REVUE DE STATISTIQUE APPLIQUÉE

A. DE FALGUEROLLES
S. JMEL

Un modèle graphique pour la sélection de variables qualitatives

Revue de statistique appliquée, tome 41, n° 2 (1993), p. 23-41
<http://www.numdam.org/item?id=RSA_1993__41_2_23_0>

Rev. Statistique Appliquée, 1993, **XLI** (2), 23-41

UN MODÈLE GRAPHIQUE POUR LA SÉLECTION DE VARIABLES QUALITATIVES

A. de Falguerolles, S. Jmel

Laboratoire de Statistique et Probabilités
U.R.A. C.N.R.S. D0745, Université Paul Sabatier
118, route de Narbonne, 31062 Toulouse Cedex

RÉSUMÉ

Nous proposons une méthode de sélection de variables qualitatives fondée sur les propriétés de certains modèles graphiques. Cette méthode est assez proche de celle proposée dans FALGUEROLLES et JMEL (1991) pour la sélection de variables en analyse en composantes principales. Nous la situons par rapport à une méthode proposée par BUI QUOC (1981). Deux exemples fournissent une illustration pratique.

Mots-clés : *Modèle loglinéaire, Modèle graphique loglinéaire, Sélection de variables.*

SUMMARY

This paper is concerned with a variable selection method based on the properties of some graphical models for qualitative variables. This method is closely related to the method which was proposed earlier in FALGUEROLLES and JMEL (1991) for variable selection in principal components analysis. We compare it to a method proposed by BUI QUOC (1981). Two examples illustrate this method.

Key-words : *Loglinear models, Graphical loglinear models, Selection of variables.*

1. Introduction

Les tables de contingence issues du croisement de nombreuses variables se prêtent rarement à une analyse par modélisation. Il est courant de les soumettre à un traitement préliminaire visant notamment à réduire le nombre de variables intervenant. La sous-table associée est alors retenue en vue d'une étude plus approfondie.

L'analyse factorielle des correspondances multiples (AFCM) est une méthode souvent utilisée dans ce contexte. On réalise des AFCM successives : à chaque étape, on supprime les variables peu influentes (au sens des contributions de

leurs modalités à la détermination des axes principaux), on regroupe les modalités voisines d'une même variable et on introduit parfois des variables obtenues par croisement de certaines variables initiales. Cette démarche très empirique donne, en général, des résultats satisfaisants. Cependant, elle ne peut être conduite de façon automatique.

Nous proposons dans cet article une procédure de sélection de variables. Basée sur un type particulier de modèles graphiques, elle a pour ambition de garantir une certaine stabilité des analyses loglinéaires effectuées sur les sous-tables obtenues en différentes dimensions.

Dans la section 2, nous introduisons brièvement les modèles loglinéaires pour variables qualitatives ; une attention spéciale est portée sur les modèles graphiques qui fondent notre approche de la sélection de variables. Dans la section 3, nous reprenons, à titre d'illustration, deux jeux de données traités dans WHITTAKER (1990). Dans la section 4, nous nous intéressons à l'aide à la construction des graphes qu'apportent, dans ce contexte, des méthodes factorielles de l'analyse exploratoire des données. Dans la section 5, nous présentons la méthode de sélection de variables et nous la situons par rapport à la méthode de BUI QUOC (1981) dans la section 6. Une dernière section est consacrée à la comparaison de ces deux méthodes sur les deux jeux de données précédemment introduits. La connaissance apportée par le modèle graphique qui les ajuste permet de contrôler la pertinence des sélections effectuées.

2. Modèles loglinéaires hiérarchiques et graphiques

Les modèles graphiques loglinéaires, introduits par DARROCH *et al.* (1980) à la suite des travaux de WERMUTH (1976), constituent une sous-classe des modèles loglinéaires hiérarchiques et sont d'une grande simplicité d'interprétation. En effet, l'analyse d'une table de contingence multiple au travers de modèles graphiques débouche sur la représentation exacte des interactions et/ou des indépendances conditionnelles par un graphe (au sens de la théorie des graphes). Le lecteur trouvera un exposé complet de ces modèles dans WHITTAKER (1990) ou se reportera à LAURITZEN (1982), SANTNER et DUFFY (1989), FINE (1992), ou à ANDERSEN (1990) pour une introduction rapide.

2.1. Définitions

Soit $X = (X_k)$, $k = 1, \ldots, q$, un vecteur aléatoire de distribution multinomiale multivariée. Soit $K = \{1, 2, ..., q\}$ l'ensemble des indices associés aux différentes variables et I_k l'ensemble des niveaux de la variable X_k. On note $I = \prod_{k=1}^{q} I_k$ et $i = (i_1 ..., i_k, ..., i_q)$ un élément de I. Si $a = \{k_1, ..., k_r\}$ est un sous-ensemble de K, on note X_a le sous-vecteur $(X_{k_1}, ..., X_{k_r})$ et i_a le r-uple $(i_{k_1}, ..., i_{k_r})$, $i_a \in I_a = \prod_{k \in a} I_k$. La loi de X est déterminée par les

probabilités $p(i)$ pour $i \in I$ avec $p(i) > 0$ et $\sum_{i \in I} p(i) = 1$. La distribution de X_a est aussi multinomiale multivariée et sa loi est définie par les probabilités marginales $p(i_a) = \sum_{i_{\overline{a}}} p(i_a, i_{\overline{a}})$ où $\overline{a}$ est le complémentaire de a dans K ($\overline{a} = K \setminus a$) et $i = (i_a, i_{\overline{a}})$. Si l'on considère un échantillon de taille n, $n(i)$ et $n(i_a) = \sum_{i_{\overline{a}}} n(i_a, i_{\overline{a}})$ désignent respectivement le nombre d'observations telles que $X = i$ et $X_a = i_a$. Dans la pratique ces effectifs s'obtiennent par des opérations de tabulation.

On définit un modèle loglinéaire par la donnée d'une décomposition additive du logarithme de la probabilité $p(i)$, supposée strictement positive, sous la forme $\sum_{a \subset K} u_a(i)$ où u_a est une fonction de i ne dépendant que de i_a. On notera abusivement $u_a(i_a) = u_a(i) = u_a(i_1, i_2, ..., i_q)$. La fonction $\log[p(i)]$ est appelée potentiel d'interaction et la fonction u_a, pour $a \subset K$, interaction entre les variables du sous-ensemble a. Si $\text{card}(a) = 1$, u_a est dite effet principal ; si $\text{card}(a) = 2$, u_a est dite interaction d'ordre 1 et, en général, si $\text{card}(a) = m$, u_a est dite interaction d'ordre $m - 1$.

