Fiche de Biostatistique – Stage 1

Représentation des données multivariées

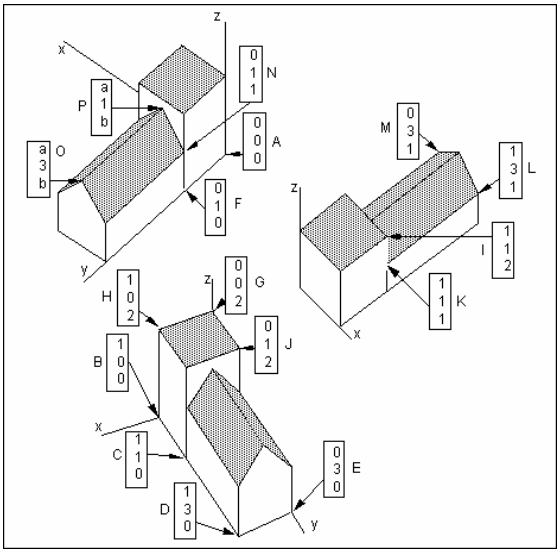
D. Chessel & A.B. Dufour

Résumé

La fiche donne les principes d'une représentation de données numériques à plus de deux composantes.

Plan

1. Représenter trois dimensions



Considérons un bâtiment formé d'une tour carrée de base 1x1 (supposons que l'unité de mesure soit fixée à un mètre) et de hauteur 2 et d'une petite maison de largeur 1, longueur 2 et hauteur 1 surmontée d'un toit symétrique à deux pentes ($a = \frac{1}{2}$, $b = \sqrt{3}$).

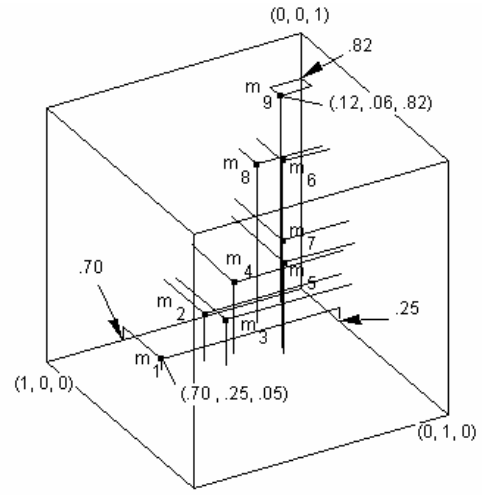
La totalité de l'information est contenue dans la matrice des coordonnées à trois dimensions de 16 points :

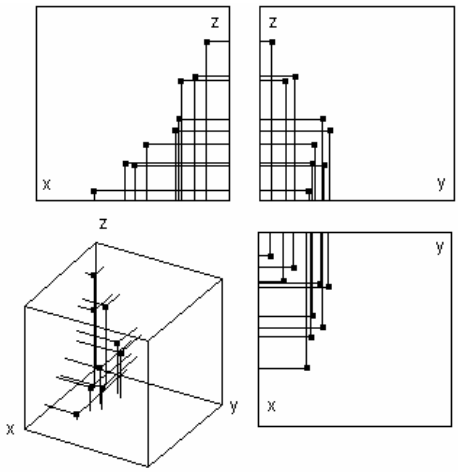
Points	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
x	0	1	1	1	0	0	0	1	1	0	1	1	0	0	a	a
y	0	0	1	3	3	1	0	0	1	1	1	3	3	1	3	1
z	0	0	0	0	0	0	2	2	2	2	1	1	1	1	b	b

La question est « comment un des dessins ci-dessus est-il calculé ? ».

Considérons la mesure de la proportion de trois catégories (Argile-Limon-Sable dans un sol ou Primaire-Secondaire-Tertiaire dans les statistiques de l'emploi ou A-B-C pour trois catégories de proies dans un estomac) dans 9 unités :

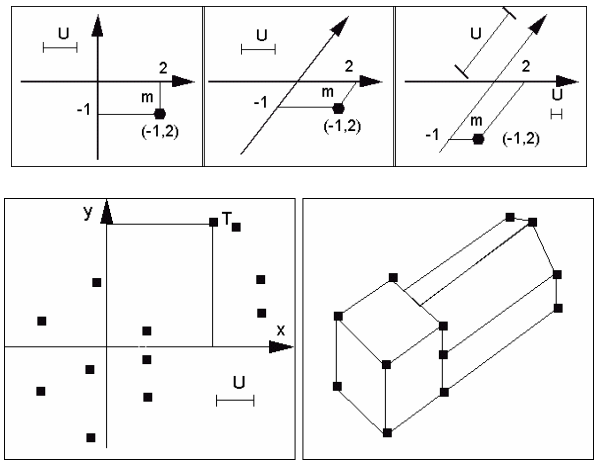
	1	2	3	4	5	6	7	8	9
X	70	54	49	43	28	18	26	25	12
Y	25	27	33	28	36	18	32	13	6
Z	5	19	18	29	36	64	42	62	82





Trois représentations bivariées

La question est la même. Pour représenter un objet à trois dimensions, il suffit de connaître deux coordonnées.



1.1. Espace euclidien

Un point ou vecteur de $\mathbb{R}^2$ est un couple de nombres réels, soit un objet du type :

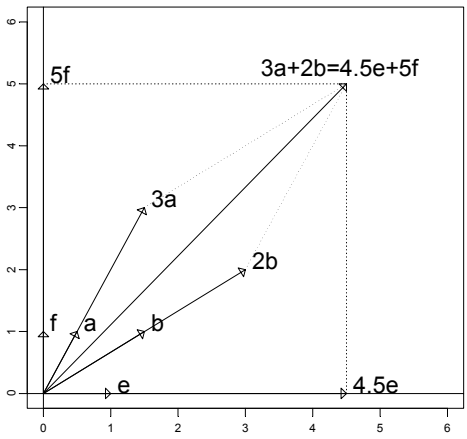
$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \Leftrightarrow \mathbf{x}^t = \begin{bmatrix} x_1 & x_2 \end{bmatrix}$$

$\mathbb{R}^2$ est un espace vectoriel.

```
tracer <- function(x,cha){
   arrows(0,0,x[1],x[2])
   text(x[1]+0.1,x[2]+0.2,cha,adj=0,cex=2)
}
relier <- function(x,y){
   segments(x[1],x[2],y[1],y[2],lty=2)
}
plot (c(0,0),xlim=c(0,6),ylim=c(0,6),type="n",xlab="",ylab="")
abline(h=0)
abline(v=0)
a <- c(0.5,1)
b <- c(1.5,1)
c <- 3*a+2*b
tracer(a,"a")
```

```
tracer(3*a,"3a")
tracer(b,"b")
tracer(2*b,"2b")
tracer (c,"")
text(locator(1),"3a+2b=4.5e+5f",adj=0.5,cex=2)
relier (2*b,c)
relier (3*a,c)

f <- c(0,1)
e <- c(1,0)
tracer(e,"e")
tracer(5*f,"5f")
tracer(f,"f")
tracer(4.5*e,"4.5e")
relier (5*f,c)
relier (4.5*e,c)
```



On y rajoute le produit scalaire canonique (PSC) défini par :

$$\langle \mathbf{x} | \mathbf{y} \rangle = x_1 y_1 + x_2 y_2$$

Pour faire les calculs :

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \quad \langle \mathbf{x} | \mathbf{y} \rangle = \mathbf{x}^t \mathbf{y} = \mathbf{y}^t \mathbf{x} = \begin{bmatrix} x_1 y_1 + x_2 y_2 \end{bmatrix}$$

Cet abus de langage (confusion entre un nombre et la matrice 1-1 qui contient ce nombre) est très pratique.

Un point ou vecteur de $\mathbb{R}^3$ est un triplet de nombres réels, soit un objet du type **x** tel que $\mathbf{x}^t = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}$. $\mathbb{R}^3$ est un espace vectoriel. On y rajoute le produit scalaire canonique (PSC) défini par :

$$\langle \mathbf{x} | \mathbf{y} \rangle = x_1 y_1 + x_2 y_2 + x_3 y_3 = \sum_{i=1}^{3} x_i y_i = \mathbf{x}^t \mathbf{y} = \mathbf{y}^t \mathbf{x} = \begin{bmatrix} x_1 y_1 + x_2 y_2 + x_3 y_3 \end{bmatrix}$$

.

Un point ou vecteur de $\mathbb{R}^s$ est un s-uple de nombres réels, soit un objet du type **x** tel que $\mathbf{x}^t = \begin{bmatrix} x_1 & x_2 & \cdots & x_s \end{bmatrix}$. $\mathbb{R}^s$ est un espace vectoriel. On y rajoute le produit scalaire canonique (PSC) défini par :

$$\langle \mathbf{x} | \mathbf{y} \rangle = \sum_{i=1}^{s} x_i y_i = \mathbf{x}^t \mathbf{y} = \mathbf{y}^t \mathbf{x} = \left[\sum_{i=1}^{s} x_i y_i \right]$$

L'application PSC vérifie les propriétés :

PS1 Symétrie $\langle \mathbf{x} | \mathbf{y} \rangle = \langle \mathbf{y} | \mathbf{x} \rangle$ pour tout **x** et **y** de $\mathbb{R}^s$

PS2a Linéarité $\langle \mathbf{x} | \mathbf{y} + \mathbf{z} \rangle = \langle \mathbf{x} | \mathbf{y} \rangle + \langle \mathbf{x} | \mathbf{z} \rangle$ pour tout **x**, **y**, **z** de $\mathbb{R}^s$

PS2b Linéarité $\langle \mathbf{x} | \alpha \mathbf{y} \rangle = \alpha \langle \mathbf{x} | \mathbf{y} \rangle$ pour tout **x** et **y** de $\mathbb{R}^s$ et pour tout α de $\mathbb{R}$

PS3 Positivité $\langle \mathbf{x} | \mathbf{x} \rangle \geq 0$ pour tout **x** de $\mathbb{R}^s$

PS4 Non dégénérescence $\langle \mathbf{x} | \mathbf{x} \rangle = 0 \Rightarrow \mathbf{x} = \mathbf{0}$

Symétrie et linéarité impliquent la bilinéarité.

1.1.1. Produits scalaires

En toute généralité, étant donnée une fonction de $\mathbb{R}^s$ x $\mathbb{R}^s$ dans $\mathbb{R}$ qui à un couple de points (**x**,**y**) associe un nombre réel $\langle \mathbf{x} | \mathbf{y} \rangle_\Phi$ on dit que c'est un produit scalaire si elle vérifie les propriétés **PS1**, ..., **PS4** :

PS1 $\langle \mathbf{x} | \mathbf{y} \rangle_\Phi = \langle \mathbf{y} | \mathbf{x} \rangle_\Phi$ pour tout **x** et **y** de $\mathbb{R}^s$

PS2a $\langle \mathbf{x} | \mathbf{y} + \mathbf{z} \rangle_\Phi = \langle \mathbf{x} | \mathbf{y} \rangle_\Phi + \langle \mathbf{x} | \mathbf{z} \rangle_\Phi$ pour tout **x**, **y**, **z** de $\mathbb{R}^s$

PS2b $\langle \mathbf{x} | \alpha \mathbf{y} \rangle_\Phi = \alpha \langle \mathbf{x} | \mathbf{y} \rangle_\Phi$ pour tout **x** et **y** de $\mathbb{R}^s$ et pour tout α de $\mathbb{R}$

PS3 $\langle \mathbf{x} | \mathbf{x} \rangle_\Phi \geq 0$ pour tout **x** de $\mathbb{R}^s$

PS4 $\langle \mathbf{x} | \mathbf{x} \rangle_\Phi = 0 \Rightarrow \mathbf{x} = \mathbf{0}$

Si une telle fonction existe, il s'en suit immédiatement qu'elle vérifie les propriétés :

PS2c $\langle \mathbf{x} + \mathbf{z} | \mathbf{y} \rangle_\Phi = \langle \mathbf{x} | \mathbf{y} \rangle_\Phi + \langle \mathbf{z} | \mathbf{y} \rangle_\Phi$ pour tout **x**, **y**, **z** de $\mathbb{R}^s$

PS2g $\langle \alpha \mathbf{x} | \mathbf{y} \rangle_\Phi = \alpha \langle \mathbf{x} | \mathbf{y} \rangle_\Phi$ pour tout **x** et **y** de $\mathbb{R}^s$ et pour tout α de $\mathbb{R}$

Quand on manipule un seul produit scalaire, si aucune confusion n'est possible, on note simplement :

$$\langle \mathbf{x} | \mathbf{y} \rangle_\Phi = \langle \mathbf{x} | \mathbf{y} \rangle$$

Mathématiquement il est plus facile de travailler sur le cas général que sur un cas particulier. Concrètement, c'est beaucoup plus utile.

