Master 1 MIM : Université de Nantes **Mai 2008**

Examen d'Analyse des données

durée 3 heures – documents interdits– calculatrice autorisée

Exercice I : DVS (4 points)

1. Rappeler le théorème sur la DVS
2. Effectuer la décomposition en valeurs singulières de A= $\begin{pmatrix} 1 & -1 \\ 1 & 1 \\ 0 & 2 \end{pmatrix}$.
3. En déduire la meilleure approximation de rang 1 de A.

Exercice II : Analyse factorielle discriminante dans le cas de deux classes (8 points)

On considère ici un tableau X décrivant pour n individus p variables quantitatives centrées. Les n individus sont répartis dans deux classes d'effectifs respectifs n_1 et n_2.
L'objectif de cet exercice est de réaliser l'analyse discriminante de ce tableau.

Partie A : Etude théorique

1. Rappeler brièvement l'objectif (et seulement l'objectif) de l'analyse discriminante (maximum 5 lignes)
2. Combien d'axes discriminants peut-on obtenir ici ? Justifier la réponse.
3. On note x_j^k la moyenne de la variable j pour la classe k. On note $B = (b_{jj'})$ la somme des carrés inter-classes.
   a) Exprimer les centres de gravité g_1 et g_2 (vecteurs colonnes) en fonction des x_j^k.
   b) Exprimer B en fonction de n_1, n_2, g_1, g_2..
   c) Justifiez que $n_1 \times g_1 + n_2 \times g_2 = 0$. En déduire g_2 en fonction de g_1.
   d) Montrer que $B = \frac{n_1 \times n_2}{n} (g_1 - g_2)(g_1 - g_2)'$.
   e) On pose $C = \sqrt{(n_1 n_2)/(n)}\ (g_1 - g_2)$. Exprimer B en fonction de C.

4. On note W la matrice des somme des carrés intra classes. On définit également $\widehat{W} = \frac{1}{n-2} W$, $\widehat{W}^{-1}$ définissant la métrique de Mahalanobis (1893-1972).

   a) Soit u une fonction discriminante. Justifiez que u vérifie $W^{-1} Bu = \lambda u$, avec λ un réel strictement positif.

   b) Montrer que $W^{-1} C$ est vecteur propre de $W^{-1} B$ et en déduire que :

   $$\lambda = {}^t C W^{-1} C = \frac{n_1 n_2}{n} {}^t(g_1 - g_2) \; W^{-1} (g_1 - g_2)$$

   c) Déterminer l'estimation de la distance au carré de Mahalanobis D_p^2 entre g_1 et g_2. En déduire que :

   $$D_p^2 = \frac{n(n-2)}{n_1 n_2} \lambda$$

   d) Montrer que $u = \widehat{W}^{-1} (g_1 - g_2)$. Cette fonction s'appelle la fonction discriminante de Fisher (1890-1962).

   e) Comment s'interpréte cette fonction en terme de projection orthogonale sur un axe et avec une métrique à préciser.

   f) La démarche de Fisher a conduit à chercher u tel que u rende maximal le rapport :

   $$\frac{{}^t g_1 u - {}^t g_2 u}{u' \widehat{W}^{-1} u}$$

   Interpréter ce rapport.

Partie B : Application

On considère deux groupes d'individus carctérisés par deux variables quantitatives X et Y.

Individu	X	Y	classe
1	-6	4	1
2	-2	1	1
3	2	-2	1
4	-2	2	2
5	2	-4	2
6	6	-1	2

1. Calculer g_1, g_2, B et W.
2. En déduire $W^{-1} = \frac{1}{880} \begin{pmatrix} 34 & 32 \\ 32 & 56 \end{pmatrix}$
3. En déduire directement l'estimation de la distance de Mahalanobis entre les deux centres de gravités.
4. En déduire la fonction discriminante de Fisher.
5. Calculer la variable discriminante et en déduire le taux d'erreur de classement.

