REVUE DE STATISTIQUE APPLIQUÉE

B. FALISSARD

Déploiement d'une matrice de corrélation sur la sphère unité de R 3

Revue de statistique appliquée, tome 43, nº 2 (1995), p. 35-48

<http://www.numdam.org/item?id=RSA_1995__43_2_35_0>

Rev. Statistique Appliquée, 1995, **XLIII** (2), 35-48

DÉPLOIEMENT D'UNE MATRICE DE CORRÉLATION SUR LA SPHÈRE UNITÉ DE $\Re^3$

B. Falissard

Unité CNRS EP 53
Pavillon Clérambault Hôpital de la Salpétrière
47 Boulevard de l'Hôpital 75651 Paris cedex 13 France

RÉSUMÉ

Une analyse en composantes principales (ACP) est fréquemment utilisée pour réduire le nombre de dimensions d'un problème. Souvent, l'analyse se fait à partir de la matrice de corrélation des variables et l'ACP autorise une représentation graphique de ces dernières en deux, trois dimensions, voire davantage. Le fait de partir d'une matrice de corrélation permet aux variables, normalisées, d'être situées sur une hypersphère de rayon un. Le présent article tire profit de cette propriété pour représenter les variables sur une sphère unité de dimension trois plutôt que sur un sous espace. Il sera montré que l'information contenue dans une telle représentation est comparable a celle d'une ACP retenant trois dimensions, alors que la lisibilité s'en trouve très améliorée. Les détails mathématiques suivis d'un exemple pratique sont proposés.

Mots-clés : *Analyse en composantes principales, Analyse multivariée non linéaire, représentation en 3 dimensions.*

SUMMARY

Principal component Analysis (PCA) is frequently used to reduce the number of dimensions of a problem. Most often, the analysis is carried on the correlation matrix of the variables and PCA provides a graphical representation of them in two, three, or even more dimensions. The use of a correlation matrix leads to a particular position of the variables : being normalised, they are on a unit hypersphere. Taking advantage of this position, the present paper suggests to represent them on a unit 3-dimensional sphere instead of a subspace. It is shown that the information contained in such a representation is comparable to the information provided by a 3-dimensional PCA, while the readability is highly improved. Mathematical details are given like a real example.

Keywords : *Principal component analysis, Principal curves, Non-linear multivariate analysis, 3-dimensional representation.*

1. Introduction

Dans de nombreuses situations expérimentales, p variables $(x_i)_{1 \leqslant i \leqslant p}$ sont mesurées sur n individus et une question centrale est de révéler les relations entre les x_i. Dans un tel cas, où les x_i peuvent être considérés comme p éléments de $\Re^n$, une analyse en composantes principales (ACP) est intéressante puisque cette méthode autorise la représentation des variables dans un espace de dimension réduite F, tout en préservant autant que possible leur variabilité globale. Souvent, quand les variables sont exprimées dans des unités hétérogènes ou ont des variances très différentes, les x_i sont normalisés et l'ACP est faite sur leur matrice de corrélation. Cela correspond à une propriété géométrique particulière : les x_i sont situés sur l'hypersphère unité centrée sur l'origine de $\Re^n$ (Lebart, Morineau et Fènelon, 1982, p.281).

Si l'on prend en compte une telle propriété, il serait plus naturel, pour représenter fidèlement les p variables x_i, de les projeter sur une sphère unité O plutôt que sur un sous espace vectoriel F. Le but de cet article est de trouver la sphère unité O de dimension 3, centrée sur l'origine, pour laquelle les projections des x_i sont aussi proches que possible des x_i (au sens des moindres carrés). Les calculs et algorithmes nécessaires à l'obtention de O sont présentés, suivis d'un exemple fondé sur une épreuve psychométrique (échelle d'hétéroévaluation en psychiatrie). Des simulations permettront ensuite de comparer la précision d'une représentation sphérique à celle d'une ACP retenant deux ou trois dimensions. En conclusion, une discussion présentera les autres méthodes non linéaires permettant de représenter des données multivariées.

2. Une représentation sphérique d'une matrice de corrélation

2.1. Généralités

Considérons un ensemble de p variables $(x_i)_{1 \leqslant i \leqslant p}$ mesurées sur n individus : $x_i = (x_{i1}, \ldots, x_{in})'$. Les x_i sont normalisés, c'est-à-dire :

$$x_{iq} = (h_{iq} - m_i) \,/\, \left(n^{1/2} s_i\right), \quad 1 \leqslant i \leqslant p, \;\; 1 \leqslant q \leqslant n$$

où h_{iq} est la $i^{\text{ème}}$ réponse mesurée sur le $q^{\text{ème}}$ individu, m_i et s_i sont les moyennes et écart-types observés obtenus à partir de l'échantillon. Nous avons ainsi :

$$R = (r_{im})_{1 \leqslant i \leqslant p,\; 1 \leqslant m \leqslant p} = (x_1, \ldots, x_p)'(x_1, \ldots, x_p),$$

ce qui implique que $||x_i||^2 = 1$: les x_i sont sur l'hypersphère unité de $\Re^n$. Si O est une sphère unité de dimension 3, la projection de x_i sur O est notée par $\text{proj}(x_i, O)$ et est définie par :

$$||x_i - \text{proj}(x_i, O)|| \underset{o \in O}{=} \min ||x_i - o||$$

où $||\;||$ désigne la norme euclidienne de $\Re^n$. En d'autres termes, $\text{proj}(x_i, O)$ est l'élément de O le plus proche de x_i. Le but de cette partie est de trouver la sphère

unité O° qui correspond à l'ajustement des moindres carrés des x_i. Plus précisément, O° est l'élément de Ω vérifiant :

$$\sum_{i=1}^{i=p} \|x_i - \text{proj}(x_i, O^\circ)\|^2 = \min_{O \in \Omega} \left(\sum_{i=1}^{i=p} \|x_i - \text{proj}(x_i, O)\|^2 \right)$$

où Ω désigne l'ensemble des sphères unités de dimension 3 centrées sur l'origine de $\Re^n$; $\text{proj}(x_i, O)$ sera maintenant noté o_i.

