La distinction variables à expliquer / variables explicatives

dans les analyses factorielles

Nicolas Robette

Tuto@Mate

vendredi 5 avril 2024

Introduction

- On oppose souvent
  - régressions : explicatives, confirmatoires
  - analyses factorielles : descriptives, exploratoires
- Pourtant
  1. cela tient aux usages plutôt qu'aux propriétés statistiques des méthodes
  2. il existe diverses manières de distinguer variables à expliquer et variables explicatives avec des analyses factorielles

Rappels sur les analyses factorielles

Objectifs

- produire des résumés (graphiques et statistiques) de tableaux
- repérer les traits saillants, faire émerger les structures
- réduire la dimensionnalité des données
- éventuellement : construire des “scores”, des “facteurs”

Questions

- Quelles sont les observations qui se ressemblent (proximités entre individus) ?
- Sur quelles variables sont fondées les ressemblances ?
- Quelles sont les relations entre les variables (proximités entre variables) ?

Un changement de repère

1. Traduire le tableau étudié dans un espace, sous la forme d'un nuage de points (nuage des individus ou nuage des variables)

2. Chercher à visualiser les nuages obtenus sur les meilleures "photos" (projections) possibles

Principe d'étalement maximum : on considère que la meilleure photo est celle où le nuage s'étale au maximum, parce qu'elle donne à voir plus d'information.

Un changement de repère

- Autant d'axes qu'il y a de variables dans le tableau étudié
- Les axes sont perpendiculaires les uns par rapport aux autres (donc non corrélés) et passent tous par le (bary)centre du nuage
- Les axes sont hiérarchisés les uns par rapport aux autres: le premier est plus important que le second, qui est plus important que le troisième, etc.
- Le premier exprime le principe de structuration majeur au sein de la population étudiée (au regard des variables prises en compte dans l'analyse)

Variantes

- Tableau de contingence
  => analyse factorielle des correspondances (AFC)
- Tableau individus x variables numériques
  => analyse en composantes principales (ACP)
- Tableau individus x variables catégorisées
  => analyse des correspondances multiples (ACM)
- et beaucoup d'autres...

Variables supplémentaires

- Les analyses factorielles en sciences sociales distinguent fréquemment deux types de variables :
  1) on réalise une ACP/ACM à partir d'un premier ensemble de variables, dites "actives"
  2) on projette sur les résultats des variables "supplémentaires" (ou "illustratives"), i.e. qui n'ont pas participé à la construction des axes factoriels

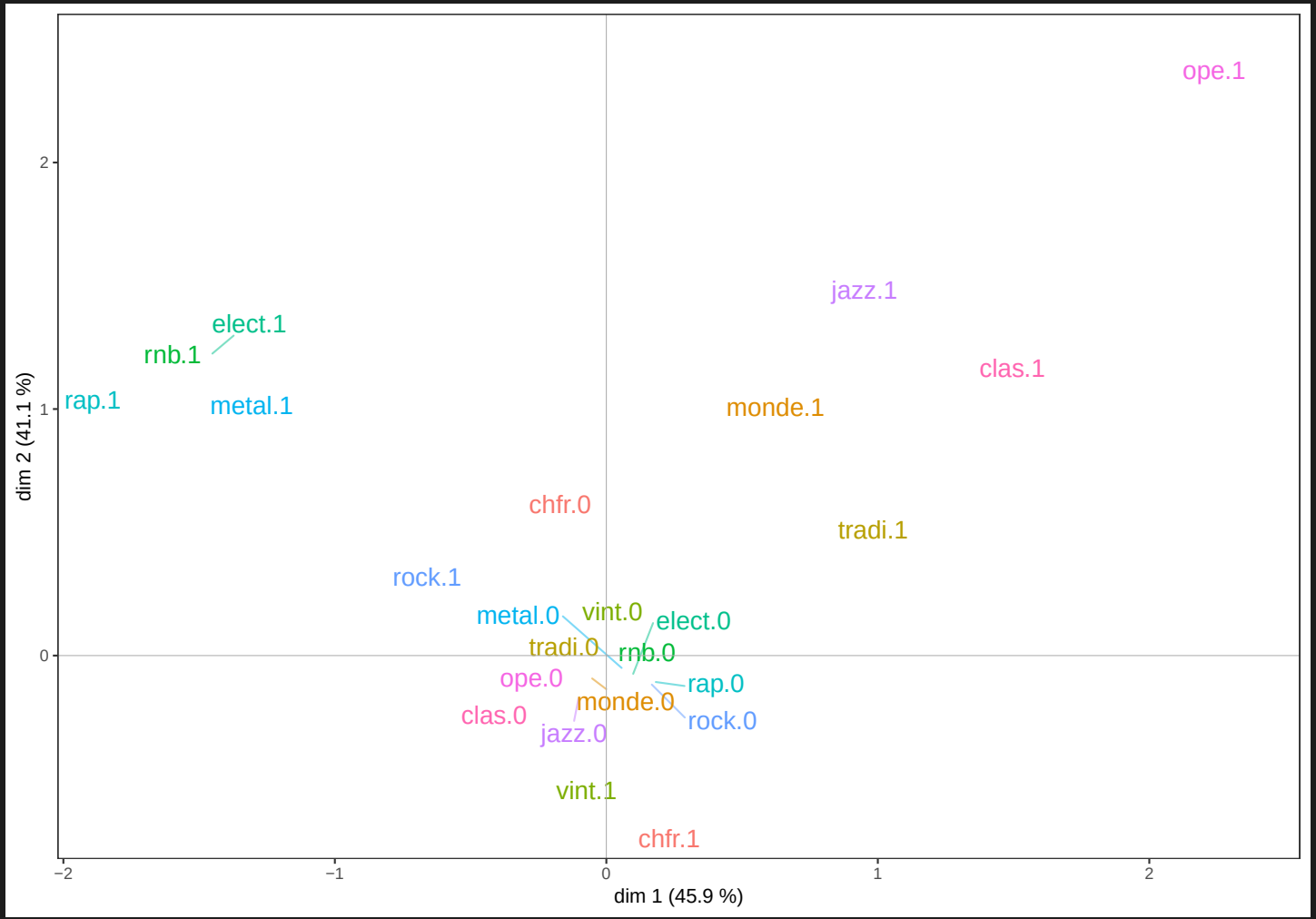
- Cette distinction recoupe généralement celle entre variables explicatives et variables à expliquer.

- Les variables explicatives peuvent être les variables actives ou supplémentaires.
  - le plus courant : propriétés sociales (diplôme, sexe, âge, PCS, etc.) projetées sur l'espace des pratiques culturelles, opinions politiques, etc.
  - mais aussi : indicateurs de pratiques ou d'opinions projetés sur l'espace social

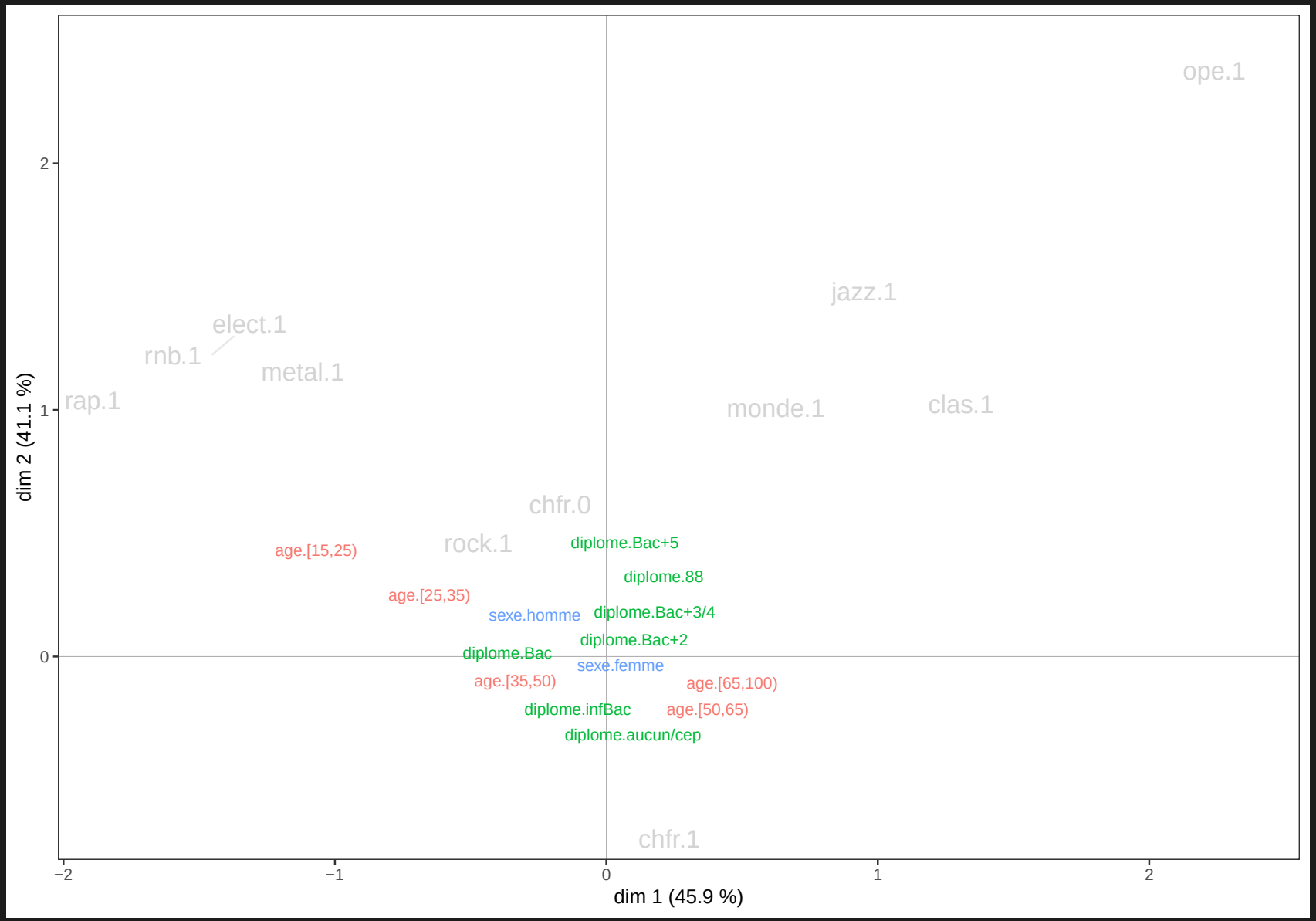
Exemple des goûts musicaux

- Enquête *Pratiques culturelles des Français* (DEPS, 2018)
- N = 9234, 15 ans et plus
- “y a-t-il des genres de musique que vous aimez particulièrement ?” (liste de 12 genres)
- ACM avec les genres musicaux en variables actives, propriétés sociales (sexe, âge, diplôme) en supplémentaires

