La revue de Modulad

TRAITEMENT DES VARIABLES INCOMPLETES EN ANALYSE DES CORRESPONDANCES MULTIPLES

Brigitte ESCOFIER

IRISA, Campus de Beaulieu 35042 Rennes Cedex
IUT rue Montaigne 56008 Vannes Cedex

Résumé

Nous présentons ici une technique qui permet de traiter des variables incomplètes en ACM. Le programme MULMD intégré dans SPAD est une variante du programme d'ACM dans lequel on peut mettre en éléments illustratifs des modalités des variables.

L'ensemble des propriétés de l'ACM reposant sur le fait que les variables sont complètes, la méthode proposée n'est pas une ACM au sens strict. Elle en conserve cependant la plupart des qualités : les pertubations entrainées par la suppression de modalités des éléments actifs sont très faibles.

Elle s'applique notamment aux données manquantes (que l'on caractérise alors par une modalité illustrative), aux modalités de faible effectif ou à des modalités particulières dont on veut neutraliser l'influence.

Mots clés :

analyse des correspondances, analyse des correspondances multiples, données manquantes

1.Introduction

1.1 Rappels

L'analyse des correspondances multiples (ACM) est une méthode d'analyse factorielle adaptée aux variables qualitatives. Elle peut être définie comme une analyse des correspondances (AFC) d'un tableau disjonctif complet (TDC) croisant les individus et les modalités.

Nous notons I l'ensemble des individus et leur nombre, Q le nombre de variables qualitatives et J le nombre total de modalités des Q variables qualitatives. L'effectif de la modalité j est noté k_j. Le terme général du tableau disjonctif complet (qui vaut 1 si l'individu i a la modalité j et 0 sinon) est noté k_{ij}. La marge sur I de ce tableau vaut Q et la marge sur J vaut k_j.

L'AFC de ce tableau aboutit à des facteurs sur les individus et des facteurs sur les modalités. Ces facteurs qui permettent une représentation simultanée des deux types d'éléments regroupent les individus qui ont des modalités semblables et les modalités qui sont possédées par les mêmes groupes d'individus.

La structure particulière du TDC induit des propriétés dont les plus importantes en pratique sont les suivantes :

- une modalité est, à une constante près, au barycentre de la population qui la possède et représente la moyenne de cette population.

- les modalités d'une même variable qualitative, qui sont affectées d'un poids égal à leur effectif, ont leur barycentre à l'origine.

- sur un facteur, la somme des inerties des modalités d'une même variable est proportionnelle au rapport de corrélation entre la variable et le facteur (quotient de l'inertie des barycentres des individus qui ont la même modalité par l'inertie du facteur). Plus ce rapport est proche de 1, plus les individus qui ont la même modalité sont regroupés sur le facteur, plus le facteur est lié à la variable qualitative et le représente bien.

- les facteurs sur I sont des variables numériques qui rendent maximum la somme des rapports de corrélation avec toutes les variables. Ceci implique que ces facteurs résument bien la structure définie par toutes les variables qualitatives sur l'ensemble des individus et que l'ACM peut être définie à partir des variables elles-même et pas seulement à partir de leurs modalités.

- les facteurs sur J peuvent être obtenus par une AFC du tableau de Burt . Ce tableau croise l'ensemble des modalités avec lui-même et contient au croisement de la modalité j et de la modalité j' l'effectif $b_{jj'}$ des individus qui ont ces deux modalités.

1.2 Les données incomplètes

Dans certains cas, on souhaite traiter des données incomplètes, c'est-à-dire des tableaux dans lesquels quelques individus n'ont aucune valeur pour une ou plusieurs variables. Le tableau de données n'est plus disjonctif complet, mais *disjonctifs incomplets*. Notons que ce tableau, contrairement au tableau disjonctif complet n'a pas une marge constante sur les individus : sa marge est égale au nombre de variables pour lesquelles l'individu a une valeur bien définie.

Citons trois exemples de cette situation.

- Le premier cas est celui des *données manquantes*.

- Le second est celui des *modalités de très faible effectif*. En effet, en ACM, les modalités rares, lorsqu'elles sont partagées par les mêmes individus, déterminent souvent les premiers facteurs de l'analyse. Ceci est gênant car l'on cherche plutôt des tendances générales. D'où l'intérêt de supprimmer ces modalités, ce qui rend les données incomplètes.

- Un troisième cas, qui recouvre en partie le second, est celui de *modalités qui dominent les facteurs* et peuvent cacher des structures plus fines mais pas forcément moins intéressantes parce que moins évidentes.

