

L'analyse factorielle multiple

à partir des macros SAS
`%AFMULT` et `%PLOTAFM`

Brigitte Gelein (Insee Poitou - Charentes)

Olivier Sautory (Cepe - Insee)

L'analyse factorielle multiple à partir des macros SAS %AFMULT et %PLOTAFM

- I. Les principes de l'AFM
- II. Les macros %AFMULT et %PLOTAFM
- III. Un exemple de données thématiques

I. Les principes de l'AFM

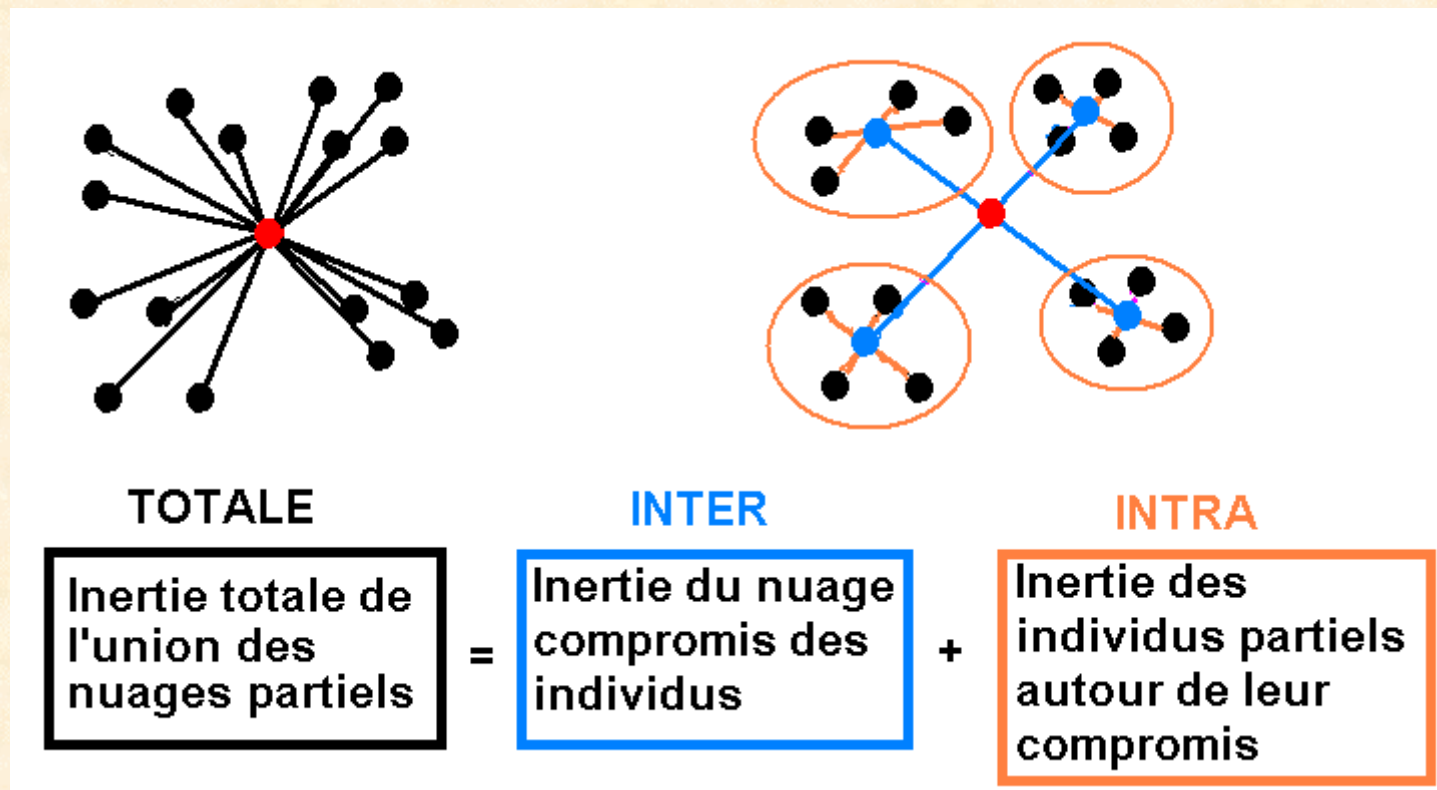
- L'AFM traite simultanément des tableaux croisant les mêmes individus * variables
- Les variables sont structurées en groupes se référant à une même date, un même thème, etc...
- L'AFM permet de traiter simultanément des groupes de variables quantitatives et des groupes de variables qualitatives

