ANNEXE 1

Quelques éléments de statistique multidimensionnelle

Les méthodes d'analyse statistique exploratoire utilisées au cours des chapitres précédents visent à mettre en forme de vastes ensembles de données, à en dégager des structures et à valider ces structures. Elles relèvent de la statistique exploratoire multidimensionnelle, de l'analyse des données, ou encore du *Data Mining*, ces trois désignations étant à peu près équivalentes dans le cadre des utilisations de cet ouvrage. Nous avons utilisé à leur propos l'expression *statistique structurale* pour marquer l'importance accordée à la phase de validation des structures. Ces méthodes généralisent la statistique descriptive classique et utilisent des outils mathématiques assez intuitifs, mais plus complexes que les moyennes, variances et coefficients de corrélations empiriques de la statistique descriptive.

Sont présentés dans cette annexe les principes des techniques utilisées ou évoquées dans les chapitres précédents, l'analyse en composantes principales étant la technique d'analyse factorielle de base des applications sémiométriques. Certains développements de l'ouvrage noté [SEM 2006][1] seront repris ; ils seront complétés par des travaux plus récents sur les méthodes de validation, et en particulier sur les techniques

1. *Statistique Exploratoire Multidimensionnelle*, [Visualisation et inférence en fouille de données] , 4ème ed. L.Lebart, M. Piron, A. Morineau. Dunod, 2006.

dites de *bootstrap*, sur les cartes de Kohonen, ou sur des techniques d'analyse moins utilisées comme l'analyse logarithmique.

A1.1 Rappel des principes des méthodes exploratoires multidimensionnelles

Les méthodes exploratoires multidimensionnelles recouvrent un grand nombre de techniques qui ont pour objectif de décrire et synthétiser l'information contenue dans de vastes tableaux de données.

A1.1.1 Représentation géométrique et nuages de points

Au départ, les données se présentent sous forme de grands tableaux rectangulaires, notés **X**. Les lignes ($i=1,...,n$) du tableau représentent les n individus, les sujets enquêtés par exemple, et les colonnes ($j=1,...m$) les m variables qui peuvent être des mesures, des caractéristiques ou encore des notes relevées sur les individus.

Afin de comprendre le principe des méthodes de statistique exploratoire multidimensionnelle, il est utile de représenter de façon géométrique l'ensemble des n individus (n lignes) et l'ensemble des m variables (m colonnes) comme deux *nuages de points*, chacun des deux ensembles étant décrit par l'autre. On définit alors, pour les deux nuages, des distances entre les points-lignes et entre points-colonnes qui traduisent les associations statistiques entre les individus (lignes) et entre les variables (colonnes).

Tableau A1.1 :
Exemple de tableau X de notes (de 1 à 7)
attribuées à : m = **7 mots, par** n = **12 répondants**

mots ***Répondants***	**arbre**	**cadeau**	**danger**	**morale**	**orage**	**politesse**	**sensuel**
R01	7	4	2	2	3	1	6
R02	6	3	1	2	4	1	7
R03	4	5	3	4	3	4	3
R04	5	5	1	7	2	7	1
R05	4	5	2	7	1	6	2
R06	5	7	1	5	2	6	5
R07	4	2	1	3	5	3	6
R08	4	1	5	4	5	4	7
R09	6	6	2	4	7	5	5
R10	6	6	3	5	3	6	6
R11	7	7	6	7	7	6	7
R12	2	2	1	2	1	3	2

Dans le cas de la sémiométrie, un mot (variable) est un point dont les coordonnées sont les notes données par les *n* individus (répondants) : le nuage des *m* mots se situe dans un espace à *n* dimensions. De même, un individu est un point dont les coordonnées sont les notes attribuées aux *m* mots ; le nuage des *n* individus se trouve dans un espace à *m* dimensions.

Les figures A1.1 et A1.2 illustrent, à partir du tableau A1.1 contenant les notes attribuées à 7 mots par 12 répondants, la représentation de ces deux nuages de points intrinsèquement liés.

Le nuage des points-mots est construit dans l'espace des individus, ici à partir seulement de deux individus, R04 et R08, car deux dimensions rendent possible un graphique dans un plan (*cf.* figure A1.1).

	arbre	**cadeau**	**danger**	**morale**	**orage**	**politesse**	**sensuel**
R01	7	4	2	2	3	1	6
R02	6	3	1	2	4	1	7
R03	4	5	3	4	3	4	3
R04	5	5	1	7	2	7	1
R05	4	5	2	7	1	6	2
R06	5	7	1	5	2	6	5
R07	4	2	1	3	5	3	6
R08	4	1	5	4	5	4	7
R09	6	6	2	4	7	5	5
R10	6	6	3	5	3	6	6
R11	7	7	6	7	7	6	7
R12	2	2	1	2	1	3	2

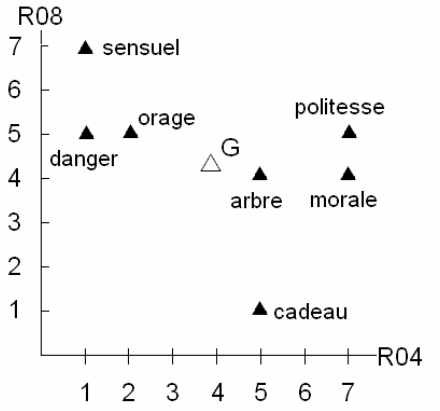


Figure A1.1 : Représentation du nuage des mots dans l'espace des deux répondants « R04 » et « R08 »

De la même façon, le nuage des 12 répondants est construit dans l'espace des variables, ici à partir de deux mots, *Morale* et *Sensuel*, c'est-à-dire dans un espace de deux dimensions (*cf.* figure A1.2).

Pour chacun des nuages est représenté le *point moyen* appelé aussi *centre de gravité*. Il s'agit de G pour le centre de gravité des notes attribuées par les répondants (*cf.* figure A1.1) et de G' pour celui des répondants ayant notés les deux mots retenus.

	arbre	**cadeau**	**danger**	**morale**	**orage**	**politesse**	**sensuel**
R01	7	4	2	2	3	1	6
R02	6	3	1	2	4	1	7
R03	4	5	3	4	3	4	3
R04	5	5	1	7	2	7	1
R05	4	5	2	7	1	6	2
R06	5	7	1	5	2	6	5
R07	4	2	1	3	5	3	6
R08	4	1	5	4	5	4	7
R09	6	6	2	4	7	5	5
R10	6	6	3	5	3	6	6
R11	7	7	6	7	7	6	7
R12	2	2	1	2	1	3	2

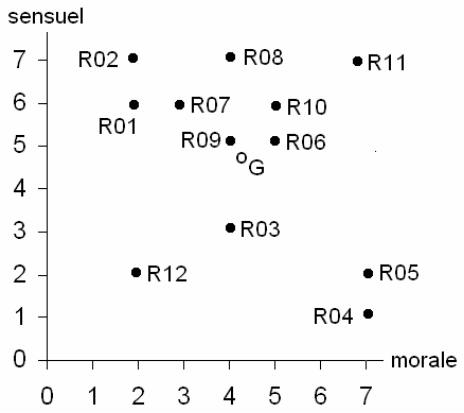


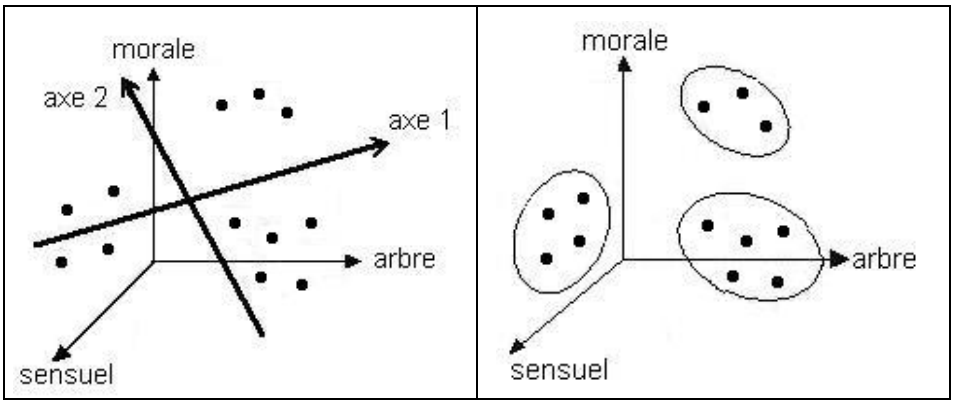
Figure A1.2 : Représentation du nuages des répondants dans l'espace des mots « S*ensuel* » et « *Morale* »

A1.1.2 Principe et méthodes d'analyse

S'il est toujours possible de calculer des distances entre les lignes et des distances entre les colonnes d'un tableau **X**, il n'est pas possible de les visualiser de façon immédiate (les représentations géométriques associées impliquant en général des espaces à plus de deux ou trois

dimensions) : il est nécessaire de procéder à des transformations et des approximations pour en obtenir une représentation plane.