Exercice III : Analyse de résultats (8 points)

Les touristes étrangers aiment bien venir en Aquitaine : ils ont passé 1,2 millions de nuits d'hôtel dans cette province en 2002. Vous me direz, ce n'est pas beaucoup car pour la même période, on enregistre 3,8 millions de nuits d'hôtel pour les français en Aquitaine. Le tableau suivant donne la répartition de ces nuitées selon un regroupement de pays (en milliers de nuits d'hôtel).

Tableau I

	Dordogne	Gironde	Landes	Lot-et-Garonne	Pyrénées-Atlantiques	Total Aquitaine
Grande-Bretagne-Irlande	74,0	93,6	25,8	11,7	102,7	307,8
Espagne-Portugal	13,8	67,6	13,8	3,7	51,2	150,1
Allemagne	23,8	59,2	20,2	4,2	39,4	146,8
Belgique-Luxembourg	42,8	27,8	13,1	5,8	31,1	120,5
États-Unis	28,4	30,5	4,1	2,6	28,3	93,8
Pays-Bas	16,3	14,2	13,9	7,4	15,8	67,5
Italie, Grèce	7,8	16,8	5,2	1,7	24,4	55,9
Suisse	7,0	13,2	5,4	1,0	19,7	46,2
Canada.	4,2	4,7	0,7	0,4	3,9	13,9
Autres Europe	4,7	23,8	4,4	1,1	15,1	49,1
Autres Etrangers	10,2	27,6	4,0	2,1	28,7	72,6
Total	232,9	379,0	110,6	41,5	360,3	1124,3

Partie A : Analyse descriptive

1. Comment s'appelle ce tableau I?
2. Proposer deux questions auxquelles pourrait répondre ce tableau.
3. Quel test statistiques peut-on proposer et pour quelle hypothèse.
4. Deux nouveaux tableaux, II et III, se déduisent du précédent (page suivante).
   a) Comment s'appellent ces tableaux et comment sont-ils construits ?
   b) Déterminer le profil ligne le plus proche du profil moyen.

Tableau II

	Dordogne	Gironde	Landes	Lot-et-Garonne	Pyrénées-Atlantiques	Aquitaine
Grande-Bretagne-Irlande	24,0	30,4	8,4	3,8	33,4	100,0
Espagne-Portugal	9,2	45,0	9,2	2,5	34,1	100,0
Allemagne	16,2	40,3	13,8	2,9	26,8	100,0
Belgique-Luxembourg	35,5	23,1	10,9	4,8	25,8	100,0
États-Unis	30,3	32,5	4,4	2,8	30,2	100,0
Pays-Bas	24,1	21,0	20,6	11,0	23,4	100,0
Italie, Grèce	14,0	30,1	9,3	3,0	43,6	100,0
Suisse	15,2	28,6	11,7	2,2	42,6	100,0
Canada.	30,2	33,8	5,0	2,9	28,1	100,0
Autres Europe	9,6	48,5	9,0	2,2	30,8	100,0
Autres	14,1	38,0	5,5	2,9	39,5	100,0
Total	20,7	33,7	9,8	3,7	32,0	100,0

Tableau III

	Dordogne	Gironde	Landes	Lot-et-Garonne	Pyrénées-Atlantiques	Aquitaine
Grande-Bretagne-Irlande	31,8	24,7	23,3	28,2	28,5	27,4
Espagne-Portugal	5,9	17,8	12,5	8,9	14,2	13,4
Allemagne	10,2	15,6	18,3	10,1	10,9	13,1
Belgique-Luxembourg	18,4	7,3	11,8	14,0	8,6	10,7
États-Unis	12,2	8,0	3,7	6,3	7,9	8,3
Pays-Bas	7,0	3,7	12,6	17,8	4,4	6,0
Italie, Grèce	3,3	4,4	4,7	4,1	6,8	5,0
Suisse	3,0	3,5	4,9	2,4	5,5	4,1
Canada.	1,8	1,2	0,6	1,0	1,1	1,2
Autres Europe	2,0	6,3	4,0	2,7	4,2	4,4
Autres	4,4	7,3	3,6	5,1	8,0	6,5
Total	100,0	100,0	100,0	100,0	100,0	100,0

Partie B : Analyse factorielle des correspondances

1. Les valeurs propres sont présentées ci-dessous.
   (a) Justifier le nombre de valeurs propres obtenues.
   (b) Combien d'axes retiendriez-vous ? Justifier votre réponse.

