Épreuve de TIPE – Partie D

Titre : Analyse en composantes principales

Temps de préparation : 2h15

Temps de présentation devant le jury : 10 minutes

Entretien avec le jury : 10 minutes

Guide pour le candidat

Le dossier ci–joint comporte :

- Document principal : 9 pages
- Documents complémentaires : 2 pages

Travail suggéré au candidat : Expliquer les principes et objectifs de la méthode d'Analyse en Composantes Principales. Insister sur la façon dont les outils d'algèbre linéaire permettent de formaliser les aspects intuitifs du problème.

Conseils généraux pour la préparation de l'épreuve :

- Lisez le dossier en entier en un temps raisonnable.
- Réservez du temps pour préparer l'exposé devant le jury.

L'Analyse en Composantes Principales (ACP) Normées propose une solution au problème de visualisation d'un nuage dans un espace de dimension finie. En effet, l'ACP permet d'obtenir une représentation du nuage de points dans un sous-espace de dimension faible. L'idée simple consiste à visualiser le nuage à travers sa projection orthogonale sur une famille de plans (ou sous-espaces de dimension 3). Dans le cadre de l'ACP, on choisit les plans (le système d'axes correspondant) de sorte que l'opération de projection déforme le moins possible les distances entre les différents points du nuage. En d'autres termes, il faudra que l'inertie du nuage (à définir) sur le sous-espace en question soit maximale.

1 Préliminaires

1.1 Définitions de base

On observe p variables sur n individus. Les résultats forment une matrice (n, p) notée X. L'observation de la j-ième variable sur le i-ème individu sera notée $x_{i,j}$. On identifie la variable j à la colonne de X correspondante et on la notera $\vec{x}_j$; on identifiera également l'individu i à la ligne de X correspondante et on le notera $\vec{e}_i$ (tout ceci relativement aux bases canoniques).

On introduit la matrice de poids D par $d_{i,j} = p_i \delta_{i,j}$ où p_i peut être s'identifier à une fréquence d'apparition de l'individu i et $\delta_{i,j} = 1$ si $i = j$ et 0 sinon. Un cas usuel correspond à $D = \frac{1}{n} I$ avec I la matrice identité (tous les individus ont le même poids).

On appelle centre de gravité le vecteur $\vec{g} = {}^t(\bar{x}_1, \ldots, \bar{x}_p)$ [1] des moyennes arithmétiques pondérées de chaque variable j ($\bar{x}_j = \sum_{i=1}^{n} p_i x_{i,j}$). En désignant par $\vec{1}_n$ le vecteur de $\mathbb{R}^n$ dont toutes les composantes sont égales à 1,

[1] ${}^t\vec{x}$ désigne la transposée de $\vec{x}$ (idem pour les matrices)

on a :

$$\vec{g} \;=\; {}^t X \; D \; \vec{1}_n$$

On introduit alors le tableau centré associé au tableau X par :

$$Y \;=\; X \;-\; \vec{1}_n \; {}^t\vec{g}$$

On appelle variance une mesure de la dispersion d'une variable autour de
30 sa moyenne et s'écrit :

$$s_j^2 \;=\; \sum_{i=1}^{n} \; p_i \; (x_{i,j} - \bar{x}_j)^2$$

s_j s'appelle l'écart-type. La covariance de deux variables $\vec{x}_j$ et $\vec{x}_k$ s'écrit :

$$cov(\vec{x}_j, \vec{x}_k) \;=\; \sum_{i=1}^{n} \; p_i \; (x_{i,j} - \bar{x}_j)(x_{i,k} - \bar{x}_k)$$

cette quantité est une mesure de la liaison entre les deux variables. Il est parfois préférable d'introduire le coefficient de corrélation linéaire (quantité sans dimension et comprise entre -1 et 1) entre deux variables :

$$cor(\vec{x}_j, \vec{x}_k) \;=\; \frac{cov(\vec{x}_j, \vec{x}_k)}{s_j \; s_k}$$

35 Ce coefficient est un indicateur de la liaison linéaire entre deux variables.

La matrice de variance-covariance $V = (cov(\vec{x}_j, \vec{x}_k))_{k,j=1}^{p}$ s'écrit :

$$\begin{aligned} V \;&=\; {}^t Y \; D \; Y \\ &=\; {}^t X \; D \; X \;-\; \vec{g} \; {}^t\vec{g} \end{aligned}$$

