REVUE DE STATISTIQUE APPLIQUÉE

A. BACCINI
A. KHOUDRAJI

Application d'un modèle d'association à l'analyse d'une table de taux

Revue de statistique appliquée, tome 40, n° 4 (1992), p. 59-75
<http://www.numdam.org/item?id=RSA_1992__40_4_59_0>

Rev. Statistique Appliquée, 1992, **XXXX** (4), 59-75

APPLICATION D'UN MODÈLE D'ASSOCIATION À L'ANALYSE D'UNE TABLE DE TAUX

A. Baccini[1], A. Khoudraji[2]

(1) Laboratoire de Statistique et Probabilités, U.R.A.-C.N.R.S. D0745
118, route de Narbonne F-31062 Toulouse Cedex
(2) Département de Mathématiques, Faculté des Sciences, Université Cadi Ayyad,
Boulevard du Prince Moulay Abdellah, B.P. S15, Marrakech, Maroc

RÉSUMÉ

Dans cet article, nous proposons d'utiliser le modèle d'association défini par L. Goodman pour analyser une table de taux à deux entrées. Après avoir rappelé les méthodes classiques permettant l'analyse d'une table de contingence standard, nous écrivons le modèle d'association pour le paramètre de la loi binomiale associée au taux observé dans chaque cellule; nous décrivons et justifions ensuite une procédure moindres carrés permettant d'obtenir les estimations des paramètres de ce modèle; nous présentons enfin un exemple réel illustrant la méthode proposée.

Mots-clés : *Analyse des correspondances, Biplot, Estimation par les moindres carrés, Modèle d'association, Tables de taux.*

ABSTRACT

In this paper, we investigate the application of Goodman's association model in analysing a two-way table of rates. In a first stage, we recall standard methods used to analyse contingency tables; then we consider the association model to modelise the probability of the binomial distribution associated with the rate in each cell, and we propose a least squares procedure to estimate the parameters of this model; finally the treatment of real data is performed to illustrate the proposed method.

Key-words : *Correspondence Analysis, Biplot, Least Squares Estimation, Association Model, Tables of Rates.*

1. Introduction

1.1. Le problème considéré

Dans cet article, on s'intéresse au problème de l'analyse statistique d'une table de taux. On entend par taux le rapport $\frac{n}{s}$ entre deux entiers (il s'agit en général d'effectifs) vérifiant $0 < s$ et $0 \leqslant n \leqslant s$. Les tables que nous considérons ici sont des tables de taux à deux entrées, correspondant au croisement de deux variables discrètes (ou catégorielles), et l'objectif est d'analyser la variation des taux selon les deux variables considérées.

Les données du type décrit ci-dessus sont assez courantes dans la pratique, mais les méthodes statistiques permettant leur analyse sont relativement peu nombreuses. D'un point de vue exploratoire, les méthodes classiques d'analyse factorielle sont mal adaptées à ce type de données ; on pourrait envisager l'analyse d'une table (celle des n) en référence à une autre table (celle des s), telle que l'ont proposée, par exemple, Domenges & Volle (1979), Qannari (1983), Escofier (1984) ou Falguerolles & Heijden (1987) ; toutefois, à notre connaissance, ces méthodes n'ont pas été mises en oeuvre pour l'analyse d'une table de taux. En ce qui concerne la modélisation d'une telle table, il est courant d'utiliser soit un modèle logistique, soit un modèle log-linéaire (voir, par exemple, Agresti, 1990).

Nous proposons en fait une extension du modèle d'association défini par Goodman (1985,1986 et 1991), ainsi que sa mise en oeuvre par des méthodes de moindres carrés. L'intérêt du modèle d'association est d'une part d'offrir une écriture assez spécifique du modèle log-linéaire (les interactions y sont plus structurées), d'autre part de permettre des représentations graphiques analogues à celles utilisées en Analyse Factorielle des Correspondances (A.F.C.).

1.2. Notations

Notons X et Y les deux variables discrètes considérées et I et J leurs nombres respectifs de modalités (ou niveaux). Dans toute la suite, aucune structure ne sera prise en compte sur ces modalités, même en présence de variables ordinales, voire numériques (regroupées en classes). La population de référence sur laquelle X et Y sont observées est notée Ω et son cardinal s. On peut donc considérer la table de contingence correspondante, de terme générique s_{ij} vérifiant $\sum_{i=1}^{I}\sum_{j=1}^{J} s_{ij} = s$.

La « variable d'intérêt » est une variable binaire, notée Z, correspondant à la présence ou à l'absence d'un certain phénomène (le cancer dans l'exemple étudié en 4). Nous noterons $n_{ij} (0 \leqslant n_{ij} \leqslant s_{ij})$ le nombre d'individus qui, aux niveaux i de X et j de Y, présentent le phénomène considéré (c'est-à-dire la valeur 1 de Z). Enfin nous poserons $r_{ij} = \frac{n_{ij}}{s_{ij}}$, $\forall (i,j) \in \{1,...,I\} \times \{1,...,J\}$; r_{ij} désigne donc le taux de présence du phénomène considéré dans la catégorie (i,j).

C'est à l'analyse de la variation des taux r_{ij} que nous nous intéressons par la suite.

2. Méthodes statistiques pour l'analyse des tables de contingence

Nous rappelons brièvement ici diverses méthodes usuelles pour l'analyse des tables de contingence ; nous ne considérons que les tables à deux entrées, toutes ces méthodes pouvant se généraliser (plus ou moins simplement) au cas de tables multidimensionnelles. Les notations sont celles introduites en 1.2, mais la variable Z n'intervient pas pour l'instant.

