UNIVERSITE DE RENNES 1

LTSI

Inserm
Institut national
de la santé et de la recherche médicale

Méthodes d'Analyses Factorielles
ACP et AFCM

Emmanuel ROUX, Alfredo HERNANDEZ et Guy CARRAULT
LTSI - INSERM U642

Introduction

- Contexte
  - Nombre important de variables et d'individus statistiques
  - Pas ou peu de connaissances préalables sur les données

- Objectifs des méthodes
  - Réduction des données
    - Identifier les variables discriminantes les plus informatives
    - Identifier des relations entre variables
  - Juger de la capacité de caractérisation des variables
  - Identifier des groupes d'individus et/ou des types de comportement

Plan de l'Exposé

- Principes communs aux méthodes factorielles
- Analyse en Composantes Principales (ACP)

  → *Exemple*
- Analyse des Correspondances
  - Simple (AFC)
  - Multiple (AFCM)

    → *Exemple*
- Étude en cours

Principes Communs

- Tableau de données

  n individus statistiques $\times$ p variables

- Objectif

  Représenter les données de manière « optimale »

- Méthode
  - Transformation $D \rightarrow X$ adaptée à l'analyse souhaitée
  - Définition d'une distance
  - Critère d'optimisation

Espace des Individus / des Variables

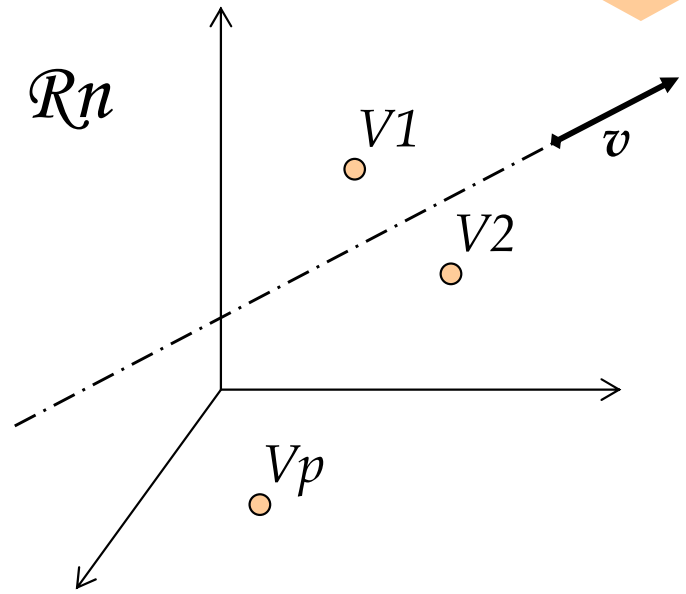
D

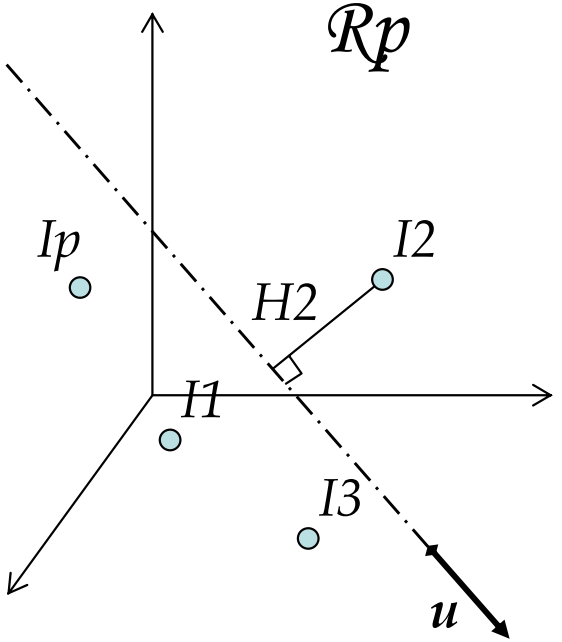
		Variables			
		V1	*V2*	...	*Vp*
Individus Statistiques	*I1*	d11	d12		
	I2				
	...				
	In				dnp

Mise en forme

X

		Variables			
		V1	*V2*	...	*Vp*
Individus Statistiques	*I1*	x11	x12		
	I2				
	...				
	In				xnp





$$\max_{\vec{u}} \left(\sum_{i=1}^{n} p_i \cdot d(OHi) \right)$$

Propriétés de Base

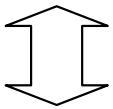
- Dans l'espace des variables $\mathcal{R}p$

  u_α = vecteurs propres de X^tX associés aux valeurs propres λ_α

- Dans l'espace des individus $\mathcal{R}n$

  v_α = *idem* avec XX^t

Les valeurs propres de X^tX est de XX^t sont égales !!



Rechercher la meilleure représentation des individus revient à chercher la meilleure représentation des variables

Analyse en Composantes Principales (ACP)

- Type de données
  - Variables continues
  - ACP normée : données centrées réduites (X)

$$(i, j) \in [1, n] \times [1, p], \quad x_{ij} = \frac{d_{ij} - \bar{d}_j}{\sigma_j}$$

- Distance euclidienne
- Critère

$$\max_{\vec{u}} \left(\sum_{i=1}^{n} d(GHi) \right)$$

- Recherche des valeurs propres de $X^t X$

Exemple ACP - Données

Activités et Tendances Culturelles : Musique Enregistrée (1998)

