

M2 – APE – Economic and Statistical Studies



Data Science - Data Analysis

Factorial methods

University year 2025-2026

R. Abdesselam

Courriel ; rafik.abdesselam@univ-lyon2.fr

<http://perso.univ-lyon2.fr/~rabdesse/Documents/>

M2 – APE – Chargé d'Etudes Economiques



Science des Données - Analyse des données

Polycopié 1 : Méthodes factorielles

Année Universitaire 2025-2026

R. Abdesselam

Courriel ; rafik.abdesselam@univ-lyon2.fr

<http://perso.univ-lyon2.fr/~rabdesse/Documents/>

Rappels - Analyse statistique des données uni-bidimensionnelles

- Introduction aux méthodes de sondage
- Les types de questions (fermée, ouverte-textuelle) - codage
- Méthodes descriptives (tris à plat - tris croisés - Tamis – analyse bivariée)
- Méthodologie et Techniques de dépouillement d'enquête - traitements statistiques

1 Analyse statistique des données multidimensionnelle

1.1 Méthodes factorielles

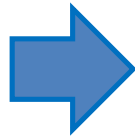
- Analyse en Composantes Principales
- Analyse Factorielle des Correspondances Simples - Multiples
- Analyse Factorielle Discriminante - Analyse Discriminante Linéaire

1.2 Méthodes de classification automatique des données

- Classification non hiérarchique : nuées dynamiques
- Classification hiérarchique ascendante – descendante
- Analyse et interprétation d'une partition

DIFFERENTS TYPES DE CARACTERES

**QUANTITATIF
CONTINU
(numérique)**



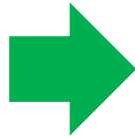
DISCRET : (valeurs entières ou discrètes)

Nombre d'enfants, N° dossier, etc...

CONTINU : (valeurs réelles)

Poids, taille, âge, C.A., PIB, taux d'inflation, etc...

**QUALITATIF
(catégorielle)**



ORDONNE : (modalités ordonnées)

Mention ou appréciation : « TB, B, AB, Passable »

Grades : « lieutenant, capitaine, commandant, etc. »

Niveaux d'études : « Bepc, Bac, Sup »

NOMINAL : (modalités incompatibles et exhaustives)

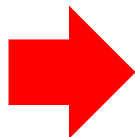
CSP : « cadre, ouvrier, employé, artisan, etc... »

INDICATRICE : (codage binaire 0/1, Oui/Non)

Sexe : « Masculin / Féminin »

Possédez-vous un abonnement Internet ? : « Oui / Non »

TEXTE



QUESTION OUVERTES :

« Que pensez-vous de la réforme LMD ? »

« Pourquoi y a-t-il du chômage en France ? »

(mots, expressions, phrases, traitement de mots)

BASE DE DONNEES - INFORMATION A RESUMER

TABLEAU DE DONNEES

p : Caractères mixtes

	x^1	x^2	...	x^j	...	x^p
x_1				x_1^j		
x_2				x_2^j		
$\vdots$				$\vdots$		
$\vdots$				$\vdots$		
$\vdots$				$\vdots$		
x_i	x_i^1	x_i^2	...	x_i^j	...	x_i^p
$\vdots$				$\vdots$		
$\vdots$				$\vdots$		
$\vdots$				$\vdots$		
x_n				x_n^j		

n objets - individus

$x^j \in \mathbb{R}^n$: $j^{\text{ème}}$ variable

$x_i \in \mathbb{R}^p$: $i^{\text{ème}}$ individu

$x_i^j \in \mathbb{R}$: observation du $i^{\text{ème}}$ individu pour la $j^{\text{ème}}$ variable

Objectifs de l'analyse factorielle

ACP - Analyse en Composantes Principales

Décrire un tableau $X_{(n,p)}$ individus $\times$ variables :

- Visualiser les corrélations entre les variables
- Interpréter les axes factoriels
- Visualiser le positionnement des individus les uns par rapport aux autres

Visualisation des données

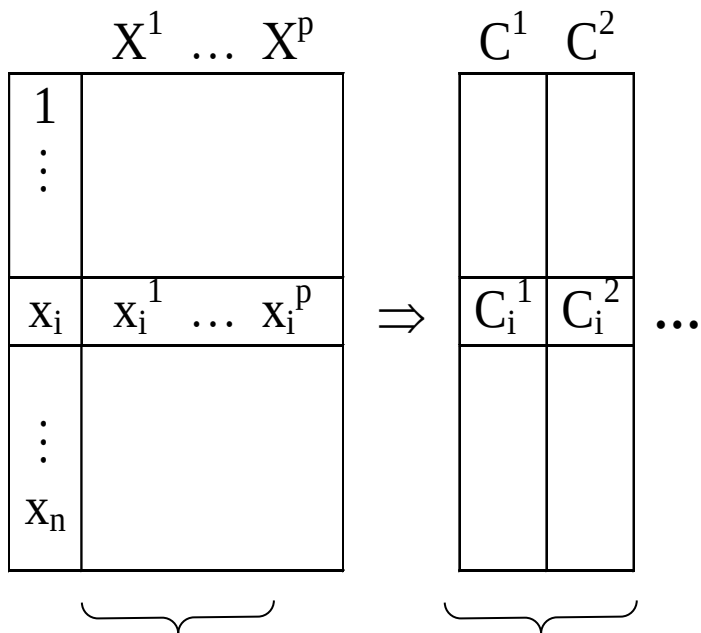
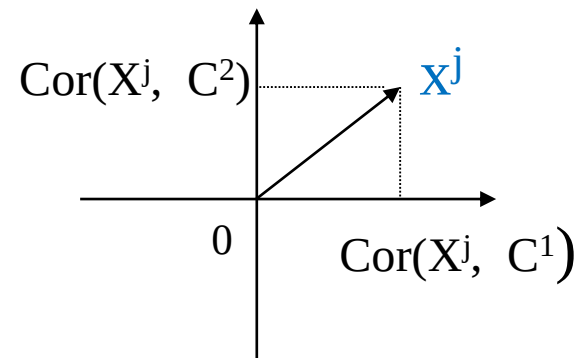
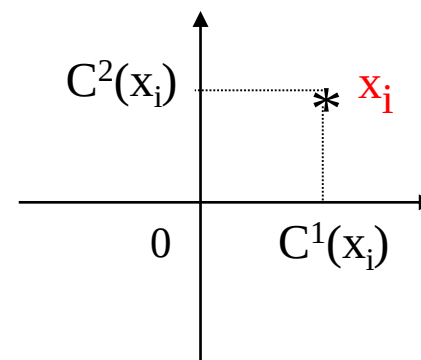


Tableau
des données
Initiales $X_{(n,p)}$

Tableau
des données
d'arrivée $C_{(n,q)}$



La carte des variables



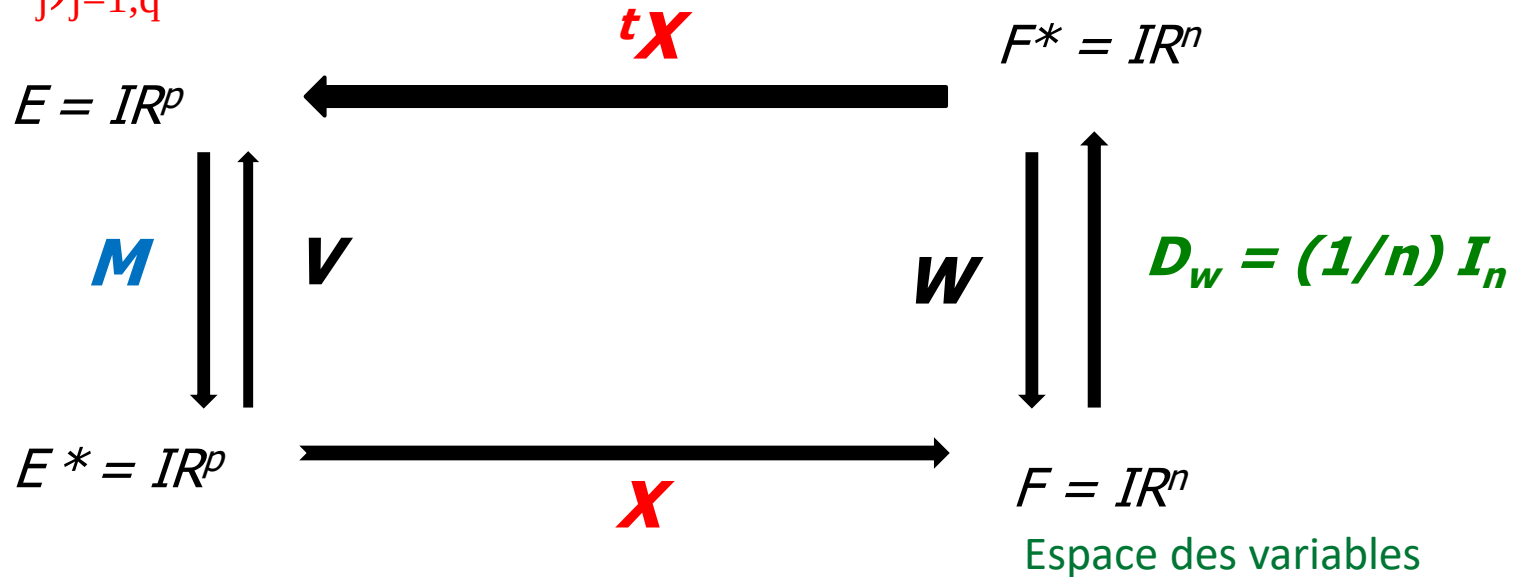
Le plan factoriel des individus

Analyse en Composantes Principales Sch ma de dualit 

- $ACP (\textcolor{red}{X}_{(n,p)} ; \textcolor{blue}{M}_{(p)} ; \textcolor{green}{D}_{w(n)})$

Espace des individus

$(c_j ; \lambda_j)_{j=1,q}$



λ_j : $j^{\text{ me}}$ valeur propre – moment principal

c_j : $j^{\text{ me}}$ vecteur propre – facteur principal

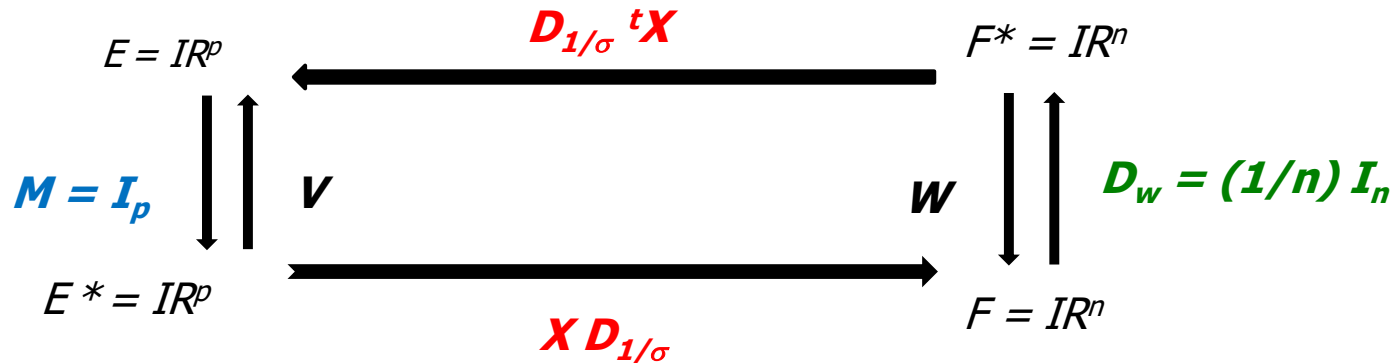
C^j : $j^{\text{ me}}$ composante principale

Diagonalisations : $V \textcolor{blue}{M} c_j = \lambda_j c_j \longleftrightarrow W D_w C^j = \lambda_j C^j$

Analyse en Composantes Principales Normée

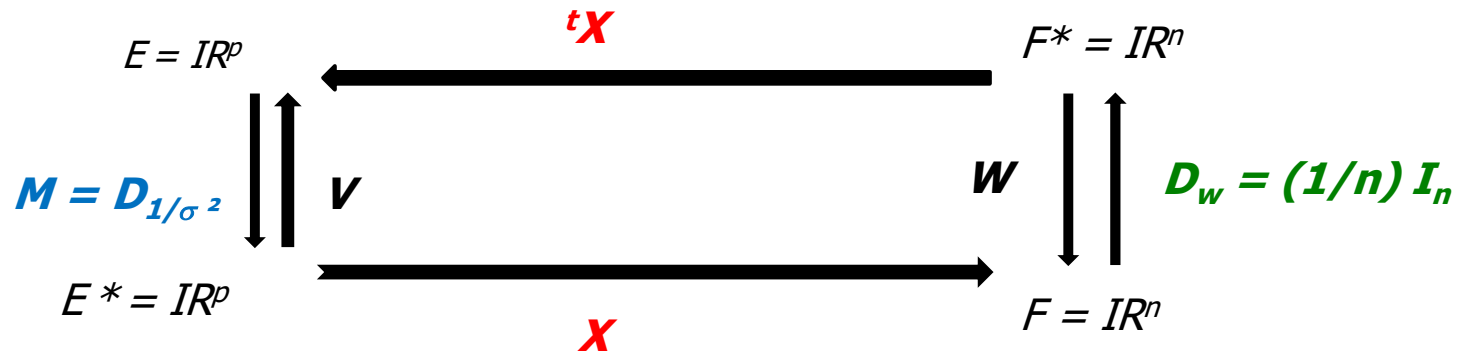
Deux métriques particulières - Schémas de dualité

- ACP1 ($X D_{1/\sigma}$; I_p ; $D_{w(n)}$)



ACP1 : $VM = D_{1/\sigma}^t X D_w X D_{1/\sigma} I_p = D_{1/\sigma} V_x D_{1/\sigma} = R$ Matrice des corrélations linéaires

- ACP2 (X ; D_{1/σ^2} ; $D_{w(n)}$)



ACP2 : $VM = tX D_w X D_{1/\sigma^2} = V_x D_{1/\sigma^2} = R$ Matrice des corrélations linéaires

Exemple d'application **Dictionnaire des variables**

-
- | | |
|--|-----------------|
| 1 . Etat membre | (3 MODALITES) |
| AUTR - Autre ZNEU - Zone Non Euro ZEUR - Zone Euro | |
| 2 . Population (%) UE-2000 | (CONTINUE) |
| POPU - Population (%) UE-28 | |
| 3 . Produit Intérieur Brut (%) UE-28 | (CONTINUE) |
| PIB - Produit Intérieur Brut | |

Thème : Finances publiques de l'UE-28 en 2014

- | | |
|--|--------------|
| 4 . Recettes totales des APU en % du PIB | (CONTINUE) |
| RECE - Recettes totales des | |
| 5 . Dépenses totales des APU en % du PIB | (CONTINUE) |
| DEPE - Dépenses totales des | |
| 6 . Solde des finances publiques en % du PIB | (CONTINUE) |
| SDFP - Solde des finances p | |
| 7 . Dette brute des APU en % du PIB | (CONTINUE) |
| DETB - Dette brute des APU | |

Thème : Activité-Emploi

- | | |
|--|--------------|
| 8 . Taux de chômage en % population active | (CONTINUE) |
| TCHO - Taux de chômage en % | |
| 9 . Ventes au détail en volume (%) | (CONTINUE) |
| VENT - Ventes au détail en | |
| 10 . Production industrielle hors bâtiment | (CONTINUE) |
| PIND - Production industrie | |
| 11 . Taux de croissance en volume du PIB | (CONTINUE) |
| TCRO - Taux de croissance | |
-

Dictionnaire des variables

SELECTION DES INDIVIDUS ET DES VARIABLES UTILES

VARIABLES NOMINALES ILLUSTRATIVES

1 VARIABLES 3 MODALITES ASSOCIEES

1 . Etat membre

(3 MODALITES)

VARIABLES CONTINUES ACTIVES : 4 VARIABLES

4 . Recettes totales des APU en % du PIB (CONTINUE)

5 . Dépenses totales des APU en % du PIB (CONTINUE)

6 . Solde des finances publiques en % du PIB (CONTINUE)

7 . Dette brute des APU en % du PIB (CONTINUE)

VARIABLES CONTINUES ILLUSTRATIVES : 4 VARIABLES

8 . Taux de chômage en % population active (CONTINUE)

9 . Ventes au détail en volume (%) (CONTINUE)

10 . Production industrielle hors bâtiment (CONTINUE)

11 . Taux de croissance en volume du PIB (CONTINUE)

INDIVIDUS

----- NOMBRE ----- POIDS -----

POIDS DES INDIVIDUS: Poids des individus, uniforme egal a 1. UNIF

RETENUS NITOT = 32 PITOT = 32.000

SELECTION APRES FILTRAGE

ACTIFS NIACT = 28 PIACT = 28 .000

SUPPLEMENTAIRES NISUP = 4 PISUP = 4.000

Analyse en Composantes Principales

ANALYSE EN COMPOSANTES PRINCIPALES

STATISTIQUES SOMMAIRES DES VARIABLES CONTINUES

EFFECTIF TOTAL : 28 POIDS TOTAL : 28.00

NUM . IDEN - LIBELLE	EFFECTIF	POIDS	MOYENNE	ECART-TYPE	MINIMUM	MAXIMUM
3 . Sold - Solde Public	28	28.00	-2.83	2.20	-8.80	1.20
4 . Dett - Dette Publique	28	28.00	73.61	36.97	10.60	177.10
5 . Dépe - Dépenses Publiques	28	28.00	46.24	6.57	34.90	58.70
6 . Rece - Recettes Publiques	28	28.00	43.44	6.51	33.40	58.50
7 . Taux - Taux Chômage	28	28.00	10.10	5.15	4.80	26.00
8 . Taux - Taux Croissance	28	28.00	1.47	1.43	-3.20	3.60
9 . Taux - Taux Inflation	28	28.00	0.43	0.61	-1.30	1.60
10 . Popu - Population-Millions	28	28.00	18.15	23.02	0.40	81.00
11 . PIB- - PIB-Milliards-Euros	28	28.00	498.53	760.71	8.10	2915.70
12 . PIB/ - PIB/Hab.-Euros	28	28.00	26752.86	11340.27	12540.00	73510.00

MATRICE DES CORRELATIONS

	Sold	Dett	Dépe	Rece
Sold	1.00			
Dett	-0.58	1.00		
Dépe	-0.19	0.40	1.00	
Rece	0.14	0.21	0.94	1.00

MATRICE DES VALEURS-TESTS

	Sold	Dett	Dépe	Rece
Sold	99.99			
Dett	-3.53	99.99		
Dépe	-1.04	2.26	99.99	
Rece	0.76	1.12	9.33	99.99

Interprétation d'une ACP :

1- Choix du nombre d'axes à conserver

□ Choisir le nombre *q* d'axes factoriels (ou de composantes principales) à retenir pour obtenir un *résumé suffisamment précis* de l'information contenue dans le tableau initial

2- Interprétation des axes factoriels

□ Donner une *signification* aux nouvelles variables (facteurs ou composantes principales)

3- Représentations graphiques

□ Construire les graphiques des *variables* et des *individus*

4- Qualité de la représentation

□ Evaluer la *qualité* de ce résumé

5- Synthèse

□ *Interprétation des résultats*

Analyse en Composantes Principales

VALEURS PROPRES

APERCU DE LA PRECISION DES CALCULS : TRACE AVANT DIAGONALISATION .. 4.0000
 SOMME DES VALEURS PROPRES 4.0000

HISTOGRAMME DES 4 PREMIERES VALEURS PROPRES

NUMERO	VALEUR PROPRE	POURCENTAGE	POURCENTAGE CUMULE	
1	2.1803	54.51	54.51	*****
2	1.4469	36.17	90.68	*****
3	0.3728	9.32	100.00	*****
4	0.0000	0.00	100.00	*

But : obtenir le maximum d'inertie conservée avec le minimum de facteurs. Deux critères empiriques pour sélectionner le nombre d'axes à retenir :

Critère de Cattell ou du "coude" : sur l'histogramme des valeurs propres, on observe un décrochement (coude) suivi d'une décroissance régulière. **On sélectionne les axes avant le décrochement.**

Critère de Kaiser : on ne retient que les axes dont l'inertie est supérieure à l'inertie moyenne $(1/p) \sum \lambda_i$. En ACP normée (centrée-réduite) $(1/p) \sum \lambda_i = 1$. On ne retiendra que les axes associés à des **valeurs propre supérieures à 1**. Les variables de départ ont des variances égales à 1 et on cherche des combinaisons linéaires de variance maximale donc supérieures.

Dans la pratique, on retient en fait les q premiers axes que l'on sait interpréter.

Analyse en Composantes Principales

Construction des nuages de points projetés :

- Chaque nuage de points (variables et individus) est construit en projection sur les **plans factoriels** : repère du plan défini par deux des q axes factoriels retenus.
Ex : Si l'on retient 3 axes, on tracera 3 graphiques pour chaque nuage: le nuage projeté sur le plan (axe1, axe2), celui projeté sur le plan (axe1, axe3) et celui projeté sur le plan (axe2,axe3).
- L'examen des plans factoriels permettra de **visualiser les corrélations** entre les variables et **d'identifier les groupes d'individus** ayant pris des valeurs proches sur certaines variables.
- *Avant de lire les graphiques, il faut interpréter les axes et s'assurer que la projection est fidèle à la réalité.*

Deux types de facteurs :

Effet taille : lorsque toutes les variables initiales sont **corrélées positivement entre elles**, toutes les variables sont du même côté du premier axe factoriel. (i.e. elles contribuent toutes dans le même sens à la formation de la première composante principale, la matrice des corrélations a tous ses termes positifs ; les individus sont rangés sur l'axe 1 par valeurs croissantes d'une moyenne pondérée des variables initiales.)

