Séance 1: Analyse en composantes principales

Révisions

Sébastien Gadat

Laboratoire de Statistique et Probabilités
UMR 5583 CNRS-UPS

www.lsp.ups-tlse.fr/gadat

Première partie I

Analyse en Composantes Principales

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 ?

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

♣Quelques questions...

- Comment ont-été trouvés les facteurs précédents ?
- Ont-ils une existence réelle ou virtuelle ?
- Que vaut l'inertie totale du nuage après passage sur les 4 facteurs ?
- En quoi ces facteurs là permettent-ils d'obtenir une représentation parcimonieuse des données ?
- Comment utiliser ces facteurs pour interpréter le jeu de données ?

Représentation avec les facteurs

Pour obtenir une interprétation des facteurs, et de l'organisation du jeu de données, on a besoin des coefficients de corrélation variables/facteurs. ♣ A quoi vont nous servir ces coefficients ?

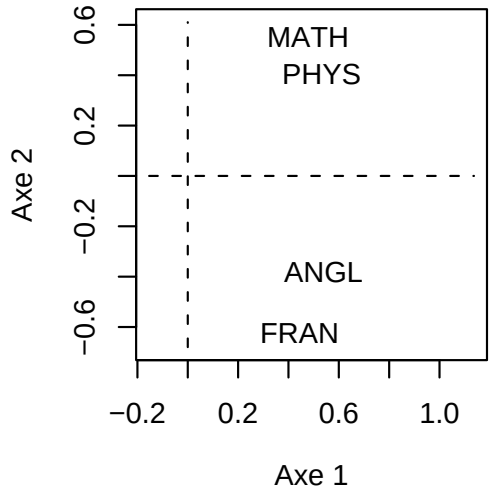

FIG.: *Données fictives : Représentation des variables*

♣Pourquoi une telle représentation ?

♣Interprétation sur les variables

- Avec qui le premier facteur est-il corrélé positivement ? Conclusion ?
- Qu'en est-il de l'axe 2 ? Interprétation ?
- Comment mieux préciser cette interprétation ?

Utilisation des facteurs principaux

Les facteurs sont en réalité une nouvelle base de l'espace vectoriel initial. On peut donc utiliser ces facteurs pour décrire les individus. Il suffit simplement pour cela de récupérer les coordonnées des individus initiaux dans la nouvelle base.

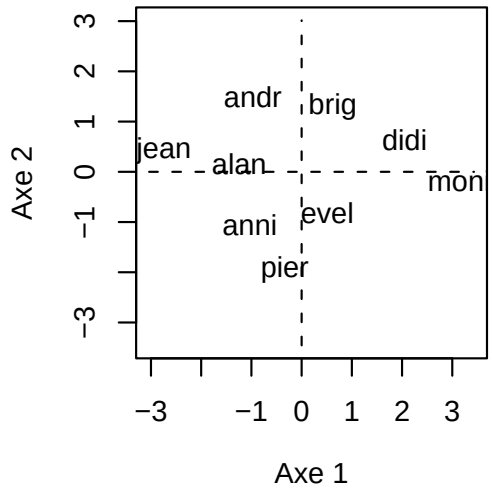

FIG.: *Données fictives : Représentation des individus*

♣Interprétation sur les individus

- Que représente le premier axe ? Comment varient les individus sur le premier axe ?
- Que représente le second axe ? Comment varient les individus sur le premier axe ?

Notations

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

Statistiques élémentaires

- ♣ Moyenne empirique ou point moyen :

$$g =$$

- ♣ Matrice centrée des données :

$$Y =$$

- ♣ Matrice de Variance/Covariance :

$$V =$$

Métrique sur les individus

En statistique, chaque dimension correspond à un caractère qui s'exprime avec son unité particulière. Comment calculer la distance entre deux individus décrits par les trois caractères : âge, salaire, nombre d'enfants ?
On utilise la formulation générale :

$$d^2(e_i, e_j) = (e_i - e_j)' M (e_i - e_j)$$

Le choix de M dépend de l'utilisateur. Les 2 choix en statistiques pour l'ACP sont :

- Cas euclidien $M =$
- Variables réduites $M =$

Intertie et corrélations

L'inertie du nuage de points pondérés vaut

$$I_g =$$

On montre que

$$I_g = \qquad =$$

On mesure la distance entre 2 variables via D :

$$< x^j, x^k >_D =$$

Inertie et corrélations

Avec des variables centrées vecteurs de $(F, \mathbf{D})$:

- ♣Que représente la "longueur" d'une variable ?

$$\|x^j\|_D{}^2 =$$

- ♣Que représente le cosinus d'un angle ?

$$\cos\theta_{j,k} = \frac{< x^j, x^k >_D}{\|x^j\|\|x^k\|} =$$

Présentation de la méthode

Objectifs :

- Représentation graphique "optimale" des individus et des variables.
- Réduction de dimension (compression).

Le principe est d'obtenir un sous-espace de dimension faible sur lequel la projection ressemble le plus au nuage initial.