2.1. L'Analyse Factorielle des Correspondances

Nous ne donnerons que quelques indications essentielles sur cette méthode largement utilisée aujourd'hui. Rappelons tout d'abord qu'elle consiste à réaliser des graphiques représentant conjointement les lignes et les colonnes de la table de contingence initiale et illustrant les liaisons entre les modalités correspondantes ; si l'on note q_{ij} la fréquence générique de cette table ($q_{ij} = \frac{s_{ij}}{s}$), une façon de construire ces graphiques consiste à réaliser la Décomposition en Valeurs Singulières (D.V.S.) de la matrice de terme général $\frac{q_{ij} - q_{i+}q_{+j}}{\sqrt{q_{i+}q_{+j}}}$, $\left(q_{i+} = \sum_{j=1}^{J} q_{ij} \right.$ et $\left. q_{+j} = \sum_{i=1}^{I} q_{ij} \right)$ et à en utiliser les résultats pour représenter le biplot selon le principe défini par Gabriel (1971).

Mentionnons également la formule dite de «reconstitution des données» :

$$q_{ij} = q_{i+}q_{+j}(1 + \sum_{k=1}^{K} \sqrt{\lambda_k} x_{ik} y_{jk}) \tag{2.1}$$

$K = \inf(I - 1, J - 1)$; les $\lambda_k, k = 1, ..., K$, sont les valeurs propres de l'A.F.C. (les $\sqrt{\lambda_k}$ sont les valeurs singulières de la D.V.S.) et vérifient $1 \geqslant \lambda_1 \geqslant ... \geqslant \lambda_K \geqslant 0$; les vecteurs $x_k = \{x_{ik}; i = 1, ..., I\}$ (resp. $y_k = \{y_{jk}; j = 1, ..., J\}$) sont orthonormés au sens de la métrique définie par D_I^{-1} (resp. D_J^{-1}), avec $D_I = \text{diag}(q_{1+}, ..., q_{I+})$ (resp. $D_J = \text{diag}\,(q_{+1}, ..., q_{+J})$) ; il s'agit des vecteurs normés associés aux composantes principales des lignes (resp. des colonnes).

La bibliographie sur l'A.F.C. est très importante ; pour des développements éventuels, nous renvoyons à Benzecri (1973), ainsi qu'à Greenacre (1984) qui constitue une bonne synthèse des approches française et anglo-saxonne de cette méthode.

Notons que l'A.F.C. ne s'adapte pas à l'analyse d'une table de taux puisque, dans une telle table, les quantités $q_{i+} et q_{+j}$ n'ont aucun sens.

2.2. Le modèle log-linéaire

Si l'on note Q_{ij} la probabilité théorique correspondant à la cellule (i,j) de la table de contingence considérée, le modèle log-linéaire consiste à écrire :

$$\log Q_{ij} = m + a_i + b_j + c_{ij}, \quad \forall (i,j) \tag{2.2}$$

m est l'effet moyen ; les $a_i, i = 1, ..., I$ (resp. les $b_j, j = 1, ..., J$) représentent les effets des lignes (resp. des colonnes) et vérifient $\sum_{i=1}^{I} a_i = 0$ (resp. $\sum_{j=1}^{J} b_j = 0$); les c_{ij} représentent les interactions entre les lignes et les colonnes et vérifient $\sum_{j=1}^{J} c_{ij} = 0, \forall i = 1, ..., I$ et $\sum_{i=1}^{I} c_{ij} = 0, \forall j = 1, ..., J$.

Le modèle (2.2) est saturé (c'est-à-dire qu'il contient autant de paramètres indépendants que la table analysée contient de cellules), mais on peut le simplifier en introduisant des hypothèses restrictives sur les interactions (ainsi, le modèle d'indépendance correspond à la nullité de toutes les interactions ; des types de contraintes « intermédiaires » seront d'autre part introduits en 2.4).

La bibliographie sur le modèle log-linéaire est également très importante ; citons, pour mémoire, Bishop et al. (1975) et Agresti (1990) ; on pourra d'autre part consulter Heijden et al. (1989) pour une étude comparative de l'A.F.C. et du modèle log-linéaire.

2.3. Le modèle de corrélation

Défini par Goodman (1985, 1986 et 1991), ce modèle consiste à écrire, pour tout entier M vérifiant $1 \leqslant M \leqslant K$,

$$Q_{ij} = Q_{i+}Q_{+j}\left(1 + \sum_{k=1}^{M} \sigma_k x_{ik} y_{jk}\right), \tag{2.3}$$

$$Q_{i+} = \sum_{j=1}^{J} Q_{ij}, Q_{+j} = \sum_{i=1}^{I} Q_{ij},$$

$$\sigma_1 \geqslant ... \geqslant \sigma_M > 0,$$

$$\sum_{i=1}^{I} Q_{i+} x_{ik} = \sum_{j=1}^{J} Q_{+j} y_{jk} = 0, \forall k = 1, ..., M,$$

$$\sum_{i=1}^{I} Q_{i+} x_{ik} x_{ik'} = \sum_{j=1}^{J} Q_{+j} y_{jk} y_{jk'} = \delta_{kk'}, \forall (k,k') \epsilon \{1, ..., M\}^2,$$

$\delta_{kk'}$ étant le symbole de Kronecker.

La formule (2.3) définit le modèle de corrélation d'ordre M. En comparant (2.3) à (2.1), on voit clairement que ce modèle est équivalent à l'A.F.C. de même ordre, $\sigma_k = \sqrt{\lambda_k}$ représentant la k-ième valeur singulière.

L'obtention des paramètres λ_k, x_{ik} et y_{jk} par une D.V.S. comme indiqué en 2.1 correspond à leur estimation par les moindres carrés; l'utilisation de la formule (2.3) soit dans le cadre d'un modèle multinomial, soit dans celui de lois de Poisson indépendantes pour les différentes cellules (i, j), permet d'obtenir les estimations maximum de vraisemblance; ces dernières présentent l'avantage d'être efficaces mais l'inconvénient de ne pas être «emboîtées» lorsque M croit. Sur l'estimation maximum de vraisemblance des paramètres, on pourra se reporter à Gilula & Haberman (1986).

