REVUE DE STATISTIQUE APPLIQUÉE

P. CAZES

Analyse factorielle d'un tableau de lois de probabilité

Revue de statistique appliquée, tome 50, n° 3 (2002), p. 5-24

<http://www.numdam.org/item?id=RSA_2002__50_3_5_0>

Rev. Statistique Appliquée, 2002, **L** (3), 5-24

ANALYSE FACTORIELLE D'UN TABLEAU DE LOIS DE PROBABILITÉ

P. CAZES

Lise Ceremade, UMR7534, Université Paris 9 Dauphine,
Place du Maréchal De Lattre de Tassigny, 75775, Paris Cedex 16.

RÉSUMÉ

On développe ici des techniques d'analyse factorielle pour analyser un tableau dont chaque case est une loi de probabilité connue. Dans le cas où le support de chacune de ces lois est fini, on obtient comme cas particulier le cas des données intervalles déjà étudié. On donne également des compléments sur l'analyse de ces données intervalles, et on discute de l'intérêt et des limites des méthodologies qui ont été développées pour traiter ce type de données.

Mots-clés : *Analyse en Composantes Principales, Données-intervalles, Lois de probabilité, Variabilité.*

ABSTRACT

We propose factorial analysis methods for study tables where each cell is a probability law. When the support of these laws is finite, we have the case of interval data which has been studied. We give also complements on these interval data and discuss methodologies which has been developped for study these types of data.

Keywords : *Interval data, Principal Component Analysis, Probability law, Variability.*

I. - Introduction

On considère un tableau rectangulaire X à n lignes et p colonnes, les lignes pouvant être considérées comme des observations ou des individus, et les colonnes comme des variables. On suppose que dans la case (i, j) du tableau précédent, on a une loi de probabilité, et on désire effectuer une analyse factorielle de ce tableau de façon à représenter sur un axe, un plan, ou de façon plus générale dans un espace à s dimensions, les lignes et les colonnes du tableau X.

Cette étude généralise le cas où dans chaque case du tableau, on a un intervalle (*cf.* Cazes *et alt.*, 1997, Tang Ahanda, 1998, Rodriguez-Rojas, 2000).Ce dernier auteur considère aussi le cas où dans chaque cellule du tableau on a un histogramme, mais il ne fait intervenir que la distribution de fréquences associée, et non pas la valeur de la variable sous-jacente (dans le cas d'une variable discrète ou d'une variable quantitative découpée en classes), contrairement à la façon d'opérer ici. Boumaza (1998) considère aussi l'Analyse en Composantes Principales d'un tableau de lois de

probabilité, mais il se place dans un cadre différent dans la mesure où chaque ligne (et non plus chaque cellule) correspond à une loi de probabilité. Lauro *et alt.* (2000) ont développé des méthodes de codage pour effectuer l'analyse factorielle discriminante de données symboliques plus générales que des données intervalles mais qui ne sont pas des lois de probabilité. Emilion (2001) traite, d'un point de vue probabiliste, du même type de tableau que ceux étudiés ici mais dans un cadre de classification et de mélange de lois. De même Aboa (2002) développe sur le même type de données des méthodes de segmentation.

Outre cette introduction, cet article comporte trois parties. Dans la première (§ 2) on définit les notations et les hypothèses tandis que dans la seconde (§ 3), on suggère un certain nombre d'analyses factorielles et on étudie leurs propriétés. On détaille en particulier la manière de représenter la variabilité (ou l'imprécision) due au fait qu'on a des lois de probabilité et non pas des valeurs numériques, en faisant référence à ce qui est fait dans le cas de données intervalles. Dans la dernière partie (§ 4), on donne des compléments sur l' analyse factorielle de données intervalles, et on compare les méthodes d'analyse factorielle déjà préconisées (méthode des centres et méthode des sommets) avec d'autres méthodes (STATIS, l'Analyse Factorielle Multiple, etc.)

II. Notations-Hypothèses

Nous supposerons que chaque individu est muni d'une masse $p_i (p_i > 0, \sum p_i = 1)$ ce qui revient à munir l'espace $F = R^n$ de la métrique diagonale des poids D_p. Nous supposerons également que l'espace R^p est muni de la métrique M.

Nous supposerons que conditionnellement à l'individu i, la variable j suit une loi de probabilité de densité $f_{ij}(x)$ et que le couple de variables j, j' $(j \neq j')$ suit une loi de densité $f_{ijj'}(x)$ (dont les marges $f_{ij}(x)$ et $f_{ij'}(x)$ sont bien sûr les lois de j et j' respectivement, conditionnellement à i) qui se réduit à $f_{ij}(x).f_{ij'}(x)$ dans le cas de l'indépendance.

Les différentes analyses considérées dans le cas de données de type intervalle (Cazes *et alt.*, 1997) correspondent au cas ou les lois $f_{ij}(x)$ sont de support fini avec l'hypothèse d'indépendance conditionnelle $(f_{ijj'}(x) = f_{ij}(x).f_{ij'}(x)$, pour $j \neq j')$.

On peut alors à chaque case (i, j) du tableau X associer, conditionnellement à i, la moyenne g_{ij} et la variance σ_{ij}^2 de j :

$$g_{ij} = \int_R x f_{ij}(x) dx$$

$$\sigma_{ij}^2 = \int_R (x - g_{ij})^2 f_{ij}(x) dx$$

On peut de même calculer conditionnellement à i la covariance entre j et $j' (j \neq j')$ qu'on notera $\sigma_{ijj'}$:

$$\sigma_{ijj'} = \int_R (x - g_{ij})(y - g_{ij'}) f_{ijj'}(x, y)\, dx\, dy$$

A chaque observation i, on peut donc associer un point moyen $\underline{g}_i$ (qui est le vecteur colonne de composantes $g_{ij}(1 \leqslant j \leqslant p)$) et une matrice variance V_i qui est la matrice des $\sigma_{ijj'}(1 \leqslant j, j' \leqslant p$, avec $\sigma_{ijj} = \sigma_{ij}^2)$.

On peut alors définir le centre de gravité général $\underline{g}$:

$$\underline{g} = \sum\{p_i \, \underline{g}_i \mid i = 1, n\}$$

ainsi que les matrices variances interclasses B, intraclasses W, et totale V :

$$B = \sum\{p_i(\underline{g}_i - \underline{g})(\underline{g}_i - \underline{g})' \mid i = 1, n\}$$
$$W = \sum\{p_i V_i \mid i = 1, n\}$$
$$V = B + W$$

Dans les expressions précédentes, le prime correspond comme c'est classique en statistique à la transposition.

On peut aussi définir la matrice carrée d'ordre n W_G de terme général :

$$\forall i, i' = 1, \ldots, n : (W_G)_{ii'} = \sum\{m_{jj'}(g_{ij} - g_j)(g_{ij'} - g_{j'}) \mid j = 1, p; j' = 1, p\}$$

$m_{jj'}$ étant le terme général de la matrice associée à la métrique M, matrice que l'on notera également M et g_j (resp. $g_{j'}$) la $j^{\text{ème}}$ (resp. $j'^{\text{ème}}$) composante de $\underline{g}$.

On supposera dans la suite sans perte de généralités qu'on a ramené le centre de gravité global $\underline{g}$ à l'origine, soit $\underline{g} = \underline{0}$.

Dans ces conditions on a :

$$B = G' D_p G$$
$$W_G = GMG'$$

G étant la matrice $n \times p$ de terme général g_{ij} et D_p la matrice diagonale, d'ordre n des p_i (*i.e.* la matrice diagonale des poids).

III. Analyses et propriétés

III.1. Analyses

On peut considérer les analyses factorielles suivantes :

– L'analyse globale qui revient à diagonaliser VM ou MV et qui correspond dans le cas de données intervalles à la méthode des sommets.