1.3 Une méthode de traitement

On souhaite donc traiter des données incomplètes. Or les propriétés de l'AFC du TDC, et donc de l'ACM, reposent sur la structure de données complètes. Il faut donc adapter la méthode pour traiter ce type de données, tout en gardant les principales propriétés qqui en font la richesse.

Pour résoudre ce problème, nous proposons une méthode que nous appelons ACM avec variables incomplètes. Nous allons la présenter à travers les trois types d'éléments qui interviennent en ACM : les individus, les modalités et les variables.

Pour simplifier la présentation, nous allons considérer que nous avons comme données un tableau disjonctif complet et que nous mettons certaines modalités de ce tableau en illustratifs. Ceci correspond au programme MULMD implanté dans SPAD. Il est cependant possible d'appliquer la technique proposée directement sur des tableaux disjonctifs incomplets.

2. Le nuage des individus

Nous allons introduire le nuage d'individus considéré dans l'ACM avec variables incomplètes à partir de celui qui est défini dans l'ACM classique du tableau disjonctif complet.

2.1 Les individus en ACM

En ACM le nuage des individus est situé dans l'espace R^J. Un individu est représenté par le profil de la ligne du tableau disjonctif complet. La marge sur I de ce tableau étant constante pour tous les individus (égale au nombre total Q de variables qualitatives), il est équivalent de considérer les lignes ou leur profil : un individu est caractérisé par les modalités qu'il possède. La distance entre deux individus i et l s'écrit :

$$d^2(i, l) = (1/Q) \Sigma_j (k_{ij} - k_{lj})^2 / k_j$$

Elle croît avec le nombre de modalités qui diffèrent pour ces individus. Une modalité j intervient dans cette distance avec un poids proportionnel à l'inverse de sa fréquence. La présence d'une modalité rare éloigne donc beaucoup plus son possesseur des autres individus que celle d'une modalité fréquente.

Le barycentre de ce nuage a pour coordonnées le profil de la marge $k_{.j}$. La distance d'un individu i au barycentre G du nuage, qui est pris comme origine des axes s'écrit :

$$d^2(i, G) = (1/Q) \Sigma_j (k_{ij} - k_{.j})^2 / k_{.j}$$

La présence d'une modalité rare éloigne son possesseur du barycentre, à la fois parce que son profil s'éloigne beaucoup du profil moyen et parce que l'inverse du poids de cette modalité est élevé.

Prenons un exemple pour concrétiser cela. Supposons qu'une question q a 3 modalités et que la répartition de la population sur ces trois modalités soit 0.8, 0.01 et 0.19. L'individu i a la modalité rare tandis que l'un individu l a la modalité fréquente:

	1	2	3
population	0.80	0.01	0.19
individu i	0.0	1.0	0.0
individu l	1.0	0.0	0.0

La part des distances induite par la question q est :

$$d^2(i, G) = (0.8)^2 / 0.8 + (0.99)^2 / 0.01 + (0.81)^2 / 0.19 + \ldots\ldots$$

$$= 0.8 + 98.01 + 3.45 + \ldots\ldots$$

$$d^2(l, G) = (0.2)^2 / 0.8 + (0.01)^2 / 0.01 + (0.19)^2 / 0.19 + \ldots\ldots$$

$$= 0.05 + 0.01 + 0.19$$

$$d^2(i, l) = (1.0)^2 / 0.8 + (1.0)^2 / 0.01 + (0.00)^2 / 0.19 + \ldots\ldots$$

$$= 1.25 + 100. + 0. + \ldots\ldots$$

2.2 Les individus dans l'ACM de variables incomplètes

Le nuage d'individus se déduit du précédent en supprimmant les coordonnées correspondant aux modalités mises en supplémentaires. L'expression des distances est la même que ci-dessus en se restreignant aux modalités restées actives.

Supposons qu'un individu i possède la modalité j d'une variable q tandis que l'individu l ne possède une autre modalité j'. Si la modalité j est supprimée, la distance de i à l (et aussi à G) est diminuée de la part de la distance induite par i ; mais de la distance entre i et l induite par la variable q reste la part induite par l'autre modalité j' de q. En effet, dans la distance d'un individu i à un individu l, interviennent non seulement les modalités possédées par i et non possédées par l, mais aussi les modalités possédées par l et non par i.