2.4. Le modèle d'association

Ce modèle a également été proposé par Goodman (1985, 1986 et 1991), parallèlement au modèle de corrélation. A l'ordre M ($1 \leqslant M \leqslant K$), il est défini par

$$Q_{ij} = \alpha_i \beta_j \exp\left(\sum_{k=1}^{M} \phi_k \mu_{ik} \nu_{jk}\right), \tag{2.4}$$

$$\left.\begin{array}{l} \text{avec} \quad \alpha_i > 0, \forall i = 1, ..., I, \beta_j > 0, \forall j = 1, ..., J, \\ \phi_1 \geqslant ... \geqslant \phi_M > 0, \\ \quad \sum_{i=1}^{I} g_i \mu_{ik} = \sum_{j=1}^{J} h_j \nu_{jk} = 0, \forall k = 1, ..., M, \\ \sum_{i=1}^{I} g_i \mu_{ik} \mu_{ik'} = \sum_{j=1}^{J} h_j \nu_{jk} \nu_{jk'} = \delta_{kk}, \\ \forall (k, k') \epsilon \{1, ..., M\}^2. \end{array}\right\} \tag{2.5}$$

Dans ce modèle, les paramètres ϕ_k sont appelés les *associations intrinsèques*, et les μ_{ik} (resp. les ν_{jk}) les *scores* des lignes (resp. des colonnes); les quantités g_i et h_j sont des poids vérifiant :

$$g_i > 0, \forall i = i, ..., I; h_j > 0, \forall j = 1, ..., J; \sum_{i=1}^{I} g_i = \sum_{j=1}^{J} h_j = 1.$$

On précisera au paragraphe 3 les poids utilisés; par ailleurs, pour une discussion générale sur le choix de ces poids, on pourra se reporter à Becker & Clogg (1989).

Le modèle (2.4) peut être considéré comme une spécification du modèle log-linéaire, dans laquelle on impose une structure commune aux interactions.

Becker (1990) a proposé une procédure d'estimation des paramètres du modèle (2.4) par maximum de vraisemblance, ainsi que le logiciel correspondant;

par ailleurs, Baccini & Khoudragji (1990) ont défini une procédure moindres carrés permettant d'obtenir les estimations à partir d'un logiciel d'A.F.C. ou de D.V.S.

Compte tenu de l'analogie entre les modèles (2.3) et (2.4), il est assez naturel d'envisager une représentation graphique conjointe des lignes et des colonnes de la table analysée, selon le principe du biplot, au moyen des μ_{ik} et des ν_{jk} (sur ce point, voir Caussinus, 1986b).

3. Analyse d'une table de taux

Les notations sont celles introduites en 1.2 ; on considère donc une table $I \times J$ de taux $r_{ij} = \dfrac{n_{ij}}{s_{ij}}$.

3.1. Le modèle choisi

Nous supposons ici que les quantités $s_{ij} (i = 1, ..., I ; j = 1, ..., J)$ sont fixées et connues (ceci correspond souvent à la réalité, comme l'illustre l'exemple présenté en 4). Nous supposons de plus que chaque n_{ij} est l'observation d'une variable aléatoire réelle (v.a.r.) N_{ij}, les N_{ij} étant mutuellement indépendantes. N_{ij} est donc une loi binomiale de paramètres s_{ij} et p_{ij}, p_{ij} désignant la probabilité que la «variable d'intérêt» Z prenne la valeur 1 sur le sous-ensemble de Ω constitué des individus ayant présenté les modalités i de X et j de Y. Les s_{ij} étant en général grands et les p_{ij} petits, on peut approximer chaque v.a.r. N_{ij} par une loi de Poisson de paramètre $\theta_{ij} = s_{ij} p_{ij}$; compte tenu que cela nous est commode dans la justification du critère (3.2) introduit au point suivant, nous ferons dorénavant cette approximation.

Dans l'analyse d'une table de taux, le problème qui nous intéresse est l'explication des probabilités p_{ij} en fonction de X et de Y ; dans ce but, nous allons écrire les p_{ij} selon le modèle d'association d'ordre M ; on pose donc :

$$p_{ij} = \alpha_i \beta_j \exp\left(\sum_{k=1}^{M} \phi_k \mu_{ik} \nu_{jk}\right), \tag{3.1}$$

les contraintes sur les paramètres étant celles écrites en (2.5).

Un des avantages de ce modèle (par rapport, par exemple, au modèle log-linéaire) est de permettre une représentation graphique des modalités de X et de Y au moyen des scores μ_{ik} et ν_{jk} (voir plus loin). Par ailleurs, dans diverses situations, ce modèle nous a paru mieux adapté que le modèle de corrélation ou l'A.F.C. (voir, notamment, Baccini et al., 1991).

Le problème qui se pose alors est l'estimation des différents paramètres introduits dans le modèle (3.1). On peut, bien entendu, envisager une procédure maximum de vraisemblance, mais il nous a paru intéressant de proposer ici une procédure moindres carrés pour diverses raisons : tout d'abord, la simplicité de sa mise en oeuvre à partir d'un programme d'A.F.C. ou de D.V.S. et la rapidité d'une telle procédure, y compris dans le cas où l'on a de grandes valeurs de I, J et

M ; ensuite le fait que, dans ce cas là, les estimations maximum de vraisemblance sont souvent assez instables ; d'autre part, l'«emboîtement» des solutions fournies par la procédure moindres carrés est intéressant au niveau de l'interprétation des paramètres du modèle ; enfin, dans la pratique, les deux ensembles d'estimations sont souvent très proches (voir Baccini & Khoudraji, 1990).

