Séance 4 : L'analyse géométrique de données

Introduction à la sociologie quantitative, niveau 1

Samuel Coavoux

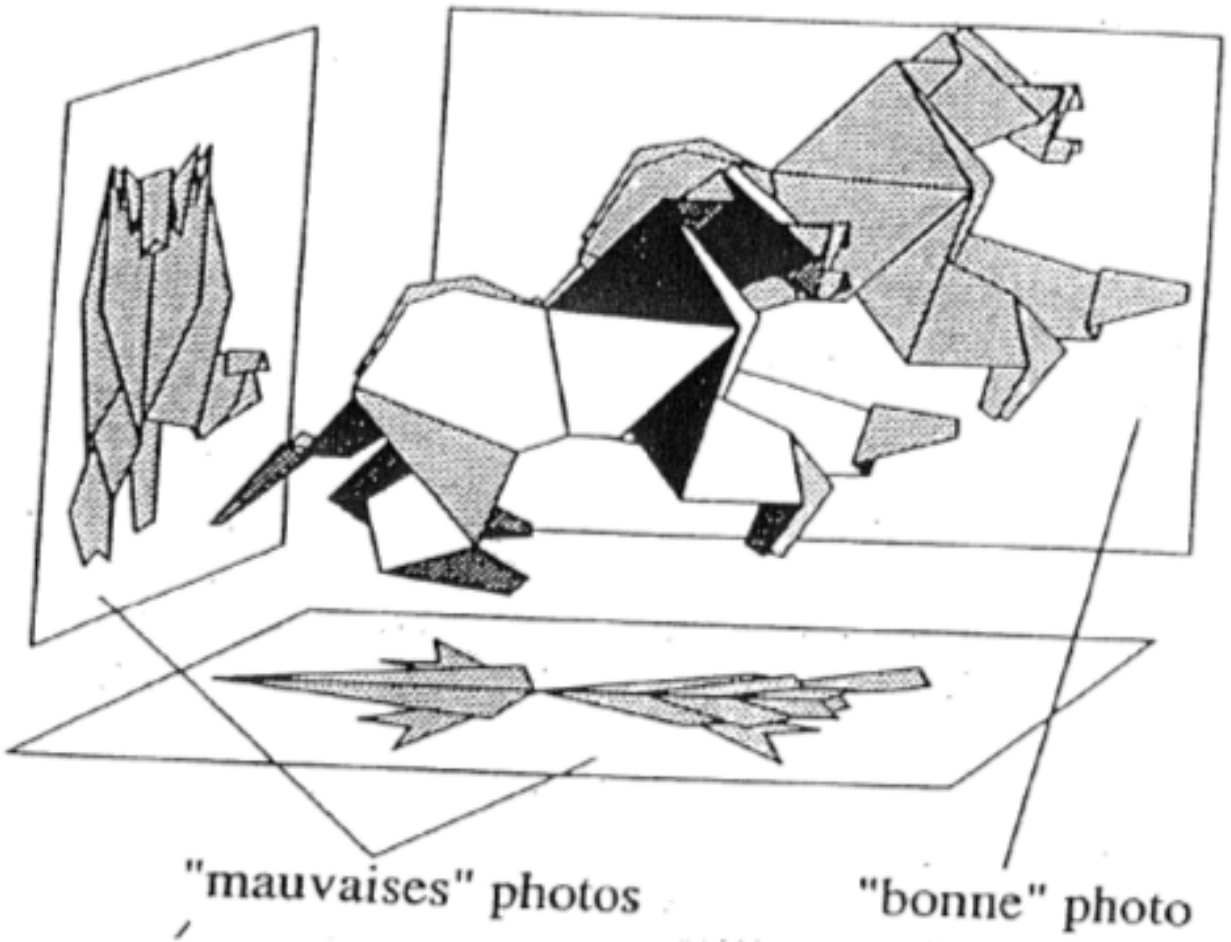
Introduction générale : une statistique géométrique

Première définition

L'**analyse géométrique des données** (AGD) est l'une des dénominations actuellement privilégiée pour désigner une famille de techniques statistiques, les **analyses factorielles**.

Ces techniques ont pour point commun de s'appuyer sur des calculs similaires pour **réduire un espace pluridimensionnel à un ensemble restreint de dimensions**, généralement entre 2 à 4. En l'occurence, les dimensions sont les variables (ou les modalités de variables) décrivant une observation statistique. En ce sens, l'AGD **extrait d'un grand nombre de variable un nombre restreint de variables synthètiques**.

La meilleure projection d'un nuage



La meilleure projection d'un nuage

Le problème: en quoi une représentation est-elle supérieure à une autre?
La réponse statistique: la meilleure projection est celle qui maximise la variance, c'est-à-dire qui maintient au mieux les différences entre les observations.

Histoire

L'analyse de données est parfois considérée comme une branche des statistique opposée à l'analyse inférentielle, et notamment à l'idée selon laquelle les modèles statistiques viendraient seulement confirmer ou infirmer des hypothèses établies par les chercheurs.

Jean-Baul Benzécri, qui se réclame de cette perspective d'analyse de données et est notamment opposé à l'analyse par régression, et ses collaborateurs produisent à partir des années 1960 les principales techniques d'analyse factorielle.

Usages en sociologie

En sociologie, le succès de l'AGD est du à la rencontre avec le modèle théorique du champ chez Pierre Bourdieu (Duval 2013). On note des "affinités electives" entre la pensée relationelle de Bourdieu et l'approche géométrique de la statistique multivariée (Rouanet, Le Roux, et Ackermann 2000).

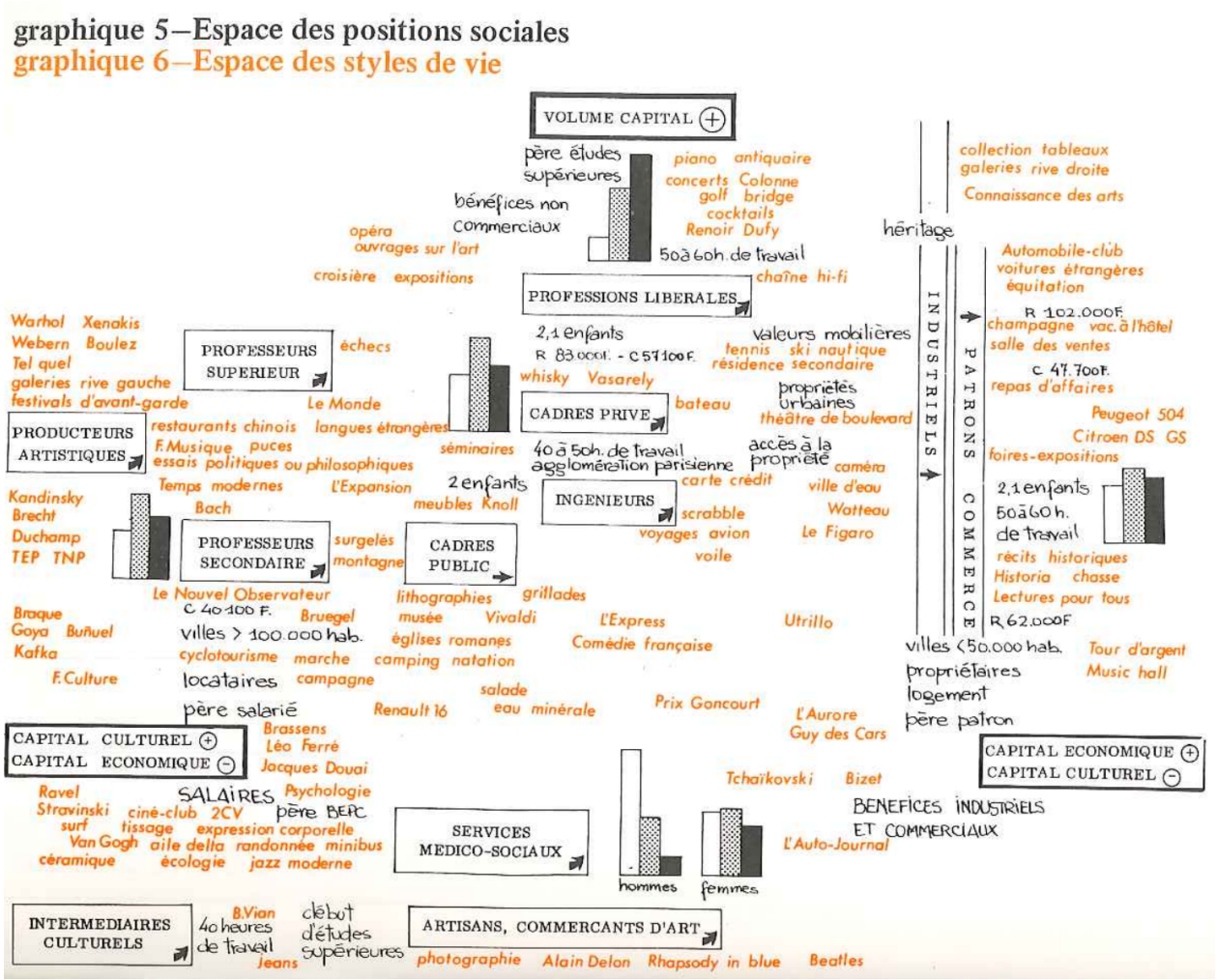
Comme Benzécri, Bourdieu est méfiant envers la statistique inférentielle, critiquant notamment la "sociologie des variables" incarnée selon lui par l'école de Paul Lazarsfeld.

Une technique française? Spearman avait développé une technique proche, encore souvent utilisée en psychologie ou en biologie.

Bourdieu et l'AGD

Ceux qui connaissent les principes de l'analyse des correspondances multiples saisiront l'affinité entre cette méthode d'analyse mathématique et la pensée en termes de champ. Ayant pris en compte l'ensemble des agents efficients (individus et, à travers eux, institutions) et l'ensemble des propriétés – ou des atouts – qui sont au principe de leur action, on peut attendre de l'analyse des correspondances qui, ainsi utilisée n'a rien de la méthode purement descriptive que veulent y voir ceux qui l'opposent à l'analyse de régression, qu'elle porte au jour la structure des oppositions, ou, ce qui revient au même, la structure de la distribution des pouvoirs et des intérêts spécifiques qui détermine, et explique, les stratégies des agents. (Bourdieu 2000)

Ceci n'est pas une ACM



Toutes choses (in)égales par ailleurs

Les oppositions entre AGD et régression sont à nuancer (Rouanet et al. 2002) notamment parce qu'elles concernent plus l'usage de ces techniques que leurs potentialités. Malgré sa réputation de technique exploratoire, l'AGD ne permet pas une approche purement inductive, ne serait-ce que parce qu'elle dépend des variables produites et analysées, qui dépendent elles-mêmes d'hypothèses plus ou moins explicitées.

Cependant, les régressions cherchent à décomposer l'effet de différentes variables et à raisonner toutes choses égales par ailleurs ; l'AGD cherche au contraire les corrélations en permet de raisonner **toutes choses inégales par ailleurs**.

Objectifs: une réduction contrôlée de la dimensionnalité

"Systématiser les démarches de l'analyse descriptive" (Volle 1997). On part d'une situation dans laquelle un individu (au sens statistique: peut-être une personne, une institution, etc.) est définit par de nombreuses variables/modalités de variables.

Il s'agit de "consentir à une perte en information afin d'obtenir un gain en signification" (Volle 1997) et tout en cherchant à limiter cette perte.

Objectifs: une réduction contrôlée de la dimensionnalité

L'AGD vise à trouver un nombre restreint de méta-variables synthétiques non corrélées, que l'on appelera axes, facteurs ou composants, tels que:

- il y ait une réduction importante de la dimensionnalité de l'information (on passe d'un grand nombre de variables à un nombre restreint de méta-variables) ;
- les méta-variables choisit sont celles qui résument le mieux les variables d'origine, c'est-à-dire celles qui **concentrent la plus grande partie de la variance** parmi toutes les méta-variables possibles.

Techniques

Il existe de nombreuses techniques dans la famille de l'AGD. Les plus connues sont les suivantes:

- Analyse factorielle des correspondances (AFC) – *correspondence analysis* (CA) : deux variables catégorielles (il s'agit de l'analyse géométrique d'un tableau croisé);
- Analyse en composantes principales (ACP) – *principal component analysis* (PCA) : uniquement des variables quantitatives (plus des variables catégorielles en illustratif);
- Analyse des correspondances multiples (ACM) – *multiple correspondence analysis* (MCA) : plusieurs variables catégorielles.

Dans ce cours, nous nous concentrerons sur l'ACM, la plus utilisée aujourd'hui dans la sociologie française, et secondairement sur l'ACP.

Pourquoi utiliser de l'AGD?

Lorsque l'on souhaite simplifier une base de données. Mais plusieurs cas possibles:

- explorer une base complexe, en repérant des associations de nombreuses variables que l'on ne peut pas nécessaire repérer deux à deux (il faut tout de même accompagner l'AGD d'autres statistiques descriptives) ;
- résumer synthétiquement l'information (réduire la dimensionnalité) pour intégrer les nouvelles variables dans d'autres analyses ;
- mettre au jour des variables latentes, inobservables.

Le troisième cas est celui de l'approche bourdieusienne classique: la mise au jour des capitaux structurant un espace social, qui sont ensuite croisé à des variables illustratives, explicatives.

Plan du cours

1. Le raisonnement géométrique: notions générales
2. Théorie de l'ACM
3. Interprétation de l'ACM
4. Théorie et interprétation de l'ACP

Un raisonnement géométrique

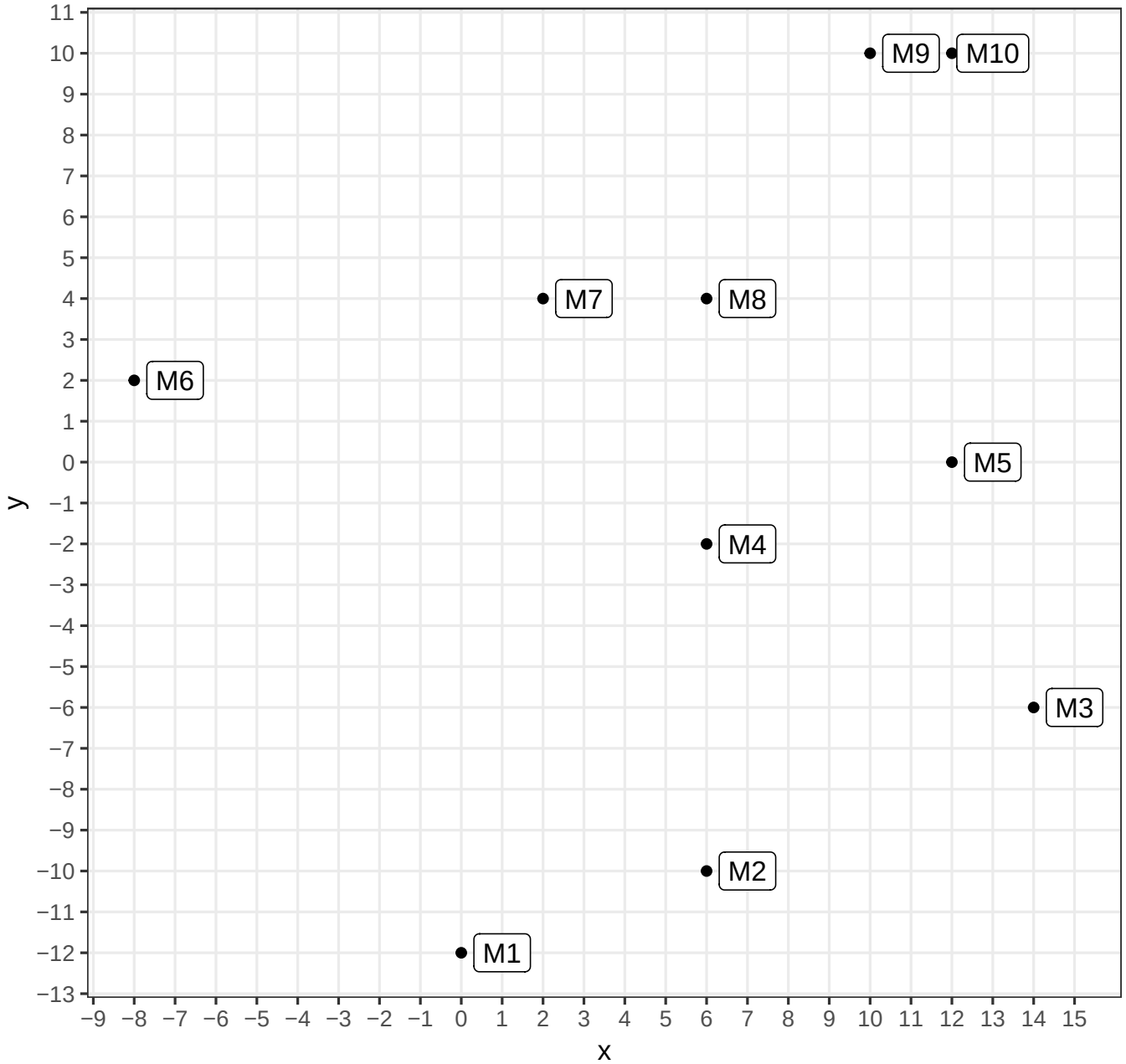
Objectifs

Dans cette première partie, nous allons chercher à comprendre à partir d'une exemple fictionnel comment il est possible de résumer l'information contenue dans un nuage de points. L'objectif est double:

- comprendre le raisonnement géométrique de l'AGD: il s'agit de dessiner un nuage puis de chercher le ou les axes qui permettent sa meilleure représentation
- définir les concepts et indicateurs principaux que l'on utilisera par la suite: coordonnées, variance, axe principal, inertie, contribution, cosinus carré.

Les données initiales

label	x	y	weight
M1	0	-12	1
M2	6	-10	1
M3	14	-6	1
M4	6	-2	1
M5	12	0	1
M6	-8	2	1
M7	2	4	1
M8	6	4	1
M9	10	10	1
M10	12	10	1

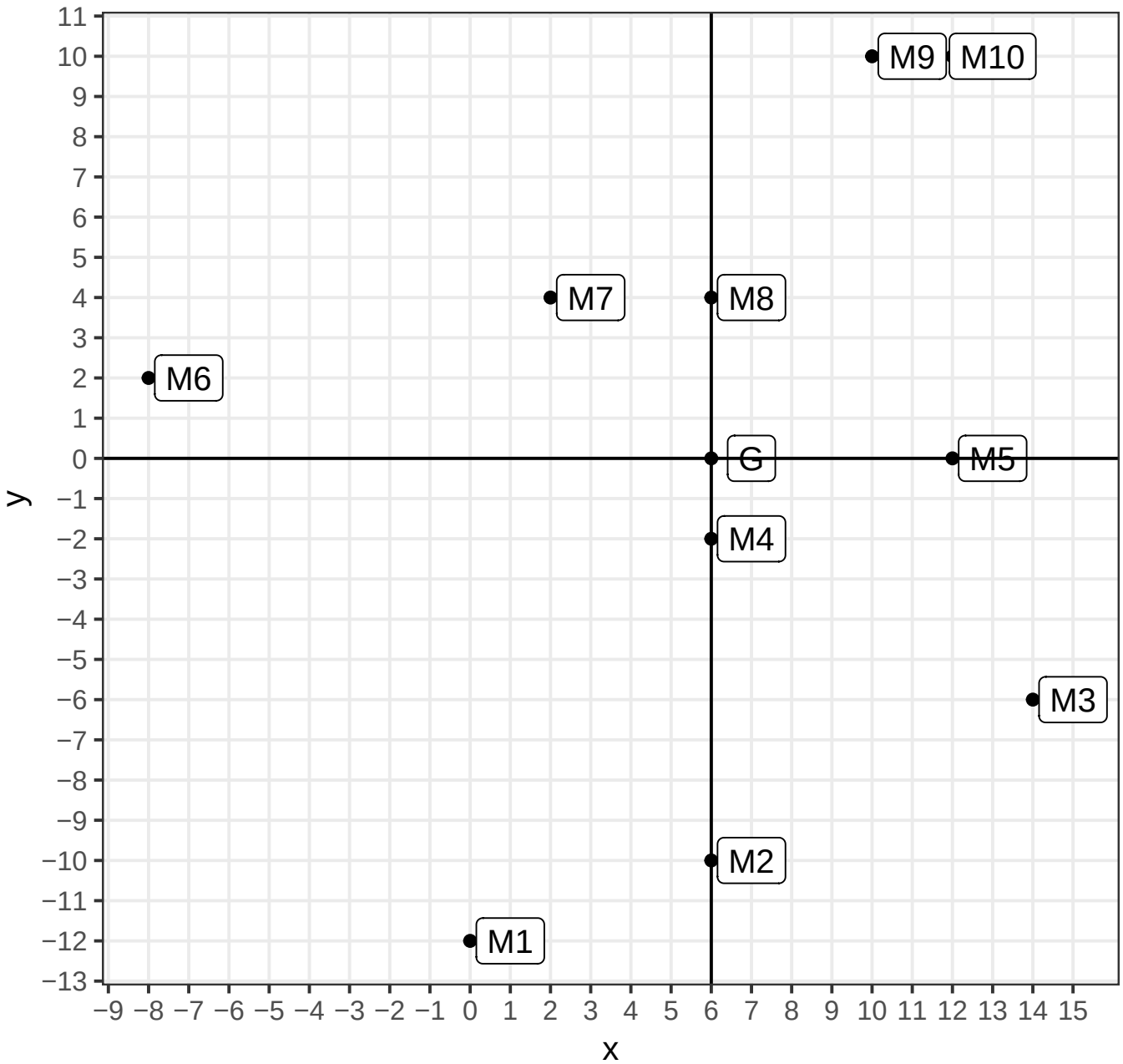

M9
M10
M7
M8
M6
M5
M4
M3
M2
M1
y
x

Trouver le centre du nuage

Le centre du nuage a pour coordonnés la moyenne pondérée des coordonnées des points.

