Reconnaissance des formes et méthodes neuronales (RCP208)

Méthodes d'analyse factorielle

Michel Crucianu
(prenom.nom@cnam.fr)
http ://cedric.cnam.fr/vertigo/Cours/ml/

Département Informatique
Conservatoire National des Arts & Métiers, Paris, France

12 octobre 2017

Plan du cours

Méthodes d'analyse factorielle

- Données : tableau de n observations décrites par d variables

Observation	X_1	X_2	...	X_d
o_1	...	...	...	...
o_2	...	...	...	...
...	...	...	...	...
o_n	...	...	...	...

- Objectif général : recherche de « facteurs » (variables dérivées) permettant de résumer les (caractéristiques des) données
  - Améliorer la « lisibilité » des données
    - Mettre en évidence des relations entre (groupes de) variables
    - Permettre une visualisation informative
  - Réduire le nombre de variables en conservant au mieux l'information utile
    - Réduire la redondance présente dans l'ensemble de variables initiales
    - Choisir un sous-espace de description plus pertinent

Méthodes d'analyse factorielle (2)

- Utilisées essentiellement dans un but d'exploration et de description
- Caractéristique importante : absence d'hypothèses préalables concernant les données
- Méthodes abordées ici (voir aussi [4, 3]) :
  1. Analyse en composantes principales (ACP) : exploratoire, variables numériques
  2. Analyse factorielle des correspondances binaires (AFCB) ou multiples (ACM) : exploratoire, variables nominales
  3. Analyse factorielle discriminante (AFD) : exploratoire (et décisionnelle), variables numériques

Plan du cours

Un exemple : feuilles d'arbres

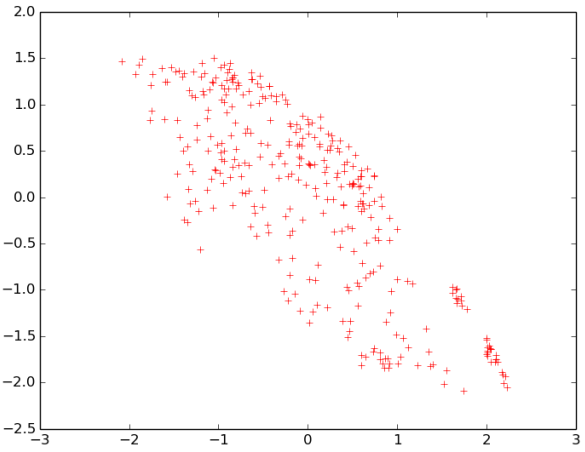
- Considérons un ensemble de feuilles de 40 espèces d'arbres (voir https ://archive.ics.uci.edu/ml/datasets/Leaf)

- Ces observations sont décrites par plusieurs variables (14 dans l'étude mentionnée)

1. Y a-t-il des variables redondantes ?
2. Y a-t-il des caractéristiques dont la variation décrit mieux les différences entre observations ?

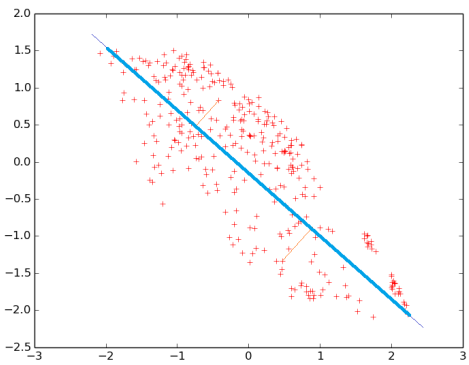
Feuilles d'arbres : lien entre deux variables

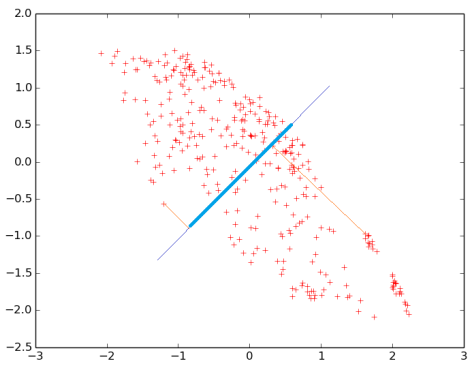
- Considérons deux des variables décrivant les feuilles
  - *Elongation* : distance maximale normalisée entre un point de la feuille et le contour
  - *Isoperimetric factor* : rapport entre l'aire (multipliée par une constante) de la feuille et le carré du périmètre
- Dans quelle mesure pouvons-nous résumer par une variable dérivée l'information qu'apportent ces deux variables initiales ?



Feuilles d'arbres : lien entre deux variables (2)

- Considérons une droite qui tourne autour du centre de gravité du nuage et les projections orthogonales des données sur cette droite
- Y a-t-il une orientation pour laquelle la variance des projections est nettement plus élevée que pour les autres orientations ?

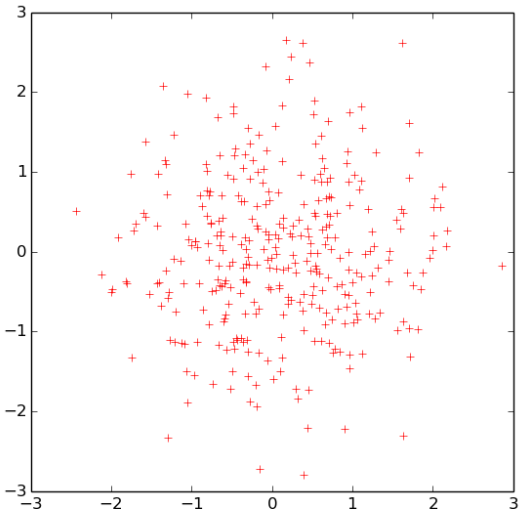




- Ici oui : l'orientation illustrée par la figure de gauche, qui correspondra à la composante principale
- Pour cette même orientation, l'écart entre les données et leurs projections est le plus faible (projections = meilleure approximation 1D du nuage 2D)

Feuilles d'arbres : lien entre deux variables (3)

Est-ce toujours le cas ?

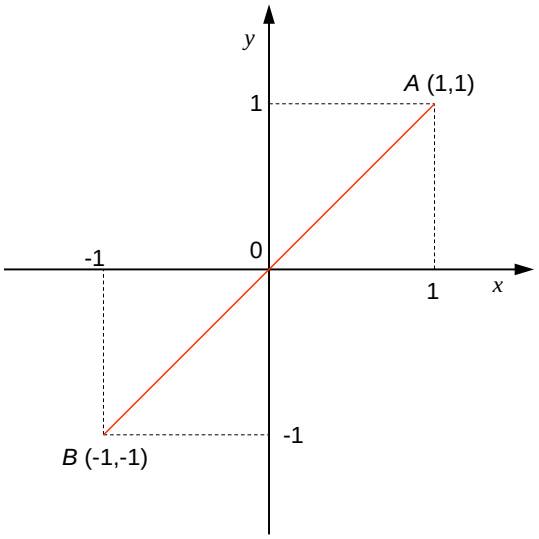


Non !

Comment trouver les composantes principales ?

- Revenons à 2 variables et utilisons 2 observations :

	X	Y
A	1	1
B	-1	-1



Comment trouver les composantes principales (2)

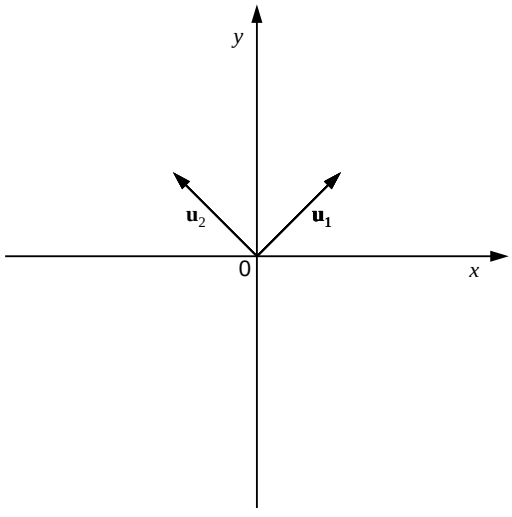
- Dans notre exemple, observations $A = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ et $B = \begin{pmatrix} -1 \\ -1 \end{pmatrix}$

$\Rightarrow$ Centre de gravité $\mathbf{g} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, matrice des corrélations empiriques $\mathbf{S} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$

- La variance des projections des données sur la droite de direction définie par le vecteur unitaire $\mathbf{u}$ est $\mathbf{u}^T\mathbf{S}\mathbf{u}$
- L'axe principal correspond au vecteur $\mathbf{u}$ tel que $\mathbf{u}^T\mathbf{S}\mathbf{u}$ (la variance des projections sur ce vecteur) soit maximale $\Rightarrow$ équation $\mathbf{S}\mathbf{u} = \lambda\mathbf{u}$ (des valeurs et vecteurs propres)
- Pour obtenir la solution il faut résoudre $\det(\mathbf{S} - \lambda\mathbf{I}) = 0$ pour déterminer λ et ensuite $\mathbf{S}\mathbf{u} = \lambda\mathbf{u}$ pour déterminer $\mathbf{u}$

Comment trouver les composantes principales (3)

$\Rightarrow$ Les résultats sont $\lambda_1 = 2$, $\mathbf{u}_1 = \begin{pmatrix} \sqrt{2}/2 \\ \sqrt{2}/2 \end{pmatrix}$ et $\lambda_2 = 0$, $\mathbf{u}_2 = \begin{pmatrix} -\sqrt{2}/2 \\ \sqrt{2}/2 \end{pmatrix}$