3.2. Le critère à minimiser

Le critère des moindres carrés considéré est le suivant :

$$\sum_{i=1}^{I}\sum_{j=1}^{J} f_{i+}f_{+j}(\operatorname{Log} r_{ij} - \operatorname{Log} p_{ij})^2. \tag{3.2}$$

On a posé ici :

$$n = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}, f_{ij} = \frac{n_{ij}}{n}, f_{i+} = \sum_{j=1}^{J} f_{ij} \text{ et } \quad f_{+j} = \sum_{i=1}^{I} f_{ij};$$

p_{ij} vérifie toujours (3.1), $\forall (i,j) \in \{1,...,I\} \times \{1,...,J\}$, mais les contraintes (2.5) sont maintenant spécifiées comme suit :

$$\left.\begin{array}{l} \alpha_i > 0, \forall i = 1,...,I, \beta_j > 0, \forall j = 1,...,J, \\ \phi_1 \geqslant ... \geqslant \phi_M > 0, \\ \displaystyle\sum_{i=1}^{I} f_{i+}\mu_{ik} = \sum_{j=1}^{J} f_{+j}\nu_{jk} = 0, \forall k = 1,...,M, \\ \displaystyle\sum_{i=1}^{I} f_{i+}\mu_{ik}\mu_{ik'} = \sum_{j=1}^{J} f_{+j}\nu_{jk}\nu_{jk'} = \delta_{kk'}, \\ \forall (k,k') \epsilon \{1,...,M\}^2. \end{array}\right\} \tag{3.3}$$

La justification du critère (3.2) et des contraintes (3.3) nécessite divers développements que nous donnons ci-dessous.

(a) Rappels sur le modèle à effets fixes.

Nous rappelons ici quelques propriétés du modèle à effets fixes dans le contexte des tables de contingence, telles qu'elles sont exposées dans Caussinus (1986a), ou dans Besse et al. (1988). Soit T le tableau $m \times p$ des données observées ; on considère alors $Y_1, ..., Y_m$, m vecteurs aléatoires de $\mathbb{R}^p$ vérifiant :

- $\forall i = 1,...,m$, $E[Y_i] = y_i \in F_q$, où F_q est un sous-espace affine inconnu de $\mathbb{R}^p$ de dimension inférieure ou égale à q $(q < p)$;
- $\operatorname{Var}(Y_i) = \dfrac{\sigma^2}{w_i}\Gamma$, où σ^2 est un paramètre d'échelle, w_i est le poids de Y_i (on suppose $w_i > 0$, $\forall i = 1,...,m$, et $\displaystyle\sum_{i=1}^{m} w_i = 1$), et Γ est une matrice carrée

d'ordre p, symétrique et définie-positive ; dans ce modèle, w_i et Γ sont supposés connus.

Le problème est alors d'estimer F_q et y_i (ainsi, éventuellement, que σ^2) en utilisant le critère des moindres carrés suivant

$$\min_{y_i \in F_q, F_q \in \mathcal{F}_q} \sum_{i=1}^{m} w_i \parallel Y_i - y_i \parallel_{\Delta}^2 \tag{3.4}$$

dans lequel $\mathcal{F}_q$ est l'ensemble des sous-espaces affines de $\mathbb{R}^p$ de dimension inférieure ou égale à q, et Δ est une matrice définissant une métrique dans $\mathbb{R}^p$ (elle est donc carrée d'ordre p, symétrique et définie-positive).

Les solutions sont rappelées ci-dessous :

- $\widehat{F}_q$ est le sous-espace affine de $\mathbb{R}^p$ passant par $\overline{Y} = \sum_{i=1}^{m} w_i Y_i$ et parallèle à $\widehat{E}_q$; $\widehat{E}_q = \text{vect}(u_1, ..., u_q)$, où $u_1, ..., u_q$ sont les vecteurs propres Δ-orthornormés de ${}^tXWX\Delta$ respectivement associés aux q plus grandes valeurs propres ; $W = \text{diag}(w_1, ..., w_m)$ définit la métrique des poids dans $\mathbb{R}^m$ et $X = T - \mathbb{1}_m {}^t\overline{Y}$ représente la matrice des données centrées ($\mathbb{1}_m$ est le vecteur de $\mathbb{R}^m$ dont toutes les coordonnées valent 1) ;
- $\widehat{y}_i$ est la projection Δ-orthogonale de y_i sur $\widehat{F}_q$;
- concernant Δ, une propriété du type Gauss-Markov (voir Besse et al., 1987), conduit à choisir $\Delta = \Gamma^{-1}$.

(b) Espérance et variance du logarithme d'une loi de Poisson.

Considérons une v.a.r. N distribuée selon une loi de Poisson de paramètre $\lambda > 0$. On sait que $\dfrac{N-\lambda}{\sqrt{\lambda}} \xrightarrow[\lambda \to +\infty]{\text{loi}} N(0,1)$; par conséquent, si l'on écrit $\sqrt{\lambda}(\text{Log}\, N - \text{Log}\, \lambda) = \sqrt{\lambda}\text{Log}\,(1 + \dfrac{N-\lambda}{\lambda})$, on voit que, pour les grandes valeurs de λ, on obtient l'approximation :

$$\sqrt{\lambda}(\text{Log}\, N - \text{Log}\, \lambda) = \frac{N-\lambda}{\sqrt{\lambda}} + O_p(\frac{1}{\sqrt{\lambda}}).$$

On en déduit :

$$\sqrt{\lambda}(\text{Log}\, N - \text{Log}\, \lambda) \xrightarrow[\lambda \to +\infty]{\text{loi}} (0,1).$$

Pour les grandes valeurs de λ, on pourra donc écrire :

$$E[\text{Log}\, N] \simeq \text{Log}\, \lambda \text{ et } \text{Var}(\text{Log}\, N) \simeq \frac{1}{\lambda}. \tag{3.5}$$

Concrètement, cette approximation pourra être utilisée pour λ atteignant quelques dizaines, ce qui est manifestement le cas dans l'exemple présenté en 4.

(c) Application du modèle à effets fixes.

