MATHÉMATIQUES ET SCIENCES HUMAINES

JEAN-FRANÇOIS BALLIF
GEORGES LERESCHE

Paramétrisation et analyse des correspondances

Mathématiques et sciences humaines, tome 65 (1979), p. 23-50
<http://www.numdam.org/item?id=MSH_1979__65__23_0>

Math. Sci. hum. (17^{e} année, n°65, 1979, p.23-50)

PARAMETRISATION ET ANALYSE DES CORRESPONDANCES

Jean-François BALLIF et Georges LERESCHE [1]

1. INTRODUCTION

Le questionnaire est, depuis longtemps, le principal outil d'investigation en sciences humaines. Depuis l'avènement de l'ordinateur se sont développées des méthodes de traitement d'autant plus puissantes que la longueur ou la difficulté des calculs ne sont plus des obstacles, pas plus que le nombre des variables (questions) ou que le grand effectif des échantillons étudiés, ce dernier élément étant un facteur important pour assurer la validité des résultats. La liberté méthodologique qui en découle rend toujours plus nécessaire la réflexion sur les méthodes elles-mêmes et sur les liens qui existent entre elles.

Instruments destinés à "mesurer" des dimension humaines, les questionnaires se distinguent des instruments ou appareils de mesure physique par l'absence d'étalon a priori. Leurs utilisateurs sont donc fréquemment confrontés à des problèmes de métrique. Un certain nombre de méthodes ont été proposées pour tourner cette difficulté, méthodes qui dépendent d'ailleurs souvent de la forme du questionnaire.

Nous considérons ici la situation suivante: une population P est soumise à un questionnaire consistant en p items I_1, I_2, ..., I_p,

---

1. Université de Lausanne, chaire de Mathématiques des Sciences humaines;
recherches sur *la genèse du choix professionnel chez les futurs bacheliers* (FNRS, n° 1'7430.72 SR; requérants: les Prof. J.-B. Dupont et G. Leresche) et sur *l'évaluation des psychothérapies brèves* (en collaboration avec la policlinique psychiatrique universitaire, Prof. P.-B. Schneider).

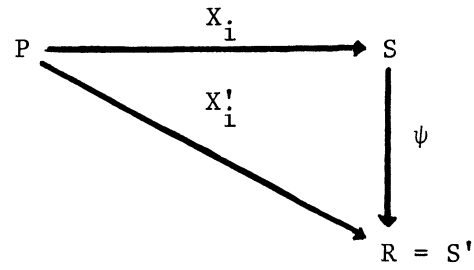
formant l'ensemble I , pour lesquels on propose le même ensemble de modalités de réponse $S = \{ S_1 , S_2 , \ldots , S_r \}$. Chaque item constitue ce que nous appellerons une *observable*. Selon la nature de S , les p observables pourront être *nominales* - lorsque S est sans structure a priori -, *ordinales* - lorsque S admet a priori une structure d'ordre -, *numériques* - lorsque S peut se plonger dans $\mathbb{R}$ (muni de sa structure).

Souvent S est une échelle destinée à mesurer un intérêt, une aptitude, un degré d'engagement, une fréquence d'utilisation etc... et admet par conséquent une structure d'ordre, laquelle peut cependant n'être que partielle, comme nous le voyons plus loin (paragraphe 5).

On aimerait voir dans quelles conditions on peut utiliser l'analyse multivariée, et notamment l'analyse factorielle, pour traiter un tel questionnaire. Or l'analyse multivariée présuppose en général que les observables soient numériques; nous devons donc définir une application ψ de S dans $\mathbb{R}$ afin de transformer nos observables ordinales en observables numériques.

Remarquons ici que notre problématique se rapproche de celle qui a débouché sur les nombreuses méthodes dites de *multidimensional scaling*, dans lesquelles une phase de "métrisation" des données précède celle de réduction ou d'analyse[1].

Paramétrer les items d'un questionnaire, c'est en fait considérer que S constitue une partition en classes d'un ensemble S' - ensemble des réactions possibles aux items-stimuli -, et attribuer à chaque classe une valeur centrale. Si l'on appelle X_i l'observable ordinale fournie par le i ème item, il lui correspond une observable implicite X_i' , numérique, dont l'ensemble des scores S' s'identifie à $\mathbb{R}$. On peut donc considérer le schéma suivant:



1. Voir par exemple Torgerson (8), pages 247 à 297.

Le problème est de définir ψ de façon que les observables paramétrées $\psi o X_i$ soient compatibles d'une manière optimale avec les hypothèses éventuelles faites sur les observables sous-jacentes X_i' ou avec les objectifs descriptifs qu'on s'est fixés.

Notons ici qu'en envisageant une paramétrisation de l'ensemble des modalités de réponse, indépendamment des items, nous supposons qu'une modalité est appréhendée de la même façon, quel que soit l'item qui lui est confronté. Cette hypothèse, si elle n'est pas formellement réaliste, - une modalité pouvant avoir, face à certains items, des connotations psychologiques différentes - n'est cependant pas trop réductive, pour autant qu'on suppose une *homogénéité de signification* des questions et une *homogénéité de compréhension* des modalités de réponse. D'autre part la variation d'interprétation des différentes modalités est certainement plus grande encore entre les individus qu'entre les items. Pourtant, notre objectif étant de comparer les individus aussi bien que les items, la nécessité d'un référentiel commun nous conduit à adopter une paramétrisation indépendante de ces deux éléments.

Par cette position nous nous distinguons du point de vue de Guttman (4) et de sa "scale analysis" (voir aussi Torgerson (8), pages 338 à 345), sur laquelle nous reviendrons et qui envisage la paramétrisation de chaque modalité pour chaque item séparément. Il est vrai que dans les situations envisagées par Guttman on ne suppose pas un ensemble S de modalités de réponse commun à l'ensemble des items; chacun de ceux-ci peut avoir son propre ensemble de modalités. D'autre part l'objectif de Guttman est bien plutôt un classement des sujets sur une échelle unidimensionnelle qu'une paramétrisation des modalités pour un usage ultérieur. Dans le cas de Guttman, la paramétrisation est une fin en soi; pour nous elle est une étape sur la voie des traitements multivariés.

En revanche, cette position nous rapproche des méthodes issues des modèles généraux de Thurstone (La loi du jugement cathégorique, voir Thorgerson (8), chapitre 10), en particulier de celles des "intervalles successifs" (voir par exemple Gulliksen (3)), lesquelles cherchent à assigner à chaque modalité de réponse - indépendamment des items - un intervalle réel. Il faut noter cependant que ces dernières méthodes reposent toutes sur des hypothèses relatives aux distributions des observables sous-jacentes X_i' - supposées généralement gaussiennes.

Or il peut arriver que de telles hypothèses ne soient pas soutenables - différences évidentes entre distributions d'items, influence prépondérante d'éléments extérieurs s'opposant au caractère normal des distributions... Dans ces conditions la paramétrisation peut être basée sur des objectifs descriptifs.

La paramétrisation étant une étape dans l'analyse des données, donc dans la description d'une certaine réalité reflétée par un questionnaire, on peut choisir, parmi les applications de S dans R , celle qui réalise le mieux certains éléments de cette description. On se trouve ici dans la situation du biologiste qui, désirant étudier au microscope une coupe d'un certain tissu, va utiliser différentes colorations, afin d'en mieux séparer les diverses composantes; comme nous, il confère à ses données un éclairage favorable, lié aux objectifs de sa recherche.

Notre propos, ici, est de présenter une méthode de paramétrisation liée à un objectif descriptif. En elle-même cette méthode n'est pas nouvelle (voir par exemple Nishisato et Arri (6)), mais, ce qui nous intéressera ici, après en avoir donné le développement théorique, c'est de mettre en évidence ses liens avec l'analyse des correspondances[1], ainsi que leurs conséquences sur le plan de l'interprétation.

Les réflexions qui vont suivre nous ont d'ailleurs été inspirées par la phrase suivante, lue dans le livre de Benzécri sur l'analyse des correspondances (1), à la page 484: *"Les distances entre notes dans le plan des deux premiers facteurs sont sans doute la meilleure mesure psychométrique de distance dont on puisse disposer."* Cherchant à approfondir cette proposition, nous sommes tombés sur la méthode que nous nous proposons d'étudier.

