MATHÉMATIQUES ET SCIENCES HUMAINES

JEAN-LUC DURAND

Taux de dispersion des valeurs propres en ACP, AC et ACM

Mathématiques et sciences humaines, tome 144 (1998), p. 15-28

<http://www.numdam.org/item?id=MSH_1998__144__15_0>

Math. Inf. Sci. hum., (36e année, n°144, *1998*, pp. 15-28)

TAUX DE DISPERSION DES VALEURS PROPRES EN ACP, AC ET ACM*

Jean-Luc DURAND[1]

RÉSUMÉ — *Nous définissons le taux quadratique de concentration d'une mesure positive, ou taux quadratique de dispersion des valeurs de sa densité élémentaire. Appliqué aux valeurs propres d'un nuage de points dans un espace euclidien, ce taux s'interprète géométriquement comme un indice de non-sphéricité du nuage, rendant compte de sa capacité à être bien résumé par le(s) premier(s) axe(s). Nous donnons, en les commentant, les expressions de la variance corrigée et du taux de dispersion des valeurs propres pour les méthodes les plus usuelles d'analyse géométrique des données : analyse en composantes principales (ACP pondérée, simple et normée), analyse des correspondances (AC) et analyse des correspondances multiples (ACM). Ces relations montrent notamment qu'en ACP normée et en ACM l'intensité moyenne des liaisons binaires entre les variables s'exprime géométriquement par la non-sphéricité des nuages de points.*

SUMMARY — Dispersion rate of eigenvalues in PCA, CA and MCA.

We define the quadratic concentration rate of a positive mesure, or quadratic dispersion rate of the values of its elementary density. When applied on the eigenvalues of a cloud of points in an Euclidean space, this rate is geometrically interpreted as an index of non-sphericity of the cloud, which accounts for its capacity to be well summarized by the first axis or axes. We provide and comment upon the expressions of corrected variance and dispersion rate of eigenvalues for the most usual methods of geometric data analysis : principal component analysis (weighted PCA, simple and standard) correspondence analysis (CA) and multiple correspondence analysis (MCA). These relationships particularly show that in standard PCA and in MCA, the average intensity of binary relations between variables is geometrically expressed by the non-sphericity of clouds of points.

1. INTRODUCTION

L'un des objectifs de l'analyse géométrique des données est de fournir un résumé de petite dimension d'un protocole multivarié. La qualité de ce résumé dépend de la proportion de la variance du nuage prise en compte par les premiers axes. Or, en analyse en composantes principales normée, il est usuel de constater que les premières valeurs propres sont d'autant plus grandes que les corrélations entre variables sont fortes. En analyse des correspondances multiples, bien que les premières valeurs propres soient en général relativement faibles, on peut

* Nous remercions Henry Rouanet et Brigitte Le Roux pour leurs remarques et commentaires sur les précédentes versions de cet article.

[1] Laboratoire d'Éthologie Expérimentale et Comparée, ESA CNRS n° 7025, Université Paris XIII, Avenue J.-B. Clément, 93430 Villetaneuse, e-mail : durand@leec.univ-paris13.fr.

également remarquer qu'elles sont d'autant plus grandes que les variables catégorisées sont fortement liées deux à deux.

Dans cette note, nous préciserons ces propriétés en exprimant la variance corrigée des valeurs propres en fonction de statistiques évaluant la liaison entre les variables. Nous commencerons par définir un indice quadratique de concentration d'une mesure fondée sur l'inégalité (ou la dispersion) de ses masses ponctuelles.

2. CONCENTRATION D'UNE MESURE POSITIVE

On rencontre dans les situations les plus diverses des données se présentant comme la répartition d'une certaine quantité sur différents éléments : répartition d'une population sur des départements, d'un impôt sur les foyers fiscaux, des fréquences de réponse sur les modalités d'une question ou de la variance d'un nuage de points sur ses directions principales. Les données constituent alors une *mesure positive* sur un ensemble[2].

Soit $x_J : J \to R^+$ une mesure positive définie sur un support J de cardinal J. Notons $T = \sum x_j$ sa masse totale (supposée strictement positive). Soit $p_J : J \to R^+$ la mesure (positive) des proportions associées aux masses de la mesure x_J, avec $p_j = \frac{1}{T} x_j$ et $\sum_{j \in J} p_j = 1$.

2.1. Coefficient d'inégalité de C. Gini

L'évaluation de la concentration d'une mesure a donné lieu à la définition de plusieurs indices. Un indice classique, utilisé dans le cas d'une échelle de rapports, est par exemple le coefficient d'inégalité de C. Gini [2]. Cet indice, noté G, est obtenu en rapportant le diamètre moyen de la mesure, noté D (moyenne des distances entre les J masses ponctuelles de la mesure, prises deux à deux) à son maximum, égal à $2T/J$ (réalisé lorsque toutes les masses ponctuelles sont nulles, sauf une, égale à la masse totale) :

$$G = \frac{JD}{2T} \tag{1}$$

avec $D = \frac{2}{J(J-1)} \sum_{jj' \in P_2(J)} |x_j - x_{j'}| = \frac{2T}{J(J-1)} \sum_{jj' \in P_2(J)} |p_j - p_{j'}|$, $P_2(J)$désignant l'ensemble des $J(J-1)/2$ paires d'éléments de J.

---

[2] Une mesure donne lieu à une sommation lorsque l'on regroupe plusieurs éléments alors qu'une variable, comme par exemple une densité de population, se dérive par moyennage. Sur la dualité entre mesures et variables, voir [4] pp. 19-40 et [5]. Cette dualité s'exprimera dans nos notations par l'utilisation d'indices bas pour les mesures et hauts pour les variables, comme dans les ouvrages de Rouanet et Le Roux, auxquels nous empruntons la plupart des notations de cet article.

