T. D. n° III - Quelques applications
[d'après exercices proposés par J. F. Durand dans
http : //www.math.univ − montp2.fr/˜durand/bibliography/polyalgmatc.pdf]

Exercice n° 1.
Soit le tableau de données

$$\mathbf{T} = \sqrt{10}\begin{bmatrix} 2 & 2 & 3 \\ 3 & 1 & 2 \\ 1 & 0 & 3 \\ 2 & 1 & 4 \\ 2 & 1 & 3 \end{bmatrix}$$

correspondant à des mesures effectuées sur 5 individus de poids statistiques égaux pour les trois variables T^1, T^2 et T^3. On va effectuer une ACP centrée-réduite sur ce tableau.

1. Calculer l'individu moyen, le vecteur $(\sigma_1, \sigma_2, \sigma_3)'$ des écarts types des variables et la matrice $\mathbf{X}$ des données centrées-réduites.
2. Calculer la matrice des corrélations $\mathbf{R}$
3. Effectuer la décomposition aux valeurs propres de $\mathbf{R}$
4. Les deux premiers vecteurs de $\mathbf{R}$ sont
$$\boldsymbol{\xi}_1' = \frac{1}{2}\left(\sqrt{2}, 1, -1\right)' \text{ et } \boldsymbol{\xi}_2' = \frac{1}{\sqrt{2}}(0,1,1)'.$$
Ils sont associés aux valeurs propres
$$\lambda_1 = 1 + \frac{\sqrt{2}}{2} \text{ et } \lambda_2 = 1.$$
Calculer les composantes principales $\mathbf{c}_1$ et $\mathbf{c}_2$ dont on vérifiera les propriétés statistiques
5. Représenter les individus dans le plan factoriel $(1, 2)$. Donner une interprétation de cette ACP

Exercice n° 2
Soit la matrice $\mathbf{X} = [\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3]$ dont les variables ont pour matrice de corrélation

$$\mathbf{R} = \begin{bmatrix} 1 & \rho & -\rho \\ \rho & 1 & \rho \\ -\rho & \rho & 1 \end{bmatrix}$$

avec $-1 \leq \rho \leq 1$. On va effectuer l'ACP centrée-réduite de $\mathbf{X}$.

1. Vérifier que $\mathbf{R}$ admet pour vecteur propre $\boldsymbol{\xi}_1 = \frac{1}{\sqrt{3}}(1, -1, 1)'$
2. Déterminer les autres vecteurs propres et valeurs propres de $\mathbf{R}$
3. Quelles sont les valeurs possibles de ρ? Justifier le fait que l'ACP a plus d'intérêt si $-1 < \rho < 0$. On se placera ensuite dans ce cas.
4. Calculer les pourcentages de variance expliquée et tracer l'éboulis de valeurs propres
5. Comment s'interprète en fonction de $\mathbf{x}^1$, $\mathbf{x}^2$ et $\mathbf{x}^3$ l'unique composante à retenir ici?

Correction exercice n°1

1. L'individu moyen est obtenu en faisant la moyenne des colonnes du tableau T, soit $\overline{\mathbf{x}} = \sqrt{10}\,(2,1,3)'$. Le vecteur des écarts types est obtenu en calculant les écarts types de chaque colonnes de T. Soit $\mathbf{T}_c$ la matrice des données centrées, $\mathbf{T}_c$= $\mathbf{T}-$ $(\overline{\mathbf{x}}', \overline{\mathbf{x}}', \overline{\mathbf{x}}', \overline{\mathbf{x}}', \overline{\mathbf{x}}')'$

$$\mathbf{T}_c = \sqrt{10}\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \\ -1 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

le vecteur $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3)'$ contient les termes en racine carrée des éléments diagonaux de la matrice $\mathbf{V} = \frac{1}{n}\mathbf{T}_c'\mathbf{T}_c$, $n = 5$, soit $\boldsymbol{\sigma}^2 = \frac{10}{5}\,(2,2,2)'$. Le calcul de la matrice $\mathbf{X}$ revient à diviser chaque colonne de $\mathbf{T}_c$ par l'écart-type de la variable correspondante :