2. POUVOIR SEPARATEUR D'UNE PARAMETRISATION

L'un des objectifs descriptifs pouvant présider à une paramétrisation est de mettre en évidence, s'il existe, un *ordre* sur l'ensemble des items. Pour le décrire au mieux on utilisera la paramétrisation qui a ce que l'on pourrait appeler le meilleur *pouvoir séparateur*, en moyenne, sur les items. En effet, comme nous le voyons plus loin (paragraphe 3), s'il existe un ordre sur

---

1. Pour une description détaillée de cette méthode d'analyse des données, voir par exemple Benzécri (1) ou Lebart et Fénelon (5).

un ensemble d'observables paramétrées, c'est au niveau des moyennes que l'on reconnaît cet ordre; l'ordre, s'il existe, sera donc d'autant plus "visible" que les moyennes des items, après paramétrisation, seront plus dispersées.

Disposant pour p items - observables ordinales -, des effectifs des choix de chacune des r modalités de réponse, nous avons le tableau T :

$$T = \begin{array}{c} \begin{matrix} S_1 & S_2 & \cdots & S_r \end{matrix} \\ \begin{pmatrix} n_{11} & n_{12} & \cdots & n_{1r} \\ n_{21} & n_{22} & \cdots & n_{2r} \\ \vdots & \vdots & \vdots & \vdots \\ n_{p1} & n_{p2} & \cdots & n_{pr} \end{pmatrix} \end{array} \begin{matrix} I_1 \\ I_2 \\ \vdots \\ I_p \end{matrix}$$

où n_{ij} est l'effectif de la réponse S_j à l'item I_i .

Par rapport à l'ensemble des $N = \sum_i \sum_j n_{ij}$ réponses données au questionnaire, chaque couple (I_i , S_j) a une fréquence $p_{ij} = \frac{n_{ij}}{N}$.

Un item I_i a alors un poids $p_{i.} = \sum_{j=1}^{r} p_{ij}$ et une modalité S_j a une fréquence d'utilisation $p_{.j} = \sum_{i=1}^{p} p_{ij}$. Notons qu'en l'absence de non-réponses, chaque item a le même poids $\frac{1}{p}$.

Nous appellerons *distribution barycentrique* la distribution de fréquences $g_S = (p_{.1} , p_{.2} , \ldots , p_{.r})$ et *distribution marginale* la distribution des poids $g_I = (p_{1.} , p_{2.} , \ldots , p_{p.})$.

Le problème est de trouver une paramétrisation ψ rendant maximale la quantité $P = Var(M)$, variance de la suite $M = (m_1 , m_2 , \ldots , m_p)$ des moyennes des observables paramétrées $\psi o X_1 , \psi o X_2 , \ldots , \psi o X_p$, munie de la distribution marginale g_I .

Appelons b_j l'image $\psi(S_j)$ d'une modalité S_j et G la suite $(b_1 , b_2 , \ldots , b_r)$ munie de la distribution barycentrique g_S.

Nous considérerons les matrices suivantes:

$$A = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1r} \\ p_{21} & p_{22} & \cdots & p_{2r} \\ \vdots & \vdots & \vdots & \vdots \\ p_{p1} & p_{p2} & \cdots & p_{pr} \end{pmatrix}$$

matrice des fréquences de chaque couple (I_i , S_j) par rapport à l'ensemble des réponses données au questionnaire

$u_I = \begin{pmatrix} 1 & 1 & \dots & 1 \end{pmatrix}$ matrice-ligne du vecteur unité dans $\mathbb{R}^p$

$u_S = \begin{pmatrix} 1 & 1 & \dots & 1 \end{pmatrix}$ matrice-ligne du vecteur unité dans $\mathbb{R}^r$

$g_I = \begin{pmatrix} p_{1.} & p_{2.} & \dots & p_{p.} \end{pmatrix}$ matrice-ligne de la distribution marginale des poids des items

$g_S = \begin{pmatrix} p_{.1} & p_{.2} & \dots & p_{.r} \end{pmatrix}$ matrice-ligne de la distribution barycentrique

$$\Delta_I = \begin{pmatrix} p_{1.} & 0 & \dots & 0 \\ 0 & p_{2.} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & p_{p.} \end{pmatrix}$$ matrice diagonale de la distribution marginale

$$\Delta_S = \begin{pmatrix} p_{.1} & 0 & \dots & 0 \\ 0 & p_{.2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & p_{.r} \end{pmatrix}$$ matrice diagonale de la distribution barycentrique

$b = \begin{pmatrix} b_1 & b_2 & \dots & b_r \end{pmatrix}$ matrice-ligne de la paramétrisation

$m = \begin{pmatrix} m_1 & m_2 & \dots & m_p \end{pmatrix}$ matrice-ligne des moyennes des items après paramétrisation

Entre ces matrices nous avons les relations suivantes:

(1) $m = \Delta_I^{-1} A \, {}^t b$

(2) $\text{Moy}(M) = g_I \, m = g_I \, \Delta_I^{-1} A \, {}^t b = \text{Moy}(G)$

(3) $g_I = u_S \, {}^t A$

(4) $g_S = u_I A$

(5) $g_I = u_I \, \Delta_I$

(6) $g_S = u_S \, \Delta_S$

D'autre part, si ψ et ψ' sont deux paramétrisations, nous avons, de façon immédiate, les deux implications suivantes:

(7) $\forall \alpha \in \mathbb{R}$, si $\psi = \alpha \psi'$ alors $P = \alpha^2 P'$ et $Var(G) = \alpha^2 Var(G')$

(8) $\forall \alpha \in \mathbb{R}$, si $\psi = \psi' + \alpha$ alors $P = P'$, $Var(G) = Var(G')$ et

$Moy(M) = Moy(M') + \alpha = Moy(G) = Moy(G') + \alpha$

Notre problème sera donc de maximiser P en gardant constante la variance de G ainsi que sa moyenne, ce dernier point équivalant à garder constante la moyenne de M . Nous nous proposons de rechercher la solution satisfaisant à la condition: G centrée et réduite; ou, si l'on préfère: $Var(G) = 1$ et $Moy(G) = 0$.

Nous devons donc maximiser la quantité

(9) $P = Var(M) = m \Delta_I {}^t m = b {}^t A \Delta_I^{-1} A {}^t b$ sous les conditions

(10) $Var(G) = b \Delta_S {}^t b = 1$

(11) $Moy(G) = g_S {}^t b = 0$

Notons que ceci revient à maximiser, sous la condition $Moy(G) = 0$, le rapport $\frac{V_E}{V_T}$, où V_E et V_T peuvent s'interpréter, dans un certain sens, comme des variances, l'une, V_E , entre les items, l'autre, V_T , totale. En effet, si l'on considère la variable aléatoire $X : \Omega \to \mathbb{R}$, où Ω est l'ensemble $I \times S$ muni des probabilités p_{ij} et X est définie comme suit:

$X (I_i , S_j) = b_j$, alors on a:

$V_T = Var(X) = Var(G)$

D'autre part, $Var(X)$ peut se décomposer en

$$Var(X) = \sum_{i=1}^{p} \sum_{j=1}^{r} p_{ij} b_j^2 = \sum_{i=1}^{p} \sum_{j=1}^{r} p_{ij} (b_j - m_i + m_i)^2 =$$

$$= \sum_{i=1}^{p} \left[\sum_{j=1}^{r} p_{ij} (b_j - m_i)^2 + \sum_{j=1}^{r} p_{ij} m_i^2 + 2m_i \underbrace{\left(\sum_{j=1}^{r} p_{ij} b_j - \sum_{j=1}^{r} p_{ij} m_i \right)}_{= 0} \right]$$

Si l'on appelle σ_i^2 la variance de l'item I_i pour la paramétrisation ψ, on a:

$$\sigma_i^2 = \sum_{j=1}^{r} \frac{p_{ij}}{p_{i.}} (b_j - m_i)^2 \quad \text{et}$$

$$\text{Var}(X) = \underbrace{\sum_{i=1}^{p} p_{i.} \sigma_i^2}_{\text{variance moyenne à l'intérieur des items}} + \underbrace{\text{Var}(M)}_{\text{variance entre les items}}$$

On peut donc écrire $V_T = V_I + V_E$ ou, si l'on préfère, $\text{Var}(G) = V_I + P$, ce qui implique en particulier:

(12) $P \leq 1$ si $\text{Var}(G) = 1$

D'autre part, la relation (7) implique que maximiser P sous la condition (10) revient à maximiser le rapport $\frac{P}{\text{Var}(G)}$, soit $\frac{V_E}{V_T}$.