Il est possible de formuler ce problème de sorte que les difficultés numériques soient considérablement réduites. La sphère O° peut être obtenue à partir d'un sous espace de dimension 3, T°, vérifiant :

$$\sum_{i=1}^{i=p} ||\text{proj}(x_i, T^\circ)|| = \max_{T \in \Theta} \left(\sum_{i=1}^{i=p} ||\text{proj}(x_i, T)|| \right)$$

où Θ est l'ensemble des sous espaces de dimension 3 de $\Re^n$; $\text{proj}(x_i, T^\circ)$ sera maintenant noté t_i. Il est remarquable qu'une propriété presque similaire définit le sous espace de dimension 3, T^*, correspondant aux 3 premières composantes principales de R :

$$\sum_{i=1}^{i=p} ||\text{proj}(x_i, T^*)||^2 = \max_{T \in \Theta} \left(\sum_{i=1}^{i=p} ||\text{proj}(x_i, T)||^2 \right)$$

Ici, une somme de carrés de normes remplace une somme de normes.

La relation reliant t_i a o_i est présentée au 2.2., alors que t_i est calculé au 2.3. Les considérations numériques sont présentées au 2.4.

2.2. Théorème : $o_i = t_i/||t_i|| \quad (1 \leqslant i \leqslant p)$

Soit O un élément de Ω et T le sous espace de dimension 3 qui le contient. Montrons dans un premier temps que :

$$\text{proj}(x_i, O) = \text{proj}(x_i, T)/||\text{proj}(x_i, T)|| \tag{1}$$

Pour plus de simplicité, adoptons pour quelques lignes les notations :

$$p_i = \text{proj}(x_i, T) \text{ et } q_i = x_i - p_i.$$

Nous avons,

$$\begin{aligned} ||x_i - p_i/||p_i||\,||^2 &= ||q_i + p_i - p_i/||p_i||\,||^2 \\ &= ||q_i||^2 + ||p_i - p_i/||p_i||\,||^2 \\ &= ||q_i||^2 + ||p_i||^2 - 2||p_i|| + 1 \end{aligned}$$

et de même, pour $o \in O$,

$$||x_i - o||^2 = ||q_i + p_i - o||^2 = ||q_i||^2 + ||p_i||^2 - 2p_i'o + 1.$$

Or, puisque $||o|| = 1, p_i'\, o \leqslant ||p_i||$ d'où $||x_i - p_i/||p_i||\,||^2 \leqslant ||x_i - o||^2$, ce qui prouve (1).

Finalement,

$$\begin{aligned} ||x_i - \text{ proj}(x_i, O)||^2 &= ||x_i||^2 - 2x_i' \text{ proj}(x_i, O) + ||\text{proj}(x_i, O)||^2 \\ &= 1 - 2x_i' \text{ proj}(x_i, T)/||\text{proj}(x_i, T)|| + 1 \\ &= 2(1 - ||\text{proj}(x_i, T)||). \end{aligned}$$

Ce qui prouve que minimiser : $\displaystyle\sum_{i=1}^{i=p} ||x_i - \text{proj}(x_i, O)||^2$

revient a maximiser : $\displaystyle\sum_{i=1}^{i=p} ||\text{proj}(x_i, T)||$

ainsi, via (1) : $p(x_i, O^\circ) = p(x_i, T^\circ)/||p(x_i, T^\circ)||$ soit $o_i = t_i/||t_i||$

2.3. Comment obtenir t_i

Chaque sous espace T peut être défini par une base orthonormée de 3 vecteurs (v_1, v_2, v_3) qui peut être complétée par $v_4, \ldots, v_n$ pour obtenir une base orthonormée V de $\Re^n$. La question est ici de trouver le triplet (v_1, v_2, v_3) qui maximise :

$$f(v_1, v_2, v_3) = \sum_{i=1}^{i=p} ||\text{proj}(x_i, T)|| = \sum_{i=1}^{i=p} \left\{(x_i'v_1)^2 + (x_i'v_2)^2 + (x_i'v_3)^2\right\}^{1/2}$$

Cela peut se faire au moyen du lagrangien

$$L = f(v_1, v_2, v_3) - \sum_{i=1}^{i=6} \lambda_i g_i(v_1, v_2, v_3)$$

où les λ_i sont les multiplicateurs de Lagrange et les g_k les contraintes :

$$\begin{gathered} g_k(v_1, v_2, v_3) = ||v_k||^2 - 1 = 0 \quad (k = 1, 2, 3) \quad g_4(v_1, v_2, v_3) = v_1 v_2 = 0\,; \\ g_5(v_1, v_2, v_3) = v_2 v_3 = 0\,; \quad g_6(v_1, v_2, v_3) = v_1 v_3 = 0 \end{gathered} \tag{2}$$

Si les coordonnées initiales des v_k sont notées $(v_{k1}, \ldots, v_{kn})$, nous avons :

$$\partial L/\partial v_{kq} = 0 \quad (k = 1, 2, 3; \quad 1 \leqslant q \leqslant n) \tag{3}$$

