

Analyse en Composantes Principales

Définitions

Les données

Le tableau des données initiales

$$X = \begin{pmatrix} \begin{matrix} X_1 \\ \downarrow \\ x_{1,1} \\ \vdots \\ x_{n,1} \end{matrix} & \dots & \begin{matrix} X_p \\ \downarrow \\ x_{1,p} \\ \vdots \\ x_{n,p} \end{matrix} \end{pmatrix} \quad \begin{matrix} \leftarrow & \text{Individu } 1 \\ & \\ & \\ \leftarrow & \text{Individu } n \end{matrix}$$

Tableau des données centrées réduites Z

On calcule pour chacune des variables X_k sa moyenne $\bar{x}_k$ et son écart-type s_k .

On calcule ensuite les $z_{i,k} = \frac{x_{i,k} - \bar{x}_k}{s_k}$.

Matrice de corrélation

Chaque élément R_{ij} est défini par $R_{ij} = r(X_i, X_j) = r(Z_i, Z_j)$.

$$R = \begin{pmatrix} 1 & r(X_1, X_2) & \dots & r(X_1, X_p) \\ r(X_1, X_2) & 1 & & r(X_2, X_p) \\ \vdots & & \ddots & \vdots \\ r(X_1, X_p) & & & 1 \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & & \lambda_p \end{pmatrix}$$

$$\text{Trace}(R) = \lambda_1 + \lambda_2 + \dots + \lambda_p = p \text{ avec } \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p.$$

Inertie

On appelle inertie totale du nuage de points la moyenne des carrés des distances des n points au centre de gravité :

$$I_G(N) = \frac{1}{n} \sum d(G, x_i)^2 = \frac{1}{n} \sum d(O, z_i)^2$$

L'inertie mesure la dispersion du nuage de points. On montre que

$$I_G(N) = \text{Trace}(R) = p = \lambda_1 + \lambda_2 + \dots + \lambda_p.$$

Analyse en composantes principales

Premier axe principal

On cherche la droite pour laquelle l'inertie du nuage projeté $I_G(\hat{N})$ est maximale.

$$\begin{aligned} I_G(N) &= I_G(\hat{N}) + \frac{1}{n} \sum_i d(z_i, \hat{z}_i)^2 \\ \lambda_1 + \dots + \lambda_p &= I_G(\hat{N}) + \frac{1}{n} \sum_i d(z_i, \hat{z}_i)^2 \end{aligned}$$

La solution est obtenue pour $I_G(\hat{N}) = \lambda_1$, la plus grande valeur propre de R .

On appelle première composante principale C^1 le vecteur formé des coordonnées des projections des individus sur l'axe principal. On calcule C^1 par

$$C^1 = u_1^1 Z_1 + u_2^1 Z_2 + \dots + u_p^1 Z_p$$

où u^1 est le vecteur propre associé à λ_1 .

$$\overline{C^1} = 0 \text{ et } Var(C^1) = \lambda_1.$$

Pourcentage d'inertie expliquée par le premier axe principal

$$\frac{I_G(\hat{N})}{I_G(N)} \times 100 = \frac{\lambda_1}{\sum \lambda_i} \times 100 = \frac{\lambda_1}{p} \times 100$$

Premier plan principal

De la même façon, on montre que le deuxième axe principal est associé à la deuxième plus grande valeur propre, λ_2 .

$$C^2 = u_1^2 Z_1 + u_2^2 Z_2 + \dots + u_p^2 Z_p$$

est la seconde composante principale, avec u^2 le vecteur propre associé à λ_2 .

On montre de même que $\overline{C^2} = 0$ et $Var(C^2) = \lambda_2$.

De plus, $r(C^1, C^2) = 0$. (Le second axe principal est orthogonal au premier).

Le pourcentage d'inertie expliqué par ce second axe est $\frac{\lambda_2}{\sum \lambda_i} \times 100 = \frac{\lambda_2}{p} \times 100$.

On appelle **premier plan principal** le plan défini par les deux premiers axes principaux. Le pourcentage d'inertie expliqué par ce premier plan principal est le rapport $\frac{\lambda_1 + \lambda_2}{\sum \lambda_i} \times 100 = \frac{\lambda_1 + \lambda_2}{p} \times 100$.

Interprétation d'une ACP

1. Examen de la matrice de corrélation

Effet taille : Lorsque toutes les variables sont corrélées positivement entre elles, la matrice R a tous ses termes positifs ; la première composante principale définit alors un “facteur de taille” ; les individus sont rangés sur l'axe 1 par valeurs croissantes d'une moyenne pondérée des variables initiales Z_i .

