REVUE DE STATISTIQUE APPLIQUÉE

F. HUSSON
J. PAGÈS

Aspects méthodologiques du modèle INDSCAL

Revue de statistique appliquée, tome 54, n° 2 (2006), p. 83-100
<http://www.numdam.org/item?id=RSA_2006__54_2_83_0>

Rev. Statistique Appliquée, 2006, **LIV** (2), 83-100

ASPECTS MÉTHODOLOGIQUES DU MODÈLE INDSCAL

F. HUSSON, J. PAGÈS

Laboratoire de Mathématiques Appliquées, Agrocampus Rennes,
65 rue de Saint-Brieuc, 35042 Rennes Cedex
E-mail : husson@agrocampus-rennes.fr

RÉSUMÉ

Le modèle INDSCAL proposé par Carroll et Chang (1970) fournit des représentations graphiques euclidiennes. Lorsque les données initiales ne sont pas euclidiennes un passage à l'euclidienneté est effectué soit en cours d'algorithme, soit lors d'un pré-traitement. Nous évaluons l'impact du pré-traitement le plus classique sur les solutions des algorithmes INDSCAL à travers le critère du *Strain* et les configurations moyennes du modèle. Nous montrons qu'il est préférable d'effectuer un pré-traitement avant d'estimer les paramètres du modèle.

Nous proposons également une nouvelle décomposition de la variabilité des matrices de produits scalaires ce qui permet de savoir ce qui contribue le plus à l'écart entre les données initiales et la solution INDSCAL. Cet écart peut-il être attribué plutôt à la non euclidienneté des données, plutôt à la multidimensionnalité des données plutôt au non ajustement à une configuration commune ou à ces trois sources de façon équivalente.

Mots-clés : *Modèle* INDSCAL*, interprétation géométrique, décomposition de l'erreur*

ABSTRACT

The INDSCAL model proposed by Carroll and Chang (1970) provides Euclidean representations. When original data are not Euclidean, an Euclidean transformation is carried out either during the algorithm, or before as a preprocessing. We evaluate the impact of the more usual preprocessing on the solutions of the INDSCAL algorithms through the *Strain* criterion and through the average configurations of the model. We show that it is preferable to carry out a preprocessing before estimating the parameters of the model.

We also propose a novel decomposition of the variability of the scalar products what allows to know what contributes more to the variation between original data and the INDSCAL solutions. Is that not Euclidean data, multidimensionnality of the data or non-adjustment to the common configuration which makes important the difference between the subject configuration of a subject and the common configuration.

Keywords : *Modèle* INDSCAL*, geometrical interpretation, residual decomposition*

Introduction

L'intérêt d'une évaluation directe et globale de la différence sensorielle entre produits alimentaires est connu depuis longtemps. Le protocole classique consiste à proposer à un ensemble d'individus (appelés juges) un ensemble de produits, paire par paire. Chaque juge évalue la différence entre deux produits à l'aide d'une note, ce qui revient finalement à fournir une matrice d'indices de dissimilarités.

Une façon d'analyser ces données est d'en proposer une (ou plusieurs) représentation(s) euclidienne(s). Parmi les méthodes disponibles, celle qui consiste à postuler le modèle INDSCAL (Carroll et Chang, 1970; Schiffman *et al.*, 1981) est tout à fait intéressante. Plusieurs algorithmes, appelés INDSCAL comme le modèle, sont disponibles (Kroonenberg, 1983, Kiers, 1989, 1997, Ten Berge et Kiers, 1991, Trendafilov, 2004).

Les représentations graphiques du modèle INDSCAL sont toujours euclidiennes. Or les données sur lesquelles le modèle INDSCAL est postulé ne sont pas nécessairement euclidieennes. Il y a donc un passage à l'euclidienneté qui est effectué au cours de l'algorithme d'estimation. Cependant, on peut se demander s'il ne serait pas plus intéressant de transformer les données initiales en données euclidiennes avant d'estimer les paramètres du modèle. Cette idée est souvent rejetée a priori au prétexte qu'elle ne respecte pas les données du fait de cette déformation. En fait, le principe de cette déformation est implicitement admis dès lors que l'on utilise le modèle INDSCAL. Simplement, il est plus visible si l'on réalise un pré-traitement, ce qui, finalement, s'avère un avantage.

L'objet de cet article est d'une part d'évaluer l'impact d'un pré-traitement sur les solutions du modèle et d'autre part de proposer des indicateurs permettant de comprendre l'écart entre la solution INDSCAL et les données initiales.

1. Méthodologie

1.1. Le modèle INDSCAL

Soit J juges et I produits; le juge j évalue globalement la dissimilarité entre les produits i et l par la note $d_j(i,l)$ variant par exemple de 0 (produits identiques) à 10 (produits très différents).

Selon le modèle INDSCAL, la dissimilarité $d_j(i,l)$ dérive d'une configuration des produits en R dimensions (soit $z_r(i)$ la coordonnée du produit i sur l'axe de rang r de cette configuration), chaque juge j accordant un poids spécifique q_r^j à la dimension r :

$$d_j^2(i,l) = \sum_{r=1}^{R} q_r^j \left(z_r(i) - z_r(l)\right)^2 + e_j(i,l) = \hat{d}_j^2(i,l) + e_j(i,l) \qquad (1)$$

où $e_j(i,l)$ est le résidu du modèle et $\hat{d}_j^2(i,l)$ le carré de la distance reconstituée par le modèle.

Si $d_j(i,l)$ est une distance euclidienne, alors on peut associer à chaque couple de produit (i,l) un produit scalaire noté $\langle i,l\rangle$ et le modèle peut s'écrire :

$$\langle i,l\rangle_j = \sum_{r=1}^{R} q_r^j \; z_r(i) \; z_r(l) + \varepsilon_j(i,l) \tag{2}$$

Le produit scalaire $\langle i,l\rangle_j$ est obtenu par la formule de Torgerson (Torgerson, 1958) :

$$\langle i,l\rangle_j = -\frac{1}{2}(d_j^2(i,l) - d_j^2(i,.) - d_j^2(.,l) + d_j^2(.,.))$$

où $d_j^2(i,.) = \frac{1}{I}\sum_{l=1}^{I} d_j^2(i,l)$ et $d_j^2(.,.) = \frac{1}{I}\sum_{i=1}^{I} d_j^2(i,.)$.
Lorsque les distances ne sont pas euclidiennes, on parle de pseudo-produits scalaires.

