

Université Toulouse 1 Capitole Ecole d'économie de Toulouse

Année universitaire 2015–2016

Session 1

Semestre 5

Licence 3 mention Économie

Epreuve : Analyse des données

Date de l'épreuve : vendredi 8 janvier 2016

Durée de l'épreuve : 1h30

Liste des documents autorisés : aucun.

Liste des matériels autorisés : calculatrice non programmable.

Nombre de pages : 11

1 Questions de cours (3 points)

1. A quel type de données est adaptée une analyse en composantes principales (ACP) ?
2. Combien peut-on calculer de composantes principales au maximum ?
3. Donner le nom de la fonction de R permettant de réaliser une ACP.
4. Donner la formule de la contribution de la variable x^k à la composante c^j avec les notations du cours.
5. Donner le nom de la procédure SAS permettant de réaliser une classification par la méthode d'agrégation au tour des moyennes mobiles.
6. Quel est le nombre maximum de nouvelles variables pouvant être calculés dans une AFC ?

2 Analyse en composantes principales (ACP) (7 points)

Les données proviennent du Groupe d'Étude et de Réflexion Inter-régional (GERI) et concernent 50 départements français. Les 8 variables considérées sont les suivantes :

- URBR : indicateur (nombre entre 0 et 1) de concentration de la population mesurant le caractère urbain ou rural d'un département,
- CHOM : taux de chômage (nombre entre 0 et 1),
- Parts (nombres entre 0 et 1) de chaque profession et catégorie socioprofessionnelle dans la population active occupée du département :
 - AGRI : agriculteurs,
 - ARTI : artisans,
 - CADR : cadres supérieurs,
 - EMPL : employés,
 - OUVR : ouvriers,
 - PROF : professions intermédiaires,

On réalise une ACP sur ces 8 variables avec **R** (package FactoMineR). Utiliser les sorties ci-après pour répondre aux questions suivantes.

1. Calculer les 2 valeurs A et B qui ont été supprimées dans les sorties et préciser à quoi correspondent ces valeurs.
2. Donner, en le justifiant, le nombre de composantes principales à retenir pour cette ACP.
3. Interpréter les composantes principales retenues à l'aide des variables initiales.
4. Commenter les départements Ge (Gers) et Pa (Paris) sur le 1er axe, les départements Va (Var) et Db (Doubs) sur le 2ème axe et le département PC (Pas de Calais) sur le 3ème axe après avoir vérifié que ces départements sont bien représentés sur les axes considérés.

resu\$eig

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	4.145	51.807	51.807
comp 2	1.645	20.562	72.369
comp 3	1.148	14.349	86.718
comp 4	0.467	5.841	92.559
comp 5	0.310	3.879	96.437
comp 6	0.169	2.119	98.556
comp 7	0.115	1.443	99.999
comp 8	0.000	0.001	100.000

resu\$var\$cor

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
urbr	0.927	-0.086	0.039	0.154	0.067
chom	0.110	0.330	0.882	0.292	-0.109
agri	-0.808	0.411	-0.262	0.065	-0.268
arti	-0.409	0.820	0.057	-0.058	0.390
cadr	0.842	0.120	-0.340	0.344	-0.012
empl	0.752	0.340	0.231	-0.460	-0.196
ouvr	-0.544	-0.745	0.356	-0.062	0.124
prof	0.941	-0.055	-0.016	-0.131	0.126

resu\$var\$contrib

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
urbr	20.714	0.449	0.135	5.080	1.448
chom	0.290	6.610	67.817	18.210	3.840
agri	15.743	10.269	5.967	0.913	23.166
arti	4.031	40.861	0.285	0.729	48.989
cadr	17.087	0.875	10.078	25.377	0.050
empl	13.649	7.036	4.635	45.186	12.428
ouvr	7.139	33.718	11.062	0.813	4.974
prof	21.347	0.181	0.022	3.692	5.105

resu\$ind\$coord

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Ge	-3.101	1.807	-2.082	0.658	-1.329
Pa	5.426	0.902	-2.823	2.896	0.155
PC	0.411	-1.328	2.024	0.607	-0.270
Va	2.034	2.763	1.999	-0.447	1.077
Db	0.039	-2.319	-0.471	0.263	0.476