Modèle

Étant donné $k < p$, on cherche F_k espace de dimension k tel que l'inertie des projections orthogonales des individus soit la plus grande possible.

Proposition : Maximiser l'inertie du nuage projeté $\Longleftrightarrow$ Minimiser la distance au sens des moindres carrés + F_k contient g.

Preuve : . . .

$$\text{Observation} = \text{Modèle} + \text{Bruit}$$

$$\mathbf{e}_i = \mathbf{f}_i + \varepsilon_i,\ i = 1, \ldots, n$$

$$\text{avec} \begin{cases} E(\varepsilon_i) = \\ \text{var}(\varepsilon_i) = \end{cases}$$

♣ Que vaut E_q et g dans l'exemple introductif ?

Emboîtement des solutions

Proposition : Soit F_k un sous-espace d'inertie maximale parmi les sous-espaces de dimension k, 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 d'inertie maximale. Les solutions sont donc "emboîtées".

Proposition : Matrice de projection sur la droite a :

$$P = a(a'Ma)^{-1}a'M$$

Proposition : L'inertie du nuage projeté sur la droite a

$$I_{proj} = \frac{a'MVMa}{a'Ma}$$

Estimation

Théorème : Le sous-espace F_k de dimension k est engendré par les k vecteurs propres de VM associés aux k plus grandes valeurs propres.

- Les vecteurs propres a_i de VM normés sont les axes principaux. Ils sont M et V^{-1} orthogonaux.
- Les facteurs principaux u_i sont les vecteurs propres M^{-1} normés de MV.
- Les composantes principales c_i sont données par

$$c_i = Xu_i$$

- La variance d'une composante principale est égale à λ_i, i^{eme} valeur propre de VM (ou MV).

Formules de reconstiturion

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

♣ Coordonnée de l'individu i projeté sur l'axe k

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

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

Représentations graphiques - Les individus

On peut représenter les individus *via* les composantes c_i^k sur l'axe k.

♣ Qualité de représentation de l'individu i sur F_k :

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

♣ Contribution de l'individu i sur F_k :

$$\gamma_i^k =$$

♣ Contribution de l'individu i à l'inertie du nuage :

$$\gamma_i =$$

Représentations graphiques

♣ Qualité de représentation des variables sur F_k :

$$r_k =$$

Une variable x_j est représentée par la projection sur le sous-espace F_k engendré par les k premiers axes factoriels. La coordonnée de $\mathbf{x}_j$ sur $\mathbf{u}^k$ est :

$$\left\langle \mathbf{x}_j, \mathbf{a}^k \right\rangle_{\mathbf{D}} = \sqrt{\lambda_k} u_j^k$$

C'est la corrélation variable j facteur k.
Qualité de la représentation de $\mathbf{x}_j$:

$$\frac{\sum_{k=1}^{q} \lambda_k (u_k^j)^2}{\sum_{k=1}^{p} \lambda_k (u_k^j)^2}.$$

Critère pour choisir q

- ♣Part d'inertie :
- ♣Règle de Kaiser :
- ♣Éboulis des valeurs propres :

Un exemple : Température de ville par mois

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

♣ Quelle est la taille de X ?

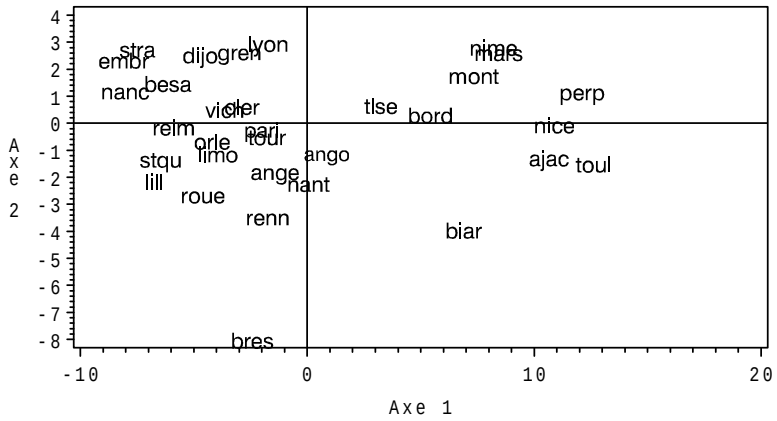

FIG.: *Températures : premier plan des individus.*

Cercle des corrélations :

