CONGRES CARME (Correspondence analysis and related methods)

Conférence invitée, Rennes, Février 2011

Some comments on correspondence analysis

Quelques réflexions sur l'analyse des correspondances

P.CAZES
CEREMADE, Université Paris Dauphine

Plan de l'exposé

CONGRES CARME (Correspondence analysis and related methods)

Some comments on correspondence analysis

Quelques réflexions sur l'analyse des correspondances

P.CAZES
CEREMADE, Université Paris Dauphine

I Introduction

Après avoir rappelé au § II pourquoi l'analyse des correspondances et de façon plus générale l'analyse des données peut être considérée comme une science expérimentale, on analyse au § III l'activité du Laboratoire de Statistique dans les années 1970-1980 avec en particulier le DEA de statistique et les publications qui en sont sorties. On revient ensuite au § IV sur l'importance du codage en analyse des correspondances avec notamment le codage flou et les codages permettant d'obtenir l'équivalence entre l'analyse des correspondances et d'autres analyses. Au § V, on rappelle que l'analyse des correspondances est un cas particulier de nombreuses méthodes classiques, tandis qu'au § VI, on détaille le cas des tableaux multiples. Au § VII, on traite des liaisons entre classification ascendante hiérarchique et analyse des correspondances puis on analyse au § VIII les liens entre analyse des correspondances et statistique classique. Le § IX montre l'intérêt de l'analyse des correspondances dans certains problèmes de modélisation tandis que le § X traite brièvement de l'utilisation de l'analyse des correspondances dans le monde du travail. Enfin, le § XI donne la bibliographie.
On n'a pas cherché à être exhaustif dans cette présentation, se contentant de donner quelques points marquants sur l'analyse des correspondances sans citer toutes les références possibles sur un sujet donné.

II L'analyse des données en tant que science expérimentale

On peut considérer dans une certaine mesure l'analyse des données comme une science expérimentale. D'une part des résultats théoriques ont été découverts et démontrés après les avoir constatés expérimentalement sur des listings d'ordinateurs. D'autre part des indices (taux d'inertie, contributions, etc.) ont été mis en place pour interpréter ou valider les résultats lus sur les sorties d'ordinateurs, ce qu'on peut mettre en liaison avec le calcul des erreurs en physique. Enfin, les techniques de codage permettent de définir le tableau adéquat à analyser ainsi que la succession des analyses à effectuer (analyses descriptives, explicatives, décisionnelles) pour traiter des données ou résoudre un problème posé ce qui est l'analogue du montage d'une expérience en physique.
Je vais revenir plus en détails sur les résultats théoriques trouvés expérimentalement en citant trois exemples.
B. Escofier, dans le cadre de sa thèse, a démontré, après l'avoir constaté sur les listings, que dans le cas d'un tableau de contingence croisant deux ensembles I et J , les moments d'inertie des nuages N_I , N_J associés à I et J respectivement étaient égaux, résultat très classique maintenant. De la même façon, elle a montré que les facteurs sur I et J issus de l'analyse du nuage N_{IxJ} des couples étaient les mêmes que ceux obtenus dans les analyses séparées de N_I et N_J [Ben 82].

En juin 1972, M. Benzécri ayant suggéré à Madame Bara de faire l'analyse des correspondances d'un tableau de 0 et de 1 dédoublé, cette dernière s'est aperçue avec J.P. Pagès que cette analyse était équivalente à l'ACP normée du tableau non dédoublé. Si ce résultat n'avait pas été constaté expérimentalement (mêmes plans factoriels dans les deux analyses), je doute fort qu'on aurait cherché à le démontrer.
De même, en 1971, J.P. Nakache arriva un jour au laboratoire en ayant analysé un tableau de 0 et de 1 (en fait un tableau disjonctif complet) et en montrant au professeur Benzécri qu'il avait obtenu des graphiques intéressants et interprétables. A partir de là M. Benzécri expliqua dans le polycopié Bin. Mult. (qui fut publié cinq ans plus tard dans les Cahiers de l'Analyse des Données, [Ben 77]) pourquoi les résultats obtenus étaient intéressants d'un point de vue pratique en montrant l'équivalence entre l'analyse des correspondances du tableau disjonctif complet et celle du tableau de Burt et cela donna lieu à un problème d'examen pour les étudiants du DEA de statistique puis au développement de l'analyse des correspondances multiples (ACM).
Je n'insisterai pas sur les problèmes de validation qui outre les indices définis pour interpréter les axes factoriels ou les classes issues d'une classification, se sont beaucoup développés avec les progrès de l'informatique et en particulier les possibilités de stockage et la vitesse des ordinateurs (bootstrap, simulations, etc.) Nous nous contenterons de rappeler que L Lebart a beaucoup œuvré pour ces techniques de validation et renverrons à une publication récente sur ce sujet ([Leb 06]) où le bootstrap est en particulier utilisé.
Quant aux problèmes de codage, compte tenu de leur importance, nous en reparlerons plus loin dans le § IV.

III Le laboratoire de statistique dans les années 1970-1980

Durant ces années le DEA de statistique qui était abrité par le laboratoire comportait entre 100 et 200 étudiants, ce qui était au niveau effectifs le plus grand DEA de France. Il en est résulté à partir des années 1974 la soutenance chaque année d'une quarantaine de thèses, dont près d'une quinzaine au mois de juin (et début juillet) où chaque lundi il y avait 3 ou 4 soutenances.
Les exemples d'application étaient très divers : biologie, écologie, économie, géologie, linguistique, médecine, physique, psychologie, sociologie, etc. La diversité des étudiants était aussi très importante : français, bien sûr, africains, argentins, égyptiens, grecs, iraniens, irlandais, libanais, syriens, turcs, vietnamiens, etc.
Il en résultait un foisonnement d'idées et un rayonnement exceptionnel du laboratoire. Ce rayonnement s'est concrétisé par la publication en 1973 des deux tomes du traité sur l'analyse des données qui reprenait la plupart des études effectuées au laboratoire entre 1968 et 1973, ainsi que les cours polycopiés du DEA. La création en 1976 des Cahiers de l'Analyse des Données (CAD) permit d'une part de compléter le traité avec des articles théoriques sur la régression, l'analyse discriminante, l'analyse des correspondances multiples et d'autre part de donner des synthèses des thèses soutenues au laboratoire. Ensuite à partir de 1980, commença la parution des ouvrages de la collection Pratique de l'Analyse des Données, le premier étant relatif à l'analyse des correspondances ([Ben80]) et le second ([Bas80]) traitant des cas modèles et reprenant avec un court rappel de cours un grand nombre de problèmes d'examen du DEA de statistique avec leur solution. Le troisième ([Ben81]) paru en 1981 était relatif à la linguistique, tandis que les deux derniers relatifs à la médecine ([Ben92]) et l'économie ([Ben86]) parurent un peu plus tard. On peut noter que ces deux derniers livres reprenaient essentiellement des articles parus dans les CAD. Signalons enfin le livre du professeur Benzécri ([Ben82]) sur Histoire et préhistoire de l'analyse des données paru en 1982 après sa publication en 4 articles dans les CAD.

