Analyse en Composantes Principales

Maxime CHAMBREUIL
maxime.chambreuil@insa-rouen.fr

22 mars 2003

Table des matières

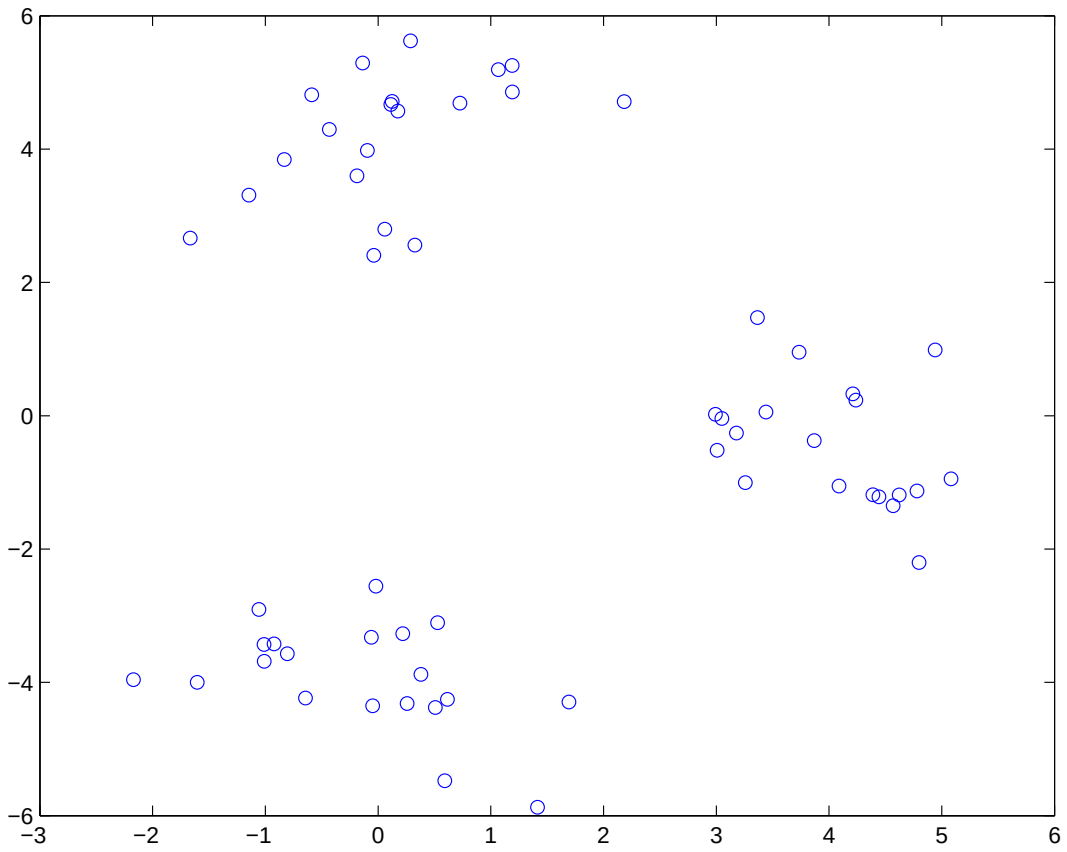
1 Explication

1.1 Préparation des données

Nous avons généré un échantillon de 60 individus à partir de 3 lois normales, d'espérance différentes :

$$\mu_1 = (0, 4)$$
$$\mu_2 = (0, -4)$$
$$\mu_3 = (4, 0)$$

et de variance la matrice identité, pour obtenir :



Nous avons ensuite multiplié notre échantillon par la matrice :

$$M = \begin{bmatrix} -1 & 1 & -2 & 5 & -3 \\ 2 & 20 & 10 & -5 & -37 \end{bmatrix}$$

pour obtenir un échantillon de 60 individus mais sur 5 variables. Nous avons aussi rajouter un bruit durant la transformation.

