Séance 4: Analyse en Composantes Principales

Sébastien Gadat

Laboratoire de Statistique et Probabilités
UMR 5583 CNRS-UPS

www.lsp.ups-tlse.fr/gadat

Quatrième partie IV

Analyse en Composantes Principales

Motivations

Exemple1 :

- Satellite photographie des zones de 10km de large.
- Chaque zone est découpée en pixels de 20 m de large.
- Sur chaque pixel sont mesurées 8 longueur d'onde différentes chacune codée sur 8 bits
- $\Longrightarrow$ Quantité importantes de données

Exemple2 :

- n relevés de notes dans différentes matières scolaires.
- Analyser les tendances entre les performances dans chacune des matières.
- $\Longrightarrow$ Interaction entre différentes variables décrivant les individus

Motivations

Objectifs :

- Représenter graphiquement l'observation de $p > 3$ variables.
- Recherche de résumés pertinents (nuages de point) dans le plan ou l'espace, respectant :
  - les distances entre individus
  - la structure des corrélations entre variables.

Présentation élémentaire de l'ACP

Notes de $n = 9$ élèves dans $p = 4$ disciplines.

	MATH	PHYS	FRAN	ANGL
jean	6.00	6.00	5.00	5.50
alan	8.00	8.00	8.00	8.00
anni	6.00	7.00	11.00	9.50
moni	14.50	14.50	15.50	15.00
didi	14.00	14.00	12.00	12.50
andr	11.00	10.00	5.50	7.00
pier	5.50	7.00	14.00	11.50
brig	13.00	12.50	8.50	9.50
evel	9.00	9.50	12.50	12.00

♣ Quelles analyses statistiques élémentaires ?

Statistiques élémentaires

Variable	Moyenne	Ecart-type	Minimum	Maximum
MATH	9.67	3.37	5.50	14.50
PHYS	9.83	2.99	6.00	14.50
FRAN	10.22	3.47	5.00	15.50
ANGL	10.06	2.81	5.50	15.00

♣ Conclusion sur la dispersion des variables ? Corrélations :

Variable	MATH	PHYS	FRAN	ANGL
MATH	1.00	0.98	0.23	0.51
PHYS	0.98	1.00	0.40	0.65
FRAN	0.23	0.40	1.00	0.95
ANGL	0.51	0.65	0.95	1.00

♣ Interprétation sur la dépendance entre les variables ?

Statistiques élémentaires

Matrice des variances-covariances :

Variable	MATH	PHYS	FRAN	ANGL
MATH	11.39	9.92	2.66	4.82
PHYS	9.92	+ 8.94	4.12	5.48
FRAN	2.66	4.12	+ 12.06	9.29
ANGL	4.82	5.48	9.29	+ 7.91

♣ Somme diagonale : 40.30. Que représente cette somme ?

L'objectif de l'ACP est de construire des "facteurs" représentant les données de façon adaptée à la structure de corrélations des variables.

FACTEUR	VAL. PR.	PCT. VAR.	PCT. CUM.
1	28.23	0.70	0.70
2	12.03	0.30	1.00
3	0.03	0.00	1.00
4	0.01	0.00	1.00

Notations - individus et variables

p variables statistiques réelles observées sur n individus de poids p_i.
On synthétise les données en une matrice

$$X = \begin{pmatrix} x_1^1 & \dots & x_1^j & \dots & x_1^p \\ \vdots & \ddots & \dots & \dots & \vdots \\ x_i^1 & \dots & x_i^j & \ddots & x_i^p \\ \vdots & \dots & \ddots & \ddots & \vdots \\ x_n^1 & \dots & x_i^j & \dots & x_n^p \end{pmatrix}$$

- ♣ À quoi correspond x_i^j ?
- ♣ Individu i : $\mathbf{e}_i$
- ♣ Variable j : x^j
- ♣ Matrice des poids : D Équi-répartion, Poids quelconques

Notations - individus et variables

Deux nuages possibles

- Nuage des individus N
  - n points
  - Espace de dimension p
  - Métrique euclidienne donnée par une matrice M
  - M symétrique définie positive
- Nuage des variables N_v
  - p points
  - Espace de dimension n
  - Métrique donnée par la matrice diagonale des poids $D = diag(p_i)$
  - Poids normalisés

Sans perte de généralité, on supposera les individus centrés, c'est-à-dire $g = 0_{\mathbb{R}^p}$.

