Data Analysis

"Data Analysis" in the broadest sense is the general term for a field of activities of ever-increasing importance in a time called the information age. It covers new areas with such trendy labels as e. g., data mining or web mining as well as traditional directions emphasizing, e. g., classification or knowledge organization. Leading researchers in data analysis have contributed to this volume and delivered papers on aspects ranging from scientific modeling to practical application.

They have devoted their latest contributions to a book edited to honor a colleague and friend, Hans-Hermann Bock, who has been active in the field for nearly thirty years.

ISBN 3-540-67731-3

9 783540 677314

http://www.springer.de

STUDIES IN CLASSIFICATION,
DATA ANALYSIS,
AND KNOWLEDGE ORGANIZATION

Data Analysis

W. Gaul
O. Opitz
M. Schader
Editors

Springer

A Geometrical Relational Model for Data Analysis

Yves Schektman[1] and Rafik Abdesselam[2]

[1] 1 bis, rue des Potiers, 31000 Toulouse, France
[2] GEMMA-LERE, Faculté de Sciences Economiques et de Gestion, Université de Caen , 14032 Caen, France

Abstract. The proposed model allows analyses which are more powerful than Factorial Discriminant or Correspondence Analyses; it may be considered as a useful complement to Multivariate Analysis of Variance. Comparing to MANOVA, statistics are not carried out from variables, but from statistical units. The statistical unit space, linked to the variable space by an isometry, contains two orthogonal subspaces associated to mean and residual values as well.

Finally, inertia of configurations of points in the unit space, used in particular to determine factorial axes, can measure either symmetrical or dissymmetrical association coefficients from explanatory variables to independent variables.

1 Introduction

In the geometrical Relational Model proposed, vectors represent statistical units (s.u), for example individuals. We can describe different configurations of s.u. points, located in a maximum of four subspaces, associated respectively to independent variables, explanatory variables, in particular dummy variables associated to the levels of a controlled factor, mean and residual variables. The Relational Model is of some interest because the distance in s.u. space is Relational (Schektman (1978)), i.e. taking into account relationships observed between variables.

Section 2 is concerned with a brief description of Relational Distances. Some proofs of the suitability of this choice are given in section 3. We show, in section 4, that the a priori general Relational Model can be simplified and we give some properties of practical interest.

To shorten this paper, notations are not described as soon as we are able to understand them, without difficulty, by analogy with similar notations defined above. We do the same for some properties and we give references for some proofs which have already been published.

2 Relational distances

$\{x^j\}$ and $\{z^k\}$ being two sets of zero mean variables, observed on the same population of statistical units (s.u.), let us denote:

- E_x [resp. E_z] the subspace of the s.u. space $E = E_x \oplus E_z \oplus \ldots$ identified, by the canonical injection Inx [resp.Inz], to an euclidean vector space E_x [resp. E_z] whose dimension is equal to the number of variables $\{x^j\}$ [resp. $\{z^k\}$],
- $\mathrm{Im}^t X \subset E_x$ [resp. $\mathrm{Im}^t Z \subset E_z$] the image of the mapping transpose of the linear mapping X [resp. Z] defined by the matrix, denote X [resp. Z], whose elements of the jth [resp. kth] column are the values of variable x^j [resp. z^k],
- $x_i \in E_x$ [resp. $z_i \in E_z$] the s.u. vector whose coordinates are the elements of ith row of matrix X [resp. Z],
- $N_x = \{x_i\} \subset E_x$ [resp. $N_z = \{z_i\} \subset E_z$] the configuration of s.u. points associated to the rows of matrix X [resp. Z],
- M_x [resp. M_z] an euclidean distance in space E_x [resp. E_z],
- D the diagonal euclidean distance, whose elements are the weights attached to s.u., in the variable space denoted F,
- $\{\lambda_j(x)\}$ [resp. $\{\lambda_k(z)\}$], $\{c_j(x) \in E_x\}$ [resp. $\{c_k(z) \in E_z\}$] and $\{C^j(x) = [\lambda_j(x)]^{-1/2} X M_x c_j(x) \in F\}$ [resp. $\{C^k(z) \in F\}$] respectively the principal inertia moments and a basis of principal vectors of N_x [resp. N_z], and the corresponding normalized principal components, with respect to the pair of distances (M_x , D) [resp. (M_z , D)],
- P_x^s [resp. P_z^s] the orthogonal projection operator, in space E_x [resp. E_z], onto the sth principal axis spanned by $c_s(x)$ [resp. $c_s(z)$].