2. Détermination du nombre de composantes principales à retenir

- Critère des valeurs propres : on retient les composantes principales pour lesquelles $\lambda > 1$.
- Critère de l'inertie : on regarde le pourcentage d'inertie expliquée par les premières composantes principales, ainsi que l'inertie cumulée.

Remarque : s'il y a un effet taille, on peut assouplir les critères.

3. Interprétation des composantes principales

On calcule les coefficients de corrélation linéaire des composantes (retenues) avec les variables initiales $r(C^i, X_j) = r(C^i, Z_j)$.

Le cercle des corrélations permet de visualiser les relations.

4. Interprétation des individus

- O est le point moyen.
- Dire que C^1 est très corrélée avec X_j signifie que les individus ayant une coordonnée élevée sur l'axe 1 sont caractérisés par une valeur de X_j nettement supérieure à la moyenne. De même pour les autres axes.

On recherche donc les individus opposés sur un axe.

- On peut observer d'éventuels regroupements d'individus.

Quelques compléments

- On peut regarder la qualité de la représentation de chaque individu, en calculant un indice de qualité de la représentation basé sur l'angle entre l'individu e_i et sa projection $\hat{z}_i$.
- Il peut être utile aussi de calculer pour chaque axe la contribution apportée par les divers individus à cet axe. Une contribution excessive ($>1/4$) est un facteur d'instabilité.
- Individus supplémentaires

Exemples et Exercices ACP

Exemple 1: les vins de Bordeaux

Source : « Probabilités et Statistiques », D. Dacunha-Castelle, M. Duflo).

Dans le cadre d’une étude de vins de Bordeaux, on recueille 34 observations climatiques des mois d’Avril à septembre des années 1924 à 1957. Les variables sont les suivantes :

- X1 : somme des températures en degrés.
- X2 : durée d’insolation (en heures).
- X3 : nombre de jours de grande chaleur.
- X4 : hauteur des pluies (en mm).
- On a de plus classé en 7 catégories les vins suivant leur qualité (allant de +++ à ---).

Simple Statistics				
	X1	X2	X3	X4
Mean	3158.470588	1248.705882	18.82352941	360.4411765
StD	142.094982	127.005256	10.01656382	91.4016084

Correlation Matrix					
		X1	X2	X3	X4
X1	X1	1.0000	0.7146	0.8622	-.4120
X2	X2	0.7146	1.0000	0.6459	-.4750
X3	X3	0.8622	0.6459	1.0000	-.4011
X4	X4	-.4120	-.4750	-.4011	1.0000

Eigenvalues of the Correlation Matrix				
	Valeur propre	Différence	Proportion	Cumulée
1	2.79234371	2.07976272	0.6981	0.6981
2	0.71258099	0.34793306	0.1781	0.8762
3	0.36464793	0.23422057	0.0912	0.9674
4	0.13042736		0.0326	1.0000

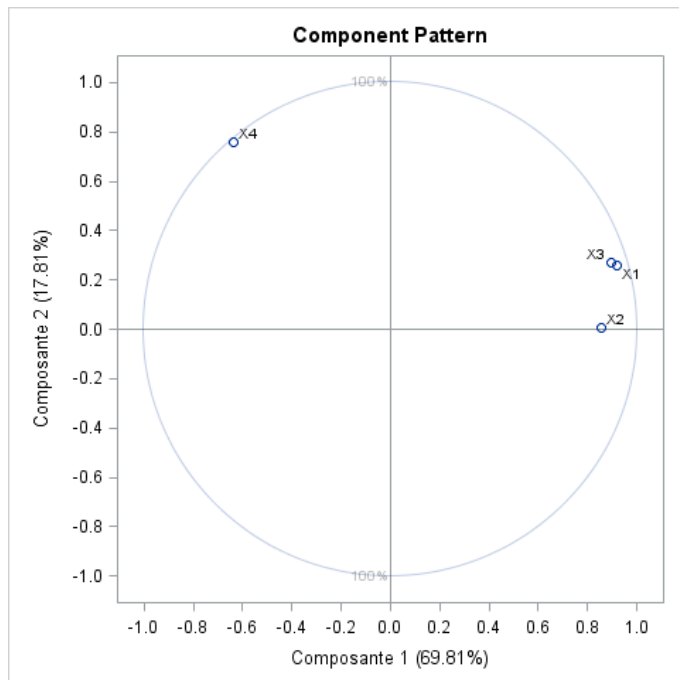
Eigenvectors					
		Prin1	Prin2	Prin3	Prin4
X1	X1	0.550369	0.303831	-.201626	-.751084
X2	X2	0.513789	0.009543	0.843967	0.153787
X3	X3	0.535518	0.319664	-.446537	0.641592
X4	X4	-.382539	0.897447	0.218340	0.024114

