La revue de Modulad

ANALYSE DES CORRESPONDANCES MULTIPLES CONDITIONNELLE

Brigitte ESCOFIER

IRISA, Campus Universitaire de Beaulieu 35042 RENNES Cédex
IUT rue Montaigne 56008 Vannes Cedex

Résumé :

L'analyse des correspondances multiples conditionnelle est une méthode qui dérive de l'analyse des correspondances multiples. Celle ci sert à étudier, d'une part, les liaisons entre plusieurs variables qualitatives définies sur une même population et d'autre part, la structure induite par l'ensemble de ces variables sur la population. L'analyse des correspondances multiples conditionnelle permet d'introduire un conditionnement par rapport à une variable extérieure et d'éliminer de ces études la part liée à cette variable extérieure. Elle possède les principales propriétés de l'analyse des correspondances multiples.

Le programme MULCO qui effectue cette analyse accepte les données manquantes pour les variables qualitatives. Il est aussi possible d'obtenir les résultats de cette analyse en appliquant un programme classique d'analyse des correspondance à un tableau obtenu par une transformation, soit du tableau disjonctif complet, soit du tableau de Burt.

Mots clés :

Analyse des correspondances multiples, conditionnement, analyse des correspondances, analyse des correspondances et modèle, analyse des correspondances et contrainte

1. Le problème et les principales propriétés de la solution

L'analyse des correspondances multiples (ACM) sert à étudier, d'une part, la liaison entre plusieurs variables qualitatives définies sur une même population et d'autre part, la structure induite par l'ensemble de ces variables sur la population.

Or, quelquefois, toutes ces variables ou un grand nombre d'entre elles sont liées à une même variable qualitative que nous noterons T en référence au temps. Par exemple, il peut y avoir une évolution temporelle importante de l'ensemble des réponses à une même enquête effectuée à des dates échelonnées dans le temps. Dans cette hypothèse de dépendance de toutes les variables à la variable T, une analyse des correspondances multiples classique fera ressortir essentiellement cette liaison, (même si la variable T n'est pas introduite dans les données). De même, une classification des individus sur les facteurs de cette analyse fera réapparaître la partition définie par T.

Si l'on s'intéresse, non pas à l'évolution temporelle, mais à l'étude des liaisons stables dans le temps, il faut éliminer l'influence de la variable T. Nous proposons ici une analyse qui réalise cette élimination tout en possédant les propriétés les plus remarquables de l'analyse des correspondances multiples :

- construction et projection de deux nuages (individus et modalités) sur leurs axes d'inertie.
- dualité et formules de transition entre les projections des individus et les projections des modalités ; celle qui relie une modalité à l'ensemble des individus est la formule barycentrique classique.
- équivalence avec l'analyse des correspondances d'un tableau analogue au tableau de Burt dans lequel les tableaux croisés sont remplacés par des tableaux croisés "conditionnés" par rapport à T.
- équivalence avec une analyse multicanonique dans laquelle on introduit une contrainte.

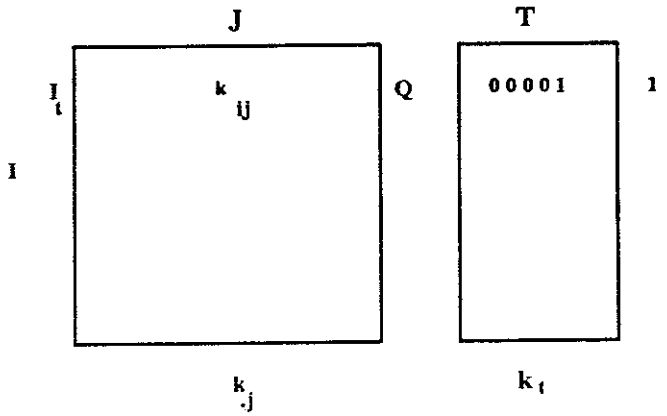
2. Les notations et le plan

Pour concrétiser les notions utilisées, et en référence aux dépouillement d'enquêtes, nous parlerons dans la suite d'individus, de questions avec leurs modalités de réponses et du temps pour la variable de conditionnement.

On note **I** l'ensemble des individus et son effectif. Sur ces individus sont définies Q+1 variables qualitatives : les **Q** variables à étudier et la variable de conditionnement. On note **J** l'ensemble des modalités des Q variables et **T** la variable de conditionnement et l'ensemble de ses modalités.

L'ACM peut être considérée comme une analyse factorielle des correspondances (AFC) d'un Tableau Disjonctif Complet (TDC). On note k_{ij} le terme général du tableau disjonctif complet qui croise I et J ; sa marge sur I est constante et vaut Q, et l'on note $k_{.j}$ sa marge sur J. On note de même k_{it} le terme général du tableau disjonctif complet, réduit à une variable, qui croise I et T ; sa marge sur I vaut 1 et sa marge sur T est notée $k_{.t}$. La variable T définit une partition de l'ensemble I, dont les classes sont notées I_t. A chaque classe I_t correspond un sous tableau et le TDC complet est la juxtaposition de ces T sous tableaux.