Signalons également les colloques de deux jours généralement conviviaux et studieux qui se sont tenus dans de nombreuses universités de province à partir de 1970 : Besançon, Marseille, Nice, Rennes, l'Arbresle près de Lyon, etc.

IV Le codage

IV 1 Codages usuels

Le codage joue un rôle fondamental en analyse des données et en particulier en analyse des correspondances pour définir le tableau à soumettre à une analyse.
Parmi les codages classiques, citons le dédoublement d'un tableau de données (cas d'un tableau de notes, d'un tableau de rangs, d'un tableau de 0-1, etc.), le codage disjonctif complet, le codage flou. Ce dernier codage a donné lieu à de nombreux articles dans les CAD dans les années 1980-1990 : codage barycentrique à 3 et r modalités d'une variable continue, codage permettant de s'affranchir de l'équation personnelle du sujet quand on a des sujets donnant un certain nombre de notes, etc. Une synthèse de ces codages est donnée dans [Caz90].
Dans les tableaux d'échange k_{IJ} où I = J désigne par exemple un ensemble de pays et où le terme général du tableau k (i, j) est égal au total des exportations de i vers j, il est d'usage d'analyser le tableau (k_{IJ} , k_{JI}) juxtaposition du tableau k_{IJ} et de son transposé. Cela permet sur une ligne i d'avoir l'ensemble des échanges du pays i vers le pays j (ensemble des exportations et des importations). Yagolnitzer [Yag77] a suggéré d'analyser le tableau suivant :

k_{IJ}	k_{JI}
k_{JI}	k_{IJ}

L'analyse des correspondances de ce tableau revient en fait à faire l'analyse des correspondances du tableau d'échanges moyen (k_{IJ} + k_{JI})/ 2 et à faire l'analyse factorielle du tableau des flux (k_{IJ} - k_{JI})/ 2 avec les pondérations (poids et métriques) données par l'analyse des correspondances du tableau (k_{IJ} + k_{JI})/ 2.
Notons que ces problèmes de codage sont importants dans l'analyse des tableaux multiples (cf. §VI).
Nous ne détaillerons pas ici l'utilité des éléments supplémentaires (éléments passifs ou illustratifs) qui permettent d'affiner l'interprétation et qui interviennent dans l'analyse des tableaux ternaires et dans certaines procédures comme l'analyse discriminante ou le scoring. Nous nous contenterons de renvoyer à [Caz82].

IV 2 Codages permettant d'obtenir l'équivalence avec d'autres analyses

IV 2. 1 Cas de l'analyse en composantes principales

Notons qu'un codage adapté permet à partir d'une analyse des correspondances de faire l'analyse en composantes principales (ACP), comme l'a montré B. Escofier [Esc79] dans le cas de l'ACP normée (ACPN). De façon générale, si x_{ij} désigne le terme général d'un tableau de données X (valeur pour l'individu i ($1 \leq i \leq n$) de la variable j ($1 \leq j \leq p$)), qu'on a centré, l'analyse des correspondances du tableau dédoublé (par rapport à une quantité positive A quelconque) $\{ [(A + x_{ij}) /2 , (A - x_{ij}) /2)] \mid 1 \leq i \leq n , 1 \leq j \leq p\}$ est équivalente à l'ACP sur matrice variance de X. Si les données sont centrées réduites, on obtient l'ACPN, B. Escofier ayant choisi pour A la valeur 1 de façon à donner même importance à chaque

variable si on analyse un mélange de variables qualitatives et quantitatives, ces dernières étant codées comme ci-dessus et les variables qualitatives suivant le codage disjonctif usuel.
Il faut noter que les tableaux précédents peuvent comporter des éléments négatifs, ce qui ne pose pas de problème dans la mesure où les marges, qui dans ce cas sont uniformes, ont tous leur termes positifs. Par contre si le tableau analysé comporte des éléments négatifs, on peut avoir des valeurs propres supérieures à 1. Pour avoir toutes les valeurs propres inférieures ou égales à 1, il suffit bien sûr de choisir la quantité A de telle sorte que tous les éléments du tableau analysé soient positifs.

IV 2. 2 Analyse par rapport à un modèle

Quand on veut comparer un tableau de fréquences f_{IJ} de marges f_I et f_J avec un tableau de référence m_{IJ}, il est d'usage d'analyser la différence f_{IJ} - m_{IJ}, avec les pondérations (poids, métriques) données par l'analyse des correspondances du tableau f_{IJ}, comme l'a suggéré B. Escofier [Esc84]. Si ces tableaux ont mêmes marges, cette dernière a montré que l'analyse précédente était équivalente à l'analyse des correspondances du tableau f_{IJ} - m_{IJ} + $f_I \otimes f_J$, $f_I \otimes f_J$ étant le tableau associé à l'hypothèse d'indépendance.
On se trouve en particulier dans ce cas quand il y a une structure a priori connue sur le tableau f_{IJ} (partition naturelle sur l'ensemble des lignes I ou l'ensemble des colonnes J ou sur les deux du tableau f_{IJ} et de façon plus générale graphe sur I ou J ou sur les deux) et si on veut faire une analyse faisant abstraction de cette structure (analyse intraclasses) i.e. regardant l'écart entre les données et le modèle associé à cette structure. On pourra en particulier consulter sur ce sujet la référence [Caz91] qui traite le cas où on a un graphe, cas qui redonne l'analyse intraclasses usuelle quand ce graphe se réduit à une partition.

V L'analyse des correspondances comme cas particulier d'autres méthodes

Par définition, l'analyse des correspondances d'un tableau k_{IJ} croisant deux variables qualitatives X et Y est une double analyse factorielle, celle du nuage des profils lignes du tableau k_{IJ} et celle du nuage des profils colonnes.
On sait aussi que cette analyse est l'analyse canonique des deux sous espaces W_X et W_Y engendrés respectivement par les variables indicatrices de X et Y. En fait cette dernière façon de voir correspond à rechercher les codages optimaux (en fait les facteurs) centrés réduits des variables X et Y ayant une corrélation maximum, et à itérer sous contrainte de non corrélation des codages suivants avec les précédents.
Comme l'a souligné L. Lebart, l'analyse des correspondances peut être aussi considérée comme une double analyse discriminante, où dans la première, la variable à expliquer est la variable qualitative Y et les variables explicatives les variables indicatrices de X, et dans la seconde, on fait l'analyse symétrique en échangeant les rôles de X et Y.
De même l'analyse des correspondances multiples (analyse du tableau disjonctif complet associé à q variables qualitatives X_1 ,…, X_q) est un cas particulier de l'analyse canonique généralisée de Carroll où les sous espaces associés sont respectivement engendrés par les variables indicatrices de X_1 ,…, X_q . On peut aussi considérer que cette analyse est l'analyse factorielle multiple (AFM, [Esc98]) du tableau disjonctif complet, chaque sous tableau correspondant aux modalités d'une des variables X_k ($1 \leq k \leq q$), puisque l'analyse des correspondances de chaque sous tableau a toutes ses valeurs propres égales à 1 et donc que la pondération de chacun des sous tableaux par l'inverse de la racine carrée de sa plus grande valeur propre (ici 1) ne change pas l'analyse.
L'analyse des correspondances correspond également à l'analyse interbatteries de Tucker ([Tuc58]) des tableaux T_X et T_Y associés respectivement aux variables indicatrices de X et

Y, avec les métriques diagonale des poids données par les marges de k_{IJ} (ou les marges colonnes de T_X et T_Y). De même l'analyse des correspondances d'un sous tableau de Burt croisant deux sous-ensembles de questions peut être considéré de plusieurs façons différentes comme une analyse de co-inertie multiple ([Che93]).
C'est cette faculté de l'analyse des correspondances d'être un cas particulier de très nombreuses méthodes qui implique sa grande importance aussi bien d'un point de vue théorique que pratique.