1.2 Application de la fonction ACP

1.2.1 Signature

A cette nouvelle matrice, nous appliquons la fonction acp :

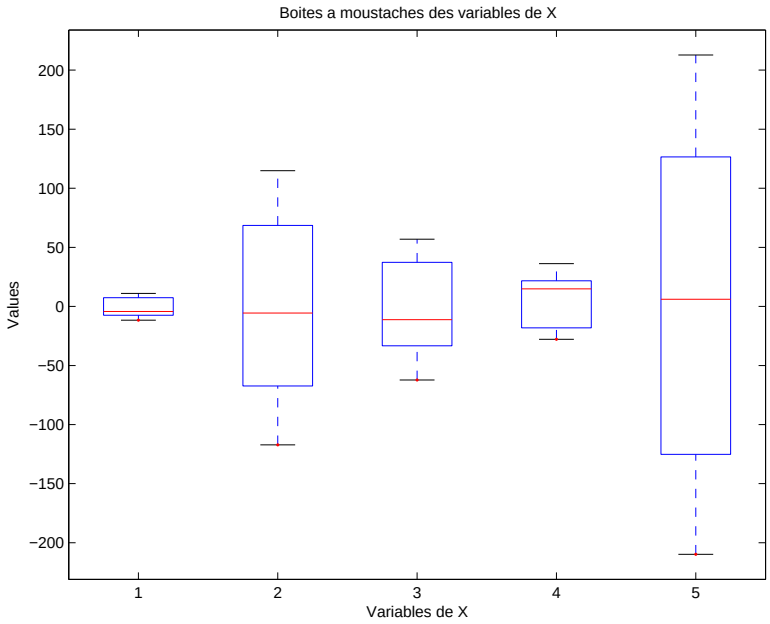
[epsilon,U] = acp(X,nbCompo)

En entrée :
X : Matrice des individus
nbCompo : Nombre de composantes principales souhaitées

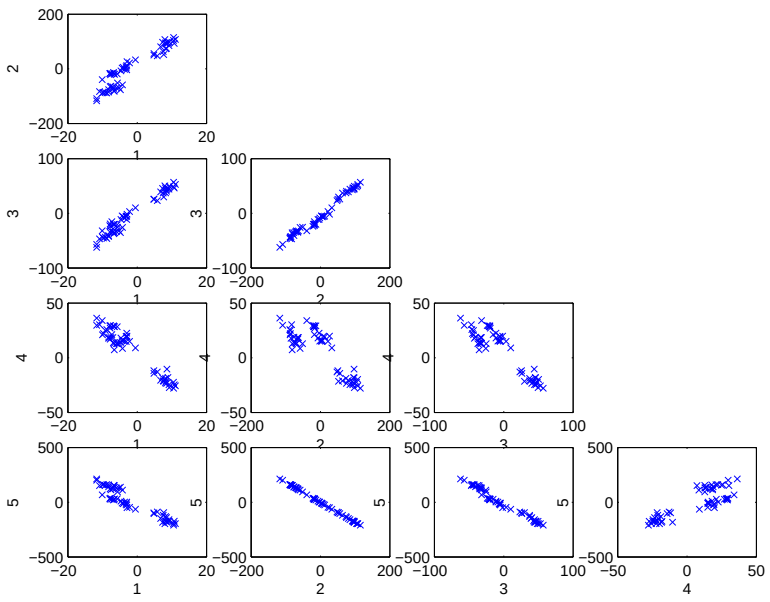
En sortie :
epsilon : X selon les nbCompo composantes principales
U : Matrice des nbCompo vecteurs propres, triés dans l'ordre

1.2.2 Déroulement de la fonction ACP

La fonction *acp* commence par afficher les boîtes à moustaches de la matrice selon chaque variable.



On applique ensuite la fonction *multiplot* qui va afficher chaque variable selon les autres et nous permettre de deviner quelques corrélations entre les variables.

