Devoir maison n° 7
BCST2

23 mars 2024

1 Problème : analyse en composantes principales

Dans ce problème, on s'intéresse à la méthode de l'analyse en composantes principales pour l'étude des données multidimensionnelles. L'analyse en composantes principales consiste à projeter des données sur un plan en conservant le maximum d'information. On part d'un exemple à deux dimensions portant sur deux caractères distincts d'une population de n individus avant de passer au cas réel suivant : En 1990, on a réalisé sur un terrain en jachère des relevés pour l'analyse chimique du sol. Six propriétés sont relevées : le pH (ph), la conductivité électrique (CE), la teneur en carbone oxydable (C), l'humidité pondérale (Hum), la teneur en NH_4^+ échangeable (NH4) et en azote potentiellement minéralisable (Nmin). On souhaite trouver des relations entre ces différentes mesures. Les analyses se font en 169 points distincts du terrain, par des outils de mesure différents : comment interpréter ces résultats? Le grand nombre de dimensions et le grand nombre de points de relevés rend l'étude à la main difficile.

Notations

— On notera ${}^t\mathrm{A}$ la transposée de la matrice A.

— On munit les espaces vectoriels $\mathbb{R}^n$ de leur base canonique $\mathscr{B} = (u_1, \cdots, u_n)$, de leur produit scalaire canonique, noté $< .,. >$ et de la norme associée notée $||\cdot||$.

Partie I/ Etude d'un exemple.

On considère les vecteurs v_1, v_2 et v_3 de $\mathbb{R}^2$ et dont les coordonnées dans la base canonique sont respectivement $(1,2)$, $(-3,-1)$ et $(2,-1)$.

1. On considère un réel m et on note pour tout $i \in [\![1,3]\!]$, v_i' le projeté orthogonal de v_i sur la droite vectorielle F_m engendrée par $w_m = u_1 + mu_2$.
   (a) Donner un vecteur a_m unitaire (de norme égale à 1) qui dirige la droite F_m.

   (b) Donner les coordonnées de v_i' dans la base canonique pour tout $i \in [\![1,3]\!]$.

   (c) Calculer en fonction de m la quantité : $\|v_1'\|^2 + \|v_2'\|^2 + \|v_3'\|^2$.

   (d) Déterminer la valeur m_0 de m pour laquelle cette quantité atteint son maximum. Ce maximum est noté λ_1.

2. On considère la matrice $\mathrm{X} = \begin{pmatrix} 1 & 2 \\ -3 & -1 \\ 2 & -1 \end{pmatrix}$.
   (a) Vérifier que λ_1 est valeur propre de la matrice $\mathrm{R} = {}^t\mathrm{X}\cdot\mathrm{X}$ et que $w_{m_0} = u_1 + m_0 u_2$ est un vecteur propre associé à λ_1.

   (b) Déterminer l'autre valeur propre de R et la comparer à λ_1.

3. On pose $\mathrm{M}_1 = (x_1, y_1) = (1,2)$, $\mathrm{M}_2 = (x_2, y_2) = (-3,-1)$ et $\mathrm{M}_3 = (x_3, y_3) = (2,-1)$.
   (a) Donner les coefficients de la droite de régression linéaire, dite aussi « d'ajustement affine », du nuage de points $\mathrm{M}_i = (x_i, y_i)$, $1 \leq i \leq 3$, c'est-à-dire les réels a et b qui rendent minimale la quantité :

$$\delta_{a,b} = \sum_{i=1}^{3} (y_i - ax_i - b)^2$$

(b) Le vecteur w_{m_0} dirige-t-il cette droite d'ajustement affine?

(c) Tracer sur un même graphique (avec Python si vous le souhaitez) le nuage de points, la droite de régression linéaire ainsi que la droite vectorielle F_m.