Les $3n + 6$ équations correspondant à (2) et (3) peuvent être réduites à un système de $3p$ équations :

$$h_{ik} = \sum_{m=1}^{m=p} t_{mk}(t_{m1}t_{i1} + t_{m2}t_{i2} + t_{m3}t_{i3} - r_{mi}) \\ \left\{(t_{m1})^2 + (t_{m2})^2 + (t_{m3})^2\right\}^{-1/2} = 0, \qquad (4)$$

$$(1 \leqslant i \leqslant p; \quad k = 1, 2, 3; \quad \textit{cf.} \text{ appendice pour plus de détails})$$

où r_{mi} est le $(m, i)^{\text{ème}}$ élément de la matrice de corrélation R et $t_{mk} = x'_m v_k (m = 1, \ldots, p; k = 1, 2, 3)$. On peut en déduire $t_i = (t_{i1}, t_{i2}, t_{i3}, 0, \ldots, 0)'$ en utilisant des méthodes de calcul numérique habituelles.

2.4. Considérations numériques

Les dérivées de $h_{i,k}$ étant facilement calculables, il est suggéré de résoudre (4) au moyen de la méthode de Newton-Raphson. Les valeurs initiales $t_{ik}(0)$ sont obtenues via les 3 premières composantes principales : si $u_1 = (u_{11}, \ldots, u_{p1})'$, $u_2 = (u_{12}, \ldots, u_{p2})'$, $u_3 = (u_{13}, \ldots, u_{p3})'$ sont les vecteurs propres normés correspondant aux 3 valeurs propres de R les plus élevées (l_1, l_2, l_3) alors :

$$t_{ik}(0) = u_{ik}(l_k)^{1/2} \quad (k = 1, 2, 3; \quad 1 \leqslant i \leqslant p)$$

Comme mentionné au § 2.1, puisque T° et T^* ont des propriétés voisines, les valeurs initiales des t_i devraient être très proches de leur valeur véritable. Ce point est montré au § 4, nous verrons d'ailleurs que les $t_{ik}(0)$ à eux seuls sont d'une précision suffisante, tout au moins pour une représentation graphique des o_i.

Finalement, si $o_i = (o_{i1}, o_{i2}, o_{i3}, 0, \ldots, 0)'$, les o_{ik} sont obtenus à partir de :

$$o_{ik} = t_{ik}/[(t_{i1})^2 + (t_{i2})^2 + (t_{i3})^2]^{1/2} \quad (k = 1, 2, 3; \quad 1 \leqslant i \leqslant p)$$

Comme nous le voyons figure 1, la représentation graphique des o_i est immédiate. Deux faces opposées de la sphère O° sont projetées sur deux plans :

– si $o_{i1} > 0$, (o_{i2}, o_{i3}) est représenté sur le premier plan, arbitrairement appelé vue de face;

– si $o_{i1} < 0$, (o_{i2}, o_{i3}) est représenté sur l'autre plan, la vue de dos.

L'impression de relief est rendue par le dessin de méridiens et de parallèles. L'orientation des plans de projection est arbitraire; au besoin, il est possible au moyen de rotations autour des 3 vecteurs v_1, v_2, v_3, de les choisir de telle sorte que la représentation apparaisse le plus clairement.