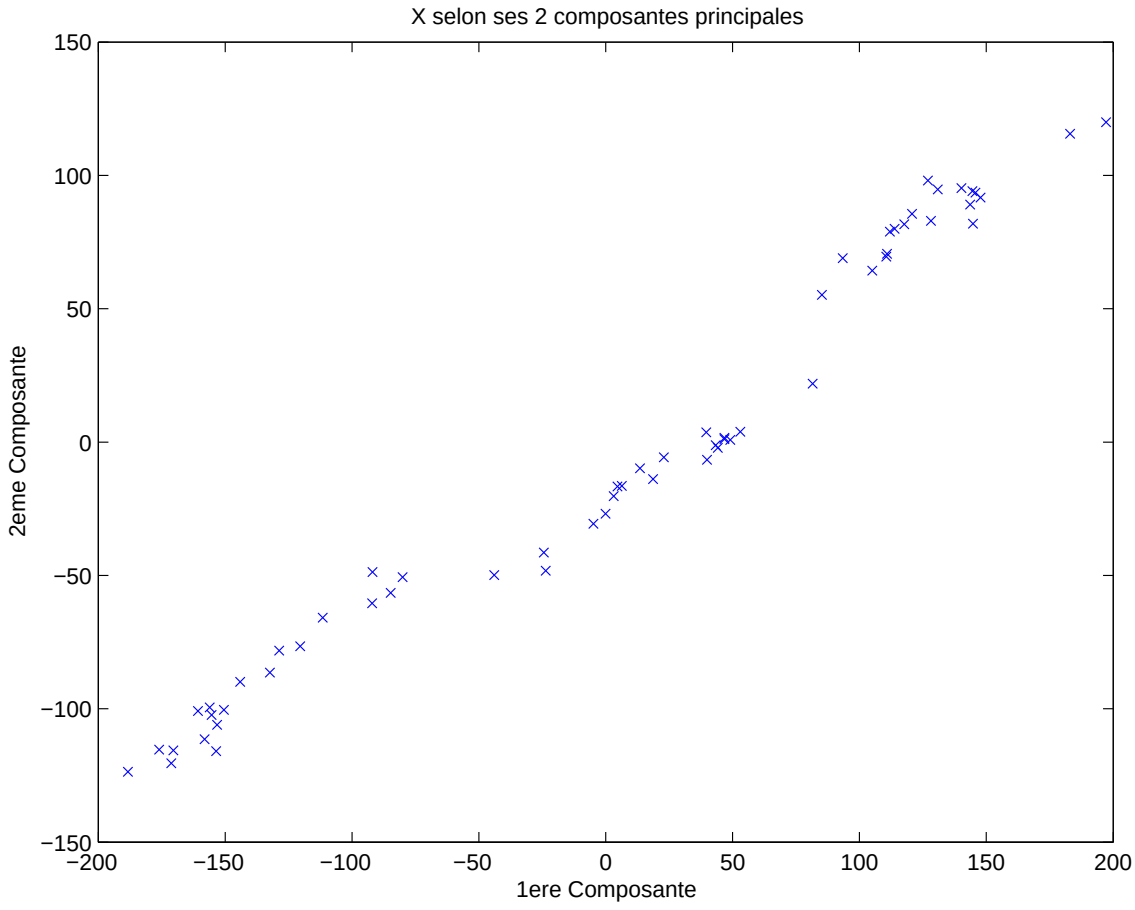


On a ensuite centré et réduit nos variables pour les ramener à la même échelle, avant de calculer les vecteurs et les valeurs propres :