Il s'agit du point tel que la somme des vecteurs entre chaque point du nuage et le centre est nulle.

$$\frac{1}{n}\sum_{i=1}^{n}\overrightarrow{GM_i} = \vec{0}$$


M9
M10
M7
M8
M6
G
M5
M4
M3
M2
M1
y
x

Distance entre deux points et variance du nuage

Le carré de la distance entre deux points est égal à la somme des carrés des distances entre leurs coordonées (Pythagore)

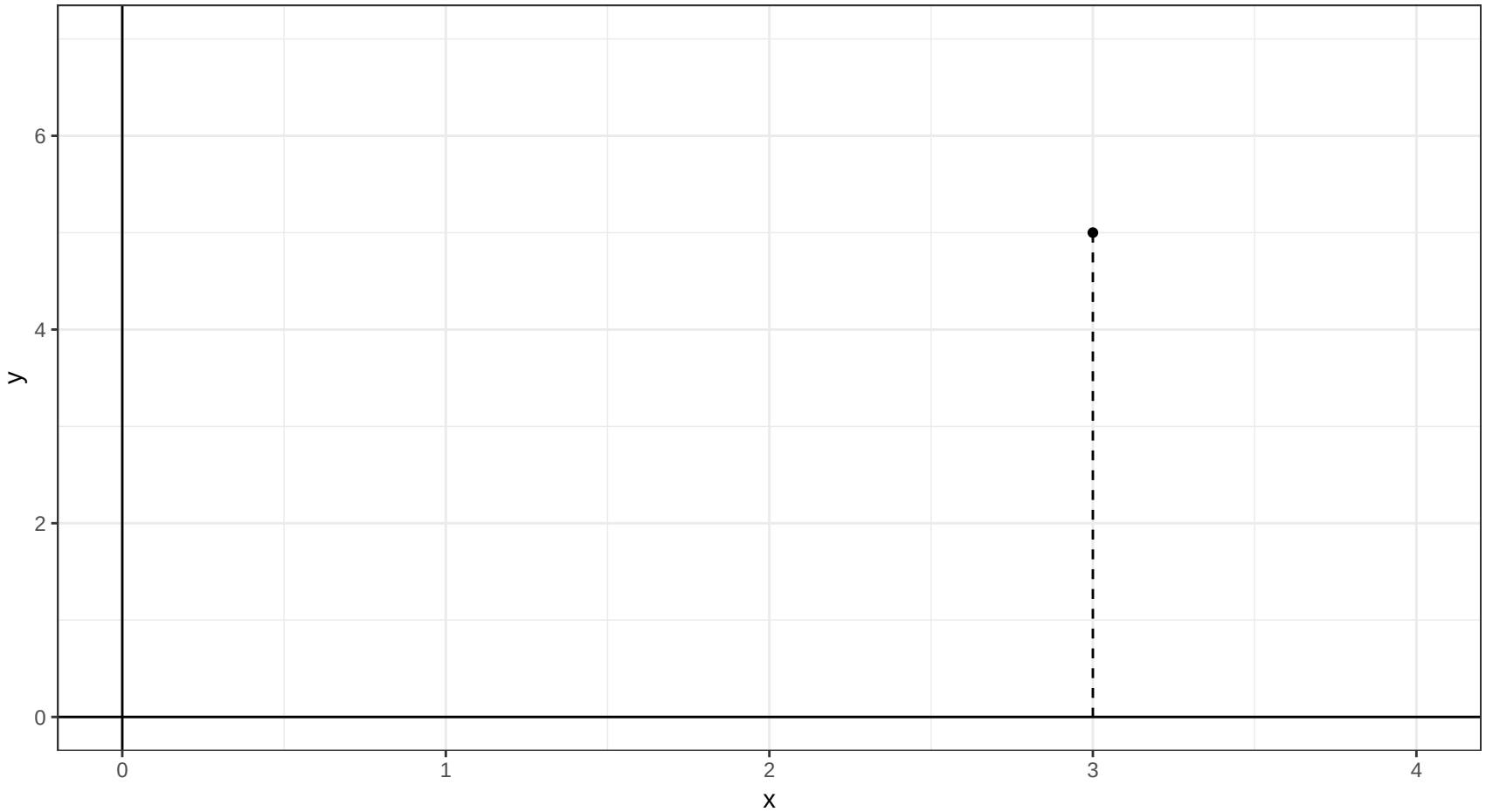
$$\delta_{ij} = (x_i - x_j)^2 + (y_i - y_j)^2$$

La variance du nuage de point est égale à la somme des distances carrés entre chaque points et le centre du nuage G.

$$\text{var} = \sum_{i=1}^{n} (x_i - x_G)^2 + (y_i - y_G)^2$$

Projection d'un point sur une droite

Les points sont définis par leurs coordonnés x et y. Il s'agit des projections orthogonales de ces points respectivement sur l'abscisse et l'ordonnée.



Variance d'un axe

La variance d'un axe est la variance de la projection des points du nuage sur cet axe.

$$var_x = \sum_{i=1}^{n} (x_i - x_G)^2$$

Contribution d'un point au nuage

On appelle contribution d'un point au nuage la part de la variance totale qui est due à ce point, c'est à dire le carré de la distance entre le point et le centre G du nuage.

$$ctr_i = \frac{(x_i - x_G)^2 + (y_i - y_G)^2}{var}$$

Par définition, la contribution est donc comprise entre 0 (si le point se confond avec le centre du nuage) et 1 (si le point est le seul point du nuage différent du centre).

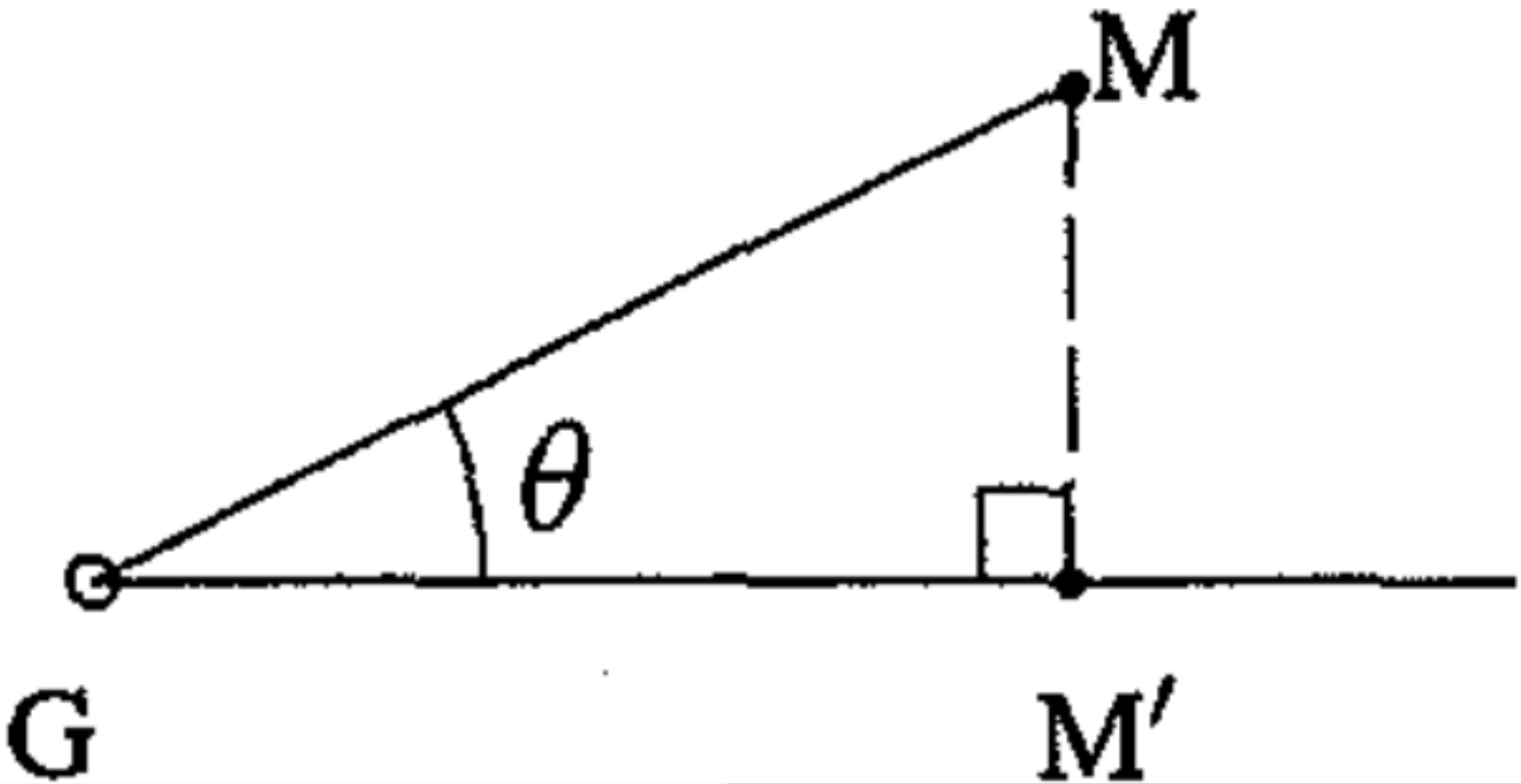
Contribution d'un point à un axe

De la même façon, on peut calculer la contribution d'un point à un axe. Il s'agit alors de la part de la variance de l'axe due à ce point.

$$ctr_{xi} = \frac{(x_i - x_G)^2}{var_x}$$

Qualité de la projection: cosinus carré

Le cosinus carré est une mesure de la qualité de la projection d'un point sur un axe.



Qualité de la projection: cosinus carré

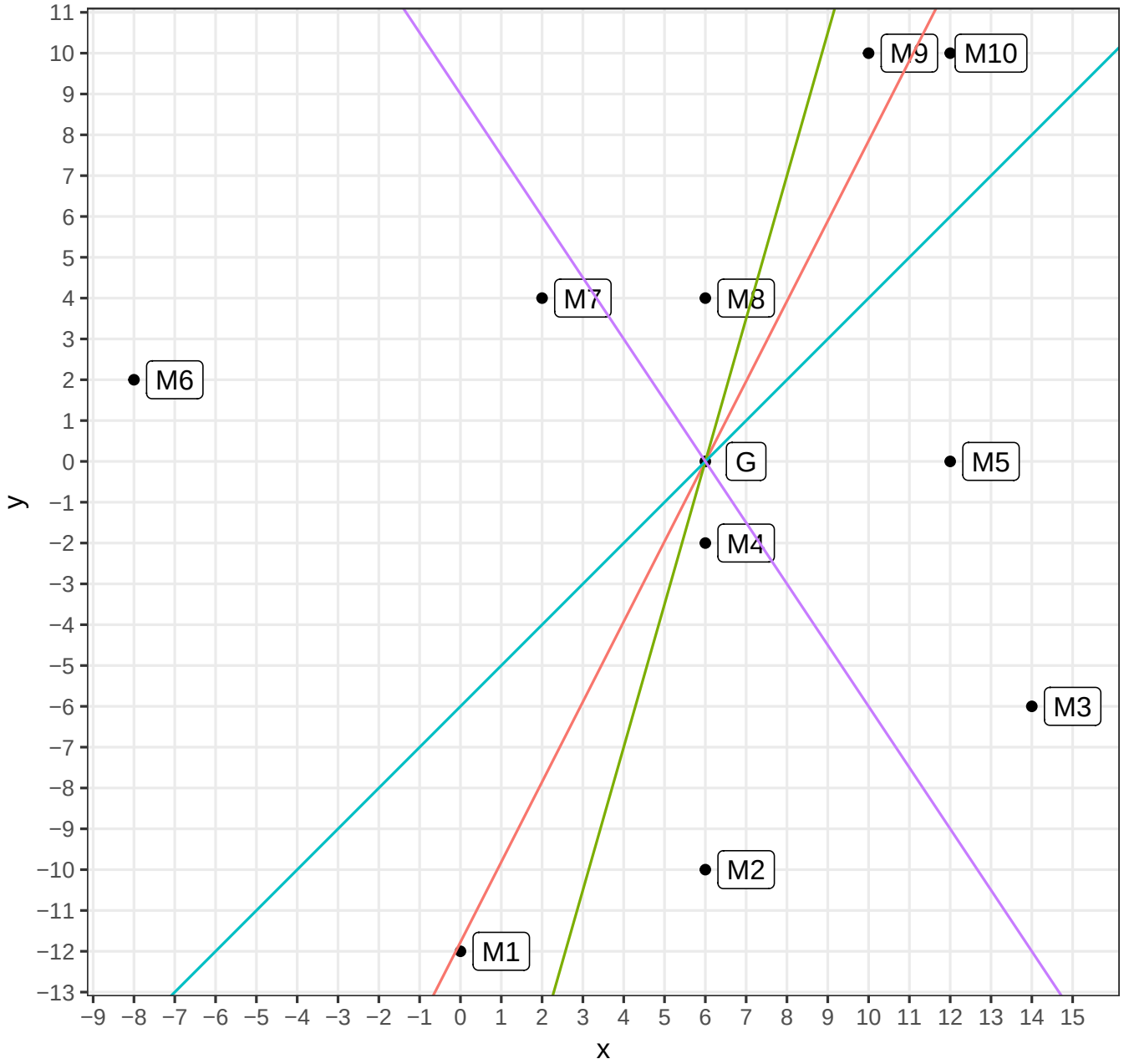
$$cos^2 = \frac{(GM')^2}{(GM)^2}$$

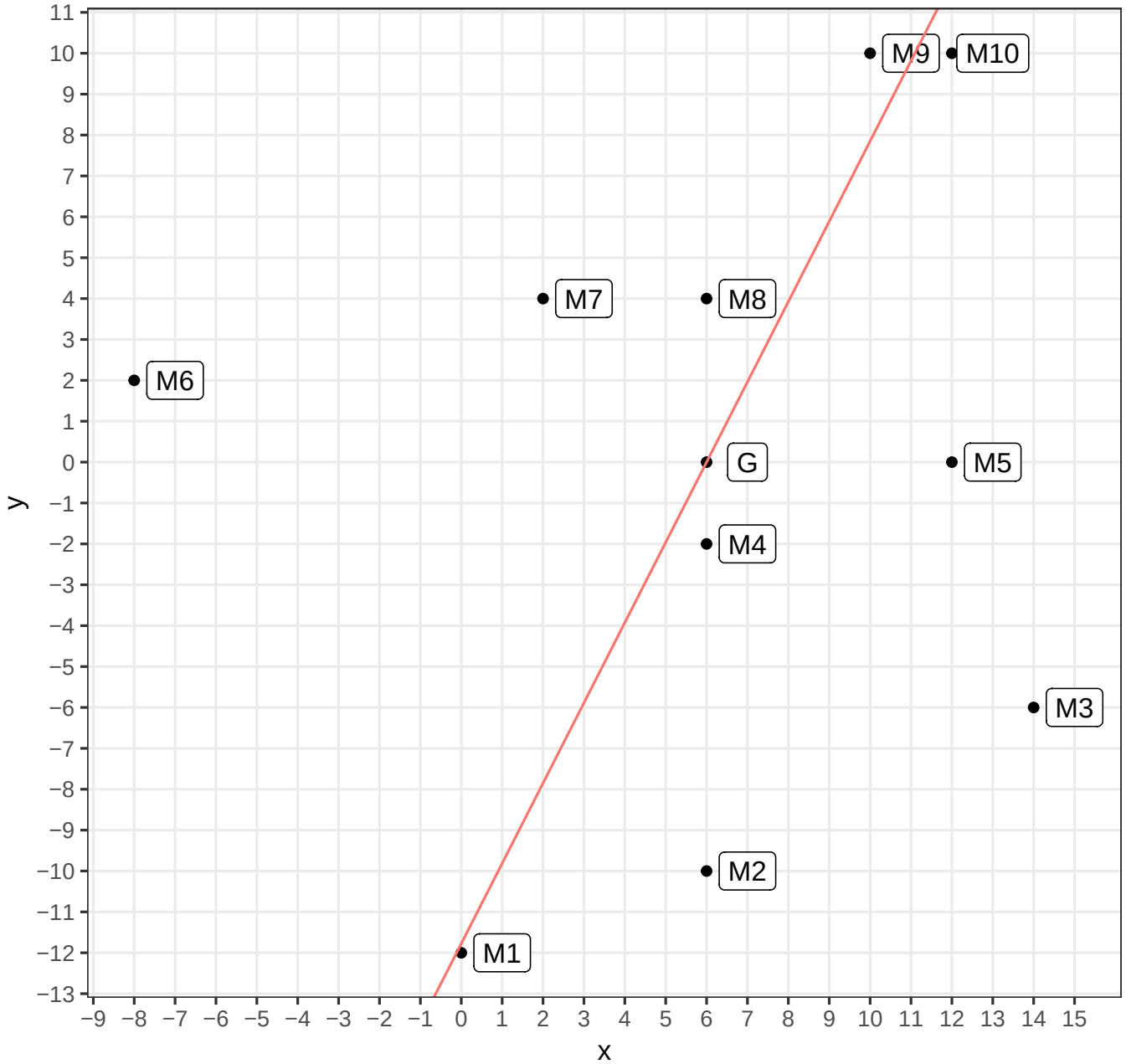
Il vaut 0 si $M' = G$, c'est-à-dire si la projection d'un point sur un axe est le centre de cet axe ; il vaut 1 si $M = M'$, c'est à dire si la projection d'un point sur un axe est le point lui-même. Ainsi, plus le cosinus carré est proche de 1 et meilleure est la représentation sur l'axe.