On introduit la matrice (n,n) $D_{\frac{1}{s}}$ définie par $d_{\frac{1}{s}}(i,j) \;=\; \frac{1}{s_i} \; \delta_{i,j}$ et le tableau des variables centrées réduites Z :

$$Z \;=\; Y \; D_{\frac{1}{s}}$$

On appelle R la matrice des coefficients de corrélation :

$$\begin{aligned} R \;&=\; D_{\frac{1}{s}} \; V \; D_{\frac{1}{s}} \\ &=\; {}^t Z \; D \; Z \end{aligned}$$

1.2 Le nuage des individus

1.2.1 L'espace des individus

Comme il y a p variables, chaque individu est défini par p coordonnées. Il peut être considéré comme un élément de l'espace vectoriel $\mathbb{R}^p$, espace des individus. L'ensemble des n individus forme un nuage de points de centre de gravité $\vec{g}$. Étant donné une matrice définie positive symétrique M de taille p, nous allons munir $\mathbb{R}^p$ du produit scalaire :

$$< \vec{e}_i, \vec{e}_j >_M = {}^t\vec{e}_i \; M \; \vec{e}_j$$

Remarque 1 – *Le choix naturel $M = I$ donne le produit scalaire usuel qui n'est pas souvent judicieux ici. On préférera $D_{\frac{1}{s^2}}$ qui présentera l'avantage de donner à chaque variable la même importance (variance unité).*

– *Il est à noter que travailler avec les données brutes (tableau X) et $M = D_{\frac{1}{s^2}}$ est équivalent à travailler avec les données centrées réduites (tableau Z) et $M = I$.*

– *Dans la suite on notera $< \ , \ >$ le produit scalaire $< \ , \ >_{D_{\frac{1}{s^2}}}$ si l'on travaille avec X et $< \ , \ >_I$ si l'on travaille avec Z. Du coup, il n'y a aucune question à se poser sur le produit scalaire.*

1.2.2 La notion d'inertie

On appelle inertie au point $\vec{a}$ du nuage de point, $I(\vec{a})$ la moyenne pondérée des carrés des distances des individus au point $\vec{a}$:

$$I(\vec{a}) \; = \; \sum_{i=1}^{n} \; p_i \; < \vec{e}_i - \vec{a} \; , \; \vec{e}_i - \vec{a} >_M$$

On a les relations suivantes :

$$I(\vec{a}) \; = \; I(\vec{g}) + \; < \vec{g} - \vec{a} \; , \; \vec{g} - \vec{a} >_M \qquad \text{Relation de Huyghens} \qquad (1)$$

$$I(\vec{g}) \; = \; trace(MV) \qquad (2)$$

Par conséquent, si $M = I$, $I(\vec{g}) \; = \; \sum_{i=1}^{n} s_i^2$ et si $M = D_{\frac{1}{s^2}}$, $I(\vec{g}) \; = \; p$.

Lemme 1 *Si deux espaces F et G sont orthogonaux, alors, en notant I_G l'inertie du nuage projeté sur G, on a :*

$$I_{F\oplus G} \;=\; I_F \;+\; I_G$$

Propriété 1 *Soit F_k un sous-espace 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, orthogonal à F_k et portant l'inertie maximale.*

Idée de la démonstration :

On considère E_{k+1} un sous espace de dimension $k+1$. En raisonnant sur les dimensions, on prouve l'existence d'un vecteur $\vec{b}$ dans $E_{k+1} \cap F_k^{\perp}$. On considère Γ un supplémentaire orthogonal de $\vec{b}$ dans E_{k+1} et $\Phi = F_k \oplus \mathbb{R}\vec{b}$. Du fait que F_k soit d'inertie maximale $I_\Gamma \;\leq\; I_{F_k}$. Une application du lemme 1 montre que :

$$I_{E_{k+1}} \;\leq\; I_\Phi$$

et par conséquent, en choisissant $\vec{b}$ tel que $\mathbb{R}\vec{b}$ soit d'inertie maximale, on montre le résultat.