Definition

M is a Relational (semi-) Distance in s.u. space E, with respect to the sets of variables $\{x^j\}$ and $\{z^k\}$, if and only if

a) $M[\mathrm{Inx}(c_j(x)), \mathrm{Inz}(c_k(z))] = D[C^j(x), C^k(z)]$ if $\lambda_j(x)\lambda_k(z) \neq 0$ (1)
b) $M[\mathrm{Inx}(u), \mathrm{Inz}(v)] = 0$ if $u \in (\mathrm{Im}^t X)^\perp$ or if $v \in (\mathrm{Im}^t Z)^\perp$. ∎

For convenience we shall use indifferently u or $\mathrm{Inx}(u)$ in the following. It is shown (Schektman (1978)-(1994)) and Croquette (1980)) the following property 1 and lemma 1.

Property 1

Given that ${}^t\mathrm{Inx} M \mathrm{Inx} = M_x$ and ${}^t\mathrm{Inz} M \mathrm{Inz} = M_z$, where (M_x, M_z) is a pair of euclidean distances, M is a Relational (semi-) Distance with respect to the sets of variables $\{x^j\}$ and $\{z^k\}$, if and only if

$${}^t\mathrm{Inx} M \mathrm{Inz} = M_x[(V_x M_x)^{1/2}]^+ V_{xz} M_z [(V_z M_z)^{1/2}]^+ \quad (2)$$

where

- $V_x = {}^t XDX, V_z = {}^t ZDZ, V_{xz} = {}^t XDZ$,
- for t equal x or z, $[(V_t M_t)^{1/2}]^+ = \sum_{\{s/\lambda_s(t) \neq 0\}} [\lambda_s(t)]^{-1/2} P_t^s$ is the Moore-Penrose generalized inverse of $(V_t M_t)^{1/2}$, weighted by M_t. ∎

Note that the use of generalized inverse is necessary when variables $\{z^k\}$ are, in particular, the zero mean dummy variables associated to the levels of a factor because dimension of $\mathrm{Im}^t Z$ equals $\dim E_z - 1$ thus $V_z M_z$ is singular.

Lemma 1

Let $U_x = X M_x[(V_x M_x)^{1/2}]^+$ and $U_z = Z M_z [(V_z M_z)^{1/2}]^+$.

a1) $(\forall s/\lambda_s(x) \neq 0)\ U_x[c_s(x)] = C^s(x)$; $\mathrm{Ker} U_x = (\mathrm{Im}^t X)^\perp$, $\mathrm{Im} U_x = \mathrm{Im} X$.
a2) U_x is a partial isometry from $\mathrm{Im}^t X \subset E_x$ onto $\mathrm{Im} X \subset F$.
b) Same properties for U_z. ∎

In the following property, U is the partial linear mapping defined by

$$(\forall (t = \mathrm{Inx}(r) + \mathrm{Inz}(s) \;/\; r \in E_x, s \in E_z))\ U(t) = U_x(r) + U_z(s) \in F.$$

Property 2

Set the following assertions:

a) $(\forall (t_1, t_2) \in (\mathrm{ImInx}\,{}^t X \oplus \mathrm{ImInz}\,{}^t Z)^2)\ M(t_1, t_2) = D[U(t_1), U(t_2)]$.
b) The image, via U, of each pair of canonical variables carried out from $(\mathrm{ImInx}\,{}^t X, \mathrm{ImInz}\,{}^t Z, M)$ is a pair of canonical variables carried out from $(\mathrm{Im}X, \mathrm{Im}Z, D)$, moreover the corresponding canonical correlation coefficients are equal.
c) The restriction of M to $\mathrm{ImInx}\,{}^t X \oplus \mathrm{ImInz}\,{}^t Z$ is an euclidean distance.