Effet forme : Deux groupes de variables opposées : celles qui contribuent positivement à l'axe, celles qui contribuent négativement.

2 Interprétation des axes factoriels : pour chaque axe retenu et chaque nuage, on regarde d'abord :

2.1 Quelles sont les variables qui « caractérisent l'axe factoriel » sont corrélées à l'axe

Outil d'aide à l'interprétation : **corrélations variables - facteurs**.

2.2 Quels sont les individus qui participent le plus à la formation de l'axe.

Outils d'aide à l'interprétation : **contributions** des points à l'inertie de cet axe et **cosinus carrés**.

Ce sont les points dont la contribution est supérieure à la moyenne qui permettent de donner un sens à l'axe.

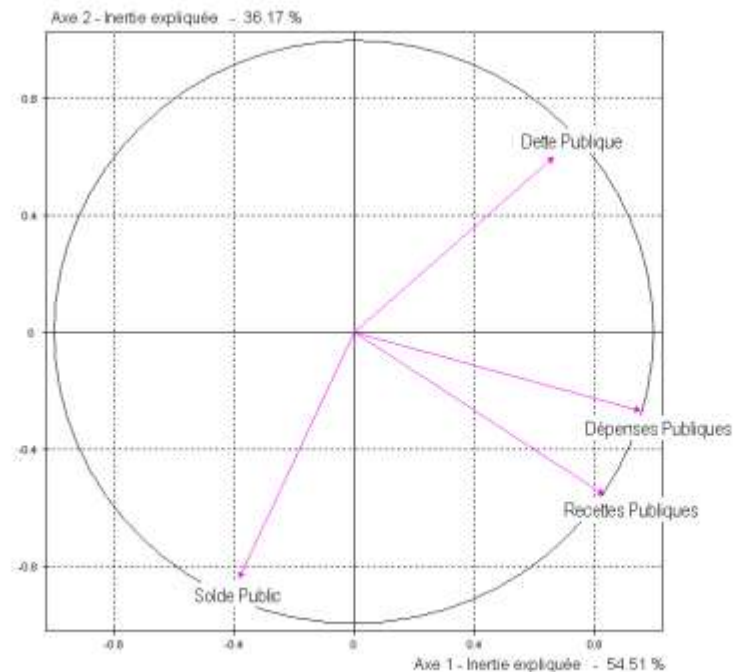
Repr sentation des variables actives – Finances publiques

Cercle de corr lation

COORDONNEES DES VARIABLES SUR LES AXES 1 A 4

VARIABLES ACTIVES

VARIABLES	COORDONNEES					CORRELATIONS VARIABLE-FACTEUR					ANCIENS AXES UNITAIRES				
	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0
Sold - Solde Public	-0.38	-0.84	-0.38	0.00	0.00	-0.38	-0.84	-0.38	0.00	0.00	-0.26	-0.70	-0.62	0.23	0.00
Dett - Dette Publique	0.66	0.60	-0.45	0.00	0.00	0.66	0.60	-0.45	0.00	0.00	0.45	0.50	-0.74	0.00	0.00
D�pe - D�penses Publiques	0.95	-0.27	0.15	0.00	0.00	0.95	-0.27	0.15	0.00	0.00	0.65	-0.22	0.24	0.69	0.00
Rece - Recettes Publiques	0.83	-0.56	0.02	0.00	0.00	0.83	-0.56	0.02	0.00	0.00	0.56	-0.46	0.03	-0.68	0.00



Etude des proximités entre points variables

- Qualité de représentation de la variable X^j sur l'axe c_k :

$$Qlt_k(X^j) = \cos^2(\theta_{kj}) = d_{jk}^2 / \|X^j\|_{Dw}^2$$

- En ACP normée « cercle de corrélations »,

$$Qlt_k(X^j) = r^2(X^j, C^k) = d_{jk}^2$$

❖ une variable est d'autant mieux représentée sur un axe qu'elle est *proche de la circonférence du cercle* des corrélations et de l'axe, d'autant plus mal représentée qu'elle est *proche de l'origine*.

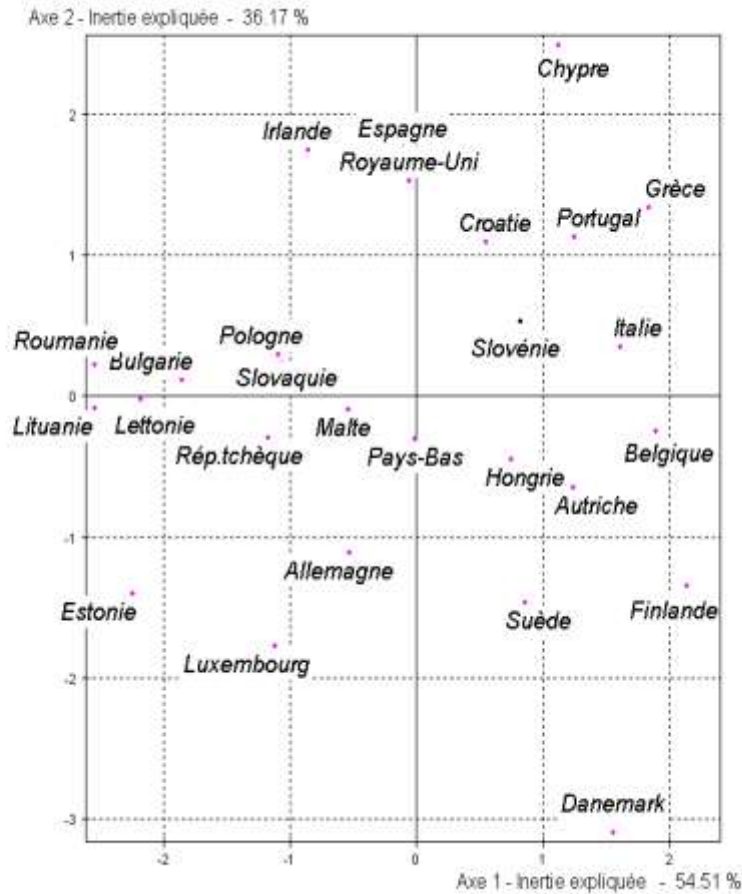
❖ les variables qui *contribuent* le plus à l'axe sont aussi celles qui sont le *mieux représentées* et inversement, donc pas besoin d'étude spécifique de la représentativité.

COORDONNEES, CONTRIBUTIONS ET COSINUS CARRES DES INDIVIDUS
INDIVIDUS ACTIFS (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
Belgique	3.57	3.71	1.90	-0.26	-0.23	0.00	0.00	5.9	0.2	0.5	0.9	0.0	0.97	0.02	0.01	0.00	0.00
Bulgarie	3.57	3.86	-1.86	0.11	0.63	0.00	0.00	5.7	0.0	3.8	1.8	0.0	0.89	0.00	0.10	0.00	0.00
Rép.tchèque	3.57	1.52	-1.18	-0.30	0.22	-0.01	0.00	2.3	0.2	0.5	12.4	0.0	0.91	0.06	0.03	0.00	0.00
Danemark	3.57	12.08	1.56	-3.10	-0.10	0.00	0.00	4.0	23.8	0.1	4.7	0.0	0.20	0.80	0.00	0.00	0.00
Allemagne	3.57	2.74	-0.53	-1.11	-1.10	0.00	0.00	0.5	3.1	11.7	2.9	0.0	0.10	0.45	0.44	0.00	0.00
Estonie	3.57	7.00	-2.25	-1.40	0.01	0.00	0.00	8.3	4.8	0.0	5.2	0.0	0.72	0.28	0.00	0.00	0.00
Irlande	3.57	4.22	-0.86	1.74	-0.67	0.00	0.00	1.2	7.5	4.3	0.0	0.0	0.18	0.72	0.11	0.00	0.00
Grèce	3.57	8.28	1.84	1.34	-1.77	0.00	0.00	5.5	4.4	29.9	0.1	0.0	0.41	0.22	0.38	0.00	0.00
Espagne	3.57	3.17	-0.10	1.76	0.24	0.00	0.00	0.0	7.7	0.5	0.0	0.0	0.00	0.98	0.02	0.00	0.00
France	3.57	5.65	2.32	-0.40	0.35	0.00	0.00	8.8	0.4	1.2	1.1	0.0	0.95	0.03	0.02	0.00	0.00
Croatie	3.57	1.91	0.55	1.09	0.65	0.00	0.00	0.5	2.9	4.0	0.1	0.0	0.16	0.62	0.22	0.00	0.00
Italie	3.57	3.57	1.61	0.35	-0.93	0.00	0.00	4.2	0.3	8.3	0.2	0.0	0.72	0.03	0.24	0.00	0.00
Chypre	3.57	8.66	1.12	2.49	1.11	0.00	0.00	2.1	15.2	11.7	0.3	0.0	0.15	0.71	0.14	0.00	0.00
Lettonie	3.57	4.75	-2.18	-0.03	-0.11	0.00	0.00	7.8	0.0	0.1	1.9	0.0	1.00	0.00	0.00	0.00	0.00
Lituanie	3.57	6.67	-2.55	-0.08	-0.40	-0.01	0.00	10.6	0.0	1.5	12.6	0.0	0.97	0.00	0.02	0.00	0.00
Luxembourg	3.57	4.42	-1.12	-1.78	-0.04	-0.01	0.00	2.1	7.8	0.0	6.9	0.0	0.28	0.72	0.00	0.00	0.00
Hongrie	3.57	0.77	0.75	-0.45	0.03	-0.01	0.00	0.9	0.5	0.0	13.8	0.0	0.73	0.27	0.00	0.00	0.00
Malte	3.57	0.34	-0.54	-0.10	-0.19	0.00	0.00	0.5	0.0	0.3	1.4	0.0	0.86	0.03	0.11	0.00	0.00
Pays-Bas	3.57	0.09	-0.01	-0.31	-0.04	0.00	0.00	0.0	0.2	0.0	1.6	0.0	0.00	0.99	0.01	0.00	0.00
Autriche	3.57	1.96	1.24	-0.65	-0.09	0.00	0.00	2.5	1.1	0.1	1.7	0.0	0.78	0.22	0.00	0.00	0.00
Pologne	3.57	1.44	-1.09	0.30	0.40	0.00	0.00	2.0	0.2	1.5	1.2	0.0	0.83	0.06	0.11	0.00	0.00
Portugal	3.57	3.13	1.25	1.13	-0.56	0.00	0.00	2.5	3.1	3.0	0.0	0.0	0.50	0.41	0.10	0.00	0.00
Roumanie	3.57	6.55	-2.55	0.22	-0.16	0.00	0.00	10.6	0.1	0.2	1.6	0.0	0.99	0.01	0.00	0.00	0.00
Slovénie	3.57	1.28	0.82	0.53	0.58	-0.01	0.00	1.1	0.7	3.2	18.3	0.0	0.52	0.22	0.26	0.00	0.00
Slovaquie	3.57	1.24	-1.06	0.23	0.24	0.00	0.00	1.8	0.1	0.6	1.2	0.0	0.91	0.04	0.05	0.00	0.00
Finlande	3.57	7.21	2.14	-1.35	0.90	0.00	0.00	7.5	4.5	7.8	3.1	0.0	0.63	0.25	0.11	0.00	0.00
Suède	3.57	3.27	0.86	-1.47	0.62	0.00	0.00	1.2	5.3	3.6	3.9	0.0	0.22	0.66	0.12	0.00	0.00
Royaume-Uni	3.57	2.50	-0.06	1.53	0.41	0.00	0.00	0.0	5.8	1.6	0.0	0.0	0.00	0.93	0.07	0.00	0.00

Représentation des individus actifs – UE 28 - Finances publiques

Premier plan principal

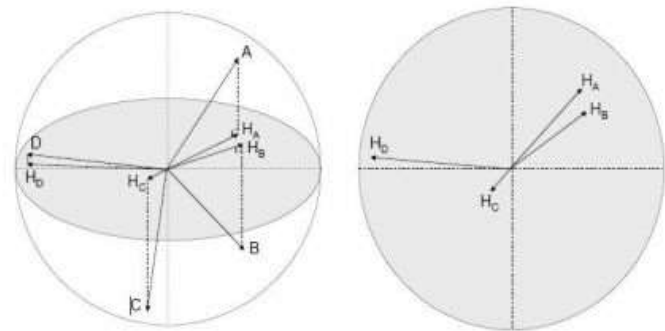
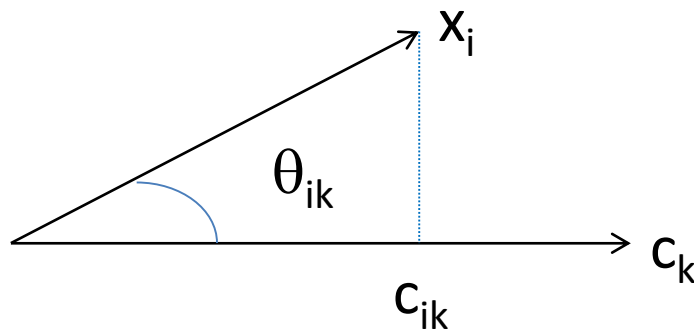


Etude des proximités entre points individus

□ Qualité de représentation de l'individu x_i sur l'axe c_k :

$$Qlt_k(x_i) = \cos^2(\theta_{ki}) = c_{ik}^2 / \|x_i\|_M^2$$

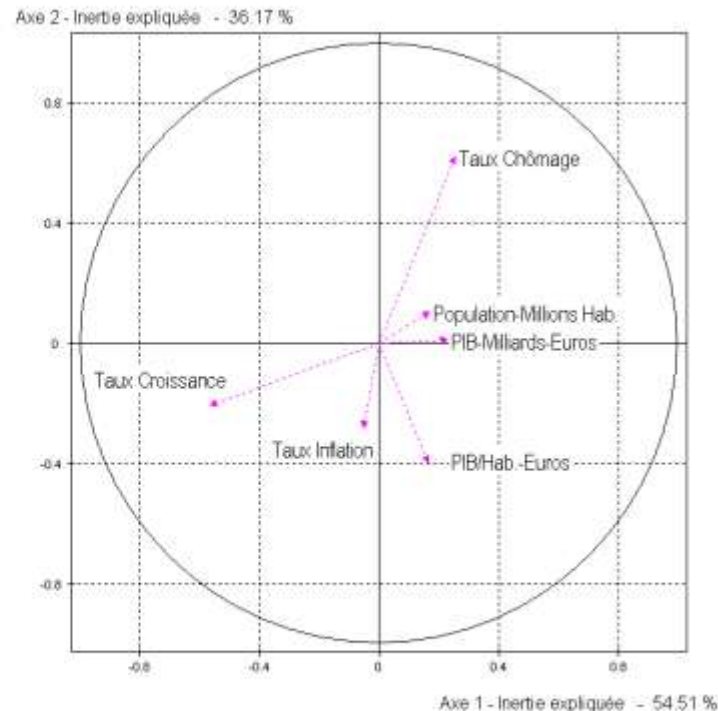
$$\text{avec } \|x_i\|_M^2 = \sum_{i=1}^p c_{ik}^2$$



❖ Un individu est bien représenté lorsque l'angle est proche de 0, c'est-à-dire que le cosinus est proche de 1. Dans le cas inverse, l'angle est proche de 90° et le cosinus est proche de zéro.

VARIABLES ILLUSTRATIVES

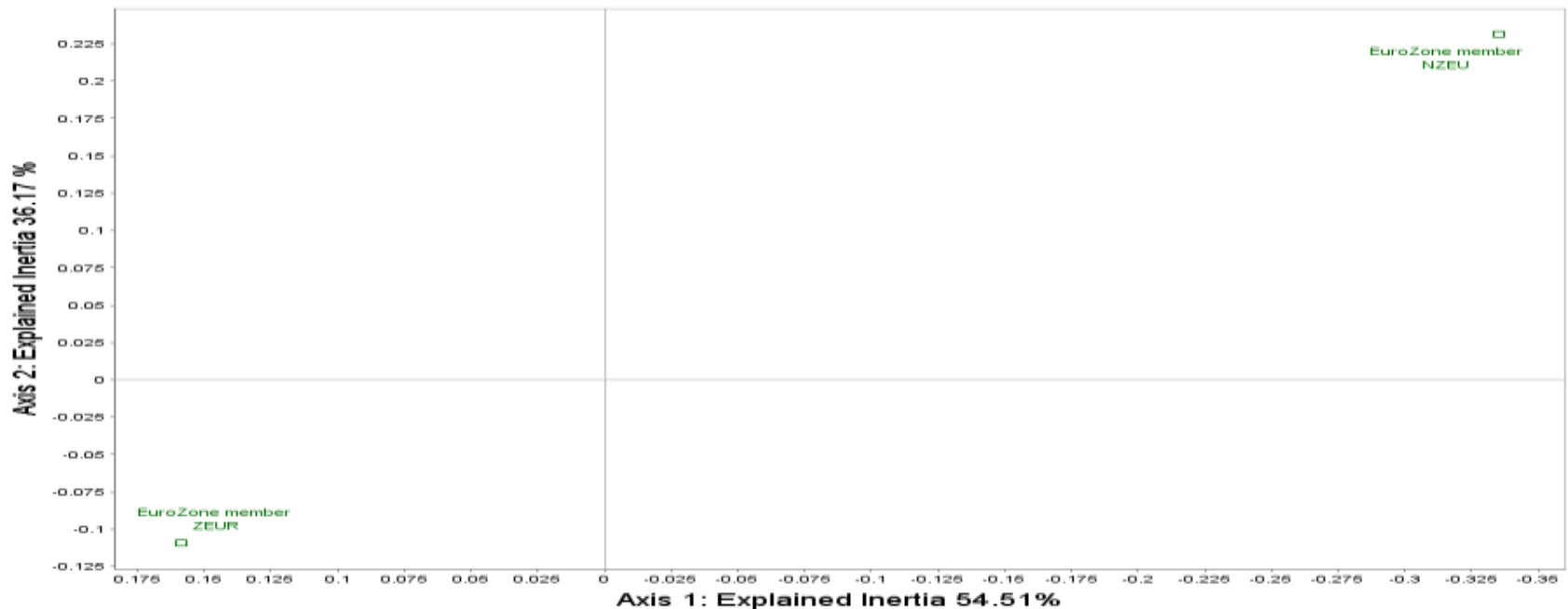
VARIABLES		COORDONNEES					CORRELATIONS VARIABLE-FACTEUR					ANCIENS AXES UNITAIRES				
IDEN	LIBELLE COURT	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0
Taux -	Taux Chômage	0.25	0.62	-0.17	-0.01	0.00	0.25	0.62	-0.17	-0.01	0.00					
Taux -	Taux Croissance	-0.56	-0.21	-0.21	-0.20	0.00	-0.56	-0.21	-0.21	-0.20	0.00					
Taux -	Taux Inflation	-0.05	-0.28	0.24	0.11	0.00	-0.05	-0.28	0.24	0.11	0.00					
Popu -	Population-Millions	0.17	0.11	-0.21	0.25	0.00	0.17	0.11	-0.21	0.25	0.00					
PIB- -	PIB-Milliards-Euros	0.23	0.01	-0.21	0.25	0.00	0.23	0.01	-0.21	0.25	0.00					
PIB/ -	PIB/Hab.-Euros	0.17	-0.40	-0.08	-0.12	0.00	0.17	-0.40	-0.08	-0.12	0.00					



Repr sentation des modalit s de la variable nominale suppl mentaire 'Etat membre' sur le premier plan principal

