Rev. Statistique Appliquée, 1987, **XXXV** (3), 7-12

SUR L'ANALYSE FACTORIELLE DES CORRESPONDANCES ET QUELQUES UNES DE SES VARIANTES

A. de FALGUEROLLES

Laboratoire de Statistique et Probabilités — U.A. — C.N.R.S. 745 Université Paul Sabatier, Toulouse, France

P.G.M. van der HEIJDEN

Department of Psychology, University of Leiden, Leiden, Pays-Bas

RÉSUMÉ

Ce court article a pour but de mettre en évidence quelques aspects de la complémentarité des méthodes d'analyse factorielle des correspondances et de modélisation.

Mots clés : *Analyse des correspondances, Modélisation.*

SUMMARY

The aim of this short article is to point out some aspects of the complementarity of correspondence analysis type methods and of modelling methods.

Introduction

L'objet de ce travail est de présenter dans un même cadre synthétique et très succinctement une démarche de l'analyse factorielle des correspondances (A.F.C.) et de quelques-unes de ses variantes ou extensions. Il n'a pas l'ambition (démesurée) de recouvrir l'ensemble des développements anciens ou récents de cette technique. Il résulte d'un effort de synthèse qui nous est apparu fécond dans la mesure où il nous permet de rendre compte simplement d'un certain nombre de variantes courantes de l'A.F.C. Il utilise le cadre formel retenu par ESCOUFIER (1985) pour présenter l'analyse factorielle des correspondances, ses propriétés et ses extensions.

Dans une première partie nous rappelons la démarche de l'A.F.C. en la situant dans une perspective qui conduit assez naturellement aux variantes ou extensions que nous étudions dans une seconde partie.

1. A.F.C.

L'A.F.C. peut être présentée comme une technique de représentation graphique simultanée des profils lignes et des profils colonnes d'une table de contingence (BENZECRI (1973)). Il apparaît en fait que l'A.F.C. fournit une représentation graphique des profils centrés (centrage par rapport au profil

marginal ligne et au profil marginal colonne). Il en résulte que l'A.F.C. peut être considérée comme une méthode de recherche d'interactions significatives dans les résidus du modèle log-linéaire d'indépendance. C'est notamment le point de vue explicité et retenu par ESCOUFIER (1985). On notera cependant que l'A.F.C. peut être replacée dans le cadre de problématiques très différentes.

1.1. Une stratégie

C'est cette articulation entre modèle et recherche de structure dans les résidus du modèle qui retient ici notre attention. L'A.F.C. nous paraît être l'expression d'une stratégie en trois étapes :

1. choix d'un modèle (le modèle d'indépendance) et ajustement à ce modèle,
2. suppression de la part expliquée des données,
3. recherche de structure encore présente dans la part inexpliquée des données.

On retrouve là une démarche courante en analyse exploratoire des données où, le cas échéant, elle est appliquée de façon itérative, l'objectif final étant d'obtenir un modèle final et des résidus sans structure apparente (cf. ANDREWS (1978)). Pour des exemples d'une telle démarche s'appuyant sur l'A.F.C. on pourra se reporter à HEIJDEN et WORSLEY (1988) et WORSLEY (1987).

Toutefois en A.F.C. classique cette procédure n'est pas répétée et s'achève en une itération par la donnée d'un modèle (souvent oublié) et d'une description de la partie significative des résidus. Il est clair que lorsque le modèle s'ajuste bien aux données, la représentation graphique des résidus est alors sans intérêt. Dans la pratique, cette situation est rarement rencontrée car le modèle retenu est trop frustre ou l'échantillon observé trop grand. Les résidus contiennent alors une information qu'il convient d'exploiter pour suppléer aux « insuffisances » du modèle.

1.2. Formules de l'A.F.C.

Nous situant dans l'optique ci-dessus, nous rappelons ci-après sans les démontrer, les formules sous-tendant les représentations graphiques de l'A.F.C. Pour plus de détails le lecteur pourra se reporter par exemple à l'article d'ESCOUFIER (1985) ou encore aux ouvrages de GIFI (1981) ou de GREENACRE (1984).

Il apparaît que ces formules peuvent être obtenues de deux manières qui sont trivialement équivalentes mais donnent lieu à des variantes différentes (voir § 2.2 ci-dessous).

Soit Y la matrice des données initiales de terme général y_{ij}; soient y_{i+}, y_{+j} et y_{++} les totaux lignes, colonnes et général. Soient R et C les matrices diagonales des fréquences marginales des lignes et des colonnes. Enfin soient A et B les matrices donnant les coordonnées des profils lignes et colonnes dans les représentations graphiques de l'A.F.C.

La première approche consiste à procéder à la décomposition en valeurs singulières généralisée du triplet $(R^{-1} X C^{-1}, C, R)$ où X est la matrice des résidus du modèle d'indépendance :

$$x_{ij} = y_{ij} - \frac{y_{i+}\, y_{+j}}{y_{++}}$$

Dans la seconde, on considère le triplet ($R^{-1/2} Z C^{-1/2}$, C, R) où Z est la matrice des résidus standardisés du modèle d'indépendance :

$$z_{ij} = \frac{y_{ij} - \dfrac{y_{i+}\, y_{+j}}{y_{++}}}{\sqrt{\dfrac{y_{i+}\, y_{+j}}{y_{++}}}}$$

On peut montrer assez facilement que, d'un point de vue technique, ces deux approches reviennent à effectuer d'abord la décomposition en valeurs singulières de Z, c'est-à-dire à écrire $Z = U \Sigma V'$ où Σ est la matrice diagonale des valeurs singulières de Z. On en déduit alors $A = R^{-1/2} V \Sigma^{\alpha}$, $B = C^{-1/2} V \Sigma^{\beta}$ et des formules de reconstitution des données. Les paramètres α et β dépendent de certaines pratiques, les valeurs 0, 1/2 ou 1 étant usuelles (cf. GOWER et DIGBY (1981)). CAUSSINUS (1986) et LEEUW et HEIJDEN (1988) discutent par exemple les enjeux du choix de ces paramètres.

Il est à noter que si l'on ne procède, ci-dessus, qu'à la décomposition en valeurs singulières de Z (et non à la décomposition en valeurs singulières généralisée des triplets (M, C, R) avec $M = R^{-1} X C^{-1} = R^{-1/2} Z C^{-1/2}$), on a $A = U \Sigma^{\alpha}$ et $B = V \Sigma^{\beta}$. Les représentations graphiques associées sont alors des cas particuliers de « biplot » de GABRIEL (1971). Elles diffèrent de celles fournies par l'A.F.C. ($A = R^{-1/2} U \Sigma^{\alpha}$ et $B = C^{-1/2} V \Sigma^{\beta}$) : l'introduction de la métrique du chi-deux et des pondérations se traduit par un « redressement » des lignes et des colonnes.