espace des modalités actives



variables supplémentaires



- Or les coordonnées d'une variable supplémentaire dans l'espace factoriel correspondent aux coefficients de corrélation de cette variable avec les axes factoriels.

- Et, comme les axes factoriels sont orthogonaux (donc indépendants), les coefficients de la régression d'une variable supplémentaire sur les axes factoriels sont les coordonnées de cette variable sur les axes.

Régression sur composantes principales

- une variable y à expliquer et des variables explicatives continues X
- étapes :
  1. ACP des X
  2. régression sur les composantes principales (coordonnées des observations sur les axes factoriels)
  3. facultatif : les composantes principales sont des combinaisons linéaires des X, on peut donc exprimer y en fonction des X (après passage par l'ACP)
- si les X ne sont pas continues mais catégorielles, ACP remplacée par ACM
- on peut ne conserver à l'étape 2 qu'une partie des composantes principales (selon la part de variance de X expliquée, les plus corrélées à y, validation croisée...)

- équivalent à l'utilisation des variables supplémentaires (y projeté sur l'espace des X)

- mais usages différents :

  - variables supplémentaires => on étudie avant tout l'analyse factorielle
  - régression sur composantes principales => on étudie avant tout les résultats de la régression

- avantages lorsqu'il y a beaucoup de variables
  - ok même si multicolinéarité
  - et même si p»n
  - réduction de dimensionnalité (avant la régression)

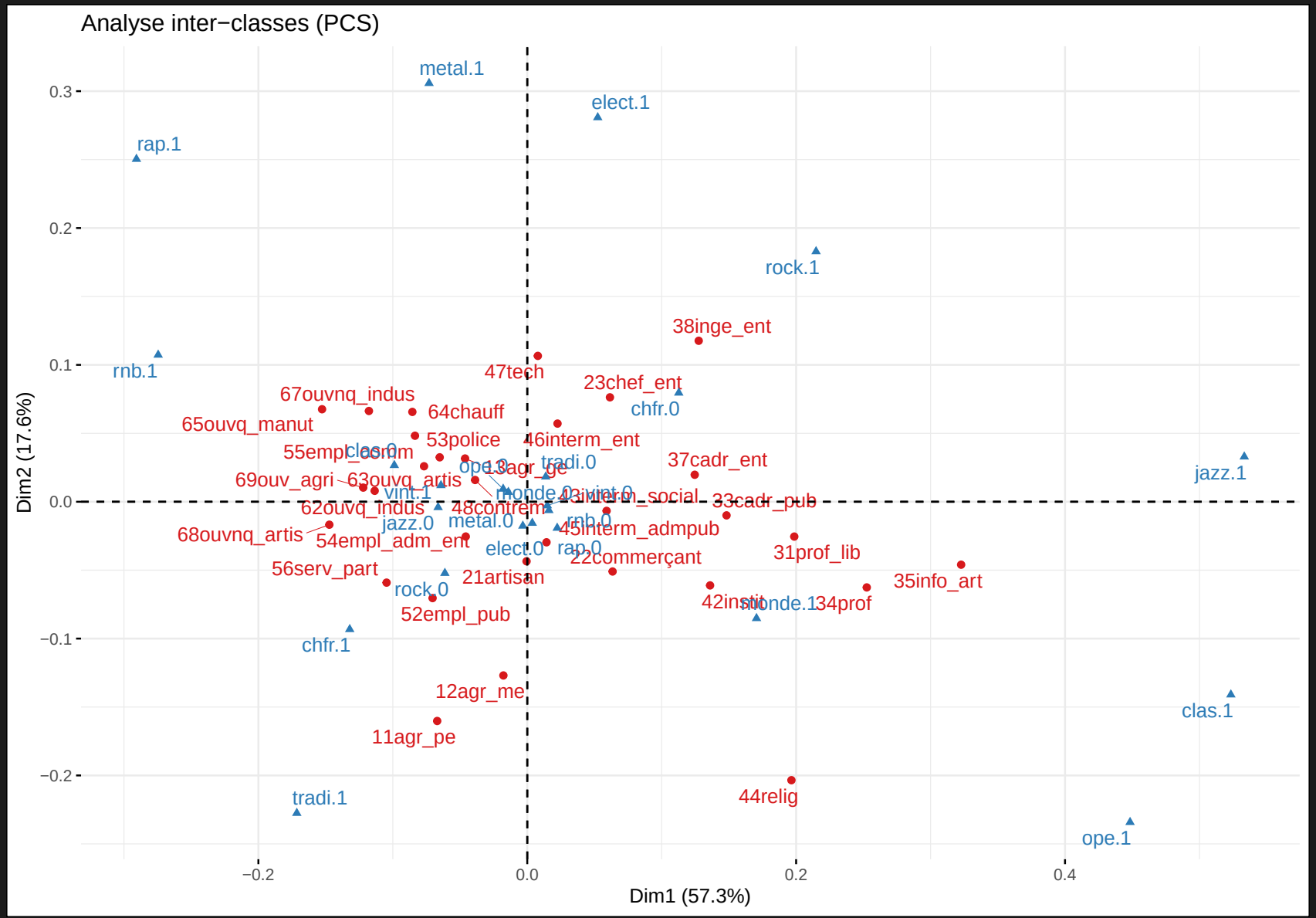
- limite
  - hypothèse implicite : les directions dans lesquelles les X varient le plus sont aussi celles qui permettent d'expliquer y. Or parfois, ce sont les dernières composantes principales qui sont associées à y.
  - autrement dit, l'espace factoriel des X est construit indépendamment des relations statistiques avec y

Analyses inter-classes

- Situation :
  - une variable à expliquer y et des variables explicatives X
  - y est catégorielle : elle partitionne les individus en groupes (ou “classes”)
- Principe de l’analyse factorielle inter-classes
  - construire un espace factoriel qui rend compte des structures qui différencient les classes, qui “explique” le mieux l’appartenance aux classes

Exemple des goûts musicaux

- à partir des 12 genres musicaux particulièrement aimés
- on construit l'espace factoriel qui "discrimine" (sépare) le mieux les PCS (en 31 modalités)


Analyse inter−classes (PCS)
metal.1
elect.1
rap.1
rock.1
38inge_ent
rnb.1
47tech
23chef_ent
67ouvnq_indus
chfr.0
64chauff
65ouvq_manut
53police
46interm_ent
55empl_comm
clas.0
tradi.0
37cadr_ent
jazz.1
69ouv_agri
63ouvq_artis
ope.0
13agr_ge
vint.1
monde.0
43interm_social
33cadr_pub
62ouvq_indus
48contrem
jazz.0
metal.0
rnb.0
45interm_admpub
68ouvnq_artis
54empl_adm_ent
elect.0
rap.0
31prof_lib
22commerçant
56serv_part
21artisan
35info_art
rock.0
42instit
monde.1
34prof
52empl_pub
chfr.1
12agr_me
clas.1
11agr_pe
44relig
tradi.1
ope.1
Dim2 (17.6%)
Dim1 (57.3%)
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.2
0.0
0.2
0.4

- Pratiquement, y peut être une variable explicative ou une variable à expliquer
  - l'espace des consommations qui discrimine le mieux la PCS (Chauvel, 1999)
  - l'espace social qui discrimine le mieux une pratique ou une opinion

- Si les variables X sont continues, ACP inter-classes

  1. Calcul du barycentre des observations pour les variables X pour chacune des classes (modalités) de y : profils-moyens
  2. ACP de l'ensemble des barycentres

- Si les variables X sont catégorielles, ACM inter-classes

  - dite aussi "analyse discriminante barycentrique" ou "analyse des correspondances discriminante"

  1. Transformation des données de X en tableau disjonctif complet
  2. Calcul du barycentre des données transformées pour chacune des classes (modalités) de y : profils-moyens
  3. AFC de l'ensemble des barycentres

- Variante de l'ACP inter-classes : Analyse Factorielle Discriminante (AFD)
  - également appelée "analyse discriminante linéaire" chez les anglo-saxons
  - seule la métrique diffère par rapport à l'ACP inter-classes