COORDONNEES ET VALEURS-TEST DES MODALITES AXES 1 A 4

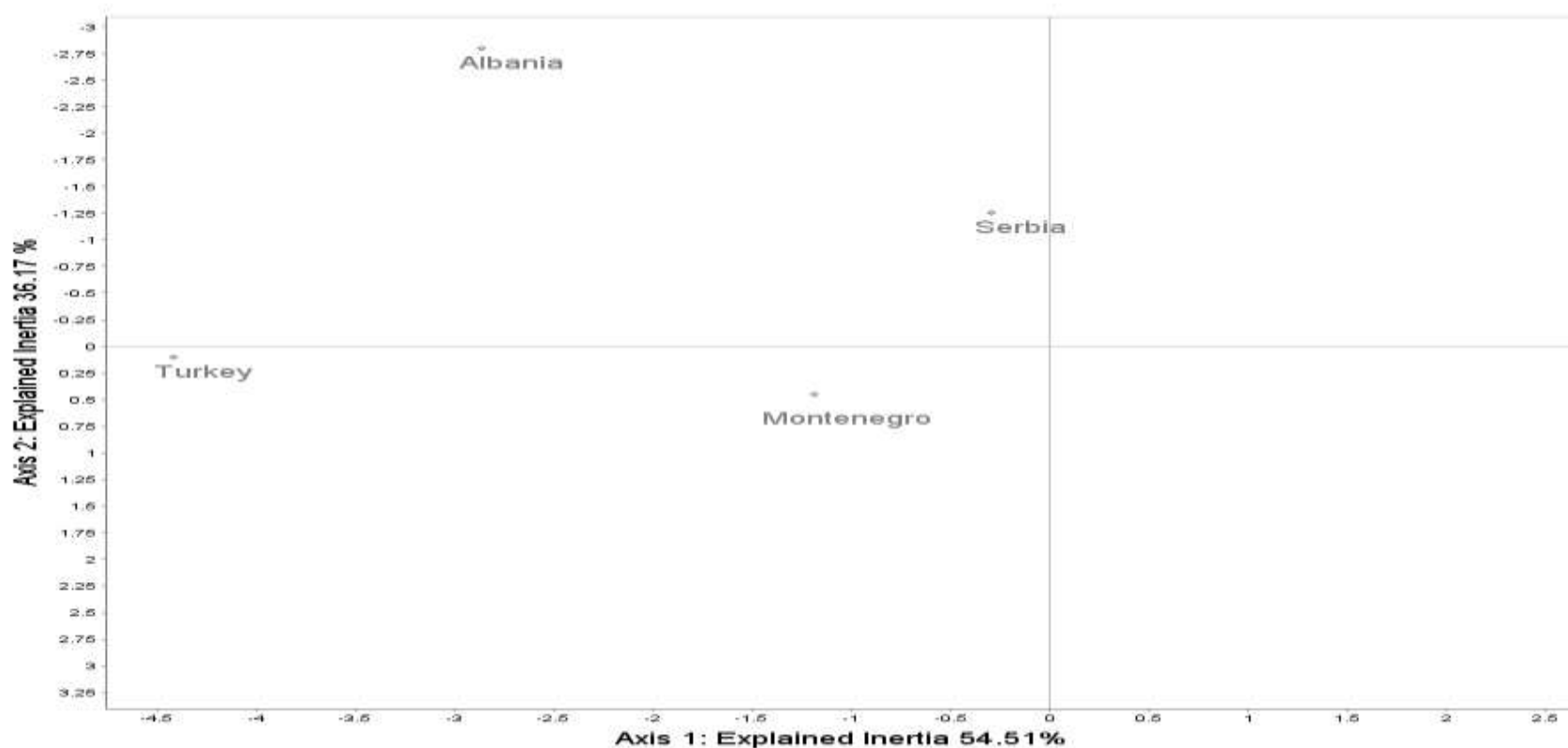
MODALITES			VALEURS-TEST					COORDONNEES					
IDEN - LIBELLE	EFF.	P.ABS	1	2	3	4	0	1	2	3	4	0	DISTO.
2 . Etat membre de la zone Euro													
ZEUR - Zone Euro	19	19.00	0.8	0.7	-1.8	0.7	0.0	0.16	0.11	-0.14	0.00	0.00	0.06
ZNEU - Non-Zone Euro	9	9.00	-0.8	-0.7	1.8	-0.7	0.0	-0.34	-0.23	0.30	0.00	0.00	0.26



Représentation des *individus illustratifs* sur le premier plan principal

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
Albanie	3.57	16.12	-2.86	2.79	0.18	0.28	0.00	0.0	0.0	0.0	0.0	0.0	0.51	0.48	0.00	0.00	0.00
Monténégro	3.57	2.66	-1.19	-0.45	-0.74	-0.71	0.00	0.0	0.0	0.0	0.0	0.0	0.53	0.08	0.21	0.19	0.00
Serbie	3.57	2.24	-0.30	1.25	0.71	0.30	0.00	0.0	0.0	0.0	0.0	0.0	0.04	0.70	0.22	0.04	0.00
Turquie	3.57	21.63	-4.42	-0.10	-1.38	-0.45	0.00	0.0	0.0	0.0	0.0	0.0	0.90	0.00	0.09	0.01	0.00



Limites de l'ACP

Principale faiblesse de l'ACP :

- Sensibilité aux **points extrêmes**. Ce manque de robustesse est notamment lié au rôle central qu'y joue le coefficient de corrélation : les points extrêmes, en perturbant les moyennes et corrélations, perturbent fortement l'analyse - on peut cependant envisager de les mettre en éléments illustratifs.
- l'ACP est inadaptée aux phénomènes **non linéaires** qui plus est en grande dimension. Pour ce genre de problème, d'autres méthodes ont été développées, comme l'ACPN (Analyse en Composantes Principales par Noyau).

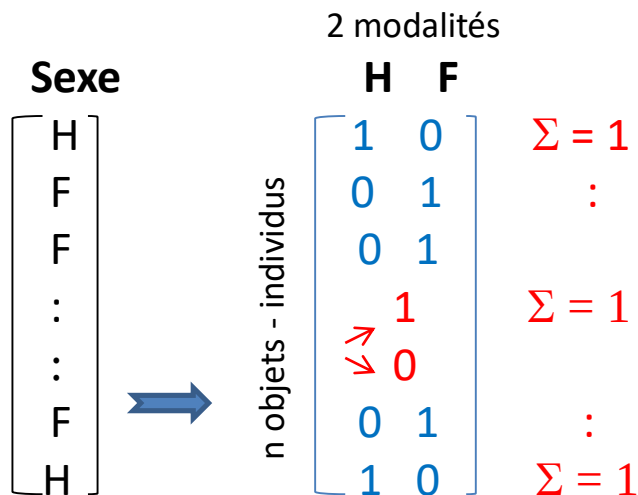
Objectifs de l'analyse factorielle

AFC - Analyse Factorielle des Correspondances

- Décrire les éventuelles associations (correspondances, liens, liaisons, dépendances) entre les modalités de **2 variables qualitatives** $X(p)$ et $Y(q)$ respectivement à p et q modalités.
- L'AFC est une méthode adaptée aux tableaux de contingence ou de dépendance (tableaux croisés) et permet d'étudier les éventuelles relations existant entre deux variables nominales.
- L'ensemble des colonnes du tableau de contingence désigne les modalités d'une variable et l'ensemble des lignes correspond à celles de l'autre variable.
- De ce fait, les lignes et les colonnes du tableau de contingence, qui désignent deux partitions (classifications) d'une même population, jouent des **rôles symétriques**
- L'AFC peut être considérée comme une ACP particulière, avec la distance dite du khi-deux.

AFC - Analyse Factorielle des Correspondances Simples

Codage d'une variable qualitative :



Codage condensé

Tableau disjonctif associé

$$X_{(n,p)} = \begin{matrix} & \begin{matrix} p \text{ modalités} \end{matrix} \\ \begin{bmatrix} 1 & 0 & \dots & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ \vdots & & \vdots & & & \\ \vdots & & \vdots & & & \\ \vdots & & \vdots & & & \\ 0 & \dots & 1 & 0 & \dots & 0 \end{bmatrix} \end{matrix}$$

$$Y_{(n,q)} = \begin{matrix} & \begin{matrix} q \text{ modalités} \end{matrix} \\ \begin{bmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 \\ \vdots & & \vdots & & & \\ \vdots & & \vdots & & & \\ \vdots & & \vdots & & & \\ 1 & 0 & 0 & \dots & 0 & 0 \end{bmatrix} \end{matrix}$$

Tableau de contingence : $TC_{(p,q)} = {}^tX Y$

AFC - Analyse Factorielle des Correspondances

Les données

Tableau de contingence : Croisement de deux variables qualitatives $X(p)$ et $Y(q)$

		Y				
		1	...	j	...	q
X	1	n_{11}		n_{1j}		n_{1q}
	:			:		
	i	n_{i1}	...	n_{ij}	...	n_{iq}
	:			:		
	p	n_{p1}		n_{nj}		n_{pq}
		$n_{.j}$				n

Tableau des effectifs observés n_{ij}

Exemple de tableau de contingence : Créations & reprises d'entreprises en France

Secteur d'activité	Qualification professionnelle du dernier emploi						Total
	A.C.C.Entps.	Cadre	P.Interméd	Employé	Ouvrier	E.S.Qualif	
I.Agro.Alimentaire	291	129	112	508	529	148	1717
Industrie	1109	1250	773	1528	1195	570	6425
Construction	1186	1011	1011	1956	4482	1107	10753
Commerce	4865	4316	3316	7927	3275	3928	27627
Transport	401	337	256	1247	557	312	3110
H.C.Restaurants	1374	684	528	3095	859	1083	7623
Serv. Entreprises	2833	5382	1654	3179	387	1670	15105
Serv. Ménages	1306	1117	1655	3265	560	1228	9131
Total	13365	14226	9305	22705	11844	10046	81491

Tableau de contingence

AFC - Analyse Factorielle des Correspondances Simples

Secteur d'activité	Qualification professionnelle du dernier emploi						Total
	A.C.C.Entps.	Cadre	P.Interméd	Employé	Ouvrier	E.S.Qualif	
I.Agro.Alimentaire	16.95	7.51	6.52	29.59	30.81	8.62	100
Industrie	17.26	19.46	12.03	23.78	18.60	8.87	100
Construction	11.03	9.40	9.40	18.19	41.68	10.29	100
Commerce	17.61	15.62	12.00	28.69	11.85	14.22	100
Transport	12.89	10.84	8.23	40.10	17.91	10.03	100
H.C.Restaurants	18.02	8.97	6.93	40.60	11.27	14.21	100
Serv. Entreprises	18.76	35.63	10.95	21.05	2.56	11.06	100
Serv. Ménages	14.30	12.23	18.13	35.76	6.13	13.45	100
Profil moyen	16.40	17.46	11.42	27.86	14.53	12.33	

Tableau des profils-lignes (%)

Exemple : répartition des créations et reprises d'entreprise dans le secteur du commerce selon les qualifications professionnelles du dernier emploi occupé. Profil-ligne du secteur « commerce » i : $f_j^i = \{n_{ij} / n_i\}$; Profil-ligne global : $f_j = \{f_{.j} = n_{.j} / n\}$

Secteur d'activité	Qualification professionnelle du dernier emploi						Profil moyen
	A.C.C.Entps.	Cadre	P.Interméd	Employé	Ouvrier	E.S.Qualif	
I.Agro.Alimentaire	2.18	0.91	1.20	2.24	4.67	1.47	2.11
Industrie	8.30	8.79	8.31	6.73	10.09	5.67	7.88
Construction	8.87	7.11	10.87	8.61	37.84	11.02	13.20
Commerce	36.40	30.34	35.64	34.91	27.65	39.10	33.90
Transport	3.00	2.37	2.75	5.49	4.70	3.11	3.82
H.C.Restaurants	10.28	4.81	5.67	13.63	7.25	10.78	9.35
Serv. Entreprises	21.20	37.83	17.77	14.00	3.27	16.62	18.54
Serv. Ménages	9.77	7.85	17.79	14.38	4.73	12.22	11.20
Total	100	100	100	100	100	100	100

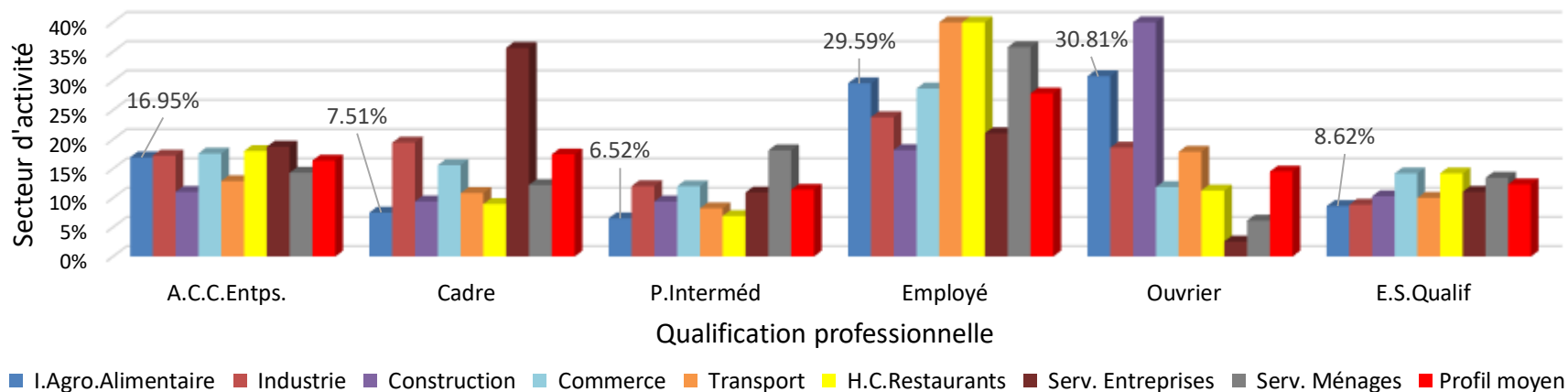
Tableau des profils-colonnes (%)

Exemple : répartition des employés créateurs et repreneurs d'entreprises selon les différents secteurs d'activité.

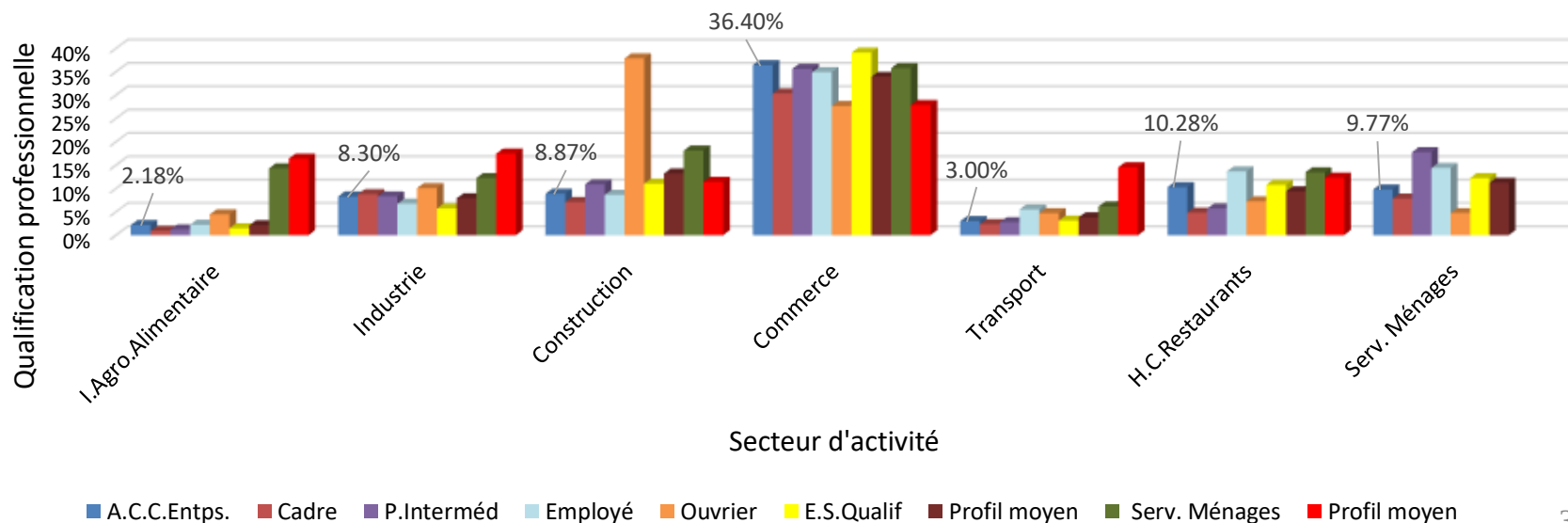
Profil-colonne de l'« employé » j : $f_i^j = \{n_{ij} / n_j\}$; Profil-colonne global : $f_i = \{f_{i.} = n_{i.} / n\}$

AFC - Analyse Factorielle des Correspondances Simples

Secteur d'activité & Qualification professionnelle – Profils-lignes (%)



Secteur d'activité & Qualification professionnelle – Profils-colonnes (%)



Test du khi-deux d'indépendance

Hypothèses statistiques :

H_0 : Les variables X et Y sont indépendantes

H_1 : Les variables X et Y sont liées entre elles

Statistique utilisée :

n_{ij} : effectif observé pour la modalité i de X et j de Y

$\hat{n}_{ij} = n_{i.} \times n_{.j} / n$: effectif théorique (attendu) sous l'hypothèse nulle d'indépendance

$\frac{n_{ij} - \hat{n}_{ij}}{\sqrt{\hat{n}_{ij}}}$: résidu standardisé (centré – réduit : moyenne = 0 et écart-type = 1)

$$\chi^2_{\text{calculé}} = \sum_{i=1}^P \sum_{j=1}^q (n_{ij} - \hat{n}_{ij})^2 / \hat{n}_{ij}$$

$v = (p - 1)(q - 1)$: degrés de liberté.

Décision :

On rejette H_0 au risque α de se tromper si la Probabilité[$\chi^2_v \geq \chi^2_{\text{calculé}}$] $\leq \alpha$

Test du khi-deux d'indépendance

Hypothèses statistiques :

H_0 : Les variables X et Y sont indépendantes

H_1 : Les variables X et Y sont liées entre elles

Conclusion :

- Si l'hypothèse nulle H_0 d'indépendance est rejetée, c'est qu'il y a un lien significatif « global » entre les variables X et Y. On effectue alors une AFC afin de « localiser » ces liens et de les représenter et de les visualiser.
- Si l'hypothèse nulle H_0 d'indépendance est acceptée, l'étude est terminée vue qu'il n'y a pas de dépendances entre les deux variables.

AFC $\Leftrightarrow$ 2 ACP particulières

ACP du tableau des profils-lignes : p modalités de X : $ACP (PL_{(p,q)} ; \chi^2_{(q)} ; D_{(p)})$

- Chaque ligne i du tableau des $PL_{(p,q)}$ a un poids $f_{i.}$ de la matrice des poids $D_{(p)}$
- Distance du $\chi^2_{(q)}$ entre les lignes i et i' : $d^2(f_{i.}, f_{i'.}) = \sum_{j=1}^p (1/f_{.j}) [(n_{ij} / n_{i.}) - (n_{i'j} / n_{i'.})]^2$
- Inertie totale : $\sum_{i=1}^p f_{i.} d^2(f_{i.}, f_{I.}) = \sum \lambda_k = \chi^2 / n = \phi^2$,

ACP du tableau des profils-colonnes : q modalités de Y : $ACP (PC_{(q,p)} ; \chi^2_{(p)} ; D_{(q)})$

- Chaque colonne j du tableau de données $PC_{(q,p)}$ a un poids $f_{.j}$ de la matrice des poids $D_{(q)}$
- Distance du $\chi^2_{(p)}$ entre les colonnes j et j' : $d^2(f_{.j}, f_{.j'}) = \sum_{i=1}^p (1/f_{i.}) [(n_{ij} / n_{.j}) - (n_{ij'} / n_{.j'})]^2$
- Inertie totale : $\sum_{j=1}^q f_{.j} d^2(f_{.j}, f_{.I}) = \sum \lambda_k = \chi^2 / n = \phi^2$

Lien entre les deux ACP

Relations de transitions entre les composantes principales

- $$C^k(i) = (1 / \sqrt{\lambda_k}) \sum_{j=1}^q (n_{ij} / n_{i.}) C'^k(j)$$

Les « lignes-secteurs d'activité » sont au barycentre des « colonnes-qualification professionnelle » à $1 / \sqrt{\lambda_k}$ près.

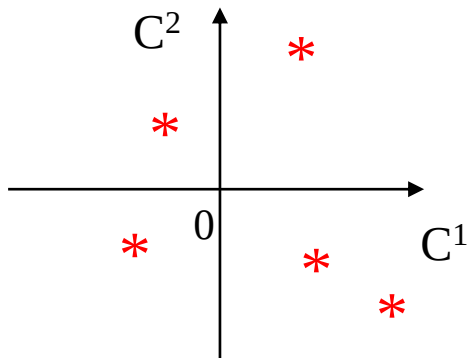
- $$C'^k(j) = (1 / \sqrt{\lambda_k}) \sum_{i=1}^p (n_{ij} / n_{.j}) C^k(i)$$

Les « colonnes-qualification professionnelle » sont au barycentre des « lignes-secteurs d'activité » à $1 / \sqrt{\lambda_k}$ près.

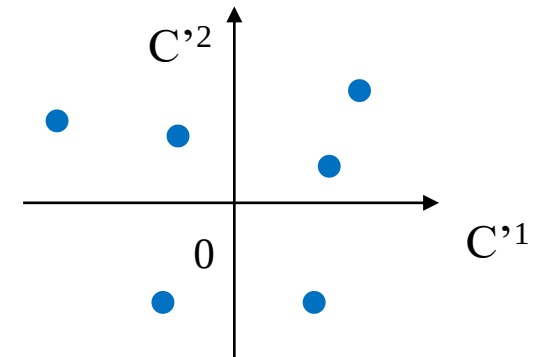
- L'AFC est une analyse symétrique, d'inertie totale « information association à résumer » est égale au $\phi^2 = \sum \lambda_k$ où, $\lambda_k = \text{Var}(C^k) = \text{Var}(C'^k)$.
- En pratique, une seule analyse est effectuée, l'autre est déduite par ces formules de transition.