– L'analyse interclasses qui revient à diagonaliser BM ou MB et qui correspond dans le cas de données intervalles à la méthode des centres. Il s'agit de l'Analyse

en Composantes Principales (ACP) du tableau G qui rappelons-le est issu du tableau X en remplaçant dans chaque case de X la loi de probabilité par sa moyenne.

– L'analyse factorielle discriminante (qui est une analyse interclasses particulière avec le choix de la métrique d'inertie $M = V^{-1}$, V étant supposée inversible) qui revient à diagonaliser BV^{-1} et $V^{-1}B$ et dont le but est de séparer au maximum les individus compte tenu de la dispersion donnée par les lois de probabilité qui les caractérisent.

– L'analyse intraclasses qui revient à diagonaliser WM ou MW. Cette dernière analyse semble moins intéressante dans la mesure où elle ne met en valeur que les différenciations entre les individus, associées à leurs dispersions individuelles autour de leur centre de gravité partiel.

Remarque : Dans les deux premières analyses ainsi du reste que dans la dernière, on peut choisir la métrique de façon à effectuer des analyses normées (indépendantes du choix des unités pour les variables considérées) ce qui revient à prendre respectivement $M = (\mathrm{Diag}V)^{-1}$, $M = (\mathrm{Diag}B)^{-1}$, $M = (\mathrm{Diag}W)^{-1}$ ($\mathrm{Diag}A$ désignant pour une matrice carrée A la matrice diagonale ayant mêmes éléments diagonaux que A).

On peut noter que dans l'analyse discriminante, on peut aussi prendre W^{-1} (W étant supposée inversible) comme métrique, auquel cas, si W est diagonale (ce qui est le cas dans la méthode des sommets dans le cas de données intervalles) l'analyse discriminante est équivalente à l'analyse factorielle usuelle (*i.e.* avec la métrique identité) du tableau de terme général $g_{ij}/(w_{jj})^{1/2}$, w_{jj} étant le $j^{\text{ème}}$ terme diagonal de W.

III.2. Représentations des individus et des variables

Soit $\underline{u}$ un vecteur axial factoriel issu d'une des analyses précédentes (*i.e.* un vecteur propre de VM, BM, BV^{-1} ou WM suivant le cas) et $\underline{\varphi} = M\underline{u}$ (avec $M = V^{-1}$ dans le cas de l'analyse discriminante) le facteur associé. Pour représenter un individu i sur l'axe factoriel associé $\Delta\underline{u}$, il semble logique de projeter le point moyen $\underline{g}_i$ associé à l'individu i. La coordonnée ψ^i de $\underline{g}_i$ sur $\Delta\underline{u}$ s'écrit alors (puisqu'on a supposé que $\underline{g} = 0$) :

$$\psi^i = \sum\{\varphi^j g_{ij} \mid j = 1, p\}$$

φ^j désignant la $j^{\text{ème}}$ composante de $\underline{\varphi}$.

On désignera par $\underline{\psi}$ le vecteur de composantes $\psi^i (1 \leqslant i \leqslant n)$, vecteur qu'on appellera composante principale, et il est immédiat de vérifier que $\underline{\psi}$ est bien centré ($\sum\{p_i\psi^i \mid i = 1, n\} = 0$).

La variance de $\underline{\psi}$ s'écrit alors :

$$\mathrm{Var}\underline{\psi} = \sigma^2_{\psi} = \sum\{p_i(\psi^i)^2 \mid i = 1, n\} = \underline{\varphi}'\, B\, \underline{\varphi}$$

Il s'agit en fait d'une variance interclasses qui dans le cas de l'analyse interclasses ou de l'analyse discriminante est égale à la valeur propre associée à $\underline{\varphi}$ (ou à $\underline{u}$).

La façon d'opérer précédente revient à projeter les lignes du tableau G sur l'axe factoriel considéré, ces lignes intervenant suivant le cas en tant qu'élément actif (analyse interclasses ou analyse discriminante, auquel cas $\underline{\psi}$ est vecteur propre de $W_G D_p$) ou supplémentaire (analyse globale ou analyse intraclasses).

Pour avoir une représentation des variables, il suffit de calculer les corrélations entre les colonnes de G et $\underline{\psi}$. Il s'agit donc de corrélations interclasses. La corrélation $r_{\psi j}$ entre $\underline{\psi}$ et la variable $\underline{j}$ s'écrit alors :

$$r_{\psi j} = (\sum\{p_i g_{ij} \psi^i \mid i = 1, n\})/((\underline{\varphi}' B \underline{\varphi}) b_{jj})^{1/2}$$

b_{jj} étant le $j^{\text{ème}}$ terme diagonal de B (*i.e.* la variance interclasses de j). Dans le cas d'une analyse interclasses, cette formule se réduit à :

$$r_{\psi j} = \lambda u_j/(\lambda b_{jj})^{1/2} = \lambda^{1/2} u_j/(b_{jj})^{1/2}$$

u_j étant la $j^{\text{ème}}$ composante de $\underline{u}$ et λ la valeur propre associée.

III.3. Représentation de la variabilité

III.3.1. Introduction

Pour représenter la variabilité due au fait que les données ne sont pas des valeurs numériques, mais des lois de probabilité traduisant une dispersion autour de la valeur moyenne, on va opérer de façon analogue à ce qui a été fait pour les données intervalles (Cazes *et alt.*, 1997).On verra au paragraphe III.4 d'autres façons de représenter la variabilité. A chaque loi de probabilité f_{ij}, on associe un intervalle (x_{ij-}, x_{ij+}). Si le support de f_{ij} est fini, x_{ij-} et x_{ij+} correspondent aux extrémités de ce support; sinon on peut prendre pour x_{ij-} et x_{ij+} les quantiles d'ordre $\alpha/2$ et $(1-\alpha/2)$ de la loi f_{ij} (on pourra prendre les valeurs classiques $\alpha = 5\%$ ou $\alpha = 1\%$; si la loi f_{ij} est mal spécifiée, mais que l'écart type σ_{ij} est connu, on pourra prendre $x_{ij-} = g_{ij} - t\sigma_{ij}$ et $x_{ij+} = g_{ij} + t\sigma_{ij}$, avec $t = 2$ ou $t = 3$). En fait, d'un point de vue pratique, pour ne pas trop surestimer la dispersion dans les procédures développées ci-dessous, qui supposent l'indépendance conditionnellement à chaque individu des variables, on aura intérêt à prendre des valeurs plus importantes de α, de l'ordre de 20 à 30 %.

III.3.2. Cas des individus

On peut associer à l'abscisse ψ^i de $\underline{g}_i$ sur l'axe factoriel $\Delta\underline{u}$ un intervalle de valeurs possibles (ψ^{i-}, ψ^{i+}) avec :

$$\psi^{i-} = \sum\{x_{ij-}\varphi^j \mid j = 1, p; \varphi^j > 0\} + \sum\{x_{ij+}\varphi^j \mid j = 1, p; \varphi^j < 0\}$$
$$\psi^{i+} = \sum\{x_{ij-}\varphi^j \mid j = 1, p; \varphi^j < 0\} + \sum\{x_{ij+}\varphi^j \mid j = 1, p; \varphi^j > 0\}$$

Posant $\varepsilon_{ij} = 1$ ou -1, $\varepsilon_i = 1$ ou -1 et assimilant le signe $+$ à 1 et le signe $-$ à -1, les formules précédentes peuvent s'écrire sous la forme condensée suivante :

$$\psi^{i\varepsilon i} = \sum\{x_{ij\varepsilon ij}\varphi^j \mid j = 1,p;\ \varepsilon_{ij}\varepsilon_i\varphi^j > 0\}\ ^{(1)}$$

Notons que l'on a :

$$\psi^{i-} = \mathrm{Min}\Big\{\sum\{x_{ij\varepsilon ij}\varphi^j \mid j = 1,p\} \mid \varepsilon_{ij} = 1 \text{ ou } -1\Big\}$$
$$\psi^{i+} = \mathrm{Max}\Big\{\sum\{x_{ij\varepsilon ij}\varphi^j \mid j = 1,p\} \mid \varepsilon_{ij} = 1 \text{ ou } -1\Big\}$$

Remarque : A la composante principale $\underline{\Psi}$, on peut associer pour chaque individu i la combinaison linéaire :

$$f_i(x) = \sum\{\varphi^j f_{ij}(x) \mid j = 1,p\}$$

qui n'est ni une loi de probabilité, ni même une mesure positive (sauf si tous les φ^j sont positifs).