1.1.2. Norme

$$\| \mathbf{x} \|_\Phi^2 = \langle \mathbf{x} | \mathbf{x} \rangle_\Phi \Leftrightarrow \| \mathbf{x} \|_\Phi = \sqrt{\langle \mathbf{x} | \mathbf{x} \rangle_\Phi}$$

$$\|\alpha\mathbf{x}\|_\Phi = |\alpha| \ \|\mathbf{x}\|_\Phi$$

Quand aucune confusion n'est possible, on note simplement $\|\mathbf{x}\|_\Phi = \|\mathbf{x}\|$.

1.1.3. Orthogonalité

Deux vecteurs de $\mathbb{R}^S$ sont Φ-orthogonaux si et seulement si $\langle \mathbf{x} | \mathbf{y} \rangle_\Phi = 0$.

CNS : $\|\mathbf{x}+\mathbf{y}\|_\Phi^2 = \|\mathbf{x}\|_\Phi^2 + \|\mathbf{y}\|_\Phi^2$ (théorème de Pythagore)

CNS : $\|\mathbf{x}+\mathbf{y}\|_\Phi^2 = \|\mathbf{x}-\mathbf{y}\|_\Phi^2$

Calcul fondamental :

$$\|\mathbf{x}+\mathbf{y}\|_\Phi^2 = \langle \mathbf{x}+\mathbf{y} | \mathbf{x}+\mathbf{y} \rangle_\Phi = \langle \mathbf{x} | \mathbf{x} \rangle_\Phi + \langle \mathbf{x} | \mathbf{y} \rangle_\Phi + \langle \mathbf{y} | \mathbf{x} \rangle_\Phi + \langle \mathbf{y} | \mathbf{y} \rangle_\Phi = \|\mathbf{x}\|_\Phi^2 + 2\langle \mathbf{x} | \mathbf{y} \rangle_\Phi + \|\mathbf{y}\|_\Phi^2$$

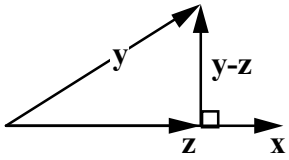
Dans la suite, on utilise un produit scalaire unique Φ.

1.1.4. Existence du projecteur Φ- orthogonal sur un vecteur

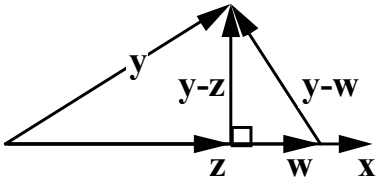
Si **x** et **y** sont deux vecteurs de $\mathbb{R}^S$ et si **x** est non nul, il existe un unique vecteur **z** de $\mathbb{R}^S$ proportionnel à **x** tel que **y** – **z** soit orthogonal à **x**. On dit que **z** est le projeté Φ-orthogonal de **y** sur **x**.

Il vaut :

$$\mathbf{z} = \frac{\langle \mathbf{y} | \mathbf{x} \rangle}{\|\mathbf{x}\|^2} \mathbf{x}$$



1.1.5. Théorème du pied de la perpendiculaire



Si **x** et **y** sont deux vecteurs de $\mathbb{R}^S$ et si **x** est non nul, le vecteur **w** de $\mathbb{R}^S$ proportionnel à **x** qui minimise $\|\mathbf{y}-\mathbf{w}\|^2$ est le projeté de **x** sur **y**.

1.1.6. Inégalité de Cauchy-Schwartz - Angle de deux vecteurs

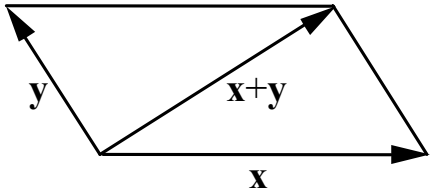
$$\left|\langle \mathbf{x}|\mathbf{y}\rangle\right| \leq \|\mathbf{x}\|\|\mathbf{y}\|$$

Si **x** et **y** sont deux vecteurs de $\mathbb{R}^s$, la mesure de l'angle de **x** et **y**, notée A(**x**,**y**) = a, est définie par $0 \leq a \leq \pi$ et :

$$\cos a = \frac{\langle \mathbf{x}|\mathbf{y}\rangle}{\|\mathbf{x}\|\|\mathbf{y}\|}$$

1.1.7. Inégalité triangulaire

$$\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$$



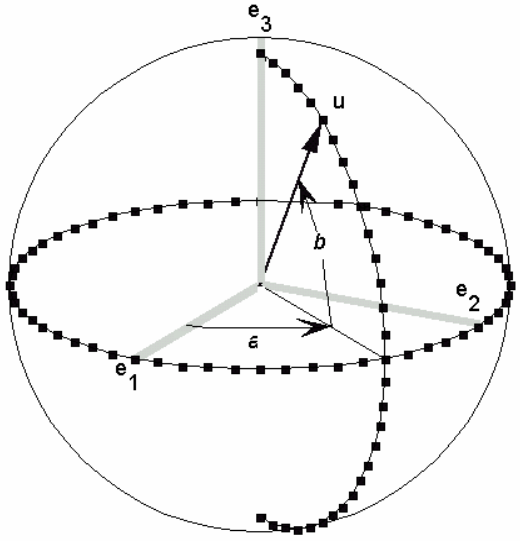
1.1.8. Distance entre deux vecteurs

Pour tout couple (**x**,**y**) de points de $\mathbb{R}^s$, on appelle distance entre **x** et **y** la quantité :

$$d_\Phi(\mathbf{x},\mathbf{y}) = d(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\| = \|\mathbf{y}-\mathbf{x}\|$$

On sait donc mesurer les angles et les distances entre points de $\mathbb{R}^s$ au sens d'un produit scalaire donné.

1.2. Base orthonormale



Définition d'une base orthonormale :

$$\mathbf{H} = \begin{bmatrix} \mathbf{u} & \mathbf{v} & \mathbf{w} \\ [\ ] & [\ ] & [\ ] \end{bmatrix} \begin{matrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{e}_3 \end{matrix} = \begin{bmatrix} \cos a \cos b & -\sin a & -\cos a \sin b \\ \sin a \cos b & \cos a & -\sin a \sin b \\ \sin b & 0 & \cos b \end{bmatrix}$$

Coordonnées dans la base canonique :

$$\mathbf{x}^t = \begin{bmatrix} x & y & z \end{bmatrix} \Leftrightarrow \mathbf{x} = x\mathbf{e}_1 + y\mathbf{e}_2 + z\mathbf{e}_3$$

Coordonnées dans la nouvelle base :

$$\mathbf{x} = \lambda\mathbf{u} + \mu\mathbf{v} + \nu\mathbf{w} = \langle \mathbf{x} | \mathbf{u} \rangle \mathbf{u} + \langle \mathbf{x} | \mathbf{v} \rangle \mathbf{v} + \langle \mathbf{x} | \mathbf{w} \rangle \mathbf{w}$$

Changement de base :

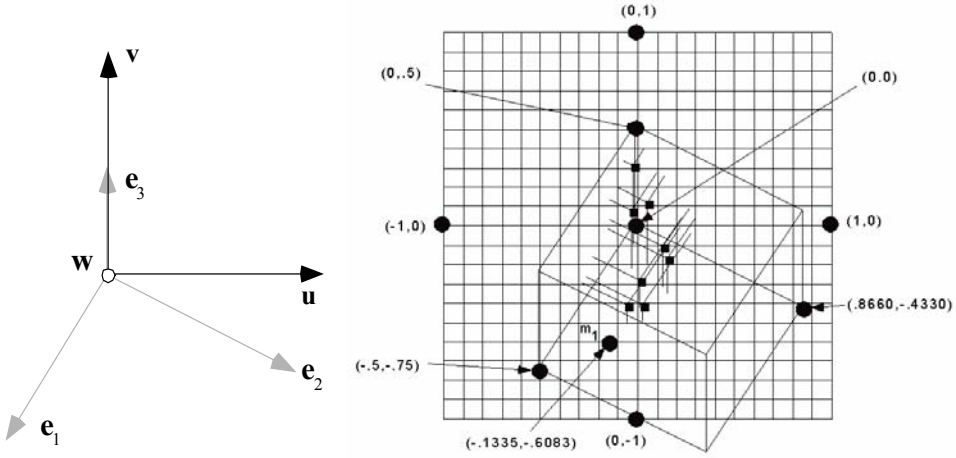
$$\begin{bmatrix} \lambda \\ \mu \\ \nu \end{bmatrix} = \begin{bmatrix} \langle \mathbf{x} | \mathbf{u} \rangle \\ \langle \mathbf{x} | \mathbf{v} \rangle \\ \langle \mathbf{x} | \mathbf{w} \rangle \end{bmatrix} = \begin{bmatrix} \mathbf{x}^t\mathbf{u} \\ \mathbf{x}^t\mathbf{v} \\ \mathbf{x}^t\mathbf{w} \end{bmatrix} = \begin{bmatrix} \mathbf{u}^t\mathbf{x} \\ \mathbf{v}^t\mathbf{x} \\ \mathbf{w}^t\mathbf{x} \end{bmatrix} = \begin{matrix} \mathbf{u}^t \\ \mathbf{v}^t \\ \mathbf{w}^t \end{matrix} \begin{bmatrix} [\ ] \\ [\ ] \\ [\ ] \end{bmatrix} [\ ] = \mathbf{H}^t \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

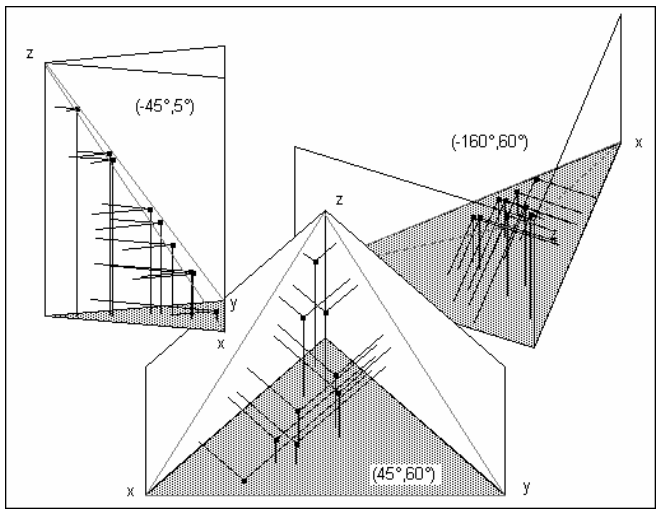
Matrice orthonormale :

$$\mathbf{H}^t\mathbf{H} = \begin{matrix} \mathbf{u}^t \\ \mathbf{v}^t \\ \mathbf{w}^t \end{matrix} \begin{bmatrix} [\ ] \\ [\ ] \\ [\ ] \end{bmatrix} \begin{bmatrix} \mathbf{u} & \mathbf{v} & \mathbf{w} \\ [\ ] & [\ ] & [\ ] \end{bmatrix} = \begin{bmatrix} \mathbf{u}^t\mathbf{u} & \mathbf{u}^t\mathbf{v} & \mathbf{u}^t\mathbf{w} \\ \mathbf{v}^t\mathbf{u} & \mathbf{v}^t\mathbf{v} & \mathbf{v}^t\mathbf{w} \\ \mathbf{w}^t\mathbf{u} & \mathbf{w}^t\mathbf{v} & \mathbf{w}^t\mathbf{w} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \mathbf{I}_3$$