Coordonnées des vins dans le plan principal

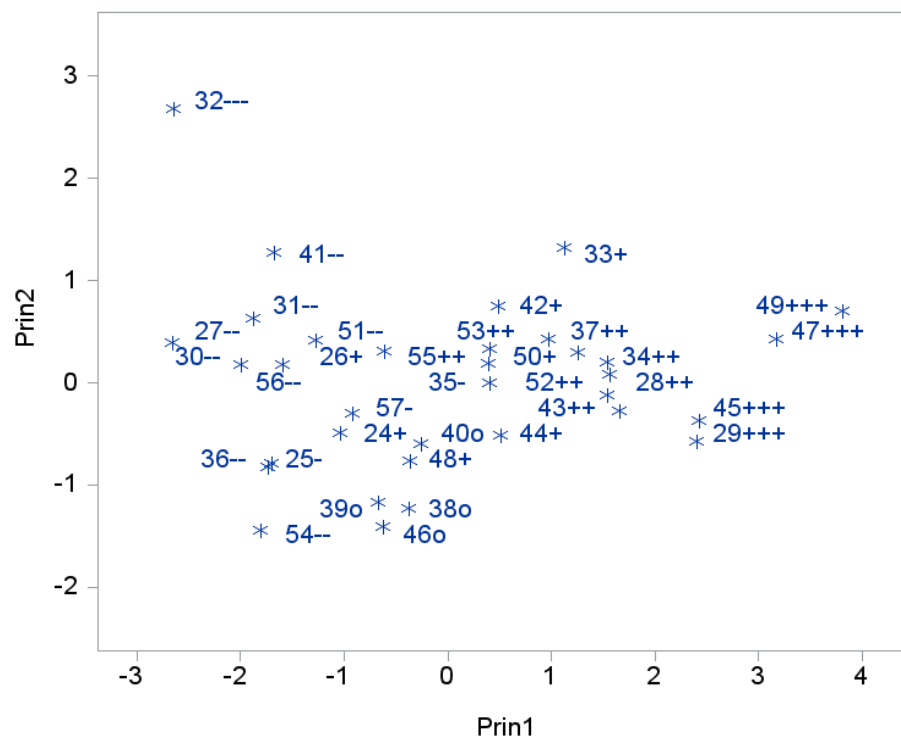
Annees	Prin1	Prin2	Annees	Prin1	Prin2
47	3.17754	0.43198	24	-1.03297	-0.48169
49	3.81321	0.70059	46	-0.62432	-1.40810
45	2.43184	-0.36733	40	-0.25342	-0.59202
29	2.40747	-0.57123	38	-0.37283	-1.22798
43	1.66681	-0.27630	39	-0.66655	-1.16658
34	1.54670	0.20643	35	0.41153	-0.00084
52	1.54458	-0.11957	57	-0.91399	-0.30022
28	1.56913	0.08151	25	-1.72986	-0.82357
55	0.40586	0.34000	30	-1.99566	0.18046
37	0.97746	0.42801	36	-1.69234	-0.77999
53	0.39202	0.19025	56	-1.58579	0.18442
44	0.51744	-0.51453	27	-2.65054	0.39516
50	1.26066	0.29566	51	-1.26947	0.42432
33	1.13136	1.32401	54	-1.80696	-1.44517
42	0.49055	0.74805	31	-1.87089	0.63547
48	-0.35671	-0.76533	41	-1.67469	1.27923
26	-0.60835	0.30920	32	-2.63881	2.68572

Coefficients de corrélation de Pearson

	X1	X2	X3	X4	Prin1	Prin2
Prin1	0.91968	0.85856	0.89487	-0.63923	1.00000	0.00000
Prin2	0.25648	0.00806	0.26984	0.75758	0.00000	1.00000



Le Système SAS



Exemple 2: Alimentation et CSP

(D'après G. Saporta – Probabilités et statistique – eds Technip)

On étudie les consommations annuelles, en 1972 et exprimées en francs, de 8 denrées alimentaires pour 8 catégories socio-professionnelles. Les données sont des moyennes par CSP . Le tableau suivant renferme les données (arrondies au franc près) ainsi que les abréviations utilisées par la suite.