On peut alors représenter les n fonctions $f_i(x)$ et voir si elles sont bien séparées ou non, ce qui peut permettre une interprétation complémentaire et intéressante de l'axe factoriel considéré.

III.3.3. Cas des variables

Dans le cas des variables, l'abscisse de la projection de la colonne j du tableau G sur la composante principale normée $\underline{\Psi}/(\underline{\varphi}'\, B\, \underline{\varphi})^{1/2}$ est donnée par :

$$\theta_j = \sum\{p_i g_{ij}\psi^i \mid i = 1,n\}/(\underline{\varphi}'\, B\, \underline{\varphi})^{1/2}$$

qui n'est autre que la covariance interclasses entre la variable j et la composante principale normée.

Quand on considère la dispersion autour de g_{ij}, θ_j varie entre les deux bornes θ_{j-} et θ_{j+} données par :

$$\theta_{j-} = \sum\{p_i x_{ij\varepsilon ij}\psi^i \mid i = 1,n; \varepsilon_{ij}\psi^i < 0\}/(\underline{\varphi}'\, B\, \underline{\varphi})^{1/2}$$
$$\theta_{j+} = \sum\{p_i x_{ij\varepsilon ij}\psi^i \mid i = 1,n; \varepsilon_{ij}\psi^i > 0\}/(\underline{\varphi}'\, B\, \underline{\varphi})^{1/2}$$

formules qui se réduisent à une seule, en considérant $\varepsilon_j = 1$ ou -1 et où $\theta_{j\varepsilon j}$ est donné par les sommations précédentes en imposant que $\varepsilon_j\varepsilon_{ij}\psi^i$ soit positif.

(1) Quand ε_{ij} (resp. ε_i) est en indice ou en exposant, on notera εij (resp. εi) pour ne pas surcharger l'écriture.

Normant les quantités θ_{j-} et θ_{j+} par $(b_{jj})^{1/2}$ pour avoir l'équivalent d'une corrélation, on obtient autour de la corrélation r_{ψ_j} la plage de variation $(\gamma_{j-} = \theta_{j-}/(b_{jj})^{1/2}, \gamma_{j+} = \theta_{j+}/(b_{jj})^{1/2})$ dont les extrémités peuvent être extérieures à l'intervalle $(-1, 1)$. On prendra donc comme intervalle de variation autour de r_{ψ_j} l'intervalle (r_{j-}, r_{j+}) donné par :

$$r_{j-} = \text{Max}(-1, \gamma_{j-})$$
$$r_{j+} = \text{Min}(1, \gamma_{j+})$$

Remarques : 1) On aurait pu, ce qui semble plus logique, rechercher directement les extrêma du coefficient de corrélation (en fait du cosinus) entre la composante principale $\underline{\psi}$ et la variable j, suivant que l'individu i est en position basse (x_{ij-}) ou en position haute (x_{ij+}) pour la variable j, ce qui donne 2^n possibilités et donc 2^n cosinus. On sera donc ramené à rechercher le minimum et le maximum de :

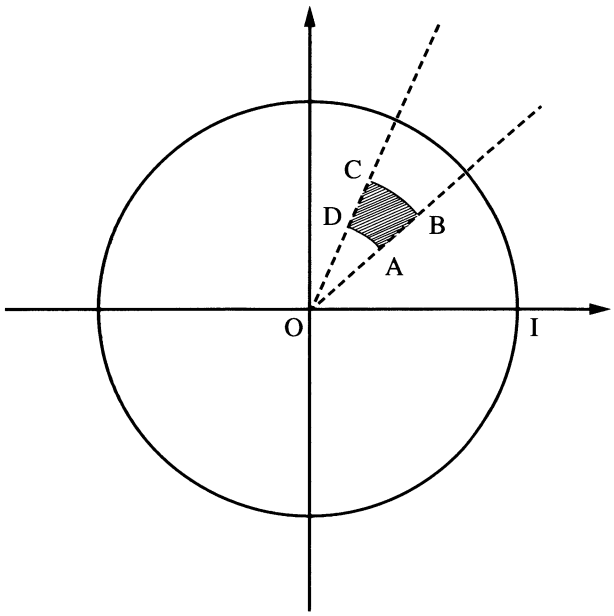
$$\frac{\sum\{p_i x_{ij\varepsilon ij}\psi^i \mid i = 1, n\}}{(\sum\{p_i(\psi^i)^2 \mid i = 1, n\} \sum\{p_i(x_{ij\varepsilon ij})^2 \mid i = 1, n\})^{1/2}} \tag{1}$$

On peut noter que la variable j associée à une suite donnée de n valeurs $\varepsilon_{ij}(1 \leqslant i \leqslant n)$ égales à 1 et -1 et qui a pour réalisations les $x_{ij\varepsilon ij}(1 \leqslant i \leqslant n)$ n'est pas en général centrée. C'est la raison pour laquelle, comme on l'a signalé ci-dessus, la quantité (1) n'est pas une corrélation, mais un cosinus.

On aurait pu aussi centrer la variable précédente, et considérer quand les ε_{ij} varient, les extrêma de la corrélation entre cette variable et la composante principale.

2) Dans le plan déterminé dans R^n par deux composantes principales, les représentations précédentes conduisent à représenter chaque variable par un rectangle. Ce rectangle n'étant pas forcément contenu à l'intérieur du cercle de centre l'origine et de rayon 1, on prendra l'intersection du rectangle et de ce cercle.

En fait la représentation précédente à partir de rectangles n'est peut être pas la plus astucieuse dans la mesure où dans une analyse factorielle sur données usuelles, les coordonnées polaires R_j, α_j d'une variable j à l'intérieur du cercle de corrélation sont au moins aussi intéressantes que ses coordonnées cartésiennes (*i.e.* ses corrélations avec les composantes principales) : R_j est en fait la corrélation multiple de la variable j par rapport aux deux composantes principales considérées. Il semble donc intéressant de considérer les valeurs extrêmales de R_j et de α_j (suivant la position basse ou haute de l'individu $i(1 \leqslant i \leqslant n)$ pour j) et de se servir des valeurs obtenues R_{j-}, R_{j+} pour R_j et α_{j-}, α_{j+} pour α_j pour représenter j à l'intérieur du cercle de corrélation. Cette représentation qui a été développée par Tang Ahanda (1998) dans sa thèse, se fait à l'intérieur du domaine délimité par deux arcs de cercle de rayons respectifs R_{j-} et R_{j+}, les angles définissant les extrémités de ces arcs de cercle étant α_{j-} et α_{j+} (*cf.* fig. 1).



Représentation de la dispersion de la variable j par l'intérieur du domaine *ABCD* dans le plan de deux composantes principales :

$$OA = OD = R_{j-} \; ; \; OB = OC = R_{j+}$$
$$(\overrightarrow{OI}, \overrightarrow{OA}) = \alpha_{j-} \; ; \; (\overrightarrow{OI}, \overrightarrow{OD}) = \alpha_{j+}$$

Figure 1

III.4. Autres analyses possibles

III.4 .1. Analyses basées sur les quantiles

On peut associer à chaque loi de probabilité f_{ij} un certain nombre de quantiles, et en déduire d'autres analyses ou d'autres représentations de la variabilité.

