

L'ANALYSE DES LIAISONS : UNE NOUVELLE MÉTHODE DE FOUILLE DE DONNÉES MULTIDIMENSIONNELLES

Jean-Luc Durand ¹

¹ *LEEC, Université Paris 13, 99 avenue J.-B. Clément, 93430 Villetaneuse
jean-luc.durand@univ-paris13.fr*

Résumé. L'une des propriétés de l'Analyse des Correspondances (AC) est que la variance des nuages de points (le carré moyen de contingence ϕ^2) fournit une mesure de la liaison entre deux variables qualitatives. Ainsi, dans une AC, une proportion de variance peut être interprétée comme une proportion de la force de cette liaison. Mais une telle interprétation de la variance n'est pas possible en Analyse en Composantes Principales (ACP) ou en Analyse des Correspondances Multiples (ACM). L'objectif de cette présentation est de proposer une méthode d'analyse de données multivariées qui permette d'interpréter la variance du nuage des individus comme une mesure globale des liaisons entre variables.

On définit l'indice de liaisons d'une variable comme une mesure, comprise entre 0 et 1, de la force des liaisons entre cette variable et les autres variables, basée sur les carrés des corrélations (r^2), les carrés des rapports de corrélation (η^2) ou les carrés moyens de contingence (ϕ^2) entre cette variable et les autres. Ensuite on utilise les indices de liaisons et, pour chaque variable qualitative, les résultats de l'AC d'une partie du tableau de Burt, pour construire un tableau numérique Individus $\times$ Variables (appelé tableau de liaisons) donnant les coordonnées brutes des points d'un nuage euclidien représentant les individus. La variance de ce nuage est une somme pondérée des indices de liaisons des variables. Enfin, les coordonnées principales des individus sont obtenues en réalisant l'ACP non normée du tableau de liaisons.

Comparés à ceux de l'ACP ou de l'ACM, les résultats de l'analyse des liaisons montrent :

- 1) une robustesse vis-à-vis de la présence de variables non pertinentes : l'ajout de variables aléatoires ne change pratiquement pas les résultats ;
- 2) pour des données quantitatives, une interprétation similaire des premiers axes ;
- 3) pour des données qualitatives, une interprétation similaire du premier axe, une meilleure prise en compte de l'information contenue dans les tableaux binaires croisant deux variables différentes, et des proportions de variance nettement plus importantes pour les premiers axes.

Mots-clés. Fouille de données, ACP, ACM.

Abstract. One of the properties of Correspondence Analysis (CA) is that the variance of the clouds of points (the mean square contingency coefficient ϕ^2) provides a measure of the relationship between two categorical variables. So, in CA, proportions of that variance can be interpreted as proportions of the strength of this relationship. But this

interpretation of the variance does not hold in Principal Components Analysis (PCA) or in Multiple Correspondence Analysis (MCA). The goal of this presentation is to propose a multivariate data analysis method that allows to interpret the variance of the cloud of individuals as a measure of global links between variables.

The linkage index of a variable is a measure, from 0 to 1, of the strength of the relations between this variable and the other variables, based on squared correlations (r^2), squared correlation ratios (η^2) and mean square contingency coefficients (ϕ^2). Then, linkage indexes and, for each qualitative variable, results of CA of a part of the Burt matrix are used to construct a numeric Individuals $\times$ Variables table (called Linkage table) giving the raw coordinates of an Euclidean cloud of points representing individuals. The variance of this cloud is a weighted sum of linkage indexes of variables. The principal coordinates of individuals are given by the covariance PCA of the Linkage table.

Compared to those of PCA or MCA, results of Linkage Analysis show :

- 1) robustness against non pertinent variables : adding random variables does not change the results ;
- 2) for quantitative data, similar interpretation of the first axes ;
- 3) for qualitative data, similar interpretation of the first axes, better consideration of the information contained in the binary tables crossing two different variables, and greater proportion of variance accounted for by first axes.

Keywords. Data mining, PCA, MCA.

1 Introduction

La conception de la méthode d'analyse de données multidimensionnelles présentée ci-dessous est motivée par l'absence de méthode fournissant à la fois :

- un nuage des individus, comme le fait l'analyse en composantes principales (ACP) pour des données quantitatives, l'analyse des correspondances multiples (ACM) pour des données qualitatives, ou l'analyse factorielle de données mixtes (AFDM) pour un mélange de variables quantitatives et qualitatives (Pages, 2002) ;
- une interprétation de la variance du nuage comme mesure de la force des liaisons entre variables, comme le fait l'analyse des correspondances (AC) d'un tableau de contingence d'effectifs ;
- et, pour les variables qualitatives, une bonne prise en compte des distances entre les profils des catégories d'une même variable, comme le fait l'AC, mais ce que ne fait pas l'ACM. En effet, dans une ACM, et c'est là son point faible, les distances entre les profils des catégories d'une variable sont prises dans les blocs diagonaux du tableau de Burt alors que l'information pertinente se trouve dans les blocs non diagonaux (Greenacre, 2017). L'ACM perd ainsi une part précieuse de l'information sur les liaisons entre variables.

