Renaud Sirdey et Franck Turmel-josek

Analyse en composantes principales

Éditions des Nik's News
www.niksnews.com/editions/

1998

L'œuvre appartient son auteur.
L'auteur est seul responsable du contenu de son œuvre.
L'auteur autorise les Éditions des Nik's News à :

– ajouter à son œuvre des informations les concernant ;

– diffuser gratuitement son œuvre ;

– choisir le ou les formats de diffusion de son œuvre.

Les Éditions des Nik's News s'engagent à ne plus publier une œuvre si son auteur le désire.

Introduction

En application du cours, nous avons réalisé l'étude d'une population d'élèves (page 7) et d'une population de crabes (page 17) à l'aide des outils statistiques élémentaires et de l'analyse en composantes principales. Méthode que nous avons programmée en utilisant le langage *matlab*.

Table des matières

Chapitre 1

Etudes des données notes

1.1 Les données

Il s'agit d'un tableau regroupant les notes obtenues par 9 élèves en mathématiques, sciences, français, latin et dessin (voir tableau 1.1).

1.2 Analyse mono-dimensionnelle

1.2.1 Statistiques élémentaires

Voir tableau 1.2.

1.2.2 « Boxplot »

Le *boxplot* des données notes (voir figure 1.1) permet de mettre en évidence un *outlier* qui correspond au 18 qu'a obtenu Evelyne en dessin.

1.3 Analyse multi-dimentionnelle élémentaire

1.3.1 Estimation de la matrice de covariance

$$v = \begin{pmatrix} 12.8125 & 11.1563 & 2.9896 & 5.4271 & 0.1250 \\ 11.1563 & 10.0625 & 4.6354 & 6.1667 & 0.0625 \\ 2.9896 & 4.6354 & 13.5694 & 10.4549 & 0.4375 \\ 5.4271 & 6.1667 & 10.4549 & 8.9028 & 0.7500 \\ 0.1250 & 0.0625 & 0.4375 & 0.7500 & 9.7500 \end{pmatrix}$$

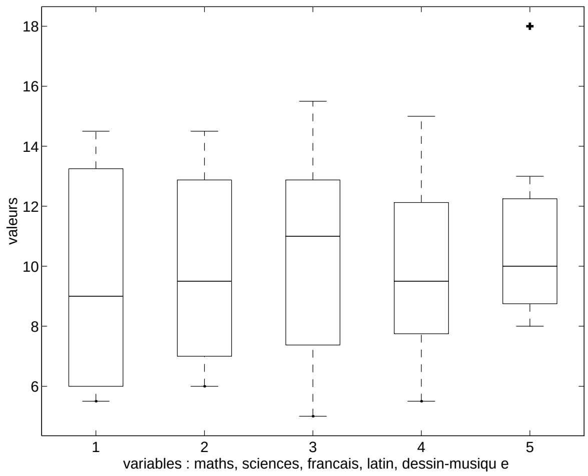
	maths	sciences	français	latin	dessin
jean	6.00	6.00	5.00	5.50	8.00
aline	8.00	8.00	8.00	8.00	9.00
annie	6.00	7.00	11.00	9.50	11.00
monique	14.50	14.50	15.50	15.00	8.00
didier	14.00	14.00	12.00	12.50	10.00
andré	11.00	10.00	5.50	7.00	13.00
pierre	5.50	7.00	14.00	11.50	10.00
brigitte	13.00	12.50	8.50	9.50	12.00
evelyne	9.00	9.50	12.50	12.00	18.00

TAB. 1.1 – Les données notes.

	maths	sciences	français	latin	dessin
moyenne	9.66	9.83	10.22	10.05	11.00
variance	12.81	10.06	13.56	8.90	9.75
écart-type	3.57	3.17	3.68	2.98	3.12
médiane	9.00	9.50	11.00	9.50	10.00
iqr	7.25	5.87	5.50	4.37	3.50
maximum	14.50	14.50	15.50	15.00	18.00
minimum	5.50	6.00	5.00	5.50	8.00

TAB. 1.2 – Statistiques élémentaires des données notes.

FIG. **1.1** – *Boxplot* des données notes.



1.3.2 Estimation de la matrice de corrélation

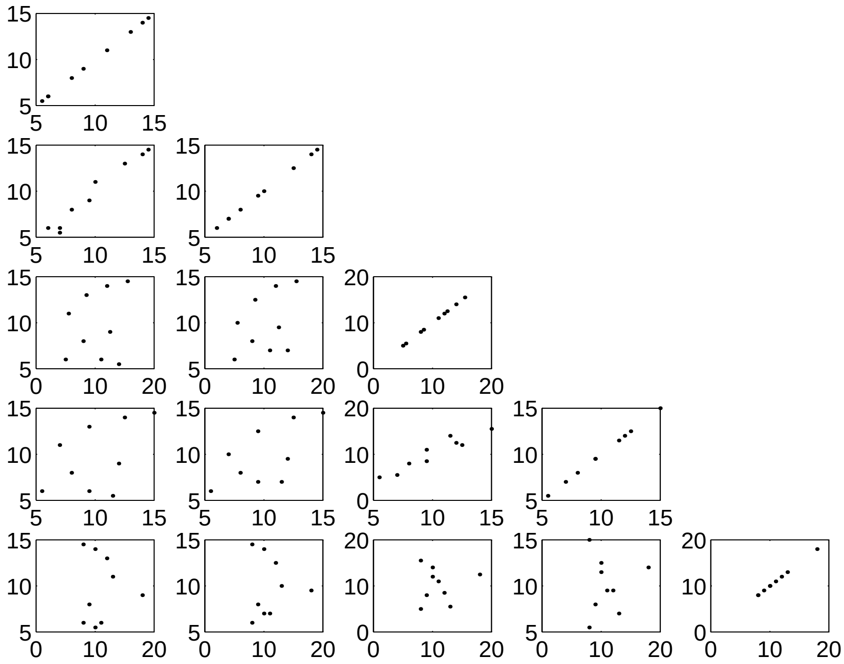
$$r = \begin{pmatrix} 1.0000 & 0.9825 & 0.2267 & 0.5081 & 0.0112 \\ 0.9825 & 1.0000 & 0.3967 & 0.6515 & 0.0063 \\ 0.2267 & 0.3967 & 1.0000 & 0.9512 & 0.0380 \\ 0.5081 & 0.6515 & 0.9512 & 1.0000 & 0.0805 \\ 0.0112 & 0.0063 & 0.0380 & 0.0805 & 1.0000 \end{pmatrix}$$

On remarque que les coefficients de la matrice sont tous positifs (c'est-à-dire que les variables évoluent à peu près de la même façon autour de leurs moyennes) et qu'il semble y avoir des dépendances linéaires entre les mathématiques et les sciences ainsi qu'entre le français et le latin.

