Cours de Statistiques

Analyse en Composantes Principales

Romain Raveaux[1]

[1]Laboratoire L3I – Université de La Rochelle
romain.raveaux01 at univ-lr.fr

Octobre 24-11, 2008

Sommaire

1. Quelques Rappels
2. Relations entre deux séries de données
   - Relations entre deux séries de données numériques
   - Relations entre deux variables ordinales
3. Analyse en Composantes Principales
   - Introduction
   - Préparation des données
   - Matrice des covariances
   - Recherche des valeurs propres
   - Taux d'Information
4. ACP : une explication géométrique
   - Décripage de la matrice des covariances
   - Recherche des composantes principales
   - Choix des r premières composantes principales
   - Cercle des corrélations
5. Conclusion

Type de variable

Numérique

- Soit l'étude de la variable X, une série de valeurs définies dans $\Re$.
- Exemple: Age, poids,...

Nominale

- Ne prend qu'un nombre limité de valeurs.
- Et que ces valeurs n'ont entre elles aucune relation apparente.
- Exemple : Le statut marital, qui pourrait prendre les valeurs "Célibataire", "Marié", "Veuf", "Divorcé", "Union libre".

Ordinale

- Ne prend qu'un nombre limité de valeurs.
- Et que ces valeurs n'ont entre elles aucune relation apparente.
- Les grades dans l'armée: "lieutenant", "capitaine", "commandant" etc...
- Par nature, les rangs sont des variables ordinales.

Il existe d'autres types de variable : Binaire, Normale,...

Variable et Espace d'étude

Une série à valeurs individuelles

- Soit l'étude de la variable X, une série de valeurs définies dans $\Re$.

Statistiques multi-dimensionnelles

- Soit l'étude d'un ensemble fini de variables (Ω), Ω est l'univers des statistiques.
- Avec $card(\Omega) = M$
- $\Omega = X_1, X_2, ..., X_m$
- $\forall X_i \in \Omega$, X_i est une série à valeurs individuelles.

Estimateurs

Soit l'étude de la variable X, une série de valeurs définies dans $\Re^+$:

Moyenne d'une série à valeurs individuelles

- $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$

Variance d'une série à valeurs individuelles

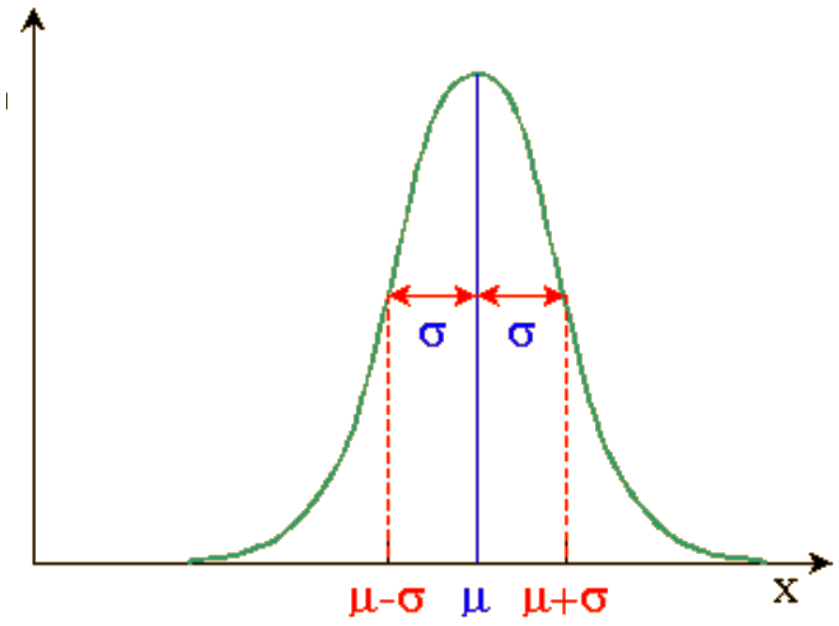
- $\theta^2 = V(X) = \sum_{i=1}^{N}(x_i - \overline{X})^2$
- L'ecart type ce déduit de la variance : $\theta = \sqrt{V(X)}$

Représentation de ces estimateurs

Soit l'étude de la variable X suivant une loi normale ($\aleph(\mu, \theta^2)$), de moyenne μ et de variance θ.