	Vente ($ US / Hab)	MusPopNat (%)	MusPopInt (%)	MusClas (%)	TxEnrPirates (%)	TxImp (%)	LectCD (Nb / 100 Hab)
Allemagne	36,6	43	47	10	3	16	75
Autriche	42,3	15	73	12	2	20	48
Belgique	36,1	20	71	9	4	21	63
Danemark	49,5	35	57	8	1	25	77
Espagne	17,1	42	51	7	2	16	47
Russie	0,6	68	26	6	70	20	2
Finlande	26,9	42	48	10	10	22	43
France	36,4	44	46	10	3	21	68
Grèce	10,9	59	37	4	25	18	22
Hongrie	5,6	32	59	9	25	25	22
Irlande	31,6	16	79	5	5	21	67
Israël	8,3	33	60	7	60	17	27
Italie	10,5	44	51	5	25	20	38
Lettonie	3,9	47	53	0	50	18	4
Norvège	62,8	19	77	4	4	23	44
Pays-Bas	35,7	27	64	9	6	18	99
Pologne	3,9	22	67	11	40	22	20
Portugal	18,7	31	65	4	3	18	30
RépTchèque	7,6	42	48	10	6	22	21
Roumanie	0,3	41	52	7	80	18	6
RoyaumeUni	49,0	48	45	7	1	18	87
Slovaquie	4,0	19	74	7	15	15	21
Suède	44,2	25	71	4	3	25	60
Suisse	45,0	8	82	10	4	8	75

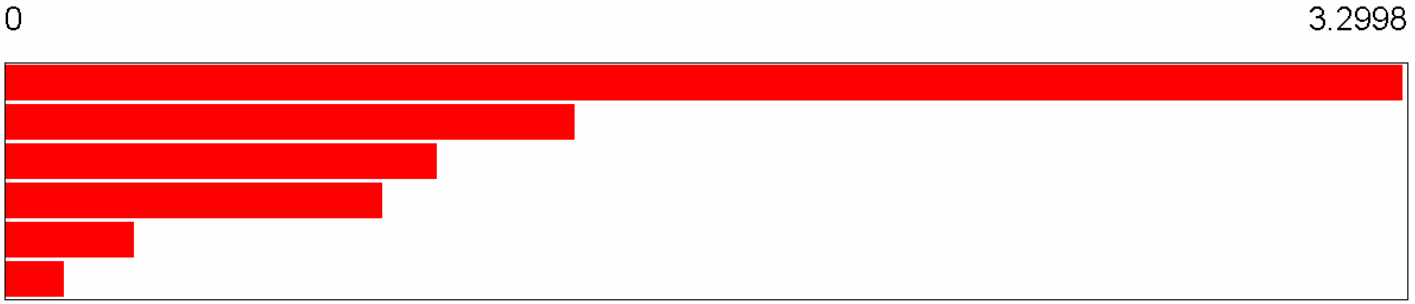
http://www.unesco.org/culture/worldreport/html_fr/stat2/table5f.pdf

Valeurs Propres

Résultats ACP	Interprétation
Valeurs propres λ_α	Variances suivant l'axe α
$\lambda_\alpha / \Sigma_\alpha \lambda_\alpha$	% d'inertie expliquée par l'axe α

n	Valeur	Pourcent	Cumul
1	3.2998	47.14	47.14
2	1.3459	19.23	66.37
3	1.0204	14.58	80.94
4	0.8897	12.71	93.66
5	0.3054	4.36	98.02
6	0.1387	1.98	100.00

