Université Paris II
L2 Sciences Économiques et Gestion
Parcours analyse économique

Analyse des données (4011)
Cours de M^{me} Chevalier

Mai-juin 2016

Aucun document n'est admis. Les calculatrices sont autorisées.
Tous les calculs doivent être justifiés.

Exercice 1

On considère des informations rassemblées sur le personnel d'une entreprise, en particulier le nombre d'années d'expérience à l'embauche et le nombre d'années d'étude. Pour simplifier, on suppose que chaque catégorie de personnel représente le même poids dans l'entreprise. Les moyennes par type d'emploi sont rassemblées dans le tableau ci-dessous :

Type de personnel	Ancienneté	Nb années d'étude
Cadre supérieur	7	7
Cadre	5	5
Ingénieur	4	6
Personnel administratif	4	3
Administrateur système	5	4

1. Déterminer le centre de gravité du nuage et la matrice des données du nuage centré.
2. Calculer la matrice d'inertie ainsi que l'inertie totale du nuage (sans calcul supplémentaire). On considérera $Q = I_2$.
3. Donner la part d'inertie expliquée par chacune des composantes principales.
4. Quel est le lien entre l'inertie du nuage et les valeurs propres de la matrice d'inertie ?

Exercice 2

On considère 6 observations A, B, C, D, E et F décrites par 2 variables quantitatives appelées x et y. Les résultats sont donnés dans le tableau suivant :

	x	y
A	4	1
B	-3	3
C	-2	1
D	3	4
E	1	3
F	2	2

1. Représenter ces observations sur un dessin.
2. Classifier ces observations en deux groupes en utilisant la méthode des centres mobiles à partir des centres initiaux D et E.
3. En utilisant la distance euclidienne usuelle, établir le tableau des distances (au carré) entre ces observations.
4. Rappeler la différence entre les méthodes utilisant le critère du diamètre, du saut minimum et de la moyenne. Quel sont les avantages de la méthode de Ward par rapport à ces dernières ?
5. En utilisant le critère du diamètre, effectuer une classification hiérarchique ascendante de ces 6 observations. Donner l'arbre hiérarchique et suggérer une coupure possible.
6. Effectuer une nouvelle classification en utilisant le critère de Ward. Donner l'arbre hiérarchique et suggérer une coupure possible. Quel pourcentage d'inertie la partition proposée explique-t-elle ?

Exercice 3

On considère un sondage comprenant les 3 questions suivantes :

- Quel sport pratiquez-vous ? (Basket, Course, Autre)
- Combien de temps par semaine en moyenne ? (Moins de 2h, Entre 2 et 4h, Plus de 4h)
- Êtes-vous intéressé par un stage sportif cet été ? (Oui/Non)

Une seule réponse est acceptée par question. Voici les résultats de ce sondage sur 5 personnes :

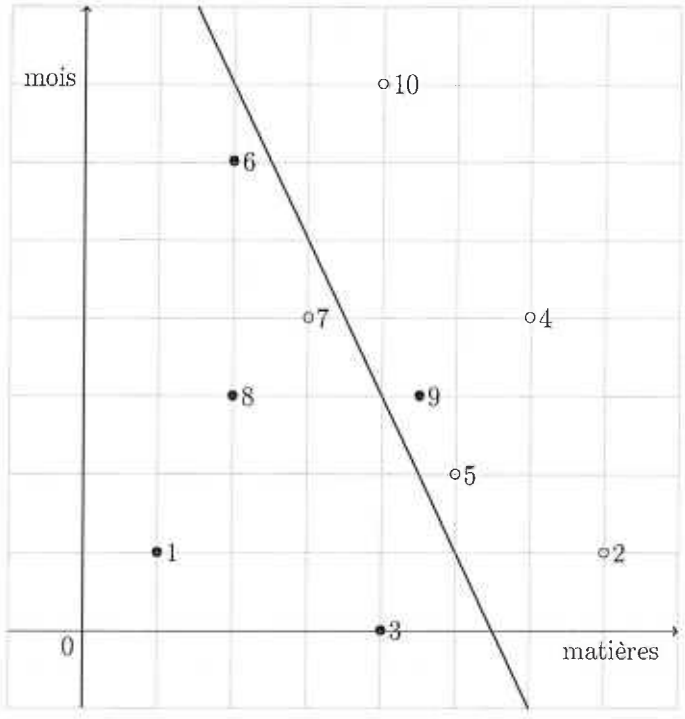
A	Basket	Entre 2 et 4h	Oui
B	Course	Plus de 4h	Oui
C	Autre	Entre 2 et 4h	Non
D	Course	Moins de 2h	Oui
E	Basket	Moins de 2h	Non

