REVUE DE STATISTIQUE APPLIQUÉE

JEAN-JACQUES DAUDIN

Analyse factorielle des dépendances partielles

Revue de statistique appliquée, tome 29, nº 2 (1981), p. 15-29
<http://www.numdam.org/item?id=RSA_1981__29_2_15_0>

ANALYSE FACTORIELLE DES DEPENDANCES PARTIELLES

Jean-Jacques DAUDIN

Département de Mathématique, Institut National Agronomique
Paris-Grignon, 16, rue Cl. Bernard, Paris

RESUME

Nous proposons des méthodes permettant d'analyser les dépendances partielles dans une table de contingence tridimensionnelle.

INTRODUCTION

Quand on dispose d'une table de contingence tridimensionnelle et que l'on s'intéresse en particulier à la dépendance entre 2 des 3 variables, la pratique courante veut que soit établie la table marginale obtenue en sommant sur la troisième variable, et que cette table soit étudiée par une analyse factorielle des correspondances. Ce faisant on perd l'information concernant les dépendances partielles.

Une autre possibilité est l'étude de toutes les tables de contingence correspondant aux différentes valeurs de la troisième variable. Cela peut être fastidieux et difficile à synthétiser si cette dernière prend de nombreuses valeurs. De plus se pose le problème des cases vides de la table tridimensionnelle s'il y a peu d'observations.

Nous proposons ici une méthode, basée sur une idée de J.N. DARROCH [1], qui permet d'étudier les dépendances partielles par 2 analyses des corrélations canoniques ou 2 analyses factorielles des correspondances sur 2 tables bidimensionnelles. J.N. DARROCH distingue 2 sortes de dépendances partielles entre 2 variables aléatoires X et Y : la dépendance "qui passe par Z" (qu'il appelle dépendance attachée) et la dépendance indépendante de Z, c'est-à-dire l'association conditionnelle (qu'il appelle dépendance détachée).

La première partie de cet article contient les définitions des dépendances attachées et détachées proposées par J.N. DARROCH dans [1] pour des variables discrètes ainsi qu'une généralisation de ces définitions au cas de variables quelconques.

Dans les deuxième et troisième parties nous proposons respectivement des méthodes d'analyses des dépendances attachées et détachées analogues de l'analyse factorielle des correspondances de la liaison marginale.

Dans la quatrième partie nous illustrons ces méthodes par des exemples.

Mots-clés : Analyse factorielle des correspondances, Dépendances attachées et détachées, Table de contingence tridimensionnelle.

I. DEPENDANCE PARTIELLE ATTACHEE ET DEPENDANCE PARTIELLE DETACHEE

A. Présentation des définitions de J.N. DARROCH

Soient X, Y, Z trois variables aléatoires discrètes et appelons

$$P_{ijk} = P(X = i\,,\ Y = j\,,\ Z = k)\ ,\ (i = 1..I\,, j = 1..J\,, k = 1..K)$$

$$P_{ij.} = \sum_{k} P_{ijk}\,, P_{.jk} = \sum_{i} P_{ijk}\,; P_{i.k}\,; P_{i.k} = \sum_{j} P_{ijk}$$

$$P_{..k} = \sum_{ij} P_{ijk}$$

X et Y sont indépendants conditionnellement à Z si (1) est vérifié :

$$\forall ijk\,,\ P_{ijk} = P_{i.k}\,P_{.jk}/P_{..k} \tag{1}$$

Définitions

DARROCH mesure la dépendance des évènements $(X = i)$ et $(Y = j)$ conditionnellement à $Z = k$ par

$$P_{ijk}/P_{..k} - (P_{i.k}/P_{..k})\,(P_{.jk}/P_{..k})$$

et la dépendance conditionnelle moyenne par (2)

$$\sum_{k} (P_{ijk}/P_{..k} - (P_{i.k}/P_{..k})\,(P_{.jk}/P_{..k}))\,P_{..k} = P_{ij.} - \Pi_{ij} \tag{2}$$

où
$$\Pi_{ij} = \sum_{k} P_{i.k}P_{.jk}/P_{..k}$$

DARROCH mesure la dépendance marginale entre les évènements $(X = i)$ et $(Y = j)$ par la quantité $P_{ij.} - P_{i..}\,P_{.j.}$

Enfin il mesure la dépendance attachée à Z par la différence entre ces 2 dernières quantités, c'est-à-dire par (3)

$$\Pi_{ij} - P_{i..}\,P_{.j.} \tag{3}$$

Cette dernière quantité mérite quelques éclaircissements : Π_{ij} peut s'interpréter comme étant la probabilité conjointe des évènement $(X = i)$ et $(Y = j)$ si ces 2 évènements sont indépendants conditionnellement à Z.

S'il n'y a pas de dépendance attachée à Z entre ces 2 évènements on doit avoir $\Pi_{ij} = P_{i..}\,P_{.j.}$. La dépendance attachée à Z rend compte d'un phénomène bien connu selon lequel il se peut que chaque table à $Z = k$ fixé indique l'indépendance entre X et Y alors que la table marginale montre une non indépendance.

La relation (4) résume les notions introduites :

$$P_{ij.} - P_{i..}P_{.j.} = (P_{ij.} - \Pi_{ij}) + (\Pi_{ij} - P_{i..}P_{.j.}) \qquad (4)$$

(4) s'exprime verbalement de la façon suivante :

Dépendance marginale = dépendance détachée de Z + dépendance attachée à Z.

Remarques

1) la dépendance détachée de Z mesure seulement une dépendance conditionnelle moyenne. Cette mesure n'est intéressante que s'il y a une relative stabilité, quand Z varie, de la dépendance conditionnelle à Z fixé. Par exemple il se peut que $\Pi_{ij} = P_{ij.}$ sans qu'il y ait obligatoirement situation d'indépendance conditionnelle.

2) Deux cas particuliers sont à remarquer :

– Si $\Pi_{ij} = P_{ij.}$ la dépendance détachée de Z est nulle, et toute la dépendance marginale est due à la composante de la dépendance attachée à Z, c'est-à-dire que la dépendance "passe" par Z.

