Module 16

MATHEMATIQUES

C07 - PROBABILITES – STATISTIQUES No 7

Analyse de données. Régressions. Analyse en composantes principales

- Comment aller au delà de la statistique descriptive en prenant en compte des données multidimensionnelles
- Régression
- Analyse en composantes principales
- Analyse factorielle des correspondances

Analyse de données

- Les outils classiques de la statistique descriptive prennent en compte, pour chaque élément observé :
  - Soit une variable (moyenne, médiane, variance, écart-type)
  - Soit deux variables (ajustement, corrélation, covariance)
- Les méthodes d'analyse de données généralisent l'observation à n variables (analyse de données multidimensionnelles)

Régression à une variable

- On peut chercher un ajustement linéaire entre le chiffre d'affaires Y (variable expliquée) et chacune des deux variables explicatives X1 (budget pub) et X2 (promotion des ventes)
- Les paramètres de cette droite sont donnés par

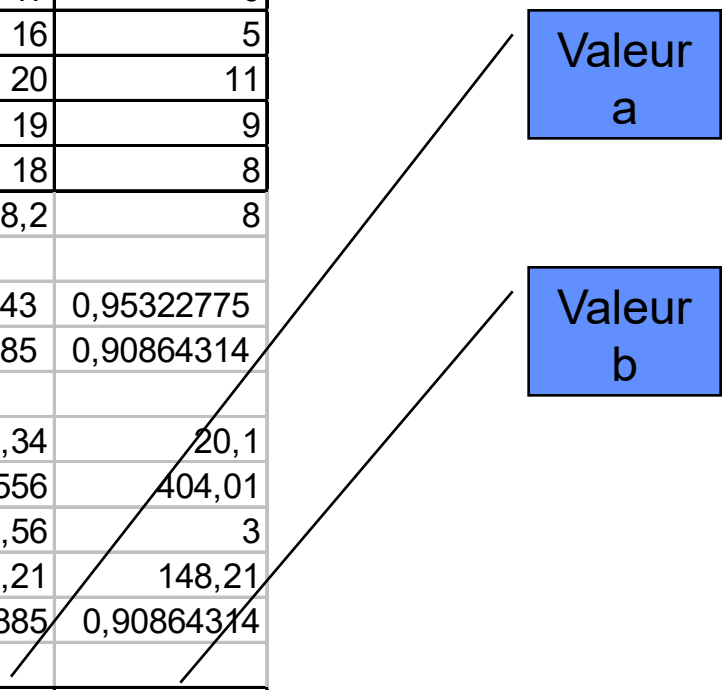
$$a = \frac{Cov(X_i, Y)}{V(X_i)}$$

$$b = \overline{Y} - a\overline{X}_i$$

- On rappelle que la covariance est la moyenne des produits des écarts pour chaque série d'observation

Régression à une variable

Période	Y	X1	X2
1	150	20	10
2	135	18	8
3	140	17	9
4	127	18	7
5	138	19	7
6	124	17	6
7	110	16	5
8	154	20	11
9	142	19	9
10	133	18	8
	135,3	18,2	8
	Coefficient correlation	0,87731343	0,95322775
	Son carré	0,76967885	0,90864314
	covariance	13,34	20,1
	covariance**2	177,9556	404,01
	V(X)	1,56	3
	V(Y)	148,21	148,21
		0,76967885	0,90864314
	Droite regression	8,55128205	-20,3333333
	Y,X1	1,65385858	30,1710225
		0,76967885	6,53221798
		26,7341095	8



Régression à une variable

Période	Y	X1	X2
1	150	20	10
2	135	18	8
3	140	17	9
	127	18	7
		19	7
	124	17	6
7	110		5
8	154		11
9	142	19	9
10	133	18	8
	135,3	18,2	8
	on	31343	0,95322775
		967885	0,90864314
	covariance	13,34	20,1
	covariance**2	177,9556	404,01
	V(X)	1,56	3
	V(Y)	148,21	148,21
		967885	0,90864314
		128205	-20,3333333
	Y,X1	1,65385858	30,1710225
		0,76967885	6,53221798
		26,7341095	8