Les tableaux de distances associés à ces représentations géométriques (simples dans leur principe, mais complexes en raison du grand nombre de dimensions des espaces concernés) peuvent être décrits par les deux grandes familles de méthodes que sont les méthodes factorielles et la classification. La première consiste à rechercher les directions principales selon lesquelles les points s'écartent le plus du point moyen. La seconde consiste à rechercher des groupes ou classes d'individus qui soient les plus homogènes possibles (figure A1.3).



Méthode factorielle (recherche des directions principales) *Méthode de classification (recherche de groupes homogènes)*

Figure A1.3 : Deux grandes familles de méthodes

Ces méthodes impliquent souvent de la même manière les individus (lignes) et les variables (colonnes). La confrontation des espaces d'individus et de variables enrichit les interprétations.

A1.2 Les méthodes factorielles : aspects techniques

Les méthodes factorielles[1] permettent de gérer simultanément des quantités importantes de données et leur système de corrélations et, par une technique réalisant une sorte de *compression*, d'en dégager la structure interne, notamment sous forme de graphique-plans.

1. Elles comprennent dans la littérature statistique française des trente dernières années toutes les techniques de représentation utilisant des « axes principaux »: analyse en composantes principales, analyse des correspondances simples et multiples, analyse factorielle dite classique (en anglais : *factor analysis*) ou analyse en facteurs communs et spécifiques.

- *Recherche des sous-espaces factoriels*

L'objectif est de rechercher des sous-espaces de dimensions réduites (entre trois et dix, par exemple) qui ajustent au mieux le nuage de points-individus et celui des points-variables, de façon à ce que les proximités mesurées dans ces sous-espaces reflètent autant que possible les proximités réelles. On obtient ainsi un espace de représentation, l'espace factoriel, défini par les axes principaux d'inertie et l'on représente les points du nuage dans ce système d'axes (*cf.* figure A1.4). Ces axes réalisent les meilleurs ajustements de l'ensemble des points selon le critère classique des moindres carrés, qui consiste à rendre minimale la somme des carrés des écarts entre les points et les axes.

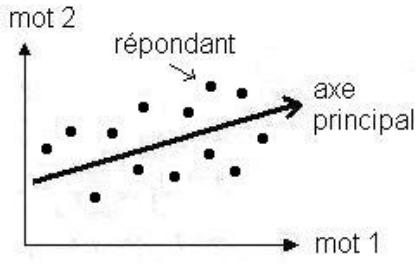


Figure A1.4 : Ajustement du nuage des points-individus dans l'espace des mots

Le premier de ces axes correspond à la droite d'allongement maximum du nuage, le second maximise le même critère en étant assujetti à être orthogonal au premier, et ainsi de suite pour les axes suivants qui sont tous orthogonaux entre eux. Cette orthogonalité traduit l'indépendance (en fait, la non-corrélation) des axes.

X désigne le tableau de données ayant subi des transformations préliminaires (variables centrées réduites, par exemple), **X'** son transposé.

Soit $\mathbf{u}_1$ le vecteur unitaire qui caractérise le premier axe. $\mathbf{u}_1$ est alors le vecteur propre de la matrice **X'X** correspondant à la plus grande valeur propre λ_1 [*cf.* SEM 2006].

Plus généralement, le sous-espace à q dimensions qui ajuste au mieux (au sens des moindres carrés) le nuage est engendré par les q premiers vecteurs propres de la matrice **X'X** correspondant aux q plus grandes valeurs propres.

La procédure d'ajustement est exactement la même pour les deux nuages. On démontre alors qu'il existe des relations simples liant les axes

calculés dans les deux espaces, celui des individus et celui des variables (*relations de transition*).

Le vecteur des coordonnées des points sur chacun des axes, appelé *facteur*, est une combinaison linéaire des variables initiales. On dénote par ψ_α et φ_α les facteurs correspondant à l'axe α respectivement dans l'espace noté $\mathbb{R}^m$ (espace dont les *n* points ont pour coordonnées sont les *m* mots) et dans l'espace noté $\mathbb{R}^n$ (espace dont les *m* points ont pour coordonnées sont les *n* individus).

Les deux nuages de points, celui des mots et celui des répondants, sont intrinsèquement liés et révèlent exactement les mêmes structures : dans un cas, les facteurs décriront les corrélations entre les mots, dans l'autre les associations entre les répondants.

Les plans factoriels de visualisation utilisés tout au long de cet ouvrage correspondent chacun à un couple de facteurs.

Le plan sémiométrique le plus utilisé est le plan (φ_2, φ_3).

Les éléments (mots ou individus) qui participent au calcul des axes sont les *éléments actifs*. On introduit aussi dans l'analyse des *éléments supplémentaires* (ou *illustratifs*) qui ne participent pas à la formation des axes, mais qui sont projetés *a posteriori* dans les plans factoriels et peuvent aider à leur interprétation (*cf.* section A1.2.4).

- *Techniques de base et méthodes dérivées*

La nature des informations, leur codage dans le tableau de données, les spécificités du domaine d'application vont introduire des variantes au sein des méthodes factorielles.

Celles qui sont utilisées ici ne sont en fait que des dérivées de deux techniques fondamentales, l'analyse en composantes principales et l'analyse factorielle des correspondances.

L'analyse en composantes principales s'applique à un tableau de mesures numériques et sera utilisée, dans le cadre de la sémiométrie, pour traiter un tableau de notes.

Les exemples d'analyse de données textuelles présentée au chapitre 4 reposent sur l'analyse factorielle des correspondances appliquée aux tableaux de contingence lexicaux.

A1.3 L'Analyse en Composantes Principales : aspects techniques

L'Analyse en Composantes Principales (Hotelling, 1933) s'applique à des variables à valeurs numériques (des mensurations, des taux, des mots etc.) représentées sous forme d'un tableau rectangulaire de mesures **R** de terme général r_{ij} dont les colonnes sont les variables et les lignes représentent les individus sur lesquels ces variables sont mesurées. En sémiométrie, les variables sont donc les mots; les lignes les répondants et les valeurs numériques, les notes.

A1.3.1 Interprétations géométriques

Les représentations géométriques entre les lignes d'une part et entre les colonnes d'autre part du tableau de données permettent de visualiser les proximités respectivement entre les individus et entre les variables (*cf.* figures A1.1 et A1.2 ci-dessus).

Dans $\mathbb{R}^m$, deux points-individus sont très voisins si, dans l'ensemble, leurs m coordonnées sont très proches. Les deux répondants concernés sont alors caractérisés par des valeurs presque égales pour chaque variable. La distance utilisée est la distance euclidienne usuelle.

Dans $\mathbb{R}^n$, si les valeurs prises par deux variables particulières sont très voisines pour tous les répondants, ces variables seront représentées par deux points très proches dans cet espace. Cela peut vouloir dire que ces variables mesurent une même chose ou encore qu'elles sont liées par une relation particulière.

Mais les unités de mesure des variables peuvent être très différentes et rendre alors nécessaire des transformations du tableau de données.

A1.3.2 Problème d'échelle de mesure et transformation des données

On veut que la distance entre deux individus soit indépendante des unités des variables pour que chaque variable joue un rôle identique. Pour cela, on attribue à chaque variable j la même dispersion en divisant chacune de ses valeurs par leur écart-type s_j avec $s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(r_{ij} - \overline{r}_j)^2$.