VI L'analyse des tableaux multiples ([Caz04])

VI 1 Principes généraux

On considère ici un ensemble de tableaux k_{ItJt} (notés aussi k_t) définis sur le produit de 2 ensembles I_t et J_t avec $t \in \{1, \ldots, t, \ldots r\} = T$ et on suppose qu'un des deux ensembles I_t ou J_t est indépendant de t. On peut sans perte de généralité supposer que I_t=I est indépendantde t et on posera alors $JT = \cup \{J_t \mid t \in T\}$ tandis que $k_{IxJT} = (k_{IJ1}, \ldots, k_{IJt}, \ldots, k_{IJr})$ designera la juxtaposition des k_{IJt}. On désignera également par k_{It} la marge sur I du tableau k_{IJt} et par k_{IT} le tableau $(k_{I1}, \ldots, k_{It}, \ldots, k_{Ir})$ juxtaposition des k_{It}. Souvent on a une série de tableaux à des instants différents et T correspond aux différentes époques où ces tableaux sont connus. Alors classiquement, on fait l'analyse des correspondances du tableau k_{IxJT} en mettant le tableau k_{IT} en supplémentaire. Il en résulte que la représentation du point t est le barycentre des éléments j_t de J_t.
Si on veut une représentation de chaque élément i de I pour chaque instant t, on peut ajouter au tableau analysé k_{IxJT} le tableau bloc diagonal k_{ITxJT} (où IT=IxT) dont le $t^{ème}$ bloc diagonal est le tableau k_{IJt}. On peut aussi représenter T comme un ensemble de lignes supplémentaires à partir du tableau bloc diagonal $k_{T\,xJT}$ dont la $t^{ème}$ ligne est nulle sauf le bloc associé à J_t qui est égal à k_{Jt}, marge sur J_t de k_{IJt}. L'intérêt de toutes les représentations précédentes réside dans le principe barycentrique déjà cité quand T est mis en colonnes supplémentaires. Par exemple, le point i en tant qu'élément actif est le centre de gravité des couples (i, t) pour t appartenant à T. De même le point ligne supplémentaire t est le centre de gravité des couples passifs (i, t) pour i appartenant à I. L'ensemble des possibilités précédentes est résumé dans la figure 1 qui est reprise de [Caz04].

k_{IJ1}	k_{IJ2}	k_{IJr}	k_{IT}
k_{IJ1}	0	0	
0	k_{IJ2}	0	
0	0	k_{IJr}	
k_{J1}	0	0	
	k_{J2}	0	
0	0	k_{Jr}	

Analyse des correspondances de k_{IxJT} avec les tableaux k_{IT} k_{ITxJT} et k_{TxJT} en supplémentaire.
Si de plus J_t =J, on peut rajouter le tableau marginal d'ordre 2 k_{IJ} en supplémentaire.
On peut aussi analyser k_{IT} avec k_{IxJT} en supplémentaire (analyse interclasses).

Figure 1

Compte tenu de la partition de JT suivant les J_t , on peut aussi faire l'analyse interclasses de k_{IxJT} (ce qui revient à faire l'analyse du tableau k_{IT}) ou son analyse intra-classes. Un exemple d'application est donné dans [Caz 94] ou [Mor00].
On peut aussi faire l'AFM ou appliquer la méthode STATIS, ce qui revient, quand les marges sur I des tableaux k_{IJt} sont proportionnelles, à faire l'analyse des correspondances du tableau issu de k_{IxJT} par une pondération adéquate de chaque tableau k_{IJt}. Des détails sur ces analyses sont donnés dans [Caz04].
D'autres analyses sont possibles en particulier si r le nombre d'éléments de T n'est pas trop élevé (r = 2 ou 3). On peut ainsi faire des analyses conditionnelles en faisant l'analyse des correspondances de chaque tableau k_{IJt} (le tableau croisant I avec JT-J_t étant bien sûr mis en supplémentaire).

VI 2 Cas où $I_t = I$ et $J_t = J$ sont indépendants de t (tableau ternaire k_{IJT}).

VI 2.1 Analyses classiques

Dans ce cas, on a un tableau k_{IJT} croisant I, J et T. On désignera respectivement par k_{IJ} , k_{JT} et k_{IT} les tableaux de marge binaires et par k_I, k_J, k_T les marges d'ordre 1 de k_{IJT}.
On peut alors faire l'analyse du tableau k_{IxJT} suggérée au § VI 1, en ajoutant en plus des tableaux supplémentaires déjà citées le tableau k_{IJ}, la représentation de j étant alors le centre de gravité des éléments (j , t) pour t appartenant à T.
En remplaçant I par J, puis I par T, on peut faire ainsi deux autres analyses analogues à celle détaillée au § VI 1.
Si I, J et T jouent des rôles symétriques, on a intérêt à faire l'analyse du tableau de Burt B_{ZZ} croisant $Z = I \cup J \cup T$ avec lui même, les blocs non diagonaux correspondant aux marges binaires du tableau k_{IJT}, tandis que les blocs diagonaux restituent les marges d'ordre un dans leur diagonale avec des zéros ailleurs.
Souvent I et J jouent des rôles symétriques contrairement à T. C'est en particulier le cas si T correspond au temps. On peut alors faire une analyse qui préserve cette symétrie en analysant k_{IJ} (analyse interclasses) et en rajoutant les tableaux k_{IxJT}, k_{IT}, $k_{ITxJ,}$ k_{TJ} en éléments supplémentaires, k_{ITxJ} correspondant au tableau k_{IJT} où les tableaux $k_{IJt} = k_t$ sont superposés et k_{TJ} la marge d'ordre 2 transposée de k_{JT}.
Les deux analyses précédentes (analyse de B_{ZZ} et analyse de k_{IJ}) ont le désavantage de ne pas tenir des interactions d'ordre supérieur à deux.

VI 2.2 Analyses approfondies. Etude des interactions.