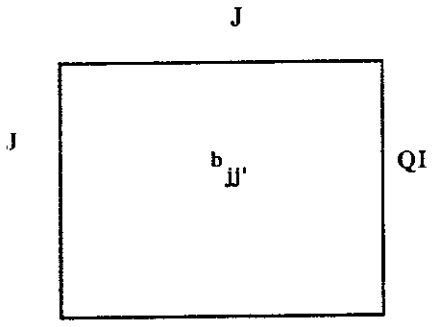
Dans un premier temps, nous présentons l'ACM conditionnelle dans cette optique. Cette technique peut alors être vue comme une étude des structures définies par l'AFC du TDC, en éliminant dans cettte étude la part de dispersion liée à la variable extérieure T. Nous donnons l'interprétation géométrique de l'élimination de la dispersion liée à T dans les nuages d'individus (paragraphe 3) et de modalités (paragraphe 4). Dans le paragraphe 5, nous donnons son expression au travers de tableaux dérivés du TDC et nous indiquons une technique simple qui permet d'obtenir les résultats de l'ACM conditionnelle avec un programme classique d'AFC.



Le tableau disjonctif complet et ses marges.

On considère aussi le tableau qui croise T et J dont le terme général est noté b_{tj}. Ce tableau se déduit du TDC en sommant les lignes associées à la même modalité t. Ses marges sur J et T sont respectivement égales à $k_{.j}$ et $Q\,k_{t.}$

On peut aussi définir l'ACM comme une AFC d'un tableau de Burt. On note $b_{jj'}$ le terme général de ce tableau. La présentation de l'ACM conditionnelle qui en dérive est l'objet du paragraphe 6.



Le tableau de Burt et ses marges.

3. Le nuage des individus en ACM conditionnelle

3.1 Le cas simple de variables numériques

Dans le cas de variables numériques, et donc de l'analyse en composantes principales,

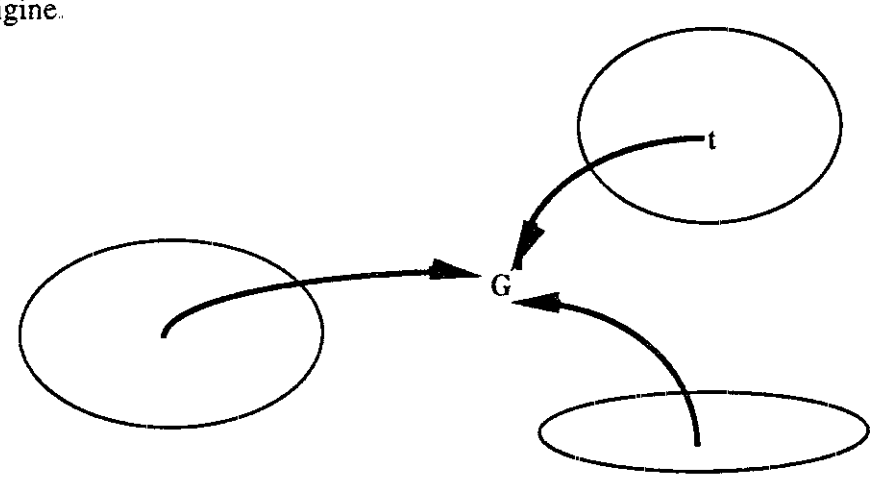
il est très simple d'éliminer une variation temporelle. Par exemple, si l'on dispose de caractéristiques financières d'un certain nombre d'agences pendant plusieurs années et que la moyenne de ces caractéristiques varient notablement d'une année sur l'autre, on peut transformer les données en centrant les variables par année. Une agence sera caractérisée, non pas par son écart à la moyenne générale, toutes années confondues, mais par son écart avec la moyenne de l'année. On obtiendra ainsi une typologie des agences dans laquelle la dérive systématique annuelle est éliminée.

En ACM conditionnelle, pour le nuage des individus, le principe est tout à fait analogue, bien que plus complexe à interpréter. Mais il ne peut s'appliquer par le biais de la même transformation des données à cause de l'ensemble des éléments entrant en jeu en ACM : profils, distance du khi2 et poids. D'autre part, l'ACM possède un ensemble de propriétés qui en font une méthode beaucoup plus riche que l'ACP et la plupart d'entre elles sont transposées en ACM conditionnelle.

3.2 Définition du nuage d'individus : analyse "intra" ou cumulée

Le nuage d'individus étudié dans l'ACM des Q variables est divisé en T sous-nuages : à chaque modalité t de T correspond le sous-nuage des individus interrogés au temps t Si, les réponses aux questions ont une déviation systématique importante avec le temps, ces sous-nuages sont assez éloignés les uns des autres.

Le nuage d'individus de l'ACM conditionnelle se déduit de celui de l'ACM classique en recentrant les T sous-nuages d'individus caractérisés par la même modalité t au centre de gravité général du nuage : chaque sous-nuage est translaté de manière à ce que son barycentre coïncide avec l'origine.