1.3.3 « Multiplot »

Le *multiplot* des données notes (voir figure 1.2) illustre les dépendances linéaires identifiées grâce à l'estimation de la matrice de corrélation. Cependant, il ne permet pas vraiment de mettre en évidence des regroupements d'élèves (sauf peut être sur le graphique de dipsersion mathématique/dessin ainsi que sur celui de sciences/dessin).

FIG. **1.2** – *Multiplot* des données notes.



1.4 Analyse en composantes principales

1.4.1 Recherche des principaux axes d'inertie

Calcul des valeurs propres de la matrice $x^T D_p x M$, sachant que D_p est une matrice diagonale (9×9) telle que $D_p = \frac{1}{9} I_{(9\times 9)}$ et M est une matrice diagonale (5×5) telle que $M = I_{(5\times 5)}$.

Valeurs propres rangées par ordre décroissant

$$\lambda_1 = 28.2532,\ \lambda_2 = 12.0747,\ \lambda_3 = 8.6157,\ \lambda_4 = 0.0217,\ \lambda_5 = 0.0099$$

Matrice des vecteurs propres rangées dans l'ordre des valeurs propres

$$u = \begin{pmatrix} -0.5145 & -0.5669 & 0.0513 & -0.2887 & -0.5725 \\ -0.5070 & -0.3720 & 0.0145 & 0.5531 & 0.5464 \\ -0.4924 & 0.6504 & -0.1081 & 0.3937 & -0.4098 \\ -0.4846 & 0.3232 & -0.0225 & -0.6742 & 0.4534 \\ -0.0306 & 0.1129 & 0.9925 & 0.0344 & -0.0127 \end{pmatrix}$$

1.4.2 Energie expliquée par les axes

Le pourcentage d'énergie expliquée par l'axe u_i est égale à

$$e_i = 100 \frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j}$$

Les pourcentages d'énergie expliquée par les axes u_1, u_2, u_3, u_4 et u_5 sont donc respectivement 57.6888%, 24.6547%, 17.5920%, 0.0444%, 0.0202%. Le pourcentage d'énergie expliquée par les trois premiers axes est égal à 99.9355%, le nuage peut par conséquent être correctement représenté dans un espace de dimension 3.

1.4.3 Calcul des composantes principales selon les individus

La $i^{\text{ème}}$ composante principale (c_i) est égale à xMu_i. Ceci peut être exprimé de façon matricielle

$$C = xMu$$

	1	2	3	4	5
jean	8.7009	-1.7027	-2.5539	-0.1495	0.1173
aline	3.9386	-0.7085	-1.8105	-0.0907	-0.0435
annie	3.2094	3.4591	-0.3007	0.1725	-0.0193
monique	-9.7557	0.2157	-3.3437	-0.1735	-0.1004
didier	-6.3714	-2.1733	-0.9571	0.0707	0.1880
andré	2.9740	-4.6509	2.6349	-0.0232	-0.1481
pierre	1.0510	6.2272	-1.6881	0.1153	-0.0428
brigitte	-1.9805	-4.0686	1.4007	0.2432	-0.0104
evelyne	-1.7662	3.4021	6.6182	-0.1649	0.0592

Tab. 1.3 – Composantes principales associées aux élèves.

où C est la matrice des composantes principales par rapport aux individus rangés par colonne (voir tableau 1.3).

Interprétation

La figure 1.3 représente les élèves dans le plan $(1,2)$. Les élèves situés sur la gauche de la figure sont relativement « bon » partout alors que ceux situés sur la droite sont relativement « mauvais » partout. Les élèves situés vers le haut de la figure sont meilleurs dans les matières littéraires alors que ceux situés vers le bas sont meilleurs dans les matières scientifiques. L'élève de « rêve » se situe donc à l'extrême gauche et à mi-hauteur (ce qui est par exemple le cas de Monique), alors que l'élève de « cauchemar » se situe à l'extrême droite et aussi à mi-hauteur (ce qui est à peu prés le cas de Jean).

La figure 1.4 représente les élèves dans le plan $(1,3)$. L'interprétation de l'axe horizontal reste la même. Les éleves situés vers le haut de la figure sont « bon » en dessin alors que ceux situés vers le bas sont « mauvais » dans cette matière.

Conclusion

L'axe 1 (première composante principale) correspond à la moyenne générale d'un élève dans les matières principales. L'axe 2 représente l'opposition scientifique/littéraire. L'axe 3 révèle le niveau d'un élève en dessin.

FIG. **1.3** – Représentation des élèves dans le plan $(1, 2)$.

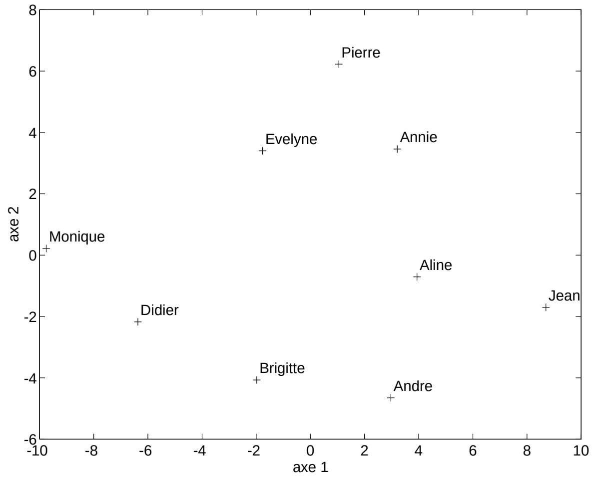
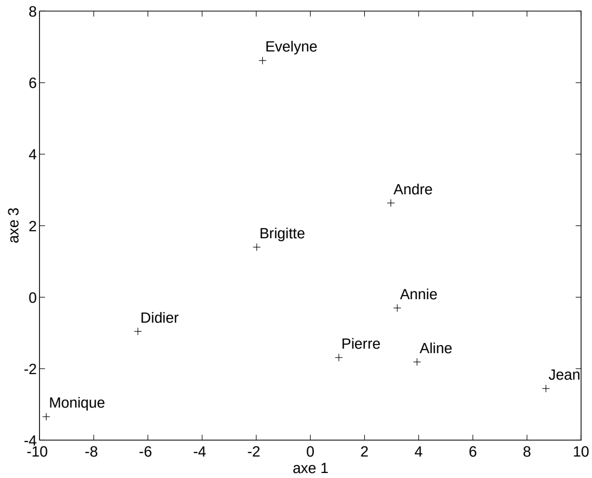


FIG. **1.4** – Représentation des élèves dans le plan $(1, 3)$.



