L'ANALYSE DES CORRESPONDANCES MULTIPLES « À LA HOLLANDAISE » : INTRODUCTION A L'ANALYSE D'HOMOGENEITE

Dominique Desbois [1]
INRA-SAE2, UMR AgroParisTech Economie publique- Bureau du RICA, Service Central des Enquêtes et Etudes Statistiques, 12, rue Henri ROL-TANGUY, TSA 70007, 93555 MONTREUIL SOUS BOIS CEDEX.
Courriel :dominique.desbois@agriculture.gouv.fr - Fax : +33 1 49 55 85 00

RESUMÉ :
L'analyse des correspondances multiples est une méthode exploratoire multidimensionnelle qui fournit une représentation synthétique des catégories issues d'une batterie de critères qualitatifs, référentiel d'un protocole d'expérimentation ou d'enquête. Cette note a pour but d'aider les utilisateurs de *SPSS* dans la mise en oeuvre de l'analyse des correspondances multiples au moyen de l'analyse d'homogénéité (procédure HOMALS du logiciel *SPSS)*. Cette mise en oeuvre concerne l'analyse de tableaux de données construits à partir de variables nominales. L'équivalence entre l'analyse d'homogénéité et l'analyse des correspondances multiples est illustrée à partir d'un exemple répertorié dans la littérature statistique. La note est complétée par un exposé algébrique consacré à l'analyse d'homogénéité.
MOT CLEFS : Analyse des correspondances multiples, analyse d'homogénéité, logiciel statistique SPSS, mise en oeuvre.

MULTIPLE CORRESPONDENCE ANALYSIS "À LA HOLLANDAISE": INTRODUCTION TO THE ANALYSIS OF HOMOGENEITY

ABSTRACT :
The multiple correspondence analysis is a multidimensional exploratory method which provides a synthetic representation of the categories issued from a battery of qualitative criteria, belonging to a reference frame of an experimentation protocol or an investigation survey. The aim of this note is to help the SPSS users in the implementation of the multiple correspondence analysis by means of the homogeneity analysis (procedure HOMALS in the SPSS software). Equivalence between the analysis of homogeneity and the multiple correspondence analysis is illustrated on the basis of an example excerpted from the statistical literature. The note is supplemented by an algebraic addendum devoted to the homogeneity analysis.
KEY WORDS:
Multiple correspondence analysis, homogeneity analysis, software statistical SPSS, implementation.

HOMALS[2] [Gifi, 1990] est une procédure itérative basée sur la technique des moindres carrés alternées permettant de réaliser une analyse d'homogénéité. L'une des options particulières de cette procédure fournit les facteurs d'une analyse des correspondances multiples. L'objectif de cette note est donc de présenter l'analyse d'homogénéité pour les utilisateurs francophones de SPSS afin qu'ils puissent utiliser plus aisément cette procédure pour dépouiller leurs données d'enquête de façon pertinente, en réalisant des analyses de correspondances multiples.

[1] *L'auteur remercie Gilbert Saporta pour ses conseils de lecture et ses remarques critiques mais reste le seul responsable des éventuelles omissions ou erreurs.*
[2] *Homogeneity Analysis by Alternating Least Squares*– Analyse d'homogénéité par les moindres carrés alternés.

1. L'ANALYSE D'HOMOGENEITE, POUR UNE REPRESENTATION OPTIMALE DES CATEGORIES.

Soit un ensemble d'observations décrivant des **objets** au moyen de **catégories** issues d'une batterie de critères qualitatifs (**variables catégorielles**). L'analyse d'homogénéité est une technique exploratoire d'analyse des données permettant de décrire les relations existant entre deux ou plusieurs de ces variables catégorielles en fournissant une représentation graphique de leurs catégories, sous la forme d'un nuage de points (**points-catégories**) projetés dans un sous-espace de faible dimension.

Cette représentation graphique, effectuée dans un système d'axes orthonormés appelés « **dimensions** » est optimale au sens où elle maximise l'écart entre les positions des différentes catégories. Dans ce sous-espace particulier, on peut également représenter les objets soumis à l'observation (**points-objets**) en liant leur représentation à celle des catégories de référence de l'étude. Pour chaque variable, les catégories d'une même variable scindent le nuage des points représentant les objets en sous-nuages de points qui rassemblent les objets partageant la même catégorie. Les points représentant les catégories sont situés au centre du sous-nuage des points représentant les objets qui appartiennent à la même catégorie. Les proximités entres objets reflètent les similarités ou les dissimilarités entre leurs configurations respectives de réponse à la batterie de critères qualitatifs. Ainsi, les objets partageant un même profil de réponse sont projetés en un même point. Cependant, la réciproque n'est pas forcément vérifiée : deux objets dont les scores (valeurs de la projection selon les dimensions) sont proches ne sont pas nécessairement similaires. Si une variable possède un bon **pouvoir discriminant**, les objets se situeront à proximité des catégories auxquelles ils appartiennent. Idéalement, les objets classés dans la même catégorie doivent se situer à proximité les uns des autres, leurs scores étant similaires. Les catégories appartenant à des variables différentes sont situées à proximité les unes des autres si elles caractérisent les mêmes sous-ensembles d'objets. Ainsi, deux objets ayant des scores similaires pour un critère particulier doivent posséder des scores similaires pour les variables qui lui sont **homogènes**.

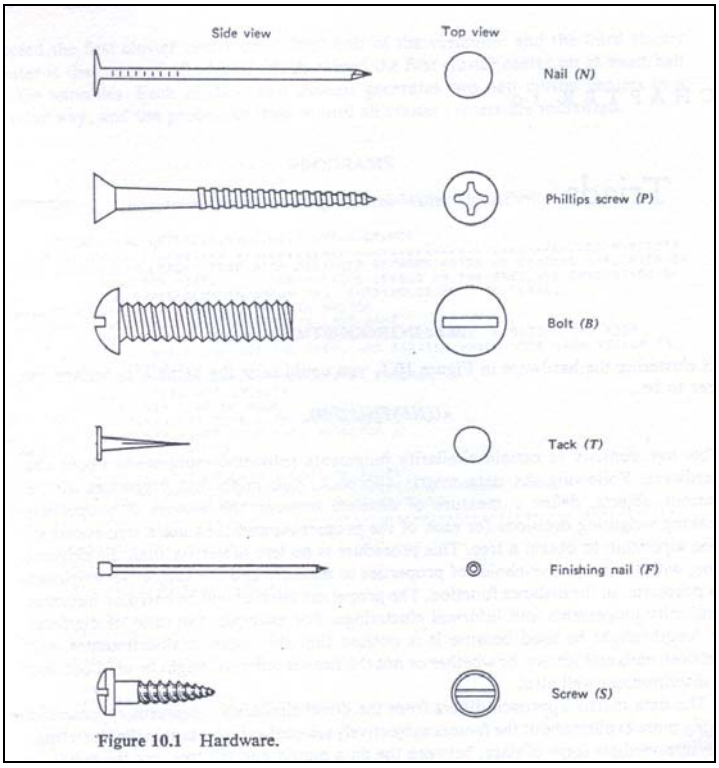
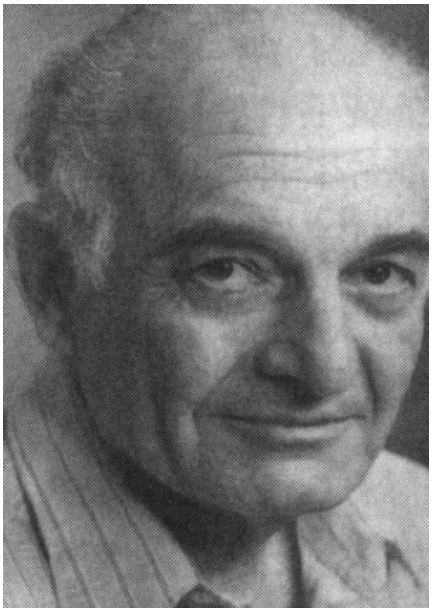


Figure 1 : *visualisation des objets, face et profil du petit matériel de quincaillerie (extrait de l'ouvrage [Hartigan 1975])***.**

Le terme d'**homogénéité** se réfère donc à une situation où les variables fournissent une partition de l'ensemble des objets selon les mêmes catégories ou des catégories similaires. Historiquement, le concept d'homogénéité est associé à un paradigme selon lequel des variables distinctes peuvent mesurer le même phénomène. Par exemple, pour les psychométriciens, les performances intellectuelles sont approchées à travers une batterie de tests qualifiés d'homogènes, au sens ou la somme des scores obtenus à un sens car elle fournit une mesure de ces performances.
De façon plus formelle, on peut définir l'analyse d'homogénéité, *stricto sensu*, comme un programme de minimisation d'une fonction-objectif, la **perte d'homogénéité** (cf. infra § 3 pour une définition), permettant d'obtenir une représentation graphique des catégories qui corresponde à la solution optimale présentée antérieurement. La généralisation de cette définition fournit un cadre méthodologique où le terme d'analyse d'homogénéité se réfère à une famille de techniques d'analyse multivariée partageant, selon différentes formes de codage des données et sous des formulations diverses du critère d'optimalité, un paradigme commun d'optimisation de l'homogénéité des variables.
L'analyse d'homogénéité peut être également présentée comme la solution d'un problème de décomposition en valeurs propres et en valeur propres singulières, et peut de ce fait être rattachée aux méthodes factorielles : ainsi, pour deux critères qualitatifs, l'analyse d'homogénéité est équivalente à l'analyse des correspondances ; pour plusieurs critères, elle

est équivalente à l'analyse des correspondances multiples. A ce titre, elle peut également être présentée comme une méthode de positionnement multidimensionnel travaillant à partir d'un tableau de « dissimilarités » constitué par les distances du Khi-Deux entre profils-lignes issus d'un tableau disjonctif complet codant, pour la **population *I*** des objets, les caractéristiques observées selon **l'ensemble *J*** des **modalités** ou catégories d'observation. L'analyse d'homogénéité peut également être considérée comme une analyse en composantes principales sur données nominales (modèle de Guttman). Lorsqu'il n'y a pas de relations linéaires entre variables ou lorsque les variables sont nominales, l'analyse d'homogénéité est préférable à une analyse en composantes principales normée (i.e. effectuées sur variables centrées et réduites).



Portrait de Louis GUTTMAN, 1916-1987
(Materials for the History of Statistics, The University of York)

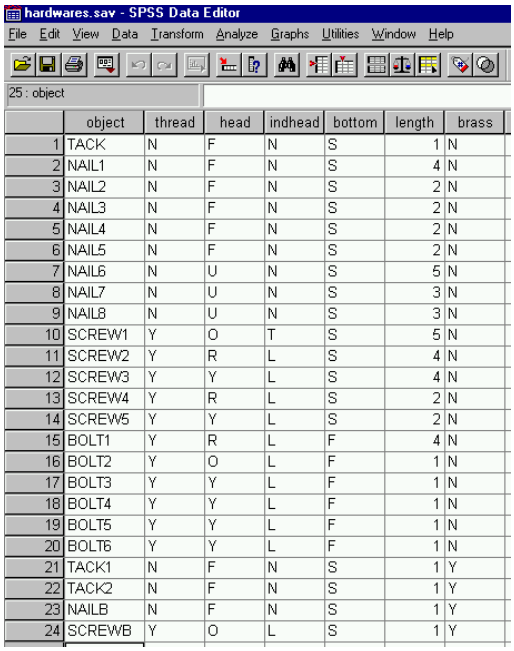
2. UN EXEMPLE D'ANALYSE D'HOMOGENEITE : les petits articles de quincaillerie.

Ce premier exemple illustratif de l'analyse d'homogénéité est basé sur des données décrivant de petits articles de quincailleries (clous, vis, boulons, etc.) à l'aide de variables catégorielles [Hartigan, 1975] décrivant leur forme et leur dimension. Il y a ***n=24*** objets ou observations et ***p=6*** variables descriptives catégorielles, la variable *OBJECT* identifiant les 24 observations.

Nom	Valeur	*Etiquette*	Position
OBJECT		*Objet*	1
THREAD		*Pointe*	2
	N	*non*	
	Y	*oui*	
HEAD		*Forme de la tête*	3
	F	*plate*	
	O	*conique*	
	R	*ronde*	
	U	*coupe*	
	Y	*cylindre*	
INDHEAD		*Indentation de la tête*	4
	L	*fente*	
	N	*aucune*	
	T	*étoile*	
BOTTOM		*Forme de la base*	5
	F	*plate*	
	S	*tranchante*	
LENGTH		*Longueur en demi-pouces*	6
	1	*0,5"*	
	2	*1"*	
	3	*1,5"*	
	4	*2"*	
	5	*2,5"*	
BRASS		*Cuivré*	7
	N	*non*	
	Y	*oui*	

Tableau 1 : *descriptif des données et détail des catégories*

Ci-dessous figure, dans l'éditeur de données *SPSS*, le tableau de ces données descriptives sous forme alphanumérique :



	object	thread	head	indhead	bottom	length	brass
1	TACK	N	F	N	S	1	N
2	NAIL1	N	F	N	S	4	N
3	NAIL2	N	F	N	S	2	N
4	NAIL3	N	F	N	S	2	N
5	NAIL4	N	F	N	S	2	N
6	NAIL5	N	F	N	S	2	N
7	NAIL6	N	U	N	S	5	N
8	NAIL7	N	U	N	S	3	N
9	NAIL8	N	U	N	S	3	N
10	SCREW1	Y	O	T	S	5	N
11	SCREW2	Y	R	L	S	4	N
12	SCREW3	Y	Y	L	S	4	N
13	SCREW4	Y	R	L	S	2	N
14	SCREW5	Y	Y	L	S	2	N
15	BOLT1	Y	R	L	F	4	N
16	BOLT2	Y	O	L	F	1	N
17	BOLT3	Y	Y	L	F	1	N
18	BOLT4	Y	Y	L	F	1	N
19	BOLT5	Y	Y	L	F	1	N
20	BOLT6	Y	Y	L	F	1	N
21	TACK1	N	F	N	S	1	Y
22	TACK2	N	F	N	S	1	Y
23	NAILB	N	F	N	S	1	Y
24	SCREWB	Y	O	L	S	1	Y

Figure 2 : *le tableau des données alphanumériques*

2.1. Pouvoir explicatif des dimensions de la solution

La représentation graphique que l'on souhaite obtenir de ces données en termes de catégories et d'objets, s'effectue dans un repère orthonormé dont on doit préciser le nombre d'axes ***a*** **,** appelé la **dimension de la solution**. La dimension maximum du sous-espace de représentation est égale soit au **nombre de catégories** (***m=19***) moins le nombre de variables sans valeurs manquantes (*p=6*), soit au nombre d'observations (*n=24*) moins un si celui-ci est inférieur, soit ***a****=min{13,23}=**13***. En pratique, le nombre d'axes utilisé pour la représentation est généralement très inférieur à ce maximum car souvent une solution comportant deux ou trois dimensions suffit pour synthétiser les traits essentiels de l'information contenue dans le tableau des données, l'information additionnelle apportée par des dimensions supplémentaires se révélant marginale.

Les **valeurs propres** permettent de rendre compte de l'importance relative de chaque dimension dans la part d'information statistique pris en compte par la solution. Ces valeurs propres prennent des valeurs dans l'intervalle $[0\,;1]$. La valeur 1 est atteinte par la valeur propre triviale qui correspond au vecteur propre reliant le centre de gravité du nuages des profils catégoriels et l'origine du repère. Les valeurs propres nulles correspondent à des directions indéterminées de la solution[3].

Eigenvalues

Dimension	Eigenvalue
1	,621
2	,368

Tableau 2 : *les deux premières valeurs propres.*

Leur rapport avec la somme totale des valeurs propres, appelé le **taux d'inertie** en analyse des correspondances, constitue une mesure pessimiste de la part de variabilité globale prise en compte. La procédure *HOMALS* de *SPSS* étant limitée à 10 dimensions, le calcul est effectué dans ce sous-espace. Néanmoins, les valeurs propres d'ordre supérieur ayant une valeur résiduelle, cette approximation ne change pas fondamentalement l'estimation des taux d'inertie.

Dimension	Valeur propre	Taux d'inertie	Inertie cumulée
1	0,621	0,287	0,287
2	0,368	0,170	0,457
3	0,328	0,151	0,608
4	0,279	0,129	0,737
5	0,197	0,091	0,828
6	0,128	0,059	0,887
7	0,086	0,040	0,927
8	0,084	0,039	0,966
9	0,056	0,026	0,991
10	0,019	0,009	1,000

Tableau 3 : *taux d'inertie associés au valeurs propres.*

[3] tout vecteur est solution de l'équation aux valeurs propres, donc vecteur propre.

Ainsi, les deux dimensions retenues permettent de prendre en compte 46% de l'inertie totale à travers une représentation graphique plane interprétable en termes de distances entre observations.

2.2. Représentation graphique des objets à partir des scores

Les ***scores*** (coordonnées des objets selon les premières dimensions de la solution) permettent de repérer les valeurs extrêmes (« *outlier* ») : l'objet projeté à l'extrémité négative de la dimension 2 (**D2<0**) peut être considéré comme une valeur atypique ou aberrante et, de ce fait, éventuellement exclu lors d'une analyse ultérieure (cf. infra).

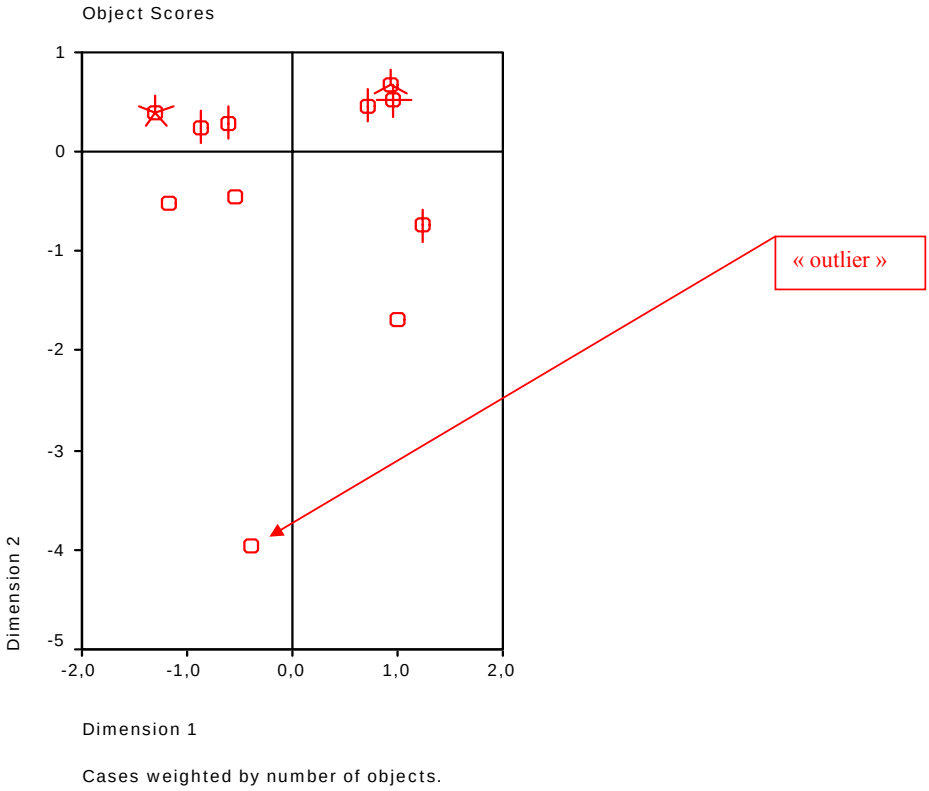


Figure 3 : *projection des objets dans le plan des deux premières dimensions.*

Cette représentation des objets sous forme de **tournesol** (le nombre de pétales du tournesol est proportionnel au nombre d'objets) est bien adaptée aux ensembles d'objets dont la cardinalité *n* est importante car elle permet de rendre compte des différences de densité au sein du nuage des points-objets.

Si le nombre d'observations est suffisamment faible, il est alors possible de projeter chacune des observations avec leur identifiant. Cela permet de vérifier la configuration de réponses fournies par des sous-ensembles particuliers d'objets. Ce graphique permet de constater que la **première dimension** (axe horizontal **D1**) sépare les vis (*screw*) et les boulons (*bolt*), qui ont un filetage (*thread*), des clous (*nail*) et des punaises (*tack*) qui n'en ont pas. De façon moins prononcée, cette première dimension instaure une séparation entre les boulons (*bolt*) qui ont

une base plate et tous les autres objets (qui ont une base pointue). La **seconde dimension** (axe vertical **D2**) sépare les objets *screw1* et *nail6* de l'ensemble des autres objets : ces deux objets sont les plus longs (cf. figure 2). Notons également que *screw1* apparaît comme l'objet le plus éloigné de l'origine : la configuration des caractéristiques de cet objet apparaît comme très spécifique puisqu'elle n'est partagée par aucun autre objet.