C'est cette interprétation qu'on trouve généralement chez les auteurs s'étant intéressés à cette méthode, et qu'ils ont baptisée "optimal scaling" (voir par exemple Nishisato et Arri (6)).

Certains, comme Bradley , Katti et Coons (2) maximisent le rapport $\frac{V_E}{V_I}$.

Revenant à notre problème nous posons:

$$b = b' \Delta_S^{-\frac{1}{2}}$$

On peut voir cette transformation comme le passage à une base orthonormée, de manière que la variance de G soit égale à $b' {}^t b'$.

(9) s'écrit alors:

(13) $$P = b' (\Delta_S^{-\frac{1}{2}} {}^t A \Delta_I^{-\frac{1}{2}}) (\Delta_I^{-\frac{1}{2}} A \Delta_S^{-\frac{1}{2}}) {}^t b'$$

tandis que (10) et (11) deviennent:

(14) $$b' {}^t b' = 1$$

(15) $$g_S \Delta_S^{-\frac{1}{2}} {}^t b' = 0$$

Il suffit maintenant de poser:

$$C = \Delta_I^{-\frac{1}{2}} A \Delta_S^{-\frac{1}{2}}$$

pour se trouver devant la situation tout-à-fait classique de maximiser $P = b' {}^tCC\, {}^tb'$ sous la condition (14), avec la condition supplémentaire (15).

On sait que P sera maximum sous la condition (14) si l'on prend pour b' le vecteur propre de ${}^tCC$ correspondant à la plus grande valeur propre λ, et que, de plus, cette valeur maximale de P sera précisément λ (voir par exemple Lebart et Fénelon (5), page 200). Assurons-nous de la compatibilité de ce résultat avec la condition (15).

Si b' est vecteur propre de ${}^tCC$ nous avons:

${}^tC\, C\, {}^tb' = \lambda\, {}^tb'$ ce qui peut s'écrire:

${}^tA\, \Delta_I^{-1}\, A\, {}^tb = \lambda\, \Delta_S\, {}^tb$ qu'on prémultiplie par u_S :

$u_S\, {}^tA\, \Delta_I^{-1}\, A\, {}^tb = \lambda\, u_S\, \Delta_S\, {}^tb$ devient, en utilisant la relation (3):

$g_I\, \Delta_I^{-1}\, A\, {}^tb = \lambda\, g_S\, {}^tb$ puis la relation (5):

$u_I\, A\, {}^tb = \lambda\, g_S\, {}^tb$ et encore la relation (3):

$g_S\, {}^tb = \lambda\, g_S\, {}^tb$ ce qui peut s'écrire

$g_S\, \Delta_S^{-\frac{1}{2}}\, {}^tb' = \lambda\, g_S\, \Delta_S^{-\frac{1}{2}}\, {}^tb'$

Ainsi, pour toute valeur propre différente de 1, la condition (15) est compatible avec le fait que b soit vecteur propre de ${}^tCC$.

Nous devons encore remarquer que la matrice ${}^tCC$ qui s'écrit:

$$({}^tC\, C)_{jk} = \sum_{i=1}^{p} \frac{p_{ij}\, p_{ik}}{p_{i.} \sqrt{p_{.j}\, p_{.k}}}$$

est celle dont les vecteurs propres nous fournissent, dans l'analyse des correspondances du tableau T - ou de la matrice A -, les facteurs sur l'ensemble des colonnes: si v est un vecteur propre ligne de ${}^tCC$ et si ${}^tvv = 1$, alors on obtient le facteur ϕ réduit correspondant par:

$$\phi = v\, \Delta_S^{-\frac{1}{2}}$$

Nous avons donc établi que *la paramétrisation offrant le meilleur pouvoir séparateur en moyenne sur les items, et correspondant à la condition " G centrée et réduite", est donnée par le premier facteur non trivial* (c'est-à-dire relatif à la plus grande valeur propre différente de 1) *réduit de l'analyse des correspondances appliquée au tableau* T *de distribution des réponses.*

D'autre part, le *principe barycentrique* (voir Benzécri (1), tome I, page 25) nous dit que *le premier facteur, non réduit, sur l'ensemble des items nous fournit la matrice* m *des moyennes des items, pour cette paramétrisation.*

Ce rapprochement entre l'analyse des correspondances et les méthodes dites d'"optimal scaling" nous paraît être d'un grand intérêt, de par le fait que l'un et l'autre s'en trouvent éclairés d'un jour nouveau. En effet, si certains résultats exposés dans le livre de Benzécri sur l'analyse des correspondances (1) permettent une interprétation plus approfondie de notre paramétrisation - nous pensons notamment à certain théorème relatif à l'ordre latéral sur lequel nous nous arrêtons plus loin -, cette dernière donne un sens nouveau à la métrique du *chi carré*, sur laquelle repose l'analyse des correspondances.

3. SUR L'INTERET DE LA RELATION ENTRE POUVOIR SEPARATEUR ET ANALYSE DES CORRESPONDANCES.

3.1. Ordre sur les items et ordre sur les modalités de réponse

S'il est toujours possible d'obtenir la paramétrisation rendant maximal le pouvoir séparateur, il peut arriver qu'elle n'ait guère de sens: destinée à mettre en évidence, d'une manière optimale, un ordre sur l'ensemble des items, elle n'a sa raison d'être que lorsque cet ordre est présent naturellement, d'une manière intrinsèque, dans les données. On devra donc s'assurer que l'ordre exhibé n'est pas un artefact lié à la méthode, mais bien un élément de la structure des données.

Avant de parler directement de ce qu'est un ordre naturel sur un ensemble d'items, il nous faut aborder la notion plus générale d'*ordre entre distributions de probabilités - ou de fréquences - sur un ensemble ordonné.*

Considérons un ensemble $S = \{ S_1 < S_2 < \ldots < S_r \}$ ordonné, et sur S deux distributions de probabilités a et b . Pour simplifier l'écriture

nous adoptons les notations :

$a (S_j) = a_j$

$b (S_j) = b_j$

Appelons F_a , respectivement F_b , l'application des probabilités cumulées de a , respectivement b ;

$$F_a : S \to [0 , 1] ; \quad F_a (S_j) = \sum_{S_k \leqq S_j} a_k$$

$$F_b : S \to [0 , 1] ; \quad F_b (S_j) = \sum_{S_k \leqq S_j} b_k$$

L'ordre usuel entre F_a et F_b , défini par :

$F_a \geqq F_b$ si et seulement si $\forall j$, $F_a (S_j) \geqq F_b (S_j)$

nous fournit un ordre sur les distributions de probabilités.

De la même manière, si S est un ensemble de modalités de réponse et si a et b sont les distributions de fréquences relatives à deux items I_a et I_b , nous pouvons définir un ordre $\leqq i$ entre les items par :

$I_a \leqq i \; I_b$ (I_b "à droite" de I_a) si et seulement si $F_a \geqq F_b$.

Si, de plus, on considère une paramétrisation ψ croissante de S , nous avons l'implication :

$$I_a \leqq i \; I_b \implies m_a \leqq m_b$$

où m_a et m_b sont les moyennes de I_a et I_b , calculées sur la base de la paramétrisation ψ .

Considérons encore X_a et X_b , les observables numériques sous-jacentes, associées - comme expliqué au paragraphe 1 - à I_a et I_b . Si X_a et X_b ne diffèrent que par la moyenne (exemple : X_a et X_b sont deux observables distribuées normalement, d'écart-type σ et de moyennes m_{X_a} et m_{X_b} respectivement), alors nous avons les deux équivalences :

$$m_{X_a} \leqq m_{X_b} \iff I_a \leqq i \; I_b \iff m_a \leqq m_b$$

Nous voyons donc que l'ordre défini à partir des fonctions de fréquences cumulées est à rapprocher de l'ordre entre les moyennes d'observables paramétrées, pour autant que la comparaison des moyennes ait tout son sens, c'est-à-dire que les autres paramètres de distribution soient invariants entre les items. Il faut toutefois remarquer que cette invariance n'est pas une condition nécessaire à la double équivalence ci-dessus; lorsqu'un ensemble d'items $\{ I_1 , I_2 , \ldots , I_p \}$ est tel que:

$\forall i , j \in \{ 1 , 2 , \ldots , p \} ; \forall \psi$, paramétrisation croissante de S ;

$$m_{X_i} \leq m_{X_j} \iff I_i \leq_i I_j \iff m_i \leq m_j$$

(où m_{X_i} et m_{X_j} sont les moyennes des observables numériques sous-jacentes associées à I_i et I_j respectivement, et m_i et m_j sont les moyennes, pour la paramétrisation ψ , des items I_i et I_j respectivement) nous dirons que les items $I_1 , I_2 , \ldots , I_p$ ne diffèrent *essentiellement* que par la moyenne.