- Axe principal donné par $\mathbf{u}_1$, la variance des projections des données sur $\mathbf{u}_1$ est $\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1 = \lambda_1 > \lambda_2$
  - Remarquer que $-\mathbf{u}_1$ est unitaire et satisfait aussi $\mathbf{S}\mathbf{u}_1 = \lambda_1\mathbf{u}_1$

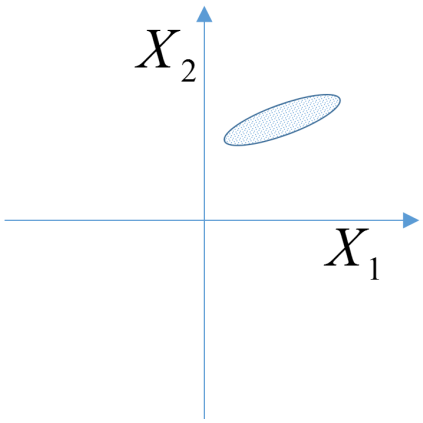
Analyse en composantes principales : données, objectifs

- Introduite par Hotelling (1933) suivant des idées de Pearson (1901)
- Données : n observations caractérisées par d variables quantitatives $\Rightarrow$ matrice de données brutes $\mathbf{R}$
- Objectif général : résumer les variables initiales par un petit nombre k de variables synthétiques (les composantes principales) obtenues comme des combinaisons linéaires des variables initiales
  - Condenser les données en conservant au mieux leur organisation globale
  - Visualiser en faible dimension l'organisation prépondérante des données
  - Interpréter les corrélations ou anti-corrélations entre multiples variables
  - Interpréter les projections des *prototypes de classes* d'observations par rapport aux variables
  - Éventuellement préparer des analyses ultérieures en diminuant la redondance

ACP générale

Appliquée directement sur la matrice de données brutes $\mathbf{R}$

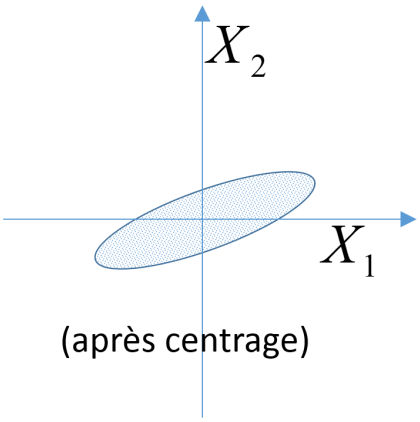
- Interviennent dans l'analyse à la fois la position du nuage d'observations par rapport à l'origine des axes et la forme du nuage
- Utilisation (rare) : analyse tenant compte du zéro naturel de certaines variables



ACP centrée

Appliquée après centrage des variables ($\rightarrow$ moyennes nulles)

- Analyse de la *forme* du nuage d'observations (par rapport à son centre de gravité)
- Utilisation : variables directement comparables

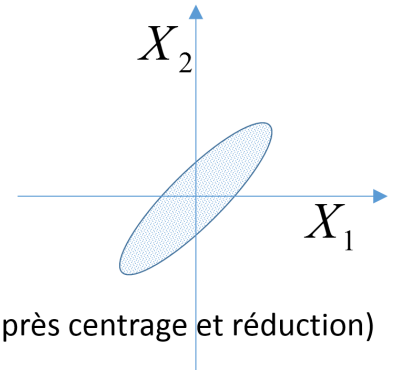


(après centrage)

ACP normée

Appliquée après normalisation des variables (moyennes nulles et écarts-types unitaires)

- Analyse de la forme du nuage après normalisation
- Utilisation : variables non directement comparables (unités de mesure différentes, intervalles de variation très différents)



(après centrage et réduction)

ACP : solution

- Lignes de $\mathbf{R}$ décrivent les observations dans l'espace des variables initiales, colonnes de $\mathbf{R}$ décrivent les variables dans l'espace des observations $\Rightarrow$ deux analyses possibles : du nuage des observations et du nuage des variables
- Analyse du nuage des observations : le sous-espace de dimension k recherché est généré par les k vecteurs propres $\mathbf{u}_\alpha$ associés aux k plus grandes valeurs propres λ_α de la matrice $\mathbf{X}^T\mathbf{X}$: $\mathbf{X}^T\mathbf{X}\mathbf{u}_\alpha = \lambda_\alpha \mathbf{u}_\alpha$, $\alpha \in \{1, \ldots, k\}$
- La matrice $\mathbf{X}$ est :
  - ACP générale : la matrice des données brutes $\mathbf{X} = \mathbf{R}$
  - ACP centrée : la matrice des données centrées (de chaque variable on retire la moyenne empirique) $\Rightarrow$ $\mathbf{X}^T\mathbf{X}$ est la matrice des *covariances* empiriques
  - ACP normée : la matrice des données normées (on divise chaque variable centrée par son écart-type) $\Rightarrow$ $\mathbf{X}^T\mathbf{X}$ est la matrice des *corrélations* empiriques
- Projection de l'observation $\mathbf{o}_i$ sur l'axe factoriel α : $\psi_{\alpha i} = \sum_j x_{ij} u_{\alpha j}$

ACP : solution (2)

- La matrice $\mathbf{X}^T\mathbf{X}$ est symétrique et (semi-)définie positive
  - $\Rightarrow$ toutes ses valeurs propres sont réelles et positives (certaines nulles si $\mathbf{X}^T\mathbf{X}$ semi-définie)
  - $\Rightarrow$ les vecteurs propres associés à des valeurs propres différentes sont orthogonaux $\Rightarrow$ axes principaux orthogonaux
- Analyse du nuage des variables : le sous-espace de dimension k recherché est généré par les k vecteurs propres $\mathbf{v}_\alpha$ associés aux k plus grandes valeurs propres μ_α de la matrice $\mathbf{X}\mathbf{X}^T$: $\mathbf{X}\mathbf{X}^T\mathbf{v}_\alpha = \mu_\alpha \mathbf{v}_\alpha$, $\alpha \in \{1, \ldots, k\}$
  - Projection de la variable $\mathbf{X}_j$ sur l'axe factoriel α : $\phi_{\alpha j} = \sum_i x_{ij} v_{\alpha i}$
  - Les valeurs propres non nulles de $\mathbf{X}\mathbf{X}^T$ sont les mêmes que celles de $\mathbf{X}^T\mathbf{X}$ et il y a des relations de transition permettant de passer des résultats d'une des analyses aux résultats de l'autre
- En général $n \gg d$ (nombre d'observations $\gg$ nombre de variables initiales) $\Rightarrow$ traiter la matrice $\mathbf{X}^T\mathbf{X}$ de dimension $d \times d$ plutôt que $\mathbf{X}\mathbf{X}^T$ de dimension $n \times n$

ACP : relations entre les deux analyses

- Les deux analyses s'intéressent aux lignes et respectivement colonnes d'une même matrice de données $\rightarrow$ quels liens entre ces analyses ?
- En (pré-)multipliant $\mathbf{X}^T\mathbf{X}\mathbf{u}_\alpha = \lambda_\alpha \mathbf{u}_\alpha$ par $\mathbf{X}$ on obtient $\mathbf{X}\mathbf{X}^T(\mathbf{X}\mathbf{u}_\alpha) = \lambda_\alpha(\mathbf{X}\mathbf{u}_\alpha)$
  $\Rightarrow$ à tout vecteur propre $\mathbf{u}_\alpha$ de $\mathbf{X}^T\mathbf{X}$ associé à une valeur propre λ_α correspond un vecteur propre $\mathbf{X}\mathbf{u}_\alpha$ de $\mathbf{X}\mathbf{X}^T$ associé à la même valeur propre
- En (pré-)multipliant $\mathbf{X}\mathbf{X}^T\mathbf{v}_\alpha = \mu_\alpha \mathbf{v}_\alpha$ par $\mathbf{X}^T$ on obtient la réciproque

$\Rightarrow$ Les valeurs propres non nulles de $\mathbf{X}\mathbf{X}^T$ et $\mathbf{X}^T\mathbf{X}$ sont toutes égales, $\lambda_\alpha = \mu_\alpha$

- En résultent les relations de transition :