Dans la pratique on se limite à la considération des modèles hiérarchiques, c'est-à-dire des modèles possédant la propriété suivante : si un ensemble d'interactions est annulé, il en est de même pour tout autre ensemble d'interactions dont les indices contiennent ceux de l'ensemble original ($u_a = 0 \Rightarrow u_b = 0$ pour tout $b \supset a$). Un modèle hiérarchique est donc défini par la donnée d'un ensemble de sous-ensembles deux-à-deux non comparables pour l'inclusion. Ces sous-ensembles déterminent les interactions maximales. Ces différents sous-ensembles sont appelés générateurs et constituent la classe génératrice du modèle associé.

Deux modèles souvent cités dans cet article sont le modèle saturé et le modèle de toutes interactions d'ordre 1. La classe génératrice du premier est K. Celle du second est donnée par l'ensemble des paires de variables et son développement en u-terme est de la forme :

$$\log[p(i)] = u_\varnothing + \sum_{1 \leqslant k \leqslant q} u(i_k) + \sum_{1 \leqslant k < k' \leqslant q} u(i_k, i_{k'}).$$

A chaque modèle loglinéaire hiérarchique peut être associé un graphe non orienté $G = (K, E)$ appelé indifféremment graphe d'interactions ou graphe d'indépendance (conditionnelle). Les sommets de ce graphe sont les éléments de K. Les arêtes se définissent de façon équivalente par les propriétés suivantes :

- toute arête entre deux sommets matérialise l'existence d'un u-terme d'indice a contenant ces sommets (approche graphe d'interactions) ;
- l'absence d'arête traduit l'indépendance des variables associées conditionnellement aux autres variables (approche graphe d'indépendance).

On peut vérifier facilement que plusieurs modèles hiérarchiques distincts peuvent admettre un même graphe. C'est le cas du modèle saturé et du modèle de

toutes interactions d'ordre 1. Toutefois certains modèles loglinéaires hiérarchiques dits graphiques se définissent de façon biunivoque à l'aide de leur graphe en identifiant classes génératrices et cliques du graphe. Le modèle saturé est un exemple assez particulier de modèle graphique dont le graphe est complet (une seule clique).

Rappelons que les cliques d'un graphe sont les sous-ensembles de sommets complets (sommets deux-à-deux connectés) et maximaux au sens de l'inclusion.

2.2. Estimation

Les estimations des u-termes et/ou des $p(i)$ d'un modèle hiérarchique sont, en général, obtenues par la méthode du maximum de vraisemblance (en imposant des contraintes d'identification pour les u-termes). Une première approche (celle du logiciel GLIM par exemple, AITKIN *et al.*, 1989) consiste à optimiser la vraisemblance par un algorithme de type NEWTON-RAPHSON. Une seconde approche (celle de MIM par exemple, EDWARDS, 1991) consiste à résoudre par réduction proportionnelle itérative le système d'équations :

$$\widehat{p}(i_c) = n(i_c)/n \quad \text{où} \quad c \in \mathcal{C} \quad \text{et} \quad i_c \in I_c,$$

pour tout système de générateurs $\mathcal{C}$ du modèle considéré. Remarquons que pour le modèle saturé, ce système se réduit aux seules équations $\widehat{p}(i) = n(i)/n$.

La qualité de l'ajustement d'un modèle hiérarchique M à des données se mesure par sa déviance par rapport à un modèle de base contenant le modèle M. En général, on choisit le modèle saturé et la déviance s'écrit :

$$\text{déviance}(M) = 2(\widehat{L}_S - \widehat{L}_M)$$

où $\widehat{L}_S$ (resp. $\widehat{L}_M$) est le maximum de la fonction log-vraisemblance, $L(p, n) \propto \sum_{i \in I} n(i)\log[p(i)]$, sous le modèle saturé (resp. sous le modèle M).

La déviance peut aussi s'écrire en fonction de l'information de KULLBACK-LEIBLER, $\mathcal{I}$, mesurant la divergence entre les probabilités $\widehat{p}_S$ et $\widehat{p}_M$:

$$\text{déviance}(M) = 2n\mathcal{I}(\widehat{p}_S, \widehat{p}_M) = 2n \sum_{i \in I} \widehat{p}_S(i) \log \frac{\widehat{p}_S(i)}{\widehat{p}_M(i)}$$

où $\widehat{p}_S$ (resp. $\widehat{p}_M$) est l'estimation maximum de vraisemblance de p sous le modèle saturé (resp. sous le modèle M).

2.3. Modèle loglinéaire collapsible sur un sous-ensemble de variables

La notion de collapsibilité explicite les relations entre les résultats obtenus pour une table et certaines de ses tables marginales. En ce sens, c'est une notion

susceptible d'éclairer certains aspects de la sélection de variables dans le cas théorique ou l'on connaîtrait le modèle ajustant les données initiales.

Etant donnés un modèle loglinéaire hiérarchique M de la table X et un sous-ensemble a de K, on considère le modèle M_a de la sous-table X_a ayant pour système générateur les intersections des générateurs de M et de a. Suivant ASMUSSEN et EDWARDS (1983), on dit que le modèle hiérarchique M est collapsible sur le sous-ensemble a de variables si $\widehat{p}_a(i_a) = \widehat{p}(i_a)$ est vérifiée pour tous les $i_a \in I_a$, p et p_a étant les probabilités associées aux modèles M et M_a respectivement.

La collapsibilité d'un modèle loglinéaire hiérarchique est caractérisée par une propriété du graphe d'interactions (ASMUSSEN et EDWARDS, 1983) : *un modèle hiérarchique est collapsible sur un sous-ensemble de variables a si et seulement si la frontière de toute composante connexe de $\overline{a}$ $(= K \setminus a)$ est contenue dans un générateur de ce modèle.*

Rappelons qu'une composante connexe d'un graphe est une classe d'équivalence de la relation d'équivalence d'existence d'une chaîne connectant toute paire de sommets.

La collapsibilité sur un sous-ensemble a se traduit au niveau des graphes d'interactions de M et de M_a : le graphe d'interactions de M_a préserve les indépendances conditionnelles constatées dans celui de M.