Partie II/ Axe principal d'inertie d'un nuage de points centré

On généralise la méthode précédente en considérant n individus sur lesquels deux variables sont mesurées, une variable dite « explicative » et l'autre dite « expliquée ». On considère que les relevés sont rassemblés dans une matrice $X \in \mathcal{M}_{n,2}(\mathbb{R})$ avec en colonne les 2 mesures (on parlera d'attributs) et en ligne les n individus. On note x_i le vecteur individu de $\mathbb{R}^2$ représenté par la ligne i et x^j le vecteur attribut de la colonne j ($1 \leq j \leq 2$), à savoir le vecteur de $\mathbb{R}^n$ dont les coordonnées sont les mesures prises par cet attribut pour chacun des individus.
L'attribut j de l'individu i est noté x_i^j et en conséquence $x_i = (x_i^1, x_i^2)$ pour tout $i \in [\![1, n]\!]$
Les valeurs de relevés x^1 et x^2 sont considérées centrées, c'est à dire que la moyenne des valeurs d'une colonne vaut zéro (on rappelle que pour centrer une série statistique x, il suffit de faire $x - \overline{x}$... aussi on peut donc facilement se ramener à cette hypothèse si jamais ce n'était pas le cas).
soit :

$$\sum_{i=1}^{n} x_i^1 = 0 \text{ et } \sum_{i=1}^{n} x_i^2 = 0$$

et

$$X = \begin{pmatrix} x_1^1 & x_1^2 \\ x_2^1 & x_2^2 \\ \vdots & \vdots \\ x_n^1 & x_n^2 \end{pmatrix}$$

On utilisera par ailleurs les notations usuelles suivante :

$$\forall 1 \leq j \leq 2, \quad \overline{x^j} = \frac{1}{n}\sum_{i=1}^{n} x_i^j, \quad \overline{(x^j)^2} = \frac{1}{n}\sum_{i=1}^{n} (x_i^j)^2 \text{ et } \sigma_{x^j}^2 = \overline{(x^j)^2} - \left(\overline{x^j}\right)^2$$

$$\overline{x^1 x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^1 x_i^2 \text{ et } \mathrm{Cov}(x^1, x^2) = \overline{x^1 x^2} - \overline{x^1} \cdot \overline{x^2}$$

1. Soit $R = {}^tX \cdot X$. Exprimer R à l'aide de la matrice $S = \begin{pmatrix} \overline{(x^1)^2} & \overline{x^1 x^2} \\ \overline{x^1 x^2} & \overline{(x^2)^2} \end{pmatrix}$.

2. Montrer que R est diagonalisable et que ses valeurs propres sont des réels positifs ou nuls (*remarque :* pour montrer que $\mathrm{Sp}(R) \subset \mathbb{R}_+$, on pourra calculer ${}^tWRW$ de deux façons différentes où w désigne un vecteur propre de R associé à la valeur propre λ et W sa matrice dans la base canonique...).
On note λ_1, λ_2 les valeurs propres de R telles que $\lambda_1 \geq \lambda_2$.
Justifier l'existence d'une base orthonormale $\mathscr{B}' = (w_1, w_2)$ de $\mathbb{R}^2$ tq : $R \cdot W_1 = \lambda_1 W_1$ et $R \cdot W_2 = \lambda_2 W_2$ où $W_i = \mathscr{M}_{\mathscr{B}}(w_i)$

3. Pour tout vecteur $u \in \mathbb{R}^2$, on pose :
   — $F_u = \mathrm{Vect}\{u\}$ la droite vectorielle engendrée par u.
   — p_u la projection orthogonale sur F_u.
   — $I(u) = \sum_{i=1}^{n} \left\| p_u(x_i) \right\|^2$. Cette quantité s'appelle l'inertie du nuage de points $\{x_i, 1 \leq i \leq n\}$ sur la droite F_u.

   (a) Pour tout vecteur **unitaire** $u \in \mathbb{R}^2$ tel que $U = \mathscr{M}_{\mathscr{B}}(u)$, on veut montrer que :

   $$I(u) = {}^tURU$$

   i. Montrer que $XU = \begin{pmatrix} (x_1|u) \\ (x_2|u) \\ \vdots \\ (x_n|u) \end{pmatrix}$.

ii. Rappeler l'expression de $||p_u(x_i)||$ en fonction du produit scalaire de x_i par u.

iii. En déduire que $\mathrm{I}(u) = {}^t\mathrm{URU}$.