		Pain ordinaire PAO	Autre pain PAA	Vin ordinaire VIO	Autre vin VIA	Pommes de terres POT	Légumes secs LEC	Raisin de table RAI	Plats préparés PLP
Exploitants agricoles	AGRI	167	1	163	23	41	8	6	6
Salariés agricoles	SAAG	162	2	141	12	40	12	4	15
Professions indépendantes	PRIN	119	6	69	56	39	5	13	41
Cadres supérieurs	CSUP	87	11	63	111	27	3	18	39
Cadres moyens	CMOY	103	5	68	77	32	4	11	30
Employés	EMPL	111	4	72	66	34	6	10	28
Ouvriers	OUVR	130	3	76	52	43	7	7	16
Inactifs	INAC	138	7	117	74	53	8	12	20

Statistiques simples

	PAO	PAA	VIO	VIA	POT	LEC	RAI	PLP
Moyenne	127.1250000	4.875000000	96.12500000	58.87500000	38.62500000	6.625000000	10.12500000	24.37500000
StD	27.9051327	3.181980515	38.79041671	31.31607301	7.87287386	2.825268635	4.45413131	12.24671968

Matrice de corrélation

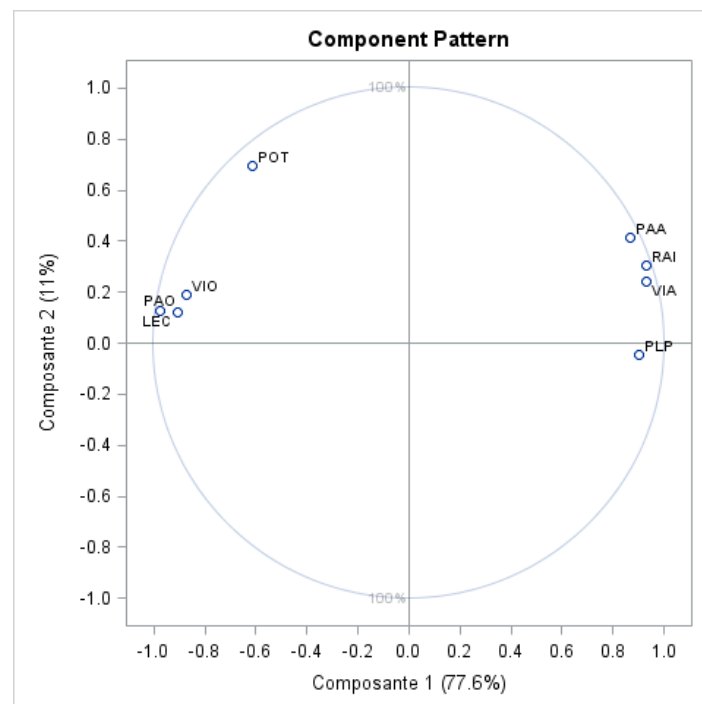
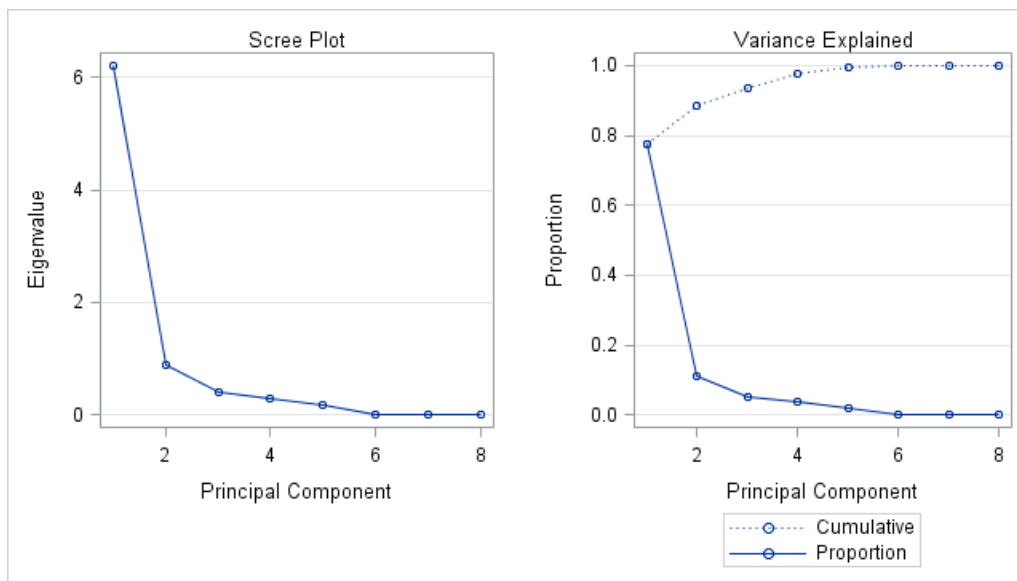
		PAO	PAA	VIO	VIA	POT	LEC	RAI	PLP
PAO	PAO	1.0000	-.7737	0.9262	-.9058	0.6564	0.8886	-.8334	-.8558
PAA	PAA	-.7737	1.0000	-.6040	0.9044	-.3329	-.6734	0.9588	0.7712
VIO	VIO	0.9262	-.6040	1.0000	-.7502	0.5171	0.7917	-.6690	-.8280
VIA	VIA	-.9058	0.9044	-.7502	1.0000	-.4186	-.8386	0.9239	0.7198
POT	POT	0.6564	-.3329	0.5171	-.4186	1.0000	0.6029	-.4099	-.5540
LEC	LEC	0.8886	-.6734	0.7917	-.8386	0.6029	1.0000	-.8245	-.7509
RAI	RAI	-.8334	0.9588	-.6690	0.9239	-.4099	-.8245	1.0000	0.8344
PLP	PLP	-.8558	0.7712	-.8280	0.7198	-.5540	-.7509	0.8344	1.0000