	1	2	3	4	5
maths	-0.7641	-0.6211	0.0409	-0.0143	-0.0182
sciences	-0.7529	-0.4075	0.0115	0.0273	0.0174
français	-0.7311	0.7124	-0.0861	0.0195	-0.0130
latin	-0.7197	0.3541	-0.0180	-0.0333	0.0144
dessin	-0.0455	0.1237	0.7908	0.0017	-0.0004

TAB. 1.4 – Composantes principales normalisées associées aux matières.

1.4.4 Calcul des composantes principales selon les variables

Les vecteurs d^i, composantes principales par rapport aux variables, sont formés des coordonnées de toutes les variables pour un mêmes axe u_i. Ces vecteurs vérifient la relation

$$d^i = \sqrt{\lambda_i} u_i$$

Finalement on préfère s'intéresser aux vecteurs normalisés c'est-à-dire tels que

$$\bar{d}^i = \frac{d^i}{\|x^i\|}, \text{ avec } \|x^i\|^2 = \sum_{j=1}^{p} (x_j^i)^2$$

Le tableau 1.4 donne les valeurs des composantes principales normalisées selon les variables. Les figures 1.5 et 1.6 représentent les matières dans les plans $(1,2)$ et $(1,3)$ des composantes principales normalisées selon les variables.

FIG. **1.5** – Représentation des matières dans le plan $(1,2)$.

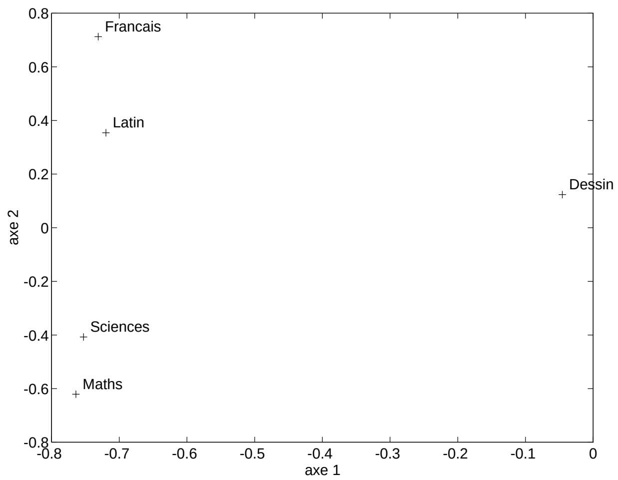
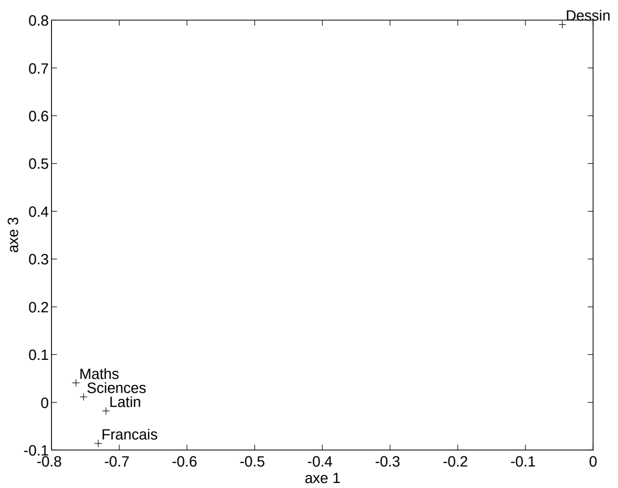


FIG. **1.6** – Représentation des matières dans le plan $(1,3)$.



Chapitre 2

Etudes des données crabes

2.1 Les données

Il s'agit de données portant sur la morphologie de 200 crabes leptograpsus provenant d'Australie (Campbell et Mahon 1974).

Les variables mesurées sur chacun des crabes sont les suivantes (voir figure 2.1) :

- FL : frontal lip of carapace (mm),
- RW : rear width of carapace (mm),
- CL : length along the midline of carapace (mm),
- CW : maximum width of carapace (mm),
- BD : body depth (mm).

2.2 Analyse mono-dimensionnelle

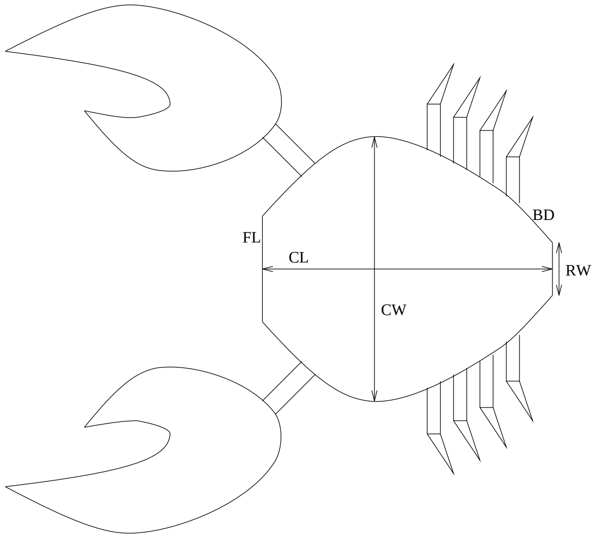
2.2.1 Statistiques élémentaires

Voir tableau 2.1.

2.2.2 « Boxplot »

Le *boxplot* des données crabes (voir figure 2.2) permet de mettre en évidence un *outlier* pour la variable RW.