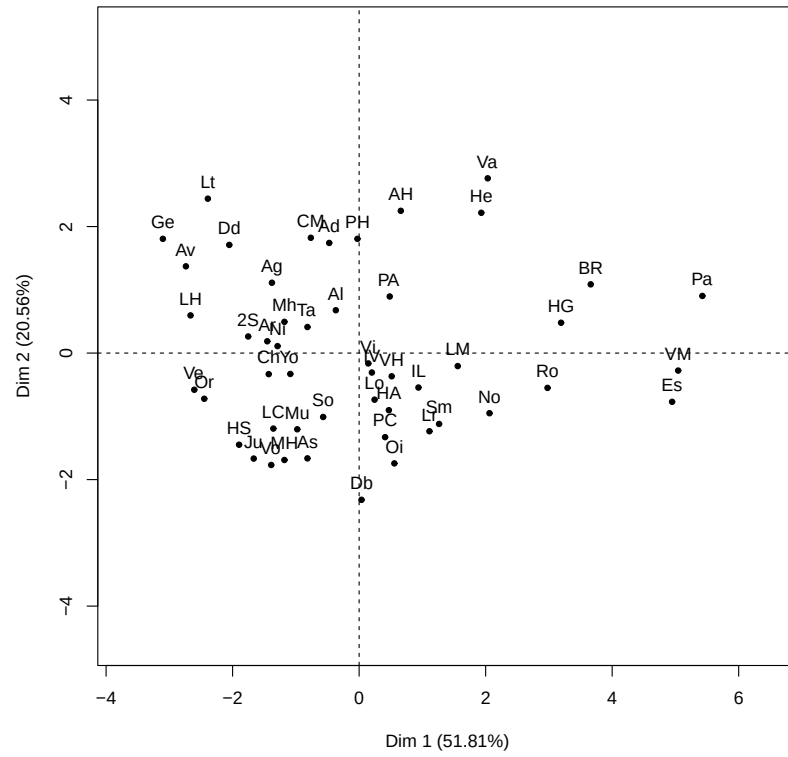
resu\$ind\$cos2

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Ge	0.482	0.164	0.217	0.022	0.089
Pa	0.608	0.017	A	0.173	0.000
PC	0.024	0.254	0.590	0.053	0.011
Va	0.239	0.441	0.231	0.012	0.067
Db	0.000	0.870	0.036	0.011	0.037

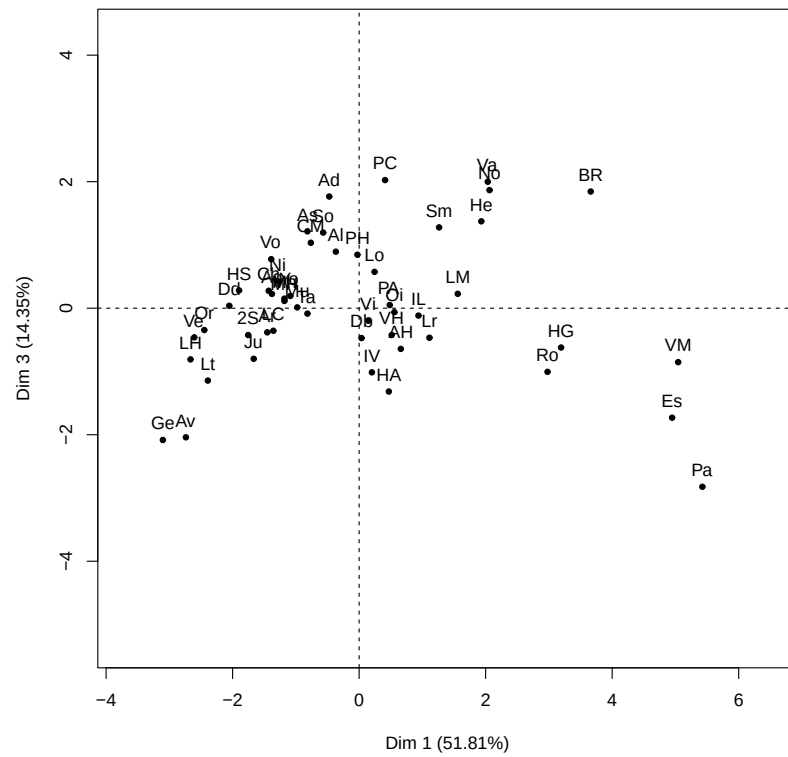
resu\$ind\$dist

	Db	Ge	PC	Pa	Va
	2.485	B	2.635	6.960	4.163

Individuals factor map (PCA)



Individuals factor map (PCA)



3 Classification (4 points)

On réalise une classification en 4 groupes des 50 départements à l'aide des 4 variables URBR, CHOM, AGRI et CADR (non standardisées) par la méthode d'agrégation autour des moyennes mobiles (AMM). Utiliser les sorties **SAS** ci-après pour répondre aux questions suivantes.

