UNIVERSITÀ DELLA CALABRIA

Dipartimento di Economia e Statistica
Ponte Pietro Bucci, Cubo 0/C
87036 Arcavacata di Rende (Cosenza)
Italy
http://www.ecostat.unical.it/

Working Paper n. 09 - 2012

WEIGHTING DISTANCE MATRICES USING RANK CORRELATIONS

Ilaria L. Amerise
Dipartimento di Economia e Statistica
Università della Calabria
Ponte Pietro Bucci, Cubo 0/C
Tel.: +39 0984 492474
Fax: +39 0984 492421
e-mail: iamerise@unical.it

Agostino Tarsitano
Dipartimento di Economia e Statistica
Università della Calabria
Ponte Pietro Bucci, Cubo 1/C
Tel.: +39 0984 492465
Fax: +39 0984 492421
e-mail: agotar@unical.it

Dicembre 2012

Weighting Distance Matrices Using Rank Correlations

Ilaria L. Amerise

Dipartimento di Economia e Statistica - Università della Calabria
Via Pietro Bucci, Cubo 0c, 87136 Rende (CS) - Italy

Agostino Tarsitano

Dipartimento di Economia e Statistica - Università della Calabria
Via Pietro Bucci, Cubo 1c, 87136 Rende (CS) - Italy

Abstract

In a number of applications of multivariate analysis, the data matrix is not fully observed. Instead a set of distance matrices on the same entities is available. A reasonable strategy to construct a global distance matrix is to compute a weighted average of the partial distance matrices, provided that an appropriate system of weights can be defined. The Distatis method developed by Abdi *et al.* (2005) is a three-step procedure for computing the global distance matrix. An important aspect of that procedure is the computation of the vector correlation coefficient (RV) to measure the similarity between partial distance matrices. The RV coefficient is based on the Pearson product moment correlation coefficient, which is highly prone to the effects of outliers. We are convinced that, in many measurable phenomena, the relationships between distances are far more likely to be ordinal than interval in nature, and it is therefore preferable to adopt an approach appropriate to ordinal data. The goal of our paper is to revise the system of weights of the Distatis procedure substituting the conventional Pearson coefficient with rank correlations that are less affected by errors of measurement, perturbation or presence of outliers in the data. In the light of our findings on real and simulated data sets, we recommend the use of a specific coefficient of rank correlation to replace, where necessary, the conventional vector correlation.

Key words: Distatis, Ordinal data, Vector rank correlation

Email addresses: iamerise@unical.it (Ilaria L. Amerise), agotar@unical.it (Agostino Tarsitano)

1. Introduction

This paper is concerned with the analysis of a fixed set of entities on the basis of multiple distance matrices where each matrix expresses a particular notion of dissimilarity of one entity to another. A few comments are in order. For one, although many settings rely on feature-based descriptions of entities, there are cases in which only the distances of the entities from each other are known. For instance, Farris (1972) observes that some comparative biochemical methods produce tables whose elements assign a value only to a comparison between a pair of entities. Legendre (2007) states that several forms of data analysis *e.g.*, clustering and ordination, are based upon distance matrices. In the same vein, Lapointe *et al.* (1999) note that laboratory methods such as comparative serology and Dna hybridization produce direct estimates of dissimilarities among entities. Ibba *et al.* (2010) state that dissimilarities or distances may be chosen when feature representations cannot be helpful in discriminating different classes of entities (*e.g.*, for binary variables), in case the experts are not able to define proper features, or if the data lies in high-dimensional spaces. In a typical setting of multidimensional scaling, we are usually only given matrices of distances instead of the original matrices of observations. Tarsitano & Falcone (2010) used distances rather than raw measurements to deal with the simultaneous presence of variables with different measurement scales. The proximity between time series offers clear examples of cases in which a dissimilarity representation might be more acceptable than a feature representation.

Several methods have been introduced for investigating relations between multiple inter-entity proximity matrices. A key issue in this context is to decide how the distance matrices should be combined to produce a global distance matrix synthesizing the information from the various sources. See, among others, de Queiroz *et al.* (1995), Lapointe & Cucumel (1997), Legendre & Anderson (1999), Schneider & Borlund (2007). In the present paper we assume that this problem has been solved in favor of an optimization-based method returning a global matrix that is "closest" to the set of input distance matrices so that they have a balanced

influence. In particular, the Distatis procedure (Abdi *et al.*, 2005) combines classical multidimensional scaling and Statis. Abdi *et al.* (2012) describe Statis as a generalization of principal component analysis whose goal is to analyze several data sets of variables collected on the same set of observations or also, several sets of observations measured on the same set of variables. In fact, the general idea behind Distatis is to transform the distance matrices into cross product matrices and then use the cross product approach, typical of Statis.

