MS BGD: MDI 343
Apprentissage non supervisé, analyse exploratoire
ACP - AC - ACM

Anne Sabourin
Telecom ParisTech

21 Décembre 2018

Apprentissage non-supervisé ?

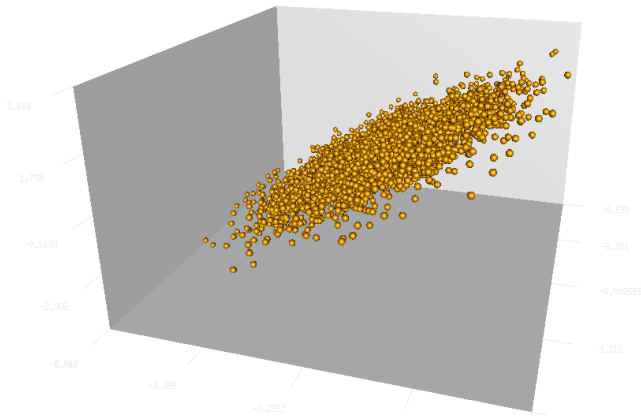
supervisé (rappel) : Données $\mathcal{D} = (x_i, y_i)_{i=1:n}$.

- x : 'features' (âge, taille, lieu de naissance, profession, ...)
- y : étiquettes (ex : revenu annuel) : la quantité d'intérêt.
- But : **prédire** $\widehat{y}$ pour une nouvelle entrée x.

non supervisé : Données $\mathcal{D} = (x_i)_{i=1:n}$: pas d'étiquettes Y_i !

- But : **construire un 'modèle'**/ une 'représentation' des x_i
- trouver des **structures** simples expliquant les données (clustering, réduction de dimension)
- détecter des **anomalies** (futures ou passées).

Réduction de dimension/analyse exploratoire : contexte



- Données en grande dimension (beaucoup d'attributs/features)
- Seules certaines combinaisons (linéaires) sont majoritaires
- Données concentrées dans un sous-espace de faible dimension

Comment trouver une représentation de faible dimension capturant la variabilité des données ?

Terminologie de l'analyse exploratoire

- $X \in \mathbb{R}^{n \times p}$ un tableau de données.
- $x_i = X_{[i,\cdot]}$: ligne i : un **individu** $\in \mathbb{R}^p$
- $X^j = X_{[\cdot,j]}$: colonne j : une **variable** (feature) $\in \mathbb{R}^n$
- Point de vue probabiliste : les 'variables' peuvent être vus comme un échantillon $(x_{1,j}, \ldots, x_{n,j})$ d'une variable aléatoire.
- Point de vue de l'analyse exploratoire : on identifie la loi de la variable aléatoire et la distribution de l'échantillon X^j : Autrement dit on identifie espérance et moyenne empirique, variance et variance empirique.

Variables continues : exemple

- Variables continues : performances (temps, hauteur/distance de saut, etc) aux dix disciplines du décathlon, pour un ensemble de sportifs.
    - Peut-on dégager des profils d'athlètes ?
    - A quel point les athlètes correspondent-ils aux profils dégagés ?
    - Comment décomposer la variabilité des performances ?

Données decathlon (exemple jouet)

	100m	Long.jump	Shot.put	High.jump	400m	110m.hurdle	Discus
SEBRLE	11.04	7.58	14.83	2.07	49.81	14.69	43.75
CLAY	10.76	7.40	14.26	1.86	49.37	14.05	50.72
KARPOV	11.02	7.30	14.77	2.04	48.37	14.09	48.95
BERNARD	11.02	7.23	14.25	1.92	48.93	14.99	40.87

	Pole.vault	Javeline	1500m	Rank	Points	Competition
SEBRLE	5.02	63.19	291.70	1	8217	Decastar
CLAY	4.92	60.15	301.50	2	8122	Decastar
KARPOV	4.92	50.31	300.20	3	8099	Decastar
BERNARD	5.32	62.77	280.10	4	8067	Decastar

- $n = 41$
- $p = 13$

Variables catégorielles (discrètes)

- 2 Variables catégorielles (= discrètes = qualitatives) : couleurs des cheveux (bruns, blonds, châtain) et des yeux (bleus, marrons, vert)
    - Certaines modalités d'une variables sont-elles particulièrement liées à une autre ?
    - comment représenter graphiquement les données croisées ?

Variables catégorielles (discrètes)

- 2 Variables catégorielles (= discrètes = qualitatives) : couleurs des cheveux (bruns, blonds, châtain) et des yeux (bleus, marrons, vert)
  - Certaines modalités d'une variables sont-elles particulièrement liées à une autre ?
  - comment représenter graphiquement les données croisées ?
- Plusieurs variables catégorielles : réponses à un sondage (=QCM avec plusieurs questions)
  - Une question = une variable j prenant m_j modalités possibles.
  - Comment représenter les individus ? les variables ? les modalités des variables ?
  - Comment résumer les relations de dépendances entre les individus/ variables /modalités ?

Représentation de données concernant 2 variables

Table de contingence : Données couleurs des yeux/cheveux.

	Brown	Blue	Hazel	Green
Black	68	20	15	5
Brown	119	84	54	29
Red	26	17	14	14
Blond	7	94	10	16

- $n = 592$ individus.
- Variable X = 'couleur des cheveux' : 4 modalités.
- Variable Y = 'couleur des yeux' : 4 modalités (hasard)

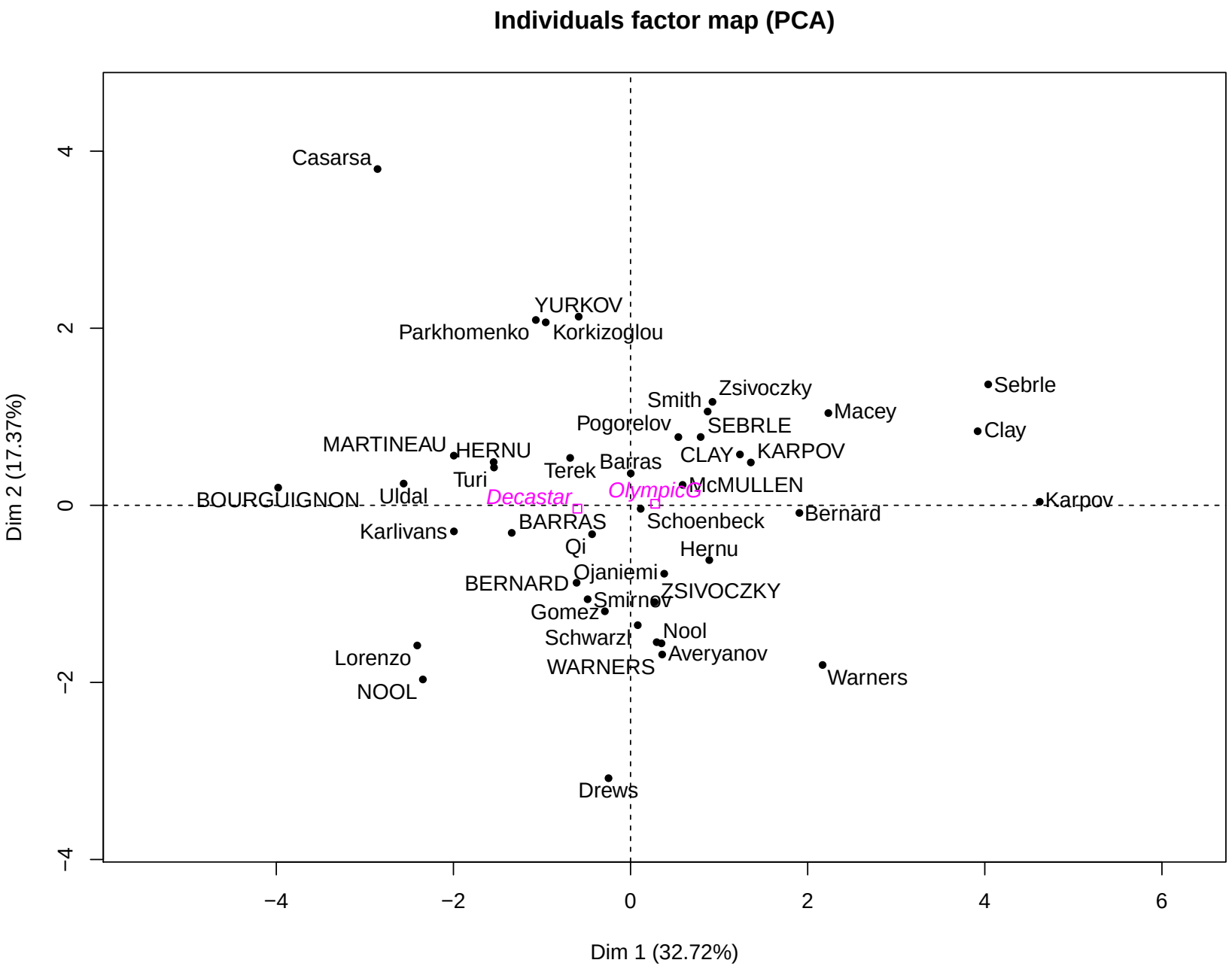
Plus de 2 variables catégorielles

Sondage de buveurs de thé

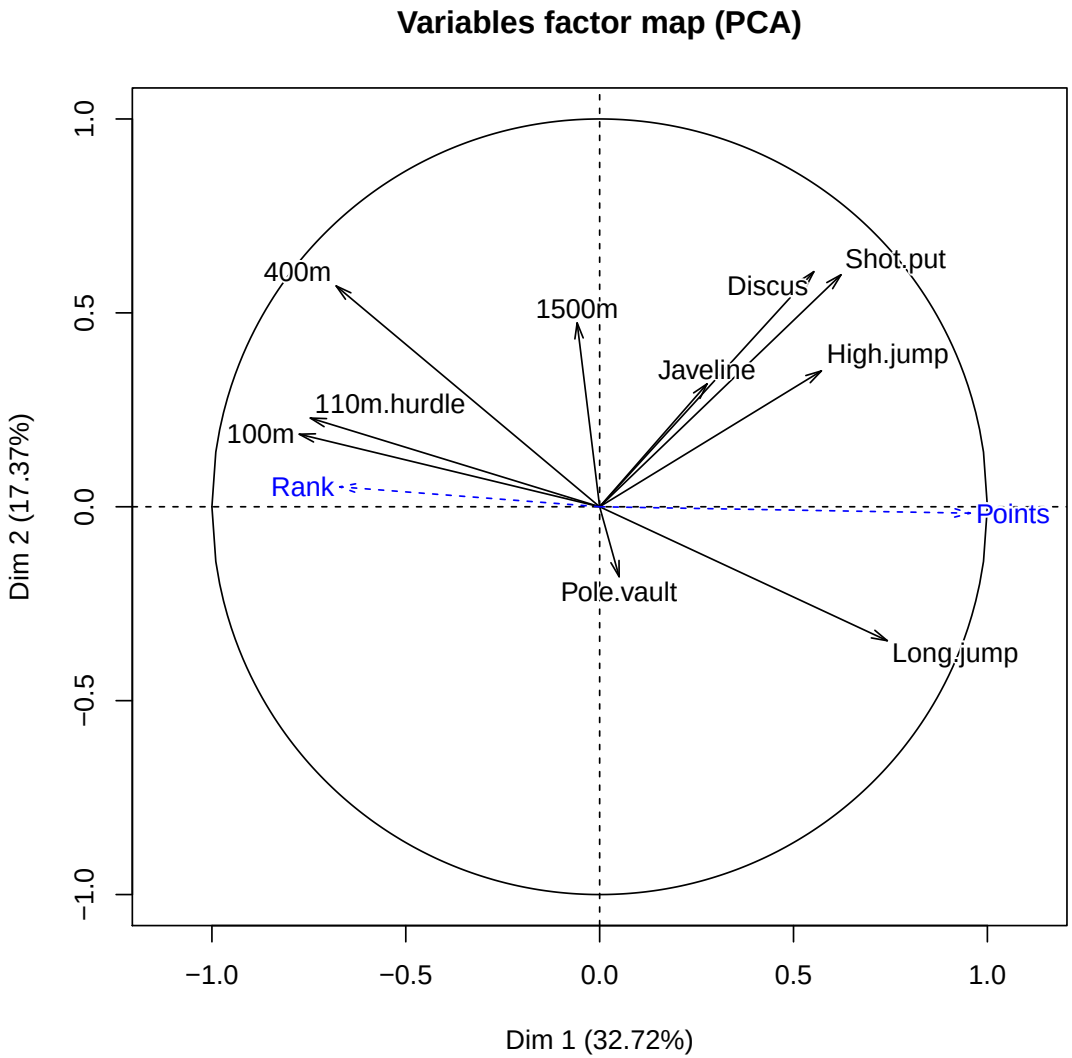
	breakfast	tea.time	evening	...	spirituality	healthy	...	sex	SPC	Sport	age_Q
1	breakfast	Not.tea time	Not.evening	...	Not.spirituality	healthy	...	M	middle	sportsman	35-44
2	breakfast	Not.tea time	Not.evening	...	Not.spirituality	healthy	...	F	middle	sportsman	45-59
3	Not.breakfast	tea time	evening	...	Not.spirituality	healthy	...	F	other worker	sportsman	45-59
4	Not.breakfast	Not.tea time	Not.evening	...	spirituality	healthy	...	M	student	Not.sportsman	15-24
5	breakfast	Not.tea time	evening	...	spirituality	Not.healthy	...	M	employee	sportsman	45-59
6	Not.breakfast	Not.tea time	Not.evening	...	Not.spirituality	healthy	...	M	student	sportsman	15-24
7	breakfast	tea time	Not.evening	...	Not.spirituality	healthy	...	M	senior	sportsman	35-44
8	Not.breakfast	tea time	evening	...	Not.spirituality	healthy	...	F	middle	sportsman	35-44
	...	...	...	...	...	...	...	...	...	...	...