Densité de probabilité d'une loi gaussienne

- $f(x) = \dfrac{1}{\theta\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\theta})^2}$

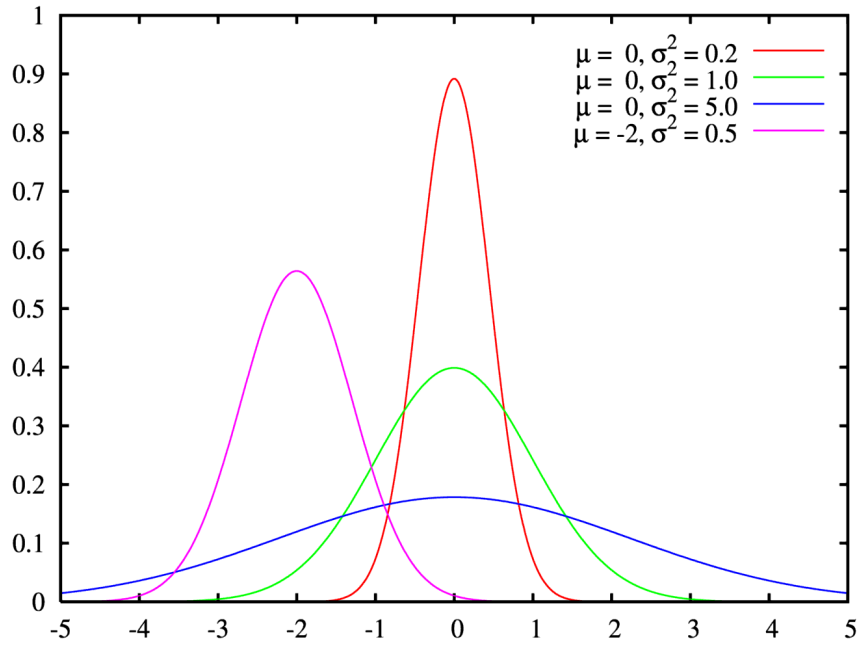


Représentation de ces estimateurs

Soit l'étude de la variable X suivant une loi normale ($\aleph(\mu, \theta^2)$), de moyenne μ et de variance θ.

Densité de probabilité d'une loi gaussienne

- $f(x) = \dfrac{1}{\theta\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\theta})^2}$



Explication intuitive de ces estimateurs

Plus la variance d'un échantillon est grande et plus les données sont éparses. Cela peut dénoter une érreur dans le phénoméne mesuré.

Relations entre deux séries de données observées

Exemples :

- Mesurer le poids ou la longueur d'un organe (variable dépendante) à différentes dates successives choisies arbitrairement (variable indépendante).
- Mesurer le rendement d'une culture (variable dépendante) en fonction de différentes doses d'engrais (variable indépendante).
- Mesurer la capacité à résoudre un problème ou à réaliser une tâche (variable dépendante) en fonction de différentes doses d'un médicament (variable indépendante).

Covariance de deux échantillons

Soit l'étude de deux variables X et Y, deux séries de valeurs définies dans $\Re$:

Covariance

- $\theta_{xy} = cov(X, Y) = \sum_{i=1}^{N}(x_i - \overline{X})(y_i - \overline{Y})$
- La fonction covariance retourne des valeurs comprises dans $[-\infty, +\infty]$
- X et Y indépendant $\Longrightarrow cov(X, Y) = 0$

Covariance de deux échantillons

Covariance

- $\theta_{xy} = cov(X, Y) = \sum_{i=1}^{N}(x_i - \overline{X})(y_i - \overline{Y})$
- Intuitivement, la covariance est une mesure de la variation simultanée de deux variables aléatoires.
- C'est-à-dire que la covariance devient plus positive pour chaque couple de valeurs qui diffèrent de leur moyenne dans le même sens, et plus négative pour chaque couple de valeurs qui diffèrent de leur moyenne dans le sens opposé.

Corrélation de deux variables aléatoires

Soit l'étude de deux variables X et Y, deux séries de valeurs définies dans $\Re$:

Corrélation de Bravais-Pearson

- $cor(X,Y) = \dfrac{\theta_{xy}}{\theta_x.\theta_y} = \dfrac{cov(X,Y)}{\sqrt{cov(X)}.\sqrt{cov(Y)}}$
- Le coéfficient de corrélation est compris entre $[-1,1]$
- $cor(X,Y) = 0 \Longrightarrow$, X et Y sont indépendant linéairement.
- $cor(X,Y) = 1$, une relation affine existe entre X et Y. L'une des variables est fonction affine croissante de l'autre variable.
- $cor(X,Y) = -1$, une relation affine existe entre X et Y. L'une des variables est fonction affine décroissante de l'autre variable.