3. Exemples

3.1. Exemple 1 : Facteurs de risque pour l'infarctus

Cet exemple a été traité par EDWARDS et HAVRANEK (1985) et repris par WHITTAKER (1990). On étudie la répartition de six facteurs de risque pour l'infarctus (six variables qualitatives binaires) sur 1841 employés d'une usine de construction automobile en Tchécoslovaquie. Les six facteurs de risque sont les suivants :

A Fumeur : oui, non
B Travail mental fatiguant : oui, non
C Travail physique fatiguant : oui, non
D Pression systolique du sang : <140, >140
E Ratio de lipoprotéines α et β : <3, >3
F Antécédents familiaux d'infarctus : oui, non

Partant du modèle saturé, on teste l'indépendance conditionnelle de chaque paire de variables conditionnellement à l'ensemble des autres. On considère alors le graphe d'indépendance constitué d'arêtes associées à des dépendances significatives. On définit et l'on teste le modèle graphique associé : il apparaît que ce modèle est rejeté. On étudie alors la réinclusion de chaque paire de variables exclue. Certaines arêtes sont donc réintroduites et on obtient le graphe reproduit en figure 1.

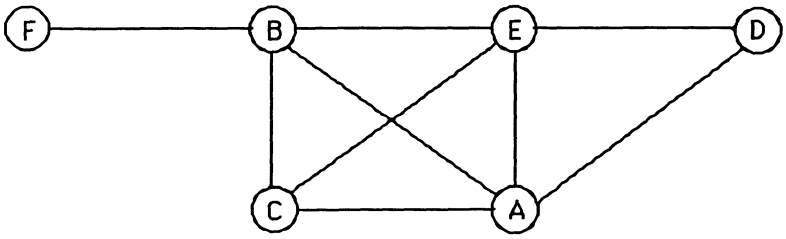

FIGURE 1
Graphe d'interactions des données de l'exemple 1

Ce modèle graphique a une déviance de 44,59 pour 42 d.d.l., ce qui permet de l'accepter avec un niveau de signification de 0,36 (pour les détails de cet ajustement, voir WHITTAKER, 1990, pp. 261-265).

3.2. Exemple 2 : Données de Rochdale

Les données considérées dans cet exemple, sont traitées dans WHITTAKER (1990) et concernent les facteurs économiques qui influencent le mode de vie à Rochdale. Un grand nombre de renseignements a été collecté auprès de 665 foyers tirés au hasard. On ne retient ici que 8 variables, toutes dichotomiques, ce qui donne une table de contingence à 2^8 (256) cellules. Ces variables sont :

A femme active : non, oui
B âge de la femme supérieur à 38 ans : non, oui
C mari au chômage : non, oui
D nombre d'enfants inférieur ou égal à 4 : non, oui
E niveau scolaire de la femme «O level» ou plus : non, oui
F niveau scolaire du mari «O level» ou plus : non, oui
G origine asiatique : non, oui
H autre membre actif : non, oui

On considère le modèle hiérarchique de toutes interactions d'ordre 1 comme modèle de départ. on teste alors l'indépendance de chaque paire de variables conditionnellement aux autres. En éliminant les interactions non significatives on obtient un modèle (non-graphique) qui est accepté. Il apparaît qu'il en est de même pour le modèle graphique déduit de son graphe d'interactions. Le graphe d'indépendance associé à ce modèle est reproduit en figure 2.

Pour les détails de cette étude, le lecteur pourra se reporter à WHITTAKER (1990, pp. 279-282).

3.3. Remarque

Dans l'exemple 1, le modèle de départ était le modèle saturé (qui est graphique). Ce choix correspond à la philosophie des modèles loglinéaires qui consiste à prendre la distribution jointe d'un ensemble de variables, considérer

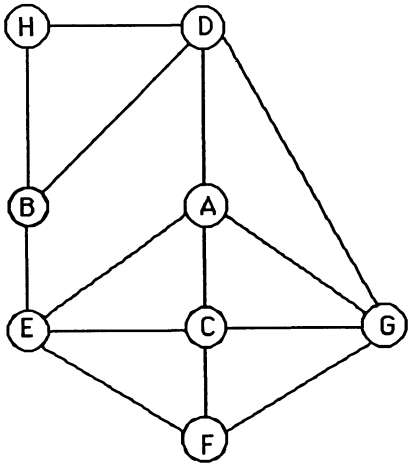

FIGURE 2
Graphe d'interactions des données de l'exemple 2

son développement en u-termes et le simplifier par des contraintes d'indépendance conditionnelle. Dans le deuxième exemple, la présence de nombreuses cellules vides (165 sur 256) ne permettait pas le choix du modèle saturé comme modèle de départ. Dans ce type de situation, on considère plutôt le modèle de toutes interactions d'ordre 1. On sait que les statistiques exhaustives de ce modèle sont toutes les tables marginales à 2 entrées, $\{n(i_k, i_{k'})\}$ pour k et $k' \in K$ (HUBER et LELLOUCH, 1974 et WHITTAKER, 1990). Il apparaît donc comme le pendant loglinéaire de l'AFCM qui analyse le tableau de BURT construit à l'aide de ces tables.

4. Aide à la construction des graphes

Le graphe d'indépendance conditionnelle visualise un modèle hiérarchique (graphique) ajusté à un ensemble de données. Cependant ce graphe est tracé de façon assez empirique. Par ailleurs, sa construction ne donne aucune information sur l'intensité des associations entre variables. Dans cette perspective, des méthodes factorielles d'analyse exploratoire, analyse factorielle des correspondances et positionnement multidimensionnel (CAILLIEZ et PAGES, 1976, GOWER, 1966 et MARDIA *et al.*, 1979), peuvent faciliter cette construction après modélisation des données (paragraphe 4.1) ou hors modélisation (paragraphe 4.2 et 4.3).

4.1. Analyse des correspondances de la matrice d'incidence

La modélisation définit au travers des u-termes significatifs la matrice d'incidence du graphe d'interactions. On peut alors utiliser une méthode classique

développée par LEBART (1984). Pour avoir une idée sur la disposition des sommets. Cette méthode consiste à effectuer une analyse des correspondances de la matrice d'incidence du graphe. On notera que, par définition, cette méthode ne peut refléter les intensités des liaisons. Les figures 3a et 3b illustrent les représentations données par cette analyse sur les deux exemples.