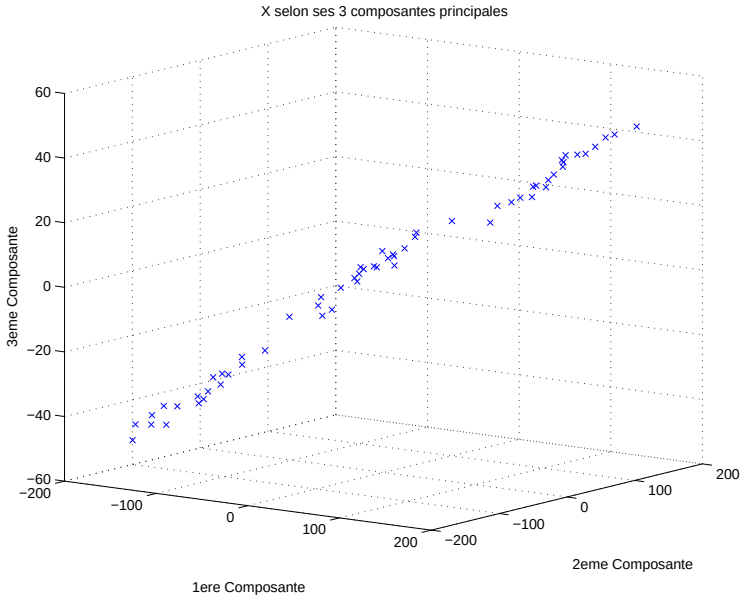
$$\nu = \begin{bmatrix} 0 & 0,01 & 0,85 & 0,26 & -0,45 \\ -0,71 & 0,34 & -0,12 & -0,40 & -0,45 \\ -0,01 & -0,86 & -0,2 & -0,12 & -0,46 \\ -0,02 & -0,23 & 0,45 & -0,75 & 0,42 \\ -0,70 & -0,32 & 0,11 & 0,44 & 0,45 \end{bmatrix}$$

$$\lambda = \begin{pmatrix} 0 \\ 0,05 \\ 0,87 \\ 17,55 \\ 276,52 \end{pmatrix}$$

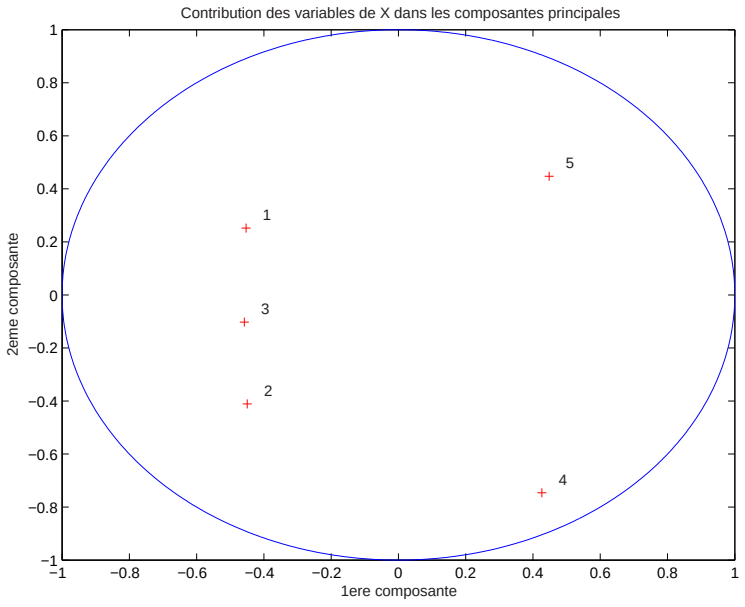
On extrait ensuite les composantes principales souhaitées, ainsi que les 2 premières pour les afficher en 2D :