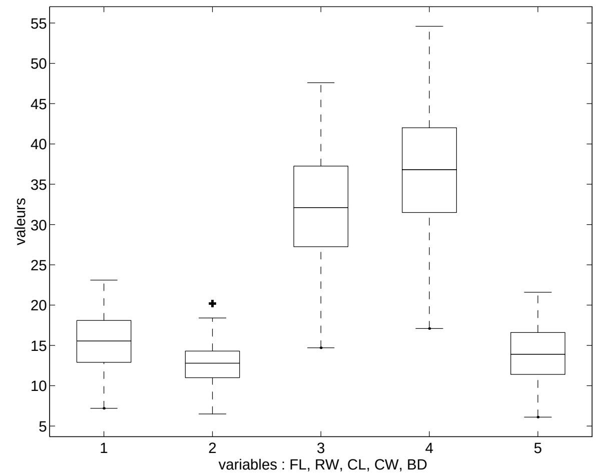
Fig. 2.1 – Mesures effectuées sur les crabes.



	FL	RW	CL	CW	BD
moyenne	15.58	12.73	32.10	36.41	14.03
variance	12.21	6.62	50.67	61.96	11.72
écart-type	3.49	2.57	7.11	7.87	3.42
médiane	15.55	12.80	32.10	36.80	13.90
iqr	5.20	3.30	10.00	10.50	5.20
maximum	23.10	20.20	47.60	54.60	21.60
minimum	7.20	6.50	14.70	17.10	6.10

Tab. 2.1 – Statistiques élémentaires des données crabes.

FIG. **2.2** – *Boxplot* des données crabes.



2.3 Analyse multi-dimentionnelle élémentaire

2.3.1 Estimation de la matrice de covariance

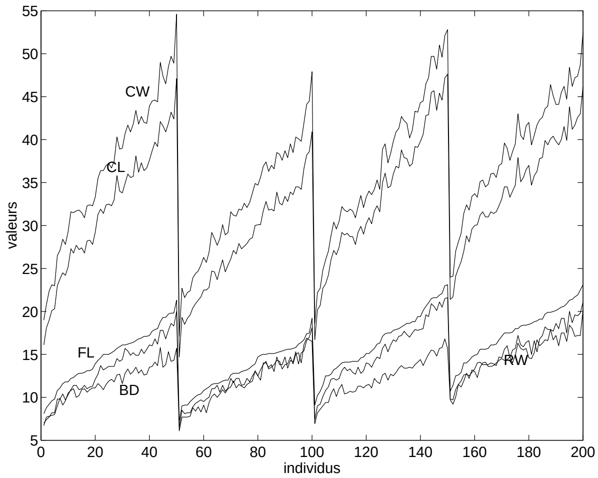
$$v = \begin{pmatrix} 12.2173 & 8.1580 & 24.3567 & 26.5508 & 11.8226 \\ 8.1580 & 6.6221 & 16.3547 & 18.2396 & 7.8367 \\ 24.3567 & 16.3547 & 50.6799 & 55.7614 & 23.9714 \\ 26.5508 & 18.2396 & 55.7614 & 61.9677 & 26.0919 \\ 11.8226 & 7.8367 & 23.9714 & 26.0919 & 11.7291 \end{pmatrix}$$

2.3.2 Estimation de la matrice de corrélation

$$r = \begin{pmatrix} 1.0000 & 0.9070 & 0.9788 & 0.9650 & 0.9876 \\ 0.9070 & 1.0000 & 0.8927 & 0.9004 & 0.8892 \\ 0.9788 & 0.8927 & 1.0000 & 0.9950 & 0.9832 \\ 0.9650 & 0.9004 & 0.9950 & 1.0000 & 0.9678 \\ 0.9876 & 0.8892 & 0.9832 & 0.9678 & 1.0000 \end{pmatrix}$$

On remarque que les coefficients de la matrice sont tous positifs (c'est-à-dire que les variables évoluent toutes de la même façon autour de leurs moyenne, voir figure 2.3) et qu'il semble y avoir des dépendances linéaires entre toutes les variables [1].

FIG. **2.3** – Les données crabes. Cette figure permet de mettre en évidence, d'une part un effet taille et d'autre part l'évolution identique des variables autour de leurs moyennes (ce qui explique que les coefficients de corrélation soient tous proches de 1).



2.3.3 « Multiplot »

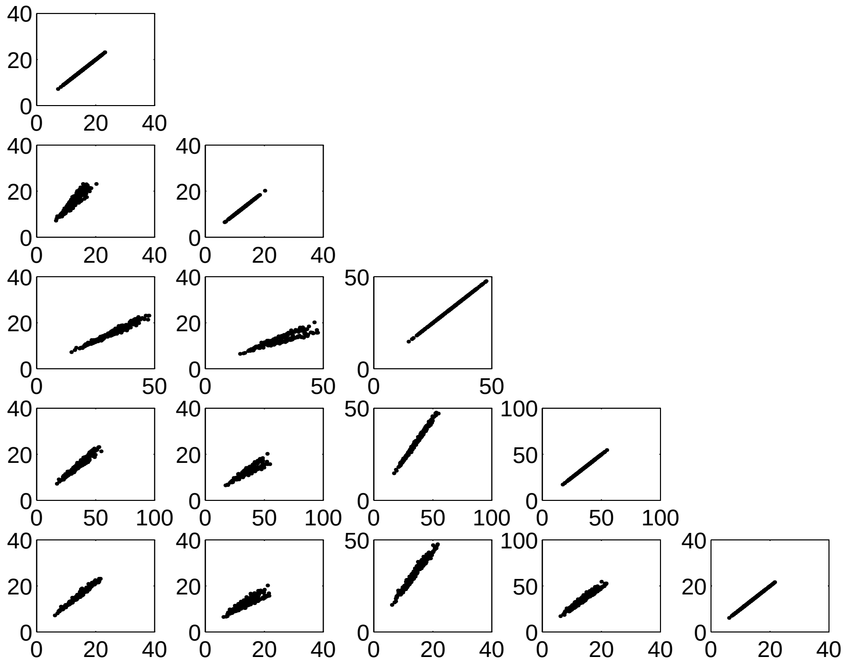
Le *multiplot* des données crabes (voir figure 2.4) illustre les dépendances linéaires identifiées grâce à l'estimation de la matrice de corrélation. Cependant, il ne permet pas de mettre en évidence des regroupements de crabes (quoique certains graphiques de dipsersion de forme « conique » laisse présager l'existence d'au moins deux groupes).