Tableau IV

HISTOGRAMME DES 4 PREMIERES VALEURS PROPRES

NUMERO	VALEUR PROPRE	POURCENT.	POURCENT. CUMULE	
1	0.0522	61.83	61.83	**********
2	0.0215	25.44	87.26	****
3	0.0087	10.29	97.55	**
4	0.0021	2.45	100.00	*

2. Les résultats sur les départements dans le tableau V.
   (a) Définir la qualité de représentation d'un département.
   (b) Décrire la qualité de représentation dans le plan F1-F2.
   (c) Les projections dans F1-F2 sont représentées dans les figures I, II et III.
      i. Dans la figure I, interpréter la répartition des profils
      ii. Dans la figure II, le symbole est proportionnel à la qualité de représentation dans le plan. Quel est l'intérêt de cette représentation ?
      iii. Dans la figure III, le symbole est proportionnel à la contribution dans le plan. Quel est l'intérêt et les limites de cette représentation ?

Tableau V

COORDONNEES, CONTRIBUTIONS DES FREQUENCES SUR LES AXES 1 A 4
FREQUENCES ACTIVES

	FREQUENCES		COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDEN	LIBELLE COURT	P.REL	DISTO	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0
C2	Dordogne	20.72	0.16	-0.37	0.13	-0.06	0.01	0.00	55.4	15.4	7.2	1.3	0.0	0.88	0.10	0.02	0.00	0.00
C3	Gironde	33.70	0.05	0.21	0.00	-0.09	-0.02	0.00	28.5	0.0	33.9	3.9	0.0	0.83	0.00	0.16	0.00	0.00
C4	Landes	9.63	0.14	-0.10	-0.36	0.00	0.08	0.00	1.8	58.4	0.0	29.9	0.0	0.07	0.89	0.00	0.04	0.00
C5	Lot-et-Gar	3.71	0.29	-0.36	-0.33	0.09	-0.19	0.00	9.2	19.3	3.5	64.2	0.0	0.45	0.39	0.03	0.13	0.00
C6	Pyrénées-Atlantiques	32.04	0.03	0.09	0.07	0.12	0.01	0.00	5.2	6.8	55.4	0.7	0.0	0.30	0.16	0.54	0.00	0.00