Quand les ensembles I, J et T jouent des rôles symétriques, on peut à partir des considérations développées ci-dessus effectuer 6 analyses :
a) l'analyse des 3 tableaux de marge binaire (avec les tableaux adéquats en supplémentaire) qui sont en fait des analyses interclassses.
b) l'analyse des 3 tableaux croisant respectivement un des trois ensembles avec le produit des deux autres (avec toujours les tableaux adéquats en supplémentaire dont les deux tableaux de marge binaire croisant le premier ensemble avec chacun des deux autres).
Pour choisir entre les analyses à effectuer, Choulakian ([Cho88]) propose de se servir de la décomposition du Φ^2 du tableau de fréquence f_{IJT} associé à k_{IJT}, décomposition qui s'écrit :

$$\Phi^2 (I, J, T) = \Phi^2 (I, J) + \Phi^2 (J, T) + \Phi^2 (I, T) + INT (I, J, T) \qquad (1)$$

Φ^2 (I, J) (resp. Φ^2 (J, T) ; Φ^2 (I, T)) étant le Φ^2 (i.e. l'inertie totale dans l'analyse des correspondances) du tableau de marge binaire f_{IJ} (resp. f_{JT} ; f_{IT}) associé au tableau f_{IJT}, Φ^2 (I, J, T) s'écrivant avec des notations évidentes :

$$\Phi^2 (I, J, T) = \Sigma \{(f_{ijt} - f_{i..} f_{.j.} f_{..t})^2 / (f_{i..} f_{.j.} f_{..t}) \mid i \in I,\ j \in J,\ t \in T\} \quad (2)$$

Rappelons que $\Phi^2 (I, J) = \Sigma \{(f_{ij.} - f_{i..} f_{.j.})^2 / (f_{i..} f_{.j.}) \mid i \in I,\ j \in J\}$, Φ^2 (J, T) et Φ^2 (I, T) se définissant de façon analogue. Le terme d'interaction (d'ordre 3) INT (I, J, T) qui intervient dans la décomposition (1) découle de cette décomposition et de la définition (2) de Φ^2 (I, J, T). En fonction des termes qui sont négligeables dans la décomposition précédente, Choulakian préconise l'analyse ou les analyses à effectuer.
Si aucun terme ne peut être négligé, Choulakian suggère une généralisation de l'analyse des correspondances à partir d'une généralisation de la formule de reconstitution faisant apparaître un terme d'interaction d'ordre 3. On peut aussi utiliser le modèle log-linéaire. Etant donné l'importance de ce dernier modèle, et de ses liens avec l'analyse des correspondances, nous en parlerons plus en détail au § IX.
D'autres analyses plus spécifiques ont été proposées dans la littérature (cf. par exemple [Abd00] , [Den00]) pour tenir compte de la structure très particulière d'un tableau ternaire : analyses intraclasses, analyses faisant ressortir les interactions, etc.
En effet, par exemple dans l'analyse du tableau k_{IxJT} croisant I avec JT = JxT, JT est muni de deux partitions (celle induite par J et celle induite par T). On a donc deux analyses interclasses qui correspondent à l'analyse des correspondances des tableaux k_{IJ} et k_{IT}. On peut donc effectuer les analyses intraclasses associées ce qui donne en tout six analyses intraclasses possible contre trois analyses interclasses.
On peut aussi construire dans l'analyse de k_{IxJT} un tableau résiduel entre k_{IxJT} et le tableau associé en l'absence d'interaction entre J et T (en considérant le tableau des profils colonne de k_{IxJT}, on peut considérer qu'on a un modèle d'analyse de variance à deux facteurs avec répétition, ce qui permet de calculer facilement le modèle sans interaction et donc le terme résiduel qui correspond à l'interaction). L'analyse de ce tableau résiduel revient à faire l'analyse des correspondances du tableau dont le terme général est donné (avec des notations évidentes) par [Den00] :

$$k_{ij.} k_{.jt} / k_{.j.} + k_{i.t} k_{.jt} / k_{..t} - k_{ijt}$$

Dans une analyse approfondie d'un tableau croisant l'origine sociale (ensemble I) avec le sexe (ensemble J) et le type d'études effectué (ensemble T = {Droit, Lettres, Sciences, Médecine, etc.}) Carlier ([Car 01]) propose une généralisation de l'algorithme TUCKALS3 dû à Kroonenberg ([Kro83]) en utilisant des métriques diagonales définies à partir des marges d'ordre un du tableau de fréquences $f_{IJT.}$ De façon précise si $X = x^{IJT}$ et $Y = y^{IJT}$ désignent deux vecteurs de R^{IJT} de composantes x^{ijt} et y^{ijt} respectivement, le produit scalaire de ces deux vecteurs est défini par :

$$<X, Y> = \sum \{f_{i..} f_{.j.} f_{..t} x^{ijt} y^{ijt} \mid i \in I,\ j \in J,\ t \in T\}$$

Posant :

$$h^{ijt} = (f_{ijt} - f_{i..} f_{.j.} f_{..t}) / (f_{i..} f_{.j.} f_{..t}) ,$$

la décomposition d'ordre (P, Q, R) h^{ijt*} de h^{ijt} fournie par TUCKALS3 (décomposition que Carlier appelle analyse des correspondances 3 voies) s'écrit :

$$h^{ijt} = h^{ijt*} + e^{ijt}$$

avec :

$$h^{ijt*} = \sum \{g_{pqr} a_{ip} b_{jq} c_{kr} \mid p=1, P ;\ q=1, Q ;\ r=1, R\}$$

et on obtiendra h^{ijt*} en écrivant que la norme du vecteur résiduel e^{IJT} de composantes e^{ijt} est

minimale sous les contraintes d'orthogonalité : $\sum\{f_{i..}\, a_{ip}\, a_{ip'} \mid i \in I\} = \delta_p^{p'} = 1$ si p=p' et zéro sinon, et de même : $\sum\{f_{.j.}\, b_{jq}\, b_{jq'} \mid j \in J\} = \delta_q^{q'}$ et $\sum\{f_{..t}\, c_{tr}\, c_{tr'} \mid t \in T\} = \delta_r^{r'}$.
Il est facile de voir que $\Phi^2(I, J, T)$ n'est autre que le carré de la norme du vecteur h^{IJT} de composantes h^{ijt}, et on a :

$$|| h^{IJT}||^2 = \Phi^2(I, J, T) = || h^{IJT*}||^2 + || e^{IJT}||^2$$

Dans le cas particulier où $f_{JT} = f_J \otimes f_T$ (indépendance des deux dernières variables), Carlier a démontré qu'avec un point de départ adéquat pour l'algorithme TUCKALS3, l'analyse des correspondances 3 voies est équivalente à l'analyse des correspondances du tableau $f_{IxJT.}$