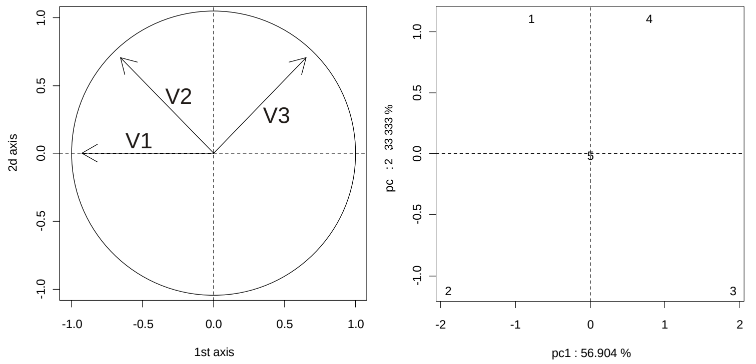
$$\mathbf{X} = \frac{\sqrt{10}}{2}\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \\ -1 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} = \frac{1}{2}\mathbf{T}_c$$

2. La matrice des corrélations
$$\mathbf{R} = \frac{1}{n}\mathbf{X}'\mathbf{X} = \frac{1}{5}\frac{10}{4}\begin{bmatrix} 2 & 1 & -1 \\ 1 & 2 & 0 \\ -1 & 0 & 2 \end{bmatrix}$$

3. L'ACP centrée-réduite de $\mathbf{T}$ nécessite le calcul des vecteurs propres de $\mathbf{R}$. On résout le système $\det(\mathbf{R} - \lambda\mathbf{I}) = 0$ soit
$$(1-\lambda)\left(\lambda - 1 - \frac{1}{\sqrt{2}}\right)\left(\lambda - 1 + \frac{1}{\sqrt{2}}\right) = 0$$
et on obtient 3 valeurs propres $\lambda_1 = 1 + \frac{1}{\sqrt{2}}$, $\lambda_2 = 1$, $\lambda_3 = 1 - \frac{1}{\sqrt{2}}$.

4. Le calcul des 2 premières composantes principales est donné par
$$\mathbf{c}^i = \mathbf{X}\boldsymbol{\xi}_i,\ i = 1,2$$
soit pour la première, associée à la valeur propre λ_1,
$$\mathbf{c}^1 = \frac{\sqrt{10}}{2} \times \frac{1}{2}\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \\ -1 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}\begin{pmatrix} \sqrt{2} \\ 1 \\ -1 \end{pmatrix} = \frac{\sqrt{10}}{4}\begin{pmatrix} 1 \\ \sqrt{2}+1 \\ -\sqrt{2}-1 \\ -1 \\ 0 \end{pmatrix}$$
et la seconde $\mathbf{c}^2 = \frac{\sqrt{10}}{2} \times \frac{1}{\sqrt{2}}\,(1,-1,-1,1,0)'$. Les propriétés de ces composantes montrent qu'elles sont orthogonales deux à deux
$$\left\langle \mathbf{c}^i, \mathbf{c}^j \right\rangle_{\mathbf{M}} = \frac{1}{5}\mathbf{c}^{i\prime}\mathbf{c}^j = 0,\ \forall i \neq j$$
et que leur norme est reliée à chaque valeur propre par
$$\left\|\mathbf{c}^j\right\|^2_{\mathbf{M}} = \frac{1}{5}\mathbf{c}^{j\prime}\mathbf{c}^j = \lambda_j,\ j = 1,2.$$

5. représentation des individus dans le plan factoriel $(1,2)$.



Le premier axe oppose les variations de V1, V2 avec V3. Le second est un axe de taille. Les individus 2 et 3 présentent de faibles valeurs de V2 et V3, l'individu 2 étant caractérisé par une forte valeur de V1. Les individus 1 et 4 sont attachés aux variables V2 et V3 respectivement. L'individu 5 est le plus consensuel puisque confondu avec le centre de gravité de l'ACP.

Correction exercice n°2

1. Si $\mathbf{R}$ admet pour vecteur propre $\boldsymbol{\xi}_1$ alors il vérifie $\mathbf{R}\boldsymbol{\xi}_1 = \lambda_1 \boldsymbol{\xi}_1,\ \lambda_1 \in \mathbb{R}^+$. On calcule $\mathbf{R}\boldsymbol{\xi}_1$:

$$\frac{1}{\sqrt{3}} \begin{bmatrix} 1 & \rho & -\rho \\ \rho & 1 & \rho \\ -\rho & \rho & 1 \end{bmatrix} \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} = \frac{1}{\sqrt{3}} \begin{pmatrix} 1-2\rho \\ 2\rho - 1 \\ -2\rho + 1 \end{pmatrix} = (1-2\rho)\, \boldsymbol{\xi}_1$$

   donc, $\boldsymbol{\xi}_1$ est bien vecteur propre de $\mathbf{R}$ pour la valeur propre $\lambda_1 = 1 - 2\rho$. Cette valeur propre étant positive (propriété de $\mathbf{R}$) on doit avoir $-1 \leq \rho \leq \frac{1}{2}$.