Figure I

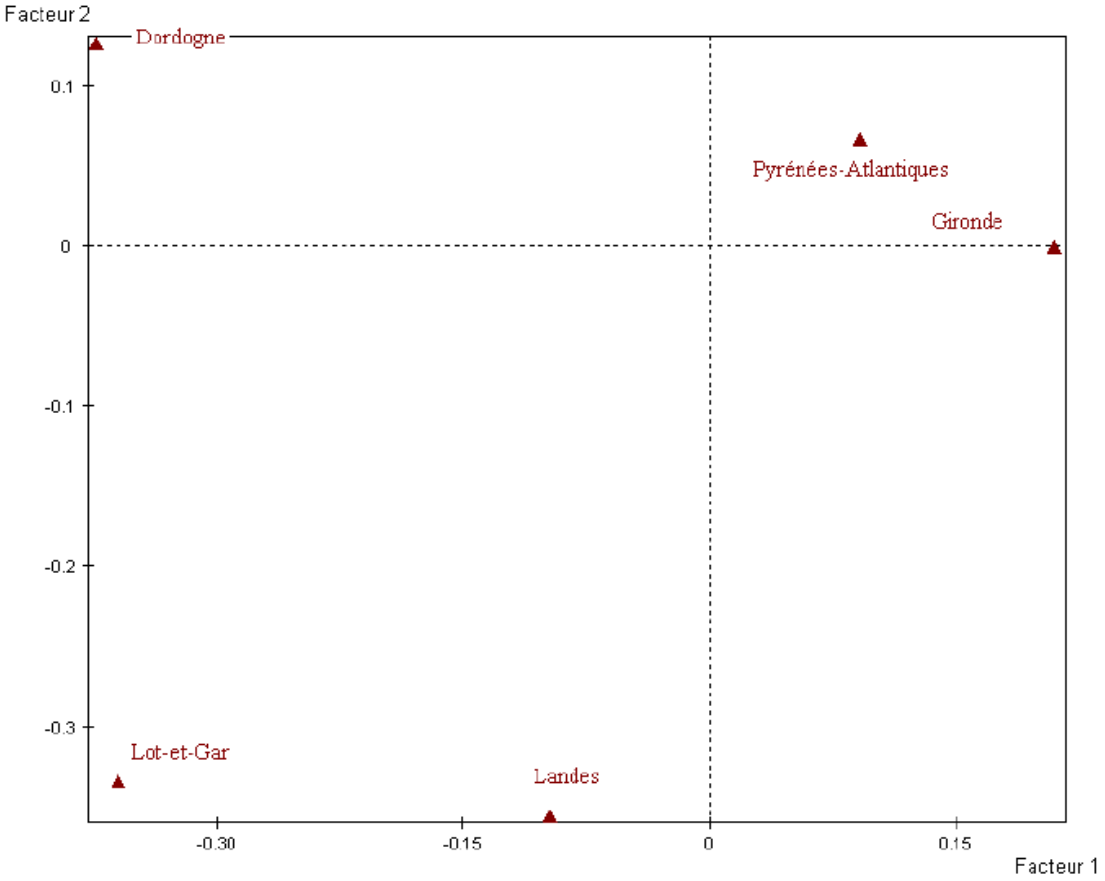


Figure II