We have : $A]$ ((1) $\Leftrightarrow$ (a) $\Leftrightarrow$ (b)),
$B]$ (a) $\Rightarrow$ ($\mathrm{Im}X \cap \mathrm{Im}Z = \{0\}$ $\Leftrightarrow$ (c)).

Proof

A] (a) $\Rightarrow$ (1) : obvious as lemma 1-a1-b holds.
(1) $\Rightarrow$ (a) : starting from expressions of t_1 and t_2 in a basis of principal vectors, then, by developing $M(t_1, t_2)$, using (1), equalities of norms and angles for corresponding c_j and C^j, and finally lemma 1-a1-b, it comes the development of $D[U(t_1), U(t_2)]$ relatively to principal vectors.
(a) $\Rightarrow$ (b) : obvious as lemma 1-a2-b holds, reasoning by absurd.
(b) $\Rightarrow$ (a) : similar proof as for ((1) $\Rightarrow$ (a)), but starting from expressions of t_1 and t_2 in a basis of canonical variables.
B] Given that D is an euclidean distance, bilinearity, symmetry and positivity of the restriction of M follow immediately from (a); furthermore, given that U is a linear mapping, $\mathrm{Im}X \cap \mathrm{Im}Z = \{0\}$ is equivalent to "U is a partial bijection from $\mathrm{ImInx}\,{}^t X \oplus \mathrm{ImInz}\,{}^t Z$ onto $\mathrm{Im}X \oplus \mathrm{Im}Z$", equivalent to $(\forall t \in (\mathrm{ImInx}\,{}^t X \oplus \mathrm{ImInz}\,{}^t Z,\ t \neq 0) \Rightarrow U(t) \neq 0)$ and finally, according to the properties of M, equivalent to (c), because (a) $\Rightarrow$ $(U(t) \neq 0 \Leftrightarrow M(t,t) \neq 0)$. ∎

It comes that if (1), i.e. property 2-a, holds then U is a partial isometry from $\mathrm{ImInx}^t X \oplus \mathrm{ImInz}^t Z$ onto $\mathrm{Im}X \oplus \mathrm{Im}Z$ if and only if $\mathrm{Im}X \cap \mathrm{Im}Z = \{0\}$.

3 Fundamental results

Let us denote:

- P_z the orthogonal projection operator onto $E_z \subset E(P_z$ is defined because the restriction of M to E_z is an euclidean distance (Schektman and Abdesselam (2000))),
- $\tilde{P}^s_x$ [resp. $\tilde{P}^s_z$] the orthogonal projection operator onto the sth canonical axis, carried out from $(\mathbf{ImInx}^t X, \mathbf{ImInz}^t Z, M)$, belonging to E_x [resp. E_z],
- $N^z_x = \{P_z(x_i) \;/\; x_i \in N_x\} \subset E_z \subset E$,
- $\tilde{N}^s_x = \{\tilde{P}^s_x(x_i) \;/\; x_i \in N_x\} \subset E_x \subset E$,
- $I[N^z_x]$ [resp. $I[\tilde{N}^s_x]$] the inertia of N^z_x [resp. $\tilde{N}^s_x$] according to its centre of gravity (origin of E),
- ρ_s the sth canonical correlation coefficient carried out from $(\mathbf{Im}X, \mathbf{Im}Z, D)$,
- Q_z the orthogonal projection operator onto $\mathbf{Im}Z$.

In property 3-a, we give a statistical and geometrical construction of N^z_x, configuration of points which plays a fundamental role in our approach.