Exemple :

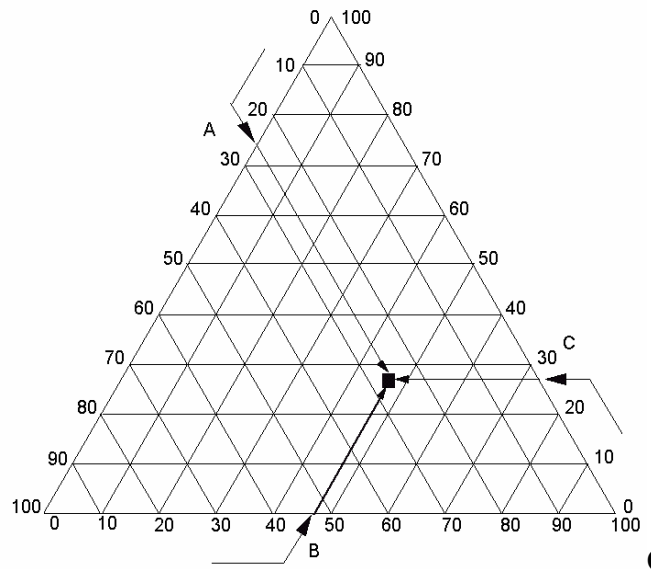
$$\begin{bmatrix} \lambda \\ \mu \\ \nu \end{bmatrix} = \mathbf{H}^t \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} .4330 & .25 & .8660 \\ -.5 & .8660 & 0 \\ -.75 & -.4330 & .5 \end{bmatrix} \begin{bmatrix} .70 \\ .25 \\ .05 \end{bmatrix} = \begin{bmatrix} .4089 \\ -.1335 \\ -.6083 \end{bmatrix}$$



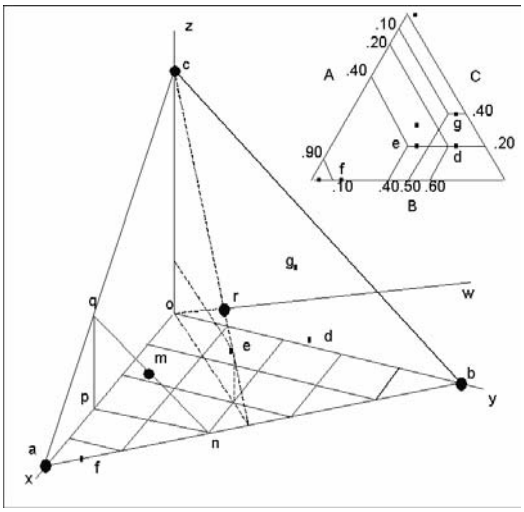


Toutes les représentations n'ont pas la même valeur.

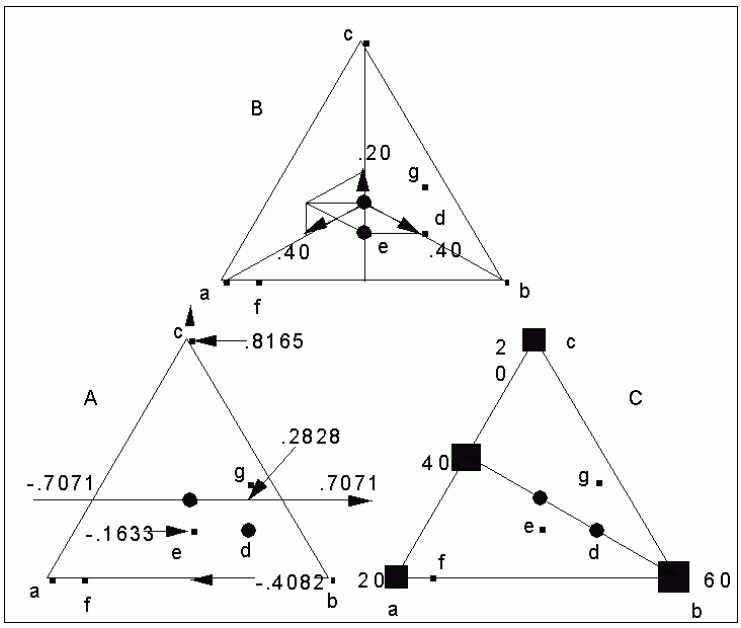
1.3. Représentation triangulaire



Ordinaire



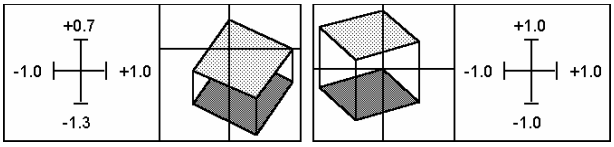
Euclidienne



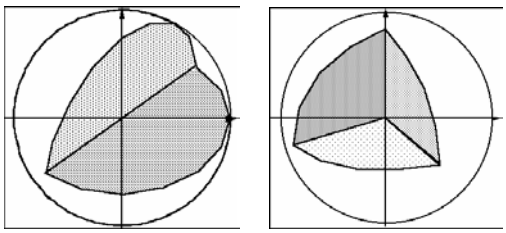
Mécanique, Informatique, Géométrique

1.4. Exercices

Soit le cube défini par les 8 points :
A(0,0,0), B(0,1,0), C(1,1,0), D(1,0,0), E(0,0,1), F(0,1,1), G(1,1,1), H(1,0,1).
Représenter cet objet vu dans les directions (30°, 60°) et (60°,30°).



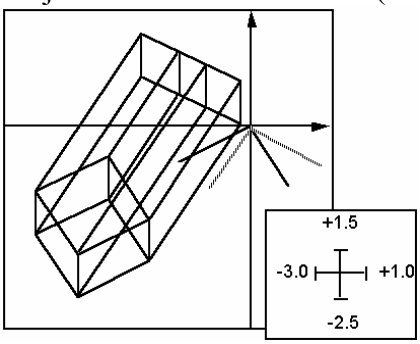
S est la sphère de centre (0,0,0) et de rayon unité. Représenter cet objet auquel on a enlevé tous les points (x,y,z) tels que y>0 et z>0 vu dans la direction (45°,45°). Représenter la même sphère de laquelle on a enlevé les points (x,y,z) tels que x>0, y>0 et z>0 vue dans la direction (60°,30°).



Soit l'objet défini par trois carrés opaques de sommets (1, 1) (1, -1) (-1, -1) et (-1, 1) respectivement dans les plans xOy, xOz et yOz. Représenter cet objet dans la direction 60°, 60°.

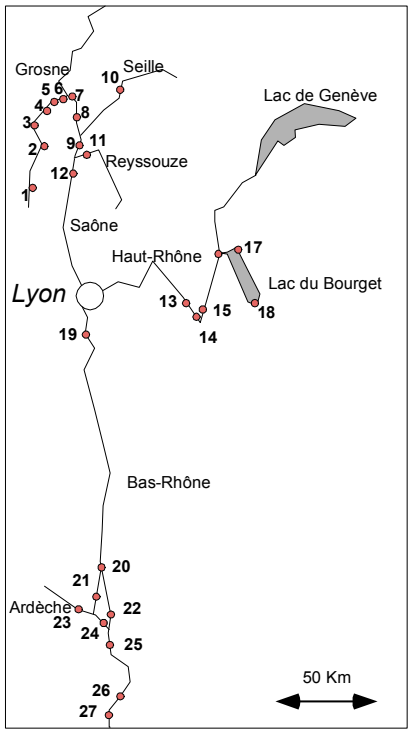
Soit l'objet défini par les 8 points :
A(2,0,0), B(3,0,0), C(3,1,0), D(2,1,0)
E(2,0,1), F(3,0,1), G(3,1,1), H(2,1,1)

Représenter cet objet vu dans la direction (60°,60°). Sur le même graphique représenter la base associée à la direction (30°,0°) et la base canonique. Sur le même graphique représenter la projection de l'objet associée à la direction (30°,0°).



Chez le Chevaine, le locus PGM-2* est polymorphe et présente trois allèles, respectivement notés 90, 98 et 100. Les fréquences alléliques sont estimées dans des échantillons prélevés dans 27 stations du bassin Rhôdanien (Guinand, B., Bouvet, Y. & Brohon, B. (1996) Spatial aspects of genetic differentiation of the European chub in the Rhone River basin. *Journal of Fish Biology* : 49, 714-726).

	Locus PGM-2*		
Stations	***90***	***98***	***100***
1	1.7	15	83.3
2	8.3	21.6	70.1
3	10.7	17.9	71.4
4	6.7	20	73.3
5	10.3	19	70.7
6	17.9	12.5	69.6
7	21.6	20	58.4
8	25	11.5	63.5
9	17.5	5	77.5
10	28.2	0	71.8
11	15.4	3.8	80.8
12	12.5	10.4	77.1
13	3.3	71.6	25.1
14	3.3	60	36.7
15	4	62	34
16	0	69	31
17	0.3	65.9	33.8
18	0	53.6	46.4
19	12.9	34.3	52.8
20	15	13.3	71.7
21	13.8	10.3	75.9
22	17.6	5.9	76.5
23	12.5	8.9	78.6
24	11.4	13.6	75
25	13.5	1	85.5
26	8.8	10.3	80.9
27	5.9	11.8	82.3

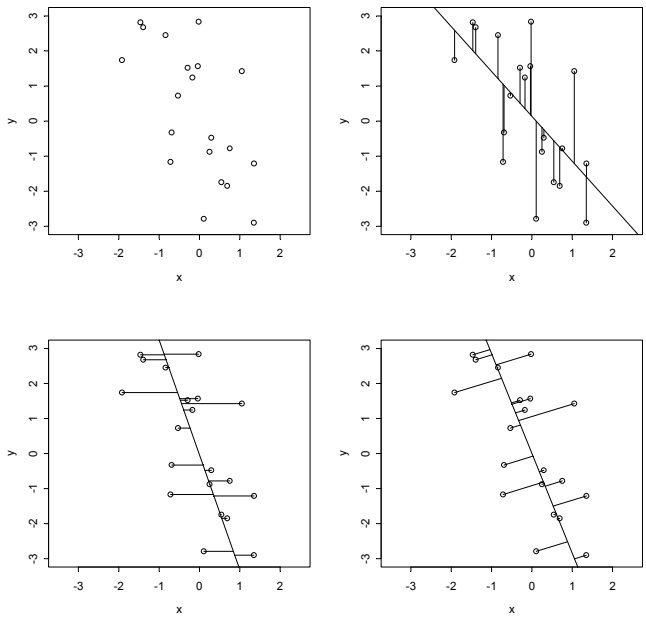


Donner une représentation triangulaire de ces données et commenter le résultat obtenu.