- $n = 300$ consommateurs $\rightarrow$ 300 lignes
- 36 questions dont âge (seule variable quantitative)

Décathlon : carte des individus

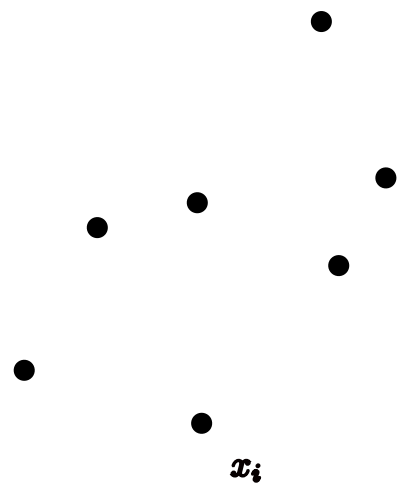


Décathlon : carte des variables (cercle des corrélations)



ACP, épisode 1 : Représentation de variance maximale

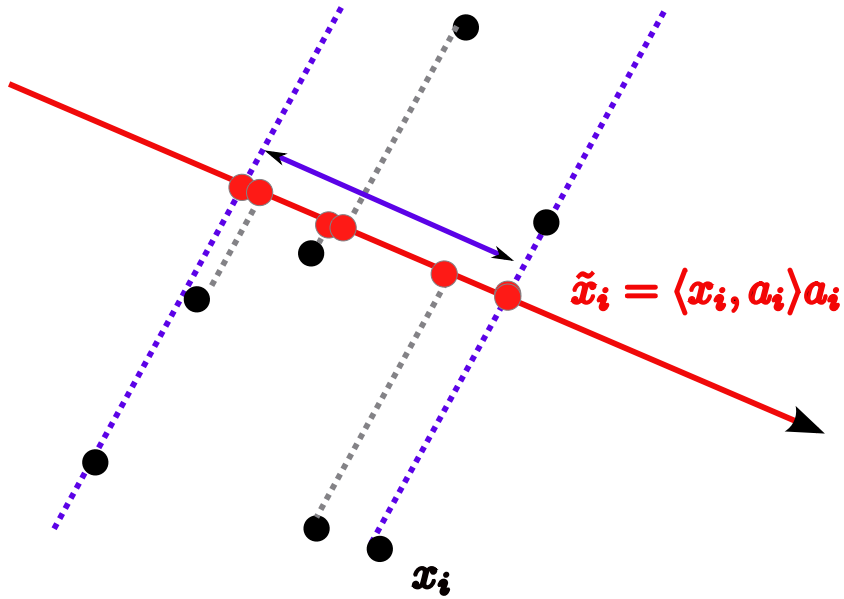
- $(x_1, \ldots, x_n)$ un échantillon, $x_i = (x_i^1, \ldots, x_i^p)$, $X \in \mathbb{R}^{n \times p}$ le tableau de données.
- Premier **axe principal** ?
  = (déf) une direction (axe principal) $a_1 \in \mathbb{R}^p$ maximisant la variance des projections



Données

ACP, épisode 1 : Représentation de variance maximale

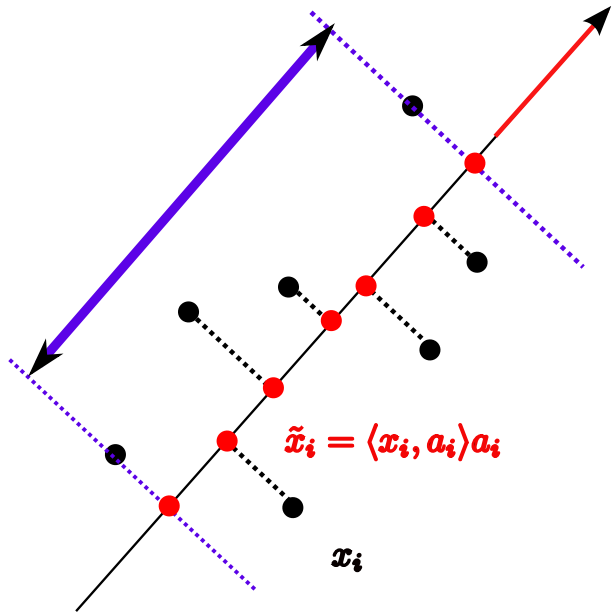
- $(x_1, \ldots, x_n)$ un échantillon, $x_i = (x_i^1, \ldots, x_i^p)$, $X \in \mathbb{R}^{n \times p}$ le tableau de données.
- Premier **axe principal** ?
  = (déf) une direction (axe principal) $a_1 \in \mathbb{R}^p$ maximisant la variance des projections



Par ici ?

ACP, épisode 1 : Représentation de variance maximale

- $(x_1, \ldots, x_n)$ un échantillon, $x_i = (x_i^1, \ldots, x_i^p)$, $X \in \mathbb{R}^{n \times p}$ le tableau de données.
- Premier **axe principal** ?
  = (déf) une direction (axe principal) $a_1 \in \mathbb{R}^p$ maximisant la variance des projections



par là ?

Un problème de maximisation de variance

- Moyenne empirique : $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
- Variance empirique de la projection sur a_1 :

$$\mathrm{var}_n(a_1^\top X) = \frac{1}{n}\sum_{i=1}^{n}(a_1^\top x_i - a_1^\top \bar{x})^2 \quad = \quad a_1^\top V a_1$$

$$\text{où} \quad V = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top$$

(Matrice de covariance empirique)

- Le premier axe principal a_1 est solution du problème d'optimisation :

$$\max_{a \in \mathbb{R}^p} \quad a^\top V a \qquad \text{s.c.} \quad \|a\| = 1.$$

Solutions du problème

- Le problème : « $\max_{a \in \mathbb{R}^p} \quad a^\top V a \qquad \text{s.c.} \quad \|a\| = 1$ » a pour solution le vecteur a_1 associé à la plus grande valeur propre λ_1 de V,

$$V a_1 = \lambda_1 a_1.$$

(**Preuve** : écrire le lagrangien du problème et annuler le gradient)

- Le vecteur a_1 est appelé le **premier axe principal** ($\in \mathbb{R}^p$).

- Les coordonnées sur a_1 des projections des x_i forment la **première composante principale** c_1 ($\in \mathbb{R}^n$)

$$c_{1,i} = \langle x_i, c_1 \rangle \; ; \quad c_1 = X a_1$$

Étapes suivantes de l'analyse

- Axes suivants :

$$\max_{a} \quad a^\top V a \qquad \text{s.c.} \quad \|a\| = 1, a \perp a_1$$

  $\rightarrow$ solution a_2, 2^{eme} axe principal, vecteur propre pour la 2^{eme} + grande v.p :

$$V a_2 = \lambda_2 a_2$$

  Deuxième composante principale : projections sur a_2,

$$c_2 = X a_2.$$

- *etc...* Recherche de a_M dans l'orthogonal de $vect(a_1, \ldots, a_{M-1})$.

ACP : Étapes

Entrée : jeu de données $(x_1, \ldots, x_n)$ en dimension p

- Calculer $\bar{x}$ et V (moyenne et covariance empiriques)
- Calculer les r premiers vecteurs propres (r = rang de X) $(a_1, \ldots, a_r)$ de V = **axes principaux**
- Calculer les **composantes principales**

$$c_k = X a_k \,, \quad k \leq r.$$

- Formule de reconstitution (écriture dans la nouvelle base)

$$x_i = \sum_{k=1}^{r} c_{i,k} a_k \,, \qquad X = C A^\top$$

- Représentation simplifiée en dimension $s \leq r$ (factorisation)

$$\tilde{X} \;=\; \tilde{C}\tilde{A}^\top \;=\; \begin{pmatrix} c_1 & \ldots & c_s \end{pmatrix} \begin{pmatrix} a_1^\top \\ \vdots \\ a_s^\top \end{pmatrix} = \tilde{C}\tilde{A}^\top$$

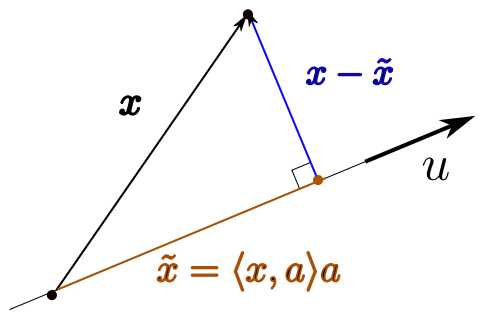
avec $\tilde{C} \in \mathbb{R}^{n \times s}$ et $\tilde{A} \in \mathbb{R}^{p \times s}$.

ACP, épisode 1 bis : minimisation de l'erreur

$\bar{x} = 0$.

- avec un axe ($s = 1$) $\tilde{X} = (a^\top X)\, a$ (projection orthogonale sur a unitaire).
- Erreur de représentation (résidu) : $\tilde{X} - X$.

$$\text{Pythagore :} \qquad \underbrace{\|X\|^2}_{\text{Constante \% } a} = \underbrace{\|\tilde{X}\|^2}_{(a^\top X)^2} + \underbrace{\|X - \tilde{X}\|^2}_{\text{Residu}}$$



- Maximiser variance $\iff$ Minimiser résidus

Variance des composantes principales, variance expliquée.

- La variance de la composante principale c associé à l'axe de valeur propre λ est (cas $\bar{x} = 0$)

$$V(c) = \frac{1}{n}\sum_k c_i^2 = \frac{1}{n}c^\top c = \frac{a^\top X^\top X a}{n} = a^\top V a = \lambda$$

- La variance expliquée par la composante c_k est le ratio

$$\frac{\lambda_k}{\sum_{j=1}^p \lambda_j}.$$

ACP : aspects computationnels

- Sensibilité à l'échelle : Avant de commencer, standardiser les données (chaque attribut doit être de même écart type (1), si possible centré).

- Calculer les vecteurs propres de V : implémenté dans `R, Python, C`... complexité en $O(p^3)$ ou au mieux $O(sp^2)$ (méthodes de puissances itérées)

- instantané pour p petit (≤ 50)

- Impossible quand p est grand !!

ACP en grande dimension ($p \gg n$) : formulation duale

$\bar{x} = 0$

- résoudre en a :

$$\frac{1}{n} X^\top X a = \lambda a \rightarrow \text{ difficile}$$

  Car $\frac{1}{n} X^\top X = V$ est de dimension p grande.

- en pré-multipliant par X :

$$\frac{1}{n} X X^\top X a = \lambda X a$$

  ie.

$$\frac{1}{n} X X^\top c = \lambda c$$

  $\rightarrow$ **c est vecteur propre de la matrice de Gram $\frac{1}{n} X X^\top$ avec même valeur propre**

- $X X^\top \in \mathbb{R}^{n \times n}$: plus facile à résoudre.

Formules de transition

- Rappel : $\frac{1}{n} X X^\top c = \lambda c$.
- En pré-multipliant par $X^\top$, on a

$$\underbrace{\frac{1}{n} X^\top X}_{V} X^\top c = \lambda X^\top c.$$

  Ainsi $X^\top c$ est vecteur propre de V, et nécessairement $X^\top c = ka$ pour un certain $k \in \mathbb{R}$.
- On obtient k en calculant les normes respectives de $\|a\|$ et $\|X^\top c\|$:
  - $\|a\| = 1$ par convention,
  - $\|X^\top c\|^2 = c^\top X X^\top c = c^\top (n\lambda c) = n^2 \lambda^2$.

  d'où $k = n\lambda$, on a montré les **formules de transition**

$$\begin{cases} c = Xa \\ a = \dfrac{1}{n\lambda} X^\top c. \end{cases} \tag{1}$$

ACP en pratique

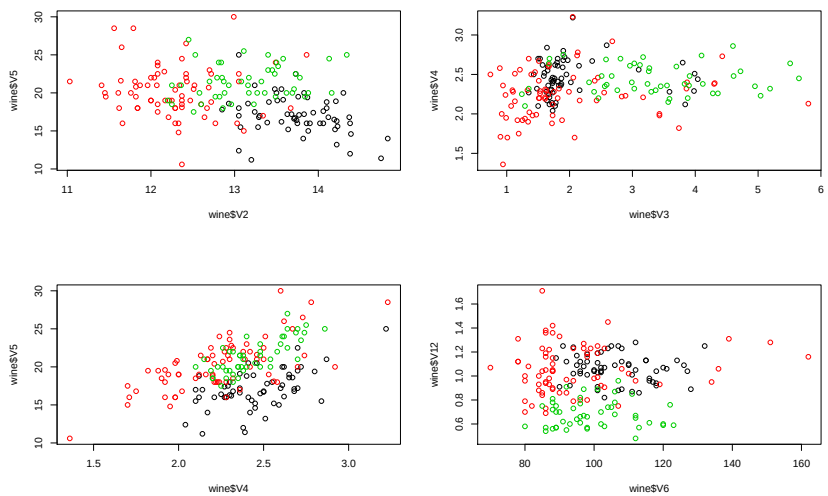
- Données de composition du vin produit par 3 producteurs : sur le dépôt UCI `http://archive.ics.uci.edu/ml/`
- Question : Peut-on retrouver des groupes correspondant à chacun des producteurs ? Quelle combinaison de features caractérise un producteur ?

```
> wine <- read.table(paste("http://archive.ics.uci.edu/ml/",
                           "machine-learning-databases/",
                           "wine/wine.data", sep=""),
                     sep=",")
> dim(wine)
[1] 178 14
```

- Première colonne : identifiant du producteur : $i \in \{1, 2, 3\}$.
- 13 suivantes : concentrations en divers composés chimiques.

Premier essai

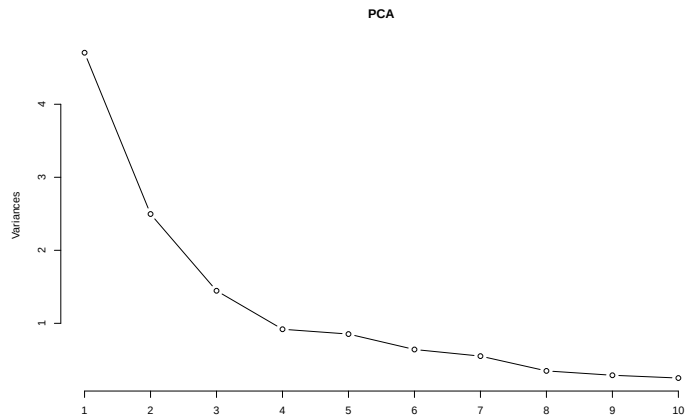
```
par(mfrow=c(2,2))
plot(wine$V2,wine$V5,col=wine$V1)
plot(wine$V3,wine$V4,col=wine$V1)
plot(wine$V4,wine$V5,col=wine$V1)
plot(wine$V6,wine$V12,col=wine$V1)
```



Couleur $\sim$ producteur : peu de structure !

