REVUE DE STATISTIQUE APPLIQUÉE

A. DE FALGUEROLLES
P. G. M. VAN DER HEIJDEN

Sur l'analyse factorielle des correspondances et quelques unes de ses variantes

Revue de statistique appliquée, tome 35, n° 3 (1987), p. 7-12
<http://www.numdam.org/item?id=RSA_1987__35_3_7_0>

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 $A = R^{-1/2} U \Sigma^{\alpha}$, $B = C^{-1/2} V \Sigma^{\beta}$ 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.