Propriétés

- Le coefficient d'inégalité de Gini est un nombre pur (sans dimension), déterminé par les proportions associées aux masses de la mesure : deux mesures proportionnelles ont un même coefficient d'inégalité.

- Le coefficient d'inégalité est compris entre 0 et 1. Il est nul lorsque les proportions sont égales (répartition uniforme). Il est égal à l'unité lorsque la mesure des proportions est une mesure de Dirac (concentration sur un seul élément du support).

- Sur un support binaire, le coefficient d'inégalité est égal à la valeur absolue de la différence entre les deux proportions :

$$\text{si} \quad \mathrm{J} = 2, \quad G = \left| p_{j1} - p_{j2} \right|. \tag{2}$$

Le coefficient d'inégalité de Gini est donc une évaluation de la concentration d'une mesure positive basée sur l'inégalité des masses ponctuelles, c'est-à-dire sur la dispersion de leurs valeurs. Plus la mesure est concentrée sur un petit nombre d'éléments du support, plus les valeurs des masses ponctuelles sont dispersées. Soulignons que les notions de concentration et dispersion portent ici sur des ensembles différents : J pour la mesure ; l'intervalle $[0, T]$ pour les valeurs de ses masses ponctuelles.

Nous nous intéresserons maintenant à l'évaluation de la dispersion des masses ponctuelles par une statistique quadratique, la variance corrigée[3]. Nous en déduirons un indice standardisé, que nous appellerons taux quadratique de concentration d'une mesure positive.

2.2. Taux quadratique de concentration

On appellera densité élémentaire d'une mesure x_J la densité de cette mesure par rapport à la mesure élémentaire $1_J : j \mapsto 1$, d'où la variable $\mathrm{x}^J : j \mapsto \mathrm{x}^j = \mathrm{x}_j$. Cette variable x^J, qui s'exprime numériquement comme la mesure x_J, nous apporte les statistiques attachées aux variables numériques, parmi lesquelles la moyenne $\bar{x} = T/\mathrm{J}$, et la variance corrigée :

$$s^2 = \mathrm{Varcor}\, x^J = \frac{1}{\mathrm{J}-1} \sum_{j \in J} \left(x^j - \bar{x} \right)^2 = \frac{1}{\mathrm{J}-1} \sum_{j \in J} \left(x^j \right)^2 - \frac{T^2}{\mathrm{J}(\mathrm{J}-1)}.$$

La variance corrigée de la densité élémentaire est minimale, de valeur nulle, lorsque ses valeurs sont toutes égales (densité constante). Elle est maximale, de valeur $s^2_{\max} = T^2/\mathrm{J}$, lorsque toutes ses valeurs sont nulles, sauf une.

Nous définissons le taux quadratique de concentration d'une mesure en rapportant à son maximum l'écart-type corrigé de sa densité élémentaire. Nous noterons $\mathrm{Cct}\, x_J$, ou en bref C, le taux quadratique de concentration d'une mesure x_J d'où $C = s\sqrt{\mathrm{J}}/T$, soit :

---

[3] L'évaluation de la concentration d'une mesure est basée sur la distribution des proportions de masses. Cette distribution, de masse totale égale à l'unité, est à J-1 degrés de liberté, d'où le choix de la variance corrigée (plutôt que la variance) dans un contexte d'analyse descriptive.

$$C^2 = (\text{Cct}\, x_J)^2 = \frac{\text{J Varcor}\, x^J}{T^2} = \text{J Varcor}\, p^J = \frac{\text{J}\sum_{j\in J}\left(p^j\right)^2 - 1}{\text{J} - 1} \qquad (3)$$

Nous réserverons le terme de concentration aux mesures positives et parlerons de dispersion pour les valeurs de leur densité. Nous noterons $\text{Dps}\, x^J$ le taux quadratique de dispersion de la variable x^J, d'où :

$$\text{Dps}\, x^J = \text{Cct}\, x_J = \frac{\sqrt{\text{J Varcor}\, x^J}}{T} = \sqrt{\text{JVarcor}\, P^J} \qquad (4)$$

Propriétés

Le taux quadratique de concentration d'une mesure positive possède les propriétés présentées ci-dessus pour le coefficient d'inégalité de Gini. On remarquera notamment que ces deux indices coïncident dans le cas d'une mesure sur un support binaire.

Pour une variable positive ayant les propriétés d'une échelle de rapports, on définit classiquement le taux de variation, quotient de son écart-type par sa moyenne : $t = \text{Ety}\, x^J/\bar{x}$. On a la relation :

$$\text{Dps}\, x^J = \frac{t}{\sqrt{\text{J} - 1}} \qquad (5)$$

Le maximum du taux de variation d'une échelle de rapports dépend du nombre J d'éléments du support : il est égal à $\sqrt{\text{J} - 1}$. Lorsque ce nombre J est fixé, le taux quadratique de dispersion apparaît comme une standardisation du taux de variation, ramenant son maximum à l'unité.

Notons enfin que l'on pourrait, de façon équivalente, définir le taux quadratique de concentration à partir de la distance du ϕ^2. Soit u_J la distribution uniforme sur J de masse totale égale à l'unité ($\forall j \in J,\ u_j = 1/\text{J}$). Considérons la distance du ϕ^2 de centre u_J entre la mesure des proportions p_J associée à une mesure x_J et la mesure uniforme u_J :

$$d(p_J, u_J) = \left(\sum_{j\in J} \frac{\left(p_j - 1/\text{J}\right)^2}{1/\text{J}} \right)^{1/2}$$

Si l'on rapporte cette distance à son maximum (égal à $\sqrt{\text{J} - 1}$ et réalisé dans le cas où la mesure des proportions est une mesure de Dirac), on retrouve le taux quadratique de concentration de la mesure x_J.