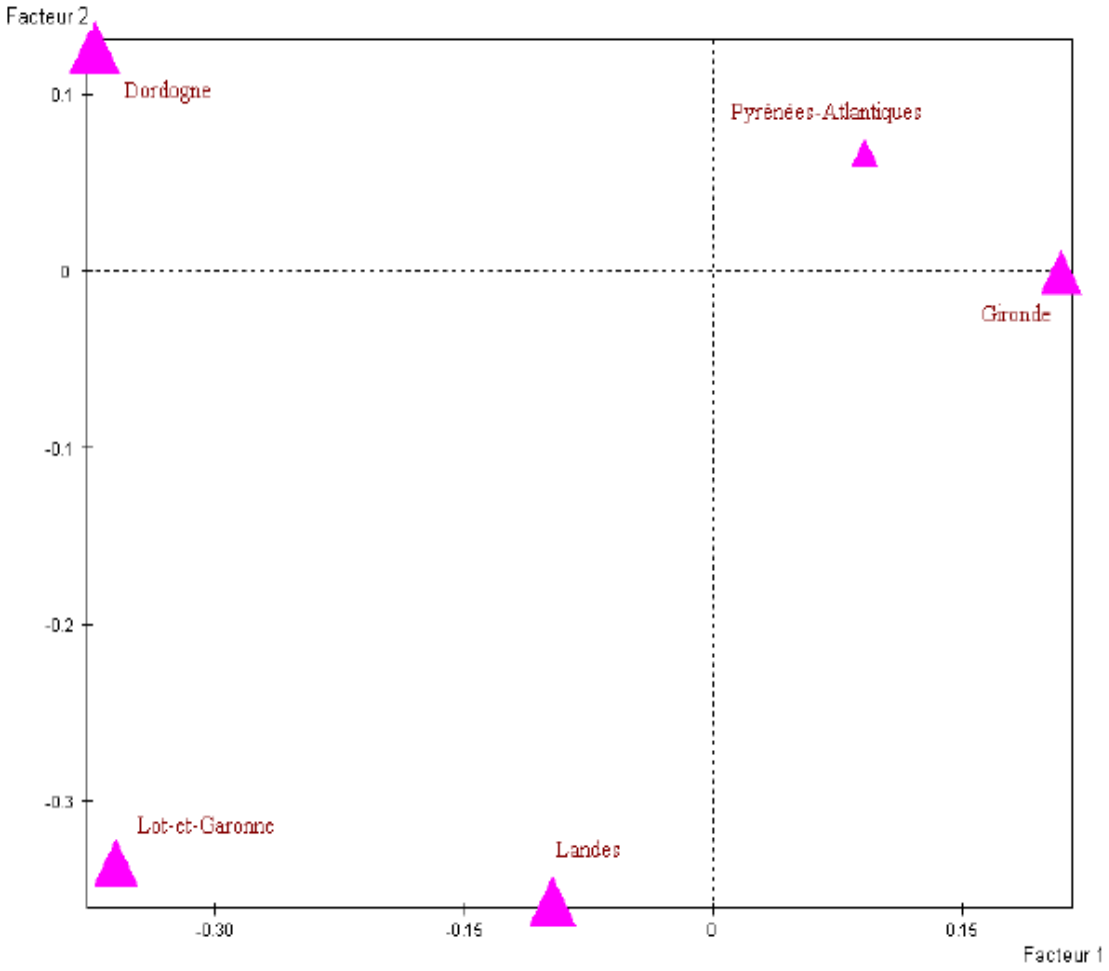
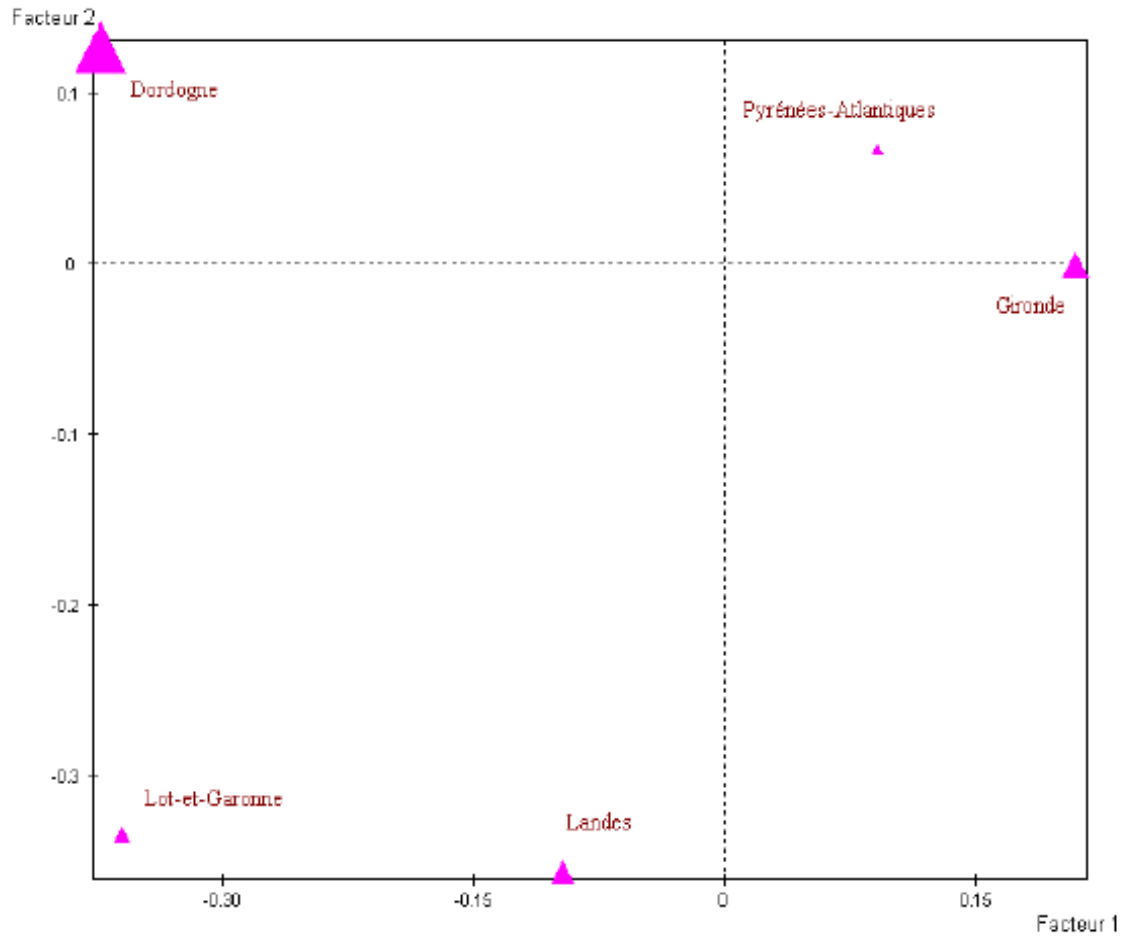