Matriciellement, le modèle peut s'écrire :

$$S_j = XW_jX' + E_j = \hat{S}_j + E_j \tag{3}$$

avec S_j (resp. $\hat{S}_j$) la matrice $(I \times I)$ des produits scalaires (resp. des produits scalaires reconstitués par le modèle) du juge j; X la matrice $(I \times R)$ des coordonnées des produits dans un espace à R dimensions et W_j une matrice diagonale $R \times R$ de poids, souvent appelée saliences (Ten Berge *et al.*, 1993).

Pour estimer les paramètres du modèle INDSCAL (coordonnées des produits dans la configuration moyenne dans un espace à R dimensions et poids affectés à chaque individu sur les R dimensions), il est nécessaire de recourir à des algorithmes car il n'existe pas de solution analytique. La méthode originale INDSCAL (Carrol et Chang, 1970) minimise, par un algorithme de type moindres carrés alternés :

$$Strain = \frac{1}{J}\sum_{j=1}^{J} \|S_j - XW_jX'\|^2 \tag{4}$$

le carré de la norme d'une matrice A de terme général a_{ij} étant : $\|A\|^2 = \sum_{i,j} a_{ij}^2$.

D'autres critères tels que le $Stress$ (équation 5) ou le $S-Stress$ (équation 6) ont été proposés par Takane *et al.* (1977); toutefois, le *Strain* reste le critère le plus intéressant en ce sens qu'il bénéficie d'une interprétation géométrique claire (Husson et Pagès, 2005) qui permet entre autre de le décomposer par juge et par produit et de l'interpréter comme un rapport d'inertie projetée sur inertie totale.

$$Stress = \sqrt{\frac{1}{J}\sum_{j=1}^{J}\left(\frac{\sum_{i,l}\left(d_j(i,l) - \hat{d}_j(i,l)\right)^2}{\sum_{i,l}\hat{d}_j(i,l)^2}\right)} \tag{5}$$

$$S - Stress = \sqrt{\frac{1}{J}\sum_{j=1}^{J}\left(\frac{\sum_{i,l}\left(d_j^2(i,l) - \hat{d}_j^2(i,l)\right)^2}{\sum_{i,l}\hat{d}_j(i,l)^4}\right)} \quad (6)$$

1.2. Pré-traitement

Chaque juge j peut être représenté par la matrice de produits scalaires S_j qui lui est associée. Si les données initiales du juge j sont constituées par un tableau individus $\times$ variables, noté V_j alors : $S_j = V_j M_j V_j'$ en notant M_j la matrice diagonale des poids des variables pour j. Si les données initiales du juge j sont constituées par un tableau de distances, non nécessairement euclidiennes, on se ramène au cas précédent en prenant pour matrice V_j les coordonnées des vecteurs propres associés aux valeurs propres positives de la matrice des produits scalaires dérivant (via la formule de Torgerson) de la matrice des distances. Ceci revient à effectuer une analyse en coordonnées principales (Gower, 1966). Notons de plus que cette représentation est optimale quand les données sont non euclidiennes (D'Aubigny, 1998). D'Aubigny note que la faible valeur absolue de valeurs propres négatives peut pousser l'utilisateur à interpréter leur présence comme le résultat d'imprécisions de la mesure des jugements de dissimilarités. Le chapitre 18 de Borg et Groenen (1997) est consacré à la transformation de dissimilarités en distances ou en distances euclidiennes. On peut considérer un juge comme un ensemble de variables définies sur les mêmes produits, ces variables étant : les variables proposées aux juges lorsque c'est le cas; sinon, les vecteurs propres de S_j, qui peuvent être considérés comme des variables latentes. En notant V_j la matrice dont les colonnes sont les vecteurs propres normés de S_j et M_j la matrice diagonale des valeurs propres correspondantes on a bien $S_j = V_j M_j V_j'$.

Par définition, le modèle INDSCAL fournit une représentation euclidienne. Si les dissimilarités initiales ne sont pas des distances euclidiennes, alors la solution du modèle INDSCAL peut être obtenue selon deux stratégies. Dans la première, on effectue le pré-traitement précédemment décrit qui approche au mieux les dissimilarités initiales par des distances euclidiennes, puis, à partir de ces distances (appelées «données transformées»), on estime les paramètres du modèle INDSCAL. L'autre possibilité est d'estimer les paramètres du modèle directement à partir des dissimilarités. Dans ce cas, la recherche d'une représentation euclidienne est effectuée au cours de l'algorithme en même temps que l'estimation des paramètres du modèle INDSCAL. Dans l'algorithme INDSCAL, une solution intermédiaire est adoptée : l'ajout d'une même constante à toutes les distances d'une matrice permet à cette matrice de vérifier toutes les inégalités triangulaires.

Cependant, l'ajout de cette constante ne suffit pas à rendre les distances euclidiennes. Dans les algorithmes INDSCAL, il y a donc un passage à l'euclidienneté qui est effectué en même temps que l'estimation des paramètres. Ainsi, pour certaines données, il y a ajout de constante **puis** passage à l'euclidienneté. Ceci n'est pas optimal comme l'a montré D'Aubigny (1998). La solution adoptée dans ces

algorithmes INDSCAL ne semble pas présenter d'avantage déterminant. Il est préférable d'effectuer un pré-traitement qui rend les distances euclidiennes puis d'estimer les paramètres du modèle INDSCAL. Cette stratégie permet de plus de comparer entre elles des méthodes qui nécessitent ou non l'euclidienneté.

1.3. Interprétation géométrique du modèle INDSCAL *dans* $\mathbb{R}^{I^2}$

On reprend l'équation 2 et on appelle z_r le vecteur des coordonnées des produits sur l'axe r. Le modèle INDSCAL s'écrit alors :

$$S_j = \sum_{r=1}^{R} q_r^j z_r z_r' + E_j$$

On considère l'espace euclidien usuel à I dimensions noté $\mathbb{R}^{I^2}$. Dans cet espace, les S_j sont des vecteurs : le modèle INDSCAL exprime que les S_j sont décomposés sur un même repère formé d'éléments symétriques de rang 1. Le poids q_r^j est la coordonnée de S_j sur l'élément $z_r z_r'$ de ce repère (voir figure 1). Le *Strain* défini en introduction peut s'écrire $Strain = \frac{1}{J}\|S_j - \hat{S}_j\|^2$ avec ces notations.