2.3. Concentration et restriction du support

La valeur d'un taux quadratique de concentration ou de dispersion dépend du nombre de modalités du support. Nous examinons d'abord les modifications apportées à la variance corrigée et au taux quadratique de concentration lorsque l'on retire du support un élément de masse ponctuelle nulle. Nous en déduirons ensuite une propriété plus générale liée à la restriction du support d'une mesure.

Soit, sur un support J de cardinal $J \geq 3$, une mesure positive x_J comportant au moins une masse ponctuelle nulle. Soit J' un support de cardinal $J' = J - 1$ obtenu en enlevant à J un élément de masse ponctuelle nulle. Notons $x_{J'}$ la restriction de la mesure x_J au support J'. Les deux mesures x_J et $x_{J'}$ sont de même masse totale T. On peut montrer que l'on a entre les moyennes, variances corrigées et taux quadratiques de dispersion de leurs densités élémentaires x^J et $x^{J'}$ les relations suivantes (ces relations seront utilisées lorsque l'on étudiera les valeurs propres de l'analyse des correspondances).

$$\text{Moy}\, x^{J'} = \frac{J}{J-1}\,\text{Moy}\, x^J \tag{6}$$

$$\text{Varcor}\, x^{J'} = \frac{J-1}{J-2}\,\text{Varcor}\, x^J - \frac{T^2}{J(J-1)(J-2)} \tag{7}$$

$$\left(\text{Dsp}\, x^{J'}\right)^2 = \frac{(J-1)^2}{J(J-2)}\left(\left(\text{Dsp}\, x^J\right)^2 - \frac{1}{(J-1)^2}\right) \tag{8}$$

Lorsqu'on passe de la variable x^J à la variable $x^{J'}$, la moyenne augmente. Le sens dans lequel la variance corrigée est modifiée dépend de la dispersion de la variable x^J. On peut montrer que la variance corrigée diminue, est inchangée ou augmente selon que le taux quadratique de dispersion de la variable x^J est respectivement inférieur, égal ou supérieur à $\sqrt{J}/(J-1)$. Enfin, on peut montrer que le taux quadratique de dispersion diminue toujours, sauf lorsqu'il est égal à l'unité, auquel cas il conserve sa valeur maximale.

On en déduit que dans tous les cas où une mesure positive comporte une ou plusieurs masses ponctuelles nulles et plusieurs masses ponctuelles non nulles la restriction de la mesure à la partie du support de masses non nulles s'accompagne d'une diminution de son taux quadratique de concentration. On retiendra que le taux quadratique de concentration n'a de sens que relativement à un support donné, de cardinal fixé.

2.4. Taux quadratique de concentration rectifié

Il peut arriver, dans certaines situations, que la définition de la mesure rende impossible la concentration de la masse totale sur un seul élément du support et que l'on ait, pour toute masse ponctuelle d'une mesure x_J, la contrainte $x_j \leq l$, cette valeur maximale l étant inférieure à la masse totale T[4]. Dans ce cas, le taux quadratique de concentration ne peut atteindre l'unité. La somme des carrés bruts maximale des masses ponctuelles, notée $SCB_{\max}$, est alors obtenue lorsqu'il n'y a pas plus d'une masse ponctuelle dans l'intervalle ouvert $]0, l[$. En notant $\mathbf{Int}(a)$ la partie entière d'un réel positif a et $\mathbf{Frac}(a) = a - \mathbf{Int}(a)$ sa partie fractionnaire, il y a alors $\mathbf{Int}(T/l)$ termes de masse maximale l et éventuellement un autre terme de masse non nulle, égale à $T - \mathbf{Int}(T/l) \times l = l \times \mathbf{Frac}(T/l)$, les autres termes (s'il y en a) étant de masse nulle. D'où :

$$SCB_{\max} = l^2 \times \mathbf{Int}(T/l) + \left(l \times \mathbf{Frac}(T/l)\right)^2 \tag{9}$$

[4] C'est le cas dans une enquête, par exemple, lorsqu'il est demandé de choisir dans une liste a modalités de réponse (a>1). La proportion maximum de réponses par modalité est $1/a$.

Si l'on veut un indice de concentration variant entre 0 et 1, on rapportera l'écart-type corrigé de la densité élémentaire de la mesure à son maximum. Nous nommerons ce rapport taux quadratique de concentration rectifié d'une mesure x_J, noté $\mathrm{Cct}_{\mathrm{rec}}\, x_J$, ou de façon équivalente taux quadratique de dispersion rectifié de la variable x^J, noté $\mathrm{Dps}_{\mathrm{rec}}\, x^J$:

$$\mathrm{Cct}_{\mathrm{rec}}\, x_J = \mathrm{Dps}_{\mathrm{rec}}\, x^J = \sqrt{\frac{\mathrm{Varcor}\, x^J}{\frac{SCB_{\max}}{\mathrm{J}-1} - \frac{T^2}{\mathrm{J}(\mathrm{J}-1)}}} \qquad (10)$$

Nous utiliserons le taux quadratique de concentration rectifié en analyse des correspondances multiples lorsque nous étudierons la dispersion des valeurs propres, toutes inférieures ou égales à l'unité (cf. §6).

2.5. Taux quadratique de concentration des valeurs propres d'un nuage euclidien

Dans les méthodes d'analyse géométrique des données, la variance d'un nuage euclidien exprime sa taille globale, et les valeurs propres, qui sont des parts de cette variance, traduisent le plus ou moins grand allongement du nuage dans ses diverses directions principales. On pourra voir les valeurs propres comme les masses d'une mesure positive (notée λ_L) sur l'ensemble L des directions principales, ou comme les valeurs d'une échelle de rapports sur L (variable notée λ^L, densité élémentaire de la mesure λ_L).