– Si $\Pi_{ij} = P_{ij.}$ la dépendance détachée de Z est nulle, et toute la dépendance marginale est égale à la dépendance détachée de Z. Dans ce cas l'analye de la dépendance marginale est donc aussi l'analyse de la dépendance détachée, c'est-à-dire de la dépendance conditionnelle moyenne.

3) A l'aide des probabilités marginales $P_{ij.}$, $P_{i.k}$, $P_{.jk}$ on peut constituer plusieurs tables $I \times J$:

a) la table dont le (i, j) ième élément est P_{ij}.

b) la table Π_{ij}.

c) la table $P_{i..}P_{.j.} + (P_{ij.} - \Pi_{ij})$

Toutes ces tables ont même marginales $P_{i..}$ et $P_{.j.}$.

On peut construire de façon identique les tables de contingence (a'), (b'), (c') correspondant aux tables des probabilités (a), (b), (c) respectivement. Pour étudier la dépendance marginale (X, Y) des variables X et Y on est amené à faire l'analyse factorielle des correspondances de la table (a'). De même pour étudier les dépendances attachées à Z et détachées de Z on est amené à faire les analyses factorielles des correspondances des tables (b') et (c').

Nous étudions les propriétés de ces 2 dernières analyses dans les 3e et 4e parties de cet article. Il est auparavant utile d'aller plus avant dans l'étude des notions de dépendance attachée et détachée dans le cas de variables aléatoires quelconques.

B. Généralisation

1) Généralisation des définitions de Darroch

Soit X, Y, Z trois variables aléatoires.

Soit dP_{XYZ} la loi de probabilité conjointe de (X, Y, Z) dP_{XY}, dP_{XZ}, dP_{YZ}, dP_X, dP_Y, dP_Z les lois marginales.

Soit L^2_X, L^2_Y, L^2_Z, les ensembles de fonction de carré sommable de respectivement X, Y et Z :

$$L^2_X = \{\psi(X) / E(\psi^2(X)) < +\infty\}$$

$$L^2_Y = \{\theta(Y) / E(\theta^2(Y)) < +\infty\}$$

$$L^2_Z = \{\zeta(Z) / E(\zeta^2(Z)) < +\infty\}$$

où $E(\psi(X))$ est l'espérance de $\psi(X)$, et où ψ, θ et ζ sont respectivement des fonctions X, Y, Z mesurables.

Soit E^Z l'opérateur espérance conditionnelle.

Pour toutes variables aléatoires $\psi(X)$ De L^2_X et $\theta(Y)$ de L^2_Y, on peut montrer que l'on a

$$\begin{aligned} E[(\psi(X) - E(\psi(X)))\,(\theta(Y) - E(\theta(Y)))] & \\ &= E[(E^Z(\psi(X)) - E(\psi(X)))\,(E^Z(\theta(Y)) - E(\theta(Y)))] \\ &+ E[(\psi(X) - E^Z(\psi(X)))\,(\theta(Y) - E^Z(\theta(Y)))] \qquad (5) \end{aligned}$$

Cette relation est évidente si on décompose chaque terme, car elle s'écrit alors :

$$\begin{aligned} E(\psi(X)\theta(Y)) - E(\psi(X))\,E(\theta(Y)) & \\ &= E[E^Z(\psi(X))\,E^Z(\theta(Y))] - E(\psi(X))\,E(\theta(Y)) \\ &+ E(\psi(X)\,\theta(Y)) - E[E^Z(\psi(X))\,E^Z(\theta(Y))] \qquad (5\text{bis}) \end{aligned}$$

La relation (5) est une généralisation de (4). En effet, si X, Y, Z sont des variables aléatoires discrètes avec les notations du début de la première partie, et si on note :

$$\psi(X) = 1_i(X) = \begin{vmatrix} 1 & \text{si} & X = i \\ 0 & \text{si} & X \neq i \end{vmatrix}$$

$$\theta(Y) = 1_j(Y) = \begin{vmatrix} 1 & \text{si} & Y = j \\ 0 & \text{si} & Y \neq j \end{vmatrix}$$

$$1_k(Z) = \begin{vmatrix} 1 & \text{si} & Z = k \\ 0 & \text{si} & Z \neq k \end{vmatrix}$$

ce qui donne :

$$E(1_i(X) \,.\, 1_j(Y)) = P_{ij}$$

$$E(1_i(X)) = P_{i..}$$

$$E(1_j(Y)) = P_{.j.}$$

$$E^Z(1_i(X)) = \sum_k (P_{i.k}/P_{..k})\,1_k(Z)$$

$$E^Z(1_j(Y)) = \sum_k (P_{.jk}/P_{..k})\,1_k(Z)$$

De ces relations et de (5bis) l'on déduit immédiatement (4).

Si on note que $E((\psi(X) - E(\psi(X)))\ (\theta(Y) - E(\theta(Y)))$ est la covariance entre $\psi(X)$ et $\theta(Y)$, que $E^Z(\psi(X) - E(\psi(X)))$ est la projection de $\psi(X) - E(\psi(X))$ sur L^2_Z, et que $\theta(X) - E^Z(\theta(X))$ est la projection de $\theta(X) - E(\theta(X))$ sur $L^{2\perp}_Z$, où $L^{2\perp}_Z$ est l'orthogonal de L^2_Z, on peut interpréter la relation (5) de la façon suivante : la covariance entre $\psi(X)$ et $\theta(Y)$ est égale à la somme des covariances entre leurs projections sur L^2_Z d'une part et sur $L^{2\perp}_Z$ d'autre part.

Nous avons montré dans [2] et [3] que toutes les corrélations entre projections sur $L^{2\perp}_Z$ de fonctions de X d'une part et de fonction de Y d'autre part mesurent la liaison moyenne conditionnellement à Z. De plus les corrélations entre projections de fonctions de X et de fonctions de Y sur L^2_Z mesurent la dépendance entre X et Y qui est liée à Z. La relation (5) est donc une généralisation de la décomposition (4) due à DARROCH (1).