On peut par exemple considérer les 3 quartiles q_{1ij}, q_{2ij}, q_{3ij}, (q_{2ij} correspondant à la médiane) et les tableaux Q_1, Q_2 et Q_3 associés.

On peut alors construire les tableaux Q et R obtenus respectivement en accolant et en superposant les tableaux Q_k :

$$Q = (Q_1, Q_2, Q_3)$$
$$R = (Q_1', Q_2', Q_3')'$$

et l'on peut envisager les trois analyses suivantes :

analyse de G avec R et Q en supplémentaire

analyse de R avec G en supplémentaire

analyse de Q avec G en supplémentaire

Dans les deux premières analyses, chaque individu i est représenté dans l'espace R^p par quatre points $\underline{g}_i$, $\underline{q}_{1i}$, $\underline{q}_{2i}$ et $\underline{q}_{3i}$, $\underline{q}'_{1i}$ (resp. $\underline{q}'_{2i}$; $\underline{q}'_{3i}$) étant la $i^{\text{ème}}$ ligne de Q_1 (resp. Q_2; Q_3), ce qui permet sur un plan factoriel de visualiser la variabilité associée à l'individu i. Notons que dans la seconde analyse, on pourrait effectuer une analyse factorielle sous contrainte, en imposant que sur chaque axe factoriel, la représentation du point médian $\underline{q}_{2i}$ se situe à l'intérieur du segment joignant les représentations des premier et troisième quartiles $\underline{q}_{1i}$ et $\underline{q}_{3i}$.

De façon analogue, dans la première et dans la troisième analyse, chaque variable est représentée par quatre points, ce qui permet à l'intérieur d'un cercle de corrélation de visualiser la variabilité de chaque variable.

Il semble aussi intéressant de considérer le tableau médian Q_2 et de l'analyser en rajoutant les tableaux Q_1 et Q_3 d'une part en lignes supplémentaires et d'autre part en colonnes supplémentaires.

Dans le cas où tous les f_{ij} ont pour support un intervalle fini (x_{ij-}, x_{ij+}) et si Q_0 (resp. Q_4) désigne le tableau des x_{ij-} (resp. x_{ij+}), on peut considérer les tableaux P et Z suivants :

$$P = (Q_0, Q, Q_4) = (Q_0, Q_1, Q_2, Q_3, Q_4)$$
$$Z = (Q'_0, R', Q'_4)' = (Q'_0, Q'_1, Q'_2, Q'_3, Q'_4)'$$

et effectuer des analyses analogues à celles préconisées ci-dessus, en remplaçant Q et R par P et Z respectivement, certaines de ces analyses pouvant être faites sous contrainte, pour chaque individu (ou variable) de respecter l'ordre naturel induit par les $Q_k (0 \leqslant k \leqslant 4)$.

On peut encore analyser Q_2 en rajoutant en supplémentaires les tableaux P et Z auxquels on a retiré Q_2.

III.4.2. Analyses basées sur les écarts-type

On peut considérer le tableau E des écarts-type σ_{ij} et effectuer l'analyse factorielle (a priori non centrée) de ce tableau avec la métrique M dans R^p et la métrique des poids dans R^n. On pourra alors dans l'analyse précédente rajouter le tableau G en lignes et colonnes supplémentaires. On pourra aussi de façon symétrique rajouter E en lignes et colonnes supplémentaires dans l'analyse de G.

Dans le cas où l'on peut associer à chaque loi f_{ij} un intervalle (x_{ij-}, x_{ij+}) (cf. § III.3.1.) on peut aussi considérer le tableau L dont le terme général $l_{ij} = x_{ij+} - x_{ij-}$ correspond à la longueur de l'intervalle (x_{ij-}, x_{ij+}).

On peut alors soit analyser L avec G en lignes et colonnes supplémentaires, soit rajouter L en lignes et colonnes supplémentaires dans l'analyse de G.

F. Gioia (2000) propose une analyse sur le tableau des $(l_{ij})^{1/2}$ et cite également une méthodologie due à Lauro et Palumbo qui combine plusieurs analyses pour étudier simultanément la position et la dispersion associées aux données intervalles. Nous ne développerons pas ici ces analyses.

IV. Rappels et compléments sur les données de type intervalle

On suppose ici que toutes les lois f_{ij} sont de support borné, ce support étant défini par l'intervalle (x_{ij-}, x_{ij+}), et on conserve les mêmes notations que précédemment. On peut donc considérer le tableau où dans chaque case (i, j) on a l'intervalle (x_{ij-}, x_{ij+}). On va discuter des méthodes déjà proposées pour traiter ce type de données (méthodes des centres, des sommets) et suggérer d'autres analyses. Par contre, nous ne dirons rien ici des méthodes d'analyse déduites de l'algèbre des intervalles (matrices intervalles, valeurs propres intervalles, etc.) qui sont développées par F. Gioia (2000).

IV.1. Les nuages de points et les analyses associées

Dans l'espace R^p, on peut considérer les deux nuages de points suivants :

– Le nuage M_G qui est le nuage des points moyens $\underline{g}_i$ affectés des masses p_i ou nuage des centres.

– Le nuage M_S qui est le nuage des sommets :

A chaque individu (ou objet) i, on associe 2^p sommets, chaque sommet $S_{ki}(1 \leqslant k \leqslant 2^p)$ étant caractérisé par un vecteur $\underline{\varepsilon}_k = (\varepsilon_{k1}, \varepsilon_{k2}, \ldots, \varepsilon_{kp})'$ dont chaque composante est égale à 1 ou -1. Le sommet S_{ki} est alors le point de R^p de composantes $x_{ij\varepsilon kj}(1 \leqslant j \leqslant p)$.

On attribuera alors à S_{ki} le poids $p_i/2^p$ et on désignera par S le tableau $n2^p \times p$ associé aux $n \times 2^p$ sommets $S_{ki}(1 \leqslant k \leqslant 2^p, 1 \leqslant i \leqslant n)$ dans R^p.

Dans l'espace R^n, on considèrera également deux nuages :

– Le nuage N_G qui est le nuage des colonnes du tableau moyen G.

– Le nuage N_T qui est le nuage des sommets (variables) :

A chaque variable j, on associe 2^n sommets, chaque sommet $T_{kj}(1 \leqslant k \leqslant 2^n)$ étant caractérisé par un vecteur $\underline{\varepsilon}k = (\varepsilon_{k1}, \varepsilon_{k2}, \ldots, \varepsilon_{kn})'$ dont chaque composante vaut 1 ou -1. Le sommet T_{kj} est alors le point de R^n de composantes $x_{ij\varepsilon ki}(1 \leqslant i \leqslant n)$, et on désignera par T le tableau $n \times p2^n$ de composantes $x_{ij\varepsilon ki}(1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p, 1 \leqslant k \leqslant 2^n)$. Si l'espace R^p est muni d'une métrique associée à une matrice définie positive M de terme général $m_{jj'}$, on munira l'espace $R^{p'}$ (avec $p' = p2^n$) de la métrique associée à la matrice dont le terme général $m_{jk,j'k'}(1 \leqslant j, j' \leqslant p, 1 \leqslant k, k' \leqslant 2^n)$ est égal à $m_{jj'}/2^n$ si $k = k'$, et à zéro sinon.

On a alors le schéma de la figure 2 qui suggère les analyses suivantes :

1) L'analyse factorielle du tableau G (i.e. l'ACP du triplet (G, M, D_p)) avec les tableaux S et T en supplémentaire, qui a l'avantage de faire jouer (à part les problèmes de centrage et de pondération) un rôle symétrique aux individus et aux variables. Cette méthode correspond à la méthode des centres.