Il nous faut encore dire que l'ordre entre distributions de probabilités, tel que nous l'avons défini à partir des fonctions de probabilités cumulées, n'est autre que l'ordre dit *latéral*, introduit par Benzécri (1) (tome I, page 261).

L'équivalence de ces deux ordres est intéressante, puisqu'elle nous permet d'appliquer à notre problème de paramétrisation le théorème suivant, démontré dans Benzécri (1) (tome I, page 279), qui peut s'énoncer ainsi:

Si les colonnes (respectivement les lignes) d'un tableau de correspondances représentent les éléments d'un ensemble ordonné S , *et que sur l'ensemble* L *des lignes (respectivement des colonnes) - dont les éléments peuvent être considérés comme des distributions de probabilités sur* S -, *l'ordre latéral est total, alors le facteur sur l'ensemble* S *(respectivement* L *), relatif à la plus grande valeur propre différente de 1 , est une fonction croissante de* S *(respectivement* L *).*

Transposant ce résultat à notre problème de paramétrisation, et nous appuyant sur les considérations faites en début de paragraphe, nous obtenons la proposition suivante:

Si l'ensemble $I = \{ I_1 , I_2 , \ldots , I_p \}$ *des items d'un questionnaire, utilisant l'ensemble ordonné* $S = \{ S_1 , S_2 , \ldots , S_r \}$ *de modalités de réponse, est tel que lui corresponde un ensemble* $X = \{ X_1 , X_2 , \ldots , X_p \}$ *d'observables numériques sous-jacentes - dont les* S_j *constituent une partition en classes - ne différant essentiellement que par la moyenne, alors l'analyse des correspondances appliquée au tableau de distribution des réponses fournit des facteurs de rang 1 qui redonnent l'ordre des modalités et l'ordre des moyennes.*

Ce résultat est d'importance puisqu'il constitue, d'une certaine manière, une généralisation de celui obtenu par Guttman (4) sur l'analyse des scalogrammes (tableau dont chaque ligne est composée uniquement de 1 et de 0); Guttman a montré en effet que lorsque, après permutation de lignes et de colonnes, un scalogramme - ou plutôt les 1 qu'il contient - a la forme d'un parallélogramme, alors l'analyse d'un tel scalogramme - qui n'est autre, comme nous le voyons plus loin, qu'une analyse des correspondances - fournit des facteurs de rang 1 qui sont des fonctions croissantes de l'ensemble des lignes (respectivement des colonnes), muni de l'ordre défini par le parallélogramme. Nous verrons plus loin (paragraphe 5) qu'il n'est pas impossible que les résultats obtenus par Guttman sur les facteurs de rang supérieur soient généralisables de la même façon.

3.2. La métrique du chi carré

Etant donné trois distributions de fréquences ou de probabilités a , b , c sur un ensemble $S = \{ S_1 , S_2 , \ldots , S_r \}$, on appellera distance du *chi carré* de centre a entre b et c , que l'on notera $d_a (b , c)$, la quantité:

$$d_a (b , c) = \left[\sum_{j=1}^{r} \frac{(b_j - c_j)^2}{a_j} \right]^{\frac{1}{2}}$$

(où l'on note a_j , b_j , c_j les fréquences - ou probabilités - $a (S_j)$, $b (S_j)$, $c (S_j)$)

Rappelons que l'analyse des correspondances est une méthode d'analyse factorielle qui mesure les distances entre lignes (respectivement colonnes) d'un tableau de dépendance par la métrique du chi carré de centre la ligne (respectivement la colonne) marginale du tableau, qui constitue d'ailleurs pour cette métrique le barycentre des distributions des lignes (res-

pectivement des colonnes).

Cette métrique, qui n'est pas toujours clairement justifiée, trouve ici une nouvelle interprétation: c'est celle qui apparaît naturellement lorsqu'on cherche à donner aux éléments des colonnes d'un tableau de contingence, des poids b_j dont les sommes, pondérées par les p_{ij} - éléments du tableau exprimés en fréquences - et en nombre égal à celui des lignes, soient de variance maximale.

Il est bon de savoir aussi que, lorsqu'on pratique l'"optimal scaling" - ou paramétrisation à meilleur pouvoir séparateur -, on mesure, en fait, les distances entre modalités de réponse par la métrique du chi carré de centre g_I , distribution marginale; ensuite, faisant l'hypothèse que l'ensemble S des modalités de réponse est unidimensionnel, on cherche le meilleur ajustement, au sens des moindres carrés, de ces distances, mesurées sur nos données, au modèle unidimensionnel.

4. LA THEORIE DE GUTTMAN

Dans ce qu'il appelle "scale analysis" Guttman (4) considère la situation suivante: N sujets sont soumis à un questionnaire - ou test - de p items I_1 , I_2 , ... I_p ; à chaque item I_i est associé un ensemble S_i de r_i modalités de réponse S_{i1} , S_{i2} , ... , S_{ir_i} ; on a donc au total $r = \sum_{i=1}^{p} r_i$ modalités. Les données constituent donc un tableau T à N lignes et r colonnes - que nous désignerons, pour simplifier les notations, par S_1 , S_2 , ... , S_r - dont un élément t_{hj} vaut 1 si le $h^{ème}$ sujet a choisi la modalité S_j et 0 sinon.

Guttman se propose alors de trouver une paramétrisation $\psi : S \to \mathbb{R}$ (où $S = \bigcup_{i=1}^{p} S_i$) qui soit la meilleure possible dans le sens suivant: pour chaque personne simultanément, les modalités retenues par cette personne devront avoir par ψ des images aussi proches que possibles les unes des autres, et aussi différentes que possible des modalités non retenues. En d'autres termes il cherche à maximiser la variance de l'observable construite

$$X(h) = \frac{1}{p} \sum_{j=1}^{r} t_{hj} \psi(S_j)$$

par rapport à la variance totale V_t du tableau T pour la paramétrisation ψ :

$$V_t = \frac{1}{Nr} \sum_{h=1}^{N} \sum_{j=1}^{r} t_{hj} \psi^2 (S_j) - \left[\frac{1}{Nr} \sum_{h=1}^{N} \sum_{j=1}^{r} t_{hj} \psi (S_j) \right]^2$$

ce qui revient à soumettre à l'analyse des correspondances le tableau T .

Guttman s'intéresse alors à ce qui se passe lorsque l'ensemble des items, ou plutôt celui des modalités de réponse - chacune d'elles pouvant être considérée comme un item dichotomique -, constitue ce qu'il appelle une échelle parfaite, c'est-à-dire que les éléments non nuls du tableau T forment - éventuellement après permutation de lignes et de colonnes - un parallélogramme; en d'autres termes, lorsque sur S , l'ordre[1] $\leqq^*$, défini par:

$S_j \leqq^* S_k$ si et seulement si $t_{hj} \leqq t_{hk}$, $\forall h$

est total, de même qu'est total sur l'ensemble $P = \{ a_1 , a_2 , \ldots , a_N \}$ des sujets, l'ordre:

$a_h \leqq^* a_q$ si et seulement si $t_{hj} \leqq t_{qj}$, $\forall j$

Il obtient les résultats suivants: le premier facteur sur S (respectivement P) est une fonction croissante de S (respectivement de P) muni de l'ordre $\leqq^*$. C'est ce que Guttman appelle la *composante métrique*. Sur S elle constitue une paramétrisation. Sur P elle donne, à une constante multiplicative près, les scores obtenus par les sujets, pour cette paramétrisation, après sommation sur l'ensemble des items.

Les facteurs suivants sont des oscillations successives; ainsi le second facteur est une fonction polynômiale de degré 2 , en $\cup$, du premier facteur; le troisième, une fonction polynômiale de degré 3 , de forme $\cup\cap$, du premier facteur; ainsi de suite. De plus, d'après Guttman, les extrema de ces fonctions, ou, si l'on préfère, leurs sommets, sont indépendants des items utilisés et des échantillons testés.