Objectifs

On souhaite construire des résumés.

- Graphiques :
  - Approcher le nuage de variables par la meilleure projection possible sur un espace vectoriel de dimension fixée.
  - Approcher le nuage d'individus par leurs coordonnées sur un espace de projection de dimension fixée.
  - Le sens du mot "meilleur" reste très largement à définir.
  - Utiliser les représentations graphiques pour interpréter les ressemblances entre individus et corrélations entre variables.
- Numériques :
  - Déterminer des axes et des nouvelles coordonnées en nombre fixé permettant de reconstruire le nuage X avec un minimum d'erreur.

Ces deux approches seront ici équivalentes

Rappels d'entités statistiques

- $\langle e_i, e_j \rangle_M = e_i' M e_j$ et $\|e_i\|^2 = e_i' M e_i$
- $d(e_i, e_j)^2 = (e_i - e_j)' M (e_i - e_j)$ et $cos(e_i, e_j)^2 = \frac{d(e_i, e_j)^2}{\|e_i\|^2 \|e_j\|^2}$
- Choix de M :
  - Identité si les données sont du même type (notes d'une copie par exemple)
  - Inverse de la matrice diagonale des écarts types dans le cas de variables non standardisées.
- Matrice de Variance/Covariance :

$$V = \sum_{i=1}^{n} p_i (e_i - g)' (e_i - g) = X' D X$$

- Proposition : L'inertie du nuage de points satisfait :

$$I_g(N) = \sum_{i=1}^{n} p_i d(e_i, g)^2 = Tr(VM) = \frac{1}{2} \overbrace{\sum_{i=1}^{n} \sum_{i=1}^{n} p_i p_i \|e_i - e_i\|^2}^{\bar{D}^2}$$

Projection sur un sous-espace

Le principe sera de projeter les individus sur un espace de petite dimension. Il faut donc calculer cette projection orthogonale pour le produit scalaire défini par M sur un espace W.
On note Π_W^M (et plus simplement Π_W) la projection sur W qui est M-orthogonale.

Lemma

Π_W *est une application linéaire, il existe donc une matrice* P *représentant cette application linéaire dans la base canonique des variables* $x^1, \dots x^p$.

Inertie d'un nuage projeté

On peut exprimer la matrice de projection associée à l'application linéaire précédente :

Lemma

Si W est engendré par k vecteurs linéairement indépendants $u_1, \dots u_k$, et si U désigne la matrice de taille $n \times k$ contenant dans chaque colonne un des u_i, la matrice P de k'application linéaire Π_W s'écrit :

$$P = U(U'MU)^{-1}U'M$$

On calcule enfin l'inertie du nuage projeté :

Theorem

L'inertie du nuage de points projeté sur W est donné par

$$I(\Pi_W(N)) = Tr(VMP)$$

Structure des espaces optimaux

L'objectif est la recherche d'un espace optimal de dimension fixé k satisfaisant le critère d'optimalité au sens des moindres carrés.

Theorem

L'espace F_k maximisant le critère des moindres carrés

$$\sum_{i=1}^{n} p_i \left(x_i - \Pi_{F_k}^M(x_i)\right)^2$$

coïncide avec l'espace maximisant l'inertie du nuage projeté

$$\sum_{i=1}^{n} p_i \|\Pi_{F_k}^M(x_i)\|_M^2$$

Structure des espaces optimaux

Le théorème suivant nous informe sur l'organisation spatiale des différents sous-espaces optimaux.