Dans l'exemple précédent, pour i et l, la part de la question q dans les distances se restreint à :

$$d^2(i,G) = 0.8 + 3.45 + \ldots\ldots$$

$$d^2(l,G) = 0.05 + 0.19 + \ldots\ldots$$

$$d^2(i,l) = 1.25 + 0. + \ldots\ldots$$

Ceci résoud donc le problème de modalités rares (ou plus généralement de modalités dont l'influence parait trop importante) : en supprimant une modalité rare on rapproche du barycentre les individus qui la possèdent, tout en gardant une part raisonnable de leur distance au barycentre et aux autres individus.

Dans la technique proposée, l'individu i n'est pas "moyen" pour la variable q, il reste éloigné du barycentre et des autres individus du fait qu'il ne possède pas les autres modalités de q. Pour rendre cet individu moyen pour q, il faudrait lui donner des coordonnées égales à celles du barycentre pour les modalités de q en mettant la répartition de la population dans les différentes modalités de la variable dans les cases correspondantes du TDC. Ce serait une situation analogue à celle qui est choisie pour des données manquantes dans des variables numériques lorsqu'une valeur inconnue est remplacée par la moyenne de la variable, mais ce n'est pas la solution proposée.

Les poids affectés aux individus restent identiques. Le nuage étudié n'est qu'une projection du nuage induit par l'ACM sur le sous-espace engendré par les modalités qui restent actives. La projection conserve les propriétés barycentriques : le nouveau nuage est centré. Comme toujours en analyse factorielle, ce nuage est projeté sur ses axes d'inertie.

Notons que ce nuage n'est pas celui qui est défini dans l'AFC du tableau disjonctif incomplet. En effet, du fait que la marge sur I de ce dernier n'est pas constante, les lignes ne sont pas équivalentes à leur profil et les poids des individus induits par l'AFC ne sont pas égaux. Dans l'AFC du tableau disjonctif incomplet, si deux individus i et l n'ont pas le même nombre de modalités actives, les marges k_i et k_l sont différentes et pour une modalité j commune aux deux individus, les valeurs différentes des profils k_{ij}/k_i et k_{il}/k_l augmentent la distance entre les deux points, ce qui est peu logique. Cet illogisme n'apparaît pas dans la technique proposée : seules les modalités qui appartiennent à un individu sans appartenir à l'autre augmentent la distance entre deux individus.

3. Les modalités

En ACM, le nuage des modalités est situé dans l'espace R^I. Chaque modalité est représentée par son profil et a un poids égal à son effectif. Le barycentre des modalités, pris comme origine des axes a toutes ses coordonnées égales à 1/I.

Le nuage des modalités considéré dans l'ACM des variables incomplètes s'en déduit en suprimmant des points actifs du nuage les modalités mises en illustratives. Les distances entre les modalités sont donc exactement celles de l'ACM, les poids aussi sont identiques. Les coordonnées du barycentre de ce nuage sont données par la marge sur I du tableau incomplet ; généralement cette marge n'est pas très différente de celle du tableau complet. L'analyse du

nuage, c'est-à-dire sa projection sur ses axes d'inertie est faite en gardant comme origine le barycentre du nuage des modalités complet, *le nuage des modalités actives n'est pas exactement centré,* mais son barycentr, noté G, n'est pas très éloigné de l'origine.
Par contre, l'ensemble de toutes les modalités, y compris celles qui sont mises en éléments supplémentaires, reste un nuage centré.

Les facteurs sur les modalités ne sont donc généralement pas centrés non plus (pour les modalités actives). Dans le programme MULMD, la projection de G sur les facteurs est systématiquement calculée. Sur le facteur d'ordre s, G_s elle vaut :

$$G_s(G) = \sum_{j \in \text{modalités actives}} (k_j / IQ)\, G_s(j)$$

Ceci permet de mesurer le décalage du nuage sur chacun des axes. Ce décalage, comme nous l'avons déja signalé est, en pratique, très faible. Il ne peut être que plus faible que dans le nuage complet, car cette différence n'intervient pas activement dans l'analyse.

Ces remarques sur le décentrage des modalités actives et le centrage du nuage complet est valable aussi pour l'ensemble des modalités d'une même variable. L'ACM sur variables incomplètes, comme l'ACM classique, oppose donc les modalités d'une même variable.

Notons la différence entre la méthode proposée et l'AFC du tableau disjonctif incomplet : contrairement à cette dernière, dans l'ACM de variables incomplètes, ni la métrique de R^I, ni l'origine des axes ne sont modifiés par la suppression de modalités.