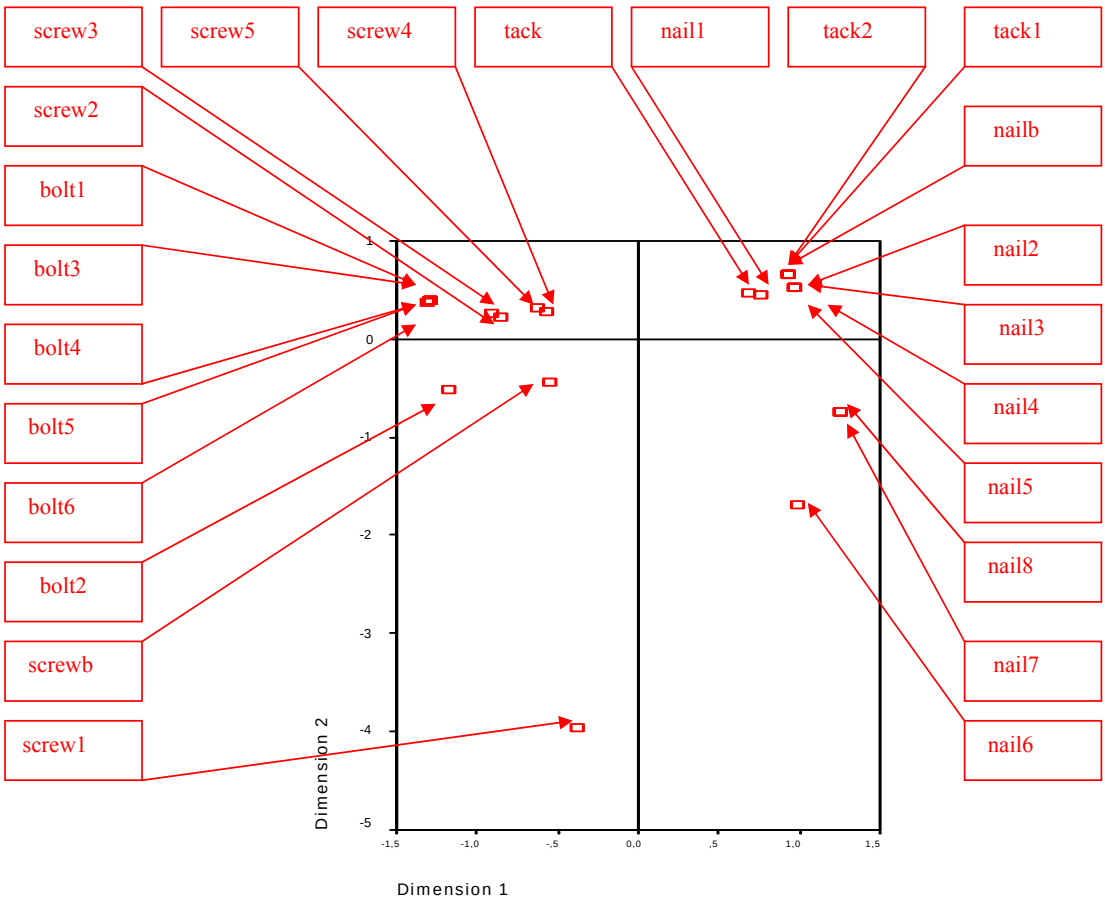


Figure 4 : *étiquetage des objets dans le plan des deux premières dimensions.*

Cependant, la pratique des variables illustratives (cf. infra § 2.5) dans l'établissement des graphiques facilite la synthèse de ces informations : pour chacun de ces graphiques illustratifs, les objets sont étiquetés à partir de la palette de valeurs catégorielles issue de la variable illustrative sélectionnée.
La procédure *HOMALS* permet de spécifier les variables illustratives utilisées pour produire une représentation graphique de la densité des différentes modalités de réponse.

2.3. Mesures du pouvoir discriminant

La mesure du pouvoir discriminant d'un critère relativement à une dimension peut se définir comme le pourcentage de variance de la dimension expliqué par ce critère. La valeur maximum de cet indicateur est égale à 1 si tous les objets se répartissent sur l'ensemble de ces catégories (caractère complet de la nomenclature des catégories) et si les objets appartenant à la même catégorie se révèlent identiques en termes de configuration descriptive relativement aux autres critères. S'il y a des données manquantes dans le tableau analysé, l'indice du pouvoir discriminant du critère peut être supérieur à 1.

Cette mesure du pouvoir discriminant étant calculée comme la moyenne pondérée, par la fréquence des catégories, des carrés des coordonnées des catégories (*quantifications*). Dans le langage de l'analyse des correspondances, il s'agit de la moyenne pondérée des qualités de représentation des modalités de cette variable sur l'axe factoriel. Le pouvoir discriminant d'un critère est d'autant plus élevée= que ses catégories présentent une dispersion importante de leurs coordonnées selon la dimension examinée. La moyenne des indices de discrimination sur l'ensemble des critères est égale pour chaque dimension à la valeur propre correspondante, exprimant ainsi la variance de cette dimension.

Les dimensions sont ordonnées dans l'ordre décroissant de leur variance , les valeurs propres étant extraites par ordre d'importance décroissant : la direction de la première dimension correspond au vecteur propre associé à la première valeur propre (la plus élevée) ; la direction de la seconde dimension correspond au second vecteur propre associé à la seconde valeur propre en importance ; etc.

Le diagramme des mesures du pouvoir discriminant indique que la première dimension est constituée par une synthèse des variables *thread* (présence d'une pointe) et *bottom* (forme de la base) : les deux variables présentent des niveaux d'indice de discrimination importants pour la 1ère dimension et faibles pour la 2nde dimension. Ainsi, les catégories de ces variables sont bien dispersées selon l'axe *D1* et peu dispersées selon l'axe *D2*.

Inversement, la variable *length* présente une valeur élevée de l'indice de discrimination selon l'axe *D2* et une valeur faible pour l'axe D1. En conséquence, l'angle entre le vecteur correspondant à cette variable et la 2nde dimension est faible, la valeur de l'indice selon l'axe *D2* correspondant au carré du cosinus de l'angle. Cet indice, assimilable au carré d'un coefficient de corrélation (R^2), exprime la similarité entre les deux directions, et reflète la ségrégation observée selon la 2nde dimension sur le diagramme des objets entre les objets les plus longs (situés dans le demi-plan *D2<0*) et l'ensemble des autres objets (situés dans le demi-plan *D2>0*).

Remarquons également que les variables concernant la forme et l'indentation de la tête présentent des valeurs importantes de leurs indices de discrimination selon les deux dimensions.

Par contre la variable *brass* située près de l'origine du graphique n'apparaît pas comme discriminante dans ce plan des deux premières dimensions, l'ensemble des objets pouvant posséder ou non le caractère cuivré. Pour la même raison, la variable *length* ne peut être liée à la 1ère dimension puisqu'elle ne discrimine les objets que dans la 2nde dimension.

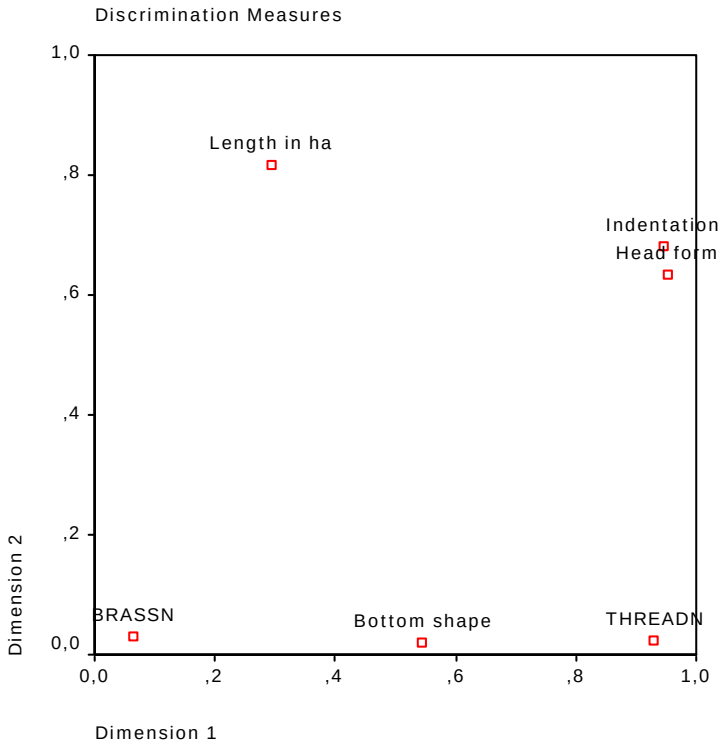


Figure 5 : *mesure du pouvoir discriminant selon les deux premières dimensions.*

Si l'indice de discrimination indique quelle est la part de variance expliquée par une variable pour chaque dimension, il ne permet pas de distinguer entre les variables dont les catégories présentent une dispersion moyenne selon une dimension et celles dont la plupart des catégories ont des coordonnées similaires à l'exception de certaines d'entre elles très différentes.

2.4. Quantifications des catégories

En revanche, les projections graphiques des catégories permettent de caractériser précisément les relations entre catégories d'une même variable mais aussi entre catégories de variables distinctes, en situant chaque catégorie sur un même graphique au moyen de leurs *quantifications* selon chaque dimension (équivalent des coordonnées factorielles des profils catégoriels dans l'analyse des correspondances multiples).

Ainsi, la variable *length* possède cinq catégories dont trois sont localisées dans la partie supérieure du graphique (demi-plan *D2>0)* et les deux autres (soit 1,5'' et 2,5'') se situent dans la partie inférieure du graphique (demi-plan *D2<0*).

En outre, la catégorie étiquetée *2_1/2_in* (soit 2,5'') située à l'extrémité négative de la 2nde dimension, se singularise très nettement par rapport à l'ensemble des autres catégories, rejoignant en cela la catégorie *STAR* (tête en étoile ou cruciforme) de la variable *Indentation of head* (indentation de la tête). En fait, la catégorie *2_1/2_in* est située au point moyen

(barycentre) des localisations des deux objets qui partagent cette spécificité, soit *screw1* et *nail6*.
La catégorie *STAR* se situe exactement au lieu géométrique de projection de l'objet *screw1* qui est le seul à présenter cette indentation cruciforme de la tête. Cette catégorie *STAR* se différencie des deux autres catégories (*SLIT* – fente et *NONE* – sans indentation) selon la 2nde dimension.

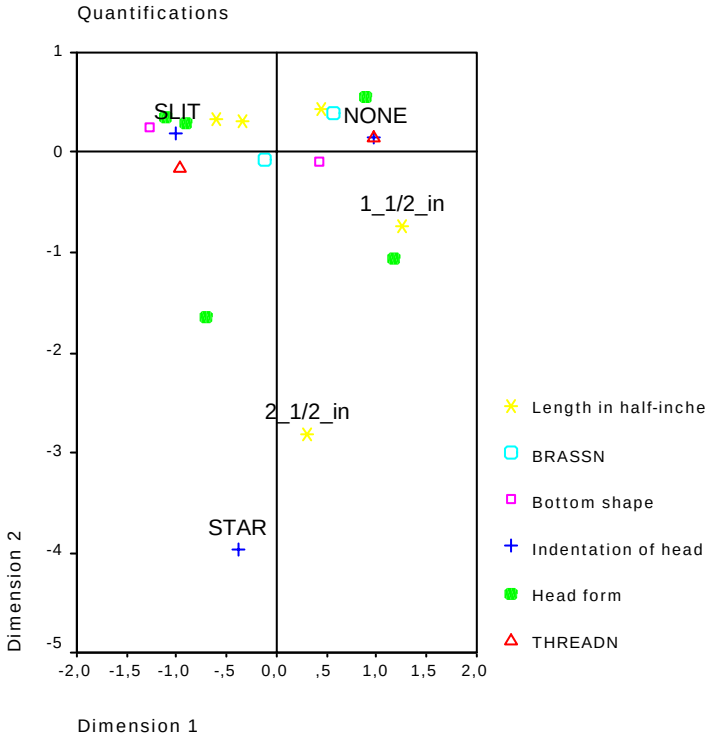


Figure 6 : *quantification des catégories.*

La dispersion des catégories d'une variable selon une dimension particulière reflète la variabilité de la configuration des réponses et constitue un indicateur de son pouvoir discriminant relatif à cette dimension.
Ainsi, selon l'axe horizontal *D1*, les catégories de la variable *THREADN* (codage numérique de la variable *thread*) sont très dispersées alors qu'elles ne le sont pas selon l'axe vertical *D2*. Il s'en suit que la variable *thread* discrimine mieux les objets selon la 1ère dimension que selon la 2nde dimension.
En revanche, les catégories de la forme de la tête (*Head form*) sont autant dispersées selon l'axe *D1* que selon l'axe *D2*. On en conclut que le pouvoir discriminant de cette variable est équivalent selon les deux dimensions.
Une variable dont les catégories sont plus dispersées selon une dimension possède un pouvoir discriminant plus important selon cette dimension qu'une autre variable dont les catégories sont projetées de façon moins dispersées. Par exemple, selon la 1ère dimension, les deux catégories de la variable *BRASSN* (codage numérique de la variable *brass* - caractère cuivré)

sont beaucoup moins dispersées que les deux catégories de la variable *THREADN*, indiquant que la variable *thread* possède un pouvoir discriminant plus important que celui de *brass* selon cette dimension (vérifiable en figure 5, d'après les niveaux relatifs de la mesure de discrimination des deux variables considérées).

2.5. Graphiques illustratifs

On peut éventuellement pousser plus loin l'analyse en consultant les différents graphiques illustratifs projetant individuellement, pour chaque variable, les objets étiquetés par le codage des catégories.

L'utilisation de ces variables illustratives montre que la 1ère dimension sépare parfaitement le groupe des articles possédant une pointe, étiquetés *Yes_Thread* et situés dans le demi-plan [*D1<0*], du groupe de ceux qui n'ont pas de pointe, étiquetés *No_Thread* et situés dans le demi-plan [*D1>0*].

Cette différenciation parfaite en fait un indicateur bien corrélé à la 1ère dimension.

Figure 7 : *projection des objets, variable illustrative THREADL (« présence d'une pointe »).*

La projection des objets étiquetés par la forme de la tête (*Head form*) montre que celle-ci discrimine bien les articles dans les deux dimensions.
Les objets à tête plate (*FLAT*) sont situés dans le quadrant supérieur droit [*D2>0 & D1>0*] tandis que les articles dont la tête est en coupe (*CUP*) sont situés dans le quadrant inférieur droit [*D2<0 & D1>0*].
Les objets à tête conique (*CONE*) sont situés dans le quadrant inférieur gauche [*D2<0 & D1<0*] mais on observe que ces objets sont beaucoup plus dispersés que dans les autres catégories.
Dans le quadrant supérieur gauche [D2>0 & D1<0], les objets à tête cylindrique (*CYLINDER*) ne peuvent être distingués des objets à tête ronde (*ROUND*).

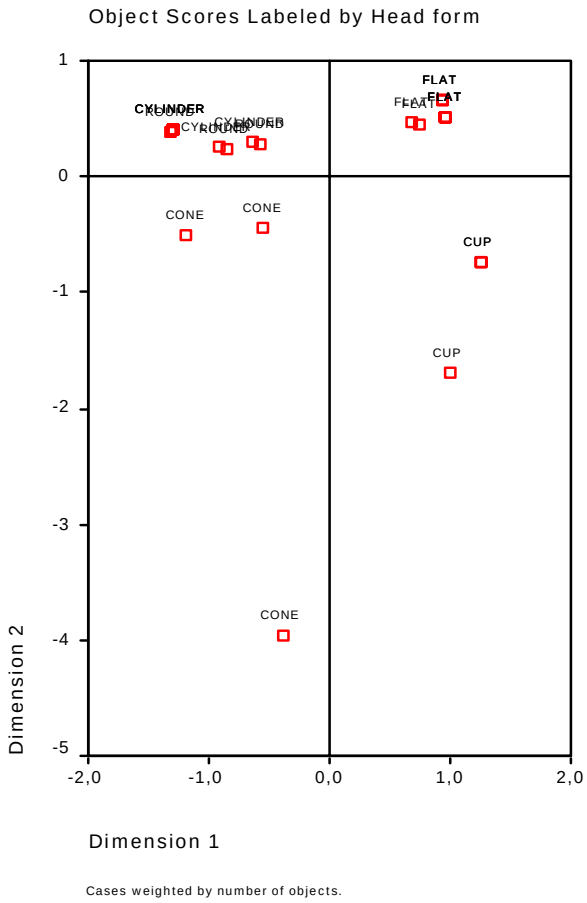


Figure 8 : *projection des objets, variable illustrative HEADL (« forme de la tête »).*

Le graphique selon les catégories de longueur montre que ces catégories se distinguent non pas selon l'axe horizontal du graphique mais plutôt selon l'axe vertical. Ce constat confirme l'analyse selon laquelle les catégories de la variable *length* ne discriminent pas les objets selon la 1ère dimension mais seulement selon la 2nde, les objets les plus courts étant situés dans le demi-plan [*D2>0*]

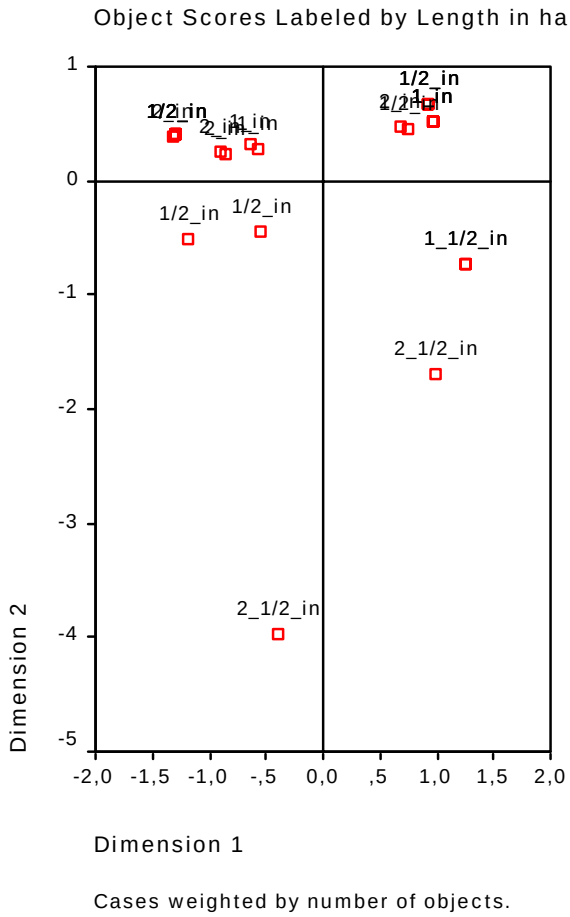


Figure 9 : *projection des objets, variable illustrative LENGHTL« longueur en pouces »*

Le graphique illustratif à partir de la variable *BRASS* (caractère cuivré ou non de l'objet) ne permet pas de mettre en évidence une différenciation nette des objets selon l'une ou l'autre des deux premières dimensions.

2.6. Filtrage des observations atypiques

Une fois identifiées les observations atypiques comportant trop de caractéristiques qui leur sont propres, on peut les exclure de l'analyse par filtrage, permettant ainsi de se focaliser sur les phénomènes dont l'occurrence n'est pas marginale. Si l'on réitère l'analyse d'homogénéité après un traitement excluant cette observation jugée atypique, on constate un léger changement au niveau des valeurs propres qui ne modifie pas de manière radicale l'ordre de grandeur de leur taux d'inertie. Pour autant, on ne doit pas conclure sans examen préalable à la quasi-équivalence des deux analyses

Le graphique des mesures de discrimination indique désormais que l'indentation de la tête (« *head indentation* ») ne discrimine plus les objets selon la 2nde dimension mais seulement selon la 1ère dimension, tandis que le caractère discriminant de la variable *brass* (cuivré ou non) se manifeste désormais selon la 2nde dimension. Les indices de discrimination des autres variables demeurent inchangés dans ces deux premières dimensions.

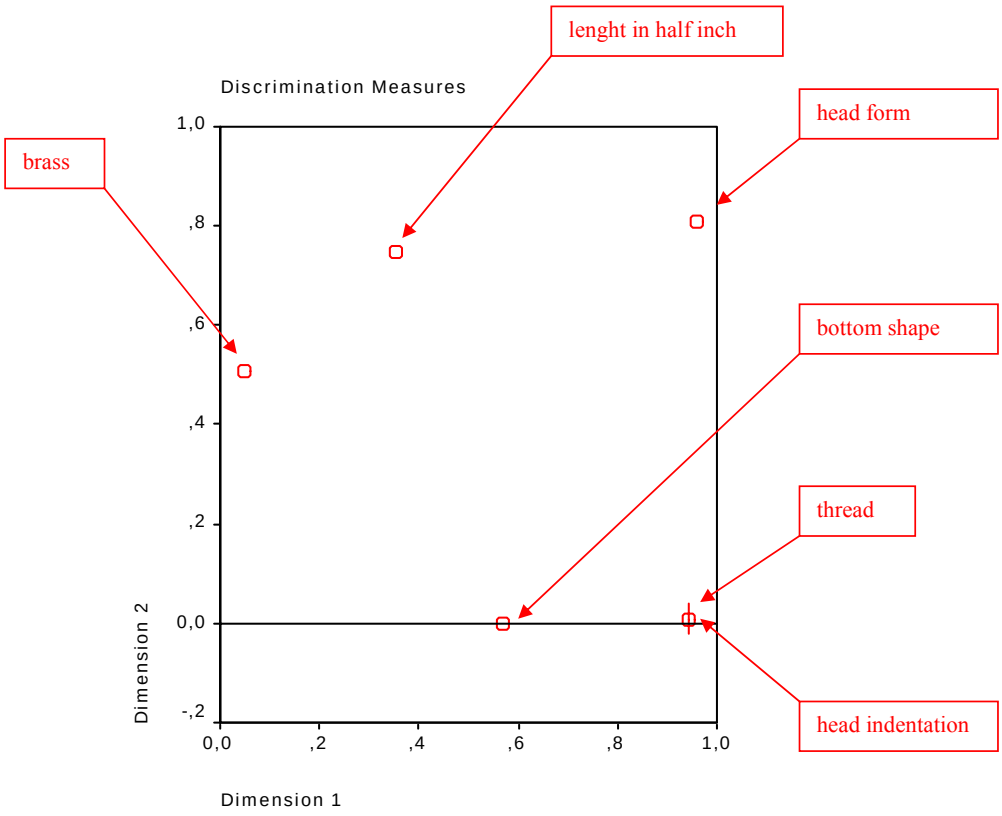


Figure 10 : *mesures de discrimination, après filtrage de l'objet atypique.*

Le graphique des objets étiquetés par la variable *brass* montre que les objets cuivrés (« *YES_Br* ») sont désormais projetés à l'extrémité négative de la 2nde dimension (zone [*-2<D2<-1*]) alors que les objets non cuivrés (« *Not_Br* ») sont projetés dans le demi-plan [*D2>-1*], confirmant ainsi le pouvoir discriminant de la variable *brass* selon la 2nde dimension.