---

1. Il s'agit, rigoureusement, d'une relation de préordre; néanmoins, dans tout ce qui suit, comme dans le paragraphe 2, nous parlons systématiquement d'ordre, pour ne pas compliquer inutilement l'exposé, la différence entre ces deux notions ne jouant aucun rôle dans ce contexte.

Les facteurs 2 , 3 et 4 ont reçu de Guttman des interprétations précises. Ainsi le second facteur serait la *composante d'intensité*. Les renseignements fournis par ce facteur, sur P , seraient donc équivalents, en quelque sorte, à la réponse à la question: " Quelle est la force de vos sentiments là-dessus ? "

Guttman interprète le facteur 3 comme la *composante de clôture* (" A quel point êtes-vous sûr que votre opinion soit bien formée ? ").

Le facteur 4 , enfin, serait la *composante d'involution* (" Avez-vous dû réfléchir longtemps avant de répondre ? ").

La question que l'on peut se poser est la suivante: dans quelle mesure ces résultats sont-ils applicables à l'analyse d'un tableau de distribution des réponses à un questionnaire ? Nous avons déjà vu (paragraphe 3.1) que, dans le cas où l'ordre latéral sur les items est total, le premier facteur est une fonction croissante de S . Au vu de certains résultats d'analyses (voir paragraphe 5.1 et Benzécri (1), tome II, page 482), on constate que les facteurs de rang supérieur à 1 constituent, comme dans la situation envisagée par Guttman, des oscillations successives. Cette particularité nous donne un moyen de contrôler le caractère unidimensionnel de l'ensemble des modalités de réponse, et de celui des items par rapport aux modalités, ou, si l'on préfère, de confirmer la présence d'un ordre naturel sur l'ensemble des items.

5. APPLICATIONS

5.1 La paramétrisation d'un questionnaire de valeurs professionnelles

Dans le cadre d'une enquête sur la *genèse du choix professionnel des futurs bacheliers en Suisse romande*[1], une série de questionnaires et de tests ont été soumis à trois reprises, en 1974, '75 et '76, à un échantillon de cinq cents candidats au baccalauréat. Parmi les instruments utilisés, certains ont été élaborés spécialement pour cette recherche. C'est le cas notamment d'un questionnaire relatif aux valeurs professionnelles, comportant quarante items et dont nous reproduisons partiellement la première page ci-dessous.

---

1. Recherche effectuée par l'Institut de psychologie appliquée de l'Université de Lausanne, et financée par le FNRS (voir la note de la première page de l'article).

Que rechercheriez-vous ou refuseriez-vous dans une activité professionnelle? (mettez une croix dans la case correspondant à votre opinion)

	je le recherche	cela m'importe peu	je le refuse	je ne sais pas
1) utiliser pleinement mes aptitudes				
2) m'épanouir et me réaliser				

Notons que l'ensemble S des modalités de réponse fait de nos 40 items des observables nominales. En effet, si nous pouvons trouver dans S deux pôles fortement opposés - "je le refuse" et "je le recherche" -, tout ce que l'on peut dire des deux modalités intermédiaires est qu'elles se situent entre ces deux pôles.

Le problème de la paramétrisation se pose aux trois niveaux suivants:

1. Voulant analyser les relations existant entre nos 40 items, et, par la suite, remplacer ces derniers par un nombre réduit de scores factoriels, en vue d'intégrer les dimensions principales des valeurs professionnelles à l'ensemble des facteurs à prendre en compte dans l'analyse du processus du choix professionnel, nous devons transformer nos 40 observables nominales en observables numériques. Ne faisant aucune hypothèse sur la forme des distributions, nous cherchons la paramétrisation qui sépare au mieux les items, par leur moyenne, afin de mettre en évidence un éventuel ordre sur les items.

2. Si un ordre sur les items existe, reste-t-il stable d'une année à l'autre ou, au contraire, se modifie-t-il au cours des ans ? En d'autres termes, la paramétrisation optimale en 1974 garde-t-elle un sens en '75 ?

3. Pour la troisième prise d'information, en 1976 donc, nous avons - sur la demande des psychologues - modifié l'ensemble S des modalités de réponse. Nous reproduisons ci-dessous le début du questionnaire, tel qu'il se présentait pour cette troisième prise d'information.

Que rechercheriez-vous (un peu, beaucoup, passionnément, pas du tout) dans une activité professionnelle ?

	un peu	beaucoup	passionnément	pas du tout
1) utiliser pleinement mes aptitudes				

Se pose, dès lors, le problème de la "reliabilité" des paramétrisations. Dans quelle mesure peut-on trouver, pour 1976, une paramétrisation comparable à celles utilisées pour les années précédentes ?

Pour répondre à ces différentes questions, nous avons soumis, pour chacune des trois prises d'information, le tableau de distribution des fréquences des réponses à l'analyse des correspondances. Il nous faut signaler ici que nous avons considéré l'absence de réponse à un item comme une cinquième modalité, et que sur les 40 items, seuls 37 composaient le tableau analysé, les 3 items abandonnés présentant des distributions par trop particulières - plus de 95% des réponses concentrées sur la seule modalité "je le recherche" - et, de ce fait, ne véhiculant, pour ainsi dire, aucune information.

Au vu de ces trois analyses, nous pouvons faire les constatations suivantes. Entre 1974 et '75, le premier facteur, tant sur l'ensemble des modalités de réponse (figure 1) que sur celui des items, reste pratiquement inchangé, seul "non-réponse" présentant une instabilité importante, due essentiellement à son très faible effectif.

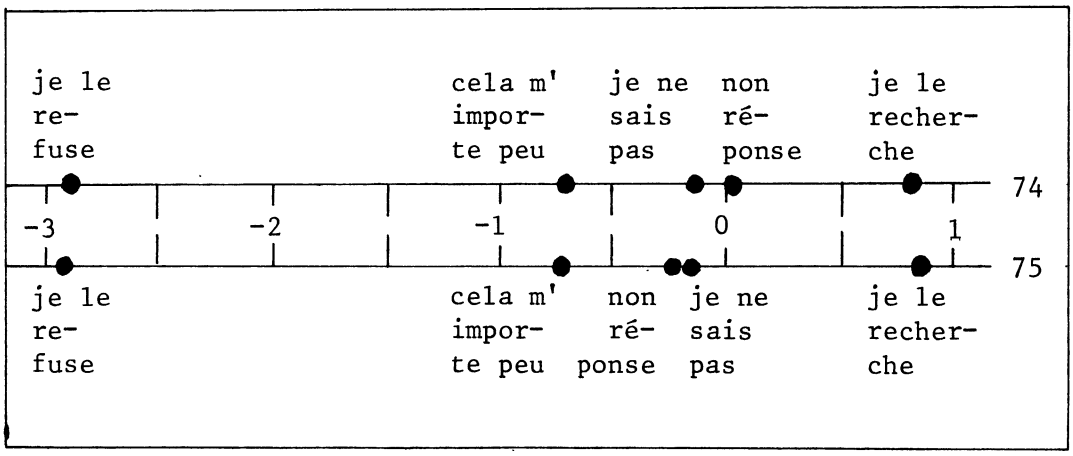


Figure 1 . Le premier facteur en 1974 et '75

Sur l'ensemble des items, on mesure, entre le premier facteur '74 et celui de '75, une congruence - ou corrélation dans notre cas - de .99 . En outre, pour ces deux années, la valeur propre correspondante est exactement la même, soit .25 . Ainsi on peut dire que la paramétrisation qui a le meilleur pouvoir séparateur, en moyenne, sur les items, est la même en 1974 et 1975, et que, de plus, les moyennes des items, pour cette paramétrisation, restent globalement inchangées.

Devant une telle stabilité on peut faire l'hypothèse qu'en 1976 les 37 items - considérés comme des observables numériques dont les p_{ij} sont des fréquences de classes - ont des distributions inchangées par rapport à celles des deux années précédentes. Si c'est le cas, on devrait pouvoir trouver une paramétrisation fournissant les mêmes moyennes qu'en 1974 et '75 et ce, malgré le changement apporté à l'ensemble S des modalités de réponse.

L'analyse des correspondances appliquée au tableau de distributions de l'année 1976 donne effectivement des résultats remarquables: la paramétrisation du nouvel ensemble de modalités, qui offre le meilleur pouvoir séparateur, en moyenne, sur les items, donne pratiquement les mêmes moyennes que celles obtenues en 1974 et '75. De plus le pouvoir séparateur - première valeur propre - reste inchangé. Présentent, en revanche, de grandes différences, les paramétrisations des deux ensembles de modalités de réponse (figure 2).