Variance totale = 7.0



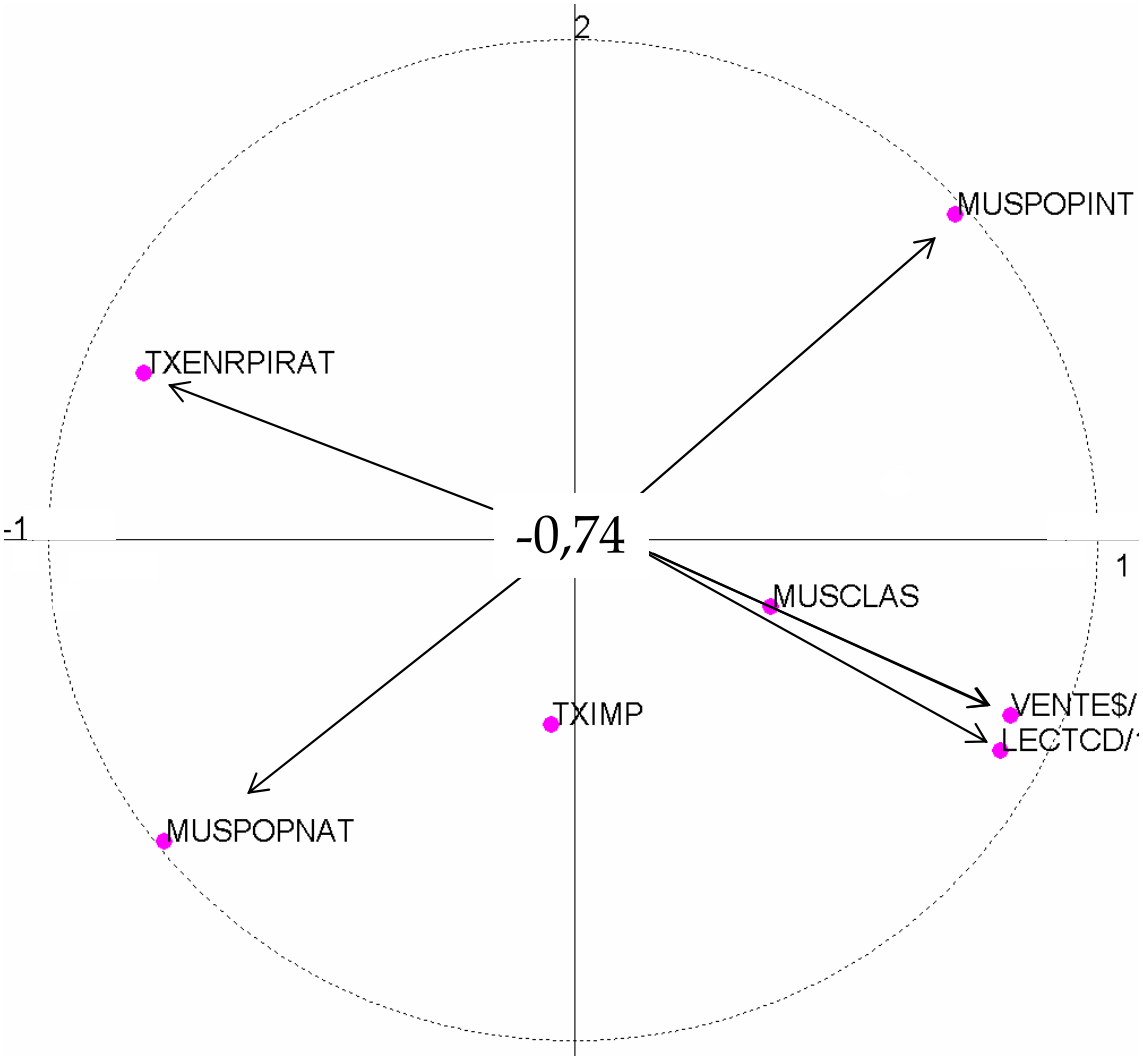
Critères de choix des axes principaux

- % inertie expliquée > 80%
- Valeurs propres > 1
- Différence significative entre 2 valeurs propres successives

Dépend du nombre de variables

Variables

Résultats ACP	Interprétation
Position absolue de la variable	Qualité de la représentation dans le plan
Cosinus angle entre variables	Corrélation



Coordonnées variables

Axe : 1

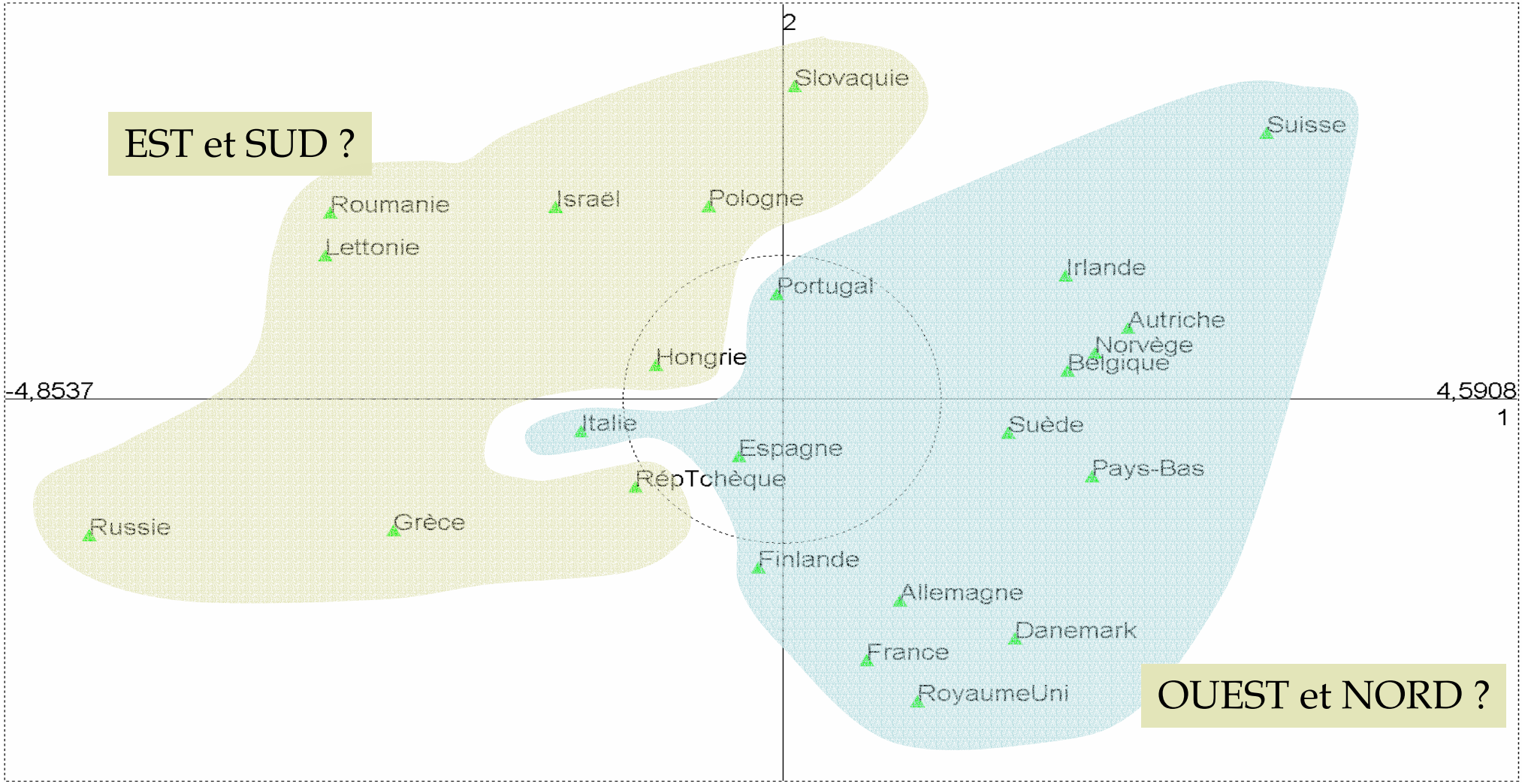
VENTE$/HAB	0,83
LECTCD/100	0,81
MUSPOPINT	0,73
MUSCLAS	0,38
TXIMP	-0,04
MUSPOPNAT	-0,78
TXENRPIRAT	-0,81

Axe : 2

MUSPOPINT	0,65
TXENRPIRAT	0,33
MUSCLAS	-0,13
VENTE$/HAB	-0,35
TXIMP	-0,37
LECTCD/100	-0,42
MUSPOPNAT	-0,60

Individus

Résultats ACP	Interprétation
Proximités entre individus	Similitudes
Contribution d'un individu à l'axe α	Contribution à l'inertie suivant l'axe α



Individus / Variables Supplémentaires (Illustratives)

- Individus : Projection dans $\mathcal{R}p$
- Variables continues : Projection dans $\mathcal{R}n$
- Variables nominales : Projection dans $\mathcal{R}p$!!