- Variante de l'ACM inter-classes : AFD sur variables qualitatives (ou "Disqual", voir Saporta 1977)
  1. ACM des données de X
  2. AFD sur les composantes (axes factoriels) de l'ACM

Variables instrumentales

- Situation
  - au lieu d'une variable catégorielle y, on a maintenant un ensemble de variables Y
- Principe de l'analyse factorielle sur variables instrumentales
  - expliquer les variables Y par l'ensemble des variables X, dites variables "instrumentales"
  - plus précisément, construire un espace factoriel des Y de manière à ce qu'il soit le mieux "expliqué" par les variables instrumentales X

- Si les variables Y sont continues : ACP sur variables instrumentales
  - aussi appelée “analyse des redondances”

  1. Calcul d’une régression linéaire par variable de Y, avec cette variable comme variable à expliquer et les variables instrumentales X comme variables explicatives
  2. ACP des valeurs de Y prédites par les régressions ($\hat{Y}$)

- Si les variables Y sont catégorielles : ACM sur variables instrumentales
  - ou “Analyse Canonique des Correspondances”

  1. ACM des variables Y, en conservant l’ensemble des dimensions de l’espace
  2. Calcul d’une régression linéaire par dimension de l’ACM, avec les coordonnées des observations sur l’axe factoriel comme variable à expliquer et les variables instrumentales X comme variables explicatives
  3. ACP des valeurs prédites par les régressions

- Analyse à compléter avec une ACP/ACM des Y : la comparaison permet de voir ce qui, dans la structure des Y, n'a pas pu être "expliqué" par les X

- NB : Les analyses inter-classes peuvent être considérées comme des cas particuliers des analyses sur variables instrumentales, avec une variable Y unique et catégorielle

- Comparaison avec régression sur composantes principales (RCP)

  - RCP => espace factoriel des variables explicatives, construit indépendamment de la relation avec la variable à expliquer
  - VI => espace factoriel des variables à expliquer *en lien* avec les variables explicatives ; plusieurs variables à expliquer y

PLS1

- Situation : une variable à expliquer continue y et un ensemble de variables explicatives X (continues ou catégorielles dichotomisées)
- PLS1 : construction d'axes factoriels à partir des X de manière à maximiser la covariance entre y et les X
- Compromis entre la régression linéaire de y sur les X et l'ACP des X

PLS2

- Situation : un ensemble de variables à expliquer Y (ou une variable à expliquer catégorielle dichotomisée) et un ensemble de variables explicatives X
- PLS2 (généralisation de PLS1) : construction d'axes factoriels à partir des X et des Y de manière à maximiser la covariance entre les X et les Y
- Compromis entre les régressions des Y sur les X, l'ACP de X et l'ACP de Y

Comparaisons avec les autres approches

- avec la projection de variables supplémentaires, celles-ci interviennent indépendamment les unes des autres (comme en régression simple)
- avec les variables instrumentales, celles-ci interviennent sous forme de combinaison linéaire (comme en régression multiple)
- avec la régression sur composantes principales, l'espace des X est construit indépendamment de y
- avec la régression PLS, c'est l'espace des X qui "explique" l'espace des Y

Usages de la régression PLS

- Les Y étant des combinaisons linéaires des axes factoriels, qui eux-mêmes sont des combinaisons linéaires des X, on peut exprimer les Y en fonction des X (après passage par la régression PLS) => "approche régression"
  - particulièrement utile si p»n
  - ou si multicolinéarité (surtout quand beaucoup de variables explicatives)

- Mais on peut aussi se concentrer sur l'interprétation des espaces factoriels des X et des Y (comme on le fait pour les ACM et ACP) => "approche analyse factorielle"

Remarques

- L'algorithme de régression PLS ne nécessite ni diagonalisation ni inversion de matrice (il n'utilise pratiquement que des régressions linéaires simples) : simple, rapide et pas de problème de convergence.

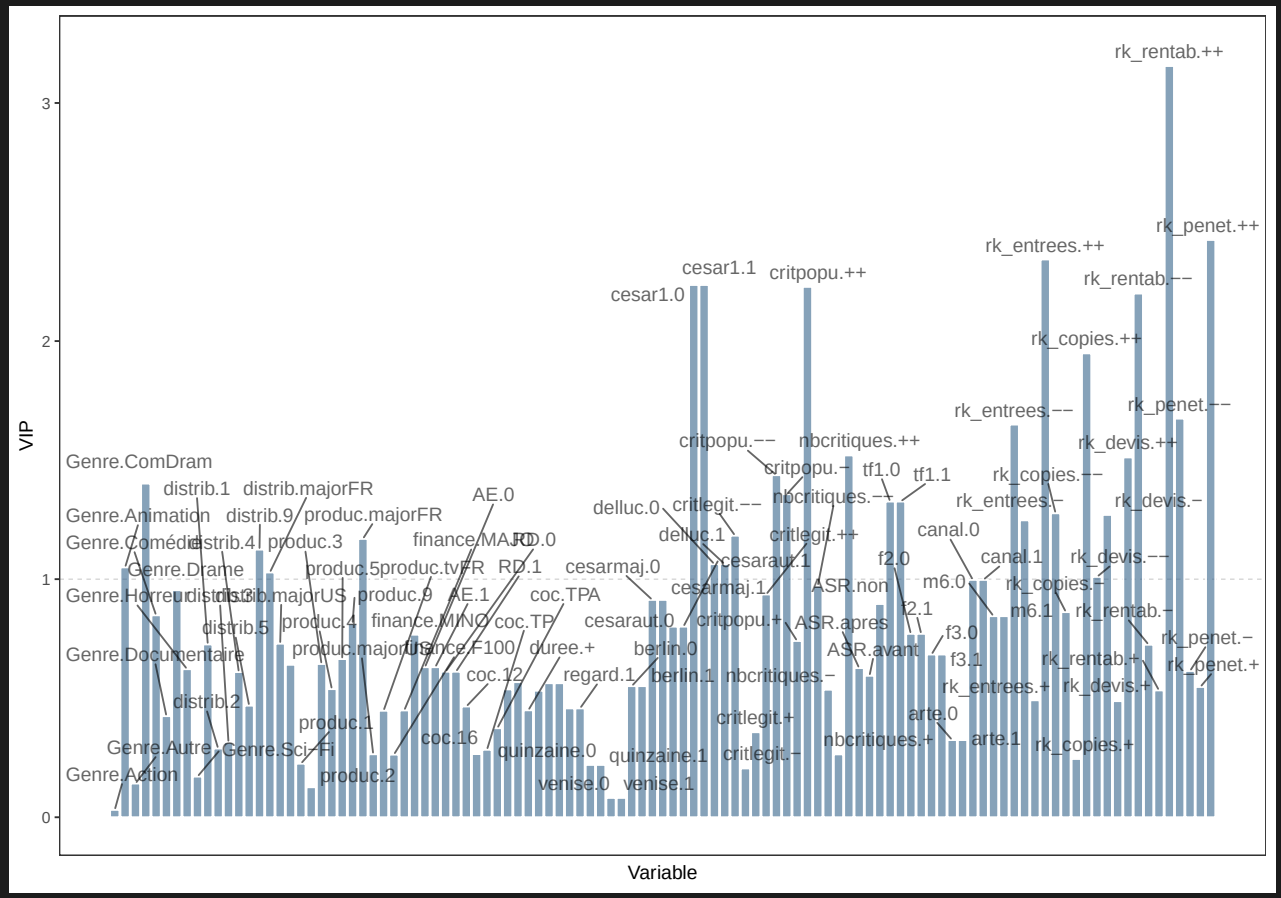
- Conséquence : les aides à l'interprétation ne sont pas les mêmes que pour les ACM et ACP, d'où petit coût d'entrée pour le nouvel utilisateur.

Exemple 1

Carrières cinématographiques

- Parmi la population des cinéastes français ayant réalisé un premier film, on chercher à déterminer les facteurs favorisant la réalisation d'un deuxième film.
- Les données :
  - 1ers films d'initiative française sortis en France entre 2000 et 2011 (N = 1088)
  - données exhaustives
  - observations non-indépendantes
  - nombreuses variables et multicolinéarité
- Spécification du problème :
  - variable à expliquer = réalisation d'un 2nd film dans les 5 ans (oui/non)
  - 32 variables explicatives (continues ou catégorielles) = caractéristiques du 1er film (production, réception)

