Analyse des correspondances multiples

L'analyse des correspondances multiples (ACM) peut être présentée comme un prolongement de l'AFC. Elle permet la representation graphique de tableaux de fréquences contnant plus de deux variables. Un exemple classique d'un tableau de fréquences avec plus de deux variables qualitatives est un tableau présentant les réponses d'individus à un questionnaire contenant Q questions à choix multiples. L'ACM est donc très utile pour visualiser les résultats d'une étude par questionnaire.

L'ACM peut aussi être vue comme une version de l'ACP quand les variables sont mixtes, i.e. comprenant à la fois des variables quantitatives et des variables qualitatives. Le traitement conjoint de ces deux types de données repose sur leur transformation préalable appelée **codage disjonctif complet**.

1 Notation

Notons n le nombre d'individus (ou d'observations) et Q le nombre de variables (ou de questions dans le cas d'un questionnaire). Chaque variable possède J_q modalités et le nombre total de modalités est égal à J.

> Définition
>
> Le tableau binaire, i.e. ne contenant que des 0 et des 1, à n lignes et J colonnes est appelé **tableau de codage disjonctif complet**. On le note Z.

Ainsi, une variable n'est pas traitée telle quelle mais à travers ses modalités. Elle est découpée en modalités et tout individus est alors codé 1 pour la modalité qu'il possède et 0 dans les autres (i.e. qu'il ne possède pas, les modalités étant exclusives). Ce codage est immédiat pour des variables qualitatives. Cependant, pour une variable qualitative, on procède en découpant au préalable la variable en classes. Ainsi, chaque individu n'appartient qu'à une seule classe. Ce processus de transformation de l'information est appelé **codage disjonctif complet**. Il s'agit bien d'un codage, car l'information initiale est transformée, disjonctif, car tout individu possède au plus une modalité, et complet, car tout individu a au moins une modalité.

Exemple

Prenons par exemple un ensemble de produits avec différents types (Hoodie, Joggers et Sneakers) et différents prix. On a 7 observations (7 produits). La variable `Type` est une variable qualitative et la variable `Prix` est une variable quantitative.

Table 1 : Jeu de données de produits.

Produit	Type	Prix ($)
Nike Tech Fleece	Hoodie	256.72
Puma Joggers	Joggers	221.26
Off-White Hoodie	Hoodie	198.45
Supreme Hoodie	Hoodie	235.50
Jordan 1 High	Sneakers	298.22
Nike Dunk Low	Sneakers	273.00
Nike Tech Fleece	Hoodie	162.38

Pour coder l'information en tableau de codage disjonctif complet, on définit trois classes de prix (prix inférieur à 200$, prix compris entre 200$ et 250$ et prix supérieur à 250$). Ainsi, on peut encoder la variable `Prix` grâce aux classes précédentes. Ainsi, le tableau disjonctif complet est donné par la tableau suivant.

Table 2 : Jeu de données de produits en codage disjonctif complet.

Produit	Hoodie	Joggers	Sneakers	$< 200$$	entre 200$ et 250$	$> 250$$
Nike Tech Fleece	1	0	0	0	0	1
Puma Joggers	0	1	0	0	1	1
Off-White Hoodie	1	0	0	1	0	0
Supreme Hoodie	1	0	0	0	1	0
Jordan 1 High	0	0	1	0	0	1
Nike Dunk Low	0	0	1	0	0	1
Nike Tech Fleece	1	0	0	1	0	0

Remarque

Lorsque l'on veut transformer des variables quantitatives en tableau de codage disjonctif complet, on perd de l'information. En effet, comme on doit découper les variables qualitatives en classes, l'appartenance à une classe est moins informatif qu'une valeur précise d'une variable. Dans l'exemple précédent, on perd de l'information sur le prix.

Propriétés

1. La somme des éléments d'une même ligne est constante et vaut Q.
2. La somme de tous les éléments du tableau est égale à nQ.
3. La somme des éléments d'une même colonne est égale à l'effectif n_j possédant la modalité j de la variable q.