Corrélation de Kendall

Soit deux variables ordinales X et Y. La corrélation de rangs rend compte d'une relation non-linéaire entre ces deux variables.
τ s'exprime de la façon suivante :

$$\tau = \frac{S}{D}$$

Où,

$$S = \sum_{i<j} (sign(x[i] - y[i]).sign(y[i] - x[i])) \tag{1}$$

et,

$$D = \frac{k(k-1)}{2} \tag{2}$$

- Mapping de R^n dans R^k. Avec k $\leq$ n.
- Projection dans un espace 2D d'un problème à n dimensions.
- Système d'axes indépendants.
- Réduction de la dimensionalité d'un problème.
- Perte d'information.
- Décomposition en valeurs propres. Transformation de Karhunen-Loève.
- Le nouvel espace est une combinaison linèaire de l'espace d'origine.

Centrer et Réduire une Matrice

- Soit S la matrice des échantillons.
- S contient k vecteurs colonnes. $X_1,\ldots,X_k$
- Soit X_i un vecteur colonne à n valeurs.
- k variables.
- n échantillons par variable.

Centrer et Réduire une Matrice (SCR)

Matrice des données brutes:

m_{11}	m_{12}	m_{13}	...	m_{1k}
m_{21}	m_{22}	m_{23}	...	m_{2k}
...	...	...	...	...
m_{n1}	...	...	...	m_{nk}

Calcul des moyennes et des variances pour chaque série de données:

M_1	M_2	M_3	...	M_k
σ_1	σ_2	σ_3	...	σ_k

Centere et réduire la matrice :

$(m_{11} - M_1)/\sigma_1$	$(m_{12} - M_2)/\sigma_2$	...	$(m_{1k} - M_k)/\sigma_k$
$(m_{21} - M_1)/\sigma_1$	$(m_{22} - M_2)/\sigma_2$	...	$(m_{2k} - M_k)/\sigma_k$
...	...	...	...
$(m_{n1} - M_1)/\sigma_1$	$(m_{n2} - M_2)/\sigma_2$	...	$(m_{nk} - M_k)/\sigma_k$

Matrice des Covariances (COVMAT)

A partir de la matrice centrée réduite.

- Construire une matrice carrée.
- Rend compte des interactions entre des pairs de variables numériques.
- Relations affines seulement.

$cov(SCR_{.0}, SCR_{.0})$	$cov(SCR_{.0}, SCR_{.1})$	$cov(SCR_{.0}, SCR_{.2})$	...	$cov(SCR_{.0}, SCR_{.k})$
$cov(SCR_{.1}, SCR_{.0})$	$cov(SCR_{.1}, SCR_{.1})$	$cov(SCR_{.1}, SCR_{.2})$	...	$cov(SCR_{.1}, SCR_{.k})$
...	...	...	...	...
$cov(SCR_{.k}, SCR_{.0})$	$cov(SCR_{.k}, SCR_{.1})$	$cov(SCR_{.k}, SCR_{.2})$	...	$cov(SCR_{.k}, SCR_{.k})$

Recherche des valeurs propres et des vecteurs propres

- Résoudre : $Det(COVMAT - \lambda I) = 0$ pour calculer les valeurs propres.
- Calcul des vecteurs propres. Résoudre : $(COVMAT - \lambda_{1\,a\,k} I)U = 0$
- Soit V la matrice des vecteurs propres.
- Projection de SCR dans l'espace ACP. $ACP = SCR.V$

$$A = \begin{pmatrix} 1 & -3 \\ -2 & 2 \end{pmatrix}$$

$$A - \lambda I = \begin{pmatrix} 1-\lambda & -3 \\ -2 & 2-\lambda \end{pmatrix}$$

$$det(A - \lambda I) = 0; (1-\lambda)(2-\lambda) - 6 = 0$$

$$\lambda^2 - 3\lambda - 4 = 0$$

$$\lambda 1 = -1;\ \lambda 2 = 4$$

Recherche des valeurs propres et des vecteurs propres

Les vecteurs propres associés à la valeur propre -1 sont ceux qui vérifient :