sélection de variables par les VIP



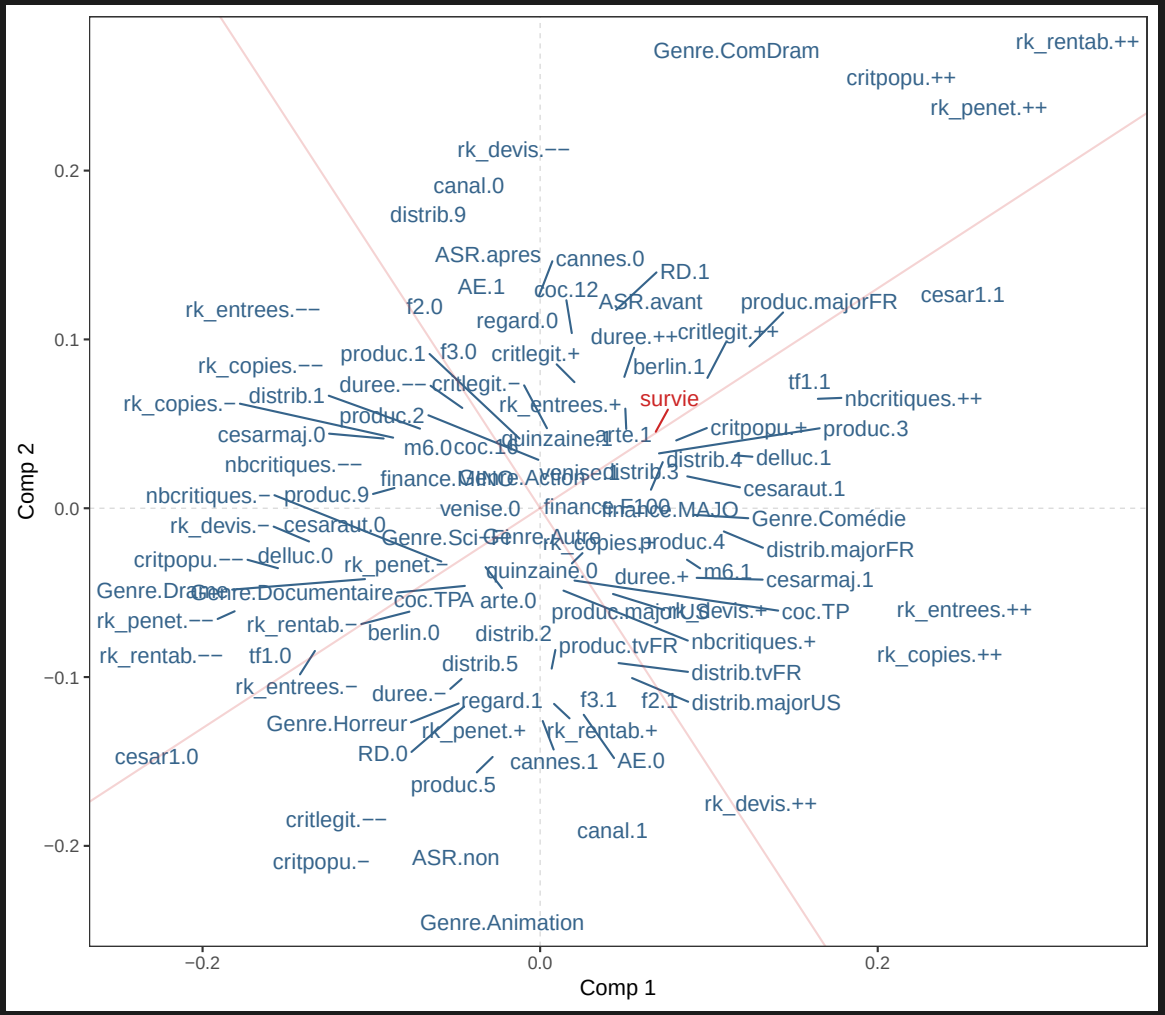
27 VIP (“Variable Importance for Prediction”) ≥ 1 (parmi les 107)

coefficients standardisés

	coefficients	std error	t-value	p-value	2.5%	97.5%
Genre.Animation	-0.01270	0.00468	-2.710	0.02398	-0.02189	-0.00351
Genre.ComDram	0.01866	0.00842	2.217	0.05384	0.00214	0.03517
distrib.9	0.00206	0.00433	0.476	0.64571	-0.00644	0.01056
distrib.majorFR	0.00671	0.00446	1.506	0.16634	-0.00203	0.01546
produc.majorFR	0.01255	0.00673	1.865	0.09507	-0.00065	0.02575
cesar1.0	-0.02225	0.00396	-5.612	0.00033	-0.03003	-0.01447
cesar1.1	0.02225	0.00396	5.612	0.00033	0.01447	0.03003
delluc.0	-0.00913	0.00597	-1.529	0.16072	-0.02085	0.00259
delluc.1	0.00913	0.00597	1.529	0.16072	-0.00259	0.02085
critlegit.–	-0.01509	0.00234	-6.459	0.00012	-0.01967	-0.01050
critpopu.–	-0.01204	0.00242	-4.967	0.00077	-0.01679	-0.00728
critpopu.-	-0.01770	0.00308	-5.750	0.00028	-0.02374	-0.01166
critpopu.++	0.02615	0.00335	7.804	0.00003	0.01957	0.03272
nbcritiques.++	0.01393	0.00526	2.648	0.02656	0.00361	0.02426
tf1.0	-0.01328	0.00434	-3.060	0.01358	-0.02180	-0.00476
tf1.1	0.01328	0.00434	3.060	0.01358	0.00476	0.02180
rk_entrees.–	-0.00598	0.00240	-2.485	0.03471	-0.01069	-0.00126
rk_entrees.-	-0.01270	0.00484	-2.625	0.02760	-0.02219	-0.00321
rk_entrees.++	0.01319	0.00285	4.628	0.00124	0.00760	0.01878
rk_copies.–	-0.00473	0.00260	-1.823	0.10157	-0.00983	0.00036
rk_copies.++	0.00912	0.00494	1.846	0.09799	-0.00057	0.01882
rk_devis.–	0.00959	0.00454	2.112	0.06391	0.00068	0.01849
rk_devis.-	-0.01010	0.00402	-2.512	0.03319	-0.01799	-0.00221
rk_devis.++	-0.00006	0.00621	-0.009	0.99271	-0.01224	0.01212
rk_rentab.–	-0.01942	0.00231	-8.389	0.00002	-0.02396	-0.01488
rk_rentab.++	0.03424	0.00344	9.967	0.00000	0.02750	0.04099
rk_penet.–	-0.01487	0.00412	-3.607	0.00569	-0.02296	-0.00678
rk_penet.++	0.02713	0.00629	4.311	0.00196	0.01478	0.03948

AUC = 0.769

plan factoriel (1,2)