2) L'analyse factorielle du tableau S avec G en supplémentaire. Cette méthode correspond à la méthode des sommets dans le cas des individus (cf. Cazes et alt., 1997).

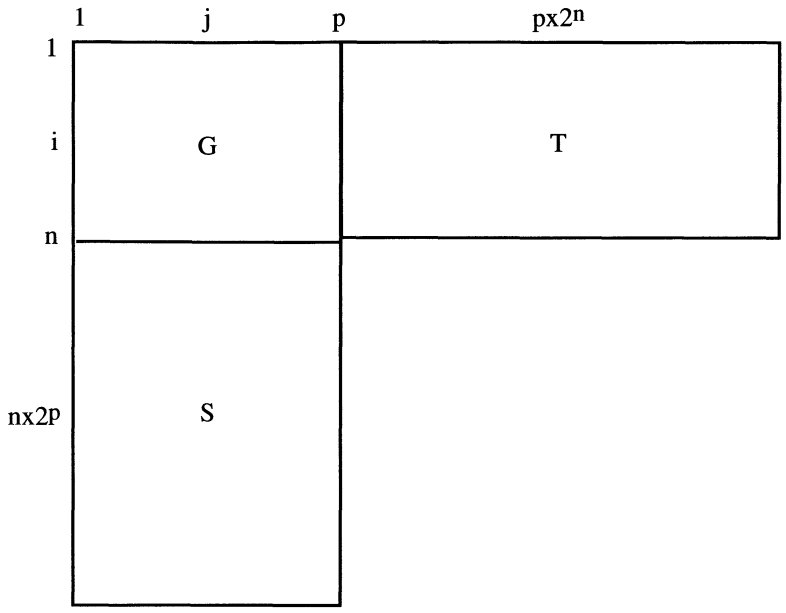


Figure 2

3) L'analyse factorielle du tableau T avec G en supplémentaire. Cette méthode qui correspond à la méthode des sommets dans le cas des variables, a été développée par Tang Ahanda (1998) dans sa thèse, ainsi que par Rodriguez-Rojas (2000).

IV.2. Etude des analyses précédentes

IV.2.1. Méthode des centres

Dans la méthode des centres envisagée par Cazes et alt. (1997) on prenait comme point moyen de l'individu i pour la variable j le centre de l'intervalle associé, soit $g_{ij} = (x_{ij+} + x_{ij-})/2$, ce qui suppose implicitement que la loi correspondante f_{ij} est symétrique. Nous nous placerons ici dans le cas d'une loi f_{ij} quelconque et donc g_{ij} ne sera pas en général le centre de l'intervalle précédent. Par contre nous supposerons toujours que le centre de gravité général est à l'origine $(g_j = \sum\{p_i g_{ij} \mid i = 1, n\} = 0, \forall j = 1, p)$.

Comme on l'a déjà dit (§ IV.1.) la méthode des centres est l'ACP du triplet (G, M, D_p) i.e. l'analyse factorielle des nuages M_G et N_G, les vecteurs axiaux factoriels, facteurs et composantes principales étant vecteurs propres respectivement de BM, MB et $W_G D_p$. La représentation de la variabilité chez les individus ou chez les variables se faisant, comme indiqué aux § § III.3.2. et III.3.3., à l'aide de rectangles dans le plan des axes factoriels ou à l'intérieur du cercle de corrélation, à partir des valeurs extrêmes des coordonnées ou des cosinus des sommets. On peut aussi dans le cas des variables, effectuer la représentation à l'aide d'arcs de cercle comme indiqué à la fin du § III.3.3. et sur la figure 1.

IV.2.2. Méthode des sommets (cas des individus)

Dans ce cas, on analyse le tableau S, et le nuage M_S. La matrice variance V associée à cette analyse a alors pour terme général (cf. Cazes et alt., 1997), si $V_{jj'}$ et $b_{jj'}$ désignent respectivement les termes généraux de V et B et si on suppose que $g_{ij} = (x_{ij+}+x_{ij-})/2$ (ce qui est en particulier réalisé si les lois f_{ij} sont symétriques) :

$$\forall j = 1,p : \ V_{jj} = b_{jj} + \sum\{p_i(x_{ij+} - x_{ij-})^2 \mid i = 1,n\}/4$$
$$\forall j,j' = 1,p; \ j \neq j' : \ V_{jj'} = b_{jj'}$$

ce qui revient à dire que la matrice variance intraclasses W est diagonale, son $j^{\text{ème}}$ terme diagonal étant égal à $\sum\{p_i(x_{ij+} - x_{ij-})^2 \mid i = 1,n\}/4$. L'expression précédente de V ne tient compte que de la configuration des sommets, et pas de l'expression des lois f_{ij}. Si l'on veut tenir compte de cette expression et si l'on suppose que conditionnellement à i les variables j et j' sont indépendantes, on a :

$$\forall j = 1,p : \ V_{jj} = b_{jj} + \sum\{p_i\alpha_{ij}(x_{ij+} - x_{ij-})^2 \mid i = 1,n\}$$
$$\forall j,j' = 1,p; \ j \neq j' : \ V_{jj'} = b_{jj'}$$

α_{ij} étant un coefficient positif ne dépendant que de la loi f_{ij}, celle-ci ayant été ramenée à avoir pour support $(-1/2, 1/2)$ (cf. Cazes et alt., op. cit.).

On peut alors représenter dans un plan factoriel chaque individu comme dans la méthode des centres par la projection de son point moyen à l'intérieur du rectangle associé sur chaque axe du plan, aux coordonnées des sommets extrêmes. Il serait plus intéressant en fait non pas de considérer les rectangles, mais les enveloppes convexes associées aux 2^p sommets caractérisant l'individu i.

Remarque : Comme on l'a souligné ci -dessus, la méthode des sommets suppose implicitement l'indépendance des variables conditionnellement à chaque individu. Si deux variables sont liées par une relation linéaire (ou de façon plus générale par une relation monotone) l'ensemble des sommets possibles associés à un individu n'est plus 2^p mais 2^{p-1}; en effet, supposons que les variables j et $j'(1 \leqslant j < j' \leqslant p)$ soient liées par une relation positive (par exemple $x_j = x_{j'}$), alors les sommets pour lesquels x_j est en position haute (resp. basse) et $x_{j'}$ en position basse (resp. haute) ne peuvent pas être atteints. La méthode des sommets, en considérant systématiquement 2^p sommets surestime donc la dispersion. De façon plus générale cette dispersion est d'autant plus surestimée qu'il y a un effet taille important.

On peut représenter les variables et leur dispersion comme indiqué au § III.3.3. à partir des composantes principales obtenues en projetant le nuage N_G en supplémentaire.

Si on veut représenter les variables à partir des éléments actifs de l'analyse, on peut, pour un axe factoriel donné, considérer la composante principale qu'on notera $\underline{\Psi}_S$ obtenue en projetant les $n\,2^p$ sommets individus (qui sont actifs) sur cet axe. Cette composante principale qui est donc une variable prenant $n\,2^p$ valeurs peut être considérée comme une variable intervalle. Il suffit d'associer à chaque individu i l'intervalle $(\Psi_S^{i-}, \Psi_S^{i+})$ (avec $\Psi_S^{i-} = \psi^{i-}, \Psi_S^{i+} = \psi^{i+}$, selon les

notations du § III.3.2.) correspondant aux projections extrêmes de ses 2^p sommets S_{ki} $(1 \leqslant k \leqslant 2^p)$. On peut alors utiliser la corrélation entre variables intervalle introduite par Rodriguez-Rojas (2000) :

Ψ_S étant une variable intervalle, désignons par T_Ψ l'ensemble de ses 2^n sommets dans R^n et par T_j l'ensemble des sommets $T_{kj} (1 \leqslant k \leqslant 2^n)$ de la variable j. Alors, on peut considérer l'intervalle de variation $(r_{\Psi_{j-}}, r_{\Psi_{j+}})$ pour la corrélation entre Ψ_S et j, avec :

$$\left.\begin{aligned} r_{\Psi_{j-}} &= \text{Min}\{r_{xy} \mid \underline{x} \in T_\Psi, \underline{y} \in T_j\} \\ r_{\Psi_{j+}} &= \text{Max}\{r_{xy} \mid \underline{x} \in T_\Psi, \underline{y} \in T_j\} \end{aligned}\right\} \qquad (2)$$

r_{xy} désignant la corrélation entre x et y.

