

TP EX12 : Tableau de notes

Pour une réaliser cet exercice, il faudra utiliser le code source compagnon de cet exercice. Si nécessaire, on pourra se reporter au document « ACPrappels.pdf ».

Ici, on va effectuer une ACP, on va s'intéresser aux étapes de la réalisation et de l'interprétation sur un jeu de données relativement simple. L'objectif de cet exercice est d'appliquer directement la méthode pour bien comprendre ce que représentent les différentes grandeurs calculées. Cette étape permettra d'appréhender les aspects essentiels à la réalisation de l'interprétation.

Les Données : Il s'agit d'un tableau de notes de 9 élèves dans 5 matières (données contenues dans le fichier **notes.dat**). Les élèves sont les individus, et les matières correspondent aux variables.

Noms	math	phys	fran	latin	dessin
Alain	6,0	6	5	5,5	8
Benoit	8,0	8	8	8	9
Cyril	6,0	7	11	9,5	11
Daisy	14,5	14,5	15,5	15	8
Emilie	14,0	14	12	12,5	10
Fanny	11,0	10	5,5	7	13
Gaétan	5,5	7	14	11,5	10
Hélène	13,0	12,5	8,5	9,5	12
Inès	9,0	9,5	12,5	12	18

Les noms des élèves ne sont pas contenus dans le fichier. Le programme doit donc associer un nom à chacune des lignes du tableau.

a) Représentation des Données

Reproduire le tableau ci-dessus et indiquez les moyennes par élève et par matière et dresser le tableau des corrélations. Par une représentation graphique on cherchera à faire ressortir des informations sur les données.

ACP sur données centrées et/ou réduites

- On veut réaliser une ACP sur les données centrées (par variable). Si on procède de même avec les données centrées réduites. Que cela change-t-il ? Expliquer.
- Soit **X** la **matrice de données centrées**. Le calcul des valeurs et des vecteurs propres peut ainsi être entrepris sur la matrice d'inertie $X^T X$. Que représentent les valeurs propres dans ce cas ?
- On considère les valeurs propres selon un ordre décroissant. Expliquer pourquoi.

b) Inertie.

- Faire une figure sur laquelle les pourcentages d'inertie des valeurs propres seront représentés ainsi que les inerties cumulées. Pourquoi s'intéresser aux inerties cumulées ?
- On cherche le plus souvent à limiter le nombre de dimensions de la représentation des données. Quel(s) critère(s) utiliser ? Les présenter.
- Indiquer le nombre de dimensions approprié à ce problème. Expliquer.
- Pour poursuivre l'exercice, on devra calculer les nouvelles variables (X_{ov}) ainsi que les nouvelles coordonnées des individus (X_u).

c) Qualités de représentation et contributions des individus.

Pour avoir des éléments supplémentaires d'appréciation des résultats en vue d'une interprétation, on doit calculer et afficher les qualités de représentation (qlt) des individus et leurs contributions (ctr) à la fabrication des axes. On rappelle les formules :

- La qualité de représentation d'un individu (\mathbf{o}_i), de norme $\|\mathbf{o}_i\|$, par un axe k est donnée par :

$$qlt_k(\mathbf{o}_i) = c_{ik}^2 / \|\mathbf{o}_i\|^2 \quad \text{avec } c_{ik} \text{ la coordonnée de l'individu } i \text{ sur l'axe } k.$$

L'individu i sera d'autant mieux représenté sur l'axe k , que cette valeur est proche de 1.

- La contribution d'un individu (\mathbf{o}_i) à la fabrication d'un axe k . C'est la part d'inertie de l'axe k prise en compte ou expliquée par l'individu i . Elle est donnée par :

$$ctr_k(\mathbf{o}_i) = q_i c_{ik}^2 / \lambda_k \quad \text{où } q_i \text{ est le poids de l'individu } i, \text{ et } \lambda_k \text{ l'inertie de l'axe } k.$$

Dans notre cas, où l'on recherche les valeurs et vecteurs propres de la matrice $\mathbf{X}\mathbf{X}^T$ où \mathbf{X} est simplement centrée, c'est-à-dire dont les données n'ont pas été pondérées, les poids q_i valent 1 quelque soit i .

On vérifiera que la somme des contributions pour chaque axe vaut 100% ainsi que la somme des qualités de représentation par individu.

- Pour chacun des axes identifier les individus ayant une contribution importante. Une contribution importante à un axe peut-elle être problématique ? Expliquer.
- Identifier les individus les mieux représentés pour chacun des trois premiers axes.

d) Etudes des représentations

On s'intéresse aux 3 axes qui cumulent le plus d'inertie pour comprendre les données. Pour chacun des plans formés en prenant ces axes deux à deux on présentera **le nuage des individus** et **le nuage des variables**.

► Faire faire une analyse détaillée de l'ensemble des données.

- Que dire de deux variables proches dans un nuage des variables ? Expliquer.
- Se servir des projections du nuage des variables pour interpréter chacun des axes.
- Pour chacun de ses axes, on devra considérer les coordonnées des individus les mieux représentés. Ces coordonnées permettent-elles de caractériser les individus.
- Que dire de deux individus proches dans un nuage des individus ? Expliquer.
- Les projections des nuages des individus permettent-elles de faire ressortir une typologie des individus ? Expliquer.