2. Axes principaux

2.1. Définition du problème en dimension 2

Quand on dispose de la mesure de deux variables sur *n* individus, on peut se poser trois questions exprimées par :



1) Prédire **y** par **x** et trouver une droite qui minimise $\frac{1}{n}\sum_{i=1}^{n}\left(y_i - ax_i - b\right)^2$

2) Prédire **x** par **y** et trouver une droite qui minimise $\frac{1}{n}\sum_{i=1}^{n}\left(x_i - a'y_i - b'\right)^2$

3) Étudier la liaison entre **x** et **y** et trouver une droite qui minimise :

$\frac{1}{n}\sum_{i=1}^{n} \mathrm{M}_i\mathrm{m}_i^2$ où m_i est la projection orthogonale de M_i sur la droite.

La première droite est dite droite de régression de Y/X, la seconde est dite droite de régression de X/Y, la troisième est dite direction principale du nuage centré.

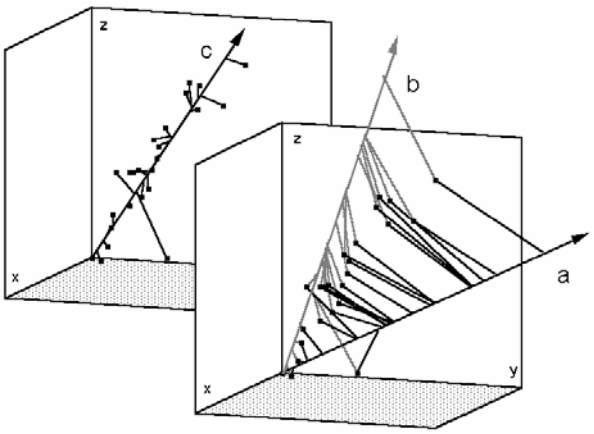
2.2. Un problème à trois variables

	X	Y	Z		X	Y	Z		X	Y	Z
1	98	81	38	9	133	102	51	17	149	107	55
2	103	84	38	10	133	105	41	18	153	107	56
3	103	86	42	11	134	100	48	19	155	115	63
4	105	86	40	12	136	102	49	20	155	117	60
5	109	88	44	13	137	98	51	21	158	115	62
6	123	92	50	14	138	99	51	22	159	118	63
7	123	95	46	15	141	105	53	23	162	124	61

24 carapaces de tortues peintes (*Chrysemis picta marginata*)

X longueur (en mm), Y largeur (en mm), Z hauteur (en mm)[1]

[1] Jolicœur, P. & Mosimann, J.E. (1960) Size and shape variation in the painted turtle. A principal component analysis. Growth : 24, 339-354.



Une tortue porte trois mesures (x_i, y_i, z_i), c'est un point de $\mathbb{R}^3$. Si la croissance est la même dans toutes les directions toutes les carapaces sont proportionnelles à une forme de référence $(x_i, y_i, z_i) = k_i (x_0, y_0, z_0)$. Si on cherche à estimer ces paramètres, il faut ajouter une contrainte car $k_i (x_0, y_0, z_0) = \frac{k_i}{h}(hx_0, hy_0, hz_0)$.

On écrit cette contrainte $x_0^2 + y_0^2 + z_0^2 = 1$ et on cherche aux moindre carrés (x_0, y_0, z_0) et k_i qui minimisent :

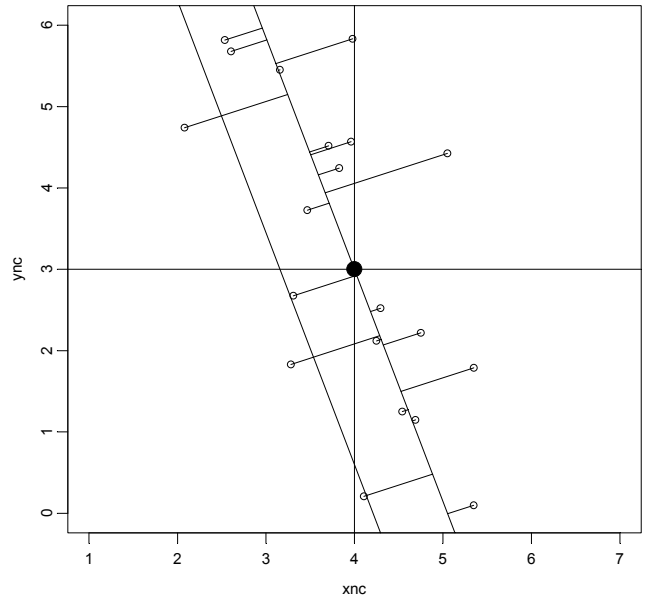
$$\sum_{i=1}^{n}\left((x_i - k_i x_0)^2 + (y_i - k_i y_0)^2 + (z_i - k_i z_0)^2\right)$$

Si on a trouvé une solution pour (x_0, y_0, z_0) d'après (1.1.5), k_i s'écrit nécessairement :

$$k_i = x_i x_0 + y_i y_0 + z_i z_0$$

Le problème se résume à trouver un vecteur $\mathbf{u} = (x_0, y_0, z_0)$ normé ($\|\mathbf{u}\|^2 = x_0^2 + y_0^2 + z_0^2 = 1$) qui minimise $\frac{1}{n}\sum_{i=1}^{n}\|\mathrm{M}_i\text{-}\mathrm{m}_i\|^2$ où m_i est la projection orthogonale de $\mathrm{M}_i = (x_i, y_i, z_i)$ sur le vecteur **u**.

Les deux problèmes qui viennent d'être discutés sont de même nature. Il présente seulement deux nuances. La première est une différence de dimensions (2 contre 3, mais cela n'a aucune importance car 2 ou 3 dimensions sont un cas particulier de *p* dimensions et, en mathématiques, 2 , 3 ou *p* dimensions posent exactement les mêmes problèmes). La seconde est une question d'origine des axes. Le point (0,0,0) dans le second problème correspond au (0,0,0) des données. Dans le premier, c'est moins clair.



Soit un nuage de *n* points de $\mathbb{R}^2$ de coordonnées (x_i, y_i). La droite qui minimise la moyenne des carrés des distances des points à cette droite passe par le point moyen de coordonnées $\left(m(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i, m(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} y_i \right)$. En effet, soient deux droites D et D' parallèles, la première passant par le point moyen (on dit aussi centre de gravité) la seconde ne passant pas par ce point moyen. Soit M_i le ième point, m_i sa projection sur D et m'_i sa projection sur D'.

$$\left\|\mathrm{M}_i - \mathrm{m}'_i\right\|^2 = \left\|\mathrm{M}_i + \mathrm{m}_i - \mathrm{m}_i - \mathrm{m}'_i\right\|^2 = \left\|\mathrm{M}_i - \mathrm{m}_i\right\|^2 + \left\|\mathrm{m}_i - \mathrm{m}'_i\right\|^2 + 2\left\langle \mathrm{M}_i - \mathrm{m}_i \middle| \mathrm{m}_i - \mathrm{m}'_i \right\rangle$$

Les vecteurs $\mathrm{m}_i - \mathrm{m}'_i$ sont tous égaux, ce qui s'écrit :

$$\left\|\mathrm{M}_i - \mathrm{m}'_i\right\|^2 = \left\|\mathrm{M}_i - \mathrm{m}_i\right\|^2 + \left\|w\right\|^2 + 2\left\langle \mathrm{M}_i - \mathrm{m}_i \middle| w \right\rangle$$

d'où : $\frac{1}{n}\sum_{i=1}^{n}\left\|\mathrm{M}_i - \mathrm{m}'_i\right\|^2 = \frac{1}{n}\sum_{i=1}^{n}\left\|\mathrm{M}_i - \mathrm{m}_i\right\|^2 + \left\|w\right\|^2 + \frac{2}{n}\sum_{i=1}^{n}\left\langle \mathrm{M}_i - \mathrm{m}_i \middle| w \right\rangle$

soit : $\frac{1}{n}\sum_{i=1}^{n}\left\|\mathrm{M}_i - \mathrm{m}'_i\right\|^2 = \frac{1}{n}\sum_{i=1}^{n}\left\|\mathrm{M}_i - \mathrm{m}_i\right\|^2 + \left\|w\right\|^2 + 2\left\langle \frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{M}_i - \mathrm{m}_i\right) \middle| w \right\rangle$

donc : $\frac{1}{n}\sum_{i=1}^{n}\left\|\mathrm{M}_i - \mathrm{m}'_i\right\|^2 = \frac{1}{n}\sum_{i=1}^{n}\left\|\mathrm{M}_i - \mathrm{m}_i\right\|^2 + \left\|w\right\|^2 > \frac{1}{n}\sum_{i=1}^{n}\left\|\mathrm{M}_i - \mathrm{m}_i\right\|^2$

C'est pourquoi on place l'origine au centre de gravité en utilisant les nouvelles coordonnées centrées :

$$\begin{cases} X_i = x_i - m(\mathbf{x}) \\ Y_i = y_i - m(\mathbf{y}) \end{cases}$$

2.3. Inertie d'un nuage de points

Les deux exemples précédents sont des cas particuliers d'un problème général. *n* points de $\mathbb{R}^p$ ont leurs coordonnées rangées dans une matrice à *n* lignes (individus) et *p* colonnes $\mathbf{X} = \left[x_{ij} \right]_{\substack{1 \le i \le n \\ 1 \le j \le p}}$. Chaque point a un poids et nous considérons pour simplifier que tous les poids sont égaux à $\frac{1}{n}$. L'ensemble des *n* points forme un nuage. L'inertie de ce nuage autour de l'origine vaut :

$$I_T = \frac{1}{n} \sum_{i=1}^{n} \left\| \mathrm{M}_i \right\|^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij}^{\ 2}$$

Cette quantité caractérise la variabilité totale de la position des points dans l'espace. Pour un nuage centré de $\mathbb{R}^2$:

$$\begin{aligned} I_T &= \frac{1}{n} \sum_{i=1}^{n} \left(x_i - m(\mathbf{x}) \right)^2 + \left(y_i - m(\mathbf{y}) \right)^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(x_i - m(\mathbf{x}) \right)^2 + \frac{1}{n} \sum_{i=1}^{n} \left(y_i - m(\mathbf{y}) \right)^2 = v(\mathbf{x}) + v(\mathbf{y}) \end{aligned}$$

Quand on prend dans $\mathbb{R}^p$ un vecteur unitaire **u**, il définit un axe. Le point M_i se projette sur cet axe en m_i. On a $\mathrm{M}_i = m_i + (\mathrm{M}_i - m_i)$ et $\left\| \mathrm{M}_i \right\|^2 = \left\| m_i \right\|^2 + \left\| \mathrm{M}_i - m_i \right\|^2$.

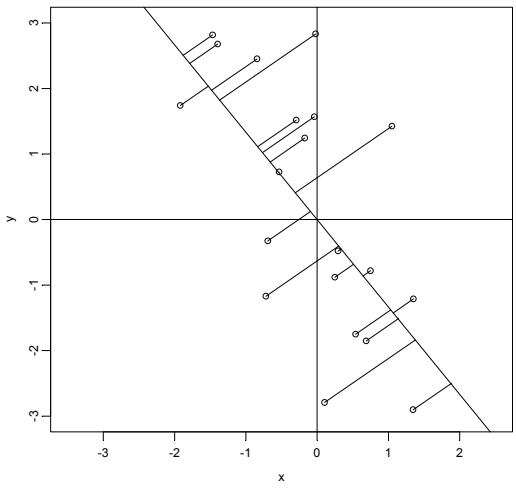
L'inertie totale se décompose en inertie projetée sur l'axe (on dit aussi inertie statistique) et inertie autour de l'axe (on dit aussi inertie mécanique) :

$$I_T = I_s(\mathbf{u}) + I_M(\mathbf{u}) = \frac{1}{n} \sum_{i=1}^{n} \left\| m_i \right\|^2 + \frac{1}{n} \sum_{i=1}^{n} \left\| \mathrm{M}_i - m_i \right\|^2$$

L'axe principal est celui qui minimise le second terme. Comme la somme des deux termes est constante et fourni par les données, il est équivalent de **maximiser l'inertie projetée** ou **minimiser l'inertie mécanique**.