CA

Budget anonceurs

Budget promotion FDV

Valeur a

Valeur b

Régression à une variable

Période	Y	X1	X2
1	150	20	10
	135,3		8
	Coefficient correlation	0,87731343	0,95322775
	Son carré	0,76967885	0,90864314
	covariance	13,34	20,1
	covariance**2	177,9556	404,01
	V(X)	1,56	3
	V(Y)	148,21	148,21
		0,76967885	0,90864314
	Droite regression	8,55128205	-20,3333333
	Y,X1	1,65385858	30,1710225
		0,76967885	6,53221798
		26,7341095	8

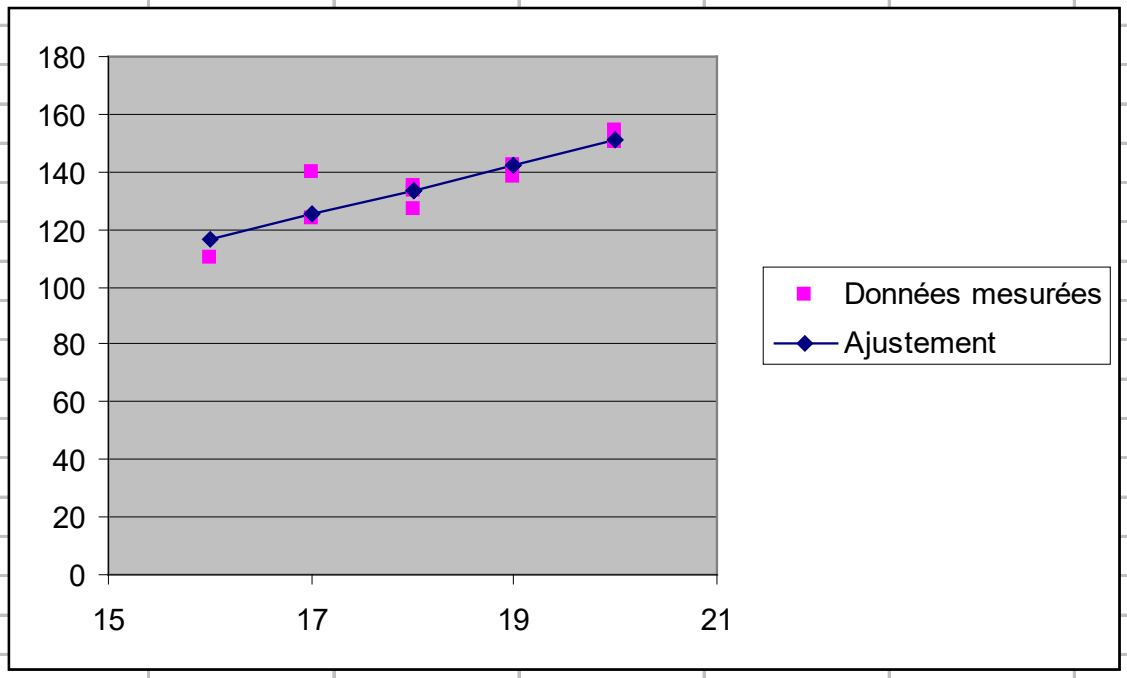
La corrélation CA - Budget promotion FDV est meilleure

Valeur a

Valeur b

Régression à une variable

	X1	DX1	Y
	16	116,487179	110
	17	125,038462	140
	17	125,038462	124
	18	133,589744	135
	18	133,589744	127
	18	133,589744	133
	19	142,141026	138
	19	142,141026	142
	20	150,692308	150
	20	150,692308	154



Régression à une variable

La fonction Y = aXi+b minimise les carrés des écarts entre les valeurs réelles de Y et les valeurs ajustée

Le coefficient de corrélation mesure la réalité de cet ajustement :

$$\rho = \frac{Cov(X,Y)}{\sigma(X)*\sigma(Y)}$$

où X et Y sont MOYENNE(matrice1) et MOYENNE(matrice2).