Nous avons supposé que les v.a.r. N_{ij} introduites en 3.1 étaient distribuées selon des lois de Poisson de paramètres respectifs $\theta_{ij} = s_{ij}p_{ij}$. Dans le cadre du modèle à effets fixes rappelé en (a), posons $Y_i = (\text{Log}\, N_{i1}, ..., \text{Log}\, N_{iJ})$, $i = 1, ..., I$. Il vient :

$$\text{Var}(\text{Log}N_{ij}) \simeq \frac{1}{\theta_{ij}} \quad \text{(d'après (3.5))}$$
$$= \frac{1}{n}\,\frac{1}{w_i}\,\frac{nw_i}{\theta_{ij}} = \frac{\sigma^2}{w_i}(\Gamma_i)_{jj}, \quad \text{avec :}$$

$\sigma^2 = \dfrac{1}{n}$ (n représente toujours le nombre total d'observations du phénomène étudié), et $\text{Var}(Y_i) = \Gamma_i = nw_i\text{diag}(..., \dfrac{1}{\theta_{ij}}, ...)$.

Pour se ramener au modèle homoscédastique défini en (a), on doit remplacer Γ_i par une matrice de covariances moyenne Γ (simplification analogue à celle faite par Caussinus, 1986a, dans le cadre de l'A.F.C.). Pour des raisons évidentes de commodité, nous choisissons ici la moyenne harmonique (pondérée par les w_i), choix par ailleurs naturel dans la mesure où les éléments diagonaux de Γ_i sont des quotients. On a donc :

$$\Gamma^{-1} = \sum_{i=1}^{I} w_i\Gamma_i^{-1} = \frac{1}{n}\sum_{i=1}^{I}\text{diag}(..., \theta_{ij}, ...) = \text{diag}(..., \frac{1}{n}\sum_{i=1}^{I}\theta_{ij}, ...),$$

où θ_{ij} est le paramètre inconnu de la loi de N_{ij}; en l'estimant par n_{ij}, on estime $\dfrac{1}{n}\displaystyle\sum_{i=1}^{I}\theta_{ij}$ par $\dfrac{n_{+j}}{n} = f_{+j}$.

On peut donc poser $\widehat{\Gamma} = \text{diag}(..., \dfrac{1}{f_{+j}}, ...)$ et $\widehat{\Delta} = \widehat{\Gamma}^{-1} = \text{diag}(..., f_{+j}, ...)$.

Le critère (3.4) s'écrit ainsi :

$$\sum_{i=1}^{I} w_i \sum_{j=1}^{J} f_{+j}(\text{Log}\, N_{ij} - E[\text{Log}\, N_{ij}])^2;$$

les lignes et les colonnes de la table de taux jouant des rôles symétriques, on doit prendre les poids w_i égaux à f_{i+}, d'où l'écriture du critère :

$$\sum_{i=1}^{I}\sum_{j=1}^{J} f_{i+}f_{+j}(\text{Log}N_{ij} - E[\text{Log}N_{ij}])^2. \qquad (3.6)$$

En utilisant encore (3.5), il vient :

$$E[\mathrm{Log}N_{ij}] \simeq \mathrm{Log}\theta_{ij} = \mathrm{Log}(s_{ij}p_{ij}).$$

En remplaçant enfin N_{ij} par sa valeur observée n_{ij}, la parenthèse de (3.6) s'écrit :

$$\mathrm{Log}\, n_{ij} - \mathrm{Log}(s_{ij}p_{ij}) = \mathrm{Log}\frac{n_{ij}}{s_{ij}} - \mathrm{Log}p_{ij} = \mathrm{Log}r_{ij} - \mathrm{Log}p_{ij}.$$

On obtient ainsi le critère (3.2), les poids g_i et h_j intervenant dans les contraintes sur les paramètres devant naturellement être remplacés par f_{i+} et f_{+j} respectivement, d'où les contraintes (3.3).

3.3. Estimation des paramètres

Nous souhaitons maintenant estimer les paramètres $\alpha_i, \beta_j, \phi_k$, μ_{ik} et ν_{jk} $(i = 1, ..., I;\ j = 1, ..., J;\ k = 1, ..., M)$ intervenant dans l'écriture (3.1) de p_{ij}, de telle sorte que le critère des moindres carrés (3.2) soit minimisé sous les contraintes (3.3).

Pour cela, considérons l'ensemble $\mathcal{M}_{I\times J}$ des matrices réelles de dimension $I \times J$ et désignons par $\| \cdot \|_{I\times J}$ la norme définie sur $\mathcal{M}_{I\times J}$ par $\| M \|^2_{I\times J} = \sum_{i=1}^{I}\sum_{j=1}^{J} f_{i+}f_{+j}m_{ij}^2$ (m_{ij} est le terme générique de $M \in \mathcal{M}_{I\times J}$).

Notons L et C les éléments de $\mathcal{M}_{I\times J}$ de termes génériques respectifs $\ell_{ij} = \mathrm{Log}r_{ij}$ et $c_{ij} = \sum_{k=1}^{M} \phi_k\mu_{ik}\nu_{jk}$. Par ailleurs, dans $\mathbb{R}^I$ (resp. $\mathbb{R}^J$), notons $\mathbb{1}_I$ (resp. $\mathbb{1}_J$) le vecteur de terme général 1 et A (resp. B) celui de terme général $\mathrm{Log}\,\alpha_i$ (resp. $\mathrm{Log}\,\beta_j$).

Le critère (3.2) peut alors se réécrire sous la forme

$$\| L - A^t\mathbb{1}_J - \mathbb{1}_I{}^t B - C \|^2_{I\times J}.$$

En suivant Gabriel (1978), on sait que la solution est obtenue en déterminant dans un premier temps $\widehat{A}$ et $\widehat{B}$ minimisant $\| L - A^t\mathbb{1}_J - \mathbb{1}_I{}^t B \|^2_{I\times J}$ (partie linéaire), puis dans un second temps $\hat{C}$ minimisant $\| L - \widehat{A}^t\mathbb{1}_J - \mathbb{1}_I{}^t \widehat{B} - C \|^2_{I\times J}$ (partie bilinéaire).