2. Les variantes de l'A.F.C.

L'idée d'un modèle complété par des traits significatifs extraits des résidus de ce modèle sous-tend de nombreuses variantes de l'A.F.C., par exemple, CAUSSINUS et FALGUEROLLES (1986), DOMENGES et VOLLE (1979), ESCOFIER (1984), HEIJDEN (1985, 1987), HEIJDEN et LEEUW (1985). Ceci conduit à évoquer brièvement le rôle des modèles en analyse des données et à présenter les formules utilisées pour représenter graphiquement les interactions lignes-colonnes dans les résidus.

2.1. Le rôle des modèles

Si le modèle est implicitement choisi en A.F.C. standard il n'en est pas de même dans les variantes de l'A.F.C. où il joue un rôle central. Schématiquement son choix nous semble alors être guidé par deux types de considérations.

Le modèle peut être introduit pour décrire les grands traits des données; par suite, il doit être adapté à la problématique des données soumises à l'analyse. En ce sens on retrouve la notion de modèle virtuel introduite par BARRA (1985) ou de « leading case » introduite par MALLOWS et TUKEY (1982). Ce point de vue est discuté par CAUSSINUS (1985). L'étude des résidus par des méthodes facto-

rielles vient alors compléter, le cas échéant, ce modèle (cf. par exemple CAUSSINUS et FALGUEROLLES (1986)).

Le modèle peut être introduit pour supprimer certains traits spécifiques ou connus des données. Autrement dit le modèle est utilisé comme un filtre. Les données filtrées sont alors l'objet principal de l'étude (cf. par exemple DOMENGES et VOLLE (1979), ESCOFIER (1984), HEIJDEN (1985, 1987), HEIJDEN et LEEUW (1985), LEEUW et HEIJDEN (1988), QANNARI (1983)).

Il est clair que les modèles log-linéaires (cf. McCULLAGH et NELDER (1983)) fournissent un réservoir important de modèles flexibles et aisément adaptables aux situations décrites ci-dessus.

2.2. Les représentations graphiques

Comme en AFC, il s'agit ici de fournir des représentations graphiques des lignes i et des colonnes j de façon à rendre compte des écarts entre les données initiales y_{ij} et celles du modèle $\hat{y}_{ij}$. Un exemple de cette démarche est donné par CAUSSINUS et FALGUEROLLES (1987) dans le même numéro de cette revue.

De façon générale, les représentations graphiques des résidus procèdent des deux approches qui ont été très artificiellement distinguées au paragraphe 1.2.

Soient donc R et C deux matrices diagonales. R et C pourront être prises simplement égales à des matrices identités ou construites à partir des effectifs marginaux lignes et colonnes de la table de contingence considérée.

Variante 1

On procède à la décomposition en valeurs singulières du triplet $(R^{-1} X C^{-1}, C, R)$ avec $x_{ij} = y_{ij} - \hat{y}_{ij}$

Variante 2

On considère, de même, le triplet $(R^{-1/2} Z C^{-1/2}, C, R)$ où

$$z_{ij} = \frac{y_{ij} - \hat{y}_{ij}}{\sqrt{\hat{y}_{ij}}}.$$

Ces deux variantes déterminent en général des coordonnées différentes pour les lignes (resp. colonnes) et produisent donc des représentations graphiques différentes. La seconde a pour principal avantage son lien immédiat avec la statistique du chi-deux de Pearson. Toutefois la première permet, dans certaines situations et pour des choix convenables de R et C, d'exploiter des propriétés spécifiques des résidus et de donner des interprétations en terme de distance du chi-deux entre lignes et colonnes.

Remarquons enfin que, pour certains modèles (cf. HABERMAN (1973)), il existe des formes alternatives de résidus asymptotiquement équivalentes aux résidus standardisés. Par exemple :

$$2\left[\sqrt{y_{ij}} - \sqrt{\hat{y}_{ij}}\right] \text{ ou } \left[\sqrt{y_{ij}} + \sqrt{y_{ij} + 1} - \sqrt{4\hat{y}_{ij} + 1}\right]$$

La première de ces formules permet de retrouver l'analyse factorielle sphérique introduite par DOMENGES et VOLLE (1979). Pour une interprétation géométrique de la trace de l'opérateur associé et une étude de son lien avec la statistique du chi-deux, on se reportera à BHATTACHARYYA (1946).

Remerciements

Ce travail a été réalisé notamment grâce au concours actif du Centre National pour la Recherche Scientifique (CNRS) pour la France, et de l'Organisation Néerlandaise pour le Développement de la Recherche Scientifique (ZWO) pour la Hollande; les auteurs du présent article leur en sont très reconnaissants.

Bibliographie

F. ANDREWS (1978). — *Data analysis, exploratory*, in International Encyclopedia of Statistics, Ed. W.H. Kruskal and J.M. Tanur, Collier Macmillan Publishers, New-York, 97-107.

J.R. BARRA (1985). — *Methodes statistiques en psychiatrie — modèles virtuels* in Model Choice, Proceedings of the 4th Franco-Belgian meeting of statistician (1983), Publication des Facultés Universitaires Saint-Louis, Bruxelles, Belgique.

A. BHATTACHARYYA (1946). — *On a measure of divergence between two multinomial populations*. Sankya, 7, 401-406.

J.P. BENZECRI (1973). — *L'analyse des données. Tome 2 : L'analyse des correspondances*. Dunod.

H. CAUSSINUS (1985). — Quelques réflexions sur la part des modèles probabilites en Analyse des Données. 4^{e} Journées Internat. An. Donn. et Inform., Versailles. Paru dans *Data Analysis and Informatics 4*, North-Holland (ed. by E. Diday), 151-165.

H. CAUSSINUS (1986). — Models and uses of principal component analysis. *In Multidimensional Data Analysis*, Ed. by J. de Leeuw *et al.*, DSWO Press, Leiden, 149-170.

H. CAUSSINUS et A. de FALGUEROLLES (1986). — *Modèle de quasi-symétrie et analyse descriptive de tableaux carrés*. In Comparaison et Evaluation des approches française et britannique de l'analyse de données complexes. Publication 02-86 du Laboratoire de Statistique et Probabilités, Université Paul Sabatier, 79-95.

H. CAUSSINUS et A. de FALGUEROLLES (1987). — Tableaux carrés : modélisation et méthodes factorielles. *Revue de Statistique Appliquée*. Vol. 35, n° 3, 35-52.

D. DOMENGES et M. VOLLE (1979). — Analyse factorielle sphérique : une exploration. *Annales de l'INSEE*, n° 35, 3-84.

B. ESCOFIER (1984). — Analyse factorielle en référence à un modèle : application à l'analyse des tableaux d'achange. *Revue de Statistique Appliquée*, Vol. 32, n° 4, 25-36.