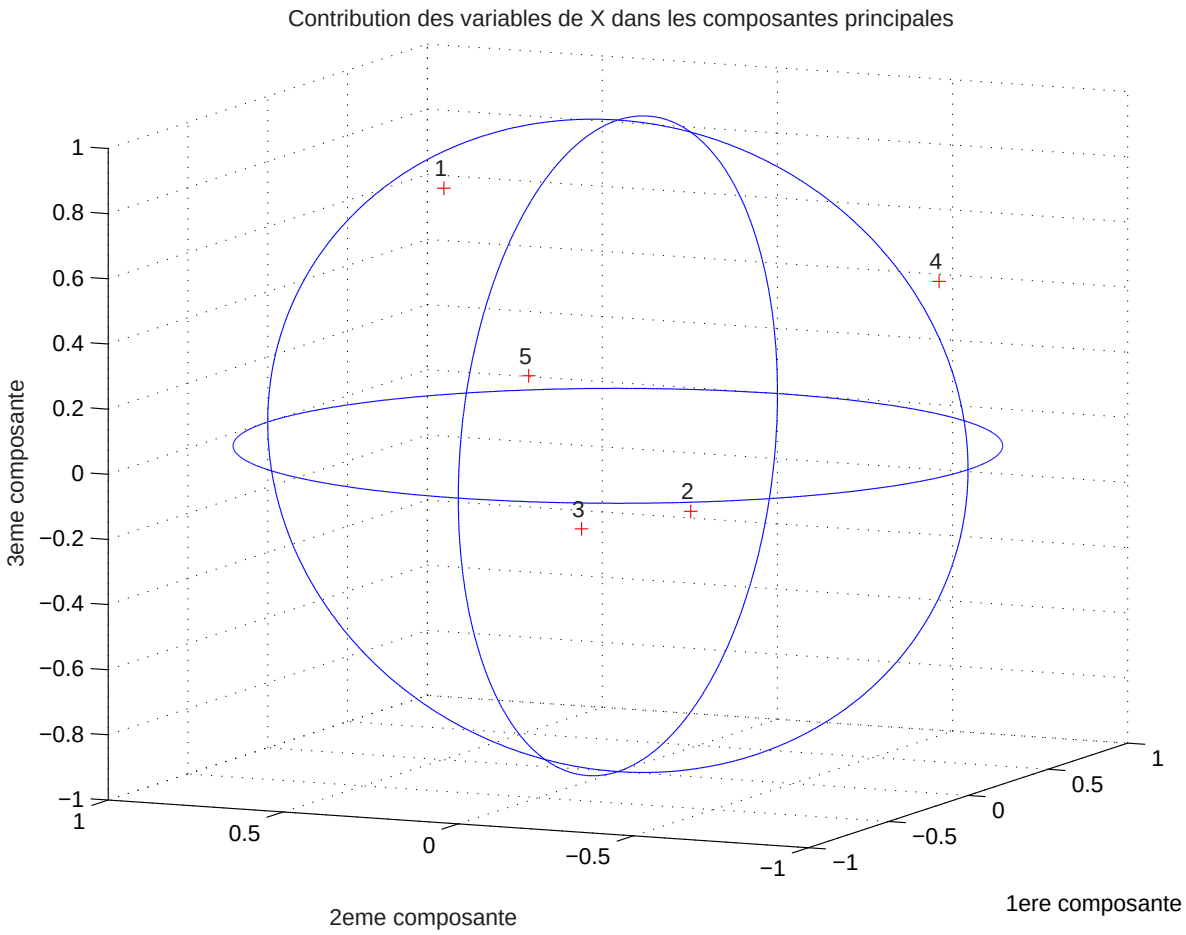
et les 3 premières pour la 3D :



Avant de retourner les variables demandées, on affiche les contributions de chacune de ces variables selon 2 composantes :



ou 3 :



2 Interprétation

Grâce à notre fonction *multiplot*, on peut noter des corrélations positives entre les couples de variables (1 , 3) et (2 , 3), ainsi que des corrélations négatives : (2 , 5), (3 , 5). Ses corrélations sont vérifiées par l'affichage des contributions : les variables 1, 2 et 3 sont proches et en même temps éloignés de la 5ème.