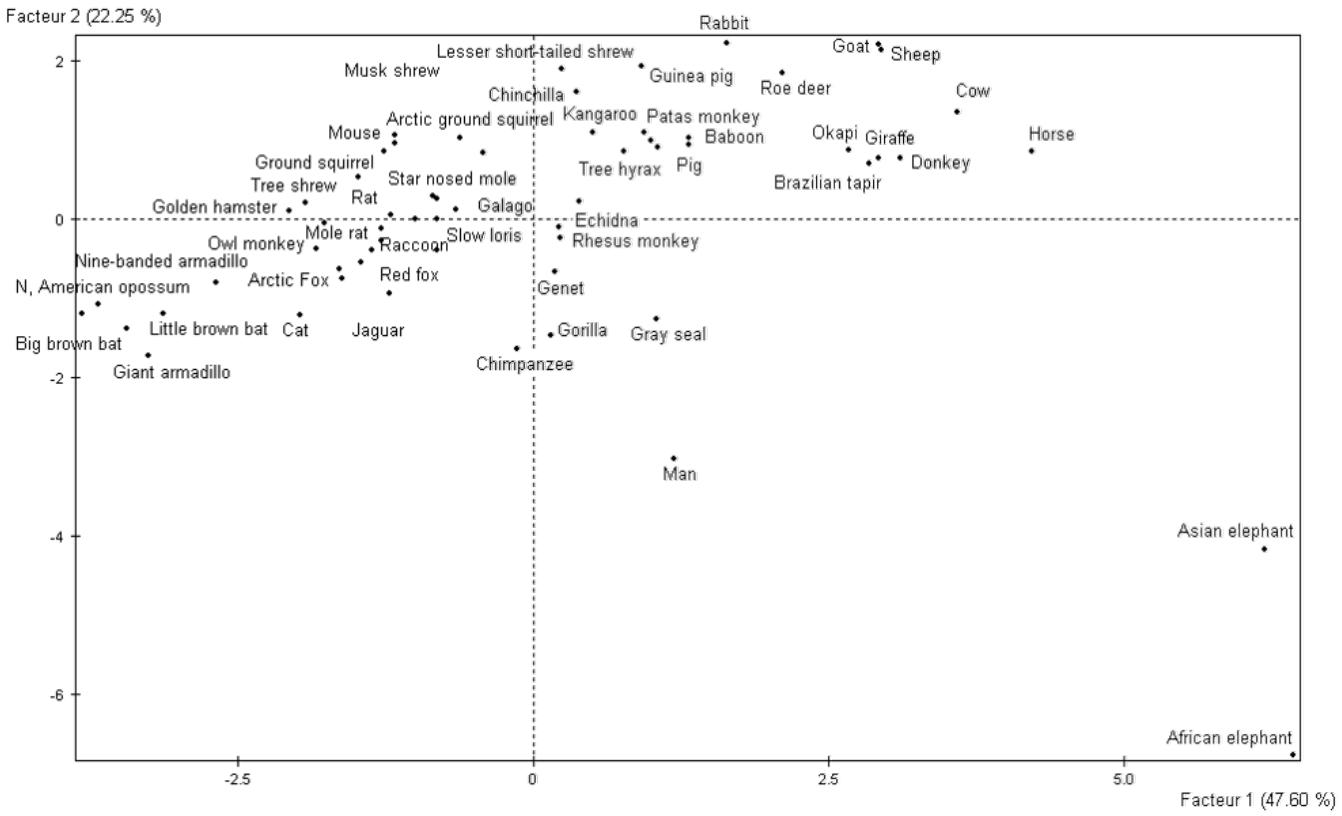
$$\psi_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^{d} x_{ij}\phi_{\alpha j}, \quad \phi_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^{n} x_{ij}\psi_{\alpha i}$$

Exemple : sommeil des mammifères

- Issu de [1] : représentants typiques de 62 espèces de mammifères, décrits par 10 variables

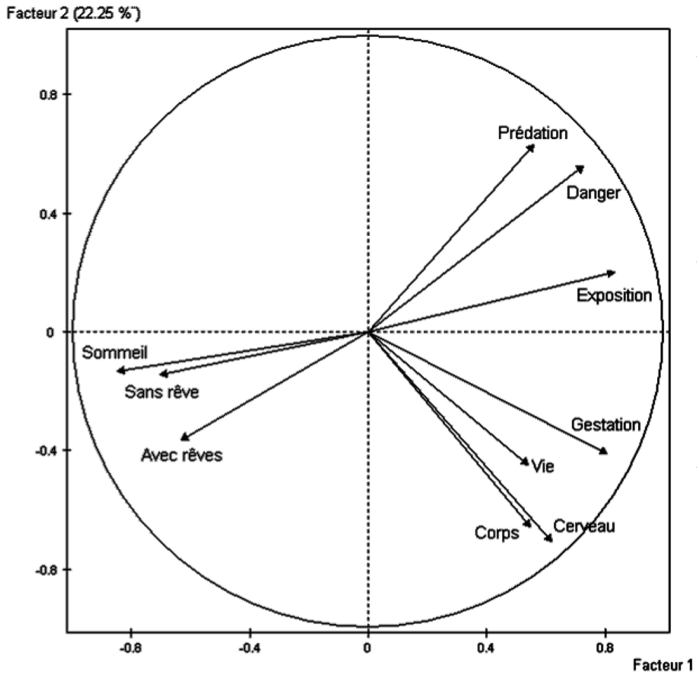
Variable	Moyenne	Écart-type
Corps	198,79	891,87
Cerveau	283,13	922,75
Sans rêve	8,67	3,63
Avec rêves	1,97	1,43
Sommeil	10,53	4,57
Vie	19,88	18,05
Gestation	142,35	145,53
Prédation	2,87	1,46
Exposition	2,42	1,59
Danger	2,61	1,43

ACP sommeil des mammifères : nuage des observations



ACP sommeil des mammifères : nuage des variables

- Groupes : « sommeil », « danger », « corps, cerveau, vie, gestation » (CCVG)
- Forte opposition entre le groupe « sommeil » et le groupe « danger »
- Opposition plus faible entre le groupe « danger » et le groupe CCVG



Règles d'interprétation

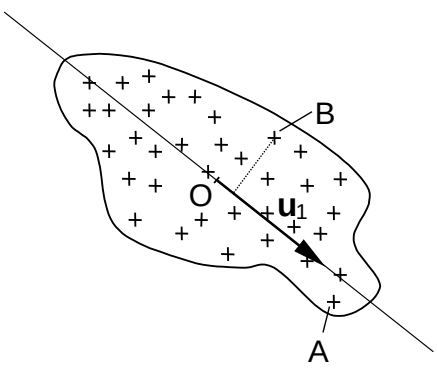
- Objectifs :
  - Révéler quelles variables sont corrélées et quelles variables s'opposent
  - Mettre en évidence des regroupements d'observations et les caractériser à partir des variables
  - Si les observations (ou les « individus ») ne sont pas anonymes, identifier des relations entre elles
- Associer une interprétation aux axes à partir des corrélations et des oppositions entre variables initiales
- Le regroupement de projections des observations sur les plans factoriels s'interprète en termes de similitudes de comportement par rapport aux variables
- Éviter d'interpréter les projections sur un axe des observations mal représentées par l'axe (contributions relatives faibles)
- Retirer, pour une nouvelle analyse, les observations ayant des contributions absolues excessives à l'orientation de certains axes

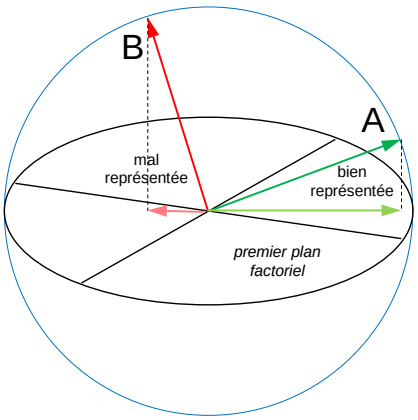
Aides à l'interprétation

- Apport d'un axe factoriel
  - Contribution relative (ou « cosinus carré » ou « qualité de représentation ») d'un axe dans l'explication de l'inertie d'une observation (ou d'une variable)

$$Cr_\alpha(i) = \cos^2(\mathbf{o}_i, \mathbf{u}_\alpha) = \frac{\psi_{\alpha i}^2}{\|\mathbf{o}_i\|^2} \text{ avec } \sum_\alpha Cr_\alpha(i) = 1, \forall i$$

  - Dans l'exemple ci-dessous, l'axe factoriel (respectivement le plan) représenté explique la majeure partie de l'inertie de A mais une faible part de l'inertie de B



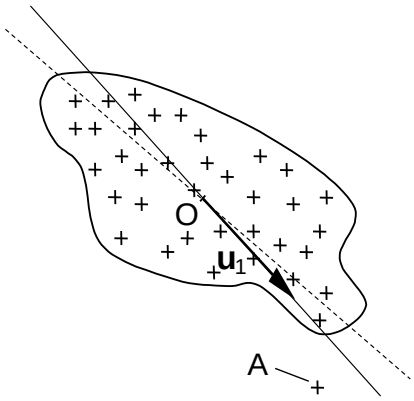


Aides à l'interprétation (2)

- Influence d'une observation (ou d'une variable)
  - Contribution absolue (ou simplement contribution) d'une observation (ou d'une variable) à la variance expliquée par un axe factoriel

$$Ca_\alpha(i) = \frac{\psi_{\alpha i}^2}{\lambda_\alpha} \text{ avec } \sum_i^n Ca_\alpha(i) = 1, \forall \alpha$$

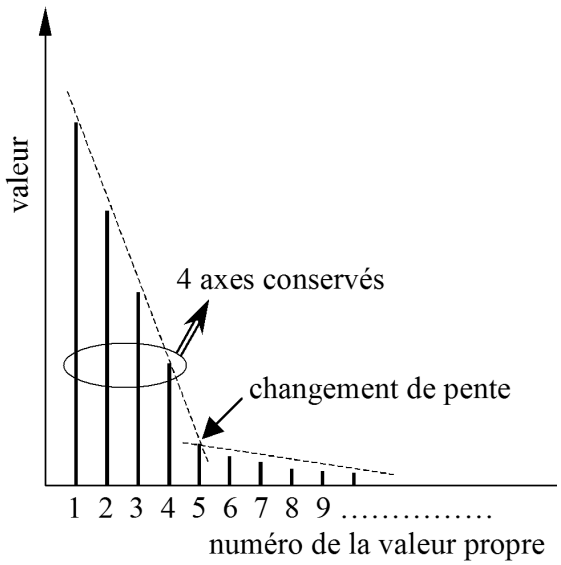
  - Une observation très excentrée aura un impact très fort sur l'orientation des axes factoriels



ACP : choix du nombre d'axes

Suivant l'objectif de l'analyse :

1. Analyse descriptive avec visualisation : à partir de quel ordre les différences entre les pourcentages d'inertie expliquée ne sont plus significatives ?
   - Empiriquement : changement de pente dans la décroissance des valeurs propres
   - Hypothèse de normalité : test statistique d'égalité entre valeurs propres successives



ACP : choix du nombre d'axes (2)

2. Compression des données : qualité d'approximation mesurée par le taux d'inertie expliquée
3. Prétraitement avant méthodes décisionnelles :
   - Simple critère de conditionnement de la matrice des covariances empiriques (ou corrélations si ACP normée)
   - Introduction du nombre d'axes comme paramètre du modèle décisionnel et emploi d'une méthode de sélection de modèle

ACP : observations et variables supplémentaires (1)

De nouvelles observations et/ou variables peuvent être projetées sur les axes factoriels afin de contribuer à l'interprétation

- Observations supplémentaires (ou illustratives) :
  1. Observations faites dans des conditions différentes, nouvelles observations
  2. Observations atypiques, éliminées de l'analyse en raison de leurs contributions absolues excessives
  3. Centre de gravité d'un groupe d'observations, par ex. celles qui possèdent une modalité d'une variable nominale

  → Projection directe sur les axes factoriels (après opérations de centrage et réduction si ACP normée)
  - Utile d'examiner la qualité de représentation !