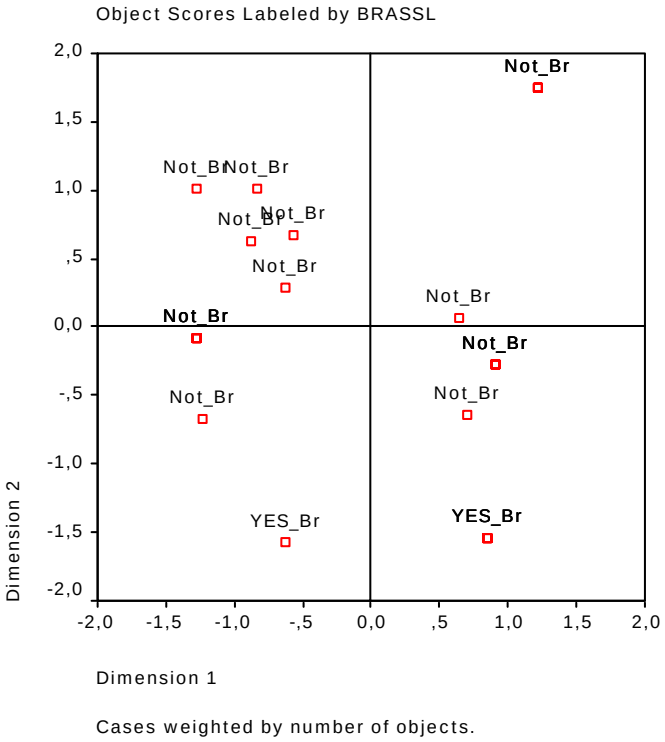
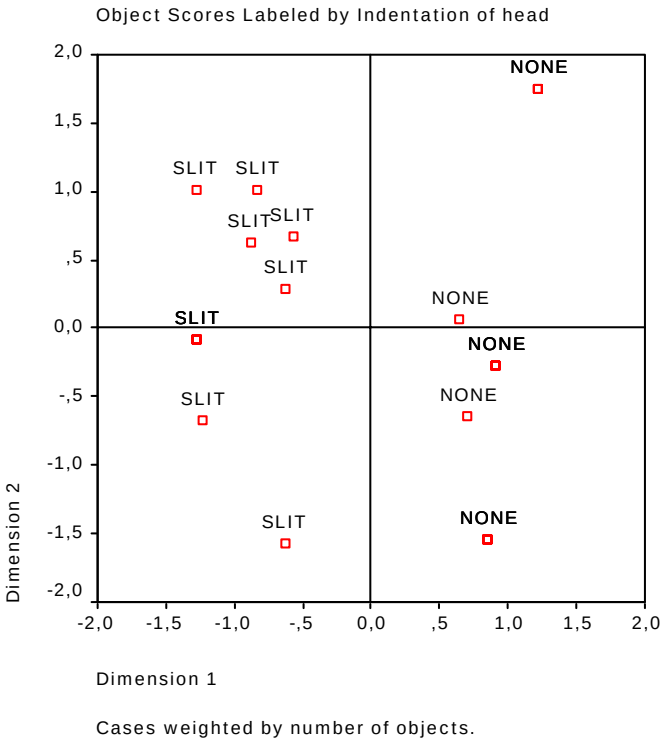


Figure 11 : *projection des objets étiquetés par BRASSL, après filtrage de l'objet atypique*

La projection illustrative des objets étiquetés par les catégories relatives à l'indentation de la tête (« *Indentation of head* ») montre que la première dimension permet de discriminer parfaitement les objets non indentés (« *NONE* ») des objets indentés (« *SLIT* »), comme dans l'analyse précédente.
Cependant, la 2nde dimension ne discrimine plus les catégories d'indentation, à l'inverse de l'analyse précédente.



.

Figure 12 : *projection des objets étiquetés par indentation de la tête (« INDHEADL»), après filtrage de l'objet atypique*

3. L'ANALYSE D'HOMOGENEITE, POUR UNE REPRESENTATION OPTIMALE DES CATEGORIES.

3.1. Le concept d'homogénéité

Développée par le groupe Albert Gifi[4], la procédure *HOMALS* se base sur le concept d'**homogénéité**, que l'on peut définir de la manière suivante.

Soit le vecteur $\mathbf{z}_j$, $j = 1, \cdots, p$, contenant les observations faites sur les *n* individus d'une population, correspondant à la variable Z_j.

Le vecteur $\mathbf{z}_j$ est ***homogène à*** $\mathbf{x}$, vecteur unitaire (de norme 1), si et seulement si après une transformation t_j de normalisation (tel que $\|t_j(\mathbf{z}_j)\| = 1$), on a $\mathbf{x} = t_j(\mathbf{z}_j)$.

Si le vecteur $\mathbf{z}_j$ n'est pas homogène à $\mathbf{x}$, on définit la ***perte d'homogénéité*** comme suit : $\sigma^2(\mathbf{x}, t) = \frac{1}{p}\sum_{j=1}^{p} {}^t(\mathbf{x} - t_j(\mathbf{z}_j))(\mathbf{x} - t_j(\mathbf{z}_j))$.

3.2. La procédure HOMALS

Soit la matrice $\mathbf{Z}_j$ des indicatrices de codage correspondant aux indicatrices de codage d'une variable Z_j qualitative à k_j modalités. La transformation t_j du vecteur $\mathbf{z}_j$ peut être définie par $t_j(\mathbf{z}_j) = \mathbf{Z}_j\mathbf{Y}_j$ où $\mathbf{Y}_j$ est une matrice à $n \times k_j$ coefficients.

La procédure *HOMALS* consiste à minimiser la fonction de perte suivante :

$$\sigma^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{p}\sum_{j=1}^{p} trace\left[{}^t(\mathbf{X} - \mathbf{Z}_j\mathbf{Y}_j)(\mathbf{X} - \mathbf{Z}_j\mathbf{Y}_j)\right]$$

sous les contraintes d'orthonormalisation ${}^t\mathbf{XX} = nI$ et de centrage $\mathbf{1X} = 0$.

3.3. Equivalence avec l'analyse des correspondances multiples

[Gifi, 1990] présente l'analyse d'homogénéité comme la résolution d'un problème de décomposition spectrale, soit en valeurs singulières, soit en valeur propres, qui fournit en fait les facteurs d'une analyse des correspondances multiples. Cette présentation est issue du travail de [Tenenhaus et Young, 1985] qui établit un cadre conceptuel commun pour analyser les relations entre différentes méthodes multivariées d'analyse de données catégorielles, montrant ainsi l'équivalence entre analyse des correspondances multiples et analyse d'homogénéité. L'analyse d'homogénéité peut également être vue comme une technique de positionnement multidimensionnel restituant une image euclidienne (à partir de graphiques-plans) des « dissimilarités » constituées par les distances du Khi-Deux entre profils-lignes.

4. EFFECTUER UNE ANALYSE D'HOMOGENEITE AVEC *SPSS*

Pour obtenir une analyse d'homogénéité sous *SPSS*, il convient de créer par recodage, à partir du tableau des données alphanumériques (cf. figure 2), un tableau numérique comportant l'ensemble des variables à analyser. Pour ce faire, il faut utiliser la procédure de recodage automatique <*Automatic Recode*> du menu de transformation <*Transform*>, créant ainsi la variable *threadn* (codage

[4] Albert Gifi fût durant quarante années le maître d'hôtel de Sir Françis Galton [Gilham, 2001] avant de devenir le nom collectif des membres du *Department of Data Theory* de l'Université de Leiden (Pays-Bas). Ce groupe, constitué autour de Jan de Leeuw a mis au point un système pour l'analyse multivariée non linéaire qui recouvre de multiples techniques factorielles allant de l'analyse en composantes principales à l'analyse canonique. Le travail de ce groupe est présenté dans l'ouvrage [Gifi, 1990]

numérique) à partir de la variable *thread* (codage alphanumérique) par transformation des catégories prises dans un ordre lexicographique croissant (cf. figure 13).

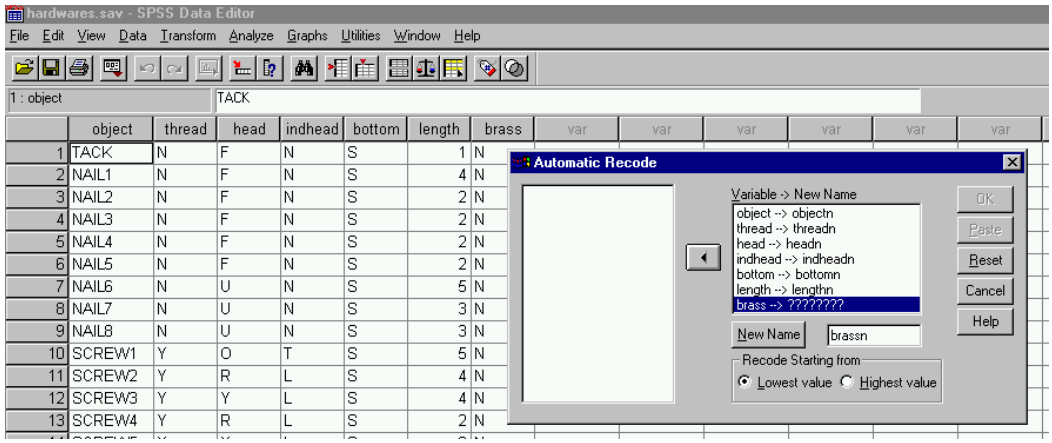
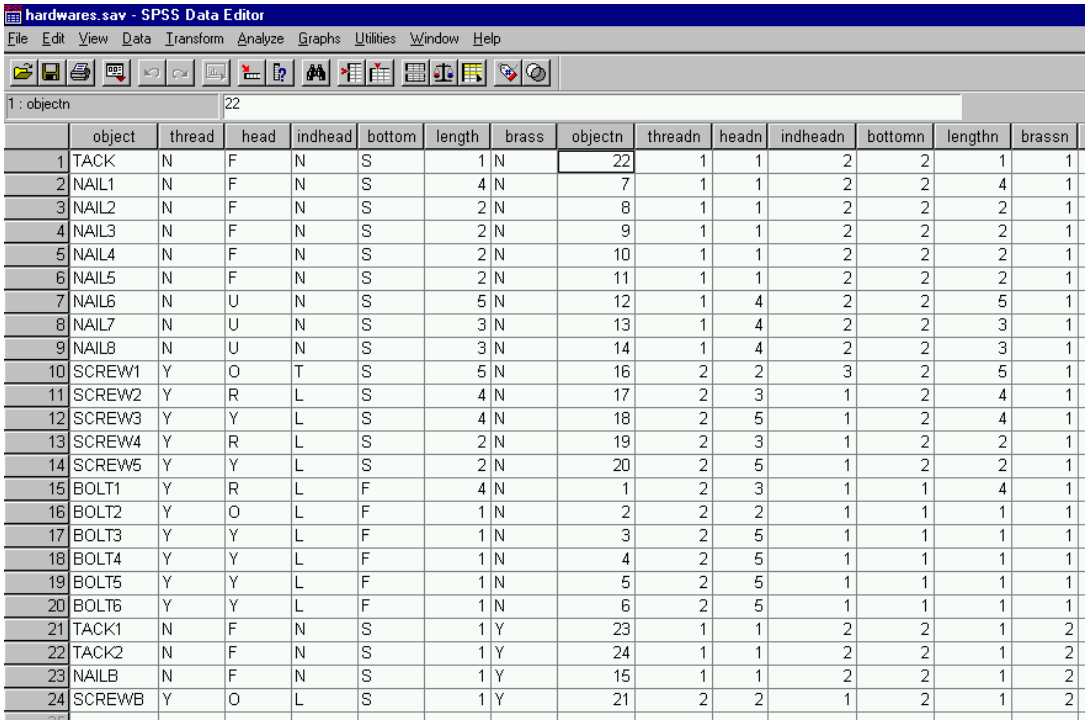


Figure 13 : *recodage des variables alphanumériques en variables numériques***.**

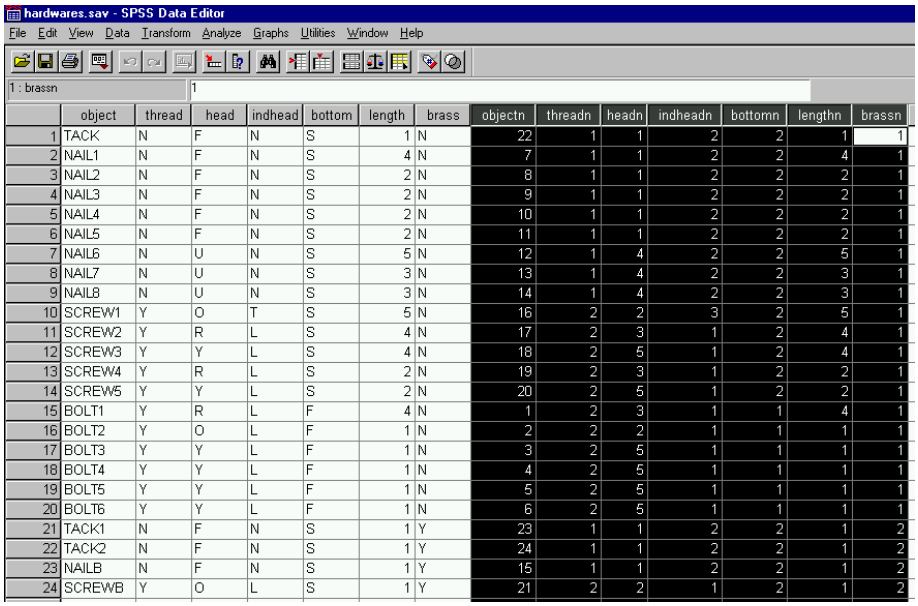


	object	thread	head	indhead	bottom	length	brass	objectn	threadn	headn	indheadn	bottomn	lengthn	brassn
1	TACK	N	F	N	S	1	N	22	1	1	2	2	1	1
2	NAIL1	N	F	N	S	4	N	7	1	1	2	2	4	1
3	NAIL2	N	F	N	S	2	N	8	1	1	2	2	2	1
4	NAIL3	N	F	N	S	2	N	9	1	1	2	2	2	1
5	NAIL4	N	F	N	S	2	N	10	1	1	2	2	2	1
6	NAIL5	N	F	N	S	2	N	11	1	1	2	2	2	1
7	NAIL6	N	U	N	S	5	N	12	1	4	2	2	5	1
8	NAIL7	N	U	N	S	3	N	13	1	4	2	2	3	1
9	NAIL8	N	U	N	S	3	N	14	1	4	2	2	3	1
10	SCREW1	Y	O	T	S	5	N	16	2	2	3	2	5	1
11	SCREW2	Y	R	L	S	4	N	17	2	3	1	2	4	1
12	SCREW3	Y	Y	L	S	4	N	18	2	5	1	2	4	1
13	SCREW4	Y	R	L	S	2	N	19	2	3	1	2	2	1
14	SCREW5	Y	Y	L	S	2	N	20	2	5	1	2	2	1
15	BOLT1	Y	R	L	F	4	N	1	2	3	1	1	4	1
16	BOLT2	Y	O	L	F	1	N	2	2	2	1	1	1	1
17	BOLT3	Y	Y	L	F	1	N	3	2	5	1	1	1	1
18	BOLT4	Y	Y	L	F	1	N	4	2	5	1	1	1	1
19	BOLT5	Y	Y	L	F	1	N	5	2	5	1	1	1	1
20	BOLT6	Y	Y	L	F	1	N	6	2	5	1	1	1	1
21	TACK1	N	F	N	S	1	Y	23	1	1	2	2	1	2
22	TACK2	N	F	N	S	1	Y	24	1	1	2	2	1	2
23	NAILB	N	F	N	S	1	Y	15	1	1	2	2	1	2
24	SCREWB	Y	O	L	S	1	Y	21	2	2	1	2	1	2

Figure 14 : *variables numériques recodées***.**

Dans une seconde étape, il faut créer par recopie autant de variables illustratives qu'il y a de critères participant à l'analyse. Pour ce faire, il suffit de sélectionner les variables recodées en

cliquant avec la touche « *Control* » maintenue enfoncée (« *Ctrl+Clic* ») sur les colonnes correspondantes de l'éditeur des données (cf. figure 15).



hardwares.sav - SPSS Data Editor

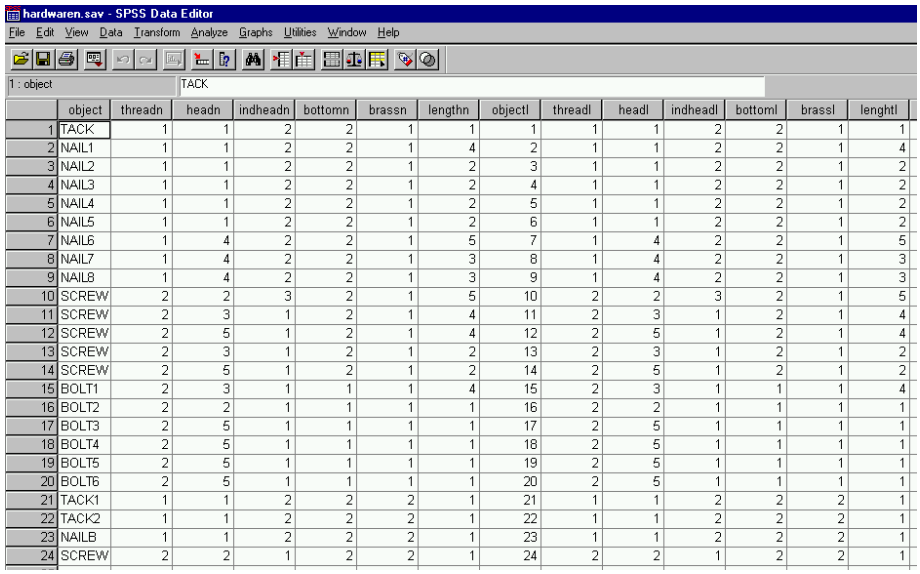
File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : brassn | 1

	object	thread	head	indhead	bottom	length	brass	objectn	threadn	headn	indheadn	bottomn	lengthn	brassn
1	TACK	N	F	N	S	1	N	22	1	1	2	2	1	1
2	NAIL1	N	F	N	S	4	N	7	1	1	2	2	4	1
3	NAIL2	N	F	N	S	2	N	8	1	1	2	2	2	1
4	NAIL3	N	F	N	S	2	N	9	1	1	2	2	2	1
5	NAIL4	N	F	N	S	2	N	10	1	1	2	2	2	1
6	NAIL5	N	F	N	S	2	N	11	1	1	2	2	2	1
7	NAIL6	N	U	N	S	5	N	12	1	4	2	2	5	1
8	NAIL7	N	U	N	S	3	N	13	1	4	2	2	3	1
9	NAIL8	N	U	N	S	3	N	14	1	4	2	2	3	1
10	SCREW1	Y	O	T	S	5	N	16	2	2	3	2	5	1
11	SCREW2	Y	R	L	S	4	N	17	2	3	1	2	4	1
12	SCREW3	Y	Y	L	S	4	N	18	2	5	1	2	4	1
13	SCREW4	Y	R	L	S	2	N	19	2	3	1	2	2	1
14	SCREW5	Y	Y	L	S	2	N	20	2	5	1	2	2	1
15	BOLT1	Y	R	L	F	4	N	1	2	3	1	1	4	1
16	BOLT2	Y	O	L	F	1	N	2	2	2	1	1	1	1
17	BOLT3	Y	Y	L	F	1	N	3	2	5	1	1	1	1
18	BOLT4	Y	Y	L	F	1	N	4	2	5	1	1	1	1
19	BOLT5	Y	Y	L	F	1	N	5	2	5	1	1	1	1
20	BOLT6	Y	Y	L	F	1	N	6	2	5	1	1	1	1
21	TACK1	N	F	N	S	1	Y	23	1	1	2	2	1	2
22	TACK2	N	F	N	S	1	Y	24	1	1	2	2	1	2
23	NAILB	N	F	N	S	1	Y	15	1	1	2	2	1	2
24	SCREWB	Y	O	L	S	1	Y	21	2	2	1	2	1	2

Figure 15 : *sélection multiple par Ctrl+Clic des variables numériques recodées***.**

Ensuite, il faut sélectionner à partir du menu <*Edit*>, la commande <*Copy*> (avec le clavier, faire un <*Ctrl+C*>), pour pouvoir coller (menu <*Edit*>, commande <*Paste*>, ou équivalent-clavier faire un <*Ctrl+V*>), après avoir effectué une sélection multiple de cinq colonnes vides :



hardwaren.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : object | TACK

	object	threadn	headn	indheadn	bottomn	brassn	lengthn	objectl	threadl	headl	indheadl	bottoml	brassl	lenghtl
1	TACK	1	1	2	2	1	1	1	1	1	2	2	1	1
2	NAIL1	1	1	2	2	1	4	2	1	1	2	2	1	4
3	NAIL2	1	1	2	2	1	2	3	1	1	2	2	1	2
4	NAIL3	1	1	2	2	1	2	4	1	1	2	2	1	2
5	NAIL4	1	1	2	2	1	2	5	1	1	2	2	1	2
6	NAIL5	1	1	2	2	1	2	6	1	1	2	2	1	2
7	NAIL6	1	4	2	2	1	5	7	1	4	2	2	1	5
8	NAIL7	1	4	2	2	1	3	8	1	4	2	2	1	3
9	NAIL8	1	4	2	2	1	3	9	1	4	2	2	1	3
10	SCREW	2	2	3	2	1	5	10	2	2	3	2	1	5
11	SCREW	2	3	1	2	1	4	11	2	3	1	2	1	4
12	SCREW	2	5	1	2	1	4	12	2	5	1	2	1	4
13	SCREW	2	3	1	2	1	2	13	2	3	1	2	1	2
14	SCREW	2	5	1	2	1	2	14	2	5	1	2	1	2
15	BOLT1	2	3	1	1	1	4	15	2	3	1	1	1	4
16	BOLT2	2	2	1	1	1	1	16	2	2	1	1	1	1
17	BOLT3	2	5	1	1	1	1	17	2	5	1	1	1	1
18	BOLT4	2	5	1	1	1	1	18	2	5	1	1	1	1
19	BOLT5	2	5	1	1	1	1	19	2	5	1	1	1	1
20	BOLT6	2	5	1	1	1	1	20	2	5	1	1	1	1
21	TACK1	1	1	2	2	2	1	21	1	1	2	2	2	1
22	TACK2	1	1	2	2	2	1	22	1	1	2	2	2	1
23	NAILB	1	1	2	2	2	1	23	1	1	2	2	2	1
24	SCREW	2	2	1	2	2	1	24	2	2	1	2	2	1