	Variables actives					**Variables supplémentaires**	
						continue	**nominale**
Individus actifs	x_{11}	…	x_{1j}	…	x_{1p}	xv^{+}_{11}	mod 1
	…	…	…	…	…	…	…
	x_{i1}	…	x_{ij}	…	x_{ip}	xv^{+}_{i1}	mod 2
	…	…	…	…	…	…	…
	x_{n1}	…	x_{nj}	…	x_{np}	xv^{+}_{n1}	mod 1
Individus supplémentaires	xi^{+}_{11}	…	xi^{+}_{1j}	…	xi^{+}_{1p}		
	xi^{+}_{21}	…	xi^{+}_{2j}	…	xi^{+}_{2p}		
	xi^{+}_{31}	…	xi^{+}_{3j}	…	xi^{+}_{3p}		
	xi^{+}_{41}	…	xi^{+}_{4j}	…	xi^{+}_{4p}		

moyenne

Individus Supplémentaires

ACP en Bref

- Variables continues
- Relations linéaires entre variables
- Pas de représentation simultanée individus –variables (en fait si)
- Possibilité de projeter individus et variables (continues et nominales) supplémentaires

Analyse Factorielle des Correspondances (AFC) - Données

- Tableau croisant deux variables nominales
- Transformations des données brutes
  - Calcul des fréquences relatives
  - Calcul des profils lignes
  - _____colonnes

Yeux \ Cheveux	Brun	Châtain	Roux	Blond	Total
Marron	68	119	26	7	220
Noisette	15	54	14	10	93
Vert	5	29	14	16	64
Bleu	20	84	17	94	215
Total	108	286	71	127	592

Profils lignes

	Brun	Châtain	Roux	Blond	Total
Marron	0,31	0,54	0,12	0,3	1
Noisette	0,16	0,58	0,15	0,11	1
Vert	0,8	0,45	0,22	0,25	1
Bleu	0,9	0,39	0,8	0,44	1
Profil moyen	0,18	0,48	0,12	0,22	1

Profils colonnes

	Brun	Châtain	Roux	Blond	Profil moyen
Marron	0,63	0,42	0,37	0,6	0,37
Noisette	0,14	0,19	0,2	0,8	0,16
Vert	0,5	0,1	0,2	0,13	0,11
Bleu	0,19	0,29	0,24	0,74	0,36
Total	1	1	1	1	1

AFC - Méthode

- Distance entre profils : distance du χ^2

$$d^2_{\chi^2}(i,i') = \sum_{j=1}^{p} \frac{1}{f_j}\left(\frac{f_{ij}}{f_i} - \frac{f_{i'j}}{f_{i'}}\right)^2$$

- Critère

$$\max_{\vec{u}}\left(\sum_{i=1}^{n} f_i \cdot d^2_{\chi^2}(O,i)\right)$$

- Recherche des valeurs propres de $F^tD^{-1}{}_nFD^{-1}{}_p$

Spécificités de l'AFC

- Lignes et Colonnes jouent le même rôle
- On s'intéresse aux distances entre profils
- L'inertie totale du nuage de points ($\Sigma_\alpha \lambda_\alpha$) reflète l'indépendance statistique entre les deux variables (test du χ^2)
- Représentation simultanée des deux nuages de points

Analyse Factorielle des Correspondances Multiples (AFCM) - Données

- Tableau croisant *n* individus statistiques et *s* « questions » ayant *ms* modalités

- Application privilégiée : enquêtes

	Sexe	**Satisfaction**	**Taille**
Individu 1	H	Très Satisfait	Grand
Individu 2	F	Moyennement Satisfait	Grand
Individu 3	H	Pas Satisfait	Petit
…	…	…	…
Individu n	F	Très Satisfait	Moyen

AFCM - Méthode

AFCM = AFC d'un tableau disjonctif complet

	Sexe	Satisfaction	Taille
Individu 1	H	Très Satisfait	Grand
Individu 2	F	Moyennement Satisfait	Grand
Individu 3	H	Pas Satisfait	Petit
…	…	…	…
Individu n	F	Très Satisfait	Moyen

Tableau Disjonctif Complet

	Sexe		Satisfaction			Taille		
	H	**F**	**Très Satisfait**	**Moyennement Satisfait**	**Pas Satisfait**	**Grand**	**Moyen**	**Petit**
Individu 1	1	0	1	0	0	1	0	0
Individu 2	0	1	0	1	0	1	0	1
Individu 3	1	0	0	0	1	0	0	0
…	…	…	…	…	…	…	…	…
Individu n	0	1	1	0	0	0	1	0

