*In memory of the professor Yves Schektman*

# Relational Distances to Analyze the Relationship Between Several Data Tables

Javier Trejos

CIMPA & School of Mathematics, University of Costa Rica.
E-Mail: javier.trejos@ucr.ac.cr

**Abstract**

Relational distances are Euclidean distances whose definition takes into account the associations existing between several sets of variables. A geometric definition of these distances is given and a fundamental property is proposed.

In this work, we present the two applications. In the first application, relational distances are used to define a Qualitative Oblique Principal Component Analysis; this analysis notably allows us to obtain partitions of a population of individuals, hierarchized by the proportion of total inertia explained by each partition. In the second application, these distances allow the design of an algorithm that discovers knowledge in the form of fuzzy assertions (or rules) within a large database.

**Keywords**: Symmetric and asymmetric association indices, qualitative principal components, classifications, clustering, knowledge discovery.

## 1 Introduction

Relational distances were proposed by professor Yves Schektman† (1944–2025) from the Paul Sabatier University, Toulouse, France. He was one of the pioneers of Data Analysis from the French School. He visited the University of Costa Rica for the first Symposium on Mathematical Methods Applied to Science in 1978, returning in 1980. In this way, he helped to

create the research groups in Data Analysis that gave rise to the Center for Research in Pure and Applied Mathematics (CIMPA).

In the Euclidean space of individuals $E = \oplus E_r$, $E_r$ being associated with the subset $G_r$ of variables, the relational distances [15] are characterized geometrically by: $\cos[c_j(r), c_k(s)] = \cos[C^j(r), C^k(s)]$, the $C^j(r) \in F_r$, where $F_r$ is the subspace spanned by the variables of $G_r$, and the $c_j(r) \in E_r$ being respectively the principal components and the principal axial vectors of the cloud $\mathcal{N}_r$, Cartesian projection of the cloud of individuals onto $E_r$. Algebraic expressions and properties of these distances have been given in [**?**, **?**, **?**, **?**].

A relational distance $M$ is defined by the blocks $M_r$ corresponding to each space $E_r$ and data table $X_r$. If $M_r$ and $M_2$ are the diagonal blocks of the relational distance matrix associated with $G_r$ and $G_s$, we have for the extra-diagonal block $M_{rs}$ the following expression:

$$M_{rs} = M_r \left[(V_{rr} M_r)^{1/2}\right]^t V_{rs} M_s \left[(V_{ss} M_s)^{1/2}\right]^t$$

where $V_{rs}$ is the covariance matrix of the group variables $G_r$ and $G_s$, and $t$ represent the weighted generalized inverse with respect to $M_r$ or $M_s$. For the first QPC, $M_r$ and $M_s$ are chosen by the user,and it is advisable to take for $M_s$ the $\chi^2$ distance when $r = 1, s = 2$. For other QPCs, only the corresponding $M_2$ should be chosen.

It is worth to note that the relational distances are such that the canonical correlations of $(E_r, E_s)$ are equal to those of $(F_r, F_s)$.

For suitable choices of relational distances, the inertia of the orthogonal projection of $\mathcal{N}_r$ onto $E_s$, denoted $I[\mathcal{N}_r/E_s]$, is equal to the numerator of the classical symmetric (Bravais-Pearson, $\Phi^2$, sum of squares of canonical correlations, ...) and asymmetric (Goodman-Kruskal, Stewart-Love) association indices between $G_r$ and $G_s$. This new formulation allows us to propose extensions of these indices [1].

Several application of these distances have been proposed. In this communication, we present two of them: i) definition of qualitative oblique principal component analysis, and ii) rule generation from large data sets.

## References


[1] Schektman, Y. (1987) *A general Euclidean approach for measuring and describing associations between several sets of variables.* Proc. of the 1st French-Japanese Sem., Tokyo, 31–42.