VII Classification ascendante hiérarchique et analyse des correspondances

Nous ne parlerons pas ici de l'enchaînement usuel analyse des correspondances- classification ascendante hiérarchique (CAH) , le cas échéant complété par la méthode des centres mobiles mais du lien d'un point de vue théorique et pratique entre ces deux méthodes.
Rappelons d'abord que quand on a un tableau de contingence k_{IJ} partitionné en blocs ($I = \cup \{ I_b \mid b=1, r \}$, $J = \cup \{ J_b \mid b=1, r \}$), les blocs non diagonaux étant nuls, la CAH de l'un des deux ensembles, I par exemple, avec la métrique du chi-2 et le critère d'agrégation de l'inertie ne permet pas, sauf dans des cas très particuliers, de retrouver la partition correspondante ([Rou89]). On peut par contre (Corr Hier, [Ben73]) construire un modèle mettant en lien analyse des correspondances et classification ascendante hiérarchique, modèle qui a été généralisé par Cazes ([Caz84]).
Si on veut essayer de conserver la symétrie entre les deux ensembles I et J, on peut essayer de faire une classification conjointe de ces deux ensembles. C'est ce que fait Govaert ([Gov83]) avec l'algorithme CROKI2, mais la classification conjointe, qui se fait avec un critère de type nuées dynamiques, ne conduit pas à des hiérarchies de I et J mais à des partitions. Denimal ([Den07b]) effectue une CAH de chaque ensemble séparément, avec la métrique du Chi-2 et un critère d'agrégation du type de celui utilisé dans VARCLUS, à savoir la minimisation de la seconde (et plus petite) valeur propre de l'analyse des correspondances du tableau dédoublé croisant les 2 classes à agréger. Ensuite il propose un élagage des deux hiérarchies H_I et H_J construites sur I et J, en tenant compte des liens entre H_I et H_J puis il donne des aides à l'interprétations intéressantes. En analyse des correspondances multiples, Denimal ([Den10]) propose une classification conjointe de l'ensemble J des modalités et de l'ensemble I des observations d'un tableau disjonctif complet k_{IJ} en adaptant la CAH conjointe de l'ensemble des variables et des observations d'un tableau de mesures ([Den07a]). Dans ce cas, on commence par faire la CAH de l'ensemble J des modalités, puis la CAH de I est construite à partir de la CAH obtenue sur J.

VIII Analyse des correspondances et statistique classique

L'analyse des correspondances et de façon plus générale l'analyse des données font appel à la statistique classique pour valider des résultats ou aider à l'interprétation. Le test du chi-2 peut en particulier être utilisé dans l'analyse des correspondances d'un tableau de contingence pour détecter le nombre de facteurs à conserver. De même, quand on a une variable explicative illustrative x (donc n'ayant pas participé à l'analyse), on peut étudier ses liaisons avec les facteurs issus de l'analyse des correspondances, ce qui peut permettre d'affiner l'interprétation. Si x est quantitative, on peut tester si cette variable est corrélée significativement à un facteur en calculant la corrélation entre x et ce facteur et en faisant le test de Student associé, ou en faisant un test non paramétrique comme le test de corrélation des range de Spearman. Si x est qualitative, on peut tester à l'aide d'un test d'analyse de

variance usuel, si un facteur peut être considéré comme un facteur interclasses (i.e. lié à la structure de partition définie par x) ou non. On peut aussi tester pour chaque modalité de x si la moyenne d'un facteur sur les individus ayant adopté cette modalité diffère significativement de la moyenne générale de ce facteur qui par construction est nulle.
Les statistiques associées aux tests précédents font appel à l'hypothèse de normalité. On peut faire abstraction de cette hypothèse en considérant que ces statistiques sont des indicateurs qui permettent de faire des comparaisons ou des classements d'un point de vue simplement descriptif.
Par ailleurs, on peut s'affranchir du cadre normal, en se plaçant dans des cadres plus généraux ([Ler98]): cadre fréqentiste, auquel est associée la notion de valeur test (cf. par exemple [Leb95]), cadre combinatoire basé sur des tests de permutation et, de façon plus générale, cadre bayésien.
On peut noter que la valeur test qui, dans l'exemple d'une modalité illustrative, traduit de façon normalisée (en se ramenant à une loi normale centrée réduite) l'écart sur un axe factoriel entre la moyenne des abscisses des individus ayant adopté cette modalité et la moyenne générale qui vaut zéro est un indicateur très utile qui est fourni par le logiciel SPAD et qui est également utilisée pour aide à l'interprétation d'une classification.

IX Analyse des correspondances et modélisation

IX 1 La formule de reconstitution considérée comme une technique de modélisation

La formule de reconstitution des données qui permet de reconstituer de façon exacte un tableau de fréquence f_{IJ} à partir des marges, des facteurs et des valeurs propres issues de l'analyse des correspondances de ce tableau peut être considérée comme un modèle quand on approche ce tableau à partir des r premiers facteurs.
Cette formule s'écrit, en désignant par f_{ij}^{*} l'approximation du terme général f_{ij} de f_{IJ} quand on conserve r facteurs :

$$f_{ij}^{*} = f_{i.}\, f_{.j} \quad (1 + \Sigma \{(\lambda_\alpha)^{1/2} \varphi^{i}_{\alpha} \varphi^{j}_{\alpha} \mid \alpha = 1, r \}) \qquad (3)$$

où $\varphi^{I}_{\alpha} = \{\varphi^{i}_{\alpha} \mid i \in I \}$ et $\varphi^{J}_{\alpha} = \{\varphi^{j}_{\alpha} \mid j \in J \}$ sont les facteurs sur I et J de variance 1 associés à la valeur propre λ_α et $f_I = \{f_{i.} \mid i \in I \}$ et $f_J = \{f_{.j} \mid j \in J \}$ les marges de f_{IJ} .
On peut noter qu'au voisinage de l'indépendance, la formule précédente est très voisine du modèle log- linéaire.
En effet (3) s'écrit approximativement en faisant un développement limité d'ordre 1 :

$$\text{Log}\,(f_{ij}^{*}) = \text{Log}\,(f_{i.}) + \text{Log}\,(f_{.j}) + \Sigma \{(\lambda_\alpha)^{1/2} \varphi^{i}_{\alpha} \varphi^{j}_{\alpha} \mid \alpha = 1, r\}$$

ce qui correspond au modèle log -linéaire

$$\text{Log}\,(f_{ij}^{*}) = \mu + \alpha_i + \beta_j + \gamma_{ij} \qquad (4)$$

avec $\mu = 0$, $\alpha_i = \text{Log}\,(f_{i.})$, $\beta_j = \text{Log}\,(f_{.j})$, et où le terme d'interaction γ_{ij} se met sous la forme $\Sigma \{(\lambda_\alpha)^{1/2} \varphi^{i}_{\alpha} \varphi^{j}_{\alpha} \mid \alpha = 1, r \}$, terme qui se réduit si r = 1 à un terme d'interaction multiplicatif $(\lambda_1)^{1/2} \varphi^{i}_{1} \varphi^{j}_{1}$. On peut noter que l'absence d'interaction correspond à l'indépendance, ce qui

fait que le modèle log- linéaire présente un intérêt restreint quand on n'a que 2 variables sauf si on modélise le terme d'interaction.
Nous allons donner deux exemples d'une telle modélisation avec un seul facteur conservé dans le premier et deux dans le second.