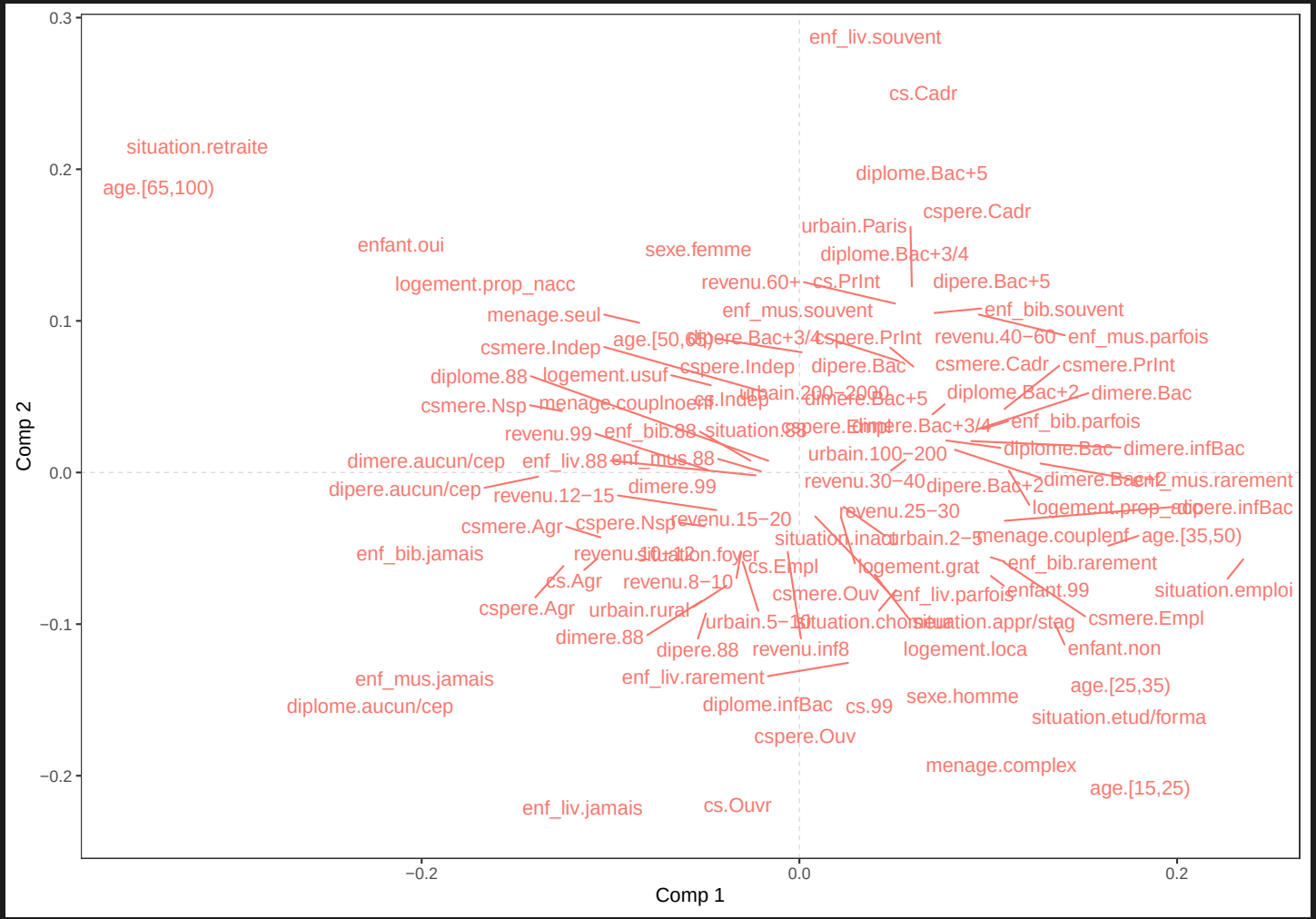


Exemple 2

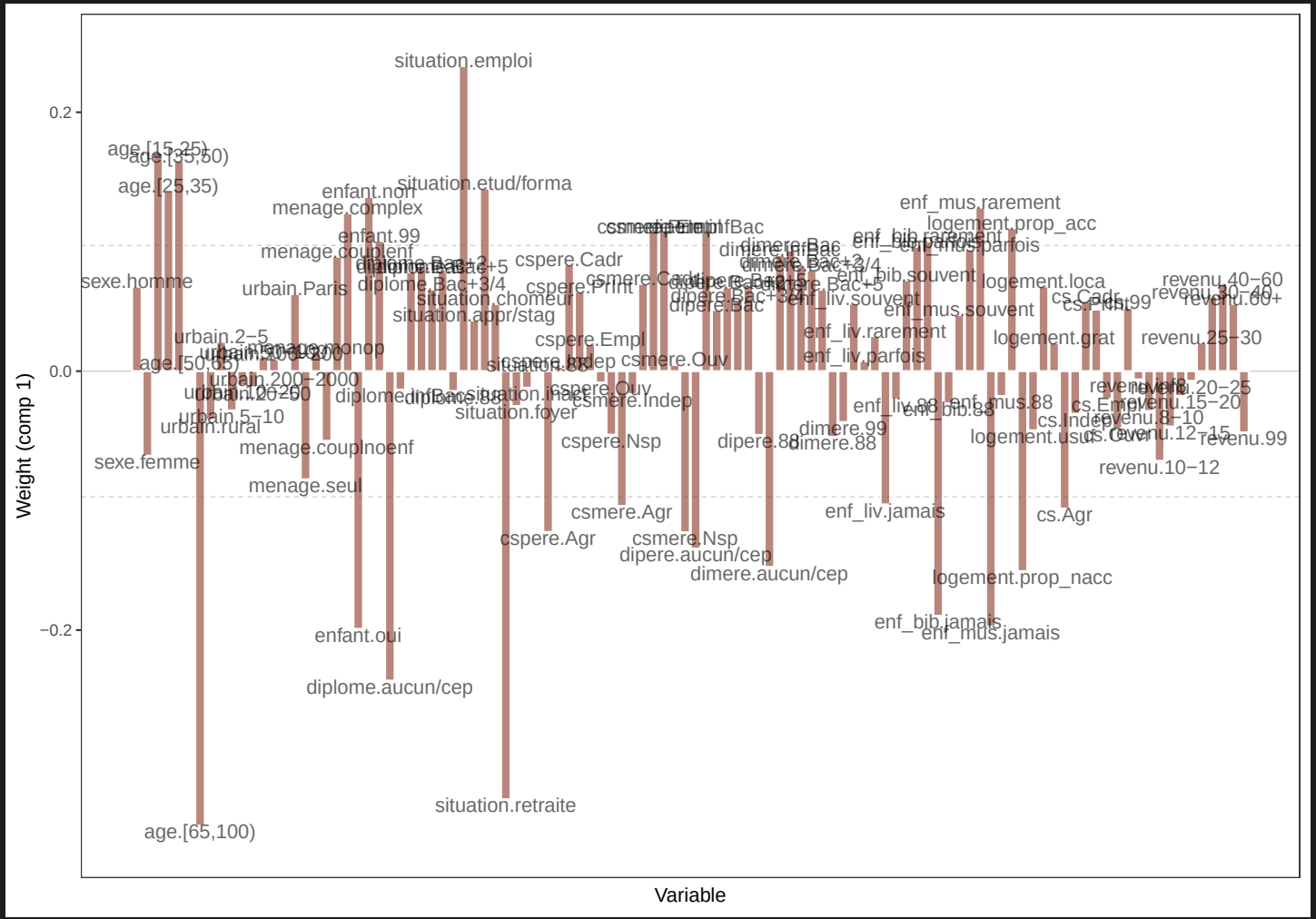
Goûts musicaux

- à partir des 12 genres musicaux particulièrement aimés + 12 genres pas aimés du tout
- propriétés sociales : sexe, âge, type de ménage, présence d'enfants, niveau de diplôme, revenus, CS, situation d'emploi, CS et diplôme de la mère, CS et diplôme du père, statut d'occupation du logement, unité urbaine, pratiques culturelles pendant l'enfance (livres, bibliothèque, musique)
- on construit l'espace factoriel des propriétés sociales qui "explique" le mieux l'espace des goûts musicaux

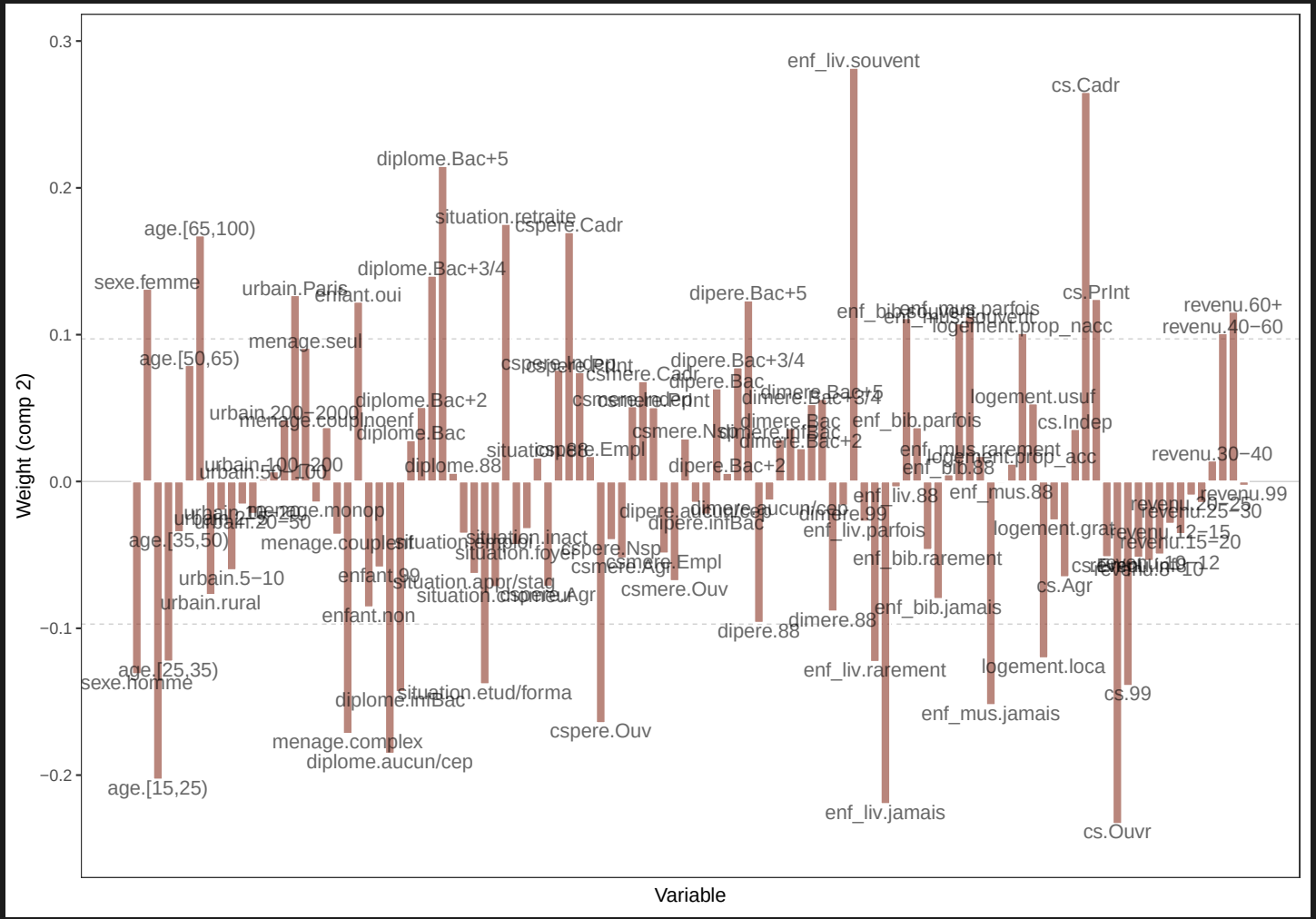
variables explicatives dans le plan (1,2)