ACP : observations et variables supplémentaires (2)

- Variables supplémentaires (ou illustratives) :
  - Quantitatives : projection directe sur les axes de l'analyse du nuage des variables (après centrage et réduction si ACP normée)
  - Nominales :
    - Pour chaque modalité, calcul du centres de gravité des observations qui la possèdent et projection sur les axes de l'analyse du nuage des observations
    - Interprétation possible si $v_{\alpha k} > 1,96$

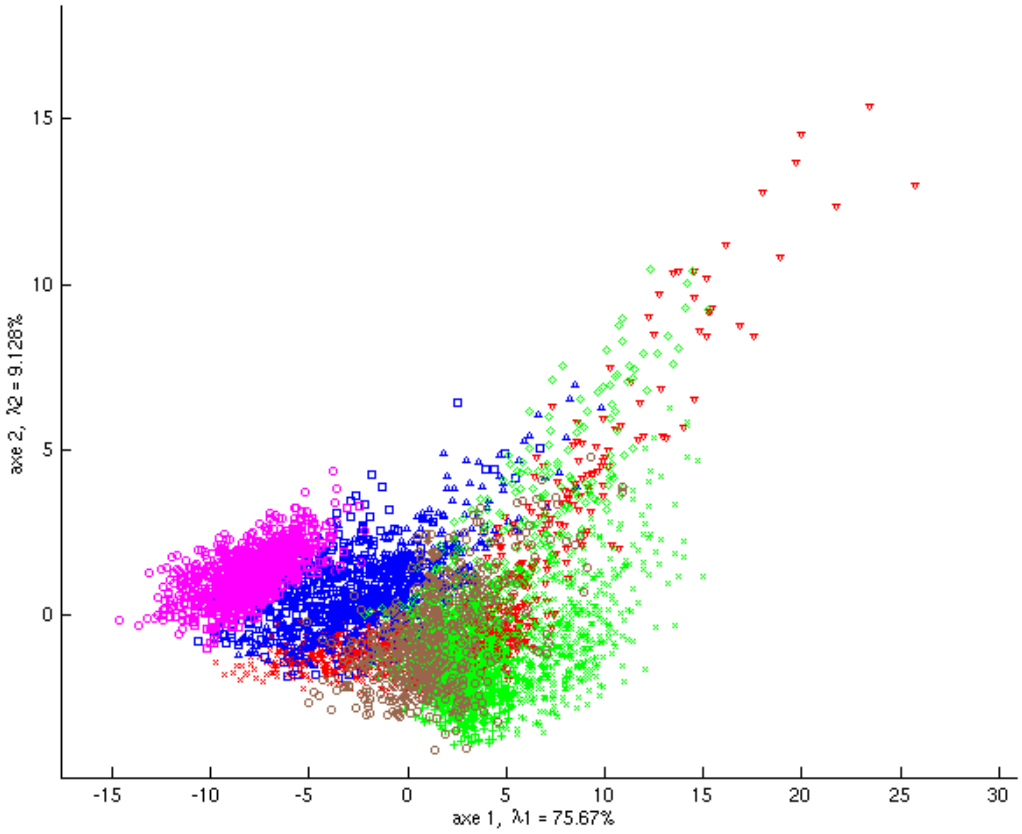
$$v_{\alpha k} = \left(\frac{1}{n_k} \sum_{i \in k} \psi_{\alpha i} \right) \frac{1}{\sqrt{\frac{n-n_k}{n-1} \frac{\lambda_\alpha}{n_k}}}$$

n_k étant le nombre d'observations qui possèdent la modalité k

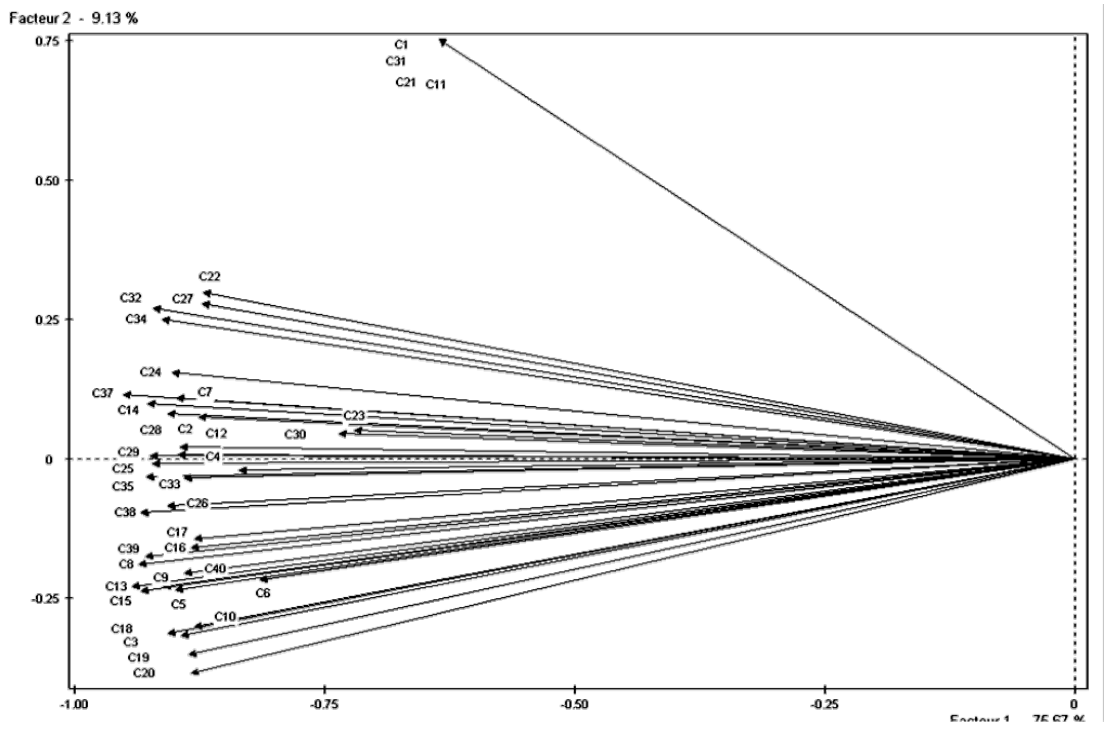
Exemple : textures

- Données mises à disposition par le LTIRF de l'INPG dans le cadre du projet ESPRIT III ELENA (No. 6891) et du groupe de travail ESPRIT ATHOS (No. 6620), voir https ://www.elen.ucl.ac.be/neural-nets/Research/Projects/ELENA/databases/REAL/texture/
- 11 classes de (micro-)textures, 500 pixels (observations « anonymes ») issus de chaque classe $\rightarrow n = 5500$
- Chaque pixel décrit par 40 variables (moments statistiques modifiés d'ordre 4, déterminés pour 4 orientations différentes $0°, 45°, 90°, 135°$, tenant compte des relations avec les voisins d'ordres 1 et 2) $\rightarrow d = 40$
- L'information de classe est ignorée par l'ACP, mais représentée dans les illustrations suivantes par la couleur et la forme du point qui correspond à chaque observation (pixel)

ACP « textures » : nuage des observations



ACP « textures » : nuage des variables



- Fort « effet taille » : corrélations fortes entre toutes les variables initiales $\rightarrow$ présence d'une variable cachée

Plan du cours

Un exemple : notes élevées ou faibles à deux matières

- Notes des 40 élèves d'une classe à deux matières : mathématiques et littérature
- Notes séparées en deux sous-ensembles : « élevées » et « faibles »

→ Tableau de contingences :

	litt. élevées	litt. faibles	somme
math. élevées	18	6	24
math. faibles	6	2	8
somme	24	8	32

- Y a-t-il un lien entre les deux variables nominales ?
  - La proportion d'élèves avec une note faible en littérature est-elle plus élevée pour les élèves avec une note élevée en mathématiques que dans la classe entière ?

Un exemple : notes élevées ou faibles à deux matières

- Notes des 40 élèves d'une classe à deux matières : mathématiques et littérature
- Notes séparées en deux sous-ensembles : « élevées » et « faibles »

→ Tableau de contingences :

	litt. élevées	litt. faibles	somme
math. élevées	18	6	24
math. faibles	6	2	8
somme	24	8	32

- Y a-t-il un lien entre les deux variables nominales ?
  - La proportion d'élèves avec une note faible en littérature est-elle plus élevée pour les élèves avec une note élevée en mathématiques que dans la classe entière ?

  → Non : $\frac{6}{24} = \frac{8}{32}$

Analyse des correspondances 31 / 56

Un exemple : notes élevées ou faibles à deux matières

- Notes des 40 élèves d'une classe à deux matières : mathématiques et littérature
- Notes séparées en deux sous-ensembles : « élevées » et « faibles »

→ Tableau de contingences :

	litt. élevées	litt. faibles	somme
math. élevées	18	6	24
math. faibles	6	2	8
somme	24	8	32

- Y a-t-il un lien entre les deux variables nominales ?
  - La proportion d'élèves avec une note faible en littérature est-elle plus élevée pour les élèves avec une note élevée en mathématiques que dans la classe entière ?