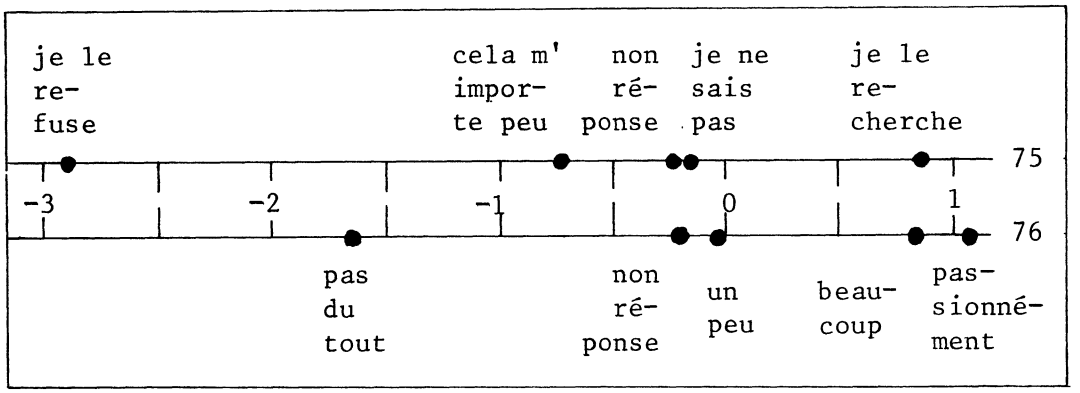


Figure 2 . Le premier facteur sur les deux ensembles de modalités

Entre la suite des moyennes obtenue en 1975 et celle fournie par la paramétrisation relative à l'année '76 on constate une congruence - ou corrélation de .97 . Il faut noter que ce coefficient peut s'interpréter comme le cosinus de l'angle entre les vecteurs - axes factoriels - sur lesquels on mesure la paramétrisation et que ces vecteurs sont situés dans un espace de dimension 37 , à chaque item correspondant un vecteur de base. Ces axes factoriels peuvent donc être interprétés comme la direction "refus-recherche" dans cet espace. A cause des fortes congruences entre les suites des moyennes des trois années, les trois axes factoriels, relatifs à 1974, '75 et '76, sont pratiquement superposés; ce qui rend comparables les termes des trois paramétrisations. Au vu de cet exemple, la suite des moyennes présentant le plus fort contraste, semble revêtir un caractère intrinsèque, peu

dépendant de l'ensemble S utilisé.

Les facteurs de rang supérieur

Nous reproduisons ci-dessous (figures 3,4 et 5) trois graphiques représentant les modalités de réponse dans les plans factoriels 1 - 2 , 1 - 3 et 1 - 4 , pour les trois années 1974, '75 et '76. Pour les années 1974 et '76 nous avons en outre séparé la population suivant les sexes. Ces graphiques mettent en évidence les résultats suivants:

1. Les facteurs de rang k sont des fonctions à k-1 extrema du premier facteur. Il faut cependant noter que cela n'apparaît pas clairement pour le facteur 4 ; cela tient certainement au rôle particulier de la "modalité" "non-réponse": de faible effectif, elle ne modifie pratiquement pas ce que seraient les trois premiers facteurs si l'on n'avait pas considéré l'absence de réponse comme une cinquième modalité; le facteur 4 , en revanche, n'est saturé essentiellement que par cette modalité, ce qui déforme considérablement, en l'aplatissant, l'oscillation que l'on pouvait attendre.

2. Les sous-populations "filles" et "garçons" fournissent des paramétrisations - ou, mieux, des suites de facteurs de correspondances - tout-à-fait semblables. Et ce, malgré le fait que ces deux strates de notre échantillon présentent, en moyenne, des différences sensibles quant aux réponses données à nos 37 items.

Ces constatations débouchent, semble-t-il, sur une possible généralisation de la théorie de Guttman à la situation qui nous intéresse ici.

Selon cette théorie, le minimum de la courbe d'intensité - fonction qui exprime le second facteur par rapport au premier (figure 3) - serait indépendant de l'échantillon étudié et des items utilisés et constituerait donc une origine naturelle - point d'intensité minimale, correspondant à un sentiment neutre - de notre paramétrisation. Peut-être serait-il donc judicieux de superposer, par translation du premier facteur, les minima des trois courbes d'intensité, et de mesurer la paramétrisation à partir de ce point, plutôt que depuis le zéro du premier facteur, qui est le centre de gravité de la paramétrisation. Notons encore qu'en superposant les minima des courbes d'intensité, on superpose également, à peu de chose près, les minima des trois courbes de clôture - fonction exprimant le troisième facteur par rapport au premier (figure 4).

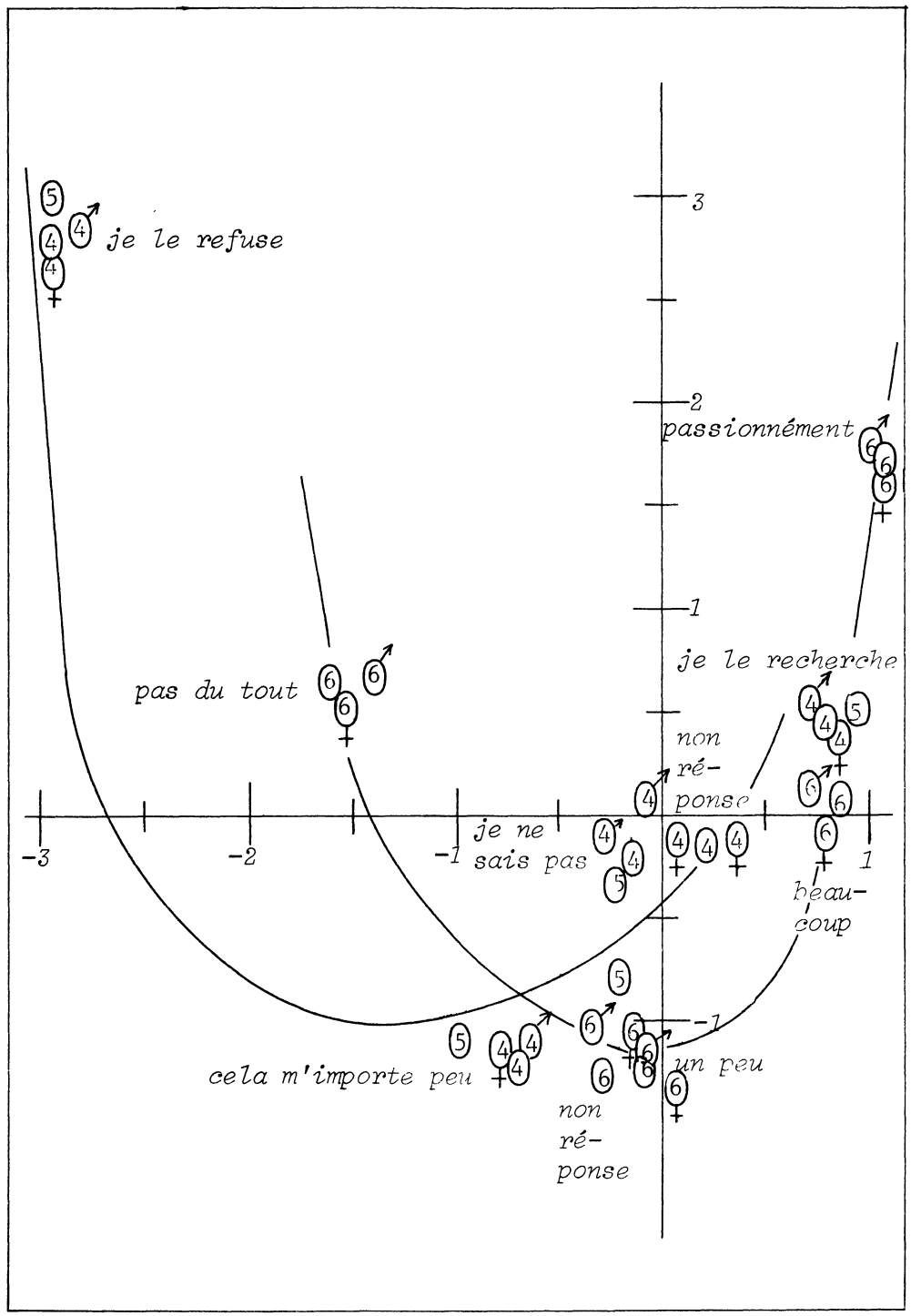


Figure 3 . Les modalités dans le plan des facteurs 1 et 2

Signification des symboles

○ = position d'une modalité pour l'analyse faite sur l'échantillon entier

○♀ = position d'une modalité pour l'analyse faite sur la strate "filles"