4. Relations de transition entre individus et modalités

Notons respectivement $F_s(i)$ et Gs(j) les valeurs des facteurs d'ordre s sur les individus et les modalités et λ_s la valeur propre associée. Les relations de transition entre les facteurs dans l'ACM de données incomplètes s'écrivent :

$$F_s(i) = (1/\sqrt{\lambda_s}) \sum_j (k_{ij}/Q - k_j/IQ)\, G_s(j)$$

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

La deuxième relation est identique à celle de l'ACM. La propriété (essentielle dans l'interprétation des résultats) qui permet de confondre sur un facteur la notion de modalité et de barycentre de la population qu'elle concerne est conservée.

La première relation diffère de l'ACM par un décalage qui correspond au décentrage du facteur défini sur les modalités. Elle s'écrit aussi :

$$F_s(i) = (1/\sqrt{\lambda_s}) \sum_j (k_{ij}/Q)\,(G_s(j) - \sum_j (k_j/IQ)\, G_s(j))$$

$$F_s(i) = (1/\sqrt{\lambda_s}) \sum_j (k_{ij}/Q)(G_s(j) - G_s(G))$$

Si l'on considère le facteur sur J recentré, la relation est celle de l'ACM : un individu est, à $1/\sqrt{\lambda_s}$ près, au barycentre des modalités qu'il possède. Le décalage du barycentre étant la plupart du temps négligeable, l'interprétation des résultats est analogue à celui de l'ACM.

Les deux formules de transition sont utilisées pour la *projection de lignes et de modalités supplémentaires*. Les modalités supplémentaires sont largement utilisées, comme en ACM, pour représenter les moyennes des populations. Dans le programme MULMD, les modalités supplémentaires qui dérivent du tableau disjonctif complet sont projetées automatiquement. Il est aussi possible d'introduire des variables supplémentaires dont toutes les modalités sont projetées ; mais il n'est pas prévu d'introduire des modalités séparées à cause de la structure des fichiers qui ont pour base les variables et non les modalités.

5. Algoritmes et démonstrations

Nous donnons ici le principe des calculs et démontrons les formules indiquées dans le paragraphe 4. Pour cela nous nous plaçons dans le cadre général d'une analyse factorielle (cf. Esc 88) que nous rappellons d'abord.

L'ensemble des données est constitué de 3 tableaux :

- un tableau X de dimension I et J. Les lignes (resp. les colonnes) de ce tableau sont exactement les coordonnées du nuage analysé.

- deux tableaux diagonaux notés D et M contenant respectivement les poids des lignes et des colonnes. Ces poids définissent à la fois les métriques des espaces R^I et R^J dans lesquels sont situés les nuages et les poids affectés dans le calcul des axes d'inertie.

Les facteurs sur I sont les vecteurs propres de la matrice XMX'D et les facteurs sur J ceux de la matrice X'DXM. Les valeurs propres de ces deux matrices sont égales. Les facteurs sur I et J sont liés par les relations de transition qui s'écrivent matriciellement :

$$F_s = (1/\sqrt{\lambda_s})\ XM\,G_s$$

$$G_s = (1/\sqrt{\lambda_s})\ X'D\,F_s$$

L'ACM rentre dans ce cadre, en prenant comme matrices X, D et M les matrices de terme général :

$$X_{ij} = \left(I\,k_{ij}/k_j - 1\right)$$

$$M_{jj'} = k_{.j} / I\,Q \qquad \text{et} \qquad D_{ii'} = 1/I$$

Pour vérifier directement cette assertion, on peut d'abord voir que les deux nuages de lignes et de colonnes définis par ces tableaux sont centrés :

$$\sum_i \left((k_{ij} / k_{.j}) - 1/I \right) = 0 \qquad \text{et} \qquad \sum_j \left((k_{ij} / Q) - (k_{.j} / I\,Q) \right) = 0$$

On peut ensuite écrire les formules de transition induites par les matrices XM et X'D. Le terme général de XM est $k_{ij}/Q - k_{.j}/IQ$ et celui de X'D est $k_{ij}/k_{.j} - 1/I$. Les formules de transition qui en découlent s'écrivent :

$$F_s(i) = (1/\sqrt{\lambda_s}) \sum_j \left(k_{ij}/Q - k_{.j}/I\,Q \right) G_s(j) = (1/\sqrt{\lambda_s}) \sum_j (k_{ij}/Q)\, G_s(j)$$

$$G_s(j) = (1/\sqrt{\lambda_s}) \sum_i \left(k_{ij}/k_{.j} - 1/I \right) F_s(i) = (1/\sqrt{\lambda_s}) \sum_i (k_{ij}/k_{.j})\, F_s(i)$$

Du fait du centrage des deux nuages, le second terme des deux formules est nul, et l'on retrouve exactement les formules de transition classique de l'ACM. Ces formules sont suffisantes pour déterminer les facteurs.