Lorsque (X, Y, Z) est un vecteur gaussien, l'espérance conditionnelle à Z étant une fonction linéaire de Z, on retrouve à partir de (5) la formule classique :

$$\rho_{XY} = \rho_{XY.Z}[(1 - \rho^2_{XZ})\ (1 - \rho^2_{YZ})]^{1/2} + \rho_{XZ}\rho_{YZ}$$

en prenant $\psi(X) = X/(\operatorname{Var} X)^{1/2}$ et $\theta(Y) = Y/(\operatorname{Var} Y)^{1/2}$.

2) Décomposition de la mesure de dépendance globale

La relation (5) concerne seulement deux variables aléatoires $\psi(X)$ et $\theta(Y)$. On peut obtenir une décomposition analogue de la dépendance globale entre X et Y.

Soit respectivement $(\psi_i)\ i \in I$ et $(\theta_j)\ j \in J$ des systèmes orthonormés totaux de L^2_X et L^2_Y, avec $\psi_0 = \theta_0 = 1$.

La liaison entre X et Y est mesurée par Φ^2_{XY} où $\Phi^2_{XY} = \sum_{ij=1}^{\infty} (E(\psi_i\theta_j))^2$.

D'après (5), on a : $\Phi^2_{XY} = A + B + C$

avec
$$A = \sum_{ij=1}^{\infty} [E(E^Z(\psi_i)\ E^Z(\theta_j)]^2$$

$$B = \sum_{ij=1}^{\infty} [E((\psi_i - E(\psi_i))\ (\theta_j - E^Z(\theta_j))]^2$$

$$C = 2\sum_{ij=1}^{\infty} E(E^Z(\psi_i)\ E^Z(\theta_j))\ E[(\psi_i - E^Z(\psi_i))\ (\theta_j - E^Z(\theta_j))]$$

A mesure la dépendance attachée à Z. $(A \geqslant 0)$

B mesure la dépendance détachée de Z. $(B \geqslant 0)$

C est un double produit qui permet de savoir si les deux formes de dépendance se renforcent l'une l'autre (cas où C est positif) ou si au contraire elles s'opposent (cas où C est négatif).

(1) Dans l'article cité, DARROCH écrit aussi (4) sous la forme d'une décomposition de la covariance dans des conditions moins générales que (5).

Ce dernier cas permet d'expliquer le paradoxe suivant :

Chaque table bidimensionnelle à Z fixé peut présenter une dépendance entre X et Y alors que la table marginale indique l'indépendance.

Notons que la liaison attachée mesurée par A possède la propriété suivante :

$$A \leqslant \Phi^2_{XZ}\,\Phi^2_{YZ} \tag{6}$$

qui précise en termes mathématiques un fait intuitif : l'intensité de la dépendance entre X et Y attachée à Z est bornée par le produit des intensités des dépendances entre X et Z d'une part et Y et Z d'autre part.

La démonstration de (6) est basée sur l'inégalité de Cauchy-Schwartz :

$$[E(E^Z(\psi_i)\,E^Z(\theta_j))]^2 \leqslant E[(E^Z(\psi_i))^2]\,E[(E^Z(\theta_j))^2]$$

et sur le fait que

$$\Phi^2_{XZ} = \sum_{i=1}^{\infty} E[(E^Z(\psi_i))^2] \quad \text{et} \quad \Phi^2_{YZ} = \sum_{j=1}^{\infty} [E(E^Z(\theta_j))^2]$$

(6) permet de retrouver le résultat (7) démontré dans [2] : Si la dépendance conditionnelle moyenne entre X et Y est nulle alors

$$\Phi^2_{XY} \leqslant \Phi^2_{ZX}\,.\,\Phi^2_{ZY} \tag{7}$$

3) Analyses canoniques des dépendances attachées et détachées

Soient H_1 et H_2 respectivement les sous espaces de L^2_Z engendrés par les variables aléatoires $(E^Z(\psi_i)\,,\ i \in I)$ et $(E^Z(\theta_j)\,,\ j \in J)$.

Définition 1

L'analyse canonique de la dépendance attachée à Z est l'analyse canonique des espaces H_1 et H_2.

Soient G_1 et G_2 respectivement les sous espaces de $L^{2\perp}_Z$ engendré par les variables aléatoires $(\psi_i - E^Z(\psi_i)\,,\ i \in I)$ et $(\theta_j - E^Z(\theta_j)\,,\ j \in J)$.

Définition 2

L'analyse canonique de la dépendance détachée de Z est l'analyse canonique des espaces G_1 et G_2.

Ces définitions seront utilisées pour des variables discrètes dans la suite de cet article.

II. ANALYSE FACTORIELLE DE LA LIAISON ATTACHEE

1. Calcul de la matrice de covariance

Considérons les indicatrices des modalités de X et de Y :

$$U_i = 1_i(X) \quad \text{et} \quad V_j = 1_j(Y)$$

$(U_0, U_1 \ldots U_{I-1})$ et $(V_0, V_1 \ldots V_{J-1})$ sont deux bases orthogonales de L^2_X et L^2_Y U_0 et V_0 étant les fonctions constantes.

