Devoir Maison n° 6
BCST2

pour jeudi 5 avril 2018

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 réel. En 1990, on a réalisé sur un terrain en jachère des relevés pour l'analyse chimique du sol. 6 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. On utilise alors la méthode d'analyse en composantes principales (ACP).

Notations
– On notera ^{t}A la transposée de la matrice A.
– On munit les espaces $\mathbb{R}^n$ de leur produit scalaire canonique, noté $<.,.>$ et de la norme associée notée $||\cdot||$.

1.1 Principe de l'ACP

On formalise ici la méthode. 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). 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 (on rappelle que pour centrée une série statistique x, il suffit de faire $x - \overline{x}$...).

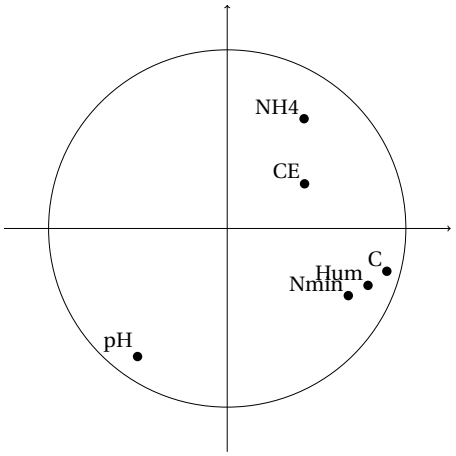
1. x et y étant deux séries statistiques de taille n, rappeler la définition de $\mathrm{Cov}(x, y)$.
   Soit R la matrice carrée d'ordre p dont le coefficient (i, j) valent $r_{ij} = \sum_{k=1}^{n} x_k^i x_k^j$.
   Justifier qu'on parle par la suite pour R de « matrice de covariance des données ».

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 $\lambda_1 \geqslant \cdots \geqslant \lambda_p$ 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 admettra que cela revient aussi à maximiser la quantité :

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

   où p_u est la projection orthogonale sur u.



4. Soit $U = \mathcal{M}_{\mathcal{B}}(u)$. Montrer que $XU = \begin{pmatrix} (x_1|u) \\ (x_2|u) \\ \vdots \\ (x_n|u) \end{pmatrix}$.

5. Rappeler l'expression de $||p_u(x_i)||$ en fonction du produit scalaire de x_i par u.
En déduire que $I(u) = {}^tURU$.

6. On écrit $u = \sum_{i=1}^{p} \mu_i w_i$. Justifier cette écriture et montrer que $I(u) = \sum_{i=1}^{p} \lambda_i \mu_i^2$.
En déduire 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 NH4 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. On peut par exemple interpréter ce cercle de la façon suivante : L'humidité d'un sol est liée à sa texture. Comme on trouve une forte corrélation de l'humidité avec l'azote minéralisable et le carbone oxydable, on en déduit que ces deux données sont liées à la texture du sol. Cet axe est donc un axe représentant la texture du sol. Le second axe est davantage lié aux propriétés chimiques du sol.

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

1.2 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 $\lambda_1 > \lambda_2 > \lambda_3 \geqslant \cdots \geqslant \lambda_n \geqslant 0$ et $w_1, w_2, \ldots, w_n$ une base orthonormée de vecteurs propres associés. On se concentre donc sur le cas particulier où $\lambda_1 > \lambda_2 > \lambda_3$.
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 :

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

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 $k \in \mathbb{N}$, $v_k = \dfrac{A^k v}{||A^k v||}$.
3. justifier que pour tout $k \in \mathbb{N}$, $A^k v = \sum_{i=1}^{n} \lambda_i^k s_i w_i$.
   En déduire $A^k v = \lambda_1^k s_1 (w_1 + \varepsilon_k)$ puis que $v_k = C_k \lambda_1^k s_1 (w_1 + \varepsilon_k)$, 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 $k \in \mathbb{N}$, $|C_k \lambda_1^k s_1|(||w_1||^2 + ||\varepsilon_k||^2) = 1$.
   Quelle est la limite de $|C_k \lambda_1^k s_1|$?
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 $R_1(v) = \dfrac{{}^t v A v}{{}^t v v}$. Exprimer $R_A(v_n)$ pour tout $n \in \mathbb{N}$ et dterminer $\lim\limits_{n\to\infty} R_A(v_n)$

1.3 Calcul numérique de λ_1 et de w_1

Chaque programme doit être commenté par une phrase détaillant le raisonnement qui a conduit à son élaboration. On 'interdit toute fonction de calcul matriciel à l'exception des opérations $*$ et + et d'une fonction `transpose(A)` *qui retourne la transposée d'une matrice* A.

1. Écrire une fonction `Norme(X)` d'argument une matrice colonne X de taille quelconque et qui renvoie le nombre ||X||.
2. Écrire une fonction `Normalise(v)` d'argument une matrice colonne $v \in \mathcal{M}_{n,1}(\mathbb{R})$ non nulle renvoie une nouvelle matrice colonne $\tilde{v}v/||v||$.
3. 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.

4. 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).
5. 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.
6. 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.

1.4 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 = Normalisee(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 = Normalisee(A*v)
7          return Normaise(A*v)
```

★ ★ ★ FIN DE L'ÉPREUVE ★ ★ ★