Réduction de dimension par ACP

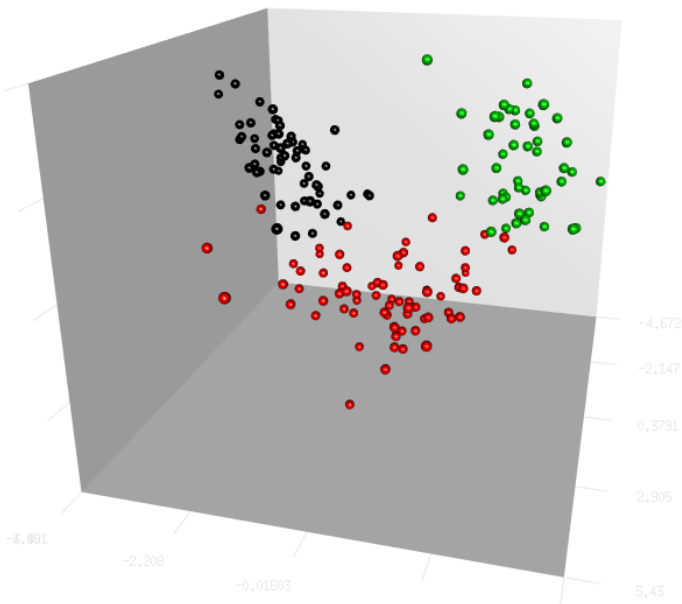
```
wine_data=wine[,-1]
Mean = apply(wine_data, 2, mean)
SD = apply(wine_data, 2, sd)
norm_data= t( (t(wine_data ) - Mean)/SD)
PCA = prcomp(norm_data)
plot(PCA,type="l")
```

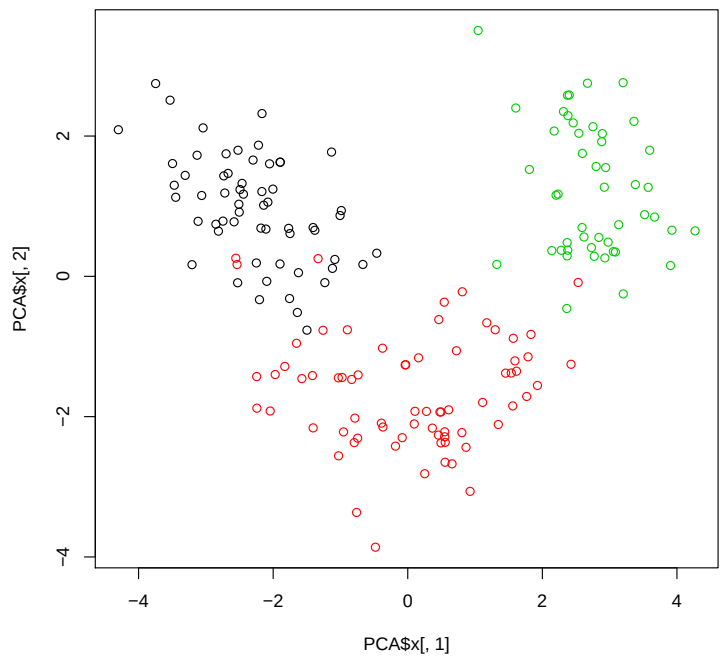


Variance de la projection sur chaque vecteur propre.
« coude » autour de $K = 4$: sélectionner 3 ou 4 composantes.

Résultats

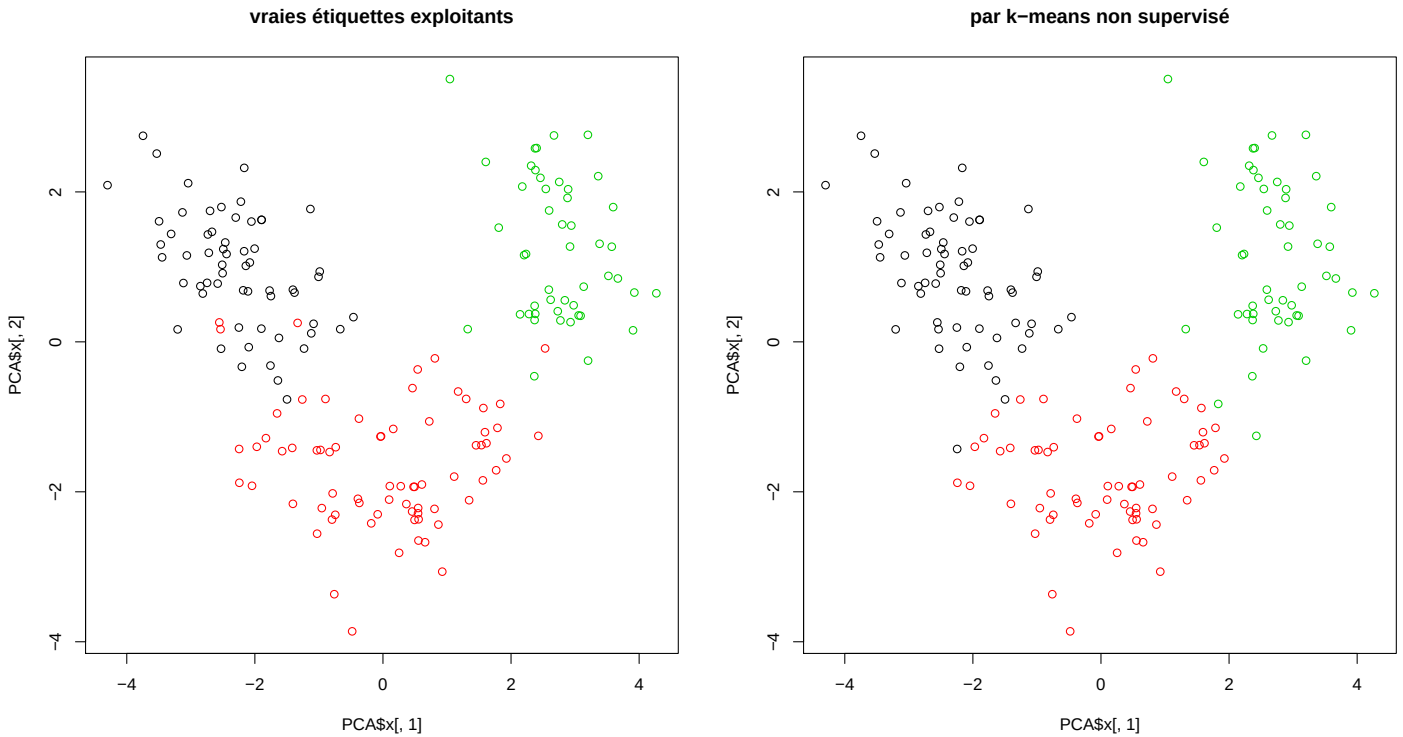
```
rgl.open()
rgl.spheres(x=PCA$x[,1], y=PCA$x[,2], z=PCA$x[,3],
            color = wine$ V1,radius=0.1)
plot(PCA$x[,1],PCA$x[,2], col=wine$V1)
```





En projetant sur 3 ou 2 composantes, on retrouve les trois groupes (producteurs)

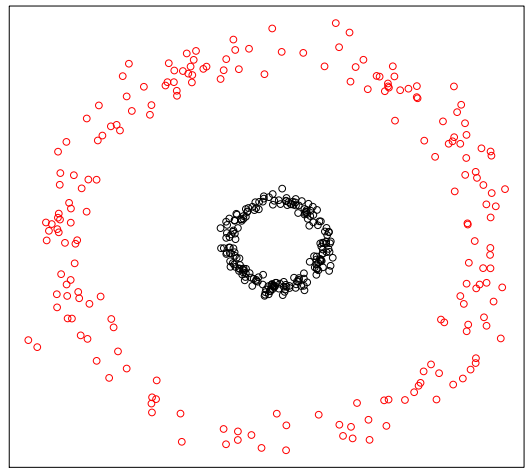
Clustering sur données réduites par ACP



nombre d'erreurs des k-means : 7 (sur 178 exemples)

Méthodes à noyaux : idée de base

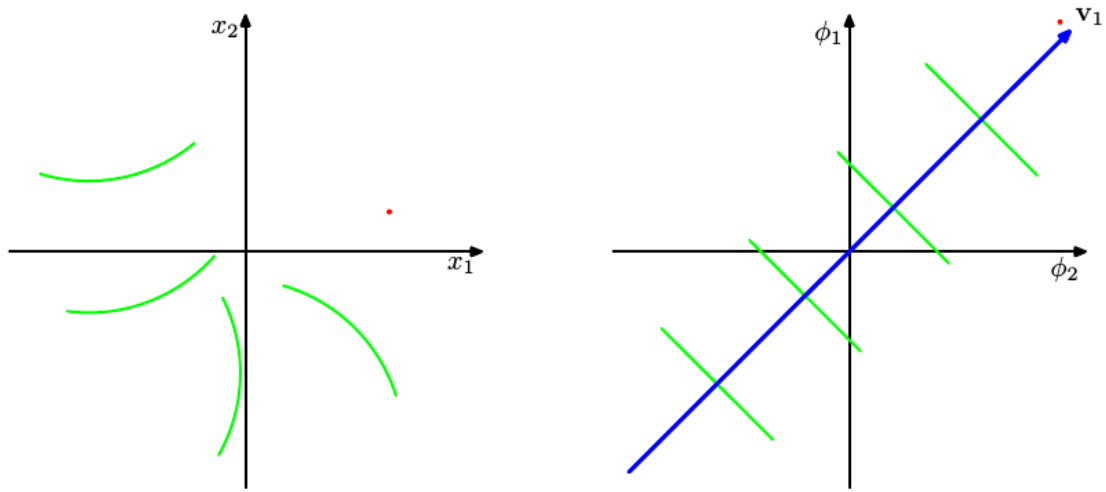
Parfois, les données n'ont pas de structure linéaire



idée : transformation Φ (non linéaire) des données, puis méthodes linéaires sur les $\Phi(x_i)$.

Méthodes à noyaux : idée de base

Parfois, les données n'ont pas de structure linéaire



(image : Bishop, Pattern recognition and ML.)

idée : transformation Φ (non linéaire) des données, puis méthodes linéaires sur les $\Phi(x_i)$.

Noyau et feature map

- On se donne une « feature map » $\Phi(x)$ ou un « noyau » (nouveau produit scalaire) $k(x, x')$, tels que

$$\Phi(x)^\top \Phi(x) = k(x, x').$$

- Φ est implicite et n'intervient pas en pratique : seul compte le noyau k.
- Exemples de noyau :
  - Linéaire $k(x, x') = x^\top x' = \langle x, x' \rangle$.
  - Gaussien : $k(x, x') = e^{-\sigma \|x - x'\|^2}$
  - Polynomial : $k(x, x') = \Big(a\langle x, x' \rangle + b\Big)^d$
- On applique l'ACP linéaire pour la grande dimension aux $\Phi(x_i)$.

ACP et astuce du noyau

- Problème initial : résoudre $Ca = \lambda a$ dans l'espace des $\Phi(x_i)$, avec

$$C = \frac{1}{n} \sum \Phi(x_i)\Phi(x_i)^\top$$

- **Rappel** : ACP en grande dimension : on résout le problème dual

$$\frac{1}{n} Kc = \lambda c, \qquad \text{où } K \in \mathbb{R}^{n \times n}, \quad K_{i,j} = x_i^\top x_j.$$

- ici : *idem* avec $x_i \leftarrow \Phi(x_i)$:

$$K_{i,j} = \Phi(x_i)^\top \Phi(x_j) = k(x_i, x_j).$$

N.B. : si $k(x_i, x_j) = \langle x_i, x_j \rangle$, ACP classique.

ACP et astuce du noyau (II)

- Sortie du solveur : $\rightarrow$ r vecteurs propres $(c_1, \ldots, c_r)$.
- Problème initial : $(a_1, \ldots, a_r)$? cf ACP grande dimension,

$$a_k = \frac{1}{n\lambda}(X^\top c_k) = \frac{1}{n\lambda}\sum_{i=1}^{n} c_{i,k}\Phi(x_i)$$

- Normalisation : $1 = \|a_k\|^2 = \frac{1}{n^2\lambda^2} c_k^\top K c_k$.
- **Réduction de dimension** : projection sur les a_k.

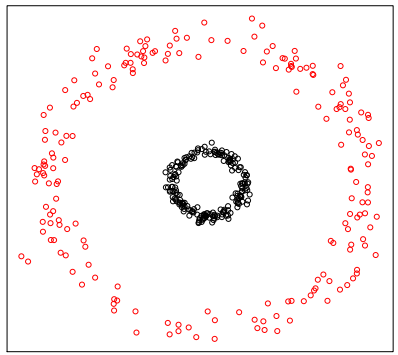
$$\Phi(x_\ell)^\top a_k = \frac{1}{n\lambda_k}\Phi(x_\ell)^\top \sum_i c_{i,m}\Phi(x_i) = \frac{1}{n\lambda_k}[Kc_k]_\ell.$$

Seule l'expression de $k(x_i, x_j)$ est nécessaire, (pas Φ) !

Exemple : KPCA avec le package `kernlab`

(génération des données :)

```
set.seed(42); N = 200
theta = runif(2*N)*2*pi;
R = exp( c( rnorm(N, 0.5, 0.1), rnorm(N, 2, 0.07)))
X = cbind(R * cos(theta), R * sin(theta))
labels=c(rep(1,N), rep(2,N))
plot(X,col = labels, xaxt="n", yaxt="n", ann=FALSE)
```



Séparation non linéaire

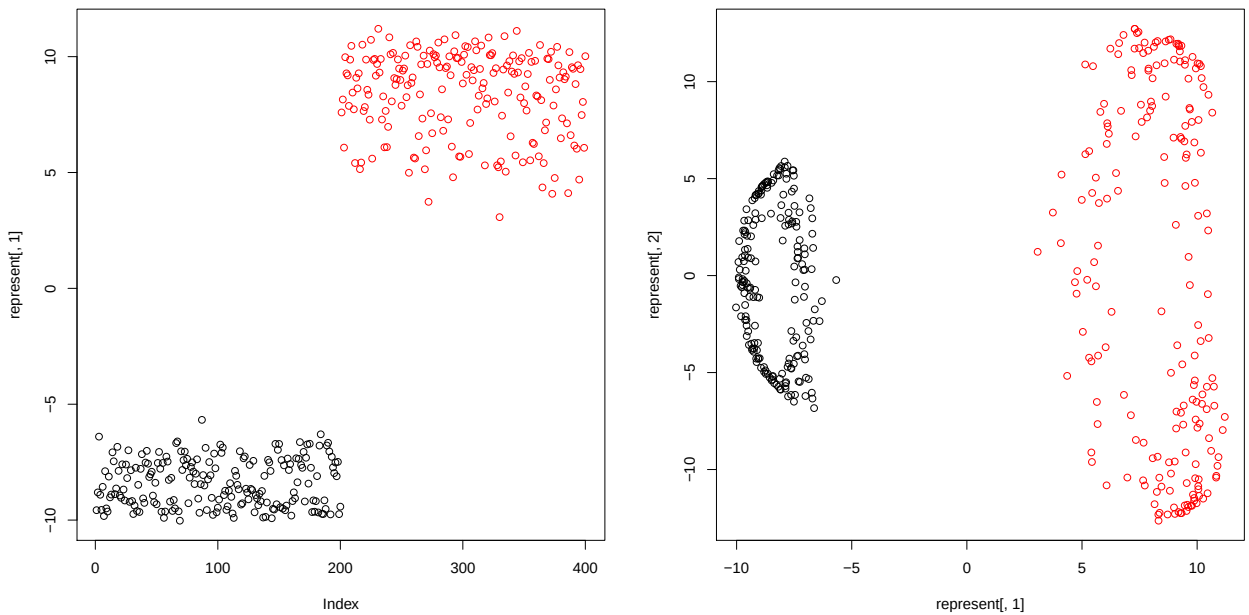
```
#ici: calcul de 3 vecteurs propres de la matrice de Gram K
Kcircle=kpca(X, kernel="rbfdot", kpar=list(sigma=1/30),
             features=3)

## projection des donnees sur les axes principaux de K
represent = rotated(Kcircle)
```

- L'objet `Kcircle` est de classe (S4) `kpca`.
- ses slots principales :
  - `Kcircle@eig` : valeurs propres de la matrice de Gram
  - `Kcircle@pcv` : matrice des vecteurs propres normalisés, ($\|v_i\|^2 = \lambda_i$, $i = 1, \ldots, M$).
  - `Kcircle@rotated` : projection des données sur les vecteurs propres (dans l'espace des features)

Projections en dimension 1 et 2

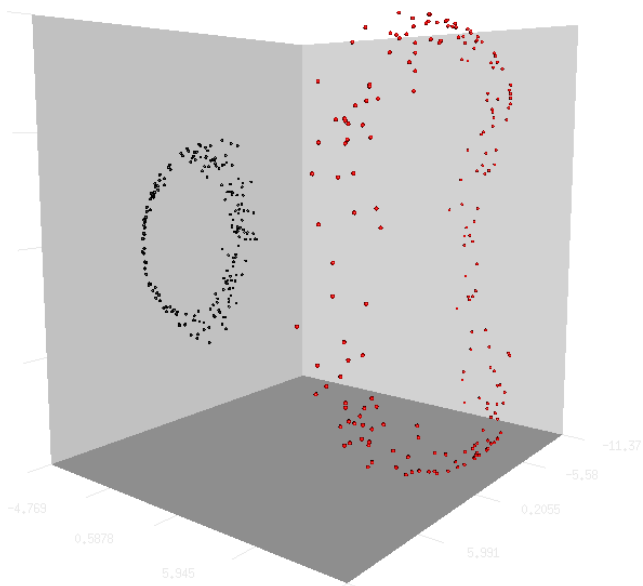
```
par(mfrow=c(1,2))
plot(represent[,1], col= labels)
plot(represent[,1],represent[,2], col= labels)
```



Dans l'espace des features, les données sont séparables linéairement, selon la première composante principale.