Les projections de U_i et V_j sur L_Z^2 sont U_i^Z et V_i^Z :

$$U_i^Z = \sum_k (P_{i.k}/P_{..k})\ 1_k(Z)$$

$$V_j^Z = \sum_k (P_{.jk}/P_{..k})\ 1_k(Z)$$

et on a

$$E(U_i^Z V_j^Z) = \sum_k P_{i.k} P_{.jk}/P_{..k}$$

$$E(U_i^Z) = P_{i..} \quad \text{et} \quad E(V_j^Z) = P_{.j.}$$

$$\mathrm{Cov}\,(U_i^Z, V_j^Z) = \sum_k (P_{i.k} P_{.jk}/P_{..k}) - P_{i..} P_{.j.}$$

$$\mathrm{Cov}\,(U_i^Z, U_{i'}^Z) = \sum_k (P_{i.k} P_{i'.k}/P_{..k}) - P_{i..} P_{i'..}$$

$$\mathrm{Cov}\,(V_j^Z, V_{j'}^Z) = \sum_k (P_{.jk} P_{.j'k}/P_{..k}) - P_{.j.} P_{.j'.}\ .$$

Soit Σ_{11} la matrice des covariances des U_i^Z, Σ_{22} celle des V_j^Z et Σ_{12} la matrice des covariances entre U_i^Z et V_j^Z. L'analyse canonique de la dépendance attachée s'obtient en diagonalisant la matrice $\Sigma_{11}^{-1}\ \Sigma_{12}\ \Sigma_{22}^{-1}\ \Sigma_{21}$. Il se peut que Σ_{11} ou Σ_{22} ne soient pas inversibles ; c'est le cas pour $K < I$ ou $K < J$. Il faut considérer pour $K < I$ une famille libre des U_i^Z et non tous les U_i^Z. Il en est de même pour les V_j^Z lorsque $K < J$.

2. Analyse factorielle de la liaison attachée

Une solution plus simple (et qui ne pose pas de problème d'inversion) consiste à faire l'analyse factorielle des correspondances de la table de contingence suivante :

$$M_{ij} = \sum_k N_{i.k} N_{.jk}/N_{..k} \tag{8}$$

où N_{ijk} est l'effectif de la table tridimensionnelle pour $X = i$, $Y = j$ et $Z = k$, et où $N_{i.k} = \sum_j N_{ijk}$, $N_{.jk} = \sum_i N_{ijk}$, $N_{..k} = \sum_{ij} N_{ijk}$, $N_{i..} = \sum_{jk} N_{ijk}$ et $N_{.j.} = \sum_{ik} N_{ijk}$.

Ceci revient à diagonaliser la matrice :

$$D_I^{-1}\ V_{12}\ D_J^{-1}\ V_{21}$$

où D_I est la matrice diagonale avec $N_{i..}$ pour ième élément de la diagonale, D_J est la matrice diagonale avec $N_{.j.}$ pour jème élément de la diagonale, V_{12} est la matrice à I lignes et J colonnes de terme général M_{ij}.

Dans cette analyse factorielle, la distance entre deux modalités i et i' de X est égale à :

$$d_{ii'}^2 = \sum_j (M_{ij}/N_{i..} - M_{i'j}/N_{i'..})^2\ N/N_{.j.}$$

L'inertie totale de l'analyse est égale à :

$$I_1 = \sum_{ij} (M_{ij} - N_{i..} N_{.j.}/N)^2 / (N_{i..} N_{.j.}).$$

III. ANALYSE FACTORIELLE DE LA DEPENDANCE DETACHEE

1. Analyse canonique de la dépendance détachée :

On considère les projections U_i^* et V_j^* de U_i et V_j sur $L_Z^{2\perp}$

$$U_i^* = U_i - \sum_k P_{i.k}/P_{..k}\, 1_k(Z)$$

$$V_j^* = V_j - \sum_k P_{.jk}/P_{..k}\, 1_k(Z)$$

on a

$$E(U_i^*) = 0$$

$$E(V_j^*) = 0$$

$$E[(U_i^*)^2] = P_{i..} - \sum_k P_{i.k}^2/P_{..k}$$

$$E(U_i^* U_{i'}^*) = -\sum_k P_{i.k} P_{i'.k}/P_{..k} \qquad (i \neq i')$$

$$E[(V_j^*)^2] = P_{.j.} - \sum_k P_{.jk}^2/P_{..k}$$

$$E(V_j^* . V_{j'}^*) = -\sum_k P_{.jk} P_{.j'k}/P_{..k} \qquad (j \neq j')$$

$$E(U_i^* V_j^*) = P_{ij.} - \sum_k P_{i.k} P_{.jk}/P_{..k} = P_{ij.} - \Pi_{ij}$$

Soit Σ_{11}^* la matrice de covariance des U_i^*

Σ_{22}^* V_j^*

Σ_{12}^* (U_i^*, V_j^*)

L'analyse canonique conduit à diagonaliser la matrice

$$\Sigma_{11}^{*-1}\, \Sigma_{12}^*\, \Sigma_{22}^{*-1}\, \Sigma_{21}^*$$

2. Analyse factorielle de la dépendance détachée

Une solution plus simple consiste à faire l'analyse factorielle des correspondances de la table de contingence suivante :

$$N_{ij}^* = N_{i..} N_{.j.}/N + (N_{ij.} - M_{ij}) \qquad (9)$$

qui a les mêmes marges $((N_{i..})$ et $(N_{.j.}))$ que les tables N_{ij} et M_{ij}. (Cette table peut éventuellement prendre quelques valeurs négatives).

Cette analyse revient à diagonaliser la matrice.

$$D_I^{-1} \Sigma_{12}^* D_J^{-1} \Sigma_{21}^*$$

Comme dans le paragraphe précédent la différence entre les 2 analyses est due à des choix différents de la métrique de référence dans L_X^2 et L_Y^2 : D_I et D_J dans l'analyse factorielle, Σ_{11}^* et Σ_{22}^* dans l'analyse canonique.

La distance entre 2 modalités i et i' de X est égale à :

$$d_{ii'}^2 = \sum_j \left(\frac{N_{ij.} - M_{ij}}{N_{i..}} - \frac{N_{i'j.} - M_{i'j}}{N_{i'..}} \right)^2 N/N_{.j.}$$

l'inertie totale de l'analyse est :

$$I_2 = \sum_{ij} (N_{ij.} - M_{ij})^2 / N_{i..} N_{.j.}$$

3. Décomposition de la mesure de dépendance globale

Les résultats du paragraphe IB2 s'appliquent ici sous la forme suivante, ou I est l'inertie de l'analyse factorielle des correspondances de la table marginale :

$$I = \sum_{ij} (N_{ij.} - N_{i..} N_{.j.})^2 / (N_{i..} N_{.j.})$$

$$A = I_1 \; ; \; B = I_2$$

$$C = 2 \sum_{ij} (M_{ij} - N_{i..} N_{.j.}/N)\,(N_{ij.} - M_{ij})/(N_{i..} N_{.j.})$$

$$= 2 \sum_{i,j} M_{ij}(N_{ij.} - M_{ij})/(N_{i..} N_{.j.})$$

En résumé, on obtient l'égalité suivante :

$$I = I_1 + I_2 + C$$

Comme C peut être de signe quelconque, il se peut que I_1 et ou I_2 soient plus grands que I.