AFC : représentations

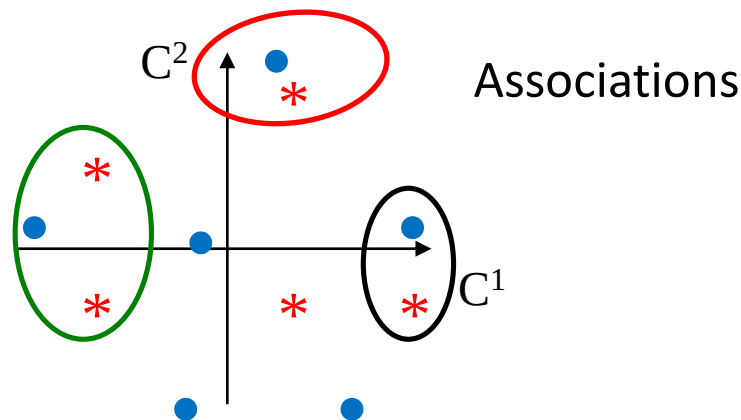
ACP des profils-lignes



ACP des profils-colonnes



Représentation simultanée « barycentrique »



AFC - Analyse Factorielle des Correspondances Simples

INSEE : Les créateurs & repreneurs d'entreprises

Secteur d'activité – Qualification professionnelle

EDITION DU TABLEAU CROISE

TABEAU 1 EN LIGNE : Secteur d'activité

POIDS TOTAL : 81491.

EN COLONNE : Qualification professionnelle du dernier emploi

POIDS % COLONNE	A.C.C.Entps.	Cadre	P.Intermédia	Employé	Ouvrier	E.S.Qualific	ENSEMBLE
% LIGNE							
I.Agro.Alimentaire	291 2.18 16.95	129 0.91 7.51	112 1.20 6.52	508 2.24 29.59	529 4.67 30.81	148 1.47 8.62	1717 2.11 100.00
Industrie	1109 8.30 17.26	1250 8.79 19.46	773 8.31 12.03	1528 6.73 23.78	1195 10.09 18.60	570 5.67 8.87	6425 1 7.88 100.00
Construction	1186 8.87 11.03	1011 7.11 9.40	1011 10.87 9.40	1956 8.61 18.19	4482 37.84 41.68	1107 11.02 10.29	10753 13.20 100.00
Commerce	4865 36.40 17.61	4316 30.34 15.62	3316 35.64 12.00	7927 34.91 28.69	3275 27.65 11.85	3928 39.10 14.22	27627 33.90 100.00
Transport	401 3.00 12.89	337 2.37 10.84	256 2.75 8.23	1247 5.49 40.10	557 4.70 17.91	312 3.11 10.03	3110 3.82 100.00
H.C.Restaurants	1374 10.28 18.02	684 4.81 8.97	528 5.67 6.93	3095 13.63 40.60	859 7.25 11.27	1083 10.78 14.21	7623 9.35 100.00
Serv. Entreprises	2833 21.20 18.76	5382 37.83 35.63	1654 17.77 10.95	3179 14.00 21.05	387 3.27 2.56	1670 16.62 11.06	15105 18.54 100.00
Serv. Menages	1306 9.77 14.30	1117 7.85 12.23	1655 17.79 18.13	3265 14.38 35.76	560 4.73 6.13	1228 12.22 13.45	9131 11.20 100.00
ENSEMBLE	13365 100.00 16.40	14226 100.00 17.46	9305 100.00 11.42	22705 100.00 27.86	11844 100.00 14.53	10046 100.00 12.33	81491 100.00 100.00

KHI2 = 14619.49 / 35 DEGRES DE LIBERTE / 0 EFFECTIFS THEORIQUES INFERIEURS A 5
 PROBA (KHI2 > 14619.49) = 0.0001 / V.TEST = 99.10

AFC - Analyse Factorielle des Correspondances Simples

Secteur d'activité	Qualification professionnelle du dernier emploi						Total
	A.C.C.Entps.	Cadre	P.Interméd	Employé	Ouvrier	E.S.Qualif	
I.Agro.Alimentaire	291	129	112	508	529	148	1717
Industrie	1109	1250	773	1528	1195	570	6425
Construction	1186	1011	1011	1956	4482	1107	10753
Commerce	4865	4316	3316	7927	3275	3928	27627
Transport	401	337	256	1247	557	312	3110
H.C.Restaurants	1374	684	528	3095	859	1083	7623
Serv. Entreprises	2833	5382	1654	3179	387	1670	15105
Serv. Ménages	1306	1117	1655	3265	560	1228	9131
Total	13365	14226	9305	22705	11844	10046	81491

Tableau de contingence

Test d'indépendance du Khi-deux :

Hypothèses statistiques :

- H_0 : indépendance du secteur d'activité et de la qualification professionnelle du dernier emploi
- H_1 : dépendance

Khi²_{calculé} (Value)

KHI2 = 14620.95 / 35 DEGREES OF FREEDOM / 0 EXPECTED FREQUENCIES LESS THAN 5
 PROB. (KHI2 > 14620.9546) = 0.0001

Probabilité critique (p-Value) < $\alpha = 5\%$: rejet de H_0

Conclusion : Au seuil de signification $\alpha = 5\%$, on peut conclure qu'il y a un lien « global » entre la qualification professionnelle du dernier emploi occupé et le secteur d'activité dans lequel ils ont créé ou repris leur entreprise. 37

AFC - Analyse Factorielle des Correspondances Simples

SIMPLE CORRESPONDENCE ANALYSIS

EIGENVALUES

COMPUTATIONS PRECISION SUMMARY : TRACE BEFORE DIAGONALISATION.. 0.1794 = $\phi^2 = \chi^2_{\text{obs}} / n$: Information à résumer « Association »

SUM OF EIGENVALUES..... 0.1794

HISTOGRAM OF THE FIRST 5 EIGENVALUES

NUMBER	EIGENVALUE	PERCENTAGE	CUMULATED PERCENTAGE	
1	0.1244	69.33	69.33	*****
2	0.0442	24.65	93.97	*****
3	0.0075	4.19	98.16	*****
4	0.0025	1.41	99.57	**
5	0.0008	0.43	100.00	*

2 axes

COORDONNEES, CONTRIBUTIONS DES FREQUENCES SUR LES AXES 1 A 5

FREQUENCES ACTIVES

FREQUENCES			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDENTIFICATEUR	P.REL	DISTO	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
ACCE - A.C.C.ENTREPRISE	16.40	0.02	0.13	0.00	0.06	0.06	0.05	2.1	0.0	6.6	20.5	54.4	0.64	0.00	0.12	0.13	0.10
CADR - CADRES	17.46	0.28	0.39	0.36	0.02	-0.03	-0.01	21.1	49.8	1.3	5.8	4.5	0.54	0.45	0.00	0.00	0.00
PINT - PROF.INTERMEDIAIRE	11.42	0.07	0.09	-0.06	-0.23	-0.02	0.02	0.7	0.8	82.3	1.3	3.4	0.12	0.05	0.83	0.00	0.00
EMPL - EMPLOYES	27.86	0.06	0.04	-0.24	0.05	-0.05	-0.01	0.4	36.7	9.3	24.7	1.1	0.03	0.90	0.04	0.03	0.00
OUVR - OUVRIERS	14.53	0.67	-0.80	0.17	0.01	-0.01	0.00	75.4	9.5	0.1	0.4	0.0	0.96	0.04	0.00	0.00	0.00
ESQU - E.SQUALIFICATION	12.33	0.03	0.05	-0.11	-0.01	0.10	-0.05	0.2	3.2	0.3	47.3	36.6	0.09	0.44	0.01	0.37	0.09

COORDONNEES, CONTRIBUTIONS ET COSINUS CARRES DES INDIVIDUS

AXES 1 A 5

INDIVIDUS			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDENTIFICATEUR	P.REL	DISTO	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
I.A. Alimentaire	2.11	0.27	-0.49	-0.02	0.14	-0.04	0.09	4.1	0.0	5.4	1.1	21.5	0.89	0.00	0.07	0.00	0.03
Industrie	7.88	0.03	-0.08	0.13	-0.02	-0.04	0.07	0.4	3.0	0.4	5.0	57.7	0.19	0.56	0.01	0.05	0.19
Construction	13.20	0.60	-0.75	0.21	-0.03	0.00	-0.02	59.1	13.2	1.5	0.0	9.9	0.92	0.07	0.00	0.00	0.00
Commerce	33.90	0.01	0.05	-0.07	-0.01	0.06	0.00	0.7	4.1	0.9	41.6	0.1	0.23	0.48	0.02	0.28	0.00
Transport	3.82	0.11	-0.16	-0.20	0.12	-0.16	-0.03	0.8	3.6	7.6	36.7	5.5	0.23	0.39	0.14	0.23	0.01
HCRestaurants	9.35	0.13	-0.01	-0.31	0.17	0.00	-0.01	0.0	20.8	38.0	0.1	0.7	0.00	0.76	0.24	0.00	0.00
S.Entreprises	18.54	0.31	0.47	0.30	0.03	-0.02	-0.01	32.8	36.7	2.1	2.3	4.0	0.71	0.28	0.00	0.00	0.00
S.Ménages	11.20	0.13	0.15	-0.27	-0.17	-0.05	-0.01	2.2	18.5	44.1	13.3	0.6	0.18	0.56	0.23	0.02	0.00

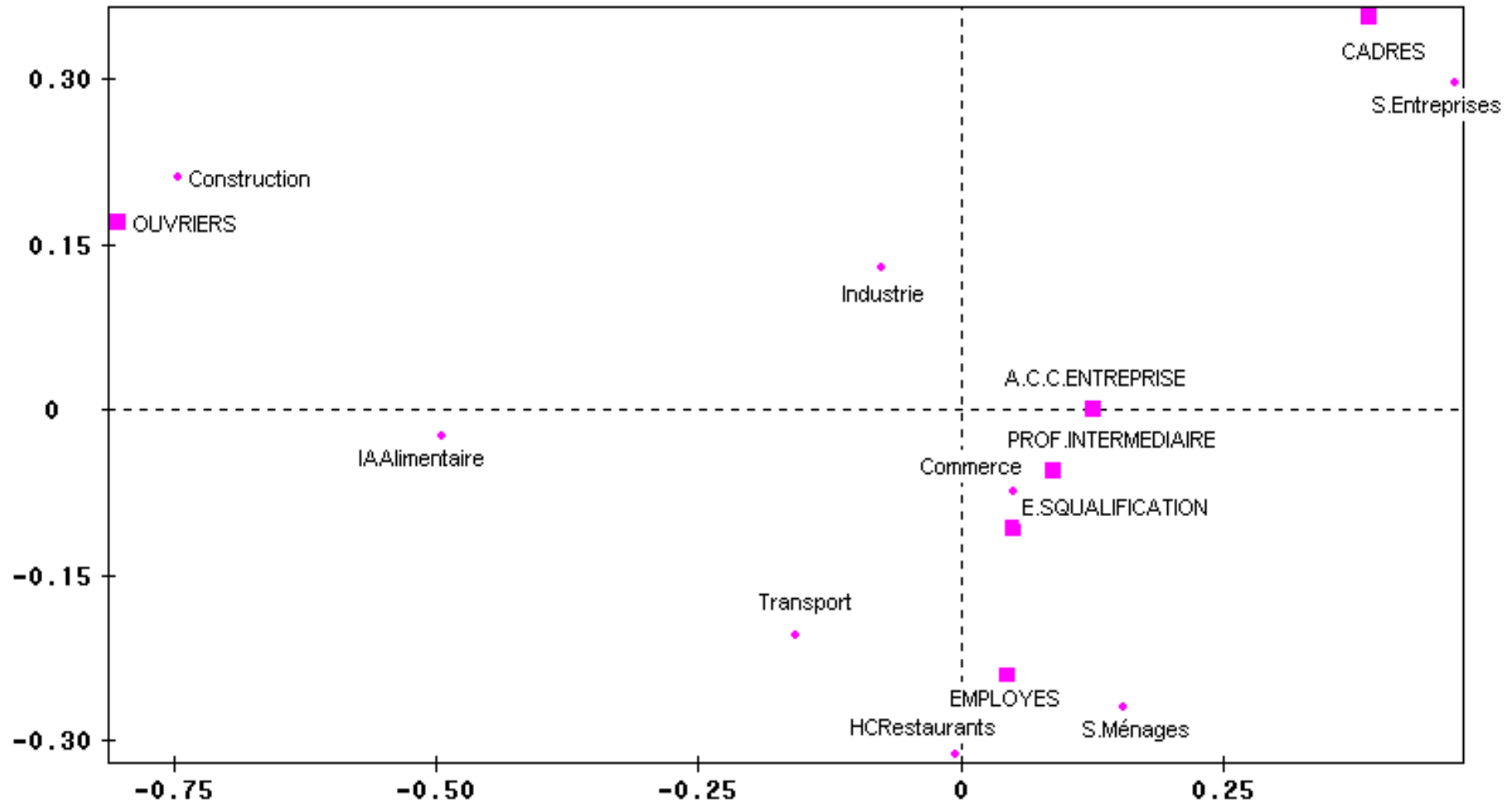
AFC - Analyse Factorielle des Correspondances Simples

Créateurs - Repreneurs d'entreprises

Qualification & Activité de l'entreprise

Représentation simultanée des modalités actives – Premier plan principal

Facteur 2 : Inertie expliquée - 24.65 %



Facteur 1 : Inertie expliquée - 69.33 %

Objectifs de l'analyse factorielle

ACM – Analyse des Correspondances Multiples

- L'ACM peut être considérée comme une extension ou une généralisation de l'AFC au cas où **plus de 2 variables qualitatives** sont mises en correspondance .
- Décrire les éventuelles associations (correspondances, liens, liaisons, dépendances) entre l'ensemble des modalités des K ($K > 2$) **variables qualitatives**.
- L'ACM est une analyse des correspondances simple appliquée non plus à une table de contingence, mais à un **tableau disjonctif complet** (**vastes tableaux de données binaires**).
- L'ACM est une méthode adaptée au **tableau de Burt** « **juxtaposition de tableaux de contingence** » et permet d'étudier les éventuelles relations existant entre l'ensemble des variables nominales.
- L'ACM est la méthode factorielle la plus utilisée, c'est notamment le cas pour l'étude des fichiers d'enquêtes socio-économiques, enquêtes par questionnaires où chaque question définit une variable qualitative dont les modalités sont les différentes réponses possibles (questions à réponse unique).

Principales matrices des données

K variables qualitatives ($K > 2$)

avec respectivement ($q_1, q_2, \dots, q_k, \dots, q_K$) modalités : $q = \sum_{K=1}^K q_K$

($y^1, \dots, y^k, \dots, y^K$)

$$\begin{aligned}
 Y_{(n, q)} &= \left[\begin{array}{c|c|c} \left[\begin{array}{c} y_1^1 \\ \vdots \\ y_i^1 \\ \vdots \\ y_n^1 \end{array} & \begin{array}{c} y_1^j \\ \vdots \\ y_i^j \\ \vdots \\ y_n^j \end{array} & \begin{array}{c} y_1^{q_1} \\ \vdots \\ y_i^{q_1} \\ \vdots \\ y_n^{q_1} \end{array} \end{array} \right] \dots \left[\begin{array}{c|c|c} \left[\begin{array}{c} y_1^1 \\ \vdots \\ y_i^1 \\ \vdots \\ y_n^1 \end{array} & \begin{array}{c} y_1^j \\ \vdots \\ y_i^j \\ \vdots \\ y_n^j \end{array} & \begin{array}{c} y_1^{q_k} \\ \vdots \\ y_i^{q_k} \\ \vdots \\ y_n^{q_k} \end{array} \end{array} \right] \dots \left[\begin{array}{c|c|c} \left[\begin{array}{c} y_1^1 \\ \vdots \\ y_i^1 \\ \vdots \\ y_n^1 \end{array} & \begin{array}{c} y_1^j \\ \vdots \\ y_i^j \\ \vdots \\ y_n^j \end{array} & \begin{array}{c} y_1^{q_K} \\ \vdots \\ y_i^{q_K} \\ \vdots \\ y_n^{q_K} \end{array} \end{array} \right] \\
 &= [Y_{(n, q_1)}^1 \dots Y_{(n, q_k)}^k \dots Y_{(n, q_K)}^K]
 \end{aligned}$$

Matrice d'indicatrices : matrices binaires associées aux q modalités

Tableau disjonctif complet :
 « juxtaposition de tableaux binaires »

ACM – Analyse des Correspondances Multiples

Principales matrices des données

$$Y_{(n, q)} = [Y_{(n, q_1)}^1 \cdots Y_{(n, q_k)}^k \cdots Y_{(n, q_K)}^K]$$

Matrice des q variables binaires

$$\forall (k, l)_{k, l=1, K} \quad k \neq l$$

Matrices symétriques :

$$N_{(q_k, q_l)} = {}^t Y^k Y^l$$

Tableau de contingence

$$B_{(q)} = {}^t Y Y$$

Tableau de Burt

$$D_w(n) = \begin{pmatrix} w_1 & & 0 \\ & \ddots & \\ 0 & & w_n \end{pmatrix} = \frac{1}{n} \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} = \frac{1}{n} I_n$$

$\forall k=1, K$

$$D_w(y^k)_{(q_k)} = \begin{pmatrix} \frac{n_1^k}{n} & & 0 \\ & \ddots & \\ 0 & & \frac{n_{q_k}^k}{n} \end{pmatrix} = {}^t Y^k D_w Y^k$$

Matrice diagonale des poids des n individus

Matrice diagonale des poids des q_k modalités de Y^k

$$D_w(Y)_{(q)} = \begin{pmatrix} D_w(y^1) & & 0 \\ & \ddots & \\ 0 & & D_w(y^K) \end{pmatrix} = {}^t Y D_w Y$$

Matrice diagonale des poids des q modalités

- Première façon : ACM peut être considérée comme une ACP particulière, elle consiste à effectuer l'ACP suivante :

$$ACM \longleftrightarrow ACP \left(\textcolor{red}{Y} \textcolor{red}{D}_w^{-1}(Y) ; \frac{1}{K} \textcolor{blue}{D}_w(Y) ; \frac{1}{n} \textcolor{green}{I}_n \right)$$

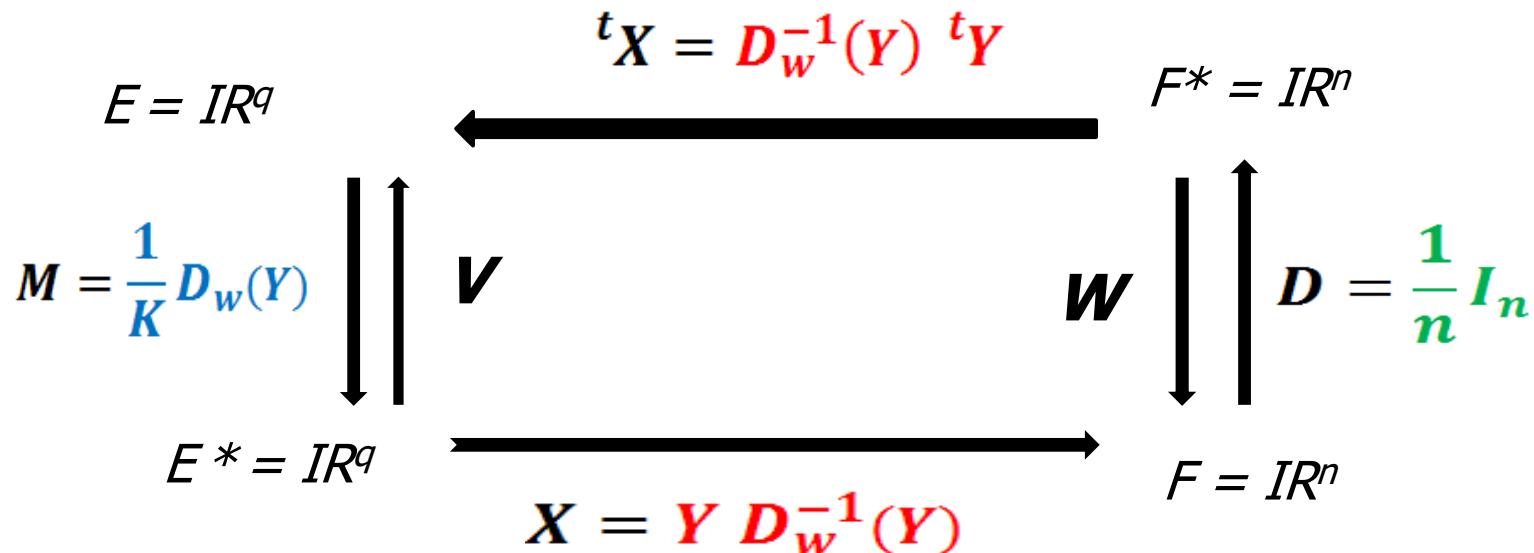


Schéma de dualité

- Inertie totale : $\sum \lambda_j = \text{trace}(WD) = \text{trace}(VM)$
 $= \text{trace}({}^t X D X M) = \dots = (1/q) \text{trace}(I_q) = q / K$
- Les **composantes principales** de l'ACM sont des **variables quantitatives** qui possèdent la propriété de maximiser $\sum_{k=1,K} \eta^2(c, y^k)$: la somme des des rapports de corrélations avec les K variables qualitatives.
- Rappelons que l'ACP normée de p variables quantitatives, avec $M = D_{1/\sigma^2}$ et d'inertie totale $= p$, les composantes principales c ont la propriété de maximiser $\sum_{j=1,p} r^2(c, x^j)$: la somme des carrés des coefficients de corrélation linéaire.
- Ainsi, l'ACM est équivalente (à un coefficient près) à une ACP où les p variables sont **qualitatives**.