4.2. Positionnement multidimensionnel de la matrice des carrés des coefficients de TSCHUPROW

On peut aussi, en dehors de toute approche modélisatrice, utiliser la représentation de variables qualitatives proposée dans SAPORTA (1976). Cette approche consiste en pratique à considérer la matrice des carrés des coefficients de TSCHUPROW entre les variables prises deux à deux. Le positionnement multidimensionnel de cette matrice qui possède les caractéristiques d'une matrice de corrélation fournit une carte des variables. Les figures 4a et 4b donnent ces représentations pour les deux exemples. Pour confronter les résultats ainsi obtenus à ceux de la modélisation effectuée à la section 3, le graphe d'interactions y est aussi reporté en pointillé. Il y apparaît encore que les positionnements obtenus peuvent être utilisés avec profit dans la construction des graphes d'interactions.

4.3. Positionnement multidimensionnel de la matrice des informations conditionnelles

Une autre approche consiste à s'inspirer de ce qui a été fait dans le cas quantitatif. Dans ce cas, WHITTAKER (1988) suggère de réaliser une analyse en composantes principales de la matrice des corrélations partielles, R_p, ou de la matrice des valeurs absolues des corrélations partielles $\mathrm{abs}(R_p)$ lorsque cette dernière est semi-définie positive. Rappelons que la matrice R_p a pour terme d'indice général (k, l) la corrélation partielle entre les variables k et l, toutes les autres variables étant fixées.

Pour exploiter cette idée dans le cadre qualitatif, nous proposons de considérer la matrice des déviances d'exclusion (cf. WHITTAKER, 1990, p. 224) ou, à un coefficient près, des informations de KULLBACK-LEIBLER conditionnelles calculées par rapport au modèle saturé (exemple 1) :

$$\mathcal{I}(X_k \perp X_l \mid X_{K\setminus\{k,l\}}) = \sum_{i\in I} \frac{n(i)}{n} \log \frac{n(i)n(i_{K\setminus\{k,l\}})}{n(i_{K\setminus\{k\}})n(i_{K\setminus\{l\}})}$$

ou au modèle de toutes interactions d'ordre 1 (exemple 2) :

$$\mathcal{I}(X_k \perp X_l \mid X_{K\setminus\{k,l\}}) = \sum_{i\in I} \frac{\widehat{n}(i)}{n} \log \frac{\widehat{n}(i)\widehat{n}(i_{K\setminus\{k,l\}})}{\widehat{n}(i_{K\setminus\{k\}})\widehat{n}(i_{K\setminus\{l\}})},$$

où $\widehat{n}(\cdot)$ désigne l'estimation du maximum de vraisemblance de $np(\cdot)$ sous ce modèle (WHITTAKER, 1990, pp. 104-108 et pp. 222-224).

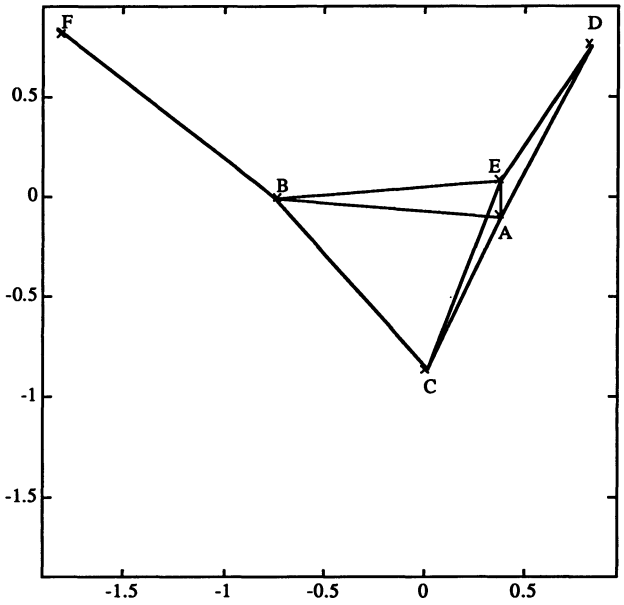

FIGURE 3a

Analyse des correspondances de la matrice d'incidence du graphe du modèle ajusté aux données de l'exemple 1.

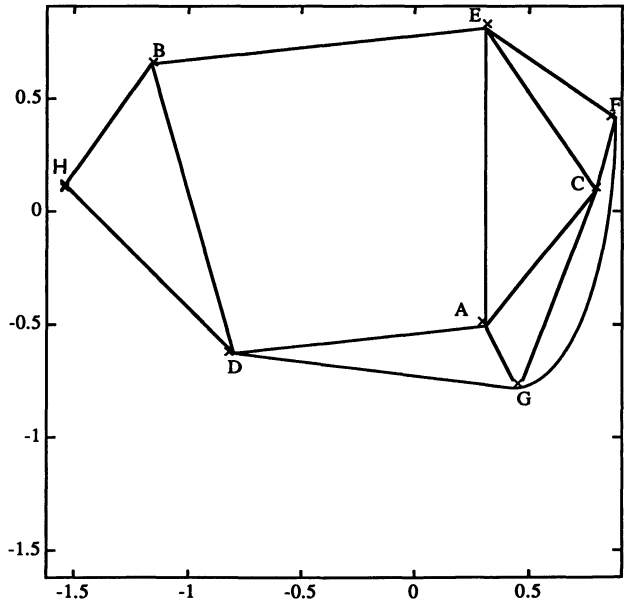

FIGURE 3b

Analyse des correspondances de la matrice d'incidence du graphe du modèle ajusté aux données de l'exemple 2.

Le positionnement multidimensionnel de cette matrice, considérée comme une matrice de similarité, fournit une ébauche spatiale du graphe qui rend compte des liaisons entre variables. Notons que ces matrices ne sont pas forcément semi-définies positives et que cela pose quelques problèmes classiques en positionnement multidimensionnel métrique. Les figures 5a et 5b illustrent cette approche sur les deux exemples.

Signalons enfin que, dans cette esprit, SAPORTA (1976) avait introduit la matrice des coefficients partiels de TSCHUPROW comme l'analogue, pour des variables qualitatives, d'une matrice de corrélations partielles. Le positionnement multidimensionnel de la matrice de ces coefficients donne des résultats intéressants bien que ces coefficients ne soient pas justifiés d'un point de vue théorique. En effet, on peut avoir indépendance conditionnelle entre deux variables sans pour autant que le coefficient de TSCHUPROW partiel correspondant soit nul, et inversement (DAUDIN, 1977).

5. Présentation d'une méthode de sélection de variables qualitatives

La méthode de sélection de variables que nous proposons est fondée sur certaines propriétés des modèles graphiques et le principe de base des modèles de structure latente (voir EVERITT, 1984 et McCUTCHEON, 1987).