Y. ESCOUFIER (1985). — L'analyse des correspondances : ses propriétés et ses extensions ISI, *Actes de la 45^{e} Section, Livraison 4*, Tome LI.

K.R. GABRIEL (1971). — The biplot-graphic display of matrices with application to principal component analysis. *Biometrika, 58*, 453-467.

A. GIFI (1981). — Non linear multivariate analysis. *Department of Data Theory*. University of Leiden.

C. GOWER and P.G.N. DIGBY (1981). — Expressing complex relationships in two dimensions. In *Interpreting Multivariate Data*, Ed. by V. Barnet, John Wiley and Sons, 83-118.

J. GREENACRE (1984). — Theory and applications of correspondence analysis. *Academic Press.*

S.J. HABERMAN (1973). — The analysis of residuals in cross-classified tables. *Biometrika 29*, 205-220.

P.G.M. van der HEIJDEN (1985). — Transition matrices, model fitting and correspondence analysis. 4[e] Journées Internat. An. Donn. et Inform., Versailles. Paru dans *Data Analysis and Informatics 4*, North-Holland (ed. by E. Diday), 221-226.

P.G.M. van der HEIJDEN (1987). — Correspondence analysis of longitudinal categorical data. DSWO Press, Leiden.

P.G.M. van der HEIJDEN et J. de LEEUW (1985). — Correspondence analysis used complementary to loglinear analysis, *Psychometrika, 50*, 429-447.

P.G.M. van der HEIJDEN et K. WORSLEY (1988). — Comment on correspondence analysis used complementary to loglinear analysis. A paraître dans *Psychometrika.*

J. de LEEUW et P.G.M. van der HEIJDEN (1988). — Correspondence analysis of incomplete tables. A paraître dans *Psychometrika.*

M. McCULLAGH et J.A. NELDER (1983). — *Generalised linear models.* Chapman and Hall. London.

C.L. MALLOWS et J.W. TUKEY (1982). — An overview of technics of data analysis emphasizing its exploratory aspects. In Some Recent Advances in Statistics. Ed. by J. Tiago de Oliveira *et al., Academic Press 111-172.*

E.M. QANNARI (1983). — Analyses factorielles de mesures, applications. *Thèse de 3[e] cycle*, Université Paul Sabatier, Toulouse.

K. WORSLEY (1987). — Un exemple de l'identification d'un modèle log-linéaire grâce à une analyse des correspondances. *Revue de Statistique Appliquée*, Vol. 35, n° 3, 13-20.

DISCUSSION

Tomàs ALUJA et Manuel MARTI

Departement de Statistique et Recherche Opérationnelle
Universitat Politècnica de Catalunya
c/Pau Gargallo 5, 08028 Barcelone, Espagne

Les 5 articles précédents représentent de remarquables contributions à l'élaboration d'une stratégie complète d'analyse des données complexes, par emploi itératif des approximations exploratoire et modélisation.

Comme l'a remarqué L. LEBART (LEBART 1985), le statisticien se trouve confronté à des situations qui peuvent être décrites selon les trois caractéristiques suivantes : données structurées par opposition à données amorphes, données univariées par opposition à données multivariées et utilisation d'une approche exploratoire par opposition à une approche confirmatoire. Habituellement les situations ne seront pas pures, comme celles qui se trouveraient aux sommets d'un cube défini par les trois caractéristiques précédentes. Dans la pratique, les données ne sont pas complètement amorphes, car on possède une certaine connaissance sur les données (soit qu'on puisse les diviser en sous-groupes homogènes, soit qu'on puisse établir une liaison temporelle ou spatiale, soit qu'on connaisse l'existence de facteurs sous-jacents qu'on voudrait confirmer, etc.). D'autre part, dans la plupart des cas l'approche n'est pas totalement exploratoire ni totalement confirmatoire, elle conjugue des éléments des 2 types d'approches, et enfin, par contre, la réalité est presque toujours multivariée.

Dans ces situations, le problème du statisticien consiste à trouver la stratégie optimale d'utilisation des méthodes (COPPI 1986) de façon à mettre à profit l'information a priori, ou acquise en cours d'analyse, pour essayer d'en savoir plus. Sur ce point notre expérience nous montre que l'approche exploratoire et la modélisation sont deux phases d'une même analyse, et correspondent à des étapes différentes de la recherche; d'abord on veut voir quelles données on a, les représenter, détecter les anomalies, et suggérer des structures possibles permanentes. Après, on désire confirmer ces hypothèses, expliquer les « patterns » détectés. Cette stratégie implique l'utilisation complémentaire des deux approches, en particulier l'utilisation des techniques de représentation d'un corpus multidimensionnel et de la réduction de la dimensionnalité comme guide et simplification pour la phase de modélisation.

Cette démarche est appliquée dans les articles qui présentent des études de cas. WORSLEY montre sur un bon exemple comment l'analyse exploratoire peut guider la formulation du modèle; il faut remarquer, cependant, le fait que, malgré que le tableau concerne trois variables, il est analysé comme un tableau de

contingence simple. Dans ce cas sont connues les analogies entre les représentations graphiques et les termes d'association du modèle log-linéaire (LAURO 1982, GOODMAN 1986). Par contre, les représentations de l'ACM, peuvent induire des associations fausses, par exemple, lorsque il existe deux associations simples entre trois variables, les représentations sur les premiers plans factoriels montreraient une association fausse entre les variables non associées (ALUJA et MARTI 1986); cela découle du fait que l'ACM analyse les tableaux marginaux de l'hyper-cube de contingence, donc, il analyse les relations entre couples de variables indépendamment du reste des variables du tableau; cette remarque est aussi valable pour les articles de BACCINI, MATHIEU et MONDOT et AITKIN, FRANCIS et RAYNAL, et explique le fait que les résultats de l'ACM et du modèle log-linéaire souvent ne coïncident pas.

Les articles de FALGUEROLLES — HEIJDEN et CAUSSINUS — FALGUEROLLES proposent un pas en avant quant à la stratégie d'utilisation des méthodes exploratoires pour l'analyse des résidus d'un modèle. CAUSSINUS — FALGUEROLLES, présentent dans leur article deux cas de mise en œuvre de cette stratégie pour l'analyse de deux tableaux carrés pour lesquels l'ACP s'avère peu satisfaisante (ce type de tableaux de données est assez fréquent, comme est le cas des matrices de flux réciproques : mobilité, échanges économiques, etc.). La modélisation proposée permet d'identifier un effet ligne, un effet colonne, un effet de similarité mutuelle entre ligne et colonne, et le résidu, qui représente la partie asymétrique de la proximité entre ligne et colonne; de plus, cette stratégie permet la représentation de la structure de ces deux matrices (similarité mutuelle et résidus). Cette stratégie est une démarche puissante et flexible pour l'analyse de ce qu'il reste dans les résidus, après avoir éliminé les effets spécifiés dans le modèle; c'est donc une façon de prendre en compte la structure connue des données dans l'analyse et cela repose sur le même principe que l'analyse des corrélations partielles.