Si un seul facteur semble suffisant pour expliquer les données et que les modalités d'une ou des deux variables que l'on croise pour obtenir le tableau f_{IJ} sont ordonnées, et que cet ordre est respecté (ou pratiquement respecté) sur le premier axe factoriel, on peut faire des hypothèses d'espacement constant entre modalités adjacentes sur cet axe. On peut aussi supposer des valeurs égales pour deux modalités proches sur cet axe (et donc dans l'espace), ce qui revient à considérer ces modalités confondues et donc à cumuler les deux lignes ou colonnes associées. On obtient ainsi un modèle plus sophistiqué (que le modèle initial de l'analyse des correspondances) dont les paramètres sont estimés à partir d'une analyse des correspondances sous contraintes , ce qui revient à un ajustement par la méthode des moindres carrés.
Un exemple d'une telle approche est fourni par Goodman [Gou85] pour étudier la liaison entre l'état mental (4 modalités) et le statut socio-économique des parents (6 modalités), ces deux variables étant mesurées sur un échantillon de 1600 individus. Dans les modélisations précédentes, Goodman utilise pour estimer les paramètres soit la méthode des moindres carrés (ce qui revient à l'analyse des correspondances) soit la méthode du maximum de vraisemblance, et il propose des tests pour valider les modèles proposés. La formule de reconstitution des données correspond au modèle d'association RC de Goodman.

Le second exemple de modélisation où deux facteurs sont utilisés est donné par Worsley ([Wor87]). Le tableau étudié croise un ensemble J de 9 modes de suicides avec un ensemble I produit du sexe par l'âge découpé en 17 tranches, soit 34 modalités en tout. Le premier axe factoriel (52% de l'inertie) oppose les hommes et les femmes, tandis que le second axe (38% de l'inertie) semble traduire pour chaque sexe un effet linéaire du temps.
La formule (4), déduite de (3) avec r = 2, peut alors se mettre sous la forme suivante, en tenant compte des constatations précédentes :

$$\log (f_{ij}^{*}) = \mu + \alpha_i + \beta_j + x_{1i}\, u_{1j} + x_{2i}\, u_{2j}$$

où $x_{1i} = 0$ pour une femme et 1 pour un homme, tandis que $x_{2i} = k$ si $i = 2k - 1$ ou $i = 2k$, 2k-1 (resp. 2k) correspondant à la $k^{ème}$ clase d'âge pour une femme (resp. un homme) ($1 \leq k \leq 17$).

IX 2 Analyse des correspondances et modèle log- linéaire dans le cas des tableaux multiples.

Nous n'avons étudié au § VI que le cas des tableaux multiples d'ordre 3, mais en pratique le nombre de variables est souvent supérieur.
Le problème qui se pose alors est de détecter les interactions significatives entre variables. L'analyse des correspondances ne permet en effet de détecter que les interactions d'ordre 2.
On peut toujours analyser des interactions d'ordre supérieur en construisant des variables produit. C'est ainsi que dans un tableau ternaire f_{IJT}, l'analyse du tableau f_{IxJT} croisant I avec JT = J x T permet d'étudier l'interaction d'ordre 3. Cette façon d'opérer revient en effet à faire l'analyse des correspondances du tableau croisant la première variable avec produit des deux autres. Notons qu'il y a trois analyses possibles suivant le choix du couple de variables dont on fait le produit. Le cas se complique quand on a 4 variables ou plus, le nombre de tableaux analysables faisant intervenir toutes les variables pour avoir l'interaction d'ordre maximal, devenant rapidement très grand (trois tableaux avec 3 variables comme dit ci-dessus, sept avec 4, quinze avec 5 , trente et un avec 6, etc.). Néanmoins dans le cas de 4

variables, Choulakian ([Cho88]) fait, comme dans les cas de 3 variables, des suggestions pour analyser les tableaux adéquats en se basant sur la décomposition du Φ^2 dont la définition est analogue à celle donnée au § VI 2.2 dans le cas de 3 variables.
Il est alors intéressant de se servir du modèle log- linéaire pour détecter les interactions significatives. L'utilisation simultanée de l'analyse des correspondances et du modèle log-linéaire peut alors être utile pour le praticien, soit en faisant l'analyse des correspondances d'un ou plusieurs tableaux suggérés par ce modèle, soit en faisant une analyse partielle tenant compte des résultats du modèle log- linéaire pour éliminer une très forte interaction et analyser plus en détail les données une fois éliminée cette interaction (cf. § IV 2.2).
La liaison entre modèle log- linéaire et analyse des correspondances est étudié dans [Leb95] qui donne une petite bibliographie sur le sujet avec en particulier le livre de Van der Heidjen ([Van87]), l'article avec discussion de Goodman ([Gou 91]) ainsi qu'un certain nombre d'articles parus dans la Revue de Statistique Appliquée dont le numéro spécial (numéro 2 de 1987) comparant les méthodologies française et anglophone d'analyse des données qualitatives.
Nous nous contenterons ici de rappeler un des premiers articles parus sur le sujet ([Dau80]), article qui étudie le lien entre 6 variables, 5 variables à 2 modalités relatives à la main d'œuvre, et la variable région (les 21 régions de la France) à partir d'une enquête effectuée sur près de 20000 exploitations agricoles françaises. Les auteurs effectuent l'analyse des correspondances du tableau de dimensions 32 (2^5) x21 croisant le produit des variables binaires avec la variable région, ce qui leur permet d'interpréter les axes 1 et 2 qui représentent les deux tiers de l'information mais pas les axes suivants. L'étude par le modèle log- linéaire des interactions d'ordre 2, 3 et 4 a permis de compléter et d'affiner les résultats obtenus par l'analyse des correspondances.

IX 3 L'analyse des correspondances comme étape intermédiaire dans un problème de modélisation

Comme toute technique d'analyse factorielle, l'analyse des correspondances intervient comme une méthode de compression des données avant la phase de modélisation. Par ailleurs l'utilisation simultanée de l'analyse des correspondances et d'un modèle permet d'avoir deux visions complémentaires des données ce qui peut permette d'affiner la modélisation.
On se contentera de citer trois exemples de l'utilisation conjointe de l'analyse des correspondances avec d'autres méthodes statistiques dans des problèmes de modélisation :
Si on dispose de l'estimation d'une loi de probabilité, par un histogramme par exemple, et que cette loi est un mélange de lois de probabilité connues (par exemple des lois de Poisson de paramètre fixé), l'estimation des proportions inconnues p_i ($1 \leq i \leq k$) peut se faire à partir d'une régression sous la contrainte que les coefficients p_i soient positifs ou nuls et de somme 1. Le problème précédent étant mal conditionné, la régression risque d'être illusoire d'où l'intérêt de compresser les données en faisant la régression sur les premiers facteurs de l'analyse des correspondances du tableau des lois de probabilité ([Caz78]).
Un exemple classique, où l'analyse des correspondances est très souvent utilisée, est le scoring où on cherche à prévoir la classe d'un individu (bon payeur - mauvais payeur ; survivant - décédé, etc.) en fonction d'un certain nombre de variables explicatives de nature quelconque. Après découpage en classes des variables explicatives numériques et construction du tableau disjonctif complet k_{IJ} de toutes les variables explicatives, on effectue l'analyse factorielle discriminante sur les facteurs issus de l'analyse des correspondances de ce tableau, en sélectionnant les facteurs associés à un pourcentage d'inertie fixé (80 ou 90%) et discriminants (ce que l'on peut tester par un test classique d'analyse de variance). Il s'agit de la méthode DISQUAL introduite par Saporta ([Sap77]).