2 Mesures standardisées des liaisons entre variables

Étant donné un ensemble V de variables (quantitatives et/ou qualitatives) définies sur un ensemble I d'individus, nous définirons deux statistiques descriptives standardisées : l'indice de liaisons d'une variable et l'indice global de liaisons entre les variables. Ces indices sont basés sur trois statistiques traditionnelles de mesure de la liaison entre deux variables v et v' : le carré du coefficient de corrélation $r_{vv'}^2$, le carré moyen de contingence $\phi_{vv'}^2$, et le carré du rapport de corrélation $\eta_{vv'}^2$.

Définition. Appelons *carré de liaison* entre les variables v et v' , et notons $L_{vv'}^2$, l'une ou l'autre des statistiques $r_{vv'}^2$, $\phi_{vv'}^2$ ou $\eta_{vv'}^2$, selon la nature des variables :

$$L_{vv'}^2 = \begin{cases} r_{vv'}^2 & \text{si } v \text{ et } v' \text{ sont quantitatives} \\ \phi_{vv'}^2 & \text{si } v \text{ et } v' \text{ sont qualitatives} \\ \eta_{vv'}^2 & \text{si l'une est quantitative et l'autre qualitative.} \end{cases}$$

2.1 Indice de liaisons d'une variable

Définition. On appelle *indice de liaisons* (*Linkage Index*) d'une variable v , le quotient, notée LI_v , de la moyenne des carrés de liaison entre la variable v et les autres variables par le carré de liaison entre la variable v et elle-même.

$$LI_v = \frac{\frac{1}{V-1} \sum_{v' \neq v} L_{vv'}^2}{L_{vv}^2}$$

L'indice de liaisons d'une variable est un indicateur direct, compris entre 0 et 1, de la force moyenne des liaisons entre cette variable et les autres variables.

2.2 Indice global de liaisons

Définition. On appelle *indice global de liaisons* (*Global Linkage Index*), le quotient, notée GLI , de la moyenne des carrés de liaison entre deux variables différentes par la moyenne des carrés de liaison entre chaque variable v et elle-même.

$$GLI = \frac{\frac{1}{V(V-1)} \sum_{v' \neq v} L_{vv'}^2}{\frac{1}{V} \sum_v L_{vv}^2}$$

L'indice global de liaisons est un indicateur direct, compris entre 0 et 1, de la force moyenne des liaisons entre toutes les variables. Étant donnée la matrice des carrés de liaisons, l'indice global de liaisons est égal au quotient de la moyenne des termes non diagonaux par la moyenne des termes diagonaux.

Définition. On appelle *poids* d'une variable v , le nombre entier, noté w_v , égal au carré de liaison entre cette variable et elle-même.

$$w_v = L_{vv}^2 = \begin{cases} 1 & \text{si } v \text{ est quantitative} \\ K_q - 1 & \text{si } v \text{ est qualitative avec } K_q \text{ catégories.} \end{cases}$$

Propriété. L'indice global de liaisons est la moyenne pondérée des indices de liaisons des variables.

$$GLI = \frac{\sum_v w_v LI_v}{\sum_v w_v}$$

L'indice global de liaisons d'un tableau multivarié détermine la dispersion des valeurs propres de l'ACP normée, de l'ACM ou de l'AFMD de ce tableau. Dans chacune de ces trois méthodes, on a la relation :

$$CV^2(\lambda) = (V - 1) GLI$$

où $CV^2(\lambda)$ est le coefficient de variation des valeurs propres et V est le nombre de variables. Cette relation généralise et simplifie les expressions de la variance des valeurs propres en ACP normée et en ACM données dans Durand (1998) et Durand et Le Roux (2017).

Nous présentons maintenant la méthode d'analyse des liaisons dans le cas de variables quantitatives, puis dans le cas de variables qualitatives.

3 Analyse des liaisons et variables quantitatives

La méthode se déroule en deux étapes. On commence par construire un tableau Individus $\times$ Variables, appelé tableau de liaisons, donnant les coordonnées brutes des points du nuage des individus dans un repère orthonormé centré sur le point moyen du nuage. On détermine ensuite les coordonnées principales des individus en réalisant une ACP non normée de ce tableau. Nous présenterons enfin la comparaison des résultats de l'ACP avec ceux de l'analyse des liaisons.

3.1 Tableau de liaisons dans le cas de variables quantitatives

Notons P le nombre de variables quantitatives et n le nombre d'individus. Pour obtenir le tableau de liaisons, chaque variable est centrée et réduite, puis multipliée par la racine carrée de son indice de liaisons. L'ACP non normée du tableau de liaisons donne un nuage des individus de dimensionnalité égale à P .

En notant M^i le point représentant un individu i et G le centre de gravité du nuage, le carré de la distance entre un individu et le centre du nuage est $(GM^i)^2 = \sum_p (z_p^i)^2 LI_p$ où $z_p^i = (x_p^i - \bar{x}_p)/\sigma_p$ est l'écart réduit de l'individu i pour la variable p .

Le carré de la distance entre deux individus i et i' est $(M^i M^{i'})^2 = \sum_p (z_p^i - z_p^{i'})^2 LI_p$.