Par ailleurs on s'intéresse à la manière dont les individus s'écartent de la moyenne. On place alors le point moyen au centre de gravité du nuage

des individus. Les coordonnées du point moyen sont les valeurs moyennes des variables notées $\overline{r}_j = \frac{1}{n}\sum_{i=1}^{n} r_{ij}$. Prendre ce point comme origine revient à soustraire pour chaque variable sa moyenne $\overline{r}_j$.

On corrige ainsi les échelles en transformant le tableau de données **R** en un nouveau tableau **X** de la façon suivante :

$$x_{ij} = \frac{r_{ij} - \overline{r}_j}{s_j \sqrt{n}}$$

Les variables ainsi réduites et centrées ont toutes une variance, $s^2(x_j)$, égale à 1 et une moyenne, $\overline{x}_j$, nulle et deviennent comparables. D'autres transformations préalables sont possibles (*cf.* section 2.5 du chapitre 2).

A1.3.3 Analyse du nuage des *n* répondants

La transformation des données amène à effectuer une translation de l'origine au centre de gravité de ce nuage et à changer (dans le cas de l'analyse dite normée) les échelles sur les différents axes.

Pour réaliser l'analyse du nuage des points-répondants dans $\mathbb{R}^m$, la matrice **X'X** à diagonaliser dans cet espace, est la matrice des corrélations (*dont la* figure A1.4 fournit un exemple) qui a pour terme général :

$$c_{jj'} = \sum_{i=1}^{n} x_{ij} x_{ij'} = \frac{1}{n} \sum_{i}^{n} \frac{(r_{ij} - \overline{r}_j)(r_{ij'} - \overline{r}_{j'})}{s_j s_{j'}}$$

$c_{jj'}$ est le coefficient de corrélation entre les variables *j* et *j'*.

Les coordonnées des *n* points-individus sur l'axe factoriel $\mathbf{u}_\alpha$ sont les *n* composantes du vecteur $\boldsymbol{\psi}_\alpha = \mathbf{X}\mathbf{u}_\alpha$.

La figure A1.4-a illustre la représentation du nuage des répondants pour le tableau de 12 répondants ayant noté 7 mots (tableau déjà présenté en section A1.1) dans le plan principal (2, 3)[1]. Les répondants R01 et R02 ont donné, de la même façon, des notes très contrastées et ont donné des notes élevées à *Arbre* et *Sensuel* et des notes faibles à *Morale* et *Politesse* ; ils sont par conséquent proches dans le plan et se différencient

1. Le plan (2, 3) a été considéré comme le plan sémiométrique principal compte tenu du caractère particulier du premier axe (axe dit *de taille*, *cf.* chapitre 5).

des répondants R05 et R04 qui se sont exprimés de façon inverse sur les mots. Le répondant R08 se distingue en ayant très bien noté *Danger* sans pour autant bien noter les autres mots, alors que R11 a bien noté tous les mots.

A1.3.4 Analyse du nuage des variables (mots)

Les coordonnées factorielles $\varphi_{\alpha j}$ des points-variables sur l'axe α sont les composantes de $\mathbf{u}_\alpha \sqrt{\lambda_\alpha}$ et l'on a :

$$\varphi_{\alpha j} = cor(j, \boldsymbol{\psi}_\alpha)$$

La coordonnée $\varphi_{\alpha j}$ d'un point-variable *j* sur un axe α n'est autre que le *coefficient de corrélation* de cette variable avec le facteur $\boldsymbol{\psi}_\alpha$ (combinaison linéaire des variables initiales) considéré lui-même comme une variable artificielle dont les coordonnées sont constituées par les *n* projections des individus sur cet axe.

Les axes factoriels étant orthogonaux deux à deux, on obtient ainsi une série de variables artificielles non corrélées entre elles, appelées *composantes principales*[1], qui synthétisent les corrélations de l'ensemble des variables initiales.

Sur la figure A1.4-b, comme sur la matrice de corrélations correspondante, *Politesse* et *Morale* sont très corrélés et dans une moindre mesure *Orage* et *Sensuel.* On retrouve bien le comportements des répondants où R01 et R02 vont dans la direction des bons noteurs d'*Arbre* et de *Sensuel* et des mauvais noteurs de *Morale* et *Politesse* à l'inverse de R04 et R05.

Les variables fortement corrélées avec un axe vont contribuer à la définition de cet axe[2]. Cette corrélation se lit directement sur le graphique puisqu'il s'agit de la coordonnée du point-variable *j* sur l'axe α.

1. L'analyse en composantes principales ne traduit que des liaisons linéaires entre les variables. Un coefficient de corrélation faible entre deux variables signifie donc que celles-ci sont indépendantes linéairement, alors qu'il peut exister une relation non linéaire.

2. L'exemple n'est bien évidemment pas suffisamment représentatif pour que le plan puisse être interprété. Il a juste vocation à rapprocher le tableau de données des résultats.

Tableau de notes (1 à 7) données à 7 mots par 12 répondants

	arbre	cadeau	danger	morale	orage	politesse	sensuel
R01	7	4	2	2	3	1	6
R02	6	3	1	2	4	1	7
R03	4	5	3	4	3	4	3
R04	5	5	1	7	2	7	1
R05	4	5	2	7	1	6	2
R06	5	7	1	5	2	6	5
R07	4	2	1	3	5	3	6
R08	4	1	5	4	5	4	7
R09	6	6	2	4	7	5	5
R10	6	6	3	5	3	6	6
R11	7	7	6	7	7	6	7
R12	2	2	1	2	1	3	2

Matrice des corrélations

	!	arbr	cade	dang	mora	orag	poli	sens
arbr	!	1.00						
cade	!	.55	1.00					
dang	!	.29	.14	1.00				
mora	!	.16	.62	.36	1.00			
orag	!	.51	.09	.54	-.01	1.00		
poli	!	.00	.63	.23	.91	-.05	1.00	
sens	!	.56	-.08	.45	-.30	.68	-.37	1.00

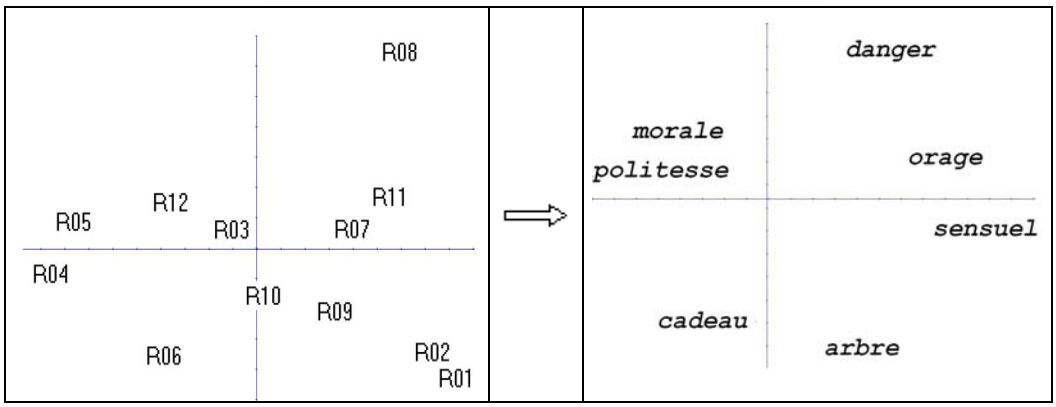


Fig. A1.4-a: Représentation des répondants dans le plan factoriel (2,3)

Fig. A1.4-b: Représentation des mots dans le plan factoriel (2,3)

Figure A1.4 : Analyse en composantes principales sur le tableau de notes de 7 mots par 12 répondants

On s'intéresse surtout aux variables présentant les plus fortes coordonnées et l'on interprétera les composantes principales en fonction des regroupements de certaines de ces variables et de l'opposition avec les autres.