Figure 16 : *fichier des variables numériques, actives et illustratives***.**

Pour obtenir une analyse d'homogénéité, il faut sélectionner à partir du menu <*Analyse*>, la procédure <*Optimal Scaling*> du menu <*Data Reduction*>, en choisissant les options correspondantes (options par défaut de la procédure, soit un seul ensemble de variables avec toutes les variables considérées comme nominales) :

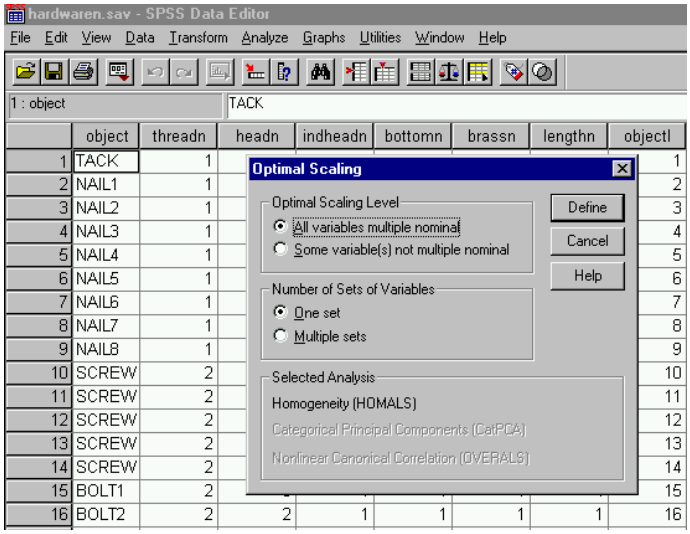


Figure 17 : *options correspondant à l'analyse d'homogénéité*

La première étape de la spécification de la procédure consiste à sélectionner les variables actives de l'analyse *(threadn*, *headn*, *indheadn*, *bottomn*, *brassn*, *lenghtn*) en définissant pour chacune d'entre-elles le nombre de modalités :

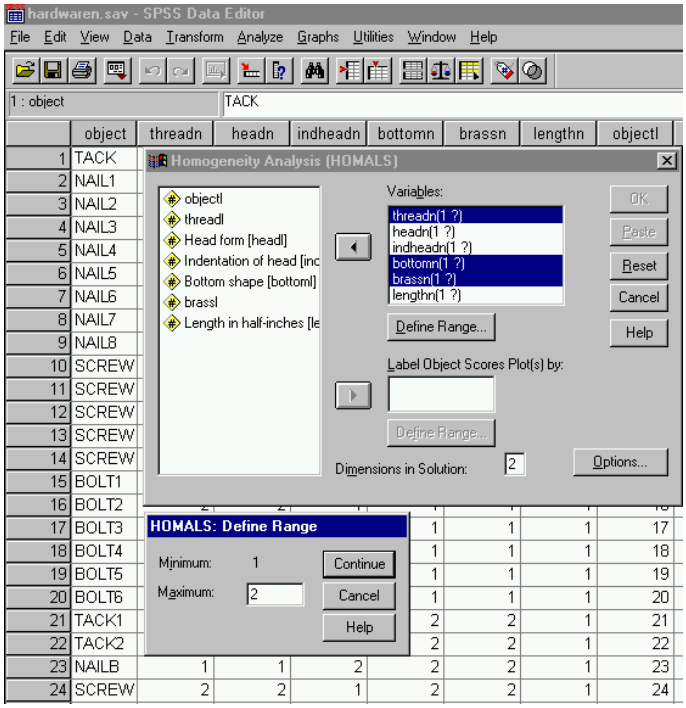


Figure 18 : *spécification des variables actives***.**

Dans la seconde étape, on spécifie les variables illustratives de l'analyse *(objectl, threadl*, *headl, brassl*, *lenghtl*) en définissant également pour chacune d'entre-elles le nombre de modalités :

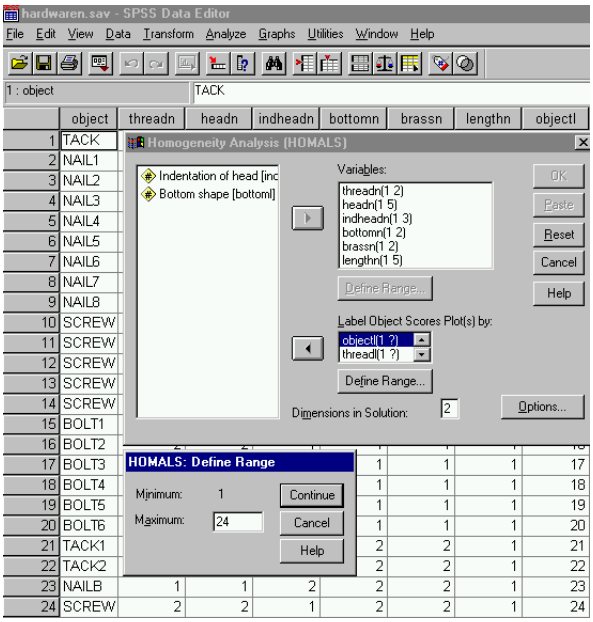


Figure 19 : *spécification des variables illustratives*

La dernière étape de cette spécification concerne le choix du nombre de dimensions (nombre d'axes factoriels) choisies pour la représentation graphique des objets, des modalités et des variables. On choisit ici une représentation graphique en deux dimensions comme solution particulière au problème d'optimisation sous contraintes que pose l'analyse formulée en terme d'homogénéité (cf. §3).
Les différentes options de traitement peuvent être choisies en utilisant le bouton <*Options…*>. Ces options portent sur les résultats (*Display*), les graphiques (*Plot*), la sauvegarde des coordonnées factorielles des objets (<*Save object scores*>) et les critères de contrôle de l'algorithme (*Criteria*).

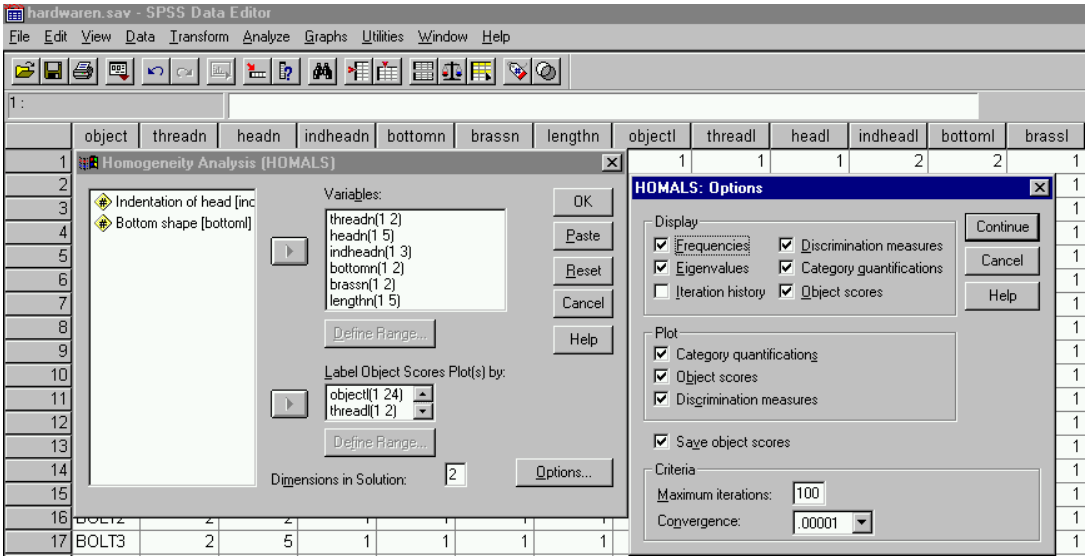
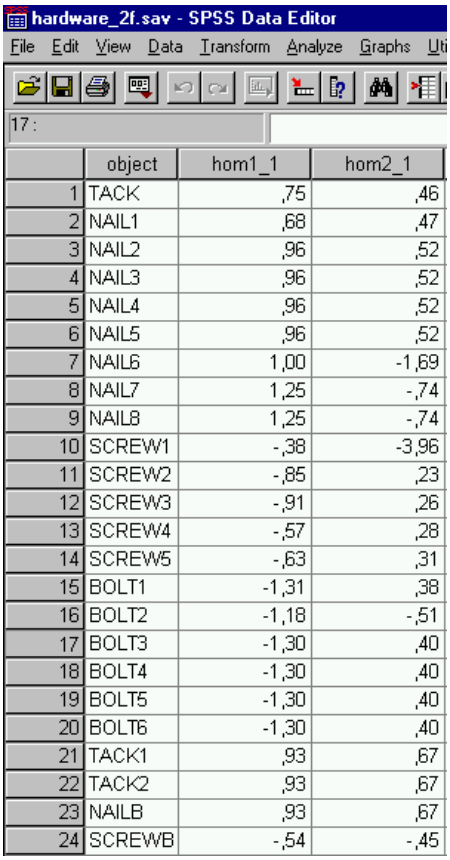


Figure 20 : *choix des options***.**

Les résultats demandés (cf. section *Display* de la figure 20) sont les distributions marginales obtenues par comptage (*Frequencies*), les valeurs propres (*Eigenvalues*), le pouvoir discriminant des variables actives (*Discrimination measures*), les coordonnées factorielles des modalités pour chaque variable (*Category quantifications*), les coordonnées factorielles des objets (*Object scores*).

Les graphiques demandés (cf. section *Plot* de la figure 20) sont le graphique factoriel des modalités de variables actives (*Category quantifications*), celui des objets (Object scores) et le diagramme du pouvoir discriminant des variables selon chacune des dimensions (*Discrimination measures*). A ces graphiques s'ajoutent autant de graphiques de densité des objets étiquetés par les modalités qu'il y a de variables illustratives.

La sauvegarde des coordonnées factorielles demandée (*Save object scores*) s'effectue dans le fichier d'origine, mais peut être ultérieurement sauvegardé dans un fichier spécifique, comme suit, pour de nouvelles analyses (classification sur axes factoriels) :


hardware_2f.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Ut

17 :

	object	hom1_1	hom2_1
1	TACK	,75	,46
2	NAIL1	,68	,47
3	NAIL2	,96	,52
4	NAIL3	,96	,52
5	NAIL4	,96	,52
6	NAIL5	,96	,52
7	NAIL6	1,00	-1,69
8	NAIL7	1,25	-,74
9	NAIL8	1,25	-,74
10	SCREW1	-,38	-3,96
11	SCREW2	-,85	,23
12	SCREW3	-,91	,26
13	SCREW4	-,57	,28
14	SCREW5	-,63	,31
15	BOLT1	-1,31	,38
16	BOLT2	-1,18	-,51
17	BOLT3	-1,30	,40
18	BOLT4	-1,30	,40
19	BOLT5	-1,30	,40
20	BOLT6	-1,30	,40
21	TACK1	,93	,67
22	TACK2	,93	,67
23	NAILB	,93	,67
24	SCREWB	-,54	-,45

Figure 21 : *sauvegarde des coordonnées factorielle des objets dans un fichier spécifique***.**

Les macro-instructions du programme *SPSS* correspondant aux options précédemment définies peuvent être sauvegardées dans un fichier de syntaxe en utilisant le bouton <*Paste*> de la boîte de dialogue :

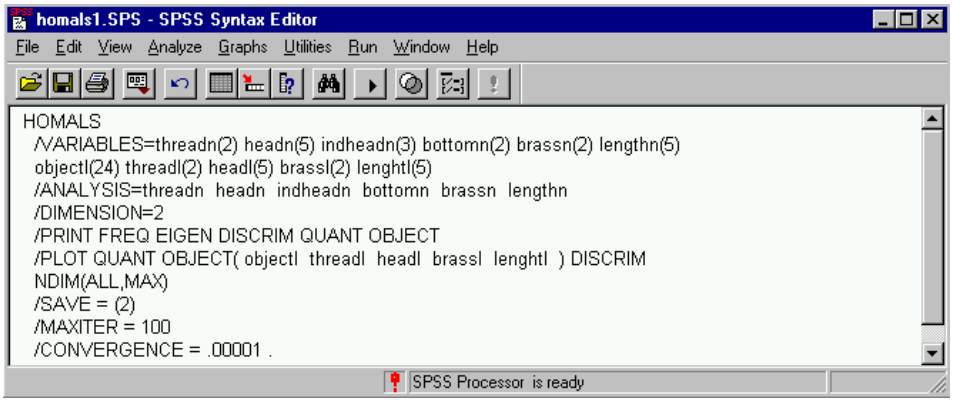


Figure 22 : *sauvegarde des macro-instructions dans un fichier programme (extension « .SPS »).*

Le seuil de convergence (*Convergence=.00001*) et le nombre maximum d'itérations (*Maximum interations=100*) permettent de contrôler l'algorithme itératif des moindres carrés alternés de la procédure *HOMALS* dans la recherche d'une solution.

Iteration History

Iteration	Fit	Difference from the Previous Iteration
1	,132757	,132757
2	,849876	,717119
3	,943649	,093773
4	,966800	,023151
5	,976822	,010022
6	,982110	,005288
7	,985104	,002993
8	,986838	,001735
9	,987851	,001013
10	,988444	,000593
11	,988793	,000349
12	,988999	,000206
13	,989122	,000123
14	,989196	,000074
15	,989241	,000045
16	,989269	,000028
17	,989287	,000018
18	,989298	,000012
19[a]	,989306	,000008

a. The iteration process stopped because the convergence test value was reached.

Tableau 4 : *historique des itérations*

Dans cet exemple, l'algorithme s'arrête à l'itération n° *19* car l'amélioration de l'indice d'ajustement (*Fit*) est devenue inférieure à la valeur du seuil de convergence.

5. L'algorithme itératif de la procédure *HOMALS de SPSS*[5]

L'algorithme itératif *HOMALS* (*Homogeneity Analysis by Means of Alternating Least Squares* – Analyse d'Homogénéité par Moindres Carrés Alternés) est la version moderne de la procédure proposée initialement par Guttman en 1941 pour l'analyse des données catégorielles. Le traitement des valeurs manquantes est basé sur l'introduction de pondérations nulles dans la fonction de perte (cf. De Leeuw & Van Rickevorsel, 1980). D'autres options pour le traitement des valeurs manquantes existent et sont basées sur le recodage (Gifi 1981, Meulman 1982).

5.1. Notations

En l'absence d'autre convention explicite, nous utilisons dans l'exposé de cet algorithme les notations suivantes :

- n nombre d'observations (ou **objets**)
- p nombre de variables (ou **critères**)
- s nombre de dimensions (ou **facteurs**)

Pour chaque critère j, $j = 1, \cdots, m$

- $\mathbf{h}_j$ vecteur $n \times 1$ des observations catégorielles
- k_j nombre de catégories (ou **modalités**)du critère j
- $\mathbf{Z}_j$ matrice $n \times k_j$ des **indicatrices de modalités** pour le critère j
- $z_{ik}^{(j)}$ élément matriciel de $\mathbf{G}_j = \begin{cases} 1 \text{ si l'observation } i \text{ appartient à la catégorie } k \text{ du critère } j \\ 0 \text{ sinon} \end{cases}$
- $\mathbf{O}_j$ matrice-filtre $n \times n$ des **indicatrices d'observations** pour le critère j
- $o_{ii}^{(j)}$ élément matriciel de $\mathbf{M}_j = \begin{cases} 1 \text{ si l'observation } i \text{ appartient à l'intervalle } [1, k_j] \\ 0 \text{ sinon} \end{cases}$
- $\mathbf{D}_j$ matrice diagonale des poids contenant les effectifs marginaux des modalités du critère j
- $\mathbf{D}$ matrice diagonale $\sum_j k_j \times \sum_j k_j$ des effectifs marginaux des modalités.

Les matrices de coordonnées factorielles sont :

- $\mathbf{X}$ matrice $n \times s$ des coordonnées factorielles des observations (**objets**) selon les s dimensions
- $\mathbf{Y}_j$ matrice $k_j \times s$ des coordonnées factorielles des modalités du critère j selon les s dimensions
- $\mathbf{Y}$ matrice concaténée $\sum_j k_j \times p$ des coordonnées factorielles de l'ensemble des modalités

[5] Cette section est une libre traduction du document technique correspondant fourni par SPSS

5.2. Formulation du programme d'optimisation de la fonction objectif

L'objectif d'HOMALS est de trouver une matrice $\mathbf{X}$ et un ensemble de matrices $\mathbf{Y}_j$ (pour $j = 1, \cdots, p$) tel que la fonction objectif :

$$\sigma(\mathbf{X}, \mathbf{Y}) = \frac{1}{p} \sum_j tr\left[\left(\mathbf{X} - \mathbf{Z}_j \mathbf{Y}_j \right)' \left(\mathbf{X} - \mathbf{Z}_j \mathbf{Y}_j \right) \right]$$

soit minimale sous la contrainte de normalisation $\mathbf{X}'\mathbf{O}_{\oplus}\mathbf{X} = np\mathbf{I}_s$,

où $\mathbf{O}_{\oplus} = \sum_j \mathbf{O}_j$ est la matrice-objet

et $\mathbf{I}_s$ est la $s \times s$ matrice identité.

L'introduction des matrices-filtres $\mathbf{O}_j$ permet de contrôler qu'aucune des valeurs observées actives pour le critère *j* ne sorte de l'intervalle $\left[1, k_j\right]$. La matrice-objet $\mathbf{O}_{\oplus}$ définit ainsi pour chaque objet *i* l'ensemble des observations actives de l'analyse ($o_{ii}^j = 1$) et l'ensemble des observations supplémentaires ($o_{ii}^j = 0$).

Les coordonnées factorielles de chaque objet sont centrées, ce qui peut s'écrire : $\mathbf{u}'\mathbf{O}_{\oplus}\mathbf{X} = 0$,

où $\mathbf{u}$ est le $n \times 1$-vecteur constant de composante scalaire égale à 1.

5.3. Algorithme itératif d'optimisation

Les principales étapes de l'algorithme d'optimisation sont les suivantes :

i) Initialisation ;
ii) Calcul des coordonnées factorielles des objets ;
iii) Orthonormalisation ;
iv) Calcul des coordonnées factorielles des modalités
v) Test de convergence : si oui, poursuivre ; si non, aller en ii) ;
vi) Rotation.

i) Initialisation

La matrice $\mathbf{X}$ des coordonnées factorielles est initialisée par tirage aléatoire sous contraintes de centrage ($\mathbf{u}'\mathbf{O}_{\oplus}\mathbf{X} = 0$) et de normalisation ($\mathbf{X}'\mathbf{O}_{\oplus}\mathbf{X} = np\mathbf{I}_s$). A partir de la matrice normalisée $\widetilde{\mathbf{X}}$, on obtient une première approximation des coordonnées factorielles des catégories du critère *j*, soit $\widetilde{\mathbf{Y}}_j = \mathbf{D}_j^{-1}\mathbf{G}'_j\widetilde{\mathbf{X}}$.

ii) Calcul des coordonnées factorielles des objets

Dans un premier temps, on définit, comme intermédiaire de calcul, une matrice $\mathbf{W}$ suivant :

$$\mathbf{W} \leftarrow \sum_j \mathbf{O}_j \mathbf{G}_j \widetilde{\mathbf{Y}}_j$$

Dans un second temps, on centre cette matrice par rapport à l'ensemble des objets actifs de l'analyse en prenant en compte le filtrage réalisé par la matrice-objet $\mathbf{O}_{\oplus}$:

$$\widetilde{\mathbf{W}} \leftarrow \left(\mathbf{O}_{\oplus} - \left[\mathbf{O}_{\oplus}\mathbf{u}\mathbf{u}'\mathbf{O}_{\oplus} / \mathbf{u}'\mathbf{O}_{\oplus}\mathbf{u}\right]\right)\mathbf{W}$$

Ces deux étapes conduisent à des solutions localement optimales si on n'applique pas de contraintes d'orthogonalité.

iii) Orthonormalisation

La procédure d'orthonormalisation consiste à trouver une matrice $\mathbf{X}^+$, $\mathbf{M}_\oplus$-orthonormale, qui soit la plus proche possible, au sens des moindres carrés, de la matrice $\tilde{\mathbf{W}}$. Cette matrice est obtenue en appliquant la procédure d'orthormalisation de Gram-Schmidt (procédure GRAM, repris de Björk et Golub, 1973), selon l'équation suivante :

$$\mathbf{X}^+ \leftarrow p^{1/2}\mathbf{M}_\oplus{}^{-1/2}\mathrm{GRAM}\left(\mathbf{M}_\oplus{}^{-1/2}\tilde{\mathbf{W}}\right)$$

ce qui, à une rotation près, conduit à la solution des moindres carrés.

iv) Calcul des coordonnées factorielles des modalités

Pour chaque critère *j*, on calcule la matrice $\mathbf{Y}_j{}^+$ des quantifications de ses modalités, comme suit :

$$\mathbf{Y}_j{}^+ = \mathbf{D}_j^{-1}\mathbf{G}'_j\tilde{\mathbf{X}}$$

v) Test de convergence

La différence $\left\{\sigma\left(\tilde{\mathbf{X}},\tilde{\mathbf{Y}}\right) - \sigma\left(\mathbf{X}^+,\mathbf{Y}^+\right)\right\}$ entre deux évaluations successives de la fonction objectif est comparée à la spécification ε du seuil de convergence, fournie par l'utilisateur. Les étapes ***ii)*** à ***iv)*** sont réitérées tant que la différence est supérieure au seuil de convergence fixé.

vi) Rotation

La fonction de perte $\sigma(\mathbf{X},\mathbf{Y})$ étant invariante par rotation simultanée de $\mathbf{X}$ et de $\mathbf{Y}$, la procédure itérative ne fournit pas nécessairement une orientation correcte pour les axes factoriels. En effet, du point de vue théorique, la solution en dimension *s* fournit les *s* premiers axes factoriels de la solution à *s+1* dimensions, ce que ne garantit pas cet algorithme itératif.
L'imbrication des différentes solutions est obtenue par extraction des vecteurs propres de la matrice

$$\frac{1}{p}\sum_j \mathbf{Y}'_j D_j Y_j$$

le calcul s'effectuant par la méthode de tridiagonalisation de Householder en utilisant l'algorithme *QL* proposé par [Wilkinson, 1965].