1. Calculer la distance entre Al (Allier) et Ar (Ardèche).
2. Déterminer les 2 valeurs C et D qui ont été supprimées dans les sorties et préciser à quoi correspondent ces valeurs.
3. Donner une typologie succincte des départements en fonction des 4 variables initiales.

4 Analyse factorielle des correspondances (AFC)(6 points)

Un chef d'entreprise souhaite faire une étude détaillée des ventes de ses produits (des biscuits) en étudiant la liaison entre le canal de distribution (variable qualitative à 5 modalités : hypermarché, supermarché, supérettes, épicerie traditionnelles et distributeurs automatiques) et la localisation divisée en 8 secteurs codés de 1 (secteur le plus urbanisé) à 8 (secteur le moins urbanisé) sur un échantillon de 624 produits vendus. Utiliser les sorties **SAS** ci-après pour répondre aux questions suivantes.

1. L'AFC est-elle justifiée ? Pourquoi ?
On donne $\chi^2_{5\%}(40) = 55,758$ et $\chi^2_{5\%}(28) = 45,419$.
2. Calculer les valeurs E et F qui ont été supprimées et préciser de façon succincte à quoi correspondent ces valeurs.
3. On souhaite 90% de l'inertie expliquée. Combien d'axes sont à retenir ?
4. Quels sont les profils ayant fortement contribué à l'apparition des axes retenus ? Justifier votre réponse.
5. Quels sont les profils bien représentés ? Justifier votre réponse.
6. Commenter l'AFC.

Extrait des
données

Obs.	depart	agri	cadr	chom	urbr
1	As	0.049	0.071	0.133	0.358
2	Al	0.086	0.07	0.135	0.464
3	AH	0.072	0.078	0.089	0.294
4	Ar	0.085	0.073	0.105	0.281

Initial Seeds				
Cluster	agri	cadr	chom	urbr
1	0.0140000000	0.1350000000	0.1690000000	0.8690000000
2	0.0860000000	0.0700000000	0.1350000000	0.4640000000
3	0.0900000000	0.0730000000	0.1200000000	0.0000000000
4	0.1780000000	0.0670000000	0.0790000000	0.2300000000

Minimum Distance Between Initial Seeds =	0.249722
--	----------

Synthèse de la classe						
Classe	Fréquence	RMS Std Deviation	Distance max. de la valeur initiale à l'observation	Rayon dépassé	Classe la plus proche	Distance Between Cluster Centroids
1	6	0.0638	0.1856		2	D
2	14	0.0482	0.1407		4	0.2642
3	9	0.0428	0.1642		C	0.1693
4	21	0.0331	0.1169		3	0.1693

Statistiques pour variables				
Variable	E-t total	Dans E-t	R-carré	RSQ/(1-RSQ)
agri	0.04573	0.02707	0.670928	2.038849
cadr	0.04341	0.02955	0.565121	1.299491
chom	0.02523	0.02525	0.059926	0.063746
urbr	0.23776	0.07303	0.911445	10.292376
OVER-ALL	0.12364	0.04352	0.883681	7.597044

Cluster Means				
Cluster	agri	cadr	chom	urbr
1	0.0098333333	0.1726666667	0.1126666667	0.8910000000
2	0.0460000000	0.0985000000	0.1196428571	0.5762857143
3	0.1348888889	0.0650000000	0.1014444444	0.1566666667
4	0.0772380952	0.0750952381	0.1146666667	0.3150000000

Cluster Standard Deviations				
Cluster	agri	cadr	chom	urbr
1	0.0093041210	0.0770419799	0.0397223699	0.0932673576
2	0.0195290710	0.0240120162	0.0290294470	0.0865878199
3	0.0473852415	0.0083666003	0.0118860329	0.0697405908
4	0.0227659939	0.0110131955	0.0216132675	0.0571987762

Distance Between Cluster Centroids				
Nearest Cluster	1	2	3	4
1	.	0.3254265986	0.7527299554	0.5880846755
2	0.3254265986	.	0.4306214253	0.2642320731
3	0.7527299554	0.4306214253	.	0.1693215859
4	0.5880846755	0.2642320731	0.1693215859	.