(b) Exprimer $\mathrm{I}(w_1)$ et $\mathrm{I}(w_2)$ à l'aide de λ_1 et λ_2.
On écrit $u = \sum_{i=1}^{2} \mu_i w_i$. Justifier cette écriture et montrer que $\mathrm{I}(u) = \sum_{i=1}^{2} \lambda_i \mu_i^2$.

(c) Pour tout vecteur $u \in \mathbb{R}^2$, de norme égale à 1, montrer que :

$$0 \leq \mathrm{I}(u) \leq \lambda_1$$

Quelle est la valeur de $\max\{\mathrm{I}(u) / u \in \mathbb{R}^2, \|u\| = 1\}$ (maximum de l'ensemble des $\mathrm{I}(u)$ lorsque u décrit l'ensemble des vecteurs de norme 1 dans $\mathbb{R}^2$) ? Pour quel vecteur u_0 ce maximum est-il atteint ?

Remarque : **La droite vectorielle engendrée par u_0 est appelée *axe principal d'inertie* du nuage de points.**

4. On appelle *inertie du nuage de points* la quantité :

$$\mathscr{I} = \sum_{i=1}^{n} \|x_i\|^2$$

(a) Pour tout $u \in \mathbb{R}^2$, de norme égale à 1, montrer :

$$\mathscr{I} = \mathrm{I}(u) + \sum_{i=1}^{n} \left\| x_i - p_u(x_i) \right\|^2$$

(b) En quoi le choix de u_0 permet d'affirmer que la droite F_{u_0} réalise une bonne approximation du nuage de points ? Cette droite coïncide-t-elle avec la droite d'ajustement affine du nuage ?

5. **Calcul effectif de l'axe principal d'inertie dans le cas d'un nuage de points non centrés.**
Dans cette question, le point moyen du nuage de points n'est pas nécessairement confondu avec l'origine du repère et on note les n individus $y_i = (y_i^1, y_i^2)$. On garde les notations statistiques introduites dans cette partie et on procède de la façon suivante :
   - Pour tout $i \in [\![1, n]\!]$, on pose $x_i^1 = y_i^1 - \overline{y^1}$ et $x_i^2 = y_i^2 - \overline{y^2}$, avec $x_i = (x_i^1, x_i^2)$.
   - On considère $u_0 \in \mathbb{R}^2$ tel que F_{u_0} soit l'axe principal d'inertie du nuage de points $\{x_i, i \in [\![1, n]\!]\}$.
   - L'axe principal d'inertie du nuage de points $\{y_i, i \in [\![1, n]\!]\}$ sera la droite passant par le point $(\overline{y^1}, \overline{y^2})$ et dirigée par u_0.

On suppose que les bibliothèques `numpy` et `matplotlib.pyplot` sont importées à l'aide de la commande :

```
import numpy as np
import matplotlib.pyplot as plt
```

Chaque programme doit être commenté par une phrase détaillant le raisonnement qui a conduit à son élaboration. On pourra utiliser les objets de type `matrix` *disponibles dans la bibliothèque numpy mais on s'interdit toute fonction de calcul matriciel à l'exception des opérations* $\star$ *et* $+$ *et d'une fonction* `transpose(A)` *ou* `A.T` *qui retourne la transposée d'une matrice* A.

6. Écrire une fonction `Norme(X)` d'argument une matrice colonne X de taille quelconque et qui renvoie le nombre ||X||.

7. Écrire une fonction `Normalise(v)` d'argument une matrice colonne $v \in \mathscr{M}_{n,1}(\mathbb{R})$ non nulle renvoie une nouvelle matrice colonne $\tilde{v} = v/||v||$.

8. Écrire une fonction python `diagonalise(S)` qui prend en argument une matrice carrée symétrique d'ordre 2 et qui donne en sortie les valeurs propres de S ainsi que des vecteurs propres normés associés.

9. Écrire une fonction `pointmoyen(X)` qui prend en argument une liste de valeurs réelles `X=[x_1,...,x_n]` et qui donne en sortie la valeur $\overline{x}$.