Le premier problème n'est pas identifiable puisque toutes les quantités $\sum_{j=1}^{J} f_{+j}\mathrm{Log}\, r_{ij} + E$ et $\sum_{i=1}^{I} f_{i+}\mathrm{Log}\, r_{ij} - E - \bar{\ell}$, avec E constante réelle et $\bar{\ell} = \sum_{i=1}^{I}\sum_{j=1}^{J} f_{i+}f_{+j}\mathrm{Log}\, r_{ij}$, fournissent le même minimum pour la partie linéaire du

critère ; pour résoudre cette difficulté, nous choisirons arbitrairement, mais de façon naturelle :

$$\mathrm{Log}\widehat{\alpha}_i = \sum_{j=1}^{J} f_{+j}\mathrm{Log}r_{ij} - \frac{1}{2}\bar{\ell},$$

$$\mathrm{Log}\widehat{\beta}_j = \sum_{i=1}^{I} f_{i+}\mathrm{Log}r_{ij} - \frac{1}{2}\bar{\ell}.$$

Pour la partie bilinéaire du critère, on obtient une solution unique en réalisant la D.V.S. généralisée (voir Greenacre, 1984) à l'ordre M de la matrice $L - \widehat{A}^t \mathbb{1}_J - \mathbb{1}_I {}^t\widehat{B}$ de terme générique

$$\mathrm{Log}\, r_{ij} - \sum_{j=1}^{J} f_{+j}\mathrm{Log}\, r_{ij} - \sum_{i=1}^{I} f_{i+}\mathrm{Log}\, r_{ij} + \sum_{i=1}^{I}\sum_{j=1}^{J} f_{i+}f_{+j}\mathrm{Log}\, r_{ij};$$

la D.V.S. généralisée est effectuée relativement aux matrices $\mathrm{diag}(f_{1+}, ..., f_{I+})$ et $\mathrm{diag}(f_{+1}, ..., f_{+J})$.

4. Exemple

Les données présentées dans ce paragraphe sont extraites du registre des cancers du Tarn ; il s'agit d'un registre à but épidémiologique contenant de nombreux renseignements concernant les personnes habitant ce département et chez lesquelles fut diagnostiqué un cancer au cours des années 1982, 1983 et 1984. On trouvera une présentation plus détaillée de ce registre dans Khoudraji (1986). Les deux variables retenues ici sont l'âge et le canton de résidence ; l'âge (au moment de la survenue du cancer) a été considéré avec trois niveaux (âge 1 = moins de 65 ans, âge 2 = entre 65 et 80 ans, âge 3 = plus de 80 ans) et les cantons avec six niveaux (notés de CTN1 à CTN6) ; il s'agit en fait des six classes obtenues à l'issue d'une classification réalisée (en utilisant la méthode des nuées dynamiques) sur les trente-six cantons du département du Tarn. On trouvera en annexe 1 la liste des cantons appartenant à chaque classe, ainsi que les trois variables (extraites du registre des cancers) à partir desquelles a été réalisée la classification.

Nous présentons en annexe 2 les tables donnant, pour chaque catégorie d'âge et pour chaque classe de cantons, d'une part la population totale au recensement de 1982 (table A1, correspondant aux s_{ij}), d'autre part le nombre total de cancers enregistrés sur la période considérée (table A2, correspondant aux n_{ij}). Les valeurs des taux ($r_{ij} = \dfrac{n_{ij}}{s_{ij}}$) sont données dans la table 1, dans laquelle on trouve également, pour mémoire, les taux marginaux.

Les estimations des paramètres du modèle (3.1) selon la méthode proposée en 3.3 sont données dans la table 2.

TABLE 1

Taux de cancers dans le département du Tarn selon la catégorie d'âge et la classe de cantons.

	CTN1	CTN2	CTN3	CTN4	CTN5	CTN6	taux marginaux
âge 1	0,0087	0,0077	0,0102	0,0142	0,0082	0,0062	0,0093
âge 2	0,0350	0,0399	0,0545	0,0517	0,0397	0,0332	0,0425
âge 3	0,0600	0,0451	0,0629	0,0807	0,0573	0,0500	0,0600
taux marginaux	0,0158	0,0146	0,0211	0,0247	0,0175	0,0142	0,0181

Remarques :

• les traitements réalisés dans cet article ont pris en compte 9 décimales (8 au delà du premier zéro);

• les taux ne sont pas directement interprétables, dans la mesure où le numérateur est un cumul sur 3 années (flux) alors que le dénominateur est un niveau à un moment donné (stock).

TABLE 2

Estimation des paramètres du modèle d'association appliqué aux taux de la table 1.

• effets principaux :

$\widehat{\alpha}_1 = 0,0591$ $\quad$ $\widehat{\beta}_1 = 0,1401$
$\widehat{\alpha}_2 = 0,2721$ $\quad$ $\widehat{\beta}_2 = 0,1347$
$\widehat{\alpha}_3 = 0,3821$ $\quad$ $\widehat{\beta}_3 = 0,1829$
$\widehat{\beta}_4 = 0,2120$
$\widehat{\beta}_5 = 0,1439$
$\widehat{\beta}_6 = 0,1162$

Dimension 1 $(M = 1)$:

• association intrinsèque : $\widehat{\phi}_1 = 0,0733$

• scores des lignes $(\widehat{\mu}_{i1})$:

$0,9954$
$-1,1186$
$0,6121$

• scores des colonnes $(\widehat{\nu}_{j1})$:

$0,9930$
$-0,9569$
$-1,0645$
$1,3906$
$-0,2018$
$-0,7284$

Dimension 2 $(M = 2)$:

• association intrinsèque : $\widehat{\phi}_2 = 0,0378$

• scores des lignes ($\widehat{\mu}_{i2}$) :

$$\begin{array}{r} 0,7610 \\ 0,1491 \\ -2,1146 \end{array}$$

• scores des colonnes ($\widehat{\nu}_{j2}$) :

$$\begin{array}{r} -0,8713 \\ 1,1108 \\ 0,7170 \\ 0,8334 \\ -0,6216 \\ -1,8937 \end{array}$$

Pour chaque valeur possible de M, la table 3 donne :

– le degré de liberté du modèle, qui vaut $(I - M - 1)(J - M - 1)$ (voir Becker & Clogg, 1989);

– la valeur de la statistique $\displaystyle\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij}-\widehat{n}_{ij})^2}{\widehat{n}_{ij}}$, où $\widehat{n}_{ij} = s_{ij}\widehat{p}_{ij}, \widehat{p}_{ij}$ étant calculé selon la formule (3.1), en utilisant les estimations données par la table 2. Asymptotiquement, cette statistique n'est pas distribuée selon une loi de khi-deux puisque les estimateurs $\widehat{n}_{ij}$ ne sont pas efficaces. Toutefois, dans la pratique, ces estimateurs sont peu différents des estimateurs du maximum de vraisemblance et nous utiliserons cette statistique à titre indicatif.

