Corrigé de la série 1 (suite)

Exercice 4 Soit la matrice des données $X={}^t\begin{pmatrix} 0 & 1 & 1 & 2 & 3 & 3 & 4 \\ 1 & 2 & 4 & 3 & 2 & 4 & 5 \end{pmatrix}$**.** Il s'agit d'une matrice d'observations de deux variables sur 7 individus. Les deux variables sont : $X^1={}^t(0,1,1,2,3,3,4)$ **et** $X^2={}^t(1,2,4,3,2,4,5)$**.**

1) Le centre de gravité du nuage est donné par : $g={}^t(\overline{X}^1,\overline{X}^2)={}^t\left(\frac{14}{7},\frac{21}{7}\right)=\begin{pmatrix} 2 \\ 3 \end{pmatrix}$.

2) La matrice centrée est la matrice $X_0={}^t\begin{pmatrix} -2 & -1 & -1 & 0 & 1 & 1 & 2 \\ -2 & -1 & 1 & 0 & -1 & 1 & 2 \end{pmatrix}$**.**

Où $X^j \to X_0^j = X^j - \overline{X}^j \times \mathrm{I}$ avec $I={}^t(1,1,1,1,1,1,1) \in \Re^7$.

3) La matrice des variances-covariances C associée à la matrice des données X_0 est donnée par :

$$C=\frac{1}{m}({}^tX_0\cdot X_0)=\frac{1}{7}({}^tX_0\cdot X_0). \quad C=\begin{pmatrix} 12/7 & 8/7 \\ 8/7 & 12/7 \end{pmatrix}=\begin{pmatrix} 1.714 & 1.143 \\ 1.143 & 1.714 \end{pmatrix}.$$

De cette matrice, nous pouvons déduire l'inertie du nuage des données autour du centre de gravité, notée I_g. Elle représente la moyenne pondérée des carrés des distances des points au centre de gravité. C'est à dire I_g **= somme des variances**. Donc $I_g = 24/7$**.**

4) Avant de déterminer les axes principaux, nous devons tout d'abord déterminer les valeurs propres de la matrice C. Ces valeurs propres sont les racines du polynôme caractéristique associé à la matrice C défini par : $P_C(\lambda)=\det(C-\lambda.Id)$.

$$P_C(\lambda)=\det(C-\lambda.Id)=\begin{vmatrix} \frac{12}{7}-\lambda & \frac{8}{7} \\ \frac{8}{7} & \frac{12}{7}-\lambda \end{vmatrix}=\left(\frac{12}{7}-\lambda\right)^2-\frac{64}{49}.$$

$$P_C(\lambda)=0 \Leftrightarrow \left(\frac{12}{7}-\lambda+\frac{8}{7}\right)\times\left(\frac{12}{7}-\lambda-\frac{8}{7}\right)=0$$

, ce qui donne

$$\begin{cases} \lambda_1=\frac{20}{7}\approx 2.8571 \\ \lambda_2=\frac{4}{7}\approx 0.5714 \end{cases}.$$

Déterminons maintenant les axes principaux qui sont les vecteurs propres (normés) associés aux valeurs propres calculées.

i) Pour λ_1, il existe $v_1 = {}^t(x,y) \in \Re^2 - \{{}^t(0,0)\}$ tel que : $C \cdot v_1 = \lambda_1 \cdot v_1$ c'est-à-dire

$\begin{pmatrix} 12/7 & 8/7 \\ 8/7 & 12/7 \end{pmatrix} \times \begin{pmatrix} x \\ y \end{pmatrix} = \frac{20}{7} \cdot \begin{pmatrix} x \\ y \end{pmatrix}$ ce qui est équivalent à $\begin{cases} \frac{12}{7}x + \frac{8}{9}y = \frac{20}{7}x \\ \frac{8}{7}x + \frac{12}{7}y = \frac{20}{7}y \end{cases}$, et donc

$\begin{cases} y = x \\ x = y \end{cases}$, par suite $v_1 = {}^t(x,x) = x\ {}^t(1,1)$ avec x quelconque dans $\Re^*$.

En général un vecteur propre engendre un sous espace vectoriel, dans notre cas il est de dimension égale à 1 et dont le générateur est le vecteur ${}^t(1,1)$. C'est pour cette raison que nous prenons $v_1 = {}^t(1,1)$ et comme il n'est pas de norme égale à 1, **alors l'axe principal** est la droite engendrée par **le vecteur unitaire** $u_1 = \frac{1}{\|v_1\|} \cdot v_1$, $u_1 = {}^t\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$.

ii) Pour λ_2, il existe $v_2 = {}^t(x,y) \in \Re^2 - \{{}^t(0,0)\}$ tel que : $C \cdot v_2 = \lambda_2 \cdot v_2$ c'est-à-dire

$\begin{pmatrix} 12/7 & 8/7 \\ 8/7 & 12/7 \end{pmatrix} \times \begin{pmatrix} x \\ y \end{pmatrix} = \frac{4}{7} \cdot \begin{pmatrix} x \\ y \end{pmatrix}$ ce qui est équivalent à $\begin{cases} \frac{12}{7}x + \frac{8}{9}y = \frac{4}{7}x \\ \frac{8}{7}x + \frac{12}{7}y = \frac{4}{7}y \end{cases}$, et donc

$\begin{cases} y = -x \\ x = -y \end{cases}$, par suite $v_2 = {}^t(x,-x) = x\ {}^t(1,-1)$ avec x quelconque dans $\Re^*$, et **l'axe principal** est la droite engendrée par **le vecteur unitaire** $u_1 = \frac{1}{\|v_1\|} \cdot v_1$,

$u_1 = {}^t\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$.