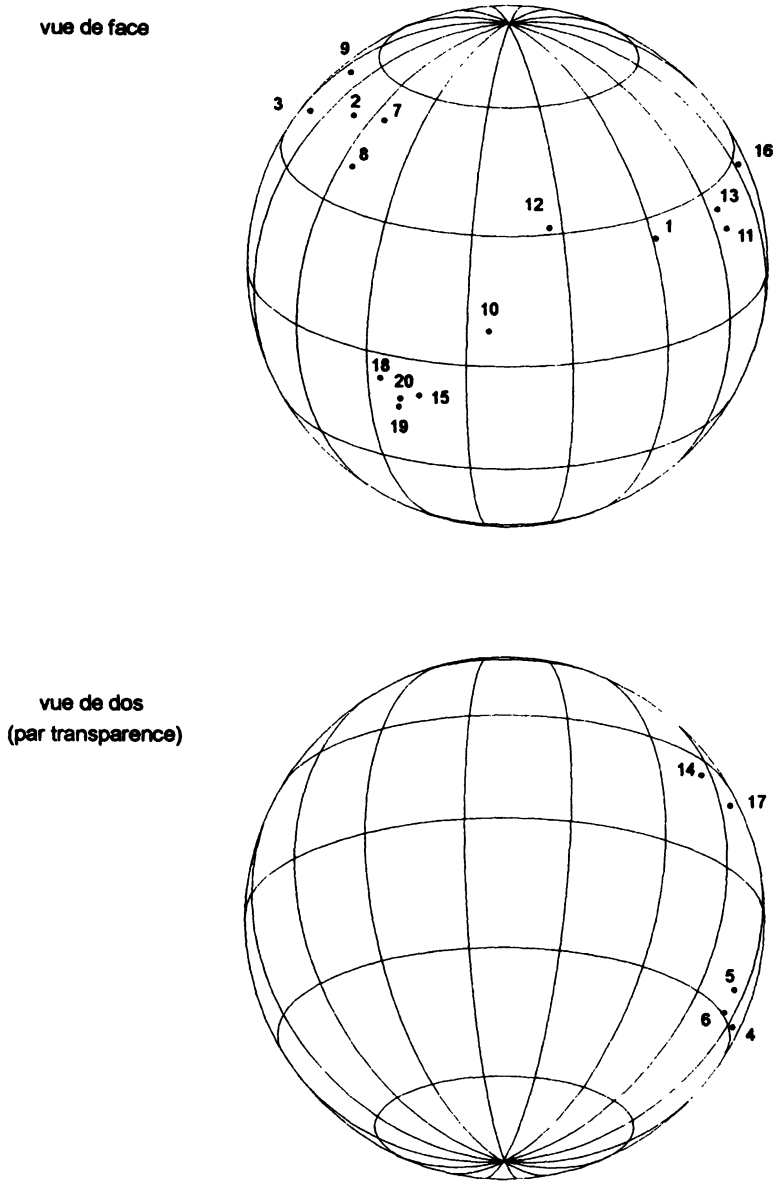

FIGURE 1

Représentation sphérique de la matrice de corrélation de l'échelle d'humeur dépressive

3. Un exemple : l'étude d'une échelle d'hétéroévaluation de l'humeur dépressive

Afin d'objectiver certains regroupements de symptômes dans la sémiologie de l'humeur dépressive, une Echelle d'Humeur Dépressive (EHD) a été développée récemment par Jouvent *et al.* (1988). Elle se présente sous la forme d'une liste de 20 items auxquels un cotateur entraîné doit affecter un entier allant de 1 a 5 en fonction de la sévérité du symptôme. Ces 20 items peuvent être résumés par : 1- Tristesse douloureuse observée, 2- Hyperexpressivité émotionnelle, 3- Instabilité émotionnelle, 4- Monotonie observée, 5- Manque d'expressivité spontanée, 6- Manque de réactivité affective, 7- Incontinence émotionnelle, 8- Hyperesthésie affective, 9- Humeur explosive observée, 10- Mimique inquiète, 11- Anhédonie observée, 12- Tristesse ressentie, 13- Anhédonie de situation, 14- Indifférence affective ressentie, 15- Hypersensibilité aux événements déplaisants, 16- Anhédonie sensorielle, 17- Monotonie affective ressentie, 18- Hyperémotivité ressentie, 19- Irritabilité ressentie, 20- Humeur explosive ressentie.

Les données présentées ici proviennent d'une étude effectuée sur 216 patients déprimés (DSM III R, American Psychiatric Association, 1987) cotés à leur entrée à l'hôpital (service de psychiatrie adulte de la Salpêtrière).

Une représentation sphérique de la matrice de corrélation des 20 items est présentée figure 1. Les regroupements sémiologiques relevés par les cliniciens ont été : 2,3,7,8,9 pour un groupe que l'on peut dénommer «hyperexpressivité», 4,5,6 pour un groupe «hypoexpressivité», 1,10,12 pour «tristesse», 11,13,14,16,17 pour «anhédonie» et 15,18,19,20 pour «hypersensibilité». Ils sont assez clairement individualisés et correspondent à une réalité clinique tangible.

Une représentation à partir d'une ACP est représentée figure 2. La représentation en 2 dimensions n'est pas suffisamment informative alors que la représentation en 3 dimensions retrouve les 5 regroupements mentionnés ci-dessus. Il est cependant notable que ces groupes apparaissent moins clairement qu'avec la représentation sphérique. La figure 3 réunit une représentation tridimensionnelle de l'ACP avec l'approximation de la représentation sphérique obtenue à partir d'une ACP (extrapolation sur la sphère unité des points variables, cf § 2.2); on retrouve la similitude entre la représentation sphérique et son approximation ainsi que la faible lisibilité de l'ACP.

4. Comparaison entre une représentation sphérique et une ACP à partir de données simulées.

Dans 8 configurations différentes, 50 matrices de corrélation sont générées aléatoirement à partir de données simulées. Pour chacune des 8 x 50 réplications sont calculés :

– $z2_i$, la projection de x_i sur le premier plan principal;

– $z3_i$, la projection de x_i sur le premier sous espace principal de dimension 3;

– $y_i = z3_i/||z3_i||$, l'approximation de o_i, ainsi qu'o_i lui même, projection de x_i sur la sphère optimale $O°$.

