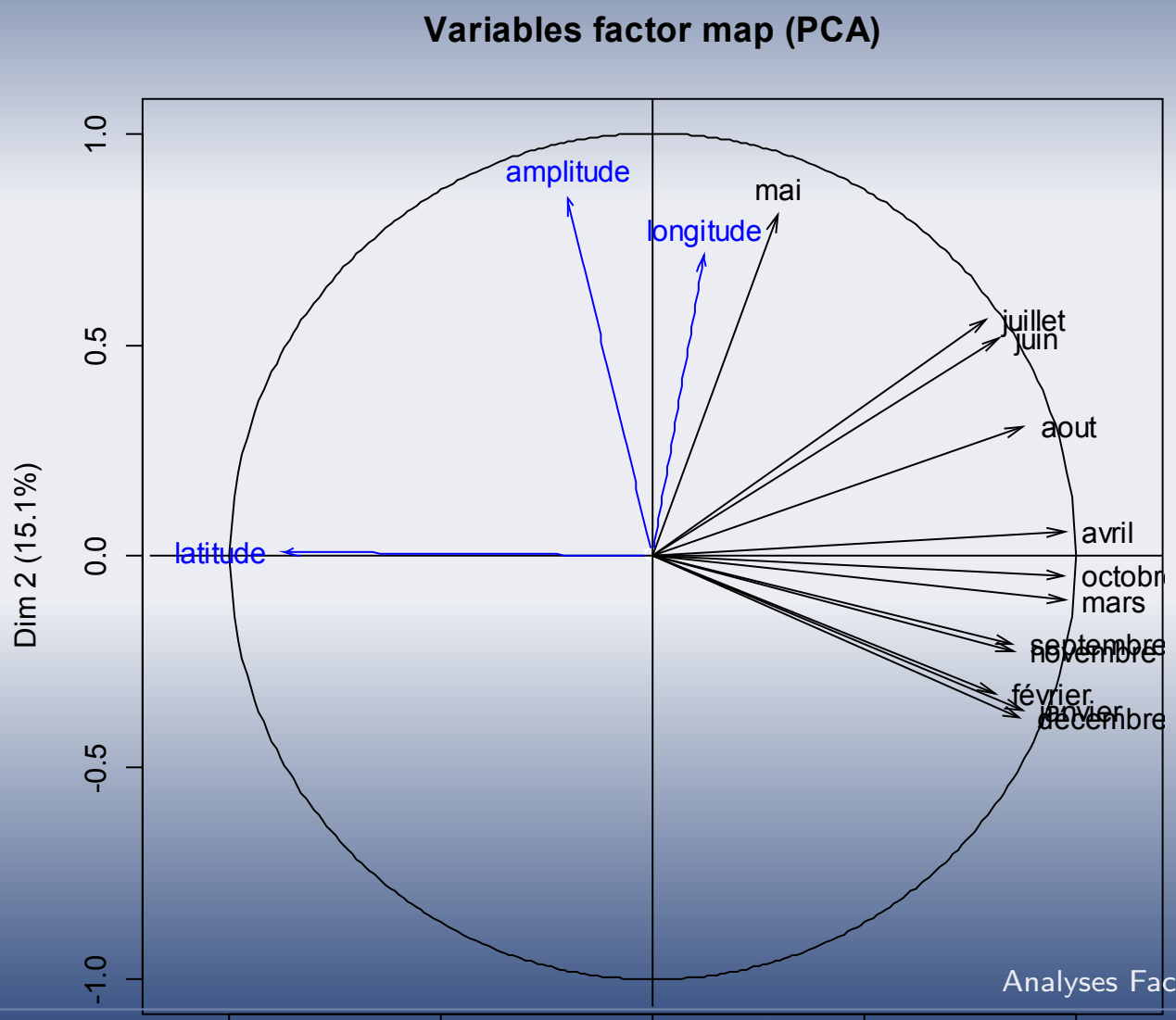


- Représentation graphique des individus en projection sur un plan
- Les projections des individus qui se ressemblent sont proches
- Centre du graphique = centre de gravité (point moyen)
- Coordonnée sur un axe = valeur de la composante principale
- Déformation possible due à la projection
- Attention aux individus mal représentés (cos^2 faible)