L'ajustement est meilleur avec X2

Régression à deux variables

On recherche de la même manière une fonction $Y = a1X1 + a2X2 + b$ qui minimise les carrés des écarts entre les valeurs réelles de Y et les valeurs ajustée

Analyse en composantes principales

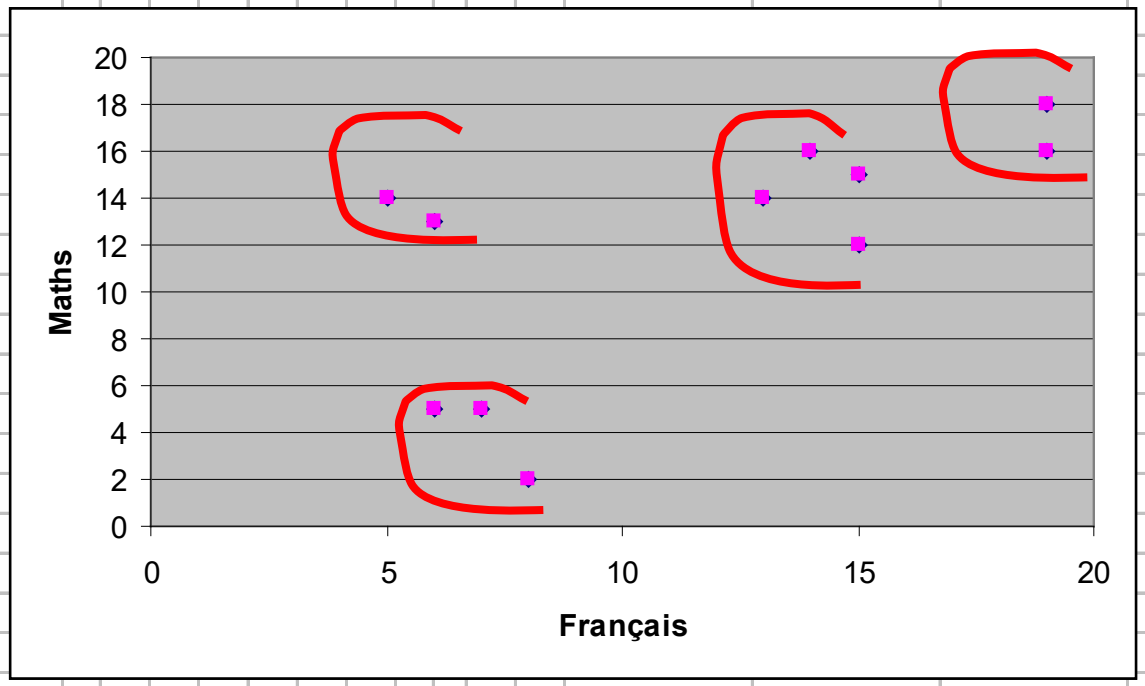
On ne cherche plus à expliquer le comportement de variables Y par les variables X.

A partir d'un certain nombre d'observations portant sur un ensemble d'individus, on cherche simplement à repérer l'existence de groupes d'individus ayant, par rapport aux variables observées, des profils communs

Analyse en composantes principales

Exemple des notes obtenues par un groupe d'élèves dans 2 matières

Individus	A	B	C	D	E	F	G	H	I	J	K
Français	6	6	19	15	19	15	5	8	7	14	13
Maths	5	13	18	12	16	15	14	2	5	16	14



Analyse en composantes principales

- Si plus de 2 variables, travail dans un espace à n dimensions
- Projection sur 2 axes, choisis comme étant ceux pour lesquels la projection du nuage de point a une variance maximale
- Nécessité logiciel spécifique

Analyse en composantes principales

- Données de base (Amerique du Sud 1996)

	PNB/Hab	Taux chômage	Taux d'inflation	Dette/PNB
Argentine (A)	8320	18,0%	0,1%	30,00%
Bolivie (BO)	800	5,8%	10,2%	79,00%
Brésil (BR)	3801	4,7%	11,0%	22,00%
Chili (CH)	4545	4,9%	6,6%	11,00%
Colombie (CO)	1910	8,6%	21,0%	30,00%
Equateur (EQ)	1390	12,0%	23,0%	73,00%
Paraguay (PA)	1690	4,8%	8,0%	30,00%
Pérou (PE)	2310	8,8%	11,0%	48,00%
Uruguay (U)	5170	10,7%	42,2%	70,00%
Vénézuela (V)	2548	13,0%	1,02	44,00%