Exercice 4 Basé sur les données de l'exercice précédent.

1) Posons: $Y = {}^tX_0 \cdot X_0$, alors : $Y = 7C$, et donc les valeurs propres de la matrice Y peuvent être calculées directement à partir de celles de la matrice C sans refaire les calculs. Elles sont définies par la relation suivante : $\lambda_i(Y) = 7 \times \lambda_i(C)$ pour $i = 1,2$.

Donc les valeurs propres sont données par : $\lambda_1 = 7(20/7) = 20$ et $\lambda_2 = 7(4/7) = 4$.

2) Calculons les valeurs propres da la matrice $X_0 \cdot {}^tX_0$. Nous remarquons que la matrice

$X_0 \cdot {}^t X_0$ n'est que la transposée de la matrice $Y = {}^t X_0 \cdot X_0$ c'est-à-dire $X_0 \cdot {}^t X_0 = {}^t Y$ avec Y matrice carrée de dimension 2 nous donnons des informations sur les des deux variables. Tandis que ${}^t Y$ nous fournit des informations sur les individus et donc c'est une matrice carrée de dimension 7 (nombre d'individus). De ce fait les valeur propres de $X_0 \cdot {}^t X_0$ sont de nombre 7.

REMARQUE **Il ya un résultat important qui dit que les valeurs propres de la matrice donnant les informations sur les individus sont données par : n valeurs propres égales à celles de la matrice d'information sur les variables et (m-n) valeurs propres nulles.**

Ainsi, les valeurs propres de la matrice $X_0 \cdot {}^t X_0$ sont : $\lambda_1 = 20$, $\lambda_2 = 4$ et $\lambda_3 = 0$ d'ordre de

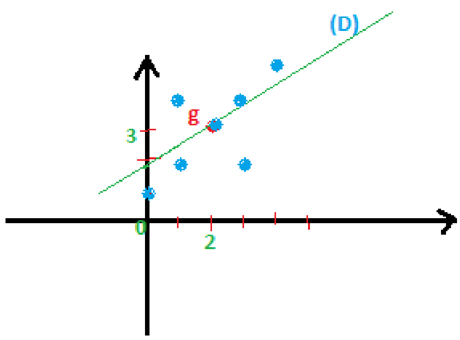
multiplicité 5 c'est-à-dire $\lambda_3 = 0 = \lambda_4 = \lambda_5 = \lambda_6 = \lambda_7$.

3) Déterminons la droite de régression de X^2 prédite par X^1 .

Notons cette droite **(D),** la droite de régression est une droite qui permet une meilleure interpolation linéaire des points (données), on dit aussi qu'il s'agit d'un ajustement linéaire qui nous permet d'estimer ou de prédire une nouvelle donnée X_o

Elle est définie par $y = a \cdot x + b$ avec $a = \dfrac{\operatorname{cov}(X^1, X^2)}{Var(X^1)}$.

De la matrice des variances-covariances déjà calculée ci-dessus (voir exo 4), nous avons : $a = \dfrac{8}{7} \times \dfrac{7}{12} = \dfrac{2}{3}$. Pour déterminer la constante b , bien-sur il suffit juste de remplacer le point g dans la droite **(D)** car cette dernière passe par ce point. En effet : $g = {}^t(2,3)$, alors : $b = 3 - \dfrac{2}{3} \times 2 = \dfrac{5}{3}$, et donc $y = \dfrac{2}{3} \cdot x + \dfrac{5}{3}$.



Du graphe présenté, nous concluons que la droite calculée est le meilleur ajustement des données, l'erreur ou bien la perte d'information (la distance perpendiculaire d'un points sur la droite) est petite pour la majorité.

Pour l'exercice 6, la matrice n'apparait pas. Apparemment tout ce qui concerne les formules mathématiques. C'est pour ça, je vous redonne l'énoncé ci-dessous.

Exercice 6 On considère la matrice de données suivante:

$$X = {}^{t}\begin{pmatrix} 1 & -1 & 0 & 1 & 0 & 1 & -1 & -1 & 0 & 0 \\ 0 & 1 & 1 & -1 & -1 & 1 & -1 & 0 & 1 & -1 \\ 1 & -1 & 0 & 1 & 0 & 1 & -1 & -1 & 0 & 0 \end{pmatrix}.$$

1) Donner le nuage des points $N(I)$.
2) Calculer le centre de gravité de ce nuage. Que peut-on déduire ?.
3) Déterminer la matrice des variances-covariances V.
4) Posons $S = {}^{t}X.X$. Montrer que S possède une valeur propre nulle (sans faire de calculs).
5) Calculer les valeurs propres de S et en déduire celles de V.
6) Déterminer le meilleur plan qui ajuste $N(I)$. (ACP non normée).