$\begin{pmatrix} 2 & -3 \\ -2 & 3 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} = 0$

Soit, $2x - 3y = 0$, E_{-1} la droite engendrée par

$$\begin{pmatrix} 2 \\ 3 \end{pmatrix}$$

$\begin{pmatrix} -3 & -3 \\ -2 & -2 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} = 0$

Soit, $-3x - 3y = 0$, E_4 la droite engendrée par

$$\begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

La matrice des vecteurs propres : $V = \begin{pmatrix} 2 & 1 \\ 3 & -1 \end{pmatrix}$

Inertie et information

Les valeurs propres nous indiquent le pourcentage d'informaion portée par chaque axe factoriel.
Taux d'information :

$$\frac{\lambda_j}{\sum_{i=1}^{k}(\lambda_i)}$$

Conservation de l'information

Il est possible de fixer un taux d'information à garantir (τ): Par exemple, nous souhaitons que dans le nouvel espace cible 90% de l'information soit conservée. Pour ce faire nous trions les valeurs propres et retenons les "p" plus grandes valeurs jusqu'à obtenir le taux d'information désiré.

- $Tri(\lambda = \{\lambda_i\}_{i=1}^{k})$
- Garder les "p" plus grandes valeurs.
- Tant que $\sum_{i=1}^{p}(\lambda_i) \leq \tau$

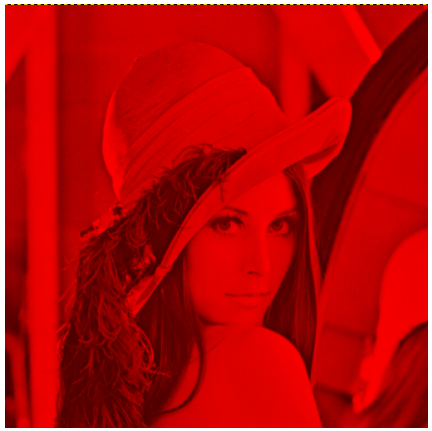
Une petite application ludique

Nous avons tous des miliers de photos sur nos PCs et cela prend beaucoup de place sur nos disques. Cela est en partie dû au fait que chaque image est en couleur: rouge ,vert, bleu (RGB). Chaque pixel, chaque site (x,y), contient une information couleur sur l'intensité du rouge, l'intensité du vert, l'intensité du bleu. Il est donc possible de diviser par trois la taille d'une image, en ne conservant qu'un seul canal ???? $(RGB - - > ACP1)$

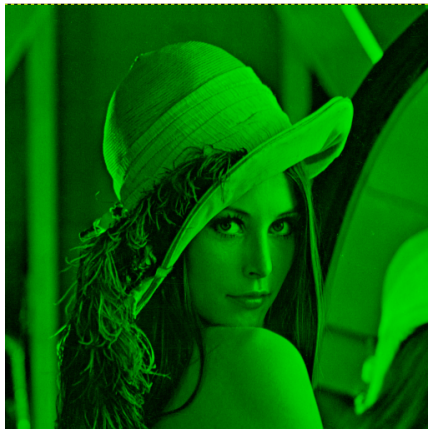
$(x, y, z) - - > x')$

Figure: L'image dont on veut réduire l'espace couleur

Image Couleur

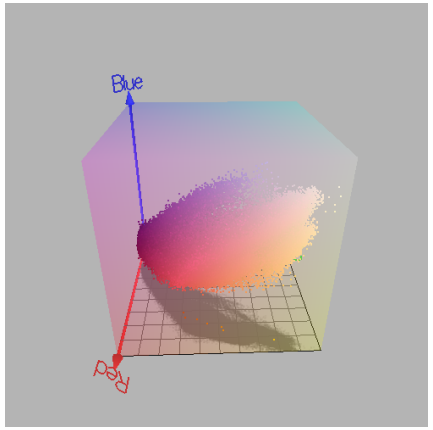


(a)



(b)

(c)



(d)

ACP sur une image couleur

Matrice de covariances:

$$P = \begin{pmatrix} var(R) & cov(R,G) & cov(R,B) \\ cov(G,R) & var(G) & cov(G,B) \\ cov(B,R) & cov(B,G) & var(B) \end{pmatrix}$$