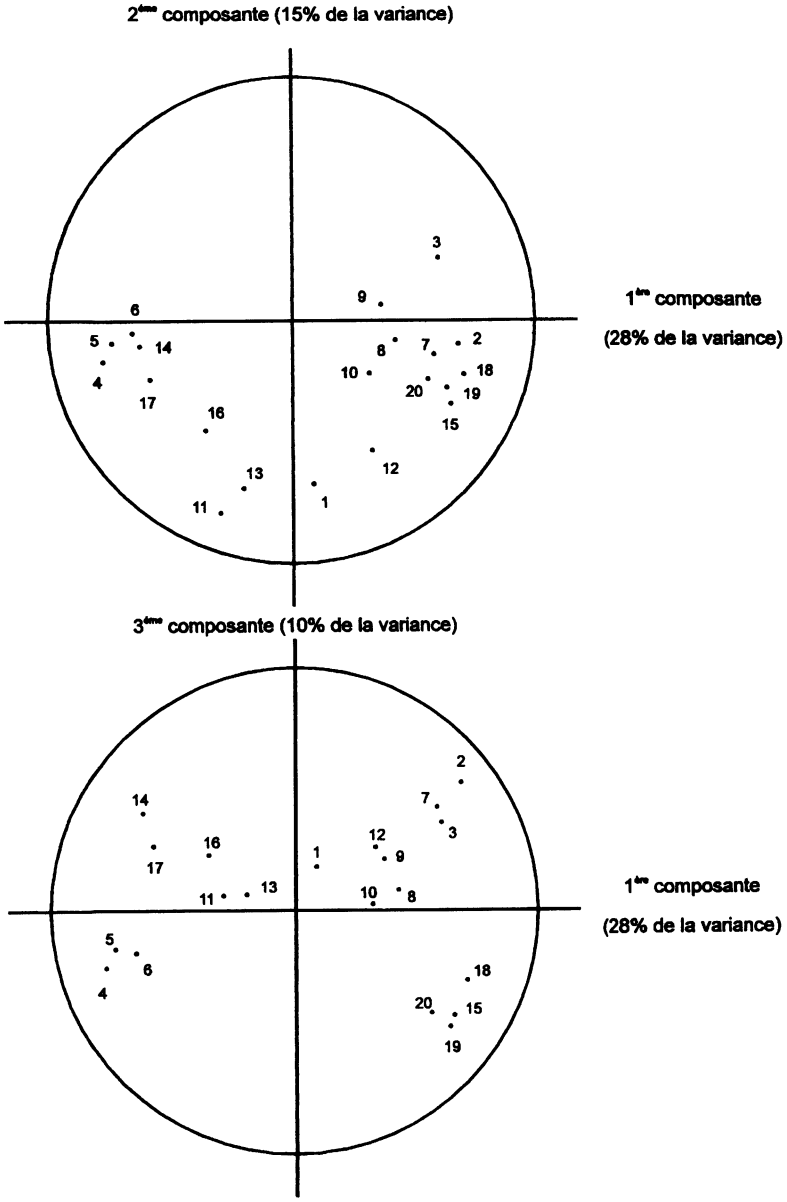

FIGURE 2

ACP de la matrice de corrélation de l'échelle d'humeur dépressive

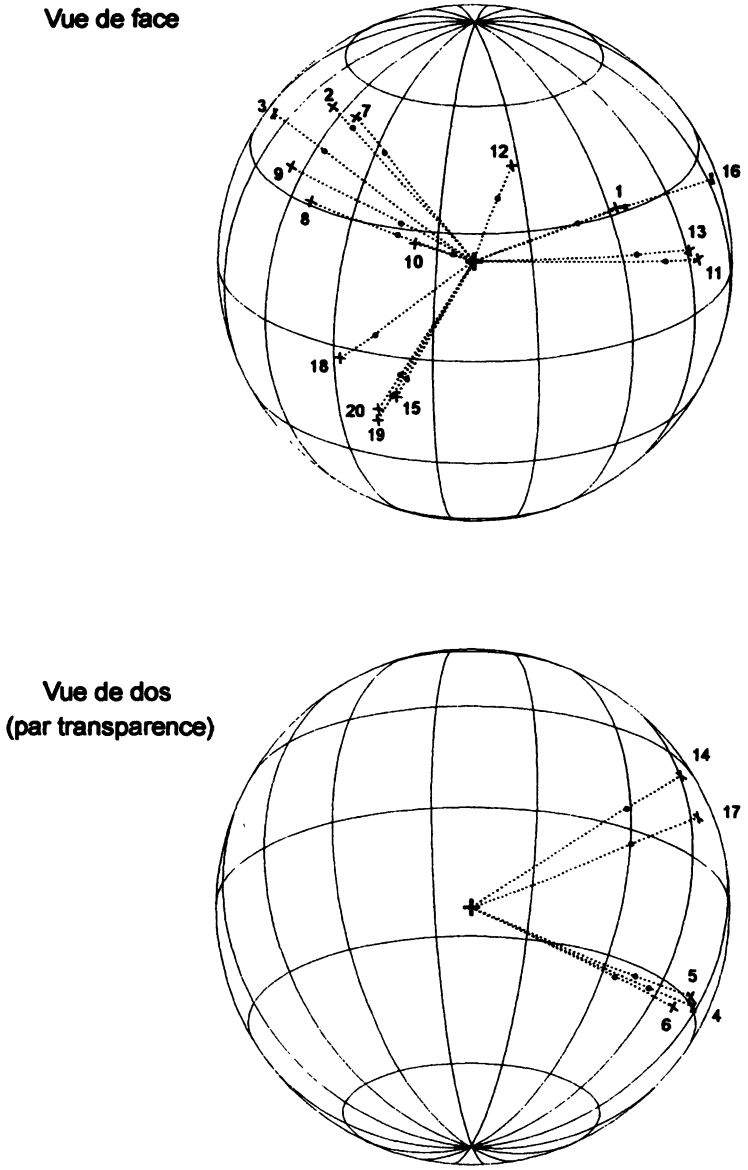

FIGURE 3

Représentation tridimentionnelle de l'ACP de la matrice de corrélation de l'échelle d'humeur dépressive (gros points à l'intérieur de la sphère), la projection de ces points sur la sphère (croix) donne une approximation de la représentation sphérique de la même matrice de corrélation.

Deux paramètres reflétant la similitude entre la configuration engendrée par les variables originales x_i et celle engendrée par leurs projections sont calculés :

– le coefficient de corrélation de Pearson entre $||x_i - x_m||^2$ et $||o_i - o_m||^2 (1 \leqslant i \leqslant p, 1 \leqslant m \leqslant p)$ est estimé et noté ρ_s, il permet d'évaluer si, quand x_i est plus proche de x_m que de x_k, il en va de même pour o_i, o_m et o_k. De la même manière ρ'_s est estimé entre $||x_i - x_m||^2$ et $||y_i - y_m||^2$, ρ_2 entre $||x_i - x_m||^2$ et $||z2_i - z2_m||^2$ et ρ_3 entre $||x_i - x_m||^2$ et $||z3_i - z3_m||^2$.