Projection en dimension 3

```
rgl.open()
rgl.spheres(x=represent[,1], y=represent[,2],
            z=represent[,3],
            color =labels,radius=0.1)
rgl.bbox(color = "lightgray",alpha=0.5)
```



Motivation

- Jusqu'ici on a utilisé :
  - la métrique euclidienne sur l'espace des individus x_i (produit scalaire $\langle x_i, x_j \rangle = x_i^\top x_j$)
  - Des poids uniformes $1/n$ pour chaque individu.
- Pour l'analyse de données discrètes, on manipulera des tables de contingences, *i.e.* tableaux du nombre de co-occurrences des différentes modalités de deux variables $\mathcal{X}_1, \mathcal{X}_2$.

$$N = (n_{ij})_{i \leq m_1, j \leq m_2} \; ; n_{ij} = \sum_{i=1}^{n} \mathbb{1}\{\mathcal{X}_1 = i, \mathcal{X}_2 = j\}.$$

- Il sera nécessaire de pondérer les lignes/ colonnes et d'utiliser une métrique appropriée.

Métriques et poids

- $X \in \mathbb{R}^{n \times p}$: données brutes.

- Choix d'une métrique $d(x_i, x_j)$ sur $\mathbb{R}^p$ (les individus) $\iff$ choix d'une matrice M symétrique, définie positive (produit scalaire). Alors

$$d^2(x_i, x_j) = \|x_i - x_j\|^2 = (x_i - x_j)^\top M (x_i - x_j)$$

$$\langle x_i, x_j \rangle = x_i^\top M x_j.$$

- Choix du poids de chaque individu pour les calculs de moyenne et de variance :

$$x_i \leftarrow \text{poids } p_i$$

On note

$$D = \begin{pmatrix} p_1 & 0 & \\ 0 & \ddots & \\ & & p_n \end{pmatrix}$$

Barycentre et inertie avec pondération

- Barycentre des x_i : $g = \sum_{i \leq n} p_i x_i = \mathbf{1}^\top DX$
- Inertie par rapport à g

$$I_g = \sum_{i=1}^{n} p_i \|x_i - g\|^2 = \mathrm{tr}\Big(\sum_{i=1}^{n} p_i (x_i - g)^\top M (x_i - g)\Big)$$

$$= \mathrm{tr}\Big(\sum_{i=1}^{n} M(x_i - g) p_i (x_i - g)^\top\Big) = \mathrm{tr}\Big(M \sum_{i=1}^{n} x_i p_i x_i^\top - M g g^\top\Big)$$

$$= \mathrm{tr}(M(X^\top DX - gg^\top))$$

$$= \mathrm{tr}(MV)$$

Avec

$$V = X^\top DX - gg^\top \text{ (matrice de covariance)}$$

- Dans la suite : on suppose que l'on a re-centré les données

$$\rightarrow g = 0.$$

Problème de l'ACP pondérée

- **But :** trouver un projecteur M-orthogonal $P \in \mathbb{R}^{p \times p}$ tel que l'inertie du nuage projeté $XP^\top$ soit maximale.
- Avec métrique M, un projecteur orthogonal vérifie

$$P^* = P \text{ avec } P^* = M^{-1}P^\top M.$$

- Inertie du nuage projeté ?

$$I_P = \mathrm{tr}(V_P M) = \mathrm{tr}(PVP^\top M) = \mathrm{tr}(PVMP) = \mathrm{tr}(VMP^2) = \mathrm{tr}(VMP).$$

- Projection sur un axe a unitaire :

$$P = aa^\top M \Rightarrow \mathrm{tr}(VMP) = \mathrm{tr}(VMaa^\top M) = a^\top MVMa = \langle a, VMa \rangle$$

- L'axe associé à la plus grande inertie est donc le vecteur propre de VM associé à la plus grande valeur propre.

$$VMa_1 = \lambda \mathbf{a}_1$$

Axes principaux

- VM est auto-adjointe pour la métrique $M \rightarrow$ diagonalisable en base orthonormée $(a_1, \ldots, a_p)$:

$$VMa_k = \lambda_k a_k, \qquad a_k^\top M a_j = \delta_{k,j}.$$

- $(a_1, \ldots, a_p)$: **Axes principaux** (rangés par ordre décroissant des valeurs propres λ_k).

Théorème : solutions de l'ACP pondérée

Le projecteur P_s de rang s maximisant l'inertie projeté I_{P_s} est le projecteur sur les s premiers axes principaux $(a_1, \ldots, a_s)$.

Facteurs principaux

- 'Presque' des axes principaux, à une multiplication par M près.
- But : pouvoir écrire un produit scalaire avec a comme un produit matriciel.
- Par définition : le facteur principal u_k est

$$u_k = Ma_k.$$

- u_k est (encore) un vecteur propre :

$$MVu = MV(Ma) = M(VMa) = \lambda Ma = \lambda u.$$

Composantes principales

- La k^e composante principale c_k est le vecteur des coordonnées sur a_k des projetés de x_i sur a_k. Ainsi $c_{i,k} = x_i^\top M a_k = x_i^\top u_k$. D'où

$$c_k = XMa_k = Xu_k.$$

- Inertie (variance) de la composante principale c_k :

$$V(c_k) = c_k^\top D c_k = a^\top M X^\top D X M a = a^\top M V M a = \lambda$$

- composantes principales = vecteurs propres de $XMX^\top D$:

$$XMX^\top Dc = XMX^\top DXu = XMVu = \lambda Xu = \lambda c$$

Axes, facteurs, composantes : formules de transition

- Équations aux valeurs propres

$$\begin{cases} (\text{composantes } x) & XMX^\top Dc = \lambda c \\ (\text{axes } a\ ) & X^\top DXMa = VMa = \lambda a \\ (\text{facteurs } u) & MVu = \lambda u \end{cases}$$

- Formules de transition

$$\begin{cases} c = XMa \\ a = \dfrac{1}{\lambda} X^\top Dc. \end{cases}$$

(deuxième formule : admise, obtenue par multiplication à gauche de l'équation aux valeurs propres pour c.)

ACP et factorisation de matrice

(On suppose X de rang r)

- Comme dans le cas standard, on a pour tout $i \leq n$
  $x_i = \sum_{k=1}^{r} \langle x, a_k \rangle a_k$

- De plus $\langle x_i, a_k \rangle = c_{i,k} (= x_i^\top M a_k = x_i^\top u_k)$.

- On a montré :

$$X = \sum_{k=1}^{r} c_k u_k^\top M^{-1} = (c_1 | \dots | c_r) \begin{pmatrix} a_1^\top \\ \vdots \\ a_r^\top \end{pmatrix} = C A^\top$$

- Cette factorisation est souvent appelée 'formule de reconstitution'.

ACP et maximisation des corrélations au carré.

- Cas de colonnes centrées
- standardisation des colonnes $\iff$ M matrice diagonale, d'entrées

$$M_{jj} = 1/\mathrm{var}(X^j) = \frac{1}{(X^j)^\top D X^j} = \frac{1}{\sum_i p_i (x_i^j)^2}$$

- Corrélation de X^j avec un vecteur c :

$$R(X^j, c) = \langle X_j, c \rangle / \sqrt{V(X_j) V(c)}.$$

Théorème

Dans le cas standard, c_1 est la combinaison linéaire des variables qui maximise la somme des carrés des corrélation avec les variables,

$$c_1 = \underset{c \in \mathrm{vect}(X^1 \ldots, X^d)}{\mathrm{argmax}} \sum_{j=1}^{p} R^2(X^j, c).$$

Représentations 2D

- Cercle des corrélations : les corrélations de chaque variable X^j avec les deux premières composantes principales,

$$\cos\theta_{j,k} = \frac{\langle X^j, c_k \rangle}{\sqrt{(X^j)^\top D X^j \times c_k^\top D c_k}},$$

  $j = 1, \ldots, p$, $k = 1, 2$, avec $\cos\theta_{j,k}$ l'angle entre la variable j et la composante k.

- les projections de chaque point sur les deux premiers axes principaux, c'est-à-dire les

$$(c_{i,1}, c_{i,2}), i = 1 \ldots n$$

Contribution d'un individu à un axe

- Comme $\lambda = \sum p_i c_i^2 = \sum p_i \langle x_i, a \rangle$ on appelle contribution à la composante c de l'individu i le ratio

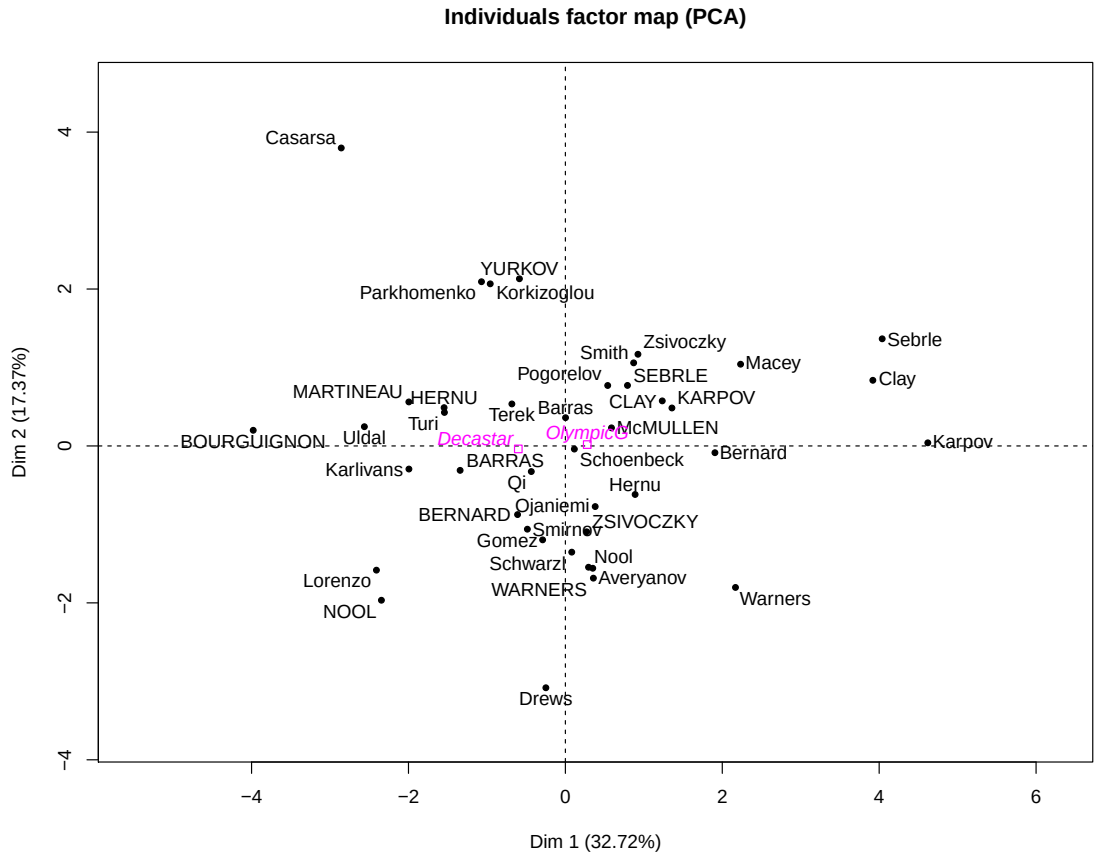
$$CTR(i, c) = \frac{p_i c_i^2}{\lambda}.$$

- Les individus ayant une très forte contribution sont potentiellement des outliers à retirer de l'analyse.