2.3.4 Effet « taille »

Comme dans le cas des papillons, il y a un effet taille. En divisant toutes les variables par FL on obtient le *multiplot* de la figure 2.5 qui permet de mettre clairement en évidence deux regroupements de crabes. Cependant, le graphique de dispersion des variables CW et RW (voir figure 2.6) semble faire apparaitre quatre regroupements. Afin de mettre ces derniers en évidence on va

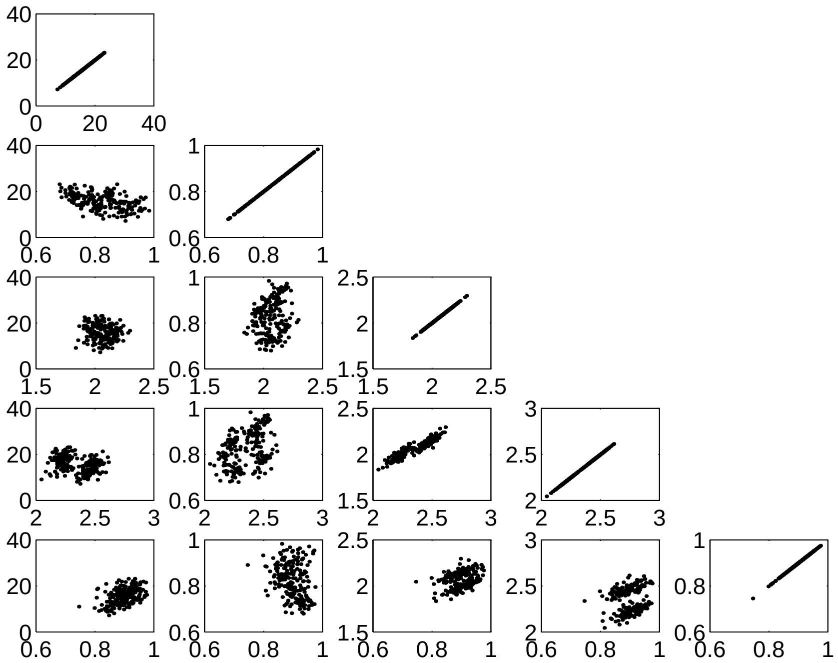
[1]. Puisque les coefficients de corrélation sont tous relativement proches de 1.

FIG. **2.4** – *Multiplot* des données crabes.



s'intéresser à l'histogramme bidimensionnel (calculé à l'aide la routine page 34) des variables CW et RW, cette fonction permet d'avoir une idée de la densité de crabes présent dans une région du plan. Un tracé de lignes de niveaux sur l'histogramme permet de faire apparaitre les quatre regroupements recherchés.

FIG. **2.5** – *Multiplot* des données crabes aprés prise en compte de l'effet taille.

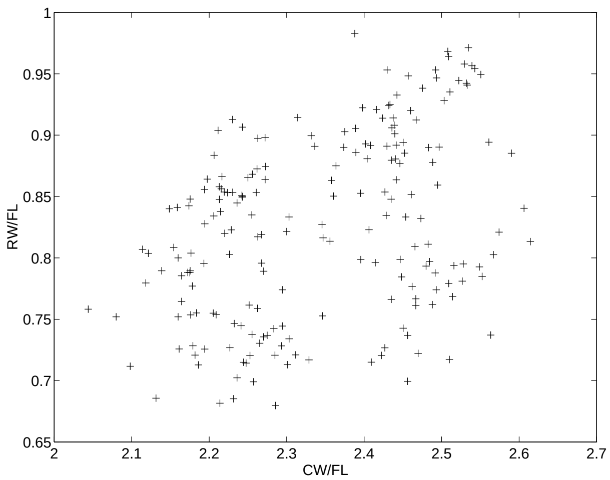


2.4 Analyse en composantes principales

2.4.1 Recherche des principaux axes d'inertie

Calcul des valeurs propres de la matrice $x^T D_p x M$, sachant que D_p est une matrice diagonale (200×200) telle que $D_p = \frac{1}{200} I_{(200\times200)}$ et M est une matrice diagonale (5×5) telle que $M = I_{(5\times5)}$.

FIG. 2.6 – Graphique de dispersion des variables CW et RW, après la prise en compte de l'effet taille. On pressent l'existence de quatre regroupements de crabes.



Valeurs propres rangées par ordre décroissant

$$\lambda_1 = 140.0022,\ \lambda_2 = 1.2904,\ \lambda_3 = 0.9953,\ \lambda_4 = 0.1346,\ \lambda_5 = 0.0775$$

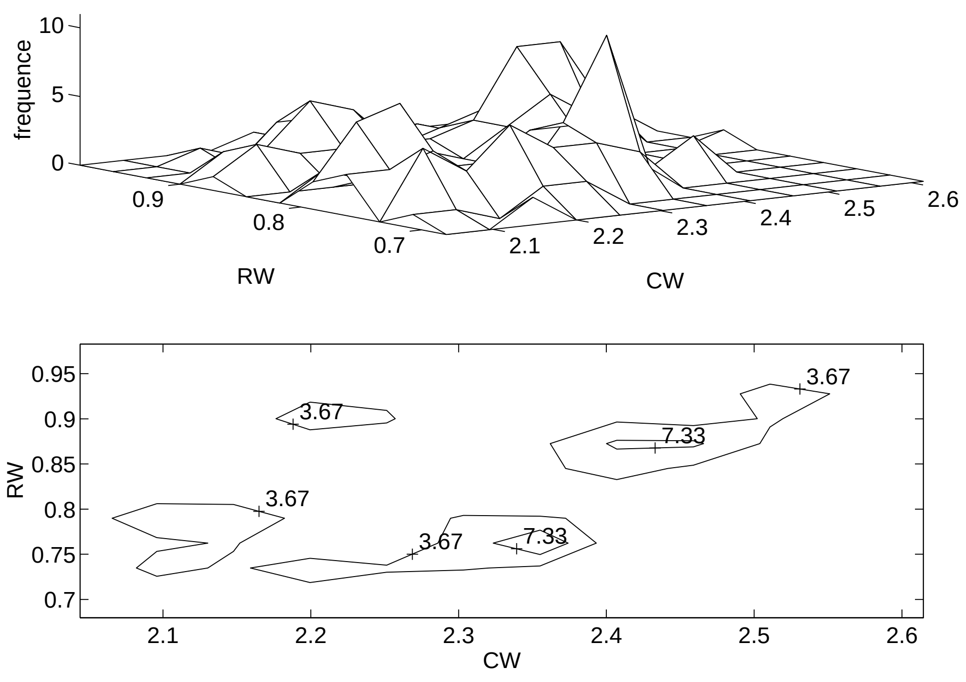
Matrice des vecteurs propres rangées dans l'ordre des valeurs propres