Le taux quadratique de concentration de la variance du nuage selon les directions principales, ou taux quadratique de dispersion des valeurs propres, s'interprète géométriquement comme indice de non-sphéricité du nuage. S'il est nul (valeurs propres égales), le nuage a la même variance dans toutes les directions et sera qualifié de nuage sphérique. Plus il est grand (*i.e.* plus les valeurs propres sont dispersées), plus le nuage est allongé dans ses premières directions principales (plus sa forme s'écarte de la sphéricité). A la limite, s'il est égal à 1 (une seule valeur propre non nulle), le nuage est porté par une droite.

3. DISPERSION DES VALEURS PROPRES D'UNE MATRICE SYMÉTRIQUE

Soit $\mathbf{A}$ une matrice symétrique $K \times K$. Notons $(a_k)_{k\in K}$ ses termes diagonaux, et $(a_{kk'})_{k\in K,\, k'\in K,\, k\neq k'}$ ses termes non-diagonaux. Notons $\mathbf{D}$ la matrice diagonale $K \times K$ définie par la famille $(\lambda_k)_{k\in K}$ de ses valeurs propres (éventuellement nulles) rangées par ordre décroissant. On peut voir les termes diagonaux de $\mathbf{A}$ comme formant une mesure sur K (dont la masse totale est la trace de la matrice), et les termes non-diagonaux comme formant une mesure sur l'ensemble des paires d'éléments de K, noté $P_2(K)$. Nous munissons les ensembles K et $P_2(K)$ de la pondération fondamentale élémentaire et notons $a^K = (a^k)_{k\in K}$ la densité élémentaire des termes diagonaux et $a^{P_2(K)} = (a^{kk'})_{kk'\in P_2(K)}$ la densité élémentaire des termes non-diagonaux.

3.1. Variance corrigée des valeurs propres

La variance corrigée des valeurs propres (variance corrigée de la variable λ^K, densité élémentaire de la mesure λ_K) est :

$$\text{Varcor}\,\lambda^K = \frac{1}{\text{K}-1}\sum_{k\in K}\lambda_k^2 - \frac{\left(\sum_{k\in K}\lambda_k\right)^2}{\text{K}(\text{K}-1)} \tag{11}$$

en notant K le cardinal de K. Or, la diagonalisation d'une matrice carrée conserve la trace :

$$\sum_{k\in K}\lambda_k = \text{tr}\,\mathbf{D} = \text{tr}\,\mathbf{A} = \sum_{k\in K} a_k$$

et, dans le cas d'une matrice symétrique, il y a également conservation de la somme des carrés des termes de la matrice :

$$\sum_{k\in K}\lambda_k^2 = \text{tr}\,\mathbf{D}^2 = \text{tr}\,\mathbf{A}^2 = \sum_{k\in K}\sum_{k'\in K} a_{kk'}^2$$

d'où :

$$\text{Varcor}\,\lambda^K = \frac{1}{\text{K}-1}\sum_{k\in K}\sum_{k'\in K} a_{kk'}^2 - \frac{(\text{tr}\,\mathbf{A})^2}{\text{K}(\text{K}-1)}$$

et, en séparant les K termes diagonaux des K(K− 1) termes non-diagonaux :

$$\text{Varcor}\,\lambda^K = \frac{1}{\text{K}-1}\sum_{k\in K} a_k^2 - \frac{(\text{tr}\,\mathbf{A})^2}{\text{K}(\text{K}-1)} + \frac{2}{\text{K}-1}\sum_{kk'\in P_2(K)} a_{kk'}^2$$

On en déduit la propriété :

$$\text{Varcor}\,\lambda^K = \text{Varcor}\,a^K + \text{K}\,\text{Moy}\left(a^2\right)^{P_2(K)} \tag{12}$$

La variance corrigée des valeurs propres d'une matrice symétrique d'ordre K est égale à la somme de la variance corrigée de ses termes diagonaux et du produit par K de la moyenne des carrés de ses termes non-diagonaux.

3.2. Taux de dispersion des valeurs propres

On suppose maintenant que la matrice $\mathbf{A}$ est semi-définie positive : $\forall k \in K \;\; \lambda_k \geq 0$. Les valeurs propres peuvent être vues comme une mesure positive sur K (de masse totale égale à la trace de la matrice) ou comme une variable sur K (densité élémentaire de la mesure) dont on peut déterminer le taux quadratique de dispersion. Les K termes diagonaux de la matrice et les K valeurs propres ayant même somme, et donc même moyenne, leur variance corrigée maximum commune s'écrit :

$$s_{\max}^2 = \text{K}\left(\text{Moy}\,a^K\right)^2 \tag{13}$$

soit, en rapportant la variance corrigée des valeurs propres à son maximum, la propriété :

$$\left(\text{Dps}\,\lambda^K\right)^2 = \left(\text{Dps}\,a^K\right)^2 + \frac{\text{Moy}\left(a^2\right)^{P_2(K)}}{\left(\text{Moy}\,a^K\right)^2} \tag{14}$$

Le carré du taux quadratique de dispersion des valeurs propres d'une matrice symétrique semi-définie positive est égal à la somme du carré du taux de dispersion de ses termes diagonaux et du quotient de la moyenne des carrés de ses termes non-diagonaux par le carré de la moyenne de ses termes diagonaux.