  → Non : $\frac{6}{24} = \frac{8}{32}$

- Et pour la classe représentée par le tableau suivant ?

	litt. élevées	litt. faibles	somme
math. élevées	14	10	24
math. faibles	6	2	8
somme	20	12	32

Un exemple : notes élevées ou faibles à deux matières

- Notes des 40 élèves d'une classe à deux matières : mathématiques et littérature
- Notes séparées en deux sous-ensembles : « élevées » et « faibles »

→ Tableau de contingences :

	litt. élevées	litt. faibles	somme
math. élevées	18	6	24
math. faibles	6	2	8
somme	24	8	32

- Y a-t-il un lien entre les deux variables nominales ?
  - La proportion d'élèves avec une note faible en littérature est-elle plus élevée pour les élèves avec une note élevée en mathématiques que dans la classe entière ?

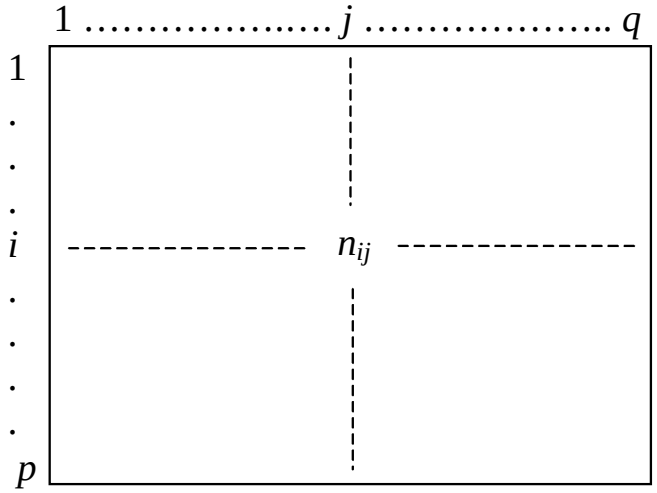
  → Non : $\frac{6}{24} = \frac{8}{32}$
- Et pour la classe représentée par le tableau suivant ? Oui, $\frac{10}{24} > \frac{12}{32}$

	litt. élevées	litt. faibles	somme
math. élevées	14	10	24
math. faibles	6	2	8
somme	20	12	32

Analyse factorielle des correspondances

→ Facile de caractériser le lien entre 2 variables nominales à 2 modalités

- Considérons 2 variables nominales, avec p et respectivement q modalités :



- Comment résumer les liens entre les modalités des deux variables ?

→ ACP des profils-lignes, ACP des profils-colonnes et représentation simultanée

  - Avec des adaptations exigées par la nature de ces données (profils)

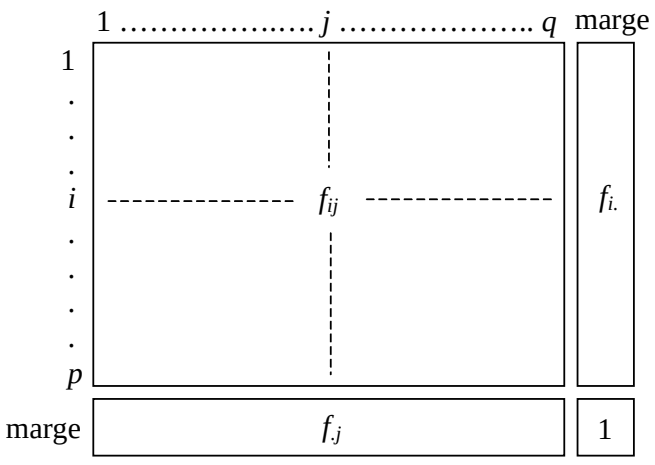
Exemple : utilisation d'Internet

- Données issues de l'enquête « Internet : accès et utilisation au Québec », Réseau Interordinateurs Scientifique Québecois, 1997
- Ici, 2 variables : ancienneté dans l'utilisation d'Internet et nombre moyen d'heures de connexion par mois
- Variables discrétisées par intervalles $\rightarrow$ 1 intervalle = 1 modalité

	<3 mois	3-6 mois	6-12 mois	1-2 ans	2-3 ans	>3 ans	total
<2 h	71	76	160	208	65	23	603
2-5 h	197	204	542	798	359	148	2248
5-10 h	251	234	491	762	427	229	2394
10-20 h	172	164	421	565	399	227	1948
>20 h	79	98	239	397	272	241	1326
total	770	776	1853	2730	1522	868	8519

Analyse factorielle des correspondances binaires (AFCB)

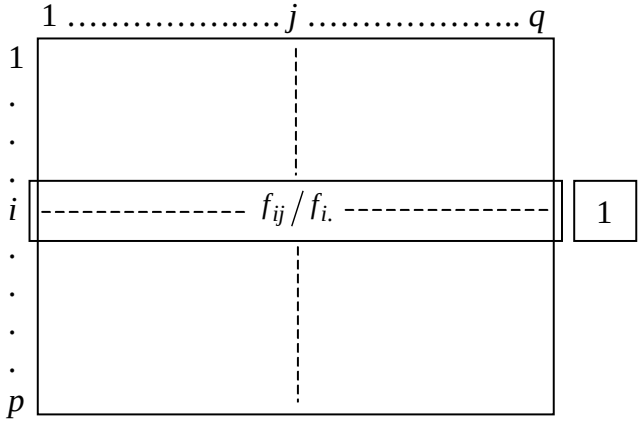
- Objectif : mettre en évidence les relations dominantes entre ces modalités ([2])
- Tableau de contingences $\rightarrow$ tableau fréquences relatives :
  $f_{ij} = \frac{n_{ij}}{n}$, $f_{i.} = \sum_j f_{ij}$, $f_{.j} = \sum_i f_{ij}$

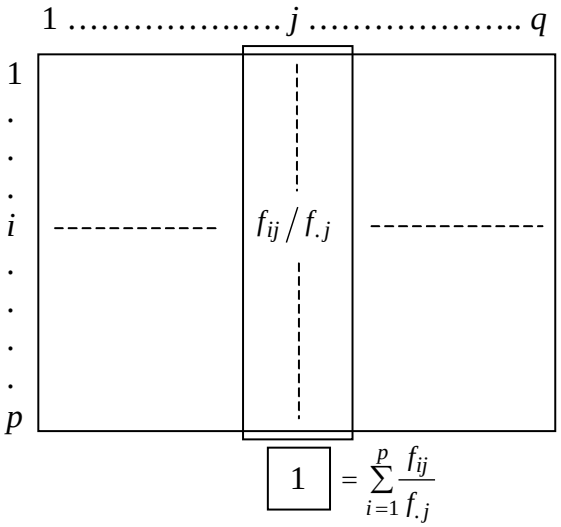


- Analyse directe de ce tableau ? Les fréquences marginales sont trop différentes...

AFCB : profils-lignes et profils-colonnes

Profil de la ligne i : $\left[\frac{f_{i1}}{f_{i.}} \ldots \frac{f_{ij}}{f_{i.}} \ldots \frac{f_{iq}}{f_{i.}}\right]$ Profil de la colonne j : $\left[\frac{f_{1j}}{f_{.j}} \ldots \frac{f_{ij}}{f_{.j}} \ldots \frac{f_{pj}}{f_{.j}}\right]$





AFCB : analyse des profils

- Pondération des profils : l'importance d'un profil dans l'analyse est proportionnelle à sa marge ($f_{i.}$ pour le profil-ligne i, $f_{.j}$ pour le profil-colonne j)
  - → dans le calcul du centre de gravité des profils
  - → dans le calcul de l'inertie par rapport au centre du nuage
- Distance entre les profils : distance du χ^2 plutôt que distance euclidienne classique, par ex. entre profils-lignes $\mathbf{i}, \mathbf{l}$:

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

  - Équivalence distributionnelle : si deux colonnes proportionnelles sont cumulées en une seule (fusion de deux modalités), les résultats de l'analyse ne changent pas
- Pour continuer à employer la distance euclidienne il faut redéfinir les profils : profil ligne i : $\frac{1}{f_{i.}}\left[\frac{f_{i1}}{\sqrt{f_{.1}}} \ldots \frac{f_{ij}}{\sqrt{f_{.j}}} \ldots \frac{f_{iq}}{\sqrt{f_{.q}}}\right]$, profil colonne j : $\frac{1}{f_{.j}}\left[\frac{f_{1j}}{\sqrt{f_{1.}}} \ldots \frac{f_{ij}}{\sqrt{f_{i.}}} \ldots \frac{f_{pj}}{\sqrt{f_{.p}}}\right]$
- Avec ces modifications des profils et l'introduction des pondérations, l'analyse des profils-lignes et des profils-colonnes se déroule comme l'ACP

AFCB : relation entre analyses, interprétation

- Comme pour l'ACP, on peut obtenir les relations de transition suivantes :

$$\psi_{\alpha i} = \frac{1}{f_{i.}\sqrt{\lambda_\alpha}} \sum_{j=1}^{q} f_{ij}\phi_{\alpha j}, \quad \phi_{\alpha j} = \frac{1}{f_{.j}\sqrt{\lambda_\alpha}} \sum_{i=1}^{p} f_{ij}\psi_{\alpha i}$$

où $\psi_{\alpha i}$ est la coordonnée du profil-ligne i sur l'axe factoriel α et $\phi_{\alpha j}$ la coordonnée du profil-colonne i sur l'axe factoriel α