Exemple AFCM

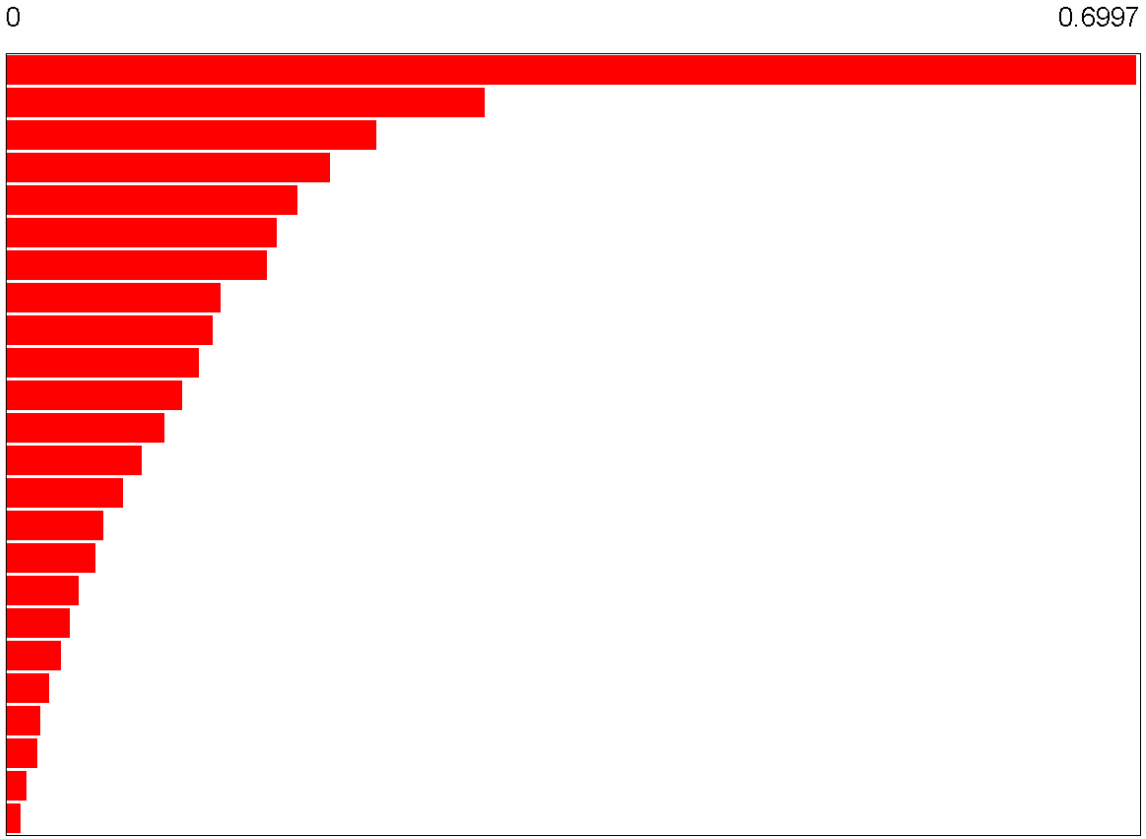
Réponses au questionnaire ASES
Évaluation fonctionnelle du membre supérieur

Individus	Mettre Un Manteau	Dormir Sur l'Epaule	Se Laver le Dos	Faire Sa Toilette	Se Peigner	Atteindre une Etagère Haute	Lever Une Charge	Lancer Une Balle	Prothèse
Pat1_1A	Facile	Facile	Facile	Facile	Assez Difficile	Facile	Facile	Facile	Totale Anatomique
Pat1_3M	Très Difficile	Très Difficile	Impossible	Facile	Assez Difficile	Impossible	Assez Difficile	Assez Difficile	Totale Anatomique
Pat1_6M	Assez Difficile	Facile	Assez Difficile	Facile	Facile	Assez Difficile	Facile	Assez Difficile	Totale Anatomique
Pat1_PO	Impossible	Impossible	Impossible	Très Difficile	Impossible	Impossible	Facile	Impossible	Totale Anatomique
Pat10_3M	Assez Difficile	Très Difficile	Impossible	Facile	Facile	Assez Difficile	Facile	Assez Difficile	Totale Anatomique
Pat10_PO	Impossible	Impossible	Impossible	Impossible	Impossible	Impossible	Impossible	Impossible	Totale Anatomique
Pat11_1A	Facile	Facile	Impossible	Facile	Facile	Facile	Facile	Assez Difficile	GERBER
Pat11_3M	Très Difficile	Facile	Impossible	Facile	Facile	Assez Difficile	Facile	Facile	GERBER
Pat11_6M	Facile	Facile	Impossible	Facile	Facile	Facile	Facile	Facile	GERBER

…
71 individus

AFCM – Valeurs Propres

n	Valeur	Pourcent	Cumul
1	0.6997	23.32	23.32
2	0.2966	9.89	33.21
3	0.2291	7.64	40.85
4	0.2005	6.68	47.53
5	0.1802	6.01	53.54
6	0.1676	5.59	59.12
7	0.1616	5.39	64.51
8	0.1328	4.43	68.94
9	0.1278	4.26	73.20
10	0.1193	3.98	77.18
11	0.1089	3.63	80.80
12	0.0979	3.26	84.07
13	0.0840	2.80	86.87
14	0.0722	2.41	89.27
15	0.0601	2.00	91.28
16	0.0554	1.85	93.12
17	0.0450	1.50	94.63
18	0.0394	1.31	95.94
19	0.0341	1.14	97.08
20	0.0264	0.88	97.96
21	0.0210	0.70	98.66
22	0.0191	0.64	99.29
23	0.0123	0.41	99.71
24	0.0088	0.29	100.00