On notera alors que tous les points-variables sont sur une sphère de rayon 1 centrée à l'origine des axes[1]. Les plans d'ajustement couperont la sphère suivant de grands cercles (de rayon 1), les *cercles de corrélations*, à l'intérieur desquels sont positionnés les points-variables. Dans cet ouvrage, les cercles ne sont pas tracés dans les plans factoriels représentant les mots pour une meilleure lisibilité des libellés (le cadrage des plans factoriels aurait en effet entraîné une forte réduction d'échelle).

A1.4 L'Analyse des correspondances

L'analyse des correspondances[2] s'applique en premier lieu à une table de contingence **K**, appelé aussi tableau croisé, à n lignes et p colonnes, qui ventile une population selon deux variables qualitatives à n et p modalités. Les lignes et les colonnes jouent donc des rôles similaires.

Dans le chapitre 4, l'analyse est appliquée à un tableau croisant les 1 191 répondants spontanément aux questions ouvertes avec les 592 mots cités au moins 4 fois comme étant agréables. Une deuxième analyse porte sur 158 mots apparaissant plus de 25 fois.

- *Notations*

Soit $k = \sum_{ij} k_{ij}$ la somme de tous les éléments k_{ij} de la table de contingence **K**. Dans le cas du tableau des questions ouvertes, k représente le nombre de fois que les 592 mots ont été cités spontanément. On note $f_{ij} = k_{ij} / k$ les fréquences relatives avec $\sum_i \sum_j f_{ij} = 1$.

---

1. L'analyse du nuage des points-variables dans $\mathbb{R}^n$ ne se fait pas par rapport au centre de gravité du nuage, contrairement à celui des points-individus mais par rapport à l'origine. La distance d'une variable j à l'origine O s'exprime par : $d^2(O, j) = \sum_{i=1}^{n} x_{ij}^2 = 1$

2. Présentée et étudiée de façon systématique comme une technique souple d'analyse exploratoire de données multidimensionnelles par J.-P. Benzécri (1973), l'analyse des correspondances s'est trouvée depuis d'autres précurseurs, en particulier C. Hayashi (1956), et a donné lieu à des travaux dispersés et indépendants les uns des autres.

On note : $f_{i.} = \sum_j f_{ij}$, $f_{.j} = \sum_i f_{ij}$, les fréquences marginales relatives.

La table de contingence **K** est transformé en un tableau de profils-lignes $f_{ij} / f_{i.}$ et un tableau de profils-colonnes $f_{ij} / f_{.j}$.

Le point i de $\mathbb{R}^m$ a pour coordonnées : $f_{ij} / f_{i.}$ pour tout $j \leq m$.

De même, le point j de $\mathbb{R}^n$ a pour coordonnées : $f_{ij} / f_{.j}$ pour tout $i \leq n$.

Notons une différence importante entre l'analyse des correspondances et l'analyse en composantes principales : les transformations opérées sur le tableau dans les deux espaces sont identiques (car les ensembles mis en correspondance jouent des rôles analogues).

- *Distance du Chi-deux et équivalence distributionnelle*

Les distances entre deux points-lignes *i* et *i'* d'une part et entre deux points-colonnes d'autre part sont données par les équations suivantes :

$$d^2(i,i') = \sum_{j=1}^{m} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2 \qquad d^2(j,j') = \sum_{i=1}^{n} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2$$

La distance du χ^2 offre l'avantage de vérifier le principe d'équivalence distributionnelle. Ce principe assure la robustesse des résultats de l'analyse des correspondances vis-à-vis de l'arbitraire du découpage en modalités des variables nominales. Il s'exprime de la façon suivante : si deux lignes (resp. colonnes) du tableau de contingence ont même profil (sont proportionnelles) alors leur agrégation n'affecte pas la distance entre les colonnes (resp. lignes). On obtient alors un nouveau point-ligne (resp. point-colonne) de profil identique et affecté de la somme des fréquences des deux points-lignes (resp. points-colonnes).

Cette propriété est importante car elle garantit une certaine invariance des résultats vis-à-vis de la nomenclature choisie pour la construction des modalités d'une variable qualitative.

A1.5 L'analyse logarithmique

L'analyse logarithmique, proposée par J.-B. Kazmierczak (1985), réalise la propriété de l'équivalence distributionnelle de l'analyse des correspondances sur des tableaux qui ne sont pas obligatoirement des tables de contingence. J.-B. Kazmierczak reprend et généralise le principe de Yule qui stipule que l'on ne change pas la distance entre

deux lignes ni la distance entre deux colonnes d'un tableau en remplaçant les lignes et les colonnes de ce tableau par d'autres lignes et colonnes qui leur sont proportionnelles (il s'agit en fait d'une généralisation du principe d'équivalence distributionnelle).

L'analyse logarithmique consiste à prendre les logarithmes des (après addition éventuelle d'une constante en cas de données négatives), puis, après les avoir centrées à la fois en ligne et en colonne, à les soumettre à une analyse en composantes principales non normée, qui coïncide ici avec une *décomposition aux valeurs singulières* [SEM, 2002].

Ainsi, si **R** est un tableau de données (n, m) et si **A** et **B** sont deux matrices diagonales respectivement de dimensions (n, n) et (p, p) à éléments diagonaux positifs, la matrice **ARB** donne lieu à la même analyse logarithmique que la matrice **R**. Cette propriété d'invariance forte a eu pour effet de supprimer le premier axe sémiométrique sans altérer la suite des axes (section 5.7 du chapitre 5).

A1.6 L'analyse factorielle en facteurs communs et spécifiques

L'analyse factorielle en facteurs communs et spécifiques (*factor analysis*) est probablement le modèle linéaire de variables latentes le plus ancien[1]. Ces modèles ont été essentiellement développés principalement par les psychologues et psychométriciens. Les développements auxquels ils donnent lieu sont complexes et diversifiés. On pourra consulter sur ce point les ouvrages classiques de Harman (1967), Mulaik (1972)[2].

Mentionnons également les travaux d'Anderson et Rubin (1956) et de Lawley et Maxwell (1963) qui ont placé l'analyse factorielle en facteurs communs et spécifiques dans un cadre inférentiel classique.

− *Le modèle de l'analyse factorielle*

Ce modèle se propose de reconstituer, à partir d'un petit nombre q de facteurs, les corrélations existant entre m variables observées. On suppose l'existence d'un modèle *a priori* :

---

3. A l'origine des principes de la méthode se trouvent Spearman (1904) (analyse monofactorielle), puis Garnett (1919) et Thurstone (1947) (analyse multifactorielle).

2. En économétrie, on distingue habituellement les modèles fonctionnels, ou à effet fixes (comme la régression multiple et le modèle linéaire dans son ensemble), et les modèles structurels ou à effet aléatoire (modèles de variables latentes).

$$\underset{(m,1)}{\mathbf{x}_i} = \underset{(m,q)}{\mathbf{\Gamma}} \underset{(q,1)}{\mathbf{f}_i} + \underset{(p,1)}{\mathbf{e}_i}$$

Dans cette écriture $\mathbf{x_i}$ représente le *i-ème* vecteur observé des *m* variables ; $\mathbf{\Gamma}$ est un tableau (*m*, *q*) de coefficients inconnus (avec $q < m$) ; $\mathbf{f_i}$ est la *i-ème* valeur du vecteur aléatoire et non observable de *q facteurs communs ;* et $\mathbf{e}_i$ *la i-ème* valeur du vecteur non observable de résidus, lesquels représentent l'effet combiné de *facteurs spécifiques* et d'une perturbation aléatoire.

On désigne par **X** le tableau (*n*,*p*) dont la *i-ème* ligne représente l'observation *i*. De même **F** désigne le tableau (*n*,*q*) non observable dont la *i-ème* ligne est $\mathbf{f}_i^{'}$ et **E** le tableau (*n*,*p*) non observable dont la *i-ème* ligne est $\mathbf{e}_i^{'}$. Le modèle liant l'ensemble des observations aux facteurs hypothétiques s'écrit :

$$\underset{(n,m)}{\mathbf{X}} = \underset{(n,q)}{\mathbf{F}} \underset{(q,m)}{\mathbf{\Gamma'}} + \underset{(n,m)}{\mathbf{E}}$$

Dans cette écriture, seul **X** est observable, et le modèle est par conséquent indéterminé.