Considérons la situation «idéale» où les variables se répartissent en deux sous-ensembles a et b tels que

(i) les variables de X_b sont deux à deux indépendantes conditionnellement aux variables de X_a,

(ii) la réunion de toute variable de X_b aux variables de X_a constitue une clique.

Il serait alors naturel de sélectionner les variables de X_a : les modèles de structure latente supposent que les associations entre variables sont expliquées par leur interaction avec les variables latentes (ici les variables de X_a).

Le principe de la méthode consiste à fixer *a priori* le nombre de variables à sélectionner et à rechercher le modèle graphique du type ci-dessus ajustant au mieux les données. Il apparaît que de tels modèles présentent une expression simple de leur déviance. La recherche d'un modèle de déviance minimum s'en trouve simplifié.

5.1. Expression de la déviance

Considérons un modèle du type précédent. Soit $X_a = (X_1, X_2, ..., X_r)$ le vecteur des variables sélectionnées et $X_b = (X_{r+1}, X_{r+2}, ..., X_q)$ celui des variables non sélectionnées. L'hypothèse (i) se traduit par

$$P(X_b = i_b \mid X_a = i_a) = \prod_{k=r+1}^{q} P(X_k = i_k \mid X_a = i_a)$$

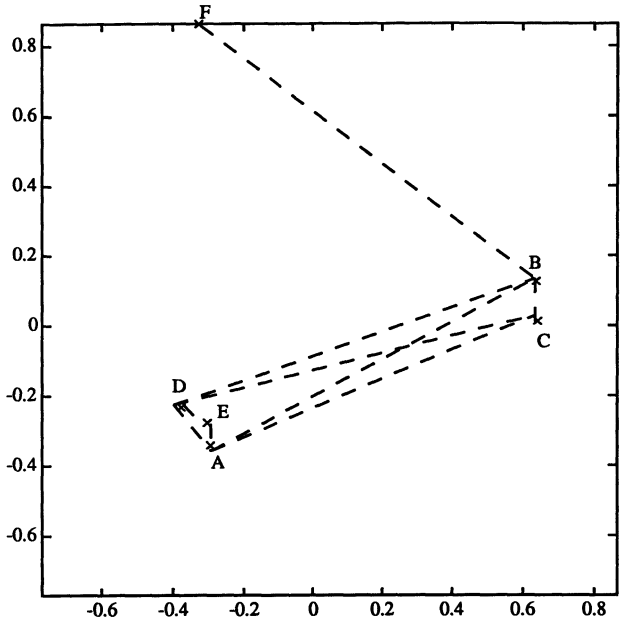

FIGURE 4a

Positionnement multidimensionnel de la matrice des carrés des coefficients de TSCHUPROW entre couples de variables de l'exemple 1. Le graphe considéré à la section 3, non fourni par la méthode, y est reporté en pointillé.

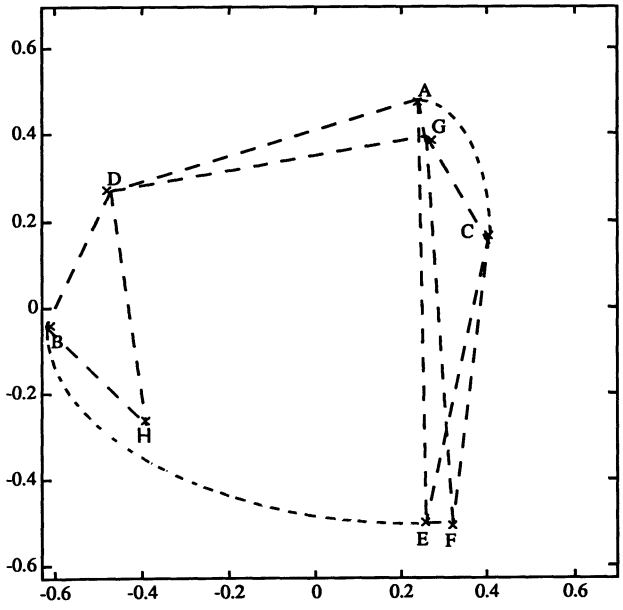

FIGURE 4b

Positionnement multidimensionnel de la matrice des carrés des coefficients de TSCHUPROW entre couples de variables de l'exemple 2. Le graphe considéré à la section 3, non fourni par la méthode, y est reporté en pointillé.

où i_a, i_b et i_k désignent respectivement des réalisations des sous-vecteurs X_a et X_b et de la variable X_k. On peut alors écrire

$$P(X = i) = P(X_a = i_a) \prod_{k=r+1}^{q} P(X_k = i_k \mid X_a = i_a)$$

et donc

$$p(i) = p(i_a) \prod_{k=r+1}^{q} \frac{p(i_{a\cup\{k\}})}{p(i_a)}.$$

Les hypothèses *(i)* et *(ii)* rendent le modèle collapsible par rapport à ses $q-r$ cliques et au sous-ensemble a. Par conséquent l'estimation de $p(i)$ pour $i \in I$ est donnée par

$$\widehat{p}(i) = \widehat{p}(i_a) \prod_{k=r+1}^{q} \frac{\widehat{p}(i_{a\cup\{k\}})}{\widehat{p}(i_a)} = \frac{n(i_a)}{n} \prod_{k=r+1}^{q} \frac{n(i_{a\cup\{k\}})}{n(i_a)}.$$

La déviance s'écrit alors

$$\begin{aligned}\text{déviance} &= 2n \sum_{i\in I} \frac{n(i)}{n} \log \frac{\dfrac{n(i)}{n}}{\dfrac{n(i_a)}{n} \displaystyle\prod_{k=r+1}^{q} \dfrac{n(i_{a\cup\{k\}})}{n(i_a)}} \\ &= 2n(H(X_a) + \sum_{k=r+1}^{q} H(X_k \mid X_a) - H(X)),\end{aligned}$$

où

$$\begin{aligned}H(X_a) &= -\sum_{i_a\in I_a} \frac{n(i_a)}{n} \log \frac{n(i_a)}{n}, \\ H(X_k \mid X_a) &= -\sum_{i_{a\cup\{k\}}\in I_{a\cup\{k\}}} \frac{n(i_{a\cup\{k\}})}{n} \log \frac{n(i_{a\cup\{k\}})}{n(i_a)}, \\ H(X) &= -\sum_{i\in I} \frac{n(i)}{n} \log \frac{n(i)}{n}.\end{aligned}$$

On retrouve ici les expressions d'entropies empiriques $H(X)$ pour la table définie par X, $H(X_a)$ pour la table définie par X_a et l'entropie conditionnelle empirique $H(X_k \mid X_a)$ pour la table définie par $X_{a\cup\{k\}}$. Ces entropies (au sens de SHANNON), compte-tenu des propriétés de collapsibilité du modèle saturé, sont aussi des estimations du maximum de vraisemblance sous n'importe quel modèle collapsible sur les tables considérées lorsqu'il en existe.