$$u = \begin{pmatrix} 0.2890 & 0.3233 & 0.5072 & 0.7343 & 0.1249 \\ 0.1973 & 0.8647 & -0.4141 & -0.1483 & -0.1409 \\ 0.5994 & -0.1982 & 0.1753 & -0.1436 & -0.7417 \\ 0.6617 & -0.2880 & -0.4914 & 0.1256 & 0.4712 \\ 0.2837 & 0.1598 & 0.5469 & -0.6344 & 0.4387 \end{pmatrix}$$

2.4.2 Energie expliquée par les axes

Les pourcentages d'énergie expliquée par les axes u_1, u_2, u_3, u_4 et u_5 sont respectivement 98.2472%, 0.9055%, 0.6984%, 0.0945%, 0.0544%. Le nuage peut quasiment être représenté dans un espace de dimension 1. Le fait de considérer les trois premiers axes permet de prendre en compte 99.8511% de l'énergie du nuage.

FIG. **2.7** – Mise en évidence des quatres regroupements de crabes à l'aide du calcul de l'histogramme bidimensionnel de CW et RW et d'un tracé de lignes de niveaux sur ce dernier.



2.4.3 Composantes principales selon les individus

Dans un premier temps, on constate que le premier axe explique approximativement 98% de l'énergie, la représentation des crabes dans l'espace de dimensions 1 (voir figure 2.8) peut donc s'avérer intéressante. En étudiant l'histogramme correspondant à cette représentation mono-dimensionnelle (qui permet d'avoir une idée de la densité de crabes présents dans une région de cet espace) on aperçoit quatres pics. Cependant, étant donné l'instabilité de l'histogramme on se gardera de conclure trop vite quand à la mise en évidence de quatre classes de crabes.

Dans un second temps, on s'intéresse à la représentation du nuages dans l'espace de dimension 3 correspondant aux trois premières composantes principales (soit quasiment 100% d'énergie expliquée, voir figure 2.9). Ceci nous permet de mettre en évidence au moins deux regroupements de crabes et de pressentir l'existence de quatre classes à l'aide d'une rotation d'axe.

Enfin, on considère le nuage dans les plans correspondants aux composantes principales prises deux à deux. Le plan $(1, 2)$ (voir figure 2.10) permet à peu près de mettre en évidence deux regroupements de crabes. Le plan $(1, 3)$ ne permet pas de conclure grand chose. Par contre, une attention particulière va être portée à la représentation du nuage dans le plan $(2, 3)$ (voir figure 2.12) qui permet quasiment de mettre en évidence les quatres regroupements que nous recherchons. Comme dans le cas de l'effet taille (voir page 20) nous allons considérer l'histogramme bidimensionnel des composantes principales 2 et 3 et mettre les classes en évidence à l'aide d'un tracé de ligne de niveaux (voir figure 2.13).

Conclusion

L'analyse en composantes principales nous a permis de mettre en évidence quatres classes de crabes. Ce qui avait aussi été observé avec des méthodes exploratoires élémentaires et en tenant compte de l'effet taille.

2.4.4 Composantes principales selon les variables

Le tableau 2.2 donne les valeurs des composantes principales selon les variables. Les figures 2.14 et 2.15 représentent les mesures dans les plans $(1, 2)$ et $(1, 3)$ des composantes principales normalisées selon les variables.

FIG. **2.8** – Représentation des crabes dans l'espace de dimension 1 correspondant à la première composante principale (soit 98.2472% d'énergie expliquée). La figure du bas donne une idée du nombre de crabes qui possèdent les caractéristiques mises en évidence à l'aide de la première composante principale.

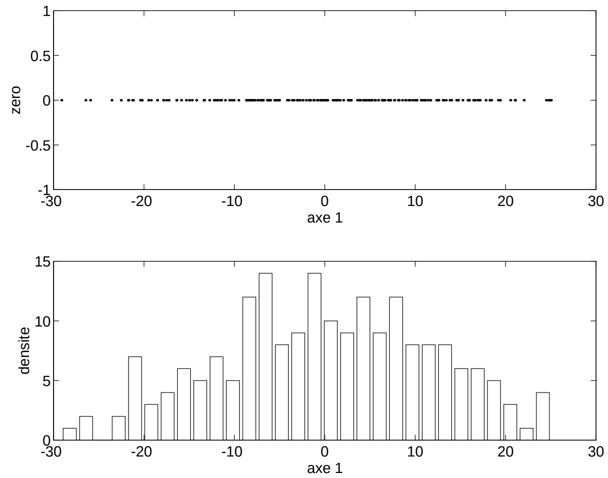


FIG. **2.9** – Représentation des crabes dans l'espace de dimension 3 correspondant aux 3 premières composantes principales (soit 99.8511% d'énergie expliquée). La figure du bas permet de pressentir l'existence de quatre regroupements de crabes.

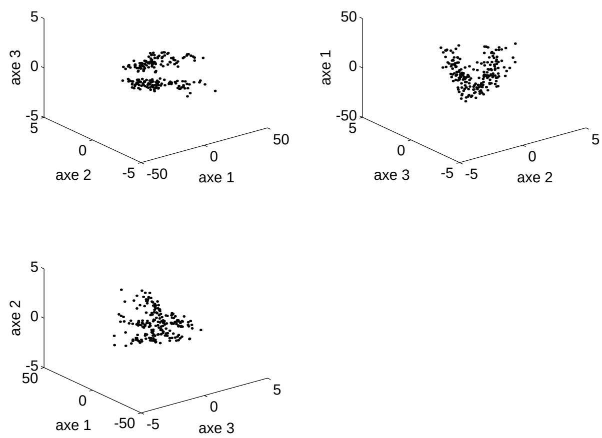


FIG. **2.10** – Représentation des crabes dans le plan $(1,2)$.