Preuve

Comme Z est un tableau disjonctif complet, on a

$$\sum_{j=1}^{J_q} z_{ij} = 1.$$

Donc, on trouve que

$$z_{i\bullet} = \sum_{j=1}^{J} z_{ij} = \sum_{q=1}^{Q} \sum_{j=1}^{J_q} z_{ij} = Q,$$

$$z_{\bullet j} = \sum_{i=1}^{n} z_{ij} = n_j,$$

$$z_{\bullet\bullet} = \sum_{i=1}^{n} \sum_{j=1}^{J} z_{ij} = nQ.$$

2 Tableau de Burt

Définition

Le **tableau de Burt**, noté B, est le produit de la transposé de Z par Z :

$$B = Z^\top Z.$$

Propriétés de B

1. Le tableau de Burt est carré et sa taille est égale au nombre total de modalités J possédées par les Q variables.
2. Les blocs diagonaux de B sont eux-mêmes des matrices diagonales. Ils sont donnés par $B_q q = Z_q^\top Z_q$ et leurs éléments diagonaux correspondent à l'effectif de chaque modalité pour la variable q.
3. Les blocs non-diagonaux de B sont donnés par $B_{qq'} = Z_q^\top Z_{q'}, q \neq q'$. Ils correspondent aux tableaux de contingence croisant les variables q et q'.
4. Le tableau de Burt est symétrique car $B_{q'q} = Z_{q'}^\top Z_q$ est la transposé de $B_{qq'} = Z_q^\top Z_{q'}$.

D'un point de vue mathématique, l'ACM est une AFC effectuée sur la matrice logique Z ou sur le tableau de Burt B. On peut démontrer que l'on obtient les mêmes facteurs, et ce, peu importe la matrice utilisé pour l'analyse.

3 Éléments propres du tableau Z

On peut calculer les éléments propres du tableau Z en utilisant la même méthode que pour l'AFC. Par analogie avec l'AFC, on cherche donc les vecteurs propres de la matrice

$$S = \frac{1}{Q} Z^\top Z D_J^{-1},$$

où D_J est la matrice diagonale de terme $n_j, j = 1, \dots, J$. On peut calculer de la même façon les coordonnées des profils-lignes sur les axes factoriels :

$$\Phi_k = nZD_J^{-1}u_k,$$

où u_k est le ke vecteur propre associé à la valeur propre λ_k de la matrice S.

On peut aussi s'intéresser à l'analyse duale du tableau Z. Toujours par analogie avec l'AFC, on cherche les vecteurs propres de la matrice

$$T = \frac{1}{Q} Z D_J^{-1} Z^\top.$$

De même, on peut calculer les coordonnées des profils-colonnes sur les axes factoriels :

$$\Psi_k = nD_J^{-1} Z^\top v_k.$$

4 Éléments propres du tableau de Burt B

Le tableau de Burt étant symétrique, l'analyse direct et l'analyse duale coïcident. On peut aussi l'analyse en analogie avec l'AFC. La somme des éléments d'une même ligne (ou d'une même colonne) de B vaut Qn_j et la somme des éléments de B est nQ^2. On cherche les vecteurs propres de la matrice

$$S' = \frac{1}{Q^2} B^\top D_J^{-1} B D_J^{-1}.$$

On remarque alors que cette matrice S' a les mêmes vecteurs propres que la matrice S. En effet,

$$S' = \frac{1}{Q^2} B^\top D_J^{-1} B D_J^{-1} = \frac{1}{Q^2} Z^\top Z D_J^{-1} Z^\top Z D_J^{-1}.$$

Et soit u et λ vérifiant $Z^\top Z D_J^{-1} = \lambda u$, alors

$$Z^\top Z D_J^{-1} Z^\top Z D_J^{-1} u = Z^\top Z D_J^{-1} \lambda u = \lambda^2 u.$$

Finalement, l'analyse de Z ou de B fournit les mêmes vecteurs propres et pour tout $k = 1, \dots, Q$, la ke valeur propre de B est la carré de la ke valeur propre de Z.

5 L'encodage des variables

L'encodage des variables, et en particulier le choix des bornes des classes, est primordiale en ACM. Pour les variables continues, les bornes devraient être pertinentes au regard du problème étudié. Par exemple, on ne va pas définir une classe > 1000\$ dans l'exemple précédent. Pour obtenir des bornes pertinentes, on peut regarder les distributions des variables, e.g. avec un histogramme. Dans certains cas particuliers, il est possible de découper la variable en modalités d'effectifs égaux. Cependant, cette approche peut conduire à des modalités peu pertinentes.

Dans le cas de variables qualitatives, le choix des classes ne se pose pas ; il est donné par la variable. Cependant, les modalités "naturelles" peuvent conduire à des effectifs (très) déséquilibrés. Dans ce cas, on doit généralement procéder à des regroupements. Ici encore, une bonne connaissance du domaine étudiée est nécessaire. En tout cas, on préferera faire des regroupements de modalités, plutôt que répartir de manière aléatoire les modalités à effectif faible dans les autres modalités (ce qui est parfois proposé dans les logiciels).