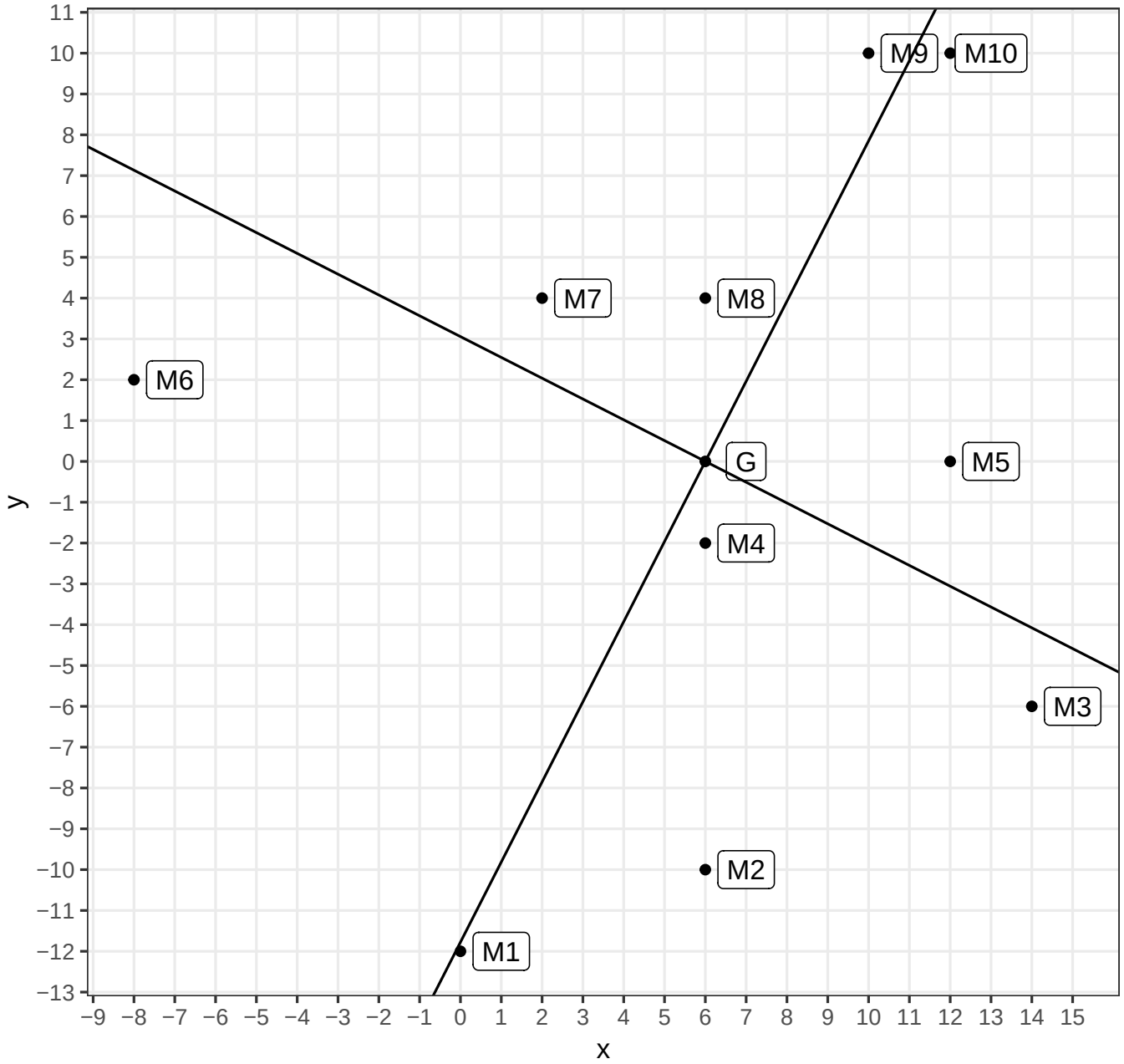
Axes principaux

On appelle axe principal du nuage la droite qui, passant par le centre du nuage, maximise la variance des projections de chaque point du nuage.

Le second axe principal est l'axe orthogonal au premier et passant par le centre du nuage qui maximise à son tour la variance des projections de chaque point.


M9
M10
M7
M8
M6
G
M5
M4
M3
M2
M1
x
y


M9
M10
M7
M8
M6
G
M5
M4
M3
M2
M1
x
y


M9
M10
M7
M8
M6
G
M5
M4
M3
M2
M1
x
y

Eigenvalue

On appelle eigenvalue de l'axe n la variance de la projection sur l'axe n du nuage. On la note λ_n. La somme des $lambda_n$ est égale à la variance totale du nuage

$$\sum_{i=1}^{n} \lambda_n = var$$

On appelle inertie la part de chaque axe dans la variance totale.

$$inertie_i = \frac{\lambda_i}{var}$$

Principe de l'analyse géométrique de données

Ce qui a été exposé est le principe général de l'AGD:

1. on part d'une représentation d'un nuage à n dimensions (dans l'exemple, deux) ;
2. on calcule le centre du nuage ainsi que sa variance ;
3. on recherche, par un algorithme, le premier axe qui maximise la variance des projections, puis le deuxième, ... le nième ;
4. on calcule l'eigenvalue de chaque axe ; et sur chaque axe, la contribution et le cosinus carré de chaque axe (qui seront nécessaires à l'interprétation).

Avec $n > 3$, ce qui est le plus souvent le cas dans l'AGD, il devient impossible ou du moins très difficile de représenter graphiquement le nuage de points, mais la logique géométrique reste inchangée.

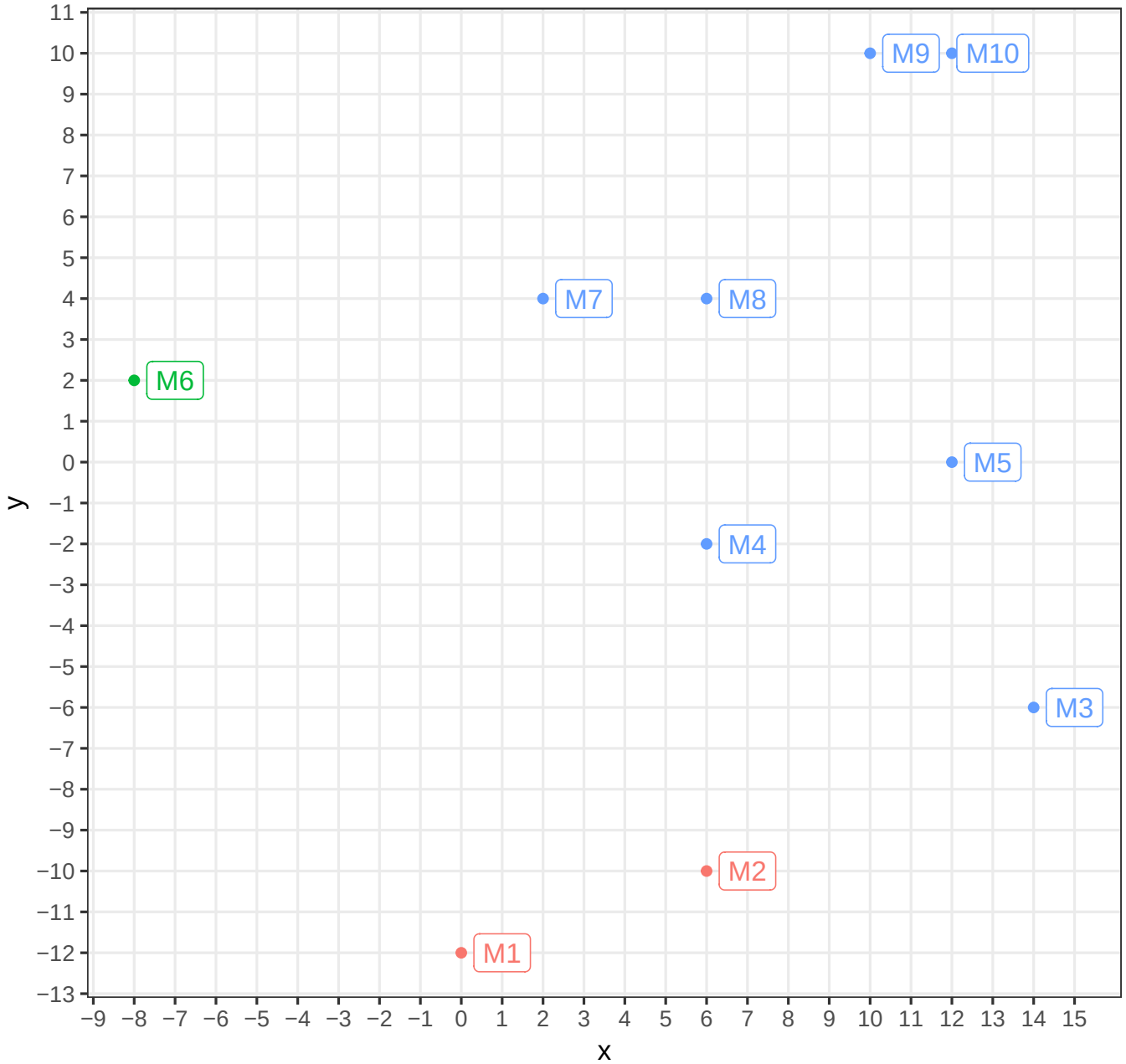
Pondération

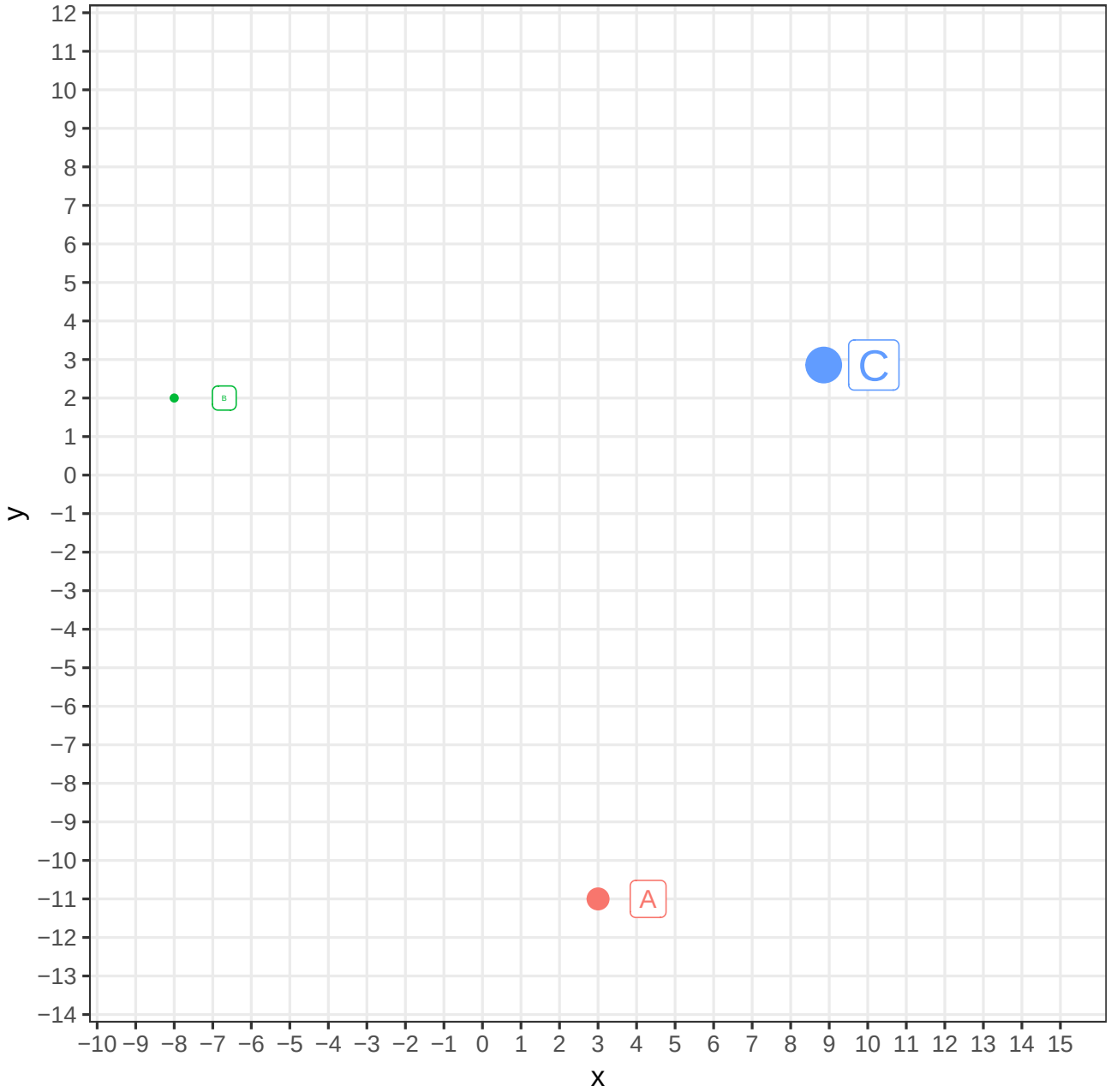
Jusque là, nous n'avons parlé que d'un nuaage dans lequel tous les points ont le même poids. Il est possible de réaliser la même opération sur des nuages de points pondérés. Dans ce cas, chaque point se voit attribuer un poids qui mesure son importance dans le nuage.

Le cas le plus fréquent est celui dans lequel on partitionne un nuage de points de poids égal en différents sous-nuages. On peut alors considérer chaque sous-nuage comme étant lui-même un point unique, dont les coordonnés sont égales à la moyenne des coordonnées des points, et dont le poids est égal au nombre de points inclus dans le sous-nuage.

Partition du nuage

Il est possible de partitionner un nuage de points en sous-ensembles. Chaque sous-ensemble aura pour barycentre le point dont les coordonnées sont la moyenne des coordonnés de ses points (le centre du sous-nuage). Le nuage composé par les barycentres de chaque sous-ensemble, ayant chacun pour poids le nombre de points dans le sous-ensemble, aura les mêmes propriétés que le nuage complet (même centre, même variance)


M9
M10
M7
M8
M6
M5
M4
M3
M2
M1
y
x


y
x
A
B
C

Variance et contribution dans un nuage pondéré

Soit un nuage de n points de poids w et W la somme des poids ($W = \sum_{i=1}^{n} w_i$). Dans ce cas, la variance (du nuage et de la projection sur un axe) et la contribution (idem) sont pondérés. Par exemple, pour un nuage à deux dimensions, la variance est:

$$\text{var} = \sum_{i=1}^{n} ((x_i - x_G)^2 + (y_i - y_G)^2) \times \frac{w_i}{W}$$

et la contribution d'un point

$$ctr_i = \frac{((x_i - x_G)^2 + (y_i - y_G)^2) \times \frac{w_i}{W}}{var}$$

L'analyse en composantes principales

Principe

L'analyse en composantes principales (principal component analysis, ACP) est une technique d'AGD qui permet de chercher les axes principaux d'un nuage de points définis par des variables quantitatives.

Préparation des données

Bien que cela ne soit pas obligatoire, il est fortement conseillé de centrer et réduire les variables avant l'ACP. Cela permet de rendre commensurable des mesures réalisées dans des unités variées. La plupart des implémentations informatique de la technique le font automatiquement.

Centrer et réduire une variable signifie la transformer de sorte à ce que sa moyenne soit 0 et son écart-type 1. On remplace ainsi chaque observation x_i de la variable x par une transformation:

$$\frac{x_i - \bar{x}}{\sigma_x}$$

où $\bar{x}$ est la moyenne de x et σ_x son écart-type.

Distances entre des individus

Soit deux observations e_i et e_j, définies par leurs coordonnées sur K axes correspondant à autant de variables quantitatives centrées et réduites x^k.

La distance carrée entre deux individus est, comme dans le cas général décrit dans la section précédente:

$$d_{ij}^2 = \sum_{k=1}^{K} (x_i^k - x_j^k)^2$$

Variance

La variance du nuage est la somme pondérée des distances carrées à l'origine (G)

$$\sum_{i=1}^{n} \frac{1}{n} \times d_{iG}^2$$

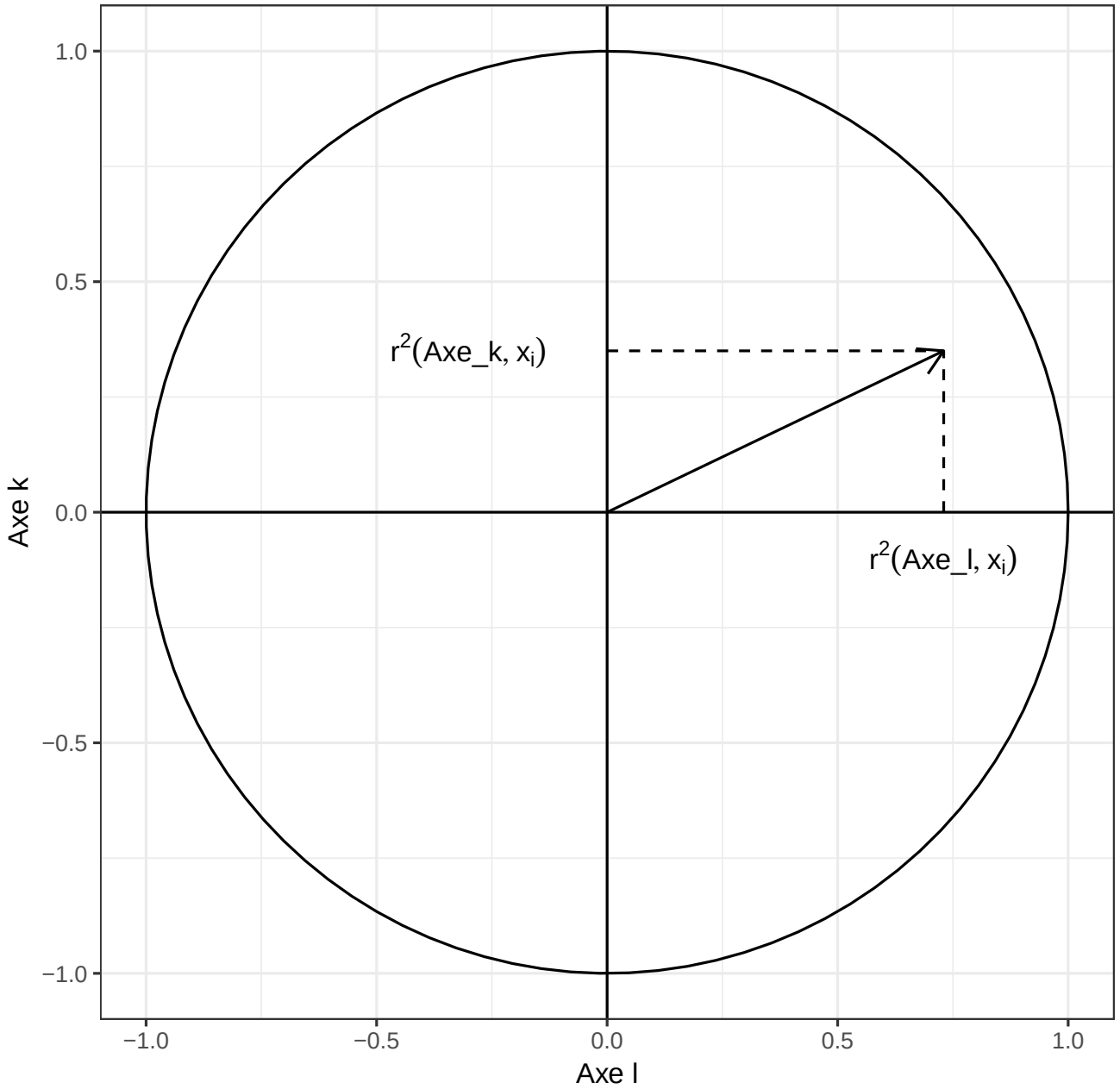
Si les variables sont centrées et réduites, leur variance respective est de 1 et la variance totale du nuage est donc égale au nombre de variables.

On calcule les axes principaux.

Représentation graphique

Des observations: Les coordonnées de chaque observation sur chaque axe principal est le résultat de sa projection sur cet axe. La contribution d'une observation à un axe est la part de la variance totale des projections des observations sur cet axe produit par la projection de cette observation. Le cosinus carré, qui mesure la qualité de la projection, est celui de l'angle entre la droite reliant le centre du nuage et l'observation et celle reliant le centre du nuage et sa projection sur l'axe.

Des variables: Les variables sont représentées sur les axes principaux par leur coefficient de corrélation avec ces axes.