In some contexts such as multi-dimensional scaling, the links between distances are more ordinal than interval in nature and it is, therefore, preferable to use measure of association appropriate to ordinal data. This decision may be justified, for example, on the grounds that the original data are outcomes of imprecise measurement, so that only their rank order is reliable (see Podani, 2005). The main motivation of our paper is that the Pearson coefficient, which is at the base of the vector correlation RV proposed by Escoufier (1973), might be substituted, at least in certain situations, by a rank correlation. Vector correlation based on ranks has already been discussed by Cléroux *et al.* (1995) and El Maache & Lepage (2003) in a context in which data are feature based and the original variable are converted into ranks before starting the Distatis procedure. In our approach, we introduce ranks at a later stage realizing a component wise ranking for each column of the principal coordinates involved in the central step of the Distatis.

The remainder of the article is organized into the following sections. In the next, we introduce a global distance matrix expressed as a linear combination of component distance matrices. In the third section, the weights of the combination are obtained so that the resulting matrix is the best representation, in a least square sense, of the whole set of matrices. In the fourth section, we introduce rank transformations of the cross product matrices and discuss arguments showing why the ranks are conceptually and methodologically appealing. The fifth section provides an evaluation of the proposed methodology using numerical example and comparing the results with those obtained from the original Distatis method. Finally, section 6 presents our conclusion and plans for future research.

2. The Distatis method

Let us suppose that K different data sources regarding a set of p distinct entities are condensed into $p \times p$ matrices $\mathbf{\Delta}_r, r = 1, 2, \ldots, K$ whose generic element $\delta_{i,j,r}$ reflects the dissimilarity or distance of entity i and j with respect to the r-th source, for $r, s = 1, 2, \ldots, p$. Note that this happens no matter what, and how many, attributes have been used in each source of data, or how the distance matrices have been constructed. We make a fairly strong assumption that $\mathbf{D}_r, r = 1, 2, \ldots, K$ are Euclidean to facilitate the principal coordinate analysis underlying Distatis (see Gower & Legendre, 1986).

In case the scale of the distances is not uniform, it is a common practice to normalize them so that their impact is not unduly influenced by numerical resolution of the measurements. We use a specific scaling factor for each matrix, which equalizes the distances up to the maximum observed range across all the sources of data.

$$\mathbf{D}_r = \left[\frac{\max_{1 \leq r \leq K} \left\{ \max_{1 \leq i,j \leq p} \left\{ \delta_{i,j,r} \right\} \right\}}{\max_{1 \leq i,j \leq p} \left\{ \delta_{i,j,r} \right\}} \right] \mathbf{\Delta}_r \quad \text{for } r = 1, 2, \cdots, K. \tag{1}$$

The new matrices are computed in such a way that the maximum distances are equal over all the matrices, but the maximum depends on the largest distance found in the data set. This type of invariance can be useful for discarding some possible heterogeneity of measurement units and precluding the consequent implicit weighting of the distances.

The global distance matrix is:

$$\mathbf{D} = \sum_{r=1}^{K} w_r \mathbf{D}_r \quad \text{with} \quad w_r \geq 0, \quad \text{and} \quad \sum_{r=1}^{K} w_r = 1 \tag{2}$$

We recall that if the component matrices are Euclidean, then $\mathbf{D}$ is also Euclidean (see Pavoine *et al.*, 2009). The cell entries of the global distance matrix are weighted averages of the distances in the component matrices. The weight w_r

expresses the trade-off between a marginal shift in one non-diagonal elements of the r-th distance matrix $\mathbf{D}_r$ and a marginal shift in one off-diagonal element of the global distance matrix $\mathbf{D}$, when all of the other distances are held constant. The positive sign of the weights and the linearity of (2) ensure that every variation in $d_{i,j}$ corresponds to an increase or decrease in one of the distances $d_{i,j,r}, r = 1, 2, \ldots, K$.

Formula (2) leaves us with the question of choosing proper weights. A number of different reasonable ways exists, *e.g.* the weights can be smaller when a matrix is considered less important or it is considered more likely to be in error. Equal weighting is evoked when the impact of the component distance matrices is unknown or not unambiguously derived since, in this case, $w_r = 1/K, r = 1, 2, \ldots, K$ is the least informative *a priori* choice for the weights. On the other hand, equal weights are the real best choice in case the matrix of vector correlations between component matrices is an equicorrelation matrix. It is worth noting that both the identity matrix and the matrix of ones (all elements are equal to 1) are special cases of equicorrelation matrices. Weights can also be computed in a way that equalizes a common characteristic of the distance matrices, such as mean, standard deviation or maximum. In absence of accepted information on what is important and what should be prioritized with regard to the partial distance matrices, we might apply a purely data-driven approach with all its inherent drawbacks.