Dans le même esprit, nous avons développé les analyses, locales et partielles (ALUJA et LEBART 1985) lorsque il existe une relation entre les individus représentable à l'aide d'un graphe; ces analyses permettent aussi de dégager d'une façon plus précise les véritables associations existantes dans une table de contingence multiple. D'autre part, l'emboitement des démarches esquissées (méthodes exploratoires préalables à la formulation du modèle et méthodes exploratoires pour l'analyse des résidus d'un modèle), signifie l'inclusion de ces méthodes dans la méthodologie statistique.

Références complémentaires

T. ALUJA et M. MARTI (1986). — « Complementarity between log-linear models and correspondence analysis », *II Catalan Symposium on Statistics*, Barcelona, Vol 1 (Invited lectures), 113-140.

T. ALUJA ET L. LEBART (1985). — « Factorial analysis upon a graph », *DLP Bullet. Tech. du CESIA*, Vol. 3, 5-19.

COPPI (1986). — « Comparaison de méthodes et comparaison de stratégies d'analyse : quelques réflexions sur l'analyse d'un gros fichier de données sociologiques » *Publications du Lab. de Stat. et Prob. n° 02-86*, pp. 27-35. Université Paul Sabatier. Toulouse.

L. LEBART (1985). — « Quelques progrès récents dans la pratique de l'analyse des données », Invited lecture at the International Meeting of Statistics in the Basque Country, *IMSIBAC III*. Bilbao.

N.C. LAURO et A. DECARLI (1982). — « Correspondence analysis and log-linear models in multiway contingency tables study. Some remarks on experimental data » *Metron n° 1-2*, pp. 213-234.

J.J. DAUDIN

Département de Mathématique et Informatique
INAPG, 16 rue Claude Bernard 75005, Paris

Les travaux effectués par les équipes du Laboratoire de statistique et de probabilités de l'Université Paul Sabatier et le centre de statistique appliquée de l'Université de Lancaster forment une contribution importante au développement de la statistique sous plusieurs aspects : non seulement ils permettent de mieux saisir les différences et les complémentarités existant entre les approches françaises et britanniques, mais aussi ils ouvrent la voie à une démarche statistique originale qui associe et coordonne modèles et analyses factorielle et/ou typologique. Si l'idée d'utiliser conjointement des outils d'analyse des données et des modèles statistiques est présente depuis quelques années, comme par exemple l'analyse des interactions d'une analyse de variance à l'aide d'une classification automatique, ou encore la conception de l'analyse factorielle des correspondances comme une analyse des résidus par rapport au modèle d'indépendance, l'exploitation systématique et raisonnée de cette idée n'avait par contre jamais été développée aussi complètement.

On pourrait opposer les approches britannique et française par la notion de modèle statistique, la première en faisant un usage intensif alors que la seconde la rejetterait. Cette opposition est trop rapide et artificielle.

D'une part, on peut présenter beaucoup de méthodes d'analyse factorielle sous la forme d'un modèle statistique (CAUSSINUS (1985)), d'autre part la notion de modèle statistique recouvre des pratiques très différentes : on peut utiliser un modèle statistique pour confirmer ou réfuter une théorie concernant la discipline scientifique pour laquelle ont été récoltées les données, mais aussi pour prédire des variables non mesurées ou encore pour décrire des données sur lesquelles on ne possède pas de théorie globale cohérente.

Lorsqu'il s'agit de données complexes il est rare que l'on dispose d'une théorie globale qui puisse être formalisée sous la forme d'un modèle statistique. On se trouve donc généralement dans la situation où ce dernier est un moyen de décrire les données de façon judicieuse, c'est-à-dire en extrayant l'information significative du bruit dû aux erreurs de mesure ou à la variabilité d'échantillonnage. Mais c'est précisément ce même but qui est poursuivi, par d'autres voies, en analyse factorielle. Dans un cas comme dans l'autre on veut décomposer les données en 2 composantes sous la forme

$$\text{Données} = \text{« effets »} + \text{bruit blanc}$$

Les deux approches sont donc concurrentes, au sens où elles cherchent toutes deux à réaliser un même objectif par des voies différentes. La différence essentielle entre les deux approches est la forme mathématique que prend la première composante : axes factoriels dans l'approche française et espace paramétrique du type du modèle linéaire dans l'approche britannique. Ce point est développé dans (DAUDIN, TRÉCOURT, TOMASSONE (1984)).

La qualité d'une méthode statistique descriptive se mesure à son aptitude à décomposer correctement les deux composantes et à la qualité de la description de la première composante qui est généralement complexe. H. CAUSSINUS et A. De FALGUEROLLES montrent clairement que l'on peut utiliser les deux types de description *conjointement sur deux éléments séparés de la première composante.*

Les mêmes auteurs développent en introduction la conception de modèle statistique dans un contexte d'analyse exploratoire. Ce point est important sur le plan méthodologique car les termes de modèle et de résidu n'ont pas, dans ce contexte, le sens qui leur est accordé classiquement. En particulier le terme de résidu a ici une acception large : il contient à la fois le résidu au sens classique et des éléments de la première composante non pris en compte par le modèle. L'intérêt de la démarche est que si les aspects les plus forts et les plus simples de la première composante des données peuvent être bien pris en compte par un petit nombre de paramètres, les aspects plus diffus et plus compliqués sont mieux synthétisés par une analyse factorielle.

L'ensemble des travaux suscite un grand nombre de questions, mais je me limite à deux commentaires supplémentaires :

La présentation des résultats se fait plutôt sous la forme de chiffres (estimation des paramètres du modèle) dans l'approche anglaise alors que l'approche française utilise des représentations graphiques issues des axes des analyses factorielles. Dans les deux cas, il s'agit de décrire, résumer, synthétiser des données complexes, mais la forme donnée à la description est différente.

Cependant, il est souvent possible d'utiliser les estimations des paramètres pour construire des graphiques comme le font CAUSSINUS et De FALGUEROLLES (voir aussi AITKIN (1984) ou DAUDIN et TRÉCOURT (1980) où une utilisation entièrement descriptive du modèle loglinéaire est faite); il semble qu'une représentation graphique soit une bonne façon de transmettre les résultats d'une analyse de données complexes, dans la mesure où les paramètres sont alors très nombreux, et où les chiffres bruts n'étant pas réutilisés dans un modèle prédictif, il n'est pas nécessaire de retenir précisément la valeur de chacun d'eux, mais plutôt les ordres de grandeur et les relations d'ordre, informations qui apparaissent bien dans un graphique.