2.4. Analyse en composantes principales centrée de deux variables

On présente la démonstration dans le cas de deux variables centrées et on généralisera sans insister sur les notations.



n point ont pour coordonnées $\begin{cases} X_i = x_i - m(\mathbf{x}) \\ Y_i = y_i - m(\mathbf{y}) \end{cases}$. Le centre de gravité du nuage est à l'origine. On prend un vecteur **u** unitaire sous la forme $\mathbf{u} = \begin{bmatrix} a \\ b \end{bmatrix}$ avec $a^2 + b^2 = 1$. La matrice **X** contient les *n* points en lignes, soit :

$$\mathbf{X} = \begin{bmatrix} x_1 - m(\mathbf{x}) & y_1 - m(\mathbf{y}) \\ \vdots & \vdots \\ x_n - m(\mathbf{x}) & y_n - m(\mathbf{y}) \end{bmatrix}$$

La matrice à *n* lignes et 1 colonne **Xu** contient les produits scalaires $\langle M_i | \mathbf{u} \rangle$. L'inertie projetée est :

$$I_s(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n} \|m_i\|^2 = \frac{1}{n}(\mathbf{Xu})^t \mathbf{Xu} = \mathbf{u}^t \left(\frac{1}{n}\mathbf{X}^t\mathbf{X}\right)\mathbf{u}$$

$$= \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} v(\mathbf{x}) & c(\mathbf{x},\mathbf{y}) \\ c(\mathbf{x},\mathbf{y}) & v(\mathbf{y}) \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = v(\mathbf{x})a^2 + 2c(\mathbf{x},\mathbf{y})ab + v(\mathbf{y})b^2$$

La matrice centrale qui contient

la variance de **x** , $v(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - m(\mathbf{x})\right)^2$

la variance de **y**, $v(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - m(\mathbf{y})\right)^2$

la covariance de **x** et **y**, $c(\mathbf{x},\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - m(\mathbf{x})\right)\left(y_i - m(\mathbf{y})\right)$

est dite matrice de variance-covariance des deux variables. Bien comprendre cette écriture efficace. On l'appelle :

$$\mathbf{C} = \begin{bmatrix} v(\mathbf{x}) & c(\mathbf{x},\mathbf{y}) \\ c(\mathbf{x},\mathbf{y}) & v(\mathbf{y}) \end{bmatrix} = \frac{1}{n}\mathbf{X}^t\mathbf{X}$$

Elle est symétrique. Son polynôme caractéristique s'écrit :

$$\left|\mathbf{C}-\lambda\mathbf{I}_2\right| = \begin{vmatrix} v(\mathbf{x})-\lambda & c(\mathbf{x},\mathbf{y}) \\ c(\mathbf{x},\mathbf{y}) & v(\mathbf{y})-\lambda \end{vmatrix} = \lambda^2 - \lambda\left(v(\mathbf{x})+v(\mathbf{y})\right) + v(\mathbf{x})v(\mathbf{y}) - c^2(\mathbf{x},\mathbf{y})$$

C'est un polynôme de degré 2 avec :

$$\Delta = \left(v(\mathbf{x})+v(\mathbf{y})\right)^2 - 4v(\mathbf{x})v(\mathbf{y}) + 4c^2(\mathbf{x},\mathbf{y}) = \left(v(\mathbf{x})-v(\mathbf{y})\right)^2 + 4c^2(\mathbf{x},\mathbf{y})$$

Le polynôme caractéristique a toujours deux racines, donc **C** a toujours deux valeurs propres et deux vecteurs propres. Les valeurs propres sont en général distinctes. Leur somme vaut $v(\mathbf{x})+v(\mathbf{y})$ et leur produit $-v(\mathbf{x})v(\mathbf{y})+c^2(\mathbf{x},\mathbf{y}) = c^2(\mathbf{x},\mathbf{y})\left(1-r^2(\mathbf{x},\mathbf{y})\right)$ est positif car la corrélation est comprise entre –1 et 1. La matrice **C** est diagonalisable. C'est en fait un cas particulier du théorème général qui dit que toute matrice symétrique admet une base de vecteurs propres orthogonaux.

Alors :

$$\mathbf{C} = \mathbf{U}\Lambda\mathbf{U}^t = \left[\begin{pmatrix} u_{11} \\ u_{12} \end{pmatrix}\begin{pmatrix} u_{21} \\ u_{22} \end{pmatrix}\right]\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}\begin{bmatrix} (u_{11} & u_{12}) \\ (u_{21} & u_{22}) \end{bmatrix}$$

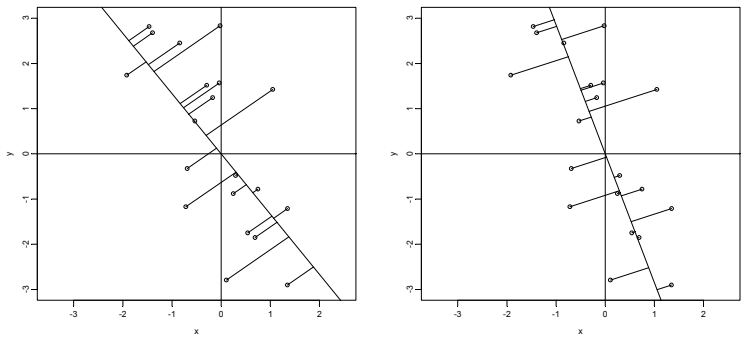
donc :

$$I_s(\mathbf{u}) = \begin{bmatrix} a & b \end{bmatrix}\left[\begin{pmatrix} u_{11} \\ u_{12} \end{pmatrix}\begin{pmatrix} u_{21} \\ u_{22} \end{pmatrix}\right]\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}\begin{bmatrix} (u_{11} & u_{12}) \\ (u_{21} & u_{22}) \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix}$$

$$= \begin{bmatrix} \alpha & \beta \end{bmatrix}\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \lambda_1\alpha^2 + \lambda_2\beta^2 \le \lambda_1\alpha^2 + \lambda_1\beta^2 = \lambda_1$$

$\begin{bmatrix} \alpha \\ \beta \end{bmatrix}$ représente les coordonnées du vecteur **u** dans la base des vecteurs propres. L'inertie ne peut dépasser la première valeur propre et l'atteint pour $\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, donc pour le premier vecteur propre.

L'axe principal d'un nuage bivarié est le premier vecteur propre de la matrice de variance-covariance des deux variables.



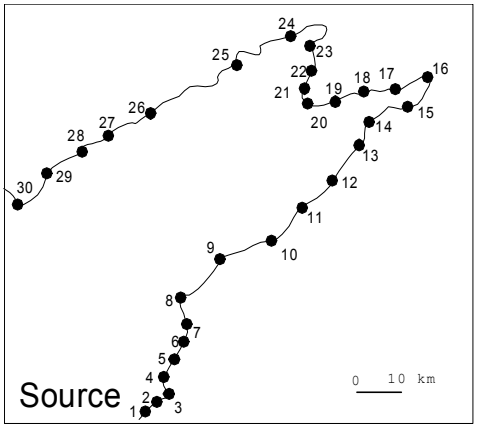
Quand l'axe tourne, il existe une position où la moyenne des carrés des distances est minimale. C'est l'axe principal (à droite).

Exercice : Montrer que lorsque les variances sont égales l'axe principal est sur une bissectrice des axes.

2.5. Analyse en composantes principales centrée

Dans la même logique, et sans plus de difficultés, on étend ce résultat en dimensions quelconques. On considère *p* variables. *p* peut être grand.

	Ablette	Anguille	Barbeau	Blageon	Bouvière	Brème_bord.	Brème_comm.	Brochet	Carpe	Chabot	Chevaine	Gardon	Goujon	Grémille	Hotu	Loche	Ombre	Perche	Perche_soleil	Poisson_chat	Rotengle	Spirlin	Tanche	Toxostome	Truite	Vairon	Vandoise
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	5	4	0
3	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	5	5	0
4	0	0	0	0	0	0	0	2	0	0	1	0	1	0	0	5	0	2	0	0	0	0	1	0	4	5	0
5	0	0	0	0	0	0	0	4	0	0	2	5	2	0	0	2	0	4	0	0	2	0	3	0	2	3	5
6	0	0	0	0	0	0	0	1	0	0	2	1	1	0	0	5	0	1	0	0	0	0	2	0	3	4	1
7	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	5	0	0	0	0	0	0	0	0	5	4	1
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	5	4	0	0	0	3	0	0	0	0	0	0	1	0	0	1	0
10	0	0	0	0	0	0	0	0	0	0	2	0	1	0	0	4	0	0	0	0	0	0	0	0	1	4	2
11	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	0	0	0	0	0	0	0	3	4	0
12	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0	4	2	0	0	0	0	0	0	0	5	4	0
13	0	0	0	2	0	0	0	0	0	2	0	0	0	0	0	2	3	0	0	0	0	0	0	0	5	5	0
14	0	0	1	3	0	0	0	1	0	3	1	0	1	0	0	4	4	0	0	0	0	0	0	0	5	5	0
15	0	0	2	4	0	0	0	0	0	3	3	0	2	0	0	5	2	0	0	0	0	0	1	0	4	4	3
16	0	0	2	5	0	0	0	1	1	2	2	1	2	0	0	5	0	1	1	0	0	1	1	4	3	3	5
17	2	1	3	2	1	0	0	1	1	1	2	2	1	0	1	4	1	2	1	0	0	4	1	4	2	4	3
18	2	1	3	1	2	0	0	1	1	1	3	2	2	1	1	3	1	3	1	0	0	3	1	3	1	3	2
19	3	1	2	1	2	1	1	1	1	0	1	5	4	1	2	5	0	1	1	0	1	2	2	3	0	3	2
20	5	2	4	0	3	2	1	2	1	0	3	5	4	2	2	2	0	2	2	0	2	3	4	2	0	1	2
21	5	2	4	0	3	3	3	3	2	0	2	5	5	3	2	1	0	3	2	1	2	2	4	2	0	1	2
22	5	2	5	0	3	4	4	3	3	0	4	5	5	4	3	1	0	4	3	2	2	1	4	2	0	0	3
23	2	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	5	0	0	0	0	1	0	0	0	0	2	2	1	2	1	0	0	0	1	0	0	0	0	0	0	0	0
25	3	0	0	0	0	0	0	1	0	0	1	1	2	1	0	0	0	0	0	0	1	0	0	0	0	0	1
26	5	2	2	0	2	2	2	2	1	0	2	4	3	4	1	1	0	1	2	1	1	1	3	0	0	0	1
27	5	3	4	0	3	4	3	4	2	0	3	5	4	5	1	1	0	1	3	2	1	1	5	1	0	0	2
28	5	4	3	0	4	5	3	3	4	0	4	5	4	5	1	1	0	2	4	3	2	1	4	1	0	0	2
29	5	4	5	1	5	4	4	5	3	0	4	5	5	5	2	1	1	4	5	4	2	3	3	2	1	1	3
30	5	5	3	0	5	5	5	4	5	0	3	5	5	5	1	0	0	5	3	5	5	5	5	2	0	0	3



30 stations – 27 espèces

Source (extrait) : Verneaux, J. (1973) Cours d'eau de Franche-Comté (Massif du Jura). Recherches écologiques sur le réseau hydrographique du Doubs. Essai de biotypologie. Thèse d'état, Besançon. 1-257.

