Polytech' lille
Département gis.
M. Marbac

Analyse en composantes principales.

Travaux pratiques (statistiques exploratoires).

Les jeux de données étudiés sont disponibles sur `http://math.univ-lille1.fr/~marbaclo/`

1 Introduction à l'ACP sous R

1. **Création de l'échantillon**. Simuler trois échantillons de taille 30 selon les lois normales bivariées ayant pour matrices de covariance $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ et pour centres respectifs $\mu_1 = (0, 0)$, $\mu_2 = (10, 0)$ et $\mu_3 = (0, 10)$ (comme la matrice de covariance est diagonale vous pouvez utiliser la fonction `rnorm()`). Stocker l'ensemble des données dans une matrice de taille 90×2. Centrer et réduire cette matrice en utilisant la fonction `scale()`. Créer un vecteur permettant d'identifier de quelle loi est issu chaque individu.
2. **Nuage de points**. Représenter les individus par un nuage de points dans le plan canonique. Les couleurs indiquent l'appartenance des individus aux lois. Ajouter les axes canoniques avec la fonction `abline()`.
3. **Projection sur les axes canoniques**. Représenter le nuage de points et la projection des individus sur l'axe des abscisses. Utiliser un symbole différent pour les individus projetés (par exemple `pch=4`). Faites la même chose sur l'axe des ordonnées. Que constatez-vous ?
4. **Mise en oeuvre de l'ACP**.
   (a) Calculer les coordonnées des axes principaux et leur inertie associée (la fonction `eigen()` permet d'obtenir les vecteurs propres et les valeurs propres d'une matrice).
   (b) Quel est le pourcentage d'inertie expliqué par le premier axe factoriel ?
   (c) Représenter les individus par un nuage de points dans le plan factoriel et représenter leur projection sur le premier axe factoriel.
5. **Formule de reconstitution**. En utilisant la formule de reconstitution vue en cours, construire la matrice reconstituée de l'échantillon à partir des deux axes factoriels. Quel est l'écart quadratique moyen entre l'échantillon et sa reconstitution ? Justifiez. Faites la même chose en utilisant uniquement le premier axe factoriel. Comparer l'écart quadratique moyen obtenu avec un le premier axe factoriel et ceux obtenu par les projections sur les axes canoniques.

2 Points absorbants sous R

À l'aide de la fonction `rmvnorm()` du package `mvtnorm`, simuler un échantillon de taille 19 selon une loi normale bivariée centrée et de matrice de covariance $\begin{pmatrix} 1 & -0.75 \\ -0.75 & 1 \end{pmatrix}$. Ajouter à cet échantillon le point de coordonnées $(20, 20)$. Faites l'ACP de ce jeu de données en utilisant la fonction `PCA()` du package `FactoMineR`.

3 Étude des consommations de denrées alimentaires sous R (fichier : denrees.txt)

L'étude concerne les consommations annuelles en 1972, exprimées en francs, de huit denrées alimentaires [1] : PAO pain ordinaire, PAA autre pain, VIO vin ordinaire, POT pommes de terres, LEC légumes

1. A. Villeneuve, « La consommation alimentaire des Français», *Collections de l'INSEE*. M 34.

sec, RAI raisin de table et PLP plats préparés. Les individus sont huit catégories socio-professionnelles et les données sont les moyennes par CSP : AGRI exploitants agricoles, SAAG salariés agricoles, PRIN professions indépendantes, CSUP cadre supérieurs, CMOY cadres moyens, EMPL employés, OUVR ouvriers et INAC inactifs. Après avoir effectué les statistiques élémentaires, faire l'ACP de ce jeu de données en justifiant votre approche.

4 ACP sous SAS (fichier notes.txt)

En utilisant la métrique identité et en donnant la même importance à chaque individu, effectuer l'ACP du jeu de données avec SAS (`proc princomp`).

1. Importer le fichier notes.txt.
2. Retrouver l'ensemble des sorties vues en cours.
3. Cette procédure produit dans sa version actuelle peu de sorties. Programmer alors en SAS pour créer l'ensemble des tableaux et des graphiques utiles en ACP.
4. Exporter l'ensemble des tableaux et des graphiques sous forme d'un fichier PDF.

5 Imagerie et ACP sous R

L'ACP permet de faire une approximation d'une matrice en réduisant sa dimension (nombre de variable). On propose d'utiliser cette technique pour réduire l'espace disque nécessaire pour stocker l'image `Lena_soderberg.ppm`. En effet, cette image est constituée de trois matrices contenant l'information sur l'intensité des trois couleurs de base (rouge, vert, bleu). On se demande si il est possible de diviser par trois l'espace mémoire nécessaire tout en conservant une représentation en couleur. Pour cela, on se propose d'effectuer les étapes suivantes :

- Importer l'image `Lena_soderberg.ppm` sous R en utilisant la fonction `read.pnm()` du package `pixmap`.
- Afficher l'image avec la fonction `plot`.
- Afficher l'image dans chacune des trois couleur de base.
- Effectuer l'ACP. Quelle métrique utilisez-vous ? Combien d'axes retenez-vous ?
- Reconstituer l'image à partir des axes retenus par l'ACP (fonction `reconst()` du package `FactoMineR`).

6 Débruitage par l'ACP

On dispose d'une base de 20 images bruitées prises par un drone au même instant. Ces images sont stockées dans l'archive `snake.tar.gz`. On souhaite débruiter les images. Pour cela, on utilise une ACP. Vous comparer une image originale et ses versions reconstituées à partir des p composantes principales $(1, 3, 6, 8, 12)$. Quelles conclusions tirez-vous ?