4. DISPERSION DES VALEURS PROPRES EN ACP

4.1. ACP bipondérée

Soit K un ensemble de K variables numériques définies sur un même support (élémentaire ou pondéré). Dans une analyse en composantes principales bipondérée[5] (forme la plus générale de l'ACP), chaque variable est affectée d'un poids (strictement positif). Notons ϖ_k le poids d'une variable k, v^k sa variance et $c^{kk'}$ la covariance entre deux variables k et k'. La matrice $\mathbf{A}$ à diagonaliser est symétrique, semi-définie positive, d'élément diagonal $a_k = \varpi_k v^k = \mathrm{Cta}_k$ (contribution absolue de la variable k à la variance du nuage des individus) et d'élément non diagonal $a_{kk'} = \sqrt{\varpi_k \varpi_{k'}}\, c^{kk'} = \sqrt{\mathrm{Cta}_k \mathrm{Cta}_{k'}}\, r^{kk'}$ (où $r^{kk'}$ est le coefficient de corrélation linéaire entre les variables k et k', *i.e.*, dans l'espace des variables, le cosinus de l'angle entre les vecteurs représentant ces variables). Notons $a^K = (a^k)_{k \in K}$ la densité élémentaire des termes diagonaux (ou densité élémentaire Cta^K des contributions absolues des variables à la variance) et $a^{P_2(K)} = (a^{kk'})_{kk' \in P_2(K)}$ la densité élémentaire des termes non-diagonaux. Notons $L = \{l \in N;\ 1 \leq l \leq \mathrm{K}\}$ (ensemble ordonné des entiers de 1 à K). La variance des K valeurs propres de l'ACP bipondérée s'écrit :

$$\mathrm{Varcor}\, l^L = \mathrm{Varcor}\, \mathrm{Cta}^K + \mathrm{K}\, \mathrm{Moy} \left(\mathrm{Cta}_k \mathrm{Cta}_{k'} (r^{kk'})^2 \right)_{kk' \in P_2(K)} \qquad (15)$$

Par ailleurs on a $\mathrm{Moy}\, a^k = \frac{1}{\mathrm{K}} \sum_{k \in K} \mathrm{Cta}_k$. Notons $\mathrm{Ctr}_k = \mathrm{Cta}_k \Big/ \sum_{k \in K} \mathrm{Cta}_k$ la contribution relative d'une variable k à la variance du nuage des individus. Les mesures Ctr_K et Cta_K étant proportionnelles, elles ont un même taux quadratique de concentration, et leurs densités élémentaires Ctr^K et Cta^K ont un même taux quadratique de dispersion, d'où le carré du taux quadratique de dispersion des valeurs propres :

$$\left(\mathrm{Dps}\, \lambda^L \right)^2 = \left(\mathrm{Dps}\, \mathrm{Ctr}^K \right)^2 + \mathrm{K}^2\, \mathrm{Moy} \left(\mathrm{Ctr}_k \mathrm{Ctr}_{k'} (r^{kk'})^2 \right)_{kk' \in P_2(K)} \qquad (16)$$

Les valeurs propres d'une ACP bipondérée sont d'autant plus dispersées que :

- les contributions des variables sont dispersées (ce qui est le cas lorsque les poids les plus forts portent sur les variables ayant les plus grandes variances) ;
- la moyenne des termes $\mathrm{Ctr}_k \mathrm{Ctr}_{k'} (r^{kk'})^2$ associés aux paires de variables est élevée (ce qui est le cas lorsque les variables contribuant le plus à la variance sont fortement corrélées).

---

[5] Cette analyse est appelée ACP pondérée dans [4] pp. 168-175. Sa mise en œuvre peut être réalisée avec le logiciel ADDAD (module ACPPON, réalisé par B. Le Roux et P. Bonnet).

4.2. ACP simple

Une ACP simple (ou non normée) est un cas particulier d'ACP bipondérée, dans lequel les variables analysées sont toutes affectées d'un même poids $\varpi_k = 1$ (et non réduites). La matrice à diagonaliser a pour éléments diagonaux les variances des variables et pour éléments non diagonaux les covariances entre variables. D'où la variance corrigée des valeurs propres :

$$\text{Varcor } \lambda^L = \text{Varcor } v^K + \text{K Moy } \left(c^2\right)^{P_2(K)} \tag{17}$$

La variance corrigée des valeurs propres d'une ACP simple sur K variables est égale à la somme de la variance corrigée des variances des variables et du produit par K de la moyenne des carrés des covariances entre les variables.

En ACP simple, la contribution absolue d'une variable étant égale à sa variance, on a $\text{Ctr}_k = \text{v}_k \Big/ \sum_{k \in K} v_k$. D'où le carré du taux quadratique de dispersion des valeurs propres :

$$\left(\text{Dps } \lambda^L\right)^2 = \left(\text{Dps } v^K\right)^2 + \text{K}^2 \text{ Moy } \left(\text{Ctr}_k \text{Ctr}_{k'} (\text{r}^{kk'})^2\right)_{kk' \in P_2(K)} \tag{18}$$

Les valeurs propres d'une ACP simple sont d'autant plus dispersées que :

- les variances des variables sont dispersées ;
- la moyenne des termes $\text{Ctr}_k \text{Ctr}_{k'} (r^{kk'})^2$ associés aux paires de variables est élevée (ce qui est le cas lorsque les variables de plus grandes variances sont fortement corrélées).

4.3. ACP normée

Une ACP normée (ou ACP standard) est un cas particulier d'ACP simple dans lequel les variables sont réduites. La matrice à diagonaliser est la matrice des corrélations entre variables : ses éléments diagonaux sont égaux à l'unité, donc de variance corrigée nulle. La variance corrigée des valeurs propres s'écrit :

$$\text{Varcor } \lambda^L = \text{K Moy } \left(r^2\right)^{P_2(K)} \tag{19}$$

La variance corrigée des valeurs propres d'une ACP normée sur K variables est égale à K fois la moyenne des carrés des corrélations entre les variables.