Variable	N	Moyenne	Ecart-type	Minimum	Maximum
agri	50	0.0707800	0.0457296	0	0.2220000
cadr	50	0.0915400	0.0434117	0.0520000	0.3210000
chom	50	0.1134400	0.0252281	0.0630000	0.1690000
urbr	50	0.4287800	0.2377649	0	1.0010000

Table de canal par secteur									
canal	secteur								
Fréquence Pourcentage Pctage en ligne Pctage en col.	1	2	3	4	5	6	7	8	Total
hypermarché	17 2.72 10.69 22.67	32 5.13 <u>E</u> 42.11	14 2.24 8.81 17.07	35 5.61 22.01 42.68	20 3.21 12.58 26.32	23 3.69 14.47 27.71	18 2.88 11.32 23.68	0 0.00 0.00 0.00	159 25.48
supermarché	15 2.40 11.63 20.00	7 1.12 5.43 9.21	23 3.69 17.83 28.05	10 1.60 7.75 12.20	19 3.04 14.73 25.00	15 2.40 11.63 18.07	22 3.53 17.05 28.95	18 2.88 13.95 24.32	129 20.67
supérette	10 1.60 11.63 13.33	13 2.08 15.12 17.11	6 0.96 6.98 7.32	5 0.80 5.81 6.10	11 1.76 12.79 14.47	12 1.92 13.95 14.46	6 0.96 6.98 7.89	23 3.69 26.74 <u>E</u>	86 13.78
épicerie	9 1.44 18.75 12.00	4 0.64 8.33 5.26	5 0.80 10.42 6.10	3 0.48 6.25 3.66	4 0.64 8.33 5.26	4 0.64 8.33 4.82	5 0.80 10.42 6.58	14 2.24 29.17 18.92	48 7.69
distributeur	24 3.85 11.88 32.00	20 3.21 9.90 26.32	34 5.45 16.83 41.46	29 4.65 14.36 35.37	22 3.53 10.89 28.95	29 4.65 14.36 34.94	25 4.01 12.38 32.89	19 3.04 9.41 25.68	202 32.37
Total	75 12.02	76 12.18	82 13.14	82 13.14	76 12.18	83 13.30	76 12.18	74 11.86	624 100.00

Décomposition de l'inertie et du Khi-2					
Valeur singulière	Inertie principale	Khi-2	Pourcentage	Pourcent. cumulé	14 28 42 56 70 -----+-----+-----+-----+-----
0.33766	0.11401	71.143	70.68	70.68	*****
0.19691	0.03877	24.195	24.04	94.72	*****
0.07332	0.00538	3.354	3.33	98.05	*
0.05603	0.00314	1.959	1.95	100.00	*
Total	0.16130	100.651	100.00		
Degrés de liberté = 28					

Sorties SAS : AFC

Coordonnées des lignes		
	Dim1	Dim2
distributeur	-0.0530	-0.1176
hypermarché	-0.4765	0.1392
supermarché	0.1834	-0.2550
supérette	0.4155	0.3237
épicerie	0.5643	0.1388

Contributions partielles à l'inertie des points des lignes		
	Dim1	Dim2
distributeur	0.0080	0.1154
hypermarché	0.5075	0.1274
supermarché	0.0610	0.3466
supérette	0.2087	0.3724
épicerie	0.2148	0.0382

Carré des cosinus pour les points des lignes		
	Dim1	Dim2
distributeur	0.1219	0.5989
hypermarché	0.9162	0.0782
supermarché	0.3144	0.6073
supérette	0.6022	0.3655
épicerie	0.8435	0.0510

Coordonnées des colonnes		
	Dim1	Dim2
1	0.1031	0.0140
2	-0.2871	0.3396
3	0.0382	-0.3267
4	-0.4555	0.0588
5	-0.0150	-0.0354
6	-0.0893	0.0250
7	-0.0216	-0.2276
8	0.7904	0.1760

Contributions partielles à l'inertie des points des colonnes		
	Dim1	Dim2
1	0.0112	0.0006
2	0.0880	0.3623
3	0.0017	0.3618
4	0.2391	0.0117
5	0.0002	0.0039
6	0.0093	0.0021
7	0.0005	0.1627
8	0.6499	0.0947

Carré des cosinus pour les points des colonnes		
	Dim1	Dim2
1	0.3846	0.0071
2	0.4141	0.5796
3	0.0129	0.9429
4	0.9532	0.0159
5	0.0107	0.0599
6	0.4355	0.0340
7	0.0076	0.8459
8	0.9523	0.0472