10. Écrire une fonction python `Delta(X1,X2)` qui prend en argument deux listes de de nombres de même longueur et qui donne en sortie la matrice R définie dans ce début de partie.

11. Dans cette question, l'utilisation des fonctions `max` et `min` n'est pas autorisée. Écrire une fonction `axe(Y1,Y2)` qui prend en entrée deux listes de nombres de même longueur (pas nécessairement centrées), qui donne en sortie le vecteur principal d'inertie du nuage de points associé à Y_1 et Y_2 et qui trace dans un même graphique ce nuage de points, ainsi que son axe principal d'inertie et la droite de régression pour laquelle on autorisera l'appel à la fonction `np.polyfit(Y1,Y2,1)`.

Partie III/ Principe général de l'ACP

3.1 Formalisation de la méthode

Soit $p \geq 2$. On considère que les relevés se trouvent dans un tableau (ou matrice) X avec en colonne les p différentes mesures (on parlera d'attributs) et en ligne les n différents points de relevés (on parlera d'individus). Comme précédemment, on note x_i le vecteur individu de $\mathbb{R}^p$ représenté par la ligne i et x^j le vecteur attribut de la colonne j, à savoir le vecteur de $\mathbb{R}^n$ dont les coordonnées sont les mesures prises par cet attribut pour chacun des individus. L'attribut j de l'individu i est noté x_i^j. Les valeurs de relevé sont considérées centrées, c'est à dire que la moyenne des valeurs d'une colonne vaut zéro.

1. Soit R la matrice carrée d'ordre p dont le coefficient (i,j) valent $r_{ij} = \sum_{k=1}^{n} x_k^i x_k^j$.
   Justifier qu'on parle par la suite pour R de « matrice de covariance des données ».

2. Montrer que $R = {}^t X X$.

3. Justifier que R est diagonalisable à valeurs propres réelles positives et qu'il existe une base orthonormée de vecteurs propres de R.

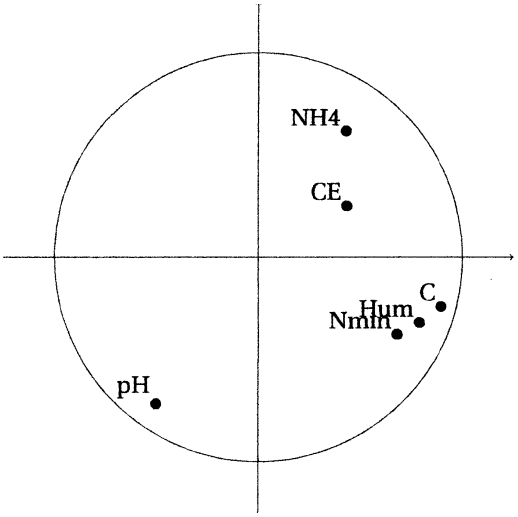
   On notera par la suite $\lambda_1, \lambda_2, \ldots, \lambda_p$ les p valeurs propres de R telles que $\lambda_1 \geqslant \ldots \geqslant \lambda_p$ et $w_1, \ldots, w_p$ les vecteurs propres orthonormés associés.
   On cherche dans un premier temps à trouver une droite F dirigée par un vecteur u unitaire, telle que la somme des carrés des distances des variables x^i à F = Vect$\{u\}$ soit minimum. Autrement dit on cherche u de norme 1 tel que Vect$\{u\}$ minimise la perte d'information...
   Remarque : Si on rappelle qu'on peut interpréter $p_u(x^i)$ comme la meilleure approximation de x^i par un vecteur de F, on a montré dans la partie II que cela revient aussi à maximiser la quantité :

   $$I(u) = \sum_{i=1}^{n} ||p_u(x_i)||^2$$

   où p_u est la projection orthogonale sur u.
4. En vous inspirant de la partie II, montrer que $I(u) = {}^t URU$ ou $U = \mathscr{M}_{\mathscr{B}}(u)$.

5. Montrer que $I(u)$ est majorée par λ_1 et que c'est une valeur maximale de $I(u)$ puisque atteinte pour u, vecteur propre associé à la valeur propre λ_1.