L'ACM de variables incomplètes se définit exactement de la même façon en se restreignant dans X et M à l'ensemble des modalités actives.

Le nuage des colonnes définis par ces matrices se confond bien évidemment avec le sous-nuage des profils des colonnes du TDC puisque coordonnées, poids et métrique sont identiques.

Le nuage des lignes défini par ces matrices est la projection du nuage des profils du tableau disjonctif complet sur le sous-espace engendré par les axes associés aux modalités actives. Ce nuage, comme celui de l'ACM est centré.

Les formules de transition déduites de XM et X'D sont identiques à celles de l'ACM restreintes aux modalités actives. Mais, comme seul le nuage des individus est centré, le second terme ne s'annule que dans la formule de G_s vers F_s.

On obtient donc bien les nuages d'individus et de modalités commentés aux paragraphes 2 et 3 et les formules de transition du paragraphe 4.

Lorsqu'il y a des variables incomplètes, il n'y a pas de facteur trivial comme en AFC classique et le rang de cette matrice n'est pas diminué d'une unité par variable comme en ACM. Les facteurs sur J s'obtiennent en diagonalisant la matrice X'DXM. Cette matrice est la restriction à l'ensemble des modalités actives, de la matrice diagonalisée en ACM (quand on n'utilise pas dans cette dernière les simplifications induites par les l'une ou l'autre des propriétés évoquées ci-dessus.

6. Les variables

6.1 Rapport de corrélation

La projection d'une modalité j étant, à $1/\sqrt{\lambda_s}$ près, au barycentre de la population qui la possède, son inertie est, à $\sqrt{\lambda_s}$ près, celle de ce barycentre affecté du poids de la population qu'il représente. Si l'on considère l'ensemble des modalités d'une variable, y compris éventuellement les modalités illustratives, la somme de leur inertie est proportionnelle au rapport de corrélation entre cette variable et le facteur. Cet indicateur de liaison, pratique pour orienter le dépouillement des résultats, se calcule donc aussi facilement qu'en ACM.

6.2 Caractérisation des facteurs par les variables

Nous allons montrer ici que, comme en ACM, les facteurs sur I peuvent se définir directement par les variables et pas seulement par leur modalités. Dans l'espace R^I, muni de la métrique 1/I, une modalité J est représentée par le point de coordonnées $k_{ij}/k_{.j}$ -1. Sa projection sur une variable normée f vaut :

$$\textit{projection du profil de j sur f} = (1/I) \sum_i \left(k_{ij}/k_{.j} - 1/I \right) f_s(i)$$

Si f est centrée, le deuxième terme s'annule. En notant G_j le barycentre des individus qui ont la modalité j, cette projection s'écrit :

$$= (1/I)\, f(G_j)$$

Le premier axe de l'analyse rend maximum l'inertie projetée du nuage des profils des modalité. C'est donc une variable centrée-réduite telle que la somme des inerties des barycentres des populations associées à chaque modalité actives soit maximum.

$$\sum_q \sum_{j \in \text{modalités actives de } q} k_{.j} \left(f(G_j) \right)^2 \qquad \textit{maximum pour une inertie totale fixée}$$

Pour une variable q, l'expression qui intervient est la part de l'inertie inter qui concerne uniquement les modalités actives, appellons la "inertie inter incomplète".

Un facteur de l'ACM sur variables incomplètes peut donc être défini comme une variable centrée qui rend maximum la somme (sur toutes les variables qualitatives) du quotient de l'inertie inter incomplète par l'inertie totale. Il résume donc l'ensemble de ces variables, en éloignant les individus qui ont des modalités différentes et en regroupant les individus qui ont la même modalité, sans tenir compte des individus qui n'ont aucune modalité.

On peut montrer aussi [Esc 87], que cette technique est une analyse multicanonique sous

contrainte.

7. Tableau de Burt

En ACM classique, les facteurs sur J peuvent être obtenus par une AFC du tableau de Burt. Ceci montre que la structure de proximité mise en évidence par l'analyse traduit aussi le fait que les profils de ces modalités dans le tableau de Burt se ressemblent. Or le profil de j dans le tableau de Burt ne représente rien d'autre que les répartitions de la population caractérisée par j dans les modalités de chacune des Q variables : deux modalités sont proches si elles s'associent de la même façon à l'ensemble des variables.