I. Les principes de l'AFM

- Étudier les rapports entre les descriptions d'individus obtenues par J groupes de variables
 - Propriétés vérifiées par la représentation simultanée des J nuages d'individus :
 - (P1) Bonne qualité de représentation de chacun des nuages
 - (P2) Proximité des points représentant le même individu dans chacun des nuages
- ⇒ ***Projection d'un nuage moyen compromis entre tous les nuages***

I. Les principes de l'AFM

- D'après le théorème de Huyghens :



I. Les principes de l'AFM

(P1) Bonne qualité de représentation de chacun des nuages

= maximiser l'inertie totale

(P2) Proximité des points représentant le même individu dans chacun des nuages

= minimiser l'inertie intra

Compromis recherché :

maximiser l'inertie inter

I. Les principes de l'AFM

- Inertie inter : inertie du nuage compromis des individus à maximiser
 - on réalise l'ACP de l'ensemble des J groupes de variables
 - on projette en supplémentaires les J nuages d'individus partiels
- Chaque variable d'un groupe est pondérée par l'inverse de la 1^{ère} valeur propre de l'analyse séparée du groupe

I. Les principes de l'AFM

- En analyse factorielle simple : un ensemble de variables a une influence proportionnelle au nombre de ses variables et à leurs corrélations
- En AFM : l'influence des groupes est équilibrée
 - structure des corrélations internes à chaque groupe conservée
 - inertie maximale d'un groupe dans une direction quelconque égale à 1
 - 2 nuages homothétiques sont confondus

I. Les principes de l'AFM

- Autre objectif de l'AFM : mettre en évidence les groupes de variables pour lesquels les structures des individus sont globalement identiques
 - les groupes de variables sont projetés dans un sous-espace dont les axes correspondent aux facteurs communs de l'AFM.
 - la ressemblance est déterminée par le calcul d'une mesure de liaison.

I. Les principes de l'AFM

- Traitement des variables qualitatives
 - Principe : l'ACM est équivalente à une ACP des indicatrices pondérées des variables qualitatives
 - 2 représentations d'une modalité qualitative : corrélation avec les facteurs et projection au centre de gravité des individus qui la possèdent

II. Les macros %AFMULT et %PLOTAFM

La macro %AFMULT

Résultats édités :

- Matrice de corrélation entre les variables actives
- Valeurs propres des analyses séparées
- Histogramme des valeurs propres de l'analyse globale
- Matrice de corrélation entre facteurs partiels

La macro %AFMULT

- Matrices des liaisons entre groupes :
Lg et Rv
- Rapport inertie inter/inertie totale
- Aides à l'interprétation des individus partiels et compromis
- Aides à l'interprétation des variables
- Aides à l'interprétation des groupes

Résultats stockés dans des tables SAS

La macro %AFMULT

L'utilisation du fenêtrage permet :

- des enchaînements d'écrans pour la saisie et la modification des paramètres
- des tests en interactif sur la saisie des paramètres.

L'utilisateur peut, s'il le souhaite, sauvegarder les paramètres saisis dans les différentes fenêtres.

La macro %PLOTAFM

L'utilisateur peut, s'il le souhaite, utiliser le module SAS-GRAPH.

Les graphiques peuvent représenter différents objets :

- les individus compromis actifs ou supplémentaires,
- les individus partiels,
- les variables actives ou supplémentaires,
- les groupes actifs ou supplémentaires.