To obtain an objective system of weights, we need a measure for how much a component matrix affects the global distance matrix. This can be expressed in terms of the amount of variation in $\mathbf{D}_r, r = 1, 2, \ldots, K$ by converting the original space into a new space in which as much variation as possible is expressed along a new axis. In the framework outlined by Abdi *et al.* (2005), we follow a three-step procedure. The initial step is the transformation

$$\mathbf{B}_r = -0.5\mathbf{C}\mathbf{D}_r^2\mathbf{C}^t, r = 1, 2, \ldots, K; \qquad \mathbf{C} = \mathbf{I}_p - p^{-1}\mathbf{u}_p\mathbf{u}_p^t \tag{3}$$

where $\mathbf{D}_r^2$ is the matrix whose (i, j)-th element is the square of the (i, j)-th element of $\mathbf{D}_r$, $\mathbf{u}_p$ is a $p \times 1$ vector of 1s and $\mathbf{I}_p$ is the identity matrix of order p. Furthermore,

$\mathbf{C}$ is a $p \times p$ symmetric and idempotent matrix that centers each column of $\mathbf{D}_r$ on the origin, that is $\mathbf{B}_r \mathbf{u}_p = \mathbf{0},\ r = 1, 2, \cdots, K$. Expression (3) converts the matrix of distances $\mathbf{D}_r$ into the matrix of cross products $\mathbf{B}_r$, which contains the same information, in terms of total sums-of squares, as $\mathbf{D}_r$ but is more suitable to the eigen-decomposition.

The central step of Distatis analyzes how the matrices $\mathbf{B}_r, r = 1, 2, \ldots, K$ resemble each other. The method used is the vector correlation proposed by Escoufier (1973),

$$RV(r,s) = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} r^2(\mathbf{b}_{i,r}, \mathbf{b}_{j,s})}{\sqrt{\sum_{i=1}^{p}\sum_{j=1}^{p} r^2(\mathbf{b}_{i,r}, \mathbf{b}_{j,r}) \sum_{i=1}^{p}\sum_{j=1}^{p} r^2(\mathbf{b}_{i,s}, \mathbf{b}_{j,s})}} \tag{4}$$

where $\mathbf{b}_{i,r}, \mathbf{b}_{j,s}$ are, respectively, the r-th and the s-th row of $\mathbf{B}_s$ and $\mathbf{B}_r$ respectively and $r^2(\mathbf{b}_{i,r}, \mathbf{b}_{j,s})$ is the square of Pearson correlation coefficient. Expression (4) implies that $0 \leq RV(r,s) \leq 1$. The fact that the columns of the $\mathbf{B}_r$ have zero mean implies that

$$r^2(\mathbf{b}_{i,r}, \mathbf{b}_{j,s}) = \begin{cases} \dfrac{\left(\mathbf{b}_{i,r}^t \mathbf{b}_{j,s}\right)^2}{\left(\mathbf{b}_{i,r}^t \mathbf{b}_{j,r}\right)\left(\mathbf{b}_{i,s}^t \mathbf{b}_{j,s}\right)} & \text{if} \quad \left(\mathbf{b}_{i,r}^t \mathbf{b}_{j,r}\right)\left(\mathbf{b}_{i,s}^t \mathbf{b}_{j,s}\right) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

The values of (4) are invariant with respect to the unit of measurement used in the various component matrices, thus no normalization of the cross product matrices $\mathbf{B}_r$ is necessary.

Let $\mathbf{\Psi}$ be the $K \times K$ matrix formed with the vector correlation between pairs of cross products matrices: $\mathbf{\Psi}_{r,s} = RV(r,s)$ for $r, s = 1, 2, \cdots, K$. Since $\mathbf{\Psi}$ is positive or, at least, non negative irreducible, the Perron-Frobenius theorem (*e.g.* Lin, 1977) ensures that there is a single eigenvalue that is positive and greater than or equal to all other eigenvalues; furthermore, the unique (up to multiplication by a non-zero constant) associated eigenvector $\mathbf{q}$ may be chosen so that all its

components are strictly positive. In the final step of the procedure we calculate the normalized eigenvector $\mathbf{w} = \left(\mathbf{u}_K^t \mathbf{q}\right)^{-1} \mathbf{q}$ whose elements $w_r,\ r = 1, 2, \cdots, K$ represent a uniquely defined differential impact measure for each source of data. The global distance matrix can now be found using

$$\mathbf{D} = \sqrt{\mathbf{b}\mathbf{u}_p^t + \mathbf{u}_p \mathbf{b}^t - 2\mathbf{B}}, \qquad \mathbf{B} = \sum_{r=1}^{K} w_r \mathbf{B}_r \tag{6}$$

where $\mathbf{b}$ is a vector of the diagonal elements of $\mathbf{B}$.

3. The Rank Distatis method

Methods based on ranks are desirable if the relationships between original distances do not follow a mathematically predictable pattern or are thought to be non-linear. Ranks, in fact, do not change when distances are monotonically transformed. For example, $1 + \log(d_{rs})$, $d_{rs}^{\alpha}, 0 < \alpha \leq 1$ or $d_{rs}/(1 + d_{rs})$ not only preserve the metric properties of d_{rs} (see Batagelj & Bren, 1995), but also generate the same ranking of distances. The invariance under monotone transformations is an attractive property when comparing distances whose actual magnitudes may be arbitrary or when there is evidence that the data set may be contaminated with outliers.