Property 3

If M is relational for variables $\{x^j\}$ and $\{z^k\}$ then

a) $P_z(x_i) = \sum_s \tilde{P}^s_z \tilde{P}^s_x(x_i)$ with $\| \tilde{P}^s_z \tilde{P}^s_x(x_i) \| = \rho_s \| \tilde{P}^s_x(x_i) \|$.

b) $I[N^z_x] = \sum_s \rho^2_s I[\tilde{N}^s_x]$.

c) $I[N^z_x] = \sum_j \lambda_j(x) \| Q_z[C^j(x)] \|^2 = \sum_j \lambda_j(x) \| P_z[c_j(x)] \|^2$

$= \sum_{(j,j')} [M_x]_{jj'} \; D[Q_z(x^j), Q_z(x^{j'})]$

where $[M_x]_{jj'}$ is the (j,j') element of matrix M_x.

Proof

As $N_x \subset \mathbf{Im}^t X$, we have $\mathbf{Inx}(x_i) = \sum_s \tilde{P}^s_x \mathbf{Inx}(x_i)$ then (a) follows from projective property of canonical axes and property 2-A. As canonical axes are orthogonal, obviously $(a) \Rightarrow (b)$. (c) is shown in (Schektman(1994)). ∎

Note 1

Let us give some more fundamental results (Schektman(1987)) which illustrate the significant information contained in N^z_x.

a) If $M_x = ({}^t XDX)^+$, where "+" denotes the Moore-Penrose generalized inverse, then $I[\tilde{N}^s_x] = 1$ and consequently $I[N^z_x] = \sum_s \rho^2_s$ (property 3-b): so $I[N^z_x]$ synthesizes, in term of inertia, the classical symmetrical association indices. Moreover, the principal axes of N^z_x and the corresponding principal components are, according to the types of variables $\{x^j\}$ and $\{z^k\}$, those of Factorial Correspondence Analysis (Benzecri(1982)) or Factorial Discriminant Analysis.

b) If M_x is the unit matrix then $I(N^z_x) = \sum_j \| Q_z(x^j) \|^2$ (property 3-c) and $I(N_x) = \sum_j \| x^j \|^2$: so $I(N^z_x) \,/\, I(N_x)$ synthesizes, in terms of inertia, the classical dissymmetrical association coefficients (Goodman-Kruskall τ, Stewart-love coefficient (1968)). These properties are used, in particular, to define a Factorial Dissymmetrical Correspondence Analysis (Abdesselam and Schektman(1996)).

4 Relational model

4.1 General model

$\{x^j\}$ and $\{y^k\}$ being respectively independent variables and explanatory variables, let:

- $\{g^j = Q_y(x^j)\}$ called mean variables if $\{y^k\}$ are the zero mean dummy variables associated to the levels of a factor, or fitted variables otherwise,
- $\{r^j = x^j - g^j\}$ called residual variables,

where Q_y is the orthogonal projection operator onto $\mathbf{Im}Y \subset F$.

Of course, we define for variables $\{y^k\}$, $\{g^j\}$ and $\{r^j\}$, the same notations $E_y, \mathbf{Im}Y, P_y, Q_y, \ldots$, as defined in sections 2 and 3 but for variables $\{z^k\}$.

We have the following classical results :

- variables $\{g^j\}$ and $\{r^j\}$ are zero means.
- $\mathbf{Im}G \subset \mathbf{Im}Y$, $\mathbf{Im}R \perp \mathbf{Im}Y$, $\mathbf{Im}X \subset \mathbf{Im}G \oplus \mathbf{Im}R$. (3)
- $V_{ry} = V_{rg} = 0$, $V_g = V_{xg} = V_{gx}$, $V_{gy} = V_{xy}$, $V_r = V_{xr} = V_{rx} = V_x - V_g$.

As for variables $\{x^j\}$, a configuration of s.u. points, denoted N_g [resp. N_r], is associated to variables $\{g^j\}$ [resp. $\{r^j\}$].