La macro %PLOTAFM

Il est possible de **sélectionner** les objets à représenter en fonction de :

- leur contribution aux plans factoriels choisis,
- leur qualité de représentation sur ces plans.

On peut ainsi alléger les graphiques.

La macro %PLOTAFM

La **taille des points** peut être uniforme ou proportionnelle à l'un de ces deux critères :

- leur contribution aux plans factoriels choisis,
- leur qualité de représentation sur ces plans.

Ce type de représentation permet de visualiser rapidement les points bien représentés ou ceux ayant joué un rôle important dans la construction des axes.

III. Un exemple de données thématiques

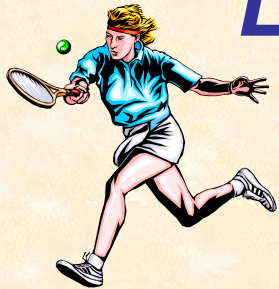
Le contexte :

La Mission Statistique du Ministère de la Jeunesse et des Sports a décidé en 2001 d'élaborer une nomenclature d'activités physiques et sportives.

Les objectifs :

- constituer un cadre d'élaboration des comptes économiques du sport, qui concernent environ 250 disciplines.
- obtenir une image simplifiée et révélatrice du domaine (9 postes).

Les familles d'activités



marche sportive, escalade - montagne - spéléo,
natation - baignade, vélo, athlétisme, boules,
bowling, gymnastique & yoga, sports de neige, sports
de glace, football, rugby, basket, volley, handball et
divers, tennis, tennis de table, badminton - squash,
muscultation, roller - skate, danse, équitation, voile
& planche, glisse sur eau, aviron - canoë kayak, judo,
autres arts martiaux, golf, sports de combat, sports
mécaniques, sports aériens, escrime, tir, pêche.



Les critères

➤ *2 critères qualitatifs :*

1. type de pratique, avec 3 modalités non ordinales :
individuel - duel - collectif
2. instrumentation, avec 4 modalités ordinales :
pas du tout - faiblement
moyennement - fortement instrumenté.

➤ **10 critères quantitatifs :**

1. effectif des pratiquants de 15 ans et plus
2. âge moyen des pratiquants de 15 ans et plus
3. part des femmes parmi les pratiquants de 15 ans et plus
4. part des jeunes de moins de 15 ans parmi les licenciés
5. pourcentage de licenciés parmi les 15 ans et plus
6. effectifs des sportifs de haut niveau
7. budget des fédérations
8. budget du secteur professionnel
9. médiatisation via les heures de télévision
10. médiatisation via l'Équipe

Ces 12 critères sont regroupés en 5 thèmes :

1. nature du sport :

type de pratique, instrumentation

2. démographie :

effectif, âge moyen, % femmes, % jeunes

3. haute performance :

budget professionnel,
effectif des sportifs de haut niveau

4. degré d'organisation :

% licenciés, budget des fédérations

5. médiatisation :

temps d'antenne télé sur chaînes hertziennes,
volume occupé dans l'Équipe

		Nature du sport		Démographie					
34	sport	instrume	pratique	effectif	age	femmes	jeunes	hautnive	B
■ 1	marche sportive	0	1	1210	40	47	13	61	
■ 2	escalade	2	1	1888	34	37	17	187	
■ 3	natation	0	1	14548	34	52	56	1063	
■ 4	velo	2	1	12739	38	41	13	726	
■ 5	athletisme	1	1	7109	32	36	39	1304	
■ 6	boules	1	2	6026	43	27	6	170	
.....									
■ 29	combat	1	2	480	20	31	35	599	
■ 30	mecaniques	3	1	399	33	13	35	280	
■ 31	sports aeriens	3	1	148	32	20	2	293	
■ 32	escrime	2	2	43	22	62	77	395	
■ 33	tir	2	1	250	38	16	11	561	
■ 34	peche	2	1	3047	44	16	12	0	



AFM pour : - équilibrer le poids des thèmes dans l'analyse
 - effectuer un lissage des données en se limitant aux premiers facteurs de l'analyse



CAH sur : coordonnées des familles d'activités sur les premiers axes factoriels de l'analyse.