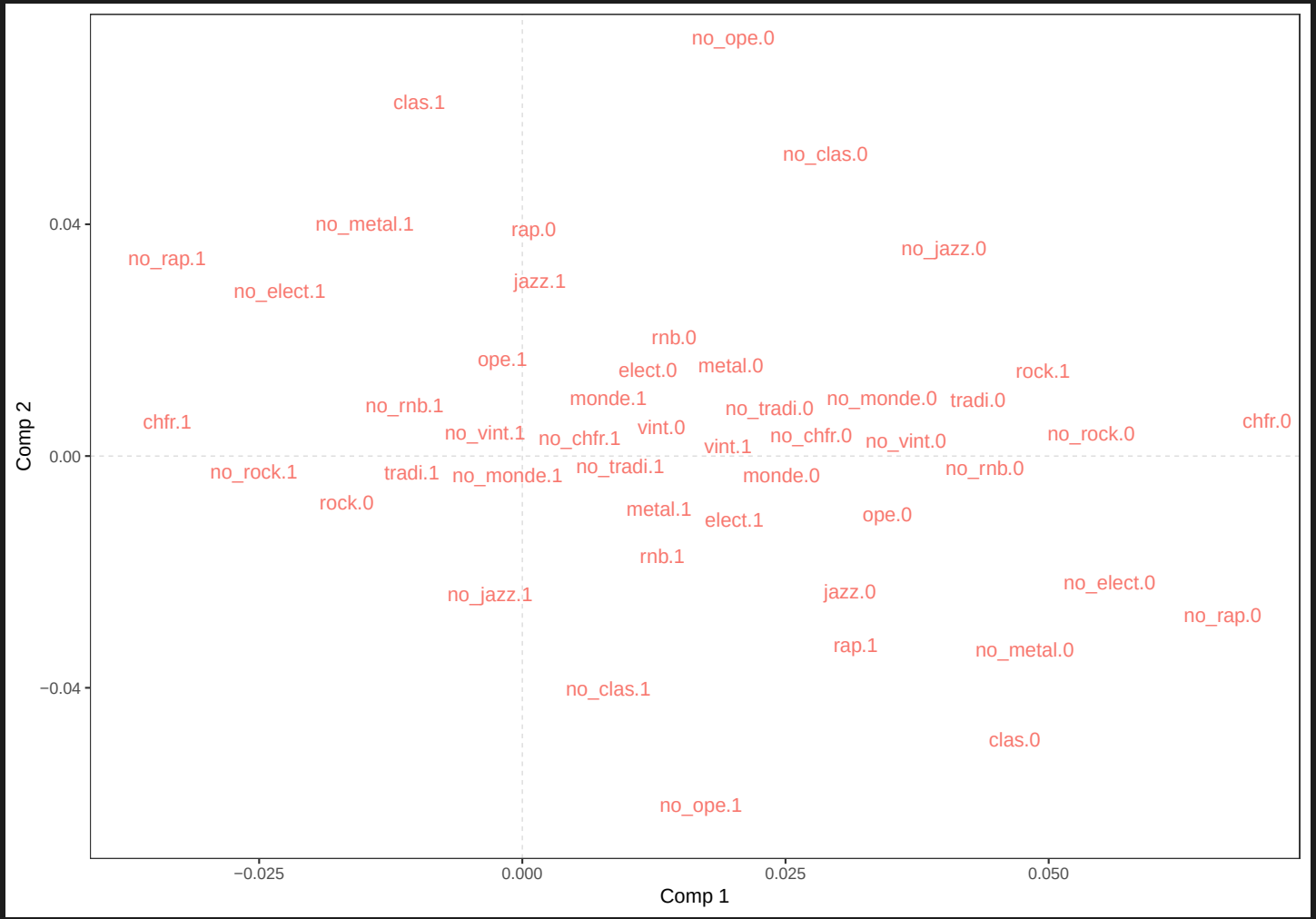
poids des variables explicatives dans l'axe 1



poids des variables explicatives dans l'axe 2



variables à expliquer dans le plan (1,2)

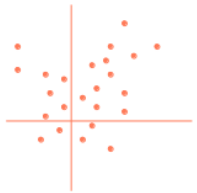
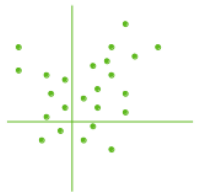


	t1	t2
rock.1	0.0849	0.0031
rap.1	0.0732	0.0585
elect.1	0.0483	0.0044
no_rock.1	0.0432	0.0000
chfr.1	0.0327	0.0002
rnb.1	0.0282	0.0143
no_rap.1	0.0282	0.0158
no_elect.1	0.0206	0.0180
metal.1	0.0170	0.0033
vint.1	0.0132	0.0000

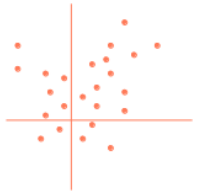
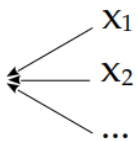
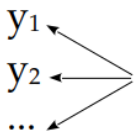
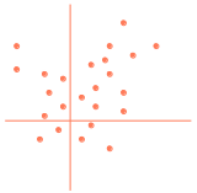
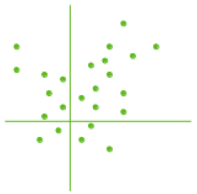
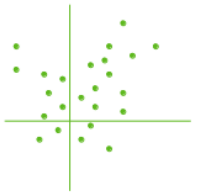
	t1	t2
clas.1	0.0071	0.0973
no_ope.1	0.0074	0.0747
no_clas.1	0.0022	0.0607
rap.1	0.0732	0.0585
jazz.1	0.0010	0.0423
no_jazz.1	0.0017	0.0250
ope.1	0.0029	0.0208
no_metal.1	0.0072	0.0197
no_elect.1	0.0206	0.0180
no_rap.1	0.0282	0.0158
rnb.1	0.0282	0.0143

Synthèse

Une seule variable à expliquer

Approche	y (variable à expliquer)	X (variables explicatives)
Variables supplémentaires ou PCR (Principal Component Regression)	continue ou catégorielle	
PLS1 (Partial Least Square Regression)	continue ou binaire	
Analyse Factorielle Inter-classes	catégorielle	

Plusieurs variables à expliquer

Approche	Y (variables à expliquer)	X (variables explicatives)
Variables supplémentaires		
Variables supplémentaires ou PCR (Principal Component Regression)		
Analyse Factorielle sur Variables Instrumentales		$\beta_1.x_1 + \beta_2.x_2 + ...$
PLS2 (Partial Least Square Regression)		

Analyses conditionnelles

- Autre approche intégrant l'analyse factorielle des données et la régression

- Principe =
  - contraindre les axes de l'analyse factorielle à être indépendants (i.e. orthogonaux) d'une ou plusieurs variables supplémentaires
  - autrement dit, construire une analyse factorielle "toute chose (de ces variables supplémentaires) égale par ailleurs" (Bry *et al*, 2016).

- La comparaison des résultats de l'analyse factorielle "classique" et de ceux de l'analyse factorielle conditionnelle est également un moyen d'étudier les effets de structure.

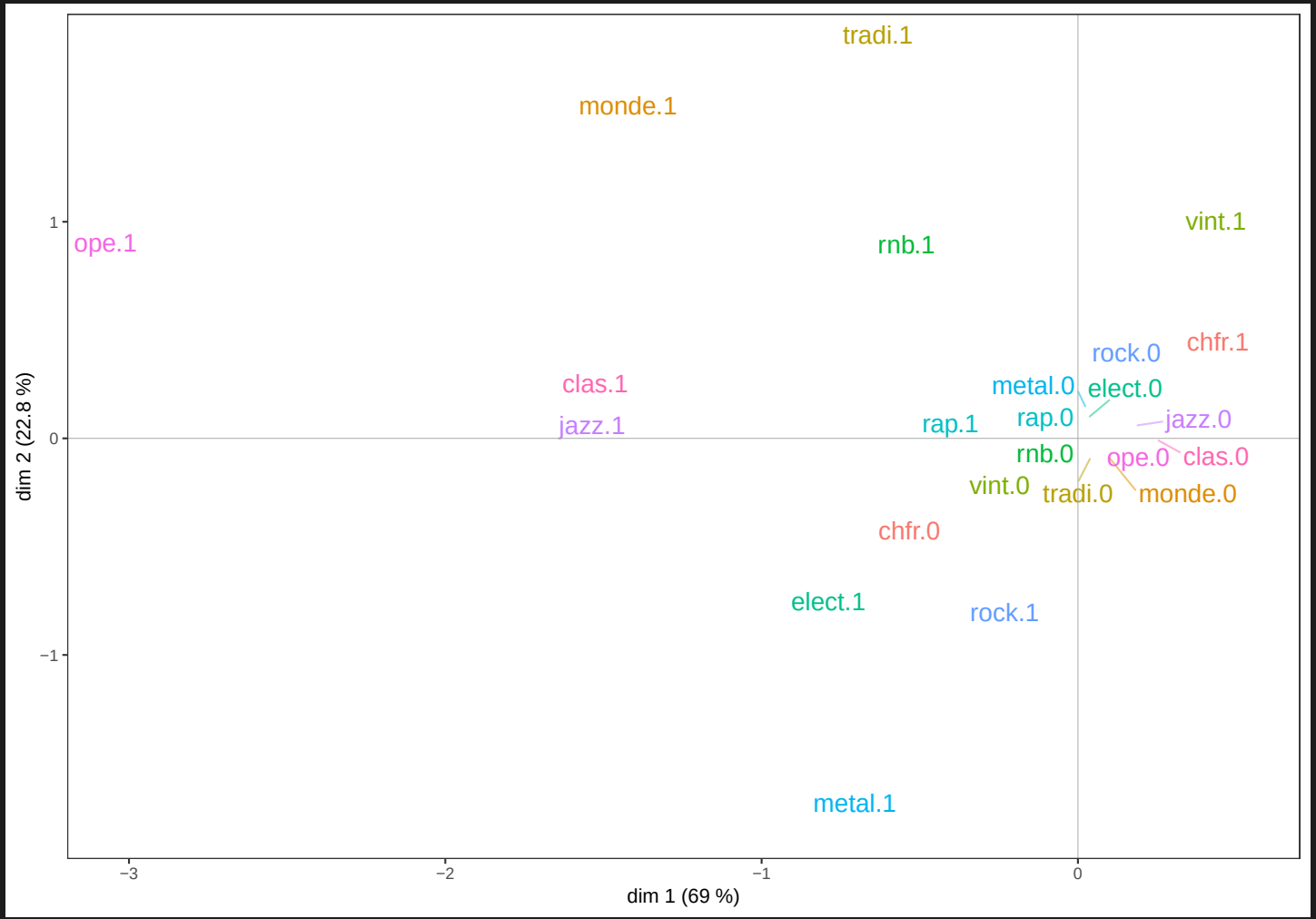
Analyses intra-classes

- Situation :
  - On a un ensemble de variables X et une variable catégorielle z dont on souhaite “éliminer l’effet”
- Principe :
  - construire l’espace factoriel des X rendant compte des structures communes aux différentes classes de la variable z
  - concrètement, les modalités de z seront toutes au centre de l’espace factoriel.