5.4. Diagnostics

***Rang maximum*[6]**

Le rang maximun $s_{\max}$ indique le nombre maximum de dimensions qui peuvent être extraites des données, soit :

$$s_{\max} = \min\left\{(n-1),\left(\left(\sum_j k_j\right) - \max(p_l, l)\right)\right\}$$

où m_l est le nombre de variables sans valeurs manquantes, k_j est le nombre de catégories distinctes du critère *j* et *n*, le nombre d'observations. Bien que le nombre de dimensions non-triviales puisse être inférieur à $s_{\max}$ lorsque $p = 2$, la procédure HOMALS permet de spécifier des cardinalités de dimension qui vont jusqu'à $s_{\max}$.

[6] Imprimé en guise d'avertissement lorsque la dimension de la solution demandée excède le rang de l'opérateur d'inertie.

Marges

Le tableau des sommes de colonne de la matrice $\mathbf{D}_j$ fournit directement les effectifs marginaux des modalités du critère *j*. La somme des éléments de la matrice $\mathbf{O}_j$ donne indirectement (en la soustrayant de *n*) le nombre de valeurs manquantes[7] pour les modalités de chaque critère *j*.

Pouvoir discriminant

Le pouvoir discriminant d'un critère *j* selon une dimension *s* est défini par :

$$\eta_{js}^2 = \frac{1}{n}\mathbf{y}_s^{(j)'}\mathbf{D}_j\mathbf{y}_s^{(j)}$$

Il est constitué par la variance de la projection du critère *j* selon la dimension *s*.

Compte tenu du fait que la trace est un opérateur invariant par changement de base, la somme des valeurs propres peut se calculer comme somme des pouvoirs discriminants sur l'ensemble des critères *j*, soit :

$$\sum_s \lambda_s = \frac{1}{p}\sum_j\sum_s \eta_{js}^2$$

La valeur minimale de la perte d'homogénéité $\sigma(\mathbf{X},\mathbf{Y})$ est égale à $s - \frac{1}{p}\sum_j\sum_s \eta_{js}^2$.

6. REFERENCES BIBLIOGRAPHIQUES

Benzécri J.-P. (1973) *L'analyse des données. Tome II L'analyse des correspondances*, Dunod, 632 p.

Bjö.rk A. et Golub G. H. (1973) « Numerical methods for computing angles between linear subspaces », *Mathematics of Computation*, 27: 579–594.

De Leeuw J. et Van Rijckevorsel, J. (1980) « HOMALS and PRINCALS—Some generalizations of principal components analysis », in: *Data Analysis and Informatics*, E. Diday et al, eds. Amsterdam: North-Holland.

Gillham N.W. (2001) *A Life of Sir Francis Galton : from African Exploration to the Birth of Eugenics*, Oxford University Press.

Gifi A. (1981) *Nonlinear multivariate analysis*, Leiden, Department of Data Theory.

Gifi A. (1990) *Nonlinear Multivariate Analysis*, Wiley, 579 p.

Guttman L. (1941) "The quantification of a class of attributes: A theory and method of scale construction". In: *The Prediction of Personal Adjustment*, P. Horst et al, eds. New York: Social Science Research Council.

Hartigan J.A. (1975) *Clustering Algorithms*. Wiley, New York, 351 p.

Lebart L (1975) « L'orientation du dépouillement de certaines enquêtes par l'analyse des correspondances multiples », *Consommation* , n°2, pp. 73-96, Dunod.

Meulman J. (1982). *Homogeneity analysis of incomplete data*, Leiden, DSWO Press.

[7] Ou encore exclues de l'analyse car les valeurs observées n'appartiennent pas à l'intervalle des catégories admises $\lfloor 1, k_j \rfloor$.

Meulman J.H., Heiser J.H. (2001) *SPSS Categories 11.0*, SPSS Inc., Chicago, 330 p.

Nishisato S. (1980).*Analysis of categorical data: Dual scaling and its application*. University of Toronto Press, Toronto.

SPSS (1994) *SPSS 6.1 Categories*, SPSS Inc., Chicago, 209 p.

Tenenhaus M., Young F.W. (1985) « An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis, and other methods for quantifying categorical multivariate data », *Psychometrika*, 50, pp. 91-119.

Wilkinson J. H. (1965) *The algebraic eigenvalue problem*, Oxford: Clarendon Press.

Portrait de Sir Francis Galton avec son maître d'hôtel Albert Gifi
Source : http://www.galton.org/photos

Annexe algébrique sur l'analyse de l'homogénéité[8]

A1 Introduction

A1.1 Le groupe Gifi

L'analyse de l'homogénéité constitue le paradigme conceptuel du système d'analyse multivarié non linéaire développé par le groupe Gifi. Albert Gifi est le nom collectif choisi par les membres du *Department of Data Theory* de l'Université de Leiden (Pays-Bas). Ce groupe, constitué autour de Jan de Leeuw a mis au point un système pour l'analyse multivariée non linéaire présenté dans l'ouvrage [Gifi, 1990]. La méthodologie développée par le groupe Gifi couvre un très large éventail de méthodes d'analyse exploratoire des données multivariées, principalement des techniques factorielles allant de l'analyse en composantes principales à l'analyse canonique.

A1.2 Le concept d'homogénéité

Le concept d'homogénéité auquel se réfère ces travaux formalise un des paradigmes fondateurs de la psychométrie selon lequel des critères différents peuvent mesurer une même caractéristique. Lorsque des variables distinctes (résultats aux tests, réponses aux questions, items choisis) semblent plus ou moins mesurer une même caractéristique, elles sont qualifiées d' « homogènes ».

A1.3 L'objectif de l'analyse d'homogénéité

Supposons que nous ayons rassemblé des données sur une population de n objets (individus, produits, régions, etc.) à partir de p critères j présentant un nombre fini de catégories selon lesquelles se distribuent les objets étudiés. L'objectif de l'analyse d'homogénéité est de représenter la structure que projette sur cette population (i.e. les profils de comportement) la batterie des critères d'observation utilisés, ceux-ci pouvant présentant des échelles de mesure différentes.

Les échelles de mesure utilisées par ces critères ou variables catégorielles j à k_j catégories peuvent être numériques (les catégories représentent des intervalles de mesure disjoints), ordinales (les catégories sont ordonnées) ou nominales (les catégories codent simplement l'appartenance à une classe).

L'objectif de l'analyse d'homogénéité est donc de représenter les objets étudiés et les critères d'étude dans un espace euclidien de faible dimension (représentation multivariée à s dimensions $s < p$) en prenant en compte les contraintes imposées par les différentes échelles de mesure utilisées. Cette représentation euclidienne constitue la **solution** du programme de maximisation de l'homogénéité associé à l'analyse d'homogénéité, $\boldsymbol{s}$ étant appelée la **dimension** de la solution.

A1.4 La méthode de représentation

Le choix de la méthode de représentation s'effectue par l'intermédiaire de l'optimisation d'une fonction-objectif mesurant l'homogénéité. Cette procédure d'optimisation permet de calculer des valeurs, *scores* et *quantifications*, utilisées pour construire une représentation géométrique dans un espace euclidien de faible dimension des relations, respectivement entre objets étudiés et entre catégories des critères d'observation.

[8] Cette annexe s'inspire très largement des ouvrages cités, en particulier de [Meulmann, 1982].

En théorie, les valeurs observées pour ces variables catégorielles distinctes mais homogènes peuvent être remplacées par la valeur unique d'une **variable synthétique x** .

A1.5 La mesure de l'homogénéité

Pour des variables catégorielles numériques, un changement d'échelle spécifique opéré par une transformation linéaire peut amener les valeurs de chaque critère à coïncider avec celles de la variable synthétique. Ces critères sont alors **homogènes**. Ce n'est pas toujours le cas, on peut alors utiliser des transformations non linéaires pour les rendre homogènes. Les critères étudiés sont alors **homogénéisables**.

En pratique, les batteries de critères étudiés ne sont pas toujours parfaitement homogénéisables. C'est souvent le cas lorsqu'elles comportent des variables ordinales voire nominales. On se contente alors d'une solution approchée pourvu que la perte d'information induite par l'agrégation des différents critères soit minimale.

Le défaut d'homogénéité peut être assimilé aux différences constatées entre les critères étudiés pour chacun des objets (**écarts internes aux objets**). Ces écarts internes aux objets doivent être distinguées des différences spécifiques entre objets constatées pour des critères homogènes (**écarts entre objets**). Une mesure possible de ce défaut d'homogénéité consiste à rapporter la mesure de ces différences internes (**somme des carrés des écarts internes aux objets**) à celle des différences spécifiques (**somme des carrés des écarts entre objets**) ou ce qui est équivalent au total des différences (**somme totale des carrés des écarts**).

En substituant une variable synthétique à la batterie de critères étudiés, on établit une relation d'équivalence entre la mesure de l'homogénéité imparfaite de ces variables catégorielles et la perte d'information liée à leur agrégation selon une échelle unique de catégories : maximiser l'homogénéité revient à minimiser la perte d'information. Pour une mesure normalisée de l'homogénéité sur un intervalle [0 ; 1], on peut formaliser cette relation d'équivalence par l'équation :

mesure d'homogénéité = 1 – perte d'information [1]

A1.6 Les principes de l'analyse d'homogénéité

A l'issue de cet exposé informel, récapitulons les principes qui constituent le fondement de l'analyse de l'homogénéité :

i) une batterie de critères d'observation numériques est dite homogène si toutes les variables qui la composent sont liés par une relation linéaire ; ces variables sont alors qualifiées d'homogènes ;

ii) une batterie de critères d'observation numériques est dite homogénéisable si elle peut-être rendue homogène au moyen de transformations portant sur ces variables numériques ;

iii) une batterie de critères formée de variables numériques, ordinales ou nominales est homogénéisable si toutes ces variables peuvent être transformées selon un processus de quantification susceptible de les rendre homogènes ;

iv) l'homogénéité d'un ensemble de variables centrées est appréciée à l'aune du rapport entre la somme des carrés des écarts entre objets (SCE_{inter}) et la somme des carrés des écarts totale (SCE_{total}); l'homogénéité parfaite correspond à la valeur 1 pour ce ratio, i.e. à une valeur nulle pour la somme des carrés des écarts interne aux objets (SCE_{intra}) ;

v) l'analyse d'homogénéité consiste à transformer les variables numériques ou à quantifier les variables ordinales ou nominales (en affectant une valeur numérique à chaque catégorie) pour maximiser la mesure de l'homogénéité.

Pour poursuivre l'analyse, il convient de donner une formulation plus précise à l'énoncé de ces principes en utilisant le cadre algébrique d'un espace vectoriel où objets et critères sont représentés par des vecteurs et leurs transformations sont représentées par des matrices.

A2 Analyse de l'homogénéité en dimension 1

A2.1 Concepts

A2.1.1 Le tableau des observations

Ainsi, le tableau des observations peut être représenté par une matrice **H** de données catégorielles concaténant les vecteurs $\mathbf{h}_j$, $j = 1,\cdots, p$, chaque vecteur $\mathbf{h}_j$ contenant les observations h_{ij} correspondant au critère ou **variable catégorielle *j*** sur l'**individu *i*** de la population des n objets observés :

$$\mathbf{H} = \left[\mathbf{h}_1,\cdots,\mathbf{h}_j,\cdots,\mathbf{h}_p\right] = \begin{bmatrix} h_{11} & \cdots & h_{1j} & \cdots & h_{1p} \\ \vdots & & \vdots & & \vdots \\ h_{i1} & \cdots & h_{ij} & \cdots & h_{ip} \\ \vdots & & \vdots & & \vdots \\ h_{n1} & \cdots & h_{nj} & \cdots & h_{np} \end{bmatrix}$$

Suivant l'équation [1], maximiser l'homogénéité revient donc à minimiser la perte d'information lorsque l'on remplace la batterie de critères $\{\mathbf{h}_1,\cdots,\mathbf{h}_j,\cdots,\mathbf{h}_p\}$ par une variable synthétique $\mathbf{x}$.

Tableau A1 : *le tableau des données catégorielles.*

id	threadn	headn	indheadn	bottomn	lengthn	brassn
1	1	1	2	2	1	1
2	1	1	2	2	4	1
3	1	1	2	2	2	1
4	1	1	2	2	2	1
5	1	1	2	2	2	1
6	1	1	2	2	2	1
7	1	4	2	2	5	1
8	1	4	2	2	3	1
9	1	4	2	2	3	1
10	2	2	3	2	5	1
11	2	3	1	2	4	1
12	2	5	1	2	4	1
13	2	3	1	2	2	1
14	2	5	1	2	2	1
15	2	3	1	1	4	1
16	2	2	1	1	1	1
17	2	5	1	1	1	1
18	2	5	1	1	1	1
19	2	5	1	1	1	1
20	2	5	1	1	1	1
21	1	1	2	2	1	2
22	1	1	2	2	1	2
23	1	1	2	2	1	2
24	2	2	1	2	1	2

Le tableau des données catégorielles ci-dessus code l'appartenance des n=24 ***objets*** *observés aux catégories de l'analyse pour les p =6* ***critères*** *d'observations retenus.*

A2.1.2 Une mesure de la perte d'information

L'élaboration d'une solution acceptable à ce problème de substitution passe en règle générale par la minimisation d'une fonction-objectif mesurant la perte d'information que l'on appelle le ***stress***, et que l'on notera $\sigma^2(\mathbf{x})$, définie par :

$$\sigma^2(\mathbf{x}) = \frac{1}{p}\sum_{j=1}^{p} \left\| \mathbf{x} - \mathbf{h}_j \right\|^2 \qquad [2]$$

où $\left\| \mathbf{x} - \mathbf{h}_j \right\|^2$ désigne une fonction quadratique des écarts (norme) entre le critère $\mathbf{h}_j$ et la variable synthétique $\mathbf{x}$.

Si cette fonction quadratique est la somme des carrés des écarts $\left\| \mathbf{x} - \mathbf{h}_j \right\|^2 = SCE\left(\mathbf{x} - \mathbf{h}_j \right)$, alors il s'agit du carré de la norme euclidienne usuelle notée $\left\| \mathbf{x} - \mathbf{h}_j \right\|_2^2$:

$$\sigma^2(\mathbf{x}) = \frac{1}{p}\sum_{j=1}^{p} SCE\left(\mathbf{x} - \mathbf{h}_j \right) = \frac{1}{p}\sum_{j=1}^{p} \left\| \mathbf{x} - \mathbf{h}_j \right\|_2^2$$

Si la norme utilisée est la racine carrée de la somme des carrés des écarts ($\left\|\mathbf{x} - \mathbf{h}_j\right\|_2$, norme euclidienne du vecteur $\mathbf{x} - \mathbf{h}_j$), l'optimum atteint par la fonction objectif est constitué par la moyenne arithmétique, notée $\overline{\mathbf{h}}$.

A2.1.3 Définition de la perte d'homogénéité

Reliant perte d'information et homogénéité sur la base de cette relation d'équivalence, le groupe Gifi a repris la méthode initialement proposée par Louis Guttman ([Guttman, 1941] pour la recherche d'un « score », échelle d'attitudes commune à une batterie de critères qualitatifs.

En synthétisant l'ensemble des observations effectuées sur les individus par une seule et même variable qui maximise l'homogénéité d'une batterie de critères, on peut introduire simultanément un opérateur de différenciation des individus en cherchant des transformations de critères (par exemple, en appliquant un système de poids $\{y_1, \cdots, y_j, \cdots, y_p\}$) qui permettent par substitution de la moyenne pondérée des variables catégorielles ainsi transformées d'obtenir des **scores individuels** (valeurs individuelles pour la variable synthétique $\mathbf{x}$ calculée après transformations) qui soient les plus différents possibles entre objets. En désignant par $\mathbf{y}$ le vecteur des poids $\{y_1, \cdots, y_j, \cdots, y_p\}$, on aboutit ainsi à un programme de minimisation d'une fonction de perte, notée $\sigma^2(\mathbf{x}, y)$, défini par :

$$\sigma^2(\mathbf{x}, \mathbf{y}) = \frac{1}{p}\sum_{j=1}^{p} \left\| \mathbf{x} - y_j\ \mathbf{h}_j \right\|_2^2 \qquad [3]$$

Les vecteurs $\mathbf{h}_j$ sont supposés centrés et le coefficient y_j permet d'effectuer une homothétie spécifique au vecteur $\mathbf{h}_j$.

Il en résulte une transformation linéaire du système de codage des catégories du critère *j*. La solution du programme de minimisation de cette fonction de perte est une moyenne pondérée des $\mathbf{h}_j$: $\tilde{\mathbf{h}}_w = \frac{1}{p}\sum_{j=1}^{p} w_j \mathbf{h}_j$ avec $w_j = \frac{y_j}{\sum_{j=1}^{p} y_j}$.

A2.2 Transformations linéaires

A2.2.1 Homogénéité et transformations linéaires

Les transformations linéaires des variables peuvent modifier à la fois les moyennes (par translation) et les variances (par homothétie). Dans un premier temps, travaillons avec des variables centrées (dont les valeurs sont des écarts à la moyenne) et procédons au moyen de pondération sur l'ensemble des variables. Soit t_j la transformation linéaire par pondération spécifique à la variable $\mathbf{h}_j$, telle que $t_j[\mathbf{h}_j] = y_j \mathbf{h}_j$ avec y_j la pondération affectée à la variable j. Les différences entre la variable synthétique x et les variables numériques $\mathbf{h}_j \quad j = 1, \cdots, p$ sont alors exprimées par la fonction de perte suivante :

$$\sigma^2(\mathbf{x};t) = \frac{1}{p}\sum_{j=1}^{p} SCE(\mathbf{x} - t_j[\mathbf{h}_j]) = \frac{1}{p}\sum_{j=1}^{p} \|\mathbf{x} - t_j[\mathbf{h}_j]\|_2^2 = \frac{1}{p}\sum_{j=1}^{p} \|\mathbf{x} - y_j\mathbf{h}_j\|_2^2 = \sigma^2(\mathbf{x};\mathbf{y})$$

Nous pouvons reformuler cette fonction de perte en l'écrivant sous forme d'une somme des produits scalaires des vecteur $(\mathbf{x} - y_j\mathbf{h}_j)$ par eux-mêmes :

$$\sigma^2(\mathbf{x},\mathbf{y}) = \frac{1}{p}\sum_{j=1}^{p} \|\mathbf{x} - y_j\mathbf{h}_j\|_2^2 = \frac{1}{p}\sum_{j=1}^{p} (\mathbf{x} - y_j\mathbf{h}_j)'(\mathbf{x} - y_j\mathbf{h}_j)$$

En développant cette expression et en remarquant qu'un scalaire est égal à son transposé ($\mathbf{x}'\mathbf{h}_j = \mathbf{h}_j'\mathbf{x}$), on peut écrire la fonction de perte sous forme matricielle si l'on définit la matrice $\mathbf{D}$ d'ordre $p \times p$ par la diagonale de la matrice $\mathbf{H}'\mathbf{H}$ ($\mathbf{D} = diag[\mathbf{H}'\mathbf{H}]$), où $\mathbf{H}$ figure la matrice des données d'ordre $n \times p$ et $\mathbf{y}$ le vecteur des poids :

$$\sigma^2(\mathbf{x};\mathbf{y}) = \mathbf{x}'\mathbf{x} + \frac{1}{p}\sum_{j=1}^{p} (y_j\mathbf{h}_j)'(\mathbf{h}_j y_j) - \frac{2}{p}\sum_{j=1}^{p} \mathbf{x}'(\mathbf{h}_j y_j) = \mathbf{x}'\mathbf{x} + \frac{1}{p}\mathbf{y}'\mathbf{D}\mathbf{y} - \frac{2}{p}\mathbf{x}'\mathbf{H}\mathbf{y}$$

En minimisant cette fonction, il convient d'imposer une contrainte sur la taille du vecteur $\mathbf{x}$ ou du vecteur $\mathbf{y}$ pour exclure la solution triviale où $\mathbf{x}$ et $\mathbf{y}$ sont nuls. Cette contrainte peut s'exprimer comme une standardisation des scores individuels $\mathbf{x}'\mathbf{x} = 1$. Une autre formulation possible de cette contrainte est la standardisation des variables transformées, soit $\mathbf{y}'\mathbf{D}\mathbf{y} = 1$. Les deux approches donnant le même résultat à un facteur d'échelle près, nous travaillerons avec la première contrainte portant sur le vecteur des scores individuels.

A2.2.2 Minimum sous contrainte de normalisation

En utilisant la technique des multiplicateurs de Lagrange, on peut déterminer le minimum de la fonction $\sigma^2(\mathbf{x};\mathbf{y})$ sous contrainte de normalisation $\mathbf{x}'\mathbf{x} = 1$: en notant μ le multiplicateur de Lagrange, cela revient à trouver le minimum de la fonction $f(\mathbf{x},\mathbf{y},\mu) = \sigma^2(\mathbf{x};\mathbf{y}) - \mu(\mathbf{x}'\mathbf{x} - 1)$. Les extrema sont obtenus en dérivant la fonction et en annulant ses dérivées partielles.
Soit :

$$\frac{\partial f(\mathbf{x},\mathbf{y},\mu)}{\partial \mathbf{x}} = 2\mathbf{x} - 2\mu\mathbf{x} - \frac{2}{p}\mathbf{H}\mathbf{y} = 0 \quad \text{et} \qquad \frac{\partial f(\mathbf{x},\mathbf{y},\mu)}{\partial \mathbf{y}} = \frac{2}{p}\mathbf{D}\mathbf{y} - \frac{2}{p}\mathbf{x}'\mathbf{H} = 0$$

($\mathbf{D}$ étant symétrique, on a $\frac{\partial \mathbf{y}'\mathbf{D}\mathbf{y}}{\partial \mathbf{y}} = 2\mathbf{D}\mathbf{y}$ et on remarquera que $\mathbf{D}\mathbf{y} = (\mathbf{D}\mathbf{y})' = \mathbf{y}'\mathbf{D}$).

On en tire les **équations normales**, respectivement : $\mathbf{H}\mathbf{y} = p(1-\mu)\mathbf{x}$ et $\mathbf{H}'\mathbf{x} = \mathbf{D}\mathbf{y}$

La matrice $\mathbf{D}$ étant diagonale, on en déduit la **transformation barycentrique** : $\mathbf{y} = \mathbf{D}^{-1}\mathbf{H}'\mathbf{x}$

Par substitution de $\mathbf{y}$ en combinant les équations précédentes, on aboutit à **l'équation aux valeurs extrémales** : $\mathbf{H}\mathbf{D}^{-1}\mathbf{H}'\mathbf{x} = p(1-\mu)\mathbf{x}$.