- Mesures de liaisons entre les groupes :
Coefficients Lg

Groupe	1	2	3	4	5
n°1 - qual act	2.13	0.40	0.57	0.51	0.39
n°2 - qual act	0.40	3.46	0.77	0.72	0.81
n°3 - qual act	0.57	0.77	3.15	0.86	1.16
n°4 - qual act	0.51	0.72	0.86	2.81	0.78
n°5 - qual act	0.39	0.81	1.16	0.78	2.48

- Mesures de liaisons entre les groupes :
Coefficient Rv normé

Groupe	1	2	3	4	5
n°1 - qual act	1.00	0.15	0.22	0.21	0.17
n°2 - qual act	0.15	1.00	0.23	0.23	0.28
n°3 - qual act	0.22	0.23	1.00	0.29	0.41
n°4 - qual act	0.21	0.23	0.29	1.00	0.30
n°5 - qual act	0.17	0.28	0.41	0.30	1.00

Validité de la représentation simultanée :

Les projections partielles des individus doivent être proches de la projection du nuage-compromis.

Si le rapport *inertie inter / inertie totale* est proche de 1 le long d'un axe, il est possible d'effectuer les projections des nuages partiels

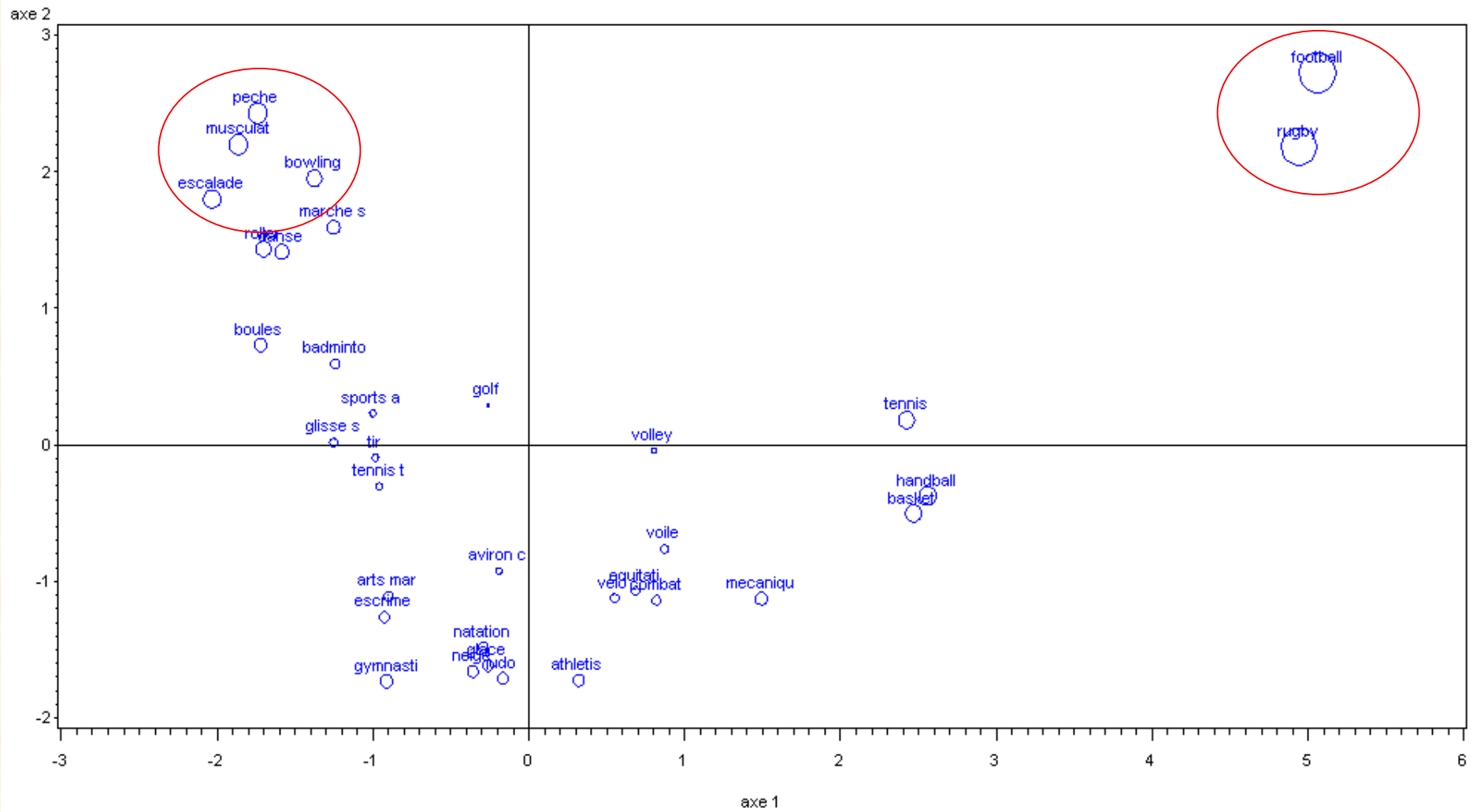
axe 1	axe 2	axe 3
0.67	0.50	0.48

Coordonnées et aides à l'interprétation des individus actifs