○♂ = position d'une modalité pour l'analyse faite sur la strate "garçons"

④ = position d'une modalité pour l'analyse relative à l'année 1974

⑤ = position d'une modalité pour l'analyse relative à l'année 1975

⑥ = position d'une modalité pour l'analyse relative à l'année 1976

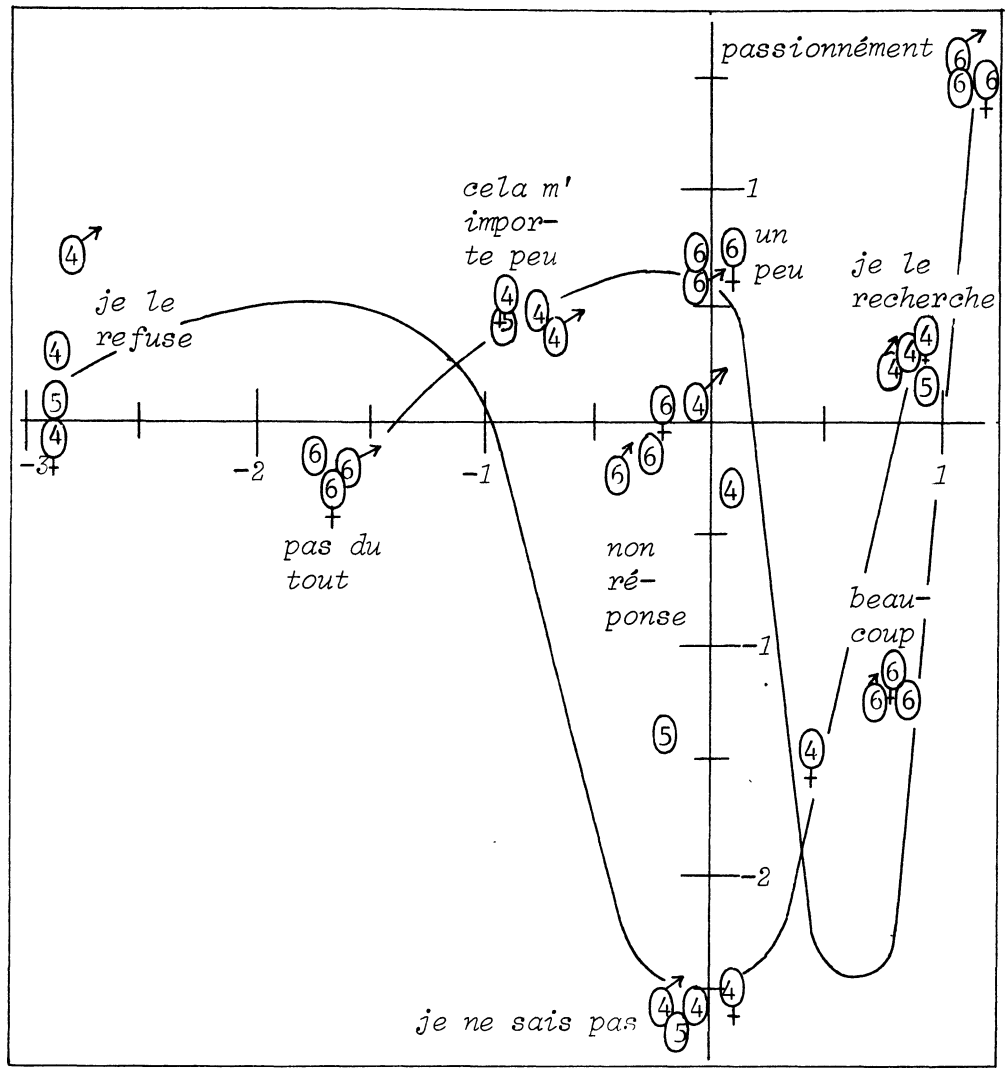


Figure 4 . Les modalités dans le plan des facteurs 1 et 3

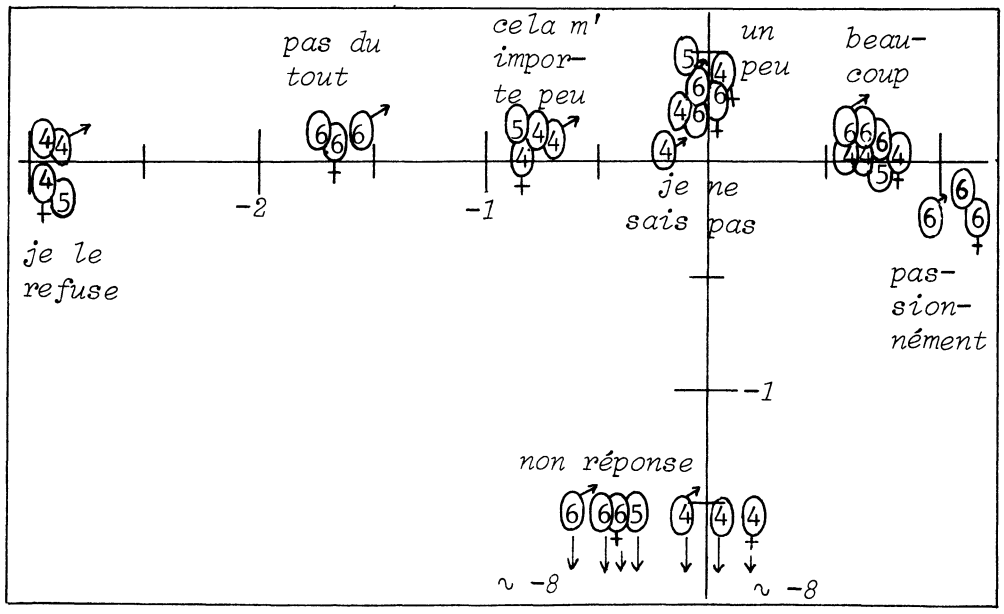


Figure 5 . Les modalités dans le plan des facteurs 1 et 4

5.2. Application au différenciel sémantique

Mis au point par Osgood (7), le différenciel sémantique se veut une méthode objective d'investigation de la pensée. Par le truchement d'une série de couples d'adjectifs antonymes, on essaye de localiser, dans ce qu'on appelle *l'espace sémantique* - qui peut être individuel ou commun au groupe des sujets observés -, un ensemble de concepts, chacun couvrant un certain champ du domaine étudié.

Elaborée dans le cadre d'une recherche sur les psychothérapies brèves[1], une forme du différenciel sémantique a été soumise à un échantillon composé de malades mentaux d'une part, et de sujets "normaux" d'autre part: onze concepts - que nous appellerons les *mots inducteurs* -, chacun couvrant un certain champ psychanalytique (exemples: MORT , MERE ...), sont à confronter tour à tour à une série de trente échelles bipolaires associées à trente couples d'adjectifs antonymes, comme, par exemple:

BON	1	2	3	4	5	6	7	MAUVAIS

Nous considérons que chaque couple sujet-mot inducteur constitue un élément de la population notée $P \times M$ - P étant l'ensemble $\{ a_1 , a_2 , \ldots , a_N \}$ des sujets et M , celui des mots inducteurs $\{ w_1 , w_2 , \ldots , w_{11} \}$. Sur cette population, les trente échelles bipolaires définissent trente observables ordinales, chaque sujet devant situer chaque mot inducteur sur chacune des trente échelles. Ce sont ces trente observables qui engendrent l'espace sémantique.

Voulant les soumettre à l'analyse factorielle afin d'exhiber les quelques dimensions de base de cet espace, nous devons préalablement paramétriser ces trente observables.