Ma deuxième remarque concerne l'utilisation du modèle loglinéaire; AITKIN, FRANCIS et RAYNAL n'ont pas pu l'utiliser sur leurs données « vues les grandes dimensions des tables de contingence et leur taux de cellules vides ». En réalité, certains logiciels mieux adaptés à ce modèle que GLIM permettent d'utiliser ce modèle sur des tables d'assez grandes dimensions dans la mesure où il n'est pas indispensable de conserver toute la table de contingence mais seulement certaines tables marginales. Pour la même raison les cellules vides ne

créent pas trop de problèmes si on se limite à des modèles raisonnables. Par contre un des problèmes de ce modèle est l'exploitation des résultats que l'on peut en faire : dans ce type de situation on obtient un grand nombre de paramètres décrivant la première composante des données et il est difficile de synthétiser cette information. On rejoint ici le problème abordé précédemment.

Références

M. AITKIN (1984). — Modélisation mathématique de l'enquête communautaire sur les forces de travail. In Développements récents dans l'analyse de grands ensembles de données, *Information de l'Eurostat, numéro spécial,* Luxembourg.

H. CAUSSINUS (1985). — Quelques réflexions sur la part des modèles probabilistes en Analyse des données. *Quatrièmes journées Internationales Analyse des données et Informatique.* INRIA. Edition provisoire.

J.J. DAUDIN et P. TRÉCOURT (1980). — Analyse factorielle des correspondances et modèle loglinéaire, comparaison des deux méthodes sur un exemple. *Revue de Statistique Appliquée,* XXVIII, 1, 5-24.

J.J. DAUDIN, R. TOMASSONE et P. TRÉCOURT (1984). — Analyse d'enquêtes à grande échelle. In Développements récents dans l'analyse de grands ensembles de données, *Information de l'Eurostat, numéro spécial,* Luxembourg.

J. De LEEUW

Department of Data Theory
Faculty of Social Sciences, University of Leiden
4 Middelstegracht, 2312 TW Leiden, Pays Bas

The papers in this issue approach the relationship between data analysis and modeling, or between correspondence analysis (CA) and log linear analysis (LLA), in at least three different ways. In the papers by WORSLEY and in that by BACCINI, MATHIEU, and MONDOT CA is used to prepare the data for model fitting, either by straightforward data reduction or by using CA results to suggest an appropriate model. In the paper by AITKIN, FRANCIS, and RAYNAL CA and modeling, in this case latent class modeling, are treated as equals and the results of the two techniques are compared. We shall concentrate, in these remarks, on the third approach to the relationship between the two classes of methods. CAUSSINUS and De FALGUEROLLES and also De FALGUEROLLES and VAN DER HEIJDEN take modeling as their starting point, and use CA to decompose the residuals that are left after a LLA is carried out. They do this in various interesting special cases. It is the purpose of these remarks to indicate what the general idea behind this "complementary approach" is, and in how far it can be generalized to other models.

Let us suppose that the data are a random sample of size n from a discrete distribution taking m possible values. They can be displayed in the form of random frequencies $\underline{n}_j$, with $j = 1, \ldots, m$. Suppose $\mathbf{E}(\underline{n}_j) = n\pi_j$, and suppose we have a

model which says that $\pi \varepsilon \Omega$, with Ω a p-dimensional differentiable manifold in S^{m-1}, the unit simplex in R^m. The log-likelihood of an observed vector of frequencies $\{n_j\}$ is $L = n \sum p_j \ln \pi_j$, where $p_j = n_j/n$. Maximum likelihood estimates are found by maximizing L over all $\pi \varepsilon \Omega$. This means that at the maximum likelihood estimate $\mathbf{p}$ we have that δ, with elements $\delta_i = p_i/\mathbf{p}_i$, is orthogonal to the tangent space $T_\Omega(\mathbf{p})$ of Ω at $\mathbf{p}$. Suppose $\{u_0, \ldots, u_{m-p-1}\}$ is a basis for the orthogonal complement in R^m of $T_\Omega(\mathbf{p})$. Without loss of generality we can assume that u_0 has all elements equal to a constant, and that the basis is unit orthogonal in the metric defined by the diagonal matrix $\mathbf{P}$. Thus $u_s' \mathbf{P} u_t = \delta^{st}$. This means that we can write

$$p = \mathbf{P}\{1 + a_1 u_1 + \cdots + a_{m-p-1} u_{m-p-1}\}. \tag{1}$$

Also

$$\mathbf{P}^{-1/2}(p - \mathbf{p}) = a_1 \mathbf{P}^{+1/2} u_1 + \cdots + a_{m-p-1} \mathbf{P}^{+1/2} u_{m-p-1}, \tag{2}$$

and consequently

$$(p - \mathbf{p})' \mathbf{P}^{-1} (p - \mathbf{p}) = (a_1)^2 + \cdots + (a_{m-p-1})^2. \tag{3}$$

These simple geometrical facts are actually the basis of the proof that the Pearson goodness of fit statistic has a ξ^2 distribution with $m - p - 1$ degrees of freedom. How are these results related to CA ? First observe that (2) defines a decomposition of the normalized residuals in terms of the orthonormal vectors $v_s = \mathbf{P}^{-1/2} u_s$, just as (3) decomposes the chi square. But in the derivation of (2) and (3) there is still some freedom, because obviously the basis $\{u_s\}$ can be chosen in many different ways. In the independence model for an $R \times C$ table the likelihood equations are of the form $\sum_r (p_{rc} - \mathbf{p}_{rc}) = 0$ and $\sum_c (p_{rc} - \mathbf{p}_{rc}) = 0$. Thus rows and columns of the residual add up to zero. Using the singular value decomposition of the residuals gives

$$p_{rc} = \mathbf{p}_{rc} + \sum_s \lambda_s x_s y_s,$$

which can be rewritten as

$$p_{rc} = p_{r+} p_{+c} \left\{1 + \sum_s \lambda_s \underline{x}_s \underline{y}_s\right\}, \tag{5}$$

with $\underline{x}_s$ and $\underline{y}_s$ suitable scaled versions of the orthonormal singular vectors x_s and y_s. This is exactly of the form (1). Thus we can decompose the residuals as in (4), and we can rescale the decomposition as in (5), which gives us a decomposition as in (1). This depends on the availability of the singular value decomposition, i.e. of a simple canonical form for two-way matrices, and on the particular form of the independence model.