Exemple : avec FactoMineR

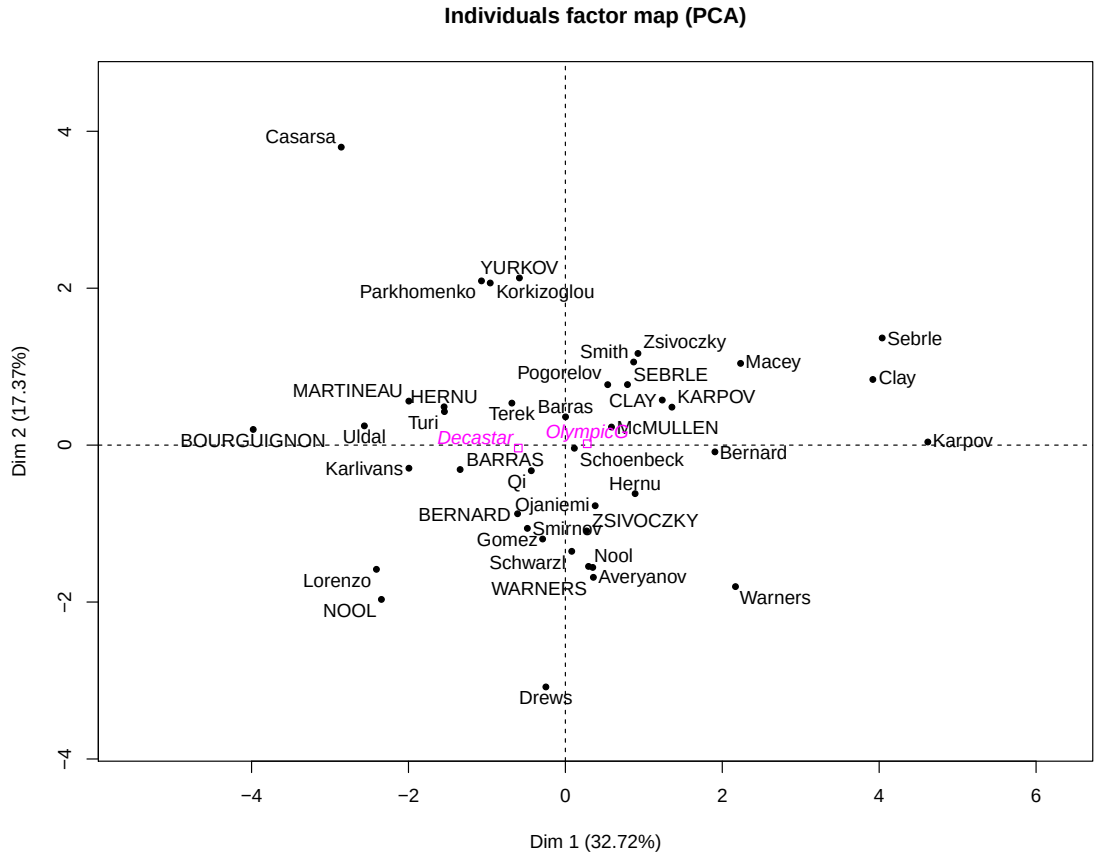
```
data("decathlon")
resPCA <- PCA(decathlon, quanti.sup = c(11,12),
              quali.sup = 13)
## Colonnes 11, 12: quantitatives, (scores, rang)
## Colonne 13: qualitative, evenement (decastar/olympique)
plot(resPCA)
```

Décathlon : carte des individus



- Paires d'individus représentatifs des deux premiers axes ?

Décathlon : carte des individus



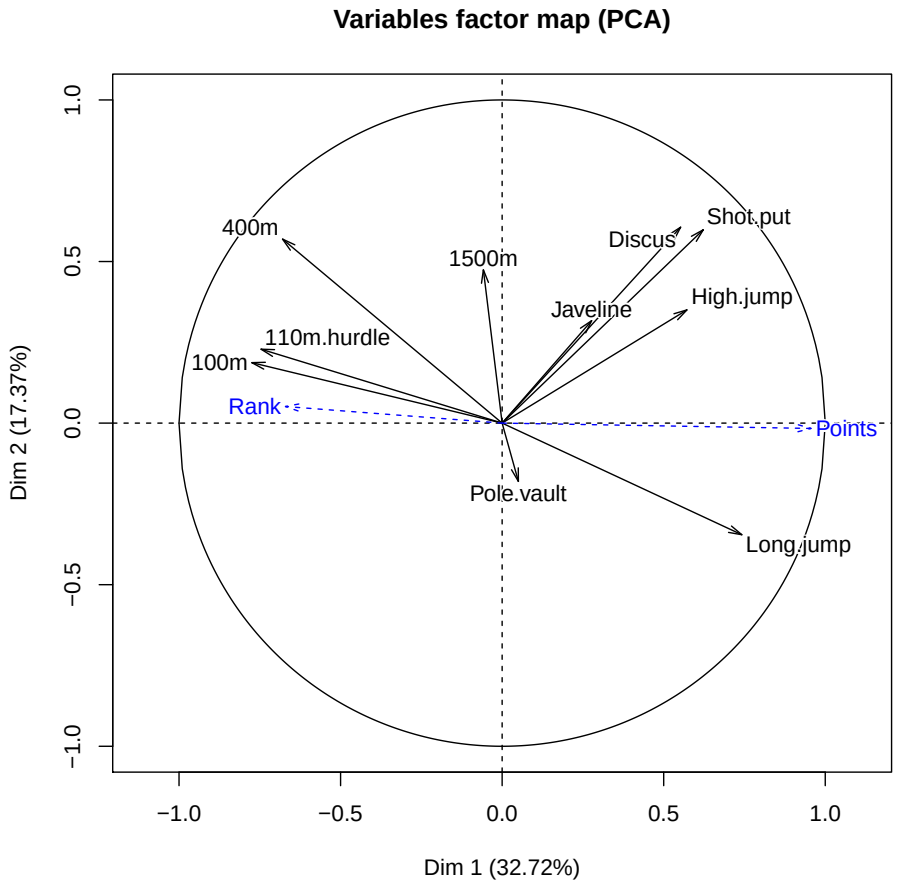
- Paires d'individus représentatifs des deux premiers axes ?
- Axe 1 : Karpov/BOURGUIGNON, Axe 2 : YURKOV/Drews.

Axe 1 : interprétation ?

```
decathlon[c(13,16),]
            100m Long.jump Shot.put High.jump 400m 110m.hurdle Discus
BOURGUIGNON 11.36    6.80    13.46     1.86 51.16      15.67  40.49
Karpov      10.50    7.81    15.93     2.09 46.81      13.97  51.65
            Pole.vault Javeline 1500m Rank Points Competition
BOURGUIGNON       5.02    54.68 291.70  13   7313   Decastar
Karpov            4.60    55.54 278.11   3   8725   OlympicG
```

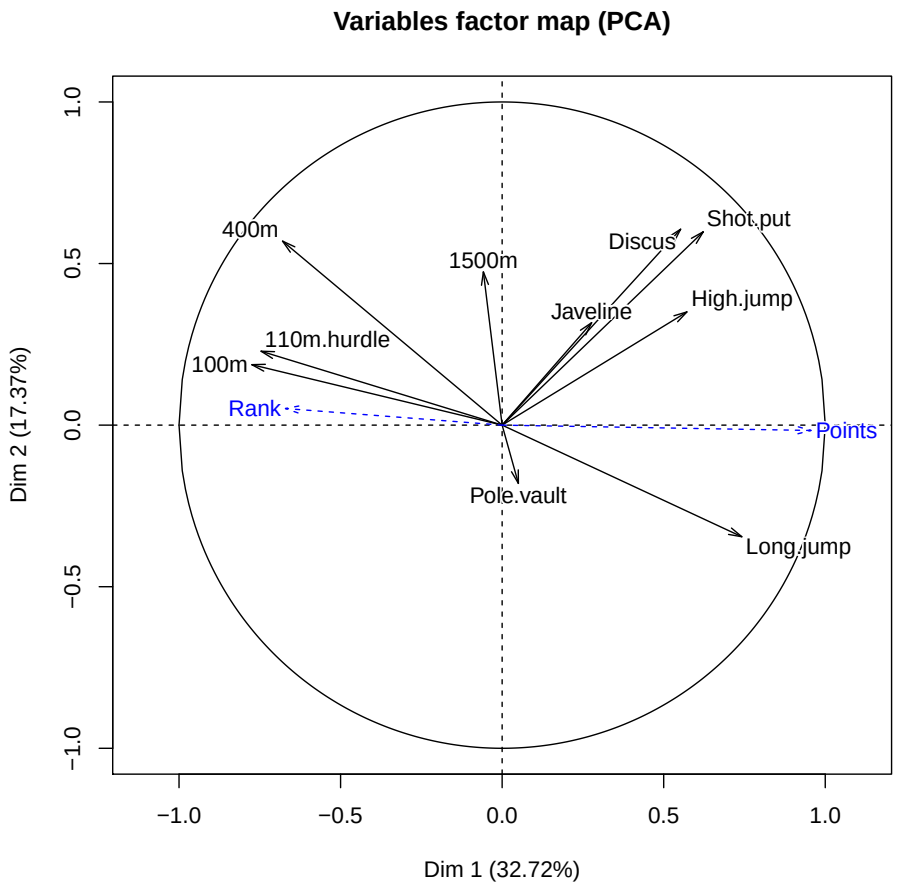
- Individus à l'opposé sur l'axe : performances globalement bonnes contre globalement (comparativement) médiocre.
- Axe 1 : performance globale.

Décathlon : carte des variables (cercle des corrélations)



Quelles sont les variables constitutives de l'axe 1 ? de l'axe 2 ?

Décathlon : carte des variables (cercle des corrélations)

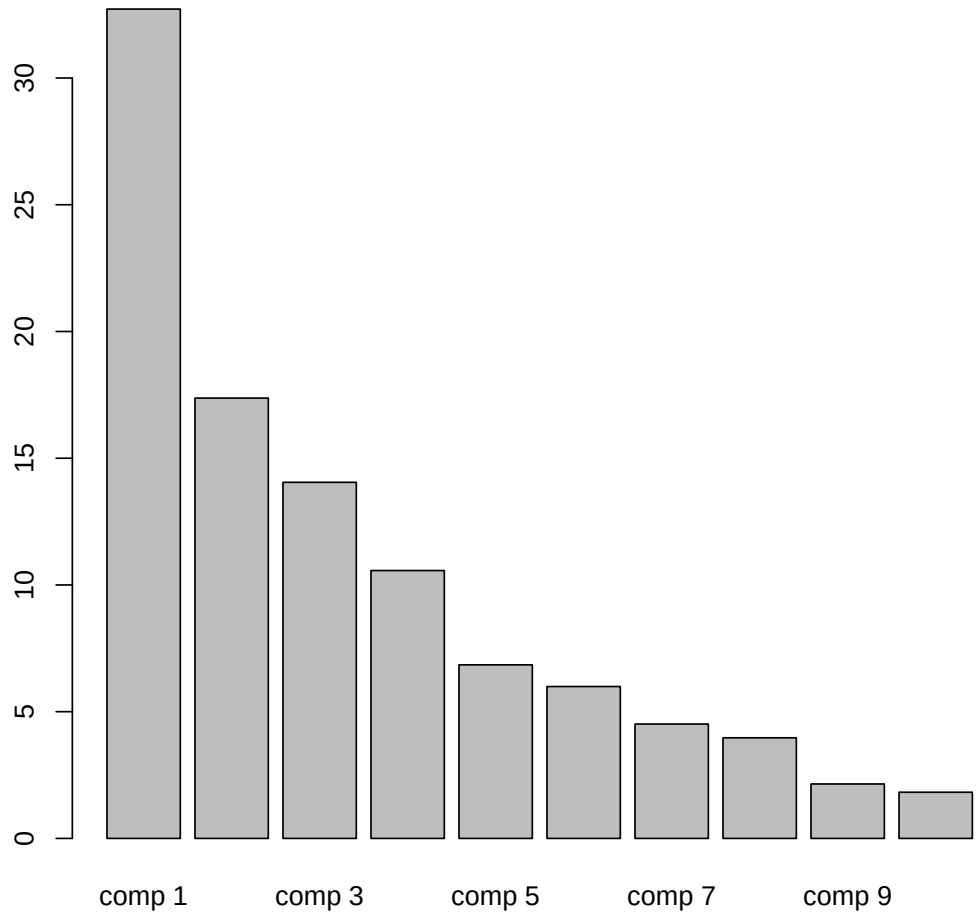


Quelles sont les variables constitutives de l'axe 1 ? de l'axe 2 ?

- 100m (temps) / saut en longueur et saut de perche/ 1500m
- Axe 1 : performance ; Axe 2 : endurance contre explosivité.

Nombre d'axes à considérer

```
> barplot(resPCA $ eig[,2])
```



Variance cumulée

```
> resPCA$eig[,3]
   comp 1    comp 2    comp 3    comp 4    comp 5    comp 6    comp 7    comp 8
 32.71906  50.09037  64.13953  74.70804  81.55577  87.54846  92.06081  96.02958
   comp 9   comp 10
 98.17773 100.00000
```

- Les 2 premiers composants n'expliquent que 50% de la variance.
- Avec 4 composants : on a presque 75%.

Bilan

- Variables supplémentaires : pas utilisées pour construire les axes mais peuvent servir à leur interprétation.

- L'interprétation des axes se fait en comparant le nuage des individus et des variables.

Variable catégorielle : test du χ^2

- X une variable discrète prenant r valeurs possibles $\{1, \ldots\}$. Soit $X^j = \mathbb{1}\{X = j\}$.
- On note $p_j = \mathbb{P}(X = j)$.
- Échantillon iid $(X_1, \ldots, X_n)$, $N_j = \sum X_i^j$.

Théorème

$$\sum_{i=1}^{n} \sum_{j=1}^{r} \frac{(N_j - np_j)^2}{np_j} \xrightarrow[n\to\infty]{loi} \chi^2_{(r-1)}$$

- Intérêt : permet de tester si X suit la loi discrète $(p_1, \ldots, p_r)$.
- Outils pour la preuve : théorème central limite et propriétés des carrés de lois gaussiennes.

Paire de variables catégorielles

- Couple de variables (X, Y), à valeurs dans $\{1, \dots, m_1\} \times \{1, \dots, m_2\}$
- Un échantillon de taille N peut être résumé par une table de contingence $(n_{i,j})_{i \leq m_1, j \leq m_2}$ avec $n_{ij} = \sum_{t=1}^{n} \mathbb{1}\{X_t = i, Y_t = j\}$.
- la somme des lignes et des colonnes vaut $n \Rightarrow$ le tableau de contingences a $(r-1)(s-1)$ degrés de liberté.
- On note $n_{i,j} = \sum_j n_{i,j}$ et $n_{.,j} = \sum_i n_{i,j}$.

Test du χ^2 d'indépendance

- Si les variables sont indépendantes, On a

$$\mathbb{P}(X = i, Y = j) = \mathbb{P}(X = i)\mathbb{P}(Y = j)$$

donc on doit avoir

$$\frac{n_{ij}}{n} \approx \frac{n_{i,}n_{,j}}{n^2}.$$

Théorème

Sous l'hypothèse d'indépendance

$$\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{n}{n_{i,.}n_{.,j}}\left(n_{ij} - \frac{n_{i,.}n_{.,j}}{n}\right)^2 \xrightarrow{loi} \chi^2_{(r-1)(s-1)}$$

- La quantité ci-dessus peut être utilisée comme statistique de test lorsque n est suffisamment grand devant la dimension (typiquement $n \geq 20$ et $n_{ij} \geq 5$ dans 80% des cas)

Exemple : couleur des yeux et des cheveux

	Brown	Blue	Hazel	Green
Black	68	20	15	5
Brown	119	84	54	29
Red	26	17	14	14
Blond	7	94	10	16

- Valeur de la statistique de test : 138.3
- p-valeur : $2e-25$ ...
- conclusion : l'hypothèse d'indépendance est largement rejetée, il y a une dépendance entre les deux variables.

- Étape suivante : comment décrire cette dépendance ?

Analyse (factorielle) des correspondances (AC)

- **Cadre :** Deux variables catégorielles X, Y.
- On considère le tableau de contingences $N \in \mathbb{R}^{m_1 \times m_2}$.
- **Buts :**
  - Savoir quels couples de catégories sont liés (ou au contraire plutôt incompatibles)
  - Représenter les 'profils' de catégories, c'est-à-dire les contributions relatives des catégories de la seconde variable aux différentes catégories de la première.