Valeurs propres de la matrice de corrélation

	Valeur propre	Différence	Proportion	Cumulé
1	6.20794684	5.32826545	0.7760	0.7760
2	0.87968139	0.46372027	0.1100	0.8860
3	0.41596112	0.10950645	0.0520	0.9379
4	0.30645467	0.13801317	0.0383	0.9763
5	0.16844150	0.15037379	0.0211	0.9973
6	0.01806771	0.01462094	0.0023	0.9996
7	0.00344677	0.00344677	0.0004	1.0000
8	0.00000000		0.0000	1.0000

Vecteurs propres

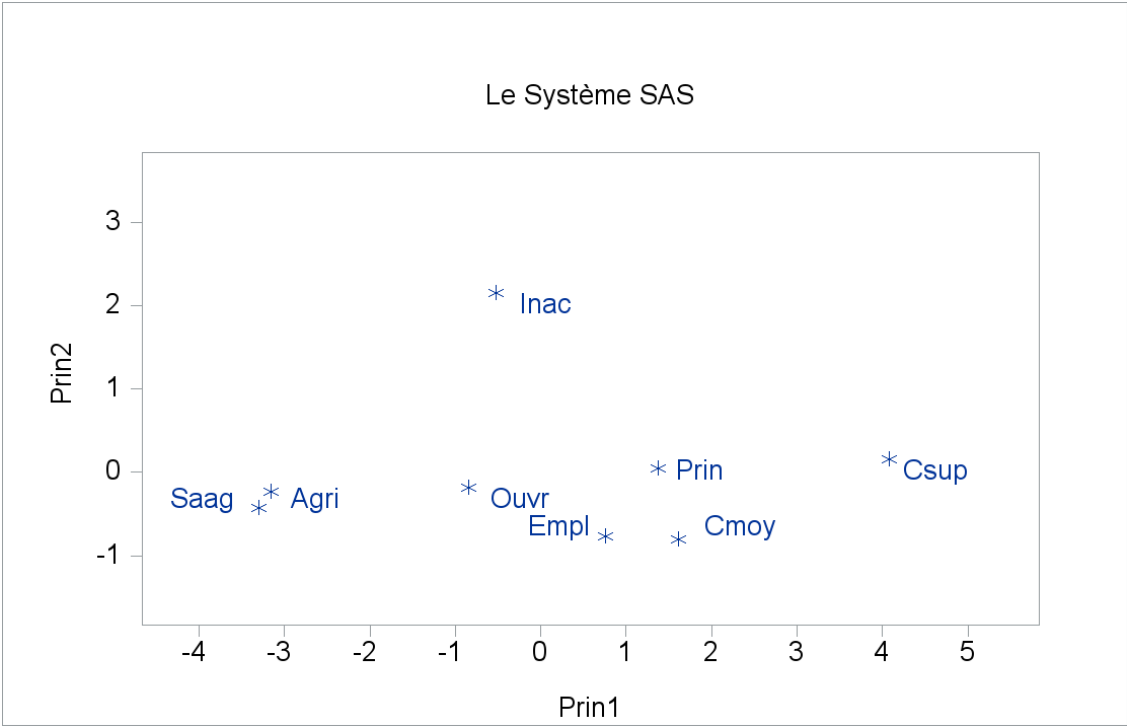
		Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8
PAO	PAO	-.391311	0.137823	0.161714	0.119350	0.294045	-.397748	0.106920	0.728963
PAA	PAA	0.348674	0.440585	0.319950	0.217909	-.265442	-.520704	-.423079	-.117773
VIO	VIO	-.349193	0.201682	0.680632	-.028883	0.245716	0.464752	-.253923	-.180130
VIA	VIA	0.373625	0.260309	0.073482	-.396545	-.345605	0.422866	-.033345	0.575000
POT	POT	-.246371	0.743826	-.557660	-.073992	0.175725	0.107747	-.093428	-.135449
LEC	LEC	-.364822	0.128021	0.032401	0.518889	-.669192	0.184942	0.313107	0.012735
RAI	RAI	0.373052	0.325980	0.254250	0.063706	0.271532	-.016265	0.765903	-.158952
PLP	PLP	0.361676	-.050227	-.161692	0.708103	0.332914	0.360245	-.224966	0.218851