Let us see what a similar analysis gives for GOODMAN'S RC-model. The likelihood equations are $\Sigma \mathbf{x}_r (p_{rc} - \mathbf{p}_{rc}) = 0$ and $\Sigma \mathbf{y}_c (p_{rc} - \mathbf{p}_{rc}) = 0$, where $\mathbf{x}_r$ and $\mathbf{y}_c$ are the maximum likelihood estimates of the scores. If follows that if we choose x_r and y_c orthogonal to these scores, then again the products $x_r y_c$ decompose the residuals of the RC-model. In fact we can choose x_r and y_c by computing the singular value decomposition of $p_{rc} - \mathbf{p}_{rc}$, because $\mathbf{x}_r$ and $\mathbf{y}_c$ are indeed singular vectors of this matrix, corresponding with singular value zero. Thus we can use the singular value decomposition of the residuals, which has rank not larger than min $(R, C) - 2$. Thus (4) is generalized very easily, but (5) becomes

$$p_{rc} = a_r b_c \left\{ \exp(\mathbf{x}_r \mathbf{y}_c) + \sum_s \lambda_s \underline{x}_s \underline{y}_s \right\}, \quad (6)$$

which is not of form (1). The products $\underline{x}_s \underline{y}_s$ are not orthogonal in the metric $\mathbf{p}_{rc}$, and thus the connection with chi square is not maintained.

For LLA, discussed in this issue by various authors, the likelihood equations are of the form $\sum g_{rcs}(p_{rc} - \mathbf{p}_{rc}) = 0$, where the G_s are known matrices. In the quasi-symmetry model, for instance, R = C, and there are R(R − 1)/2 elementary symmetric matrices G_s (one upper diagonal element and the corresponding lower diagonal element + 1). There are an additional R + C matrices G_s which take care of the marginals, in the sense that they have one row or one column filled with + 1 while all other elements are zero. It follows that the residuals are antisymmetric, in the sense that elements above and below the diagonal add to zero. Moreover the diagonal is filled in such a way that rows and columns add to zero. This already indicates one decomposition, but it is not a very satisfactory one in terms of separate scores for rows and columns. The more satisfactory decomposition, in this respect, is to make

$$p_{rc} = \mathbf{p}_{rc} + a_{rc}\delta^{rc} + b_{rc}, \quad (7)$$

where $b_{rc} + b_{cr} = 0$ for all c, r, and thus $b_{rr} = 0$ for all r. Residuals are decomposed in a diagonal matrix and an antisymmetric matrix. The antisymmetric part can then be decomposed by using the familiar Gower-decomposition of antisymmetric matrices. Again this generalizes the idea to use the singular value decomposition to study residuals, but it again does not preserve the close connection with chi-square of the independence model.

It seems that the complementary approach to modeling works especially nicely with the independence model, because of the availability of a simple canonical form, and because of the simple product form of the model which matches this canonical form. There are no other examples in which the complementary model works out so elegantly, except perhaps the quasi-independence model mentioned briefly by DE FALGUEROLLES and VAN DER HEIJDEN. In other cases the chi square geometry of the maximum likelihood method, and the unweighted Euclidean geometry of the singular value decomposition, cannot be matched.

Y. ESCOUFIER

Unité de Biométrie, 9, Place Pierre Viala, 34060 Montpellier Cedex

Disons d'entrée que l'entreprise que représente cette publication me paraît intéressante. Je défends trop souvent l'idée qu'il n'y a pas de recherche en statistique sans traitement effectif de données pour ne pas me réjouir des travaux qui me sont soumis. Bien sûr, ce type de préoccupation donne à certains textes une fome qui est plus celle d'un compte-rendu d'activités que celle d'un article scientifique traditionnel. Il me semble bon d'accepter au moins de temps en temps une telle présentation de nos connaissances. Elle aide ceux qui ne peuvent accéder aux méthodes par la compréhension mathématique de leurs fondements à y accéder par la reproduction des démarches décrites. N'est-ce pas là une approche pédagogique reconnue ?

Pour avoir une vue générale des textes proposés à ma lecture, je prendrai pour observatoire une situation que je connais bien, l'Analyse des Correspondances. Parmi de nombreuses présentations possibles de cette méthode je choisis celle qui la voit comme une approximation au sens des moindres carrés des écarts des fréquences observées aux fréquences attendues sous l'hypothèse d'indépendance. Il y a donc trois choix. On parle de fréquences et non de logarithmes ou de racines carrées des fréquences; on se situe par rapport aux fréquences attendues sous l'hypothèse d'indépendance et non par rapport à un modèle quelconque; on utilise le critère des moindres carrés alors que d'autres parlent de déviance, de moindres valeurs absolues ou de critère minimax. Tous ces choix sont discutables et des choix alternatifs ont déjà été expérimentés. Les travaux de VOLLE ou ESCOFIER cités dans les textes en sont des exemples. Le travail de C. LAURO sur l'Analyse non symétrique des Correspondances relève de la même approche. Le modèle Log-linéaire est un exemple de choix différents : dans un contexte probabiliste, on choisit de travailler avec les logarithmes des fréquences pour le critère du maximum de vraisemblance.

Deux points me paraissent importants dans cette remise en cause des choix initiaux. Le premier concerne la cohérence qui parait nécessaire entre les éléments qui concourent à construire la démarche. Prenons un exemple : Ajuster des P_{ij} par des

$$\hat{P}_{ij} = \alpha_i \beta_j \exp\left(\sum_{\alpha=1}^{k} \sqrt{\lambda_\alpha}\, \Psi_{\alpha i}\, \phi_{\alpha j}\right)$$

au sens du maximum de vraisemblance puis étudier les écarts $P_{ij} - \hat{P}_{ij}$ par une décomposition en valeurs singulières c'est-à-dire au sens des moindres carrés me paraît curieux. Pourquoi ne pas conserver les paramètres trouvés et ajuster des

$$\hat{\hat{P}}_{ij} = \alpha_i \beta_j \exp\left(\sum_{\alpha=1}^{k} \sqrt{\lambda_\alpha}\, \Psi_{\alpha i}\, \phi_{\alpha j} + \sum_{\alpha=k+1}^{v} \sqrt{\lambda_\alpha}\, \Psi_{\alpha i}\, \phi_{\alpha j}\right)$$

au sens du maximum de vraisemblance ?

Le second point important est, me semble-t-il, de ne pas restreindre les approches concevables à celles rendues possibles aujourd'hui par des programmes disponibles. Prenons ici aussi un exemple directement tiré de l'article de K.J. WORSLEY : par ses formules (1) et (2) l'auteur rappelle des liens qui existent entre les solutions de l'Analyse des Correspondances et celles du modèle log-multiplicatif. Analysant les résultats de l'Analyse des Correspondances, il en déduit que des paramètres qu'il note u_{1i} et u_{2i} pourraient dans un but de simplicité être contraints à ne prendre que certaines valeurs. Il introduit alors ces contraintes dans le modèle log-linéaire, c'est-à-dire dans une approximation fondée sur le critère du maximum de vraisemblance. Pourquoi ne pas le faire directement dans l'approximation au sens des moindres carrés ? Je crains que la réponse soit simplement l'absence actuelle de programme. C'est une lacune des méthodes factorielles telles qu'elles sont largement pratiquées. Nous nous sommes endormis dans le confort des solutions en vecteurs propres et valeurs propres des problèmes aux moindres carrés sans contraintes.