L'identification de ce modèle et l'estimation des paramètres posent des problèmes complexes. Une cascade d'hypothèses *a priori* supplémentaires permet cette identification. L'application dans le cadre de cet ouvrage concerne la section 2.6 du chapitre 2.

A1.7 Méthodes de classification hiérarchique

Les techniques de classification automatique[1] sont destinées à produire des groupements d'objets ou d'individus décrits par un certain nombre de variables ou de caractères. Les circonstances d'utilisation sont sensiblement les mêmes que celles des méthodes d'analyse factorielle descriptive présentées aux sections précédentes. Dans le chapitre 3, la classification est réalisée sur l'ensemble des 210 mots à partir des coordonnées de ces mots sur les axes principaux.

Il existe plusieurs familles d'algorithmes de classification : les *algorithmes hiérarchiques* qui fournissent une hiérarchie de partitions des objets et les algorithmes conduisant directement à des *partitions* comme les méthodes d'agrégation autour de centres mobiles. Les principes communs aux diverses techniques de classification ascendante hiérarchique sont simples. Il s'agit de créer, à chaque étape de

1. La classification est une branche de l'analyse des données qui constitue une étape fondamentale dans beaucoup de disciplines scientifiques. Elle a donné lieu à des publications nombreuses et diversifiées dont : Sokal et Sneath (1963) et Benzécri (1973).

l'algorithme, une partition obtenue en agrégeant deux à deux les éléments les plus proches.

- *L'algorithme de la Classification Hiérarchique*

L'algorithme de base de la classification ascendante hiérarchique produit une hiérarchie en partant de la partition dans laquelle chaque élément à classer constitue une classe, pour aboutir à la partition formée d'une seule classe réunissant tous les éléments.

Pour *n* éléments à classer, il est composé de *n* étapes. A la première étape, il y a donc *n* éléments à classer. On construit la matrice de distances entre les *n* éléments et l'on cherche les deux plus proches, que l'on agrège en un nouvel élément.

On construit une nouvelle matrice des distances qui résultent de l'agrégation, en calculant les distances entre le nouvel élément et les éléments restants. On se trouve dans les mêmes conditions qu'à l'étape *1*, mais avec seulement (*n-1*) éléments à classer.

On cherche de nouveau les deux éléments les plus proches, que l'on agrège. On réitère le processus jusqu'à n'avoir plus qu'un seul élément regroupant tous les objets et qui constitue la dernière partition.

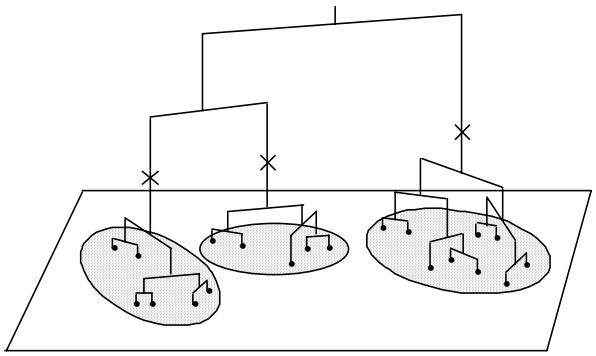


Figure A1.6: Dendrogramme ou arbre hiérarchique

L'algorithme ne fournit pas une partition en *q* classes d'un ensemble de *n* objets mais une *hiérarchie de partitions,* se présentant sous la forme *d'arbres* appelés également *dendrogrammes* et contenant *n - 1* partitions (*cf.* figure A1.6). L'intérêt de ces arbres est qu'ils peuvent donner une idée du nombre de classes existant effectivement dans la population. Chaque coupure d'un dendrogramme fournit une partition.

Dans le tableau 3.1 du chapitre 3, on a retenu trois coupures du dendrogramme en 36, 24 et 12 classes. La coupure la moins fine en 12

classes figure en première colonne, celle en 24 classes en seconde colonne et la coupure la plus fine en 36 classes en troisième colonne. Elle est suivie de la liste des mots qui composent ces 36 classes.

A1.8 Les cartes auto-organisées de Kohonen

L'objectif des cartes auto-organisées de Kohonen[1] est de classer un ensemble d'observations de façon à conserver la topologie initiale de l'espace dans lesquelles elles sont décrites. Comme les réseaux de neurones auxquelles elles sont rattachées, ces cartes obtiennent de bonnes performances pour la reconnaissance de formes. Elles ont été utilisées aux chapitres 3 (section 3.1) et 4 (section 4.2).

- *Le principe*

Les cartes de Kohonen cherchent à représenter dans un espace à deux (parfois trois) dimensions les lignes ou les colonnes d'un tableau en respectant la notion de voisinage dans l'espace des éléments à classer. Tout comme l'analyse en composantes principales, il est utile d'imaginer au départ l'ensemble des données (les mots) comme un nuage de points dans un espace de grande dimension (celui des individus ou répondants).

Le principe est de considérer une carte comme une grille rectangulaire (parfois hexagonale) aux mailles déformables, laquelle, une fois dépliée épouse au mieux les formes du nuage de points. Les nœuds de la grille sont les *neurones* de la carte. Chaque point du nuage est projeté sur le nœud dont il est le plus proche. De fait, chaque point, décrit initialement dans un espace multidimensionnel est représenté à la fin par deux coordonnées donnant la position du *neurone* sur la carte : l'espace est réduit. L'ensemble des points affectés à un même *neurone* sont proches dans l'espace initial. Ils décrivent et regroupent des individus semblables.

On définit *a priori* une notion de voisinage entre classes et les observations voisines dans l'espace des variables de dimension q appartiennent après classement à la même classe ou à des classes voisines. Ces voisinages peuvent être choisis de diverses manières mais en général on les suppose directement contigus sur la grille rectangulaire (ce qui représente alors 8 voisins pour un *neurone*).

---

1. Introduites en 1981 par Teuvo Kohonen, elles font partie des méthodes dites *neuronales* (*cf.* Kohonen, 1989). Elles donnent lieu à plusieurs applications relevant par exemple de l'analyse de textes, les diagnostics médicaux et industriels, les contrôles de processus, la robotique.

- *L'algorithme*

L'algorithme d'apprentissage pour classer m points est itératif[1]. L'initialisation consiste à associer à chaque classe k un centre provisoire $\mathbf{C}_k$ à q composantes choisi de manière aléatoire dans l'espace à q dimensions contenant les m mots à classer. A chaque étape on choisit un mot i au hasard que l'on compare à tous les centres provisoires et l'on affecte le mot au centre $\mathbf{C}_{k_0}$ le plus proche au sens d'une distance donnée *a priori*. On rapproche alors du mot i le centre $\mathbf{C}_{k_0}$ et les centres voisins sur la carte ce qui s'exprime à l'étape t par :

$$\mathbf{C}_k(t+1)=\mathbf{C}_k(t)+\varepsilon(i(t+1)-\mathbf{C}_k(t))$$

où i(t+1) est le mot présenté à l'étape t+1, ε un paramètre d'adaptation positif et inférieur à 1. Cette expression n'intervient que pour le centre $\mathbf{C}_{k_0}$ et ses voisins.

Cet algorithme est analogue à celui des centres mobiles [SEM 2006], mais dans ce dernier cas, il n'existe pas de notion de voisinage entre classes et on ne modifie à chaque étape que la position du centre $\mathbf{C}_{k_0}$.

L'auto-organisation de la carte de Kohonen est la conséquence de la notion de voisinage. Comme l'algorithme des centre mobiles, cet algorithme est très adapté aux applications où les données sont importantes et où il n'est pas utile de les stocker.

A1.9 Outils de validation

Tout au long du présent ouvrage ont été utilisées les notions de *valeur-test* et de *variable supplémentaire*.

Les *valeurs-test* (section A1.9.1) sont un outil d'inférence statistique élémentaire, mais polyvalent et très utile, surtout si l'utilisateur est averti des problèmes de *comparaisons multiples* qui ne manquent pas d'intervenir (section A1.9.2).