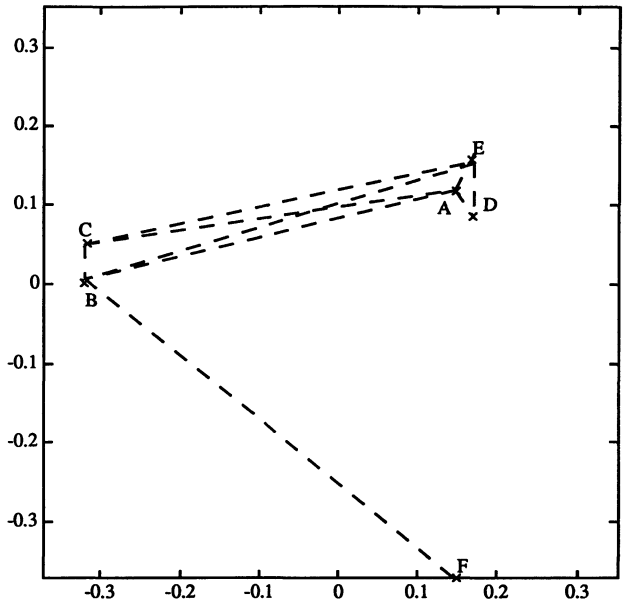

FIGURE 5a

Positionnement multidimensionnel de la matrice des informations conditionnelles de KULLBACK-LEIBLER (exemple 1). Le graphe considéré à la section 3, non fourni par la méthode, y est reporté en pointillé.

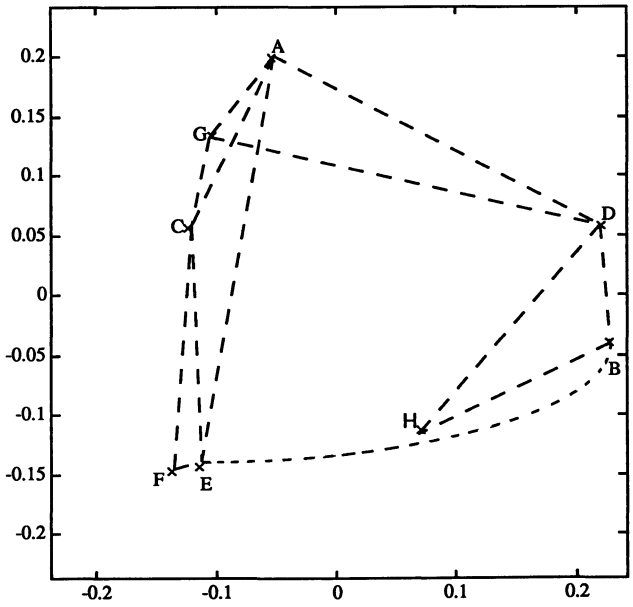

FIGURE 5b

Positionnement multidimensionnel de la matrice des informations conditionnelles de KULLBACK-LEIBLER (exemple 2). Le graphe considéré à la section 3, non fourni par la méthode, y est reporté en pointillé.

On notera enfin que cette déviance et ces différentes entropies sont aisément calculées à partir de tabulations élémentaires des données.

5.2. Procédure de sélection

La procédure de sélection consiste, pour $1 \leqslant r \leqslant q-2$ fixé, à rechercher la meilleure approximation du modèle proposé. *On ajuste donc tout ou partie de l'ensemble des modèles graphiques loglinéaires définis comme ci-dessus et on retient le sous-ensemble de variables réalisant la plus petite déviance.*

On notera d'une part, que r n'est pas connu *a priori* et que la sélection s'obtient donc en faisant varier r de 1 à $q-2$; d'autre part, que pour r fixé, on peut choisir ou non de ne pas remettre en cause les sélections de l'étape $r-1$. Dans le premier cas, on obtient des sélections emboîtées. C'est la démarche que nous avons adoptée dans cet article.

6. Méthode de BUI QUOC

Dans cette section, nous présentons la méthode de BUI QUOC (1981) et une de ses variantes en nous attachant à expliciter les liens avec la méthode que nous proposons.

6.1. Réduction proportionnelle de l'incertitude

Soient X_k et X_l deux variables qualitatives. Classiquement, on mesure la proportion de variation de X_k expliquée par X_l par la quantité :

$$\mathrm{PRU}(X_k \mid X_l) = \frac{H(X_k) - H(X_k \mid X_l)}{H(X_k)}$$

où PRU est une abréviation de "proportional reduction in uncertainty" (réduction proportionnelle de l'incertitude) et H désigne l'entropie (au sens de SHANNON). Il n'est fait l'hypothèse d'aucun modèle loglinéaire pour X autre que le modèle saturé. On se réfère à des entropies empiriques aisément calculées.

Le PRU peut prendre toute les valeurs entre 0 et 1. Il prend la valeur 0 si X_k et X_l sont indépendantes. Il prend la valeur 1 lorsque la variable X_k est «pleinement prédictible» à partir de X_l. En particulier $\mathrm{PRU}(X_k \mid X_k) = 1$. Pour plus de détail sur le PRU on peut se reporter à STRONKHORST et PANNEKOEK (1984) ou à ISRAELS *et al.* (1984).

Cette mesure peut être étendue au cas du conditionnement par plusieurs variables formant le sous-vecteur X_a. Elle devient alors

$$\mathrm{PRU}(X_k \mid X_a) = \frac{H(X_k) - H(X_k \mid X_a)}{H(X_k)}.$$

Remarquons qu'on trouve dans l'expression du PRU les différentes entropies intervenant dans le calcul de la déviance des modèles graphiques de sélection considérés en section 5.

6.2. Méthode de BUI QUOC

Le principe de sélection séquentielle de BUI QUOC consiste à sélectionner une variable qui réalise le plus grand PRU moyen calculé sur les variables non sélectionnées.