Vector correlation based on ranks are not new in the literature of Distatis. As an example, Cléroux *et al.* (1995) converted into ranks the original variables before starting the procedure. In the present investigation, we realize a component wise ranking for the columns of the cross product matrices $\mathbf{B}_r, r = 1, 2, \cdots, K$

$$\mathbf{G}_r = \begin{bmatrix} g_{1,1,r} & g_{1,2,r} & \cdots & g_{1,p,r} \\ g_{2,1,r} & g_{2,2,r} & \cdots & g_{2,p,r} \\ \cdots & \cdots & \cdots & \cdots \\ g_{p,1,r} & g_{p,2,r} & \cdots & g_{p,p,r} \end{bmatrix}, \qquad r = 1, 2, \ldots, K \tag{7}$$

where the $(p \times 1)$ vectors $\mathbf{b}_{i,r}$ are replaced by the $(p \times 1)$ vectors of ranks $\mathbf{g}_{i,r}$

obtained transforming each value of the original column in the corresponding rank ranging from 1 to p, in ascending sequence.

The similarity between matrices can be measured using a non-parametric statistic of rank relation in place of the Pearson correlation

$$RV_h(r,s) = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} r_h^2\left(\mathbf{g}_{i,r},\mathbf{g}_{j,s}\right)}{\sqrt{\sum_{i=1}^{p}\sum_{j=1}^{p} r_h^2\left(\mathbf{g}_{i,r},\mathbf{g}_{j,r}\right)\sum_{i=1}^{p}\sum_{j=1}^{p} r_h^2\left(\mathbf{g}_{i,s},\mathbf{g}_{j,s}\right)}}. \tag{8}$$

This definition is given by (Cléroux *et al.*, 1995) and derive from the application of the RV coefficient to the matrices formed with the rank correlations. By far, the most frequent recommendation is to use Spearman rank-order correlation, the argument being that this coefficient would be more valid than Pearson coefficient when certain assumptions are violated. Cléroux *et al.* (1995) applied, in fact, the Spearman coefficient to the association between feature-based matrices. However, several other coefficients have been developed for more than a century. Our review of literature has highlighted the four rank correlations reported in Table 1, which have been studied extensively in Tarsitano & Lombardo (2011).

Table 1: A selection of rank correlations.

Name	Formula
Spearman	$r_1 = \left(\frac{3}{p^3-p}\right)\left(\sum_{i=1}^{p}\lvert\eta_i^*-\pi_i\rvert^2-\sum_{i=1}^{p}\lvert\eta_i-\pi_i\rvert^2\right)$
Kendall	$r_2 = \left(\frac{2}{p^2-p}\right)\sum_{i<j} sgn\left(\eta_j-\eta_i\right)\left(\pi_j-\pi_i\right)$
Gini	$r_3 = \left(\frac{2}{n^2-k_p}\right)\left(\sum_{i=1}^{p}\lvert\eta_i^*-\pi_i\rvert-\sum_{i=1}^{p}\lvert\eta_i-\pi_i\rvert\right), k_p = p \bmod 2$
	$r_4 = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p}\left[g_i\left(\eta,\pi^*\right)g_j\left(\eta^*,\pi\right)-g_i\left(\eta^*,\pi^*\right)g_j\left(\eta,\pi\right)\right]}{\left[k_p+2\sum_{i=1}^{\lfloor p/2\rfloor}\frac{p+1-i}{i}\right]^2-p^2}$

Here π and η are permutations of order p; $\pi^* = p+1-\pi$ and $\eta^* = p+1-\eta$ are the reverse permutations formed by the anti-ranks of π and η, respectively; $\lfloor x \rfloor$ indicates the lowest integer greater than x. Finally

$$g_i(\boldsymbol{\eta}, \boldsymbol{\pi}) = \begin{cases} \pi_i/\eta_i & \text{if} \quad \eta_i < \pi_i \\ \eta_i/\pi_i & \text{if} \quad \eta_i \geq \pi_i \end{cases} \qquad i = 1, 2, \cdots, p \tag{9}$$

The indices in Table 1 vary in the range: $-1 \leq r_i(\eta, \pi) \leq 1$ with $r_i(\eta, \eta) = 1$, $r_i(\pi, \pi) = 1$, $r_i(\eta, \eta^*) = -1$ and $r_i(\pi, \pi^*) = -1$. The larger $r_i(\pi, \pi)$, ignoring sign, the stronger the association between the two rankings. At the other extreme, an $r_i(\pi, \pi)$ of zero implies that the two rankings are not linearly related. Other properties of $r_i(\eta, \eta) = 1$ are the symmetry: $r_i(\eta, \pi) = r_i(\pi, \eta)$ and the zero expected value under independence.

Let $\boldsymbol{\Psi}_h, h = 1, 2, 3, 4$ be the $K \times K$ matrix formed with the rank vector correlations in (8). These matrices have the same properties as the $\boldsymbol{\Psi}$ of the previous section. The first principal component $\mathbf{q}_h$ derived from $\boldsymbol{\Psi}_h$ is determined as a new variable which orders the rows of the matrices $\mathbf{G}_r, r = 1, 2, \cdots, K$ in such a way that the weighted sum of squares of rank-correlation coefficients between the original variables and the new one is maximal. For $h = 1, 2, 3, 4$, corresponding to the four rank correlations in Table 1, the normalized eigenvector $\mathbf{w}_h = \left(\mathbf{u}_K^t \mathbf{q}_h\right)^{-1} \mathbf{q}_h$ belonging to the largest eigenvalue of $\boldsymbol{\Psi}_h$ can be used to assign weights as in formula (6).