Y est le tableau brut. On calcule la moyenne de chacune des variables :

$$m_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij}$$

X est le tableau centré. Il a les mêmes dimensions que **Y** et son terme général est :

$$x_{ij} = y_{ij} - m_j$$

La matrice $\mathbf{C} = \frac{1}{n}\mathbf{X}^t\mathbf{X}$ contient à la ligne *j* et à la colonne *k* la covariance des variables *j* et *k* :

$$c_{jk} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik}$$

La matrice **C** est symétrique. Elle a *p* lignes et *p* colonnes (dans l'exemple *p* = 30). Elle admet une base de vecteurs propres orthonormés du type :

$$\mathbf{C} = \frac{1}{n}\mathbf{X}^t\mathbf{X} = \mathbf{U}\Lambda\mathbf{U}^t$$

Remarque : La diagonalisation d'une matrice 30-30 n'est possible qu'avec un ordinateur mais la théorie est connue depuis 1901 **1** et utilisée depuis les années 50 **2**. Le premier vecteur propre normé est un vecteur de $\mathbb{R}^p$ qui maximise l'inertie projetée. Le second vecteur propre normé est un vecteur de $\mathbb{R}^p$ orthogonal au premier qui maximise à nouveau l'inertie projetée. Et ainsi de suite. Si $\mathbf{u}_k$ est le vecteur propre de rang *k*, les coordonnées des projections des *n* points sont obtenues simplement par :

$$\mathbf{l}_k = \begin{bmatrix} l_{1k} \\ \vdots \\ l_{nk} \end{bmatrix} = \begin{bmatrix} \langle M_1 | \mathbf{u}_k \rangle \\ \vdots \\ \langle M_n | \mathbf{u}_k \rangle \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{p} x_{1j}u_{jk} \\ \vdots \\ \sum_{j=1}^{p} x_{nj}u_{jk} \end{bmatrix} = \mathbf{X}\mathbf{u}_k$$

Le vecteur $\mathbf{u}_k$ est appelé **l'axe principal** de rang *k*. Le vecteur $\mathbf{l}_k$ est un vecteur de $\mathbb{R}^n$. On l'appelle le vecteur des **coordonnées** sur l'axe principal de rang *k*. Ce vecteur est une nouvelle variable artificielle qui a une moyenne nulle, à cause du centrage :

$$m(\mathbf{l}_k) = \frac{1}{n}\sum_{i=1}^{n} l_{ik} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}u_{jk} = \sum_{j=1}^{p} u_{jk}\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$$

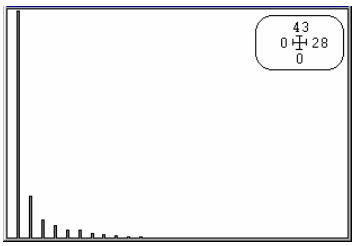
Sa variance vaut :

$$v(\mathbf{l}_k) = \frac{1}{n}\sum_{i=1}^{n} l_{ik}^2 = \frac{1}{n}(\mathbf{X}\mathbf{u}_k)^t\,\mathbf{X}\mathbf{u}_k = \mathbf{u}_k^t\mathbf{C}\mathbf{u}_k = \lambda_k\mathbf{u}_k^t\mathbf{u}_k = \lambda_k\|\mathbf{u}_k\|^2 = \lambda_k$$

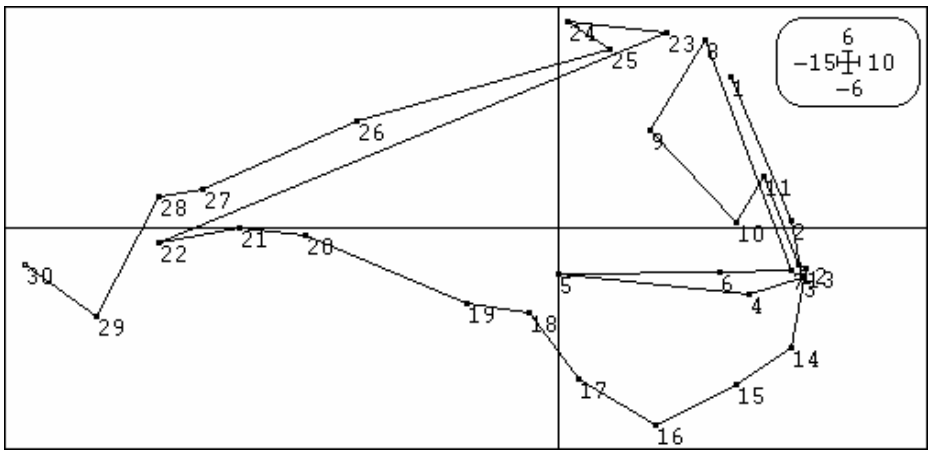
1 Pearson, K. (1901) On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* : 2, 559-572.

2 Goodall, D.W. (1954) Objective methods for the classification of vegetation III. An essay in the use of factor analysis. *Australian Journal of Botany* : 2, 304-324.

La coordonnée sur l'axe de rang *k* est centrée de **variance la valeur propre associée**. La somme des valeurs propres est l'inertie totale et vaut la somme des variances des variables de départ. Le **graphe des valeurs propres** exprime donc la manière dont la variabilité des données se répartit dans l'espace. Dans l'exemple, on trouve :



La **carte factorielle** est la représentation du nuage projeté sur un couple d'axes principaux. C'est une manière de voir de l'information multidimensionnelle. Sur la carte factorielle 1-2, chaque point est positionné par ses deux coordonnées (l_{i1}, l_{i2}). C'est l'image de la projection du nuage sur le plan qui représente une part maximale de la variabilité.



Le nuage de points a dans l'espace une structure simple. 77% de la variabilité totale peut être exprimée à l'aide des deux premières coordonnées. Le **tableau de décomposition de l'inertie** est :

Variables	Variances	Taux
Ablette	0.850	0.013
Anguille	4.023	0.061
Barbeau	3.796	0.057
Blageon	3.579	0.054
Bouvière	0.983	0.015
Brème_bord.	1.632	0.025
Brème_comm.	0.707	0.011
Brochet	1.649	0.025
Carpe	2.179	0.033
Chabot	1.782	0.027
Chevaine	2.979	0.045
Gardon	1.890	0.029
Goujon	3.272	0.050
Grémille	2.222	0.034
Hotu	2.293	0.035
Loche	2.623	0.040
Ombre	1.899	0.029
Perche	1.277	0.019
Perche_soleil	1.739	0.026
Poisson_chat	2.917	0.044
Rotengle	2.249	0.034
Spirlin	1.640	0.025
Tanche	3.462	0.052
Toxostome	4.690	0.071
Truite	2.832	0.043
Vairon	4.890	0.074
Vandoise	2.023	0.031
Somme	66.078	1.000

Axes	Variances	Taux
1	42.746	0.647
2	8.158	0.123
3	3.723	0.056
4	2.641	0.040
5	1.862	0.028
6	1.703	0.026
7	1.175	0.018
8	0.967	0.015
9	0.635	0.010
10	0.490	0.007
11	0.441	0.007
12	0.352	0.005
13	0.328	0.005
14	0.243	0.004
15	0.194	0.003
16	0.103	0.002
17	0.099	0.001
18	0.073	0.001
19	0.060	0.001
20	0.029	0.000
21	0.018	0.000
22	0.016	0.000
23	0.010	0.000
24	0.006	0.000
25	0.005	0.000
26	0.001	0.000
27	0.001	0.000
	66.078	1.000

2.6. ACP générale

On dit Analyse en composantes principales centrée quand la représentation porte sur les données centrées. La théorie s'applique à d'autres transformations initiales des données.

On retiendra la démarche générale. **Y** est le tableau brut. **X** est le tableau transformé. Il a les mêmes dimensions que **Y** et son terme général est :

$$x_{ij} = f\left(y_{ij}\right)$$

La matrice $\mathbf{C} = \frac{1}{n}\mathbf{X}^t\mathbf{X}$ contient à la ligne *j* et à la colonne *k* le terme :

$$c_{jk} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik}$$

La matrice **C** est symétrique. Elle a *p* lignes et *p* colonnes et elle admet une base de vecteurs propres orthonormés du type :

$$\mathbf{C} = \frac{1}{n}\mathbf{X}^t\mathbf{X} = \mathbf{U}\Lambda\mathbf{U}^t$$

La projection du nuage sur l'ensemble des axes principaux se fait simultanément par :

$$\mathbf{L} = \mathbf{XU}$$

On garde seulement les premiers axes pour voir la forme du nuage en exécutant des cartes factorielles.

Une enquête sur la pratique du sport a conduit à estimer la répartition des pratiquants d'un sport donné en fonction de la taille des agglomérations. On obtient le tableau ci-dessous (9 lignes et 6 colonnes). En ligne, figurent les sports principaux :
1 — Gymnastique 2 — Jogging 3 — Cyclisme 4 — Football 5 — Marche
6 — Natation 7 — Ski 8 — Tennis 9 — Autres

En colonne, on note le type de résidence des sportifs :

a- communes rurales b- cantons ruraux partiellement urbains
c- petites villes (≤20 000 hab) d- villes moyennes (≤100 000 hab)
e- grandes villes (sauf Paris) f- agglomération parisienne

	a	b	c	d	e	f
1	**0.074**	0.136	0.143	0.129	0.320	0.198
2	0.049	0.096	0.125	0.139	0.305	0.286
3	0.082	0.173	0.171	0.134	0.264	0.176
4	0.122	0.188	0.133	0.117	0.278	**0.162**
5	0.044	0.114	0.178	0.104	0.329	0.231
6	0.037	0.094	0.156	0.114	0.325	0.274
7	0.035	0.167	0.149	0.119	0.337	0.193
8	0.057	0.125	0.139	0.123	0.345	0.211
9	0.068	0.140	0.156	0.121	0.306	0.209

Le tableau indique donc que 7.4 % des pratiquants de la gymnastique habitent une commune rurale et 16.2 % des pratiquants du football sont parisiens. On estime qu'il serait bon de tenir compte de la répartition totale de la population entre les 6 types d'habitat. Le dernier recensement donne :

Total	0.1197	0.1465	0.1527	0.1249	0.2848	0.1713

La transformation initiale s'écrit $x_{ij} = y_{ij} - t_j$ (ACP sur modèle).

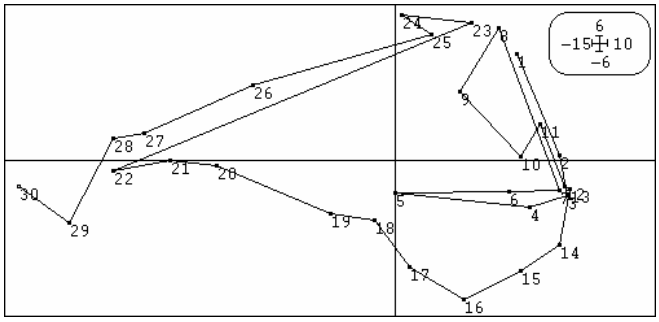
3. L'interprétation

Interpréter un tableau de données par une analyse statistique, c'est énoncer des faits concernant la réalité transcrite dans le tableau (matériel relevant d'une discipline expérimentale) en s'appuyant sur les résultats d'une procédure fondée sur des théorèmes (outil mathématique). Cette opération suppose qu'on porte l'intérêt à une réalité (sociale, économique, politique, biologique, ...) à laquelle on accède par le biais d'un enregistrement numérisé (sondage, enquête, dépouillement d'archives, expérience, ...). Cette opération suppose qu'on maîtrise la procédure et son principe qui génère pour partie des résultats imposés (conséquences de l'algorithme) et des résultats particuliers (conséquences de la structure des données).
L'interprétation n'est pas une technique mais une rencontre entre la logique d'un outil et la signification d'une réalité, entre l'unicité d'un raisonnement et la diversité infinie des observations. L'outil permet d'aller plus vite dans la lecture des données des petits tableaux, ce qui permet d'apprendre la démarche, mais n'est indispensable que pour

l'approche des grands volumes de données numériques. L'outil met en évidence l'information mais ne la crée pas.