Considérons sur $P \times M$ la stratification $\left\{ P \times \{ w_1 \} , P \times \{ w_2 \} , \ldots , P \times \{ w_{11} \} \right\}$; si l'on veut étudier les *configurations spatiales* - situations des mots inducteurs dans l'espace sémantique -, il est cohérent de chercher la paramétrisation qui rende, en moyenne, ces configurations les plus claires possible, en séparant au mieux les mots inducteurs. Pour chaque échelle séparément, il s'agit donc de rendre maximale la variance des

---

1. Voir la note de la première page.

moyennes des classes $P \times \{ w_j \}$, sous la condition " variance totale (calculée sur $P \times M$) = 1 "; ou, si l'on préfère, de maximiser la variance "interclasse" en gardant constante la variance totale. On obtient ce résultat en analysant, pour une échelle bipolaire donnée, disons la $i^{ème}$, le tableau T_i

$$T_i = \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,7} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,7} \\ \vdots & \vdots & \vdots & \vdots \\ p_{11,1} & p_{11,2} & \cdots & p_{11,7} \end{pmatrix}$$ où $p_{i,j}$ est la fréquence d'apparition du niveau k pour le mot w_j .

Ayant procédé aux trente analyses des correspondances, on distingue, selon les échelles, trois types de résultats:

1. Les échelles pour lesquelles le premier facteur, sur l'ensemble ordonné des niveaux $S = \{ 1 , 2 , 3 , 4 , 5 , 6 , 7 \}$, est une fonction strictement croissante de S .

2. Les échelles pour lesquelles on constate des permutations légères aux extrémités (entre les niveaux 1 et 2 ou 6 et 7). Dans cette situation les différences entre $\psi (1)$ et $\psi (2)$ ou entre $\psi (6)$ et $\psi (7)$ ne sont pas significatives et l'on doit considérer que ψ identifie les niveaux extrèmes 1 et 2 ou 6 et 7 .

3. Enfin quatre échelles présentent de fortes permutations qui rendent impossible l'utilisation du premier facteur comme paramétrisation. Il s'agit des échelles LARGE - ETROIT , LENT - RAPIDE , MOU - DUR , CHAOTIQUE - RYTHME . On en déduit que, pour ces échelles-là, les différentes classes $P \times \{ w_j \}$ présentent, au niveau des observables numériques sous-jacentes, des différences importantes quant à la forme des distributions, et que, par conséquent, chercher à mettre en évidence un ordre sur les moyennes de ces classes n'est pas raisonnable. On constate, en outre, que, pour chacune de ces quatre échelles, l'utilisation de la case centrale 4 - globalement forte - varie beaucoup, quantitativement, d'un mot inducteur à l'autre. Si bien que, dans $\mathbb{R}^{11}$, le point représentatif de 4 se trouve, avec un poids relativement grand, éloigné de l'origine, et attire vers lui, en quelque sorte, le premier axe factoriel. Ces quatre échelles, investies très différemment suivant les mots inducteurs auxquels elles sont confrontées, présen-

tent donc un caractère spécifique, qui pourrait remettre en question la présence de ces quatre couples d'adjectifs dans cette forme du différenciel sémantique.

Notons que nous aurions pu, à l'instar de Bradley, Katti et Coons (2), ou de Nishisato et Arri (6), maximiser la variance "interclasse" - ou entre les mots inducteurs - en imposant une contrainte d'ordre. Nous ne l'avons pas fait, préférant considérer que notre idée de départ était trop ambitieuse, puisqu'elle supposait, pour chaque couple d'adjectifs, un ordre naturel sur les mots inducteurs. Aussi avons-nous opté pour une méthode plus globale.

La présentation du différenciel sémantique supposant qu'on situe chaque mot inducteur par rapport à des échelles *bipolaires* - donc simultanément, et d'une manière symétrique par rapport au deux pôles (hypothèse certainement utopique) -, nous voulons obtenir une paramétrisation symétrique, donc satisfaisant aux conditions:

$\psi (1) = - \psi (7)$

$\psi (2) = - \psi (6)$

$\psi (3) = - \psi (5)$

$\psi (4) = 0$

Ce qui revient à considérer qu'une réponse du type:

BON | 1 | 2 | ~~3~~ | 4 | 5 | 6 | 7 | MAUVAIS

correspond à la double prise de position:

BON | 1 | 2 | ~~3~~ | 4 | 5 | 6 | 7 |

MAUVAIS | 1 | 2 | 3 | 4 | ~~5~~ | 6 | 7 |

Cette attitude, délibérément réductive, admet plusieurs justifications:

1. Si l'hypothèse ci-dessus, que nous avons qualifiée d'utopique, devait être vraie, le dédoublement de chaque prise de position serait alors implicitement compris dans nos données et l'ajouter explicitement ne changerait rien.

2. Si cette hypothèse est fausse, en dédoublant chacune des échelles, on contrebalance, en quelque sorte, les effets, sur les distributions des réponses, dus au seul fait que les sujets observés ne se situent pas simultanément et de manière symétrique par rapport aux deux pôles de l'échelle, mais bien plutôt par rapport à un seul pôle privilégié.

3. Enfin il paraît cohérent de rechercher une paramétrisation indépendante de l'orientation des échelles.

Pratiquement on obtiendra une paramétrisation symétrique unique, commune à l'ensemble des échelles et des mots inducteurs, en analysant le tableau dédoublé T :

$$T = \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,7} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,7} \\ \vdots & \vdots & \vdots\vdots & \vdots \\ p_{30,1} & p_{30,2} & \cdots & p_{30,7} \\ p_{31,1} & p_{31,2} & \cdots & p_{31,7} \\ \vdots & \vdots & \vdots\vdots & \vdots \\ p_{60,1} & p_{60,2} & \cdots & p_{60,7} \end{pmatrix}$$

où

$$p_{j,k} = \begin{cases} \text{si } j \leq 30 : \text{fréquence d'apparition du niveau } k \text{ (calculée sur } P \times M) \text{ pour la } j^{\text{ème}} \text{ échelle} \\ \text{si } j > 30 : \text{fréquence du niveau } 8-k \text{ pour la } (j-30)^{\text{ème}} \text{ échelle (calculée sur } P \times M) \end{cases}$$

Il faut remarquer que, d'un point de vue théorique, cette procédure n'est pas équivalente à celle qui consisterait à maximiser le pouvoir séparateur en moyenne sur les échelles, avec la condition supplémentaire de symétrie. Une telle procédure serait réalisable par une méthode du genre de celles proposées par Bradley, Katti et Coons (2), ou par Nishisato et Arri (6). Sur le plan des résultats, en revanche, on ne devrait s'en écarter que de très peu.

Nous obtenons ainsi la paramétrisation suivante:

-13	-13	-7	0	7	13	13

Ce qu'on peut interpréter de la manière suivante. Les niveaux extrêmes, situés aux deux pôles de l'échelle (0 , 1 et 6 , 7) apportent globalement la même contribution à la "séparation" des échelles. Cela provient du fait que l'utilisation de l'échelle en 7 points se fait de façon très différenciée entre les sujets; ainsi, certains n'utilisent pratiquement que les cases 1 , 4 et 7 , se refusant à toute nuance, alors que d'autres, au contraire,

hésitent à prendre position de façon très nette, et utilisent essentiellement les cases intérieures de l'échelle. Dans un certain sens, donc, les 6 de certains sujets ont autant d'"intensité", si ce n'est plus, que les 7 d'autres.

Nous considérerons donc que 1 et 2 d'une part, 6 et 7 d'autre part, sont "investis", en moyenne, de façon semblable, et qu'il n'y a pas lieu de les distinguer au plan du traitement; ce qui ne remet nullement en question l'échelle en 7 points, pour ce qui est de la présentation de l'instrument et de son administration.

BIBLIOGRAPHIE

(1) BENZECRI J.P. et coll., *L'analyse des données* (2 tomes), Paris, Dunod, 1973.

(2) BRADLEY R.A., KATTI S.K. et COONS I.J., "Optimal scaling for ordered categories", *Psychometrika*, 27 (1962), 335-374.

(3) GULLIKSEN H., "A least squares solution for successive intervals assuming inequal standard deviations", *Psychometrika*, 19 (1954), 117-139.

(4) GUTTMAN L. "The principal components of scale analysis", in STOUFFER S.A. et al., *Measurement and prediction*, Princeton, princeton University Press, 1950.

(5) LEBART L. et FENELON J.P., *Statistique et informatique appliquée*, Paris, Dunod, 1975.

(6) NISHISATO S. et ARRI P.S., "Nonlinear programming approach to optimal scaling of partially ordered categories", *Psychometrika*, 40 (1975), 525-548.

(7) OSGOOD C.E., SUCI G.J. et TANNENBAUM P.H., *The measurement of meaning*, Chicago, University of Illinois Press, 1957.

(8) TORGERSON W.S., *Theory and methods of scaling*, New-York, Wiley, 1958.