IV. EXEMPLES

1er exemple

On considère 3 variables X, Y, Z a respectivement 5,5 et 2 modalités. On considère les tables de contingence suivantes :

Z	1	2	3	4	5
1	0	10	50	90	100
2	100	90	50	10	0

X

Z	1	2	3	4	5
1	100	90	50	10	0
2	0	10	50	90	100

Y

		1	2	3	4	5
	1	24	19	19	19	19
	2	19	24	19	19	19
Y	3	19	19	24	19	19
	4	19	19	19	24	19
	5	19	19	19	19	24

Tables marginales $Z \times X$, $Z \times Y$ et $Y \times X$

A l'aide des tables $Z \times X$ et $Z \times Y$ on construit les tables suivantes :

0	4	20	36	40
4	7.2	20	32.8	36
20	20	20	20	20
36	32.8	20	7.2	4
40	36	20	4	0

44	35	19	3	– 1
35	36.8	19	6.2	3
19	19	24	19	19
3	6.2	19	36.8	35
– 1	3	19	35	44

La première table est calculée par (8), la seconde par (9).

La première table permet d'analyser la dépendance attachée, la seconde la dépendance détachée.

On observe que la dépendance marginale est faible, $I = 0.01$; par contre les dépendances attachées et détachées sont élevées $I_1 = .43$ $I_2 = .49$ et de sens contraire par C est négatif : $C = -.91$.

Il n'est pas nécessaire de procéder à une analyse factorielle des correspondances sur ces deux tables pour décrire les dépendances attachées et détachées : on observe directement sur les tables que les premières modalités de X sont fortement associées aux dernières modalités de Y dans la liaison attachée, alors qu'au contraire elles sont fortement associées aux premières modalités de Y dans la liaison détachée. Le cumul de ces 2 associations contradictoires donne une dépendance marginale non significative.

2^{e} exemple

Cet exemple montre l'intérêt des méthodes proposées même si l'ordre de grandeur de la dépendance est faible et si la dépendance détachée est relativement semblable à la dépendance marginale.

On considère la table de contingence où sont répartis les salariés par tranche de salaire annuel et âge de fin d'études. Les classes de salaire annuel sont les suivantes : (nous avons regroupé certaines classes à partir du découpage initial). 0 à 5 000, 5 000 à 10 000 ; 10 000 à 15 000, 15 000 à 20 000, 20 000 à 50 000, 50 000 à 80 000, plus de 80 000.

Les modalités de la variable âge de fin d'étude sont les suivantes : moins de 15 ans, 15 ans, 16 ans, 17 ans, 18 ans, 19 ans, 20-24 ans 25-29 ans, plus de 30 ans.

En fait on dispose de 2 tables de contingence : l'une pour les individus de 16 à 34 ans, l'autre pour ceux de plus de 35 ans.

Les données sont issues de l'INSEE [4] p. 185 et datent de 1970. Le tableau des données est dans la table 1 (on a regroupé en 2 classes les 6 premières classes de salaires.)

1) Les résultats de l'analyse factorielle des correspondances sur la table cumulée (individus de 16 à 34 ans et individus de plus de 35 ans) sont les suivants :

valeur propre	pourcentage	pourcentage cumulé
.214	84,4	84,4
.027	10,5	94,9
.012	4,6	99,5
...	...	...

Le plan principal de cette analyse est représenté par la figure 1. On observe une étroite dépendance entre salaire et âge de fin d'étude ; les différentes modalités de salaire et d'âge de fin d'étude s'alignent sur une courbe de salaire et d'âge de fin d'étude croissants. La seule anomalie consiste dans le fait que la modalité des salaires de moins de 5 000 F se trouve au-dessus de celle des salaires de 5 000 à 10 000 ; cette anomalie est gênante pour l'interprétation du graphique et mérite une explication.

2) Les résultats de l'analyse factorielle des correspondances sur la table représentant la dépendance détachée de l'âge (qui est une variable à 2 modalités), dépendance donnée par l'application de la formule (9) sont les suivants :

valeur propre	pourcentage	pourcentage cumulé
.238	85,6	85,6
.028	10,2	95,7
.011	3,8	99,6

le plan principal de cette analyse est représenté dans la figure 2. On observe cette fois que l'ordre des modalités 0-5 et 5-10 a été rétabli.

3) On peut construire la table de contingence de la dépendance attachée, en utilisant (8) et en faire l'analyse par une analyse factorielle des correspondances. Cependant cette étude ne donne qu'une valeur propre non nulle ce qui correspond au fait que la variable Z n'a que 2 modalités. Il est donc plus rapide de considérer directement les 2 tables marginales $X \times Z$ et $Y \times Z$.

On observe que les individus de plus de 35 ans ont en moyenne des salaires plus élevés et un âge de fin d'étude moins élevé que les individus de 16 à 34 ans.