1. En modélisant ces résultats par trois variables qualitatives numériques (représentant les résultats aux trois questions), donner ces résultats sous la forme d'un tableau de données chiffrées à 5 lignes et 3 colonnes.
2. Donner le tableau disjonctif complet correspondant.
3. Écrire le tableau de Burt associé à ces données.
4. Expliquer soigneusement comment déterminer à partir du tableau de Burt :
   - le nombre de personnes qui courent plus de 4h par semaine.
   - le nombre de personnes qui font du basket et sont intéressées par un stage cet été ;
   - le nombre de personnes qui souhaitent s'inscrire à un stage cet été.

   Vérifier ces valeurs sur les données initiales.

Exercice 4

1. Expliquer brièvement le principe de l'analyse discriminante, en particulier ses deux grands types : descriptif et décisionnel.
2. Un organisme indépendant a conduit une étude sur des étudiants, afin de savoir s'ils avaient obtenu ou non le stage convoité pendant leurs études, en fonction du nombre de matières validées avec plus de 14 et du nombre de mois de stages en entreprise déjà effectués au cours de leur scolarité. On a réalisé une analyse discriminante sur ces données et les résultats (restreints à 10 étudiants) sont représentés graphiquement ci-dessous (un • signifie que le stage n'a pas été obtenu, et un ∘ qu'il l'a été). La droite dessinée a pour équation $y = -2x + 11$.



(a) Un étudiant ayant validé 5 matières avec plus de 14 et effectué 3 mois de stages s'interroge sur ses chances d'embauche dans le stage qu'il souhaiterait. Que prédit le modèle ?

(b) Même question pour un étudiant ayant cumulé 5 mois de stage, mais ayant eu seulement 2 notes avec plus de 14.

(c) Que peut-on dire au sujet du modèle sur les étudiants numéros 7 et 9 ?

Exercice 5

Un institut de sondage a recueilli les données suivantes pour étudier la relation entre la catégorie socio-professionnelle (CSP) d'une personne (agriculteur : AGRI, cadre supérieur : CSUP, cadre moyen : CMOY, employé : EMPL, ouvrier : OUVR, retraité : RETR et chômeur : CHOM) et sa principale source d'information concernant les problèmes environnementaux (télévision : TEL, journaux : JOU, radio : RAD, livres : LIV, associations : ASS et mairie : MAI).

CSP	TEL	JOU	RAD	LIV	ASS	MAI	Total
AGRI	26	18	9	5	4	6	68
CSUP	19	49	4	16	5	3	96
CMOY	44	87	4	39	14	3	191
EMPL	83	87	13	24	5	1	213
OUVR	181	107	16	31	7	7	349
RETR	167	95	29	15	7	7	320
CHOM	27	9	4	2	2	2	46
Total	547	452	79	132	44	29	1283

1. Justifier et commenter l'analyse ci-jointe produite sur ces données. Vous préciserez en particulier le nombre d'axes principaux choisis et les raisons qui vont ont conduit à faire ce choix.
2. En notant n le nombre d'invidus et p et le nombre de variables, on a vu que l'analyse par composantes principales déterminait tout d'abord les axes principaux (dirigés par les vecteurs propres notés U_i, dans $\mathbb{R}^p$), puis les facteurs principaux (notés C_i, dans $\mathbb{R}^n$), puis les composantes principales (notées D_i, dans $\mathbb{R}^p$).