Montrons que la fonction de perte $\sigma^2(\mathbf{x};\mathbf{y})$ atteint son minimum pour la valeur de μ à l'extrémum.

En substituant $\mathbf{H}'\mathbf{x}$ à $\mathbf{D}\mathbf{y}$, on obtient :

$$\sigma^2(\mathbf{x};\mathbf{y}) = \mathbf{x}'\mathbf{x} + \frac{1}{p}\mathbf{y}'\mathbf{H}'\mathbf{x} - \frac{2}{p}\mathbf{x}'\mathbf{H}\mathbf{y} = \mathbf{x}'\mathbf{x} - \frac{1}{p}\mathbf{x}'\mathbf{H}\mathbf{y}\,.$$

En remplaçant $\mathbf{H}\mathbf{y}$ par $p(1-\mu)\mathbf{x}$, on en déduit :

$$\sigma^2(\mathbf{x};\mathbf{y}) = \mathbf{x}'\mathbf{x} - (1-\mu)\mathbf{x}'\mathbf{x} = \mu\,\mathbf{x}'\mathbf{x} = \mu\,.$$

La fonction de perte $\sigma^2(\mathbf{x};\mathbf{y})$ atteint donc son minimum pour la valeur de μ à l'extremum.

A2.2.3 Décomposition en valeur singulières

Soit la **décomposition en valeur singulières de rang *r***, éventuellement **complétée** (par *p-r* valeurs nulles) de la matrice $\mathbf{Z} = \mathbf{H}\mathbf{D}^{-1/2}$ d'ordre $n \times p$ définie par : $\mathbf{Z}\,U = V\,\Lambda^{1/2}$ où

$\mathbf{V}$ est une matrice d'ordre $n \times r$, orthonormale , i.e. telle que $\mathbf{V}'\mathbf{V} = \mathbf{I}_p$

$\mathbf{U}$ est une matrice orthonormale d'ordre $p \times r$, i.e. telle que $\mathbf{U}'\mathbf{U} = \mathbf{I}_p$

Λ est une matrice diagonale d'ordre $r \times r$ dont les r valeurs diagonales positives $\sqrt{\lambda_1},\cdots,\sqrt{\lambda_\rho},\cdots,\sqrt{\lambda_r}$ sont appelées **valeurs singulières**.

Commentaire [DD1] : I il s'agit de la décomposition en valeur singulières d'une matrice de rang *r*, « complétée » par des valeurs singulières nulles pour atteindre l'ordre *p* [décomposition en valeurs singulières « maigre » (DSV1, jusqu'au rang *r*) et décomposition en valeurs singulières pleine (DVS2, complétée jusqu'à l'ordre *p*, cf. Jean-François Durand, *Eléments de calcul matriciel et différentiel pour l'analyse factorielle des données*, Université de Montpellier II, polyalgmatcomp.pdf]

A2.2.4 Décomposition spectrale de la matrice des produits scalaire entre objets (analyse dans $\Re^n$)

Montrons que le vecteur des scores individuels $\mathbf{x}$ est le vecteur propre de la matrice $\mathbf{H}\mathbf{D}^{-1}\mathbf{H}'$ associé à sa plus grande valeur propre $\lambda_1 = p(1-\mu)$.

En utilisant l'orthonormalité de $\mathbf{U}$, on en déduit $\mathbf{Z} = \mathbf{V}\,\Lambda^{1/2}\,\mathbf{U}'$ et la décomposition spectrale de $\mathbf{H}\mathbf{D}^{-1}\mathbf{H}' = \mathbf{Z}\mathbf{Z}'$:

$$\mathbf{H}\mathbf{D}^{-1}\mathbf{H}' = \mathbf{Z}\mathbf{Z}' = \mathbf{V}\Lambda\mathbf{V}'$$

D'où une nouvelle formulation de l'équation aux valeurs extrémales

$$p(1-\mu)\mathbf{x} = \mathbf{H}\mathbf{D}^{-1}\mathbf{H}'\mathbf{x} = \mathbf{V}\Lambda\mathbf{V}'\mathbf{x} = \lambda\mathbf{x}$$

Si les valeurs singulières sont rangées par ordre de grandeur décroissant, la solution de cette équation minimisant $\sigma(\mathbf{x};\mathbf{y}) = \mu$ est donnée par λ_1 la plus grande valeur propre de l'opérateur symétrique $\mathbf{Z}\mathbf{Z}'$ et $\mathbf{x} = \mathbf{v}_1$ le vecteur propre associé (les colonnes de la matrice $\mathbf{V}$ sont vecteurs propres de $\mathbf{Z}\mathbf{Z}'$) :

$$\mathbf{H}\mathbf{D}^{-1}\mathbf{H}'\mathbf{x} = \lambda_1\,\mathbf{x}$$

A2.2.5 Scores individuels optimaux

Ainsi, le vecteur $\mathbf{x}$ des scores individuels optimaux est le vecteur propre de la matrice $\mathbf{HD}^{-1}\mathbf{H}'$ associé à la plus grande valeur propre $\lambda_1 = p(1-\mu)$.

La fonction de perte $\sigma(\mathbf{x};\mathbf{y})$ est minimisée pour le vecteur propre $\mathbf{x}$ correspondant à la plus grande valeur propre de la matrice $\mathbf{HD}^{-1}\mathbf{H}'$ et elle atteint donc son minimum en $\mu = 1 - \frac{\lambda_1}{p}$.

A2.2.6 Décomposition spectrale de la matrice des produits scalaires entre critères (analyse dans $\Re^p$)

En utilisant les équations normales $\mathbf{Hy} = p(1-\mu)\mathbf{x}$ et $\mathbf{H}'\mathbf{x} = \mathbf{Dy}$, on en déduit l'équation aux valeurs extrémales pour le vecteur des quantifications : $\mathbf{H}'\mathbf{Hy} = \lambda_1 \mathbf{Dy}$.

Les vecteurs colonnes de la matrice $\mathbf{U}$ sont vecteurs propres de la matrice $\mathbf{Z}'\mathbf{Z}$ des produits scalaires entre critères, soit pour le premier vecteur propre associé à λ_1 : $\mathbf{Z}'\mathbf{Z}\,\mathbf{u}_1 = \lambda_1 \mathbf{u}_1$.

Le vecteur $\mathbf{y}$ des **quantifications catégorielles** est donné par : $\mathbf{y} = \sqrt{\lambda_1}\mathbf{D}^{-1/2}\mathbf{u}_1$

Commentaire [DD2] :
$\mathbf{H}'\mathbf{Hy} = p(1-\mu)\mathbf{H}'\mathbf{x} = \lambda$

Commentaire [DD3] : on peut vérifier que ce vecteur $\mathbf{y}$ vérifie l'équation aux valeurs extrêmes :
$\mathbf{H}'\mathbf{Hy} = \mathbf{H}'\mathbf{H}\left(\sqrt{\lambda_1}\mathbf{D}^{-1/2}\mathbf{u}\right.$
$\sqrt{\lambda_1}\mathbf{D}^{1/2}\mathbf{Z}'\mathbf{Z}\mathbf{u}_1 = \sqrt{\lambda_1}\mathbf{D}^{1/}$

A2 .2.8 L'algorithme du centrage réciproque

Plutôt que de calculer la décomposition en valeurs singulières, l'analyse d'homogénéité utilise l'algorithme du centrage réciproque (*Reciprocal Averaging - RA*) déjà mentionné dans [Fisher, 1940]. Un tel algorithme, également appelé « moindres carrés alternés » (*Alternating Least Squares – ALS*), peut être vu comme un algorithme de la puissance itérée pour calculer la décomposition aux valeurs singulières ([Nishisato 1980] en donne une preuve). L'utilisation d'un tel algorithme était initialement justifiée par sa faible complexité en taille mémoire et son efficacité dans la recherche de la valeur propre maximale.

Pour minimiser la perte de pouvoir discriminant, il faut à chaque itération *l* calculer les quantifications catégorielles $\tilde{\mathbf{y}}$ comme moyenne des scores individuels initiaux $\mathbf{x}^{(0)}$ (arbitrairement choisis en première instance sous la condition $\mathbf{1}'\mathbf{x}^{(0)} = 0$), puis calculer les nouveaux scores $\tilde{\mathbf{x}}$ sur la base des quantifications catégorielles obtenues, enfin normaliser ces scores individuels ce qui termine l'itération.

Pour l'itération *l*, on a donc les étapes suivantes :

1 : $\tilde{\mathbf{y}} := \mathbf{D}^{-1}\mathbf{H}'\mathbf{x}^{(l)}$

2 : $\tilde{\mathbf{x}} := \frac{1}{p}\mathbf{H}\tilde{\mathbf{y}}$

3 : $\mathbf{x}^{(l+1)} := \tilde{\mathbf{x}}\left(\tilde{\mathbf{x}}'\tilde{\mathbf{x}}\right)^{-1/2}$, sous la condition $\mathbf{1}'\tilde{\mathbf{x}} = 0$

La réitération de ce cycle produit des scores $\tilde{\mathbf{x}}$ et des quantifications $\tilde{\mathbf{y}}$ qui, au bout d'un certain nombre d'itérations, ne se modifient plus de manière détectable. Ce couple de vecteurs stationnaires $(\tilde{\mathbf{x}}^*, \tilde{\mathbf{y}}^*)$ représente alors l'optimum recherché et la norme $\left\|\tilde{\mathbf{x}}^*\right\|_2 = \left(\tilde{\mathbf{x}}^{*\prime}\,\tilde{\mathbf{x}}^*\right)^{1/2}$ du

Commentaire [DD4] : Cet algorithme connu également sous le terme de *dual scaling* est signalé par [Saporta, 1990] sous le terme de « méthode des moyennes réciproques » (cf. p.215)

Commentaire [DD5] : L'algorithme de la puissance itérée est souvent utilisé en pratique pour rechercher la valeur propre dominante .

vecteur $\tilde{\mathbf{x}}^*$ stationnarisé nous fournit la plus grande valeur propre λ_1 et donc la valeur minimale de la fonction de perte de pouvoir discriminant.

A2.3 Transformations non-linéaires

A2.3.1 Extension aux transformations non -linéaires

Le concept d'**homogénéité** est étendu aux transformations non linéaires ([De Leeuw & Van Rijckevorsel, 1980]) de la manière suivante :

le vecteur $\mathbf{h}_j$ est dit ***homogène*** au score $\mathbf{x}$, vecteur unitaire (de norme 1) représentant la variable synthétique ciblée, si et seulement si après une transformation τ_j de normalisation spécifique à chaque critère j (c.à.d. telle que $\left\|\tau_j\left[\mathbf{h}_j\right]\right\|_2 = 1$) , on obtient l'égalité $\mathbf{x} = \tau_j\left[\mathbf{h}_j\right]$.

Le vecteur transformé $\tau_j\left[\mathbf{h}_j\right]$ constitue une **quantificatio**n du critère qualitatif j.

Si le vecteur h_j n'est pas homogène à $\mathbf{x}$, on définit la **perte d'homogénéité** comme une fonction quadratique des écarts au score, que nous pouvons écrire vectoriellement :

$$\sigma^2(\mathbf{x},t) = \frac{1}{p}\sum_{j=1}^{p}\left\|\mathbf{x} - \tau_j\left[\mathbf{h}_j\right]\right\|_2^2 = \frac{1}{p}\sum_{j=1}^{p}\left(\mathbf{x} - \tau_j\left[\mathbf{h}_j\right]\right)'\left(\mathbf{x} - \tau_j\left[\mathbf{h}_j\right]\right) \qquad [4]$$

en utilisant le produit du vecteur-colonne $\mathbf{x} - \tau_j\left[\mathbf{h}_j\right]$ par son transposé, le vecteur-ligne $\left(\mathbf{x} - \tau_j\left[\mathbf{h}_j\right]\right)'$

La minimisation de cette fonction objectif sur l'ensemble des critères analysés revient à rechercher un score $\mathbf{x}$ et des transformations non linéaires τ maximisant l'homogénéité de la batterie de critères proposés.

A2.3.2 Indépendance vis à vis du codage

Le passage des transformations linéaires aux transformations non linéaires s'effectue en fait par la recherche d'une quantification des catégories qui constitue une ***solution invariante***, c'est à dire indépendante du codage utilisé initialement. Cette indépendance vis à vis du codage est obtenue en analyse de l'homogénéité par l'intermédiaire des indicatrices de codage ([Guttman, 1941]).

Les **indicatrices de codage** sont des variables logiques indiquant pour chaque catégorie d'un critère qualitatif quels sont les objets lui appartenant :

si la variable vectorielle $\mathbf{h}_j$ codant le $j^{ème}$ critère à k_j catégories comporte n observations codées par un ensemble de catégories variant de *1* à k_j, on crée une **matrice indicatrice** $\mathbf{G}_j$ à n x k_j éléments g_{ik}^j définis par :

$g_{ik}^j = 1$ *si* $h_{ij} = k$ « l'objet *i* appartient à la catégorie *k* du critère *j* »

$g_{ik}^j = 0$ *si* $h_{ij} \neq k$ « l'objet *i* n'appartient pas à la catégorie *k* du critère *j* »

A2.3.3 Tableau disjonctif complet

Afin de pouvoir opérer sur l'ensemble des critères qualitatifs, on concatène les matrices indicatrices $\mathbf{G}_j$ dans un tableau booléen, matrice globale des indicatrices, notée $\mathbf{G}$:

$$\mathbf{G} = \left[\mathbf{G}_1 \middle| \cdots \middle| \mathbf{G}_j \middle| \cdots \middle| \mathbf{G}_p\right]$$

Tableau A2 : *tableau disjonctif complet codant les n=24 objets observés selon les catégories de l'analyse pour les p =6 critères d'observations retenus*

G	G_1		G_2					G_3			G_4		G_5					G_6		
i	thread1	thread2	head1	head2	head3	head4	head5	indhead1	indhead2	indhead3	bottom1	bottom2	lenght1	lenght2	lenght3	lenght4	lenght5	brass1	brass2	tot
1	1	0	1	0	0	0	0	0	1	0	0	1	1	0	0	0	0	1	0	6
2	1	0	1	0	0	0	0	0	1	0	0	1	0	0	0	1	0	1	0	6
3	1	0	1	0	0	0	0	0	1	0	0	1	0	1	0	0	0	1	0	6
4	1	0	1	0	0	0	0	0	1	0	0	1	0	1	0	0	0	1	0	6
5	1	0	1	0	0	0	0	0	1	0	0	1	0	1	0	0	0	1	0	6
6	1	0	1	0	0	0	0	0	1	0	0	1	0	1	0	0	0	1	0	6
7	1	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	1	1	0	6
8	1	0	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	6
9	1	0	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	6
10	0	1	0	1	0	0	0	0	0	1	0	1	0	0	0	0	1	1	0	6
11	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	1	0	6
12	0	1	0	0	0	0	1	1	0	0	0	1	0	0	0	1	0	1	0	6
13	0	1	0	0	1	0	0	1	0	0	0	1	0	1	0	0	0	1	0	6
14	0	1	0	0	0	0	1	1	0	0	0	1	0	1	0	0	0	1	0	6
15	0	1	0	0	1	0	0	1	0	0	1	0	0	0	0	1	0	1	0	6
16	0	1	0	1	0	0	0	1	0	0	1	0	1	0	0	0	0	1	0	6
17	0	1	0	0	0	0	1	1	0	0	1	0	1	0	0	0	0	1	0	6
18	0	1	0	0	0	0	1	1	0	0	1	0	1	0	0	0	0	1	0	6
19	0	1	0	0	0	0	1	1	0	0	1	0	1	0	0	0	0	1	0	6
20	0	1	0	0	0	0	1	1	0	0	1	0	1	0	0	0	0	1	0	6
21	1	0	1	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	1	6
22	1	0	1	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	1	6
23	1	0	1	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	1	6
24	0	1	0	1	0	0	0	1	0	0	0	1	1	0	0	0	0	0	1	6
tot	12	12	9	3	3	3	6	11	12	1	6	18	10	6	2	4	2	20	4	144

Ce tableau utilise un codage booléen de l'information qualifié de « disjonctif complet » dans la mesure où chaque individu se situe dans une catégorie unique (la marge ligne de chaque matrice indicatrice $\mathbf{G}_j$ est égale à *1*). La somme de chaque ligne du tableau G est égale au nombre de critères qualitatifs *p* et la somme de chacune de ses colonnes donne le poids de chacune des catégories. Le poids total du tableau est égal au produit du nombre d'objets observés par le nombre de critères d'observation : $n \times p$.

A2.3.4 Tableau de Burt

Le tableau de contingence généralisé $\mathbf{C} = \mathbf{G}'\mathbf{G}$ (ou encore tableau de Burt) croisant l'ensemble des critères contient la structure des inter-relations entre les catégories des différents critères. Il est imprimé ci-contre sous une forme triangulaire car ce tableau est symétrique et possède donc $\sum_{j=1}^{p} k_j$ lignes et $\sum_{j=1}^{p} k_j$ colonnes.

Bien qu'il soit utilisé dans les calculs de la procédure *HOMALS*, le tableau de Burt C n'est pas fourni par le logiciel *SPSS* .

A2.3.5 Blocs diagonaux : matrices de pondération

Les blocs diagonaux du tableau de Burt sont constitués par des matrices $\mathbf{D}_j = \mathbf{G}'_j\mathbf{G}_j$, issues du produit matriciel de $\mathbf{G}_j$ par son transposé $\mathbf{G}'_j$. Elles sont diagonales avec une diagonale constituée par la marge-colonne de $\mathbf{G}_j$ (donnant le nombre d'objets appartenant à chaque catégorie *k* du critère *j*). $\mathbf{D}_j$ est la matrice de pondération des effectifs marginaux des catégories du critère *j*.

Tableau A3 : blocs diagonaux du tableau de Burt.

D_1	thread1	thread2
thread1	12	0
thread2	0	12

D_2	head1	head2	head3	head4	head5
head1	9	0	0	0	0
head2	0	3	0	0	0
head3	0	0	3	0	0
head4	0	0	0	3	0
head5	0	0	0	0	6

D_3	idh1	idh2	idh3
idh1	11	0	0
idh2	0	12	0
idh3	0	0	1

Tableau A4 : *tableau de contingence généralisé ou tableau de Burt pour n=24 observations et p=6 critères.*

C	thread1	thread2	head1	head2	head3	head4	head5	indhead1	indhead2	indhead3	bottom1	bottom2	lenght1	lenght2	lenght3	lenght4	lenght5	brass1	brass2
thread1	12	0																	
thread2	0	12																	
head1	9	0	9	0	0	0	0												
head2	0	3	0	3	0	0	0												
head3	0	3	0	0	3	0	0												
head4	3	0	0	0	0	3	0												
head5	0	6	0	0	0	0	6												
indhead1	0	11	0	2	3	0	6	11	0	0									
indhead2	12	0	9	0	0	3	0	0	12	0									
indhead3	0	1	0	1	0	0	0	0	0	1									
bottom1	0	6	0	1	1	0	4	6	0	0	6	0							
bottom2	12	6	9	2	2	3	2	5	12	1	0	18							
lenght1	4	6	4	2	0	0	4	6	4	0	5	5	10	0	0	0	0		
lenght2	4	2	4	0	1	0	1	2	4	0	0	6	0	6	0	0	0		
lenght3	2	0	0	0	0	2	0	0	2	0	0	2	0	0	2	0	0		
lenght4	1	3	1	0	2	0	1	3	1	0	1	3	0	0	0	4	0		
lenght5	1	1	0	1	0	1	0	0	1	1	0	2	0	0	0	0	2		
brass1	9	11	6	2	3	3	6	10	9	1	6	14	6	6	2	4	2	20	0
brass2	3	1	3	1	0	0	0	1	3	0	0	4	4	0	0	0	0	0	4

En ne retenant que les blocs diagonaux $\mathbf{D}_j$ du tableau de Burt $\mathbf{C}$, on obtient une matrice de pondération $\mathbf{D}$, possédant également $\sum_{j=1}^{p} k_j$ lignes et $\sum_{j=1}^{p} k_j$ colonnes, qui constitue la matrice de pondération globale.

A2.3.6 Blocs non diagonaux : tableaux de contingence simples

Les blocs non diagonaux du tableau de Burt sont constitués par des matrices $\mathbf{C}_{jj'} = \mathbf{G}'_j \mathbf{G}_{j'}$, issues du produit matriciel de $\mathbf{G}_{j'}$ par le transposé $\mathbf{G}'_j$. Le tableau de contingence non diagonal $\mathbf{C}_{jj'}$ correspond au tri croisé des critères j et j'.

Tableau A5 : *bloc non-diagonal et sommation en ligne*

C_{42}	head1	head2	head3	head4	head5
bottom1	0	1	1	0	4
bottom2	9	2	2	3	2

→

D_4	bottom1	bottom2
bottom1	6	0
bottom2	0	18

La somme de la $k^{\text{ème}}$ ligne d'une matrice non diagonale $\mathbf{C}_{jj'}$ est égale au $k^{\text{ème}}$ élément diagonal de la matrice $\mathbf{D}_j$ quelque soit le critère de croisement j'.

Tableau A6 : *bloc non-diagonal et sommation en colonne*

D_1	thread1	thread2
thread1	12	0
thread2	0	12

↑

C_{21}	thread1	thread2
head1	9	0
head2	0	3
head3	0	3
head4	3	0
head5	0	6

De façon symétrique, la somme de la $k^{\text{ème}}$ colonne d'une matrice non diagonale (tri croisé) $\mathbf{C}_{jj'}$ est égale au $k^{\text{ème}}$ élément diagonal de la matrice $\mathbf{D}_{j'}$ quelque soit le critère de croisement j'.

A2.3.7 Projecteurs orthogonaux

Ultérieurement dans cet exposé, nous serons amené à utiliser le **projecteur orthogonal** $\mathbf{P}_j$ qui permet de projeter les objets dans le sous-espace engendré par les variables indicatrices du codage du critère *j* que sont les k_j vecteurs booléens de $\mathbf{G}_j$:

$$\mathbf{P}_j = \mathbf{G}_j \left(\mathbf{G}'_j \mathbf{G}_j\right)^{-1} \mathbf{G}'_j = \mathbf{G}_j \mathbf{D}_j^{-1} \mathbf{G}'_j$$

qui est une matrice d'ordre $n \times n$.