4) Interprétation des résultats. La dépendance marginale obtenue en cumulant les 2 tables de contingence est la somme de 2 types de dépendances : celle positive entre âge de fin d'étude et salaire : c'est la dépendance détachée

TABLE 1

Répartition des salariés par tranche de salaire annuel et âge de fin d'études

Hommes salariés, occupés à temps complet, sans activité secondaire (a)

Source : tableau 4025

Age au 1/1/71	Tranche de salaire annuel / Age de fin d'études	de 0 à 2.999 F	de 3.000 à 3.999 F	de 4.000 à 4.999 F	de 5.000 à 5.999 F	de 6.000 à 7.999 F	de 8.000 à 9.999 F	de 10.000 à 14.999 F	de 15.000 à 19.999 F	de 20.000 à 49.999 F	de 50.000 à 79.999 F	80.000 F et plus	Total
Hommes de 16 à 34 ans	Moins de 15 ans	50.572 *2,9*	17.040 *1,0*	16.779 *1,0*	25.275 *1,5*	191.378 *11,2*	413.745 *24,2*	720.442 *42,1*	209.719 *12,3*	62.689 *3,7*	1.587 *0,1*	630 ε	1.709.856 *100,0*
	15 ans	34.376 *8,0*	13.905 *3,2*	7.859 *1,8*	7.814 *1,8*	52.526 *12,1*	67.035 *15,5*	167.690 *38,7*	49.507 *11,4*	29.730 *6,9*	1.361 *0,3*	1.126 *0,3*	432.929 *100,0*
	16 ans	24.297 *7,8*	3.108 *1,0*	3.342 *1,1*	10.324 *3,3*	39.241 *12,6*	45.881 *14,7*	102.065 *32,8*	46.247 *14,9*	35.165 *11,3*	873 *0,3*	738 *0,2*	311.281 *100,0*
	17 ans	6.998 *1,4*	2.540 *0,5*	1.152 *0,2*	4.591 *0,9*	30.875 *6,1*	90.803 *17,8*	218.845 *43,0*	102.315 *20,1*	50.857 *10,0*	259 ε	217 ε	509.452 *100,0*
	18 ans	1.572 *0,4*	966 *0,3*	– –	3.853 *1,0*	15.349 *4,2*	52.069 *14,3*	155.817 *42,7*	76.235 *20,9*	57.650 *15,8*	1.077 *0,3*	548 *0,1*	365.136 *100,0*
	19 ans	1.261 *0,6*	923 *0,5*	– –	– –	8.557 *4,3*	21.146 *10,6*	86.402 *43,3*	47.144 *23,6*	32.924 *16,5*	815 *0,4*	511 *0,2*	199.683 *100,0*
	20-24 ans	1.103 *0,3*	– –	– –	614 *0,2*	7.704 *2,2*	11.035 *3,2*	106.267 *30,5*	94.646 *27,2*	118.636 *34,1*	7.262 *2,1*	891 *0,2*	348.158 *100,0*
	25-29 ans	– –	– –	– –	– –	– –	920 *1,7*	7.502 *13,9*	12.909 *23,9*	28.800 *53,2*	3.260 *6,0*	729 *1,3*	54.120 *100,0*
	30 ans et plus	– –	– –	– –	– –	– –	– –	– –	193 *12,0*	1.415 *88,0*	– –	– –	1.608 *100,0*
	TOTAL	120.179 *3,1*	38.482 *1,0*	29.132 *0,7*	52.471 *1,3*	345.630 *8,8*	702.634 *17,9*	1.565.030 *39,8*	638.915 *16,3*	417.866 *10,6*	16.494 *0,4*	5.390 *0,1*	3.932.223 *100,0*
Hommes de 35 ans et plus	Moins de 15 ans	11.181 *0,3*	8.791 *0,3*	25.769 *0,8*	21.361 *0,7*	266.027 *8,3*	584.255 *18,1*	1.394.218 *43,3*	531.910 *16,5*	359.613 *11,2*	11.301 *0,3*	7.304 *0,2*	3.221.730 *100,0*
	15 ans	991 *0,3*	– –	2.278 *0,6*	1.121 *0,3*	11.803 *3,1*	36.771 *9,8*	128.062 *34,1*	88.502 *23,6*	98.249 *26,2*	5.821 *1,6*	1.648 *0,4*	375.246 *100,0*
	16 ans	731 *0,2*	1.110 *0,4*	– –	1.056 *0,3*	5.737 *1,8*	23.002 *7,3*	82.551 *26,2*	75.328 *23,9*	114.089 *36,3*	8.989 *2,9*	2.255 *0,7*	314.848 *100,0*