Profils lignes et colonnes

- Le i^e profil ligne est la distribution empirique de Y sachant $X = i$, *i.e.* le vecteur des contributions relatives des catégories de Y à la ligne i de N,

$$\mathcal{L}_i = \frac{1}{n_{i\cdot}}(n_{i,1}, \ldots, n_{i,m_2}) = \frac{1}{n_{i\cdot}} N_{i,\cdot}.$$

- On note $D_1 = \mathrm{diag}(N\,\mathbf{1})$ la matrice diagonale formée des sommes des lignes de N. Le tableau des profils lignes est alors

$$N = D_1^{-1} N.$$

  Chaque ligne appartient au simplexe unité de $\mathbb{R}^{m_2}$.

- On définit de même les profils colonnes en remplaçant N par $N^\top$.

$$D_2 = \mathrm{diag}(N\,\mathbf{1})\,; \quad \mathcal{C} = D_2^{-1} N^\top$$

Exemple : Hair-Eye

- Profils lignes

	Brown	Blue	Hazel	Green
Black	0.63	0.19	0.14	0.05
Brown	0.42	0.29	0.19	0.10
Red	0.37	0.24	0.20	0.20
Blond	0.06	0.74	0.08	0.13

- Profils colonnes

	Brown	Blue	Hazel	Green
Black	0.31	0.09	0.16	0.08
Brown	0.54	0.39	0.58	0.45
Red	0.12	0.08	0.15	0.22
Blond	0.03	0.44	0.11	0.25

Poids et métrique pour les lignes

- Chaque ligne i a pour poids $p_i = \frac{n_{i\cdot}}{n}$ (fréquence marginale de la modalité i)

$$n^{-1}D_1 : \text{ matrice des poids pour les lignes}$$

- Distance entre les lignes i, k : **métrique du** χ^2.

$$d(\ell_i, \ell_k) = \sum_{j=1}^{m_2} \frac{n}{n_{\cdot j}} (\ell_{i,j} - \ell_{k,j})^2 = \ell_i^\top n D_2^{-1} \ell_k.$$

  - Matrice de la métrique du χ^2 : $M = nD_2^{-1}$.
  - Produit scalaire : $\langle \ell_i, \ell_k \rangle = \ell_i^\top n D_2^{-1} \ell_k$
  - Pondération de chaque modalité j par l'inverse de sa fréquence marginale
  - Les catégories rares des colonnes sont les plus importantes pour comparer les profils lignes.
  - On ne change pas la distance entre les lignes en regroupant deux profils colonnes identiques en une seule.

Poids et métrique pour les colonnes

- Chaque **colonne** j a pour poids $p_j = \frac{n_{\cdot,j}}{n}$ (fréquence marginale de la modalité j)

$$n^{-1}D_2 : \text{ matrice des poids pour les } \textbf{colonnes}$$

- Distance entre les colonnes j, k : **métrique du** χ^2.

$$d(\mathcal{C}_j, \mathcal{C}_k) = \sum_{i=1}^{m_1} \frac{n}{n_{i,\cdot}} (\mathcal{C}_{i,j} - \mathcal{C}_{i,k})^2 = \mathcal{C}_i^\top n D_1^{-1} \mathcal{C}_k.$$

  - Matrice de la métrique du χ^2 : $M = nD_1^{-1}$.
  - Produit scalaire : $\langle \mathcal{C}_j, \mathcal{C}_k \rangle = \mathcal{C}_j^\top n D_1^{-1} \mathcal{C}_k$
  - Pondération de chaque modalité i par l'inverse de sa fréquence marginale

Dualité ligne/colonne

- La matrice des poids des lignes est l'inverse de la métrique sur les colonnes.
- (la matrice de poids des colonnes est l'inverse de la métrique sur les lignes.)

Barycentre et inertie (lignes)

- barycentre du nuage de lignes ('profil marginal') :

$$g_1 = \sum_i p_i \ell_i = \sum_i \frac{n_{i\cdot}}{n} \frac{1}{n_{i\cdot}} N_{i,\cdot} = \frac{1}{n} \sum_i N_{i,\cdot} = \frac{1}{n}(n_{\cdot 1}, \ldots, n_{\cdot, m_2})$$

  D'où $D_2 = n\mathrm{diag}(g_1)$.

- Indépendance exacte : $n_{ij} = n_{i\cdot} n_{j\cdot}/n$ donc chaque ligne égale $(n_{\cdot 1}/n, \ldots, n_{\cdot m_2}/n) = g_1$. L'inertie des profils lignes et nulle.

- En général l'inertie du nuage de lignes par rapport à g_1 est

$$J_1 = \sum_{i=1}^{m_1} \frac{n_{i\cdot}}{n} \sum_{j=1}^{m_2} \frac{n}{n_{\cdot j}} \left(\frac{n_{ij}}{n_{i\cdot}} - \frac{n_{\cdot j}}{n} \right)^2$$

$$= \frac{1}{n} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \frac{\left(n_{ij} - \frac{n_{\cdot j} n_{i\cdot}}{n} \right)^2}{\frac{n_{i\cdot} n_{\cdot j}}{n}}$$

  On reconnaît la statistique du χ^2 (divisée par n).

ACP : Centrer ou ne pas centrer ? (les lignes)

- On fait l'ACP de $X = D_1^{-1}N$, avec poids $D = \frac{1}{n}D_1$ et métrique $M = nD_2^{-1}$.
- La matrice à diagonaliser est

$$VM = X^\top DXM - g_1 g_1^\top M$$

- En pratique on travaillera avec $X^\top DXM$ plutôt que VM car ...

ACP : Centrer ou ne pas centrer ? (les lignes)

Théorème

Les vecteurs et valeurs propres de VM *sont les mêmes que ceux de* $X^\top DXM$*, à l'exception de* g_1 *qui est à la fois*

- *vecteur propre de* VM *avec valeur propre* 0 *(dans le noyau de* VM*)*
- *vecteur propre de* $X^\top DXM$ *avec valeur propre* 1*.*

- **Conclusion** : les axes principaux d'inertie sont donnés par les vecteurs propres de $X^\top DXM$ (ACP non centrée) et les valeurs propres associées donnent l'inertie de la projection du nuage sur chaque axe, à l'exception de $\lambda = 1$ associée à g qu'il faut enlever de l'analyse.

- Dans la suite on note $V = X^\top DXM$, on n'oubliera pas de remplacer la valeur propre $\lambda = 1$ par $\lambda = 0$ à la fin.

ACP des profils lignes

Rappel : $X = D_1^{-1}N$, avec poids $D = \frac{1}{n}D_1$ et métrique $M = nD_2^{-1}$.

- Les **composantes principales** c sont les vecteurs propres de

$$XMX^\top D = D_1^{-1}NnD_2^{-1}N^\top D_1^{-1}D_1/n,$$

ie. ils vérifient

$$D_1^{-1}ND_2^{-1}N^\top \, c = \lambda \, c. \tag{2}$$

- Les **axes principaux** a sont vecteurs propres de $VM = N^\top D_1^{-1}(D_1/n)D_1^{-1}NnD_2^{-1}$, avec les mêmes valeurs propres,

$$N^\top D_1^{-1}ND_2^{-1} \, a = \lambda \, a. \tag{3}$$

- les **facteurs principaux** u sont les vp. de MV,

$$D_2^{-1}N^\top D_1^{-1}N \, u = \lambda \, u. \tag{4}$$

Précisions

- **normalisation** On choisit $\|a\|_M = 1$, ce qui impose $\|c\|_D^2 = \lambda$ (cf. ACP), ie.

$$\begin{aligned} n a^\top D_2^{-1} a &= 1 \\ n^{-1} c^\top D_1 c &= \lambda. \end{aligned} \tag{5}$$

- **Premier triplet axe, composante, facteur**
  g_1 est axe principal de $X^\top DXM$ associé à la vp 1. Pour cette première composante de l'analyse on a
  - $a = g$
  - $c = XMa = D^{-1}NnD_2^{-1}g = D^{-1}N\mathbf{1} = \mathbf{1}$.
  - $u = Ma = Mg = \mathbf{1}$.

ACP non centrée des profils colonnes

- On note respectivement (γ, α, v) les (composantes, axes et facteurs) principaux.

- Par la symétrie mentionnée plus haut on a directement

$$\begin{cases} \text{composantes} & D_2^{-1} N^\top D_1^{-1} N \, \gamma = \lambda \, \gamma \\ \text{facteurs} & D_1^{-1} N D_2^{-1} N^\top \, v = \lambda \, v \\ \text{axes} & \alpha = M^{-1} v = n D_1^{-1} v \\ \text{normalisation} & n \alpha^\top D_1^{-1} \alpha = 1 \; ; \quad n^{-1} \gamma^\top D_2 \gamma = \lambda. \end{cases}$$

d'où (raisonnement sur les normes pour trouver les constantes de proportionnalité)

$$\begin{cases} \gamma = \sqrt{\lambda} u, \\ c = \sqrt{\lambda} v. \end{cases}$$

Formules de transition

- Appelées 'propriétés barycentriques' des facteurs et composantes : les coordonnées des composantes principales des lignes sont les barycentres des composantes principale des colonnes (à $\sqrt{\lambda}$ près).

- En prémultipliant à gauche par $D_2^{-1}N^\top$ et $D_1^{-1}N$ on montre :

$$\begin{cases} \gamma(=\sqrt{\lambda}u) = \dfrac{1}{\sqrt{\lambda}} D_2^{-1} N^\top c \\ c(=\sqrt{\lambda}v) = \dfrac{1}{\sqrt{\lambda}} D_1^{-1} N \gamma \end{cases}$$

Représentation graphique 2D

- Point représentatif de la catégorie i de la variable X : le point $x_i = (c_{1,i}, c_{2,i})$ (projection de l'individu 'profil ligne' sur les deux premiers axes principaux)

- on exclut le premier axe de l'analyse non centrée correspondant au barycentre.
- Les nuages sont centrés :

$$\sum_i p_i c_i = c^\top D_1 / n\mathbf{1} = \langle c, c_{g_1} \rangle_D = 0$$

(les composantes de l'ACP sont D-orthogonales).

Contributions à l'inertie

- Une modalité contribue significativement à l'inertie du nuage projeté sur un axe si sa coordonnée correspondante est grande. En effet

$$\lambda = V(c) = c^\top Dc = c^\top \frac{1}{n} D_1 c = \sum_i \frac{n_{i\cdot}}{n} c_i^2$$

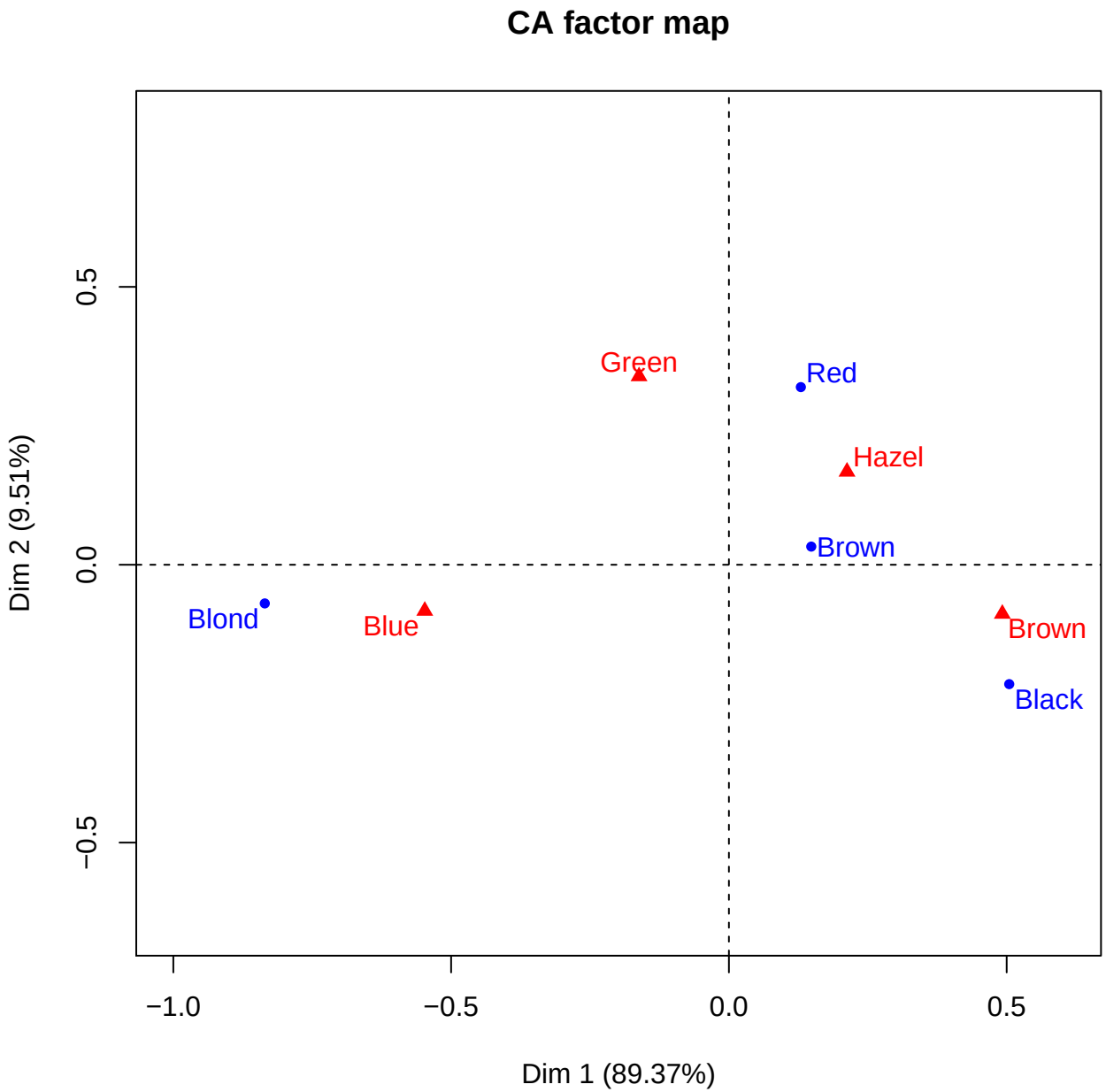
- La contribution du profil ligne i à l'inertie sur la composante c est donc

$$CTR(i) = \frac{\frac{n_{i\cdot}}{n} c_i^2}{\lambda}$$

Représentation jointe lignes/colonnes

- De même on trace souvent les points représentatifs $(\gamma_{i,1}, \gamma_{i,2})$ des colonnes.
- Usage courant : superposer les deux graphiques.
- Les modalités i e X et j de Y sont particulièrement dépendantes si les points représentatifs correspondants sont proches

Exemple


CA factor map
Green
Red
Hazel
Brown
Blond
Blue
Brown
Black
0.5
0.0
−0.5
Dim 2 (9.51%)
−1.0
−0.5
0.0
0.5
Dim 1 (89.37%)

Justification de la représentation jointe

- X_i est liée positivement à Y_j si $\frac{n_{ij}}{n} \gg \frac{n_{i,}\, n_{\cdot,j}}{n}$
- Formule de reconstitution (générale) : $X = \sum_k c_k u_k^{-1} M_{\chi^2}^{-1}$
- dans notre cas ($X = D_1^{-1}N, M = nD_2^{-1}, D = n^{-1}D_1$)

$$\frac{n_{ij}}{n_{i\cdot}} = \sum_k c_{k,i} u_{k,j} \frac{n_{\cdot j}}{n} \qquad (u = \gamma/\sqrt{\lambda})$$

$$= \sum_k \frac{c_{k,i}\gamma_{k,j}}{\sqrt{\lambda_k}}$$

$$\Rightarrow \frac{n_{ij}}{n} = \frac{n_{i,}\, n_{\cdot,j}}{n} \sum_k \frac{c_{k,i}\gamma_{k,j}}{\sqrt{\lambda_k}}$$

Justification de la représentation jointe

- En commençant la numérotation à la deuxième composante (car la première composante correspond au barycentre et le ratio vaut 1) :

$$\frac{n_{ij}}{n} = \frac{n_i n_j}{n}\left(1 + \sum_k \frac{c_{k,i}\gamma_{k,j}}{\sqrt{\lambda_k}}\right)$$

- Avec les notations de la section précédente on a bien

$$\frac{n_{ij}}{n} \approx \frac{n_i n_j}{n}\left(1 + x_{i1}y_{j1}/\sqrt{\lambda}_1 + x_{i1}y_{j1}/\sqrt{\lambda}_2\right).$$

(produit scalaire entre les deux vecteurs représentatifs, à une pondération par les valeurs propres près)

Cadre de l'ACM : AC généralisée pour $p \geq 2$ variables

- On considère p variables qualitatives $\mathcal{X}_1, \ldots \mathcal{X}_p$ et un jeu de données sous forme de tableau disjonctif $X = (X_1|X_2|\ldots|X_p)$. avec $X_j \in \{0,1\}^{n \times m_j}$

- Exemple : Un sondage avec p questions et une réponse possible par question.