Ce projecteur est un opérateur symétrique ($\mathbf{P}_j = \mathbf{P}'_j$) et idempotent ($\mathbf{P}_j = \mathbf{P}_j{}^2$).

On peut en dériver une notion de **projecteur moyen**, noté $\mathbf{P}_0$, en effectuant la moyenne de ces opérateurs sur l'ensemble des critères qualitatifs : $\mathbf{P}_0 = \frac{1}{p}\sum_{j=1}^{p} \mathbf{P}_j$.

Commentaire [DD6] : $\mathbf{P}_0$ orthogonal suppose $\mathbf{P}_j \mathbf{P}_{j'} = 0 \qquad \forall\ j \neq j'$, ce qui n'est pas le cas.

A2.3.8 Discrétisation et scores induits

L'opération de discrétisation des variables revient à remplacer le vecteur transformé $\tau_j\left[\mathbf{h}_j\right]$ par le produit matriciel $\mathbf{G}_j \mathbf{y}_j$ où $\mathbf{y}_j$ est le vecteur des **quantifications** pour les k_j catégories du critère qualitatif *j*.

La fonction de perte s'écrit alors :

$$\sigma^2(\mathbf{x}, \mathbf{y}) = \frac{1}{p}\sum_{j=1}^{p} \left\| \mathbf{x} - \tau\left[\mathbf{h}_j\right] \right\|_2^2 = \frac{1}{p}\sum_{j=1}^{p} \left\| \mathbf{x} - \mathbf{G}_j \mathbf{y}_j \right\|_2^2$$

Le vecteur $\mathbf{q}_j = \mathbf{G}_j \mathbf{y}_j$ à *n* éléments contient les résultats numériques de la transformation du critère qualitatif *j* pour chacun des objets, ces éléments étant appelés les **scores induits**. Le vecteur **x** contient également *n* éléments caractérisant chacun des objets, appelés les **scores individuels**. La fonction de perte mesure alors le défaut d'ajustement entre les scores induits et les scores individuels.

A2.3.9 Optimisation globale de la fonction de perte

Le but de l'analyse d'homogénéité peut être formulé selon deux points de vue distincts :

- d'une part, remplacer les *p* vecteurs $\mathbf{q}_j$ par un vecteur unique $\mathbf{x}$, avec une perte d'homogénéité minimale. Idéalement, cela revient à choisir les vecteurs $\mathbf{y}_j$ tels que les vecteurs $\mathbf{q}_j$ soient tous identiques. Dans ce cas, le vecteur $\mathbf{x}$ des scores induits représente une échelle unidimensionnelle commune dont les *j* critères qualitatifs constituent des représentations homogènes ;
- d' autre part, en partant du vecteur **x** des scores individuels, l'objectif de minimisation de la perte d'homogénéité sera atteint si nous choisissons ces scores de façon à ce que tous les objets d'une même catégorie partagent le même score, ce qui implique que les *p* vecteurs $\mathbf{q}_j$ soient identiques.

Ainsi, le problème d'optimisation vu sous ces deux angles différents conduit à une solution où les scores individuels contenus dans $\mathbf{x}$ et les quantifications des catégories contenues dans les $\mathbf{y}_j$ soient parfaitement cohérentes, au sens suivant :

$$\mathbf{x} = \mathbf{G}_1 \mathbf{y}_1 = \cdots = \mathbf{G}_j \mathbf{y}_j = \cdots = \mathbf{G}_p \mathbf{y}_p$$

Ces deux points de vue conduisent à deux formulations distinctes du problème d'optimisation : la première formulation en termes de perte d'homogénéité, la seconde en termes de perte de pouvoir discriminant.

La **perte d'homogénéité** s'observe lorsqu'il n'existe pas de système de quantifications catégorielles $\{\mathbf{y}_1 \cdots \mathbf{y}_j \cdots \mathbf{y}_p\}$ tel que $\mathbf{x} = \mathbf{G}_1\mathbf{y}_1 = \cdots = \mathbf{G}_j\mathbf{y}_j = \cdots = \mathbf{G}_p\mathbf{y}_p$. Cette formulation suppose de partir des quantifications catégorielles $\mathbf{y}_j$ et de tester leur homogénéité en construisant un vecteur $\mathbf{x}$ des scores individuels.

La **perte de pouvoir discriminant** intervient lorsqu'il n'existe pas de vecteur de scores individuels $\mathbf{x}$ tel que $\mathbf{x} = \mathbf{G}_1\mathbf{y}_1 = \cdots = \mathbf{G}_j\mathbf{y}_j = \cdots = \mathbf{G}_p\mathbf{y}_p$. Cette formulation suppose de partir des scores individuels $\mathbf{x}$ et de tester leur pouvoir discriminant avec un système de quantifications catégorielles $\mathbf{y}_j$.

Compte-tenu des différentes échelles de mesure utilisées par l'ensemble des critères d'observation, un ajustement parfait des scores individuels et des quantifications catégorielles n'est pas réalisable en général : la résolution approchée du problème de représentation passe donc par la minimisation de la fonction globale de perte d'homogénéité définie pour les *p* critères. Si la norme est la somme des carrés des écarts, la fonction de perte globale s'écrit alors :

$$\sigma^2(\mathbf{x},\mathbf{y}) = \frac{1}{p}\sum_{j=1}^{p}\left\|\mathbf{x} - \mathbf{G}_j\mathbf{y}_j\right\|_2^2 = \frac{1}{p}\sum_{j=1}^{p}\left(\mathbf{x} - \mathbf{G}_j\mathbf{y}_j\right)'\left(\mathbf{x} - \mathbf{G}_j\mathbf{y}_j\right)$$

A2.3.10 Minimisation de la perte de pouvoir discriminant

Si l'on minimise la fonction de perte globale pour un vecteur de scores $\mathbf{x}$ donné relativement à un système inconnu $\mathbf{y}$ de quantifications, on est conduit d'après ce qui précède à définir la perte minimale de pouvoir discriminant par :

$$\sigma^2(\mathbf{x},*) = \min_{\mathbf{y}}\left\{\sigma^2(\mathbf{x},\mathbf{y})\middle|\mathbf{y}\right\}$$

En annulant la dérivée partielle par rapport à $\mathbf{y}$ et en résolvant vectoriellement le système d'équations normales, on trouve la solution suivante à ce problème d'optimisation :

$$\mathbf{y}_j = \left(\mathbf{G}_j'\mathbf{G}_j\right)^{-1}\mathbf{G}_j'\mathbf{x}$$

que l'on peut récrire sous la forme : $\mathbf{y}_j = \mathbf{D}_j^{-1}\mathbf{G}_j'\mathbf{x}$ ce qui indique clairement que la quantification catégorielle optimale $\mathbf{y}_j$ constitue une moyenne pondérée (par l'inverse des éléments diagonaux de la matrice diagonale $\mathbf{D}_j = \mathbf{G}_j'\mathbf{G}_j$) des scores individuels pour les objets appartenant aux catégories correspondantes du critère j (sur la base des valeurs de $\mathbf{G}_j'\mathbf{x}$).

En substituant la valeur optimale $\mathbf{y}_j$ à l'inconnue $\mathbf{y}$, on obtient :

$$\sigma^2(\mathbf{x},*) = \frac{1}{p}\sum_{j=1}^{p}\left\|\mathbf{x} - \mathbf{G}_j\left(\mathbf{G}_j'\mathbf{G}_j\right)^{-1}\mathbf{G}_j'\mathbf{x}\right\|_2^2$$

et en remarquant que $\mathbf{G}_j\left(\mathbf{G}_j'\mathbf{G}_j\right)^{-1}\mathbf{G}_j' = \mathbf{P}_j$, on aboutit à : $\sigma^2(\mathbf{x},*) = \frac{1}{p}\sum_{j=1}^{p}\left\|\mathbf{x} - \mathbf{P}_j\mathbf{x}\right\|_2^2$.

Développons cette expression :

$$\sigma^2(\mathbf{x},*) = \frac{1}{p}\sum_{j=1}^{p}\left(\mathbf{x} - \mathbf{P}_j\mathbf{x}\right)'\left(\mathbf{x} - \mathbf{P}_j\mathbf{x}\right) = \frac{1}{p}\sum_{j=1}^{p}\left(\mathbf{x}'\mathbf{x} + \mathbf{x}'\,\mathbf{P}_j'\mathbf{P}_j\mathbf{x} - 2\mathbf{x}'\mathbf{P}_j\mathbf{x}\right)$$

et utilisons le fait que $\mathbf{P}_j$ est symétrique et idempotent,

nous en tirons : $$\sigma^2(\mathbf{x},*) = \frac{1}{p}\sum_{j=1}^{p}\left(\mathbf{x}'\mathbf{x} - \mathbf{x}'\mathbf{P}_j\mathbf{x}\right) = \mathbf{x}'\mathbf{x} - \mathbf{x}'\mathbf{P}_0\mathbf{x} = \mathbf{x}'\mathbf{x}\left(1 - \frac{\mathbf{x}'\mathbf{P}_0\mathbf{x}}{\mathbf{x}'\mathbf{x}}\right)$$

Afin d'écarter les solutions triviales du problème d'optimisation

$$\sigma^2(*,*) = \min_{\mathbf{x}}\left\{\mathbf{x}'\mathbf{x} - \mathbf{x}'\mathbf{P}_0\mathbf{x}\right\}$$

qui revient à maximiser $\mathbf{x}'\mathbf{P}_0\mathbf{x}$

nous imposons des contraintes de normalisation, soit :

$\mathbf{1}'\mathbf{x} = 0$ (pour éviter la solution triviale $\mathbf{x} = \mathbf{1}$ et $\mathbf{y}_j = \mathbf{1}$)

et

$\mathbf{x}'\mathbf{x} = 1$ (pour éviter la solution triviale $\mathbf{x} = \mathbf{0}$)

où $\mathbf{1}$ est le vecteur dont toutes les composantes sont égales à 1 et $\mathbf{0}$ le vecteur dont toutes les composantes sont nulles.

La maximisation de $\mathbf{x}'\mathbf{P}_0\mathbf{x}$ sous la contrainte $\mathbf{x}'\mathbf{x} = 1$ équivaut à maximiser le rapport $\frac{\mathbf{x}'\mathbf{P}_0\mathbf{x}}{\mathbf{x}'\mathbf{x}}$.

L'ensemble des catégories d'un critère qualitatif induisant une partition en k_j groupes, la décomposition des sommes de carrés conduit à distinguer la somme des carrés inter-groupes de la somme des carrés intra-groupes, comme suit :

Somme des carrés inter-catégories : $SC_B = \mathbf{x}'\mathbf{G}_j\mathbf{D}_j^{-1}\mathbf{G}_j'\mathbf{x} = \mathbf{x}'\mathbf{P}_j\mathbf{x}$

Somme des carrés intra-catégories : $SC_W = \mathbf{x}'\left(\mathbf{I} - \mathbf{G}_j\mathbf{D}_j^{-1}\mathbf{G}_j'\mathbf{x}\right) = \mathbf{x}'\left(\mathbf{I} - \mathbf{P}_j\mathbf{x}\right)$

Somme des carrés totale : $SC_T = \mathbf{x}'\mathbf{x}$

La maximisation de $\mathbf{x}'\mathbf{P}_0\mathbf{x}$ sous la contrainte de normalisation $\mathbf{x}'\mathbf{x} = 1$ peut donc s'interpréter comme la maximisation du rapport de la variance inter-groupes à la variance totale.

La recherche d'un vecteur de scores individuels maximisant $\mathbf{x}'\mathbf{P}_0\mathbf{x}$ correspond ainsi à un objectif de discrimination globale des groupes d'objets induits par les catégories du critère *j*.

En utilisant des opérateurs de centrage $\mathbf{J}_M = \left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{\mathbf{1}'\mathbf{1}}\right)$ et $\mathbf{J}_D = \left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'\mathbf{D}}{\mathbf{1}'\mathbf{D}\mathbf{1}}\right)$, on montre ([Meulman, 1982]) que la fonction de perte de pouvoir discriminant peut s'écrire : $\sigma^2(\mathbf{x},*) = 1 - \mathbf{x}'\mathbf{Z}\mathbf{Z}'\mathbf{x}$

avec $\mathbf{Z} = p^{-1/2}\mathbf{J}_M{}'\,\mathbf{G}\,\mathbf{J}_D\,\mathbf{D}^{-1/2}$ opérateur réalisant la projection du vecteur objet sur les sous-espaces engendrés par les indicatrices du codage de l'ensemble des critères orthogonalement aux solutions triviales, que nous appellerons **projecteur-objet**.

Cette reformulation montre que la recherche d'une solution au problème de maximisation du ratio $\frac{\mathbf{x}'\mathbf{P}_0\mathbf{x}}{\mathbf{x}'\mathbf{x}}$ est équivalente au plan algébrique à un problème de recherche de vecteurs propres et de valeurs propres.

En effet, si $\mathbf{V}\Lambda\mathbf{V}'$ est la décomposition spectrale de la matrice réelle symétrique $\mathbf{Z}\mathbf{Z}'$, la solution $\mathbf{x}$ maximisant $\mathbf{x}'\mathbf{V}\Lambda\mathbf{V}'\mathbf{x}$ est le vecteur propre correspondant à la plus grande valeur propre de la matrice $\mathbf{Z}\mathbf{Z}'$.

Commentaire [DD7] : $\frac{\mathbf{1}\mathbf{1}'}{(\mathbf{1}'\mathbf{1})}$ on reconnaît l'expression d'un projecteur **I**-orthogonal sur le vecteur $\mathbf{1}$.

Commentaire [DD8] : $\frac{\mathbf{1}\mathbf{1}'\mathbf{I}}{(\mathbf{1}'\mathbf{D}}$, projecteur **D**-orthogonal sur $\mathbf{1}$

Commentaire [DD9] : Il s'agit d'un problème de maximisation du quotient de deux formes quadratiques : le rapport $\frac{\mathbf{x}'\mathbf{P}_0\mathbf{x}}{\mathbf{x}'\mathbf{I}\,\mathbf{x}'}$ est maximal pour le vecteur propre $\mathbf{v}_1$ de $\mathbf{I}^{-1}\,\mathbf{P}_0$ associé à sa plus grande valeur propre λ_1 (cf. [Saporta 90], p.484)

Le minimum de $\sigma^2(\mathbf{x},*)$ est donc égal à $1-\lambda_1$ où λ_1 est la valeur propre maximum de la matrice $\mathbf{ZZ}'$.

A2.3.11 Minimisation de la perte d'homogénéité

Si l'on minimise maintenant la fonction de perte globale pour un système $\mathbf{y}$ donné de quantifications catégorielles relativement à un vecteur de scores individuels $\mathbf{x}$ inconnu, nous exprimons alors le programme de minimisation de la perte d'homogénéité comme suit :

$$\sigma^2(*;\mathbf{y}) = \min\{\sigma(\mathbf{x};\mathbf{y})|\mathbf{x}\},$$

En dérivant désormais la fonction de perte par rapport à $\mathbf{x}$ et en résolvant le système d'équations normales, on trouve la solution à ce problème d'optimisation, $\mathbf{y}$ étant donné :

$$\mathbf{x} = \frac{1}{p}\sum_{j=1}^{p}\mathbf{G}_j y_j = \frac{1}{p}\mathbf{G}\mathbf{y}$$

Les scores objets optimaux sont constitués par la moyenne des quantifications des catégories correspondantes.

En substituant la solution optimale à l'équation définissant la fonction de perte globale, nous en déduisons :

$$\sigma^2(*;\mathbf{y}) = \frac{1}{p}\sum_{j=1}^{p}\left(\frac{1}{p}\mathbf{G}\mathbf{y}-\mathbf{G}_j\mathbf{y}_j\right)'\left(\frac{1}{p}\mathbf{G}\mathbf{y}-\mathbf{G}_j\mathbf{y}_j\right)$$

$$= \frac{1}{p}\sum_{j=1}^{p}\left(\frac{1}{p^2}\mathbf{y}'\mathbf{G}'\mathbf{G}\mathbf{y}-\mathbf{y}_j{}'\mathbf{G}_j{}'\mathbf{G}_j\mathbf{y}_j-2\frac{1}{p}\mathbf{y}'\mathbf{G}'\mathbf{G}_j\mathbf{y}_j\right)$$

en remplaçant $\mathbf{G}'\mathbf{G}$ par $\mathbf{C}$ et $\mathbf{G}_j'\mathbf{G}_j$ par $\mathbf{D}_j$, cela nous conduit à :

$$\sigma^2(*;\mathbf{y}) = \frac{1}{p}\sum_{j=1}^{p}\left(\frac{1}{p^2}\mathbf{y}'\mathbf{C}\mathbf{y}+\mathbf{y}_j{}'\mathbf{D}_j\mathbf{y}_j-2\frac{1}{p}\mathbf{y}'\mathbf{G}'\mathbf{G}_j\mathbf{y}_j\right)$$

$$= \frac{1}{p^2}\mathbf{y}'\mathbf{C}\mathbf{y}+\frac{1}{p}\mathbf{y}'\mathbf{D}\mathbf{y}-2\frac{1}{p^2}\mathbf{y}'\mathbf{G}'\mathbf{G}\mathbf{y} = \frac{1}{p}\mathbf{y}'\mathbf{D}\mathbf{y}-\frac{1}{p^2}\mathbf{y}'\mathbf{C}\mathbf{y}$$

$$= \frac{1}{p}\mathbf{y}'\mathbf{D}\mathbf{y}\left(1-\frac{\mathbf{y}'\mathbf{C}\mathbf{y}}{p\mathbf{y}'\mathbf{D}\mathbf{y}}\right)$$

Nous allons maintenant minimiser $\sigma^2(*;\mathbf{y})$ sur l'ensemble des $\mathbf{y}$ satisfaisant la condition $\mathbf{y}'\mathbf{D}\mathbf{y} = p$, ce qui revient à maximiser $\mathbf{y}'\mathbf{C}\mathbf{y}$.

Ce problème d'optimisation peut être interprété en termes d'analyse de la variance. Nous pouvons décomposer les scores induits $\mathbf{q}_j = \mathbf{G}_j\mathbf{y}_j$ en un composante $\mathbf{q}_{i\bullet}$ *inter-objets* et une composante $\mathbf{q}_{ij}-\mathbf{q}_{i\bullet}$ *intra-objets :*

Somme des carrés inter-objets : $SC_B = p\sum_{i=1}^{n}\mathbf{q}_{i\bullet}{}^2 = \mathbf{y}'\mathbf{C}\mathbf{y}$

Somme des carrés intra-objets : $SC_W = \sum_{i=1}^{n}\sum_{j=1}^{p}(\mathbf{q}_{ij}-\mathbf{q}_{i\bullet})^2 = \mathbf{y}'\left(\mathbf{D}-\frac{1}{p}\mathbf{C}\right)\mathbf{y}$

Somme des carrés totale : $SC_T = \sum_{i=1}^{n}\sum_{j=1}^{p}\mathbf{q}_{ij}{}^2 = \mathbf{y}'\mathbf{D}\mathbf{y}$

Ainsi, nous maximisons le ratio de la somme des carrés inter-objets sur la somme des carrés totale, sous la contrainte $\mathbf{y}'\mathbf{D}\mathbf{y} = p$. Cela s'interprète comme la recherche des quantifications catégorielles qui maximisent la somme des covariances pour les scores induits $\mathbf{q}_j$, tout en gardant la somme des variances constante. Il s'agit de minimiser la somme des carrés intra-objets, partant la perte d'homogénéité, et en conséquence de maximiser l'homogénéité du système de quantifications donné a priori pour les catégories des critères retenus dans l'analyse.
En utilisant la condition de normalisation $\mathbf{y}'\mathbf{D}\mathbf{y} = p$ et l'opérateur de projection orthogonale à la solution triviale $\mathbf{J}_D$, il vient :

$$\sigma^2(*;\mathbf{y}) = 1 - \frac{1}{p^2}\mathbf{y}'\mathbf{J}_D{}'\mathbf{C}\mathbf{J}_D\mathbf{y}$$

Si l'on exprime la perte d'homogénéité en fonction de $\mathbf{u} = p^{-1/2}\,\mathbf{D}^{1/2}\mathbf{y}$, on trouve alors :

$$\sigma^2(*;\mathbf{u}) = 1 - \frac{1}{p}\mathbf{u}'\mathbf{D}^{-1/2}\mathbf{J}_D{}'\mathbf{C}\mathbf{J}_D\mathbf{D}^{-1/2}\,\mathbf{u}$$

En utilisant le projecteur-objet $\mathbf{Z}$, on montre ([Meulman, 1982]) que la fonction de perte d'homogénéité peut s'écrire : $\sigma(*,\mathbf{u}) = 1 - \mathbf{u}'\mathbf{Z}'\mathbf{Z}\mathbf{u}$
La valeur minimale de $\sigma(*,\mathbf{a})$ est donc : $1 - \lambda_1(\mathbf{Z}'\mathbf{Z})$
où λ_1 est la plus grande valeur propre de la matrice $\mathbf{Z}'\mathbf{Z}$ (également valeur propre maximale de la matrice $\mathbf{Z}\mathbf{Z}'$) associé au vecteur propre , ce qui nous conduit, à un facteur d'échelle près, à une solution algébrique homologue du problème de minimisation de la perte de pouvoir discriminant.