Obs.	CSP	Prin1	Prin2	QLT
1	Agri	-3.15382	-0.22994	0.88914
2	Saag	-3.29426	-0.41850	0.91255
3	Prin	1.37696	0.05473	0.57551
4	Csup	4.07727	0.16473	0.94336
5	Cmoy	1.60711	-0.80132	0.94006
6	Empl	0.75444	-0.75631	0.85769
7	Ouvr	-0.84103	-0.17122	0.37555
8	Inac	-0.52667	2.15782	0.98745

(contributions en pourcentage)

Obs.	ctr1	ctr2
1	22.8891	0.8586
2	24.9729	2.8443
3	4.3631	0.0487
4	38.2554	0.4407
5	5.9436	10.4277
6	1.3098	9.2891
7	1.6277	0.4761
8	0.6383	75.6149



Exemple 3: Crimes US

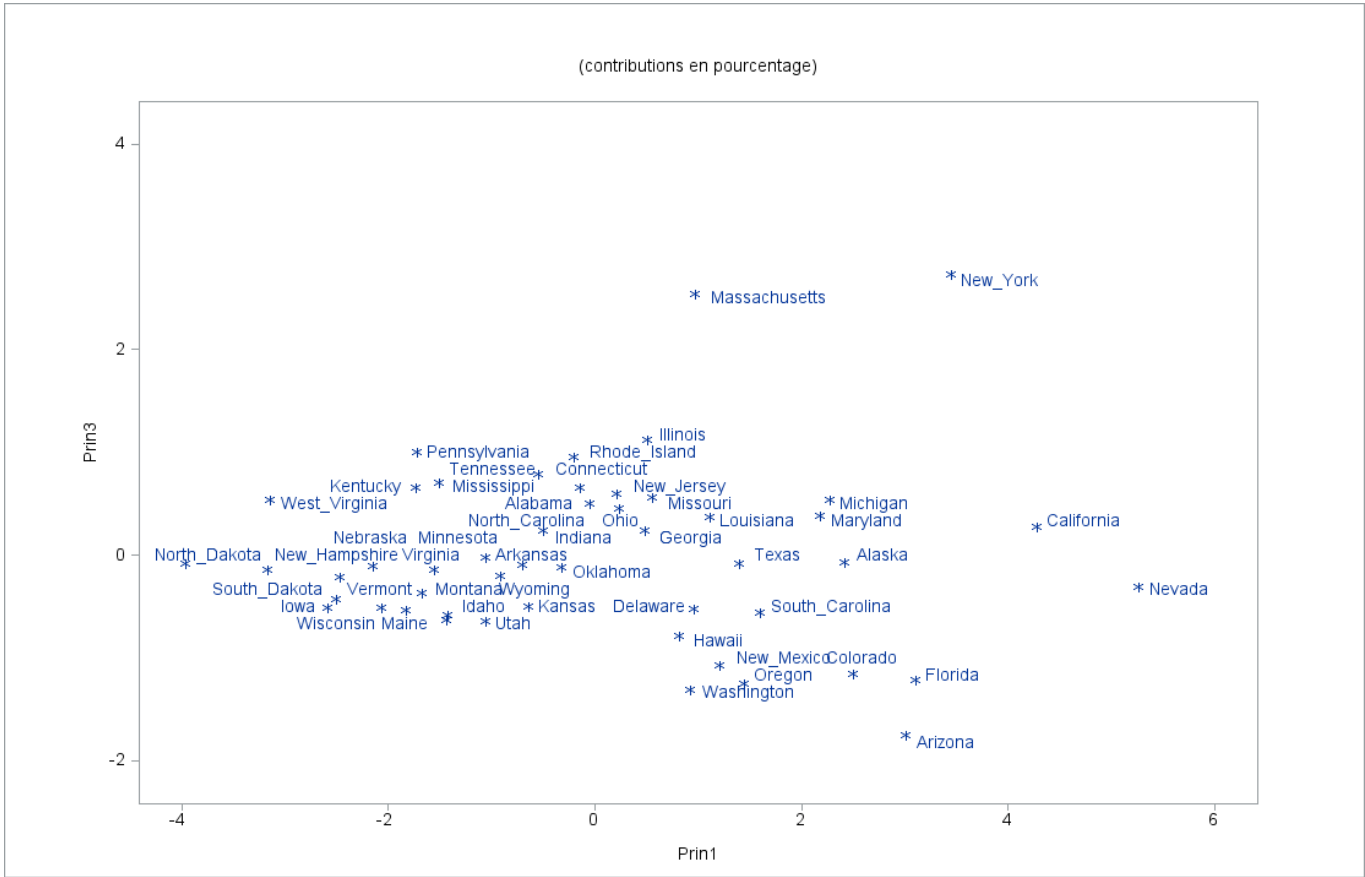
Les données sont les taux de crimes pour 100.000 personnes dans sept catégories et pour chacun des 50 états des US en 1977. Les variables sont les suivantes : Murder (meurtre), Rape (viol), Assault (agression), Robbery (vol qualifié), Burglary (cambriolage), Larceny (larcin) et Auto-thieft (vol de voiture).