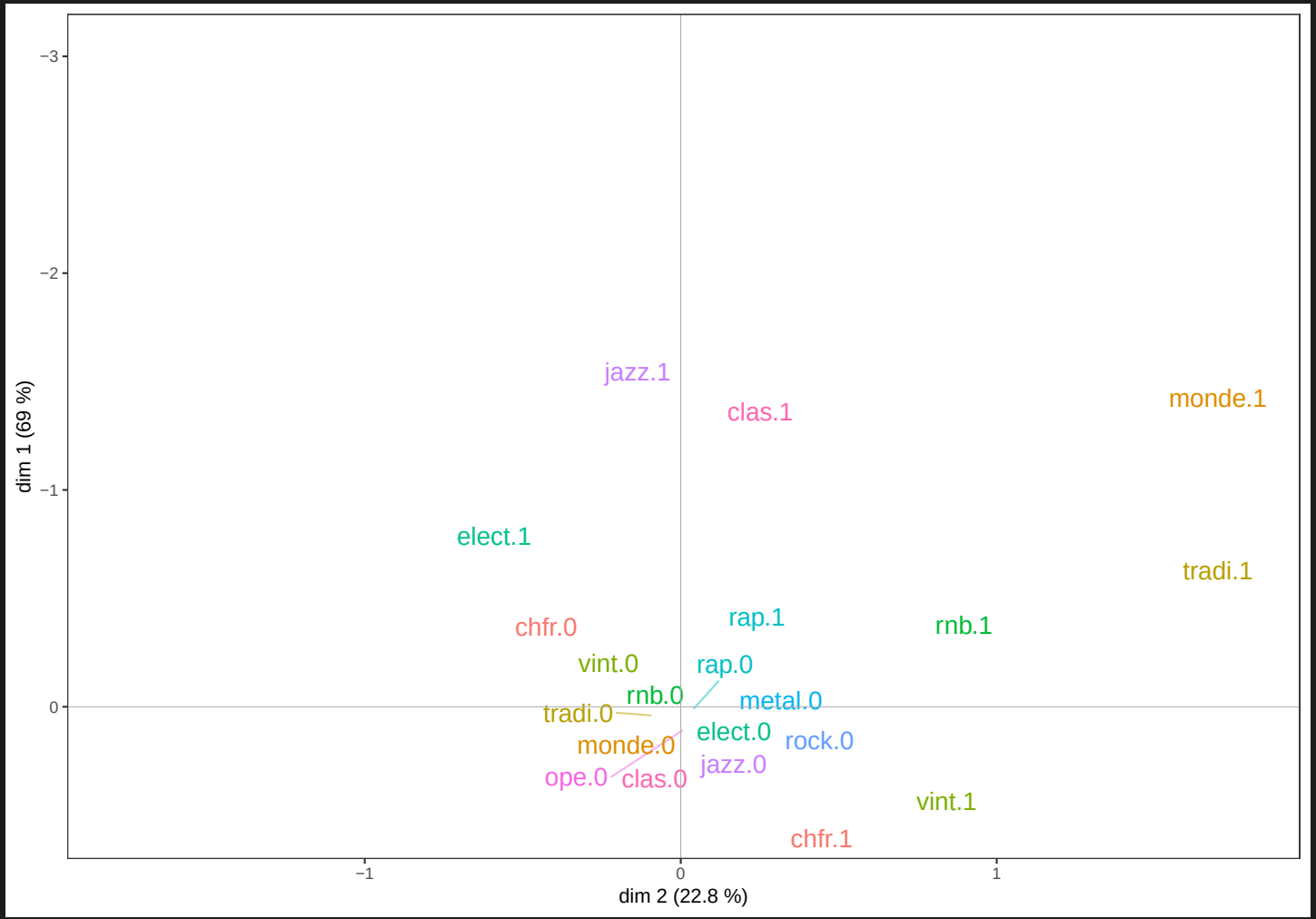
Exemple des goûts musicaux

- à partir des 12 genres musicaux particulièrement aimés
- on souhaite éliminer les liaisons avec l'âge, pour mieux voir apparaître celles avec le diplôme et du sexe : espace factoriel conditionnellement à l'âge