The proposed Relational Model must satisfy the following hypotheses:

H1) $E = E_x \oplus E_y \oplus E_g \oplus E_r$.

H2) M is a relational (semi-) distance in E for each of the six pairs of sets of variables defined just above.

H3) Distances in spaces E_g and E_r are equal to euclidean distance M_x in E_x : it is indeed reasonable to "see" N_g and N_r in the same way as N_x.

According to Note 1, E_y is the "explanatory" subspace upon which we shall project s.u. $\{x_i\}$. So the nature of euclidean (semi-) distance in E_y is of no importance; however, we shall opt for the Moore-Penrose generalized inverse of V_y, denoted V_y^+, for its use simplifies calculations. Note that we can opt for the chi-square distance if $\{y^k\}$ are associated to a factor.

Property 4

a) M is an euclidean semi-distance; its restriction to $\mathrm{ImIng}^t G \oplus \mathrm{ImInr}^t R$ is a distance.
b) $E_g \perp E_r$.

Proof

a) These results follow from (3) and property 2-B.
b) Using (2), it follows that $V_{rg} = 0 \Rightarrow {}^t\mathrm{Inr} M \mathrm{Ing} = 0$. ∎

4.2 Simplified model

Lemma 2

a) $(\forall g \in E_g) \quad \| g - P_y(g) \| = 0$.
b) $(\forall x \in E_x) \quad \| P_g(x) - P_y(x) \| = 0$.
c) $(\forall x \in E_x) \quad \| x - (P_g + P_r)(x) \| = 0$.

Proof

a) We have
$$\begin{aligned} P_g\mathrm{Iny} &= \mathrm{Ing} M_x^{-1}\, {}^t\mathrm{Ing} M \mathrm{Iny} \\ &= \mathrm{Ing}[(V_g M_x)^{1/2}]^+ V_{gy} V_y^+ [(V_y V_y^+)^{1/2}]^+ \quad \text{using (2)} \\ &= \mathrm{Ing}[(V_g M_x)^{1/2}]^+ V_{gy} V_y^+ V_y V_y^+ \\ &= \mathrm{Ing}[(V_g M_x)^{1/2}]^+ V_{gy} V_y^+ . \end{aligned}$$

Similarly
$$P_y\mathrm{Ing} = \mathrm{Iny} V_{yg} M_x [(V_g M_x)^{1/2}]^+ . \tag{4}$$

It follows
$$P_g P_y \mathrm{Ing} = \mathrm{Ing}[(V_g M_x)^{1/2}]^+ V_{gy} V_y^+ V_{yg} M_x [(V_g M_x)^{1/2}]^+ = \mathrm{Ing}$$
for
$$V_{gy} V_y^+ V_{yg} = {}^tGDYV_y^+\, {}^tYDG = {}^tGDQ_yG = {}^tGDG = V_g . \tag{5}$$

Thus $\| g - P_y(g) \|^2 = M[g,g] - M[g, P_y(g)] = 0$

for $M[g, P_y(g)] = M[P_g(g), P_y(g)] = M[g, P_g P_y(g)] = M[g,g]$.

b) We have
$$\begin{aligned} P_g\mathrm{Inx} &= \mathrm{Ing}[(V_g M_x)^{1/2}]^+ V_{gx} M_x [(V_x M_x)^{1/2}]^+ \\ &= \mathrm{Ing}(V_g M_x)^{1/2} [(V_x M_x)^{1/2}]^+ \quad \text{since } V_{gx} = V_g . \end{aligned} \tag{6}$$

Using (6), (5) and $V_{yg} = V_{yx}$, it follows
$$P_y P_g \mathrm{Inx} = \mathrm{Iny} V_{yx} M_x [(V_x M_x)^{1/2}]^+ = P_y \mathrm{Inx} . \tag{7}$$

thus $\| P_g(x) - P_y(x) \| = \| P_g(x) - P_y P_g(x) \| = 0$ using (a).