2. Pour déterminer les autres éléments propres de $\mathbf{R}$, on résout $\det(\mathbf{R} - \boldsymbol{\lambda}\mathbf{I}) = 0$, ce qui équivaut à

$$(1-\lambda)\left((1-\lambda)^2 - \rho^2\right) - 2\rho^2\,(1 - \lambda + \rho) = 0$$

$$(1-\lambda+\rho)\left[(1-\lambda)(1-\lambda-\rho) - 2\rho^2\right] = 0$$

$$(1-\lambda+\rho)\left[\lambda^2 - \lambda(2-\rho) + 1 - \rho - 2\rho^2\right] = 0$$

   On sait que $\lambda_1 = 1-2\rho$ est valeur propre de $\mathbf{R}$. Ceci permet de calculer par identification la racine du polynôme ci-dessus. On montre que $\lambda = 1+\rho$ est racine double. On peut maintenant déterminer les vecteurs propres pour cette valeur propre. Soit $\boldsymbol{\xi} = (x, y, z)'$ un vecteur vérifiant $\mathbf{R}\boldsymbol{\xi} = \lambda\boldsymbol{\xi}$. En développant, on obtient le système

$$\begin{cases} -\rho x + \rho y - \rho z = 0 \\ \rho x - \rho y + \rho z = 0 \\ -\rho x + \rho y - \rho z = 0 \end{cases} .$$

   Il nous faut maintenant trouver des valeurs arbitraires de x, y et z qui vérifient ce système. On en trouve facilement 2 tiercés avec $(1,1,0)$ et $(1,0,-1)$ qui ne soient pas combinaison linéaire l'un de l'autre. En normalisant ces vecteurs, on obtient finalement les deux vecteurs propres $\boldsymbol{\xi}_2 = \frac{1}{\sqrt{2}}(1,1,0)'$ et $\boldsymbol{\xi}_3 = \frac{1}{\sqrt{2}}(1,0,-1)'$. Finalement, la matrice des corrélations $\mathbf{R}$ peut être décomposée sous la forme $\mathbf{R} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}'$ avec $\mathbf{P} = [\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \boldsymbol{\xi}_3]$, matrice des vecteurs propres et $\boldsymbol{\Lambda}$, matrice des valeurs propres de termes diagonaux $(1-2\rho, 1+\rho, 1+\rho)$.

3. Nous avons déjà vu que les valeurs possibles de ρ sont $-1 \leq \rho \leq \frac{1}{2}$ pour assurer la positivité des valeurs propres. Supposons maintenant que $-1 < \rho < 0$. On peut ranger les valeurs propres par ordre décroissant avec $1 - 2\rho > 1 + \rho$. On se rend alors compte que l'espace initial à 3 variables peut être réduit à une seule variable, combinaison linéaire des 3 variables initiales. En effet, si l'on considère le sous-espace propre de dimension 2 associé à la valeur propre double, l'information du nuage de points résumé dans cet espace est identique dans les deux directions. Cela n'apporte rien de les conserver.

4. Les pourcentages d'inertie expliquée sont donnés, dans chaque direction propre, par le rapport d'une valeur propre sur la somme totale des valeurs propres, égale dans ce cas à 3, puisqu'elle correspond à l'inertie totale calculée à partir de la matrice des corrélations (variables réduites). L'éboulis correspond au tracé, sur le même graphique, de barres de hauteur $(1-2\rho)/3$, $(1+\rho)/3$ et $(1+\rho)/3$.

5. A partir du premier vecteur propre et du tableau $\mathbf{X}$ centré-réduit noté $\mathbf{X}_{cr}$, on peut calculer la composante principale
$$\mathbf{c}^1 = \mathbf{X}_{cr}\boldsymbol{\xi}_1.$$
En notant que $\mathbf{X}_{cr} = \left[\mathbf{x}_{cr}^1, \mathbf{x}_{cr}^2, \mathbf{x}_{cr}^3\right]$ et que $\boldsymbol{\xi}_1 = (\xi_{11}, \xi_{21}, \xi_{31})'$ on peut exprimer la composante en fonction des variables initiales à un centrage et une réduction près comme
$$\mathbf{c}^1 = \xi_{11}\mathbf{x}_{cr}^1 + \xi_{21}\mathbf{x}_{cr}^2 + \xi_{31}\mathbf{x}_{cr}^3.$$
Une composante principale est donc une nouvelle variable, combinaison linéaire des variables initiales.