4. Experimental results

The experiments presented here look for evidence that incorporation of ranks into the core of the Distatis procedure can lead to an effective mechanism of weighting distance matrices.

4.1. Real data sets

Spielman (1973) gives $K = 4$ matrices about geographical, genetic, anthropometric and SFA (Serological For Anthropometric) distances among $p = 19$ villages of the Yanomami Indians (southern Native Americans). All the matrices were found not

to be Euclidean. To overcome this problem Cailliez (1983) suggested transforming the distance as follows $[d_{i,j,r} - 2\lambda_p(\mathbf{B}_r)]^{0.5}$ where $\lambda_p(\mathbf{B}_r) < 0$ is the smallest eigenvalue of $\mathbf{B}_r$. Of course the results of the successive analyses are sensitive, though perhaps not very sensitive, on this transformation.

Table 2 shows the weights of the four partial distance matrices obtained using the measures of association presented in the previous sections (all the analyses were done using the statistical software R, version 2.15.1).

Table 2: Weights for Yanomami villages distance matrices.

Coefficient	Geographic	Genetic	Anthropometric	SFA
RV	0.267	0.260	0.274	0.200
RV_1	0.239	0.268	0.234	0.259
RV_2	0.235	0.271	0.235	0.259
RV_3	0.239	0.270	0.233	0.257
RV_4	0.233	0.263	0.239	0.265

The usual RV coefficient yields a two-block configuration of the weights. On the one hand there are tree larger weighs, almost equal among themselves. On the other hand stands a weight, much smaller than the average of the other weights, associate with the SFA distances. The weights generated by $RV_1, \ldots, RV_4$ have a configuration reflecting the equal influence of the various distances. There were no appreciable differences among any of the weighting systems, thus making it appear that outliers or other anomalies were not a major cause of errors. Spielman (1973) excluded SFA from the successive analyses, but this exclusion can be considered legitimate from the point of view of the association between matrices only if the association is measured by RV.

The second example is in Abdi *et al.* (2005). This data set includes $K = 4$ matrices of order 6×6 containing measures of proximity between four algorithms compared using a distance matrix between six face. The data are available as $faces2005$ data set in *ExPosition* package of R. The matrices are not Euclidean and hence we have applied the transformation discussed in the previous example. In Table 3 we report the weights obtained by Distatis and rank Distatis procedures.

Table 3: Six faces analyzed by different "algorithms".

Coefficient	Pixels	Measures	Ratings	Pairwise
RV	0.272	0.243	0.237	0.249
RV_1	0.270	0.255	0.249	0.227
RV_2	0.282	0.249	0.256	0.214
RV_3	0.277	0.253	0.264	0.207
RV_4	0.267	0.250	0.254	0.229

All the weighting systems show small divergences between the observed and the expected weights under the hypothesis of equicorrelation: $(1/4, \ldots, 1/4)$, which reveals that, even though the original distance matrices were proposed for entirely different purposes, they share a common underlying structure.

4.2. Artificial data sets

In the first study on simulated distances, we use the data set $mat5Mrand$ of the package *ape* in R, which includes $K = 5$ genetic distance matrices of order 50×50 computed from simulated DNA (see Campbell *et al.*, 2009 for details). The matrices are not Euclidean, but become such when one applies the square root transformation $d_{i,j,r}^{0.5}$.

In order to assess the potentiality of rank transformation to attenuate the impact of outliers, we have perturbed one of the distance matrices (otherwise supposed to be free of noisy data or idiosyncratic entities) by multiplying a randomly chosen row (and the corresponding column) by a factor of 10 and keeping the rest of the selected matrix unchanged. The matrix to be perturbed is also chosen at random. In table 4 we present the weights given by the Distatis procedures and their standard deviation, before and after the alteration. The figures are the average values over 10000 random selection of the pair $(\text{matrix}, \text{entity})$. The column headed $\sigma(\mathbf{w})$ refers to the standard deviation of the weights as a measure of their uncertainties. The final column shows the percentage of variation in $\sigma(\mathbf{w})$.

Table 4: Genetic distance matrices.

		$\mathbf{D}_1$	$\mathbf{D}_2$	$\mathbf{D}_3$	$\mathbf{D}_4$	$\mathbf{D}_5$	$\sigma(\mathbf{w})$	$PV\sigma$
RV	Original	0.1894	0.2121	0.2192	0.1844	0.1949	0.0134	
	Altered	0.1978	0.2096	0.2169	0.1873	0.1884	0.0325	143.0
RV_1	Original	0.1669	0.2422	0.2181	0.1761	0.1968	0.0275	
	Altered	0.1771	0.2359	0.2091	0.1808	0.1971	0.0336	22.2
RV_2	Original	0.1680	0.2430	0.2169	0.1769	0.1953	0.0273	
	Altered	0.1775	0.2365	0.2079	0.1819	0.1962	0.0369	35.2
RV_3	Original	0.1720	0.2409	0.2219	0.1757	0.1895	0.0270	
	Altered	0.1805	0.2352	0.2120	0.1805	0.1918	0.0356	31.8
RV_4	Original	0.1680	0.2416	0.2105	0.1778	0.2020	0.0259	
	Altered	0.1784	0.2341	0.2062	0.1822	0.1992	0.0257	-0.9