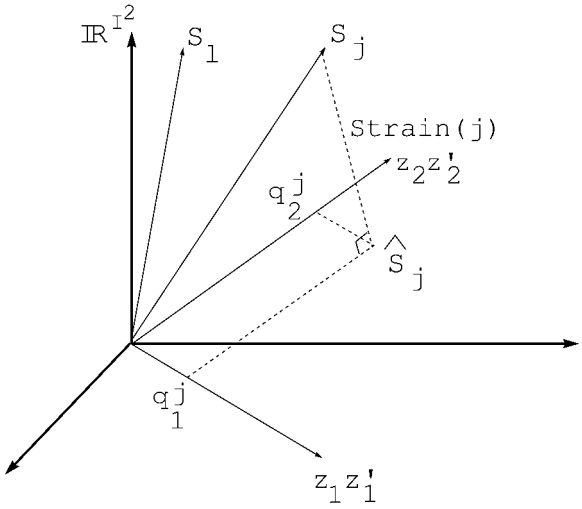

FIGURE 1
Interprétation géométrique du modèle INDSCAL *dans* $\mathbb{R}^{I^2}$

Par analogie avec l'analyse factorielle, cette interprétation géométrique suggère d'ajuster les données au modèle en cherchant dans $\mathbb{R}^{I^2}$ une suite de vecteurs, normés ($\|z_r z_r'\| = 1$, avec $\|A\|^2 = \sum_{i,j} a_{ij}^2$) ce qui permet l'identifiabilité du modèle, associés chacun à une matrice symétrique de rang 1, et à projeter le nuage des S_j sur

cette suite de vecteurs. Il découle des propriétés du produit scalaire dans $\mathbb{R}^{I^2}$ que la non-corrélation entre les facteurs z_r entraîne l'orthogonalité entre les éléments $z_r z_r'$.

L'interprétation géométrique permet :

- d'éclairer la remarque de Carroll et Chang (1970) sur la non corrélation des z_r et la décomposition des juges : Carroll et Chang disent que lorsque les dimensions sont non corrélées, le carré de la distance d'un point à l'origine mesure le pourcentage de variance des produits scalaires expliquée pour ce juge. En effet, si les z_r sont non corrélés alors les carrés des coordonnées d'un juge (donc les carrés des poids) s'additionnent d'un axe à l'autre (Pythagore) et leur somme sur les axes est une mesure de la qualité de représentation du juge;
- de montrer que si une configuration de produits vérifie le modèle INDSCAL, alors une rotation appliquée à cette configuration des produits ne vérifiera pas le modèle INDSCAL (excepté pour des rotations d'angle $k\ \pi/2$ avec k entier). En effet, une combinaison linéaire de z_1 et z_2 (variables dans $\mathbb{R}^I$) n'est pas associée à une combinaison linéaire de $z_1 z_1'$ et $z_2 z_2'$ dans $\mathbb{R}^{I^2}$:

$$(a_1 z_1 + a_2 z_2)(a_1 z_1 + a_2 z_2)' = a_1^2 z_1 z_1' + a_2^2 z_2 z_2' + a_1 a_2 z_1 z_2' + a_1 a_2 z_2 z_1' \,. \quad (7)$$

Le terme de droite dépend de $z_1 z_2'$ et $z_2 z_1'$. Ainsi, le plan de $\mathbb{R}^{I^2}$ n'est pas stable à une rotation (dans $\mathbb{R}^I$) de la configuration moyenne. Inversement, chaque plan de $\mathbb{R}^{I^2}$ généré par un ensemble $\{z_r z_r',\ r = 1, .., R\}$ d'éléments de rang 1 ne contient pas d'autres éléments de rang 1. Quand le modèle INDSCAL s'ajuste parfaitement, la solution est unique.

L'interprétation géométrique dans $\mathbb{R}^{I^2}$ suggère également de prendre comme critère le rapport entre l'inertie projetée et l'inertie totale :

$$\frac{\text{Inertie projetée}}{\text{Inertie totale}} = \frac{\sum_j \|\hat{S}_j\|^2}{\sum_j \|S_j\|^2}$$

Ceci correspond au critère usuellement utilisé en analyse factorielle. Il revient au même de maximiser ce critère ou de minimiser le *Strain* dès lors que $\|S_j\| = 1$. Ceci suggère une généralisation du *Strain* lorsque les S_j sont non normés :

$$Strain = \frac{\sum_j \|S_j - \hat{S}_j\|^2}{\sum_j \|S_j\|^2}$$

Dans l'algorithme INDSCAL, la dernière étape est l'estimation des poids : par conséquent, la dernière régression effectuée minimise $\|S_j - \hat{S}_j\|^2$ pour chaque juge j, et donc $\|S_j - \hat{S}_j\|^2 = \|S_j\|^2 - \|\hat{S}_j\|^2$. On a alors :

$$Strain = \frac{1}{\sum_{j=1}^{J} \|S_j\|^2} \sum_{j=1}^{J} \left(1 - \frac{\|\hat{S}_j\|^2}{\|S_j\|^2}\right) \|S_j\|^2$$

On relie le critère avec la qualité de représentation du juge j $\dfrac{\|\hat{S}_j\|^2}{\|S_j\|^2}$. Cette écriture du *Strain* suggère deux décompositions de ce critère :

- une décomposition par juge : le *Strain* est égal à la somme des sinus carrés, somme éventuellement pondérée par $\|S_j\|^2$. La contribution d'un juge au *Strain* correspond donc au sinus carré, ce qui permet de détecter les juges qui contribuent le plus à l'écart au modèle INDSCAL.
- une décomposition du critère global axe par axe si les facteurs sont non corrélés. À la différence des deux autres critères, le *Strain* prend une place naturelle dans le cadre d'une interprétation globale du modèle.

1.4. Nouvelle décomposition de l'erreur

Comment interpréter une mauvaise reconstitution des données par le modèle INDSCAL ? Pour répondre à cette question, il est intéressant de décomposer la variabilité des données en plusieurs termes : la variabilité expliquée par la solution du modèle INDSCAL, la variabilité expliquée par le passage des données brutes à des données euclidiennes, la variabilité expliquée par le non ajustement sensu stricto des données rendues euclidiennes à un modèle INDSCAL.