BUI QUOC propose un algorithme de sélection en bloc en trois étapes. Etant construite la matrice $q \times q$ des PRU entre toutes les paires de variables, on procède comme suit :

a) calcul pour chaque variable du total-ligne des PRU correspondants.

b) sélection d'une variable qui réalise le plus grand total-ligne des PRU.

c) suppression de la ligne et de la colonne correspondant à la variable sélectionnée et retour à l'étape *a)*.

A chaque étape r, un pourcentage de l'information cumulée PIC est calculé à partir de la relation de récurrence :

$$\text{PIC}(r) = \text{PIC}(r-1) + [1 - \text{PIC}(r-1)] \frac{\text{SPRU}(r)}{N - r + 1}$$

où $\text{SPRU}(k)$ est la somme des PRU de la variable sélectionnée à l'étape k et avec $\text{PIC}(0) = 0$.

6.3. Variante de la méthode de BUI QUOC

La méthode de BUI QUOC, dans sa version simplifiée, ne tient plus compte des variables dès lors qu'elles sont sélectionnées. Une extension naturelle de cette méthode consiste à considérer, à chaque étape r, les variables sélectionnées $X_{a_{r-1}}$ au côté des variables candidates X_l, $l \in K \setminus a_{r-1}$. Considérant la généralisation du PRU associé à $X_{a_{r-1}}$ on recherchera une variable X_l qui maximise la quantité

$$\sum_{k \in K \setminus a_{r-1} \cup \{l\}} \text{PRU}(X_k \mid X_{a_{r-1} \cup \{l\}}) = \sum_{k \in K \setminus a_{r-1} \cup \{l\}} \frac{H(X_k) - H(X_k \mid X_{a_r})}{H(X_k)}.$$

où $a_r = a_{r-1} \cup \{l\}$.

7. Comparaison des deux méthodes

Dans cette section, nous comparons notre méthode à celle de BUI QUOC (version initiale et variante) sur les deux jeux de données introduits à la section 3.

7.1. Comparaison à la méthode de BUI QUOC

A chaque étape, la méthode de BUI QUOC ne reconsidère plus la variable sélectionnée à l'étape précédente. Tout se passe donc comme si l'on travaillait sur la sous-table obtenue en marginalisant la table courante par rapport à cette variable. Pour que cette comparaison se fasse dans des conditions équitables, à chaque étape, notre méthode est appliquée à la sélection d'une seule variable dans la sous-table obtenue de façon analogue.

Les résultats de cette comparaison sont consignés dans les tableaux 1a (exemple 1 : 6 variables) et 1b (exemple 2 : 8 variables). A titre indicatif, on donne, pour les différentes tailles de sélection r $(r = 1, ..., q-2)$, la déviance et le nombre de degrés de liberté du modèle graphique défini par les variables sélectionnées (selon l'une ou l'autre méthode) ainsi que celles obtenues sans la contrainte d'emboîtement.

TABLEAU 1a

	Méthode basée sur les modèles graphiques de sélection			Méthode de BUI QUOC				
étape	variable retenue	déviance	d.l.l.	variable retenue	PIC	déviance	d.d.l.	déviance optimale
1	C	113,566	52	C	0.22	113,566	52	113,566
2	C, E	67,500	44	C, E	0,38	67,500	44	66,475
3	C, E, A	39,271	32	C, E, A	0,53	39,271	32	35,475
4	C, E, A, F	12,226	16	C, E, A, B	0,69	18,345	16	12,226

TABLEAU 1b

	Méthode basée sur les modèles graphiques de sélection			Méthode de BUI QUOC				
étape	variable retenue	déviance	d.l.l.	variable retenue	PIC	déviance	d.d.l.	déviance optimale
1	D	505,17	240	D	0,18	505,17	240	505,17
2	D, G	338,86	228	D, A	0,34	347,00	228	292,14
3	D, G, E	194,50	208	D, A, E	0,48	216,73	208	164,97
4	D, G, E, B	117,64	176	D, A, E, B	0,60	125,01	176	95,57
5	D, G, E, B, C	61,90	128	D, A, E, B, C	0,72	64,03	128	46,26
6	D, G, E, B, C, F	20,68	64	D, A, E, B, C, F	0,81	12,35	64	12,35

Les deux méthodes donnent des sélections assez proches. A l'étape 4 du tableau 1a, il se trouve que la sélection par la méthode «graphique» est optimale (plus petite déviance pour tout modèle graphique de sélection de 4 variables). C'est aussi le cas pour la méthode de BUI QUOC à l'étape 6 du tableau 1b.

7.2.Comparaison à la variante de BUI QUOC

On compare la méthode «graphique» à la variante de BUI QUOC. On effectue donc une sélection pas à pas en conservant et tenant compte, à chaque étape, des variables sélectionnées à l'étape précédente pour l'optimisation des critères considérés.

TABLEAU 2a

Méthode basée sur les modèles graphiques de sélection	Variante de la méthode de BUI QUOC

étape	variable retenue	déviance	d.l.l.	variable retenue	déviance	d.l.l.	déviance optimale
1	**C**	113,566	52	**C**	113,566	52	113,566
2	C, **E**	67,500	44	C, **D**	82,715	44	66,475
3	C, E, **D**	38,915	32	C, D, **E**	38,915	32	35,475
4	C, E, D, **B**	21,305	16	C, D, E, **F**	22,651	16	12,226

TABLEAU 2b

Méthode basée sur les modèles graphiques de sélection	Variante de la méthode de BUI QUOC

étape	variable retenue	déviance	d.l.l.	variable retenue	déviance	d.l.l.	déviance optimale
1	**D**	505,17	240	**D**	505,17	240	505,17
2	D, **E**	315,05	228	D, **E**	315,05	228	292,14
3	D, E, **A**	216,73	208	D, E, **A**	216,73	208	164,97
4	D, E, A, **B**	125,01	176	D, E, A, **F**	141,83	176	95,57
5	D, E, A, B, **F**	53,35	128	D, E, A, F, **H**	63,70	128	46,26
6	**D, E, A, B, F, C**	12,35	64	D, E, A, F, H, **C**	21,05	64	12,35

Les résultats sont consignés dans les tableaux 2a (exemple 1) et 2b (exemple 2) en indiquant les valeurs indicatives de la déviance et du nombre de degrés de liberté associés aux modèles graphiques définis par les différentes sélections.

Les deux méthodes donnent encore des résultats très proches qui, de plus, ne diffèrent pas trop de ceux obtenus lors de la première comparaison. On notera qu'à l'étape 6 du tableau 2b, la sélection (variables en caractères gras) par la méthode «graphique» est optimale.