Statistiques simples							
	Murder	Rape	Robbery	Assault	Burglary	larceny	Auto_thieft
Moyenne	7.444000000	25.73400000	124.0920000	211.3000000	1291.904000	2671.288000	377.5260000
StD	3.866768941	10.75962995	88.3485672	100.2530492	432.455711	725.908707	193.3944175

Matrice de corrélation								
		Murder	Rape	Robbery	Assault	Burglary	larceny	Auto_thieft
Murder	Murder	1.0000	0.6012	0.4837	0.6486	0.3858	0.1019	0.0688
Rape	Rape	0.6012	1.0000	0.5919	0.7403	0.7121	0.6140	0.3489
Robbery	Robbery	0.4837	0.5919	1.0000	0.5571	0.6372	0.4467	0.5907
Assault	Assault	0.6486	0.7403	0.5571	1.0000	0.6229	0.4044	0.2758
Burglary	Burglary	0.3858	0.7121	0.6372	0.6229	1.0000	0.7921	0.5580
larceny	larceny	0.1019	0.6140	0.4467	0.4044	0.7921	1.0000	0.4442
Auto_thieft	Auto-thieft	0.0688	0.3489	0.5907	0.2758	0.5580	0.4442	1.0000

Valeurs propres de la matrice de corrélation					
	Valeur propre	Différence		Proportion	Cumulé
1	4.11495951	2.87623768		0.5879	0.5879
2	1.23872183	0.51290521		0.1770	0.7648
3	0.72581663	0.40938458		0.1037	0.8685
4	0.31643205	0.05845759		0.0452	0.9137
5	0.25797446	0.03593499		0.0369	0.9506
6	0.22203947	0.09798342		0.0317	0.9823
7	0.12405606			0.0177	1.0000

Vecteurs propres								
		Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7
Murder	Murder	0.300279	-.629174	0.178245	-.232114	0.538123	0.259117	0.267593
Rape	Rape	0.431759	-.169435	-.244198	0.062216	0.188471	-.773271	-.296485
Robbery	Robbery	0.396875	0.042247	0.495861	-.557989	-.519977	-.114385	-.003903
Assault	Assault	0.396652	-.343528	-.069510	0.629804	-.506651	0.172363	0.191745
Burglary	Burglary	0.440157	0.203341	-.209895	-.057555	0.101033	0.535987	-.648117
larceny	larceny	0.357360	0.402319	-.539231	-.234890	0.030099	0.039406	0.601690
Auto_thieft	Auto-thieft	0.295177	0.502421	0.568384	0.419238	0.369753	-.057298	0.147046



Exemple 4 : Les notes scolaires

On considère les notes sur 20 en maths, physique, français et anglais de 15 élèves.