A2.3.12 L'algorithme du centrage réciproque
Plutôt que de calculer la décomposition en valeurs singulières, l'analyse d'homogénéité utilise l'algorithme du centrage réciproque (*Reciprocal Averaging - RA*) déjà mentionné dans [Fisher, 1940]. Un tel algorithme, également appelé « moindres carrés alternés » (*Alternating Least Squares – ALS*), peut être vu comme un algorithme de la puissance itérée pour calculer la décomposition aux valeurs singulières ([Nishisato 1980] en donne une preuve). L'utilisation d'un tel algorithme était initialement justifiée par sa faible complexité en taille mémoire et son efficacité dans la recherche de la valeur propre maximale.
Pour minimiser la perte de pouvoir discriminant, il faut à chaque itération *l* calculer les quantifications catégorielles $\tilde{\mathbf{y}}$ comme moyenne des scores individuels appropriés (arbitrairement choisis en première instance), puis calculer les nouveaux scores $\tilde{\mathbf{x}}$ sur la base des quantifications catégorielles obtenues, enfin normaliser ces scores individuels ce qui termine l'itération.
Pour l'itération *l*, on a donc les étapes suivantes :

1 : $\tilde{\mathbf{y}} := \mathbf{D}^{-1}\mathbf{G}\mathbf{x}^{(l)}$

2 : $\tilde{\mathbf{x}} := \frac{1}{p}\mathbf{G}\tilde{\mathbf{y}}$

3 : $\mathbf{x}^{(l+1)} := \tilde{\mathbf{x}}(\tilde{\mathbf{x}}'\tilde{\mathbf{x}})^{-1/2}$, sous la condition $\mathbf{1}'\mathbf{x} = 0$

La réitération de ce cycle produit des scores $\tilde{\mathbf{x}}$ et des quantifications $\tilde{\mathbf{y}}$ qui, au bout d'un certain nombre d'itérations, ne se modifient plus de manière détectable. Ce couple de vecteurs stationnaires $(\mathbf{x}^*, \mathbf{y}^*)$ représente alors l'optimum recherché et la norme $(\tilde{\mathbf{x}}'\tilde{\mathbf{x}})^{1/2}$ du vecteur $\tilde{\mathbf{x}}$

Commentaire [DD10] : Cet algorithme connu également sous le terme de *dual scaling* est signalé par [Saporta, 1990] sous le terme de « méthode des moyennes réciproques » (cf. p.215)

Commentaire [DD11] : L'algorithme de la puissance itérée est souvent utilisé en pratique pour rechercher la valeur propre dominante .

stationnarisé nous fournit la plus grande valeur propre λ_1 et donc la valeur minimale de la fonction de perte de pouvoir discriminant.
Pour la minimisation de la perte d'homogénéité, ce sont exactement les mêmes étapes qui alternent :

1 : $\quad \widetilde{\mathbf{x}} := \frac{1}{p}\mathbf{G}\mathbf{y}^{(l)}$

2 : $\quad \widetilde{\mathbf{y}} := \mathbf{D}^{-1}\mathbf{G}\widetilde{\mathbf{x}}$

mais cette fois ci la normalisation porte sur $\mathbf{y}$ au moyen de la transformation :

3 : $\quad \mathbf{y}^{(l+1)} := \widetilde{\mathbf{y}}\left(\widetilde{\mathbf{y}}'\,\mathbf{D}\,\widetilde{\mathbf{y}}\right)^{-1/2}$, sous la condition $\mathbf{1}'\,\mathbf{y} = 0$

Le schéma d'alternance duale des transformations opérées dans l'analyse d'homogénéité est issu de ces deux formulations distinctes du problème d'optimisation : la première formulation en termes de perte d'homogénéité ; la seconde en termes de perte de pouvoir discriminant. Il n'est cependant pas possible pour des raisons tenant à la géométrie de la solution d'obtenir une normalisation simultanée des deux vecteurs $(\mathbf{x}^*, \mathbf{y}^*)$ optimaux.

Commentaire [DD12] : cf. les relations pseudo-barycentriques.

A2.3.13 L'algorithme des moindres carrés alternés

L'algorithme itératif des moindres carrés alternés (**MCA**) impliqué dans la procédure *HOMALS* utilise la première version de cette méthode pour converger vers une solution stationnaire $(\mathbf{x}^*, \mathbf{y}^*)$ qui minimise la perte globale d'homogénéité :

$$\sigma^2(\mathbf{x},\mathbf{y}) = \frac{1}{p}\sum_{j=1}^{p}\left(\mathbf{x}-\mathbf{G}_j\mathbf{y}_j\right)'\left(\mathbf{x}-\mathbf{G}_j\mathbf{y}_j\right)$$

sous la contrainte de normalisation $\mathbf{x}\,\mathbf{x}' = 1$.

Si la recherche des solutions doit s'effectuer en théorie sous la contrainte $\mathbf{x}\,\mathbf{x}' = 1$, elle s'effectue en pratique sous la contrainte $\mathbf{x}\,\mathbf{x}' = n$, en utilisant la normalisation $\widetilde{\mathbf{x}}^{(l)} := \sqrt{n}\,\left(\widetilde{\mathbf{x}}'\,\widetilde{\mathbf{x}}\right)^{-1/2}\,\widetilde{\mathbf{x}}$.

L'implantation de l'algorithme MCA sous *HOMALS* se déroule donc selon les étapes suivantes :

0) *initialisation* :
   - *a.* tirage aléatoire du vecteur initial $\widetilde{\mathbf{x}}^{(0)}$ dans une loi uniforme de moyenne nulle et de variance *n*
   - *b.* calcul du vecteur initial $\quad \widetilde{\mathbf{y}}^{(0)} := \mathbf{D}^{-1}\mathbf{G}'\,\widetilde{\mathbf{x}}^{(0)}$

1) *calcul du vecteur des scores* : $\quad \widetilde{\mathbf{x}} \leftarrow \frac{1}{p}\,\mathbf{G}\,\widetilde{y}^{(l-1)}$

2) *normalisation (*de variance *n)* $\quad \widetilde{\mathbf{x}}^{(l)} := \sqrt{n}\,\left(\widetilde{\mathbf{x}}'\,\widetilde{\mathbf{x}}\right)^{-1/2}\,\widetilde{\mathbf{x}}$

3) *calcul du vecteur des quantification :* $\quad \widetilde{\mathbf{y}}^{(l)} := \mathbf{D}^{-1}\mathbf{G}'\,\widetilde{\mathbf{x}}^{(l)}$

4) *test de convergence :* $\quad \left\|\widetilde{\mathbf{x}}^{(l)} - \widetilde{\mathbf{x}}^{(l-1)}\right\| \le \varepsilon$ et/ou $\left\|\widetilde{\mathbf{y}}^{(l)} - \widetilde{\mathbf{y}}^{(l-1)}\right\| \le \varepsilon$

L'algorithme converge vers la solution stationnaire $(\mathbf{x}^*, \mathbf{y}^*)$:

$$\mathbf{x}^* = \mathbf{v}_1$$
$$\mathbf{y}^* = \sqrt{\lambda_1}\,\mathbf{D}^{-1/2}\,\mathbf{u}_1$$

En raison de la transformation $\tilde{\mathbf{y}}^{(l)} := \mathbf{D}^{-1}\mathbf{G}'\,\tilde{\mathbf{x}}^{(l)}$ appliquée à chaque itération, cette solution satisfait la « condition de stationnarité » exprimée par l'équation $\mathbf{D}_j\mathbf{y}_j = \mathbf{G}_j\mathbf{x}$.

A3 Analyse multivariée de l'homogénéité

A3.1 Représentation en plusieurs dimensions

Compte-tenu de la complexité des inter-relations entre objets et variables catégorielles intervenant dans certains phénomènes, un ajustement des scores individuels et des quantifications catégorielles réalisé selon une seule dimension peut apparaître comme insuffisant pour rendre compte des phénomènes observés. Une solution plus satisfaisante du problème de représentation peut alors être fournie par l'utilisation d'une image euclidienne à plusieurs dimensions. La recherche des différentes dimensions de la solution est réalisée en effectuant successivement l'analyse de l'homogénéité en dimension 1 orthogonalement aux solutions trouvées précédemment. Ainsi, on recherchera une variable synthétique $\mathbf{x}_2$, orthogonale à la première direction identifiée par la variable synthétique $\mathbf{x}_1$, correspondant à la plus grande valeur propre λ_1. Il convient alors de rechercher le minimum de la fonction de perte $\sigma(\mathbf{x}_2, \mathbf{y}_2)$ sous contrainte d'orthogonalité $\langle \mathbf{x}_1 | \mathbf{x}_2 \rangle = 0$, ce que nous exprimons vectoriellement par $\mathbf{x}_1'\mathbf{x}_2 = 0$. Pour rechercher une troisième variable synthétique $\mathbf{x}_3$, on minimisera la fonction de perte $\sigma(\mathbf{x}_3, \mathbf{y}_3)$ sous les contraintes $\mathbf{x}_1'\mathbf{x}_3 = 0$ et $\mathbf{x}_2'\mathbf{x}_3 = 0$.

Cette recherche peut s'effectuer en utilisant les solutions fournies par la décomposition en valeurs singulières (DVS) car la matrice $\mathbf{V}$ est une matrice orthogonale ($\mathbf{V}'\mathbf{V} = \mathbf{I}_p$). Les vecteurs-colonnes de $\mathbf{V}$ satisfont les contraintes d'orthogonalité et sont, comme vecteurs propres, solutions successives des différentes étapes du programme d'optimisation que l'on peut mener jusqu'au rang *r* de la matrice $\mathbf{Z}$ (DVS « stricte ») ou jusqu'à *p* solutions incluant les vecteurs propres correspondant aux valeurs propres nulles $\mathbf{ZZ}'$ (DVS « étendue »).

A3.2 Analyse de l'homogénéité selon plusieurs dimensions

Cet ajustement, que nous avons effectué jusqu'ici en utilisant un sous-espace de dimension 1 (le vecteur $\mathbf{x}$ des scores), peut être réalisé sans perte de généralité dans un sous-espace à *s* dimensions, $1 < s \leq d$, en utilisant une **matrice $\mathbf{X}$ des scores individuels** (coordonnées des *n* objets selon les *s* dimensions) à $n \times s$ coefficients et une matrice $\mathbf{Y}$ à $p \times s$ coefficients appelée **matrice des quantifications catégorielles** (coordonnées des *p* critères *j* selon les *s* dimensions).

En formant à partir des *s* couples solutions $(\mathbf{x}_j ; \mathbf{y}_j)$ le couple de matrices $(\mathbf{X}; \mathbf{Y})$, on définit une approximation d'ordre *s* de la matrice $\mathbf{Z}$ par : $\mathbf{Z} = \mathbf{HD}^{-1/2} = \mathbf{V}\,\Lambda^{1/2}\,\mathbf{U}' \cong \mathbf{XY}'\mathbf{D}^{1/2}$.

Commentaire [DD13] : $\mathbf{Z} =$. Cette approximation est-elle la meilleure approximation de rang *s* au sens du critère des moindres carrés (Théorème d'Eckart-Young) ?

Les vecteurs-colonnes de la matrice $\mathbf{V}$ forment une base orthonormée de l'espace vectoriel généré par les critères étudiés, représentés par les vecteurs-colonnes de $\mathbf{H}$. Les *s* premiers vecteurs-colonnes de $\mathbf{V}$ constituant la matrice $\mathbf{X}$ forment une base orthonormée de l'espace vectoriel généré par les vecteurs-colonnes de $\mathbf{Y}$ issus des transformations linéaires appliquées aux critères étudiés.

On réalise ainsi une **analyse de l'homogénéité en dimension *s***.

A3.3 Qualité de l'ajustement

La qualité de l'ajustement réalisé par l'approximation d'ordre *s* peut être mesurée en utilisant la norme de Frobenius ($\|\mathbf{A}\|_F^2 = \sum_{i,j} |a_{ij}|^2$) comme norme matricielle. On peut alors comparer $\mathbf{Z} = \mathbf{H}\mathbf{D}^{-1/2}$ à son approximation d'ordre *s*, $\mathbf{X}\mathbf{Y}'\mathbf{D}^{1/2}$ en formant le rapport de leurs normes respectives :

$$\|\mathbf{Z}\|_F^2 = tr(\mathbf{Z}\mathbf{Z}') = tr(\mathbf{H}\mathbf{D}^{-1}\mathbf{H}') = tr(\mathbf{V}\,\mathbf{\Lambda}\,\mathbf{V}') = \sum_{j=1}^{p} \lambda_j$$

et

$$\left\|\mathbf{X}\mathbf{Y}'\mathbf{D}^{1/2}\right\|_F^2 = tr(\mathbf{X}\,\mathbf{Y}'\,\mathbf{D}\,\mathbf{Y}\,\mathbf{X}') = tr(\mathbf{Y}'\,\mathbf{D}\,\mathbf{Y}\,) = \sum_{j=1}^{s} \lambda_j \ .$$

Les vecteurs-colonnes de $\mathbf{Z} = \mathbf{H}\mathbf{D}^{-1/2}$ étant standardisés ($\mathbf{D} = diag[\mathbf{H}'\,\mathbf{H}]$), on en conclut que : $\|\mathbf{Z}\|_F^2 = \sum_{j=1}^{p} \lambda_j = p$.

A3.4 Opérateurs de transformation

La matrice $\mathbf{G}_j$, d'ordre $n \times k_j$, regroupant les vecteurs booléens associés aux indicatrices de codage du critère qualitatif j à k_j modalités, la transformation t_j du vecteur $\mathbf{h}_j$ peut se définir matriciellement par $t_j(\mathbf{h}_j) = \mathbf{G}_j\,\mathbf{Y}_j$ où $\mathbf{Y}_j$ est une matrice à $k_j \times s$ coefficients appelée **matrice des quantifications catégorielles** (coordonnées des catégories du critère *j* selon les *s* dimensions).

A3.5 Formulation globale de la fonction de perte multivariée

L'analyse multivariée de l'homogénéité consiste alors à minimiser une fonction globale de perte multivariée, s'exprimant en fonction des matrices $\mathbf{Y}_j$ et $\mathbf{X}$ en étendant la définition de la fonction de perte globale univariée (cf. supra partie I):

$$\sigma^2(\mathbf{x},\mathbf{y}) = \frac{1}{p}\sum_{j=1}^{p} \left\|\mathbf{x} - \mathbf{G}_j\mathbf{y}_j\right\|_2^2 = \frac{1}{p}\sum_{j=1}^{p} \left(\mathbf{x} - \mathbf{G}_j\mathbf{y}_j\right)'\left(\mathbf{x} - \mathbf{G}_j\mathbf{y}_j\right)$$

En utilisant la norme de Frobenius comme extension de la norme euclidiennne à $\Re^{n\times s}$, on définit une norme quadratique pour la matrice $(\mathbf{X} - \mathbf{G}_j\mathbf{Y}_j)$ d'ordre $n \times s$ qui permet de spécifier la fonction de perte globale multivariée comme une fonction quadratique

$$\sigma^2(\mathbf{X},\mathbf{Y}) = \frac{1}{p}\sum_{j=1}^{p} \left\|\left(\mathbf{X} - \mathbf{G}_j\mathbf{Y}_j\right)\right\|_F^2 = \frac{1}{p}\sum_{j=1}^{p} trace\left[\left(\mathbf{X} - \mathbf{G}_j\mathbf{Y}_j\right)'\left(\mathbf{X} - \mathbf{G}_j\mathbf{Y}_j\right)\right]$$

à minimiser sous les contraintes

d'ortho-normalisation $\mathbf{X}'\mathbf{X} = n\,I_s$

de centrage $\mathbf{1}'\mathbf{X} = \mathbf{0}$,

permettant d'éviter les solutions triviales dans la résolution de ce système d'équations correspondant à $\mathbf{X} = \mathbf{0}$ et $\mathbf{Y}_j = \mathbf{0}$ pour chacun des critères *j*.

L'opérateur *trace* utilisé pour la matrice $\left(\mathbf{X} - \mathbf{G}_j\mathbf{Y}_j\right)'\left(\mathbf{X} - \mathbf{G}_j\mathbf{Y}_j\right)$ d'ordre $s \times s$, défini comme la somme des éléments diagonaux d'une matrice carré symétrique, à l'avantage d'être

invariant par changement de base orthonormée, ce qui nous assure une solution indépendante du système de représentation choisi à un déplacement près (translation et/ou rotation).

A3.6 Algorithme matriciel des moindres carrés alternés
La solution à ce problème d'optimisation est fournie par l'algorithme des moindres carrés alternés (*ALS – Alternating Least Squares*) qui consiste à minimiser alternativement la fonction de perte conditionnellement aux matrices $\mathbf{Y}_j$ et $\mathbf{X}$:

Initialisation : tirer une matrice aléatoire $\widetilde{\mathbf{X}}^{(l=0)} := \mathbf{X}_0$ telle que $\mathbf{1}'\mathbf{X}_0 = \mathbf{0}$ et $\mathbf{X}_0'\mathbf{X}_0 = n\,\mathbf{I}_s$

Itération *l* :

Etape *l.1* : minimiser la fonction de perte conditionnellement à $\mathbf{Y}_j$ pour $\mathbf{X}$ fixé. avec la transformation $\widetilde{\mathbf{Y}}_j := \mathbf{D}_j^{-1}\mathbf{G}_j{}'\widetilde{\mathbf{X}}^{(l)} \qquad j \in J$ [5]
où $\mathbf{D}_j = \mathbf{G}_j'\mathbf{G}_j$ est la matrice diagonale d'ordre $k_j \times k_j$ contenant les effectifs marginaux des catégories du critère j.

Etape *l.2* : minimiser la fonction de perte conditionnellement à $\mathbf{X}$ en ayant fixé les matrices $\widetilde{\mathbf{Y}}_j$ avec la transformation : $\mathbf{Z} := \frac{1}{p}\sum_{j=1}^{P}\mathbf{G}_j\widetilde{\mathbf{Y}}_j$

Etape *l.3* : centrer la matrice $\mathbf{Z}$ avec la transformation $\widetilde{\mathbf{Z}} := \mathbf{Z} - \frac{1}{n}\mathbf{1}(\mathbf{1}'\mathbf{Z})$

Etape *l.4* : orthonormaliser la matrice $\widetilde{\mathbf{Z}}$ avec la transformation de Gram-Schmidt $\widetilde{\mathbf{X}}^{(l+1)} := \sqrt{n}\;GRAM\left(\widetilde{\mathbf{Z}}\right)$

Test : tester la stationnarité de la solution $\left\|\widetilde{\mathbf{X}}^{(l+1)} - \widetilde{\mathbf{X}}^{(l)}\right\| \le \varepsilon$ **?**

Les étapes 2 à 4 sont réitérées jusqu'à ce que la matrice $\mathbf{X}$ satisfasse au test de stationnarité, indiquant qu'un optimum est atteint.
L'algorithme converge ainsi vers un couple de matrices stationnaires qui est la solution fournie par la procédure *HOMALS* pour le problème d'optimisation défini précédemment.

A3.7 Invariance de la solution par rotation
Si l'on utilise une base différente dans l'espace des colonnes de la matrice $\mathbf{X}$ issue de la solution originale par rotation $\mathbf{R}$ (transformation telle que $\mathbf{R}'\mathbf{R} = \mathbf{R}\,\mathbf{R}' = \mathbf{I}_d$), alors les matrices issues de cette transformation $\breve{\mathbf{X}} = \mathbf{X}\,\mathbf{R}$ et $\breve{\mathbf{Y}}_j = \mathbf{D}_j{}^{-1}\mathbf{G}_j'\breve{\mathbf{X}}$ sont également optimales pour la fonction de perte d'homogénéité.

A3.8 Normalisation
Les colonnes de la matrice des scores individuels $\mathbf{X}$ sont centrées : soustraction de la moyenne $\bar{x}_{.d}$ à chaque valeur x_{id}, soit matriciellement $\widetilde{\mathbf{X}} = \mathbf{X} - \frac{1}{n}\mathbf{1}(\mathbf{1}'\mathbf{X})$. Puis la matrice $\widetilde{\mathbf{X}}$ est orthonormalisée par la procédure d'orthogonalisation de Gram-Schmidt stabilisée (*MGS - Modified Gram-Schmidt*) ou par une factorisation *QR*.

En raison de ces exigences modestes en ressources de calcul, la procédure d'orthogonalisation de Gram-Schmidt stabilisée est en règle générale utilisée par les programmes implantant la procédure *HOMALS*. Cependant, si l'orthonormalité est requise de manière critique dans la construction de la solution optimale, il est préférable de recourir à la méthode *QR* de factorisation, méthode « moderne » de résolution des problèmes de moindres carrés où *Q* désigne une matrice orthogonale et *R* une matrice triangulaire inférieure.

En fixant $\mathbf{X}^* = \sqrt{n}\widetilde{\mathbf{X}}$, l'étape de normalisation conduit à $\mathbf{X}^{*'}\mathbf{X}^* = n\,\mathbf{I}_s$.

A3.9 Décomposition en valeurs singulières

Le problème d'optimisation de la fonction multivariée de perte d'homogénéité :

$$\sigma^2(\mathbf{X},\mathbf{Y}) = \frac{1}{p}\sum_{j=1}^{p} tr\left[\left(\mathbf{X}-\mathbf{G}_j\mathbf{Y}_j\right)'\left(\mathbf{X}-\mathbf{G}_j\mathbf{Y}_j\right)\right]$$

sous les contraintes

d'ortho-normalisation $\mathbf{X}'\mathbf{X} = n\,I_d$

de centrage $\mathbf{1}'\mathbf{X} = \mathbf{0}$,

peut être assimilé à un problème de décomposition en valeurs singulières. En effet, on peut montrer que la matrice $\mathbf{X}^*$ est constituée par les vecteurs singuliers à gauche de la matrice $p^{-1/2}\left(\mathbf{I}-\frac{\mathbf{1}\mathbf{1}'}{n}\right)\mathbf{G}\,\mathbf{D}^{-1/2}$, où $\mathbf{G} = \left[\mathbf{G}_1\middle|\cdots\middle|\mathbf{G}_j\middle|\cdots\middle|\mathbf{G}_p\right]$ est le tableau disjonctif complet corrigé.

La décomposition complète en valeurs singulières possède $d = \sum_{j=1}^{p} k_j - p$ dimensions.

Après extraction des scores individuels, on calcule les quantifications catégorielles en utilisant l'équation : $\mathbf{Y}_j = \mathbf{D}_j^{-1}\mathbf{G}_j'\mathbf{X}$.

L'avantage de l'algorithme des moindres carrés alternés est de pouvoir travailler seulement sur les *s* premières dimensions requises en supposant que *s* soit très petit devant *d* ($s << d$), minimisant les besoins en taille mémoire pour améliorer ainsi l'efficacité algorithmique assez médiocre des moindres carrés alternés.

Logo du Data Theory Group (Université de Leiden, Faculté des Sciences Sociales et du Comportement, Prof. Jacqueline Meulmann):

Source : http://www.socialsciences.leidenuniv.nl/educationandchildstudies/datatheory/