3.1. Sur un tableau faunistique

On reprend la carte factorielle :



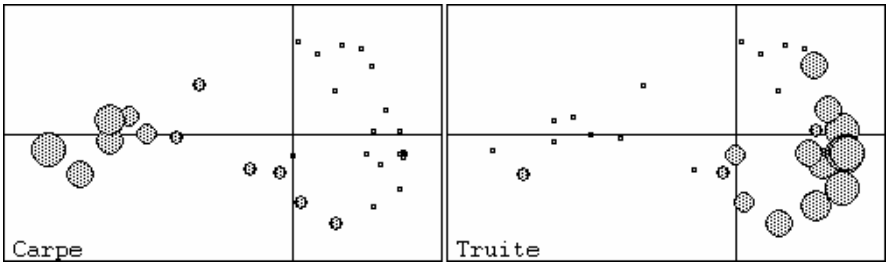
Que signifie cette figure ? Elle donne la position relative de deux stations. Deux stations proches sur ce plan sont-elles proches dans leur composition faunistique ? Pas forcément puisqu'il s'agit d'une projection. Il convient de savoir si un point est proche du plan. Deux points voisins sur la carte et voisins du plan sont proches dans l'espace. Pour ce faire, on calcule le carré de la distance d'un point à l'origine :

$$d_i^2 = \sum_{j=1}^{p} x_{ij}^2$$

On sait que la moyenne de ces quantités est l'inertie totale. On appelle **contribution à la trace** le rapport d_i^2 / I_T . C'est la participation du point à la variabilité totale. Dans ce problème, ces contributions sont équilibrées. Le dernier point est plus original. Cette distance se décompose en somme de carrés par le théorème de Pythagore et on a $d_i^2 = \sum_{k=1}^{p} l_{ik}^2$. Le rapport l_{ik}^2 / d_i^2 est le carré du cosinus de l'angle entre le point et l'axe. On dit que c'est la **contribution relative** de l'axe *k* à la représentation du point *i*.

```
----------Relative contributions------------
|Num  |Fac 1|Fac 2||Remains| Weight | Cont.|
-----------------------------------------------
|    1| 4547| 3504|| 1947  |  333   |  239 |
|    2| 8225|    7|| 1767  |  333   |  242 |
|    3| 7571|  314|| 2113  |  333   |  288 |
|    4| 6096|  720|| 3183  |  333   |  222 |
|    5|    0|  283|| 9716  |  333   |  286 |
|    6| 6386|  465|| 3147  |  333   |  149 |
|    7| 8030|  259|| 1710  |  333   |  251 |
|    8| 3204| 5143|| 1652  |  333   |  251 |
|    9| 1421| 1587|| 6991  |  333   |  223 |
|   10| 6884|    8|| 3107  |  333   |  169 |
|   11| 7930|  519|| 1550  |  333   |  195 |
|   12| 8200|  192|| 1607  |  333   |  259 |
|   13| 7199|  201|| 2598  |  333   |  312 |
|   14| 5736| 1496|| 2766  |  333   |  348 |
|   15| 3920| 3036|| 3042  |  333   |  298 |
|   16| 1067| 4363|| 4568  |  333   |  326 |
|   17|   75| 4391|| 5533  |  333   |  193 |
|   18|  263| 2302|| 7433  |  333   |  114 |
|   19| 1657| 1143|| 7198  |  333   |  184 |
|   20| 7581|    8|| 2410  |  333   |  311 |
|   21| 9205|    0||  793  |  333   |  405 |
|   22| 9222|   11||  765  |  333   |  639 |
|   23| 2160| 6928||  910  |  333   |  203 |
|   24|   16| 7364|| 2618  |  333   |  212 |
|   25|  610| 7075|| 2313  |  333   |  165 |
|   26| 6817| 1899|| 1282  |  333   |  219 |
|   27| 8771|  107|| 1120  |  333   |  529 |
|   28| 8960|   55||  984  |  333   |  658 |
|   29| 8910|  324||  764  |  333   |  886 |
|   30| 8695|   39|| 1264  |  333   | 1210 |
-----------------------------------------------
```

La **contribution relative cumulée** sur deux axes donne le carré du cosinus de l'angle entre le point et le plan des deux axes. Quand il est élevé, le point est proche du plan. Les points 5, 9, 17, 18 et 19 sont loin du plan. On dit qu'ils sont mal représentés. La plus grande partie est bien représentée. La station 5 est tout à fait anormale. On y trouve de la Truite et du Vairon, du Gardon et du Brochet. C'est en fait une retenue traversée par le Doubs qui regroupe les espèces du début de bassin et des espèces d'eaux calmes. On aurait pu l'enlever avant l'analyse. Pour comprendre le plan, on peut représenter l'abondance de chaque espèce sur ce plan :



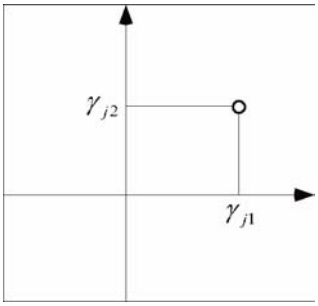
On peut alors caractériser une espèce par ses covariances avec les coordonnées :

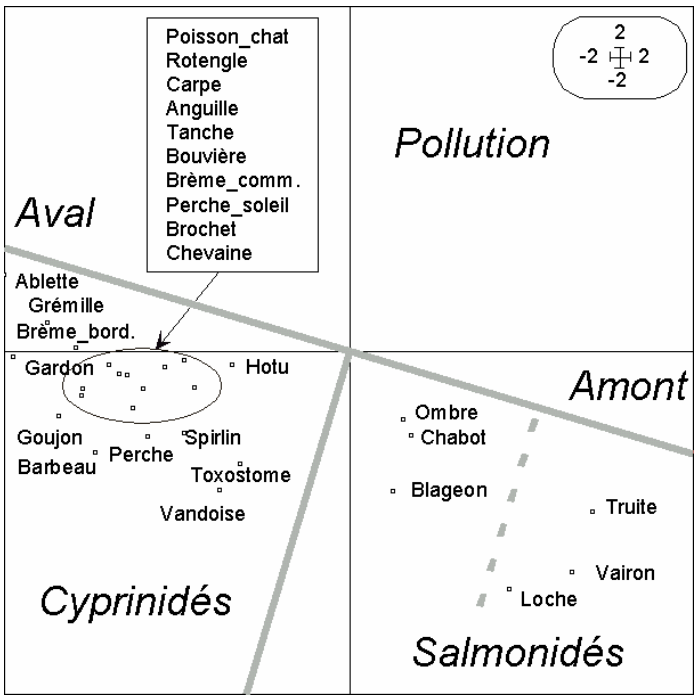
$$\mathbf{K}_1 = \frac{1}{n}\mathbf{X}^t\mathbf{X}\mathbf{U} = \mathbf{C}\mathbf{U} = \mathbf{U}\Lambda$$

On peut caractériser aussi une espèce par son coefficient dans la combinaison linéaire qui donne les coordonnées des lignes, donc par la composante associée des vecteurs de **U**.

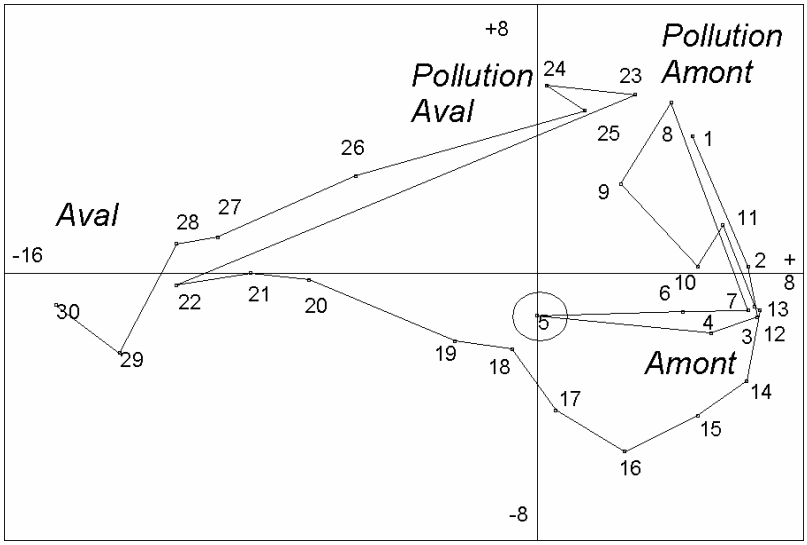
On peut enfin utiliser les covariances avec les coordonnées normalisées (divisées par leur écart-type $\sqrt{\lambda_k}$) $\mathbf{K} = \frac{1}{n}\mathbf{X}^t\mathbf{X}\mathbf{U}\Lambda^{-\frac{1}{2}} = \mathbf{C}\mathbf{U} = \mathbf{U}\Lambda^{\frac{1}{2}}$. Suivant les logiciels, on utilise l'un ou l'autre de ces systèmes très proches.

Dans la matrice **K** on trouve à la ligne (espèce) *j* et à la colonne (axe) *k* la covariance γ_{jk} entre la coordonnée normalisée $\mathbf{l}_k$ et la variable (colonne) *j* du tableau de données. On donnera plus tard une autre signification à cette quantité. On peut appeler provisoirement ces quantités **coordonnées des colonnes**. On peut représenter directement ces nombres dans une représentation cartésienne :





L'organisation des communautés le long du gradient amont-aval et l'exclusion des espèces des stations polluées sont alors claires.



3.2. Mode Q et R …

On considère un tableau comportant 10 lignes et 3 colonnes consignant l'avis fourni par 10 consommateurs sur la qualité de 3 objets à l'aide du code -1 (avis défavorable) 0 (sans opinion) et +1 (avis favorable)

Les 3 objets sont notés A,B,C et les 10 personnes interrogées sont numérotées de 1 à 10.

	A	B	C
1/	-1	-1	-1
2/	-1	0	1
3/	-1	0	0
4/	1	1	1
5/	0	0	1
6/	-1	-1	0
7/	0	0	0
8/	0	1	1
9/	-1	0	1
10/	-1	1	1

Le tableau est traité par une analyse en composantes principales centrée (diagonalisation de la matrice de covariance) qui donne:

```
MOYENNES ET VARIANCES DU FICHIER DE DEPART
A: -.50    .45
B: +.10    .49
C: +.50    .45

MATRICE DES COVARIANCES

         A       B       C
      ------- ------- -------
A     0.4500  0.2500  0.1500
B     0.2500  0.4900  0.3500
C     0.1500  0.3500  0.4500

VALEURS PROPRES

1:.9760E+00/0.70220/0.70220
2:.3112E+00/0.22392/0.92611
3:.1027E+00/0.07389/1.00000



NOMBRE DE FACTEURS CONSERVES    2

      COORDONNEES DES LIGNES        COORDONNEES DES COLONNES
         *      1         2                    1          2
         *    ------    ------          -------   -------
      1*     1.833   -0.531      A*    -0.4732     -0.4683
      2*     0.017    0.665      B*    -0.6537      0.0854
      3*     0.594    0.144      C*    -0.5699      0.2909
      4*    -1.602   -0.861
      5*    -0.462   -0.174
      6*     1.256   -0.009
      7*     0.115   -0.696
      8*    -1.123   -0.021
      9*     0.017    0.665
      10*   -0.644    0.818
```

a) Indiquer brièvement quelle information est apportée par cette analyse. Quelle question n'est manifestement pas abordée par cette opération ?

b) On décide alors de transposer le tableau et de considérer qu'il comporte 3 individus et 10 variables.