La technique des variables supplémentaires (section A1.9.3) est un outil fondamental de valorisation des méthodes factorielles, qui permet une validation *externe* des résultats, à la fois épreuve de cohérence et enrichissement des interprétations.

1. On se réfère dans la présentation de l'algorithme au cours de P.Letremy et M.Cottrell (SAMOS-MATISSE, Université Paris I). Voir aussi Thiria et *al.* (1997).

Les deux autres outils de validation utilisés dans cet ouvrage sont les *intervalles de confiance d'Anderson* et les procédures de reéchantillonnage *bootstrap.*

Les *intervalles de confiance d'Anderson* (section A1.9.4) sont utilisés dans la section 2.2 du chapitre 2 pour valider des valeurs propres.

Les procédures de reéchantillonnage *bootstrap* (section A1.9.5) sont utilisées aux sections 2.3 et 2.4 du chapitre 2 pour mettre en évidence la stabilité des structures sémiométriques, et, dans la section 4.4 du chapitre 4, pour valider une analyse textuelle.

A1.9.1 Qu'est-ce qu'une valeur-test ?

La valeur-test est un critère qui permet d'apprécier rapidement si une modalité d'une variable nominale (*i.e.* : une catégorie de répondants) a une position *significative* sur un axe. Pour cela, on teste l'hypothèse selon laquelle un groupe d'individus, correspondant à une modalité donnée d'une variable nominale supplémentaire (comme la modalité *profession libérale, cadre supérieur* pour la variable nominale *catégorie socio-professionnelle*, par exemple), peut être considéré comme tiré au hasard, sans remise, dans la population.

Dans le cas d'un véritable tirage au hasard, le centre de gravité du sous-nuage représentant le groupe (*i.e.* : la modalité) s'éloigne peu du centre de gravité du nuage global correspondant à tout l'échantillon.

On convertit alors la coordonnée de cette modalité sur l'axe en une *valeur-test* qui est, sous cette hypothèse, la réalisation d'une variable normale centrée réduite. Autrement dit, dans l'hypothèse selon laquelle une modalité a une composition *aléatoire*, la valeur-test correspondante a 95% de chances d'être comprise dans l'intervalle [-1.96, +1.96].

On considère alors comme occupant une *position significative* les modalités dont les valeurs-test sont supérieures à 2 (pour 1.96) en valeur absolue, ce qui correspond approximativement au seuil usuel de probabilité de 5%.

Souvent les valeurs-test sont largement supérieures à ce seuil. On les utilise alors pour trier les modalités, des plus significatives au moins significatives. La valeur-test systématise la notion de *t-value* souvent utilisée dans la littérature anglo-saxonne.

Supposons qu'une modalité *j* concerne n_j individus. Si ces n_j individus sont tirés au hasard (c'est ce qu'on appelle l'hypothèse nulle H_0) parmi les *n* individus analysés (tirage supposé sans remise), la moyenne de n_j coordonnées tirées au hasard dans l'ensemble fini des *n* valeurs $\psi_{\alpha i}$ (coordonnée du répondant *i* sur l'axe α) est une variable aléatoire $X_{\alpha j}$: $X_{\alpha j} = \frac{1}{n_j} \sum_{i \in I(j)} \psi_{\alpha i}$ avec pour espérance $E(X_{\alpha j})=0$ et pour variance[1] $Var_{H_0}(X_{\alpha j}) = \frac{n - n_j}{n-1} \frac{\lambda_\alpha}{n_j}$

Dans la formule donnant $X_{\alpha j}$, *I(j)* est le sous-ensemble des répondants caractérisés par la modalité j de la variable nominale.

La coordonnée $\varphi_{\alpha j}$ de la modalité *j* est proportionnelle à la variable aléatoire $X_{\alpha j}$ et s'écrit ainsi : $\varphi_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} X_{\alpha j}$

On a donc $E(\varphi_{\alpha j}) = 0$ et $Var_{H_0}(\varphi_{\alpha j}) = \frac{n - n_j}{n - 1} \frac{1}{n_j}$

La quantité $t_{\alpha j}$: $t_{\alpha j} = \sqrt{n_j \frac{n-1}{n-n_j}} \varphi_{\alpha j}$ mesure en *nombre d'écart-types* la distance entre la modalité *j*, c'est-à-dire le quasi-barycentre des n_j individus, et l'origine, sur l'axe factoriel α. On appelle cette quantité « *valeur-test* ». D'après le théorème de la limite centrale (*central limit theorem*), sa distribution tend vers une loi de Laplace-Gauss centrée réduite.

On doit noter que les valeurs-test n'ont de sens que pour les modalités supplémentaires (*cf.* section suivante), ou des modalités actives ayant des contributions absolues faibles, c'est-à-dire se comportant en fait comme des modalités supplémentaires[2].

Lorsque l'on dispose d'un nombre important de modalités supplémentaires, les valeurs-test permettent de repérer rapidement les modalités utiles à l'interprétation d'un axe ou d'un plan factoriel.

1. Il s'agit de la formule classique donnant la variance d'une moyenne lors d'un tirage sans remise de n_j objets parmi *n*, en fonction de la variance totale λ_α, qui est aussi, dans le cas des coordonnées factorielles, la valeur propre correspondant à l'axe α.

2. Les coordonnées sur un axe des individus correspondant à une modalité active ne peuvent être considérées comme tirée au hasard, puisque cette modalité aura contribué à construire l'axe.

A1.9.2 Le problème des comparaisons multiples

Le calcul simultané de plusieurs valeurs-test ou de plusieurs seuils de probabilités se heurte à l'écueil des *comparaisons multiples,* bien connu des statisticiens ; *cf.* O'Neill et Wetherill (1971), Saville (1990), Westfall et Young (1993), Westfall *et al.* (1999), Hsu (1996).

Supposons que l'on projette 100 modalités *supplémentaires* (*cf.* section suivante A1.9.3) qui soient vraiment tirées au hasard. Les valeurs-test attachées à ces modalités sont alors toutes des réalisations de variables aléatoires normales centrées réduites indépendantes.

Dans ces conditions, *en moyenne*, sur 100 valeurs-test calculées, cinq seront en dehors de l'intervalle [-1.96, +1.96] et seront, en apparence seulement, significatives. Le seuil de 5% n'a de sens en fait que pour un seul test, et non pour des tests multiples.

On résout en pratique cette difficulté en choisissant un seuil plus sévère[1]. Le seuil le plus sévère et pessimiste que l'on puisse imaginer est le « seuil de *Bonferroni* » (on divise le seuil initial par le nombre de tests : dans le cas de 210 tests : $0.05 / 210 = 2.4\ 10^{-4}$). La valeur-test unilatérale correspondante est de 3.49. Cette valeur nous fournit un garde-fou prudent à l'excès[2].

Comme cela a été signalé dans le corps du texte (*cf.*, par exemple, les notes de la section 7.2 du chapitre 7), l'interdépendance des mots ne permet pas d'appliquer aveuglément les résultats concernant les comparaisons multiples. Que conclure, en effet, lorsque plusieurs mots de sens voisins ont simultanément des valeurs-test de l'ordre de 1.96 ? Celles-ci ne sont pas significatives une par une en retenant le seuil de *Bonferroni*, mais elles se confirment et se valident mutuellement.

Une solution pragmatique (cas multidimensionnel) : le bootstrap.

La technique de validation par *bootstrap* dont il sera question plus loin dans cette annexe apporte une contribution intéressante au difficile problème des comparaisons multiples, car les réplications d'échantillons permettent de prendre en compte simultanément toutes les variables, et donc l'interdépendance des variables.

Il s'agit d'un test global, et non plus de tests séparés pour chaque variable. Une illustration en est donnée par la figure 4.4 du chapitre 4 qui

1. Les valeurs-test permettent surtout de *classer* les modalités supplémentaires par ordre d'intérêt décroissant, ce qui constitue une aide précieuse à l'interprétation des facteurs.

2. *Cf.*, par exemple, Hochberg (1988), Perneger (1998).

représente les zones de confiance simultanées des mots, dont certains apparaissent comme significativement distincts. Dans ce cas, les tests ne sont pas réalisés isolément ni en série, mais simultanément.