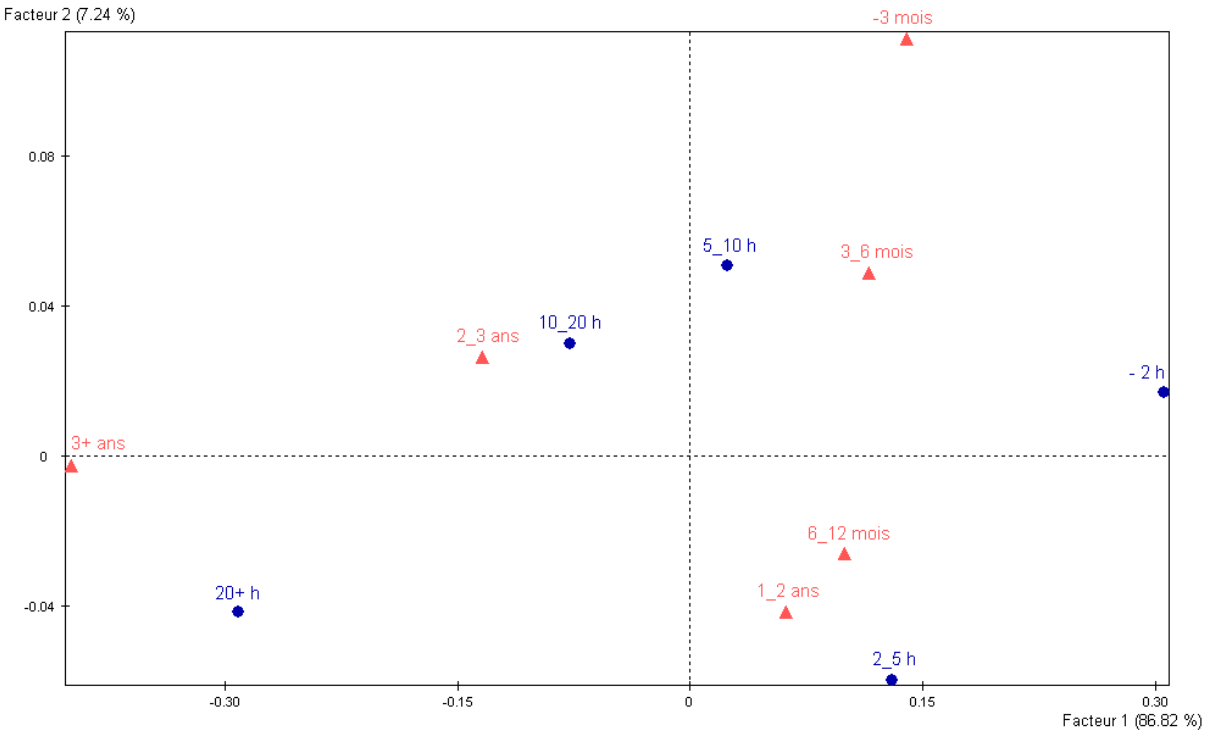
$\Rightarrow$ rend possible l'interprétation à partir d'une représentation simultanée

- Les outils employés pour l'ACP restent valables : qualité de représentation (contributions relatives), influence de chaque profil (contributions), choix du nombre d'axes, profils supplémentaires

- Variables nominales indépendantes $\Rightarrow$ toutes les projections sont confondues

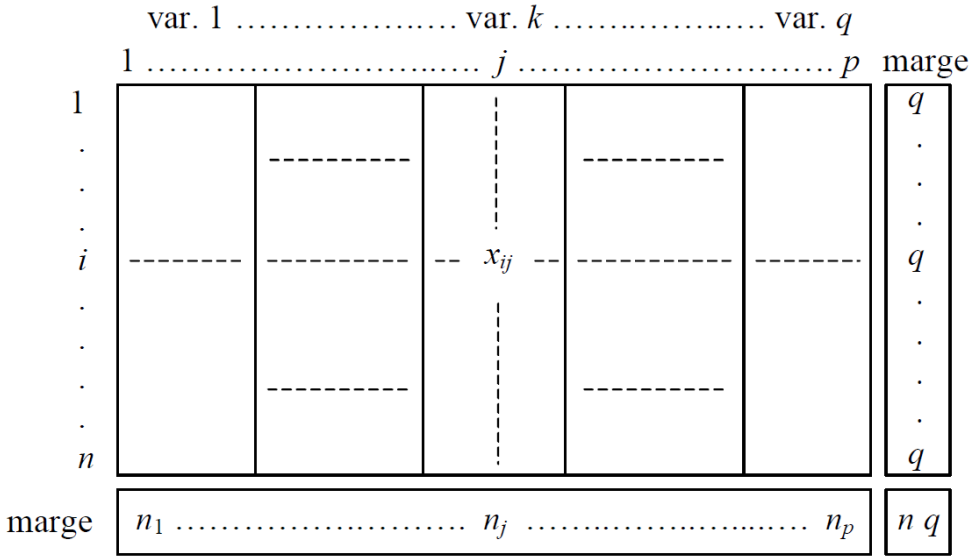
$\Rightarrow$ appliquer avant analyse un test du χ^2 d'indépendance entre les variables !

Utilisation Internet : résultats AFCB



Analyse des correspondances multiples (ACM)

- Plus de 2 variables nominales $\rightarrow$ analyse des correspondances multiples
- Données : n observations caractérisées par q variables *nominales* (ou à modalités), représentées par un *tableau disjonctif complet* (TDC) ; chaque observation possède exactement une modalité pour chaque variable



ACM : objectifs, utilisations

- Objectif général : mettre en évidence les relations dominantes entre modalités des variables nominales initiales
- Utilisations
  - Traitement d'enquêtes basées sur des questions fermées à choix multiples
  - Résumer un grand nombre de variables nominales par un faible nombre de variables quantitatives
  - Lorsque le nombre d'observations est faible, peut mettre en évidence des relations intéressantes entre observations et variables
  - Possibilité d'inclure des variables quantitatives dans l'analyse, après leur transformation en variables nominales

ACM : exemple

- Données issues de l'enquête « Les étudiants et la ville », sous la direction de S. Denèfle, Université de Tours, 2001 (voir [3])
- Questions (variables nominales) choisies :
  - Habitez-vous : seul(e), en colocation, en couple, avec les parents
  - Quel type d'habitation occupez-vous : cité U, studio, appartement, chambre chez l'habitant, autre
  - Si vous vivez en dehors du foyer familial, depuis combien de temps : moins d'1 an, de 1 à 3 ans, plus de 3 ans, non applicable
  - A quelle distance de la fac vivez-vous : moins d'1 km, de 1 à 5 km, plus de 5 km
  - Quelle est la surface habitable de votre logement : moins de $10m^2$, de $10m^2$ à $20m^2$, de $20m^2$ à $30m^2$, plus de $30m^2$
- Observation : sur les 5 variables, 3 sont issues de variables quantitatives discrétisées !
- Nombre élevé de modalités $\Rightarrow$ le TDC (ou le tableau de Burt $=$ concaténation des tableaux de contingences par paires de variables) est peu lisible $\Rightarrow$ synthèse par ACM nécessaire

ACM : solution

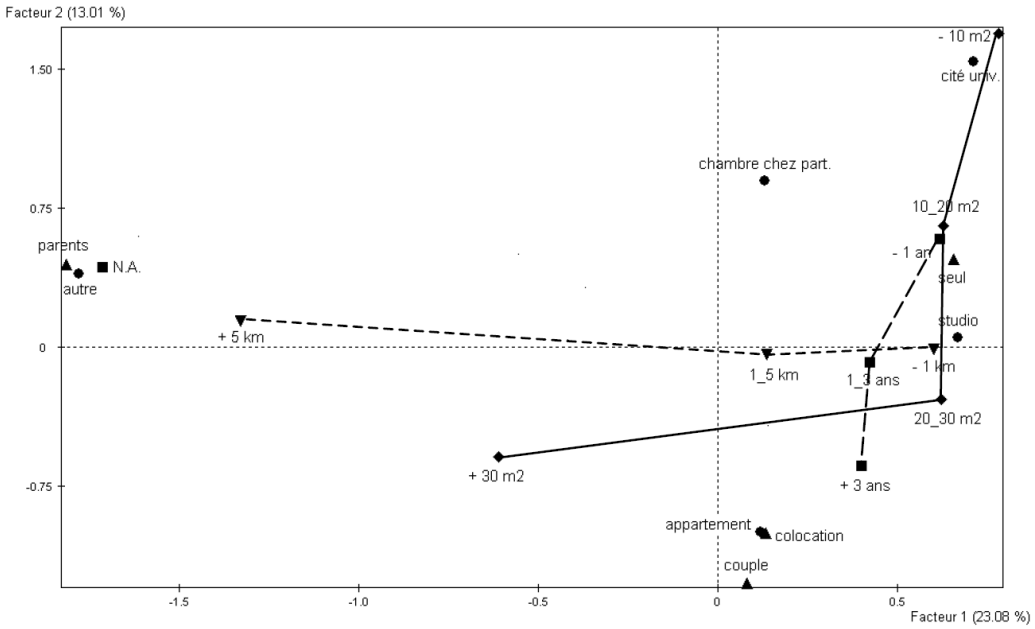
- Analyse des modalités (colonnes TDC) et analyse des observations (lignes du TDC)
  - Grand nombre d'observations (n) $\rightarrow$ on s'intéresse surtout à l'analyse des modalités
  - Emploi de la distance du χ^2 : l'influence de chaque coordonnée est pondérée par l'inverse de son poids
- Analyse des modalités :
  - Pondération de chaque modalité (colonne du TDC) par sa fréquence relative $\Rightarrow$ les modalités rares ont un impact faible sur l'analyse des observations
  - Distance du χ^2 : pas d'impact ici car toutes les observations ont le même poids
- Analyse des observations :
  - Pondération des observations : sans impact, toutes les observations ont le même poids
  - Distance du χ^2 entre observations : oui, car les modalités ont des poids différents
- Comme pour l'ACP, on cherche un espace de dimension k ($\ll$ nombre colonnes du TDC) qui résume le mieux la dispersion du nuage analysé