   (a) Que représentent concrètement ces différents vecteurs ?

   (b) L'analyse ci-jointe effectuée par le logiciel SAS procure 4 tableaux : la matrice de corrélation (*correlation matrix*), les valeurs propres (*eigenvalues*), les vecteurs propres (*eigenvectors*) et enfin un dernier (*the SAS system*). Deux d'entre eux correspondent aux familles de vecteurs rappelés ci-dessus, lesquels ?

   (c) Expliquer comment obtenir la troisième famille de vecteurs en utilisant les données jointes et calculer explicitement les deux premiers vecteurs de cette famille.

   (d) Comment calculer la qualité de la représentation des individus sur les axes principaux ? La calculer pour le premier individu.

   (e) Comment calculer la qualité de la représentation des variables sur les axes principaux ? La calculer pour la première variable.

The SAS System

The PRINCOMP Procedure

Observations	7
Variables	6

Simple Statistics

	TEL	JOU	RAD	LIV	ASS	MAI
Mean	78.14285714	64.57142857	11.28571429	18.85714286	6.285714286	4.142857143
StD	68.91644354	39.22523483	9.15995425	13.40930736	3.817254062	2.478478796

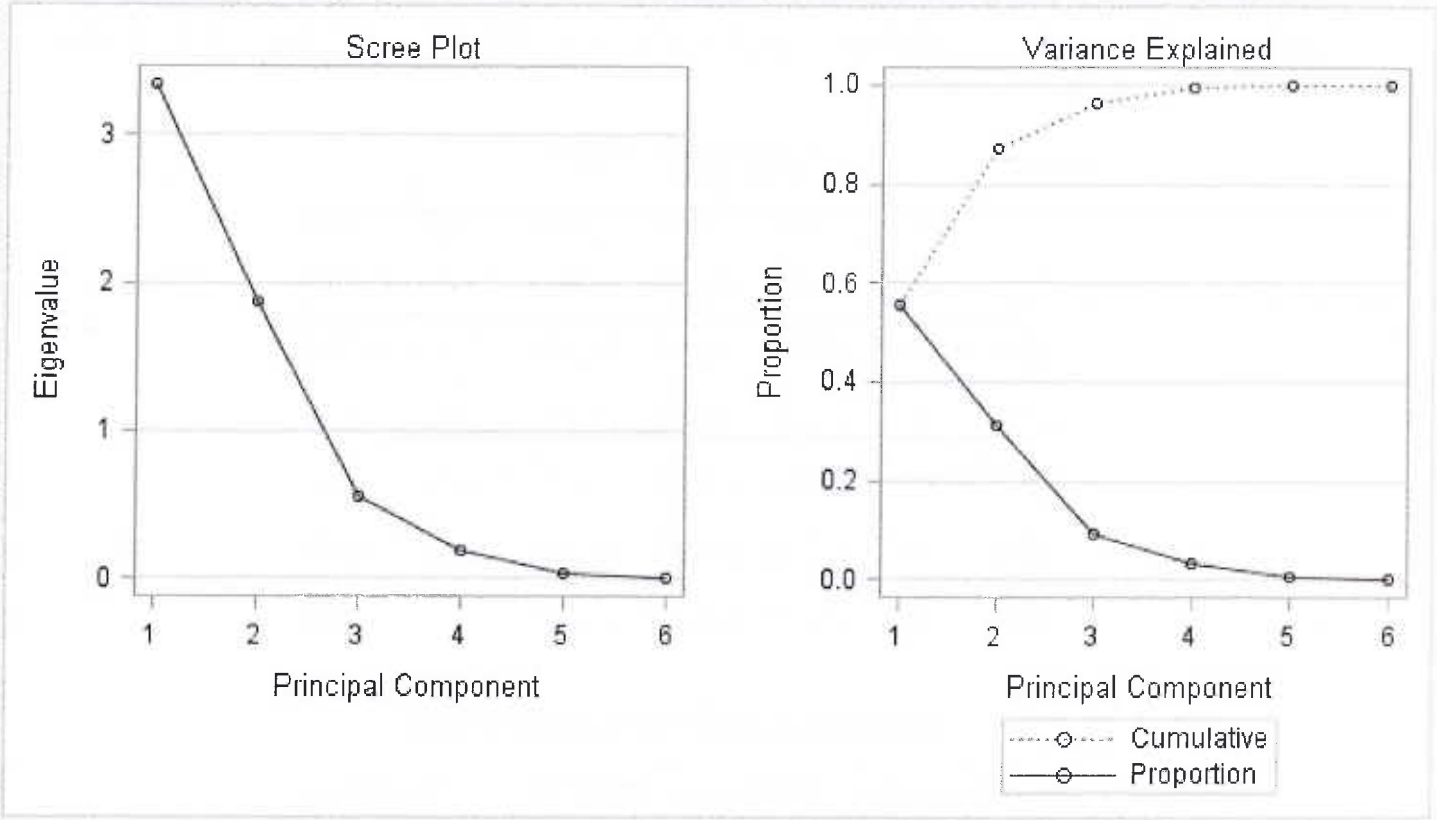
Correlation Matrix

	TEL	JOU	RAD	LIV	ASS	MAI
TEL	1.0000	0.7770	0.8551	0.3602	0.1785	0.6361
JOU	0.7770	1.0000	0.5747	0.8212	0.6243	0.2785
RAD	0.8551	0.5747	1.0000	0.0357	0.0163	0.6366
LIV	0.3602	0.8212	0.0357	1.0000	0.8540	-.0093