- *Deuxième façon : Une autre façon plus pratique de présenter l'ACM, est d'appliquer l'AFC sur le **tableau symétrique de Burt** B .*
- *Maximum $r = q - K$ dimensions (axes factoriels) pour les représentations graphiques de l'ACM. L'inertie totale est égale à $\sum \lambda_j = (q / K) - 1$. Les valeurs propres ou singulières $\lambda_1, \dots, \lambda_r$, sont les moments principaux d'inertie.*
- *L'analyse des correspondances du tableau d'indicateurs Y (première façon) **est équivalente** à celle effectuée sur le tableau de Burt B (deuxième façon). Cependant, les moments principaux λ_j ($j=1, r$) de l'analyse du tableau de Burt sont les carrés de ceux de l'analyse du tableau Y .*

Quelques remarques (1)

- Les *variables actives* doivent être *homogènes*; elles concernent un même thème à décrire (*interprétation interne*). Ne pas considérer comme variables actives des variables de différents thèmes. Par exemple, des variables signalétique (âge, sexe, profession, etc.) et des variables d'opinions ou de comportements.
- Toutes ou une partie des autres variables peuvent être éventuellement positionnées en éléments-variables illustratives (*interprétation externe*) .
- Les variables illustratives peuvent être *mixtes* (qualitatives et/ou quantitatives) et non nécessairement homogènes. Elles *ne participent pas à l'analyse*.

Quelques remarques (2)

- Les variables illustratives sont projetées *a posteriori* sur les plans factoriels déterminés par les variables actives. Ainsi, l'interprétation de chaque variable illustrative doit être comparée à l'ensemble des variables actives.
- Comme pour l'AFC, les *outils d'aide à l'interprétation* des résultats sont : la somme des contributions sur chaque axe, les contributions des modalités et les cosinus carrés sur chaque axe factoriel. L'interprétation des variables illustratives s'effectue comme en ACP.
- Compte tenu de la nature particulière du tableau disjonctif qui conduit à un nuage très éparpillé, les *pourcentages d'information* apportés par chaque axe sont en général *très faibles*. Cela rend un peu plus délicat le choix du nombre pertinent d'axes factoriels à retenir.

ACM – Analyse des Correspondances Multiples

Statistiques sommaires & Distributions marginales des variables qualitatives

Exemple

Source:
Les créations d'entreprises –
INSEE - France

Variables actives
Signalétique des
créateurs-repreneurs
d'entreprises
'Thème homogène'

Variables
supplémentaires

Variable	Effectif	Pourcentage	Effectif Cumulé	Pourcentage Cumulé
1- GENRE				
Masculin	56742	69,63	56742	69,63
Féminin	24749	30,37	81491	100,00
2- AGE DE L'ENTREPRENEUR				
Moins de 25 ans	7019	8,61	7019	8,61
25-39 ans	42678	52,37	49697	60,98
40 ans et plus	31794	39,02	81491	100,00
3- NIVEAU D'EDUCATION				
Unqualified	17519	21,50	17519	21,50
Intermediate Level	7054	8,66	24573	30,15
Second.School	23826	29,24	48399	59,39
Technical Bachelor	7446	9,14	55845	68,53
Generalist Bachelor	5264	6,46	61109	74,99
Technology diploma	6411	7,87	67520	82,86
After 2 years univ.	3254	3,99	70774	86,85
After 3 years univ.	10717	13,15	81491	100,00
4- QUALIFICATION DERNIER EMPLOI				
Artisan-Commerçant-Chef d'entreprise	13365	16,40	13365	16,40
Cadre	14226	17,46	27591	33,86
Profession Intermédiaire	9305	11,42	36896	45,28
Employé	22705	27,86	59601	73,14
Ouvrier	11844	14,53	71445	87,67
Sans Qualification	10046	12,33	81491	100,00
5- SITUATION AVANT CREATION-REPRISE				
Actif	34270	42,05	34270	42,05
Sans emploi - Inactif	47221	57,95	81491	100,00
6- SECTEUR D'ACTIVITE				
Indus.Agro.Alimentaire	6425	7,88	6425	7,88
Industrie	10753	13,20	17178	21,08
Construction	27627	33,90	44805	54,98
Commerce	3110	3,82	47915	58,80
Transports	7623	9,35	55538	68,15
Hôtels-Cafés-Restaurants	9131	11,20	64669	79,36
Services aux entreprises	15105	18,54	79774	97,89
Services aux ménages	1717	2,11	81491	100,00

ACM – Analyse des Correspondances Multiples

Statistiques sommaires & Distributions marginales des variables qualitatives

Exemple

Source:
Les créations d'entreprises –
INSEE - France

Variables actives
Signalétique des
créateurs-repreneurs
d'entreprises
'Thème homogène'

Variables
supplémentaires

Modality label	Frequency	Percent	Cumulative Frequency	Cumulative Frequency
1- GENDER				
Man	56742	69,63	56742	69,63
Woman	24749	30,37	81491	100,00
2- AGE OF THE ENTREPRENEUR				
Under 25 years	7019	8,61	7019	8,61
25-39 years old	42678	52,37	49697	60,98
40 years old – over	31794	39,02	81491	100,00
3- EDUCATIONAL LEVEL				
Unqualified	17519	21,50	17519	21,50
Intermediate Level	7054	8,66	24573	30,15
Second.School	23826	29,24	48399	59,39
Technical Bachelor	7446	9,14	55845	68,53
Generalist Bachelor	5264	6,46	61109	74,99
Technology diploma	6411	7,87	67520	82,86
After 2 years univ.	3254	3,99	70774	86,85
After 3 years univ.	10717	13,15	81491	100,00
4- QUALIFICATION LAST JOB				
Craftsman-Manager	13365	16,40	13365	16,40
Executive	14226	17,46	27591	33,86
Skilled worker	9305	11,42	36896	45,28
Employee	22705	27,86	59601	73,14
Worker	11844	14,53	71445	87,67
Inactive people	10046	12,33	81491	100,00
5- SITUATION BEFORE CREATION				
Active	34270	42,05	34270	42,05
Unemployed-Inactive	47221	57,95	81491	100,00
6- BRANCH OF INDUSTRY				
Food industry	6425	7,88	6425	7,88
Industry	10753	13,20	17178	21,08
Construction	27627	33,90	44805	54,98
Trade	3110	3,82	47915	58,80
Transports	7623	9,35	55538	68,15
Catering	9131	11,20	64669	79,36
Business Services	15105	18,54	79774	97,89
Household Services	1717	2,11	81491	100,00

ACM – Analyse des Correspondances Multiples

Résultats (1)

		GENDER		AGE			NIVEAU D'ÉDUCATION								QUALIFICATION DERNIER EMPLOI											
		Masculin	Féminin	Moins de 25 ans	25-39 ans	40 ans et plus	Unqualified	Intermediate Level	Second School	Technical Bachelor	General Bachelor	Technology diploma	After 2 years univ.	After 3 years univ.	Art_Comm-Chef Ent	Cadre	Prof. Intermédiaire	Employé	Ouvrier	Sans Qualification						
GENRE	Masculin	56742		Table de contingence Age x Sexe																						
	Féminin	0	24749																							
AGE	Moins de 25 ans	4778	2241																		7019					
	25-39 ans	30307	12371	0	42670																					
	40 ans et plus	21657	10137	0	0	31794																				
NIVEAU D'ÉDUCATION	Unqualified	12004	5515	1506	7221	8792	17519																			
	Intermediate Level	4381	2673	671	3507	2076	0	7054																		
	Second.School	17772	6054	2527	13782	7517	0	0	23826																	
	Technical Bachelor	5290	2156	811	3745	2890	0	0	0	7446																
	General Bachelor	3042	2222	462	2000	2002	0	0	0	0	5264															
	Technology diploma	4597	1014	605	3758	2040	0	0	0	0	0	6411														
	Bac+2 univ.	1867	1307	151	1570	1533	0	0	0	0	0	0	3254													
	Bac+3 et plus	7789	2928	286	6295	4136	0	0	0	0	0	0	0	10717												
QUALIFICATION DERNIER EMPLOI	Art-Com-Chef Ent	10220	3145	290	5080	7995	291	1109	1186	4065	401	1374	2833	1306	13365											
	Cadre	11460	2766	65	6338	7823	129	1250	1011	4316	337	604	5382	1117	0	14226										
	Prof. Intermédiaire	6009	2496	355	4754	4196	112	773	1011	3316	256	528	1654	1655	0	0	9305									
	Employé	12675	10030	2341	14453	5911	508	1528	1956	7927	1247	3095	3179	3265	0	0	0	22705								
	Ouvrier	10049	995	1336	7426	3002	529	1195	4482	3275	557	859	387	560	0	0	0	0	11844							
	Sans Qualification	4729	5317	2632	4627	2707	148	570	1107	3928	312	1083	1670	1228	0	0	0	0	0	10046						

Table de contingence
Age x Sexe

Résultats (2)

Histogramme des 15 premières valeurs propres

Numéro	Valeur propre	Pourcentage	Pourcentage Cumulé	Histogramme
1	0,5479	14,61	14,61	*****
2	0,4616	12,31	26,92	*****
3	0,3683	9,82	36,74	*****
4	0,3116	8,31	45,05	*****
5	0,2651	7,07	52,12	*****
6	0,2411	6,43	58,55	*****
7	0,2348	6,26	64,81	*****
8	0,2344	6,25	71,06	*****
9	0,2198	5,86	76,92	*****
10	0,2108	5,62	82,54	*****
11	0,1898	5,06	87,60	*****
12	0,1789	4,77	92,37	*****
13	0,1354	3,61	95,98	*****
14	0,0949	2,53	98,51	*****
15	0,0559	1,49	100,00	***

Nombre total d'axes
factoriels :
 $q - K = 19 - 4 = 15$

Inertie = Somme des valeurs
propres = $(q / K) - 1 = 3,75$

ACM – Analyse des Correspondances Multiples

Résultats (3)

Variables actives - Signalétique

Genre - Age - Diplôme - CSP du dernier emploi

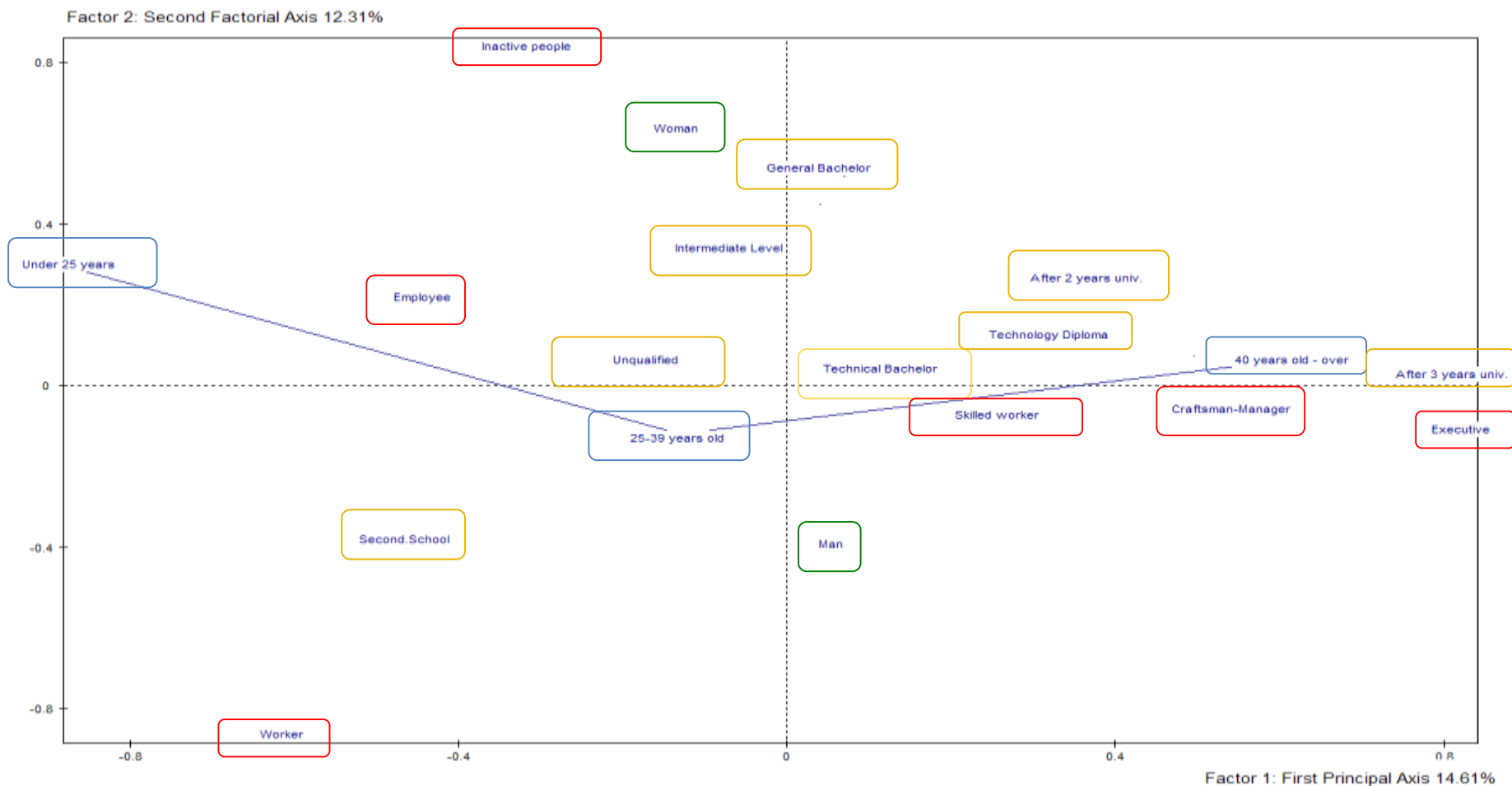
ACTIVE MODALITIES			COORDINATES				CONTRIBUTIONS				SQUARED COSINES			
Labels	Relative Weight (%)	Squared Distance to the Origin	Axis 1	Axis 2	Axis 3	Axis 4	Axis 1	Axis 2	Axis 3	Axis 4	Axis 1	Axis 2	Axis 3	Axis 4
1- GENRE														
Masculin	17,41	0,12	0,06	-0,38	-0,40	-0,70	0,40	13,70	0,20	1,20	0,03	0,62	0,01	0,05
Féminin	7,59	0,64	-0,13	0,63	0,08	0,17	0,90	24,60	0,50	2,70	0,03	0,62	0,01	0,05
Cumulated Contribution =							1,30	38,30	0,70	3,90				
2- AGE														
Under 25 years	2,15	2,88	-0,87	0,29	-0,30	-0,82	11,00	1,20	1,90	17,50	0,28	0,06	0,03	0,23
25-39 years old	13,09	0,24	-0,13	-0,12	0,30	-0,01	1,00	1,40	12,20	0,00	0,07	0,06	0,38	0,00
40 years old - over	9,75	0,43	0,56	0,05	-0,34	0,20	9,90	0,20	11,60	4,70	0,32	0,01	0,27	0,09
Cumulated Contribution =							21,90	2,80	25,70	22,20				
3- EDUCATIONAL LEVEL														
Unqualified	5,37	0,96	-0,17	0,06	-0,68	0,08	1,00	0,10	25,30	0,50	0,03	0,00	0,48	0,01
Intermediate Level	2,16	2,65	-0,06	0,34	-0,06	0,24	0,10	2,00	0,10	1,50	0,00	0,04	0,00	0,02
Second.School	7,31	0,66	-0,55	-0,47	0,15	0,04	10,10	8,30	1,60	0,10	0,30	0,26	0,03	0,00
Technical Bachelor	2,28	2,50	0,13	0,04	0,19	0,18	0,30	0,00	0,90	0,90	0,01	0,00	0,02	0,01
General Bachelor	1,61	3,66	0,07	0,52	0,31	0,43	0,10	3,60	1,60	3,60	0,00	0,20	0,03	0,05
Technology diploma	1,97	3,00	0,33	0,12	0,57	-0,38	1,00	0,20	6,60	3,50	0,04	0,01	0,11	0,05
After 2 years univ.	1,00	6,11	0,37	0,26	0,16	0,18	2,30	1,70	0,20	0,40	0,05	0,03	0,00	0,01
After 3 years univ.	3,29	1,82	0,81	0,02	0,14	-0,55	14,80	0,00	0,70	11,90	0,36	0,00	0,01	0,17
Cumulated Contribution =							29,70	15,90	37,00	22,40				
4- QUALIFICATION LAST JOB														
Craftsman-Manager	4,10	1,34	0,54	-0,06	-0,57	0,43	5,40	0,10	13,30	9,00	0,30	0,00	0,24	0,14
Executive	4,36	1,36	0,83	-0,11	0,26	-0,37	20,60	0,40	3,10	7,10	0,51	0,01	0,05	0,10
Skilled worker	2,85	1,97	0,26	-0,07	0,32	0,13	1,30	0,10	2,90	0,60	0,03	0,00	0,05	0,01
employee	6,97	0,71	-0,41	0,21	0,36	0,36	8,20	2,50	9,30	11,10	0,29	0,06	0,18	0,19
Worker	3,63	1,68	-0,61	-0,87	-0,31	-0,18	9,20	22,10	3,60	1,40	0,28	0,45	0,06	0,02
Inactive people	3,08	1,96	-0,35	0,84	-0,36	-0,78	2,60	17,60	4,20	22,40	0,06	0,36	0,07	0,31
Cumulated Contribution =							47,30	42,80	36,40	51,60				

ACM – Analyse des Correspondances Multiples

Active Demographic variables

Gender – Age – Educational Level – Qualification last job

Representation on the first principal plane



ACM – Analyse des Correspondances Multiples

Quelques remarques

- *Interprétation des modalités supplémentaires : **cosinus carrés** ou **valeurs test** des tables. Les modalités de chaque variable supplémentaire sont interprétées relativement aux variables actives.*
- ***Valeurs test** : qualité de la représentation de la modalité supplémentaire sur l'axe factoriel, elle est considérée comme significative si la valeur test est supérieure en valeur absolue à 1.96 (fratiles de la loi normale avec un risque d'erreur de 5%).*
- ***Effet Guttman** : apparaît lorsque la variable est ordinale ou discontinue (variable continue mise en classes de valeurs). Dans ce cas, le premier axe factoriel oppose généralement les modalités (classes) extrêmes. Les modalités intermédiaires se retrouvent sur le second axe factoriel.*
- *Le graphique suivant illustre la projection des modalités des deux variables supplémentaires (situation préalable et secteur d'activité) sur le premier plan principal de la signalétique des créateurs et entrepreneurs d'entreprises en France.*

Results (4)