Les nuages d'individus correspondant à chaque modalité t sont recentrés.

On peut décomposer l'inertie du nuage d'individus de l'ACM suivant Huygens : en inertie inter (inertie des T barycentres) et inertie intra (inertie de chacun des T sous-nuages autour de leur barycentre). Recentrer les sous-nuages supprime la dispersion inter T, il ne subsiste que la dispersion intra. L'ACM conditionnelle est une analyse intra.

On peut aussi considérer l'ACM conditionnelle commme une analyse cumulée des différentes époques. En effet, dans l'ACM séparée de chacune des populations, le nuage d'individus est centré ; dans l'ACM conditionnelle, on étudie un nuage qui n'est autre que

l'union des T nuages centrés considérés dans les T ACM séparées.

3.3 Interprétation des distances dans le nuage d'individus

Cette translation des sous-nuages revient à représenter un individu, non pas par son écart avec la moyenne générale de la population, comme dans l'ACM classique, mais par son écart avec la population questionnée au même instant t que lui. L'interprétation de cet écart est plus complexe que pour des variables numériques où seules entrent en jeu des différences numériques. Ici, la référence est la répartition de la sous-population suivant les différentes modalités de toutes les variables et non une simple moyenne. Précisément, un individu i ayant la modalité t aura comme coordonnée sur l'axe j la valeur :

$$k_{ij} / Q - b_{tj} / Q \; k_{.t}$$

Prenons un exemple pour concrétiser cette notion. Supposons qu'une des questions ait 4 modalités de réponses et que l'ensemble de la population enquêtée se répartisse également entre les quatre modalités. En ACM, deux individus i et i' qui ont choisi la première modalité auront comme coordonnées (à Q près) sur les 4 axes associés à cette question :

individus i et i'	1.00	0.00	0.00	0.00
poulation totale	0.25	0.25	0.25	0.25
individus i et i' centrés	0.75	- 0.25	- 0.25	- 0.25

Si la sous-population enquêtée à la même date que i a choisi systématiquement cette première modalité, alors, en ACM conditionnelle, les coordonnées de cet individu sur ces 4 axes seront nulles.

population de i	1.00	0.00	0.00	0.00
individu i	0.00	0.00	0.00	0.00

Par contre, si la sous-population de i' se répartit dans les 4 modalités suivant les pourcentages 0.1 0.0 0.4 0.5, les coordonnées de i' sur ces 4 axes (à Q près) seront :

population de i'	0.10	0.00	0.40	0.50
individu i'	0.90	0.00	- 0.40	- 0.50

Il est clair qu'un individu est d'autant plus éloigné de l'origine qu'il diffère de la sous-population enquêtée au même temps t ; plus précisément, en ACM conditionnelle, on trouvera aux extrémités des axes les individus qui ont choisi des réponses rarement choisies par leur sous-population. Comme en ACM, du fait de la métrique, les réponses globalement rares éloignent plus de l'origine que les réponses globalement fréquentes.

La distance entre deux individus appartenant à la même sous-population est la même que dans l'ACM : ils sont éloignés si leurs réponses sont différentes surtout si leurs réponses sont des réponses rares.

Pour deux individus de deux sous-nuages distincts, la situation est plus complexe. Pour qu'ils soient très proches deux cas sont possibles pour chaque question.

- Soit ils sont tous deux cohérents avec leur sous-population de référence : ils ont choisi une réponse très fréquente pour elle. A la limite, si chacune des deux sous-populations a choisi exactement la même réponse, quelle que soit cette réponse les coordonnées correspondantes sont nulles (cf. individu i de l'exemple ci-dessus).

- Soit ils sont atypiques de leur sous-population : ils ont choisi une réponse rare dans celle-ci (cf. individu i' de l'exemple ci-dessus). Dans ce cas, pour se ressembler, il faut deux conditions. D'abord que leur réponse soit identique car la coordonnée sur un axe j n'est positive que si l'individu i a choisi la modalité j. Ensuite que les deux sous-populations aient des répartitions de réponses assez proches car ces répartitions interviennent dans les coordonnées négatives. La situation n'est donc pas équivalente à celle des variables numériques où seule la différence avec la population de référence intervient.

Notons que deux individus très proches en ACM peuvent être éloignés en ACM conditionnelle. En effet, une réponse identique j éloigne deux individus si cette réponse correspond à la situation de référence de l'un et non de l'autre (coordonnées toutes nulles pour le premier, fortement positive sur j et négatives sur les autres modalités pour l'autre).

3.4 Recentrage des sous-nuages ou projection orthogonale

L'évolution temporelle se traduit par le déplacement des centres de gravité. Au lieu d'éliminer cette influence en recentrant les sous-nuages, on pourrait songer à l'éliminer en se plaçant dans l'orthogonal du sous-espace engendré par ces centres de gravité. On analyserait alors la projection du nuage sur ce sous-espace. Cette projection élimine toute la dispersion dans la direction de l'évolution temporelle, même celle des sous-nuages. Ceci peut faire perdre une information importante sur les structures à un temps donné, si par malheur la direction de dispersion des sous-nuage correspond à celle de l'évolution globale. Cette solution est donc plus dangereuse que celle de l'ACM conditionnelle.