Theorem

Soit F_k un sous-espace de dimension k portant l'inertie maximale, alors le sous-espace de dimension $k+1$ portant l'inertie maximale est la somme directe de F_k et du sous-espace de dimension 1 M-orthogonal à F_k portant l'inertie maximale :

$$F_{k+1} = F_k \oplus_M Vect(a_{k+1})$$

Les solutions sont « emboîtées ».

Résolution de l'ACP

- Le théorème 2 assure qu'il faut rechercher les axes « un par un »et reconstruire F_k par

$$F_k = Vect(a_1) \oplus \cdots \oplus Vect(a_k)$$

- Ici, a_i sont des vecteurs maximisant l'inertie sur un espace M-supplémentaire à $Vect(a_1) \oplus \cdots \oplus Vect(a_{i-1})$ dans $E = \mathbb{R}^p$.
- Il suffit donc de trouver mathématiquement comment déterminer u_i.

Theorem

L'espace F_k est engendré par les k vecteurs propres de VM associés aux k plus grandes valeurs propres de VM.

Remarque : Il y a deux points à démontrer :

- On obtient des vecteurs propres pour les a_i.
- Ces facteurs sont classés par ordre d'importance croissante en fonction des valeurs propres qui sont toutes positives ou nulles.

Facteurs principaux et composantes principales

À tout axe a M-normé à 1 correspond un facteur principal $u \in \mathbb{R}^p$ donné par

$$u = Ma$$

Theorem

Les facteurs principaux sont les vecteurs propres associés aux mêmes valeurs propres de MV et sont M^{-1} orthonormaux.

Les variables c_i (éléments de $\mathbb{R}^n$) sont les composantes principales :

$$c_i = Xu_i$$

Theorem

Si $\lambda_1, \dots \lambda_p$ désignent les p valeurs propres, on a alors

$$\forall i \in \{1, \dots p\} \qquad Var(c_i) = \lambda_i$$

Ces c_i sont également vecteurs propres d'une matrice car :

Formules de reconstitution

Connaissant les composantes principales et les facteurs principaux, il est possible de reconstituer le nuage initial par le bais du résultat suivant :

Theorem

$$X = \sum_{j=1}^{p} c_j u_j' M^{-1}$$

On peut également obtenir les égalités :

Theorem

$$MV = \sum_{j=1}^{p} \lambda_j u_j u_j' M^{-1}$$

et

$$VM = \sum_{j=1}^{p} \lambda_j a_j a_j' M$$

Résumé

- L'ACP revient à remplacer les variables $x^1, \dots x^p$ qui sont *a priori* corrélées par de nouvelles variables $c^1, \dots c^p$ non corrélées.
- Ces variables sont appelées composantes principales, elles sont combinaisons linéaires des x^j.
- Les composantes principales sont de variance maximale (elles ont été obtenues par maximisation d'inertie).
- Elles sont en plus les plus liées, au sens des moindres carrés, aux variables x^j.

Présentation rapide du « protocole »
Choix du nombre d'axes
Corrélations entre composantes et variables initiales
Représentation des individus
Représentation simultanée variables/individus
Un exemple : température des villes
Un autre exemple : le tableau de notes
Effet « taille »

Présentation rapide du « protocole »

L'ACP fournit de nouvelles variables qui ne sont que virtuelles, ainsi que des représentations graphiques permettant de visualiser les relations entre variables ainsi que l'existence éventuelle de groupes d'individus et groupes de variables. L'interprétation est censée répondre à des questions simples du type :

- Combien d'axes retenir dans l'interprétation ?
- Y a-t-il des groupes d'individus voisins ou de variables corrélées ?
- Y a-t-il des individus « atypiques »
- Quels individus ou variables s'opposent ?
- Quelle fiabilité d'interprétation a-t-on sur tels axes, variables ou individus ?

Choix du nombre d'axes

- ♣ Règle de la part d'inertie : on décide de garder q tel que la part d'inertie expliquée est supérieure à un seuil fixé a priori

$$r_q = \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k}$$

- ♣Règle de Kaiser : on ne conserve que les axes dont la valeur propre est supérieure à leur moyenne car ce sont les seules jugées informatives.
- ♣Éboulis des valeurs propres : on sélectionne q en localisant le « coude »dans l'éboulis des valeurs propres.