1.0
0.5
0.0
−0.5
−1.0
Axe k
r^2(Axe_k, x_i)
r^2(Axe_l, x_i)
−1.0
−0.5
0.0
0.5
1.0
Axe l

Interprétation

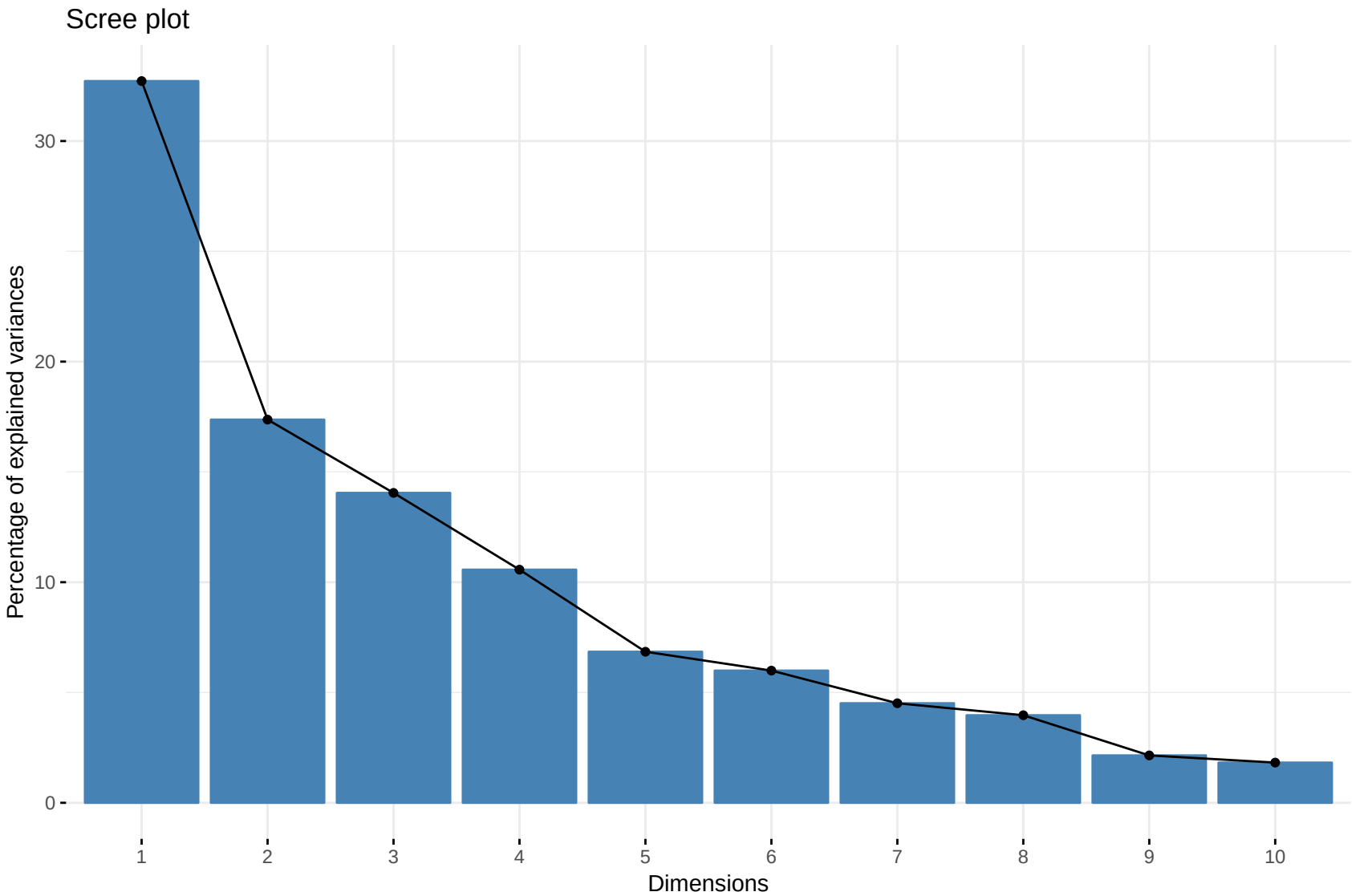
1. Choix du nombre d'axes
2. Interprétation par les coordonnées des variables
3. Interprétation par les observations (selon le jeu de données)
4. Interprétation par les variables supplémentiares

On prend pour exemple un jeu de données mesurant les performances de 41 ahtletes à chacune des dix épreuves de deux compétitions de décathlon.

Examen de l'inertie

On commence par examiner l'inertie des axes, c'est à dire la part de variance qu'ils résument. On peut pour cela les représenter graphiquement.

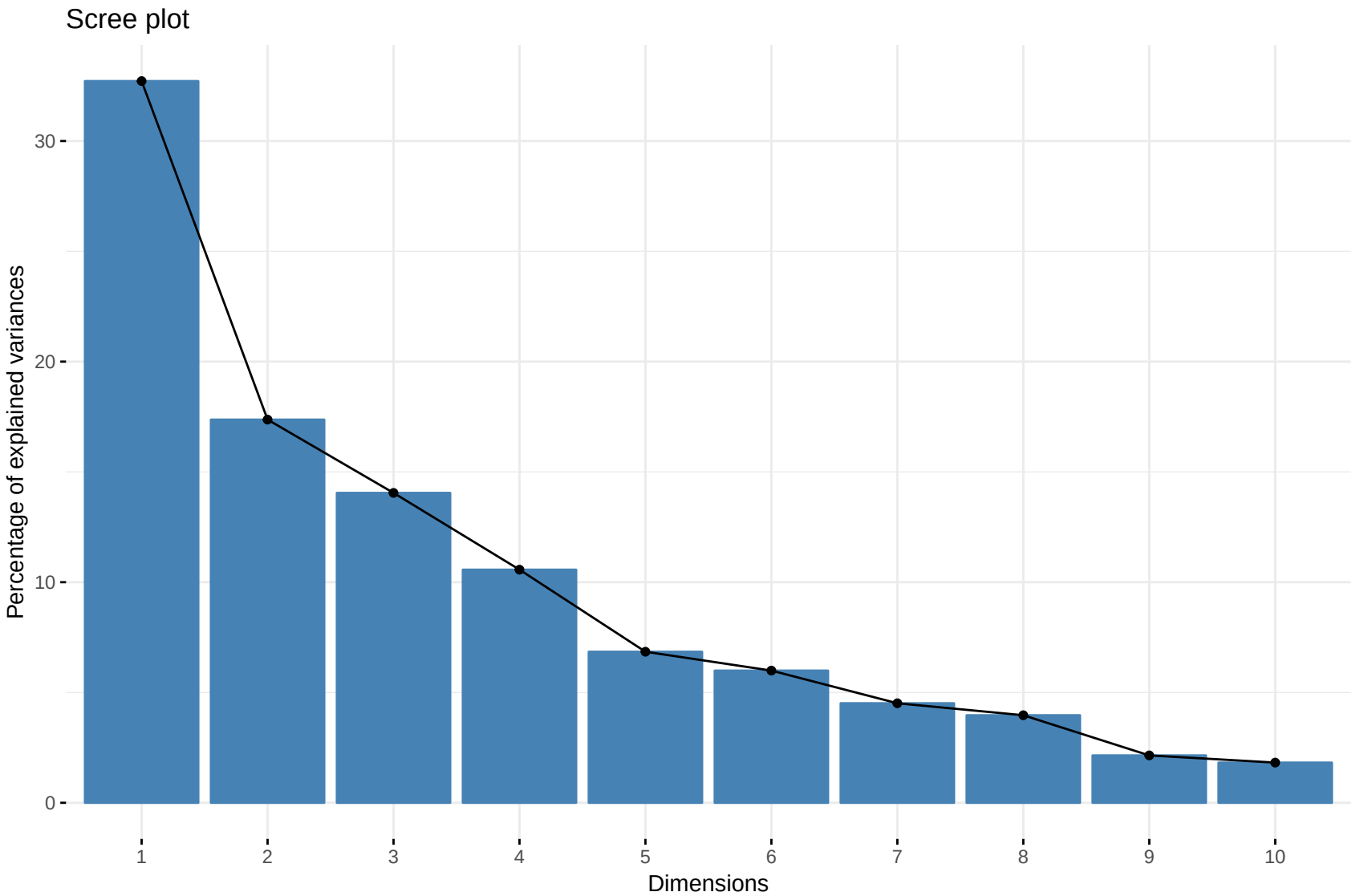
Cet examen a pour objectif de déterminer le nombre d'axe à retenir. On cherche à réduire la dimensionnalité des données, et donc à faire en sorte d'avoir un nombre de méta-variables plus faible que le nombre de variables d'origines. On va donc décider à ce stade de conserver seulement les quelques axes les plus importants.

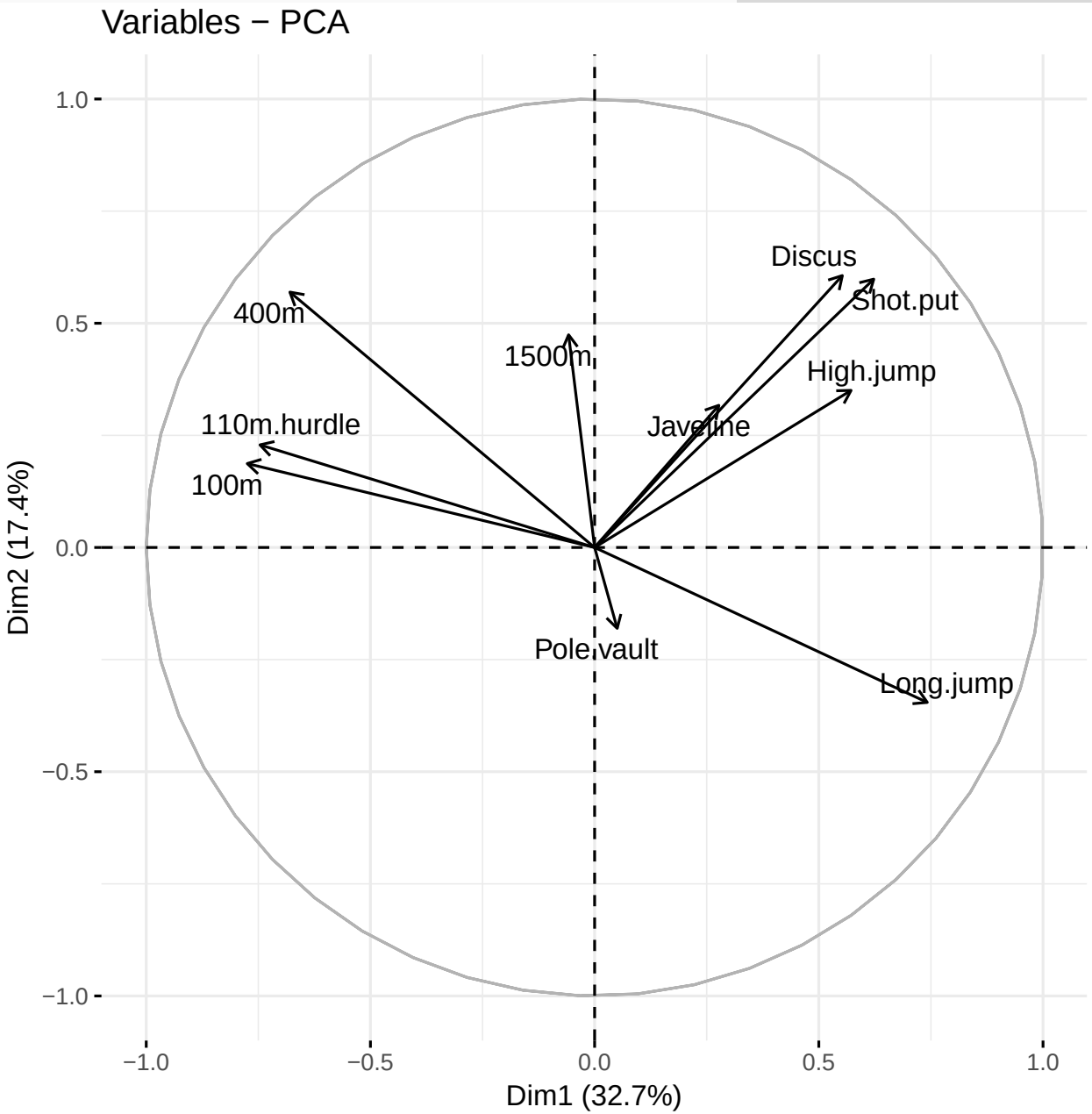

Scree plot
Percentage of explained variances
30
20
10
0
1
2
3
4
5
6
7
8
9
10
Dimensions

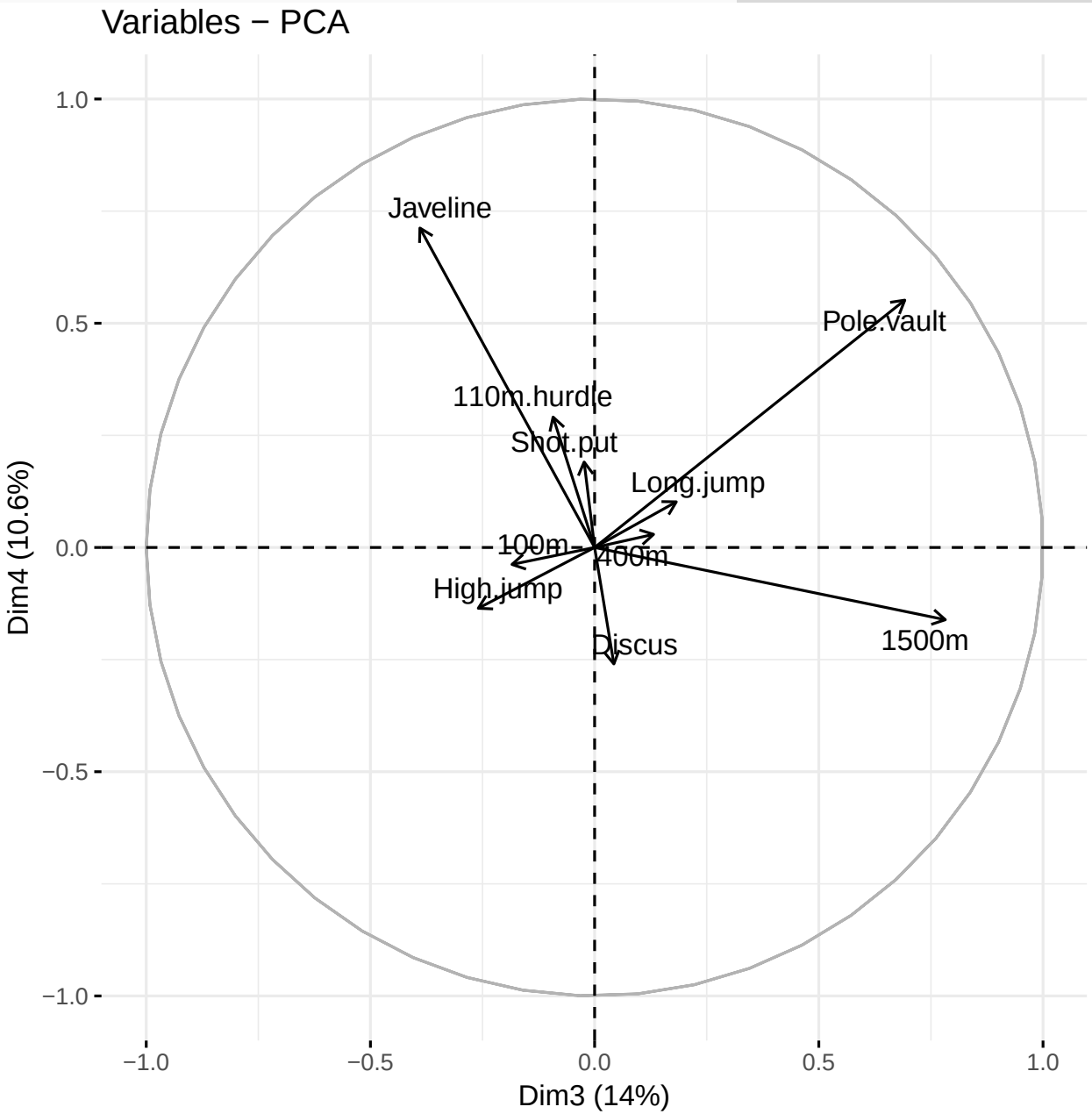
Choix du nombre d'axes à retenir

Une bonne manière de faire est de rechercher les "coudes" dans le diagramme en barre de l'inertie, c'est-à-dire les écarts importants d'inertie entre deux axes successifs. En effet, il serait injustifié de couper après un axe arbitraire si l'axe suivant à une inertie très proches (et donc qu'il est autant représentatif des données).

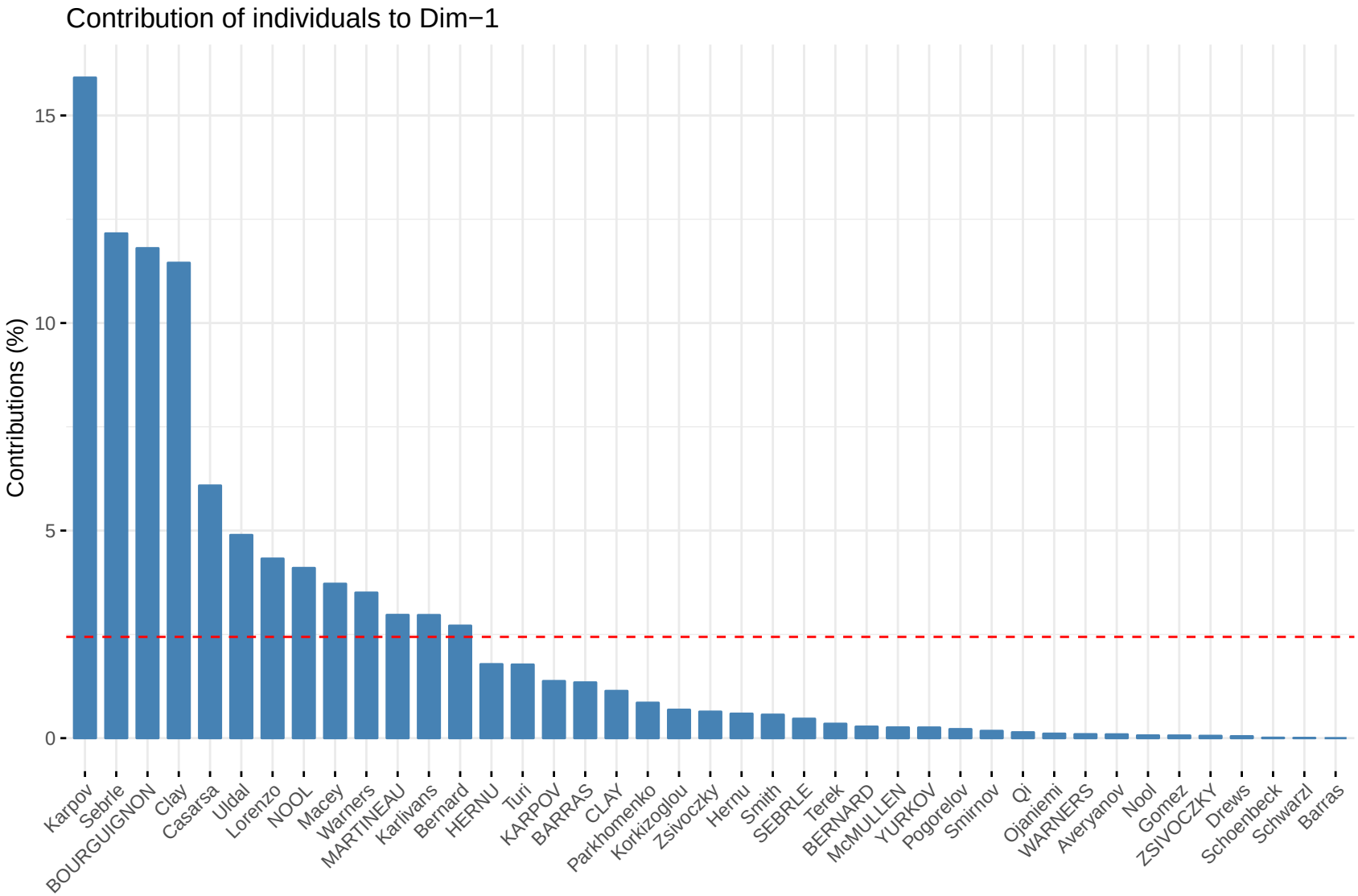
L'autre critère à prendre en compte est le degré de finesse que l'on souhaite conserver: préfère-t-on résumer très fortement, et donc plus caricaturalement, l'information des données originelles, ou veux-t-on plus de précision. Plus on souhaite résumer et moins on choisit d'axes.