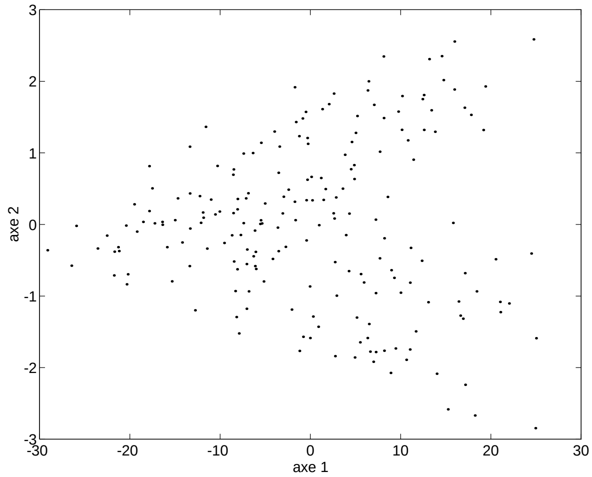


FIG. **2.11** – Représentation des crabes dans le plan $(1,3)$.

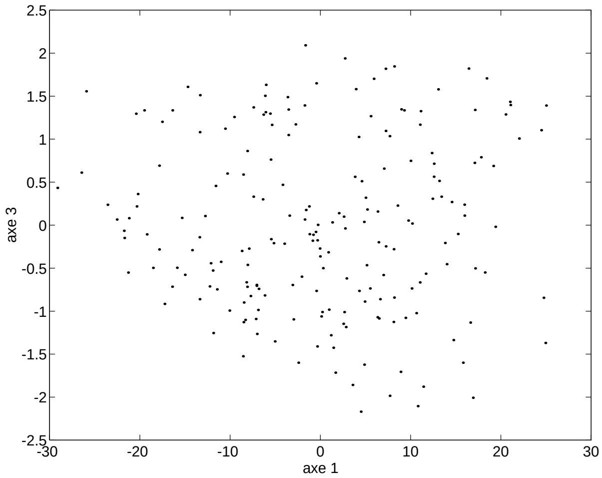
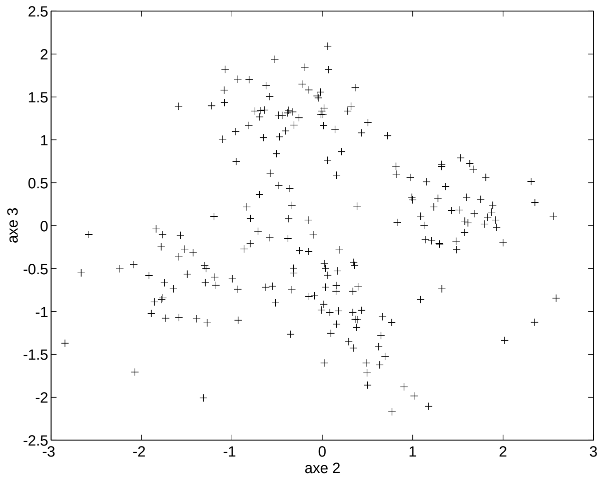


FIG. **2.12** – Représentation des crabes dans le plan $(2, 3)$, on pressent l'existence de quatres regroupements.



	1	2	3	4	5
FL	0.9782	0.1427	0.0711	0.0342	0.0102
RW	0.6678	0.3817	-0.0580	-0.0069	-0.0115
CL	2.0291	-0.0875	0.0246	-0.0067	-0.0603
CW	2.2398	-0.1271	-0.0689	0.0059	0.0383
BD	0.9605	0.0706	0.0766	-0.0296	0.0357

TAB. 2.2 – Composantes principales normalisées associées aux mesures effectuées sur les crabes.

FIG. **2.13** – Mise en evidence des quatres regroupements de crabes à l'aide du calcul de l'histogramme bidimensionnel des $2^{ème}$ et $3^{ème}$ composantes principales et d'un tracé de lignes de niveaux sur ce dernier.

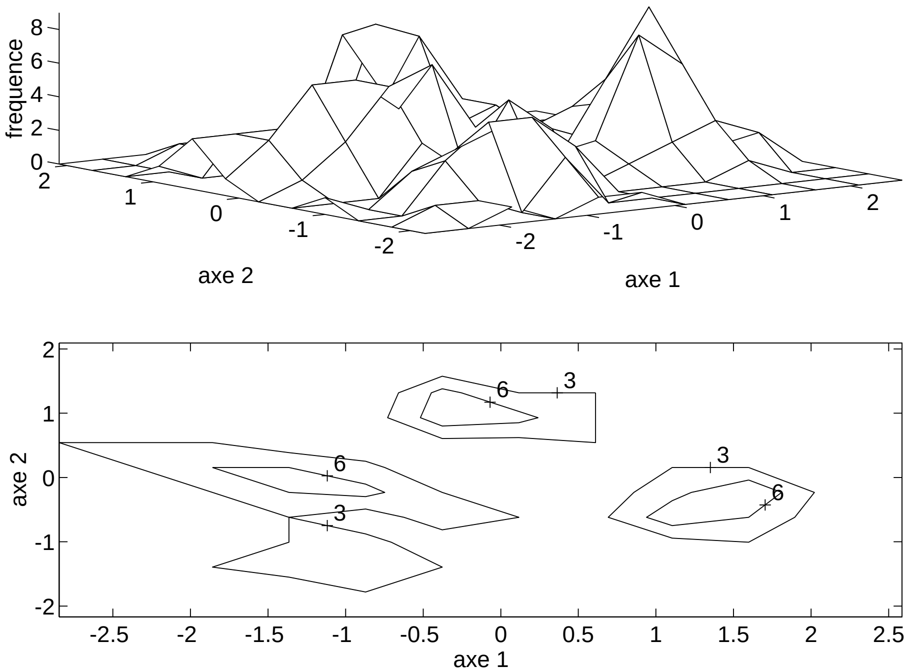


Fig. **2.14** – Représentation des composantes principales associées aux mesures effectuées dans le plan $(1,2)$.