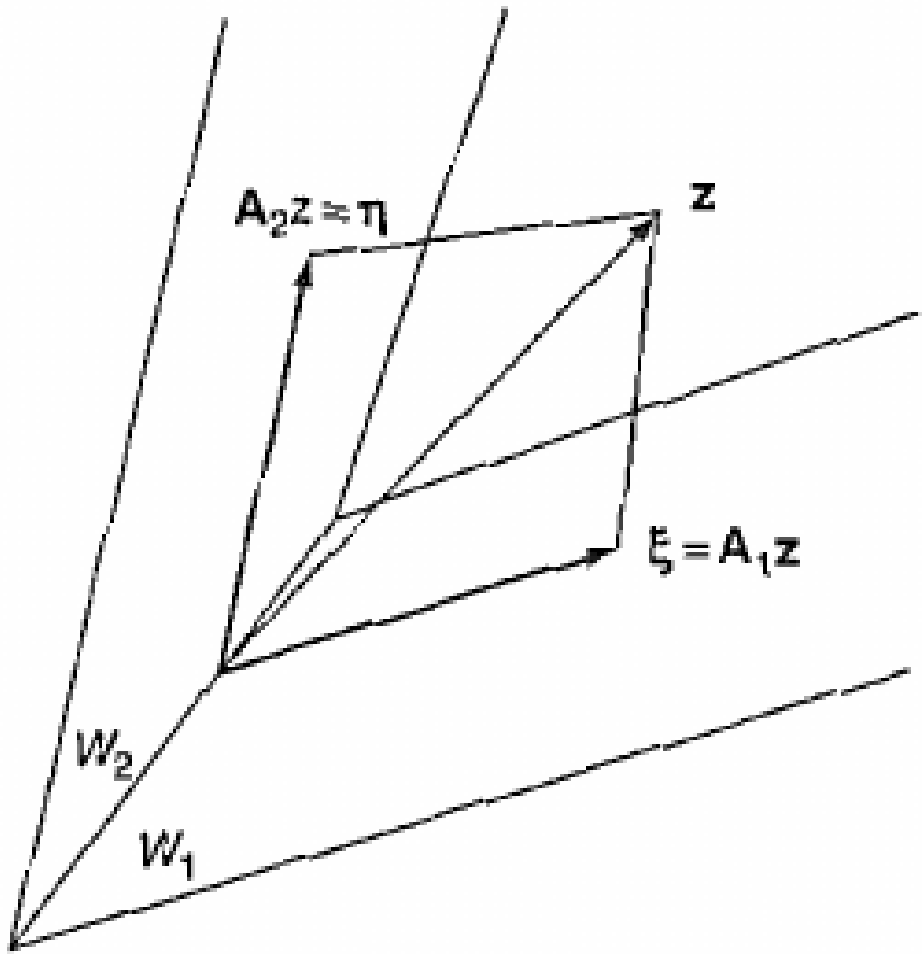
- notation : X_j^m : m^e colonne (binaire) du groupe j (réponses m à la question j) : un vecteur de taille n.

- nombre de colonnes : $\tilde{p} = \sum_{j=1}^{p} m_j$.

- On a $X\mathbf{1}_{\tilde{p}} = p$ et $X_j\mathbf{1}_{m_j} = 1$.

Objectif heuristique

Trouver des axes de projection z dans l'espace des variables ($vect(X_1, \ldots, X_p) \subset \mathbb{R}^n$, espace engendré par les colonnes), maximisant un résumé numérique de la corrélation avec l'ensemble des variables.



Corrélations multiple

- **But heuristique** : maximiser la 'corrélation avec l'ensemble des variables' ?
- Corrélation multiple entre $z \in \mathbb{R}^n$ et l'espace engendré par les colonnes $(X_j^1, \ldots, X_j^{m_j})$ de la variable j :

$$R(z,j) = \frac{\langle c, A_j c \rangle}{\|c\| \cdot \|A_j c\|}$$

  Avec A_j : projecteur orthogonal sur l'espace engendré par les colonnes $(X_j^1, \ldots, X_j^{m_j})$.
- **But formel** : maximiser la somme des corrélations au carré :

$$\max_{z \in \mathbb{R}^n} \sum_{j=1}^{p} R^2(z,j)$$

Solutions et représentations des variables

On montre (cf Saporta, chap. 8) que

- les axes solutions sont les vecteurs propres de $A_1 + \cdots + A_p$ avec A_j le projecteur sur W_j

- Ils sont solutions de $XMX^\top Dz = \mu z$, avec $M = \text{diag}(V_{11}^{-1}, \ldots, V_{pp}^{-1})$, $V_{ii} = X_i^\top D X_i$ et D une matrice (diagonale) de poids quelconque.

- les composantes s'écrivent

$$z = Xb \qquad (b : \text{facteur principal})$$

et b vérifie

$$MVb = \mu b$$

si X est de rang plein, avec μ : valeur propre pour z.

- On peut alors projeter z sur chaque sous espace pour obtenir des composantes principales $\xi_j = A_j z$

Analyse des correspondance formelle du tableau disjonctif

- On 'peut' voir le tableau disjonctif X comme une table de contingence (creuse)
- Somme des lignes $= p$ (1 réponse par question)
- Somme des colonnes $= d_j^m$, $m \leq m_j, j \leq p$.
- Dans le cadre de l'AC, on note $\tilde{n} = \sum_{i,m} X_{i,m} = p * n$,
- $\tilde{D}_1 = \text{diag}(X\mathbf{1}_{\tilde{p}}) = pI_{\tilde{p}}$,
- $\tilde{D}_2 = \text{diag}(\mathbf{1}_n^\top X) = \text{diag}(D_1, \ldots, D_p) := \tilde{D}$.

Profils lignes et colonnes

- profils lignes :
  - matrice de données $\tilde{D}_1^{-1} X = \frac{1}{p} X$,
  - Matrice diagonale des poids $\tilde{D}_1 / \tilde{n} = \frac{1}{n} I_n$,
  - Métrique du χ^2 : $\tilde{n} \tilde{D}_2^{-1} = np \; diag(D_1^{-1}, \ldots, D_p^{-1})$.
- profils colonnes :
  - matrice de données $\tilde{D}_2^{-1} X^\top = diag(D_1^{-1}, \ldots, D_p^{-1}) X^\top$,
  - poids $\tilde{D}_2 / \tilde{n} = \frac{1}{np} diag(D_1, \ldots, D_p) X^\top$,
  - métrique du χ^2 : $np \tilde{D}_1^{-1} = n I_{\tilde{p}}$.

Différence avec l'AC ordinaire : la symétrie lignes/colonnes est rompue.

ACP des profils colonnes

- *cf.* AC : composantes principales=vecteurs propres $\tilde{\gamma}$ de

$$\tilde{D}_2^{-1} X^\top \tilde{D}_1^{-1} X = \frac{1}{p} \tilde{D}^{-1} \tilde{X}^\top \tilde{X} = \frac{1}{p} \begin{pmatrix} I_{m_1} & D_1^{-1} N_{12} & \dots & D_1^{-1} N_{1p} \\ D_2^{-1} N_{21} & I_{m_2} & \dots & D_2^{-1} N_{2p} \\ \vdots & & \ddots & \\ D_p^{-1} N_{p1} & \dots & & I_{m_p} \end{pmatrix}$$

  avec $N_{ij} = (X_i)^\top X_j$ la table de contingence des variables X_i et X_j.

- Cette matrice par blocs est le **Tableau de Burt**

- On prend comme convention de normalisation $\tilde{\gamma} \frac{1}{\tilde{n}} \tilde{D}_2 \gamma = \mu$, ie.

$$\frac{1}{np} \tilde{\gamma}^\top \tilde{D} \gamma = \mu.$$

Compatibilité avec l'AC pour $p = 2$

- Pour $p = 2$ les vecteurs propres du tableau de Burt (a, b) vérifient

$$\begin{cases} a + D_1^{-1} N b & = \mu a \\ D_2^{-1} N^\top a + b & = \mu b \end{cases} \iff \begin{cases} D_1^{-1} N b & = (\mu - 1) a \\ D_2^{-1} N^\top a & = (\mu - 1) b \end{cases}$$

  d'où par substitution

$$\begin{cases} D_2^{-1} N^\top D_1^{-1} N b & = (\mu - 1)^2 b \\ D_1^{-1} N D_2^{-1} N^\top a & = (\mu - 1)^2 a \end{cases}$$

  on retrouve les équations aux valeurs propres de l'AC

- Conséquence : (a, b) est de type (c, γ) où c et γ sont des composantes principales lignes et colonnes de l'AC de la table de contingence N.

ACP des profils lignes

- composantes principales $\tilde{c}$ du nuage des lignes = vecteurs propres de

$$\tilde{D}_1^{-1} X \tilde{D}_2^{-1} X^\top = \frac{1}{p}(X_1|\ldots|X_p) diag(D_1^{-1}, \ldots, D_p^{-1}) \begin{pmatrix} X_1^\top \\ \vdots \\ X_p^\top \end{pmatrix}.$$

- Rappel : $D_j = X_j^\top X_j = V_{jj}$ (j^e bloc de la matrice de covariance).
- En développant, la matrice à diagonaliser s'écrit comme une somme de projecteurs :

$$\frac{1}{p} \sum_{j \leq p} X_j V_{jj}^{-1} X_j^\top = \frac{1}{p} \sum_{j \leq p} A_j.$$

- **Conclusion** : Les composantes principales du nuage de lignes de l'ACM maximisent la somme des $R^2(z, j)$ (problème initial).
- **Conclusion 2** : l'AC formelle du tableau X permet de résoudre le problème de départ.

Sur l'interprétation des résultats de l'ACM

- Représentations graphiques similaires à celles de l'ACP et de l'AC (projection des individus sur les axes principaux),

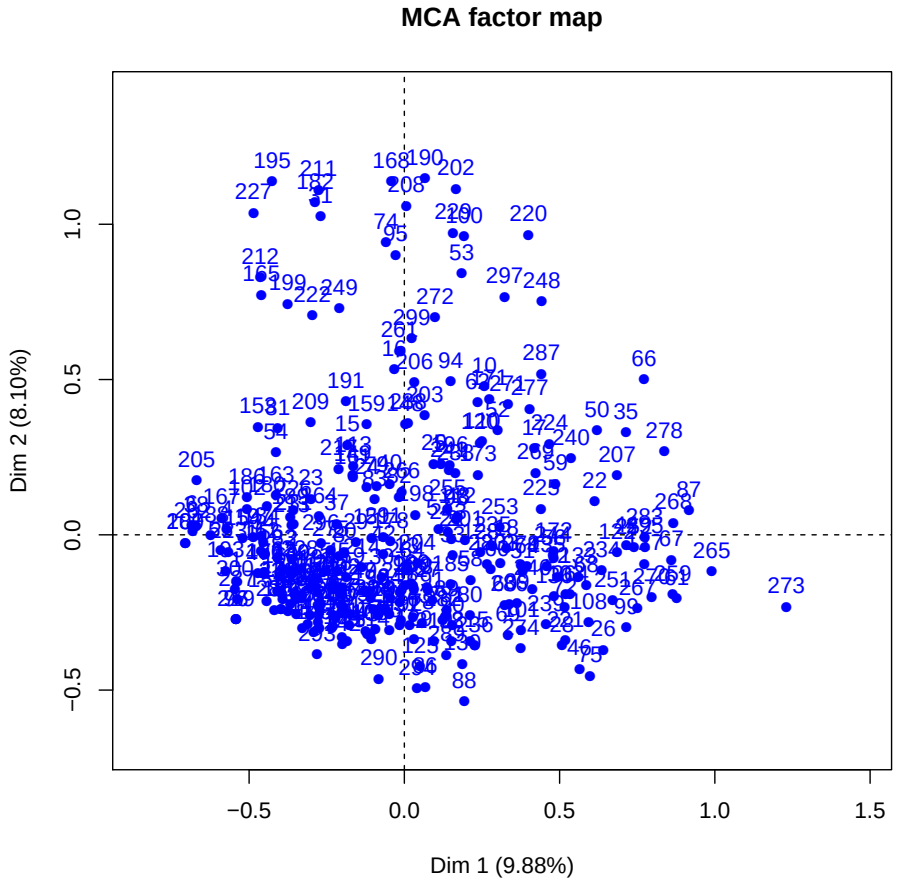
- en plus : une représentation possible des composantes principales 'globales' de chaque groupe (les $(A_i c_1, A_i c_2)$ avec (c_1, c_2) composantes principales du nuage de lignes, A_i projecteur sur le groupe i.