4. Le nuage des modalités en ACM conditionnelle

4.1 Dualité centrage-projection : le cas des variables numériques

Pour des variables numériques, en ACP, on sait que le centrage des variables, qui se traduit pour le nuage d'individus par un déplacement de l'origine, se traduit pour les variables par une projection sur le sous-espace orthogonal à la première bissectrice. Cette propriété se généralise : le recentrage de T sous-nuages d'individus (caractérisés par les modalités t) se traduit dans le nuage des variables par une projection sur l'orthogonal du sous-espace engendré par les T variables indicatrices des modalités t.

En effet, à la décomposition du nuage d'individus en T sous-nuages correspond dualement la décomposition de l'espace R^I en T sous-espaces orthogonaux, notés R^{It}, engendrés par les axes associés aux individus de chacune des sous-populations. Se restreindre à la sous-population associée à t revient à projeter le nuage de variables sur le sous-espace R^{It}. Le nuage des variables se décompose donc en T composantes orthogonales qui se confondent avec les nuages de variables définis dans les ACP des T sous-populations. En centrant une sous-population I_t on projette le nuage de variables situé dans R^{It} sur la première bissectrice de ce sous-espace. Cette première bissectrice n'est autre que la variable indicatrice de la modalité t. En recentrant les variables sur chacune des sous-populations, on projette chacune desT

En recentrant les variables sur chacune des sous-populations, on projette chacune desT composantes dans R^{It} des variables sur l'orthogonal de l'indicatrice de t. Les variables indicatrices de ces T modalités sont orthogonales entre elles et engendrent un sous espace de R^I de dimension T, notés R^T. Appliquer une ACP à des données recentrées par classe revient donc à étudier une projection du nuage complet sur le sous-espace orthogonal à R^T.

Cette technique de projection n'est pas spécifique de l'analyse "intra". Une étude systématique des techniques qui consistent à étudier les projections de nuages de variables sur des sous-espaces déterminés par un autre tableau de données pour analyser la part de variabilité qui dépend (ou ne dépend pas) de l'autre tableau est faite dans la thèse de Sabatier [Sab 87] qui reprend la terminologie de Rao [Rao 64] "analyses sur variables instrumentales".

Pour des variables qualitatives, la situation est tout à fait analogue, mais son interprétation est plus riche.

4.2 Dualité centrage- projection pour des variables qualitatives

En ACM, le nuage des profils de modalités J est situé dans l'espace R^I. Ce nuage se décompose exactement comme dans le cas des variables numériques suivant T composantes orthogonales. La projection de ce nuage sur R^{It} se confond avec le nuage *non centré* des profils des modalités défini dans l'ACM séparée de la sous-population I_t. La variable indicatrice de la modalité t n'est autre que la "première bissectrice" de ce sous-espace. Dans l'ACM de la sous-population le nuage des profils des modalités est à la fois centré et orthogonal à cette indicatrice. Comme dans le cas de l'ACP, au recentrage des sous-nuages d'individus, correspond la projection du nuage des modalités sur l'orthogonal de R^T, le sous-espace engendré par les indicatrices.

4.3 Définition du nuage des modalités : analyse intra

Calculons la projection du profil $k_{ij}/k_{.j}$ de j, sur R^I. Les indicatrices des modalités t de T étant orthogonales entre elles, cette projection est la somme des projections sur ces indicatrices. La coordonnée de $k_{ij}/k_{.j}$ sur l'indicatrice k_{it} est égale à leur produit scalaire divisé par la norme de k_{it} :

$$\sum_i (k_{ij}\, k_{it} / k_{.j}) / \sum_i (k_{it})^2 = b_{tj} / k_{.j}\, k_{t.}$$

Le carré de la distance entre les projections des profils de j et j' vaut donc :

$$D^2(\text{proj j, proj j'}) = \sum_t \sum_{i \in It} ((b_{jt}/k_j k_t) - (b_{j't}/k_{j'} k_t))^2\, I$$
$$= \sum_t (b_{jt}/k_{.j} - b_{j't}/k_{j'})^2\, (I / k_{t.})$$

C'est exactement la distance du "khi 2" entre les profils de j et j' dans le tableau b_{tj} qui croise T et J. Ce tableau traduit la structure induite par le temps T sur l'ensemble J. Pour éliminer dans le nuage N(J) la part de la distance induite par T, il suffit donc de le projeter sur l'orthogonal de R^T. La colonne j sera alors représentée par le point de coordonnées :

$$(1/k_{.j})(k_{ij} - b_{jt}/k_{t.}) \quad \text{si } i \in I_t$$

A la décomposition de l'inertie du nuage des individus du TDC en part "inter T" et "intra T" correspond donc une décomposition duale du nuage des modalités. Cette décomposition se fait au niveau des carrés des distances entre les profils des modalités. Elle correspond géométriquement aux projections orthogonales du nuage de ces profils sur deux sous espaces de R^I : le sous-espace de dimension T engendré par les indicatrices des modalités de la variable T pour la part inter et son orthogonal pour la part intra.