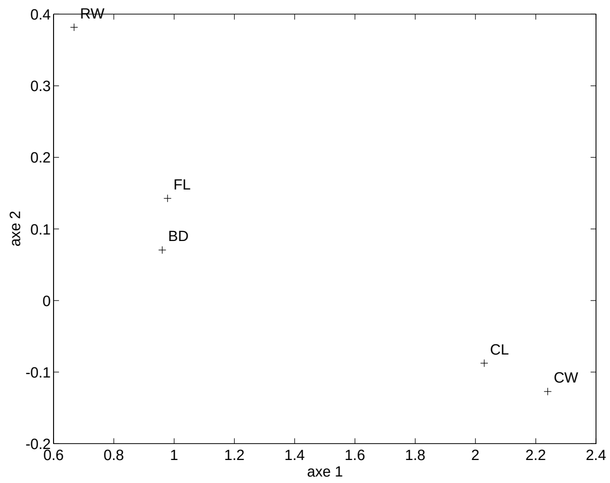
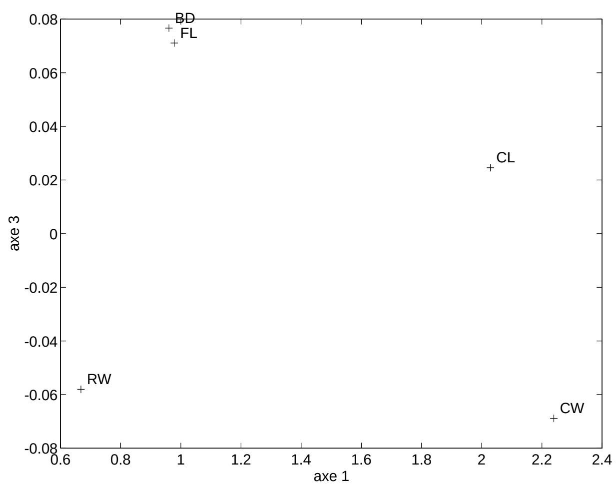


Fig. **2.15** – Représentation des composantes principales associées aux mesures effectuées dans le plan $(1,3)$.



Conclusion

Finalement, l'analyse en composantes principales est une méthode systématique permettant d'identifier des groupements dans un échantillon multidimensionnel. Les méthodes d'analyse statistique multidimensionnelle élementaires permettent aussi d'indentifier des groupements. Mais, à l'inverse de l'analyse en composantes principales, ces méthodes (et notament la prise en compte d'un effet taille) ne sont pas systématiques et nécessitent un « travail manuel » (aux conclusions empiriques) impossible à réaliser avec l'échantillon d'un vecteur aléatoire de dimension 200.

Annexe A

Programmation de l'analyse en composantes principales

A.1 Paramétrage

A.1.1 Entrées

- x : tableau de données $(n \times p)$,
- $D_p \in \mathcal{M}_{(n\times n)}$: matrice diagonale de poids des individus, en général on choisira $D_p = \frac{1}{n} I_{(n\times n)}$,
- $M \in \mathcal{M}_{(p\times p)}$: matrice diagonale de poids des variables, en général on choisira $M = I_{(p\times p)}$.

A.1.2 Sorties

- l : matrice diagonale $(p \times p)$ contenant les valeurs propres de la matrice $x^T D_p x M$ ordonnées par ordre décroissant,
- u : matrice $(p\times p)$ des vecteurs propres associés aux valeurs propres précédement calculées,
- e : vecteur colonnes (p) contenant les pourcentages d'énergie expliquée par les axes,
- C_α : matrice $(n \times p)$ des composantes principales selon les individus,
- D_α : matrice $(p \times p)$ des composantes principales selon les variables.

A.2 Implémentation avec *matlab*

A.2.1 Centrage en colonne

```
x = x-(ones(size(x))*diag(mean(x)));
```

A.2.2 Calcul des valeurs propres de $x^T D_p x M$

```
[u,l] = eig(x´*Dp*x*M);
```

A.2.3 Tri des valeurs propres et des vecteurs propres

```
[l,i] = sort(diag(l));
l     = flipud(l);
i     = flipud(i);
l     = diag(l);
u     = u(:,i);
```

A.2.4 Calcul du pourcentage d'énergie expliquée par les axes

```
e = 100*diag(l)/trace(l);
```

A.2.5 Calcul des composantes principales selon les individus

```
Ca = x*M*u;
```

A.2.6 Calcul des composantes principales selon les variables

```
Da = ones(size(l))*sqrt(l).*u;
% normalisation
Da = Da./(ones(size(Da))*diag(std(x)));
```

A.3 Calcul de l'histogramme bidimensionnel d'un vecteur aléatoire de dimension 2

```
function [h]=hist2(x1,x2,n,xl,yl)
```

```
if ~(size(x1,1) == size(x2,1))
    disp('size() error');
    return;
end;

dx1 = max(x1)-min(x1);
dx2 = max(x2)-min(x2);
```

A.3.1 Bornes des classes

```
cx1 = linspace(0,1,n);
cx1 = (cx1 * dx1 + min(x1))'; % vecteur colonne
cx2 = linspace(0,1,n);
cx2 = (cx2 * dx2 + min(x2))'; % vecteur colonne

h = zeros(n,n); % init. histogramme
```

A.3.2 Comptage

```
for i=1:size(x1,1)
    ix1 = find(cx1<=x1(i,1));
    ix2 = find(cx2<=x2(i,1));
    ix1 = ix1(size(ix1,1),1);
    ix2 = ix2(size(ix2,1),1);
    h(ix1,ix2)=h(ix1,ix2)+1;
end;
```

A.3.3 Representation de la surface

```
subplot(2,1,1);
mesh(cx1,cx2,h);
axis([min(x1) max(x1) min(x2) max(x2) 0 max(max(h))]);
hidden on;
xlabel(xl);
ylabel(yl);
zlabel('frequence');
```

A.3.4 Representation des lignes de niveaux

```
subplot(2,1,2);
```

```
c=contour(cx1,cx2,h,2);
axis([min(x1) max(x1) min(x2) max(x2)]);
clabel(c);
xlabel(xl);
ylabel(yl);
```