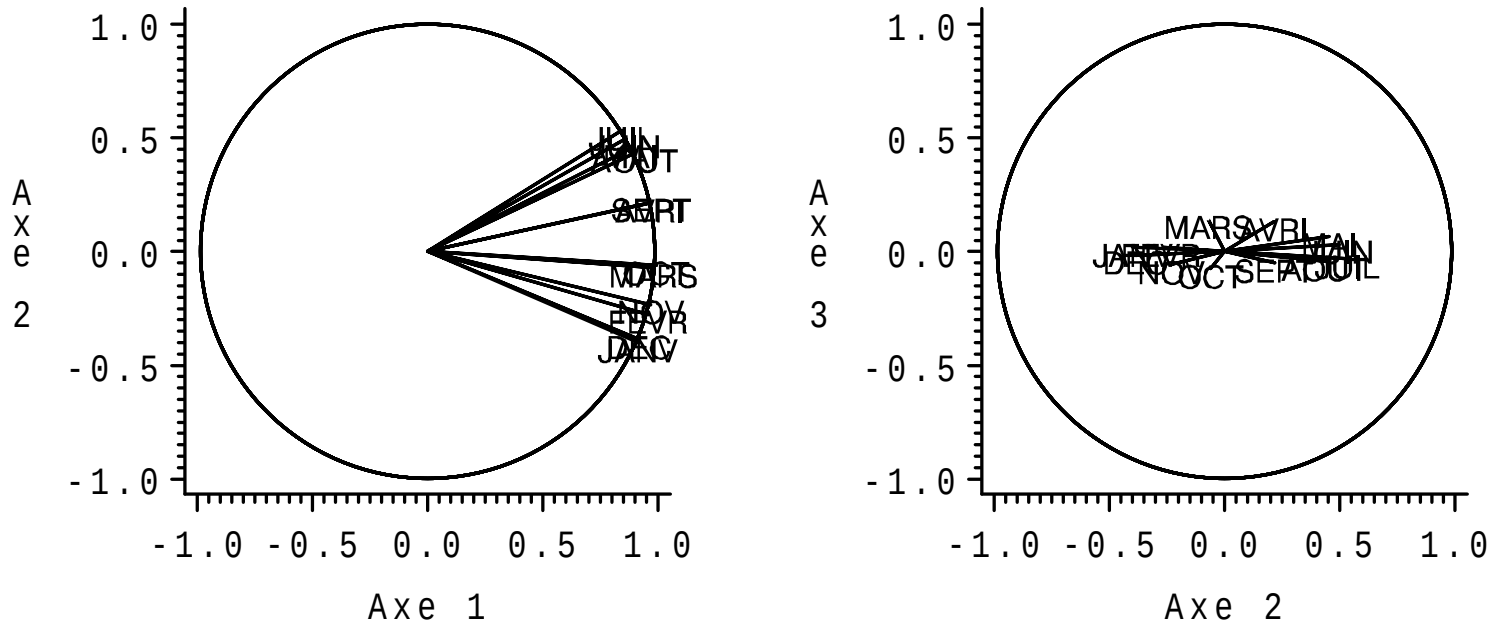

FIG.: *Températures : Premier et deuxième plan des variables.*

Représentation simultanée :

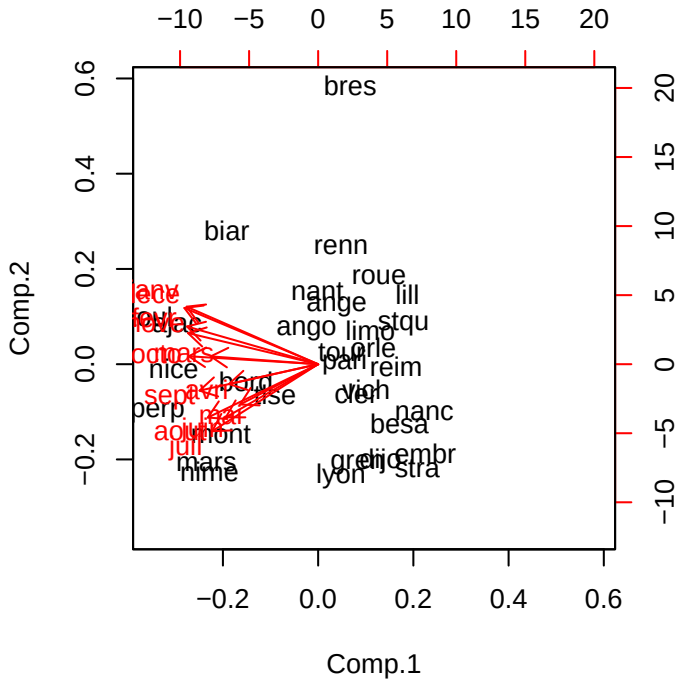

FIG.: Températures : Représentation simultanée ou biplot du premier plan.

Interprétation :

- ♣ Comment peut on interpréter les premiers axes par rapport aux variables,individus ? Est-il intéressant ?
- ♣Combien d'axes retenir ?

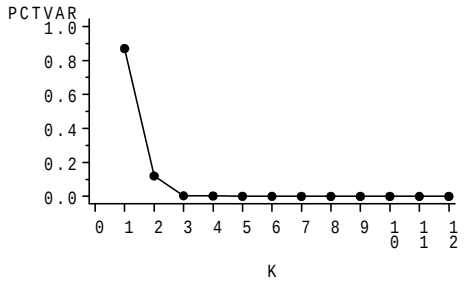

FIG.: *Températures : éboulis des valeurs propres.*