Figure III



3. Les résultats sur les pays sont présentés dans le tableau VI.
   (a) Décrire la qualité de représentation dans le plan F1-F2.
   (b) Interpréter chacun des axes F1 et F2.
   (c) Justifier la représentation simultanée de la figure IV.
   (d) Rédiger une analyse synthétique des résultats.

Tableau VI

COORDONNEES, CONTRIBUTIONS DES FREQUENCES SUR LES AXES 1 A 4
FREQUENCES ACTIVES

	FREQUENCES			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDEN	LIBELLE	COUNT	P.REL	DISTO	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0
C2	Grande-Bretagne-Irla	27.37	0.01		0.07	-0.07	-0.03	0.01	0.00	2.9	5.9	3.1	0.6	0.0	0.49	0.41	0.09	0.01	0.00
C3	Espagne-Portugal	13.35	0.11		-0.32	0.05	0.03	0.02	0.00	26.7	1.3	1.4	3.5	0.0	0.97	0.02	0.01	0.01	0.00
C4	Allemagne	13.05	0.05		-0.11	0.14	0.12	-0.06	0.00	3.1	11.8	20.0	24.8	0.0	0.25	0.40	0.27	0.08	0.00
C5	Belgique-Luxembourg	10.72	0.16		0.39	-0.05	0.05	-0.04	0.00	30.7	1.2	3.3	6.6	0.0	0.96	0.02	0.02	0.01	0.00
C6	États-Unis	8.35	0.08		0.14	-0.23	0.08	0.03	0.00	3.0	20.3	6.1	4.0	0.0	0.24	0.67	0.08	0.01	0.00
C7	Pays-Bas	6.01	0.34		0.37	0.44	-0.07	0.08	0.00	15.5	53.4	3.1	16.4	0.0	0.40	0.57	0.01	0.02	0.00
C8	Italie, Grèce	4.97	0.07		-0.14	-0.02	-0.22	-0.03	0.00	1.8	0.1	28.3	2.3	0.0	0.27	0.01	0.71	0.01	0.00
C9	Suisse	4.12	0.07		-0.10	0.01	-0.21	-0.12	0.00	0.8	0.0	20.7	26.5	0.0	0.16	0.00	0.65	0.20	0.00
C10	Canada.	1.24	0.07		0.14	-0.20	0.12	0.03	0.00	0.4	2.3	2.0	0.6	0.0	0.25	0.54	0.19	0.01	0.00
C11	Autres Europe	4.37	0.13		-0.34	0.05	0.11	0.03	0.00	9.7	0.5	6.5	2.4	0.0	0.88	0.02	0.10	0.01	0.00
C12	Autres	6.46	0.07		-0.21	-0.10	-0.09	0.06	0.00	5.5	3.1	5.4	12.0	0.0	0.67	0.16	0.11	0.06	0.00