	17 ans	511 *0,1*	– –	– –	619 *0,2*	5.802 *1,9*	12.331 *3,9*	71.931 *23,1*	68.582 *22,0*	145.309 *46,6*	5.576 *1,8*	1.156 *0,4*	311.817 *100,0*
	18 ans	– –	– –	– –	515 *0,2*	1.054 *0,3*	9.607 *3,2*	53.702 *18,3*	77.866 *26,5*	137.906 *46,9*	9.329 *3,2*	4.098 *1,4*	294.077 *100,0*
	19 ans	– –	– –	– –	– –	1.006 *0,6*	2.717 *1,7*	21.460 *13,5*	24.259 *15,3*	96.030 *60,5*	10.714 *6,7*	2.626 *1,7*	158.812 *100,0*
	20-24 ans	– –	– –	– –	216 *0,1*	1.700 *0,5*	6.703 *2,0*	13.798 *4,2*	35.899 *10,9*	189.599 *57,3*	59.342 *17,9*	23.448 *7,1*	330.705 *100,0*
	25-29 ans	– –	– –	– –	– –	– –	1.018 *1,3*	3.556 *4,6*	2.560 *3,3*	37.652 *48,5*	22.930 *29,6*	9.855 *12,7*	77.571 *100,0*
	30 ans et plus	– –	– –	– –	– –	1 073 *5,0*	– –	5.221 *24,4*	1.447 *6,8*	7.629 *35,7*	4.821 *22,5*	1.201 *5,6*	21.392 *100,0*
	TOTAL	13.414 *0,3*	9.901 *0,2*	28.047 *0,5*	24.888 *0,5*	294.202 *5,8*	676.404 *13,2*	1.774.499 *34,8*	906.353 *17,8*	1.186.076 *23,2*	138.823 *2,7*	53.591 *1,0*	5.106.198 *100,0*
Hommes de tous âges	Moins de 15 ans	61.753 *1,2*	25.831 *0,5*	42.548 *0,9*	46.636 *0,9*	457.405 *9,3*	998.000 *20,2*	2.114.660 *42,9*	741.629 *15,0*	422.302 *8,6*	12.888 *0,3*	7.934 *0,2*	4.931.586 *100,0*
	15 ans	35.367 *4,4*	13.905 *1,7*	10.137 *1,3*	8.935 *1,1*	64.329 *8,0*	103.806 *12,8*	295.752 *36,6*	138.009 *17,1*	127.979 *15,8*	7.182 *0,9*	2.774 *0,3*	808.175 *100,0*
	16 ans	25.028 *4,0*	4.218 *0,7*	3.342 *0,5*	11.380 *1,8*	44.978 *7,2*	68.883 *11,0*	184.616 *29,5*	121.575 *19,4*	149.254 *23,8*	9.862 *1,6*	2.993 *0,5*	626.129 *100,0*
	17 ans	7.509 *0,9*	2.540 *0,3*	1.152 *0,1*	5.210 *0,6*	36.677 *4,5*	103.134 *12,6*	290.776 *35,4*	170.897 *20,8*	196.166 *23,9*	5.835 *0,7*	1.373 *0,2*	821.269 *100,0*
	18 ans	1.572 *0,2*	966 *0,1*	– –	4.368 *0,7*	16.403 *2,5*	61.676 *9,3*	209.519 *31,8*	154.101 *23,4*	195.556 *29,7*	10.406 *1,6*	4.646 *0,7*	659.213 *100,0*
	19 ans	1.261 *0,3*	923 *0,2*	– –	– –	9.563 *2,7*	23.863 *6,7*	107.862 *30,1*	71.403 *19,9*	128.954 *36,0*	11.529 *3,2*	3.137 *0,9*	358.495 *100,0*
	20-24 ans	1.103 *0,2*	– –	– –	830 *0,1*	9.404 *1,4*	17.738 *2,6*	120.065 *17,7*	130.545 *19,2*	308.235 *45,4*	66.604 *9,8*	24.339 *3,6*	678.863 *100,0*
	25-29 ans	– –	– –	– –	– –	– –	1.938 *1,5*	11.058 *8,4*	15.469 *11,7*	66.452 *50,5*	26.190 *19,9*	10.584 *8,0*	131.691 *100,0*
	30 ans et plus	– –	– –	– –	– –	1.073 *4,7*	– –	5.221 *22,7*	1.640 *7,1*	9.044 *39,3*	4.821 *21,0*	1.201 *5,2*	23.000 *100,0*
	TOTAL	133.593 *1,5*	48.383 *0,5*	57.179 *0,6*	77.359 *0,9*	639.832 *7,1*	1.379.038 *15,3*	3.339.529 *36,9*	1.545.268 *17,1*	1.603.942 *17,7*	155.317 *1,7*	58.981 *0,7*	9.038.421 *100,0*