A1.9.3 Utilité des éléments supplémentaires

L'analyse factorielle permet de trouver des sous-espaces de représentation des proximités entre points-individus ou entre points-variables. Elle s'appuie pour cela sur des éléments (individus ou variables) dits *actifs*.

Il est possible d'introduire en supplémentaire d'autres points (ou éléments) que l'on ne souhaite pas faire intervenir dans la composition et définition des axes mais dont on veut connaître les positions dans les espaces factoriels[1]. On projette alors ces points après la construction des axes factoriels dans ce nouveau repère. Cette projection se fait de façon très simple en utilisant les formules dites *de transition*, que ce soit en analyse en composantes principales ou en analyse des correspondances.

C'est le cas lorsque l'on veut positionner les mots, variables numérique, dans l'espace des variables de contrôle (*cf.* section 5.4). On calcule, *a posteriori*, leurs coordonnées sur les axes factoriels.

C'est également le cas lorsque l'on souhaite caractériser les axes sémiométriques[2] par les critères socio-démographiques (variables nominales) de la population enquêtée (*cf.* section 1.4).

Ces critères définissent en fait des groupes d'individus et sont considérés soit comme des modalités de variables nominales, soit comme des individus, mis en éléments supplémentaires.

Ce sont les centres de gravité de ces groupes qui sont positionnés dans l'espace des variables. La valeur-test permet d'en apprécier la significativité sur l'axe.

Cette procédure pourrait être utilisée comme méthode alternative pour comparer des sous-populations dans le chapitre 7.

---

1. On peut citer trois raisons qui peuvent susciter la mise en supplémentaire d'un point : 1) enrichir l'interprétation des axes par des variables (de nature ou de thématique différente de celle des éléments actifs) n'ayant pas participé à leur construction ; 2) adopter une optique de prévision en projetant les variables supplémentaires dans l'espace des individus. Celles-ci seront « expliquées » par les variables actives ; 3) faire ressortir l'essentiel d'une structure masquée par l'existence d'un point actif, de faible masse, mais très excentré qui pourrait déformer le nuage.

2. Ces axes, rappelons-le, sont définis par les variables actives que sont les mots.

A1.9.4 Intervalles de confiance d'Anderson

Anderson (1963) a calculé les lois limites des valeurs propres d'une analyse en composantes principales sans nécessairement supposer que les valeurs théoriques correspondantes sont distinctes.

L'ampleur de l'intervalle donne une indication sur la stabilité de la valeur propre vis-à-vis des fluctuations dues à l'échantillonnage supposé laplacien. L'empiétement des intervalles de deux valeurs propres consécutives suggérera donc l'égalité de ces valeurs propres. Les axes correspondants sont alors définis à une rotation près. Ainsi l'utilisateur pourra éviter d'interpréter un axe instable selon ce critère.

Si les valeurs propres théoriques λ_α de $\mathbf{\Sigma}$ sont distinctes, les valeurs propres $\hat{\lambda}_\alpha$ de la matrice des covariances empiriques **S** suivent asymptotiquement des lois normales d'espérance λ_α et de variance $2\lambda_\alpha^2/(n-1)$ où n est la taille de l'échantillon.

On en déduit les intervalles de confiance approchés au seuil 95% :

$$\lambda_\alpha \in \left[\hat{\lambda}_\alpha\left(1-1.96\sqrt{2/(n-1)}\right)\ ;\ \hat{\lambda}_\alpha\left(1+1.96\sqrt{2/(n-1)}\right)\right]$$

Les intervalles de confiance d'Anderson concernent en fait aussi bien les valeurs propres des matrices des covariances que des matrices de corrélations. Les simulations entreprises montrent que les intervalles de confiance obtenus sont en général « prudent » : le pourcentage de couverture de la vraie valeur est le plus souvent supérieur au seuil de confiance annoncé.

Dans tous les cas, la nature asymptotique des résultats et l'hypothèse sous-jacente de normalité[1] font considérer les résultats comme indicatifs.

A1.9.5 Les techniques de *bootstrap*

Face aux résultats d'une analyse factorielle, certaines questions sur la validité des axes obtenus se posent naturellement : Existe-t-il des critères pour tester la stabilité d'une structure et la valider ? Quelle est la part de l'échantillonnage des individus mais aussi, notion plus difficile, celle du choix ou de la sélection des variables ?

Nous avons vu au chapitre 2 que pour tenter de répondre partiellement à ces questions, on avait recours aux méthodes empiriques de validation. Elles consistent à perturber le tableau initial par des ajouts ou retraits

1. Muirhead (1982) a montré que l'hypothèse d'existence des quatre premiers moments pour la loi théorique de l'échantillon suffisait pour valider ces intervalles.

d'éléments du tableau, individus ou variables (poids, codage, etc.). L'hypothèse est la suivante : si les perturbations effectuées sur les échantillons n'affectent pas les configurations observées dans les sous-espaces, celles-ci sont supposées stables et la structure mise en évidence est alors « significative ».

Les méthodes de reéchantillonnage se proposent de systématiser cette démarche[1]. Celle du *bootstrap*, non paramétrique, est bien adaptée au problème de la validité des formes observées dans un plan factoriel ; elle calcule, à partir de simulations, des zones de confiance pour les positions des points-lignes et des points-colonnes.

- *Principe du bootstrap*

La technique du *bootstrap*, introduite par Efron (1979), consiste à simuler *s* (*s* est généralement supérieur à *30*) échantillons de même taille *n* que l'échantillon initial. Ils sont obtenus par tirage au hasard *avec remise* parmi les *n* individus observés au départ, ceux-ci ayant tous la même probabilité *1/n* d'être choisis. Certains individus apparaîtront plusieurs fois et auront de ce fait un poids élevé *(2, 3,...)* alors que d'autres seront absents (poids nul).

Cette méthode est employée pour analyser la variabilité de paramètres statistiques simples en produisant des intervalles de confiance de ces paramètres. Elle peut aussi être appliquée à de nombreux problèmes pour lesquels on ne peut pas estimer analytiquement la variabilité d'un paramètre. Ceci est le cas pour les caractéristiques des méthodes multidimensionnelles où les hypothèses de multinormalité sont rarement vérifiées. L'analyse en composantes principales est un domaine d'application qui a donné à un grand nombre de travaux utilisant les méthodes de reéchantillonnage de *bootstrap*.

Prenons l'exemple de l'estimation du coefficient de corrélation *r* entre deux variables ou entre une variable et un facteur. Le principe consiste à calculer le coefficient de corrélation pour chaque échantillon répliqué (pour lequel on effectue un tirage avec remise des *couples* d'observations). On établit alors la distribution des fréquences du coefficient de corrélation (représentée par l'histogramme des *s* valeurs du coefficient *r* correspondant aux *s* réplications). Puis on calcule à partir de

1. Ce sont des méthodes de calculs intensifs qui reposent sur des techniques de simulations d'échantillons à partir d'un seul échantillon. Rendues possibles par la puissance de calcul des ordinateurs, ces techniques se substituent dans certains cas aux procédures plus classiques qui reposent sur des hypothèses contraignantes. Elles sont les seules procédures possibles lorsque la complexité analytique du problème ne permet pas d'inférence classique.

l'histogramme la probabilité pour que le coefficient de corrélation d'un échantillon soit compris dans différentes fourchettes de valeurs définissant ainsi les intervalles de confiance. On obtient une estimation de la précision de la valeur de *r* obtenue sur l'échantillon de base sans faire l'hypothèse d'une distribution normale des données. Les bornes de l'intervalle de confiance peuvent être estimées directement par les quantiles de la distribution simulée.

Pour estimer les coordonnées factorielles issus d'une analyse en composantes principales, le principe est le même que pour le coefficient de corrélation ; on effectue sur chaque échantillon simulé, une analyse en composantes principales puis on établit une distribution de fréquences pour chacune des composantes[1].