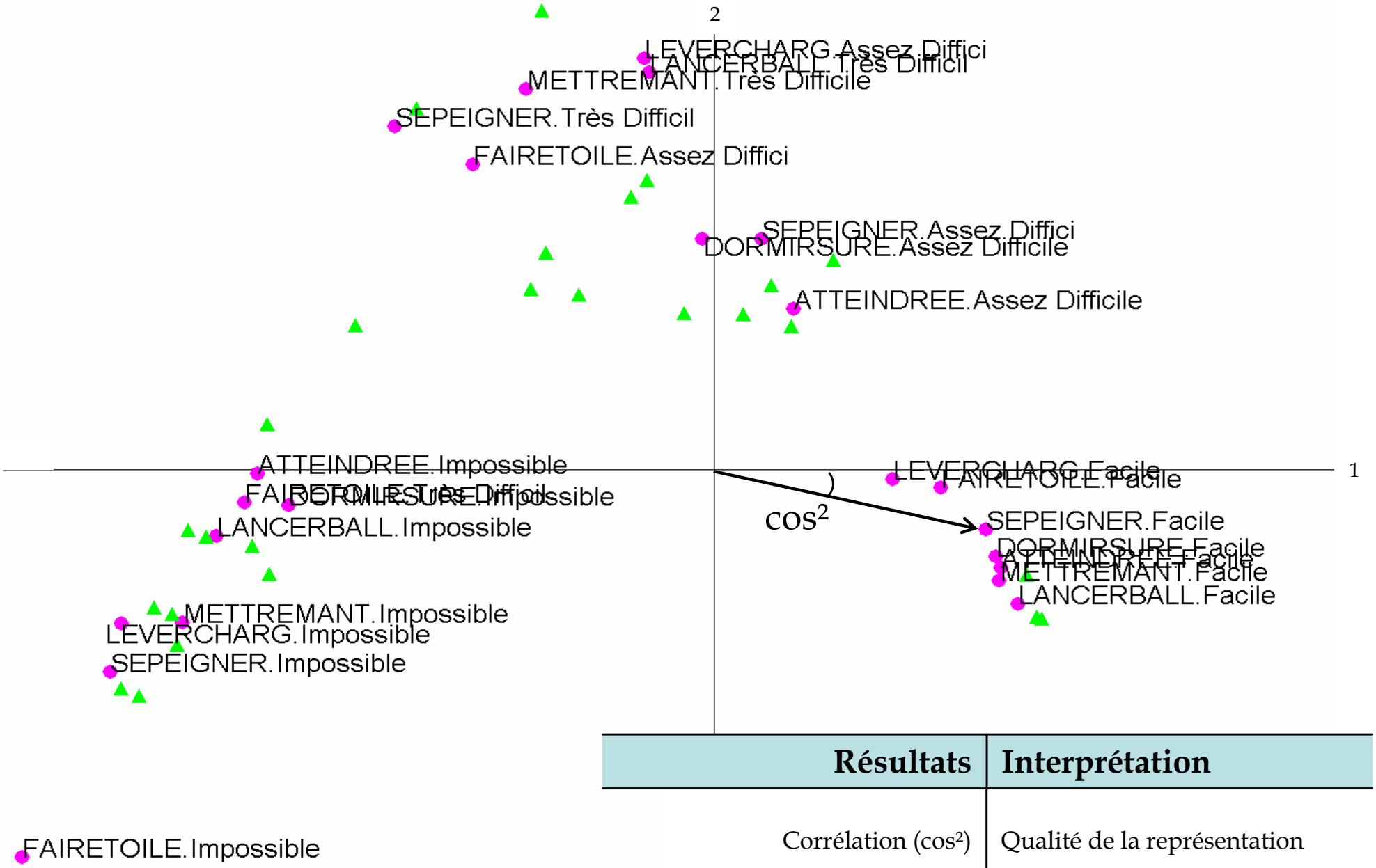
Variance totale = 3.0

% cumulé

0.57463697
0.67789221
0.73949778
0.7866822
0.82479575

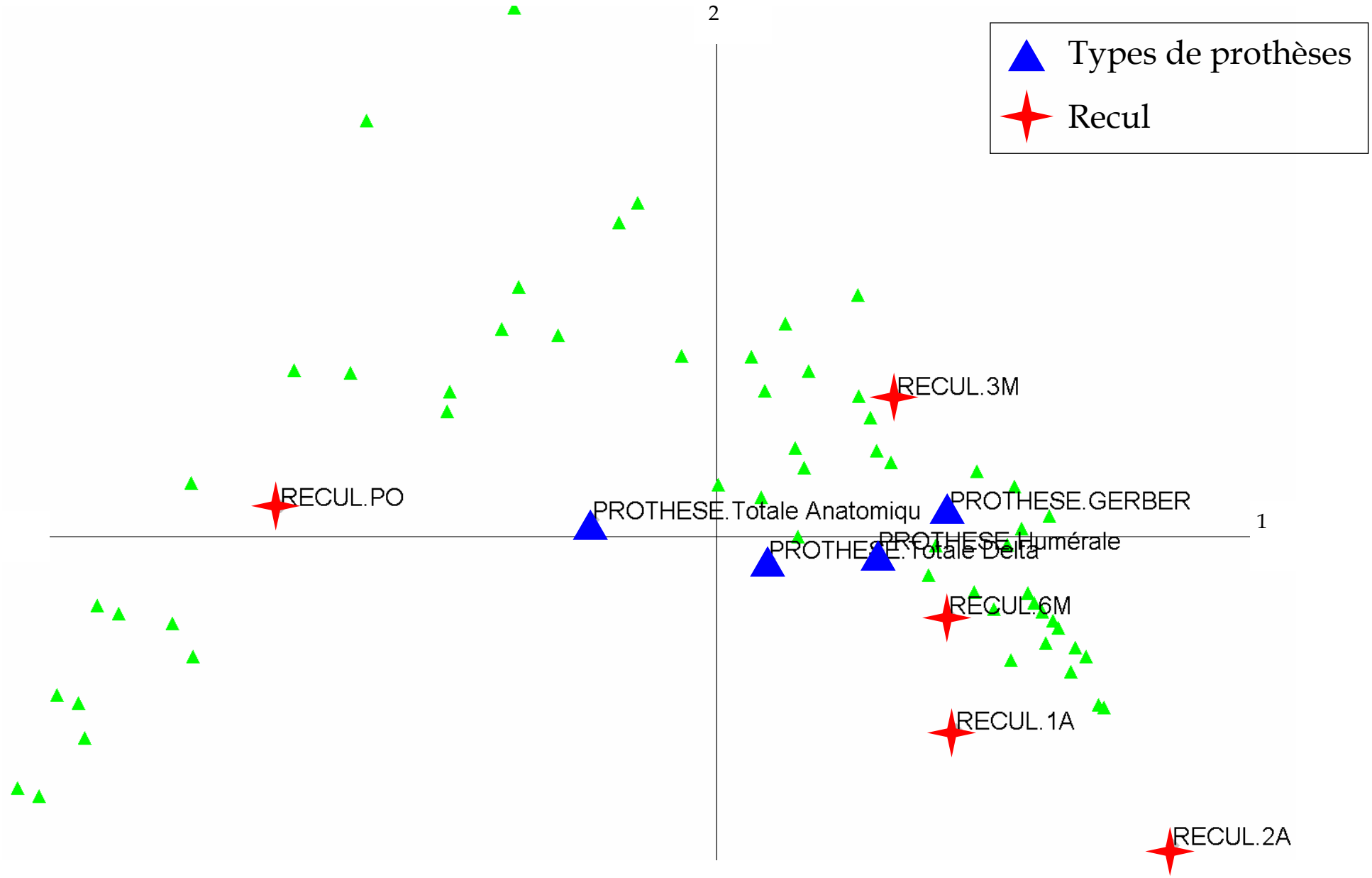
% : indice pessimiste de l'information extraite
Autre critère d'appréciation de l'inertie : (Valeurs Propres) ^{2}