Scree plot
Percentage of explained variances
30
20
10
0
1
2
3
4
5
6
7
8
9
10
Dimensions

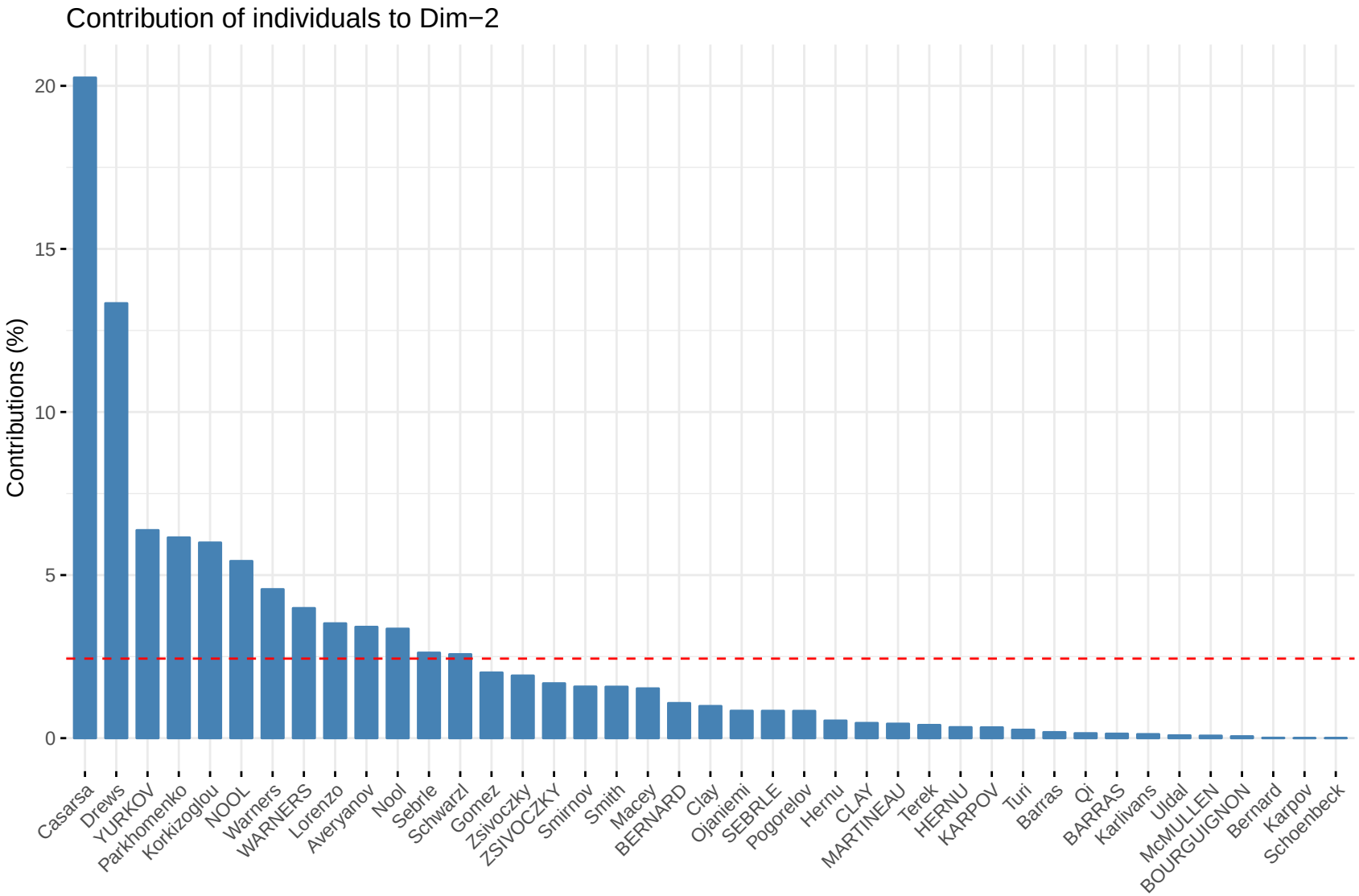

Variables − PCA
Discus
Shot.put
400m
1500m
High.jump
110m.hurdle
Javeline
100m
Pole.vault
Long.jump
1.0
0.5
0.0
−0.5
−1.0
Dim2 (17.4%)
Dim1 (32.7%)


Variables − PCA
Javeline
Pole.vault
110m.hurdle
Shot.put
Long.jump
100m
400m
High.jump
Discus
1500m
1.0
0.5
0.0
−0.5
−1.0
Dim4 (10.6%)
Dim3 (14%)

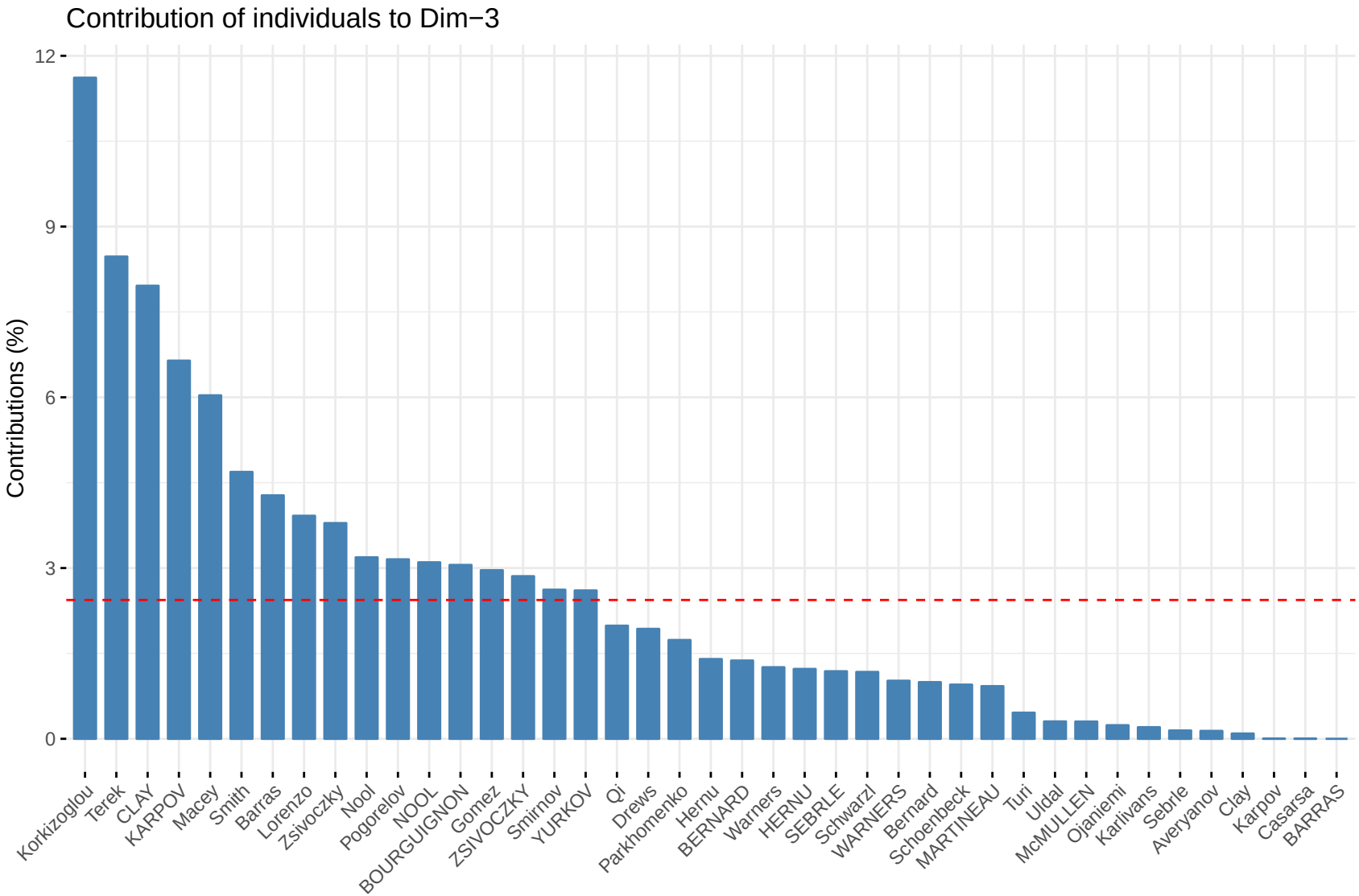
Interprétation des axes: contribution des observations



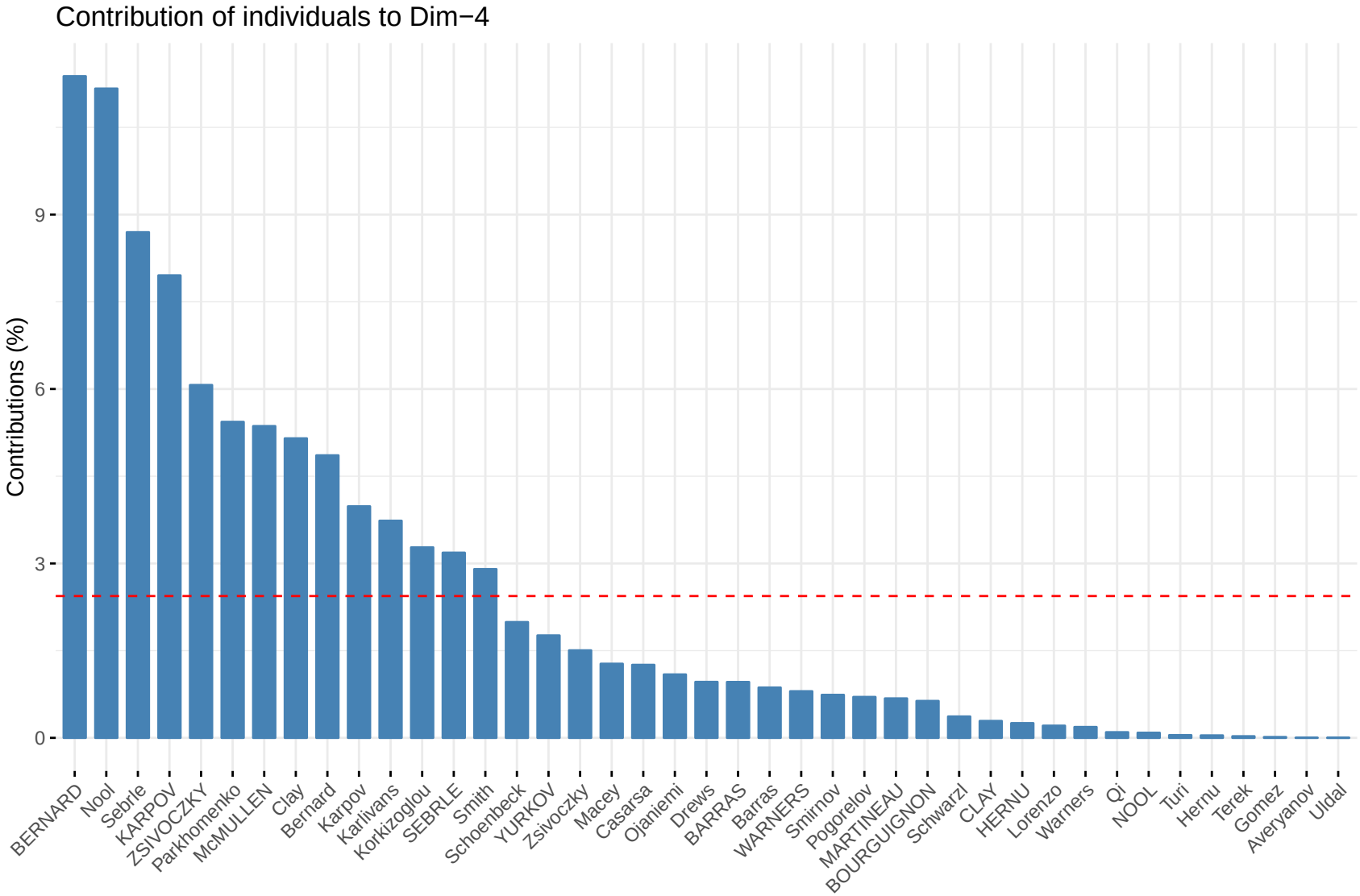
Interprétation des axes: contribution des observations



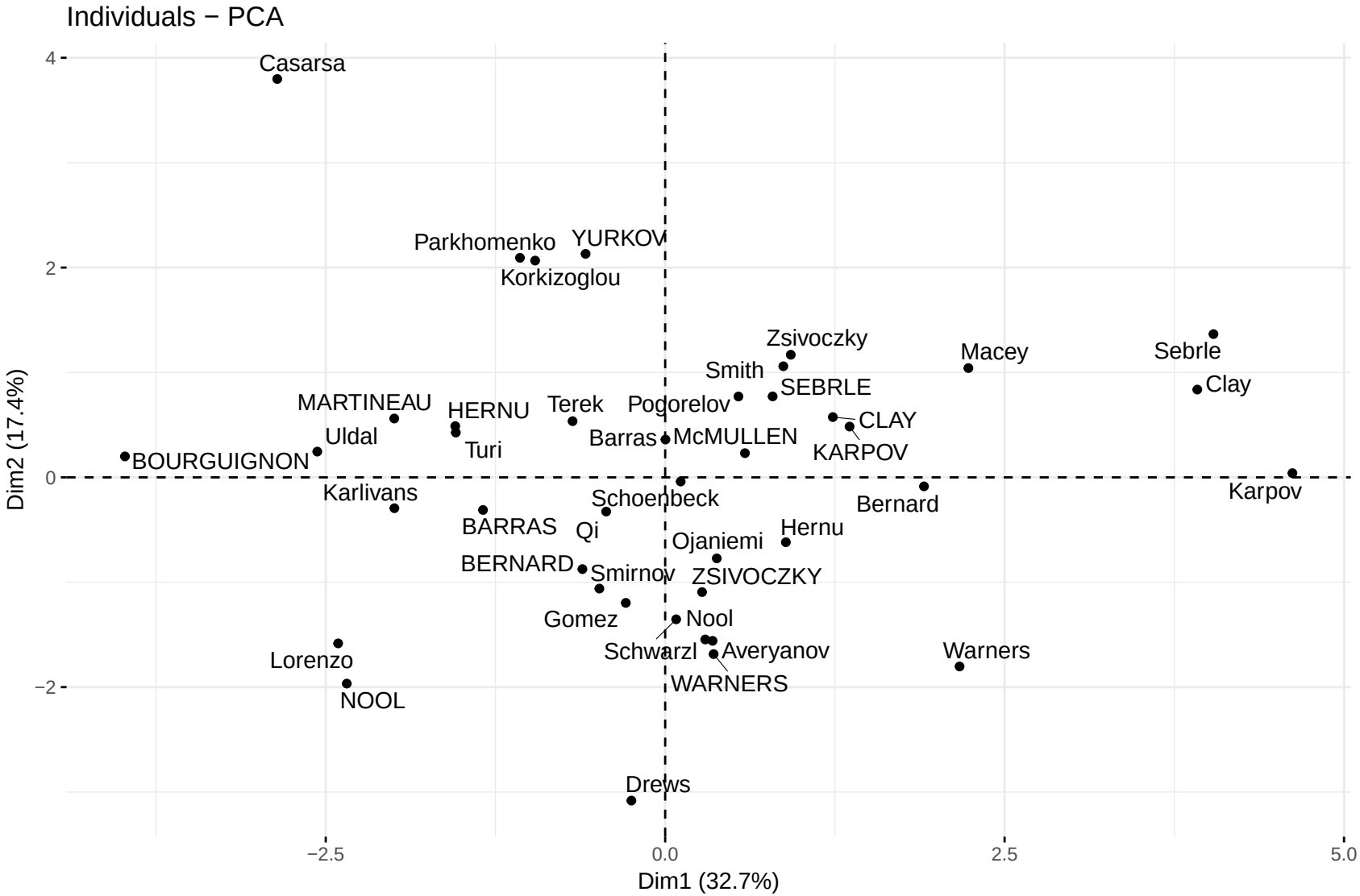
Interprétation des axes: contribution des observations



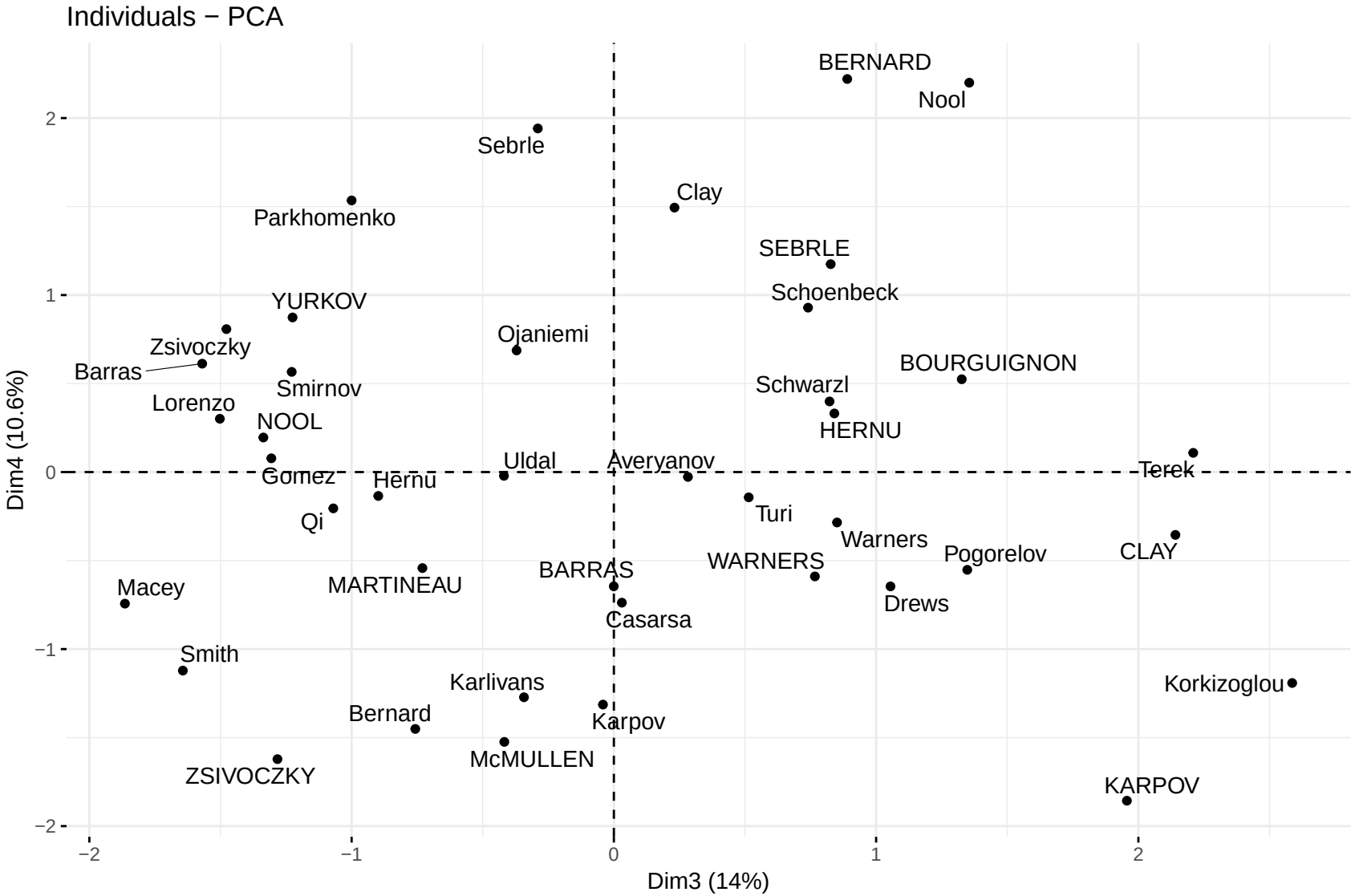
Interprétation des axes: contribution des observations

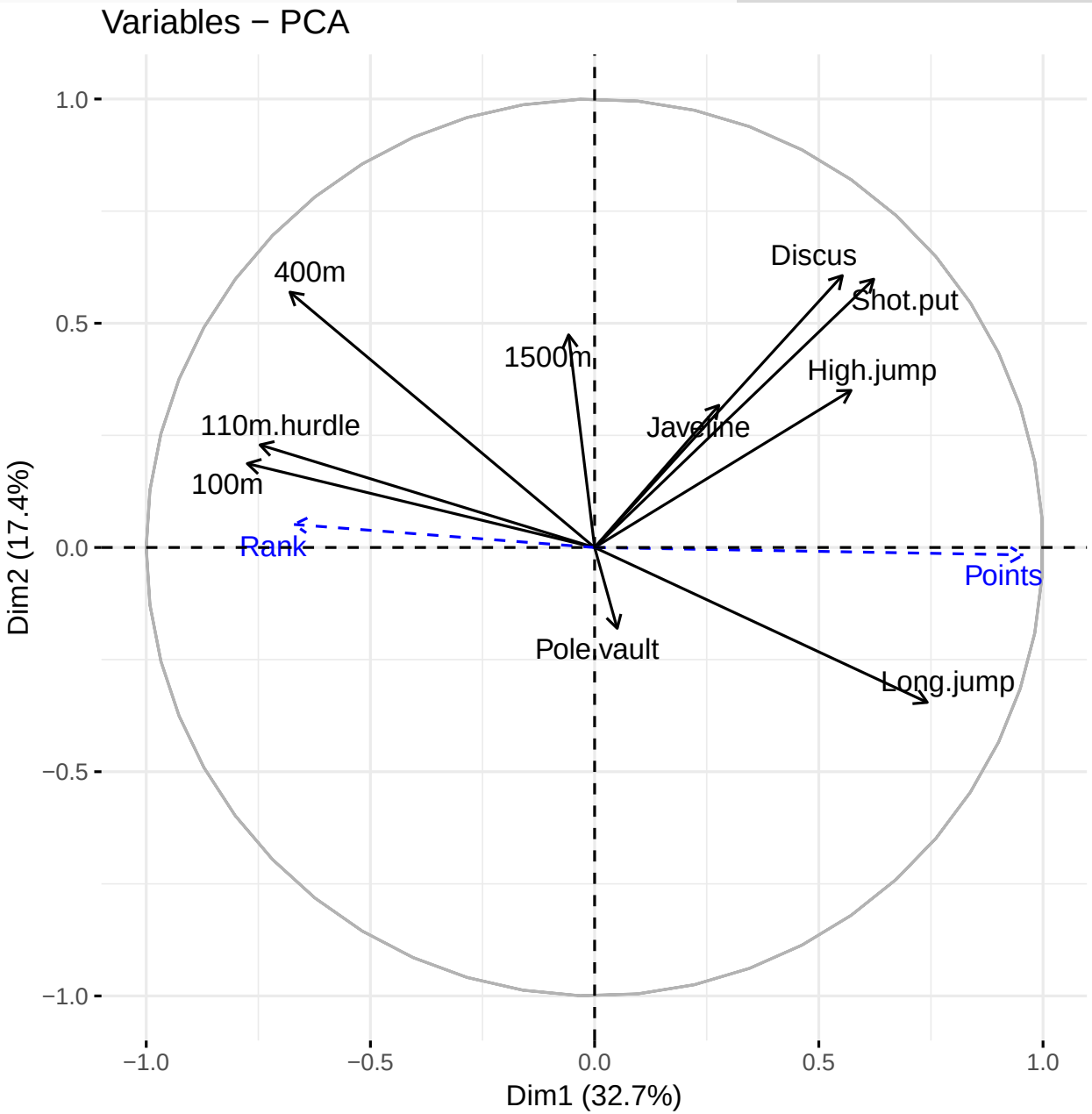


Interprétation des axes: coordonnées des observations



Interprétation des axes: coordonnées des observations




Variables − PCA
1.0
0.5
0.0
−0.5
−1.0
Dim2 (17.4%)
400m
Discus
Shot.put
1500m
High.jump
110m.hurdle
Javeline
100m
Rank
Points
Pole.vault
Long.jump
−1.0
−0.5
0.0
0.5
1.0
Dim1 (32.7%)

Théorie de l'analyse des correspondances multiples

Principe

L'analyse des correspondances multiples est une technique d'AGD qui permet de résumer un ensemble de variables catégorielles.