c) We have
$$\begin{aligned} P_x\mathrm{Ing} &= \mathrm{Inx}[(V_x M_x)^{1/2}]^+ (V_g M_x)^{1/2} \quad \text{since } V_{xg} = V_g \\ P_x\mathrm{Inr} &= \mathrm{Inx}[(V_x M_x)^{1/2}]^+ (V_r M_x)^{1/2} \quad \text{since } V_{xr} = V_r \\ P_r\mathrm{Inx} &= \mathrm{Inr}(V_r M_x)^{1/2} [(V_x M_x)^{1/2}]^+ \quad \text{since } V_{rx} = V_r . \end{aligned} \tag{8}$$

It follows
$$P_x P_r \mathrm{Inx} = \mathrm{Inx}[(V_x M_x)^{1/2}]^+ V_r M_x [(V_x M_x)^{1/2}]^+$$

$$P_x P_g \mathrm{Inx} = \mathrm{Inx}[(V_x M_x)^{1/2}]^+ V_g M_x [(V_x M_x)^{1/2}]^+ \quad \text{using (6);}$$

then adding $P_x(P_g + P_r)\mathrm{Inx} = \mathrm{Inx}$, since $V_r + V_g = V_x$.

Therefore $\| x - (P_g + P_r)(x) \|^2 = M[x,x] - M[x, (P_g + P_r)(x)] = 0$

for. $M[x, (P_g + P_r)(x)] = M[x, P_x(P_g + P_r)(x)] = M[x,x]$. ∎

According to lemma 2-b, euclidean representations of N_x^g and N_x^y are identical, so the Relational Model can be simplified by taking $E = E_x \oplus E_g \oplus E_r$. Hence variables $\{y^k\}$ only serve to calculate variables $\{g^j\}$, and E_g replaces E_y. Notice that E_g is of a richer nature than E_y since $E_g \supset N_g$. This simplification can be confirmed analytically: for, it follows

- from (6) that the principal components and principal inertia moments associated to principal axes of N_x^g, are characteristic elements of $X\, {}^t[(V_x M_x)^{1/2}]^+\, {}^t(V_g M_x)^{1/2} M_x (V_g M_x)^{1/2} [(V_x M_x)^{1/2}]^+\, {}^tXD$ equal to $X M_x [(V_x M_x)^{1/2}]^+ V_g M_x [(V_x M_x)^{1/2}]^+\, {}^tXD$. (9)
- from (7) that the corresponding operator, but with N_x^y, is $X M_x [(V_x M_x)^{1/2}]^+ \; V_{xy} V_y^+ V_{yx} \; M_x [(V_x M_x)^{1/2}]^+\, {}^tXD$ equal to expression (9) since $V_{yx} = V_{yg}$ and (5).

According to lemma 2-c euclidean representations of N_x and $N_{g+r} = \{P_g(x_i) + P_r(x_i) \,/\, x_i \in N_x\}$ are identical; so, the Model can be once more simplified by taking $E = E_g \oplus E_r$ and replacing N_x by N_{g+r}.

Note 2

Using (6) and (8), the two following partitioned matrices,

$$X\, {}^t([(V_x M_x)^{1/2}]^+)\, {}^t\begin{pmatrix} (V_g M_x)^{1/2} \\ (V_r M_x)^{1/2} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} M_x & 0 \\ 0 & M_x \end{pmatrix}$$

are respectively associated to N_{g+r} and to distance M in $E = E_g \oplus E_r$.

4.3 Some properties

Property 5

Principal axes and principal inertia moments of N_x^g and N_g [resp. N_x^r and N_r] are identical; moreover, the principal components associated to principal axes of N_x^g [resp. N_x^r] belong to $\mathrm{Im}X$.