Remarque : On notera que l'analyse statistique réalisée ici nécessite que soient connus, en plus des taux r_{ij}, au moins les fréquences marginales du phénomène étudié (c'est-à-dire les f_{i+} et les f_{+j}); de plus, pour calculer la statistique ci-dessus, on doit disposer de toutes les quantités s_{ij} et n_{ij}.

TABLE 3

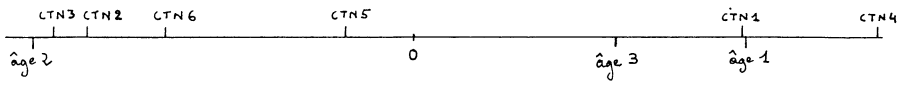
M	d.d.ℓ.	χ^2
0	10	30,74
1	4	6,75
2	0	0

Le modèle d'ordre 0 (indépendance) n'est pas satisfaisant, le modèle d'ordre 1 l'étant davantage; il fournit une reconstitution convenable de la table initiale.

L'interprétation des effets principaux est immédiate : ils sont quasiment proportionnels aux taux marginaux correspondants; ainsi, si l'on rapporte les $\widehat{\alpha}_i$ (resp. les $\widehat{\beta}_j$) aux taux marginaux des lignes (resp. des colonnes), on trouve, en ne conservant qu'une seule décimale, trois fois 6,4 (resp. des rapports compris entre 8,2 et 9,2). Les associations intrinsèques donnent une idée de l'importance relative des dimensions successives du modèle (ici, 66% pour la dimension 1 et 34% pour la dimension 2). Enfin, l'interprétation directe des scores, en termes d'interactions, n'est pas aisée; par contre, une représentation graphique simultanée des lignes et

GRAPHIQUE 1

Représentation des scores des lignes et des colonnes en dimension 1.



des colonnes de la table initiale au moyen des scores facilite grandement cette interprétation. Le graphique 1 fournit cette représentation en dimension 1.

L'interprétation en est la suivante : deux modalités des deux variables représentées proches l'une de l'autre et éloignées de l'origine correspondent à une interaction positive, c'est-à-dire à un taux plus élevé que celui attendu, en l'absence d'interaction, à la ligne et à la colonne correspondantes (exemple : âge 1 et CTN4) ; deux modalités simultanément éloignées de l'origine, et situées de part et d'autre, correspondent à une interaction négative, c'est-à-dire à un taux plus faible que celui attendu (exemple : âge 3 et CTN3) ; les modalités situées près de l'origine correspondent à des interactions faibles, voire négligeables. On retrouve donc un principe d'interprétation analogue à celui de l'A.F.C., ce qui fait l'un des intérêts de cette méthode.

5. Conclusion

Le modèle d'association proposé par L. Goodman nous paraît incontestablement un outil très performant dans l'analyse des tables de contingence. En particulier, au niveau de l'étude des interactions, il permet d'en donner une structure claire et interprétable, bien au delà de ce que permettent de faire, par exemple, le modèle log-linéaire ou le modèle logistique ; de plus, cette structure conduit à une représentation graphique simple qui s'interprète de la même façon qu'une A.F.C.

Dans le cas spécifique considéré ici, celui d'une table de taux, l'adaptation du modèle d'association a pu se faire sans difficultés, alors qu'il n'en va pas de même pour d'autres techniques telles que l'A.F.C.

La procédure moindres carrés proposée pour l'estimation des paramètres du modèle considéré présente le double avantage d'être très simple à mettre en oeuvre (la mise en œuvre se faisant par l'intermédiaire d'une D.V.S.) et de fournir des estimations stables. Toutefois, on peut, lorsqu'on le souhaite, estimer ces paramètres par maximum de vraisemblance en utilisant un logiciel tel que GLIM.

Pour l'avenir, il nous paraît intéressant d'envisager une généralisation du modèle proposé ici à des tables de taux à plus de deux entrées ; en particulier, cela nous semble pouvoir se faire sans trop de difficultés pour une table à trois entrées.

Annexe 1

Liste des cantons du département du Tarn appartenant à chacune des six classes considérées :

CTN1 = classe 1 : ALBI ;

CTN2 = classe 2 : CASTRES ;

CTN3 = classe 3 : CASTELNAU DE MONTMIRAL, GAILLAC, GRAULHET, LAVAUR, VILLEFRANCHE D'ALBIGEOIS ;

CTN4 = classe 4 : ALBAN, CARMAUX, MAZAMET ;

CTN5 = classe 5 : CORDES, DOURGNE, LABRUGUIERE, LACAUNE, LISLE SUR TARN, PAMPELONNE, PUYLAURENS, RABASTENS, REALMONT, SAINT AMANS SOULT ;

CTN6 = classe 6 : ANGLES, BRASSAC, CADALEN, CUQ TOULZA, LAUTREC, MONESTIES, MONTREDON LABESSONIE, MURAT SUR VEBRE, ROQUECOURBE, SALVAGNAC, SAINT PAUL CAP DE JOUX, VABRE, VALDERIES, VALENCE D'ALBIGEOIS, VAOUR, VIELMUR SUR AGOUT.