Analyse en composantes principales

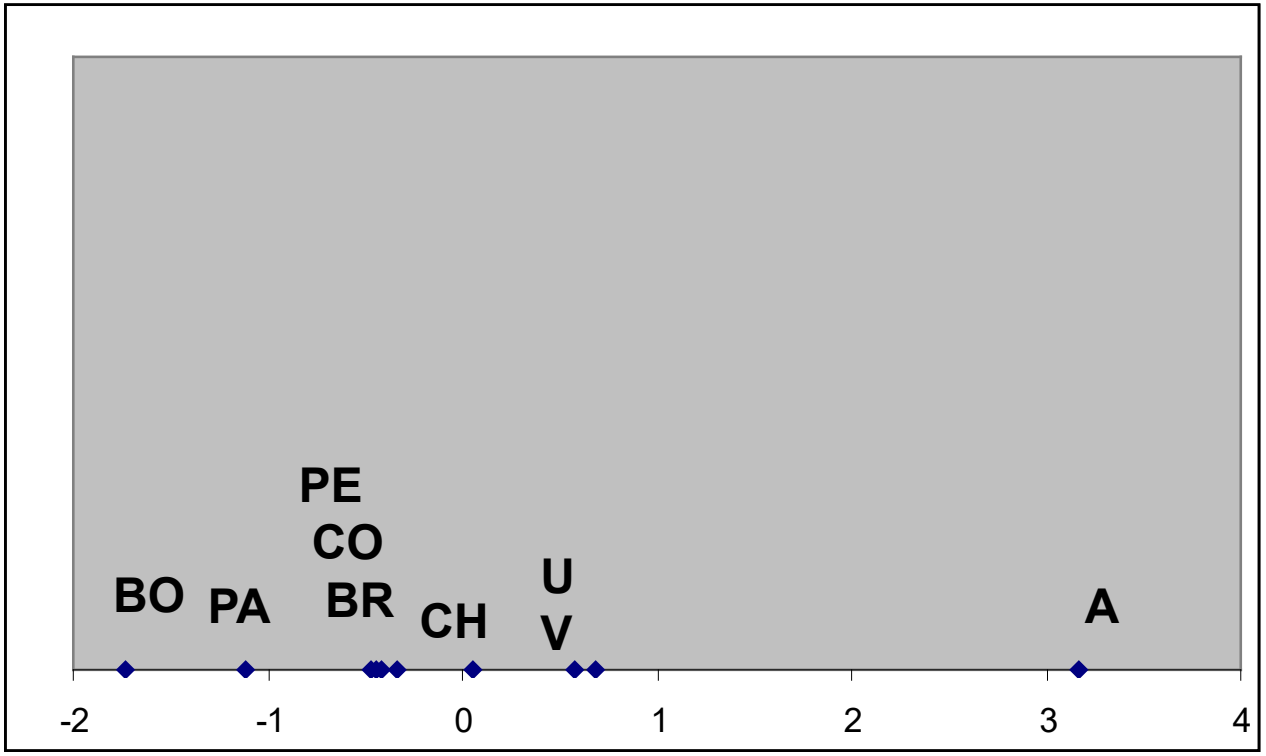
- Résultats programme ACP

Composantes sur axes principaux		
	Axe 1	Axe 2
PNB/Hab	0,727	-0,237
Taux chômage	0,647	0,403
Taux d'inflation	0,076	0,619
Dette/PNB	-0,215	0,631
Part des axes dans variance totale		
Axe 1	0,393	
Axe 2	0,376	

Coordonnées des pays sur les axes		
	Axe 1	Axe 2
Argentine (A)	3,17	-0,60
Bolivie (BO)	-1,73	0,66
Brésil (BR)	-0,33	-1,38
Chili (CH)	0,05	-1,85
Colombie (CO)	-0,41	-0,35
Equateur (EQ)	-0,47	1,31
Paraguay (PA)	-1,11	-0,98
Pérou (PE)	-0,44	-0,08
Uruguay (U)	0,69	1,10
Vénézuela (V)	0,58	2,17