Supplementary variables

Situation before the creation – Branch of Industry

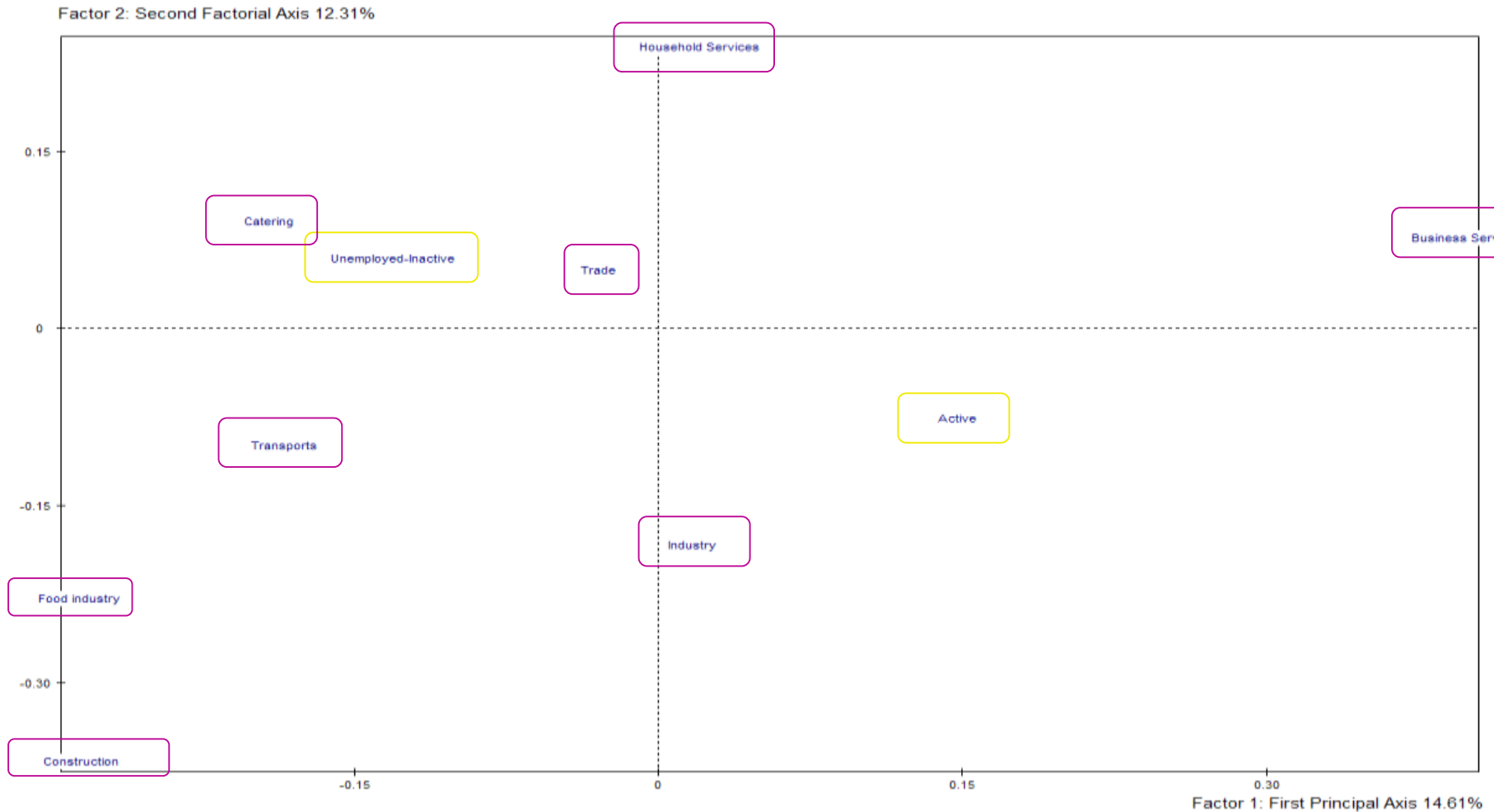
SUPPLEMENTARY MODALITIES			COORDINATES				CONTRIBUTIONS				SQUARED COSINES			
Labels	Relative Weight (%)	Squared Distance to the Origin	Axis 1	Axis 2	Axis 3	Axis 4	Axis 1	Axis 2	Axis 3	Axis 4	Axis 1	Axis 2	Axis 3	Axis 4
5- SITUATION BEFORE THE CREATION														
Active	10,51	0,05	0,14	-0,08	-0,01	0,11	0,00	0,00	0,00	0,00	0,39	0,12	0,00	0,21
Unemployed-Inactive	14,49	0,03	-0,11	0,06	0,01	-0,08	0,00	0,00	0,00	0,00	0,39	0,12	0,00	0,21
6- BRANCH OF INDUSTRY														
Food industry	0,53	0,15	-0,29	-0,23	-0,04	0,05	0,00	0,00	0,00	0,00	0,53	0,36	0,01	0,02
Industry	1,97	0,02	0,02	-0,18	-0,01	0,01	0,00	0,00	0,00	0,00	0,01	0,64	0,01	0,01
Construction	3,30	0,27	-0,29	-0,37	-0,15	-0,07	0,00	0,00	0,00	0,00	0,32	0,52	0,09	0,02
Trade	8,48	0,01	-0,03	0,05	-0,04	0,05	0,00	0,00	0,00	0,00	0,07	0,10	0,13	0,17
Transports	0,95	0,08	-0,17	-0,10	-0,03	0,07	0,00	0,00	0,00	0,00	0,38	0,14	0,01	0,07
Catering	2,34	0,08	-0,18	0,09	-0,08	0,13	0,00	0,00	0,00	0,00	0,40	0,11	0,08	0,22
Business Services	2,80	0,22	0,40	0,08	0,13	-0,14	0,00	0,00	0,00	0,00	0,71	0,03	0,08	0,08
Household Services	2,80	0,13	0,02	0,24	0,17	0,01	0,00	0,00	0,00	0,00	0,00	0,42	0,22	0,00

ACM – Analyse des Correspondances Multiples

Supplementary variables

Situation before the creation – Branch of industry

Representation on the first principal plane



Objectifs de l'analyse factorielle

AFD - Analyse Factorielle discriminante

- L'objectif de l'analyse discriminante dite « décisionnelle » est de mettre en évidence les liaisons entre une **variable qualitative** $Y(q)$ à expliquer (à discriminer) à q modalités (classes, groupes) et un ensemble de **p variables explicatives** $(X^1, X^2, \dots, X^p)$ **quantitatives** (discriminantes).
- L'analyse va alors chercher à mettre en évidence les variables quantitatives dites « **discriminantes** » qui **séparent et différencient aux mieux** ces q groupes d'individus.
- Elle permet également de définir **une règle de décision** permettant de classer un **individu anonyme** dans un des q groupes possibles à partir de ces données explicatives.
- L'AFD peut être considérée comme une **ACP particulière**, avec la distance dite de Mahalanobis.
- **Nombreux domaines d'application** : médical pour l'aide au diagnostic, en économie et finance, les banques et les assurances avec les techniques de scoring, en biomédical, le management pour décider de l'avenir d'une entreprise en fonction de ses diverses mesures comptables, de la vente pour sélectionner les clients potentiels les plus intéressants, etc.

AFD - Analyse Factorielle discriminante

Objectifs

On distingue donc deux objectifs pour l'analyse discriminante :

- **Objectif descriptif - “Factoriel”** : rechercher quels sont les *variables dites discriminantes* qui permettent de séparer le mieux possible les classes (groupes) d'individus et donner des représentations graphiques qui rendent compte au mieux de cette séparation.
- **Objectif décisionnel - “Modélisation”** : Un nouvel individu (anonyme) se présente, pour lequel on ne connaît que les valeurs des variables descriptives discriminantes, il s'agit alors de décider *dans quelle classe il faut l'affecter*.

AFD - Analyse Factorielle discriminante

Schematically

Explanatory variables

**p quantitative
variables**

$(x^1 \ x^2 \ \dots \ x^j \ \dots \ x^p)$

Explain variable

**qualitative variable
with q modalities**

y

q binary variables
 $(y^1 \ y^2 \ \dots \ y^k \ \dots \ y^q)$

$$x_i \rightarrow \begin{pmatrix} x_1^1 & \dots & x_1^j & \dots & x_1^p \\ \vdots & & \vdots & & \vdots \\ x_i^1 & \dots & x_i^j & \dots & x_i^p \\ \vdots & & \vdots & & \vdots \\ x_n^1 & \dots & x_n^j & \dots & x_n^p \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ \vdots \\ 3 \\ q \\ 2 \end{pmatrix} \Leftrightarrow y_i \rightarrow \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ \vdots & & \vdots & & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 1 \\ 0 & 1 & 0 & 0 & \dots & 0 \end{pmatrix}$$

AFD - Analyse Factorielle discriminante

Notations - Main Data Matrices

$$\mathbf{X}_{(n,p)} = \begin{pmatrix} x_1^1 & \dots & x_1^j & \dots & x_1^p \\ \vdots & & \vdots & & \vdots \\ x_i^1 & \dots & x_i^j & \dots & x_i^p \\ \vdots & & \vdots & & \vdots \\ x_n^1 & \dots & x_n^j & \dots & x_n^p \end{pmatrix} ; \quad \mathbf{Y}_{(n,q)} = \begin{pmatrix} y_1^1 & \dots & y_1^k & \dots & y_1^q \\ \vdots & & \vdots & & \vdots \\ y_i^1 & \dots & y_i^k = \begin{smallmatrix} 0 \\ 1 \end{smallmatrix} & \dots & y_i^q \\ \vdots & & \vdots & & \vdots \\ y_n^1 & \dots & y_n^j & \dots & y_n^q \end{pmatrix}$$

Explanatory p continuous variables

Explain qualitative variable (q dummy variables)

$$\mathbf{V}_{\mathbf{x}(p)} = \begin{pmatrix} \text{var}(x^1) & \dots & \text{cov}(x^1, x^j) & \dots & \text{cov}(x^1, x^p) \\ \vdots & & \dots & & \vdots \\ & & \text{var}(x^j) & & \text{cov}(x^j, x^p) \\ \vdots & & \dots & & \vdots \\ & & \dots & & \text{var}(x^p) \end{pmatrix} = {}^t\bar{\mathbf{X}} \mathbf{D}_w \bar{\mathbf{X}} \quad \mathbf{V}_{\mathbf{xy}(p,q)} = \begin{pmatrix} \text{cov}(x^1, y^1) & \dots & \text{cov}(x^1, y^k) & \dots & \text{cov}(x^1, y^q) \\ \vdots & & \dots & & \vdots \\ \text{cov}(x^j, y^1) & \dots & \text{cov}(x^j, y^k) & \dots & \text{cov}(x^j, y^q) \\ \vdots & & \dots & & \vdots \\ \text{cov}(x^p, y^1) & \dots & \text{cov}(x^p, y^k) & \dots & \text{cov}(x^p, y^q) \end{pmatrix} = {}^t\bar{\mathbf{X}} \mathbf{D}_w \bar{\mathbf{Y}}$$

Variance-Covariance matrix

$$\text{var}(x^j) = \frac{1}{n} \sum_{\{i=1\}}^n (x_i^j - \bar{x}^j)^2 \quad ; \quad \text{cov}(x^j, x^l) = \frac{1}{n} \sum_{\{i=1\}}^n (x_i^j - \bar{x}^j)(x_i^l - \bar{x}^l) \quad ; \quad \text{cov}(x^j, y^k) = \frac{1}{n} \sum_{\{i=1\}}^n (x_i^j - \bar{x}^j)(y_i^k - \bar{y}^k)$$

AFD - Analyse Factorielle discriminante

Main Data Matrices

$$\mathbf{G}_{(q,p)} = \begin{pmatrix} \bar{x}^1(1) & \dots & \bar{x}^p(1) \\ \vdots & \dots & \vdots \\ \bar{x}^1(k) & \dots & \bar{x}^p(k) \\ \vdots & \dots & \vdots \\ \bar{x}^1(q) & \dots & \bar{x}^p(q) \end{pmatrix} \begin{matrix} \leftarrow G(1) \\ \\ \leftarrow G(k) \\ \\ \leftarrow G(q) \end{matrix}$$

Table of coordinates of the centres of gravity

$$\mathbf{G} = \mathbf{D}_w^{-1}(\mathbf{G}) \quad {}^t\mathbf{Y} \mathbf{D}_w \mathbf{X} = \chi_y^2 \quad {}^t\mathbf{Y} \mathbf{D}_w \mathbf{X} \quad ; \quad \bar{x}^j(k) = \frac{1}{n} \sum_{\{i / y_i^k=1\}}^n x_i^j$$

$$\mathbf{D}_w(\mathbf{G})_{(q)} = \begin{pmatrix} w(1) & & 0 \\ & \ddots & \\ 0 & & w(q) \end{pmatrix} = \begin{pmatrix} \frac{n_1}{n} & & 0 \\ & \ddots & \\ 0 & & \frac{n_q}{n} \end{pmatrix} = {}^t\mathbf{Y} \mathbf{D}_w \mathbf{Y} \quad ; \quad \mathbf{D}_{w(n)} = \begin{pmatrix} w_1 & & 0 \\ & \ddots & \\ 0 & & w_n \end{pmatrix} = \frac{1}{n} \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} = \frac{1}{n} \mathbf{I}_n$$

Diagonal weight matrix of the q modalities

Diagonal weight matrix of the n individuals

Main Data Matrices

$$V_{x(p)} = \begin{pmatrix} \text{var}(x^1) & \dots & \text{cov}(x^1, x^j) \dots & \text{cov}(x^1, x^p) \\ \vdots & & \dots & \vdots \\ & \text{var}(x^j) & & \text{cov}(x^j, x^p) \\ \vdots & & \dots & \vdots \\ & & & \text{var}(x^p) \end{pmatrix} = {}^t\bar{X} D_w \bar{X}$$

Variance-Covariance matrix of explanatory variables

$$V_{xy(p,q)} = \begin{pmatrix} \text{cov}(x^1, y^1) & \dots & \text{cov}(x^1, y^k) \dots & \text{cov}(x^1, y^q) \\ \vdots & & \dots & \vdots \\ \text{cov}(x^j, y^1) & \dots & \text{cov}(x^j, y^k) \dots & \text{cov}(x^j, y^q) \\ \vdots & & \dots & \vdots \\ \text{cov}(x^p, y^1) & \dots & \text{cov}(x^p, y^k) \dots & \text{cov}(x^p, y^q) \end{pmatrix} = {}^t\bar{X} D_w \bar{Y}$$

Covariance matrix between variables

$$V_x^{-1}$$

Mahalanobis distance matrix

$$\chi_y^2 = \begin{pmatrix} \frac{n}{n_1} & 0 \\ & \frac{n}{n_k} \\ 0 & \frac{n}{n_q} \end{pmatrix} = D_w^{-1}(G)$$

Chi-square matrix

$$G_{(q,p)} = \chi_y^2 V_{xy} \quad ; \quad V_x^{-1}(p) \quad ; \quad D_w(G)_{(q)}$$

Data X:
Centre of gravity Matrix

Distance in the individual space M:
Mahalanobis distance

Distance in the variable space D:
Diagonal weights matrix

AFD - Analyse Factorielle discriminante

Définition : l'AFD peut être considérée comme une ACP particulière, elle consiste à effectuer l'ACP suivante :

$$AFD \longleftrightarrow ACP (\mathbf{G} ; \mathbf{V}_x^{-1} ; \mathbf{D}_w(\mathbf{G}))$$

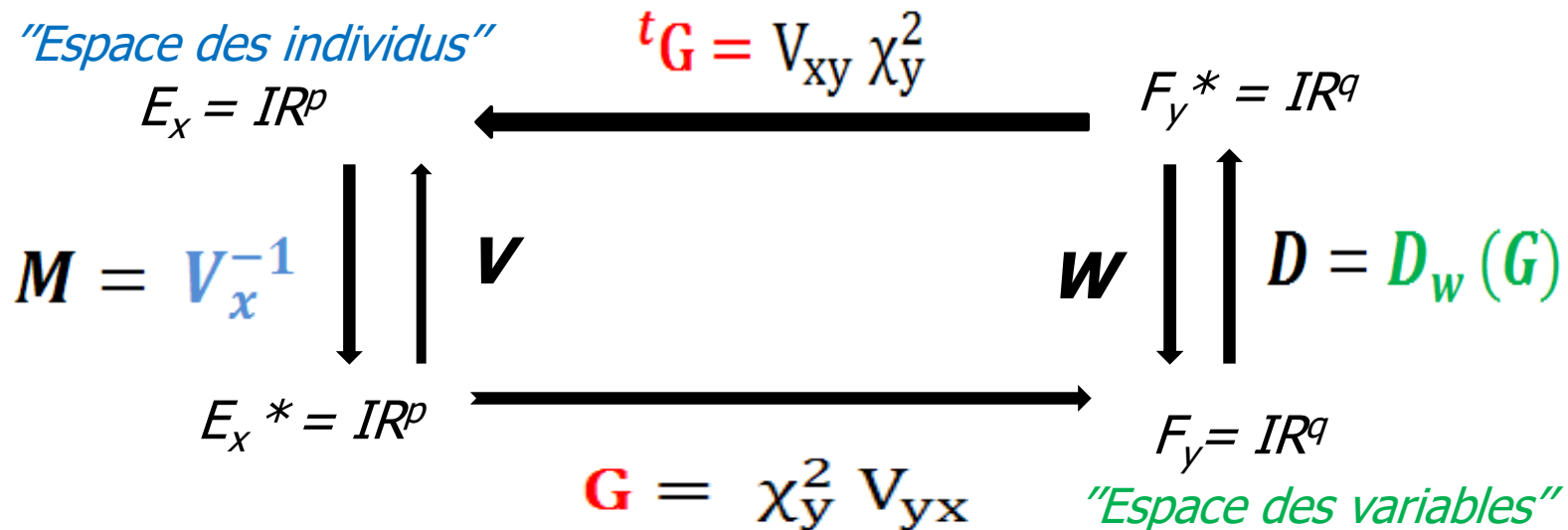
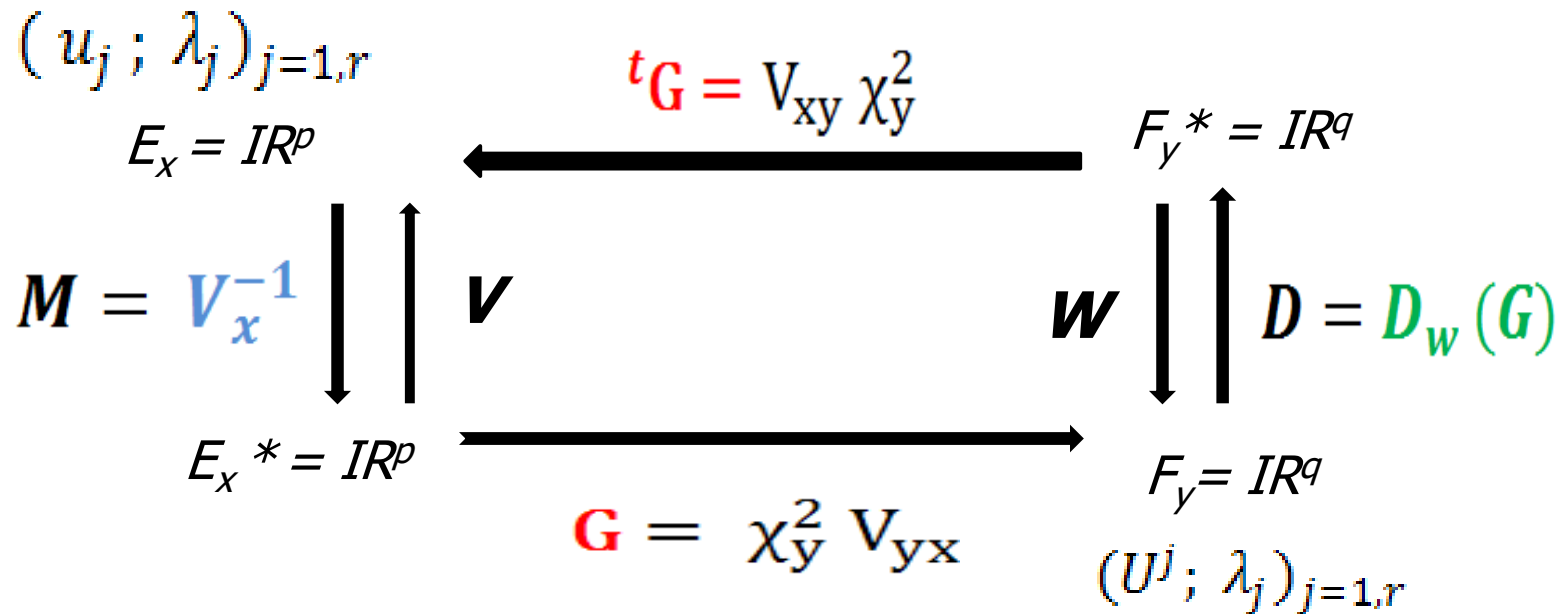


Schéma de dualité

AFD - Analyse Factorielle discriminante

Some notations



- $\forall j = 1, r = \min(q - 1 ; p) = q - 1 \quad (q \ll p)$

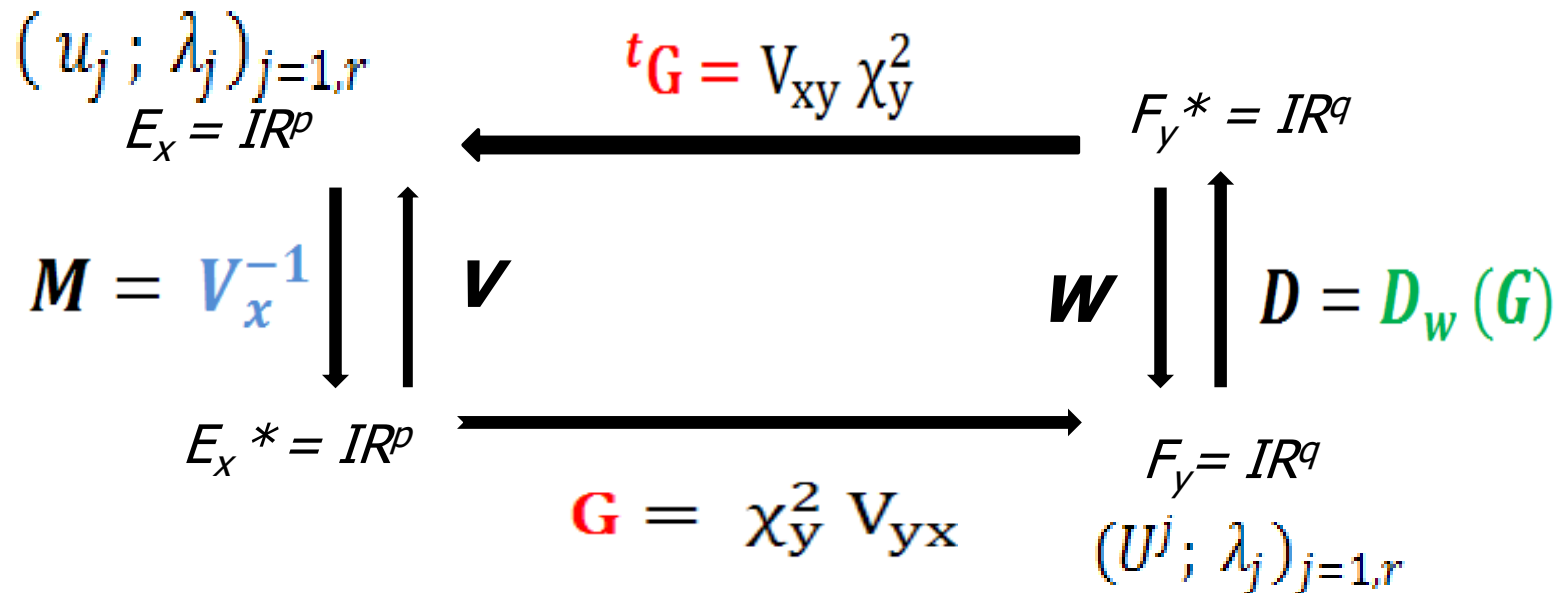
λ_j : j th principal moment (discriminating power)

u_j : j th principal axis (discriminant axis)

U^j : j th principal component (discriminant factor)
- eigenanalysis : $VMu_j = \lambda_j u_j \quad \leftrightarrow \quad WD U^j = \lambda_j U^j$

AFD - Analyse Factorielle discriminante

Some properties and relations



- $V = {}^tG D_w(G) G$ (Between classes variance matrix)
- Maximum $r = q - 1$ dimensions for graphical display in FDA.
- $\sum \lambda_j = \text{trace}(VM)$ corresponds to the Between classes inertia.
- $U^j = G M u_j = \chi_y^2 V_{yx} V_x^{-1} u_j$
- $u_j = \frac{1}{\lambda_j} V_{xy} U^j$

AFD - Analyse Factorielle discriminante

Décomposition de l'inertie totale à expliquer

$$\begin{aligned}
 I(N_y) &= \sum_{i=1}^n w_i d^2(x_i ; G) = \sum_{k=1}^q \sum_{x_i \in N(k)} w_i [d^2(x_i ; G(k)) + d^2(G(k) ; G)] \\
 &= \sum_{k=1}^q w(k) d^2(G(k) ; G) + \sum_{k=1}^q \frac{w_i}{w(k)} d^2(x_i ; G(k)) \\
 &= \sum_{k=1}^q w(k) d^2(G(k) ; G) + \sum_{k=1}^q w(k) I(N_k)
 \end{aligned}$$

$$\begin{array}{ccccc}
 I_T & = & I_B & + & I_W \\
 \text{Totale} & & \text{Inter « Between » classes} & & \text{Intra « Within » classes}
 \end{array}$$

↑
 Constante

↑
 Maximiser

↑
 Minimiser

Séparer au mieux les centres des classes afin de visualiser les différences entre les classes.