Choix du nombre d'axes

On observe des températures moyennes mensuelles de 32 villes francaises, moyennes effectuées sur 10 ans.

♣ Quelle est la taille de X ?

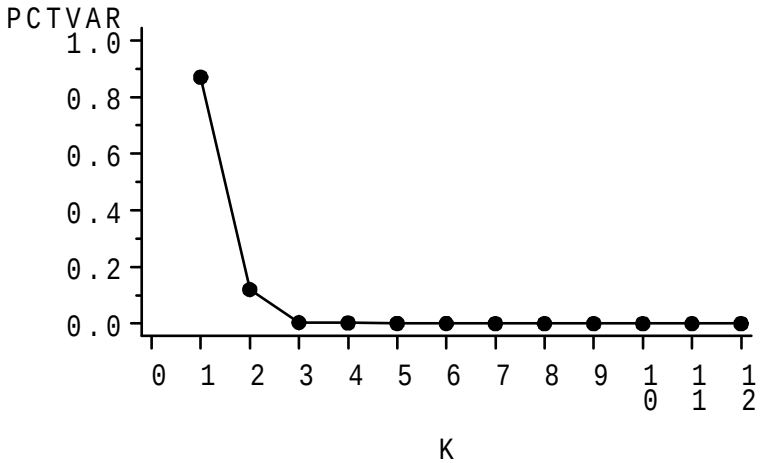


FIG.: *Températures : éboulis des valeurs propres.*

Présentation rapide du « protocole »
Choix du nombre d'axes
Corrélations entre composantes et variables initiales
Représentation des individus
Représentation simultanée variables/individus
Un exemple : température des villes
Un autre exemple : le tableau de notes
Effet « taille »

Corrélations entre composantes et variables initiales

- On donne une signification aux composantes principales en reliant ces composantes aux variables initiales.
- Le plus simple est de calculer les coefficients de corrélation linéaire entre une composante c et une variable x^j

$$r(c, x^j)$$

Theorem (Corrélations composantes/variables)

Dans le cas le plus fréquent où l'on utilise les données centrées réduites($M = D_{1/s^2}$) les corrélations sont données par

$$r(c, x^j) = \sqrt{\lambda} u_j$$

où λ est la valeur propre associée à la composante c et u le facteur principal associé (i.e. $c = \bar{X}_{1/s^2} u$)

Utilisation des corrélations entre composantes et variables initiales

On effectue ces calculs pour des composantes principales 2 par 2.
On synthétise en général les résultats dans des « cercles de corrélation »ou par des « biplots » :

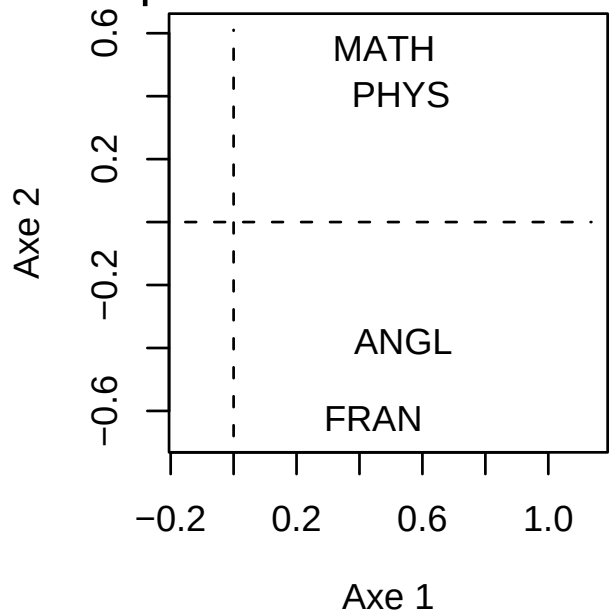


FIG. : Projection des variables « notes »sur les 2 premiers axes de l'acp

Qualité de représentation des variables sur un axe

Il se peut qu'une variable soit plus ou moins bien projetée sur un axe de l'ACP. Dans ce cas, l'interprétation doit être effectuée avec précaution :

- Si une variable est bien représentée sur un axe ou un plan, on peut interpréter de façon fiable.
- Sinon, l'interprétation est plus qu'hasardeuse.