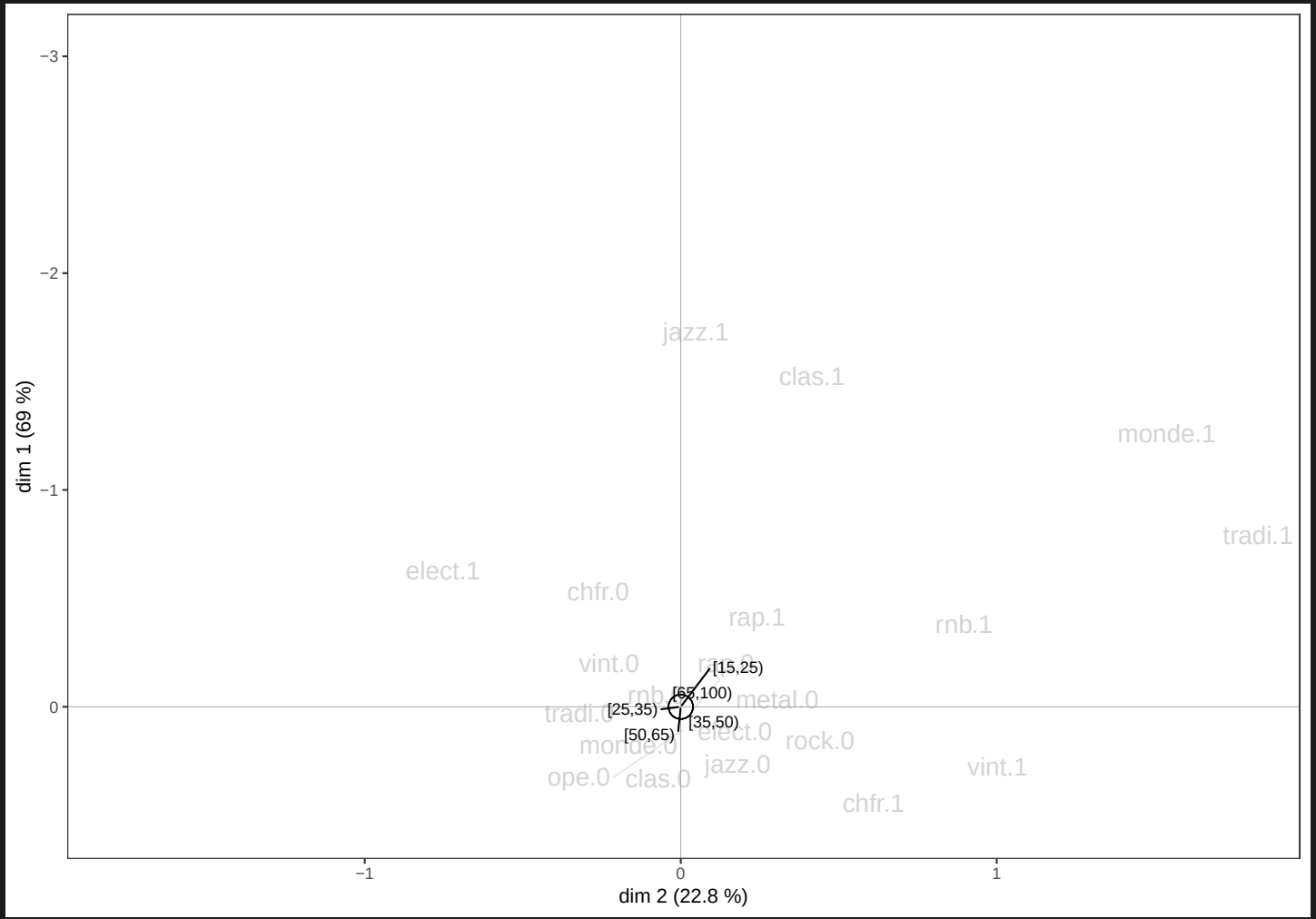
espace des modalités actives



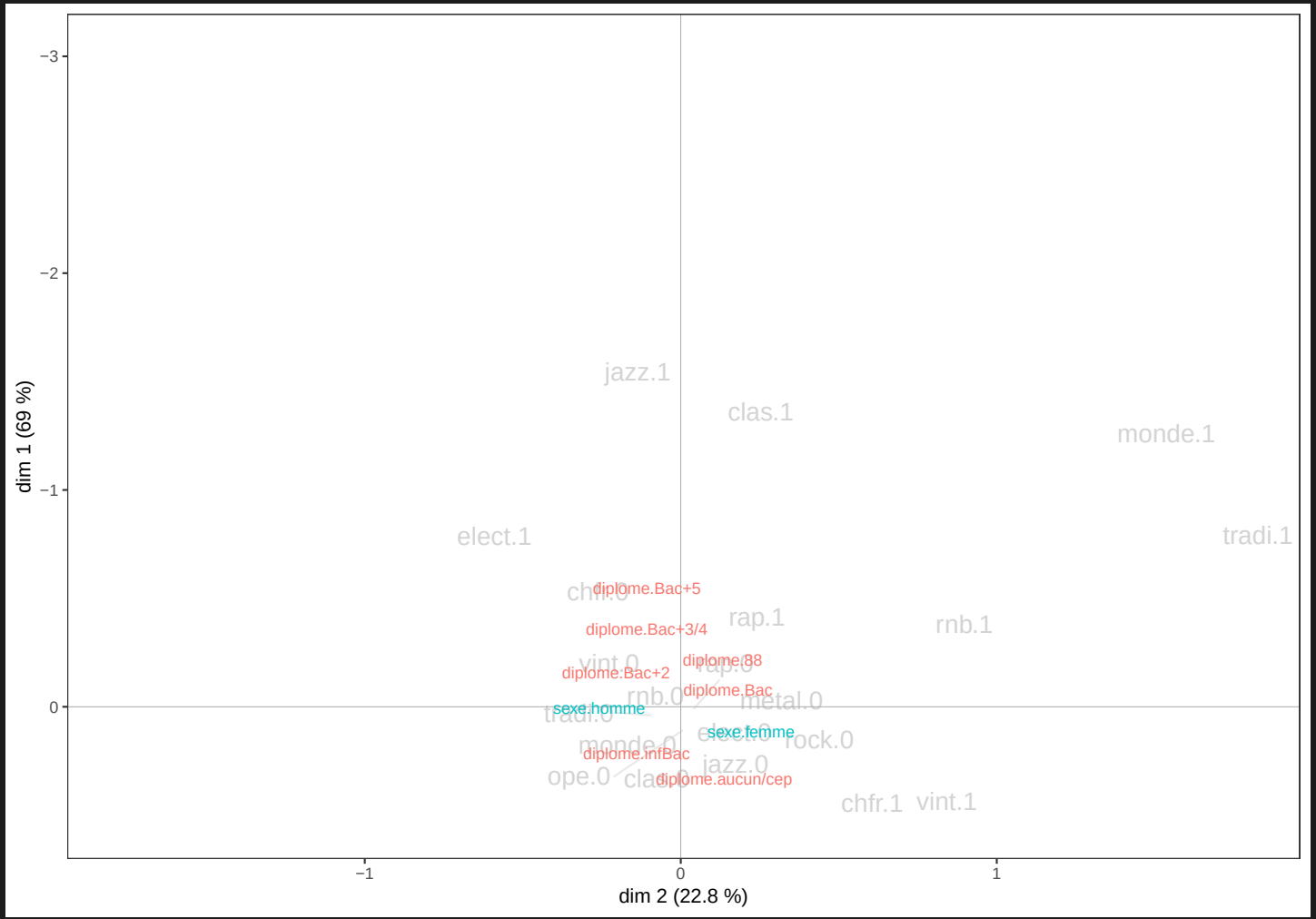
après une rotation à 90°



projection de l’âge



autres variables supplémentaires



- Si les variables X sont continues, ACP intra-classes
  - l'ACP est réalisée en centrant les variables de X, pour chaque observation, par rapport à la moyenne de la classe de l'observation, au lieu de la moyenne globale
- Si les variables X sont catégorielles, ACM intra-classes
  - aussi appelée "ACM conditionnelle" (Escofier, 1990)
  - l'ACM est réalisée en centrant les modalités des variables de X, pour chaque observation, par rapport à la moyenne de la classe de l'observation (fréquences conditionnelles), au lieu de la moyenne globale (fréquences marginales).

Variables instrumentales orthogonales

- Situation :
  - Au lieu d'une variable catégorielle z, on a un ensemble de variables Z dont on veut éliminer l'influence.
- NB : Les analyses intra-classes peuvent être considérées comme des cas particuliers des analyses sur variables instrumentales orthogonales, avec une variable instrumentale orthogonale unique et catégorielle

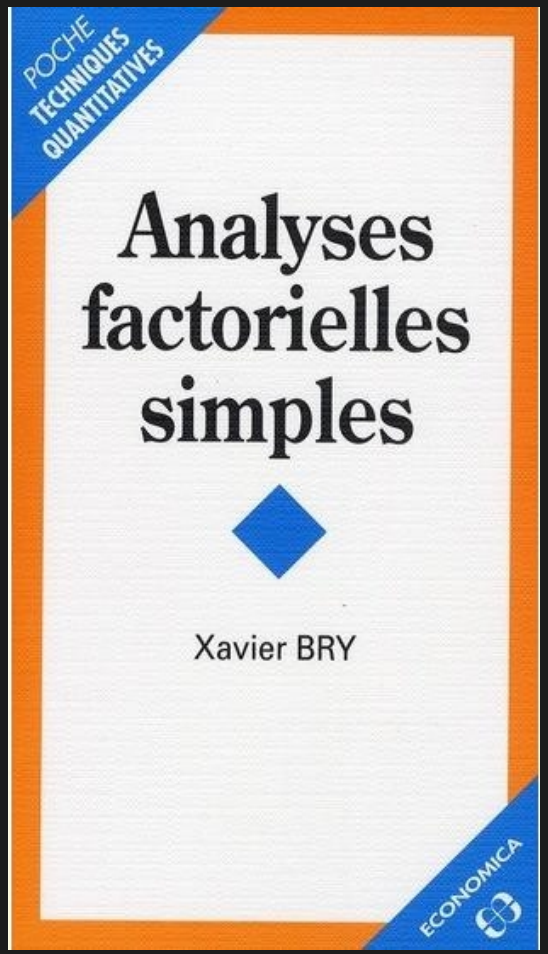
- Si les variables X sont continues : ACP sur variables instrumentales orthogonales
    1. Calcul d'une régression linéaire par variable de X, en prenant cette variable comme variable à expliquer et les variables Z comme variables explicatives
    2. ACP de l'ensemble des résidus des régressions (X "purgées" des effets des Z, i.e. partie des X non-expliquées par les Z)

- Si les variables X sont catégorielles : ACM sur variables instrumentales orthogonales
    1. ACM des X, en conservant l'ensemble des dimensions de l'espace
    2. Calcul d'une régression linéaire par dimension de l'ACM, avec les coordonnées des observations sur l'axe factoriel comme variable à expliquer et les variables Z comme variables explicatives
    3. ACP de l'ensemble des résidus des régressions

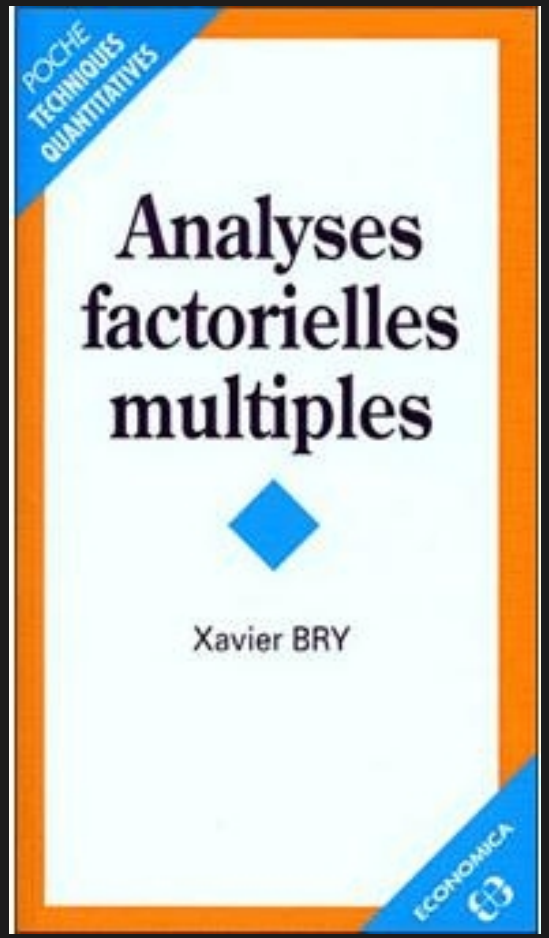
Usages

- Analyses intra-classes : les **structures communes** à différents groupes d'observations
- Analyses inter-classes : les **structures qui différencient** plusieurs groupes d'observations
- Très utiles pour comparer des groupes sociaux (classes sociales, classes d'âge, etc.), des unités géographiques (régions, pays, etc.), des périodes historiques…
- … et pour étudier les **homologies structurales**

Bibliographie



+



Autres références

Abdi H., 2007, “Discriminant Correspondence Analysis”, In: Neil Salkind (Ed.), *Encyclopedia of Measurement and Statistics*, Thousand Oaks (CA): Sage. [lien]

Bry X., Robette N., Roueff O., 2015, “A dialogue of the deaf in the statistical theater? Addressing structural effects within a geometric data analysis framework”, *Quality & Quantity*. [version fr]

Escofier B., 1990, “Analyse des correspondances multiples conditionnelle”, *La revue de Modulad*, 5, 13-28. [lien]

Lebart L., Morineau A., Piron M., 2000, *Statistique exploratoire multidimensionnelle*, Dunod. [lien]

Rouanet H., Lebaron F., Le Hay V., Ackermann W. et Le Roux B., 2002, “Régression et analyse géométrique des données : réflexions et suggestions”, *Mathématiques et sciences humaines*, 160. [lien]

Tenenhaus M., 1998, *La Regression PLS. Théorie et Pratique*, Editions Technip, Paris.

Application avec R

- packages GDAtools ou ade4 pour les analyses factorielles contraintes
- avec éventuellement explor (de Julien Barnier) et factoextra pour les interprétations
- packages pls et morepls pour les régressions PLS
- voir le tutoriel “TutoMate-Facto-Tutoriel.html” pour des exemples