   On vient de trouver la direction sur laquelle projeter pour avoir le moins de perte d'information mais on veut désormais projeter sur un plan et on s'intéresse à la seconde direction maximale : On admettra que celle ci est donnée par un vecteur propre associé à λ_2.
   On obtient ainsi un plan sur lequel on peut projeter chaque individu, ou encore visualiser les projections des vecteurs qui représentent les mesures initiales. Cette projection des vecteurs initiaux forme ce que l'on appelle le cercle de corrélation. Typiquement, on obtient dans notre cas un cercle de corrélation représenté sur la figure ci-dessus.
   En observant ce cercle, on s'aperçoit que l'attribut NH_4^+ et l'attribut ph sont opposés, alors que les trois attributs C, Hum et Nmin sont presque confondus. Les deux directions ainsi définies sont approximativement orthogonales. En ACP, on dira que l'on obtient deux directions principales : la première est portée par les attributs NH4 et ph, la deuxième par les attributs C, Hum, Nmin.



6. Expliquer pourquoi il est naturel que les attributs ph et NH4 soient opposés ?

7. Comment interprétez-vous ces résultats : que représentent ces deux directions ?

3.2 Recherche de la plus grande valeur propre et du vecteur propre associé

Dans la méthode de l'analyse en composantes principales, on a besoin d'effectuer le calcul de la première et de la deuxième valeur propre, ainsi que des vecteurs propres associés. Dans la suite, on demandera d'écrire des algorithmes en Python qui permettent de réaliser cette recherche.
On se propose ici de présenter la méthode des puissances itérées pour le calcul de la première valeur propre et d'un vecteur propre associé. On se donne une matrice A symétrique réelle positive de taille n. On note $\lambda_1, \lambda_2, \ldots, \lambda_n$ les n valeurs propres de A telles que $\lambda_1 > \lambda_2 > \lambda_3 \geqslant \ldots \geqslant \lambda_n \geqslant 0$ et $w_1, w_2, \ldots, w_n$ une base orthonormée de vecteurs propres associés.
Soit v un vecteur quelconque normé de $\mathbb{R}^n$ non orthogonal à w_1 ni à w_2, on écrit $v = s_1 w_1 + \cdots + s_n w_n$ (ainsi $s_1 \neq 0$ et $s_2 \neq 0$).
On définit la suite de vecteurs v_k par :

$$\begin{cases} v_0 = v \\ v_{k+1} = \dfrac{1}{||\mathrm{A}v_k||}\mathrm{A}v_k \end{cases}$$

Remarque : Dans la suite de cette partie, on notera **de la même façon** les vecteurs v_k et leur représentation matricielle.

1. Soit $k \in \mathbb{N}$, que vaut $||v_k||$?

2. Montrer que pour tout $k \in \mathbb{N}$, $v_k = \dfrac{\mathrm{A}^k v}{||\mathrm{A}^k v||}$.

3. justifier que pour tout $k \in \mathbb{N}$, $\mathrm{A}^k v = \sum_{i=1}^{n} \lambda_i^k s_i w_i$.
   En déduire $\mathrm{A}^k v = \lambda_1^k s_1 (w_1 + \varepsilon_k)$ puis que $v_k = \mathrm{C}_k \lambda_1^k s_1 (w_1 + \varepsilon_k)$, où C_k est un réel dépendant de k que l'on précisera et ε_k est un vecteur dont la norme tend vers 0 et orthogonal à w_1.
   Ainsi, la direction de v_k tend vers la direction de w_1.

4. Montrer que pour tout $k \in \mathbb{N}$, $\left|\mathrm{C}_k \lambda_1^k s_1\right| (||w_1||^2 + ||\varepsilon_k||^2) = 1$.
   Quelle est la limite de $|\mathrm{C}_k \lambda_1^k s_1|$?