AFCM – Individus et Variables



Résultats	Interprétation
Corrélation (cos^2)	Qualité de la représentation

AFCM –Variables Supplémentaires



AFCM En Bref

- Variables nominales et/ou continues avec codage (flou)
- Relations non linéaires entre variables
- Représentation simultanée individus –variables
- Possibilité de projeter individus, modalités et variables (nominales et continues) supplémentaires

Étude en Cours – Projet CEPICA

- Population d'étude

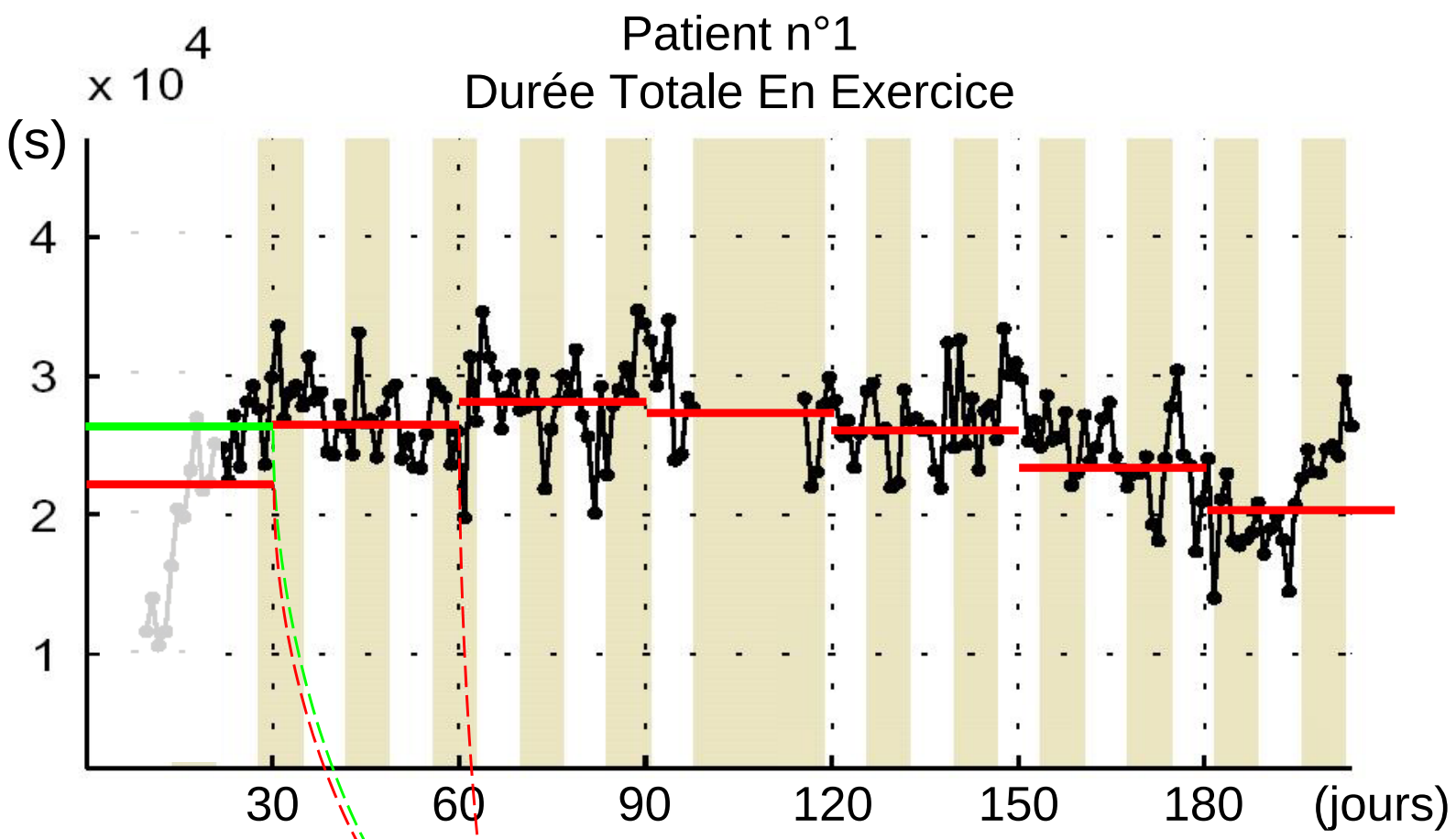
  Patients avec pacemaker bi-ventriculaire

- Objectif

  Les données « physiologiques » recueillies permettent-elles
  - Le suivi de l'état de santé des patients ?
  - De différentier les répondeurs des non-répondeurs à la stimulation ?