L'ensemble des matrices semi-définies positives ne forme pas un sous-espace vectoriel mais un espace convexe (un cône positif, i.e. l'ensemble des combinaisons linéaires à coefficients positifs ou nuls d'un système générateur de ce cône). Lorsque les valeurs propres négatives des matrices de produits scalaires sont « ramenées » à 0, alors ces matrices sont projetées sur ce cône, et plus précisément sur une de ses faces (voir figure 2). Cette face du cône est une « portion » de sous-espace vectoriel et on peut donc écrire :

$$\|S_j\|^2 = \|S_j^{euclid}\|^2 + \|S_j - S_j^{euclid}\|^2 \tag{8}$$

avec S_j la matrice des produits scalaires du juge j et S_j^{euclid} son approximation euclidienne.

Par ailleurs, l'interprétation géométrique du modèle INDSCAL dans $\mathbb{R}^{I^2}$ suggère la décomposition :

$$\|S_j^{euclid}\|^2 = \|\hat{S}_j^{euclid}\|^2 + \|\hat{S}_j^{euclid} - S_j^{euclid}\|^2 \tag{9}$$

avec $\hat{S}_j^{euclid}$ les produits scalaires de la solution du modèle INDSCAL obtenue à partir de produits scalaires euclidiens.

On peut donc écrire :

$$\|S_j\|^2 = \|\hat{S}_j^{euclid}\|^2 + \|\hat{S}_j^{euclid} - S_j^{euclid}\|^2 + \|S_j - S_j^{euclid}\|^2 \qquad (10)$$

Variabilité totale = variabilité du modèle + erreur d'ajustement du modèle INDSCAL aux données rendues euclidiennes + distance entre la matrice des données brutes et son approximation euclidienne.

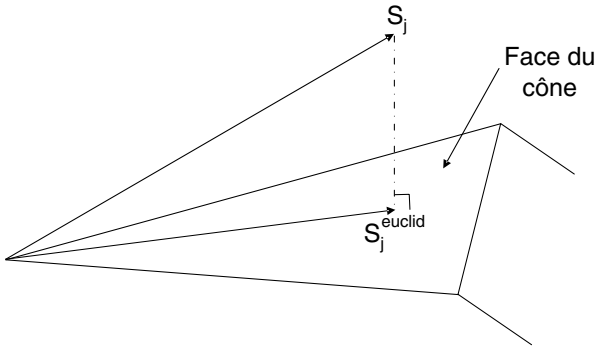

FIGURE 2
Projection de la matrice de produits scalaires du juge j (S_j) sur le cône

Cette décomposition présente une difficulté : dans le processus qui consiste à rendre les données euclidiennes puis à construire le modèle INDSCAL, réside une normalisation. Pour exprimer les variabilités en pourcentage de la variabilité initiale, il est nécessaire de tenir compte de cette normalisation. D'après la figure 3, on voit que la matrice de produits scalaires S_j, est dans un premier temps rendue euclidienne(S_j^{euclid}) puis, afin de construire le modèle INDSCAL, cette matrice de produits scalaires est normalisée ($S_j^{euclid^n}$) et enfin, la solution du modèle INDSCAL est obtenue ($\hat{S}_j^{euclid^n}$). La solution ($\hat{S}_j^{euclid^n}$), pour être comparable à S_j doit être dilatée en $\hat{S}_j^{euclid}$. Notons que l'équation 10 peut être calculée sur les termes normalisés (avec les indice n sur le graphe 3) ou non.

La contribution du juge j au *Strain*, que l'on notera $strain_{euclid_j}$ et qui correspond à la distance entre la solution INDSCAL (obtenue à partir des données rendues euclidiennes) et les produits scalaires rendus euclidiens, est égale à $\|\hat{S}_j^{euclid} - S_j^{euclid}\|^2$. Ce critère peut être fourni par les algorithmes INDSCAL puisqu'il suffit d'ajuster le modèle INDSCAL sur les données euclidiennes. En revanche, si on utilise le programme INDSCAL en travaillant directement sur les S_j (non euclidiens), la décomposition de la variabilité proposée à l'équation 10 n'est pas possible.

Par ailleurs, on peut prendre en compte la dimension de la solution INDSCAL pour apprécier la qualité d'ajustement d'un juge. L'erreur d'ajustement sensu stricto

($\|S_j^{euclid} - \hat{S}_j^{euclid}\|^2$) du modèle peut être comparée à la distance entre la représentation euclidienne des données et la représentation euclidienne en les k premières dimensions. En effet, si le modèle INDSCAL est en k dimensions et que les représentations individuelles sont en Q dimensions (avec $Q > k$), alors le modèle INDSCAL ne pourra pas récupérer toute l'information contenue dans les données. Mieux, le pourcentage d'inertie récupéré par le modèle est maximisé par le pourcentage d'inertie contenue dans les k premières dimensions. Ceci fournit une aide à l'interprétation intéressante du non ajustement au modèle INDSCAL. Dans tous nos exemples, nous avons pris $k = 2$ dimensions pour le modèle INDSCAL.

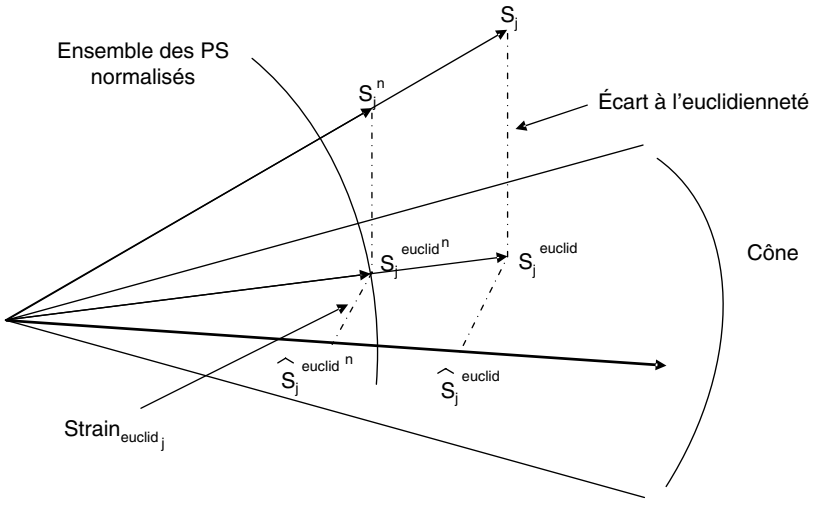

FIGURE 3
Décomposition de la variabilité
^{n} signifie que les produits scalaires sont normés par la norme des S_j^{euclid} :
*$\hat{S}_j^{euclid^n}$ signifie que c'est le vecteur des produits scalaires d'*INDSCAL
normé par la norme des produits scalaires euclidiens