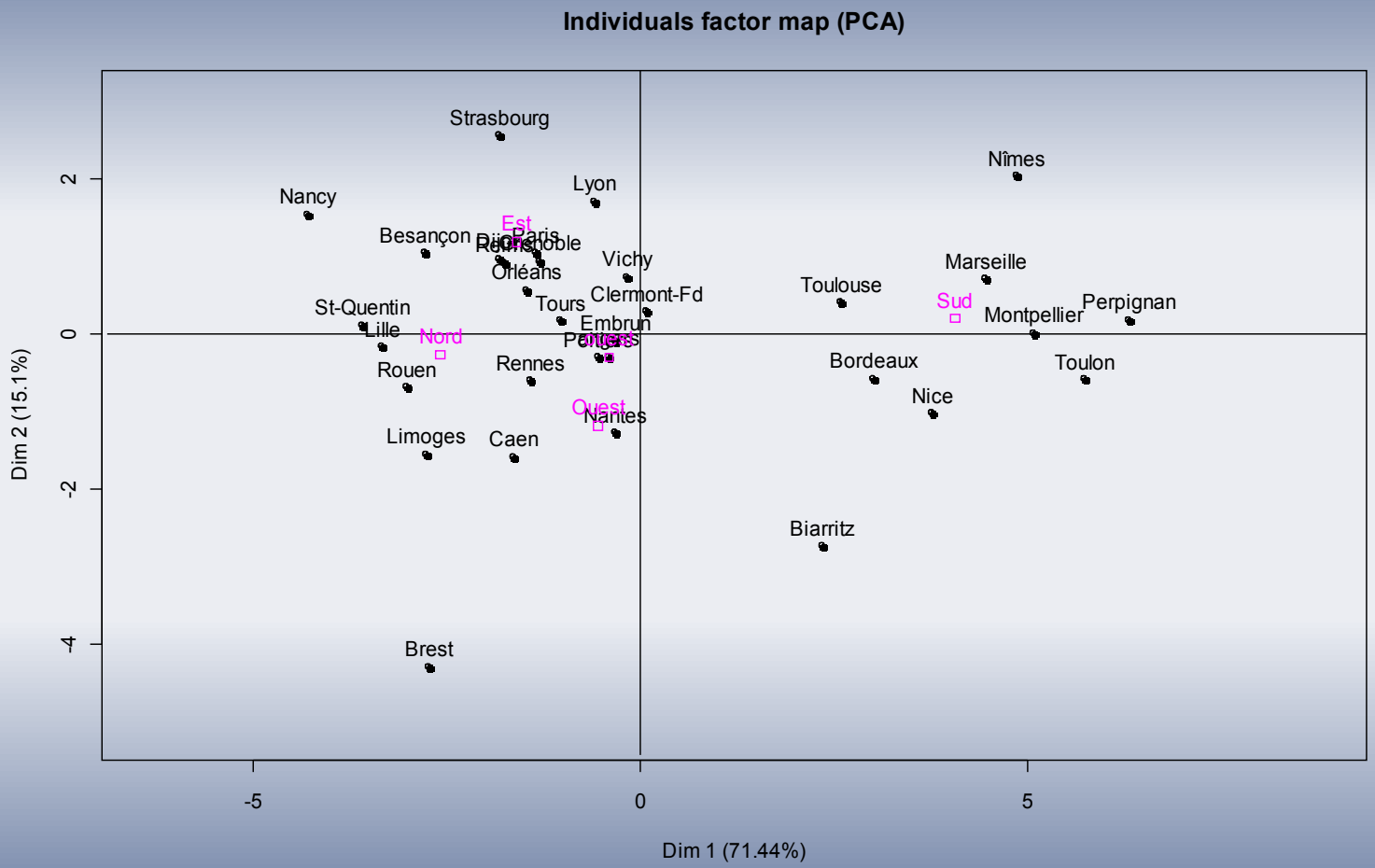
Elements illustratifs

- variables ou individus supplémentaires jouant un rôle particulier
  - Ex : géographie
  - Ex : amplitude, latitude, longitude
- qui n'interviennent pas dans la construction des axes
- représentés en projection sur les plans définis par les éléments actifs
- permettent de compléter les interprétations
- représentés dans la vue adéquate (variables ou individus)