ACM : interprétation

- Sur la base de similarités (ou oppositions) entre projections de modalités :
  - De variables différentes : mêmes populations d'observations
  - D'une même variable (donc mutuellement exclusives) : populations similaires par rapport aux *autres* variables
- Observations importantes :
  - Le centre de gravité des modalités d'une même variable se confond avec le centre de gravité du nuage de toutes les modalités
  - Plus une modalité est rare, plus elle est éloignée du centre de gravité

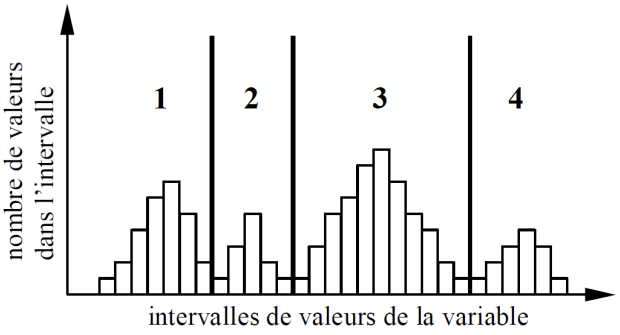
ACM : résultats sur l'exemple [3]

- Oppositions fortes entre modalités : « seul » *vs* « parents », « +5 km » *vs* « -1 km »
- Pour faciliter l'interprétation, les modalités successives de variables ordinales (issues de variables quantitatives discrétisées) sont reliées entre elles



ACM : inclusion de variables quantitatives

- Intérêt :
  - Trouver des relations entre modalités de variables qualitatives et *intervalles* de valeurs de variables quantitatives
  - Mettre en évidence des relations *non linéaires* entre intervalles de variables quantitatives
- Découper en intervalles le domaine de variation de la variable quantitative (sur la base de connaissances *a priori*, à partir de l'histogramme, etc.) ; chaque intervalle sera une modalité de la nouvelle variable qualitative



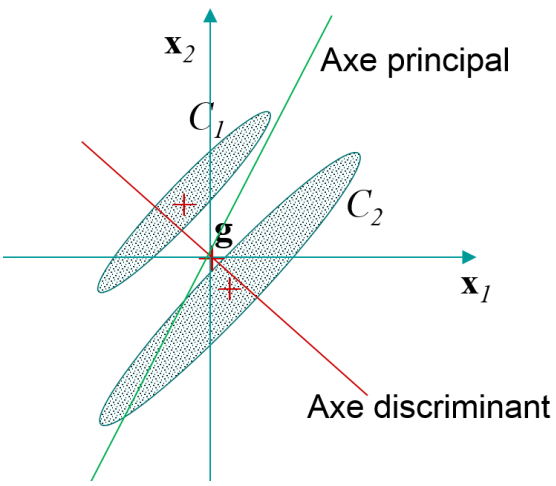
Plan du cours

AFD : données, objectifs

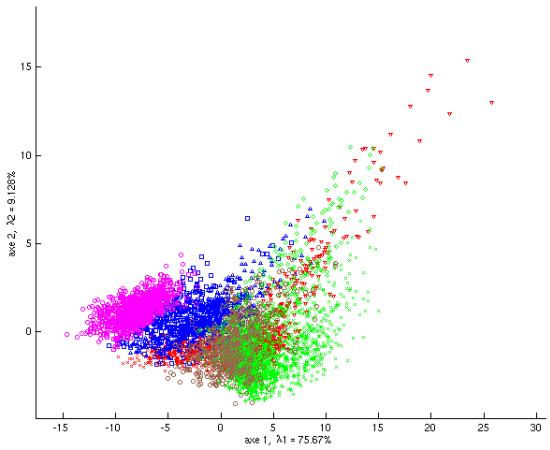
- Données : n observations caractérisées par m variables quantitatives (matrice de données $\mathbf{X}$) *et* une variable nominale « classe » $Y \in \{1, \ldots, q\}$
- Objectifs :
  - Étape descriptive : identifier des « facteurs discriminants » (combinaisons linéaires de variables explicatives) qui permettent de différencier au mieux les classes
  - Étape décisionnelle : sur la base des valeurs prises par les variables explicatives, décider à quelle classe affecter une nouvelle observation
- Utilisations :
  - Descriptive : condenser la représentation des données en conservant au mieux la séparation entre les classes
  - Décisionnelle : classer de nouvelles observations à partir de leur projection sur le sous-espace linéaire qui optimise la séparation

AFD *versus* ACP

- L'ACP maximise la *variance* des projections sur le sous-espace
- L'AFD maximise la différenciation entre les classes dans le sous-espace

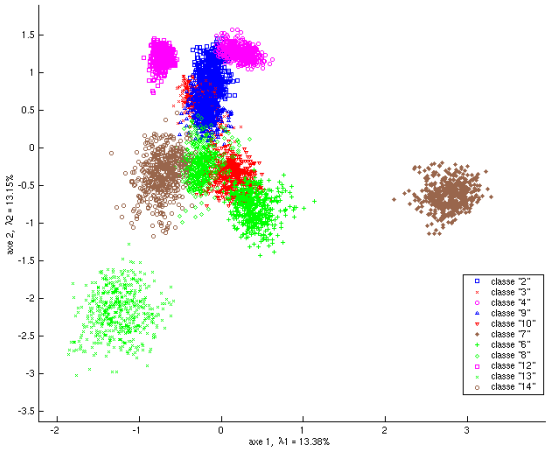


AFD *versus* ACP : illustration sur Textures



ACP

(proj. sur les 2 premiers axes principaux)

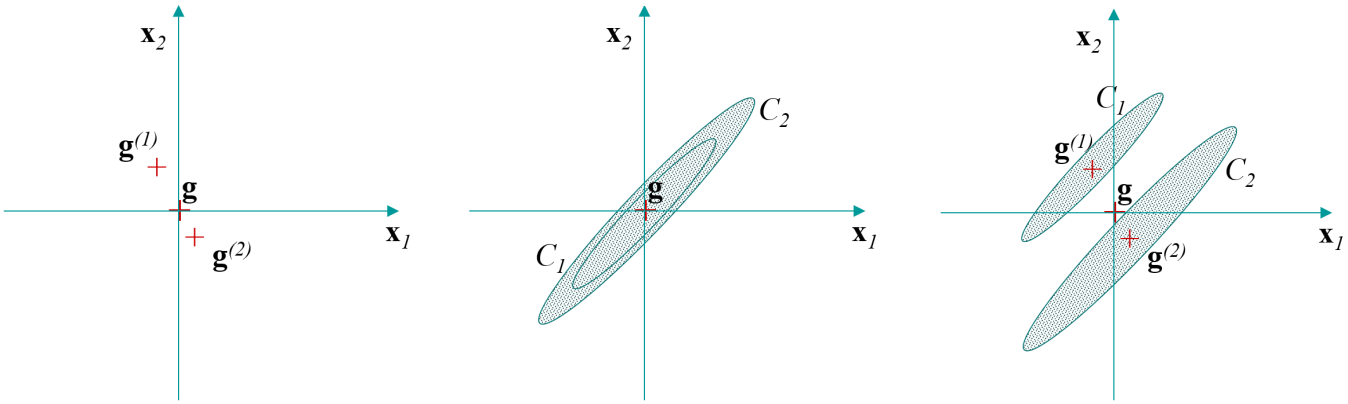


AFD

(proj. sur les 2 premiers axes discriminants)

AFD : solution

- Covariances entre variables :
  1. Inter-classes : $\mathbf{E}$, calculée en considérant que les observations sont les centres de gravité des classes
  2. Intra-classes : $\mathbf{D}$, calculée sur les observations de départ, en centrant chaque classe sur son centre de gravité
  3. Totale : $\mathbf{S}$, calculée sur les observations de départ ; relation de Huygens $\mathbf{S} = \mathbf{E} + \mathbf{D}$

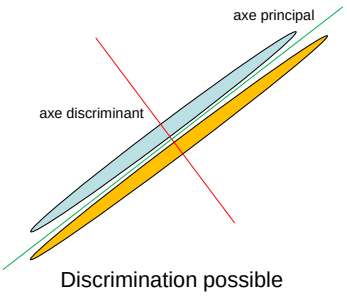


AFD : solution (2)

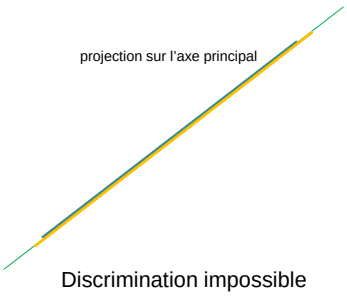
- Le sous-espace de dimension k recherché est généré par les k vecteurs propres $\mathbf{u}_\alpha$ associés aux k plus grandes valeurs propres λ_α de l'équation de valeurs et vecteurs propres généralisée $\mathbf{E}\mathbf{u}_\alpha = \lambda_\alpha \mathbf{S}\mathbf{u}_\alpha$, $\alpha \in \{1, \ldots, k\}$
  - Il est possible de résoudre plutôt $\mathbf{E}\mathbf{u}_\alpha = \lambda_\alpha \mathbf{D}\mathbf{u}_\alpha$ si le rang de $\mathbf{D}$ n'est pas inférieur à celui de $\mathbf{S}$
  - Si $\mathbf{S}$ est inversible, il est préférable de résoudre $\mathbf{S}^{-1}\mathbf{E}\mathbf{u}_\alpha = \lambda_\alpha \mathbf{u}_\alpha$
- La matrice $\mathbf{S}^{-1}\mathbf{E}$ n'est pas symétrique en général $\Rightarrow$ les axes factoriels discriminants ne sont pas orthogonaux (vecteurs propres associés à des valeurs propres différentes)
- q classes $\Rightarrow \mathrm{rang}(\mathbf{E}) \leq q - 1$ (car $\mathbf{E}$ est calculée à partir des seuls centres de gravité des q classes) $\Rightarrow$ au maximum $q - 1$ axes discriminants !
  - Problème à 2 classes $\Rightarrow$ 1 axe discriminant