Remarques : 1) Les variables de T_Ψ et de T_j ne sont pas en général centrées. On aurait donc pu remplacer dans les formules (2) les corrélations entre x et y (qui reviennent à calculer les cosinus entre x et y après centrage) par les cosinus entre x et y (sans centrer).

2) Dans le plan associé à deux composantes principales $\underline{\Psi}_{1S}$ et $\underline{\Psi}_{2S}$, on peut représenter la dispersion des variables par le domaine défini par l'intersection du cercle de corrélation avec le rectangle déterminé à partir des valeurs extrêmes fournies par (2) pour chacune des deux composantes principales considérées.

3) La plage de variation $r_{\Psi_{j-}}$, $r_{\Psi_{j+}}$ risque d'être importante et donc peu intéressante d'un point de vue pratique dans la mesure où dans (2) les extrêma du coefficient de corrélation r_{xy} entre x et y se calculent à la fois sur x et y, ce qui donne un domaine de recherche des extrêma $T_\Psi \times T_j$ très grand.

En comparaison, la plage de variation (ψ^{i-}, ψ^{i+}) d'un individu i sur un axe factoriel qui est obtenue à partir des abscisses extrêmes des projections des sommets associés à i correspond à la recherche de valeurs extrêmes quand on considère le domaine Si (et non plus un produit de domaines) des sommets de l'individu i.

4) Si deux individus sont identiques, l'ensemble des sommets possibles associés à une variable n'est plus 2^n mais 2^{n-1}. En considérant 2^n sommets on surestime donc la dispersion.

IV.2.3. Méthode des sommets (cas des variables)

Dans ce cas, on considère le tableau T à n lignes et $p2^n$ colonnes, et dans R^n les composantes principales sont vecteurs propres de la matrice $W_T D_p$, où la matrice W_T (cf. annexe) peut s'obtenir uniquement à partir des x_{ij+} et x_{ij-} sans avoir à balayer l'ensemble des 2^n sommets associés à chaque variable. De façon précise le terme général $(W_T)_{ii'}$ de W_T s'écrit, en posant $x^c_{ij} = (x_{ij+} + x_{ij-})/2$:

$$\forall i, i' = 1, n, i \neq i' : (W_T)_{ii'} = \sum \{m_{jj'} x^c_{ij} x^c_{i'j'} \mid j = 1, p; j' = 1, p\}$$

$$\forall i = 1, n : (W_T)_{ii} = \sum \{m_{jj'} x^c_{ij} x^c_{ij'} \mid j = 1, p; j' = 1, p\} +$$

$$(1/4) \sum \{m_{jj'} (x_{ij+} - x_{ij-})(x_{ij'+} - x_{ij'-}) \mid j = 1, p; j' = 1, p\}$$

Désignant par C la matrice $n \times p$ des x_{ij}^c, et posant :

$$W_C = CMC'$$

On a :

$$W_T = W_C + W_W$$

W_W désignant la matrice diagonale de $i^{\text{ème}}$ terme diagonal

$$(1/4)\sum\{m_{jj'}(x_{ij+} - x_{ij-})(x_{ij'+} - x_{ij'-}) \mid j = 1,p; j' = 1,p\}.$$

Notons que dans le cas où les lois f_{ij} sont symétriques, on a $C = G$ et donc $W_C = W_G$.

Sur un axe factoriel $\Delta\underline{u}$ de $R^{p'}$ (où $p' = p2^n$), chaque individu i est représenté de façon classique par l'abscisse de sa projection Ψ^i sur $\Delta\underline{u}$, le vecteur $\underline{\Psi}$ de composante $\Psi^i (1 \leqslant i \leqslant n)$ étant, rappelons-le, la composante principale qui est vecteur propre de $W_T D_p$.

La représentation de la variabilité pour l'individu i peut se faire de façon analogue à ce qui a été vu au § IV.2.2. Il suffit d'intervertir le rôle des individus et des variables et d'adapter les formules (2) en ne considérant plus des corrélations ou des cosinus, mais des abscisses de projection. Néanmoins cette façon de procéder a peu d'intérêt car elle risque de conduire à des plages de variation très grandes.

En ce qui concerne les variables, on peut adopter les représentations définies dans les remarques du § III.3.3. en recherchant les extrêma de (1) ou des représentations planes analogues à celles fournies par la figure 1. Il faut noter ici que $\underline{\Psi}$, contrairement aux § § III.2. et III.3., n'est pas une composante principale interclasses; par ailleurs $\underline{\Psi}$ (cf. remarque ci-dessous) n'est pas en général centrée; par contre le carré de sa norme (pour la métrique des poids D_p) est égal à la valeur propre associée à l'axe considéré.

Remarques : 1) Contrairement aux tableaux G et S, le tableau T n'est pas centré, et l'analyse factorielle préconisée ci-dessus est une analyse non centrée, afin de pouvoir comparer les analyses de T et G.

T étant non centré, son analyse risque de fournir un premier axe contenant un pourcentage d'explication très important, axe relativement trivial car joignant l'origine au point moyen de T. Les axes suivants risquent alors d'être peu interprétables car noyés dans le bruit du premier axe. On peut aussi centrer T, ce qui permet d'avoir une analyse plus classique, mais non comparable de façon aussi évidente à celle de S. Par contre on peut en centrant parler de corrélation (et non plus de cosinus), ce qui est plus parlant pour un statisticien.

2) Comme on l'a déja dit (cf. remarque 4) à la fin du paragraphe précédent), si deux individus sont identiques (ou proportionnels), l'ensemble des sommets possibles associés à une variable n'est plus 2^n mais 2^{n-1}. La méthode des sommets variables, en considérant dans tous les cas de figure 2^n sommets pour chaque variable surestime donc la dispersion. Par contre, si deux variables sont identiques, leurs sommets sont

confondus, ce qui ne pose pas de problème, contrairement à ce qui se passe dans la méthode des sommets individus.

IV.3. Autres analyses

IV.3.1. Analyses normées

Ces analyses sont obtenues en prenant $M = (\mathrm{Diag}(B))^{-1}$ dans la méthode des centres, et $M = (\mathrm{Diag}(V))^{-1}$ dans la méthode des sommets (individus ou variables).

Dans le premier cas, cela revient à faire l'analyse factorielle usuelle (i.e. avec la métrique usuelle $M = \mathrm{Id}_p$) du tableau de terme général $g_{ij}/(b_{jj})^{1/2}$, tandis que dans les deux méthodes des sommets, cela revient à raisonner sur le tableau $S_n(n2^p \times p)$ de terme général $x_{ij\varepsilon kj}/(V_{jj})^{1/2}(1 \leqslant k \leqslant 2^p)$ dans le cas des sommets individus avec la métrique usuelle de R^p et sur le tableau $T_n(n \times p2^n)$ de terme général $x_{ij\varepsilon ki}/(V_{jj})^{1/2}(1 \leqslant k \leqslant 2^n)$ dans le cas des sommets variables avec la métrique $\mathrm{Id}_{p'}/2^n$ dans $R^{p'}$ (avec $p' = p2^n$). Notons que dans ces deux cas, le tableau moyen correspondant est le tableau $G_n(n \times p)$, de terme général $g_{ij}/(V_{jj})^{1/2}$, associé à la métrique usuelle de R^p.