- Données disponibles
  - Données « physiologiques » journalières sur 1 ou 3 mois, tous les 3 mois
  - 37 variables
  - Trop peu d'individus (8)

ACP – Codage des Données

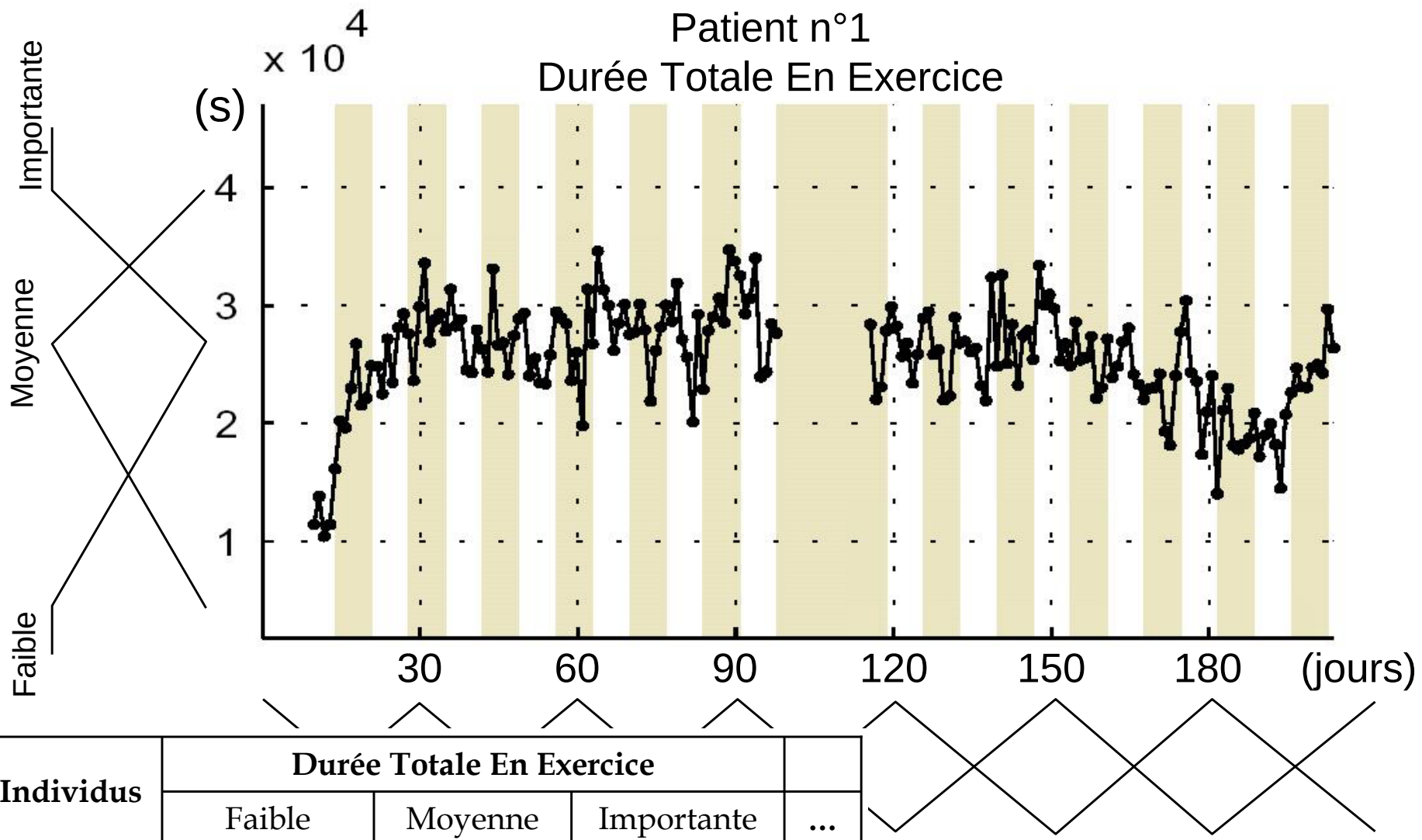


Individus	Durée Totale En Exercice	…
Pat1_30	2,6 . 10e4	…
Pat1_60	2,6 . 10e4	…
…	…	…

ACP

AFCM – Codage des Données

(Loslever et Bouilland, Fuzzy sets and systems, 1999)



Individus	Durée Totale En Exercice			
	Faible	Moyenne	Importante	...
Pat1_Début	0.4	0.5	0.1	…
Pat1_30j	0.3	0.5	0.2	…
...	…			…

AFCM

ACP et AFCM en Bref

– Peu adaptées à la prise de décision mais étape préalable pour
  - $\Rightarrow$ Tests d'hypothèses
  - $\Rightarrow$ Classification non-supervisée / supervisée …

\+ Méthodes descriptives, exploratoires

\+ Pas d'hypothèse préalable sur les données (non Paramétriques)

\+ Synthèse rapide de l'information