En notant $D_I(j,j')$ la distance induite par le TDC entre les deux modalités j et j', $D_T(j,j')$ celle qui est induite par le tableau b_{tj} et $D_{I/T}(j,j')$ celle qui est considérée en ACM conditionnelle, on a la relation suivante :

$$D^2_I(j,j') = D^2_T(j,j') + D^2_{I/T}(j,j')$$

Ce que l'on peut encore écrire :

$$D^2_{Totale} = D^2_{Inter} + D^2_{Intra}$$

4.4 Interprétation des distances entre modalités

Citons quelques propriétés des distances entre modalités en les situant par rapport à celles de l'ACM.

- Deux modalités proches dans l'ACM restent proches dans l'ACM conditionnelle.

- Deux modalités éloignées en ACM se rapprochent dans l'ACM conditionnelle si leur distance provient essentiellement du fait qu'elles ont été choisies à des époques différentes. Si, à la même époque, elles ont été choisies par des populations distinctes, elles restent éloignées.

- Une modalité choisie (ou exclue) en bloc par chacune des sous-populations se situera à l'origine des axes en ACM conditionnelle.

- La projection orthogonale conserve les propriétés barycentriques. En ACM conditionnelle, comme en ACM, les modalités d'une même variable sont centrées à l'origine.

5. Analyse des écarts du TDC à un tableau "modèle"

5.1.Généralisation de l'analyse des correspondances

L'AFC classique analyse les écarts entre un tableau de fréquence croisant deux variables qualitatives et un tableau "modèle", défini par le produit des deux marges du tableau de fréquence et qui correspond à une situation d'indépendance totale entre les deux variables. Cette analyse peut être généralisée (cf. Esc 87) à l'écart entre un tableau de données et un tableau "modèle" quelconque.

Pour l'analyse conditionnelle, nous allons considérer un tableau "modèle" du TDC dont les deux marges sont égales à celles du TDC et qui va traduire exactement la part de

dispersion définie par T, i.e. la structure induite par le tableau b_{tj}.

5.2 Le modèle

Le terme de la ligne i (appartenant à la classe t) et de la colonne j du tableau "modèle" s'écrit b_{tj}/k_t. Rappelons que b_{tj} note le nombre d'individus qui ont à la fois la modalité j et la modalité t, (i.e. le terme général du tableau qui croise J et T) et k_t le nombre d'individus qui ont la modalité t (i.e. l'effectif de la classe définie par t).

Dans ce "modèle", toutes les lignes qui correspondent aux individus i de I qui ont la même modalité t de la variable extérieure T sont identiques. Ces lignes représentent les profils de répartition de la classe définie par t dans les différentes modalités de chacune des Q variables. Ce tableau n'est pas disjonctif complet, mais correspond à ce que l'on appelle couramment un "codage flou" : pour chaque ligne, la somme des valeurs qui correspondent aux modalités d'une même variable est toujours 1.

Vérifions que les nuages de lignes et de colonnes définis dans l'AFC de ce tableau se déduisent de ceux qui sont définis dans l'AFC du TDC en éliminant la part de dispersion inter-T.

Tout d'abord, il est clair que la marge sur I de ce tableau est, comme celle du TDC, constamment égale à Q et que sa marge sur J est aussi égale à celle du TDC. Les métriques et les poids considérés dans son AFC, tant dans le nuage des lignes que dans celui des colonnes sont donc identiques à ceux qui sont considérés dans l'AFC du TDC.

Considérons maintenant le nuage des lignes associé au TDC. Dans le nuage des lignes du tableau "modèle", tous les individus ayant la même modalité t sont confondus. Le profil de ces lignes étant celui de la classe définie par t, ces individus sont situés au barycentre de cette classe. Le poids total de ces éléments est proportionnel à l'effectif de la classe. Ce nuage représente donc la part "inter T" de la dispersion du nuage des lignes du TDC.

L'inertie du nuage des colonnes, de par la dualité fondamentale de l'analyse des correspondances, est égale à celle des lignes et donc à l'inertie "inter T". Il est facile de vérifier directement que le nuage des colonnes est identique à celui qui est défini dans l'AFC du tableau b_{tj}. En effet, la distance utilisée en AFC satisfait au principe d'équivalence distributionnelle : si l'on cumule dans un tableau des lignes proportionnelles, la distance entre les profils des colonnes est inchangée. Or, dans le tableau modèle, toutes les lignes qui correspondent à la même modalité t sont identiques ; si on les cumule, le tableau obtenu est exactement le tableau croisant J et T, de terme général b_{tj}.