id	groupe	disto	coord1	coord2	coord3	ctr1	ctr2	ctr3	cos1	cos2	cos3
marche s		13.09	-1.25	1.59	1.07	1.5	4.0	2.0	12.0	19.4	8.8
	1	-13.6	0.88	1.04	-0.25	1.8	0.1	0.5	-5.7	-8.0	-0.5
	2	136.7	-1.81	1.10	0.94	0.1	0.1	0.0	2.4	0.9	0.7
	3	61.22	-2.95	3.10	-0.57	1.1	0.7	0.8	14.2	15.7	0.5
	4	88.57	0.06	-0.69	3.98	0.7	1.6	2.6	0.0	0.5	17.9
	5	54.33	-2.45	3.41	1.26	0.6	1.0	0.0	11.1	21.4	2.9
escalade		11.22	-2.03	1.80	-0.09	3.9	5.1	0.0	36.8	28.8	0.1
	1	-23.0	-1.08	-0.06	-1.39	0.3	1.1	0.5	-5.1	-0.0	-8.5
	2	133.0	-1.38	1.19	2.03	0.2	0.1	1.4	1.4	1.1	3.1
	3	61.22	-2.95	3.10	-0.57	0.3	0.5	0.1	14.2	15.7	0.5
	4	54.98	-2.30	1.35	-1.77	0.0	0.1	0.9	9.6	3.3	5.7
	5	54.33	-2.45	3.41	1.26	0.1	0.8	0.6	11.1	21.4	2.9
natation		14.53	-0.29	-1.48	-1.76	0.1	3.4	5.3	0.6	15.1	21.3
	1	19.36	0.88	1.04	-0.25	0.5	2.0	0.7	4.0	5.6	0.3
	2	160.1	-0.91	-3.20	-1.73	0.1	0.9	0.0	0.5	6.4	1.9
	3	55.85	-0.24	-0.34	-1.90	0.0	0.4	0.0	0.1	0.2	6.5
	4	70.04	-0.33	-2.10	-3.36	0.0	0.1	0.8	0.2	6.3	16.2
	5	57.90	-0.87	-2.80	-1.55	0.1	0.5	0.0	1.3	13.5	4.2