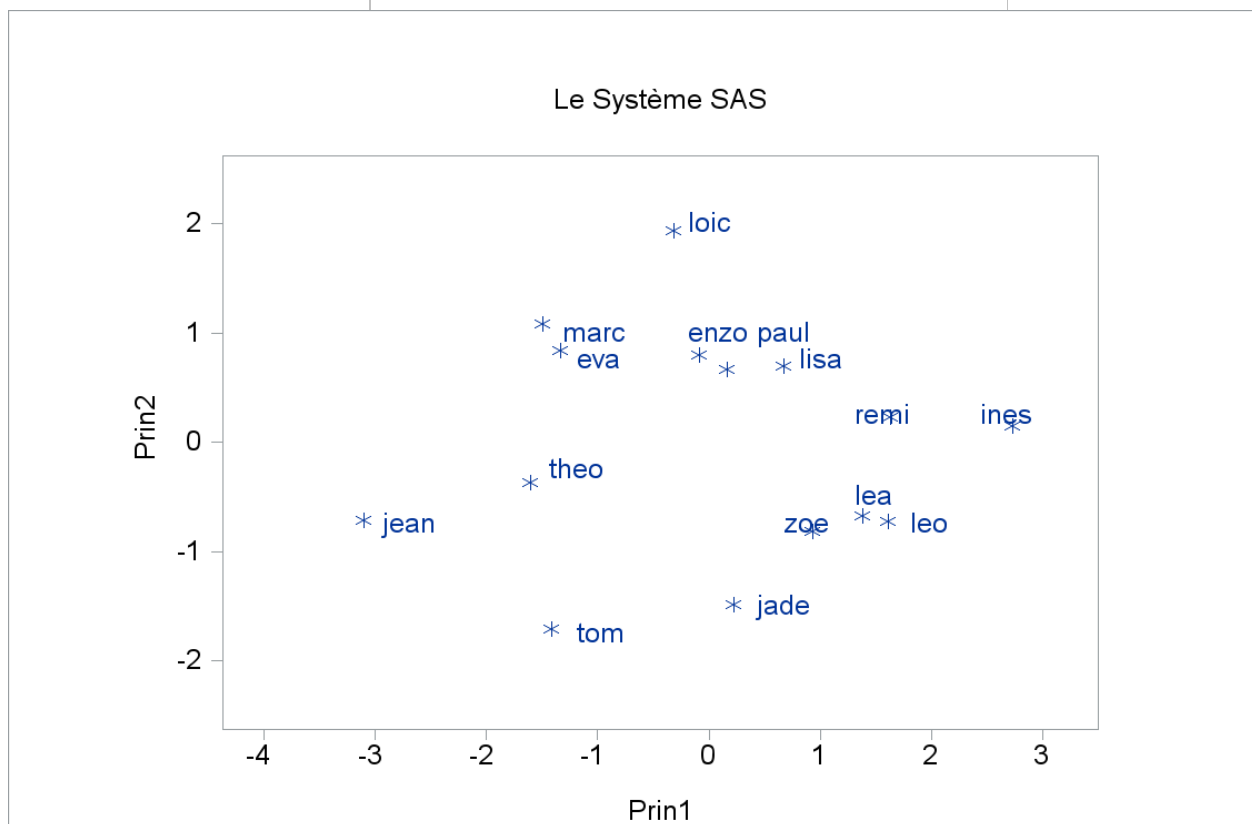
Simple Statistics				
	MATHS	PHYSIQ	FRAN	ANGL
Mean	10.30000000	9.786666667	10.73333333	11.26666667
StD	3.86744656	2.880443088	3.16152465	2.91465917

Correlation Matrix					
	MATHS	PHYSIQ	FRAN	ANGL	
MATHS	MATHS	1.0000	0.7092	0.1311	0.5785
PHYSIQ	PHYSIQ	0.7092	1.0000	0.3306	0.4012
FRAN	FRAN	0.1311	0.3306	1.0000	0.7350
ANGL	ANGL	0.5785	0.4012	0.7350	1.0000

Eigenvalues of the Correlation Matrix				
	Valeur propre	Différence	Proportion	Cumulé
1	2.45814791	1.41085051	0.6145	0.6145
2	1.04729740	0.61425114	0.2618	0.8764
3	0.43304625	0.37153781	0.1083	0.9846
4	0.06150844		0.0154	1.0000

Eigenvectors					
		Prin1	Prin2	Prin3	Prin4
MATHS	MATHS	0.501718	-.522436	-.421940	0.545257
PHYSIQ	PHYSIQ	0.496498	-.424051	0.677288	-.339045
FRAN	FRAN	0.438212	0.663723	0.344125	0.499020
ANGL	ANGL	0.556542	0.326669	-.494799	-.581998

	MATHS	PHYSIQ	FRAN	ANGL
Prin1	0.78662	0.77843	0.68705	0.87257
Prin2	-0.53465	-0.43396	0.67924	0.33431
Prin3	-0.27766	0.44570	0.22646	-0.32561
Prin4	0.13523	-0.08409	0.12376	-0.14434



1. Examinez la matrice de corrélation, qu'en pensez-vous ?
2. Combien de composantes principales allez-vous conserver ? Justifier ;
3. Tracer le cercle des corrélations. Interpréter les deux premiers axes principaux.
4. Interprétez la représentation des individus.