La méthode de *bootstrap* donne dans la plupart des cas une bonne image de la précision statistique de l'estimation sur un échantillon. Les recherches théoriques menées par Efron, en particulier, montrent que, pour de nombreux paramètres statistiques, l'intervalle de confiance correspondant à la distribution simulée par *bootstrap* et celui correspondant à la distribution réelle sont généralement de même amplitude.

- *Mise en œuvre et calcul des zones de confiance*

Il existe plusieurs procédures pour tester, par la méthode de bootstrap, la stabilité des coordonnées factorielles. Gifi (1981), Meulman (1982), Greenacre (1984) ont réalisé des premiers travaux dans le contexte de l'analyse des correspondances simples ou multiples. Dans le cas de l'analyse en composantes principales, Diaconis et Efron (1983), Holmes (1989), Stauffer et *al*. (1985), Daudin et *al*. (1988) ont posé le problème du choix du nombre d'axes pertinent et ont proposé des intervalles de confiance pour les points du sous-espace défini par les principaux axes. Les paramètres correspondant sont calculés à partir des échantillons répliqués et supposent des contraintes qui dépendent de ces échantillons.

Pour pallier ces difficultés, il faut se référer à un espace factoriel commun. Plusieurs variantes sont possibles.

Nous nous sommes basés dans le chapitre 2 sur deux techniques appelées ici le *bootstrap total* et le *bootstrap partiel*.

Le *bootstrap total* consiste à réaliser autant d'analyses en composantes principales qu'il y a de réplications, moyennant une série

1. On trouvera des compléments sur l'intérêt et les limites de cette méthode dans les travaux de Diaconis et Efron (1983) et de Young (1994).

de transformations afin de retrouver des axes homologues au cours des diagonalisations successives des *s* matrices de corrélation répliquées $\mathbf{C}_k$ ($\mathbf{C}_k$ correspond à la *k-ème* réplication). Ces transformations sont des changements de signe des axes, rotations ou permutations d'axes. Cette méthode, proposée par Milan et Whittaker (1995) est en défaut s'il existe des valeurs propres très voisines.

Dans le bootstrap partiel, proposé par Greenacre (1984) dans le cas de l'analyse des correspondances, il n'est pas nécessaire de calculer les valeurs et vecteurs propres pour l'ensemble des simulations : les axes principaux calculés sur les données originales non perturbées, jouent un rôle privilégié (la matrice des corrélations initiale **C** est en effet l'espérance mathématique des matrices perturbées $\mathbf{C}_k$).

Le bootstrap partiel se fonde sur la projection en tant qu'*éléments supplémentaires* des points répliqués sur les sous-espaces de référence fournis par les axes principaux de la matrice de corrélation **C=X'X**, provenant de l'échantillon initial, donnés par :

$$\mathbf{u}_q = \frac{1}{\sqrt{\lambda_q}} \mathbf{X}'\mathbf{v}_q$$

où $\mathbf{u_q}$, $\mathbf{v_q}$ sont respectivement les *q-èmes* vecteurs propres de **X'X** et **XX'** et λ_q la valeur propre associée.

La projection[1] de la *k-ème* réplication des *m* variables (mots) est donnée par le vecteur $\mathbf{u}_q(k)$ de $\mathbb{R}^m$ tel que :

$$\mathbf{u}_q(k) = \frac{1}{\sqrt{\lambda_q}} \mathbf{X}'\mathbf{D}_k\mathbf{v}_q$$

et $\mathbf{D_k}$ désigne la matrice diagonale (*n, n*) des *poids bootstrap* associée à la *k-ème* réplication[2].

Dans le cas du bootstrap partiel, les analyses des matrices $\mathbf{C}_k$ ne sont en aucun cas nécessaires puisque les vecteurs propres sont obtenus à partir de l'analyse en composantes principales de la matrice **C**.

---

1. La projection des réplications Bootstrap, dans le contexte de l'analyse en composantes principales, consiste à utiliser le fait que la coordonnée d'une variable sur un axe factoriel n'est autre que son coefficient de corrélation avec la variable « coordonnées des individus sur l'axe ». On calcule donc les réplications de ce coefficient, ce qui revient à repondérer, pour chaque réplication, les individus avec les *poids Bootstrap* qui caractérisent un tirage sans remise. On obtient, comme sous-produit, des réplications de la variance sur l'axe, qui sont évidemment distinctes de ce que seraient des réplications des valeurs propres.

2. *Cf.* Chateau et Lebart (1996).

La variabilité bootstrap s'observe donc mieux sur le repère fixe initial, qui est d'ailleurs le moins mauvais, étant le seul à n'avoir pas été perturbé. Cette technique, éprouvée empiriquement, répond parfaitement aux préoccupations des utilisateurs dans le cas de l'analyse en composantes principales.

- *Bootstrap sur l'ensemble des variables*

Classiquement les réplications sont obtenues par des tirages avec remises dans l'ensemble des *n* individus. Pour tester la stabilité des structures vis-à-vis de l'ensemble des mots, nous proposons de répliquer cet ensemble par la méthode du *bootstrap total*.

Nous supposons ainsi implicitement que l'ensemble des mots du questionnaire constitue un échantillon de *m* mots extrait aléatoirement de l'ensemble des mots « sémiométrisables » de la langue française.

Nous cherchons à perturber cet échantillon de mots selon les mêmes principes que le *bootstrap* opéré sur les individus.

Pour cela, on appelle $\mathbf{B_k}$ la matrice diagonale (*m, m*) dont les éléments diagonaux sont les poids des mots de la *k-ème* réplication Bootstrap (*1,0,2,0,…*). La matrice **X** d'ordre (*n, n*) initiale étant supposée centrée, la matrice à diagonaliser est la matrice $\mathbf{T_k}$ qui vaut :

$$\mathbf{T}_k = \mathbf{X}\mathbf{B}_k\mathbf{X'} = \mathbf{X}\mathbf{B}_k^{1/2}\mathbf{B}_k^{1/2}\mathbf{X'}$$

On obtient donc :

$$\mathbf{X}\mathbf{B}_k\mathbf{X'}\mathbf{v}_q(k) = \lambda_q\mathbf{v}_q(k)$$

en multipliant chaque terme par $\mathbf{B}_k^{1/2}\mathbf{X'}$ on a :

$$\mathbf{B}_k^{1/2}\mathbf{X'}\mathbf{X}\mathbf{B}_k^{1/2}\mathbf{B}_k^{1/2}\mathbf{X'}\mathbf{v}_q(k) = \lambda_q\mathbf{B}_k^{1/2}\mathbf{'X'}\mathbf{v}_q(k)$$

et en posant $\mathbf{u}_q(k) = \mathbf{B}_k^{1/2}\mathbf{X'}\mathbf{v}_q(k)$ alors :

$$\mathbf{B_k}^{1/2}\mathbf{X'}\mathbf{X}\mathbf{B_k}^{1/2}\mathbf{u}(k) = \lambda_q\mathbf{u}(k)$$

$\mathbf{T}_k=\mathbf{X}\mathbf{B}_k^{1/2}\mathbf{B}_k^{1/2}\mathbf{X'}$ a les mêmes valeurs propres non nulles que la matrice $\mathbf{T}_k^* = \mathbf{B}_k^{1/2}\mathbf{X'}\mathbf{X}\mathbf{B_k}^{1/2}$. On diagonalisera la matrice $\mathbf{T}_k^*$ de dimension (*m, m*)

En pratique, on remplace les *poids bootstrap* nuls par des poids infinitésimaux, de façon à ce que les variables absentes d'une réplication apparaissent quand même avec le statut de variable supplémentaire.

Cette épreuve de validation est évidemment très sévère. On montre en effet que le tirage sans remise suscite approximativement, en moyenne, l'abandon d'un tiers des éléments (ici, des mots !) à chaque réplication.

La figure 2.10 de la section 2.4 du chapitre 2 illustre les effets de ces fortes perturbations sur la position des points variables.

Tous les calculs des chapitres 1 à 5 (analyses factorielles diverses, classifications, cartes de Kohonen, zones de confiances *bootstrap*), ont été réalisés à l'aide du logiciel académique *DtmVic* spécialisé dans la *fouille de données numériques et textuelles*.