◀

1.2.3 Ajustement du nuage des individus

On cherche à représenter le nuage en le déformant le moins possible, c'est à dire que l'on cherche le sous-espace qui rend minimale la distance entre un point-individu M_i et son projeté H_i sur le sous-espace : on veut donc

$$\sum_{i=1}^{n} M_i {H_i}^2 \quad \text{minimale.} \tag{3}$$

Comme GM_i est fixe, on a

Propriété 2 *La condition* (3) *revient à la condition que l'inertie au centre de gravité du nuage projeté soit maximale ($\sum_{i=1}^{n} {GH_i}^2$ maximale).*

Ceci signifie aussi que la distance entre les projetés est la plus proche possible de la distance entre les points.

En vertu de ce résultat, la question revient à rechercher un sous-espace d'inertie maximale. Or, le résultat 1 nous donne une méthode itérative de recherche du sous-espace en question. En effet, si on sait déterminer un sous espace de dimension 1, on saura trouver un sous-espace de dimension quelconque.

Propriété 3 *La droite de $\mathbb{R}^p$ passant par $\vec{g}$ et maximisant l'inertie du nuage projeté est engendrée par un vecteur propre associé à la plus grande valeur propre de R.*

Idée de la démonstration :

Soit $\vec{a}$ un vecteur unitaire ($< \vec{a}, \vec{a} >= 1$) porté par la droite en question. On remarque d'abord que :

$$\begin{aligned} \sum_{i=1}^{n} {GH_i}^2 &= < Z\vec{a}, Z\vec{a} > \\ &= {}^t\vec{a}\ {}^tZZ\vec{a} \end{aligned}$$

Par conséquent, pour trouver $\vec{a}$ il faut trouver le maximum de :

$${}^t\vec{a}\ {}^tZZ\ \vec{a} \qquad \text{sous la contrainte} \quad {}^t\vec{a}\vec{a} = 1$$

Cela peut se faire (mais ce n'est pas demandé) en utilisant les propriétés de la matrices ${}^tZZ$, qui est symétrique définie positive.

◀

1.3 Le nuage des variables

1.3.1 Espace des variables

Chaque variable $\vec{x}_j$ est une liste de n valeurs numériques, on la considérera comme un point de $\mathbb{R}^n$: l'espace des variables. Le choix du produit scalaire :

$$\begin{aligned} < \vec{x}_j, \vec{x}_k >_D &= {}^t\vec{x}_j\ D\ \vec{x}_k \\ &= cov(\vec{x}_j, \vec{x}_k) \end{aligned} \tag{4}$$

est motivé par la relation (4) puisque, pour des données centrées :

- la longueur coïncide avec l'écart-type ($< \vec{x}_j, \vec{x}_j >_D = s_j^2$) ;
- l'angle $\theta_{i,j}$ entre les variables $\vec{x}_j$ et $\vec{x}_k$ est caractérisé par

$$cos(\theta_{j,k}) \ = cor(\vec{x}_j, \vec{x}_k)$$

le coefficient de corrélation linéaire.