Variables quantitatives supplémentaires



Variables qualitatives supplémentaires



Quelques outils utiles

- dimdesc : description automatique des axes
- dynGraph : pour les graphiques

Conclusion

- réduction de la dimension ; parfois jusqu'à une seule (utilisation en traitement de l'image)
- base orthonormée
- visualisation graphique

Part II

Analyse factorielle des correspondances : AFC

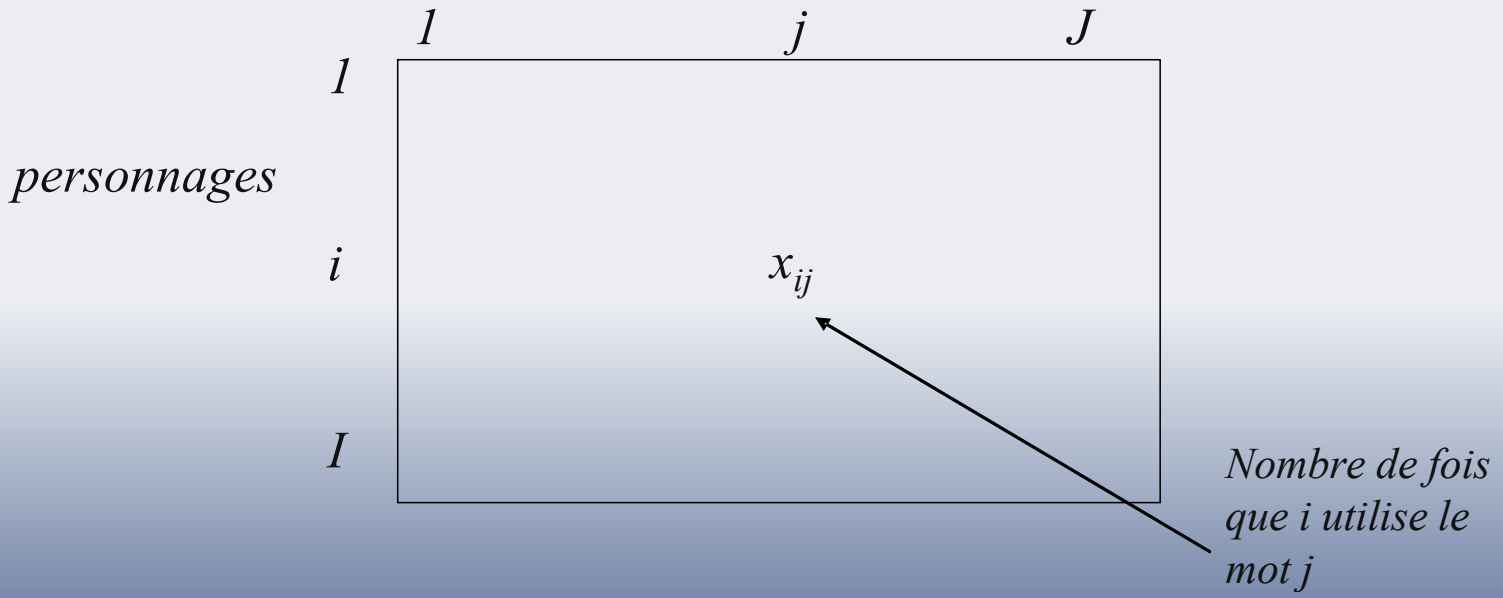
plan

A l'origine...