Concernant les implications des variables dans les composantes principales, on peut noter que 4 et 5 jouent un rôle important dans la première composante, 1 et 5 sont prépondérants dans la 2ème et 1, 4 et 5 dans la 3ème. Donc 5 apparait dans 3 composantes et 4 dans 2. Cela vérifie bien nos résultats sur les valeurs propres :

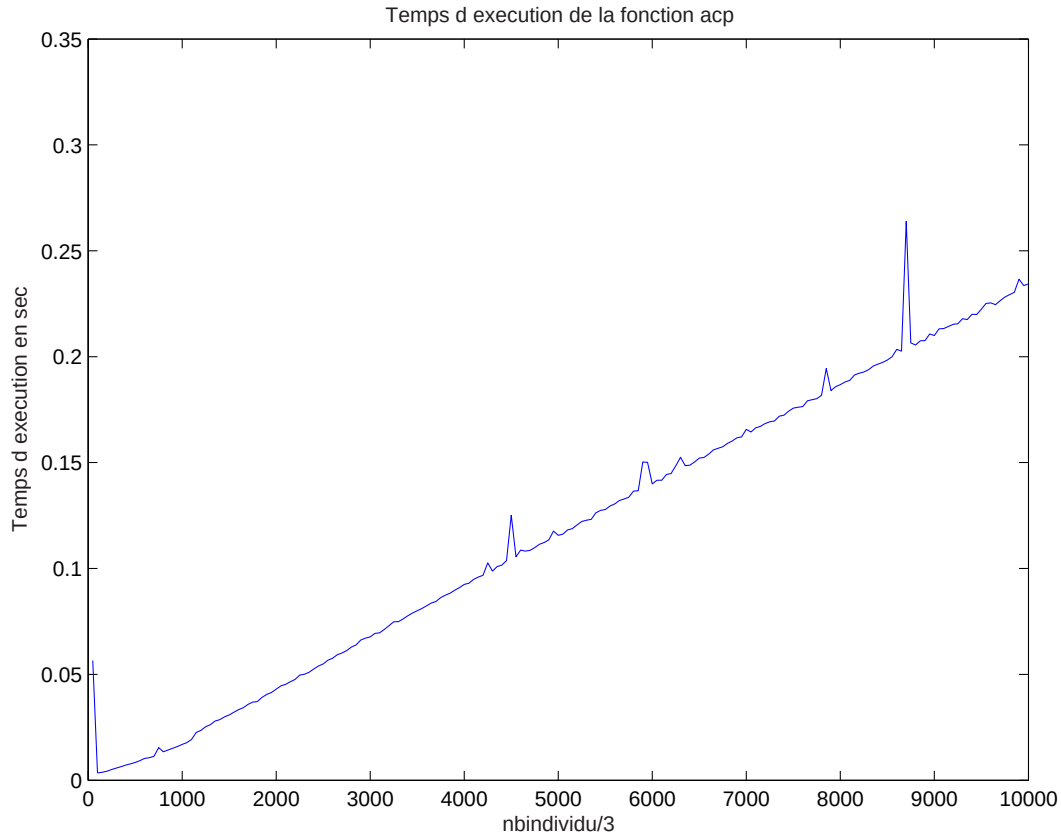
$$\lambda_5 > \lambda_4$$

On remarque sur les courbes selon les composantes principales qu'on retrouve 3 nuages de points, correspondants à nos lois normales d'origine.

3 Conclusion

Ce TP m'a permis de mieux comprendre l'utilité de savoir calculer les valeurs propres d'une matrice. On peut ainsi retirer intelligemment les variables qui n'ont pas grande influence sur les individus. Ceci doit être très utile lorsqu'on traite un gros volume de données.

De plus, l'ACP est assez efficace : Voici la courbe du temps d'exécution de la fonction *acp* en fonction de nbindividu/3, avec 5 variables.



Je n'ai pas eu le temps de tester son efficacité par rapport à l'augmentation du nombre de variables. Mais je pense que cela ne doit pas plus le déranger que celle des individus.