1.3.2 Ajustement du nuage des variables

La démarche est rigoureusement la même et consiste dans un premier temps à rechercher le vecteur unitaire $\vec{b}$ qui ajuste au mieux le nuage de points. Cela conduit à nouveau à rendre maximale la somme des carrés des p projections sur $\vec{v}$ qui sont les p composantes du vecteur ${}^tZ\vec{v}$. On est donc contraint à maximiser la quantité :

$$\vec{v}\ {}^tZZ\ {}^t\vec{v} \qquad \text{avec la contrainte} \ \ \vec{v}\ {}^t\vec{v} = 1$$

Comme précédemment, nous sommes amenés à retenir les q vecteurs propres associés au q plus grandes valeurs propres de ${}^tZZ$, matrice (n, n).

1.4 Formule de transition

Il y a un lien très fort entre les relations des deux parties précédentes qui débouche sur la propriété suivante :

Propriété 4 (Formule de transition entre les espaces $\mathbb{R}^p$ et $\mathbb{R}^n$) *Les matrices ${}^tZZ$ et $Z\ {}^tZ$ ont les mêmes valeurs propres et l'on a la relation suivante entre le α-ième vecteur propre unitaire $\vec{u}_\alpha$ de ${}^tZZ$ et $\vec{v}_\alpha$ vecteur propre unitaire de $Z\ {}^tZ$:*

$$\vec{v}_\alpha \ = \ \frac{1}{\sqrt{\lambda_\alpha}}\ Z\vec{u}_\alpha$$

2 L'analyse en composantes principales normée

2.1 Principe

120 On part d'un tableau X de taille (n, p) de données brutes. On centre les variables et on les réduit. On obtient un nouveau tableau Z des données centrées réduites :

$$z_{i,j} \;=\; \frac{x_{i,j} - \bar{x}_j}{\sqrt{n}\; s_j} \qquad \left(\vec{z}_j \;=\; \frac{\vec{x}_j - \bar{x}_j \vec{1}_n}{\sqrt{n}\; s_j}\right)$$

(voir définitions page 3). L'introduction du $\sqrt{n}$ est pratique parce que l'on a :

$${}^t\vec{z}_j \vec{z}_j \;=\; \frac{var(\vec{x}_j)}{s_j^2} \;=\; 1$$

Analyse des individus Conformément au paragraphe 1.2, on doit diagonaliser la matrice ${}^tZZ$ qui n'est autre que la matrice des corrélations.

125

On appelle α-ième axe principal le α-ième vecteur propre unitaire $\vec{u}_\alpha$ de ${}^tZZ$ associé à la valeur propre λ_α classées dans l'ordre décroissant. Le vecteur des coordonnées des n points individus sur l'axe principal $\vec{u}_\alpha$ sont données par :

$$\vec{\psi}_\alpha \;=\; Z\; \vec{u}_\alpha$$

On a les propriétés suivantes :

$$\sum_{i=1}^{n} \psi_{\alpha i} \;=\; 0 \tag{5}$$

$$Var(\vec{\psi}_\alpha) \;=\; \lambda_\alpha \tag{6}$$

130 On représente enfin les individus dans les divers plans $(\vec{u}_{\alpha_1}, \vec{u}_{\alpha_2})$.

Analyse des variables Conformément au paragraphe 1.3, on doit diagonaliser la matrice $Z\ {}^tZ$. Les formules de transition nous assure que c'est

inutile, elle se déduit du paragraphe précédent.

On appelle α-ième composante principale le α-ième vecteur propre $\vec{v}_\alpha$ de $Z\ {}^tZ$ associé à la valeur propre λ_α classées dans l'ordre décroissant. Le vecteur des coordonnées des p points variable sur la composante principale $\vec{v}_\alpha$ sont données par :

$$\vec{\phi}_\alpha \ = \ {}^tZ\ \vec{v}_\alpha$$

On a les propriétés suivantes :

$$\phi_{\alpha,j} \ = \ cor(\vec{z}_j, \vec{\psi}_\alpha) \tag{7}$$

On représente enfin les individus dans les divers plans $(\vec{v}_{\alpha_1}, \vec{v}_{\alpha_2})$. Les coordonnées étant des coefficients de corrélation, les points sont à l'intérieur d'un disque de rayon 1 appelé cercle des corrélations.