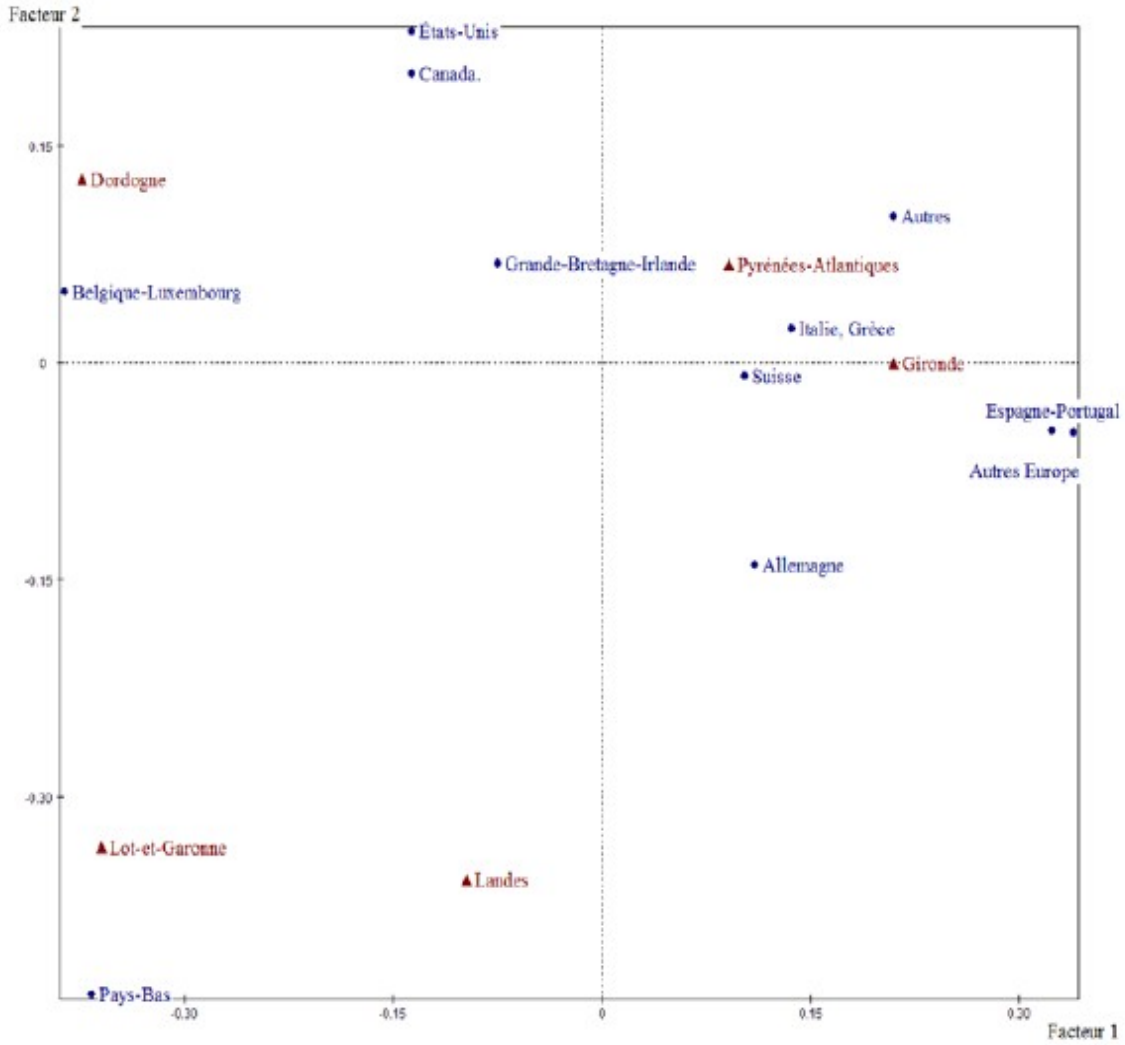


Figure IV

4. On utilise un profil supplémentaire, celui de la France (Tableau VII).
   (a) Définir la notion de profil supplémentaire.
   (b) Préciser son intérêt.
   (c) Comment est obtenu le tableau VIII ? Le commenter.
   (d) Représenter la france dans la figure IV.

Tableau VII

	Dordogne	Gironde	Landes	Lot-et-Garonne	Pyrénées-Atlantiques	Aquitaine
France	12,6	29,6	18,6	3,9	35,2	100,0

Tableau VIII

INDIVIDUS ILLUSTRATIFS (AXES 1 A 4)

INDIVIDUS			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDENTIFICATEUR	P.REL	DISTO	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0
France	338.61	0.12	0.07	-0.27	0.14	0.14	0.00	0.0	0.0	0.0	0.0	0.0	0.04	0.63	0.15	0.18	0.00