5.3 Analyse de l'écart à un modèle

Lorsque, comme c'est le cas ici, le tableau et le modèle ont les mêmes marges, pour analyser la différence entre le tableau et le modèle, il suffit d'appliquer un programme classique d'analyse des correspondances au tableau suivant :

Données - Modèle + Produit des marges / effectif total

Les termes de ce tableau ne sont pas forcément positifs, il ne s'agit donc pas exactement d'une AFC, mais le programme et l'interprétation géométrique des résultats restent identiques. Les facteurs sont les projections de nuages sur leurs axes principaux d'inertie ; les

facteurs sur les lignes et les colonnes sont liés par des formules de transition qui expriment la dualité entre les deux nuages. Les deux nuages centrés de l'A.F.C. représentent alors la différence entre les profils des lignes (resp. des colonnes) du tableau et du modèle. Nous allons le vérifier dans notre cas particulier.

Pour analyser la différence entre le TDC et le "modèle", i.e. la part intra-T de la dispersion des nuages d'individus et de modalités, nous appliquons donc un programme d'analyse des correspondances à un tableau de dimension I x J, dont le terme général est :

$$k^*_{ij} = k_{ij} - (b_{jt} / k_t) + (k_{.j} / I) \quad \text{si } i \in I_t$$

Ce tableau a les mêmes marges, sur I et sur J que le tableau disjonctif complet. Les métriques induites sur R^I et R^J sont donc les mêmes que dans l'A.F.C. de ce dernier. Le profil d'une ligne est le point de R^J de coordonnées :

$$(k_{ij}/Q) - (b_{tj} / Qk_t) + (k_{.j} / IQ)$$

Le centre de gravité du nuage des lignes a pour coordonnées $k_{.j} / IQ$. En prenant ce point comme origine, les coordonnées de i sont exactement celles du point du nuage d'individus recentré par classe, introduit au paragraphe 3. Les facteurs sur I de l'A.F.C. du tableau k^*_{ij} sont donc les projections, sur ses axes principaux d'inertie, de ce nuage.

Un calcul analogue sur les colonnes montre que le point j défini dans l'A.F.C. du tableau k^*_{ij} a pour coordonnées :

$$(k_{ij}/k_{.j}) - (b_{tj}/k_{.j}k_t) + (1/I)$$

Soit, avec comme origine le centre de gravité du nuage :

$$(k_{ij}/k_{.j}) - (b_{tj}/k_{.j}k_t)$$

Le nuage des colonnes est donc le nuage de modalités introduit au paragraphe 4. Les facteurs sur J de l'A.F.C. de k^*_{ij} sont ses projections sur ses axes principaux d'inertie.

5.4 Formules de transition

Les facteurs sur I, notés F_s et les facteurs sur J, notés G_s étant issus de l'analyse d'un même tableau k^*_{ij}, ces facteurs sont liés par des formules de transition que nous écrivons et commentons ci-dessous :

(1) $$F_s(i) = (1/\sqrt{\lambda_s}) \sum_j ((k_{ij} / Q) - (b_{jt} / Qk_t)) \; G_s(j) \quad \text{si } i \in I_t$$

(2) $$G_s(j) = (1/\sqrt{\lambda_s}) \sum_t \sum_{i \in It} ((k_{ij}/k_{.j} - (b_{jt} / k_{.j}k_t)) \; F_s(i)$$

Ce ne sont pas des formules barycentriques classiques, les coefficients qui apparaissent dans les parenthèses ne sont pas toujours positifs.

Dans (1), le premier terme du second membre correspond à la formule de transition classique du TDC. C'est à $1/\sqrt{\lambda_s}$ près, le barycentre des modalités choisies par i. Le deuxième terme est, à $1/\sqrt{\lambda_s}$ près, le barycentre des modalités choisies par i la classe t de i L'interprétation de cette formule est analogue à celle de la formule classique en se référant à la moyenne de la classe au lieu de la moyenne générale : un individu est attiré par les modalités qu'il a choisi et ceci d'autant plus que la modalité a été peu choisie dans sa classe. Il est repoussé par celles qu'il n'a pas choisies surtout si sa classe les a généralement choisies.

La formule (2) de transition de F_s vers G_s se simplifie :

$$G_s(j) = (1/\sqrt{\lambda_s}) \{ \sum_i (k_{ij}/k_{.j}) F_s(i) - \sum_t b_{jt}/k_t \sum_{i \in I_t} F_s(i)$$

$$= (1/\sqrt{\lambda_s}) \sum_i (k_{ij}/k_j) F_s(i)$$

En effet, $\sum_{i \in I_t} F_s(i) = 0$ car chaque sous-nuage est centré. C'est donc la formule de transition classique de l'analyse des correspondances multiples. Chaque modalité J est à $(1/\sqrt{\neg_s})$ près, au centre de gravité des individus qui la possèdent

5.5 Aides à l'interprétation

Les interprétations géométriques étant valables, les aides à l'interprétation classiques en analyse factorielle existent en ACM conditionnelle. On obtient donc les qualités de représentation et les contributions à l'inertie des individus et des modalités.

