Analyse de Données – ID
L'ACP par l'exemple

Philippe LERAY
philippe.leray@univ-nantes.fr

Equipe COnnaissances et Décision
Laboratoire d'Informatique de Nantes Atlantique – FRE 2729
Site de l'Ecole Polytechnique de l'université de Nantes

Polytech'Nantes
École d'ingénieurs de l'université de Nantes

UNIVERSITÉ DE NANTES

lina LABORATOIRE D'INFORMATIQUE DE NANTES ATLANTIQUE

Les données (Besse & Baccini, 2005)

Notes (de 0 à 20) obtenues par 9 élèves dans 4 disciplines

	MATH	PHYS	FRAN	ANGL
jean	6.00	6.00	5.00	5.50
alan	8.00	8.00	8.00	8.00
anni	6.00	7.00	11.00	9.50
moni	14.50	14.50	15.50	15.00
didi	14.00	14.00	12.00	12.50
andr	11.00	10.00	5.50	7.00
pier	5.50	7.00	14.00	11.50
brig	3.00	12.50	8.50	9.50
evel	9.00	9.50	12.50	12.00

Stats descriptives classiques : 1D, 2D ... ici 4D ?

Analyse en Composantes Principales (ACP) = réduire l'espace en déformant le moins possible la réalité (les données)

Statistiques préliminaires

Variable	Moyenne	Ecart-type	Minimum	Maximum
MATH	9.67	3.37	5.50	14.50
PHYS	9.83	2.99	6.00	14.50
FRAN	10.22	3.47	5.00	15.50
ANGL	10.06	2.81	5.50	15.00

Grande homogénéité des 4 variables considérées

Corrélation linéaire - Covariance

	MATH	PHYS	FRAN	ANGL
MATH	1.00	0.98	0.23	0.51
PHYS	0.98	1.00	0.40	0.65
FRAN	0.23	0.40	1.00	0.95
ANGL	0.51	0.65	0.95	1.00

	MATH	PHYS	FRAN	ANGL
MATH	11.39	9.92	2.66	4.82
PHYS	9.92	8.94	4.12	5.48
FRAN	2.66	4.12	12.06	9.29
ANGL	4.82	5.48	9.29	7.91

Dispersion totale des individus =

$11.39 + 8.94 + 12.06 + 7.91 = 40.30$

Valeurs propres et variances expliquées

Calculons les valeurs propres de la matrice de covariance.

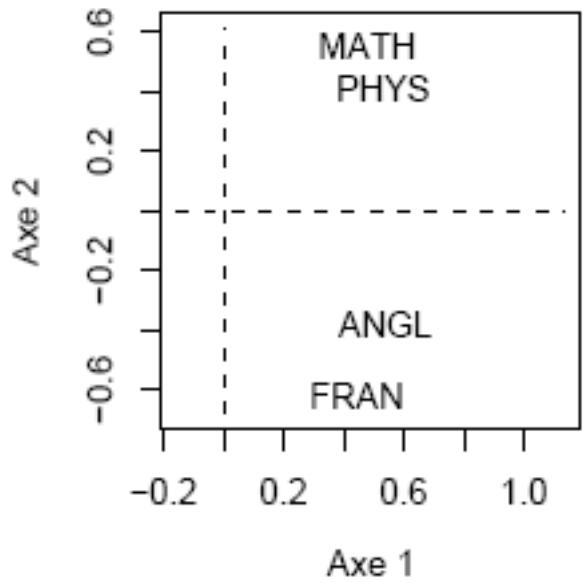
Composante	λ_i	% variance	% cumulé
1	28.23	0.70	0.70
2	12.03	0.30	1.00
3	0.03	0.00	1.00
4	0.01	0.00	1.00
Total	40.30	1.00	-

- somme des valeurs propres $= 40.30 =$ dispersion totale
- les composantes 1 et 2 représententent la quasi totalité de la dispersion, les composantes 3 et 4 sont négligeables
- $\Rightarrow$ Ici 2 composantes suffisent à expliquer les données

Comment les variables initiales et les composantes sont elles liées ?