(a) 4 080 hommes ayant déclaré n'avoir pas terminé leurs études au moment de l'enquête, bien qu'étant salariés, sont exclus de ce tableau.

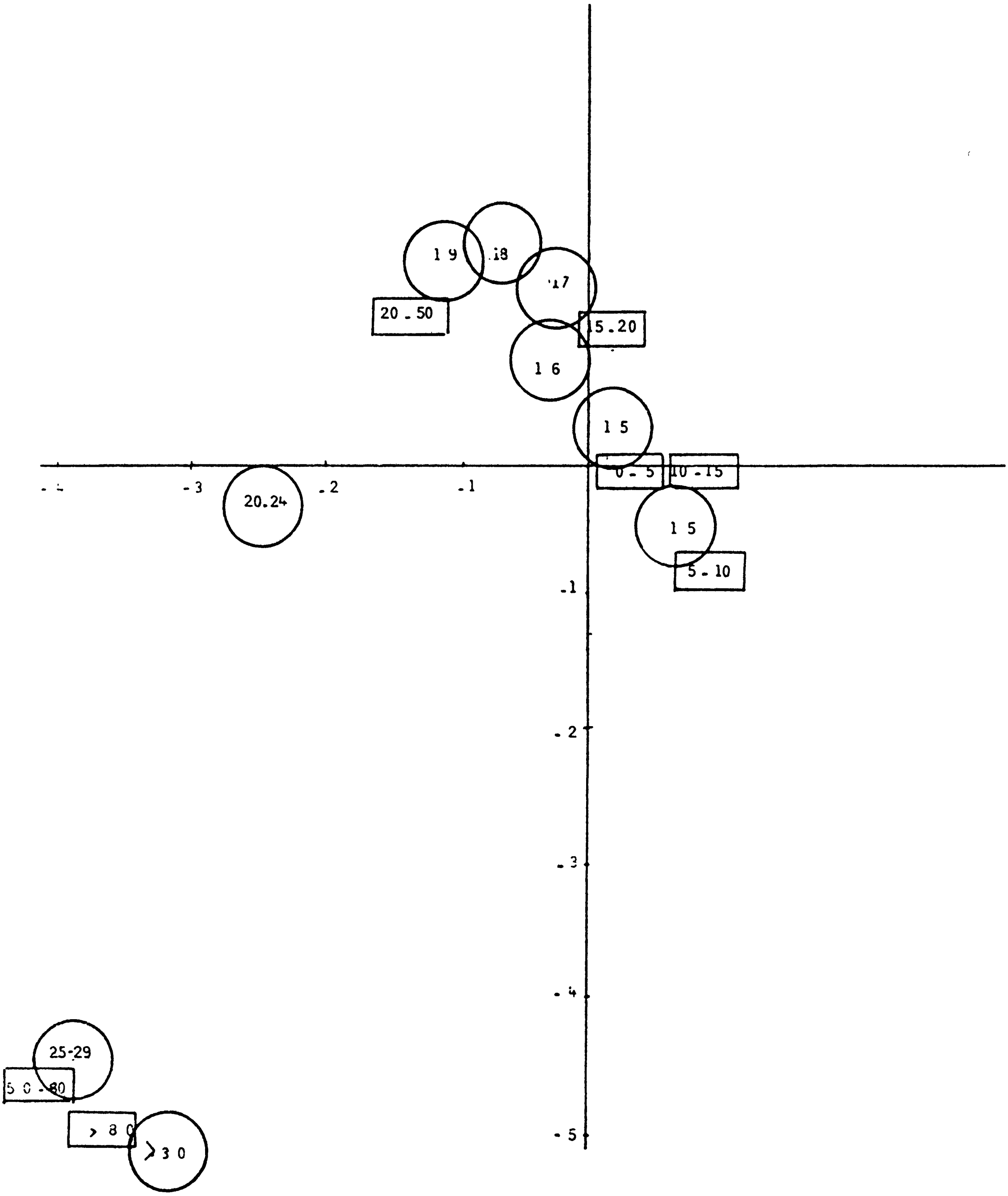

Figure 1. – Représentation simultanée des classes de fin d'étude et des tranches de salaires : plan principal de l'Analyse Factorielle des Correspondances de la liaison marginale. Les tranches de salaires sont représentées par des rectangles. Les classes d'âge de fin d'étude sont représentées par des cercles.

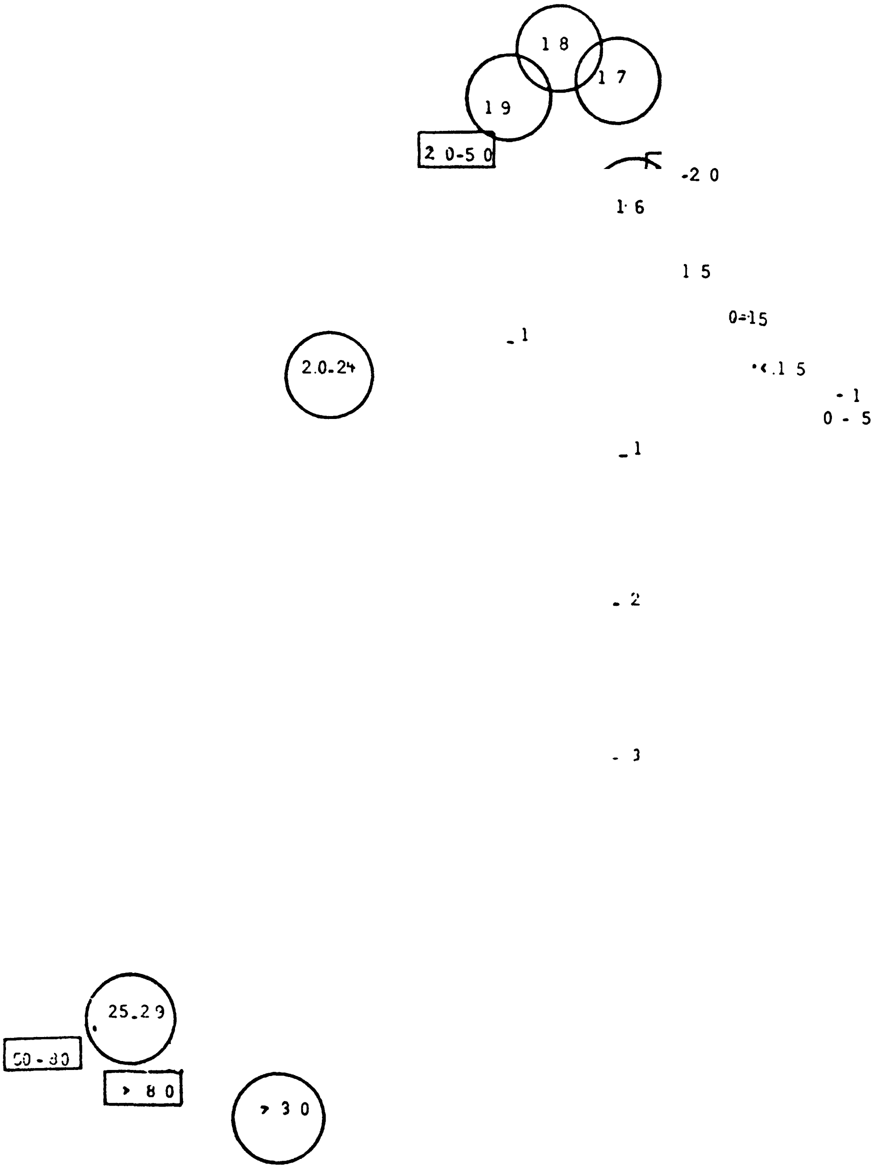

Figure 2. – Représentation simultanée des classes d'âge de fin d'étude et tranches de salaires : plan principal de l'Analyse Factorielle des Correspondances de la liaison détachée de l'âge. (les légendes sont identiques à celles de la figure 1).

(pour une classe d'âge fixée) celle négative entre âge de fin d'étude et salaire : c'est la dépendance attachée à la classe d'âge. Les ordres de grandeurs de ces différentes liaisons sont :

$$I = .265 \qquad I_2 = 0.282 \qquad I_1 = 0.003 \qquad C = -0.020$$

Bien que I_1 et C soient faibles cela suffit pour perturber les résultats de l'analyse factorielle de la dépendance marginale comme on l'a vu sur la figure 1. On peut interpréter l'interversion des 2 modalités de salaires comme un effet de la dépendance attachée aux classes d'âge : les salariés gagnant moins de 5 000 F sont en moyenne plus jeunes que les autres et ont un âge moyen de fin d'études relativement élevé.

BIBLIOGRAPHIE

[1] J.N. DARROCH. – "Marginal and Conditional Dependance and Covariance Analysis" Unpublished report 1979. Flinders University of South Australia.

[2] J.J. DAUDIN. – "Etude de la liaison entre variables aléatoires. Régression sur variables qualitatives". Thèse 3e cycle – Paris XI, 1978.

[3] J.J. DAUDIN. – "Partial association measure and an application to qualitative regression" *Biometrika,* (1980), 67, 3, pp. 581-90.

[4] INSEE. – "Enquête Formation professionnelle de 1970". Collection de l'INSEE D-32.