On mesure cette qualité de représentation sur l'axe k de l'acp par

$$r_k(x^j) = \frac{\lambda_k \left(u_j^k\right)^2}{\sum_{i=1}^{p} \lambda_k \left(u_i^k\right)^2}$$

Ici, u^k désigne le k^{ieme} facteur associé à la valeur propre λ_k.

Remarques sur la qualité de représentation des variables sur un axe

En général, si l'ACP donne des résultats fiables, les variables ne sont bien représentées que sur les premiers axes de l'ACP.
Souvent, les variables bien représentées sur un plan de l'ACP se situent « loin »de l'origine du biplot sur ce plan.
Rappel : la qualité de représentation de la totalité des variables sur les k premiers axes de l'acp (espace appelé F_k) vaut

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$$

L'appréciation du pourcentage d'inertie doit faire intervenir le nombre de variables initiales : un 10% n'a pas le même intérêt sur un tableau

Représentation des individus

- Si les individus ne sont pas « anonymes », ils aident à l'interprétation des axes principaux et des composantes principales. On pourra par exemple rechercher les individus opposés le long d'un axe.
- On peut calculer la contribution d'un individu i sur une composante c^k par

$$C(c^k, i) = \frac{p_i {c^k}_i^2}{\lambda_k}$$

  Ici, c_i^k désigne la valeur de la composante c^k pour l'individu i.
- Une contribution excessive d'un individu à une composante est source d'instabilité et doit être évitée, par exemple en supprimant l'individu de l'analyse.

Représentation des individus

- Les composantes c_i^k permettent de positionner les individus sur les axes de l'ACP.
- On peut également calculer la qualité de représentation de l'individu i sur F_k par

$$\frac{\sum_{j=1}^{k}(c_i^j)^2}{\sum_{j=1}^{p}(c_i^j)^2}$$

- Et enfin la contribution de l'individu i à l'inertie du nuage

$$\frac{\sum_{j=1}^{k}(c_i^j)^2}{\sum_{j=1}^{p}\lambda_j}$$

Représentation simultanée variables/individus

- De nombreux logiciels scientifiques permettent de représenter simultanément variables et individus sur les mêmes biplots.
- Cette représentation est *a priori* dénuée de sens puisque ces deux entités mathématiques n'appartiennent pas au même espace.
- On ne peut pas parler de « proximité »entre un individu et une variable !
- Les deux interprétations se complètent néanmoins mais ne peuvent être superposées.

Un exemple : température des villes

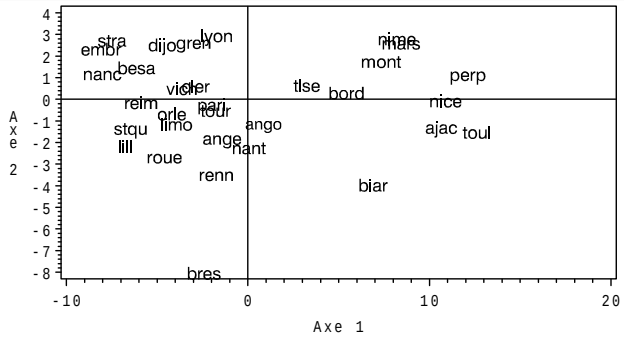
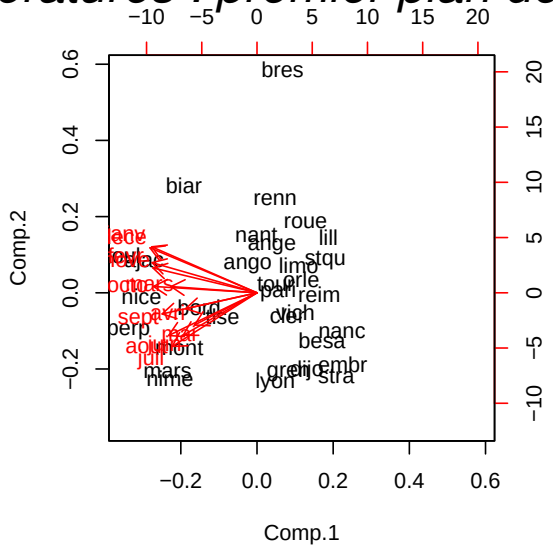