Soit une base de données comprenant Q variables; les modalités de la variable q sont désignées par K_q; et K désigne l'ensemble des modalités (des catégories), soit la somme des K_q possibles.

Dans l'ACM, on construit un espace de $K - Q$ dimensions. En effet, une variable à n modalités peut être exprimée graphiquement par $n - 1$ axes.

Principe

Dans cet espace, on projette deux nuages de points:

- le nuage des individus: chaque point est une observation de la base de données originale, de poids 1;
- le nuage des modalités: chaque point est une modalité, dont les coordonnées sont égales à la moyenne des coordonnées des observations qui prennent cette modalité (le barycentre du sous-nuage des individus) et dont le poids est égal au nombre d'observations dans cette catégorie.

Ces deux nuages sont équivalents (même variance, même poids relatifs, même axes principaux). On cherche alors les axes principaux qui résument le mieux le nuage.

Principe

Enfin, vient la phase d'interprétation des résultats. On décide alors du nombre d'axes qu'ils convient de commenter à l'aide de l'inertie. On nomme alors la structure latente décrite par les axes (les méta-variables) à l'aide des coordonnées, des contributions et des cosinus carrés.

Distance

La distance carrée entre deux individus i et i' relative à une question q est la somme de l'inverse des fréquences des modalités respectives.

$$d_q^2(i,i') = \frac{1}{f_k} + \frac{1}{f_{k'}}$$

Elle vaut évidemment 0 si les deux individus ont la même modalité.

La distance carrée totale entre deux individus est la somme pondérée de la distance carrée sur chaque question.

$$d^2(i,i') = \frac{1}{Q} \sum_{q \in Q} d_q^2(i,i')$$

Variance du nuage des individus

La distance d'un point M_i au centre du nuage G est:

$$(GM^i)^2 = (\frac{1}{Q} \sum_{k \in K_i} \frac{1}{f_k}) - 1$$

La variance du nuage est la somme pondérée des carrés des distances:

$$var = \frac{1}{n} \times \sum_{i=1}^{n} (GM^i)^2 = \frac{K}{Q} - 1$$

Variance du nuage des catégories

La distance carrée entre deux catégories k et k' dépend de l'effectif de chacune, respectivement n_k et $n_{k'}$ et de l'effectif des individus présents dans les deux catégories, $n_{kk'}$.

$$M^k M^{k'} = \frac{n_k + n_{k'} - 2n_{kk'}}{\frac{n_k \times n_{k'}}{n}}$$

Elle est d'autant plus faible que les deux catégories sont souvent choisies en même temps (elle vaut 0 si $n_k + n_{k'} = 2n_{kk'}$, c'est-à-dire si toutes les observations de catégorie k sont également de catégorie k' et vice-versa). Elle est d'autant plus grande qu'il y a peu d'observations ayant les deux catégories entre les deux catégories et que ces catégories ont des effectifs importants.

Variance du nuage des catégories

La distance carré du point de la catégorie k du centre du nuage est:

$$(GM^k)^2 = \frac{1}{f_k} - 1$$

$=>$ plus une catégorie est peu choisie et plus elle est loin du centre du nuage

La variance du nuage des catégories est la même que la variance du nuage des observations, $\frac{K}{Q} - 1$.

Contribution des catégories

Le point moyen des modalités d'une variable est G.

La contribution d'une catégorie au nuage de point dépend de sa fréquence ; la contribution d'une question de son nombre de modalités

$$ctr_k = \frac{1 - f_k}{K - Q}$$

$$ctr_q = \frac{K_q - 1}{K - Q}$$

Inertie

L'eigenvalue λ_l de l'axe principal l, est égale à la variance de la projection du nuage sur cet axe. La somme de toutes les eigenvalue est égale à la variance totale du nuage.

Inertie modifiée

On présente le plus souvent non pas l'eigenvalue brute, mais le taux de variaux τ_l, c'est à dire la part de la variance totale capturée par l'axe principal l. $\tau_l = \frac{\lambda_l}{var}$.

Pour les nuages de grandes dimensionnalité, il est normal que ce taux soit relativement faible (de l'ordre de 0.1 pour le premier axe par exemple). C'est un effet de la construction de la méthode.

Benzecri a donc proposé d'utiliser le taux modifié.

Inertie modifiée

Le taux modifié s'appuie sur la pseudo-eigenvalue λ'_l.

$$\lambda'_l = (\frac{Q}{Q-1})^2 \times (\lambda_l - \bar{\lambda})^2$$

où $\bar{\lambda}$ désigne la moyenne des eigenvalue des axes.

On peut alors obtenir un taux d'inertie modifié:

$$\tau'_l = \frac{\lambda'_l}{\sum_{l=i}^{l_{max}} \lambda'_l}$$

Ce taux peut faciliter la comparaison entre modèles (il pondère l'inertie par la dimensionnalité pour ne pas sanctionner les nuages les plus complexes), mais son usage n'est pas consensuel.

Elements illustratifs

Il est possible de projeter des individus et des variables supplémentaires.

- une observation supplémentaire est une observation qui ne participe pas à la construction des axes;
- une variable supplémentaire est une variable qui ne participe pas au calcul de la distance entre deux individus.

Elements illustratifs

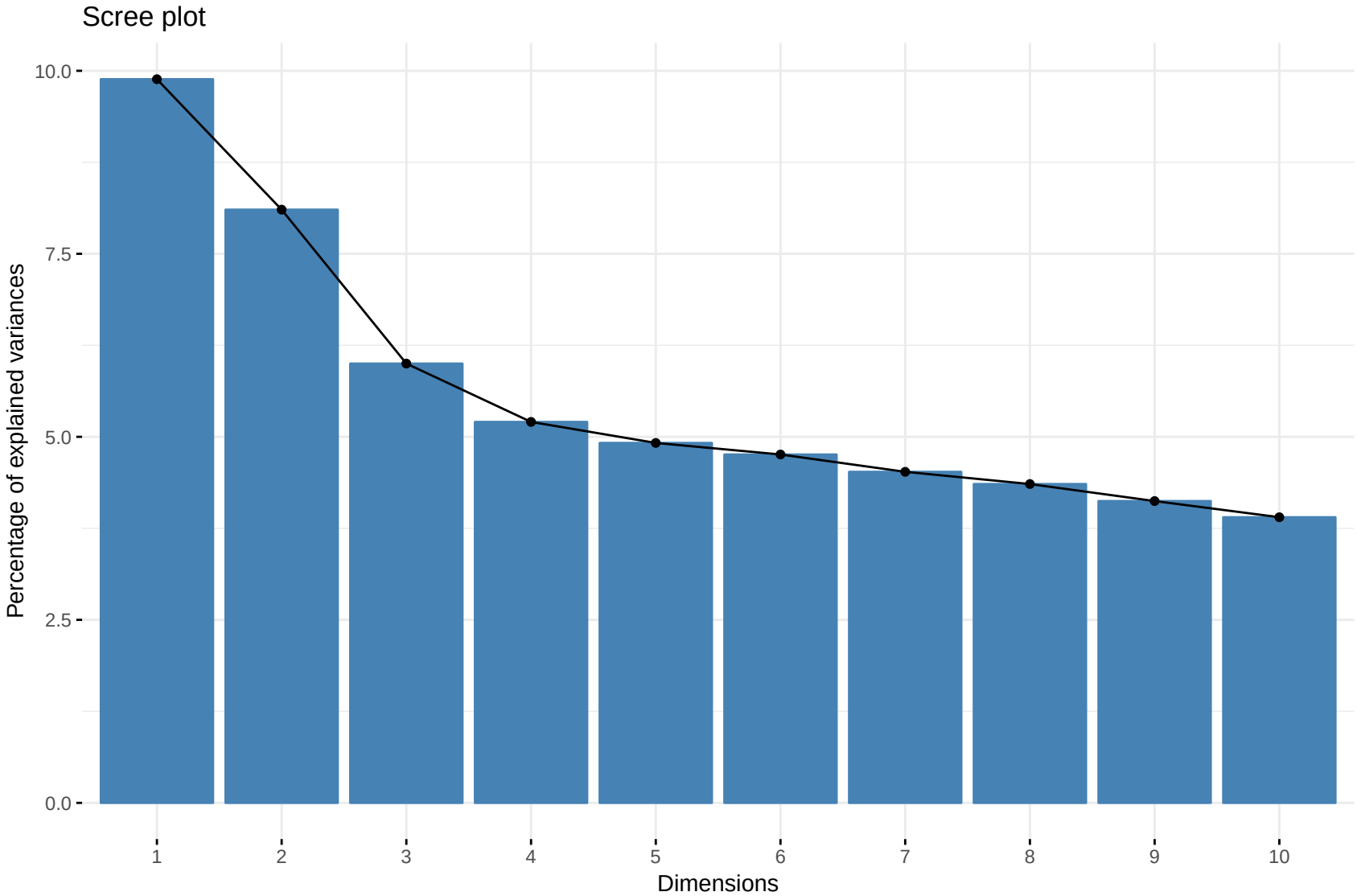
On calcule les coordonnées sur les axes principaux des observations supplémentaires à partir des modalités de variables actives (elles seront proches des observations dont les modalités sont proches) ; et les coordonnées des modalités supplémentaires sont les barycentres des sous-nuage d'observations dans cette modalité. Les individus et modalités supplémentaires ont des coordonénes et des cosinus carrés qui permettent de mesurer la qualité de leur représentation ; mais la notion de contribution ne fait pas sens pour eux, car ils ne participent pas à la variance du nuage.

Construction et interprétation de l'ACM

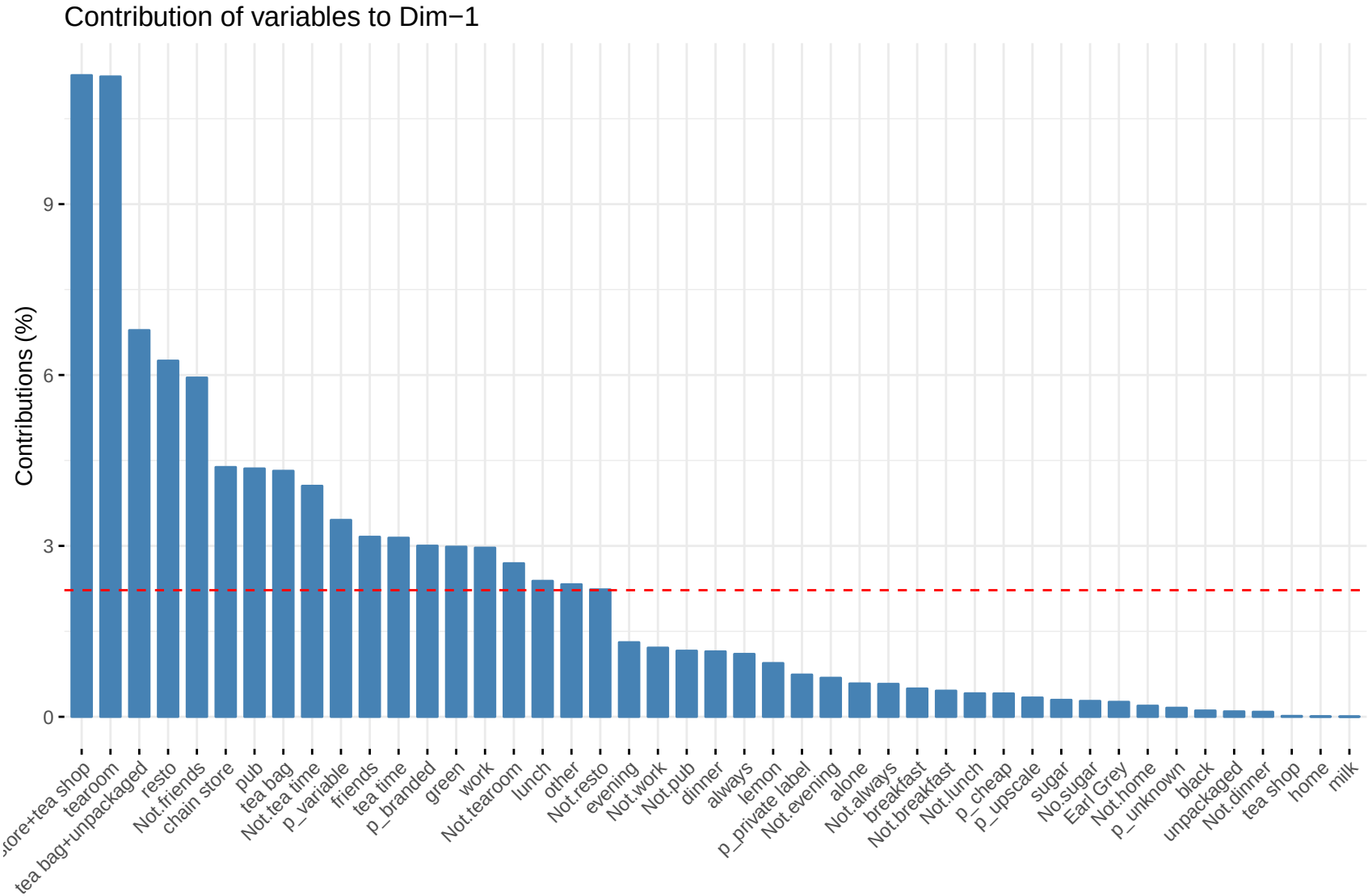
Interprétation de l'ACM

Le jeu de données d'exemple rassemble le résultat d'un questionnaire portant sur les pratiques de boisson du thé.

Choix du nombre d'axes



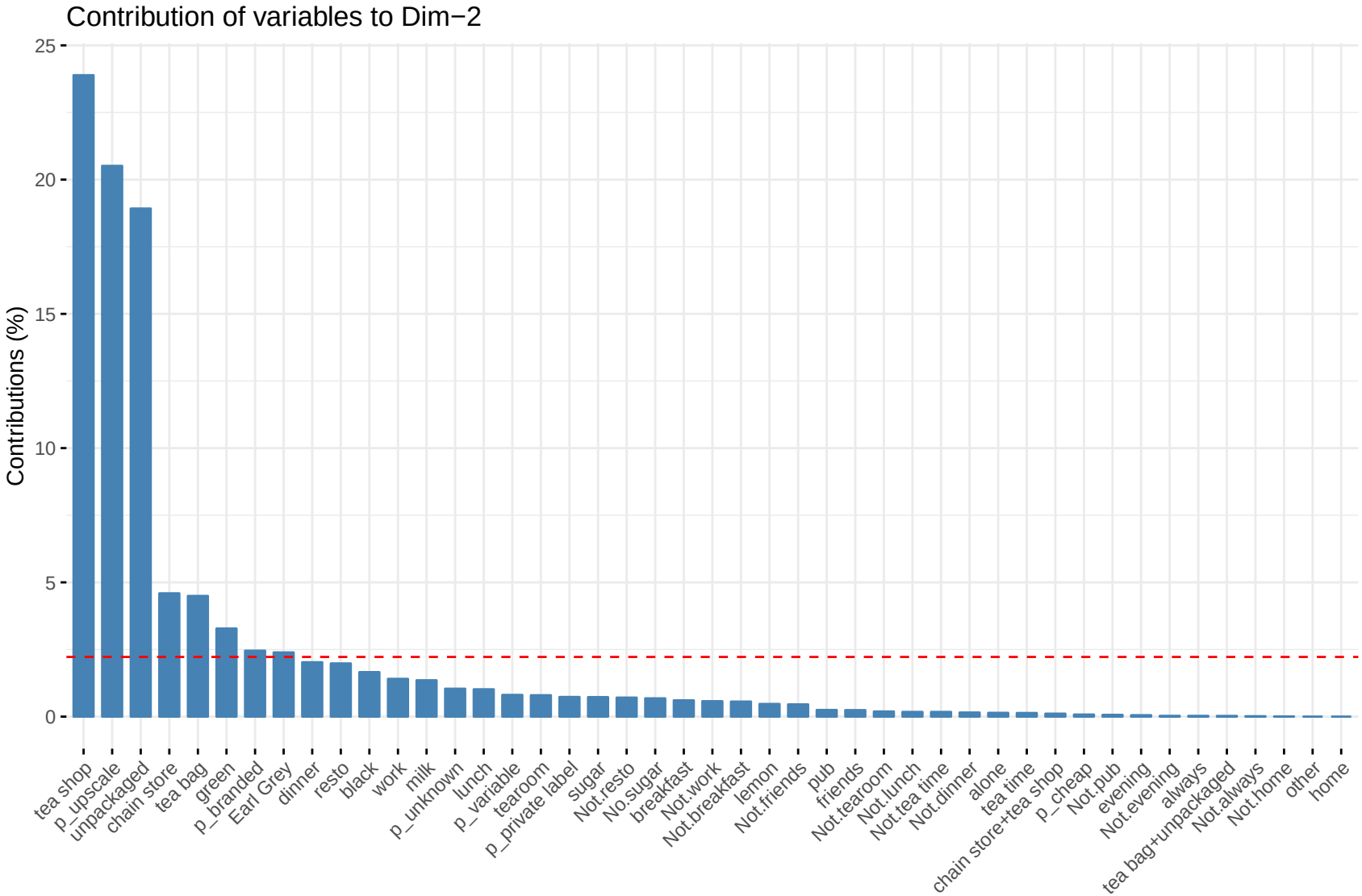
Interprétation de l'axe 1



Interprétation de l'axe 1

Mod	Dim 1 coord	Dim 1 contrib	Dim 1 cos2
tearoom	1.2456659	11.240125	0.3718911
chain store+tea shop	1.0752464	11.262911	0.4062166
resto	0.7959497	6.250839	0.2264676
tea bag+unpackaged	0.7603104	6.786562	0.2637804
pub	0.7442050	4.357785	0.1472236
p_variable	0.4969407	3.454357	0.1471192
friends	0.3592278	3.158901	0.2431995
Not.tearoom	-0.2985480	2.693914	0.3718911
chain store	-0.4275216	4.382858	0.3249328
tea bag	-0.4509027	4.316727	0.2658711
Not.tea time	-0.4977386	4.053348	0.1920380
p_branded	-0.5030448	3.002458	0.1172690
Not.friends	-0.6770063	5.953314	0.2431995
green	-0.8509626	2.984523	0.0895001