Matrice des vecteurs propres:

$$V = \begin{pmatrix} 0.614 & 0.588 & 0.526 \\ -0.581 & -0.114 & 0.806 \\ 0.5346 & -0.801 & 0.271) \end{pmatrix}$$

Matrice des valeurs propres:

$$D = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}$$

$$= \begin{pmatrix} 2718 & 0 & 0 \\ 0 & 110 & 0 \\ 0 & 0 & 11) \end{pmatrix}$$

La conservation de l'axe principale permet d'expliquer plus 90% de l'information:

$$\frac{\lambda_1}{\sum_{i=1}^{3}(\lambda_i)} > \tau(0.90)$$

ACP sur une image couleur

Projection des données originales sur les axes factoriels:

$$P = \begin{pmatrix} RACP1 \\ RACP2 \\ RACP3 \end{pmatrix} = \begin{pmatrix} 0.614 & 0.588 & 0.526 \\ -0.581 & -0.114 & 0.806 \\ 0.5346 & -0.801 & 0.271) \end{pmatrix} . \begin{pmatrix} R \\ G \\ B \end{pmatrix}$$

Exemple : pour l'axe principal (λ_1)

$$RACP1 = R * 0.614 + G * 0.588 + B * 0.526$$

Image Couleur

(a) (b) (c)

Figure: (a) : Image originale. (b) Image projetée sur les trois axes de l'ACP. (c) Image projetée sur l'axe principale.

Recherche des composantes principales

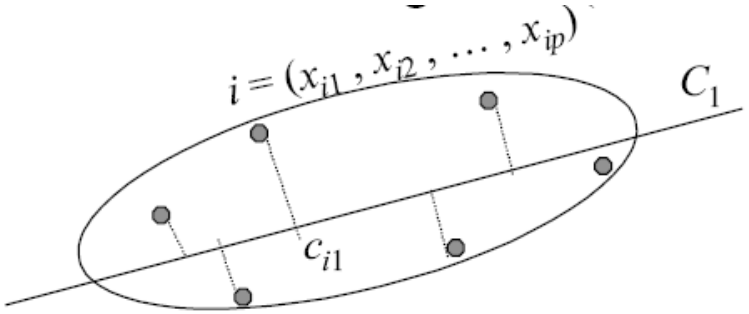
- Analyse de la structure de la matrice variance-covariance; c-à-d de la variabilité, dispersion des données.
- Objectif de l'ACP: décrire à l'aide de $q \leq p$ composantes un maximum de cette variabilité.
- Ce qui permet :
  - une réduction des données à q nouveaux descripteurs
  - une visualisation des données à 2 ou 3 dimensions (si q = 2 ou 3)
  - une interprétation des données : liaisons inter-variables

Recherche des composantes principales

- Composantes : $C_1, C_2, ..., C_k, ..., C_q$
- C_k = nouvelle variable = combinaison linéaire des variables d'origine $X_1, ..., X_p$:
- $C_k = a_{1k}X1 + a_{2k}X2 + \ldots\ldots + a_{pk}Xp$. coefficients a_{jk} à déterminer. Ceux sont les composantes des vecteurs propres.
- telle que les C_k soient:
  - 2 à 2 non corrélées,
  - de variance maximale,
  - d'importance décroissante.

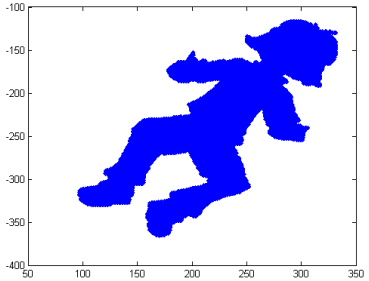
Recherche des composantes principales

- C_1 = 1ère composante principale doit être de variance maximale.
- Géométriquement : C_1 détermine une nouvelle direction dans le nuage de points qui suit l'axe d'allongement (étirement) maximal du nuage.
- c_{i1} = coordonnée du point i sur l'axe C_1 projection de x_i sur C1.
- $c_{i1} = \sum_{j=1}^{p} a_{1j} x_{ij}$.
- C_1 de variance maximale les projections c_{i1} sont les plus dispersées possible.
- C_1 = droite passant par le centre de gravité réalisant le meilleur ajustement possible du nuage
  - c-à-d : qui conserve au mieux la distance entre les points (après projection) =¿ droite de projection assurant une distorsion minimale.
- C_2 = 2ème composante, orthogonale à C1 et de variance maximale.