Remarque : Comme T, le tableau T_n n'est pas centré. De plus il n'est pas réduit (i.e. le carré de la norme de chacune de ses colonnes pour la métrique D_p n'est pas égal à 1). On peut donc envisager l'analyse du tableau réduit ou du tableau centré réduit associé à T (ou T_n); mais ces analyses ne sont pas directement comparables à celles des tableaux G_n et S_n.

IV.3.2. Analyse discriminante

Si l'on veut essayer de discriminer au maximum les individus de telle sorte que sur un plan factoriel les rectangles associés aux différents individus se recoupent le moins possible, on fera l'analyse discriminante du tableau S, ce qui revient à faire l'analyse factorielle du tableau G avec la métrique $M = W^{-1}$. Compte tenu de la structure symétrique des hypercubes associés à l'ensemble des sommets d'un individu, la matrice W est diagonale, comme on l'a déjà remarqué au § IV.2.2. L'analyse discriminante est alors équivalente à l'analyse factorielle usuelle du tableau de terme général $g_{ij}/(w_{jj})^{1/2}$ comme on l'a déjà signalé à la fin du § III.1.

IV.3.3. Statis

Le tableau S des sommets peut se mettre sous la forme :

$$S = (S_1', S_2', , S_n')'$$

le tableau S_i (à 2^p lignes et p colonnes) étant le tableau donnant les coordonnées des 2^p sommets associés à l' individu i.

On peut donc envisager d'appliquer la méthode STATIS-DUAL (cf. Lavit (1988) , Dazy-Le Barzic (1996)).

La matrice variance V_i associée à S_i est, du fait de la structure parallélépipédique des sommets, diagonale, de $j^{\text{ème}}$ terme diagonal $a_{ij} = (x_{ij+} - x_{ij-})^2/4$.

Le terme général $c_{ii'}$ de la matrice C des produits scalaires (en supposant R^p muni de la métrique usuelle) entre les matrices variances V_i et $V_{i'}$ s'écrit alors :

$$c_{ii'} = \text{Trace}(V_i V_{i'}) = \sum\{a_{ij}a_{i'j} \mid j = 1, p\}$$

Soit :

$$C = AA'$$

A étant la matrice de terme général $a_{ij}((1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p)$.

La diagonalisation de C permet alors à partir des vecteurs propres associés aux plus grandes valeurs propres d'obtenir une représentation des V_i, donc des individus dans un plan ou dans un espace de faible dimension.

Par contre si $\underline{u}$ désigne le vecteur propre normé de C associé à sa plus grande valeur propre, vecteur dont toutes les composantes ont le même signe, que l'on choisira positif, et si u_i désigne la $i^{\text{ème}}$ composante de $\underline{u}$, le compromis $\sum\{u_i V_i \mid i = 1, n\}$ ne possède que peu d'intérêt pour représenter les sommets puisqu'il est, comme les V_i diagonal; la représentation se fait en effet dans le plan des deux variables associées aux deux termes diagonaux les plus importants de ce compromis.

IV.3.4. L'analyse factorielle multiple (AFM) (Escofier-Pagès, 1998)

On va effectuer l'AFM duale du tableau S dans la mesure où l'on ne considère pas comme dans l'AFM usuelle des groupes de variables dans R^n, mais des groupes d'individus (en fait des groupes de sommets) dans R^p. Dans cette AFM, on fait l'ACP du tableau déduit de S en pondérant chaque S_i par l'inverse de la racine carrée de la plus grande valeur propre issue de l'ACP de S_i, ce qui revient à pondérer S_i par $1/(\max_j a_{ij})^{1/2}$ puisque $V_i = \text{Diag}(a_{ij})$ (avec les notations du paragraphe IV.3.3). Puisque $a_{ij} = (x_{ij+} - x_{ij-})^2/4$, ceci revient à ramener à deux la longueur du plus grand intervalle associé à l'individu i.

En fait, dans la façon de procéder ci-dessus, on a supposé implicitement que chaque tableau S_i était centré (le centrage se faisant par rapport au centre de l'hypercube associé à S_i) de façon à avoir une ACP classique (i.e centrée), ce qui implique la diagonalisation de la matrice variance diagonale V_i. L'analyse des sommets considérée est donc ici une analyse intraclasses, tandis que l'AFM est une analyse intraclasses pondérée.

IV.3.5. Autres analyses déduites du tableau S

Soit $x_{ij\varepsilon kj}(1 \leqslant j \leqslant p)$ la ligne du tableau S (de dimensions $n2^p \times p$) associée au $k^{\text{ième}}$ $(1 \leqslant k \leqslant 2^p)$ sommet de l'individu $i(1 \leqslant i \leqslant n)$, ce sommet étant caractérisé par le vecteur dont les composantes $\varepsilon_{kj}(1 \leqslant j \leqslant p)$ sont égales à 1 ou -1.

On peut disposer de façon différente les éléments du tableau S de façon à obtenir un tableau U_1 à n lignes et $p2^p$ colonnes (cf. figure 3 où $n = 3$ et $p = 2$). Il est alors facile de voir en cumulant les colonnes identiques de U_1 que l'analyse de U_1

Tableau S', transposé de S dans le cas où $p = 2$ et $n = 3$
(on a noté x pour x_1 et y pour x_2)

	1er individu				2ème individu				3ème individu			
k	1	2	3	4	1	2	3	4	1	2	3	4
x	x_{1+}	x_{1+}	x_{1-}	x_{1-}	x_{2+}	x_{2+}	x_{2-}	x_{2-}	x_{3+}	x_{3+}	x_{3-}	x_{3-}
y	y_{1+}	y_{1-}	y_{1+}	y_{1-}	y_{2+}	y_{2-}	y_{2+}	y_{2-}	y_{3+}	y_{3-}	y_{3+}	y_{3-}
	$S'1$				$S'2$				$S'3$			

Tableau U_1 à $n = 3$ lignes et $p2^p = 8$ colonnes
obtenu par redisposition des éléments de S

$k=1$		$k=2$		$k=3$		$k=4$	
x	y	x	y	x	y	x	y
x_{1+}	y_{1+}	x_{1+}	y_{1-}	x_{1-}	y_{1+}	x_{1-}	y_{1-}
x_{2+}	y_{2+}	x_{2+}	y_{2-}	x_{2-}	y_{2+}	x_{2-}	y_{2-}
x_{3+}	y_{3+}	x_{3+}	y_{3-}	x_{3-}	y_{3+}	x_{3-}	y_{3-}

Tableau $U/2^{p-1} = U/2$, dont l'analyse est équivalente à celle de U_1,
U étant déduit de U_1 par cumul des colonnes identiques de U_1

x_{1+}	y_{1+}	x_{1-}	y_{1-}
x_{2+}	y_{2+}	x_{2-}	y_{2-}
x_{3+}	y_{3+}	x_{3-}	y_{3-}

Figure 3

est identique à celle du tableau U de dimensions $n \times 2p$, dont le terme général $u_{ij\varepsilon}$ (avec $\varepsilon = 1$ ou -1) est donné par :

$$u_{ij+} = 2^{p-1} x_{ij+}$$
$$u_{ij-} = 2^{p-1} x_{ij-}$$

Il s'agit donc (au facteur 2^{p-1} près) du tableau où on a remplacé chaque variable x_j par deux variables, la variable maximale x_{j+} et la variable minimale x_{j-}.