Apparently, RV sterilizes the effect of a single outlier, at least on the average, but the standard deviations of the weights are trying to tell a different story. In fact, the value of $\sigma(\mathbf{w})$ for RV, after the perturbation, has increased indicating that occasionally the importance of a data source might be seriously affected by the presence of even a single outlying entity. In absence of perturbations, rank correlations determine weights in which the importance of $\mathbf{D}_1$ and $\mathbf{D}_4$ is diminished, and meanwhile, the role of the other matrices has equally grown by the consequent increase in their weights. This result is quite unexpected in the light of the fact that the matrices of the data set $mat5Mrand$ were built to be no more similar to each other then randomly selected matrices would be (which would lead to equal weighting). In all probability, ranks alter the interdependencies contained within the different matrices. When RV incorporates ranks, the weights of $\mathbf{D}_2, \mathbf{D}_3$ and $\mathbf{D}_5$ are reduced and the charge is transferred mostly to $\mathbf{D}_1$. Nonetheless, the overall structure of the weights replicated quite well before and after the perturbation, which can be interpreted as a symptom of robustness against outliers. It is perhaps useful to observe that weights determined by RV_4 have the lowest deviations among rank correlations for the original distances and that they achieve the minimum $\sigma(\mathbf{w})$ among all coefficients for the altered distances.

The second experiment on artificial data addresses the behavior of the vector correlations in randomly created data sets, as the number of entities p increases. To determine random distance matrices we first generate random correlation matrices $\mathbf{H}_r, r = 1, 2, \cdots, K$ for $K = 7$. Then these are converted into distance matrices using the transformation $d_{i,j,r} = [0.5\,(1 - h_{i,j,r})]^{0.5}$. There are a number of techniques used to produce a random correlation matrix (see Numpacharoen & Atsawarungruangkit, 2012 for a recent review). The algorithm proposed by Rousseeuw & Molenberghs (1993) in terms of goniometric functions of angles uniformly distributed in $[0, \pi]$ has been particularly effective in our experiments. Due to rounding errors, there are sometimes very small negative eigenvalues. In this case the vector of random angles is rejected and a new one is tried. However, if the number of consecutive rejections of the positive definiteness of $\mathbf{H}_r$ exceeded five, all the negative eigenvalues are set to zero and the correlation matrix obtained with the eigenvalue method (see Rousseeuw & Molenberghs, 1993). The distance matrices so generated have low correlation and, as a consequence, the information that they convey should compete for weight on a more-or-less equal basis.

Table 5: Vector correlation for random distance matrices.

p		$\sigma(\mathbf{w})$	Mean	Min	Max	p		$\sigma(\mathbf{w})$	Mean	Min	Max
25	RV	0.0318	0.0803	0.0182	0.2276	50	RV	0.0326	0.0409	0.0101	0.1151
	RV_1	0.0301	0.0845	0.0214	0.2318		RV_1	0.0316	0.0423	0.0114	0.1179
	RV_2	0.0304	0.0629	0.0162	0.1736		RV_2	0.0317	0.0313	0.0085	0.0873
	RV_3	0.0316	0.0749	0.0183	0.2117		RV_3	0.0326	0.0383	0.0103	0.1089
	RV_4	0.0216	0.1139	0.0454	0.2582		RV_4	0.0190	0.0678	0.0326	0.1429
100	RV	0.0302	0.0207	0.0053	0.0582	200	RV	0.0313	0.0104	0.0026	0.0293
	RV_1	0.0297	0.0212	0.0056	0.0588		RV_1	0.0305	0.0106	0.0028	0.0294
	RV_2	0.0297	0.0156	0.0042	0.0434		RV_2	0.0306	0.0078	0.0021	0.0216
	RV_3	0.0304	0.0194	0.0052	0.0546		RV_3	0.0315	0.0097	0.0026	0.0277
	RV_4	0.0147	0.0424	0.0239	0.0799		RV_4	0.0114	0.0279	0.0183	0.0471

The first column of each block in Table 5 shows the standard deviations of the weights. The last three columns contain mean, minimum and maximum of the off-diagonal entries in the matrix $\mathbf{\Psi}$ of correlations between component matrices. To stabilize the results, each reported value is the average of the corresponding observed values in 250 independent replications of the simulation process.

Two facts emerge clearly as the number of entities gets large: (1) The weights become closer and closer to the expected weights $\mathbf{w} = (1/K, \cdots, 1/K)$, as confirmed by the diminishing $\sigma(\mathbf{w})$ for all the weighting systems. (2) The entries concerning the effects of the number of entities do not confirm an RV increasing with p in case of random matrices (Smilde *et al.*, 2009, Mayer *et al.*, 2011). On the contrary, all the coefficients show a common trend towards zero. In passing, we note that RV_4 presents the smallest standard deviations of the weights and the slowest decay of the correlations.