L'AFC est née à Rennes en 1965.

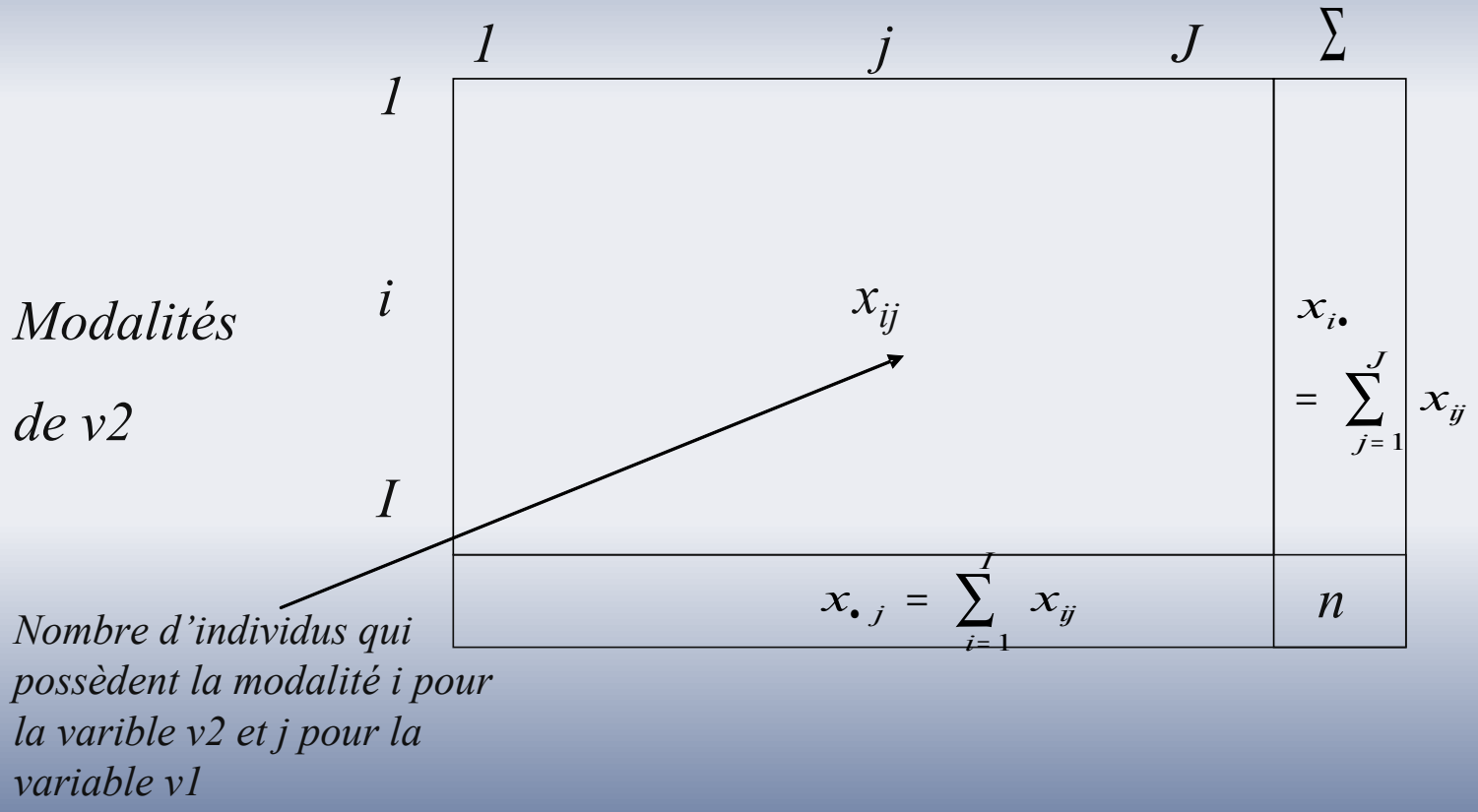
- JP Benzécri
- B. Escofier

Pour étudier les textes de Phèdre



les données

Individus décrits par deux variables qualitatives (ex : couleur des yeux et couleurs des cheveux)