2. Applications

2.1. Données

À partir d'un exemple fictif et de deux exemples réels, nous allons évaluer pratiquement les différences entre les solutions du modèle INDSCAL lorsque des données initialement non euclidiennes sont rendues euclidiennes avant la construction du modèle INDSCAL et lorsque la construction du modèle est réalisée directement sur les données brutes.

Exemple. – Le triangle tordu

Soit 4 points A, B, C et D tels que le triangle ABD est rectangle en B avec AB = 3, AD = 5 et BD = 4 et C est le symétrique de A par rapport à la droite (BD) (*cf* figure 4). Cette configuration est euclidienne. En maintenant fixes toutes les autres

distances, nous faisons varier, dans la matrice de distances, la distance CD de 1 à 7. En procédant ainsi, la matrice de distances obtenue vérifie l'inégalité triangulaire mais n'est plus euclidienne. Cette contrainte qui consiste à vérifier toutes les inégalités triangulaires est liée au fait que, dans le programme original INDSCAL, il existe un pré-traitement (méthode de la constante additive) pour satisfaire cette inégalité. La comparaison entre les procédures est donc plus claire.

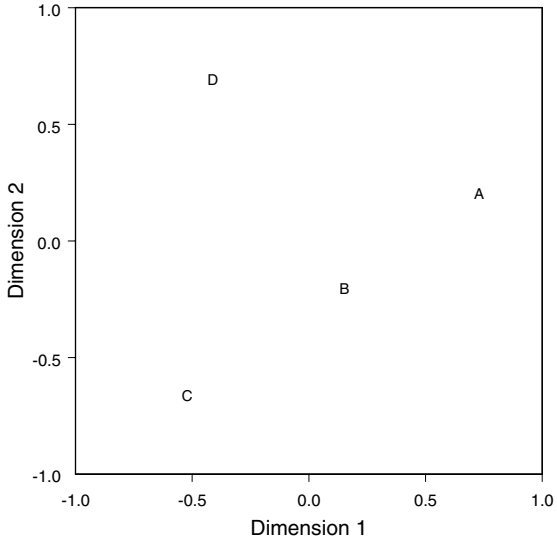

FIGURE 4
Exemple de distances non euclidiennes vérifiant l'inégalité triangulaire. Si la distance CD vaut 5, alors les distances sont euclidiennes (et vérifient l'inégalité triangulaire); si $CD \in [1, 7]$ *(et* $CD \neq 5$*) les distances sont non euclidiennes mais vérifient l'inégalité triangulaire; si* $CD \notin [1, 7]$ *les distances ne vérifient pas l'inégalité triangulaire (et donc ne sont pas euclidiennes)*

Les matrices de données sont non euclidiennes, sauf pour la distance CD = 5. On évalue l'écart à l'euclidienneté par le rapport de la somme des carrés des valeurs propres négatives sur la somme des carrés de l'ensemble des valeurs propres pour chacune des matrices (voir tableau 1).

TABLEAU 1
Rapport de la somme des carrés des valeurs propres négatives sur la somme des carrés de l'ensemble des valeurs propres en fonction de la distance CD

CD	1	2	3	4	5	6	7
Rapport	0.0163	0.0087	0.0023	0.0002	0	0.0001	0.0015

Exemple des données de Schiffman et al. (1981). – On considère les données de 10 colas présentées par Schiffman *et al.* (1981). Les stimuli sont des colas et les 10 sujets sont des étudiants âgés de 18 à 21 ans – 5 hommes et 5 femmes. Les dissimilarités entre colas sont mesurées sur une échelle allant de 0 (colas identiques) à 100 (colas très différents).

Pour ce jeu de données également, les matrices sont non euclidiennes. À titre indicatif, lors de la diagonalisation des 10 matrices de produits scalaires, les 10 racines carrés des rapports de la somme des carrés des valeurs propres négatives sur la somme des carrés de l'ensemble des valeurs propres sont compris entre 0.176 et 0.377.

Exemple des crèmes. – Les distances entre 5 crèmes de compositions différentes ont été évaluées par 4 sujets. Les évaluations sont des notes comprises entre 0 (= crèmes identiques) et 20 (= très diférentes).

Dans cet exemple, les distances sont non euclidiennes : à titre indicatif, lors de la diagonalisation des 4 matrices de produits scalaires, les 4 racines carrés des rapports de la somme des carrés des valeurs propres négatives sur la somme des carrés de l'ensemble des valeurs propres valent : 0.116, 0.138, 0.088, 0.316.

2.2. Comparaison des solutions INDSCAL *avec et sans pré-traitement*

Pour comparer les solutions du modèle INDSCAL, nous comparons d'une part les critères d'évaluation classiques puis les configurations moyennes (i.e. la configuration commune, les juges étant différenciés par les poids qu'ils accordent à chaque dimension).

2.2.1. Du point de vue des critères

Le tableau 2 donne les trois critères classiques : le *Strain*, le $Stress$ et le $S-Stress$ pour chaque jeu de données d'une part avec un pré-traitement qui consiste à rendre les données euclidiennes avant de construire le modèle INDSCAL et d'autre part sans pré-traitement.