– La somme $S_s = \Sigma(||x_i - x_m|| - ||o_i - o_m||)^2 \quad (1 \leqslant i \leqslant p,\ 1 \leqslant m \leqslant p)$ est aussi calculée, elle reflète la qualité de la conservation des distances entre 2 points et entre leurs projections. S'_s, S_2, S_3, sont définis de façon similaire à partir de $(x_i, y_i), (x_i, z2_i)$ et $(x_i, z3_i)$,...

L'intérêt qu'il faut porter à la conservation de $||x_i - x_m||$ vient du fait que $r_{im} = 1 - (1/2)||x_i - x_m||^2$: deux variables sont d'autant plus corrélées que les points qui leurs correspondent sont proches (Lebart *et al.*, 1982, p. 287).

Les 8 configurations de matrices de corrélation $R_{p,k,r}$ sont telles que :

$$R_{p,k,r} = \left\| \begin{array}{cccccc} \begin{vmatrix} 1 & & (r1) \\ & \ddots & \\ (r1) & & 1 \end{vmatrix} & & & & & \\ & \begin{vmatrix} 1 & & (r1) \\ & \ddots & \\ (r1) & & 1 \end{vmatrix} & & & (r2) & \\ & & \ddots & & & \\ & & & \begin{vmatrix} 1 & & (r1) \\ & \ddots & \\ (r1) & & 1 \end{vmatrix} & & \\ & (r2) & & & & \end{array} \right\| \quad p \text{ lignes}$$

avec : $p = 10, 30$; $k = 2, 5$ et $r = 1$ si $(r_1, r_2) = (0.9, 0.3)$, et $r = 2$ si $(r_1, r_2) = (0.5, 0)$. Les notations (r_1) et (r_2) dans la matrice $R_{p,k,r}$ ci-dessus sont utilisées pour exprimer que les termes non diagonaux (pour r_1) et les blocs non diagonaux (pour r_2) sont tous égaux a r_1 ou r_2.

Pour chaque p, k, r , 50 matrices $R_{p,k,r}(l)$ $(1 \leqslant l \leqslant 50)$ sont générées, chacune à partir de $90 \times p$ données simulées :

$$x'_{iq} = an_q + bn_{<(i-1)/k>q} + n_{iq} \quad (1 \leqslant i \leqslant p; \quad 1 \leqslant q \leqslant 90)$$

où :

– n_q et n_{iq} sont des réalisations indépendantes d'une variable aléatoire normale d'espérance 0 et de variance unité;

– < > signifie «partie entière de»,

– $b = \left\{[(q_1)^2 - (q_2)^2]/[1 + (q_2)^2]\right\}^{1/2}$ et $a = q_2 \left\{[1 + (q_1)^2]/[1 + (q_2)^2]\right\}^{1/2}$

avec $q_1 = [r_1/(1-r_1)]^{1/2}$ et $q_2 = [r_2/(1-r_2)]^{1/2}$. Cela garantit que $E\{R_{p,k,r}(I)\} = R_{p,k,r}$.

Les résultats sont présentés tableau 1, ils montrent que ρ_s est supérieur à ρ_2 et même légèrement supérieur à ρ_3. De même, S_s est inférieur à S_2 et à S_3. Une représentation sphérique conserve donc une information comparable à celle que l'on trouve dans une ACP retenant 3 dimensions. On remarquera que ce point est d'autant plus vrai que le pourcentage d'inertie contenu dans les trois premières composantes principales est élevé; ceci est naturel puisque dans une telle situation les points variables représentés par une ACP sont proches de la sphère unité. Finalement, ρ_s est très proche de ρ'_s (et S_s de S'_s). Cela signifie que T^* est si proche de T° qu'en pratique, l'approximation $o_i = y_i$ est acceptable et qu'un simple logiciel qui calcule une ACP sera suffisant pour obtenir une représentation sphérique.

TABLEAU 1

A partir de simulations et de deux indices de similitude (ρ, S), *comparaison d'une représentation sphérique* (ρ_s, S_s) *avec son approximation à partir d'une ACP* (ρ'_s, S'_s), *une ACP retenant 2 dimensions* (ρ_2, S_2) *et une ACP retenant 3 dimensions* (ρ_3, S_3). *(Voir § 4 pour une définition de* ρ_i, S_i, p, k *et* r*).* l_1, l_2 *et* l_3 *sont les 3 valeurs propres les plus élevées de* $R_{p,k,r}$.

	$p=10$				$p=30$			
	$k=2$		$k=5$		$k=2$		$k=5$	
	$r=1$	$r=2$	$r=1$	$r=2$	$r=1$	$r=2$	$r=1$	$r=2$
S_s	3.32	13.4	9.67	11.5	47.5	195	97.2	157
S'_s	3.32	13.7	10.2	12.1	47.5	196	102.0	162
S_2	7.09	36.4	38.66	40.7	74.9	404	374.7	508
S_3	3.57	19.4	15.22	22.9	51.0	288	153.8	339
ρ_s	.996	.922	.794	.803	.997	.938	.798	.785
ρ'_s	.996	.922	.789	.797	.997	.937	.793	.779
ρ_2	.994	.890	.574	.685	.997	.921	.586	.655
ρ_3	.996	.914	.770	.801	.997	.930	.774	.769
l_1/p	.610	.300	.430	.150	.603	.267	.403	.100
l_2/p	.310	.300	.130	.150	.303	.267	.103	.100
l_3/p	.010	.050	.130	.150	.003	.016	.103	.100
$(l_1+l_2+l_3)/p$	.930	.650	.690	.450	.910	.550	.610	.300

5. Conclusion

Il est temps maintenant de comparer l'utilité de la représentation sphérique aux autres méthodes graphiques permettant de représenter des données multivariées. Nous envisagerons successivement l'ACP, les techniques de «Non linear mapping» et enfin la méthode des surfaces principales.