Corrélations variables-composantes

Composantes	c_1	c_2
MATH	0.81	-0.58
PHYS	0.90	-0.43
FRAN	0.75	0.66
ANGL	0.91	0.40



- c_1 est corrélée positivement et assez fortement, avec chacune des 4 variables initiales
- c_2 oppose le français et l'anglais (corrélations positives) et les mathématiques et la physique (corrélations négatives)

Résultats sur les individus

	POIDS	c_1	c_2
jean	1/9	-8.61	-1.41
alan	1/9	-3.88	-0.50
anni	1/9	-3.21	3.47
moni	1/9	9.85	0.60
didi	1/9	6.41	-2.05
andr	1/9	-3.03	-4.92
pier	1/9	-1.03	6.38
brig	1/9	1.95	-4.20
evel	1/9	1.55	2.63

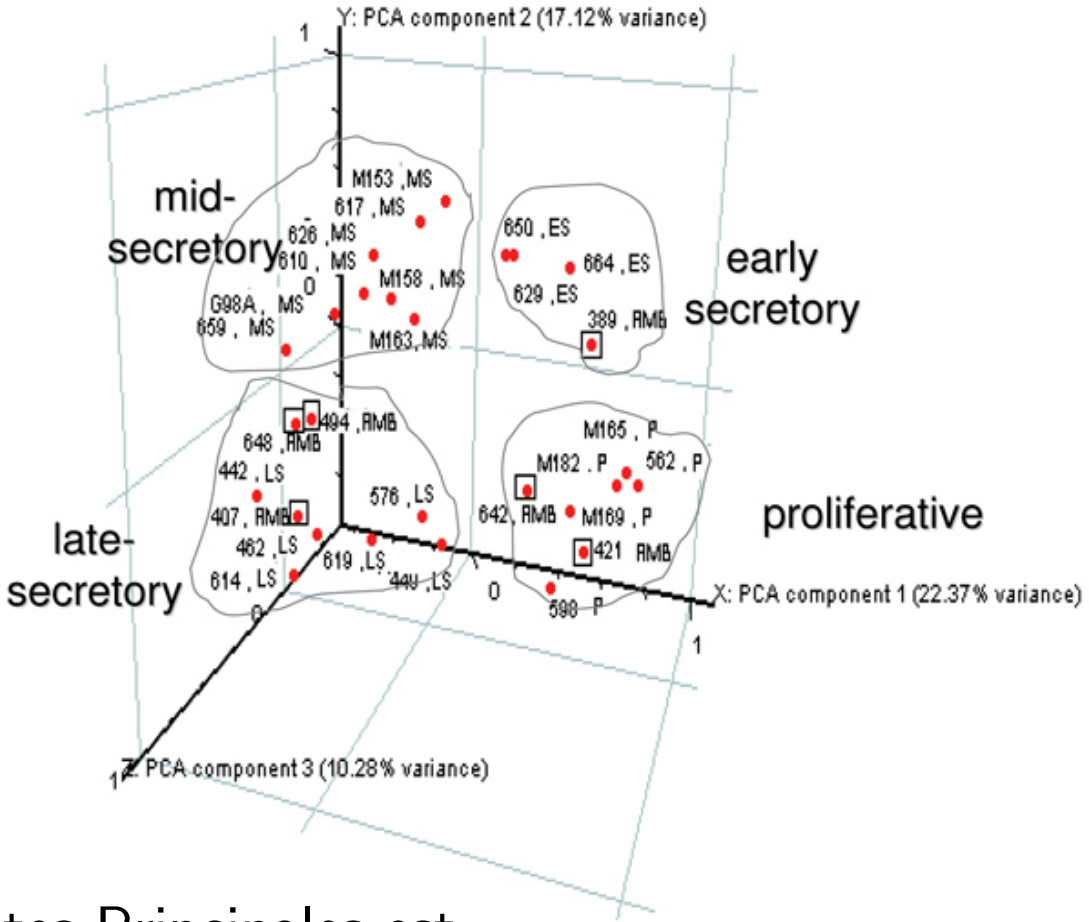
- c_1 = résultat d'ensemble des élèves
- c_2 = constraste scientifique/littéraire
- On peut représenter les individus dans le plan (c_1, c_2)