On peut aussi introduire des individus et des variables supplémentaires. Le grand intérêt de ces dernières est le même qu'en ACM : obtenir les projections sur les axes des barycentres de n'importe quelle population puisque la formule (3) est la formule barycentrique de l'ACM

6. Tableau de Burt et conditionnement

6.1 Tableau de Burt

Classiquement, à l'analyse d'un tableau disjonctif complet, on associe l'analyse équivalente du tableau de Burt. Le tableau de Burt est une juxtaposition de sous-tableaux croisant deux variables et les marges de ces tableaux sont proportionnelles à celles du tableau entier. Appliquer une AFC à un tableau de Burt, c'est encore analyser son écart au tableau correspondant à l'indépendance, $k_{.j}k_{j'.}$, produit de ses marges. Ce dernier est aussi la juxtaposition des sous-tableaux correspondant chacun à l'hypothèse d'indépendance des couples de variables. L'analyse du tableau de Burt peut donc s'interpréter comme l'analyse conjointe de l'écart à l'indépendance de tous les couples de variables. Dans cette optique, les tableaux diagonaux qui croisent une variable avec elle même posent un problème : par nature ils diffèrent de l'indépendance. Mais on peut montrer qu'ils n'influent pas sur l'analyse, en ce sens que l'AFC du tableau de Burt est égale (aux valeurs des valeurs propres près) à celle d'un tableau où

les sous-tableaux diagonaux sont remplacés par le produit de leurs marges.

6.2 Analyse de l'écart à un modèle du tableau de Burt

L'équivalence entre l'AFC d'un TDC et celle du tableau de Burt découle d'une équivalence plus générale : celle d'un tableau k_{ij} quelconque et celle du tableau symétrique qui s'en déduit et qui a pour terme général $b_{jj'} = \sum_i k_{ij} k_{ij'} / k_i$ (cf. Benz 73 p. 243). Les facteurs sur J sont les mêmes et les valeurs propres du second sont les carrés de celles du premier.

L'AFC du tableau k^*_{ij} est donc équivalente à l'AFC d'un tableau de même dimension que le tableau de Burt dont le terme général noté $b^*_{jj'}$ s'écrit :

$$b^*_{jj'} = \sum_i k^*_{ij} k^*_{ij'}$$

En effet, la marge sur I de k^*_{ij} est constante et l'AFC de deux tableaux proportionnels est la même.

En remplaçant k^*_{ij} par sa valeur en fonction de k^*_{ij}, et en simplifiant les résultats avec les termes qui s'annulent, on obtient finalement :

$$b^*_{jj'} = b_{jj'} - \sum_t (b_{jt} b_{j't} / k_t) + k_j k_{j'} / I$$

Le premier terme est le terme général du tableau de Burt. Le deuxième est le terme général d'un tableau symétrique dont les marges sont égales à celles du tableau de Burt, c'est-à-dire Qk_j. Le dernier terme est, à l'effectif total $Q^2 I$ près, le produit des marges $Qk_j Qk_{j'}$. En divisant par l'effectif total du tableau, on retrouve l'expression :

Données - Modèle + Produit des marges / effectif total

L'analyse des correspondances conditionnelles s'interprète donc comme l'analyse des écarts du tableau de Burt à un modèle.

6.3 Le modèle : Tableau de Burt conditionné

Chacune des sous-populations I_t définit un tableau de Burt. Notons $b^t_{jj'}$ le terme général du tableau défini uniquement par la classe I_t. Considérons le sous-tableau croisant les modalités de deux variables q et q'. Son effectif total vaut k_t.

Si il y avait indépendance entre q et q' pour la population I_t on aurait pour tout j et tout j' l'égalité :

$$b^t_{jj'} / k_t = (b_{jt} / k_t) (b_{j't} / k_t)$$

Si cette condition était réalisée pour tout t, on aurait :

$$\Sigma_t b^t_{jj'} = \Sigma_t (b_{jt} b_{j't}) / k_t$$

Cette expression est moins forte que l'indépendance conditionnelle de q et de q' pour tout t. Nous dirons qu'il y a "indépendance conditionnelle moyenne" lorsque cette dernière égalité est vérifiée pour tout j, j' et t. Le tableau modèle du tableau de Burt en analyse des correspondances multiples conditionnelle est donc une juxtaposition de sous-tableaux (croisant les variables deux à deux) qui correspondent à l'hypothèse d'indépendance conditionnelle moyenne.

Dans le sous-tableau du modèle qui croise deux variables q et q', on retrouve l'expression introduite par Daudin (Dau 81). Il dit que dans ce cas la liaison entre q et q' "passe entièrement par T". Dans le même sous-tableau de la différence *données - modèle - produit des marges* on retrouve le tableau auquel Daudin propose d'appliquer une AFC pour analyser la part de la liaison entre q et q' "qui ne passe par T". L'analyse des correspondances multiples conditionnelle généralise cette étude à plusieurs variables (comme l'ACM généralise l'AFC). C'est une étude simultanée des liaisons binaires entre les Q variables en conditionnant par rapport à t. D'où le nom d'analyse des correspondances multiple conditionnelle.