	1	2	3	4	5	6	7	8	9	10
A	-1	-1	-1	1	0	-1	0	0	-1	-1
B	-1	0	0	1	0	-1	0	1	0	1
C	-1	1	0	1	1	0	0	1	1	1

Calculer les moyennes et les variances de l'ACP centrée du nouveau tableau (Exprimer les résultats sous forme de fractions). Quand on veut diagonaliser la matrice de covariances on obtient les messages:

```
1:.2731E+01/0.87796/0.87796          VALEUR PROPRE NEGATIVE 7    -6.07772E-15
2:.3797E+00/0.12204/1.00000          VALEUR PROPRE NEGATIVE 8    -2.27279E-08
3:.6508E-06/0.00000/1.00000          VALEUR PROPRE NEGATIVE 9    -1.29848E-07
4:.7040E-07/0.00000/1.00000          VALEUR PROPRE NEGATIVE 10   -1.18605E-06
5:.1052E-13/0.00000/1.00000
6:.4450E-14/0.00000/1.0000
```

Expliquer pourquoi. Calculer la matrice de dimension minimale qu'il suffit de diagonaliser pour faire cette analyse (Exprimer les résultats sous forme de fractions). Quelles sont les dimensions de cette matrice et son rang ? Calculer ses valeurs propres en utilisant ce qui précède.

c) Les coordonnées factorielles des lignes fournies par le programme après cette diagonalisation sont :

```
     1          2
   -------  -------
A* +2.1900  +0.3044
B* -0.3879  -0.8593
C* -1.8021  +0.5549
```

Peut-on considérer que la diagonalisation de la matrice de covariance s'est normalement déroulée ? En utilisant les coordonnées des colonnes de cette analyse consignées ci-dessous, décrire brièvement l'information apportée par cette analyse.

```
     1          2
* -------   -------
1* +0.0000  +0.0000
2* -0.8052  +0.1355
3* -0.4417  -0.1647
4* -0.0000  -0.0000
5* -0.3635  +0.3002
6* -0.3635  +0.3002
7* +0.0000  +0.0000
8* -0.4417  -0.1647
9* -0.8052  +0.1355
10* -0.8834    -0.3294
```

Lequel des consommateurs est-il le plus représentatif de l'opinion collective ?

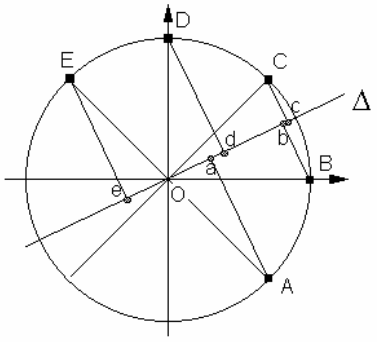
Quand les deux ACP centrées possibles pour un tableau sont utilisables on dit traditionnellement que l'une est en mode Q et l'autre en mode R.

3.3. Exercices

1) $A = (1, 0)$ $B = \left(\sqrt{3}/2, 1/2\right)$ $C = (0, 1)$ sont trois points du plan et a,b, et c sont leur projection sur un vecteur unitaire du plan (métrique canonique). Trouver le vecteur qui maximise :

$$Oa^2 + Ob^2 + Oc^2$$

2) On considère 5 points A, ..., E sur le cercle unité définis par les angles polaires $-\pi/4$, 0, $\pi/4$, $\pi/2$ et $3\pi/4$. Soit Δ une droite passant par l'origine et a, ..., e les projections orthogonales des 5 points initiaux sur Δ. Quelle est la droite Δ qui maximise la quantité $f(\Delta)=Oa^2+Ob^2+Oc^2+Od^2+Oe^2$? Quel est le maximum atteint ? Quelle est la droite Δ qui minimise la quantité $f(\Delta)$? Quel est le minimum atteint ?



3) Calculer et dessiner le premier axe principal du nuage de 8 points de $\mathbb{R}^2$ défini par :

$$\begin{bmatrix} 2 & 5 & 8 & 3 & -2 & -5 & -8 & -3 \\ -4 & 0 & 4 & 4 & 4 & 0 & -4 & -4 \end{bmatrix}$$

4) On considère les 9 points de $\mathbb{R}^2$ de coordonnées :

	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9
	−4	−3	−2	−1	0	1	2	3	4
	−3	−4	−1	−2	0	4	3	2	1

Soit O l'origine du plan, **u** un vecteur unitaire, m_i la projection orthogonale de M_i sur **u**, h(**u**) la somme des carrés des longueurs de Om_i. Le vecteur qui maximise h(**u**) est-il porté par la première bissectrice ? La valeur maximale atteinte est-elle 118 ?

5) Soient les deux variables à 6 valeurs $x = (0,1,0,1,1,0)$ et $y = (0,1,1,0,1,0)$. Calculer moyennes, variances et covariances. Dessiner le nuage centré et placer les deux droites de régression et les deux axes principaux.

6) Soient deux variables **x** et **y** mesurées sur *n* individus. Existe-t-il deux réels *a* et *b* tels que $a^2+b^2=1$ qui maximisent la variance de la variable $a\mathbf{x}+\boldsymbol{b}\mathbf{y}$?

7) On considère les 10 points M_i de coordonnées :

i	1	2	3	4	5	6	7	8	9	10
x_i	–2	–1	0	1	2	–2	–1	0	1	2
y_i	1	1	1	1	1	–1	–1	–1	–1	–1

Calculer $m(\mathbf{x})$, $m(\mathbf{y})$, $v(\mathbf{x})$, $v(\mathbf{y})$ et $c(\mathbf{x},\mathbf{y})$. Calculer les deux droites de régression et l'axe principal. Tracer le nuage et ses droites caractéristiques.

On considère les 10 points P_i de coordonnées :

$$z_i = x_i \cos\alpha + y_i \sin\alpha$$
$$t_i = x_i \sin\alpha - y_i \cos\alpha$$

Calculer $m(\mathbf{z})$, $m(\mathbf{t})$, $v(\mathbf{z})$, $v(\mathbf{t})$ et $c(\mathbf{z},\mathbf{t})$ pour α quelconque. Tracer le nouveaux nuages et ses trois droites caractéristiques pour $\alpha = 60°$. Montrer que, en général, l'axe principal est conservé par rotation du nuage.

8) Considérons le tableau X **1** comportant 6 lignes et 3 colonnes.

On y trouve l'opinion des chefs d'entreprises industrielles sur l'état de leur carnet de commandes. L'INSEE interroge chaque mois un échantillon de 2000 chefs d'entreprises et donne la fréquence des réponses exprimées en trois catégories : a-bien garni, b-normalement garni, c-peu garni. Par exemple en janvier 1968, 44% des

1 Grais B., 1979, *Statistiques descriptives*, 2° édition, Bordas, Paris,275 pp.

personnes interrogées pensent que le carnet de commande de leur entreprise est peu garni.

1968	a	b	c
janvier	.09	.47	.44
mars	.09	.48	.43
mai	.18	.49	.33
juillet	.21	.49	.30
septembre	.28	.50	.22
novembre	.36	.48	.16

a) Construire la représentation triangulaire associée au tableau X.

b) Calculer les moyennes et variances du tableau considéré comme présentant 6 individus et 3 variables. Choisir deux des trois variables pour représenter au mieux l'évolution de l'opinion des chefs d'entreprises. Quel pourcentage d'inertie du nuage est-il alors exprimé ?

c) Calculer la matrice des covariances du tableau **X**. Noter $\mathbf{X}_0$ le tableau centré, **e** le vecteur de $\mathbb{R}^3$ de composantes toutes égales à 1. Quel est le rang de la matrice **C** ?

d) Sachant que 0.00010372 est valeur propre de **C**, diagonaliser **C**. Représenter, sur le plan des deux premiers axes principaux, la projection de la base canonique de $\mathbb{R}^3$. Sur le même graphique, représenter la projection des points dont les coordonnées sont les lignes de $\mathbf{X}_0$.

e) Sur la figure précédente représenter les lignes de **X** comme centres de gravité des sommets des vecteurs de la base canonique pondérés par les distributions de fréquence associées à chaque date. Ajouter sur le même principe le centre de gravité du nuage des 6 points. Comparer avec la figure obtenue en a) et expliquer le résultat. Cette pratique a été employée pour la première fois dans **1**.

9) On considère le tableau :

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Faire l'ACP centrée de ce tableau en donnant les moyennes, les variances, la matrice des covariances, les valeurs propres, la répartition de l'inertie entre les axes, les coordonnées factorielles et les cartes factorielles.

Indiquer comment cette analyse rend compte de la structure du tableau. On donne la matrice des axes principaux :

---

1 Gabriel K.R., The biplot graphic display of matrices whith application to principal component analysis, *Biometrika*, 58, 453-467, 1971.

$$\mathbf{U} = \begin{bmatrix} 1/2 & .3825 & -1/2 & .1426 & -.5773 \\ 1/2 & -.2740 & 1/2 & .6626 & 0 \\ 0 & -.7651 & 1 & -.2852 & .5773 \\ -1/2 & -.2740 & -1/2 & .6626 & 0 \\ -1/2 & .3825 & 1/2 & .1426 & -.5773 \end{bmatrix}$$

10) Valeur propre double **1**

a) Diagonaliser les matrices du type :

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

b) On dispose du classement de 11 individus pour trois matières du certificat d'études (nouveau programme) :

Maths	1	2	3	4	5	6	7	8	9	10	11
Musique	6	1	4	5	3	2	9	7	8	10	11
Sanscrit	2	6	5	3	4	1	8	9	7	10	11

Effectuer l'ACP normée du tableau 11 individus - 3 variables. La base des axes principaux est-elle unique ? Tracer le graphe canonique.

c) Un auditeur libre a eu des notes qui lui auraient donné les rangs 8 en maths, 2 en musique et 1 en sanscrit. Le situer par rapport à l'ensemble des individus.

d) Le lendemain, on dispose des classements des mêmes candidats en :

Natation	3	4	5	6	7	8	9	10	11	1	2

Positionner le point natation sur la première composante principale. Aurait-on obtenu le même résultat en analysant un tableau 11 individus - 4 variables ?

1 Cehessat, R. (1976) Exercices commentés de Statistique et Informatique Appliquée. Dunod, 418 p.

Annexe : Script pour refaire la figure du paragraphe 2.1

```
covv <- matrix(c(1,-1.6,-1.6,4),2,2)
par(mfrow=c(2,2))
z1 <- mvrnorm(20,mu=rep(0,2),Sigma=covv)     #ou rmvnorm(20,,covv)
x <- z1[,1]
y <- z1[,2]
mx <- sum(x)/20
my <- sum(y)/20
cxy <- sum((x-mx)*(y-my))/20
vx <- sum((x-mx)*(x-mx))/20
vy <- sum((y-my)*(y-my))/20
a <- cxy/vx
b <- my-a*mx

plot (x,y, xlim=c(-3.5,2.5),ylim=c(-3,3))

plot (x,y, xlim=c(-3.5,2.5),ylim=c(-3,3))
abline(b,a)
for (i in 1:20) {
   segments(x[i],y[i],x[i],a*x[i]+b)
}

a <- cxy/vy
b <-  mx-a*my
alpha <- 1/a
beta <- -b/a

plot (x,y, xlim=c(-3.5,2.5),ylim=c(-3,3))
abline(beta,alpha)
for (i in 1:20) {
   segments(x[i],y[i],a*y[i]+b,y[i])
}

u1 <- eigen(var(z1))$vectors[1,]
plot (x,y, xlim=c(-3.5,2.5),ylim=c(-3,3))
abline(0,u1[2]/u1[1])
for (i in 1:20) {
   ps <- x[i]*u1[1]+y[i]*u1[2]
   segments(x[i],y[i],ps*u1[1],ps*u1[2])
}
```