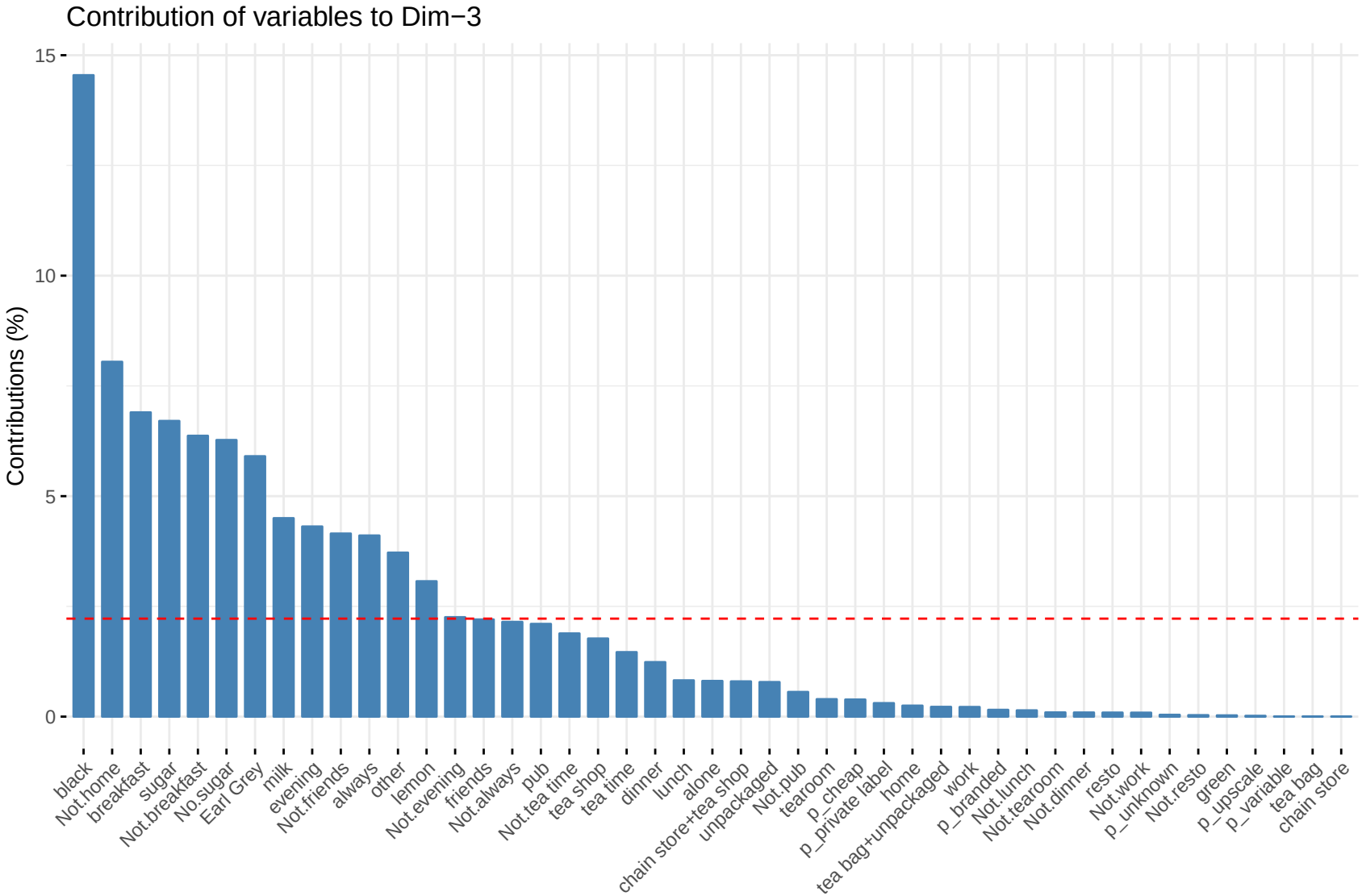
Interprétation de l'axe 2



Interprétation de l'axe 2

Mod	Dim 2 coord	Dim 2 contrib	Dim 2 cos2
tea shop	2.2861316	23.8883796	0.5807109
unpackaged	1.8574192	18.9227980	0.4704554
p_upscale	1.5937593	20.5108815	0.5450350
green	0.8079160	3.2817790	0.0806743
dinner	0.7955838	2.0251355	0.0476417
black	0.3833276	1.6566659	0.0481131
tearoom	0.3001838	0.7962796	0.0215967
Earl Grey	-0.2851164	2.3903682	0.1466284
work	-0.3257533	1.4065652	0.0433428
milk	-0.3754510	1.3530380	0.0374713
chain store	-0.3962053	4.5920307	0.2790731
resto	-0.4056790	1.9808661	0.0588301
p_branded	-0.4118665	2.4552718	0.0786109
tea bag	-0.4165666	4.4944945	0.2269209

Interprétation de l'axe 3



Interprétation de l'axe 3

Mod	Dim 3 coord	Dim 3 contrib	Dim 3 cos2
Not.home	2.0846272	8.046686	0.1344022
lemon	0.6724012	3.069648	0.0558804
sugar	0.4740954	6.705284	0.2102654
evening	0.4510684	4.311612	0.1063790
Not.breakfast	0.4454728	6.369196	0.2149831
always	0.4402403	4.107092	0.1013329
Earl Grey	0.3856250	5.904803	0.2682279
Not.evening	-0.2358378	2.254294	0.1063790
Not.friends	-0.4404704	4.151304	0.1029463
No.sugar	-0.4435086	6.272685	0.2102654
breakfast	-0.4825955	6.899962	0.2149831
milk	-0.5892367	4.500262	0.0922936
black	-0.9774692	14.546402	0.3128452
other	-1.4171801	3.718864	0.0621154

Diagramme des variables, axes 1 et 2

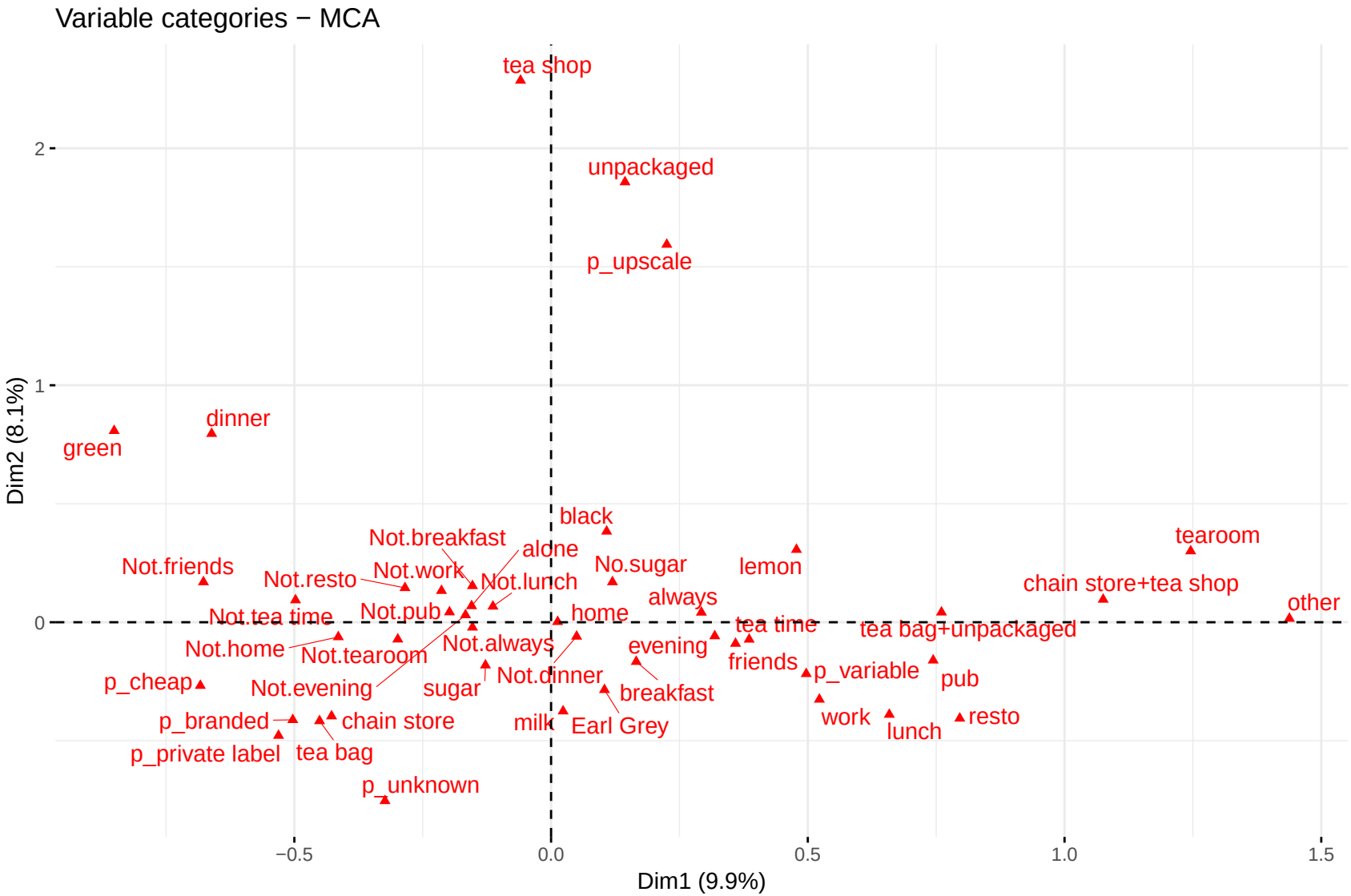


Diagramme des variables, axes 1 et 3

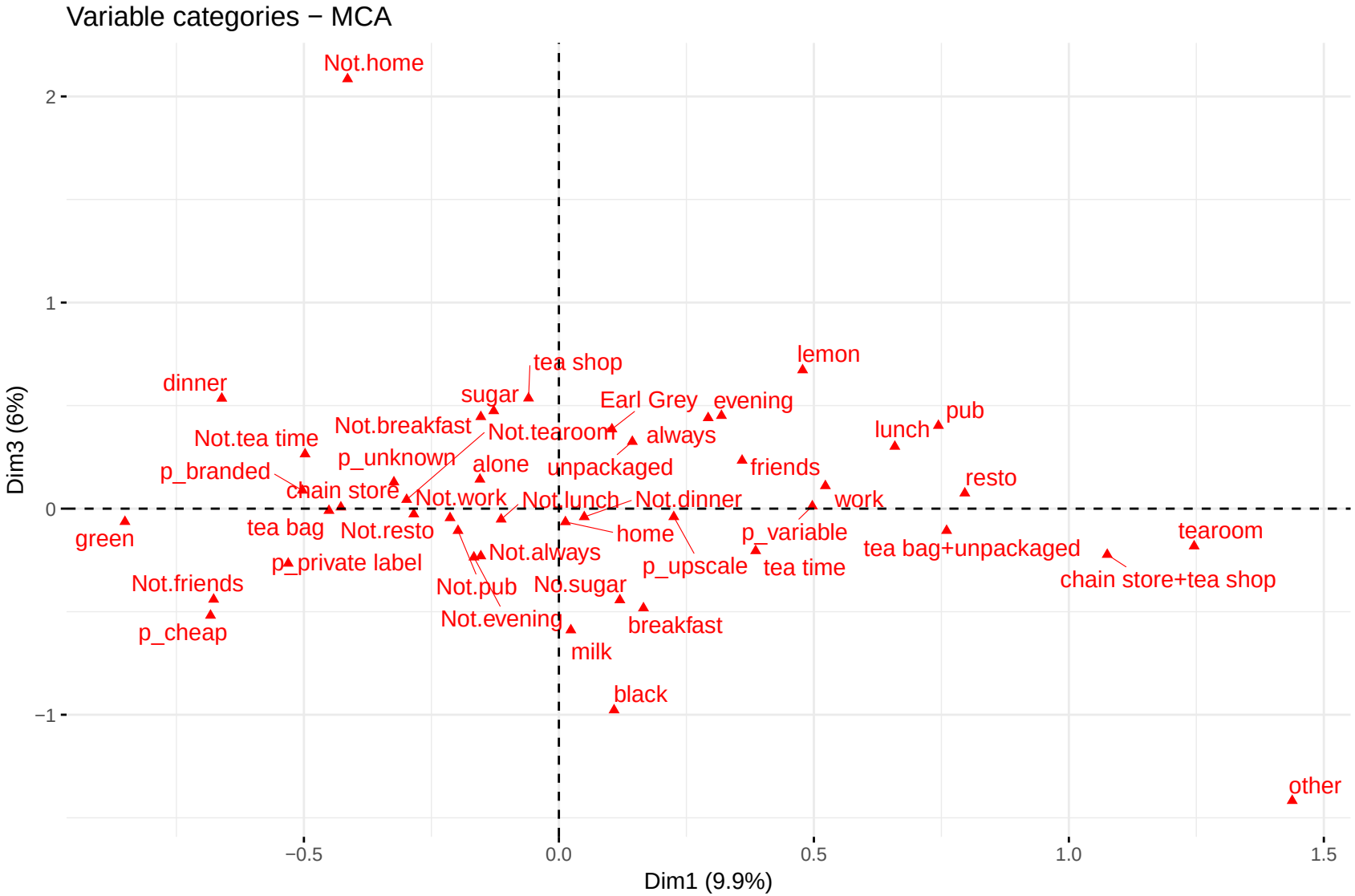


Diagramme des observations

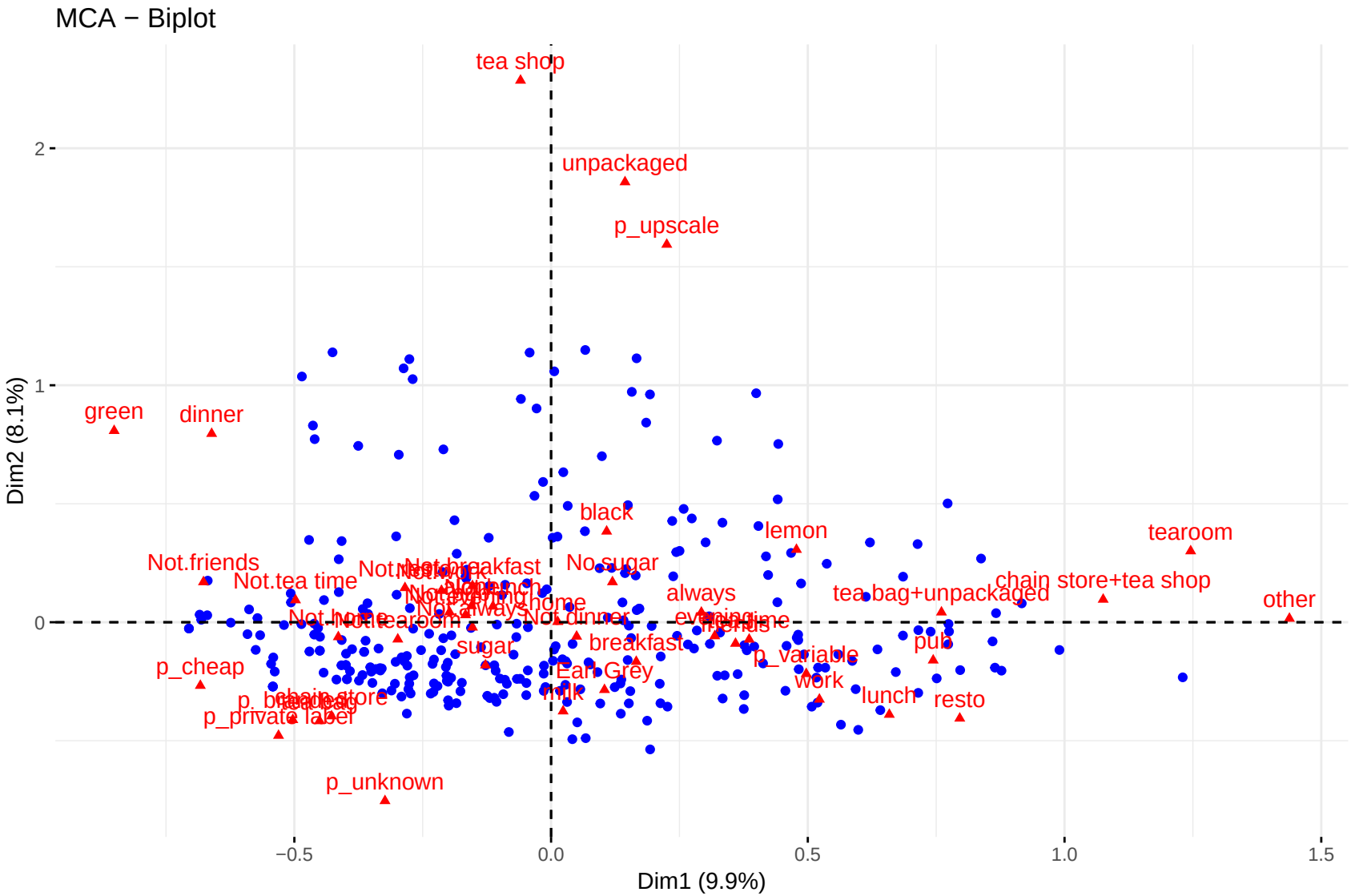
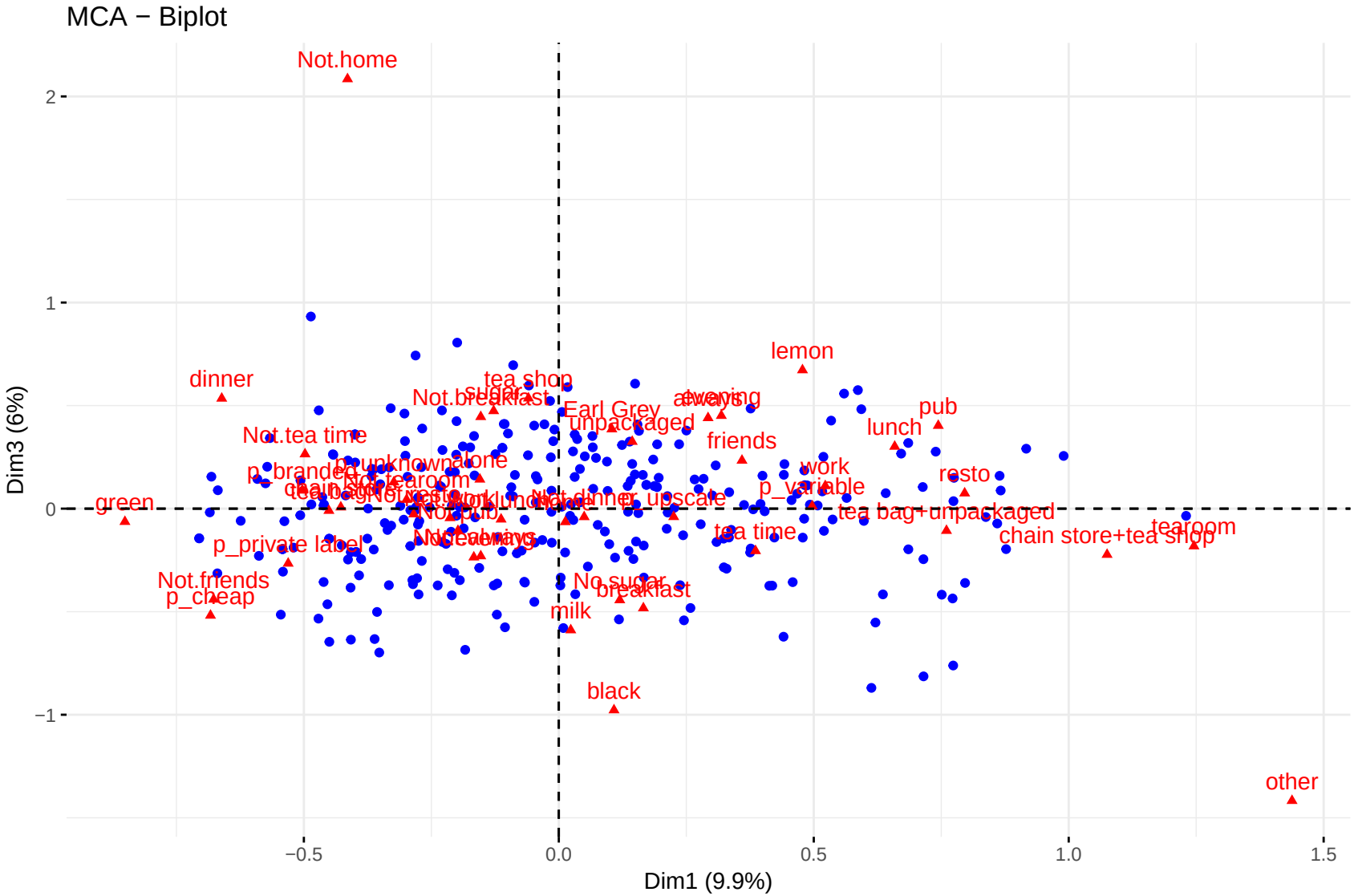
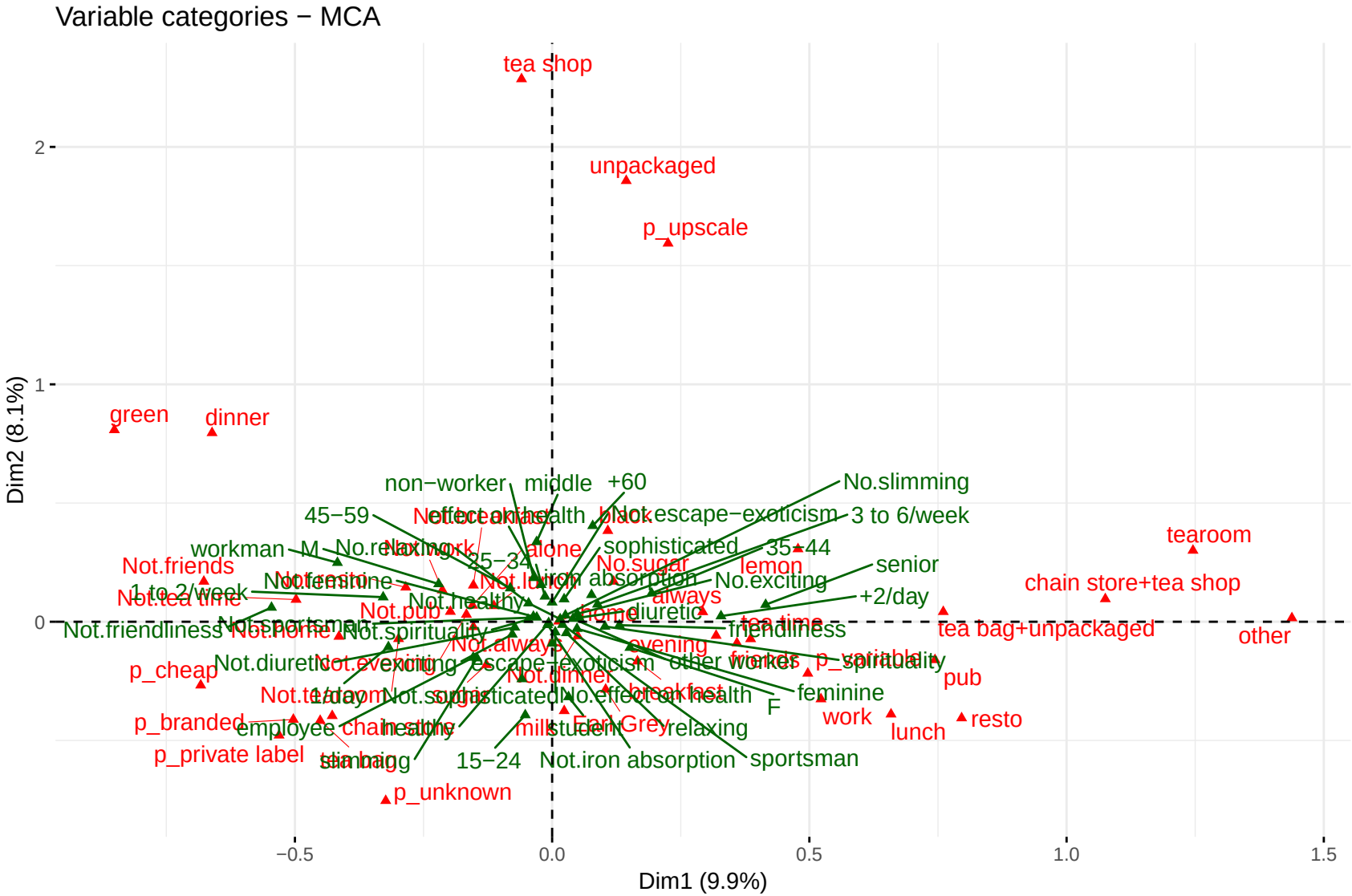