TABLEAU 2
Critères pour les solutions du modèle INDSCAL *obtenues à partir des données brutes ou des données avec pré-traitement*

Jeu de données	Méthode	Strain	Stress	S-Stress
Le triangle tordu	Sans pré-traitement	0.03094	0.09634	0.12638
	Avec pré-traitement	0.03134	0.10177	0.12746
Cola	Sans pré-traitement	0.48289	0.56801	0.81142
	Avec pré-traitement	0.48915	0.52470	0.72617
Crème	Sans pré-traitement	0.13488	0.26333	0.33948
	Avec pré-traitement	0.13643	0.25912	0.33271

On peut remarquer que pour les trois exemples, les critères obtenus avec ou sans pré-traitement sont très proches. Le *Strain* est légèrement inférieur lorsque la recherche de la solution INDSCAL est faite en même temps que le passage à l'euclidienneté. Ceci est attendu car le *Strain* est le critère minimisé. En revanche, pour les deux autres critères, il n'y a rien de systématique.

2.2.2. Du point de vue des configurations moyennes

Pour comparer les estimations des paramètres du modèle INDSCAL fournies par les différents algorithmes, nous focalisons l'attention sur les représentations des individus, laissant de côté les poids. En effet, si les configurations des individus fournies par deux algorithmes coïncident axe par axe, alors les poids coïncideront aussi. *A contrario*, si les configurations des individus ne coïncident pas, alors la comparaison des poids des juges n'a pas de sens.

Pour cela, deux configurations étant données, nous calculons :

– les corrélations entre facteurs de même rang;

– les coefficients RV entre représentations de même dimension (Robert et Escoufier, 1976).

Exemple du triangle tordu. – Le tableau 3 montre que la configuration moyenne obtenue avec un pré-traitement (distances préalablement rendues euclidiennes) et la configuration moyenne obtenue sans pré-traitement (le passage à l'euclidienneté étant effectué lors de l'estimation des paramètres) sont très proches. La figure 5 représente la configuration moyenne avec pré-traitement (celle sans-pré-traitement étant trop proche pour être superposée) et la configuration des poids avec pré-traitement et sans pré-traitement. Sur le graphe seuls les poids des juges ayant pris la distance $CD = 1$ et $CD = 2$ sont suffisamment différents d'une stratégie à l'autre pour se différencier. Le poids accordé à l'axe 2 est très faible pour le groupe 1 qui est le groupe ayant évalué la distance CD à 1. Or la deuxième dimension du graphe des stimuli oppose principalement les points C et D. *A contrario*, le groupe 7 qui évalue la distance CD à 7 (distance la plus grande) a une forte (la plus forte) coordonnée sur l'axe 2.

TABLEAU 3

Comparaison entre configurations moyennes. $r(F_1, F_1)$ *est le coefficient de corrélation entre le facteur 1 de la solution* INDSCAL *obtenue à partir des données brutes et le facteur 1 de la solution* INDSCAL *obtenue à partir des données rendues euclidiennes par un pré-traitement,* $r(F_2, F_2)$ *est le coefficient de corrélation entre les facteurs 2 de ces solutions, et RV le coefficient RV entre les configurations des 2 méthodes à deux dimensions*

Exemple	$r(F_1, F_1)$	$r(F_2, F_2)$	RV
Le triangle tordu	0.99998	0.99999	0.999991
Cola	0.9999	–0.9997	0.9995
Crème	0.9996	0.9998	0.9998

On peut noter de plus que que les poids sont ordonnés suivant la distance CD (figure 5).

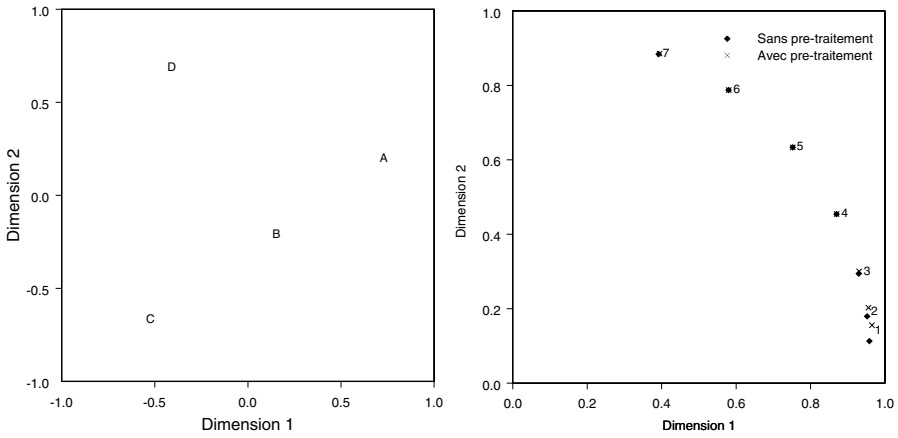

FIGURE 5

Triangle tordu.
Configuration moyenne et configuration des poids lorsque le modèle INDSCAL
est estimé sur les données brutes et les données rendues euclidiennes
après pré-traitement

Exemple des colas. – Les solutions de l'algorithme INDSCAL originel appliqué directement ou après avoir rendu les données euclidiennes, se sont avérées équivalentes. En effet les configurations moyennes sont extrêmement proches et leur interprétation ne dépend pas de la présence ou de l'absence d'un pré-traitement (à une inversion de l'axe 2 près). Les coefficients de corrélation entre les facteurs respectifs de la représentation des colas issue de l'algorithme INDSCAL avec et sans pré-traitement sont de 0.9999 pour le facteur 1 et –0.9997 pour le facteur 2. Une seule configuration moyenne est donnée et les deux configurations des poids sont fournies figure 6.

Exemple des crèmes. – Ici encore, les solutions de l'algorithme originel INDSCAL appliqué directement ou après avoir rendu les données euclidiennes, sont équivalentes ($RV = 0.9998$, *cf* tableau 3). Les coefficients de corrélation entre les facteurs homologues de la représentation des crèmes issue de l'algorithme INDSCAL avec et sans pré-traitement sont proches de 1 (0.9998 et 0.9996). La figure 7 montre que les configurations sont extrêmement proches et leur interprétation ne dépend pas de la présence ou l'absence d'un pré-traitement.

Pour tous les exemples que nous avons traités, les critères et les solutions du modèle INDSCAL obtenus sur les données brutes ou sur les données rendues euclidiennes sont très proches. Ces deux stratégies sont très comparables d'un point de vue opérationnel. Le paragraphe suivant permettra de voir qu'il est donc intéressant d'opter pour la stratégie qui consiste à rendre les données euclidiennes avant la construction du modèle.