On notera que les regroupements n'ont pas été faits en fonction de critères géographiques, mais en fonction des ressemblances entre les personnes atteintes d'un cancer, du point de vue de l'âge, de la catégorie socioprofessionnelle et du type de cancer (variables figurant dans le registre des cancers du Tarn et s'étant révélées parmi les plus importantes à l'issue d'une Analyse des Correspondances Multiples).

Annexe 2

TABLE A1 (s_{ij})

population du département du Tarn au recensement de 1982 selon la catégorie d'âge et la classe de cantons :

	CTN1	CTN2	CTN3	CTN4	CTN5	CTN6	Total
âge 1	36 160	27 784	36 624	30 408	33 156	26 908	101 040
âge 2	8 352	5 692	8 824	7 408	9 480	7 348	47 104
âge 3	2 200	1 552	2 512	2 108	2 480	2 140	12 992
total	46 712	35 028	47 960	39 924	45 116	36 396	251 136

TABLE A2 (n_{ij})

nombre de cancers enregistrés dans le département du Tarn en 1982, 1983, et 1984 selon la catégorie d'âge et la classe de cantons :

	CTN1	CTN2	CTN3	CTN4	CTN5	CTN6	Total
âge 1	314	213	374	432	273	166	1 772
âge 2	292	227	481	383	376	244	2 003
âge 3	132	70	158	170	142	107	779
total	738	510	1 013	985	791	517	4 554

Bibliographie

AGRESTI A. (1990), "Categorical Data Analysis", Wiley, New York.

BACCINI A., CAUSSINUS H., & FALGUEROLLES A. de (1991), "RC Models in Exploratory Data Analysis", Discussion of the Paper by L.A. Goodman, *Journal of the American Statistical Association,* 86, 1115-1117.

BACCINI A., & KHOUDRAJI A. (1990), "A Least Squares Procedure for Estimating the Parameters in an Association Model", *Publications du Laboratoire de Statistique et Probabilités,* N° 05-90, Toulouse.

BECKER M.P. (1990), "Algorithm AS 253 : Maximum Likelihood Estimation of the RC(M) Association Model", *Applied Statistics,* 39, 152-167.

BECKER M.P., & CLOGG C.C. (1989), "Analysis of Sets of Two-Way Contingency Tables Using Association Models", *Journal of the American Statistical Association,* 84, 142-151.

BENZÉCRI J.P. (1973), «L'analyse des données», tome 2 : l'analyse des correspondances, Dunod, Paris.

BESSE P., CAUSSINUS H, FERRÉ L., & FINE J. (1987), «Sur l'utilisation optimale de l'Analyse en Composantes Principales», *note aux C.R.A.S.*, 304, série I, Paris.

BESSE P., CAUSSINUS H., FERRÉ L., & FINE J. (1988), "Principal Components Analysis and Optimization of Graphical Displays", *Statistics,* 19, 301-312.

BISHOP Y.M., FIENBERG S.E., & HOLLAND P.W. (1975), "Discrete Multivariate Analysis : Theory and Practice", MIT Press, Cambridge, Massachusetts.

CAUSSINUS H. (1986a), «Quelques réflexions sur la part des modèles probabilistes en analyse des données», in *Proceedings of the 4th International Symposium on Data Analysis and Informatics,* Diday et al. eds, North-Holland, Amsterdam, 151-165.

CAUSSINUS H. (1986b), discussion of the paper by L.A. Goodman, *International Statistical Review*, 54, 274-278.

DOMENGES D., & VOLLE M. (1979), «Analyse factorielle sphérique : une exploration», *Annales de l'INSEE,* 35, 3-83.

ESCOFIER B. (1984), «Analyse factorielle en référence à un modèle ; application à l'analyse de tableaux d'échange», *Revue de Statistique Appliquée,* 32, 4, 25-36.

FALGUEROLLES A. de, & HEIDJEN P.G.M. van der (1987), «Sur l'Analyse Factorielle des Correspondances et quelques unes de ses variantes», *Revue de Statistique Appliquée,* 35, 3, 7-12.

GABRIEL K.R. (1971), "The Biplot Graphic Display of Matrices with Application to Principal Component Analysis", *Biometrika,* 58, 453-467.

GABRIEL K.R. (1978), "Least Squares Approximation of Matrices by Additive and Multiplicative Models", *Journal of the Royal Statistical Society,* Series B, 40, 186-196.

GILULA Z. & HABERMAN J. (1986), "Canonical Analysis of Contingency Tables by Maximum Likelihood", *Journal of the American Statistical Association,* 81, 780-788.

GOODMAN L.A. (1985), "The Analysis of Cross-Classified Data Having Ordered and/or Unordered Categories : Association Models, Correlation Models, and Asymmery Models for Contingency Tables with or without Missing Entries", *Annals of Statistics,* 13, 10-69.

GOODMAN L.A. (1986), "Some Useful Extensions of the Usual Correspondence Analysis Approach and the Usual Log-Linear Models Approach in the Analysis of Contingency Tables", *International Statistical Reviews,*54,243-270.

GOODMAN L.A. (1991), "Measures, Models, and Graphical Displays in the Analysis of Cross-Classified Data", *Journal of the American Statistical Association*, 86, 1085-1111.

GREENACRE M.J. (1984), "Theory and Applications of Correspondence Analysis", Academic Press, Londres.

HEIJDEN P.G.M. van der, FALGUEROLLES A. de, & LEEUW J. de (1989), "A Combined Approach to Contingency Table Analysis Using Correspondence Analysis and Log-Linear Analysis", *Applied Statistics,* 38, 249-273.

KHOUDRAJI A. (1986), «Analyse statistique sur des données épidémiologiques relatives aux cancers observés dans le département du Tarn», note technique, Laboratoire de Statistique et Probabilités, Université Paul Sabatier, Toulouse.

KHOUDRAJI A. (1988), «Analyse des Correspondances et mise en œuvre du modèle de Goodman», Thèse de 3^{e} cycle, Université Paul Sabatier, Toulouse.

QANNARI E.M. (1983), «Analyses factorielles de mesures; applications», Thèse de 3^{e} cycle, Université Paul Sabatier, Toulouse.