Il est théoriquement possible de représenter des données à partir d'une ACP en retenant 2, 3 ou davantage de dimensions. Avec 2 dimensions, un simple diagramme

plan est techniquement suffisant et l'information est facilement accessible. Pour 3 dimensions ou plus, il est nécessaire de combiner plusieurs plans et il est souvent difficile de réarranger mentalement les différents diagrammes (quelques logiciels essaient de donner l'illusion d'une troisième dimension, mais ils ne sont pas très convaincants). Comme nous l'avons vu au § 4, (même si les indices utilisés pour la comparaison peuvent paraître arbitraires) une représentation sphérique semble aussi informative qu'une ACP retenant 3 composantes, mais elle permet une meilleure visualisation des variables.

Puisqu'une représentation sphérique apparaît comme une méthode permettant de représenter des similitudes entre des variables à partir d'un ensemble de points dans un ensemble de dimension 3, il est nécessaire de la comparer aux techniques de non linear mapping. Si les buts du non linear mapping et d'une représentation sphérique sont proches, ils ne sont pas identiques. L'objectif spécifique des techniques de non linear mapping est de présenter graphiquement des similitudes entre variables. L'objectif d'une représentation sphérique est plutôt de représenter graphiquement une matrice de corrélation. Ainsi les variables peuvent elles être regroupées, mais aussi opposées (si elles sont diamétralement opposées) ou même vues non corrélées (si leurs rayons sont perpendiculaires). De plus, la plupart des méthodes de non linear mapping fournissent une représentation graphique dans un sous espace, ce qui n'est pas le cas de la représentation sphérique.

Cette remarque nous amène tout naturellement à comparer la représentation sphérique aux autres méthodes qui essaient d'approximer un ensemble de points par des nappes non linéaires. La sphère optimale O° peut être considérée comme une «Principal Surface» comme les définissent Hastie et Stuetzle (1989). Cependant, ce dernier article considère de façon très générale des surfaces non paramétrées interpolant un ensemble de points. Comme nous le mentionnions dans l'introduction, la paramétrisation utilisée dans une représentation sphérique est, elle, justifiée par la position des points variables x_i sur l'hypersphère unité de $\Re^n$.

On peut aussi mentionner les méthodes de régression non linéaires, pour lesquelles une des variables doit cependant avoir un rôle privilégié... Voire même les méthodes d'analyse factorielle non linéaires (McDonald 1962), mais ici l'idée d'un modèle avec variables latentes est prépondérante.

Un des avantages substantiels de la représentation sphérique par rapport a ces différentes méthodes est certainement sa simplicité numérique. Comme nous l'avons vu au §4, un logiciel estimant des composantes principales suffit à calculer des projections sphériques.

Une limite à cette méthode mérite cependant, pour terminer, d'être mentionnée. La précision d'une représentation graphique par une ACP peut être estimée par l'importance des valeurs propres associées aux composantes retenues pour l'analyse. Ce n'est pas le cas pour une représentation sphérique : il n'y a pas de moyen rigoureux d'évaluer jusqu'à quel point la représentation graphique correspond à la situation réelle sur l'hypersphère. On peut seulement conjecturer que puisqu'une représentation sphérique semble aussi informative qu'une ACP avec 3 dimensions, il est possible de considérer la somme des 3 valeurs propres les plus élevées de R comme un indice d'adéquation.

Remerciements

L'auteur tient à remercier R. Jouvent pour l'utilisation de ses données dans l'exemple ainsi que M. Berthier, M. Chavance, L. Lebart, J. Fermanian et un rapporteur pour leurs commentaires.

Bibliographie

AMERICAN PSYCHIATRIC ASSOCIATION (1987). *Diagnostic and Statistical Manual of Mental Disorders, third edition, revised (DSM-III-R).* American Psychiatric Association, Washington, DC.

HASTIE T., & STUETZEL W. (1989). Principal Curves. *Journal of the American Statistical Association*, 84, p. 502-516.

JOUVENT R., VINDREAU C., MONTREUIL M., BUNGENER C., & WIDLOCHER D. (1988). La clinique polydimensionnelle de l'humeur dépressive. Nouvelle version de l'échelle EHD. *Psychiatrie et Psychobiologie*, 3, p. 245-253.

LEBART L., MORINEAU A., & FÉNELON J. P. (1982). *Traitement des données statistiques.* Paris : Dunod.

MCDONALD R. P. (1962). A general approach to non linear factor analysis. *Psychometrika*, 27, p. 397-415.

SAMMON, J. W. (1969). A non-linear mapping for data structure analysis. *IEEE Trans. Comput.*, C-18, p. 401-9.