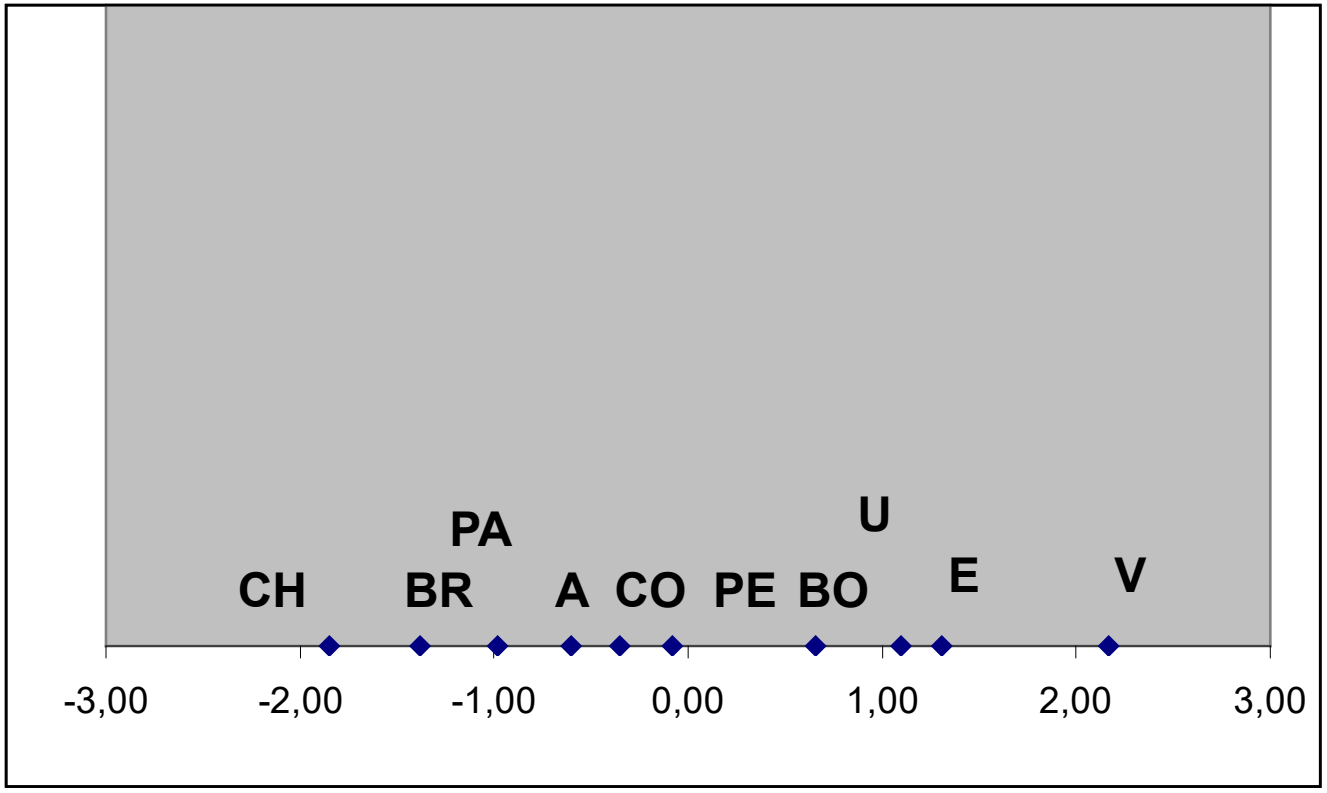
Analyse en composantes principales

- Interprétation résultats
- Axe 1, 39% de la variance totale, marqué par le poids de la composante PNB/Hab (0,727) => niveau développement