Je terminerai par une remarque de vocabulaire concernant l'article de M. AITKIN *et al.* Pour moi, l'algorithme décrit en 4.2. est un algorithme de nuées dynamiques; ce que les auteurs appellent Etape E et Etape M correspond aux fonctions d'affectation et aux fonctions de représentations du livre de E. DIDAY

et al. Ce que les auteurs appellent Nuées Dynamiques n'est que le cas très particulier dit des moyennes mobiles. Ceci dit je m'interroge sur la signification des résultats de la dernière table du paragraphe 5. Comment un même profil peut-il donner des individus dans des classes différentes ?

Références

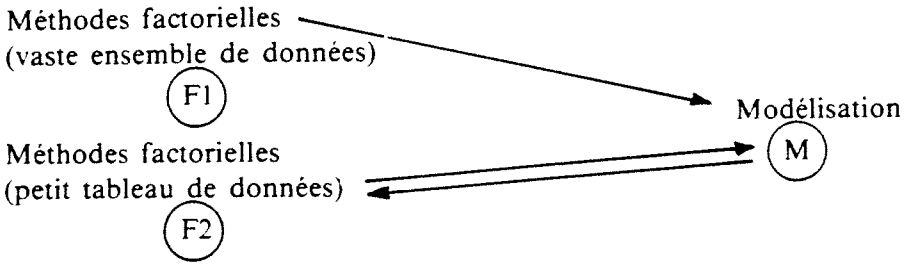
N. LAURO, L. D'AMBRA (1983). — L'analyse non symétrique des correspondances. 3e Journées Internationales d'Analyse de Données et Informatique, Versailles. Paru dans *Data Analysis and Informatics 3*, North-Holland, Amsterdam, 433-446.

A. LECLERC

Unité 88, INSERM, 91 Boulevard de l'Hôpital, 75013 Paris

Les articles publiés dans ce numéro constituent un ensemble très riche et complet sur la confrontation entre approches de type modélisation, et méthodes factorielles. Ceci n'aurait pas été possible sans un travail en commun important, entre équipes de formation différente. Ceux qui ont eu, pendant longtemps, à regretter l'existence d'un fossé entre ces deux approches ne peuvent que se réjouir : d'abord il est satisfaisant pour l'esprit de savoir que de nombreuses passerelles existent entre des méthodes habituellement considérées comme éloignées; ensuite, du point de vue du praticien, une sorte de « code » d'un usage complémentaire des méthodes semble se dessiner. C'est ce point que je voudrais développer ici : que peut-on conseiller à un praticien, à la lecture de ces articles, sur une stratégie d'analyse combinant approches de type modélisation, et méthodes factorielles ?

Schématiquement, les différentes propositions rentrent dans le cadre suivant :



Les articles de M. AITKIN, de A. BACCINI, et de leurs collaborateurs, concluent à l'intérêt d'une stratégie F1 → M , pour l'analyse de vastes ensembles de données.

Les articles de A. FALGUEROLLES, P.G.M. VAN DER HEIJDEN, et H. CAUSSINUS suggèrent des stratégies de type M → F2 , et insistent sur le fait que la modélisation est implicite au départ de l'application d'une méthode factorielle telle que l'Analyse des Correspondances. Enfin, WORSLEY suggère la possibilité d'une itération entre méthodes factorielles et modélisation, en donnant un exemple de stratégie de type F2 → M .

Il n'y a pas contradiction entre ces propositions, car elles concernent des situations différentes :

Dans le premier cas, les données sont nombreuses, la situation-type est celle d'une enquête où sont éventuellement distinguées variables « à expliquer » et « explicatives »; les méthodes factorielles utilisées seraient principalement l'Analyse des Correspondances multiples ou la classification. La stratégie d'analyse est bien résumée par les auteurs : « Les méthodes d'Analyse des données, telles que l'AFCM ou la Classification automatique, sont les mieux adaptées à l'étude globale et exploratoire de gros fichiers de données » (M. AITKIN *et al.*) »; Une fois les données simplifiées, diverses méthodes de modélisation peuvent alors être envisagées (A. BACCINI *et al.*). Notre propre expérience nous avait amené aux mêmes conclusions (cf. A. LECLERC et alt.). Ce qui n'est peut-être pas assez souligné par les auteurs, c'est que dans beaucoup de cas l'étape de modélisation est également indispensable : un modèle logistique ou linéaire permet de quantifier les effets propres des variables explicatives (autrement dit, de tenir compte du fait que les variables explicatives ne sont pas indépendantes entre elles); or ceci est un objectif à atteindre en présence de variables à expliquer, dans beaucoup de domaines d'étude, comme la Santé ou l'Éducation. Une méthode comme l'Analyse des Correspondances ne fournit pas de résultats assez précis; c'est aussi l'idée avancée par WORSLEY, concernant l'intérêt d'« un modèle qui peut fournir des conclusions quantitatives complétant utilement les représentations graphiques de l'Analyse des Correspondances ».

Dans le second cas, où l'approche de type « modèle » est première, il s'agit de tableaux de contingence à plusieurs dimensions, dont la taille n'est pas excessive. L'Analyse des Correspondances, ou une méthode proche, peut être utilisée pour décrire les résidus d'un modèle : « La modélisation est utile pour tenir compte de la problématique précise du cas étudié. Les méthodes d'analyse multidimensionnelle peuvent ensuite intervenir » (H. CAUSSINUS). Ceci correspond à une situation où la modélisation est possible d'emblée (la taille des données le permet), mais pas intéressante : il faudrait choisir entre deux modèles, l'un « trop simple », l'autre « trop riche » (modèle saturé, par exemple). La séquence M → F2 permet de commencer par un modèle sous-dimensionné, de décrire les résidus, et éventuellement de revenir à un modèle plus complexe, comme le suggère WORSLEY.

En conclusion, l'ensemble de ces articles clarifie bien la question de la complémentarité de différentes approches, et propose des stratégies opérationnelles d'analyse. Il est frappant, par ailleurs, de constater à quel point les distinctions classiques entre approches « exploratoires » ou descriptives, et approches « confirmatoires » (test d'hypothèses...) sont devenues floues. La pratique du statisticien, liée aux possibilités que lui offre l'ordinateur, ne lui permet plus toujours de se situer clairement par rapport à ce qu'il a appris en statistique. Ceci pourrait être un sujet intéressant à aborder dans la Revue de Statistique Appliquée.

Références

A. LECLERC, A. CHEVALIER, D. LUCE et M. BLANC (1985). — Analyses des Correspondances et modèle logistique : possibilités et intérêt d'approches complémentaires. *Revue de statistique appliquée*. Vol. 33, n° 1, 25-40.