En ACP normée, la contribution absolue de chaque variable étant égale à l'unité, on a pour toute variable k : $\text{Ctr}_k = 1/\text{K}$. D'où le taux quadratique de dispersion des valeurs propres :

$$\text{Dps } \lambda^L = \sqrt{\text{Moy } \left(r^2\right)^{P_2(K)}} \tag{20}$$

Le taux quadratique de dispersion des valeurs propres d'une ACP normée est égal à la moyenne quadratique des corrélations entre les variables.

Les valeurs propres d'une ACP normée sont d'autant plus dispersées que les variables sont corrélées. En ACP normée, la taille du nuage des individus est indépendante des données (la variance du nuage est égale au nombre de variables) et c'est sa non-sphéricité (évaluée par le

taux quadratique de dispersion des valeurs propres) qui exprime l'intensité des liaisons linéaires entre les variables. Cette propriété éclaire l'influence du choix des variables sur le pourcentage de variance expliqué par les premiers axes d'une ACP ([3] pp. 370-371).

Dans le cas particulier d'une ACP normée sur deux variables, en notant r leur coefficient de corrélation, le taux de dispersion des deux valeurs propres s'écrit (comme le coefficient d'inégalité de Gini) : $\frac{\lambda_{l1} - \lambda_{l2}}{2} = |r|$, d'où l'expression des valeurs propres : $\lambda_{l1} = 1 + |r|$ et $\lambda_{l2} = 1 - |r|$.

5. DISPERSION DES VALEURS PROPRES EN AC

Soit $n_{JK} = (n_{jk})_{jk \in J \times K}$ un tableau de contingence d'effectifs. Notons $n_J = (n_j)_{j \in J}$ et $n_K = (n_k)_{k \in K}$ ses distributions marginales (supposées strictement positives), avec $n_j = \sum_{k \in K} n_{jk}$ et $n_k = \sum_{j \in J} n_{jk}$, et n son effectif total. Notons $f_K^j = (f_k^j)_{k \in K}$ le profil sur K d'une modalité j de J (avec $f_k^j = n_{jk}/n_j$) et $f_K = (f_k)_{k \in K}$ le profil moyen sur K (avec $f_k = n_k/n$). Notons $d^{jK} = (d^{jk})_{k \in K}$ la densité du profil de j sur K par rapport au profil moyen sur K, avec $d^{jk} = f_k^j / f_k = f_{jk}/f_j f_k$ (densité de la fréquence conjointe par rapport à la fréquence-produit). Le tableau des densités $d^{JK} = (d^{jk})_{jk \in J \times K}$ peut être vu de trois façons : comme une famille de J variables sur K (muni de la pondération f_K) pondérées par f_J ; comme une famille de K variables sur J (muni de la pondération f_J) pondérées par f_K ; ou enfin comme une variable sur $J \times K$ muni de la pondération des fréquences-produits $\hat{f}_{JK} = (f_j f_k)_{j \in J, k \in K}$. Ces variables sur J, sur K et sur $J \times K$ ont toutes 1 pour moyenne (les taux de liaison de Rouanet et Le Roux [4] sont obtenus par centrage de ces variables : $t^{jk} = d^{jk} - 1$). La variance de la variable d^{JK} est le carré moyen de contingence, noté ϕ^2.

Dans l'analyse des correspondances du tableau n_{JK}, les deux nuages euclidiens $M^J = (M^j)_{j \in J}$ et $M^K = (M^k)_{k \in K}$, représentant respectivement les J profils sur K et les K profils sur J, sont de même dimension, inférieure ou égale à $\mathrm{L} = \min(\mathrm{J} - 1, \mathrm{K} - 1)$, de même variance ϕ^2, et possèdent la même famille de valeurs propres. Notons $L = \{l \in N ;\ 1 \leq l \leq \mathrm{L}\}$.

Les résultats de l'AC du tableau n_{JK} sont équivalents à ceux des L premières directions principales de l'ACP bipondérée du tableau des densités d^{JK} (muni des pondérations f_J sur J et f_K sur K) ou de son transposé. On pourra donc exprimer la dispersion des valeurs propres de l'analyse des correspondances à partir des variances et covariances des J densités sur K et de la pondération f_J (ou de façon symétrique, comme nous le faisons ci-dessous, à partir des variances et covariances des K densités sur J et de la pondération f_K).

Nous supposerons que l'on a $\mathrm{J} \geq \mathrm{K}$, d'où $\mathrm{L} = \mathrm{K} - 1$. Notons v^k la variance de la densité d^{kJ} du profil d'une modalité k sur J, $\mathrm{Cta}_k = f_k v^k$ la contribution absolue au ϕ^2 de cette modalité, $c^{kk'}$ et $r^{kk'}$ la covariance et le coefficient de corrélation linéaire entre les deux variables d^{kJ} et $d^{k'J}$. Dans l'espace des profils sur J, en notant G le point moyen du

nuage M^K, v^k est le carré de la longueur du vecteur GM^k représentant la variable d^{kJ} et $r^{kk'}$ est le cosinus de l'angle entre les vecteurs GM^k et $GM^{k'}$).