2.2 Qualité de la représentation d'un individu

La qualité de représentation de l'individu i par l'axe $\vec{u}_\alpha$ est mesurée par le rapport :

$$QLT_\alpha(i) \ = \ \frac{\text{inertie de la projection de } i \textit{ sur } \vec{u}_\alpha}{\text{inertie totale de } i}$$

c'est le cosinus carré de l'angle entre l'axe $\vec{u}_\alpha$ et le vecteur reliant le centre de gravité et l'individu i.

2.3 Reconstitution

On peut reconstruire le tableau des données centrées réduites Z et donc S au moyen des composantes principales et des axes principaux :

Théorème 1 *On a :*

$$Z \ = \ \sum_{\alpha=1}^{p} \ \sqrt{\lambda_\alpha}\ \vec{v}_\alpha\ {}^t\vec{u}_\alpha$$

3 Exemple

3.1 Résultats numériques (annexe 1)

L'exemple considéré est basé sur le poids, la taille, l'âge et la note moyenne de 10 élèves de quatrième. Il est issu du livre de T. Foucart « L'analyse de données (P.U.R., 1997) ». Vue la taille de l'échantillon considéré, cet exemple est purement pédagogique et n'a aucune vocation sociologique.

Les données, les moyennes et écarts-type sont consignés dans l tableau 1. Le tableau 2 correspond aux données centrées réduites (Z). Le tableau 3 est la matrice de corrélation $\frac{1}{10}\ {}^tZZ$. La diagonalisation de cette matrice fournie la matrice diagonale Λ et la matrice de passage U des tableaux 4 et 5. On remarque également les contributions des valeurs à la variation totale et l'on peut tout de suite noter que les deux premières composantes représentent presque 80 % de l'inertie du nuage original. On déduit le tableau des coordonnées des individus dans la base des vecteurs propres (tableau 6) autrement dit le tableau des coordonnées des individus sur les différents axes. La qualité de la représentation des individus selon ces axes figurent dans le tableau 7. Enfin le tableau 8 est la matrice de corrélation entre les composantes principales et les variables.

3.2 Résultats graphiques (annexe 2)

Le tableau 6 conduit au diagramme de projection sur le plan principal (1,2) (figure 1).
Le tableau 8 débouche sur le cercle des corrélations (figure 2).

3.3 Quelques éléments d'interprétation

Tout d'abord il faut remarquer que la représentation par un sous-espace de dimension 2 est bonne puisque les deux premières composantes représentent presque 80 % de l'inertie du nuage original.

3.3.1 Le plan principal (1,2)

Cette représentation permet de grouper les individus selon des caractéristiques communes.

L'axe 1 oppose un premier groupe composé des individus A,B,C et un deuxième groupe composé des individus D,E,F,G les individus H et I sont eux mal représentés par l'axe 1.

L'axe 2 sert principalement la caractéristique de l'individu I.

C'est résultats sont confortés par les résultats du tableau 7.

3.3.2 Le cercle des corrélations

Le cercle de corrélation nous donne des informations sur les liens entre les différentes variables :

Tout d'abord il est bon de noter que les variables sont proches de la périphérie ce qui indique une bonne représentation des variables sur le plan des composantes principales (1,2). On remarque aussi que les direction des points-variables sont proches de la composante 1 ce qui met en évidence la forte contribution de cette composante au phénomène (résultat conforté par la contribution de la valeur propre).

La première composante met en évidence l'opposition entre Note d'une part et les caractéristiques physiques d'autre part.

La seconde composante met en quelque sorte l'accent sur l'opposition entre les variables Age, Poids et Note d'une part et Taille de l'autre.

3.3.3 Compilation des résultats des deux graphiques

Il suffit au moyen des composantes de faire le lien entre les individus et les variables. Par exemple l'individu I qui monopolise la composante 2 met en évidence un individu plutôt petit, âgé assez lourd avec de bons résultats (ou l'inverse, le sens est reconnu par les données initiales).