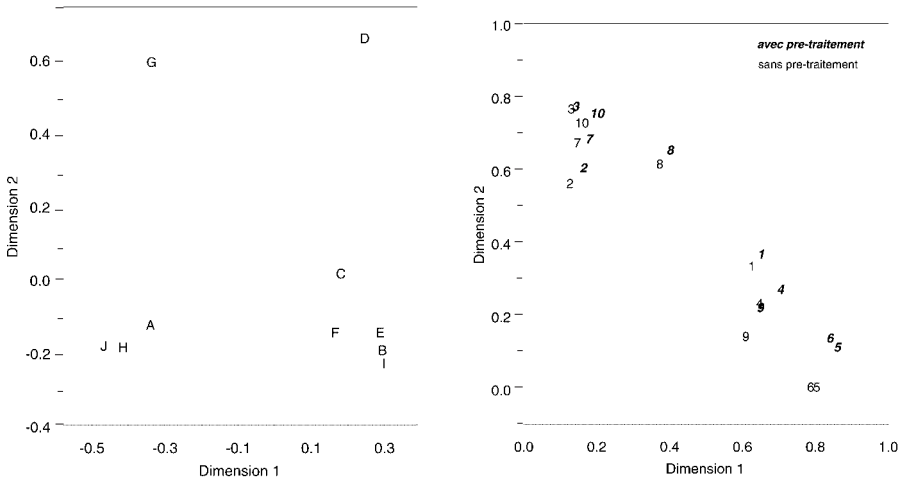

FIGURE 6
Exemple des colas. Comparaison des solutions du modèle INDSCAL *avec ou sans pré-traitement. Représentation moyenne obtenue avec pré-traitement (la représentation moyenne obtenue sans pré-traitement est trop proche pour être représentée sur le même graphe) et représentation des poids. Les colas sont notés de A à J et les juges sont numérotés de 1 à 10*

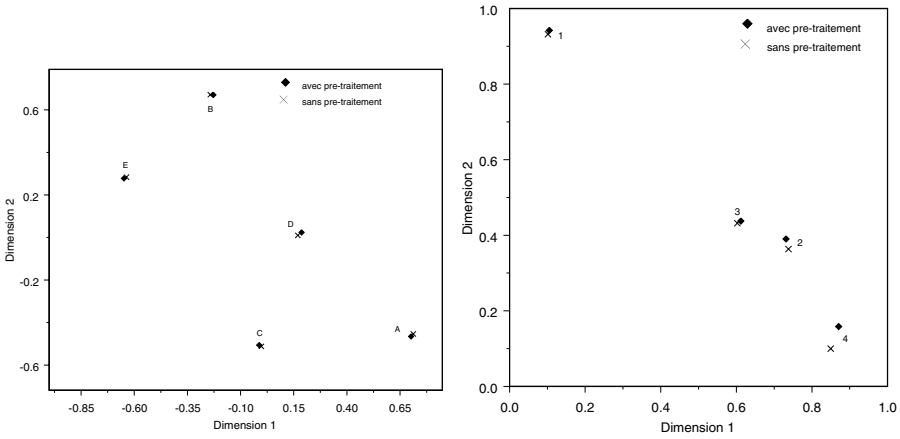

FIGURE 7
Exemple des crèmes. Configuration moyenne et configuration des poids lorsque le modèle INDSCAL *est estimé sur les données brutes et les données rendues euclidiennes après pré-traitement*

2.2.3 Écart entre la solution INDSCAL *et les données brutes*

Lorsqu'on décompose la variabilité des produits scalaires en plusieurs termes, nous sommes en présence de situations différentes suivant les exemples. Le tableau 4 fournit, par juge, les proportions associées à la décomposition de la variabilité de l'équation 10.

TABLEAU 4

Proportions, pour chaque juge puis pour l'ensemble des juges, associées à la décomposition de la variabilité des produits scalaires dans l'équation 10

Exemple triangle tordu Pourcentage de variabilité expliquée par	Gr1	Gr2	Gr3	Gr4	Gr5	Gr6	Gr7	Moy
le passage à l'euclidienneté des données	0.0163	0.0087	0.0023	0.0002	0	0.0001	0.0015	0.00417193
le passage à 2 dimensions	0	0	0	0	0	0	0	0
le modèle INDSCAL	0.9490	0.9581	0.9714	0.9883	0.9974	0.986	0.9607	0.972981
l'erreur d'ajustement du modèle INDSCAL	0.0347	0.0331	0.0263	0.0116	0.0026	0.0139	0.0378	0.02284710

Exemple crème Pourcentage de variabilité expliquée par	Gr1	Gr2	Gr3	Gr4	Moy
le passage à l'euclidienneté des données	0.0134	0.019	0.0077	0.0996	0.03495092
le passage à 2 dimensions	0	8.393e-05	0.0771	0.0016	0.01969038
le modèle INDSCAL	0.9686	0.9096	0.7851	0.8094	0.8681886
l'erreur d'ajustement du modèle INDSCAL	0.0179	0.0713	0.2072	0.091	0.09686047

Exemple cola Pourcentage de variabilité expliquée par	Gr1	Gr2	Gr3	Gr4	Gr5	Gr6
le passage à l'euclidienneté des données	0.0863	0.1422	0.031	0.1362	0.123	0.1168
le passage à 2 dimensions	0.1461	0.2477	0.1695	0.1954	0.1033	0.1097
le modèle INDSCAL	0.5195	0.3356	0.596	0.5014	0.6734	0.6527
l'erreur d'ajustement du modèle INDSCAL	0.3942	0.5222	0.373	0.3624	0.2036	0.2305

Exemple cola Pourcentage de variabilité expliquée par	Gr7	Gr8	Gr9	Gr10	Moy
le passage à l'euclidienneté des données	0.0328	0.0936	0.1035	0.0619	0.092741
le passage à 2 dimensions	0.2480	0.1890	0.2017	0.1883	0.1798685
le modèle INDSCAL	0.4817	0.5354	0.4302	0.5656	0.5291575
l'erreur d'ajustement du modèle INDSCAL	0.4854	0.3709	0.4663	0.3725	0.3781015

Exemple le triangle tordu. – Dans cet exemple, l'erreur est faible et le modèle INDSCAL représente bien les données (97.3 % de la variabilité des produits scalaires est récupéré par le modèle). L'essentiel de l'erreur provient d'une erreur d'ajustement au modèle INDSCAL, le passage à l'euclidienneté représentant (0.0417 / (1-0.973) =) 15% de l'erreur (voir tableau 4). Mais les différents groupes ont des écarts à l'euclidienneté plus ou moins important (cet exemple avait été construit justement pour cela).