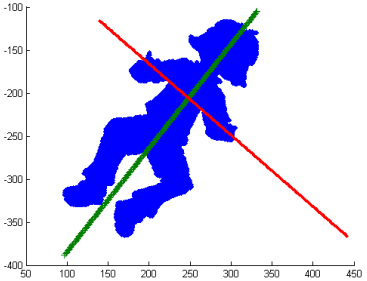


Exemple : Un cas d'étude.

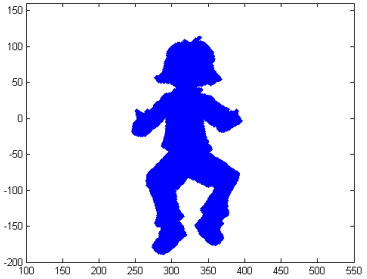
(a)



(b)



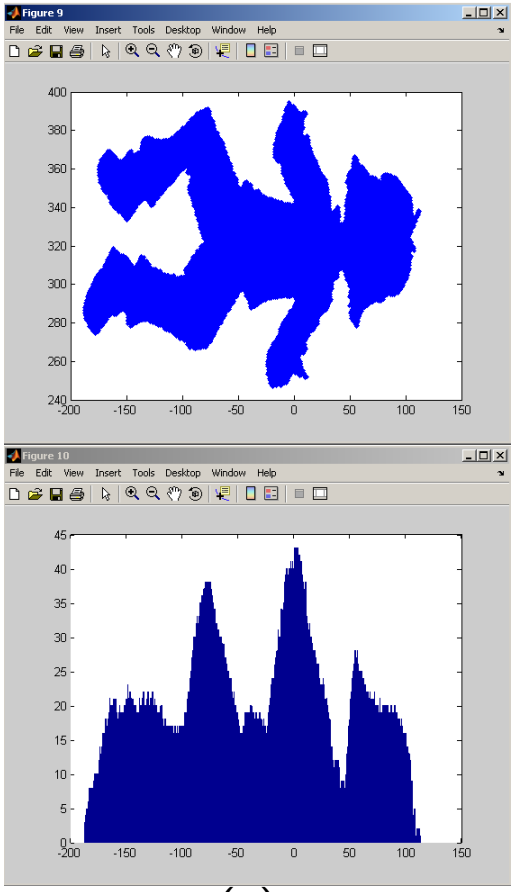
(c)



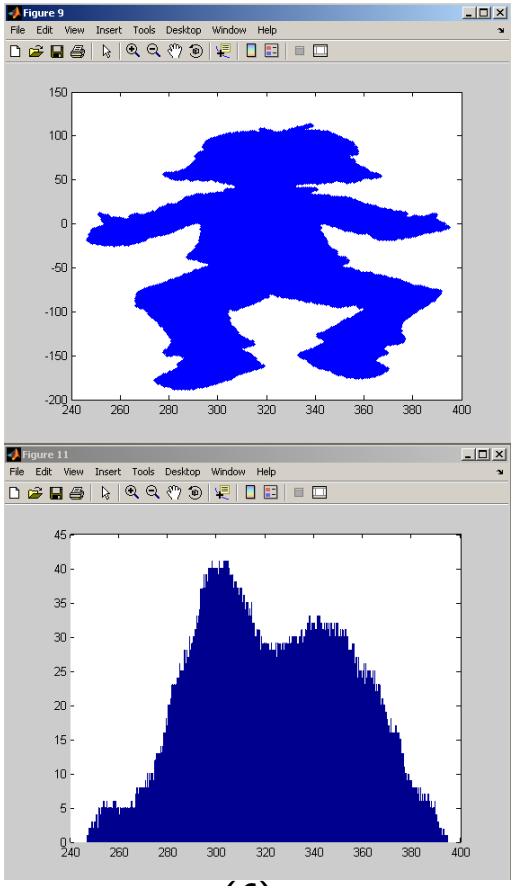
(d)

(a) : Image originale. (b) Image matricielle. (c) Axes principaux. (f) Projection sur C1 et C2.

Exemple : Un cas d'étude.