ASS	0.1785	0.6243	0.0163	0.8540	1.0000	0.1007
MAI	0.6361	0.2785	0.6366	-.0093	0.1007	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	3.34296696	1.46627345	0.5572	0.5572
2	1.87669351	1.32525447	0.3128	0.8699
3	0.55143904	0.36283621	0.0919	0.9618
4	0.18860283	0.15160794	0.0314	0.9933
5	0.03699489	0.03369212	0.0062	0.9994
6	0.00330277		0.0006	1.0000

Eigenvectors						
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6
TEL	0.482196	-.281956	-.240531	-.348144	-.711974	-.044962
JOU	0.513715	0.163761	-.321066	-.079799	0.469458	-.615985
RAD	0.394737	-.444602	-.229573	0.640353	0.183380	0.387463
LIV	0.378392	0.513874	-.035476	-.343804	0.191132	0.660880
ASS	0.327685	0.502925	0.446734	0.520516	-.371542	-.176457
MAI	0.311784	-.423098	0.765198	-.270624	0.253967	-.022688



The SAS System

Obs	CSP	TEL	JOU	RAD	LIV	ASS	MAI	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6
1	CHOM	27	9	4	2	2	2	-1.37430	-0.37519	-0.40747	-0.12925	-1.69132	-0.26473
2	AGRI	26	18	9	5	4	6	-0.78041	-0.74411	1.29682	0.09025	0.78406	1.29441
3	CSUP	19	49	4	16	5	3	-0.69268	0.32615	0.02791	-0.35957	1.27946	-1.67021
4	EMPL	83	87	13	24	5	1	0.02242	0.40496	-1.85552	0.19726	0.67054	1.02317
5	CMOY	44	87	4	39	14	3	0.45271	1.87626	0.82799	0.63961	-0.54874	0.14761
6	RETR	167	95	29	15	7	7	1.14609	-1.19551	-0.03670	1.40901	-0.16066	-0.62260
7	OUVR	181	107	16	31	7	7	1.22616	-0.29256	0.14698	-1.84731	-0.33334	0.09234