Le cas de deux variables

On sait (cf. Leb 79) que dans le cas de deux variables seulement, l'analyse des correspondances multiples classique est équivalente à l'AFC du tableau croisant ces deux variables. Qu'en est-il pour l'Analyse des correspondances multiples conditionnelles ? Cette propriété n'est pas vérifiée : l'AFC du sous-tableau n'est pas équivalente à celle du tableau de Burt conditionné, les tableaux diagonaux ont ici une influence dans l'analyse.

Les tableaux diagonaux

Dans le tableau $b^*_{jj'}$, les tableaux qui croisent une variable avec elle-même n'ont généralement pas la forme diagonale. Un simple calcul montre que si on leur applique une AFC, leurs facteurs sont les mêmes que ceux de l'AFC du tableau qui croise la variable en question avec T, mais dans l'ordre inverse. La somme des deux valeurs propres associées au même facteur vaut 1.

Si la variable q est indépendante de T, les valeurs propres de l'analyse de b_{tj} sont nulles, celles du tableau conditionné valent 1 comme en correspondances multiples, le conditionnement ne retire aucune structure au tableau qui reste diagonal.

Etude plus fine du cas de deux variables

Dans le cas de deux variables seulement, il est possible d'étudier l'écart à l'indépendance conditionnelle simple, plus stricte que l'indépendance conditionnelle moyenne. Mais on doit alors considérer T tableaux croisant I et J et analyser un tableau qui les juxtapose. C'est ce que nous avons avons proposé dans [Dro 83]. Notons que dans la juxtaposition, les rôles de I et de J ne sont pas symétriques et que deux "analyses intra" distinctes sont possibles.

8. Analyse multicanonique

L'analyse des correspondances multiples est une analyse multicanonique des Q groupes de variables indicatrices de chaque question. Les facteurs sur I sont les variables générales, elles maximisent la somme des carrés des cosinus des angles avec les sous-espaces engendrés par les variables indicatrices de chaque groupe. Ces cosinus carrés sont égaux aux rapports de corrélation entre le facteur et les Q variables qualitatives. Ils maximisent donc la somme des rapports de corrélation avec les Q questions et résument donc ces variables en rapprochant les individus ayant les mêmes modalités.

On peut montrer que l'analyse des correspondances multiples conditionnelle se traduit aussi en termes d'analyse multicanonique sous contrainte. Les facteurs sur I sont les variables qui satisfont à la contrainte d'avoir une valeur nulle sur les barycentres de chaque classe I_t ou, ce qui est équivalent, d'être orthogonales au sous-espace engendré par les indicatrices des modalités t. Sous cette contrainte, elles maximisent la somme des rapports de corrélation avec les Q variables étudiées. Ce sont donc des variables, centrées sur chaque sous-population, qui résument "au mieux" les Q variables indicatrices.

9. Conclusion

La convergence de plusieurs points de vue enrichit l'interprétation de cette analyse et lui donne cohérence et solidité.

La technique est simple à mettre en oeuvre. On peut appliquer un programme classique d'analyse des correspondances sur l'un ou l'autre des tableaux k^*_{ij} et $b^*_{jj'}$ qui s'obtiennent tous deux par des transformations de tableau. Dans le programme MULCO qui s'intègre dans SPAD, il suffit d'indiquer parmi les variables qualitatives codées sous la forme condensée classique, quelle est parmi les variables, celle qui conditionne les données. Ce programme [Ben 87] admet les données manquantes suivant la technique que nous avons introduit en analyse des correspondances multiples.

REFERENCES

Ben 87 BENALI H. Données manquantes et modalités à faible effectif en analyse des correspondances multiples et conditionnelle. Data analysis and informatics. p. 257--266

Benz 72 BENZECRI J.P. L'analyse des données Dunod

Dau 81 DAUDIN J.J. : Analyse factorielle des dépendances partielles. Revue de statistique appliquée, 1981, vol. XXIX, n° 2.

Dro 83 DROUET D. , ESCOFIER B. Comparaison de plusieurs tableaux de fréquence, Cahiers de l'analyse des données, 1983, n° 3

Esc 83 ESCOFIER B. : Analyse de la différence entre deux mesures définies sur le produit de deux mêmes ensembles. Cahiers de l'analyse des données, 1983, vol. VIII, n° 3

Esc 85 ESCOFIER B. : Analyse factorielle en référence à un modèle. Application au traitement des tableaux d'échanges Revue de statistique appliquée, 1985, n°2

Esc 88 ESCOFIER B. , PAGES J. Analyses factorielles simples et multiples. Dunod 1988

Leb 72 LEBART L., MORINEAU A., TABARD N. Techniques de la description statistiques Dunod 1972

Rao 64 RAO C.R. The use and interpretation of Principal Component Analysis in applied research. Sankhya, A 26,4 p.329-358

Sab 87 SABATIER R. Méthodes factorielles en analyse des données. Approximation et prise en compte de variables concomitantes. Thèse 1987 Université deMontpellier