FIG.: *Températures : premier plan des individus.*



Un exemple : température des villes

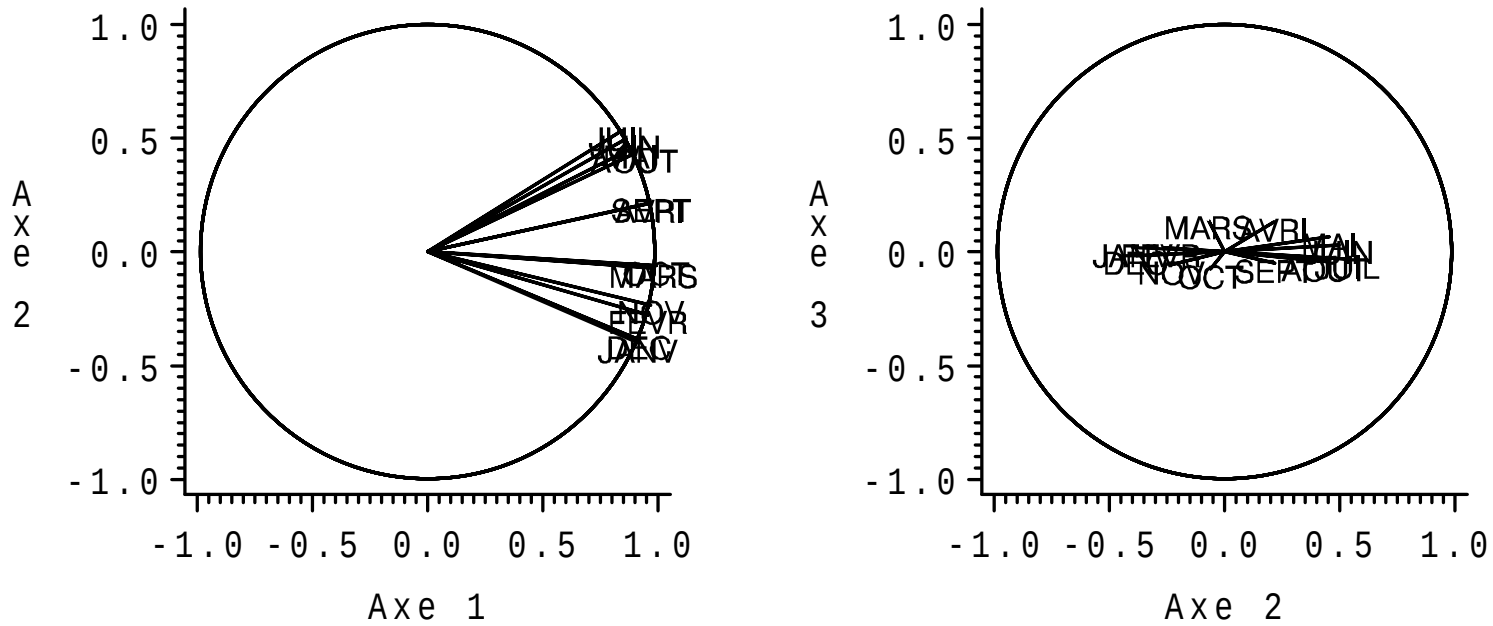


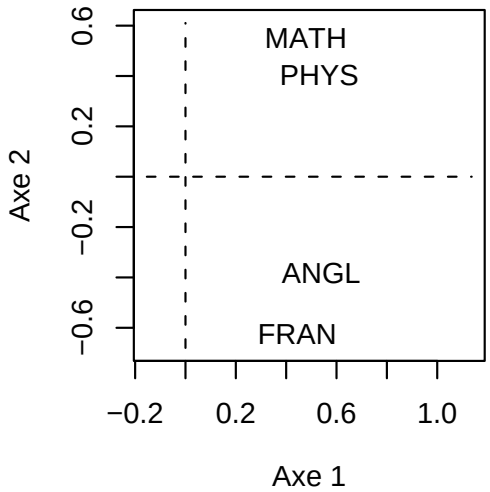
FIG.: *Températures : Premier et deuxième plan des variables.*

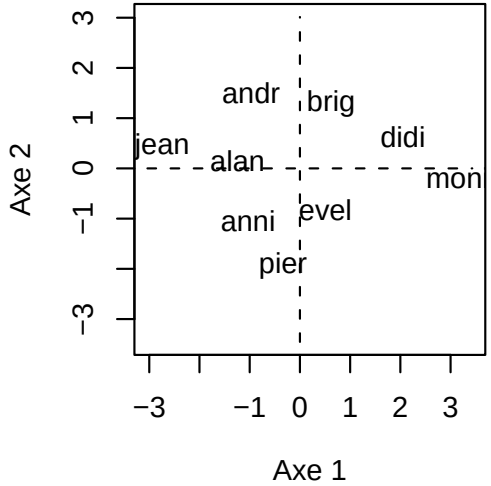
Un autre exemple : le tableau de notes

	MATH	PHYS	FRAN	ANGL
jean	6.00	6.00	5.00	5.50
alan	8.00	8.00	8.00	8.00
anni	6.00	7.00	11.00	9.50
moni	14.50	14.50	15.50	15.00
didi	14.00	14.00	12.00	12.50
andr	11.00	10.00	5.50	7.00
pier	5.50	7.00	14.00	11.50
brig	13.00	12.50	8.50	9.50
evel	9.00	9.50	12.50	12.00

FACTEUR	VAL. PR.	PCT. VAR.	PCT. CUM.
1	28.23	0.70	0.70
2	12.03	0.30	1.00
3	0.03	0.00	1.00
4	0.01	0.00	1.00

Un autre exemple : le tableau de notes





Effet taille

- Lorsque toutes les variables sont corrélées positivement entre elles, la première composante principale définit un « facteur taille ».
- Dans ce cas, la composante principale est proportionnelle à la moyenne

$$\frac{1}{p}\sum_{j=1}^{p} x^j$$

- La seconde composante principale différencie alors les individus de taille semblable, c'est le facteur de « forme ».
- Identifier ces effets dans l'exemple des notes précédents...

Analyse factorielle d'un tableau de distance

- On souhaite analyser un tableau de distance entre individus
- Ce tableau se présente sous la forme

$$\mathcal{D} = \begin{pmatrix} d(1,1) & \dots & d(1,j) & \dots & d(1,n) \\ \vdots & & \dots & & \vdots \\ d(i,1) & \dots & d(i,j) & \dots & d(i,n) \\ \vdots & & \dots & & \vdots \\ d(n,1) & & d(n,j) & & d(n,n) \end{pmatrix}$$

- D est symétrique, à termes positifs
- Question : À quelle condition la matrice D est-elle un tableau de distances euclidiennes entre points ?
- Comment représenter ces points ?

Introduction

Notations

- On suppose donnés n individus.
- p variables de valeurs inconnues pour chaque individu.
- Indice de dissimilarité/distance entre chacun des individus connus.
- ♣ Comment construire une représentation euclidienne dans un espace de taille réduite fidèle aux données ?

Exemples

- ♣ La matrice de distances est-elle euclidienne ?
- ♣ L'approximation euclidienne est-elle satisfaisante ?
- La MDS est une technique factorielle (nécessité de déterminer un nombre de dimension).
- Possibilité d'observer graphiquement les données à travers différentes optiques.