On peut aussi construire le tableau k_{CJ} croisant l'ensemble C des deux modalités de la variable à expliquer avec l'ensemble J de toutes les modalités explicatives, faire l'analyse des correspondances de ce tableau qui comporte un seul axe opposant les deux classes à prévoir. La projection sur cet axe, en éléments supplémentaires, de l'ensemble I des individus, caractérisés par les modalités explicatives, donne le score qui permet d'affecter chaque individu à une classe. C'est l'analyse discriminante barycentrique (ADB) qui peut aisément se généraliser, comme DISQUAL, au cas où la variable à expliquer comporte plus de deux modalités.
Si dans la méthode DISQUAL, on garde tous les facteurs, on obtient le même axe discriminant dans l'espace R_J des modalités explicatives que dans l'ADB, à savoir l'axe joignant les centres de gravité des deux modalités à prévoir. Par contre les scores discriminants diffèrent, car dans le premier cas on projette les individus avec la métrique d'inertie, ce qui revient à raisonner (avec la métrique usuelle) sur les facteurs sur J de variance 1, tandis que dans le second cas on se sert de la métrique du chi-2 issu de l'analyse des correspondances de k_{CJ} (ou k_{IJ}) ce qui revient à raisonner (avec la métrique usuelle) sur les facteurs usuels, i.e. les facteurs de variance la valeur propre.
Notons enfin, qu'au lieu de faire l'analyse factorielle discriminante, on peut aussi effectuer la régression logistique sur les facteurs issus de l'analyse des correspondances du tableau disjonctif complet k_{IJ} .
Dans un troisième exemple ([Mor94]), l'utilisation simultanée de l'analyse des correspondances, de la CAH, de l'analyse en composantes principales (ACP) et de la modélisation de Box-Jenkins a permis de prévoir la quantité de ventes de périodiques d'un certain nombre de grossistes avec une précision meilleure que celle fournie par les experts. La donnée de départ est un tableau k_{IT} croisant un ensemble I de 1577 grossistes avec un ensemble T de 157 semaines, le terme général de ce tableau étant le total des ventes k (i, t) du grossiste i la semaine t. Une analyse des correspondances du tableau des profils lignes du tableau k_{IT} suivie d'une CAH a permis de définir des classes de grossistes. Pour chaque classe, on fait alors l'ACP du sous tableau de k_{IT} correspondant et la formule de reconstitution des données permet d'exprimer le terme général k(i, t) en fonction de la moyenne m (t) des ventes dans la classe et des facteurs F_α (i) et F_α (t), la précision étant suffisante si on garde 2 facteurs . La modélisation de m(t), F_1 (t) et F_2 (t) par un processus SARIMA a permis de prévoir k (i, t) avec une bonne précision comme on l'a dit ci dessus.

X L'utilisation de l'analyse factorielle dans le monde du travail

Dans le secteur public, l'analyse factorielle (analyse des correspondances, ACP) est essentiellement utilisée dans les laboratoires universitaires (laboratoires de mathématiques appliquées, de psychologie, de sociologie, etc.), dans les départements de mathématiques appliquées des grands organismes publics de recherche (INRIA, INRETS, etc.) et des écoles d'ingénieurs, en particulier les écoles d'agronomie.
Si dans le secteur privé, l'ACP est la technique factorielle qui semble la plus utilisée, c'est essentiellement l'analyse des correspondances multiples (et non pas binaires) qui est utilisée soit comme un chaînon intermédiaire d'une méthodologie de prévision (par exemple le scoring), soit pour dépouiller des enquêtes dans des services de marketing ou géo marketing par exemple.
Parmi les techniques d'analyse des données le plus souvent utilisées par les étudiants de l'ISUP ou de masters seconde année (ex DESS) en statistique lors de leur stage de fin d'études, citons la régression, l'analyse discriminante, la régression logistique, la régression PLS et de façon plus générale l'approche PLS, l'ACP, l'AFM, les méthodes de classification,

les réseaux de neurones, et en particulier les cartes de Kohonen, etc. La régression PLS s'est beaucoup développée depuis 15 ans pour étudier des données issues de la spectrométrie proche infra rouge, ce type de données intervenant souvent en contrôle de qualité et étant devenues nombreuses avec le développement de spectromètres performants. Notons également que des extensions ou des méthodologies intéressantes basées sur l'analyse factorielle ont été développées dans le cadre de l'analyse sensorielle qui a pris une grande importance depuis quelques années.
Néanmoins il nous semble que de façon générale, l'analyse des données et en particulier les méthodes factorielles restent encore insuffisamment utilisées en France, alors qu'elles pourraient rendre de grands services dans de nombreux cas aux praticiens disposant de bases de données de plus en plus volumineuses et complexes.
Signalons pour terminer le rapprochement récent entre la communauté de l'analyse des données et celle de l'apprentissage, rapprochement qui a donné lieu au symposium SLdS (Statistical Learning and Data Sciences) 2009 qui s'est tenu à l'Université Paris-Dauphine début avril 2009. Les principales communications de cet événement viennent d'être publiées dans un numéro spécial (numéro 42, été 2010) de la revue MODULAD.

XI Bibliographie

[Abd00] ABDESSEMED, L., ESCOFIER, B. (2000) : Analyse de l'interaction et de la variabilité inter et intra dans un tableau de fréquence ternaire, in L'analyse des correspondances et les techniques connexes. Approches nouvelles pour l'analyse statistique des données, Eds. J. MOREAU, P.A. DOUDIN, P. CAZES, Springer, pp. 145- 164.
[Bas80] BASTIN, Ch., BENZECRI, J.P., BOURGARIT, Ch., CAZES, P. (1980) : Pratique de l'analyse des données, Tome 2 : Abrégé théorique. Etude de cas modèles, Dunod, 477 pages.
[Ben73] BENZECRI, J.P. (1973) : L'analyse des données, Tome 1 : La taxinomie, 627 pages Tome 2 : L'analyse des correspondances, Dunod, 619 pages.
[Ben77] BENZECRI, J.P. (1977): Sur l'analyse des tableaux binaires associés à une correspondance multiple, [Bin Mult.], CAD, Vol. 2, n° 1, pp. 55, 71.
[Ben80] BENZECRI, J.P., BENZECRI, F. (1980) : Pratique de l'analyse des données, Tome 1 : Analyse des correspondances. Exposé élémentaire, Dunod, 432 pages.
[Ben81] BENZECRI, J.P. (1981) : Pratique de l'analyse des données, Tome 3 : Linguistique & lexicologie, Dunod, 575 pages.
[Ben82] BENZECRI, J.P. (1982) : Histoire et préhistoire de l'analyse des données, Dunod, 159 pages.
[Ben86] BENZECRI, J.P., BENZECRI, F. (1986) : Pratique de l'analyse des données, Tome 5 : Economie, Dunod, 543 pages.
[Ben92] BENZECRI, J.P., BENZECRI, F., MAITI, G.D. (1992) : Pratique de l'analyse des données, Tome 4 : Médecine, pharmacologie physiologie clinique, Statmatic, 542 pages.
[Car 01] CARLIER, A. (2001) : Examples of 3-ways correspondence analysis (non publié).
[Caz78] CAZES, P. (1978) : Estimation de la statistique de multiplication du premier étage d'un photomultiplicateur à dynodes, [Photomultiplicateur], CAD, Vol.3, n° 4 pp. 393, 417.
[Caz82] CAZES, P. (1982) : Note sur les éléments supplémentaires en analyse des correspondances, I : [El Sup.1], Pratique et utilisation, CAD, Vol.7, n° 1, pp. 9-23 ; II : [El Sup.2], Tableaux multiples, CAD, Vol.7, n° 2, pp. 133-154.
[Caz84] CAZES, P. (1984) : Correspondances hiérarchiques et ensembles associés, Cahiers du B.U.R.O., n^{os} 43-44, pp. 43-142