Cette propriété demeure dans l'ACM sur variables incomplètes. On peut montrer que les facteurs obtenus sont identiques à ceux d'une analyse du sous-tableau de Burt réduit aux modalités actives. Mais, comme pour le tableau disjonctif complet, cette analyse n'est pas exactement une AFC. Une AFC de ce tableau n'est d'ailleurs pas une analyse des profils de répartition des populations car les marges de ce sous-tableau ne sont pas proportionnelles aux effectifs des modalités. L'analyse modifiée permet d'avoir une représentation de ces répartitions en excluant les modalités illustratives. Précisons rapidement le principe d'une démonstration.

En se plaçant dans le cadre de l'analyse générale, notons Y la matrice (symétrique) des coordonnées associée au tableau de Burt. On a $Y = X'DX$. Les matrices des poids sont toutes deux égales à M. En restreignant Y et M aux modalités actives, on définit une analyse différente de l'AFC. On a les mêmes relations entre cette analyse et celle de l'ACM sur variables incomplètes qu'entre l'ACM et l'AFC du tableau de Burt.

8.Compléments pratiques

Pour conclure, nous allons donner quelques indications sur le programme implanté dans SPAD et donner quelques résultats sur son efficacité réelle.

8.1 Le programme MULMD

Ce programme est écrit à partir de l'étape MULTC qu'il peut remplacer complètement. En effet, dans le cas de données complètes, la méthode proposée se confond bien heureusement avec l'ACM classique. Les possibilités de MULTC sont enrichies et le traitement des modalités de faible effectif est modifié.

Le tableau de données est donc sous la forme classique du codage condensé pour les variables qualitatives.

Les modalités mises en supplémentaires peuvent être :

- les modalités de faible effectif définies, comme dans MULTC, par un seuil de pourcentage précisé par l'utilisateur.

- des modalités particulières de certaines variables indiquées par l'utilisateur.

-une modalité par variable, celle qui correspond au numéro le plus élevé (pratique si les données manquantes ont été codées par ce numéro).

La matrice à diagonaliser se déduit du sous-tableau de Burt réduit aux modalités actives. L'ensemble des modalités actives d'une variable n'étant pas toujours centré, le rang de la matrice à diagonaliser n'est pas systématiquement diminué d'une unité par variable. L'économie qui résulte de l'exploitation de cette propriété par MULTC, ne peut être reproduite ici.

8.2 Stabilité de la méthode

Une étude a été faite dans [Ben 85] pour tester la stabilité de cette méthode et son efficacité à traiter les données manquantes et la suppression des modalités de faible effectif en l'étudiant sur des fichiers de données réels.

Pour les modalités de faible effectif, la stabilité a été étudiée en calculant les corrélations entre les facteurs obtenus pour plusieurs valeurs de l'effectif minimum de suppression de modalités. Dans les exemples traités, les corrélations entre les facteurs obtenus pour les seuils de 2 % et 3 % sont assez élevées. Les résultats ont été comparés à ceux obtenus par le programme MULTC qui remplace aléatoirement la modalité de faible effectif par une autre modalité. Les facteurs obtenus par cette dernière sont beaucoup plus sensibles au choix du seuil que ceux de la méthode que nous proposons.

Pour tester la stabilité vis à vis des données manquantes, des valeurs manquantes ont été générées aléatoirement dans des fichiers de données (de 5 % à 20 % des données sur certaines variables, puis 10 % dans l'ensemble des variables). Techniquement, une modalité particulière a été crée pour les variables concernées et déclarée illustrative. Les corrélations entre les facteurs obtenus pour ces différents pourcentage et les facteurs du tableau complet sont très élevées pour les premiers facteurs, même dans le cas de 20 % de données manquantes. La pratique montre donc l'efficacité de cette méthode pour le traitement des données manquantes aléatoires, même très nombreuses.

Ben 85 Benali H. Stabilité devl'analyse en composantes principales et de l'analyse des correspondances multiples en présence de certains types de pertubations. Thése 3° cycle Rennes 1985

Esc 88 Escofier B., Pagès J Analyses factorielles simples et multiples. Objectifs, méthodes et interprétation Dunod 1988

Esc 87 Escofier B Traitement des questionnaires avec non réponse, analyse des correspondances avec marge modifiée et analyse multicanonique avec contrainte. Pub Inst. Stat Univ. XXXII, fasc 3, 1987, 33 --69