La matrice **A** à diagonaliser est symétrique, semi-définie positive, d'élément diagonal $a_k = f_k v^k = \mathrm{Cta}_k$ et d'élément non diagonal $a_{kk'} = \sqrt{f_k f_{k'}} c^{kk'} = \sqrt{\mathrm{Cta}_k \mathrm{Cta}_{k'}} r^{kk'}$. La variance corrigée des $\mathrm{L}' = \mathrm{K}$ valeurs propres de l'ACP pondérée s'écrit :

$$\mathrm{Varcor}\ \lambda^{L'} = \mathrm{Varcor}\ \mathrm{Cta}^K + \mathrm{K}\ \mathrm{Moy}\left(\mathrm{Cta}_k \mathrm{Cta}_{k'} (r^{kk'})^2\right)_{kk' \in P_2(K)} \tag{21}$$

La dernière des K valeurs propres de l'ACP bipondérée étant nulle, la variance corrigée des $\mathrm{L} = \mathrm{K} - 1$ valeurs propres de l'AC s'écrit, en utilisant l'équation (7) :

$$\mathrm{Varcor}\ \lambda^L = \frac{\mathrm{K}-1}{\mathrm{K}-2}\,\mathrm{Varcor}\ \mathrm{Cta}^K + \mathrm{K}\ \mathrm{Moy}\left(\mathrm{Cta}_k \mathrm{Cta}_{k'} (r^{kk'})^2\right)_{kk' \in P_2(K)} - \frac{\left(\phi^2\right)^2}{\mathrm{K}(\mathrm{K}-1)(\mathrm{K}-2)} \tag{22}$$

et le carré du taux quadratique de dispersion des valeurs propres de l'AC :

$$\left(\mathrm{Dps}\ \lambda^L\right)^2 = \frac{(\mathrm{K}-1)^2}{\mathrm{K}(\mathrm{K}-2)}\left(\left(\mathrm{Dps}\ \mathrm{Ctr}^K\right)^2 + \mathrm{K}^2\ \mathrm{Moy}\left(\mathrm{Ctr}_k \mathrm{Ctr}_{k'} (r^{kk'})^2\right)_{kk' \in P_2(K)} - \frac{1}{(\mathrm{K}-1)^2}\right) \tag{23}$$

Les valeurs propres d'une AC sont d'autant plus dispersées que :

- les contributions des modalités à la variance sont dispersées (ce qui est le cas lorsque les modalités les plus fréquentes possèdent les profils les plus éloignés du profil moyen) ;
- la moyenne des termes $\mathrm{Ctr}_k \mathrm{Ctr}_{k'} (r^{kk'})^2$ associés aux paires de modalités est élevée (ce qui est le cas lorsque les profils des modalités contribuant le plus à la variance ont des densités fortement corrélées).

En AC, l'intensité de la liaison globale entre les deux variables catégorisées s'exprime géométriquement par la taille des nuages de points (leur variance, somme des valeurs propres, est aussi la variance des taux de liaison, cf. [4] pp. 207, 213). La dispersion des valeurs propres est déterminée par la dispersion des contributions des modalités à la variance et par les corrélations entre les densités des profils des modalités ayant les plus fortes contributions.

6. DISPERSION DES VALEURS PROPRES EN ACM

Soit Q un ensemble de questions et K l'ensemble de toutes les modalités de ces questions. Notons Q et K leurs cardinaux respectifs. Notons $L = \{l \in N;\ 1 \leq l \leq \mathrm{K} - \mathrm{Q}\}$ (ensemble ordonné des entiers de 1 à K - Q). Considérons l'analyse des correspondances multiples d'un protocole à valeurs dans le produit cartésien de ces questions, c'est-à-dire l'analyse des correspondances du tableau disjonctif complet.

La variance corrigée des K - Q valeurs propres non triviales est :

$$\text{Varcor}\,\lambda^L = \frac{1}{\text{K}-\text{Q}-1}\sum_{l\in L}\lambda_l^2 - \frac{\left(\sum_{l\in L}\lambda_l\right)^2}{(\text{K}-\text{Q})(\text{K}-\text{Q}-1)} \tag{24}$$

Or en ACM la somme des valeurs propres est :

$$\sum_{l\in L}\lambda_l = \frac{\text{K}-\text{Q}}{\text{Q}} \tag{25}$$

et la somme des carrés des valeurs propres est égale au ϕ^2 du tableau de Burt, lequel est la moyenne des ϕ^2 des Q^2 sous-tableaux composant le tableau de Burt ([4] p. 256) :

$$\sum_{l\in L}\lambda_l^2 = \Phi^2_{\text{Burt}} = \frac{1}{\text{Q}^2}\left(\text{K}-\text{Q}+2\sum_{qq'\in P_2(Q)}\left(\phi^2\right)^{qq'}\right) \tag{26}$$

où $\left(\phi^2\right)^{qq'}$ désigne le carré moyen de contingence des deux questions q et q'.

D'où la variance corrigée des valeurs propres de l'analyse des correspondances multiples :

$$\text{Varcor}\lambda^L = \frac{\text{Q}-1}{\text{Q}(\text{K}-\text{Q}-1)}\,\text{Moy}\left(\phi^2\right)^{P_2(Q)} \tag{27}$$

En ACM, la variance corrigée des valeurs propres est proportionnelle à la moyenne des carrés moyens de contingence des tableaux binaires. On en déduit la relation :

$$\text{Moy}\left(\phi^2\right)^{P_2(Q)} = \sum_{l\in L}\frac{\text{Q}}{\text{Q}-1}\left(\lambda_l - \frac{1}{\text{Q}}\right)^2 \tag{28}$$

La force moyenne des liaisons binaires entre les questions ne s'exprime pas par la somme des valeurs propres (qui, égale à (K - Q)/Q, est indépendante des données) mais par les carrés des écarts des valeurs propres à leur moyenne (égale à 1/Q). On peut rapprocher cette propriété de la formule de calcul des taux modifiés, proposée par Benzécri ([1] p. 306) dans le but de mieux apprécier l'importance relative des directions principales d'une ACM. À chaque valeur propre λ_l supérieure à la valeur propre moyenne 1/Q on associe la quantité $\lambda_l' = \left(\frac{\text{Q}}{\text{Q}-1}\right)^2\left(\lambda_l - \frac{1}{\text{Q}}\right)^2$. Le taux modifié de Benzécri pour l'axe l est proportionnel à la contribution de cet axe à la moyenne des ϕ^2 binaires (en se limitant toutefois aux axes de valeurs propres supérieures à la moyenne).