[Caz90] CAZES, P. (1990) : Codage d'une variable continue en vue de l'analyse des correspondances, RSA, Vol. 38, n° 3, pp. 33-51.
[Caz91] CAZES, P., MOREAU, J. (1991): Analysis of a contingency table in which the rows and columns have a graph structure, in Symbolic-Numeric Data Analysis and Learning, Eds. E. DIDAY, Y. LECHEVALLIER, Nova Sciences Publishers, pp. 271-280.
[Caz94] CAZES, P., MOREAU, J., DOUDIN, P.A. (1994) : Etude des variabilités interindividuelles et intraindividuelles dans un questionnaire où toutes les questions ont le même nombre de modalités. Application à une recherche sur le développement de l'intelligence, RSA, Vol. 42, n° 2, pp. 5-25.
[Caz04] CAZES, P. (2004) : Quelques méthodes d'analyse factorielle d'une série de tableaux de données, la revue de MODULAD, n° 31, pp. 1-31.
[Che93] CHESSEL, D., MERCIER, P. (1993) : Couplage de triplets statistiques et liaisons espèces – environnement, Eds. LEBRETON, J.D., ASSELAIN, B., Masson, Paris, pp. 15-44.
[Cho88] CHOULAKIAN, V. (1988) : Analyse factorielle des correspondances de tableaux multiples, RSA, Vol. 36, n° 4, pp. 33-42.
[Dau80] DAUDIN, J.J., TRECOURT, P. (1980) : Analyse factorielle des correspondances et modèle log- linéaire : Comparaison des deux méthodes sur exemple, RSA, Vol.28, n° 1, pp. 5-24.
[Den00] DENIMAL, J.J. (2000) : L'analyse factorielle des interactions, in L'analyse des correspondances et les techniques connexes. Approches nouvelles pour l'analyse statistique des données, Eds. J. MOREAU, P.A. DOUDIN, P. CAZES, Springer, pp. 165- 180.
[Den07a] DENIMAL, J.J. (2007a) : Classification factorielle hiérarchique optimisée d'un tableau de mesures, JSFdS - RSA, Vol. 148 n° 2, pp. 29, 63.
[Den07b] DENIMAL, J.J. (2007b) : Classification factorielle hiérarchique optimisée des lignes et des colonnes d'un tableau de contingence, JSFdS - RSA, Vol. 148 n° 3, pp. 37, 70.
[Den10] DENIMAL, J.J. (2010): Extension aux correspondances multiples de la classification hiérarchique optimisée, JSFdS, à paraître.
[Esc79] ESCOFIER, B. (1979) : Traitement simultané de variables qualitatives et quantitatives en analyse factorielle, [Qualitatives et Quantitatives], CAD, Vol. 4 n° 2, pp. 137-146.
[Esc84] ESCOFIER, B. (1984) : Analyse factorielle en référence à un modèle. Application à l'analyse de tableaux d'échanges, RSA, Vol. 32 n° 4, pp. 25-36.
[Esc98] ESCOFIER, B., PAGES, J. (1998) : Analyses factorielles simples et multiples. Objectifs, méthodes et interprétation, 3éme éd., Dunod, 300 pages.
[Gou85] GOODMAN, L.A. (1985): Correspondence analysis models, log-linear models, and log-bilinear models for the analysis of contingency tables, Proceedings of the 45th session of ISI, Amsterdam.
[Gou91] GOODMAN, L.A. (1991): Measures, models, and graphical displays in the analysis of cross-classified data (with discussion), J. of Amer. Stat. Assoc., vol.86, pp.1085-1138.
[Gov83] GOVAERT, G. (1983) : Classification croisée, thèse d'état, Un. Paris 6.
[Kro83] Kroonenberg, P.M. (1983): Three-mode principal component analysis: Theory and applications, Leiden, DSWO Press.
[Leb95] LEBART, L., MORINEAU, A., PIRON, M. (1995) : Statistique exploratoire multidimensionnelle, Dunod, 456 pages.
[Leb06] LEBART, L. (2006) : Validation Techniques in Multiple Correspondence Analysis, in Correspondence Analysis and Related Methods, Ed. M. GREENACRE, J. BLASIUS, Chapman & HALL, pp.179, 195.
[Ler98] LEROUX, B. (1998) : Inférence combinatoire en analyse géométrique des données, Math. Inf. Sci. Hum., Vol. 36, numéro 144, pp. 5-14.

[Mor00] MOREAU, J., DOUDIN, P.A., CAZES, P. (2000) : Etude de la variabilité intraindividuelle par l'analyse des correspondances, in L'analyse des correspondances et les techniques connexes. Approches nouvelles pour l'analyse statistique des données, Eds. J. MOREAU, P.A. DOUDIN, P. CAZES, Springer, pp. 106- 119.
[Mor94] MORINEAU, A., SAMMARTINO, A.E., GETTLER-SUMMA, M., PARDOUX, C. (1994) : Analyse des données et modélisation des séries temporelles. Application à la vente de périodiques, RSA, Vol. 42, n° 4, pp. 61-81.
[Rou89] ROUSSEAU, R. (1989) : Reconnaissance de la structure de blocs d'un tableau de correspondance par la classification hiérarchique (suite), [Rec.Bloc.II], CAD, Vol 14, n° 3, pp. 257-266.
[Sap77] SAPORTA, G. (1977) : Une méthode et un programme d'analyse discriminante sur variables qualitatives, Premières Journées Internationales, Analyse des Données et Informatique, INRIA, Versailles.
[Tuc58] TUCKER, L.R. (1958): An inter-battery method of factor analysis, Psychometrica, Vol. 23, n° 2, pp. 111-136.
[Van 87] VAN DER HEIDJEN, P.G.M. (1987): Correspondence analysis of longitudinal categorical data, DSWO Press, Leiden, 282 pages.
[Wor87] WORSLEY (1987) : Un exemple d'identification d'un modèle log- linéaire grâce à une analyse des correspondances, RSA, Vol. 35, n° 3, pp.13-20.
[Yag77] YAGOLNITZER (1977) : Comparaison de deux correspondances entre les mêmes ensembles, [Compar. Corr.], CAD, Vol. 2, n° 3, pp. 251, 264.