Représentation des individus compromis actifs

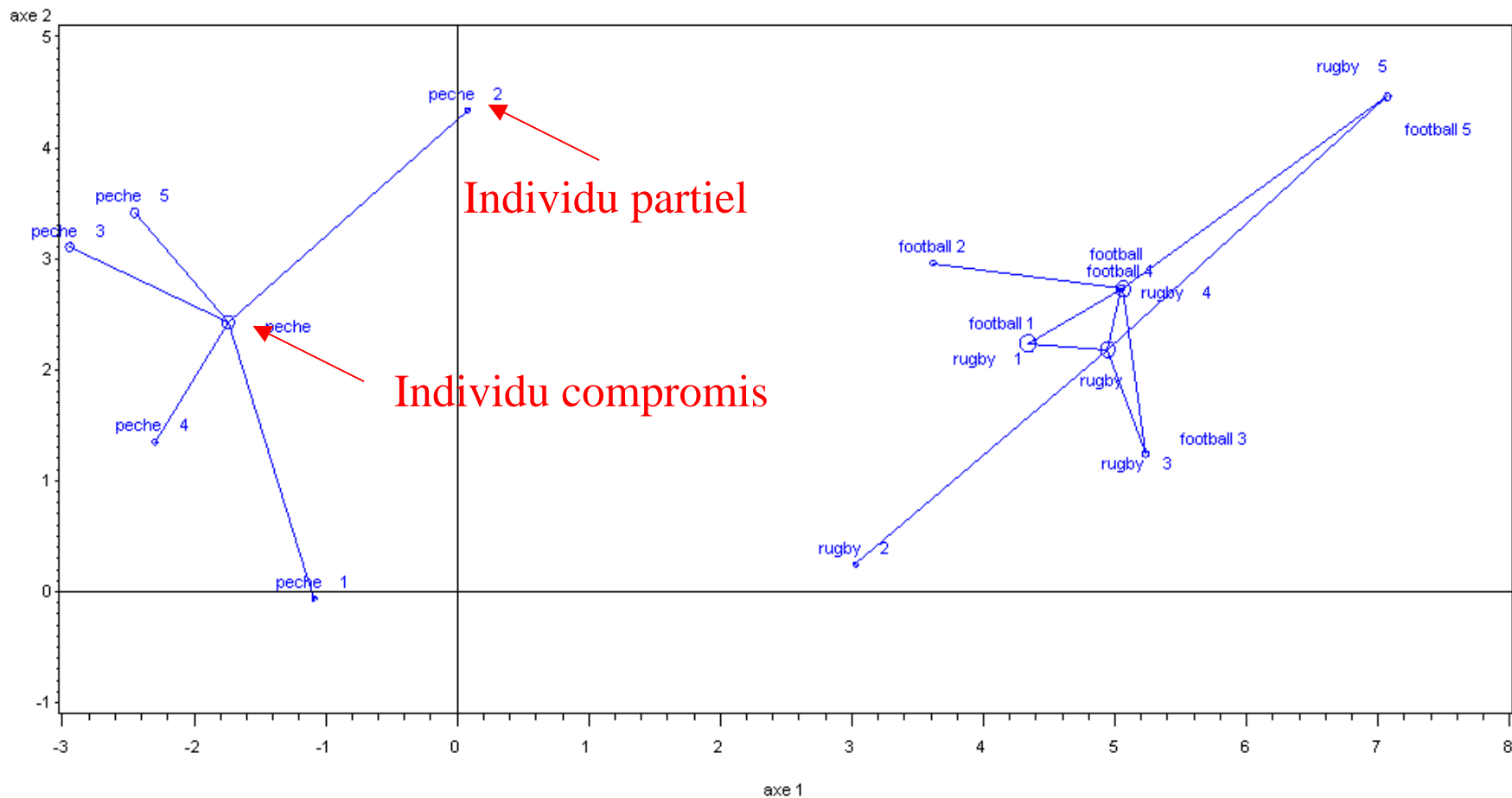
proportionnelle à la contribution à l'inertie expliquée par le plan