Conclusion

Il ressort de cette comparaison que ces méthodes donnent des sélections de qualités comparables. En effet, disposant pour ces exemples d'un modèle graphique ajustant les données initiales, on constate que les variables sélectionnées «couvrent» le graphe associé au modèle : toute variable non sélectionnée est connectée à des variables sélectionnées. C'est ce que l'on constate en particulier pour les données de l'exemple 2 et la sélection des variables B, D, E et G.

Il apparaît aussi que les données initiales ne sont pas toujours collapsibles sur les sous-ensembles de variables sélectionnées. C'est le cas du deuxième exemple et du sous-ensemble des variables B, D, E et G. On notera cependant que la contrainte de collapsibilité pourrait être introduite sans trop de difficulté dans les méthodes de sélection.

Les sous-tables obtenues par sélection peuvent à leur tour être soumises à l'analyse loglinéaire. La figure 6 donne les graphes d'interactions obtenus pour les données de l'exemple 2 et la sélection des variables B, D, E et G. On y notera

comment les interactions BG et EG, introduites au côté de celles conservées, traduisent la complexité des liaisons dans la table initiale. Leur interprétation demeure délicate.

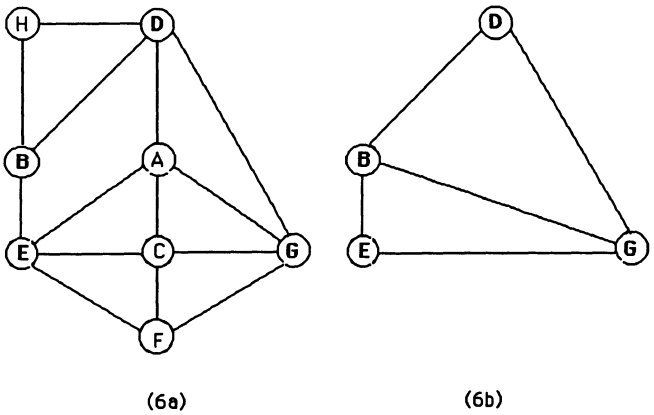

FIGURE 6
Graphes d'interactions des données de Rochdale (exemple 2) : le graphe (6a) est celui du modèle des données initiales (cf. §3.2) tandis que le graphe (6b) est celui du modèle ajusté à la sélection de 4 variables (cf. § 5 et tableau 1b).

Bibliographie

AITKIN, M., ANDERSON, D., FRANCIS, B. et HINDE, J. (1989). Statistical Modelling in GLIM. Clarendon Press. Oxford.

ANDERSEN, E.B. (1990). The statistical analysis of categorical data. Springer-Verlag.

ASMUSSEN, S. et EDWARDS, D. (1983). Collapsibility and response variables in contingency tables. Biometrika, 70, 567-578.

BUI QUOC, T. (1981). A data reduction based on information theory. In Contingency Table Analysis for Road Safety Studies. G.A. FLEISCHER (edi.). Series E : Applied Sciences. N° 42.

CAILLIEZ, F. et PAGES, J.P. (1976). Introduction à l'Analyse des Données. SMASH, Paris.

DARROCH, J.N., LAURITZEN, S.L. et SPEED, T.P. (1980). Markov field and loglinear interaction models for contingency tables. Annals of Statistics, 8, 522-539.

DAUDIN, J.J. (1977). Coefficient de TSCHUPROW partiel et indépendance conditionnelle. Statistique et Analyse des Données, 3, 55-58.

EDWARDS, D. (1991). A guide to MIM 2.0, Statistical Research Unit, University of Copenhagen, Danemark.

EDWARDS, D. et HAVRANEK, T. (1985). A fast procedure for model search in multidimensional contingency tables. Biometrika, 72, 2, 339-351.

EVERITT, B. (1984). An Introduction to Latent Variable Models. Chapman & Hall, Londres.

FALGUEROLLES, A. de et JMEL, S. (1991). Un critère de choix de variables en analyse en composantes principales fondé sur des modèles graphiques gaussiens particuliers. A paraître dans la Revue Canadienne de Statistique.

FINE, J. (1992). Modèles graphiques d'associations. Dans Modèles pour l'analyse des données multidimensionnelles. Economica, Paris, 267-313.

GOWER, J.C. (1966) Some distance properties of latent root and vecteur methods used in multivariate analysis. Biometrika 53, 325-338.

HUBER, C. et LELLOUCH, J. (1974). Estimation dans les tableaux de contingence a un grand nombre d'entrées. Int. Stat. Rev., Vol. 42, N°. 2, 193-203.

ISRAELS, A.Z., BETHLEHEM, J.G., Van DRIEL, J., JANSEN, M.E., PANNEKOEK, J. et REE, S.J.M. de (1984). Méthodes d'analyse multidimensionnelle pour variables discretes. In Développements récents dans l'analyse de grands ensembles de données. Information de l'Eurostat, numéro spécial, Luxembourg.

LAURITZEN, S.L. (1982). Lectures on contingency tables. University of Aalborg Press : Aalborg.

LEBART, L. (1984). Correspondence analysis of graph structure. Bulletin Technique du CESIA, 2, N° 1-2, 5-19.

MARDIA, K.V., KENT, J.T. et BIBBY, J.M. (1979). - Multivariate Analysis. Academic Press, New York, NY.

McCUTCHEON, A.L. (1987). Latent Class Analysis. Sage University Paper on Quantitative Applications in the Social Sciences. Beverly Hills, CA.

SANTNER, T.J. et DUFFY, D.E. (1989). The Statistical Analysis of Discrete Data. Springer-Verlag.

SAPORTA, G. (1976). Quelques applications des opérateurs d'ESCOUFIER au traitement des variables qualitatives. Statistique et Analyse des Données. 1, 38-46.

STRONKHORST, H. et PANNEKOEK, J. (1984). Passage de l'enseignement primaire à l'enseignement secondaire en 1964 et 1977, une analyse loglinéaire. In Développements récents dans l'analyse de grands ensembles de données. Information de l'Eurostat, numéro spécial, Luxembourg.

WERMUTH, N. (1976). Analogies between multiplicative models in contingency tables and covariance selection. Biometrics, 32, 95-108.

WHITTAKER, J. (1988). E.S.R.C. Workshop : "Graphical modelling : Transparencies". Technical Report. Department of Statistics, University of Lancaster.

WHITTAKER, J. (1990). Graphical Models in Applied Multivariate Statistics. Wiley.