La variance du nuage est $\text{Var}(M^I) = \frac{1}{n} \sum_i (GM^i)^2 = \sum_p LI_p = P \times GLI$.

La moyenne des P valeurs propres est égale à l'indice global de liaisons.

La contribution absolue d'une variable p à la variance du nuage est égale à LI_p .

Une première conséquence de ces propriétés est que le taux de variance pris en compte par un axe s'interprète comme une proportion de la force globale des liaisons entre les variables. Une seconde conséquence est que cette méthode est robuste vis-à-vis de la présence, parmi les variables initiales, de variables non pertinentes, c'est-à-dire non corrélées, ou très peu corrélées, avec les autres variables. Leurs indices de liaisons étant quasi nuls, ces variables ne créent pas de distance supplémentaire entre les individus et ne modifient pas les résultats de l'analyse, alors qu'en ACP normée, chaque variable non corrélée aux autres crée un axe de valeur propre égale à 1, à la variance duquel elle est seule à contribuer.

3.2 ACP et analyse des liaisons : comparaison des résultats

Sur une dizaine d'exemples pris dans divers manuels d'analyse des données, on peut observer que lorsque les données ne contiennent pas de variables peu corrélées aux autres, les résultats de l'ACP et ceux l'analyse des liaisons sont très proches. Les taux de variance des premiers axes sont légèrement plus grands avec l'analyse des liaisons, et les variables principales sont très fortement corrélées, axe par axe, d'où des interprétations analogues.

4 Analyse des liaisons et variables qualitatives

La construction du tableau de liaisons est précédée, pour chacune des variables, d'une analyse des correspondances donnant les coordonnées principales de ses catégories.

4.1 Coordonnées principales des catégories d'une variable

On note Q le nombre de variables qualitatives et K le nombre de total de catégories. Pour chaque variable q , à K_q catégories, on commence par réaliser l'analyse des correspondances de la bande de Burt de cette variable, privée de son bloc diagonal ; cette AC donne un nuage des catégories de q ayant pour variance la moyenne, égale à $(K_q - 1)LI_q$, des ϕ^2 calculés entre cette variable et les autres variables. Les $K_q - 1$ variables principales de ce nuage vont être utilisées pour construire le tableau de liaisons.

4.2 Tableau de liaisons dans le cas de variables qualitatives

Le tableau de liaisons est construit en associant à chaque individu, pour chaque variable q à K_q catégories, les $K_q - 1$ coordonnées principales de la catégorie choisie par cet individu. On obtient ainsi un tableau contenant $K - Q$ variables centrées, avec une ligne par individu.

L'ACP non normée du tableau de liaisons donne un nuage des individus de dimensionnalité égale à $K - Q$.

La variance du nuage est égale à la somme pondérée des indices de liaisons des variables, soit $(K - Q)GLI$.

La moyenne des $K - Q$ valeurs propres est égale à l'indice global de liaisons.

La contribution absolue d'une variable à la variance du nuage est égale à son indice de liaisons pondéré, soit $(K_q - 1)LI_q$.

On retrouve les mêmes propriétés que dans le cas de variables quantitatives :

- interprétation des taux de variance comme proportions de la force globale des liaisons entre les variables ;
- robustesse vis-à-vis de la présence éventuelle de variables non liées aux autres, car elles ne créent pas de distance entre les individus.

4.3 ACM et analyse des liaisons : comparaison des résultats

La comparaison, sur une dizaine d'exemples pris dans divers manuels d'analyse des données, entre les résultats de l'ACM et ceux de l'analyse des liaisons appliquée à des variables qualitatives montre en outre les trois propriétés suivantes :

- les proportions de variance expliquées par les premiers axes sont beaucoup plus importantes en analyse des liaisons qu'en ACM ;
- pour chaque variable à plusieurs catégories, le nuage des points moyens des catégories a dans l'espace une forme proche de celle obtenue par l'AC de la bande de Burt de cette variable (privée du bloc diagonal) et très différente de celle donnée par l'ACM : l'analyse des liaisons respecte l'information contenue dans les blocs non diagonaux du tableau de Burt.
- le premier axe de l'ACM et le premier axe de l'analyse des liaisons sont fortement corrélés, d'où des interprétations analogues du premier axe. Mais les axes suivants sont généralement peu corrélés : l'analyse des liaisons peut révéler des informations contenues dans les tableaux binaires, mais masquées par l'ACM.

Bibliographie

- Durand, J.-L. (1998). Taux de dispersion des valeurs propres en ACP, AC et ACM. *Mathématiques Informatique et Sciences humaines*, 144, pp. 15-28.
- Durand, J.-L. et Le Roux, B. (2017). Linkage Index of Variables and its Relationship with Variance of Eigenvalues in PCA and MCA, *Statistica Applicata - Italian Journal of Applied Statistics*, 29 (2-3), pp. 123-135.
- Greenacre, M. (2017). *Correspondence Analysis in Practice*, Chapman and Hall/CRC, New York.
- Pagès, J. (2002). Analyse factorielle de données mixtes, *Revue de Statistique Appliquée*, 52 (4), pp. 93-111.