Utilisation

- A l'origine : l'AFC sert à analyser des tableaux de contingence

  Exemple : âge du chef d'exploitation (en classes) et SAU (en classes)

- Technique très proche de celle de l'ACP (la métrique diffère)

- Par extension : l'AFC peut s'adapter à tout tableau de mesures positives

  Exemple : présence-absence (0-1) d'espèces dans des stations de relevés. L'AFC permettra de décrire la représentation des espèces suivant les stations

plan

Exemple

Exemple : enquête 1972, CREDOC
Attitude des femmes à l'égard du travail des femmes; 1724 enquêtées; On sort 2 questions.

La famille idéale est celle où :	Activité de famille convenant le mieux à une mere de famille quand les enfants vont à l'école				
	Reste au foyer	Trav. à mis-temps	Trav à pl. temps	Σ	prob marg..
Les deux conjoints travaillent également	13	142	106	261	.151
Le mari a un métier plus absorbent que celui de sa femme	30	408	117	555	.322
Seul le mari travaille	241	573	94	908	.527
Σ	284	1123	317	1724	1.

Probabilité marginale	.165	.651	.184	1.

Profils

Indépendance : $\frac{f_{ij}}{f_{i.}} = f_{.j}$
tous les profils lignes sont égaux et sont égaux au marginal

	1	j	J	$\sum$
1				
i		$f_{ij} = \frac{x_{ij}}{n}$		$f_{i\bullet}$
I				
		$f_{\bullet j}$		1

Principe de l'AFC

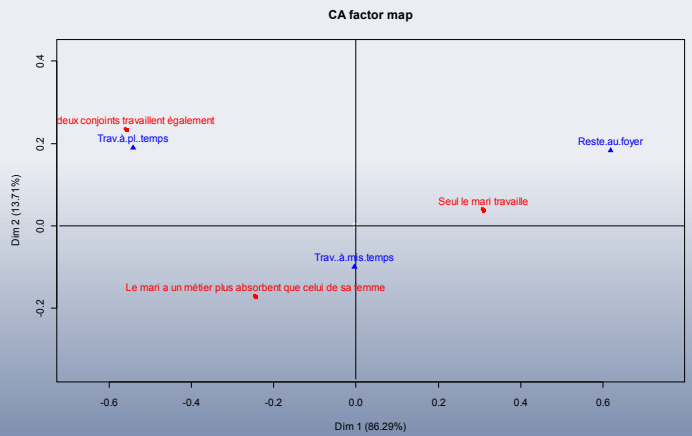
- Dans l'AFC, les lignes et les colonnes jouent le même rôle
- L'AFC est une double ACP utilisant la distance du χ^2 : ACP sur les lignes et ACP sur les colonnes, dont on superpose les graphiques
- On raisonne en fait sur les " profils " des lignes et des colonnes
- Les valeur du "cos^2 " pour un point mesure la qualité de la représentation de ce point en projection sur l'axe considéré
- Les " contributions relatives " permettent de détecter les éléments qui sont les plus explicatifs de l'axe considéré
- Les graphiques permettent de visualiser les proximités entre les points

plan

Représentation graphique

Objectif de l'analyse

Mettre en évidence des attractions et des répulsions entre des éléments d'un même ensemble.



- L'axe 1 est un axe d'attitude vis-à-vis du travail féminin. Il ordonne les modalités de la moins favorable à la plus favorable au travail féminin.
- Le graphique suggère que la modalité "le métier du mari est plus absorbant " (modalité de la question " la famille idéale est celle où ") est déjà plus favorable au travail féminin.

References I

Escofier, B. et Pagès, J. (2008). **Analyses factorielles simples et multiples Objectifs, méthodes et interprétation**. Dunod.

François, H., Lê, S., et Pagès, J. (2009). **Analyse de données avec R**. Presse Universitaire de Rennes.