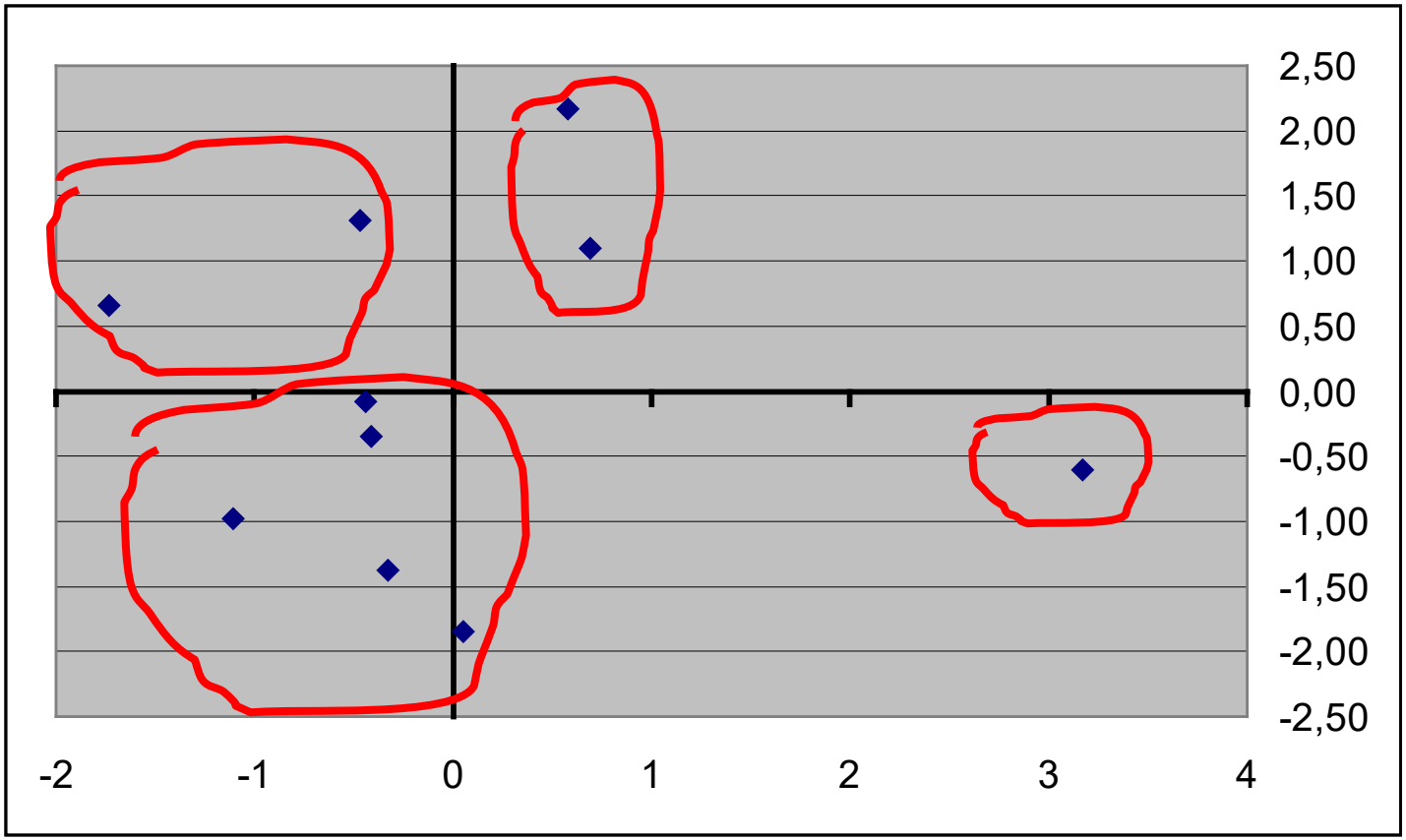
Analyse en composantes principales

- Interprétation résultats
- Axe 2, 37% de la variance totale, marqué par le poids des composantes taux d'inflation (0,619) et dette/PNB (0,631) => Conformité FMI



Analyse en composantes principales

- Projection sur le plan des deux axes principaux



Analyse factorielle des correspondances

- Même principe appliqué à des tableaux de contingence répartissant une population statistique en fonction de 2 variables, dans un tableau à double entrée
- Les totaux verticaux et horizontaux ont alors une signification (Vérif 100 % en valeur relative)