Exemple des crèmes. – Pour les crèmes, l'erreur est plus importante puisque le modèle INDSCAL ne récupère que 86.8% de la variabilité des produits scalaires. La non-euclidienneté des données engendre 26.5 % de l'erreur, le reste (73.5 %) correspond à l'erreur d'ajustement au modèle *sensu stricto*.

Exemple des colas. – Dans cet exemple, l'erreur est importante car elle représente 47.1% de la variabilité des produits scalaires. Si on décompose cet erreur, l'écart à l'eulidienneté représente environ 20% de l'erreur.

Il est surtout intéressant dans le tableau 4 de noter les écarts par groupe. Ceci permet de savoir quel groupe (ou sujet) donne une représentation éloignée de la configuration moyenne (erreur d'ajustement du modèle INDSCAL importante) et ceux qui donne des représentations très non euclidiennes. Dans l'exemple des colas, le sujet 4 donne une représentation non euclidienne mais relativement proche de la configuration moyenne fournie par INDSCAL puisque l'écart au modèle est un des plus faible. En revanche, le sujet 7 donne une représentation proche de l'euclidienneté (écart à l'euclidienneté de 0.0328) mais relativement éloignée de la configuration moyenne fournie par INDSCAL. Autrement dit, ce sujet donne des évaluations «cohérentes» (fournit une configuration euclidienne) mais différentes des représentations fournies par les autres sujets.

Lorsque pour un sujet l'erreur d'ajustement au modèle INDSCAL est importante (et non due à la non euclidienneté de la configuration proposée par le sujet), on peut se demander pourquoi ce sujet fournit une configuration différente de la configuration moyenne. Une explication possible est que le sujet a fourni une configuration très multidimensionnelle. Or la solution INDSCAL à laquelle on compare la représentation du sujet est en 2 dimensions, donc une représentation du sujet en plus de 2 dimensions ne pourra pas être expliquée par le modèle INDSCAL par construction (par le choix de la dimension du modèle). Ainsi, résumer la configuration du sujet par une configuration en 2 dimensions comme le propose le modèle INDSCAL induit inévitablement une perte d'information importante. Si on reprend le sujet 7 de l'exemple des colas, ce sujet est relativement éloigné de la configuration moyenne, mais ceci peut s'expliquer par le fait que la configuration de ce sujet est très multidimensionnelle (le pourcentage de variabilité expliqué par l'écart entre sa représentation euclidienne et sa représentation euclidienne en 2 dimensions égal à 0.248, voir tableau 4).

3. Conclusion

La solution du modèle INDSCAL est euclidienne par construction, il est donc préférable de transformer les distances brutes en distances euclidiennes avant d'estimer les paramètres du modèle INDCSAL puisque les solutions du modèle obtenues

sur les données brutes ou sur les données rendues euclidiennes sont proches. Il est très fréquent que les distances initiales soient non euclidiennes lorsque le recueil de données a été effectué à partir de comparaison par paire.

L'intérêt de travailler sur les distances rendues euclidiennes est que l'on peut décomposer l'erreur d'ajustement et quantifier la part de l'erreur qui est due au fait que la configuration proposée par un sujet est non euclidienne et la part due au fait que ce sujet fournit une configuration différente de la moyenne. Travailler sur les distance euclidiennes plutôt que sur les données brutes permet également d'utiliser des méthodes descriptives qui nécessitent des distances euclidiennes.

Références

BORG I., & GROENEN P. (1997), *Modern Multidimensional Scaling : theory and applications*, Berlin : Springer-Verlag.

CARROLL J.D. & CHANG J.J. (1970), Analysis of individual differences in multidimensional scaling via an N-way generalization of «Eckart-Young» decomposition, *Psychometrika*, **35** : 283-319.

D'AUBIGNY G. (1998), Vers un renouveau des méthodes de positionnement multidimensionnel, 4^e *journées MODULAD organisées par le CISIA*.

GOWER J.C. (1966), Some distance properties of latent root and vector methods in multivariate analysis, *Biometrika, 53,* 325–338.

HUSSON F. & PAGÈS J. (2005, SOUS PRESSE), Indscal Model : geometrical interpretation and methodology, *Computational Statistics and Data Analysis*.

KIERS H.A.L. (1989), A Computational Short-Cut for INDSCAL with Orthonormality Constraints on Positive Semi-Definite Matrices of Low Rank, *Computational Statistics Quartely*, **5**, 119-135.

KIERS H.A.L. (1997), A modification of the SINDCLUS algorithm for fitting the ADCLUS and INDCLUS models, *Journal of classification*, **14**, 297-310.

KROONENBERG P.M. (1983), *Three mode principal component analysis : Theory and applications*, Leiden : DSWO press.

ROBERT P. et ESCOUFIER Y. (1976), A Unifying Tool for Linear Multivariate Statistical Methods : the RV-Coefficient, *Applied Statistics*, 29 (3), 257-265.

SCHIFFMAN S.S., REYNOLDS M.L. & YOUNG F.W. (1981), *Introduction to Multidimensional Scaling*, Academic Press.

TAKANE Y., YOUNG F.W. & DE LEEUW J. (1977), Nonmetric individual differences multidimensional scaling : an alternating least square method with optimal scaling features, *Psychometrika*, **42**, 7-67.

TEN BERGE J.M.F. & KIERS H.A.L. (1991), Some clarification of the CANDECOMP algorithm applied to INDSCAL, *Psychometrika*, **56** : 317-326.

TEN BERGE J.M.F., KIERS H.A.L. & KRIJNEN W.P. (1993), Computational Solutions for the problem of Negative Saliences and Nonsymmetry in INDSCAL, *Journal of classification*, **10** : 115-124.

TORGERSON W. S. (1958), Theory and Methods of Scaling, Wiley, New York.

TRENDAFILOV N. (2004), Orthonormality-constrained INDSCAL with Nonnegative Saliences. Full and Off-diagonal Fitting, *Computational Science and Its Applications*, **3044 / 2004** pp 952-960, Springer-Verlag Heidelberg.