AFD : solution (3)

- Approche fréquente si $\mathbf{S}$ est singulière :
  1. Réduire dimension avec l'ACP pour que dans l'espace réduit $\mathbf{S}'$ soit de rang complet
  2. Résoudre $\mathbf{S}'^{-1}\mathbf{E}'\mathbf{u}_\alpha = \lambda_\alpha \mathbf{u}_\alpha$ dans l'espace réduit
- Si $\mathbf{S}$ est mal conditionnée ($\frac{\|\lambda_{\max}\|}{\|\lambda_{\min}\|} > \theta$, $\lambda_{\min}$ 0), lors de son inversion l'imprécision est amplifiée de façon excessive ; la borne θ dépend de la précision de représentation et de l'algorithme. Solutions :
  1. Réduire la dimension avec une ACP pour rendre dans l'espace réduit $\mathbf{S}'$ bien conditionnée (et non seulement de rang complet) $\Rightarrow$ risque d'élimination de variables discriminantes !



Discrimination possible



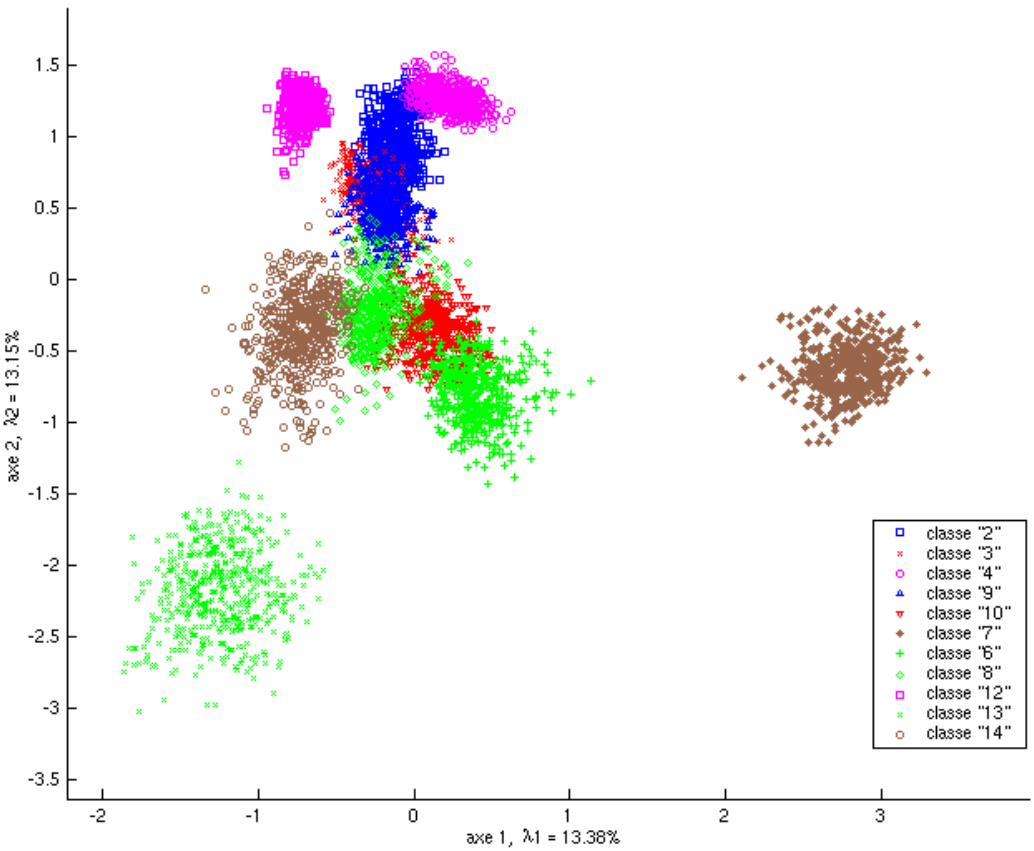
Discrimination impossible

  2. Régularisation : remplacer $\mathbf{S}$ par $\mathbf{S} + r\mathbf{I}_m$ pour r assez grand ($\frac{\|\lambda_{\max}\|+r}{\|\lambda_{\min}\|+r} < \theta$)

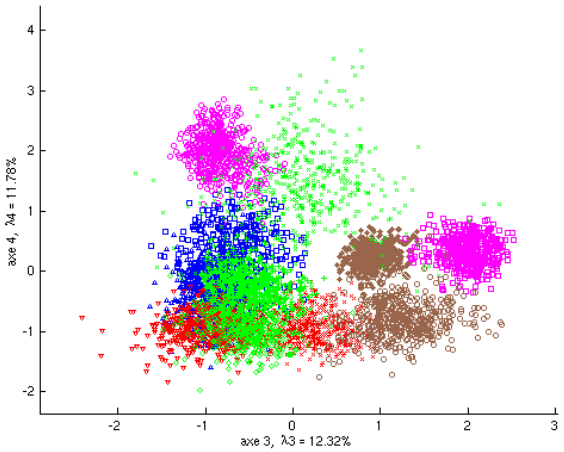
AFD : choix du nombre d'axes

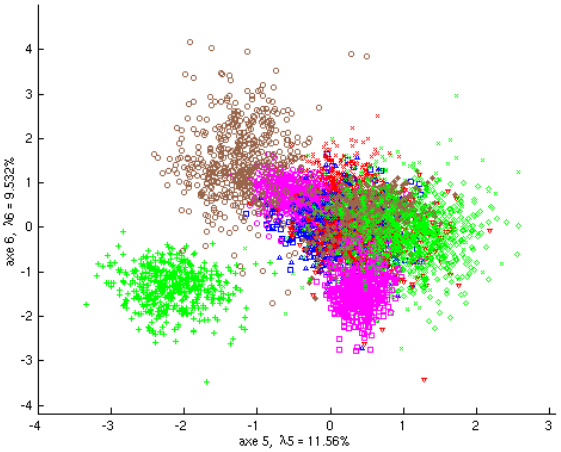
1. Tests statistiques (hypothèses : classes issues de lois normales) :
   1. Test de Rao : test d'égalité à 0 de la i-ème valeur propre
   2. Test du Lambda de Wilks : apport significatif des axes au-delà du i-ème ?
   3. Test incrémental : apport significatif du $i+1$-ème axe ?
2. Méthode de l'échantillon-test (si utilisation décisionnelle de l'AFD ou pré-traitement par AFD avant application d'un modèle décisionnel) :
   1. Extraire (par tirages aléatoires) un échantillon-test
   2. Répéter pour différentes valeurs du nombre d'axes : appliquer l'AFD sur les données restantes, développer le modèle décisionnel sur ces mêmes données, évaluer le modèle décisionnel sur l'échantillon-test
   3. Choisir les paramètres (dont nombre d'axes d'AFD) qui donnent le meilleur résultat
   - La capacité de généralisation du modèle retenu sera estimée à partir de données non utilisées pour réaliser l'AFD ou pour choisir les paramètres

AFD sur « textures » : projections sur les 2 premiers axes



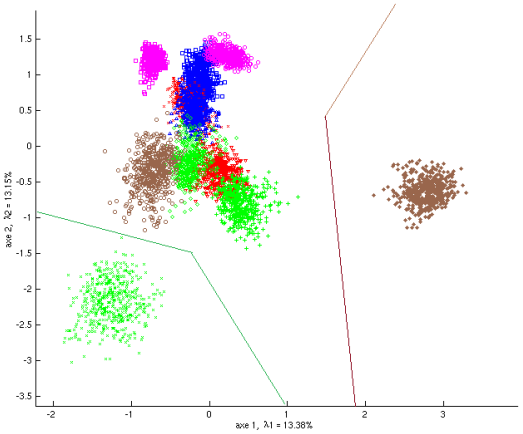
AFD sur « textures » : projections sur les axes 3-4 et 5-6





Etape décisionnelle de l'AFD

- Objectif : classer une nouvelle observation dans une des classes sur la base des valeurs prises par les variables explicatives
- Méthode : affecter l'observation à la classe la plus « proche »
  - Étape descriptive $\rightarrow$ projection sur le sous-espace le plus discriminant
  - Proximité à une classe = distance entre observation et le centre de la classe
    - Même métrique (donnée par $\mathbf{S}'^{-1}$) pour toutes les classes $\Rightarrow$ frontières linéaires
    - Métrique spécifique à chaque classe $\Rightarrow$ frontières quadratiques



Références I

T. Allison and D. Cicchetti.

Sleep in mammals : ecological and constitutional correlates.
Science, 194 :732–734.

J.-P. Benzécri.

L'analyse des données, tome 1, La taxinomie, tome 2, L'analyse des correspondances.
Dunod, Paris, 1973.

M. Crucianu, J.-P. Asselin de Beauville, and R. Boné.

Méthodes d'analyse factorielle des données : méthodes linéaires et non linéaires.
Hermès, Paris, 2004.

G. Saporta.

Probabilités, Analyse des Données et Statistique.
Technip, Paris, 2011.