Résultats sur les individus

	POIDS	$CONT_G$	$CONT_1$	$CONT_2$	COS_1^2	COS_2^2
jean	1/9	20.99	29.19	1.83	0.97	0.03
alan	1/9	4.22	5.92	0.23	0.98	0.02
anni	1/9	6.17	4.06	11.11	0.46	0.54
moni	1/9	26.86	38.19	0.33	1.00	0.00
didi	1/9	12.48	16.15	3.87	0.91	0.09
andr	1/9	9.22	3.62	22.37	0.28	0.72
pier	1/9	11.51	0.41	37.56	0.03	0.97
brig	1/9	5.93	1.50	16.29	0.18	0.82
evel	1/9	2.63	0.95	6.41	0.25	0.73

- $CONT_G$ = contribution du point à la dispersion totale (en %)
- $CONT_i$ = contribution du point à la dispersion en c_i
- COS_i^2 = qualité de la représentation de l'individu par c_i = (cos^2 de l') angle entre le vecteur initial et sa projection en c_i

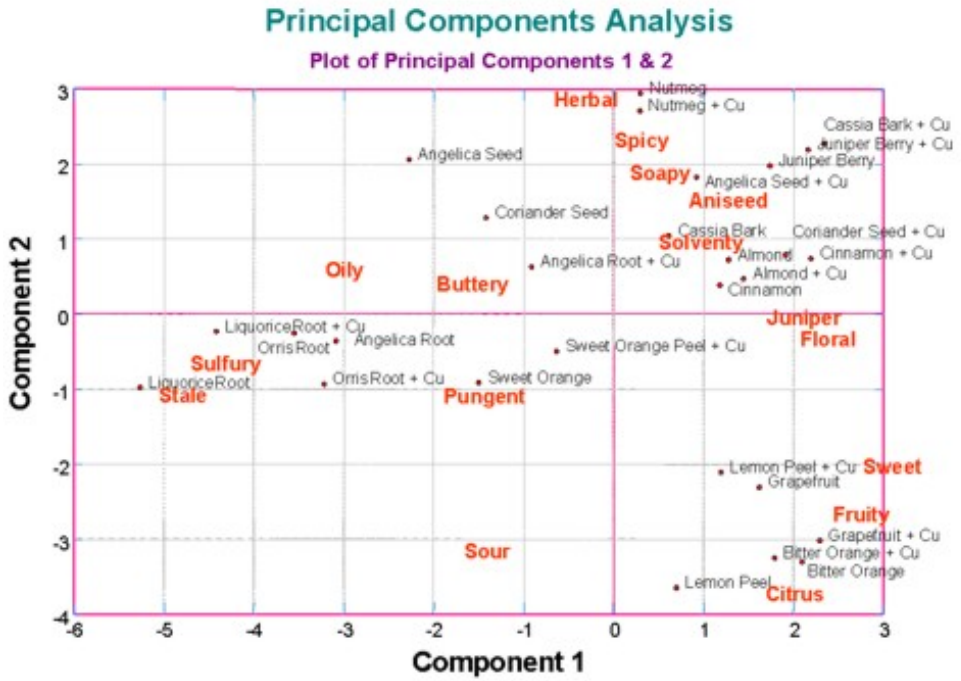
En résumé



L'Analyse en Composantes Principales est

- une méthode de réduction de dimension
- une méthode de visualisation de données

En résumé



Une composante principale

- vecteur propre de la matrice de covariance
- (valeur propre = importance relative de la composante)
- combinaison linéaire des variables initiales
- orthogonale aux autres composantes