Proof

It follows from (6) that principal axes and principal inertia moments of N_x^g are characteristic elements of

$$(V_gM_x)^{1/2}\,[(V_xM_x)^{1/2}]^{+\,t}XDX\,M_x[(V_xM_x)^{1/2}]^{+}\,(V_gM_x)^{1/2} = V_gM_x$$

and the property of associated principal components is a consequence of (9). Same proofs for N_x^r and N_r, use in particular (8). ■

Let $U_g = GM_x[(V_gM_x)^{1/2}]^+$, $U_g^+ = [(V_gM_x)^{1/2}]^{+\,t}GD$, $U_r = RM_x[(V_rM_x)^{1/2}]^+$ and $U_r^+ = [(V_rM_x)^{1/2}]^{+\,t}RD$. Obviously, U_r and U_g have the same properties (lemma 1) as U_x; in particular

$$\mathrm{Im}U_r = \mathrm{Im}R \text{ and } \mathrm{Im}U_g = \mathrm{Im}G. \qquad (10)$$

It is easy to show the following lemma 3 (Schektman (1994)).

Lemma 3

a1) U_g^+ is a partial isometry from $\mathrm{Im}G$ onto $\mathrm{Im}\,{}^tG$.
a2) U_g^+ is the Moore-Penrose generalized inverse of U_g, weighted by the pair of distances (M_x, D).
b) Same properties for U_r^+. ■

Lemma 4

$$\mathbf{Ing}U_g^+Q_gU_x = P_g\mathbf{Inx} \text{ and } \mathbf{Inr}U_r^+Q_rU_x = P_r\mathbf{Inx}.$$

Proof

$$\begin{aligned}\mathbf{Ing}U_g^+Q_gU_x &= \mathbf{Ing}U_g^+U_gU_g^+U_x = \mathbf{Ing}U_g^+U_x \quad \text{(lemma 3-a2 and (10))}\\ &= \mathbf{Ing}[(V_gM_x)^{1/2}]^{+\,t}GDXM_x[(V_xM_x)^{1/2}]^+\\ &= \mathbf{Ing}M_x^{-1\,t}\mathbf{Ing}M\mathbf{Inx} = P_g\mathbf{Inx} \quad \text{using (2).}\end{aligned}$$

Similar proof for the second expression. ■

Let U^* be the partial linear mapping defined by
$(\forall(t = u + w\,/\,u \in \mathrm{Im}G, w \in \mathrm{Im}R))\ U^*(t) = \mathbf{Ing}U_g^+(u) + \mathbf{Inr}U_r^+(w) \in E$.

Property 6

$E = E_g \oplus E_r$ can be enriched with the images, via U^*, of $Q_g[C^s(x)], Q_r[C^s(x)]$, $g^j = Q_g(x^j), r^j = Q_r(x^j)$ and $x^j = \sum_s a_s^jC^s(x)$ i.e. respectively with $P_g[c_s(x)]$, $P_r[c_s(x)]$, $\sum_s a_s^jP_g[c_s(x)]$, $\sum_s a_s^jP_r[c_s(x)]$ and $\sum_s a_s^j[P_g + P_r]c_s(x)$.

Proof

It follows from lemma 4 and lemma 1-a that $P_g[c_s(x)] = U^*Q_g[C^s(x)]$ and $P_r[c_s(x)] = U^*Q_r[C^s(x)]$. Moreover, as $\mathrm{Im}G \perp \mathrm{Im}R, E_g \perp E_r$ and according to lemma 3-a1-b thus U^* is a partial isometry from $\mathrm{Im}G \oplus \mathrm{Im}R$ onto $\mathrm{Im}\mathbf{Ing}\,{}^tG \oplus \mathrm{Im}\mathbf{Inr}\,{}^tR$. ■

Note 3

According to Note 1, symmetrical (Benzecri(1982)) or dissymmetrical (Abdesselam and Schektman(1996)) Correspondence Analyses, with simultaneous representation of modalities of variables, are equivalent to Principal Component Analysis (PCA) of $(N_x^g \cup N_g)$, suitable distances being chosen in E_g. So, the Relational Model leads us naturally to enrich the results, provided by these analyses, with those of PCA of $(N_x^r \cup N_r)$.