Appendice : démonstration de l'équation (4)

Il s'agit de maximiser

$$f(v_1, v_2, v_3) = \sum_{i=1}^{i=p} ||\operatorname{proj}(x_i, T)|| = \sum_{i=1}^{i=p} \left\{(x_i'v_1)^2 + (x_i'v_2)^2 + (x_i'v_3)^2\right\}^{1/2} \quad (1')$$

sous la contrainte que les vecteurs (v_1, v_2, v_3) sont orthonormés :

$$g_k(v_1, v_2, v_3) = ||V_k||^2 - 1 = 0 \quad (k = 1, 2, 3)$$

$$g_4(v_1, v_2, v_3) = v_1 v_2 = 0; \;\; g_5(v_1, v_2, v_3) = v_2 v_3 = 0; \;\; g_6(v_1, v_2, v_3) = v_1 v_3 = 0$$

Le maximum de f est obtenu pour $T = T^\circ$, où $x_i'v_k$ vaut t_{ik} $(1 \leqslant i \leqslant p;\; k = 1, 2, 3)$.

Si l'on pose : $L = f(v_1, v_2, v_3) - \sum_{i=1}^{i=6} \lambda_i g_i(v_1, v_2, v_3)$ où les λ_i sont les multiplicateurs de Lagrange, les équations :

$$\partial L / \partial v_{1q} = 0 \quad (1 \leqslant q \leqslant n), v_{1q} \text{ sont les coordonnées de } v_1)$$

conduisent à : $\partial f/\partial v_{1q} = 2\lambda_1 v_{1q} + \lambda_6 v_{3q} + \lambda_4 v_{2q}$ (2′)

d'où

$$\begin{aligned}\Sigma_{1\leqslant q\leqslant n}(\partial f/\partial v_{1q})v_{1q} &= 2\lambda_1\Sigma_{1\leqslant q\leqslant n}(V_{1q})^2 + \lambda_6\Sigma_{1\leqslant q\leqslant n}(v_{3q}v_{1q}) \\ &\quad + \lambda_4\Sigma_{1\leqslant q\leqslant n}(v_{2q}v_{1q}) = 2\lambda_1 \qquad (3')\\ \Sigma_{1\leqslant q\leqslant n}(\partial f/\partial V_{1q})v_{3q} &= \lambda_6 \\ \Sigma_{1\leqslant q\leqslant n}(\partial f/\partial v_{1q})v_{2q} &= \lambda_4\end{aligned}$$

Or de (1′) on obtient : $\partial f/\partial v_{1q} = \Sigma_{1\leqslant m\leqslant p} x_{mq} t_{m1}/R_m$ (4')

où $R_m = \left[(t_{m1})^2 + (t_{m2})^2 + (t_{m3})^2\right]^{1/2} \quad (1 \leqslant m \leqslant p)$

Ainsi à partir de (3′) nous obtenons :

$$\begin{aligned}2\lambda_1 &= \Sigma_{1\leqslant q\leqslant n}\left(\Sigma_{1\leqslant m\leqslant p} x_{mq} t_{m1}/R_m\right) v_{1q} \\ &= \Sigma_{1\leqslant m\leqslant p}(\Sigma_{1\leqslant q\leqslant n} x_{mq} v_{1q}) t_{m_1}/R_m = (\Sigma_{1\leqslant m\leqslant p}(t_{m1})^2/R_m \qquad (5')\\ \lambda_4 &= \Sigma_{1\leqslant m\leqslant p}(t_{m2}t_{m_1})/R_m \\ \lambda_6 &= \Sigma_{1\leqslant m\leqslant p}(t_{m3}t_{m1})/R_m\end{aligned}$$

De même, à partir de (4′) :

$$\begin{aligned}\Sigma_{1\leqslant q\leqslant n}(\partial f/\partial v_{1q})x_{iq} &= \Sigma_{1\leqslant q\leqslant n}\left(\Sigma_{1\leqslant m\leqslant p} x_{mq} t_{m1}/R_m\right) x_{iq} \\ &= \Sigma_{1\leqslant m\leqslant p}\left(\Sigma_{1\leqslant q\leqslant n} x_{mq} x_{iq}\right) t_{m1}/R_m \qquad (6')\\ &= \Sigma_{1\leqslant m\leqslant p}(R_{mi} t_{m1})/R_m\end{aligned}$$

où r_{mi} est le $(m,i)^{\text{ème}}$ élément de la matrice de corrélation R.

Alors que (2′) :

$$\begin{aligned}\Sigma_{1\leqslant q\leqslant n}(\partial f/\partial v_{1q})x_{iq} &= 2\lambda_1(\Sigma_{1\leqslant q\leqslant n} v_{1q} x_{iq}) + \lambda_6(\Sigma_{1\leqslant q\leqslant n} v_{3q} x_{iq}) \\ &\quad + \lambda_4(\Sigma_{1\leqslant q\leqslant n} v_{2q} x_{iq}) \\ &= 2\lambda_1 t_{i1} + \lambda_6 t_{i3} + \lambda_4 t_{i2} \qquad (7')\\ &= (\Sigma_{1\leqslant m\leqslant p}(t_{m1})^2 t_{i1}/R_m) + (\Sigma_{1\leqslant m\leqslant p} t_{m3} t_{m1} t_{i3}/R_m) \\ &\quad + (\Sigma_{1\leqslant m\leqslant p} t_{m2} t_{m1} t_{i2}/R_m)\end{aligned}$$

(en utilisant (5'))

Et finalement, puisque (6') = (7') :

$$h_{i1} = \Sigma_{1\leqslant m\leqslant p} t_{m1}(t_{m1}t_{i1} + t_{m2}t_{i2} + t_{m3}t_{i3} - r_{mi})/R_m = 0$$

(h_{i2} et h_{i3} sont obtenues de façon similaires)