Définitions

Rappelons quelques propriétés et définitions élémentaires mais basiques à propos de la notion de distance.

- Une matrice $(n \times n)$ $\mathcal{D}$ est appelée matrice de distance si elle est symétrique et si :

$$d_j^j = 0 \text{ et } \forall (j,k), j \neq k, d_j^k \geq 0.$$

- Une matrice $(n \times n)$ $\mathcal{C}$ est appelée matrice de similarité si elle est symétrique et si

$$\forall (j,k), c_j^k \leq c_j^j.$$

- ♣ On peut transformer une matrice de similarité en $(c_j^k)_{j,k}$ en matrice de distance *via* :

$$d_j^k = (c_j^j + c_k^k - 2c_j^k)^{-1/2}$$

Définitions

Une matrice de distance est dite euclidienne s'il existe une configuration de vecteurs $\{x_1, \ldots, x_n\}$ vérifiant

$$(d_j^k)^2 = \langle x_j - x_k ; x_j - x_k \rangle$$

On note A la matrice issue de $\mathcal{D}$ de terme général

$$(a_i^j) = -\frac{(d_i^j)^2}{2}$$

D désigne la matrice des poids des individus.

Définitions

Proposition :

- La matrice de projection D-orthogonale au vecteur $\mathbf{1}$ est donnée par
$$H = Id - \mathbf{1}\mathbf{1}'\mathbf{D}$$
- Une matrice de distance $\mathcal{D}$ est euclidienne si et seulement si $B = HAH'$ est symétrique définie positive. (B est la matrice obtenue par double centrage de A).
- Si la matrice de similarité $\mathcal{C}$ est positive, alors la matrice de distance $\mathcal{D}$ déduite de $\mathcal{C}$ est euclidienne.

Recherche d'une configuration de points

- Positionnement multidimensionnel : recherche d'une configuration de points dans un espace euclidien qui admette une matrice de distances :
  - égale à $\mathcal{D}$ si celle-ci satisfait la proposition précédente
  - meilleure approximation possible pour un rang de matrice donné (en général 2) de $\mathcal{D}$
- Il n'y a jamais unicité d'une telle représentation : si $(x_i)_i$ est une solution, alors

$$(z_i)_i = (Fx_i + b)_i$$

  est une solution lorsque F orthogonale et b quelconque.
- Une solution est donc définie à rotation et translation près.

Algorithme MDS

On se donne $\mathcal{D}$ matrice de distance et B la matrice centrée des lignes et colonnes, calculée comme précédemment.

- Si $\mathcal{D}$ est une matrice de distance euclidienne de points $\{x_1, \ldots, x_n\}$, alors B s'écrit en fait

$$b_{i,j} = \langle x_i - \bar{x}, x_j - \bar{x} \rangle$$

et B se met sous la forme $B = (HX)'(HX)$ et est appelée matrice des produits scalaires de la configuration centrée.

- Réciproquement, si $\mathbf{B}$ est positive de rang p, on écrit sa décomposition spectrale

$$\mathbf{B} = \mathbf{U}\mathbf{\Delta}\mathbf{U}'$$

Les vecteurs sont les lignes de $\mathbf{X} = \mathbf{U}\mathbf{\Delta}^{1/2}$ qui fournissent les coordonnées des vecteurs de la représentation euclidienne.

Relation entre ACP et MDS

Supposons que Y soit un nuage de n points $(x_1, \dots x_n)$ décrits par p variables (connues). On définit $\mathcal{D}$ comme la matrice des distances entre les n individus :

$$d_i^j = \|x_i - x_j\|_M$$

La représentation graphique obtenue par MDS sur $\mathcal{D}$ est identique à la réalisation d'une ACP sur $(Y, M, 1/nId)$.

Exemples

- Données : tableau contenant les distances à parcourir par route entre différentes villes (en km).
- Les "coordonnées" des villes sont inconnues.