The objective of last simulations is to look at the recovery rates of clustering algorithms based on the global distance matrix determined by the five variants of Distatis considered in this paper. Here, we simulate $K = 6$ random correlation matrices of order $p \times p$. The first five matrices are generated according the same scheme of the previous experiment. The sixth matrix is built in blocks by applying the algorithm described in Hardin *et al.* (2012). More specifically, a Toeplitz scheme is used as a model for five correlation sub matrices of order $p/5 \times p/5$ assuming that pairs of adjacent entities are correlated and that the correlation between entity i-th and j-th decays exponentially with $|i - j|$. Correlation matrices are converted into distance matrices using the transformation of the previous experiment.

For the number of entities $p \in (25, 50, 100)$, Table 6 reports in the first column the adjusted Rand index (ARI) between two classifications. The former is the true clustering of the entities, whereas the other one is computed using a hierarchical aggregative algorithm after cutting the tree at five clusters. We apply the seven links included in the command *hclust* of the package *stats* in R but used only the best result for the given simulation. The second column shows the ARI obtained

by the PAM algorithm of the package *cluster* in *R* for the five-cluster solution. For this iterative scheme we have applied three different methods to establish the starting medoids and retained only the best solution. Finally, note that each entry in the table is an average across 250 experiments of the same type.

Table 6: Recovery rate of average link and PAM for Distatis procedures

	p=25		p=50		p=100	
	Hier.	PAM	Hier.	PAM	Hier.	PAM
RV	0.5662	0.5801	0.5011	0.5196	0.3881	0.3599
RV_1	0.5405	0.5541	0.5700	0.5864	0.4978	0.4485
RV_2	0.5256	0.5470	0.5781	0.5881	0.5129	0.4520
RV_3	0.5680	0.5710	0.5962	0.6041	0.5151	0.4517
RV_4	0.4784	0.5052	0.4665	0.4868	0.4078	0.3774

The recovery rate of the clustering algorithms decreases, as expected, with increasing p because of the disturbances due to the growing randomness of the distance matrices. According to the results reported in Table 6, the clusters of the well-structured matrix $\mathbf{D}_6$ are obscured by the other five noisy matrices even when $p = 25$. The performance of the PAM algorithm exceeded that of hierarchical links, but the discrepancy becomes narrower as the number of entities increases, independently of the weighting system. For $p = 100$ the results are reversed. The precise reasons for this are not clear, but one conceivable explanation is that it is due to a greater robustness of hierarchical methods against noisy distances. When the weights are computed with RV the performance of the two clustering methods are undoubtedly worse than using the rank correlations.

5. Conclusion and future research

We have explored the question of substituting a correlation for the conventional Pearson correlation in the coefficient RV of the Distatis procedure (Abdi *et al.*, 2005). In particular, we have studied four different non-parametric statistics of rank relation: RV_1 (Spearman), RV_2 (Kendall), RV_3 (Gini cograduation index) and RV_4 (a measure based on ratios of ranks and anti-ranks) and determined

which ordinal measure of association is the most plausible proxy for the Pearson correlation and under what circumstances.

The essence of our findings is that, in the presence of contaminated data, rank correlations attenuate adverse effects of anomalies and, in case of clean and faultless data, yield weighting systems which generally conform to those obtained from RV. Comparisons among the four rank-based variants of the Distatis method reveals that the results produced by RV_4 most closely resemble those obtained by RV and for this reason it can be a valid substitute for it in cases where ordinal relationships between distances appear more relevant for judging clustering effectiveness than the numerical magnitude representation of the distances.

A limitation of the present study concern the derivation of the large samples distribution of vector correlations under the null hypothesis of no association, without being in possession of the original data matrices. The problem would be at least partially solved if the distributional assumptions, usually imposed when dealing with the variance/covariance matrices of the characters of the entities, were met by the cross products matrices reconstructed from the inter-entity distances. See Fraser (1956), Zegers & ten Berge (1985), Cléroux *et al.* (1995). However, more research (for example, along the lines of Oliveira & Mexia, 2007 and Abdi *et al.*, 2009) is needed to determine the sampling and asymptotic behavior of the vectorial statistics, which are discussed in the present paper.

References

Abdi, H. and O'Toole, A. J. and Valentin, D. and Edelman, B. "DISTATIS: the analysis of multiple distance matrices" *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Diego, CA, (USA), 42–47 (2005).

Abdi, H. and Dunlop, J.P. and Williams, L.J. "How to compute reliability estimates and display confidence and tolerance intervals for pattern classifiers using the Bootstrap and 3-way multidimensional scaling (DISTATIS)" *NeuroImage* 45, 89–95 (2009).

Abdi, H., and Lynne, J.W., and Valentin, D., and Bennani-Dosse, M. "STATIS and DISTATIS: optimum multitable principal component analysis and three way metric multidimensional scaling" *Wiley Interdisciplinary Reviews: Computational Statistics* 4, 124-167 (2012).