Les individus d'une même classe doivent être le plus concentrés possible autour du centre de la classe.

AFD - Analyse Factorielle discriminante

Décomposition de l'inertie totale à expliquer

- $I(N_y) = \text{tr}(V_y M_y) = \text{tr}(V_y \chi_y^2) = q - 1$
 $= I_T \quad (\text{Total Inertia})$
- $I(N_G) = \text{tr}(VM) = \text{tr}({}^t G D_w(G) G V_x^{-1}) = \text{tr}(V_{xy} \chi_y^2 V_{yx} V_x^{-1})$
 $= \sum_{j=1}^r \lambda_j \quad : \text{generalized correlation ratio}$
 $= I_B \quad (\text{Between classes Inertia})$
- $\frac{I_B}{I_T} = \frac{I(N_G)}{I(N_y)} = \frac{\sum_{j=1}^r \lambda_j}{q - 1} \quad : \text{explained inertia ratio}$

Remarques méthodologiques (1) Efficacité d'une analyse discriminante

- *Un bon classement est associé à un taux d'erreur faible. Il s'agit donc de comparer le risque d'erreur associé à une prédiction sans utiliser l'analyse discriminante avec un risque d'erreur associé à une prédiction en utilisant l'analyse discriminante. Sans analyse discriminante, nous avons 1 chance sur 2 d'effectuer un mauvais classement.*
- *Lorsque la taille de l'échantillon observé n est grande, il est conseillé de subdiviser (aléatoirement) cet échantillon en deux :*
 - *un **échantillon de base ou d'apprentissage de taille n_1** pour caractériser le modèle de discrimination,*
 - *un **échantillon test de taille n_2** pour juger de la qualité de la discrimination du modèle dans l'affectation de ces individus connus a priori et qui n'ont pas participé à l'élaboration du modèle.*
- *On obtient la **matrice de confusion – table des individus bien classés** par le modèle aussi bien sur l'échantillon d'apprentissage que sur l'échantillon test, avec $n = n_1 + n_2$ and $n_1 \gg n_2$ (par exemple, $n_1 \approx 80\%$ et $n_2 \approx 20\%$).*

Remarques méthodologiques (2)

- Fonctions de densité normales : on suppose l'égalité des matrices de variances-covariances intra-classes. La règle de classement qui en résulte est linéaire : *fonction discriminante linéaire de Fisher*.
- Etablir les matrices de confusion (apprentissage - test), puis comparer les pourcentages d'individus bien classés.
- A noter que le pourcentage de bien classés, calculé sur l'échantillon d'apprentissage, ayant servi à estimer les paramètres, est généralement trop optimiste. Il faudrait normalement le comparer à celui obtenu sur l'échantillon test.

Exemple illustratif - 1

- On s'intéresse ici à un échantillon de voitures de moins de 3,80 mètres. On dispose du **prix** en milliers de francs belges, de la **consommation** urbaine, de la **cylindrée**, de la **vitesse** maximum, du **volume** maximum du coffre, du **rapport poids/puissance** et de la **longueur** de 27 voitures.
- L'objectif est de mettre en évidence les caractéristiques qui différencient ces voitures concurrentes selon la marque du constructeur.
- On désire déterminer dans quelle mesure la **marque du constructeur** est liée aux caractéristiques mesurées par les six variables quantitatives explicatives. La variable qualitative "**cible**" à **discriminer** est la marque du constructeur (**Française, Etrangère**) à $q = 2$ modalités – **Analyse discriminante linéaire**.

ADL - Analyse Discriminante Linéaire

Example: 1 Belgium market of small cars (Source: Lambin J. 1990)

Name	Price	Consump tion	Cubic Capacity	Speed	Volume	Weight/ Power	Length		Mark (2)
AS2	239.900	6.200	998.000	140.000	955.000	23.200	3.400		Foreign
CI4	265.500	5.600	954.000	145.000	1170.000	19.400	3.500		French
PE6	292.500	6.700	993.000	145.000	1151.000	20.800	3.610		French
FI3	242.000	6.300	999.000	140.000	1088.000	21.800	3.640		Foreign
FI5	269.500	6.200	999.000	145.000	968.000	21.500	3.640		Foreign
FI8	500.100	8.900	1301.000	200.000	968.000	11.000	3.640		Foreign
FID	356.900	7.700	1302.000	165.000	968.000	16.000	3.640		Foreign
FO1	261.000	7.000	1117.000	137.000	900.000	22.700	3.640		Foreign
RE7	434.800	9.300	1597.000	180.000	973.000	12.000	3.640		French
NI1	248.000	6.400	988.000	140.000	375.000	17.000	3.640		Foreign
OP1	261.000	7.200	993.000	143.000	845.000	22.400	3.620		Foreign
PE1	265.200	6.800	954.000	134.000	1200.000	23.800	3.700		French
PE3	315.600	5.800	1124.000	142.000	1200.000	21.400	3.700		French
DA2	379.300	9.200	1360.000	170.000	1200.000	13.900	3.700		Foreign
PE9	503.500	8.700	1580.000	190.000	1200.000	11.200	3.700		French
RE1	259.600	6.300	956.000	115.000	950.000	33.100	3.670		French
RE3	276.100	6.300	1108.000	120.000	950.000	28.400	3.670		French
RE4	283.100	5.800	1108.000	143.000	915.000	20.600	3.590		French
FO9	345.000	7.900	1397.000	167.000	915.000	13.800	3.590		Foreign
RE8	506.300	8.700	1397.000	200.000	915.000	10.200	3.590		French
SE4	385.600	8.800	1461.000	175.000	1200.000	14.700	3.630		Foreign
SE9	219.300	7.300	903.000	131.000	1088.000	23.400	3.460		Foreign
SZ2	242.300	6.400	993.000	145.000	400.000	18.400	3.580		Foreign
SZ3	293.100	6.500	1324.000	163.000	400.000	14.000	3.580		Foreign
TO1	280.000	6.100	999.000	150.000	202.000	19.500	3.700		Foreign
TO3	337.000	6.800	1295.000	170.000	202.000	15.000	3.700		Foreign
VW3	360.900	7.800	1272.000	170.000	1040.000	14.300	3.650		Foreign



ADL - Analyse Discriminante Linéaire

Summary statistics of the Continuous variables & Categorical Variable

Label	Frequency	Mean	Std. Deviation	Minimum	Maximum
Price (MF- Belgium)	27	319.374	82.806	219.300	506.300
Urban Consumption (liter)	27	7.137	1.120	5.600	9.300
Cubic Capacity (cm3)	27	1165.630	204.170	903.000	1597.000
Maximum Speed	27	154.259	21.935	115.000	200.000
Maximum Volume of trunk	27	901.408	301.668	202.000	1200.000
Power Weight Ratio	27	18,648	5.424	10.200	33.100
Length of the car (meter)	27	3.623	0.072	3.400	3.700

Explanatory variables

Brand carmaker				
Modality label	Frequency	Percent	Cumulative Frequency	Cumulative Percent
French	10	37.04	10	37.04
Foreign	17	62.96	27	100.00

Explain variable

ADL - Analyse Discriminante Linéaire

Univariate statistics - Comparisons of means

Variable Label		G1: French Frequency: 10	G2: Foreign Frequency: 17	T Student	Probability
Price (MF- Belges)	Mean	340,220	307,112	0,984	0,335
	Standard deviation	95,476	71,570		
Urban Consumption (liter)	Mean	7,000	7,218	0,471	0,641
	Standard deviation	1,305	0,987		
Cubic Capacity (cm3)	Mean	1177,100	1158,880	0,216	0,831
	Standard deviation	241,184	178,511		
Maximum Speed	Mean	151,400	155,941	0,502	0,620
	Standard deviation	27,431	17,718		
Maximum Volume of trunk	Mean	1062,400	806,706	2,243	0,034 *
	Standard deviation	123,670	333,652		
Power Weight Ratio	Mean	20,090	17,800	1,041	0,308
	Standard deviation	7,052	3,941		
Length of the car (meter)	Mean	3,637	3,615	0,752	0,459
	Standard deviation	0,062	0,077		

**Significance less or equal than 1%

* Significance 1% - 5%

ADL - Analyse Discriminante Lin aire

Results (1)

CLASSIFICATION TABLE

ORIGINAL GROUPS	CLASSIFICATION COUNTS AND (PERCENTAGES)		
	Well classified	Misclassified	TOTAL
French	9.00 (90.00)	1.00 (10.00)	10.00 (100.00)
Foreign	17.00 (100.00)	0.00 (0.00)	17.00 (100.00)
TOTAL	26.00 (96.30)	1.00 (3.70)	27.00 (100.00)

LIST OF MISCLASSIFIED CASES

GROUP french : RE7

GROUP foreign : NO MISCLASSIFIED CASE

Percentage of cars well classified

ADL - Analyse Discriminante Linéaire

Results (2)

DISCRIMINANT LINEAR FUNCTION

VARIABLES NUM LABELS	CORRELATIONS VARIABLES AVEC D.L.F. (THRESHOLD= 0.39)	COEFFICIENTS DISCRIM. FUNCTION	REGRESSION	STANDARD DEVIATION (RES. TYPE REG.)	T STUDENT	PROBA
1 Price	0.195	+0.1988	0.0306	0.0067	+4.57	0.000**
2 Consumption	-0.095	-4.0235	-0.6201	0.2180	-2.84	0.010*
3 Cubic capacity	0.043	0.0004	0.0001	0.0014	0.05	0.963
4 Speed	-0.101	-0.7739	-0.1193	0.0377	-3.17	0.005**
5 Volume	0.413	0.0040	0.0006	0.0005	1.14	0.270
6 Weight/Power	0.206	-0.9570	-0.1475	0.0929	-1.59	0.129
7 Length	0.150	-15.1890	-2.3411	2.1097	-1.11	0.281
CONSTANT		152.184906	23.645252	10.7812	2.1932	0.0409
.....						
R2 = 0.70447		F = 6.47020	PROBA = 0.001			
D2 = 9.46492		T2 = 59.59396	PROBA = 0.001			

**Significance less or equal than 1%

* Significance 1% - 5%

ADL - Analyse Discriminante Linéaire

SPAD Software Results

LIST OF GROUP ASSIGNMENT AND RELATED PROBABILITIES

CASE	ORIG	ASSIGNMENT	PROB.	FICHER FUNCTION	
TO3	FOREIGN	==	1.000	-10.015	I *..... I
OP1	FOREIGN	==	1.000	-9.278	I ..*..... I
SZ2	FOREIGN	==	1.000	-8.656	I ...*..... I
SZ3	FOREIGN	==	1.000	-8.539	I*..... I
TO1	FOREIGN	==	0.999	-7.483	I*..... I
SE4	FOREIGN	==	0.998	-6.886	I*..... I
SE9	FOREIGN	==	0.997	-6.283	I*..... I
DA2	FOREIGN	==	0.997	-6.219	I*..... I
FI3	FOREIGN	==	0.995	-5.874	I*..... I
FO9	FOREIGN	==	0.987	-4.835	I*..... I
VW3	FOREIGN	==	0.982	-4.540	I*..... I
FO1	FOREIGN	==	0.974	-4.149	I*..... I
AS2	FOREIGN	==	0.973	-4.112	I*..... I
FI5	FOREIGN	==	0.972	-4.064	I*..... I
NI1	FOREIGN	==	0.942	-3.326	I*..... I
FID	FOREIGN	==	0.907	-2.811	I*..... I
FI8	FOREIGN	==	0.720	-1.475	I*..... I
RE7	FRENCH	FOREIGN	0.704	-1.399	I*..... I
PE1	FRENCH	==	0.622	1.030	I*..... I
RE8	FRENCH	==	0.921	1.917	I*..... I
CI4	FRENCH	==	0.953	2.474	I*..... I
RE4	FRENCH	==	0.978	3.253	I*..... I
RE1	FRENCH	==	0.997	5.137	I*..... I
PE9	FRENCH	==	1.000	7.681	I*..... I
RE3	FRENCH	==	1.000	9.110	I*..... I
PE3	FRENCH	==	1.000	9.190	I*..... I

ADL - Analyse Discriminante Linéaire

Affectation of the anonymous individuals-cars
Decisional objective of the model

Name	Price	Consumption	Cubic Capacity	Speed	Volume	Weight/Power	Length	➡	Mark (2)
AC1	262,50	6,5	954	135	870	22	3,60		?
AC2	315,50	6,3	1100	150	750	19,8	3,60		?
AC3	385,80	8,2	1295	175	900	14,7	3,70		?

LINEAR DISCRIMINANT ANALYSIS ON THE SAMPLE: ANONYMOUS
 BETWEEN THE 2 GROUPS: **FRENCH** and **FOREIGN**
 VARIABLE NUMBER OF GROUP : MARK CONSTRUCTOR

LIST OF THE AFFECTATIONS AND THE CORRESPONDING PROBABILITIES

IND.	ORIG	AFFEC	PROBA AFFEC	FONCT. FISHER	
AC1	?	FOREIGN	0.998	-6.758	I * I
AC2	?	FRENCH	0.791	0.797	I * I
AC3	?	FRENCH	0.940	2.220	I * I

Affectation of the anonymous individuals-cars

Predictive Models

Apprentissage Automatique – Machine Learning

Comparisons

Predictive model	Abreviation	% Well classified	AUC - ROC
Discriminant Analysis	DA	96.30	1.00
Logistic Regression	LR	100.00	1.00
Random Forest	RF	74.07	1.00
Support Vector Machine	SVM	96.30	1.00
Neuronal Networks	NN	100.00	
Decision Tree	DT	92.59	

Curve ROC – Receiver Operating Characteristic

AUC – Area Under the Curve

Illustrative Example-2

- *We use the same sample of 27 cars but the aim here is to highlight the characteristics that differentiate these competing cars in relation to the Fiscal Power.*
 - *We want to determine how the Fiscal Power is linked to the characteristics measured by the seven quantitative variables (price, consumption, capacity, speed, volume, weight/power ratio and length).*
 - *The qualitative discriminant variable is the Fiscal Power (4FP, 5FP, 6FP), here takes three modalities ($q = 3$, $q - 1 = 2$ discriminant factors).*
 - *Since we discriminate three groups, there will be no more than two factors. One discriminant plane from the point of view graphic representation.*

Factorial Discriminant Analysis

Example 2: Belgium market of small cars

(Source: Lambin J. 1990)

Name	Price	Consump tion	Cubic Capacity	Speed	Volume	Weight/ Power	Length		Fiscal Power (3)
AS2	239.900	6.200	998.000	140.000	955.000	23.200	3.400		4FP
CI4	265.500	5.600	954.000	145.000	1170.000	19.400	3.500		4FP
PE6	292.500	6.700	993.000	145.000	1151.000	20.800	3.610		4FP
FI3	242.000	6.300	999.000	140.000	1088.000	21.800	3.640		4FP
FI5	269.500	6.200	999.000	145.000	968.000	21.500	3.640		4FP
FI8	500.100	8.900	1301.000	200.000	968.000	11.000	3.640		6FP
FID	356.900	7.700	1302.000	165.000	968.000	16.000	3.640		6FP
FO1	261.000	7.000	1117.000	137.000	900.000	22.700	3.640		4FP
RE7	434.800	9.300	1597.000	180.000	973.000	12.000	3.640		6FP
NI1	248.000	6.400	988.000	140.000	375.000	17.000	3.640		4FP
OP1	261.000	7.200	993.000	143.000	845.000	22.400	3.620		4FP
PE1	265.200	6.800	954.000	134.000	1200.000	23.800	3.700		4FP
PE3	315.600	5.800	1124.000	142.000	1200.000	21.400	3.700		5FP
DA2	379.300	9.200	1360.000	170.000	1200.000	13.900	3.700		6FP
PE9	503.500	8.700	1580.000	190.000	1200.000	11.200	3.700		6FP
RE1	259.600	6.300	956.000	115.000	950.000	33.100	3.670		4FP
RE3	276.100	6.300	1108.000	120.000	950.000	28.400	3.670		5FP
RE4	283.100	5.800	1108.000	143.000	915.000	20.600	3.590		5FP
FO9	345.000	7.900	1397.000	167.000	915.000	13.800	3.590		6FP
RE8	506.300	8.700	1397.000	200.000	915.000	10.200	3.590		6FP
SE4	385.600	8.800	1461.000	175.000	1200.000	14.700	3.630		6FP
SE9	219.300	7.300	903.000	131.000	1088.000	23.400	3.460		4FP
SZ2	242.300	6.400	993.000	145.000	400.000	18.400	3.580		4FP
SZ3	293.100	6.500	1324.000	163.000	400.000	14.000	3.580		5FP
TO1	280.000	6.100	999.000	150.000	202.000	19.500	3.700		4FP
TO3	337.000	6.800	1295.000	170.000	202.000	15.000	3.700		5FP
VW3	360.900	7.800	1272.000	170.000	1040.000	14.300	3.650		6FP



Factorial Discriminant Analysis

Summary statistics

Label	Frequency	Mean	Std. Deviation	Minimum	Maximum
Price (MF- Belgium)	27	319.374	82.806	219.300	506.300
Urban Consumption (liter)	27	7.137	1.120	5.600	9.300
Cubic Capacity (cm3)	27	1165.630	204.170	903.000	1597.000
Maximum Speed	27	154.259	21.935	115.000	200.000
Maximum Volume of trunk	27	901.408	301.668	202.000	1200.000
Power Weight Ratio	27	18,648	5.424	10.200	33.100
Length of the car (meter)	27	3.623	0.072	3.400	3.700

Explanatory continuous variables

Fiscal Power				
Modality label	Frequency	Percent	Cumulative Frequency	Cumulative Percent
GFP4	13	48.15	13	48.15
GFP5	5	18.52	18	66.67
GFP6	9	33.33	27	100.00

Explain target categorical variable

Factorial Discriminant Analysis

Main Data Matrices

$G =$

Groups Of Fiscal Power	Price	Consumption	Cubic Capacity	Speed	Volume	Weight/Power	Length
GFP4	3,108	5,804	4,840	6,347	2,879	4,070	49,753
GFP5	3,635	5,572	5,837	6,729	2,431	3,665	50,416
GFP6	5,062	7,640	6,893	8,191	3,454	2,399	50,336

Data: Coordinates of the centres of gravity

$V_x^{-1} =$

	Price	Consumption	Cubic Capacity	Speed	Volume	Weight/Power	Length
Price	19.813						
Consumption	-0.061	3.831					
Cubic Capacity	-3.889	-1.517	5.138				
Speed	-24.889	-2.223	3.476	43.906			
Volume	-2.366	-0.838	0.304	2.575	1.699		
Weight/Power	-11.190	-0.620	3.191	23.260	0.143	16.331	.
Length	-2.453	-0.071	-0.182	2.849	0.555	0.688	1.498

Distance in variable space: Mahalanobis Distance

$D_w(G) =$

	GFP4	GFP5	GFP6
GFP4	13/27	0	0
GFP5	0	5/27	0
GFP6	0	0	9/27

Distance in individual space: Diagonal weights matrix of the centre of gravity


 $PCA (G ; V_x^{-1} ; D_w(G))$

Factorial Discriminant Analysis

Results (1)