Représentation des individus compromis et partiels actifs

proportionnelle à la qualité de représentation sur le plan

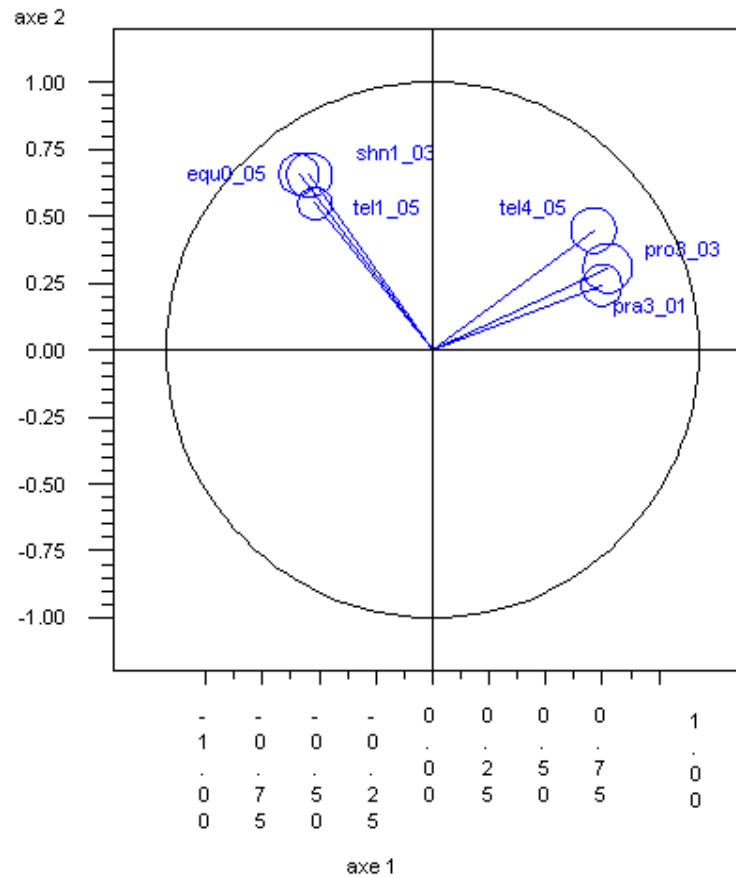
3 individus actifs ayant la plus forte contribution à l'inertie expliquée par le plan



Représentation des variables actives

proportionnelle à la contribution à l'inertie expliquée par le plan

6 variables ayant la plus forte contribution à l'inertie expliquée par le plan



Coordonnées et aides à l'interprétation des groupes actifs

groupe	disto	coord1	coord2	coord3	ctr1	ctr2	ctr3	cos1	cos2	cos3
groupe 1	2.13	0.49	0.10	0.19	15.7	5.4	10.9	11.2	0.5	1.6
groupe 2	3.46	0.35	0.36	0.29	11.5	19.0	16.9	3.6	3.7	2.4
groupe 3	3.15	0.80	0.44	0.57	25.9	23.4	33.3	20.4	6.1	10.3
groupe 4	2.81	0.61	0.32	0.48	19.7	17.3	28.2	13.2	3.7	8.3
groupe 5	2.48	0.84	0.66	0.18	27.2	34.9	10.7	28.5	17.3	1.4
Ensemble					100.0	100.0	100.0	14.8	6.2	5.1

N°	sigle	NAPS Libellé des 9 classes	Activités physiques et sportives retenues dans l'analyse statistique en 34 familles
1	LEA	Loisirs équipées autonomes	Escalade-montagne-spéléo ; Bowling & quilles ; Musculation ; Roller-skate Pêche
2	SOL	Socio-ludiques	Marche sportive ; Boules et billard ; Tennis de table ; Badminton-squash Danse
3	HDO	À haut degré d'organisation	Judo et divers ; Autres arts martiaux ; Sports aériens ; Escrime ; Tir
4	INI	Individuelles instrumentées	Sports de glace ; Glisse sur eau ; Aviron-canoë ; Golf
5	DEQ	D'équipage	Équitation ; Voile & planche
6	IDM	Individuelles de masse	Natation-baignade ; Vélo ; Athlétisme-footing ; Gymnastique & yoga ; Sports de neige
7	MEC	Mécaniques	Sports mécaniques
8	SPR	Semi-professionnelles	Basket-ball ; Volley-ball ; Handball & co ; Tennis ; Sports de combat
9	PRO	Professionnelles	Football Rugby