(e)



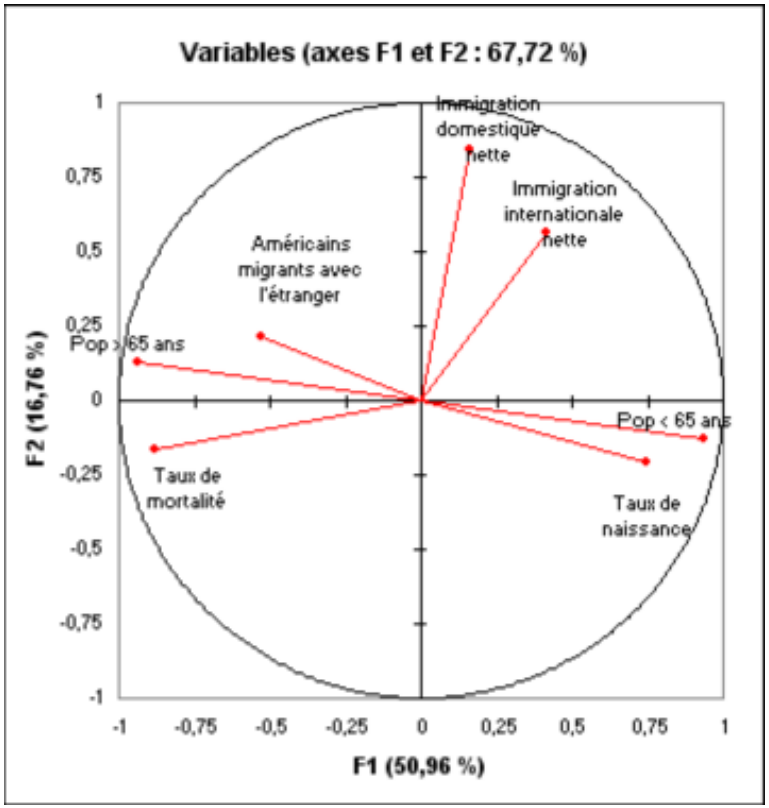
(f)

(e) Projection sur C1. (f) Projection sur C2.

Choix des r premières composantes principales

- Objectif : garder un maximum d'information des données initiales.
- Mesure de cette information : le % de variance expliquée
- $$\frac{\sum_{i=1}^{r} Var(C_i)}{Intertie\ totale}$$
-
- Si les variables originales sont fortement corrélées entre elles, un nombre réduit de composantes permet d'expliquer 80% à 90% de variance !

Notion

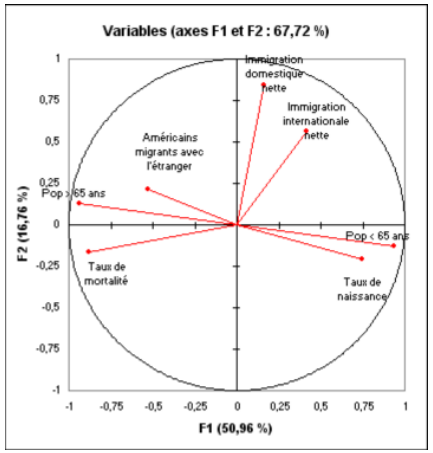


Cette notion est fondamentale en ACP: Pour chaque variable, on évalue la corrélation entre les données dans la base d'origine et les données dans la base cible.

$$Exemple : x = cor(X, XACP1) \; ; \; y = cor(X, XACP2)$$

Plus les variables sonct proches du bord du cercle et plus les variables sont bien représentées par le plan factoriel, c'est-à-dire que la variable est bien corrélée avec les deux facteurs constituant ce plan.

Interprétation



L'angle entre 2 variables, mesuré par son cosinus, est égal au coefficient de corrélation linéaire entre les 2 variables: cos(angle) = r(X1,X2)

- si les points sont très proches (angle proche de 0) : cos(angle) = r(X1,X2) = 1 donc X1 et X2 sont très fortement corrélés positivement.
- si a est égal à 90° , cos(angle) = r(X1,X2) = 0 alors pas de corrélation linéaire entre X1 et X2
- si les points sont opposés, a vaut 180°, cos(angle) = r(X1,X2) = -1 : X1 et X2 sont très fortement corrélés négativement

Attention, on ne peut interpréter que les variables situées au bord du cercle

Conclusion

- ACP pour l'analyse de séries de données numériques.
  - Etude de matrice de covariances.
  - Visualisation dans un espace décorrélé.
  - Description des interactions entre les variables.
- Les inconvénients ?
  - N'appréhende que les relations affines entre les variables. (phénoméne linèaire)