Batagelj, V. and Bren, M. "Comparing resemblance measures" *Journal of Classification* 12, 73–90 (1995).

Cailliez, F. "The analytical solution of the additive constant problem" *Psychometrica*, 48, 305-308 (1983).

Campbell, V. and Legendre, P. and Lapointe, F.-J. "Assessing congruence among ultrametric distance matrices" *Journal of Classification* 26, 103–117 (2009).

Cleroux, R. and Lazraq, A. and Lepage, Y. "Vector correlation based on rank and a nonparametric test of no association between vectors" *Communications in Statistics. Theory Methods* 24, 713–733 (1995).

de Queiroz, A. and Donoghue, M.J. and Kim, J. "Separate versus combined analysis of phylogenetic evidence" *Annual Review of Ecology and Systematics* 26, 657–681 (1995).

El Maache, H. and Lepage, Y. "Spearman's rho and Kendall1's tau for multivariate data sets" *IMS-Lecture Notes–Monograph Series* 42, 113–130 (2003).

Escoufier, Y. "Le traitement des variables vectorielles" *Biometrics* 29, 751-760 (1973).

Farris, J.S. "Estimating phylogenetic trees from distance matrices", *The American Naturalist* 106, 645-668 (1972).

Fraser, D.A.S. "A Vector form of the Wald–Wolfowitz–Hoeffding theorem" *The Annals of Mathematical Statistics* 27, 540–543 (1956).

Gower, J. C., and Legendre, P., "Metric and Euclidean properties of dissimilarity coefficients" *Journal of Classification* 3, 5–48 (1986).

Harding , J. and Garcia, R. and Golan, D. "A method for generating realistic correlation matrices". Available at http://arxiv.org/pdf/1106.5834v2.pdf

Ibba, A. and Duin, R. P. W. and Lee, W.-J., "A study on combining sets of differently measured dissimilarities" *Proceedings of the 2010 20th International*

Conference on Pattern Recognition, 3360–3363 (2010).

Lapointe, F.-J. and Cucumel G."The average consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa" *Systematic Biology* 46, 306–312 (1997).

Lapointe, F.-J. and Kirsch, A.W. and Hutcheon, J.M. "Total evidence, consensus, and bat phylogeny: a distance-based approach" *Molecular Phylogenetics and Evolution* 11, 55–66 (1999).

Legendre, P. "Comparison of permutation methods for the partial correlation and partial mantel tests" *Journal of Statistical Computation and Simulation* 67, 37-73 (2007).

Legendre, P. and Anderson, M.J. "Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments" *Ecological Monographs* 69, 1–24 (1999).

Lin, K.Y. " An elementary proof of the Perron-Frobenius theorem for non-negative symmetric matrices" *Chinese Journal of Phisics* 15, 283–285 (1977).

Mayer, C-D., and Lorent, J. and Horgan, G.W. "Exploratory analysis of multiple omics datasets using the adjusted RV coefficient" *Statistical Applications in Genetics and Molecular Biology* (2011).

10, Numpacharoen, K. and Atsawarungruangkit, A. "Generating correlation matrices based on the boundaries of their coefficients" *PloS one* 7, (2012). Available at http://europepmc.org/abstract/MED/23152816

Oliveira, M.M. and Mexia, J.T. "Modelling series of studies with a common structure" *Computational Statistics & Data Analysis* 51, 5876–5885 (2007).

Pavoine, S. and Vallet, J. and Dufour, A-B and Gachet, S. and Herve, D. "On the challenge of treating various types of variables: application for improving the measurement of functional diversity" *Oikos* 18, 391–402 (2009).

Podani, J. "Multivariate exploratory analysis of ordinal data in ecology: pitfalls, problems and solutions" *Journal of Vegetation Science* 16, 497–510, (2005).

Rousseeuw, P. and Molenberghs, G. "Transformation of nonpositive semidefinite correlation matrices" *Communications in Statistics. Theory and Methods* 22,

965–984 (1993).

Schneider, J.W. and Borlund, P. "Matrix comparison, part 1: motivation and important issues for measuring the resemblance between proximity measures or ordination results", *Journal of the American Society for Information Science and Technology* 58, 1586–1595 (2007).

Smilde, A. K. and Kiers, H. A. L. and Bijlsma, S. and Rubingh, C. M. and van Erk, M. J. "Matrix correlations for high-dimensional data: the modified rv-coefficient" *Bioinformatics* 25, 401–405 (2009).

Spielman, R.S. "Differences among Yanomama Indian villages: do the patterns of allele frequencies, anthropometrics and map locations correspond?" *American Journal of Physical Anthropology* 39, 461–480 (1973).

Tarsitano, A. and Falcone, M. "Missing values adjustment for mixed-type data. *Journal of Probability and Statistics* (2011).

Tarsitano, A., and Lombardo, R. "A Coefficient of correlation based on ratios of ranks and anti-ranks", *The Journal of Economics and Statistics* . Forthcoming 2012.

Zegers, F. E. and ten Berge, J.M.F: "A family of association coefficients for metric scales" *Psychometrika* 50, 17–24 (1985).