On peut aussi, au lieu de superposer les tableaux S_i, pour obtenir S, les juxtaposer ce qui donne le tableau Z à 2^p lignes et np colonnes suivant :

$$Z = (S_1, S_2, ., S_n)$$

tableau que l'on peut analyser, mais dont l'analyse semble à priori moins intéressante que celle du tableau U

Remarques : 1) On peut opérer de façon analogue avec le tableau des sommets variables T. L'équivalent du tableau U est alors un tableau Y de dimensions $2n \times p$ et de terme général $y_{ij\varepsilon}$ (avec $\varepsilon = 1$ ou -1) tel que :

$$y_{ij+} = 2^{n-1} x_{ij+}$$
$$y_{ij-} = 2^{n-1} x_{ij-}$$

Il s'agit (au facteur 2^{n-1} près) d'un sous-tableau de S où au lieu de considérer les 2^p sommets associés à un individu, on ne considère que les deux sommets extrêmes, celui pour lequel l'individu est en position haute pour toutes les variables et celui pour lequel il est en position basse.

2) Comme le tableau T, le tableau U n'est pas centré. On aura ici intérêt à centrer (et le cas échéant réduire) le tableau U, puisque son analyse n'est pas directement comparable à celles des tableaux S et G.

3) Les tableaux U et Y sont à une constante près des sous-tableaux des tableaux P et Z considérés au § III.4.1.

IV.4. Conclusion

Dans toutes les méthodes précédentes, la représentation des variables et en particulier de l'imprécision associée pose problème dans la mesure où le tableau T des sommets associés aux variables dans R^n n'est pas centré, même si les tableaux G et S le sont. De même, si on norme les variables (soit par l'écart type interclasses $(b_{jj})^{1/2}$, soit par l'écart type total $(V_{jj})^{1/2}$ pour la variable j), le tableau T n'est pas normé, ce qui peut aussi poser problème. Par ailleurs, comme on l'a déjà dit, les plages de variation, ou les domaines proposés pour représenter la dispersion des variables risquent d'être très grands, ce qui limite leur intérêt.

Nous préconisons pour avoir une analyse symétrique d'effectuer l'analyse du tableau G avec les tableaux S et T en supplémentaire. Les composantes principales associées étant centrées, la représentation des sommets des variables (i.e. des colonnes du tableau T) sur ces composantes principales normées est indépendante du fait qu'on ait ou non centré T. Par contre, si on représente ces sommets, à l'intérieur du cercle de centre l'origine et de rayon 1, on obtient en centrant des corrélations qui en valeur absolue sont supérieures aux cosinus que l'on aurait obtenus sans centrer.

Au niveau de la variabilité, la meilleure représentation dans un plan nous semble être l'enveloppe convexe de la projection des sommets individus dans R^p et variables dans R^n. A défaut, la représentation de l'imprécision par des rectangles dans R^p et à partir d'arcs de cercle dans R^n comme préconisé sur la figure 1, nous semble la plus indiquée.

Néanmoins, si des variables (ou des individus) sont liées , les représentations précédentes surestiment la dispersion puisque ces représentations supposent implicitement qu'il n'y a pas de liaison.

En conclusion, il nous semble important d'expérimenter sur de nombreux lots de données intervalles afin d'affiner la stratégie d'analyse et d'en déduire un programme convivial et utile pour les praticiens.

Annexe : Calcul de W_T

La matrice N associée à la métrique dans $R^{p'}$ (avec $p' = p2^n$) étant la matrice diagonale par blocs de $k^{\text{ème}}$ bloc diagonal $M/2^n (1 \leqslant k \leqslant 2^n)$ (soit en notations tensorielles : $N = 2^{-n}\, M \otimes \mathrm{Id}_a$ (avec $a = 2^n$), Id_a étant la matrice unité d'ordre a) le terme général $(W_T)_{ii'}$ de la matrice W_T qui est égale à TNT' s'écrit :

$$(W_T)_{ii'} = (2^{-n}) \sum \{m_{jj'} x_{ij\varepsilon ki} x_{i'j'\varepsilon i'ki'} \mid k = 1, 2^n; j = 1, p; j' = 1, p\},$$

avec $\varepsilon_{ki} = 1$ ou -1 et de même pour $\varepsilon_{ki'}$, les signes de ε_{ki} et $\varepsilon_{ki'}$ étant déterminés par le sommet k considéré.

Pour $i \neq i'$, le produit des coordonnées de chacun des quatre sommets du quadrilatère défini par (x_{ij-}, x_{ij+}) et $(x_{i'j'-}, x_{i'j'+})$ apparaît 2^{n-2} fois d'où après mise en facteurs :

$$(W_T)_{ii'} = (2^{n-2}/2^n) \sum \{m_{jj'}(x_{ij-} + x_{ij+})(x_{i'j'-} + x_{i'j'+}) \mid j = 1, p; j' = 1, p\}$$

Soit : $(W_T)_{ii'} = \sum \{m_{jj'} x^c_{ij} x^c_{i'j'} \mid j = 1, p; j' = 1, p\}$

en posant $x^c_{ij} = (x_{ij-} + x_{ij+})/2$ et de même pour $x^c_{i'j'}$.

Raisonnant de façon analogue pour $i = i'$, on obtient :

$$(W_T)_{ii} = (1/2) \sum \{m_{jj'}(x_{ij+} x_{ij'+} + x_{ij-} x_{ij'-}) \mid j = 1, p; j' = 1, p\}$$

Compte de l'égalité $aa' + bb' = (1/2)[(a+b)(a'+b') + (a-b)(a'-b')]$, vérifiée par tout système de nombre réels a, a', b, b', $(W_T)_{ii}$ s'écrit encore :

$$(W_T)_{ii} = \sum \{m_{jj'} x^c_{ij} x^c_{ij'} \mid j = 1, p; j' = 1, p\} + (1/4) \sum \{m_{jj'}(x_{ij+} - x_{ij-})(x_{ij'+} - x_{ij'-}) \mid j = 1, p; j' = 1, p\}$$

Remerciements

Je remercie Ludovic Lebart qui m'a incité à publier cet article dans la Revue de Statistique Appliquée, et Jérôme Pagès pour l'avoir relu et fourni des remarques pour l'améliorer.

Bibliographie

(1) ABOA Y. J. P. (2002), Méthodes de segmentation sur un tableau de variables aléatoires, Thèse, Université Paris 9 Dauphine.

(2) BOUMAZA R. (1998), Analyses en composantes principales de distributions gaussiennes multidimensionnelles, RSA, vol. 46, n° 2, pp. 5-20.

(3) CAZES P., CHOUAKRIA A., DIDAY E., SCHEKTMAN Y. (1997), Extension de l'analyse en composantes principales à des données de type intervalle, RSA, vol. 45 n° 3, pp. 5-24.

(4) DAZY F., LE BARZIC J.F. (1996), L'analyse des données évolutives. Méthodes et applications, GERI, Ed. Technip, 250 pages.

(5) EMILION R. (2001), Clustering and mixtures of stochastic processes, Cahier du Ceremade no 0111.

(6) ESCOFIER B., PAGÈS J. (1998), Analyses factorielles simples et multiples, Dunod, 300 pages.

(7) GIOIA F. (2000), Factorial Analysis for Interval Data (séminaire LISE-CEREMADE), novembre 2000.

(8) LAURO N.C., VERDE R., PALUMBO F. (2000), Factorial discriminant analysis on symbolic objects, in "Analysis of symbolic data. Exploratory methods for extracting statistical information from complex data", édité par H.H. BOCK et E. DIDAY, pp. 212, 233, Springer-Verlag.

(9) LAVIT C. (1988), Analyse conjointe de tableaux quantitatifs, Masson, 254 pages.

(10) RODRIGUEZ ROJAS O. (2000), Classification et modèles linéaires de l'analyse des données symboliques, Thèse, Université Paris 9 Dauphine.

(11) TANG AHANDA B. (1998), Extension de méthodes d'analyse factorielle sur des données symboliques, Thèse, Université Paris 9 Dauphine.