RÉPONSE DES AUTEURS

Lorsque les résultats d'un travail collectif sont présentés sous la forme de cinq articles distincts, il est difficile de faire ressortir une conclusion générale. Heureusement, les commentaires précédents pallient cette carence dans une très large mesure en apportant les divers ingrédients d'une telle conclusion : éléments de synthèse, précisions sur le degré d'achèvement (et, partant, d'inachèvement) du travail, pistes pour des réflexions nouvelles.

Un premier point que nous relèverons est que, même si chaque commentateur l'exprime avec ses nuances personnelles, un large consensus se dégage sur l'intérêt d'utiliser sans exclusive des techniques statistiques variées, dans une approche ouverte et flexible. La stratégie d'analyse peut varier selon les cas, et nous avons cherché à montrer comment sur des exemples de natures diverses; le schéma présenté par A. LECLERC est extrêmement utile pour synthétiser ces différentes possibilités.

Mais la discussion met aussi en lumière bien des limites de nos propositions, ce qui nous satisfait encore car l'un des objectifs était de sortir de divers conformismes et donc de soulever des interrogations. Nous allons maintenant passer en revue quelques-unes de ces questions.

Un point important est celui des rapports entre Maximum de Vraisemblance et Moindres Carrés. Est-il cohérent de faire cohabiter ces techniques ? On peut noter d'abord qu'elles sont moins éloignées qu'il peut paraître : par exemple, les estimateurs de Maximum de Vraisemblance et ceux des Moindres Carrés convenablement pondérés sont asymptotiquement équivalents. Remarquons aussi que les commentaires de J. DE LEEUW apportent quelques éléments de réponse (positive ou négative selon les cas) aux interrogations d'Y. ESCOUFIER. Mais, au delà de la technique des Moindres Carrés, c'est la décomposition en valeurs singulières elle-même qui pourrait être mise en cause comme unique méthode d'analyse des résidus proposée : elle ne s'applique directement qu'à des tableaux à deux entrées et elle cherche systématiquement des approximations en termes d'effets ligne et d'effets colonne. Ses limites sont donc évidentes, et c'est en partie pour cela qu'une analyse par rapport à un modèle, tenant compte d'effets conjoints, peut s'avérer utile. Mais il est clair que ces questions sont très ouvertes et fournissent un vaste champs de réflexions futures.

Il faut ensuite noter l'interrogation sur le rôle et le concept même de modèle qui se dégage d'une façon ou d'une autre de tous les commentaires. Sans vouloir revenir sur ce que certains d'entre nous ont dit à ce sujet ici ou ailleurs, soulignons cependant quelques points. Il semble en premier lieu que, pour beaucoup de statisticiens, le modèle soit de plus en plus considéré d'un point de vue purement utilitariste, comme un outil permettant une approche simplifiée, une représentation « à grands traits », d'une réalité sous-jacente dont l'appréhension complète est illusoire. Il est alors clair qu'une approche ne saurait être entièrement « confirmatoire » au sens fort du terme (comme celui qui est évoqué par le concept de tests d'hypothèses) puisqu'il s'agira au mieux de « confirmer des approximations », pratique qui reste à préciser. Il est donc naturel que la division entre approches confirmatoire et exploratoire ne puisse être que floue (A. LE-

CLERC) ou que la situation pratique soit le plus souvent intermédiaire (T. ALUJA et M. MARTI). En fait, la notion même de modèle, et d'adéquation d'un modèle, variera en fonction de la richesse des données disponibles et cette question est aujourd'hui particulièrement sensible car cette richesse a beaucoup crû ces dernières années, trop vite pour que les cadres de référence évoluent toujours de façon convenable (rappelons à ce sujet que l'essentiel de nos efforts a porté sur des exemples assez volumineux). D'autre part, il faut noter que la part de modélisation variera selon les buts de l'analyse (compréhension ou prévision, par exemple) comme le souligne justement J.-J. DAUDIN; on peut rappeler à ce propos que la qualité prévisionnelle d'un modèle est autre chose que sa bonne adéquation aux données.

Nous répondrons maintenant à quelques remarques plus spécifiques.

T. ALUJA et M. MARTI indiquent à juste titre que, dans certains cas particuliers d'interactions entre les variables, l'Analyse des Correspondances Multiples peut suggérer des liaisons artificielles. Il nous semble que c'est précisément dans de telles situations qu'une modélisation (log-linéaire ou autre) peut à sa suite permettre de corriger les erreurs susceptibles d'être commises en utilisant uniquement la première méthode. La modélisation joue alors typiquement un rôle confirmatoire au sens affaibli dont il est question ci-dessus.

Y. ESCOUFIER remarque que les étapes E et M de l'algorithme E.M. (article de AITKIN et al.) correspondent aux déterminations respectives des fonctions d'affectation et de représentation de la méthode de classification des Nuées Dynamiques; c'est exact, mais il faut préciser que le modèle des Classes Latentes conduit à calculer des probabilités d'affectation des individus aux classes (au moyen de l'algorithme E.M.), alors que les Nuées Dynamiques affectent les individus de façon déterministe : ce n'est donc que dans un cas particulier du modèle de Classes Latentes (affectations avec probabilités 0 ou 1) que l'on retrouve les Nuées Dynamiques (pour plus de détails, voir N. RAYNAL (1987)). A ce même sujet, deux individus aux profils identiques auront les mêmes probabilités d'affectation aux classes dans l'utilisation du modèle des Classes Latentes : la question d'Y. ESCOUFIER sur la possible différence de leur classification pose donc le principe même de la légitimité d'une affectation probabiliste (pour notre part, nous avons décidé de ne pas l'exclure).

Répondons pour terminer aux remarques de J.-J. DAUDIN et Y. ESCOUFIER concernant les logiciels utilisés. L'étude comparative que nous nous étions proposé de mener comportait un certain nombre d'a priori, en particulier l'idée d'utiliser des logiciels standards largement disponibles : SICLA, SPAD et des logiciels du même type pour les méthodes d'analyse des données, GLIM (essentiellement mais non uniquement) pour la modélisation. Il n'en reste pas moins que d'autres seraient mieux adaptés pour tel ou tel point et que le développement de l'aspect logiciel nous est apparu nécessaire : voir le rapport L.S.P.-C.A.S. (1986).

C'est finalement un grand plaisir de remercier très chaleureusement tous ceux qui ont bien voulu participer à cette discussion, sans oublier la Revue de Statistique Appliquée et P. CAZES qui l'ont organisée. Elle a été pour nous extrêmement bénéfique, et il est au plus haut point encourageant que la première étape de ce travail, entrepris sous le signe du dialogue, s'achève ainsi dans les mêmes conditions. Surtout quand on mesure tout ce qui reste à faire...