Number	Eigenvalue	Percentage	Cumulative Percentage	Histogram
1	0.888	62.31	62.31	*****
2	0.537	37.69	100.00	*****

$$\Sigma \lambda_j = 1.425$$

Histogram of the eigenvalues

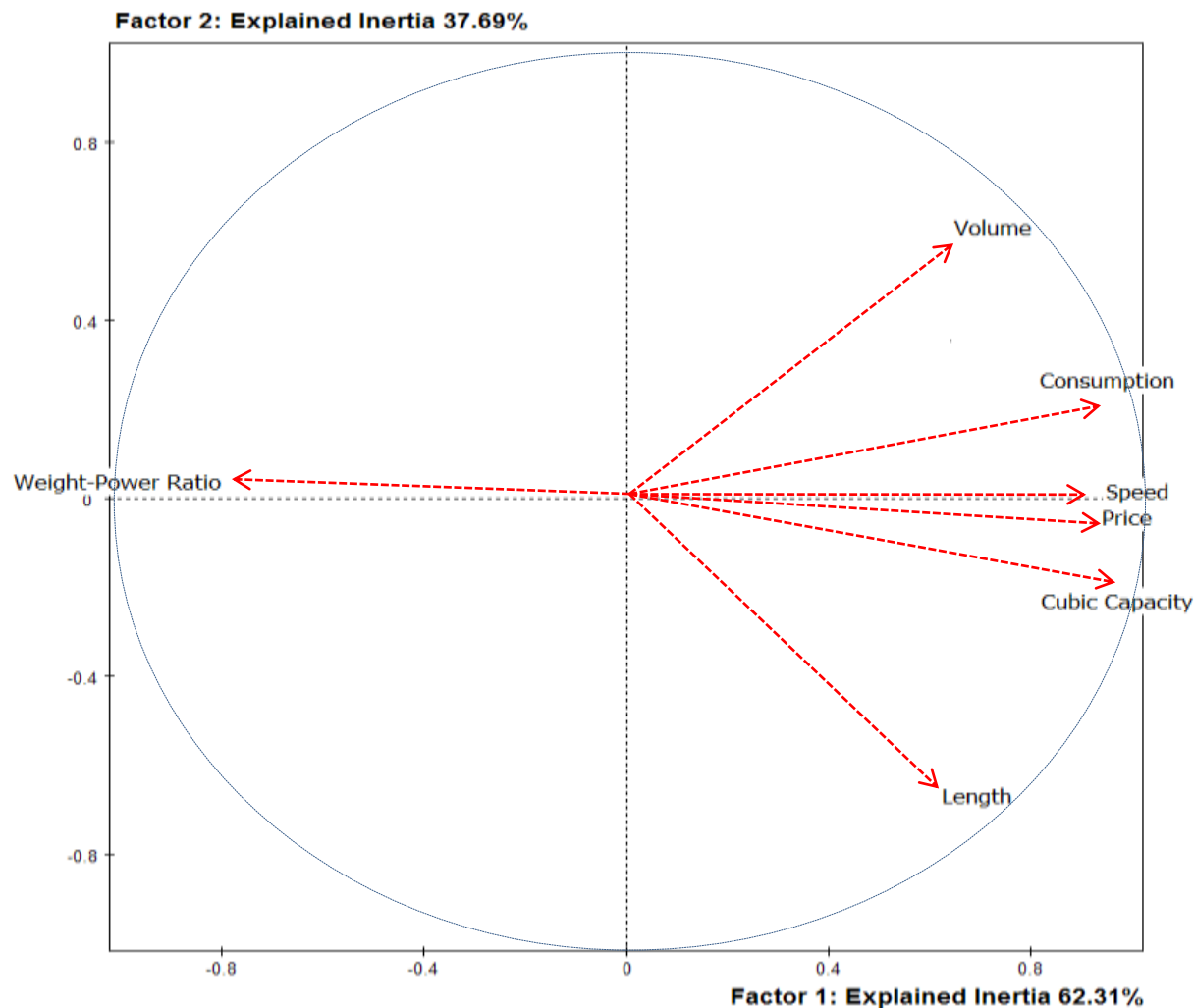
Total Discriminant axes:
 $\text{Min}(q - 1 = 2 ; p = 7)$
 Since $q = 3$, there are at most
 two discriminating factors.

Total Inertia: =1.425
 "Between-classes Inertia"

Criterion	Inertia
T-Total	2.000
B-Between classes	1.425
W-Within class	0.575
Explained Inertia	71.25%

Factorial Discriminant Analysis

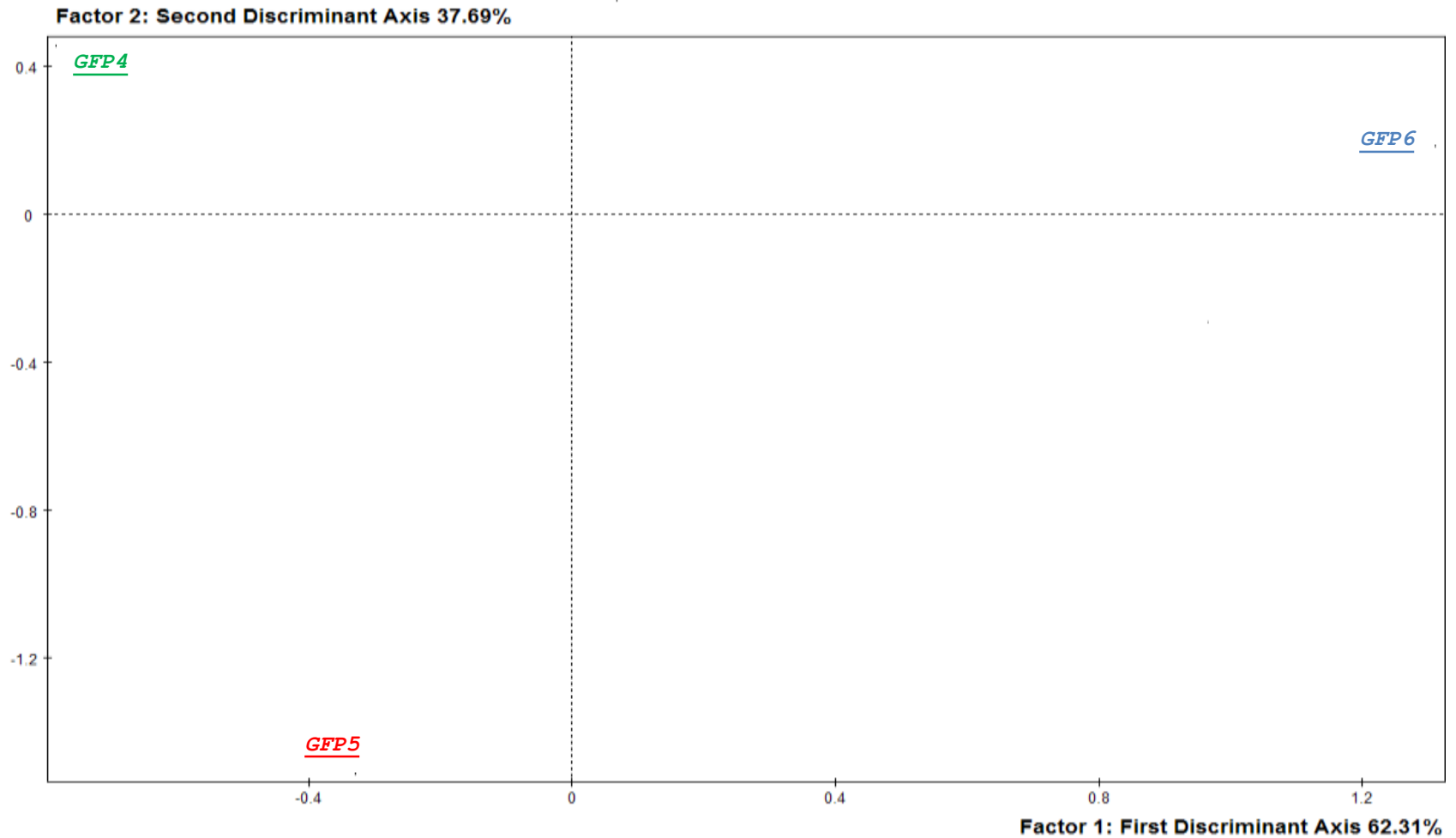
Representation of the variables on the first discriminant plane
 Circle of Correlations



Correlation	F1	F2
Price	0.999	-0.044
Consumption	0.963	0.269
Cubic Capacity	0.973	-0.231
Speed	1.000	0.009
Volume	0.803	0.597
Weight/Power	-1.000	0.021
Length	0.743	-0.669

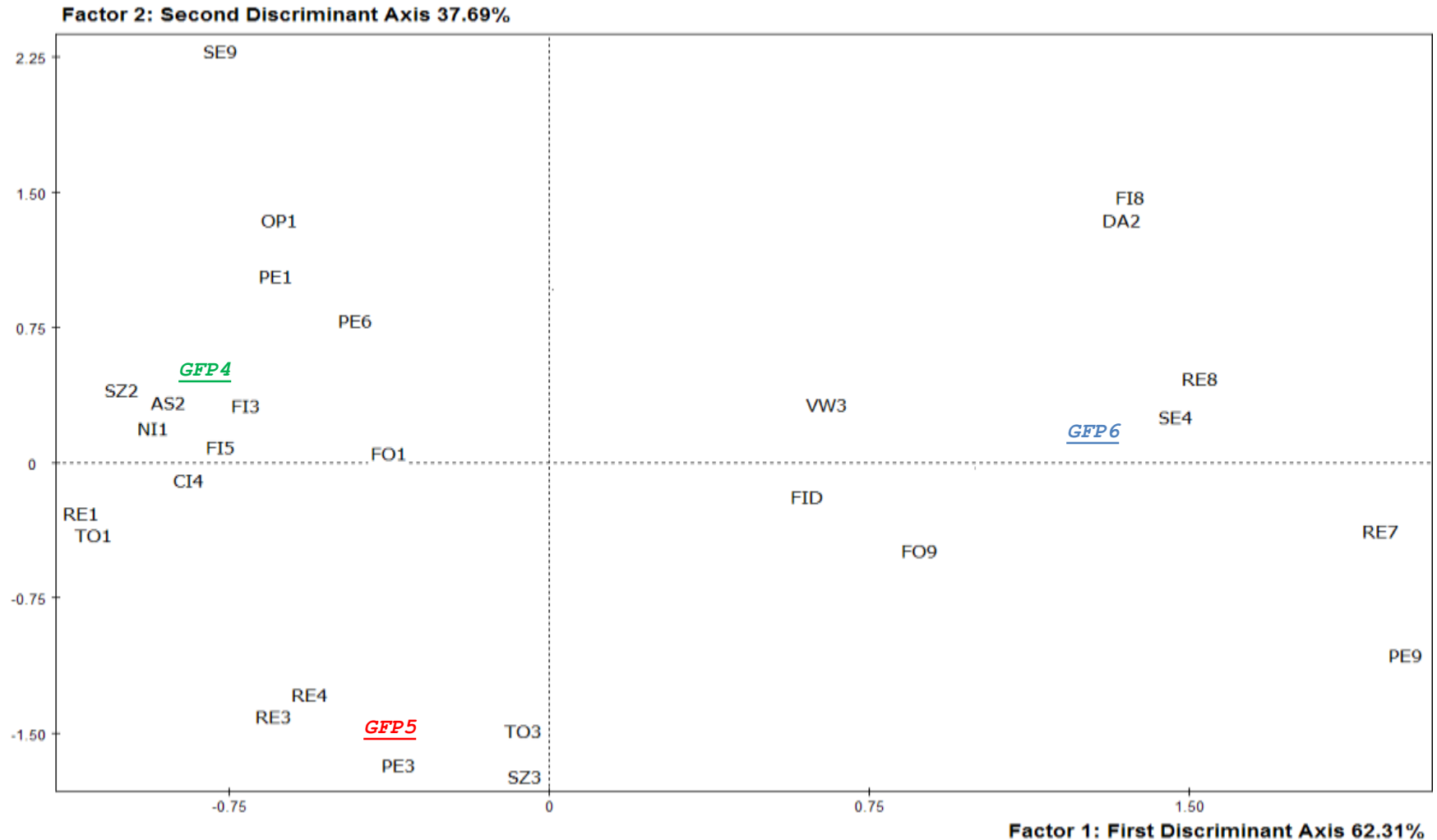
Factorial Discriminant Analysis

Representation of the centres of gravity on the first discriminant plane



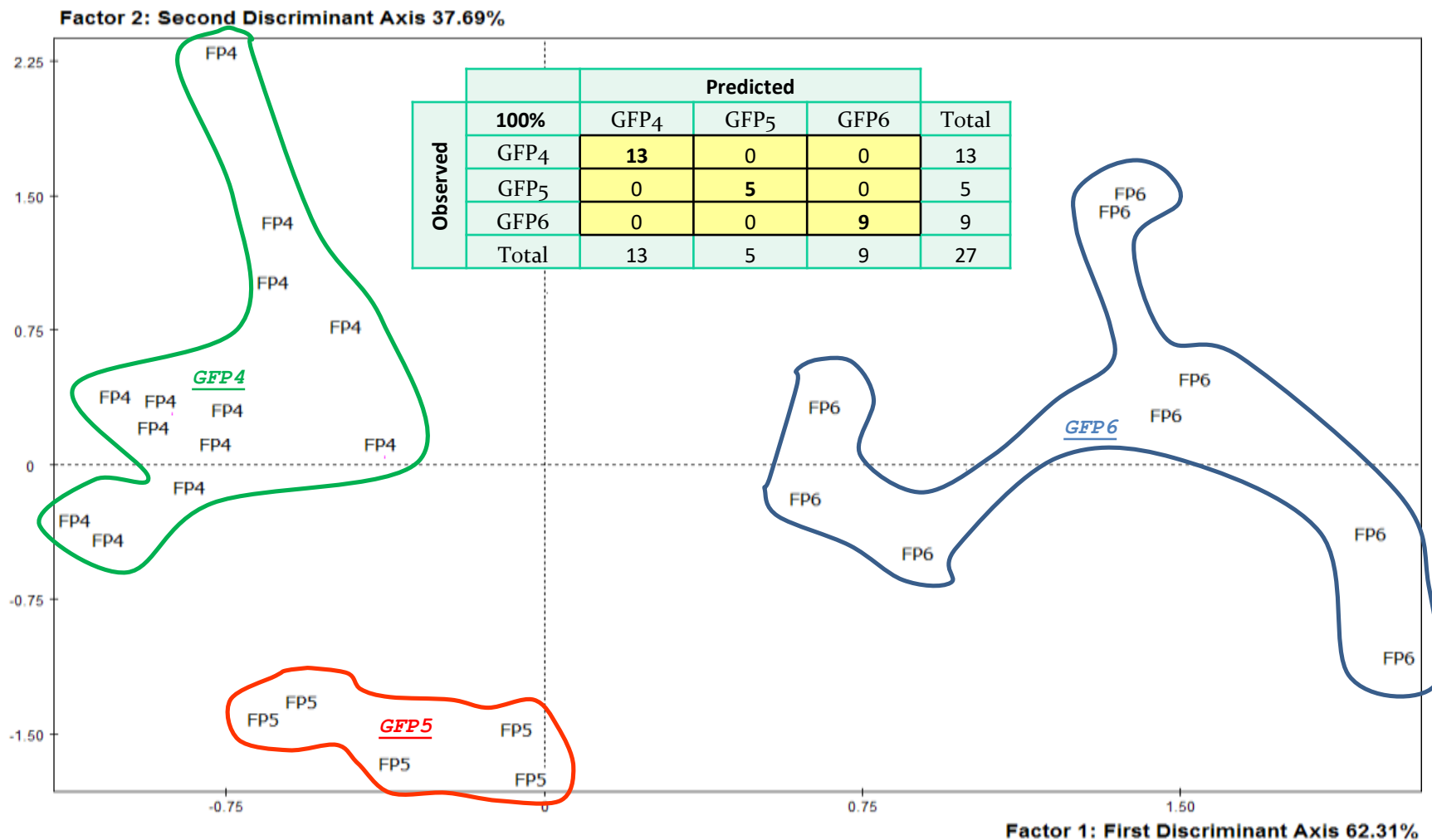
Factorial Discriminant Analysis

Representation on the first discriminant plane
 Centres of gravity and Individuals-cars



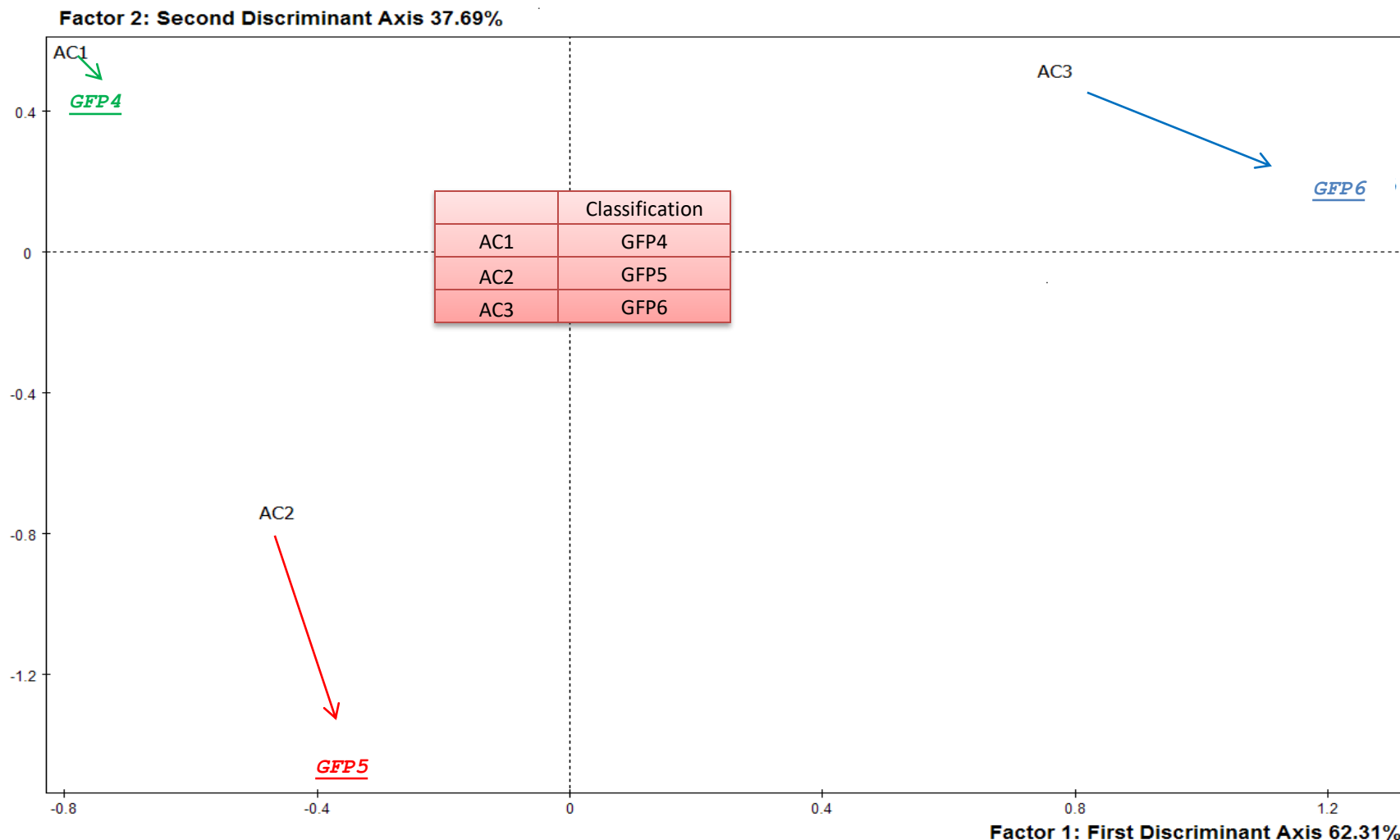
Factorial Discriminant Analysis

Representation on the first discriminant plane
 Individuals-cars and centres of gravity



Factorial Discriminant Analysis

Representation on the first discriminant plane
Anonymous individuals-cars and centres of gravity



Remarques et conclusion

- L'analyse discriminante dans sa version usuelle n'utilise que des **variables explicatives quantitatives** (Fisher 1938).
- Depuis, une méthodologie appelée "**Méthode DISQUAL**" (Saporta 1977) permet d'étendre le contexte de l'analyse discriminante à des variables explicatives qualitatives. On peut également effectuer une analyse discriminante sur variables explicatives mixtes (quantitatives et qualitatives) "**Méthode DISMIX**" (Abdesselam 2007)
 - Dans le premier cas, l'AFD peut être considérée comme une analyse discriminante sur les **composantes principales de l'ACM**.
 - Dans le second cas, l'AFD peut être considérée comme une analyse discriminante sur les **composantes principales de l'ACP Mixte**.
- AFDM Analyse Factorielle sur Données Mixtes, B. Escoffier & J. Pagès, 1979 et 2004, développée sous R.
- A noter que le logiciel SPAD ne propose que l'analyse discriminante à $q = 2$ modalités-classes. Lorsque q est supérieur à 2, on peut utiliser d'autres logiciels (SAS, SPSS, etc.)
- Enfin, il est intéressant de comparer les performance de l'AFD avec celles d'autres méthodes de classement (Régression logistique).