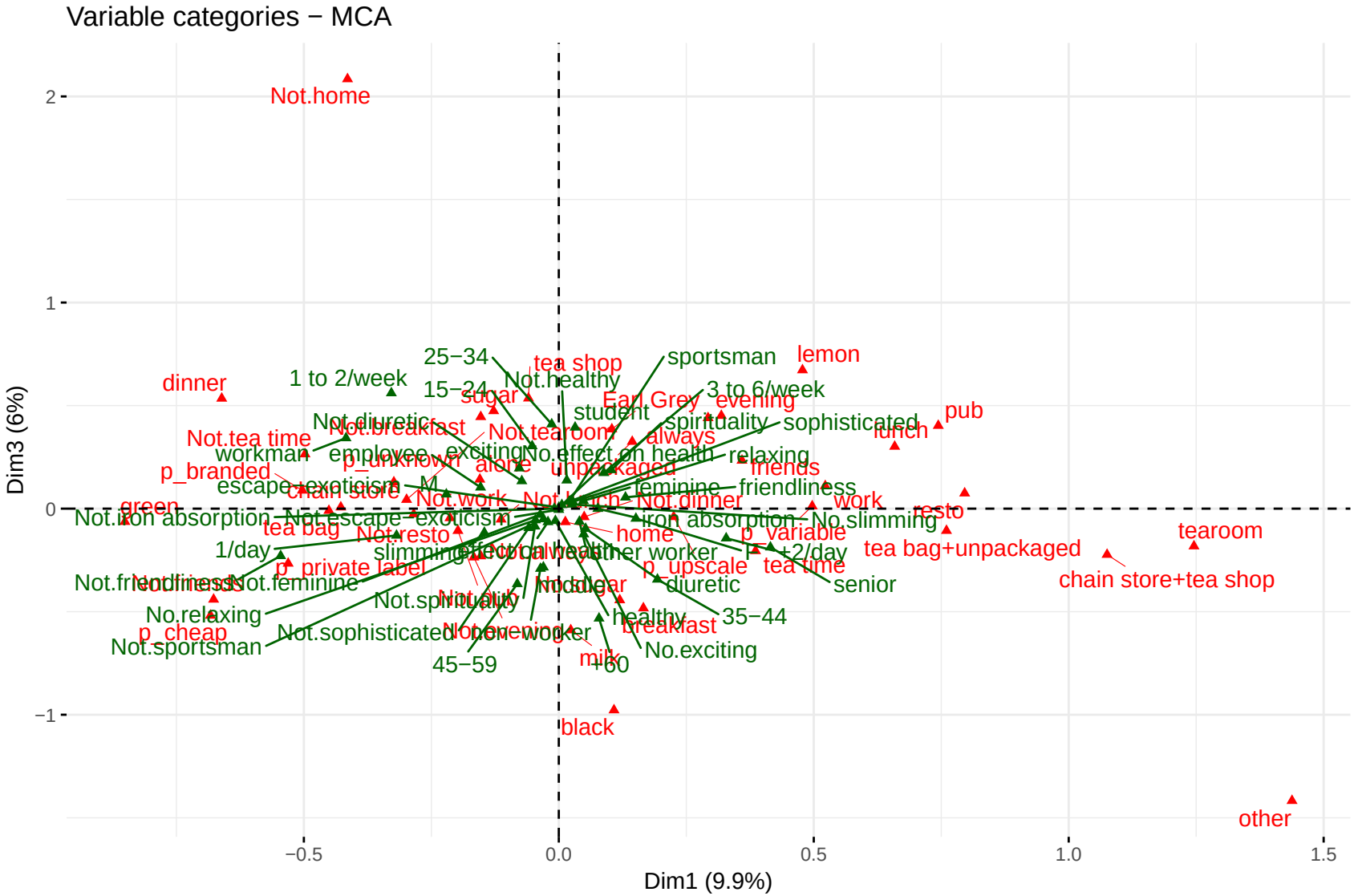
Diagramme des observations



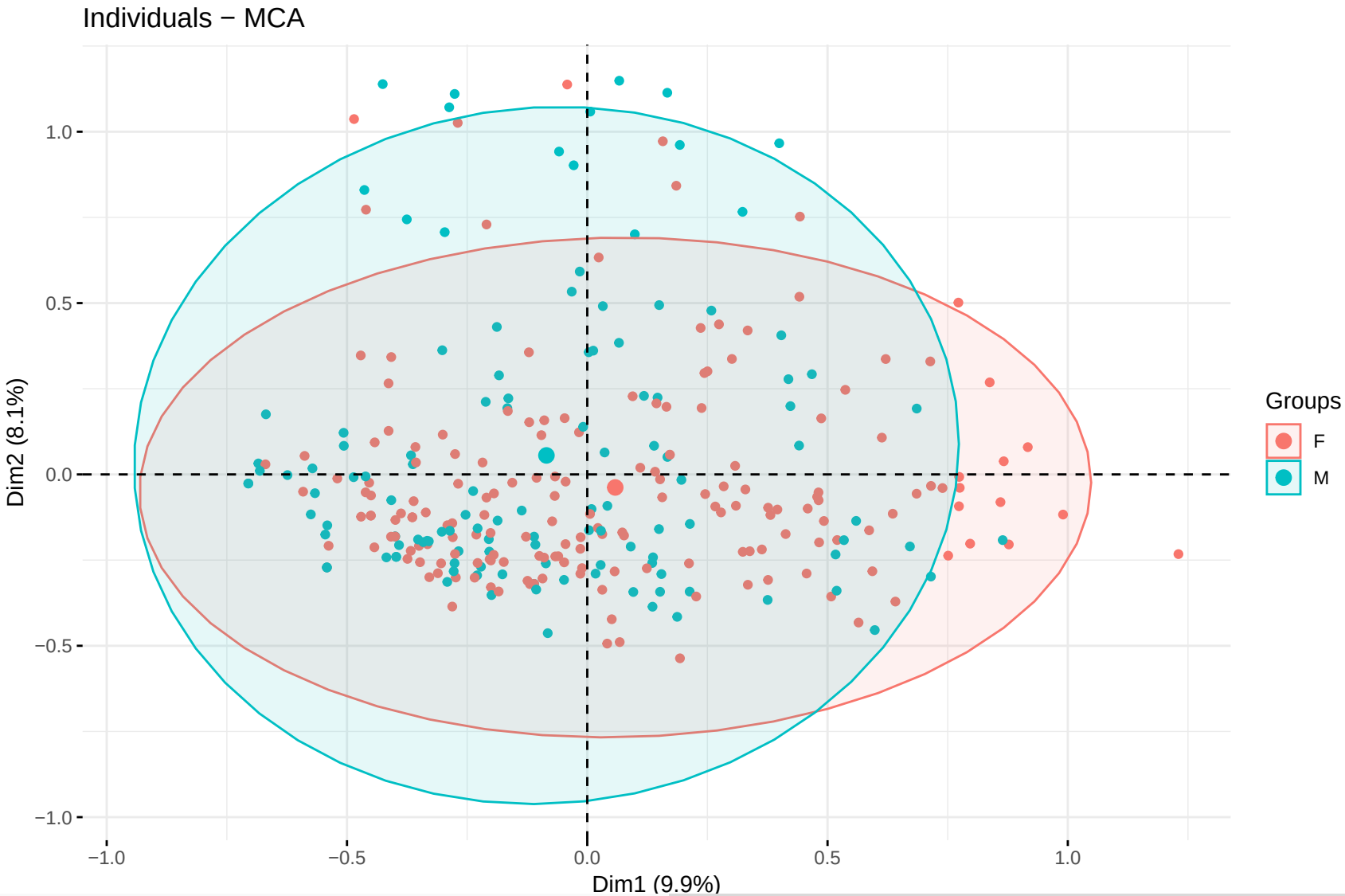
Variables supplémentaires



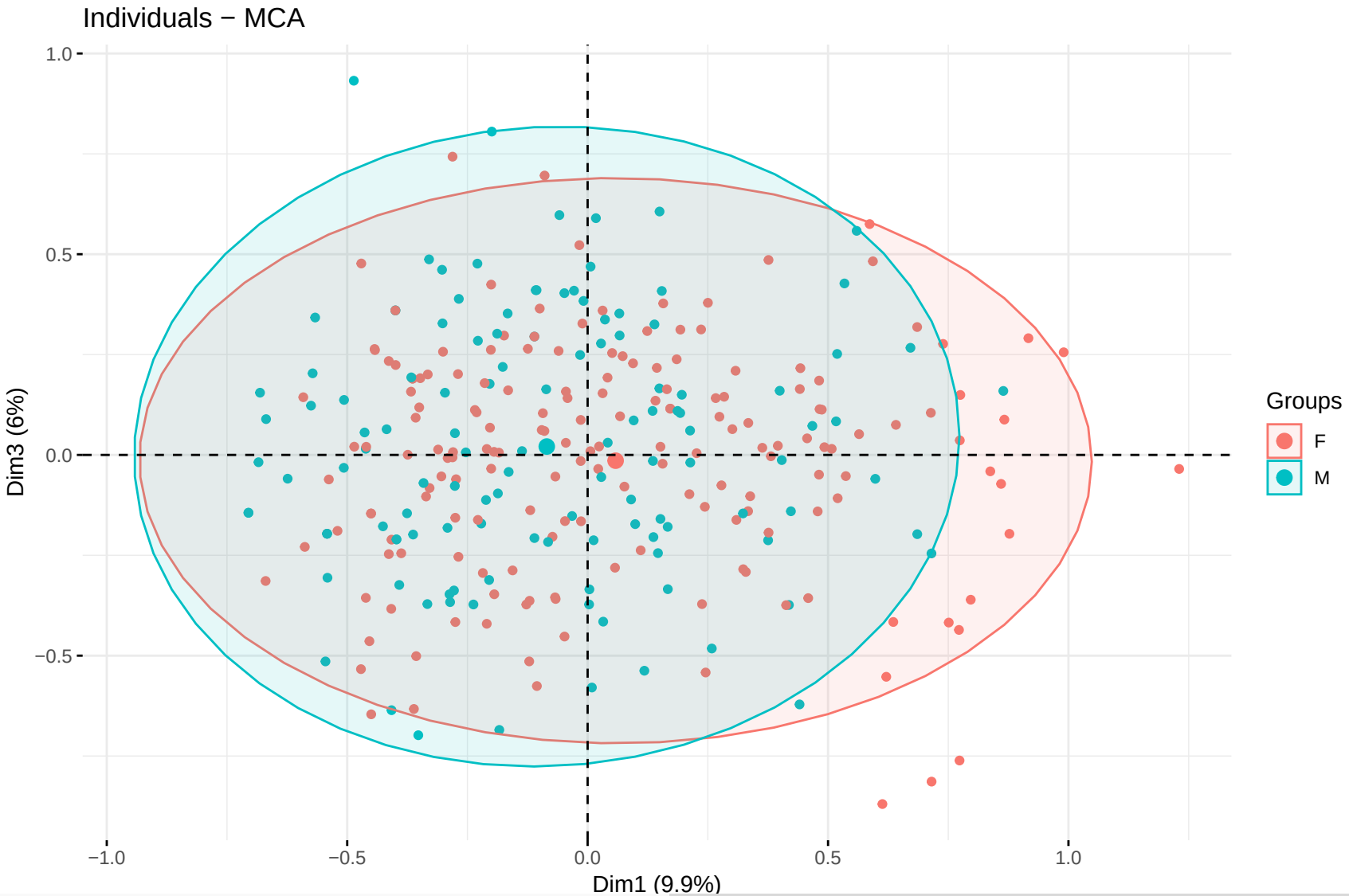
Variables supplémentaires



Variables supplémentaires



Variables supplémentaires



Construction: conseils pratiques I

La part de la variance dûe à une variable dépend uniquement du nombre de modalités. Par conséquent, si les variables ont un nombde de modalités très différentes, celles qui ont le plus de modalités vont mécaniquement peser plus lourd dans la construction du nuage et donc dans la détermination des axes principaux.

Il est donc recommandé de ne pas mettre de variables avec un très grand nombre de modalités et de limiter les écarts de nombre de modalité entre variables.

Plus généralement, si l'on fait l'hypothèse qu'il existe des variables latentes structurant le nuage, qui correspondent à des combinaisons de variables observées, il importe d'équilibrer le nombre supposé de modalités de chaque variable latente.

Construction: conseils pratiques II

Exemple: on cherche à cartographier l'espace social et l'on pense le faire avec des capitaux culturels et économiques. S'il y a seulement une question sur le diplôme avec trois modalités, mais plusieurs questions sur le capital économique (patrimoine immobilier en trois niveau, patrimoine financier en trois niveaux, revenus en trois niveaux)... le nuage sera nécessairement surdéterminé par le capital économique.

La distance entre deux individus dépend de l'inverse de la fréquence de leurs modalités ; par conséquent, elle s'accroît considérablement à mesure que la fréquence d'une modalité décroît. Les modalités rares produisent des coordonnées extrêmes dans le nuage et le surdétermine. Dans la mesure du possible, il convient de limiter l'usage de ces modalités à moins qu'elles ne soient effectivement très discriminante.

Construction: conseils pratiques III

Ces deux contraintes invitent en fait à recoder les questions pour fusionner les modalités trop proches, en particulier lorsqu'elles sont rares. Il ne s'agit pas d'un impératif: les hypothèses de recherche priment. Si une modalité rare est très discriminante et incommensurable avec d'autres (la possession d'un titre de noblesse par exemple), il faut la conserver.

L'ACM ne peut venir qu'après une étape de description des données. On cherchera notamment, en premier lieu: à identifier les non-réponses et les manières d'y remédier ; à identifier les liens entre les variables et à établir des hypothèses sur leurs associations dans une approche multivariée.

Construction: conseils pratiques IV

Le choix des variables actives et supplémentaires dépend des traditions de recherche et des questions posées. Le plus souvent, on met en variables actives les variables "dépendantes" et en variables supplémentaires les variables "indépendantes". Ainsi, le nuage est structuré par les pratiques étudiées (les consommations culturelles ; les rapports au travail ; etc.) et on teste l'homologie avec un espace social en regardant si les variables indépendantes sont correlées avec les axes principaux.

Cependant, les variables sociodémographiques sont parfois mises en variables actives, notamment lorsque l'enjeu de l'ACM est de cartographier un espace social pluridimensionnel (par exemple les travaux sur les élites (Denord, Lagneau-Ymonet, et Thine 2011)).

Construction: conseils pratiques V

Attention à un point durant l'interprétation: la proximité des projections sur un axe principal ne signifie pas que les points soient effectivement proches. En effet, ils ne permettent d'affirmer une proximité que sur cet axe en particulier. Ainsi, l'interprétation porte d'abord sur les oppositions plutôt que sur les proximités.

Bibliographie I

Bourdieu, Pierre. 2000. *Les structures sociales de l'économie*. Paris: Seuil.

Denord, François, Paul Lagneau-Ymonet, et Sylvain Thine. 2011. « Le champ du pouvoir en France ». *Actes de la recherche en sciences sociales* 190: 24-57.

Duval, Julien. 2013. « L'analyse des correspondances et la construction des champs ». *Actes de la recherche en sciences sociales* 200. CAIRN: 110-23. doi:10.3917/arss.200.0110.

Ollion, Étienne. 2011. « De la sociologie en Amérique. Éléments pour une sociologie de la sociologie états-unienne ». *Sociologie* 2 (3): 277-94. http://www.cairn.info.bibliotheque-nomade2.univ-lyon2.fr/revue-sociologie-2011-3-page-277.htm.

Bibliographie II

Rouanet, Henry, Frédéric Lebaron, Viviane Le Hay, Wemer Ackermann, et Brigitte Le Roux. 2002. « Régression et analyse géometrique des données. Reflexions et suggestions ». *Mathématiques et sciences humaines* 160: 13-45.

Rouanet, Henry, Brigitte Le Roux, et Wemer Ackermann. 2000. « The Geometrical Analysis of Questionnaires. The Lesson of Bourdieu's *La Distinction* ». *Bulletin de Méthodologie Sociologique* 65: 5-15.

Volle, Michel. 1997. *Analyse des données*. Paris: Economica.