- Attention à l'inertie des lignes : pas d'interprétation 'physique' car

$$I = \sum_k \mu_k = \mathsf{tr}(XMX^\top \tilde{D}) = \mathsf{tr}(\text{tableau de burt}) = \frac{\sum_{j \leq p} m_j}{p}$$

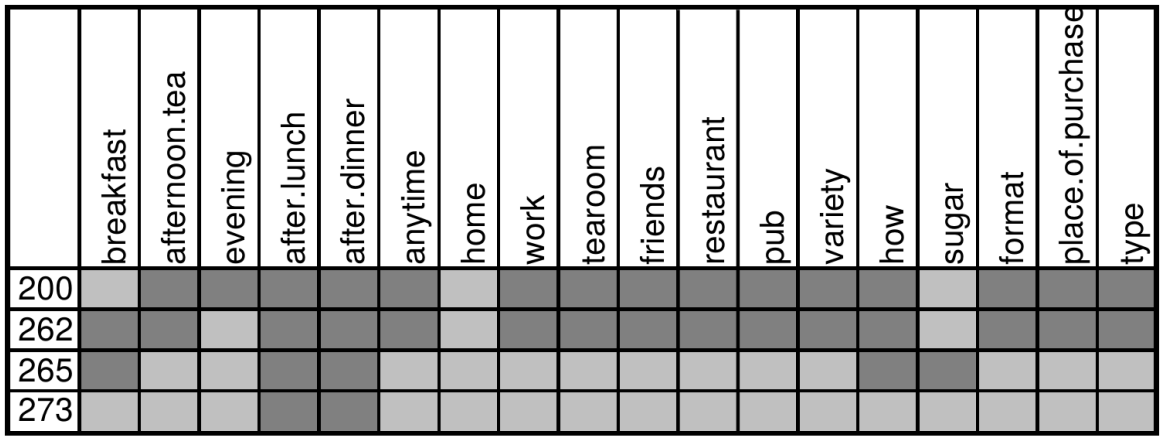
(c'est la moyenne du nombre de catégories par variable)

Tea : carte des individus



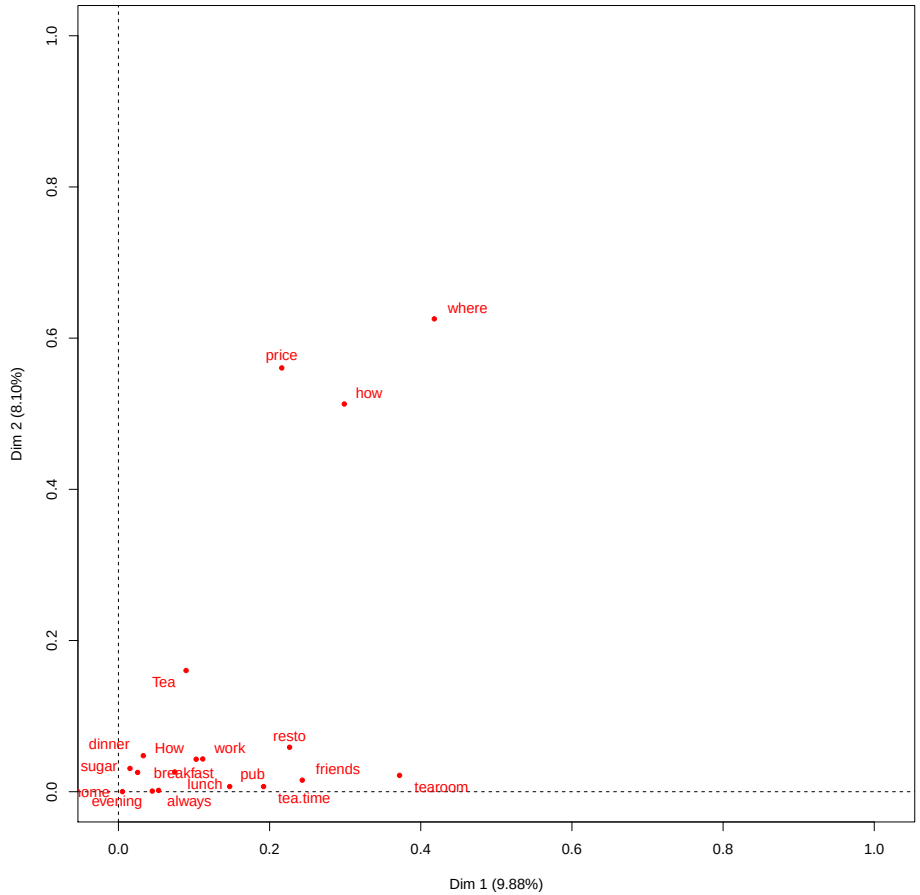
- Pas de structure particulière

Tea : individus proches/opposés sur l'axe 1



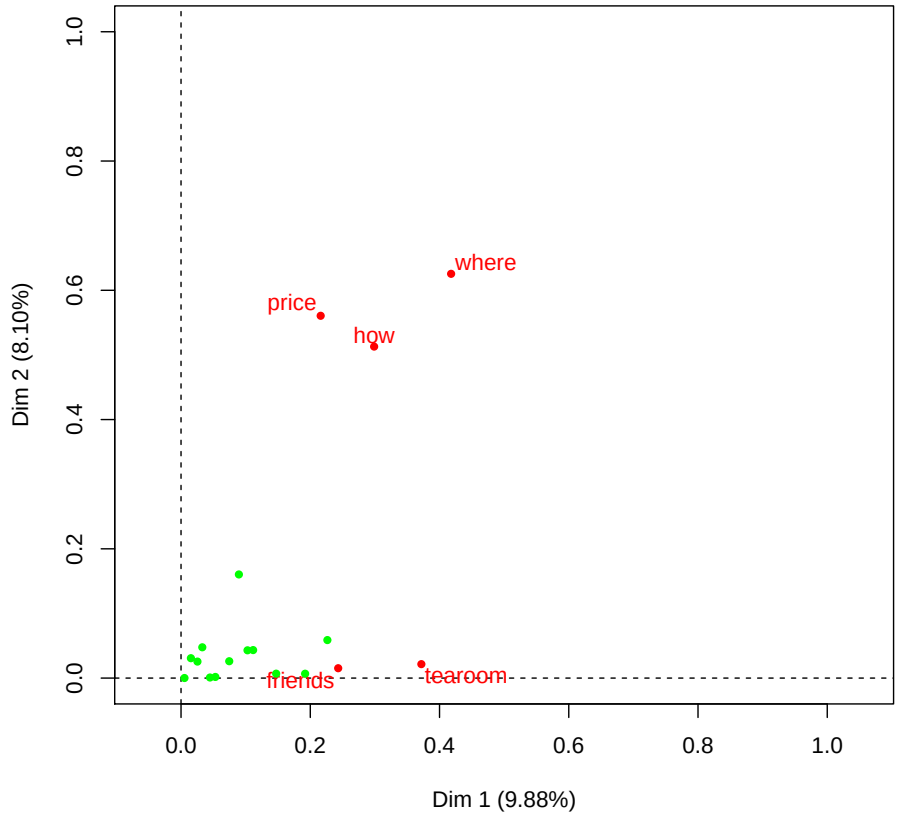
- individus 200, 262 : à l'extremité gauche
- individus 265,273 : à droite
- Variables considérées : avec 2 catégories (oui/non)
- Interprétation de l'axe 1 : Consommation toute la journée (265,273) ou seulement à la maison, le matin et/ou le soir.

Tea : carte des variables (corrélations au carré)



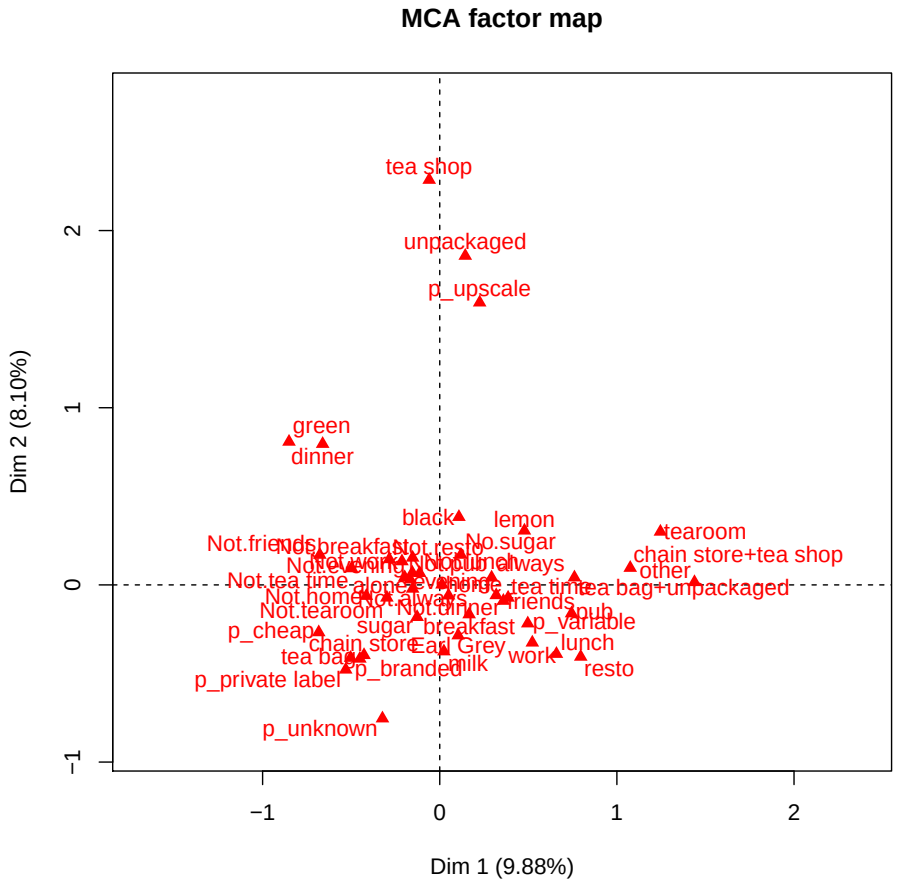
- Corrélations $(R^2(j, z_1), R^2(j, z_2))$
- Beaucoup de variables, Certaines peu corrélées aux composantes principales
- Permet de concentrer l'analyse sur les variables les plus corrélées

Tea : carte des variables (corrélations au carré)



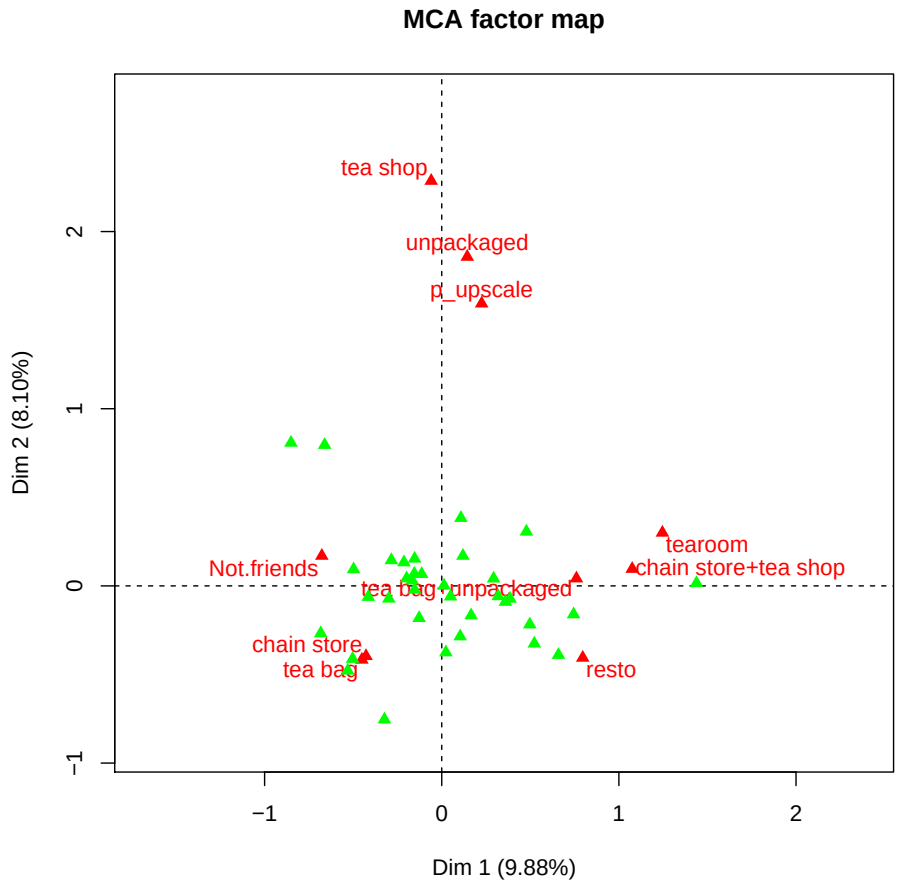
- Corrélations $(R^2(j, z_1), R^2(j, z_2))$
- Beaucoup de variables, Certaines peu corrélées aux composantes principales
- Permet de concentrer l'analyse sur les variables les plus corrélées

Tea : carte des catégories



- Projection des catégories sur les 2 premiers axes
- Sélection possible des catégories selon leurs contributions à l'inertie sur l'axe, (ou leurs corrélations au carré, etc ...)

Tea : carte des catégories



- Projection des catégories sur les 2 premiers axes
- Sélection possible des catégories selon leurs contributions à l'inertie sur l'axe, (ou leurs corrélations au carré, etc ...)

Analyse de panier, règles d'associations

- Cadre : analyse de panier de consommateur.
- Données = tickets de caisse. Lignes = clients, colonnes : chaque produit.
- $X_{i,j} = 1$ si client i a acheté produit j, 0 sinon.
- **but général** : trouver des 'règles d'association' du type
  - Si un client a acheté tel produit, va-t-il probablement en acheter tel autre ?
  - Quels sont les groupes de produits (item-sets) susceptibles d'êtres achetés simultanément par un client ?

Algorithme Apriori (Agrawal, Srikant, 1994.)

- **But** : trouver des 'item-sets' fréquents dans la base de données.
- Problème : pour p produits il y a $2^p - 1$ sous groupes non vides ! !
- Approche de l'algorithme Apriori :
  1. Rechercher les paires les plus fréquentes.
  2. Pour qu'un triplet (i, j, k) soit fréquent, il faut que chacune des paires $(i, j), (i, k), (k, j)$ le soient :
     $\rightarrow$ recherche des triplets fréquents seulement parmi un ensemble restreint de candidats
  3. Même chose avec les quadruplets
  4. ...
  5. On s'arrête aux groupes de taille m lorsqu'on n'a plus de candidats de taille $m + 1$.

Pour aller plus loin : ACP, AC, ACM, Apriori

- Livre : G. Saporta, *Probabilités, Analyse de données et Statistiques*, 2006, Chapitres 6 à 10 → Description mathématique complète et plutôt détaillée

- Package `FactoMineR` :
  `http://factominer.free.fr/index_fr.html`
- Mooc de François Husson (vidéos sur l'analyse de données , sur sa page enseignement)
  → description intuitive, moins de maths, plus d'interprétations, exemples en `R`
- Livre : Husson, Lê, Pagès, *Exploratory Multivariate Analysis by example using R*, 2011.

- Package `ade4` et ressources associées
  `http://pbil.univ-lyon1.fr/ade4/`

- Règles d'association : Vignette du package `R arules`