5. On suppose dans cette question uniquement que s_1 est positif.
   (a) En exprimant $v_k - w_1$, montrer qu'alors v_k converge vers w_1.
   (b) Soit $\mathrm{R}_\mathrm{A}(v) = \dfrac{{}^t v \mathrm{A} v}{{}^t v v}$. Exprimer $\mathrm{R}_\mathrm{A}(v_n)$ pour tout $n \in \mathbb{N}$ et déterminer $\lim\limits_{n\to\infty} \mathrm{R}_\mathrm{A}(v_n)$.

3.3 Calcul numérique de λ_1 et de w_1

1. Écrire une fonction `InitialeV(n)` qui, étant donné un entier n, renvoie une matrice colonne de $\mathcal{M}_{n,1}(\mathbb{R})$ dont les coefficients sont pris au hasard dans l'intervalle $[0,1]$.

   On se donne à présent une matrice $A \in S_n(\mathbb{R})$. Soit v_0 un élément quelconque de $\mathcal{M}_{n,1}(\mathbb{R})$. En supposant qu'aucun des termes n'est dans le noyau de A, on peut former la suite $(v_n)_{n\geq 0}$ de $\mathcal{M}_{n,1}(\mathbb{R})$ définie en I.2.

2. Écrire en Python une fonction `puissancesIterees(A,n)` qui étant donnée une matrice symétrique A et un entier naturel n, détermine la taille de A, choisit aléatoirement une matrice colonne $v_0 \in \mathcal{M}_{n,1}(\mathbb{R})$, puis calcule et renvoie la matrice colonne v_n (en supposant que tous les termes de la suite sont bien définis).
3. On se propose d'écrire maintenant une fonction `VecteurPropre(A,e)` qui étant donnée une matrice symétrique A d'ordre n et un nombre $e > 0$, calcule les termes de la suite $(v_n)_{n\geq 0}$ jusqu'à ce que deux termes successifs vérifient $||v_n - v_{n+1}|| < e$, et renvoie alors la matrice colonne v_{n+1}.
   On trouvera page suivante trois propositions de programmes. Indiquer lequel est (ou lesquels sont correct(s). Pour chaque programme *incorrect* on indiquera succinctement ce qui ne va pas.

4. Pour chaque fonction correcte, compléter le `return` afin que la fonction retourne non seulement le vecteur propre w_1 mais aussi la plus grande valeur propre λ_1. Tester votre programme sur deux matrices symétriques réelles positives de votre choix, l'une d'ordre 2 (par exemple la matrice R de la partie I) et l'autre d'ordre 3.

3.4 Recherche de la seconde plus grande valeur propre et du vecteur propre associé

1. On définit $B = A - \lambda_1 w_1 {}^t w_1$. Montrer que 0 est valeur propre de B et que les w_i sont aussi vecteurs propres pour $i \geqslant 2$, associés aux valeurs propres initiales. Quelle est la valeur propre maximale de B ?

2. En déduire une méthode de calcul de la deuxième valeur propre et d'un vecteur propre associé.

3. Écrire une fonction en Python nommée `deflation(A)` qui calcule cette deuxième valeur propre et un vecteur propre normé associé.

   Maintenant que l'on a accès aux deux vecteurs propres w_1, w_2, l'analyse en composante principales n'est plus qu'une question de projections... Sauriez-vous écrire le projeté des vecteurs x^j dans la base (w_1, w_2) ?

```
1  def VecteurPropre(A,e) :
2          d = A.shape
3          v = initialiseV(d[0])
4          v = Normalise(v)
5          w = Normalise(A*v)
6          while Norme(v-w)>=e :
7                  v = w
8                  w = Normalisee(A*v)
9          return w
```

```
1  def VecteurPropre(A,e) :
2          d = A.shape
3          v = initialiseV(d[0])
4          v = Normalise(v)
5          w = Normalise(A*v)
6          ecart = Norme(v-w):
7          while ecart>=e :
8                  v = w
9                  w = Normalise(A*v)
10         return w
```

```
1  def VecteurPropre(A,e) :
2          d = A.shape
3          v = initialiseV(d[0])
4          v = Normalise(v)
5          while Norme(v-Normalise(A*v))>=e :
6                  v = Normalise(A*v)
7          return Normalise(A*v)
```

★ ★ ★ FIN DE L'ÉPREUVE ★ ★ ★