Dans le cas particulier où les Q questions sont toutes binaires, la somme des valeurs propres est égale à l'unité, et leur taux de dispersion s'écrit :

$$\text{Dps}\,\lambda^L = \sqrt{\text{Moy}\left(\phi^2\right)}^{P_2(Q)} \tag{29}$$

Le taux de dispersion des valeurs propres d'une ACM portant sur des questions binaires est égal à la moyenne quadratique des ϕ^2 des tableaux binaires.

Dans les autres cas, la somme des valeurs propres est supérieure à l'unité, alors que les valeurs propres sont toutes inférieures à 1. Nous pourrons donc, pour avoir un indice variant entre 0 et 1, calculer le taux de dispersion rectifié. La valeur maximale pour la somme des carrés bruts est :

$$SCB_{\max} = \mathrm{Int}\left(\frac{K}{Q} - 1\right) + \left(\frac{K}{Q} - 1 - \mathrm{Int}\left(\frac{K}{Q} - 1\right)\right)^2 \tag{30}$$

Nous calculerons donc le taux quadratique de dispersion rectifié des valeurs propres :

$$\mathrm{Dps}_{rec}\,\lambda^L = \sqrt{\frac{\mathrm{Varcor}\,\lambda^L}{\frac{SCB_{\max}}{K - Q - 1} - \frac{K - Q}{Q^2(K - Q - 1)}}} \tag{31}$$

Dans le cas particulier de questions ayant toutes un même nombre k de modalités (avec Qk = K), la variance corrigée maximale des valeurs propres s'écrit :

$$s^2_{\max} = \frac{(Q-1)(K-Q)}{Q^2(K-Q-1)} \tag{32}$$

d'où le taux de dispersion rectifié :

$$\mathrm{Dps}_{rec}\,\lambda^L = \sqrt{\frac{1}{K-1}}\,\mathrm{Moy}\left(\phi^2\right)^{P_2(Q)} \tag{33}$$

soit encore, en notant $\left(V^2\right)^{qq'} = \left(\phi^2\right)^{qq'} / (K-1)$ le carré du coefficient de contingence de Cramer entre deux questions q et q' :

$$\mathrm{Dps}_{rec}\,\lambda^L = \sqrt{\mathrm{Moy}\left(V^2\right)^{P_2(Q)}} \tag{34}$$

Dans le cas de questions ayant toutes un même nombre de modalités, le taux de dispersion rectifié des valeurs propres d'une ACM est la moyenne quadratique des coefficients de Cramer associés aux tableaux binaires.

7. DISCUSSION

Nous avons montré qu'en ACP normée comme en ACM, le taux de dispersion des valeurs propres est proportionnel à la liaison moyenne entre les variables étudiées (moyenne des carrés des corrélations en ACP normée et moyenne des carrés moyens de contingence en ACM).

L'examen de la matrice des corrélations (en ACP normée) ou de la matrice des ϕ^2 (en ACM) permet de prévoir le taux de dispersion des valeurs propres et de s'attendre à une plus ou moins grande proportion de variance expliquée par les premiers axes[6]. Si les variables sont très peu liées, les valeurs propres seront proches de leur moyenne, (nuages de points de formes

[6] Il serait souhaitable que les logiciels d'analyse de données fournissent la valeur de la moyenne des carrés des corrélations (en ACP standard) ou celle de la moyenne des carrés moyens de contingence (en ACM), ou encore celle du taux de dispersion des valeurs propres.

proches de la sphéricité), et la proportion de variance expliquée par les premiers axes sera relativement faible. À l'opposé, si les variables sont fortement liées entre elles, les valeurs propres seront très dispersées (nuages de formes éloignées de la sphéricité) et la proportion de variance expliquée par les premiers axes sera élevée.

En ACP normée et en ACM, la variance des nuages de points est indépendante de l'intensité des liaisons entre les variables. Cette caractéristique distingue ces deux méthodes de l'analyse des correspondances d'un tableau de contingence binaire qui produit deux nuages ayant pour variance le carré moyen de contingence, indice global de liaison entre les deux questions. L'intensité de la liaison entre les variables s'exprime géométriquement par la taille des nuages (somme des valeurs propres) en AC et par leur non-sphéricité (taux de dispersion des valeurs propres) en ACP normée et en ACM.

BIBLIOGRAPHIE

[1] BENZÉCRI J.-P., & Coll., *Pratique de l'analyse des données*, tome 1 : *Analyse des correspondances, exposé élémentaire*, Paris, Dunod, 2ème édition, 1984.

[2] BARBUT M., "Diamètres et écarts, une décomposition du coefficient d'inégalité de C. Gini", *Mathématiques, Informatique et Sciences humaines*, 93, (1986), 61-69.

[3] LEBART L., MORINEAU A., PIRON M., *Statistique exploratoire multidimensionnelle*, Paris, Dunod, 1995.

[4] ROUANET H., LE ROUX B., *Analyse des données multidimensionnelles*, Paris, Dunod, 1993.

[5] ROUANET H., LEPINE D., "Structures linéaires et analyse des comparaisons", *Mathématiques, Informatique et Sciences humaines*, 56, (1976), 5-46.