5 Conclusion

Notes 1 and 3 clearly describe that Relational Model is a formal tool useful (i) to synthesize well known Factorial Analyses, (ii) to enrich provided results with those extracted from residual configurations of s.u. points, and (iii) to extend the area of these analyses to dissymmetrical association coefficients. Concerning this latter new approach, which is often more appropriate to the observed reality, you can find criteria, a tool, an example and references in (Abdesselam and Schektman (1996)) to know, in particular, how to choose a reasonable dissymmetrical association coefficient and for what benefits.

The fundamental utility of the Relational Model is to propose the orthogonal decomposition $x_i = P_g(x_i) + P_r(x_i)$ of each s.u. vector, according to mean and residual subspaces. Thus we hold in the s.u. space E what classically exits for each variable $x^j = Q_g(x^j) + Q_r(x^j) = g^j + r^j$, in the variable space F. Moreover, $E = E_g \oplus E_r$ and F being linked by an isometry, we can enrich the representation of s.u. points on principal planes, with respect either to fitted (or mean) variables or residual variables, with elements of F, as indicated in Property 6.

These results may be useful, in MANOVA, if we really must try, for ourself, to understand, with more details, variations observed on data. In this case, we shall notice that

- design matrix Y can either correspond to dummy variables associated to the levels of a factor or be deduced from a null hypothesis on parameters,
- using Property 3-c, $I[N_x^g] = \sum_j \lambda_j(x) \parallel Q_g[C^j(x)] \parallel^2 = \mathrm{trace}[V_gM_x]$.

Notice that $I[N_x^g]$ is equal to Pillai criteria if $M_x = V_x^+$.

Obviously, the Relational Model can be also of pratical interest in clustering or classification, where explanatory variables $\{y^k\}$ are quantitatives and variables $\{x^j\}$ are dummy variables.

Finally, as explanatory subspace $\mathrm{Im}Y$ in the variable space, proved its utility for independent variables, we hope that a large scale use of corresponding subspace $\mathrm{Im}\,{}^tY$ (or $\mathrm{Im}\,{}^tG$), in the Relational Model, will prove the same, but for statistical units or individuals.

References

ABDESSELAM, R. and SCHEKTMAN, Y. (1996) : Une Analyse Factorielle de l'Association Dissymtrique entre deux Variables Qualitatives. *Revue de Statistique Applique*. Paris, XLIV(2), 5–34.

BENZECRI, J.P. (1982) : L'Analyse des donnes : L'Analyse des Correspondances. Edition Dunod, Paris.

CROQUETTE, A. (1980) : Quelques Rsultats Synthtiques en Analyse des Donnes Multidimensionnelles : Optimalit et Mtriques Effets Relationnels. Thse 3-ime cycle, Universit Toulouse III.

SCHEKTMAN, Y. (1978) : Contribution la Mesure en Facteurs dans les Sciences Exprimentales et la Mise en Oeuvre Automatique des Calculs Statistiques. Thse d'Etat, Universit Toulouse III.

SCHEKTMAN, Y. (1987) : A General Euclidean Approach for Measuring and Describing Associations between Several Sets of Variables. In : C. Hayashi and al. (Eds.) : *Recent Developments in clustering and Data Analysis*. Academic Press, Inc, Tokyo, 37–48.

SCHEKTMAN, Y. (1994) : Proprits des Produits Scalaires Relationnels. Note Interne, DIEM-LEMME, Universit Toulouse III (47 pages).

SCHEKTMAN, Y. and ABDESSELAM, R. (2000) : Semi-Produits Scalaires Relationnels. Note Interne, GEMMA-LERE, Universit de Caen (42 pages).

STEWART, D. and LOVE, W. (1968) : A general canonical correlation index. Psychological Bull, 70, 160–163.